DMAGaze: Gaze Estimation Using Feature Disentanglement and Multi-Scale Attention⋆

Haohan Chen[a], Hongjia Liu[b], Shiyong Lan[a,*], Wenwu Wang[c], Yixin Qiao[a], Yao Li[a], Guonan Deng[a]

[a]Sichuan University, 24 South Section 1, 1st Ring Road, Chengdu, 610065, Sichuan Province, China
[b]Aalto University, Maarintie 8, Espoo, Helsinki, 02150, Uusimaa, Finland
[c]University of Surrey, Stag Hill, Guildford, GU2 7XH, Surrey, UK

Abstract

Gaze estimation, which predicts gaze direction, commonly faces the challenge of interference from complex gaze-irrelevant information in face images—a key bottleneck limiting its accuracy in real-world scenarios. In this work, we propose DMAGaze, a novel gaze estimation framework that exploits information from facial images in three aspects: gaze-relevant global features (disentangled from facial image), local eye features (extracted from cropped eye patch), and head pose related features, to improve overall performance. Firstly, we design a new continuous mask-based Disentangler to separate gaze-relevant and gaze-irrelevant information in facial images through reconstructing the eye and non-eye regions using a dual-branch architecture. Furthermore, we introduce a new attention module, called Multi-Scale Global Local Attention Module (MS-GLAM), to fuse the global and local information at multiple scales via a customized attention structure, thereby further enhancing the information from the Disentangler. Finally, we combine the global gaze-relevant features, with head pose and local eye features, and pass them through the detection head for high-precision gaze estimation. Our proposed DMAGaze has been evaluated extensively on two widely used public datasets: obtaining a gaze estimation error of 3.74° on MPIIFaceGaze and 6.17° on RT-GENE, outperforming SOTA methods.

Keywords: Gaze estimation, Feature disentanglement, Gaussian similarity, Multi-scale attention

1. Introduction

Gaze estimation, the task of predicting gaze direction, crucial for measuring human attention, is widely applied in areas like saliency detection [1, 2], virtual reality [3], driver distraction monitoring [4], human-computer interaction [5], and autism diagnosis [6]. Recently, gaze estimation has shifted from model-based methods to appearance-based methods. Model-based methods [7] aim to estimate gaze direction by capturing the physical characteristics of the human eye and pupil with specialized sensors and 3D reconstruction techniques. In contrast, appearance-based methods [8, 9, 10, 11, 12] aim to achieve gaze estimation utilizing features extracted from facial images with techniques such as deep learning.

In appearance-based methods, the input eye and facial images are used to estimate gaze independently [9, 10, 11, 12], in fusion [13], or through mutual assistance [14, 15], as shown in the gray pathway on the left part of Fig. 1, a typical framework taken by conventional methods. With only eye images as input, gaze estimation can be overly influenced by the states of eyes. Using only face images as input, gaze estimation can be degraded by various information, such as facial appearances, expressions, and poses.

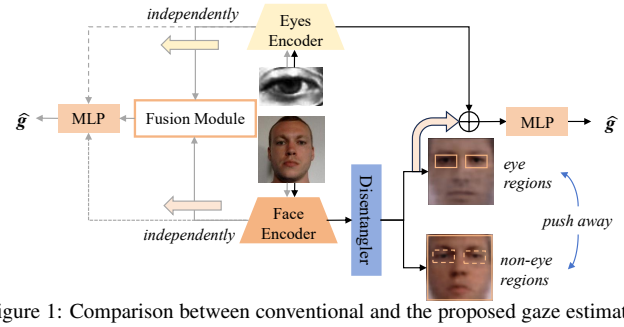


Figure 1: Comparison between conventional and the proposed gaze estimation methods. The gray pathway represents the typical framework of traditional methods, while the black pathway represents our proposed gaze estimation method.

To overcome the aforementioned limitations, numerous approaches have been proposed. These include applying convolutions with spatial weights [10], utilizing dilated convolutions [14], and adopting coarse-to-fine adaptive networks [15] to extract informative features from facial and eye regions. Other methods leverage cross-attention mechanisms to fuse features from both sources [13], integrate transformers into the gaze estimation framework [11], or introduce facial text prompts [16, 17] inspired by the contrastive language–image pre-training (CLIP) model [18] to guide the estimation process. Furthermore, some studies jointly learn static and dynamic gaze cues from images and videos under weak supervision using gaze-following labels [19]. Despite these advancements, deep learning–based gaze estimation continues to face challenges in unconstrained environments and suffers from limited robustness due to entangled feature representations [20]. Consequently, given the highly intertwined nature of various

⋆The code is available at https://github.com/Sajelhhh/DMAGaze.

*Corresponding author: lanshiyong@scu.edu.cn.
Email addresses: sajel@stu.scu.edu.cn (Haohan Chen), hongjia.liu@aalto.fi (Hongjia Liu), lanshiyong@scu.edu.cn (Shiyong Lan), w.wang@surrey.ac.uk (Wenwu Wang), Qiaoyixin@stu.scu.edu.cn (Yixin Qiao), liyao518@stu.scu.edu.cn (Yao Li), 2023223045102@stu.scu.edu.cn (Guonan Deng)

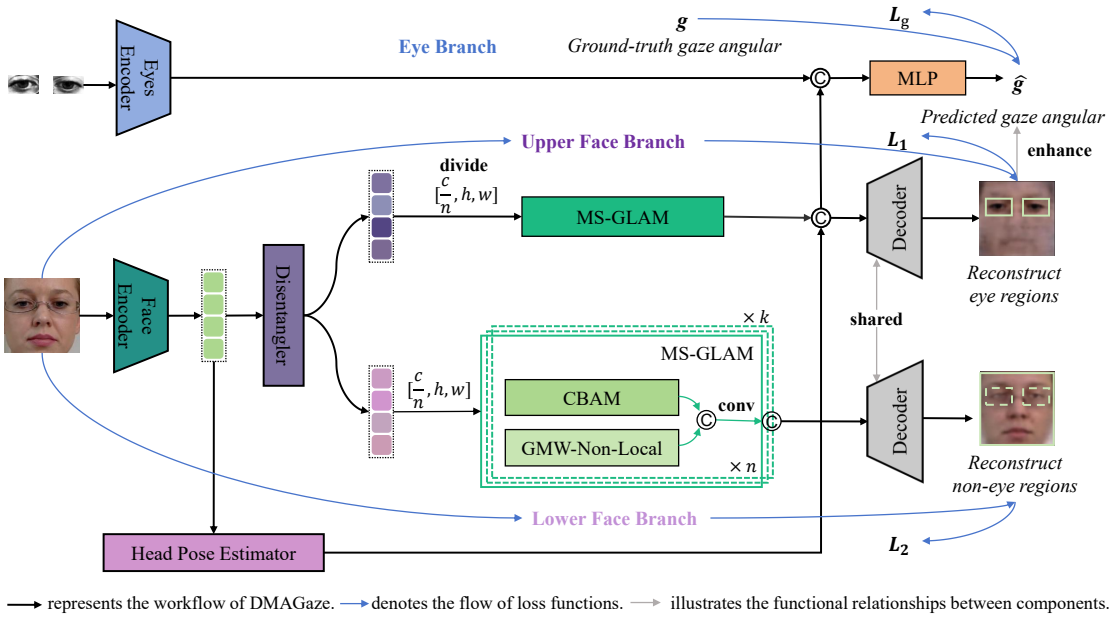


Figure 2: The overall architecture of our proposed DMAGaze, along with its integration of framework and loss functions.

facial attributes, such as expressions, head poses, and appearances, effectively disentangling gaze-relevant information from irrelevant components in facial representations remains a major challenge for achieving robust gaze estimation.

To address this challenge, we introduce DMAGaze, illustrated in the black pathway of Fig. 1 and Fig. 2. DMAGaze builds on the observation that the non-stationary gazing process [21] is predominantly driven by eye movements and head pose, and that facial information useful for eye reconstruction inherently encodes gaze-relevant cues. At the core of DMAGaze is a Disentangler module, inspired by feature disentanglement techniques [22]. This module takes facial features as input and separates them into two sub-branches, each emphasizing either the eye region or the non-eye regions of the face. Unlike the binary masks used in [22], our Disentangler employs continuous masks, enabling a more flexible representation of subtle variations between features. This design facilitates effective separation of gaze-relevant and irrelevant information. The distilled global gaze-relevant facial features further complement the limited local cues extracted solely from eye images, thereby enhancing the robustness of the final gaze estimation.

Although integrating facial and eye features is essential, feature representations still require refinement via attention mechanisms. The Convolutional Block Attention Module (CBAM) [23, 24] enhances representations through channel and spatial weighting, while the Non-Local Network [25, 26] captures global dependencies crucial for modeling cross-regional gaze interactions. However, used independently, they fail to disentangle gaze-relevant and irrelevant information, limiting global–local gaze modeling. While [27] introduces a multi-bit attention framework for 12-bit RAW data, it is hardware-specific. Thus, a more general multi-scale approach to jointly capture global and local dependencies remains an open challenge.

We propose a Multi-Scale Global–Local Attention Module (MS-GLAM) that links CBAM and Non-Local through a cascaded attention structure, effectively integrating global and local information across multiple scales. Traditional Non-Local operations rely on dot-product similarity, which is essentially linear and may fail to capture complex nonlinear dependencies among gaze-relevant features. To address this, we introduce Gaussian Modulated Weighting (GMW), which incorporates a Gaussian similarity-based distance metric into the Non-Local operation (Fig. 3), enabling stable and expressive modeling of nonlinear global feature dependencies.

Our main contributions are summarized as follows:

- We present the DMAGaze framework, which employs a Disentangler to separate global gaze-relevant and gaze-irrelevant facial features. By combining the extracted gaze-relevant facial cues with estimated head pose and local eye information, the framework achieves improved gaze estimation.
- We propose MS-GLAM, featuring a customized cascaded attention structure to effectively capture and integrate global and local information. Additionally, we enhance the Non-Local module with GMW, which leverages Gaussian similarity-based distance metrics to improve nonlinear modeling of feature dependencies.
- Finally, we conduct extensive comparisons on mainstream datasets, demonstrating the superiority of our gaze estimation method.

2. Related work

With the development of deep learning, appearance-based methods have become the mainstream. They typically take facial or eye images as input, extract features through neural networks, and then perform gaze estimation. For example, Zhang et al. [8, 10] were the first to introduce convolutional neural network (CNN) into the gaze estimation task and later used the attention mechanism to weight facial features. Fischer et al. [28] used two VGG networks to process two eye images.

Cheng et al. [15] first estimated gaze from facial images and then refined it using eye images. Krafka et al. [29] proposed a multi-channel network that takes eye images, full-face images, and facial mesh information as inputs. Cheng et al. [26] used generative adversarial networks to purify gaze-related features in facial images. Abdelrahman et al. [30] used pitch and yaw dual branches for cross-entropy and regression iterations on facial image features. Cătrună et al. [13] fused facial and eye features through cross-attention for gaze estimation.

Recent studies have made progress in addressing gaze estimation within specific scenarios. Pathirana et al. [31] introduced a dual-path model tailored for retail single-user gaze estimation; however, it depends on manually annotated object channels and does not effectively disentangle gaze-relevant features. Senarath et al. [32] developed a depth-based dual attention (DDA) model to improve depth perception in retail environments, yet its multi-scale fusion mechanism overlooks local–global feature interactions, leading to errors under minor head movements. Murthy et al. [33] proposed the lightweight MAGE-Net for distraction detection in automotive settings, but its cross-scenario generalization remains limited. Although their PARKS-Gaze dataset [34] provides extensive coverage of extreme head poses, models trained on it still suffer from noticeable cross-domain errors due to inadequate feature adaptation.

In previous studies, fusion methods for multi-input gaze estimation [13, 14, 15] have primarily relied on simple concatenation or basic attention mechanisms, which struggle to suppress noise and irrelevant information unrelated to gaze. Moreover, two key research directions—cross-person adaptation and model explainability—remain insufficiently addressed in current work. To address these challenges, our proposed DMAGaze framework explicitly disentangles gaze-relevant and gaze-irrelevant features through complementary reconstruction. Within this framework, the MS-GLAM module performs effective feature fusion across multiple scales, integrating both global and local information. In addition, the nonlinear similarity measurement introduced in GMW further refines the selection of gaze-relevant features, ultimately leading to more competitive and robust gaze estimation performance.

3. Methodology

A high-level overview of our proposed DMAGaze structure is shown in Fig. 2. Eye images are processed through an eye encoder in the eye branch to extract eye features, while facial images are passed through a face encoder in the face branch. Gaze-relevant features are then distilled through the Disentangler and MS-GLAM by reconstructing the eye and non-eye regions of the full-face images using the decoders. The distilled gaze-relevant features are subsequently combined with head pose features extracted by simple convolutions and local eye features to predict gaze.

We input left and right eye images $\mathbf{I}^l$ and $\mathbf{I}^r$ of size $h^e \times w^e \times 3$ into the eyes encoder Enc^e, and the facial image $\mathbf{I}$ of size $h^f \times w^f \times 3$ into the face encoder Enc^f, where 3 denotes the RGB color channels. The eyes encoder extracts the eye features $\mathbf{F}^e \in \mathbb{R}^{n\times d}$, where n is the number of features and d is the dimension of the features, while the face encoder extracts the initial full-face features $\mathbf{F}^f \in \mathbb{R}^{c\times h\times w}$, where c is the number of channels, h is the height and w is the width. This process can be formalized as:

$$\mathbf{F}^e = \mathrm{Enc}^e(\mathbf{I}^l, \mathbf{I}^r) \tag{1}$$

$$\mathbf{F}^f = \mathrm{Enc}^f(\mathbf{I}) \tag{2}$$

3.1. Disentangler

The initial facial features $\mathbf{F}^f$ extracted by the facial branch are disentangled by the Disentangler into two complementary components—gaze-relevant features $\mathbf{F}^r$ and gaze-irrelevant features $\mathbf{F}^{ir}$, where $\mathbf{F}^r$ provides global information from the full face to complement the local eye representations. This process is defined as follows:

$$\mathbf{F}^r = \mathbf{K} \odot \mathbf{F}^f \tag{3}$$

$$\mathbf{F}^{ir} = (\mathbf{1} - \mathbf{K}) \odot \mathbf{F}^f \tag{4}$$

where $\odot$ denotes element-wise multiplication, $\mathbf{K} \in [0,1]^{c\times h\times w}$ and $\mathbf{1}-\mathbf{K} \in [0,1]^{c\times h\times w}$ serve as learnable multi-variable weight matrices with $\mathbf{1}$ being a tensor of all ones. The mask $\mathbf{K}$ acts as a region-aware feature gate that selectively controls the information flow from $\mathbf{F}^f$ and is learned jointly with the network parameters through backpropagation under the joint supervision of reconstruction and gaze estimation objectives.

3.2. Multi-scale global-local attention module

The disentangled facial features $\mathbf{F}^r$ and $\mathbf{F}^{ir}$ are passed to the MS-GLAM module. To enhance the modeling for the long-range dependencies and further disentangle gaze-relevant and irrelevant cues, we introduce GMW into Non-Local [25], using a Gaussian similarity-based distance metric to replace the dot product calculation, indirectly capture non-linear relationships in a high-dimensional space, enhancing the ability to capture global feature dependencies, as shown in Fig. 3. The Gaussian similarity used in GMW is defined as follows:

$$G(\mathbf{q}, \mathbf{k}) = \exp(-\frac{\|\mathbf{q}-\mathbf{k}\|^2}{2\sigma^2}) \tag{5}$$

where $\mathbf{q}$ and $\mathbf{k}$ are the input feature vectors, and σ is the variance parameter.

Convolutional block attention module. CBAM combines channel and spatial attention mechanisms for feature selection. Assuming the input feature is $\mathbf{F} \in \mathbb{R}^{c\times h\times w}$, the entire process can be expressed as follows:

$$\mathbf{F}' = \mathrm{M}_c(\mathbf{F}) \odot \mathbf{F} \tag{6}$$

$$\mathbf{F}'' = \mathrm{M}_s(\mathbf{F}') \odot \mathbf{F}' \tag{7}$$

where M_c and M_s are the channel attention and spatial attention operations, respectively. $\mathbf{F}'$ and $\mathbf{F}''$ represent the channel-refined feature and CBAM feature, respectively. The channel attention module includes average pooling and max pooling layers along the spatial dimension:

$$\mathbf{F}^c_{avg} = \mathrm{avgpooling}(\mathbf{F}) \tag{8}$$

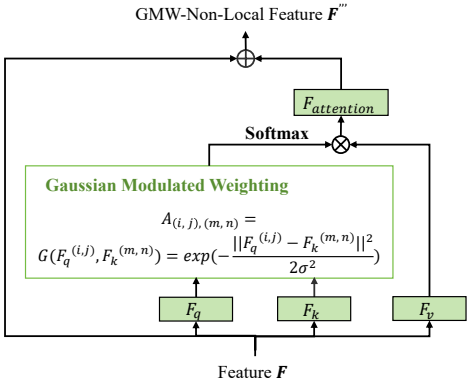


Figure 3: Structure of the proposed Multi-Scale Global–Local Attention Module (MS-GLAM).

$$\mathbf{F}^{c}_{max} = \text{maxpooling}(\mathbf{F}) \tag{9}$$

$$\mathbf{F}^{'} = \text{sigmoid}(\text{MLP}(\mathbf{F}^{c}_{avg}) + \text{MLP}(\mathbf{F}^{c}_{max})) \tag{10}$$

where sigmoid$(\cdot)$ is a sigmoid activation function, MLP$(\cdot)$ is a multi-layer perceptron. Then, the average and maximum values are taken along the channel dimension. The initial spatial attention map output from the convolution layer is used to produce a spatial attention map as follows:

$$\mathbf{F}^{s}_{avg} = \frac{1}{C}\sum_{c=1}^{C}\mathbf{F}^{'}[:, C, :, :] \tag{11}$$

$$\mathbf{F}^{s}_{max} = \max_{c=1...C}\mathbf{F}^{'}[:, C, :, :] \tag{12}$$

$$\mathbf{F}^{''} = \text{sigmoid}(\text{conv}(\text{cat}(\mathbf{F}^{s}_{avg}, \mathbf{F}^{s}_{max}))) \tag{13}$$

where conv$(\cdot)$ is the convolution operation, cat$(\cdot)$ is the feature concatenation operation.

Gaussian modulated weighting-non-local. Non-Local [25, 26] utilizes non-local operators to capture long-distance feature dependencies. However, the similarity measure used in this method is based on dot-product computation, which tends to assign high similarity to large but gaze-irrelevant activations and thus struggles to model complex nonlinear dependencies among eye appearance, head pose, and facial cues. To address this, we introduce the GMW, as shown in Fig. 3, where Gaussian modulated weighting is used with Non-Local operation to perform nonlinear similarity measure based on distances, leading to GMW-Non-Local, where large Euclidean differences trigger rapid exponential decay. We take the input feature $\mathbf{F} \in \mathbb{R}^{c\times h\times w}$ and apply convolutional layers to transform it:

$$\mathbf{F}_q = \text{conv}_q(\mathbf{F}) \tag{14}$$

$$\mathbf{F}_k = \text{conv}_k(\mathbf{F}) \tag{15}$$

$$\mathbf{F}_v = \text{conv}_v(\mathbf{F}) \tag{16}$$

In this case, conv$_q(\cdot)$, conv$_k(\cdot)$ and conv$_v(\cdot)$ represent the convolutional transformations. The attention weight matrix $\mathbf{A} \in \mathbb{R}^{(h_q\times w_q)\times(h_k\times w_k)}$ between the query $\mathbf{F}_q$ and the key $\mathbf{F}_k$ is defined as:

$$\begin{aligned}\mathbf{A}_{(i,j),(m,n)} &= G(\mathbf{F}_q^{(i,j)}, \mathbf{F}_k^{(m,n)}) \\ &= \exp(-\frac{\|\mathbf{F}_q^{(i,j)} - \mathbf{F}_k^{(m,n)}\|^2}{2\sigma^2})\end{aligned} \tag{17}$$

where h_q, w_q and h_k, w_k are the spatial dimensions of $\mathbf{F}_q$ and $\mathbf{F}_k$, respectively, i, j and m, n denote the spatial position indices of $\mathbf{F}_q$ and $\mathbf{F}_k$, respectively, and $\mathbf{F}_q^{(i,j)}$ and $\mathbf{F}_k^{(m,n)}$ are the input feature vectors of GMW. This Gaussian similarity projects features into a high-dimensional kernel space, enabling the model to capture nonlinear interactions beyond dot-product similarity. By emphasizing distance-based correlations, GMW suppresses gaze-irrelevant activations and highlights subtle, gaze-consistent dependencies. The final output $\mathbf{F}''' \in \mathbb{R}^{c\times h\times w}$ is given by:

$$\mathbf{F}''' = \text{softmax}(\mathbf{A}) \times \mathbf{F}_v + \mathbf{F} \tag{18}$$

where softmax$(\cdot)$ denotes normalization.

Cascaded attention structure. We design a cascaded structure for the above two attention modules, which not only integrates local and global information effectively but also captures the multi-scale representations. We take the input feature $\mathbf{X} \in \mathbb{R}^{c\times h\times w}$ and divide it into n groups along the channel dimension. $\mathbf{X}_i \in \mathbb{R}^{\frac{c}{n}\times h\times w}$ represents the grouped feature and $i \in \{0, 1, ..., n-1\}$ represents the i-th group. After applying convolutional transformation to each group of features, the CBAM and GMW-Non-Local operations are applied. This process is repeated for each group until all groups have been processed. The entire process can be formalized as:

$$\mathbf{x}^{i}_{sub} = \text{conv}_{\text{sub}}(\mathbf{X}_i) \tag{19}$$

$$\mathbf{a}_i = \begin{cases}\text{CBAM}(\mathbf{x}^{i}_{sub}), & i = 0 \\ \text{CBAM}\left(\text{cat}(\mathbf{z}_{i-1}, \mathbf{x}^{i}_{sub})\right), & 1 \le i \le n-1\end{cases} \tag{20}$$

$$\mathbf{b}_i = \begin{cases}\text{GMW}(\mathbf{x}^{i}_{sub}), & i = 0 \\ \text{GMW}\left(\text{cat}(\mathbf{z}_{i-1}, \mathbf{x}^{i}_{sub})\right), & 1 \le i \le n-1\end{cases} \tag{21}$$

$$\mathbf{z}_i = \text{conv}_{\text{tail}}(\text{cat}(\mathbf{a}_i, \mathbf{b}_i)) \tag{22}$$

$$\mathbf{Z} = \text{cat}(\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_{n-1}) \tag{23}$$

where conv$_{\text{sub}}(\cdot)$ represents the convolutional operation for subgroups, CBAM$(\cdot)$ and GMW$(\cdot)$ represent the CBAM operation and GMW-Non-Local operation, respectively, and conv$_{\text{tail}}(\cdot)$ represents the convolutional operation at the end. Traversing all subgroups constitutes one round. We choose to carry out k rounds. The initial input for each round is as follows:

$$\mathbf{X}^{j} = \begin{cases}\mathbf{X}, & j = 0 \\ \text{conv}(\text{cat}(\mathbf{Z}^{j-1}, \mathbf{X}^{j-1}), & 1 \le j \le k-1\end{cases} \tag{24}$$

where $\mathbf{X}^j$ is the initial input before the j-th round of grouping, and $\mathbf{Z}^{j-1}$ is the output after the $(j-1)$-th round of MS-GLAM.

3.3. Loss function

The facial dual-branch features $\mathbf{F}^r$ and $\mathbf{F}^{ir}$ are processed through MS-GLAM and then decoded with the aim of disentangling gaze-relevant and gaze-irrelevant information. Based on the cognitive premise that facial features beneficial for eye

region reconstruction inherently contain gaze-relevant information, the upper and lower face branches are trained with region-specific reconstruction losses. The propagation of the loss function in our overall framework is shown in Fig. 2. The reconstruction loss for the eye regions L_1 is defined as follows:

$$L_1 = \text{MSE}(\hat{\mathbf{I}}^l, \mathbf{I}^l) + \text{MSE}(\hat{\mathbf{I}}^r, \mathbf{I}^r) \tag{25}$$

where $\hat{\mathbf{I}}^l$ and $\hat{\mathbf{I}}^r$ represent the reconstructed images of the left and right eyes, respectively, while $\mathbf{I}^l$ and $\mathbf{I}^r$ denote the corresponding original images. MSE$(\cdot)$ represents the mean squared error. The reconstruction loss for the non-eye regions L_2 is defined as follows:

$$L_2 = \text{MSE}(\hat{\mathbf{top}}, \mathbf{top}) + \text{MSE}(\hat{\mathbf{mid}}, \mathbf{mid}) + \text{MSE}(\hat{\mathbf{bot}}, \mathbf{bot}) \tag{26}$$

where, $\hat{\mathbf{top}}$, $\hat{\mathbf{mid}}$ and $\hat{\mathbf{bot}}$ correspond to the reconstructed top, middle, and bottom face regions after removing the eye region, while **top**, **mid** and **bot** denote their respective original regions.

The distilled facial upper-branch features are combined with local eye features and head pose features to jointly predict gaze angles. The loss of gaze estimation L_g, which measures the discrepancy between the predicted gaze direction and the ground truth, is defined as:

$$L_g = \frac{1}{N} \sum_{i=1}^{N} |\hat{\mathbf{g}}_i - \mathbf{g}_i| \tag{27}$$

where N is the number of samples, $\hat{\mathbf{g}}_i$ and $\mathbf{g}_i$ represent the predicted and ground truth gaze angles for the i-th sample, respectively, and $|\cdot|$ takes the absolute value of its argument.

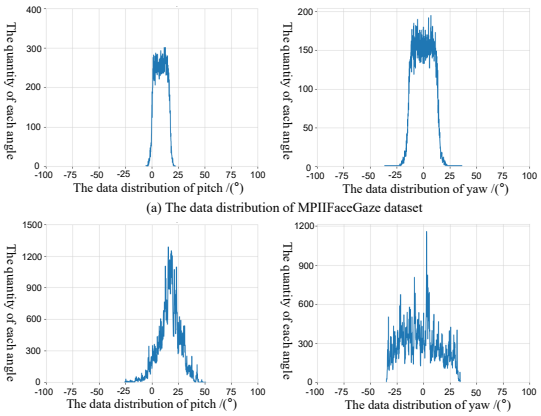
The loss term L_1 guides the mask $\mathbf{K}$ to retain gaze-relevant information, whereas L_2 encourages $(\mathbf{1} - \mathbf{K})$ to reconstruct non-eye regions and suppress irrelevant noise. Meanwhile, L_g further refines this disentanglement by highlighting features most indicative of gaze direction. Collectively, these objectives jointly optimize $\mathbf{K}$ into a region-aware weighting map that effectively separates gaze-relevant from gaze-irrelevant components.

4. Experiments

4.1. Datasets

To evaluate the performance of our model, experiments are conducted on publicly available gaze estimation datasets: MPIIFaceGaze [10] and RT-GENE [28]. The MPIIFaceGaze dataset is smaller in scale, while the RT-GENE dataset is larger. All datasets are preprocessed as described in [35] for fairness, and eye images are cropped using eye corner position information. To compute the reconstruction loss, the eye corner position information in MPIIFaceGaze and RT-GENE is added to the label files for model training.

MPIIFaceGaze. The MPIIFaceGaze dataset is an extension of the MPIIGaze [8] dataset, providing 45K facial images captured over several months from 15 subjects using a laptop camera. It includes various lighting conditions, facial expressions, and head poses in an unconstrained environment.



(a) The data distribution of MPIIFaceGaze dataset

(b) The data distribution of Rt-Gene datasets

Figure 4: The data distribution of MPIIFaceGaze and RT-GENE dataset.

Table 1: Comparison with the state-of-the-art methods.

Method	Year	Angular error		FLOPs	Params	Time
		MPIIFaceGaze	RT-GENE			
FullFace [10]	2017	4.93°	10.00°	2.99G	196.6M	-
RT-GENE [28]	2018	4.66°	8.60°	30.81G	82.0M	-
Dilated-Net [14]	2018	4.42°	8.38°	3.14G	3.9M	23ms
Gaze360 [12]	2019	4.06°	7.06°	7.29G	11.9M	49ms
CA-Net [15]	2020	4.27°	8.27°	15.60G	34.1M	-
AGE-Net [37]	2021	4.09°	7.44°	-	-	-
GazeTR [11]	2022	4.00°	6.55°	1.84G	11.4M	19ms
GazeCaps [38]	2023	4.06°	6.92°	1.82G	11.7M	-
EM-Net [39]	2024	3.88°	6.27°	**0.31G**	**2.9M**	18ms
GazeSerMerge [40]	2024	3.88°	6.46°	3.03G	65.3M	-
DMAGaze (Ours)	-	**3.74**	**6.17**	5.65G	36.7M	**7ms**

Notes: **-** denotes no results released in these baselines.

RT-GENE. The RT-GENE dataset contains 122K images from 15 subjects wearing eye-tracking glasses. The subjects are positioned between 0.5 and 2.9 meters from the camera, with notable variations in head pose.

4.2. Setup

Training. Our experiments are implemented with PyTorch and conducted on NVIDIA GeForce RTX 4090. We train DMAGaze with a batch size of 48 and for 40 epochs on the MPIIFaceGaze datasets, and with a batch size of 64 and for 40 epochs on the RT-GENE dataset. The AdamW optimizer [36] is used for model optimization, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We adjust the learning rate using MultiStepLR [36] with an initial learning rate of $1e-4$, milestones at $[10, 25]$ and a learning rate decay factor of 0.1.

Implementation details. The input consists of normalized face images of size $224 \times 224 \times 3$ and eye images of size $36 \times 60 \times 3$. The face encoder in DMAGaze utilizes a pre-trained ResNet34. In MS-GLAM, the number of rounds k for the cascaded attention structure is set to 4, and the number of input feature groups n is set to 4. The variance parameter σ in GMW is set to 1.0. We evaluate the performance using the widely used metric, i.e. the angular error, which measures the angular difference between the predicted $\mathbf{g}^*$ and ground truth 3D gaze vectors $\mathbf{g}$:

$$\text{Angular Error} = \arccos\left(\frac{\mathbf{g} \cdot \mathbf{g}^*}{\|\mathbf{g}\| \cdot \|\mathbf{g}^*\|}\right) \tag{28}$$

A smaller angular error indicates better gaze estimation performance.

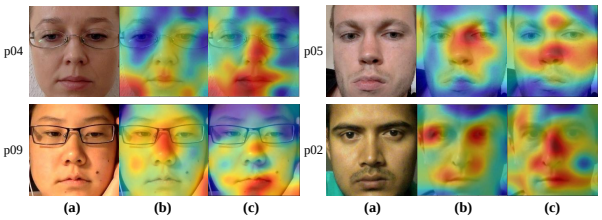


Figure 5: Visualization of the attention maps of the upper and lower face branches of the proposed DMAGaze. (a) Input images from the MPIIFaceGaze dataset. (b) Attention maps from the upper face branch. (c) Attention maps from the lower face branch.

4.3. Comparison with the state-of-the-art methods

We compare DMAGaze with state-of-the-art methods on the MPIIFaceGaze and RT-GENE datasets, both of large appearance variations and unbalanced gaze distributions, as shown in Fig. 4. The sample points are mainly concentrated within angular ranges of ±25° and ±50°, posing a challenge for gaze estimation. This makes them suitable for validating the effectiveness of our proposed method. As shown in Table 1, our method achieves angular errors of 3.74° and 6.17° on these datasets, outperforming all compared methods by 3.61% and 1.59%, respectively. This result demonstrates the effectiveness of the DMAGaze framework we proposed, which disentangles gaze-relevant and gaze-irrelevant facial information using the Disentangler and MS-GLAM, and integrates global gaze-relevant features with head pose and local eye features for accurate gaze estimation.

DMAGaze achieves a strong balance between accuracy and efficiency, requiring only 5.65G FLOPs, 36.7M parameters, and 7 ms per sample—suitable for real-time gaze estimation. As shown in Fig. 5, the upper face branch focuses on eye regions while leveraging nearby facial cues, whereas the lower branch attends to non-eye areas such as the nose, mouth, and contours. These complementary patterns confirm the framework's ability to disentangle gaze-relevant and irrelevant information.

4.4. Performance analysis

Cross-subject evaluation. To further evaluate the adaptation capability of DMAGaze, we conduct leave-one-subject-out validation on the MPIIFaceGaze dataset. Table 2, shows that DMAGaze achieves best (in bold) or second-best (underlined) performance on 12 out of 15 subjects, with notable improvements for individuals exhibiting distinct facial morphology variations (e.g., p01, p06, and p09). This demonstrates that the disentanglement mechanism is effective in separating gaze-relevant cues from subject-specific factors, generalizing well across subjects.

Angle-wise evaluation. To assess DMAGaze's robustness across gaze angles, test samples are divided into yaw intervals from −50° to +30° (Table 3). DMAGaze attains the lowest errors of 3.60° and 6.03° in the (−10° to +10°) and (−20° to −10°) ranges on the MPIIFaceGaze and RT-GENE datasets, respectively, where frontal eye visibility is highest. Performance declines at larger yaw angles ($|yaw| > 20°$) due to reduced iris–sclera visibility and partial occlusion. These results show that the proposed disentanglement framework, MS-GLAM, and GMW modules maintain robustness under moderate head poses, while extreme orientations warrant further exploration.

Table 2: Comparison with the state-of-the-art methods on individual subjects.

Subject	FullFace [10]	Dilated-Net [14]	GazeTR [11]	DMAGaze (Ours)
p00	3.04°	2.74°	**2.13°**	**2.13°**
p01	5.66°	2.91°	4.12°	**2.41°**
p02	4.81°	6.35°	5.31°	**5.28°**
p03	4.36°	3.00°	**2.42°**	2.67°
p04	4.91°	3.64°	2.74°	**2.59°**
p05	5.76°	4.77°	4.06°	**3.94°**
p06	4.18°	3.90°	3.44°	**2.97°**
p07	5.61°	**4.17°**	4.04°	4.18°
p08	4.70°	4.80°	**4.05°**	4.17°
p09	5.12°	4.58°	4.23°	**3.63°**
p10	4.30°	4.84°	5.45°	**4.85°**
p11	4.42°	4.92°	5.30°	**4.94°**
p12	6.57°	3.92°	**3.60°**	**3.60°**
p13	4.19°	5.41°	3.69°	**3.52°**
p14	6.32°	6.29°	5.38°	**5.33°**
Average	4.93°	4.42°	4.00°	**3.74°**

Table 3: Performance across different yaw intervals.

Yaw interval	MPIIFaceGaze	RT-GENE
−50° to −30°	-	8.75°
−30° to −20°	11.95°	7.08°
−20° to −10°	4.10°	**6.03°**
−10° to +10°	**3.60°**	6.20°
+10° to +20°	3.85°	8.23°
+20° to +30°	12.51°	10.17°

4.5. Selection of attention modules

We evaluate attention module selection and combination strategies in the proposed MS-GLAM framework through comparative experiments on the MPIIFaceGaze dataset. As summarized in Table 4, the analysis focuses on two aspects: hybrid channel–spatial attention and global contextual attention. For the former, we test CBAM [24], GAM [41], and SCSA [42], representing different attention paradigms. For the latter, we compare Non-Local [25], Agent Attention [43], and our improved GMW-enhanced Non-Local module. All modules are integrated into the same MS-GLAM position with identical input/output dimensions and iteration number ($k = 4$) for fair evaluation.

We evaluated three configurations: independent replacement, simple concatenation, and the combined configuration. In the independent replacement setup, each attention module was individually embedded into the cascaded attention framework to assess its standalone effect. As shown in Table 4, GMW-Non-Local outperformed the standard Non-Local by 1.04%, demonstrating its advantage in capturing gaze-relevant dependencies and its effectiveness in feature selection. In the simple concatenation setup, where the proposed cascaded attention was replaced with a basic concatenation scheme, GMW-Non-Local still achieved the best result of 3.83°, though overall performance decreased, underscoring the benefit of the cascaded design. Finally, the combined configuration, integrating CBAM with GMW-Non-Local within the cascaded framework, achieved the best overall performance of 3.74°, confirming the complementary strengths of global and local attention.

4.6. Parameter analysis

We further analyze the effect of the hyperparameter rounds k in the cascaded attention structure under $\sigma = 1.0$. As shown in Table 5(a), the best performance occurs at $k = 4$ with an angular error of 3.74°, balancing feature representation and generalization. Fewer iterations ($k = 2$) limit feature extraction, while

Table 4: Comparative experiments on MPIIFaceGaze dataset.

Trial number	Multi-scale	Attention module selection	Angular error
Independent replacement			
1	✓	CBAM	3.82°
2	✓	GAM	3.86°
3	✓	SCSA	3.83°
4	✓	Non-Local	3.85°
5	✓	GMW-Non-Local	**3.81°**
6	✓	Agent Attention	3.84°
Simple concatenation			
7	×	CBAM	3.85°
8	×	GAM	3.89°
9	×	SCSA	3.85°
10	×	GMW-Non-Local	**3.83°**
11	×	Agent Attention	3.91°
Combined configuration			
12	✓	CBAM + GMW-Non-Local	**3.74°**
13	✓	GAM + GMW-Non-Local	3.82°
14	✓	SCSA + GMW-Non-Local	3.83°
15	✓	CBAM + Agent Attention	3.83°
16	✓	GAM + Agent Attention	3.87°
17	✓	SCSA + Agent Attention	3.82°

* **Multi-scale** stands for whether the attention module is applied to our proposed cascaded attention structure.

excessive ones ($k = 8$) cause overfitting. For σ with $k = 4$, the lowest error is achieved at $\sigma = 1.0$, suggesting an optimal balance between local and global information for gaze-relevant feature capture.

4.7. Ablation study

We conducted ablation studies on the MPIIFaceGaze dataset to explore the impact of different modules. (1) **Baseline**: gaze estimation is performed using only the eye branch. (2) **Face branch (w/o MS-GLAM)**: the face branch is introduced to disentangling gaze-relevant and gaze-irrelevant features using the Disentangler, where "w/o MS-GLAM" indicates the absence of MS-GLAM. (3) **Head pose**: head pose features extracted using simple convolution are incorporated into the final gaze estimation. (4) **MS-GLAM (w/o GMW)**: the MS-GLAM is added, where "w/o GMW" indicates the exclusion of GMW. (5) **GMW**: the GMW is added to the Non-Local module in MS-GLAM, representing the full model.

Table 5: Ablation study of DMAGaze.

(a) Ablation study on parameter k and σ.

Rounds k	Variance σ	Angular error
2	1.0	3.87°
4	1.0	**3.74°**
8	1.0	3.81°
4	0.5	3.87°
4	1.0	**3.74°**
4	2.0	3.84°

(b) Ablation on each module.

Method	Angular error
Baseline	5.02°
+ Face branch (w/o MS-GLAM)	4.04°
+ Head pose	3.86°
+ MS-GLAM (w/o GMW)	3.79°
+ GMW (total model)	**3.74°**

(c) Ablation on binary and continuous masks.

Method	Mask type	Angular error
-w/o MS-GLAM	Binary	3.91°
DMAGaze (total model)		3.80°
-w/o MS-GLAM	Continuous	3.86°
DMAGaze (total model)		**3.74°**

Table 5(b) summarizes the ablation results. The baseline yields an angular error of 5.02°. Adding the face branch lowers the error by 19.52%, validating the effectiveness of our cognitively inspired framework that disentangles gaze-relevant facial information via separate reconstruction of eye and non-eye regions. Incorporating head pose features further reduces the error by 4.46%, confirming their contribution. Adding MS-GLAM (w/o GMW) achieves 3.79°, highlighting the cascaded attention's ability to fuse global and local information. Finally, introducing GMW brings an additional 1.32% improvement, confirming the advantage of Gaussian similarity over dot-product computation. Fig. 6 shows the gaze estimation performance of each component in the ablation studies under various conditions, such as glasses occlusion, variable lighting, and head pose variations. For clarity, gaze directions are visualized with a standardized origin located at the center of the eye, where the green line represents the ground-truth gaze and the red line denotes the estimated gaze.

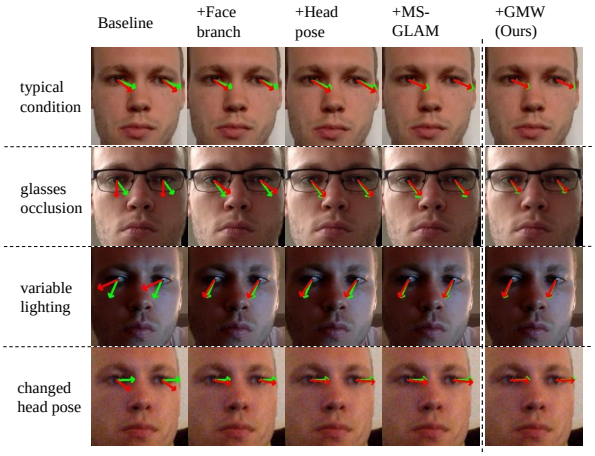


Figure 6: Visualization of gaze estimation results in different components of our gaze estimation model from ablation studies.

Table 5(c) shows the ablation study results on the impact of mask type. When using binary masks [22], the model achieves an angular error of 3.91°. In contrast, replacing binary masks with our continuous masks reduce the error to 3.86°. For the total model, binary masks achieves 3.80°, while continuous masks further lower it to 3.74°. This consistent gain confirms that continuous masks, by enabling smooth rather than rigid feature weighting, enhance both the disentanglement process and MS-GLAM for finer separation of gaze-relevant and irrelevant features.

5. Conclusion

We have presented DMAGaze, a gaze estimation framework that leverages facial and eye region relationships for accurate predictions. A Disentangler with a continuous mask separates gaze-relevant facial features, which are combined with local eye and head pose features via a cascaded attention structure integrating global and local information at multiple scales. We further enhance non-local operations with Gaussian modulated weighting to capture nonlinear dependencies. Experiments show state-of-the-art performance with angular errors of 3.74° and 6.17°, and ablation studies confirm the effectiveness of each module. DMAGaze demonstrates robust performance, highlighting its potential for human–computer interaction and related applications.

References

[1] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (8) (2019) 1913–1927.

[2] W. Wang, J. Shen, Deep visual attention prediction, IEEE Transactions on Image Processing 27 (5) (2017) 2368–2378.

[3] A. M. Soccini, Gaze estimation based on head movements in virtual reality applications using deep learning, in: 2017 IEEE Virtual Reality, 2017, pp. 413–414.

[4] P. K. Sharma, P. Chakraborty, A review of driver gaze estimation and application in gaze behavior understanding, Engineering Applications of Artificial Intelligence 133 (2024) 108117.

[5] M. Lombardi, E. Maiettini, D. D. Tommaso, A. Wykowska, L. Natale, Toward an attentive robotic architecture: Learning-based mutual gaze estimation in human–robot interaction, Frontiers in Robotics and AI 9 (2022).

[6] J. Li, Z. Chen, Y. Zhong, H. K. Lam, J. Han, G. Ouyang, et al., Appearance-based gaze estimation for ASD diagnosis, IEEE Transactions on Cybernetics 52 (7) (2022) 6504–6517.

[7] Z. H. Wan, C. H. Xiong, W. B. Chen, H. Y. Zhang, Robust and accurate pupil detection for head-mounted eye tracking, Computers & Electrical Engineering 93 (2021).

[8] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4511–4520.

[9] Y. Yu, J. M. Odobez, Unsupervised representation learning for gaze estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7314–7324.

[10] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, It's written all over your face: Full-face appearance-based gaze estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 51–60.

[11] Y. Cheng, L. Feng, Gaze estimation using transformer, in: 2022 26th International Conference on Pattern Recognition, IEEE, 2022, pp. 3341–3347.

[12] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, A. Torralba, Gaze360: Physically unconstrained gaze estimation in the wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6912–6921.

[13] A. Cătrună, A. Cosma, E. Rădoi, CrossGaze: A strong method for 3d gaze estimation in the wild, in: 2024 18th IEEE International Conference on Automatic Face and Gesture Recognition, 2024, pp. 500–507.

[14] Z. Chen, B. E. Shi, Appearance-based gaze estimation using dilated-convolutions, in: Asian Conference on Computer Vision, Springer, 2018, pp. 309–324.

[15] Y. Cheng, S. Huang, F. Wang, C. Qian, F. Lu, A coarse-to-fine adaptive network for appearance-based gaze estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10623–10630.

[16] P. Yin, G. Zeng, J. Wang, D. Xie, CLIP-Gaze: Towards general gaze estimation via visual-linguistic model, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 6729–6737.

[17] J. Wang, H. Ruan, M. Wang, C. Zhang, H. Li, J. Zhou, GazeCLIP: Towards enhancing gaze estimation via text guidance, arXiv preprint arXiv:2401.00260 (2023).

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[19] P. Vuillecard, J. Odobez, Enhancing 3D gaze estimation in the wild using weak supervision with gaze following labels, in: Proc. IEEE/CVF CVPR, 2025, pp. 13508–13518.

[20] P. Pathirana, S. Senarath, D. Meedeniya, S. Jayarathna, Eye gaze estimation: A survey on deep learning-based approaches, Expert Systems with Applications 199 (2022) 116894.

[21] S. Wang, Y. Huang, J. Xie, F. Chen, Z. Wang, Cross-dataset gaze estimation by evidential inter-intra fusion, arXiv preprint arXiv:2409.04766 (2024).

[22] X. Yu, H. H. Tseng, S. Yoo, H. Ling, Y. Lin, INSURE: An information theory inspired disentanglement and purification model for domain generalization, IEEE Transactions on Image Processing (2024).

[23] Y. Zhang, J. Li, G. Ouyang, Gaze estimation with multi-scale attention-based convolutional neural networks, in: Proceedings of the 29th International Conference on Mechatronics and Machine Vision in Practice, IEEE, 2023, pp. 1–6.

[24] S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.

[25] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[26] Y. Cheng, Y. Bao, F. Lu, PureGaze: Purifying gaze feature for generalizable gaze estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 436–443.

[27] K. Tian, H. Wang, Z. Wu, Y. Lyu, M. I. Vai, Y. Liu, Multibit attention fusion for gaze estimation using 12-bit RAW data from CMOS sensors, IEEE Trans. Instrum. Meas. 74 (2025) 1–13.

[28] T. Fischer, H. J. Chang, Y. Demiris, RT-GENE: Realtime eye gaze estimation in natural environments, in: The European Conference on Computer Vision, 2018, pp. 334–352.

[29] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, A. Torralba, Eye tracking for everyone, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2176–2184.

[30] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, L2CS-Net: Fine-grained gaze estimation in unconstrained environments, in: 2023 8th International Conference on Frontiers of Signal Processing, IEEE, 2023, pp. 98–102. doi:10.1109/ICFSP59764.2023.10372944.

[31] P. Pathirana, S. Senarath, D. Meedeniya, S. Jayarathna, Single-user 2d gaze estimation in retail environment using deep learning, in: 2022 2nd International Conference on Advanced Research in Computing (ICARC), 2022, pp. 206–211.

[32] S. Senarath, P. Pathirana, D. Meedeniya, S. Jayarathna, Customer gaze estimation in retail using deep learning, IEEE Access 10 (2022) 64904–64919.

[33] M. L.R.D, A. Mukhopadhyay, P. Biswas, Distraction detection in automotive environment using appearance-based gaze estimation, in: Companion Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22 Companion, Association for Computing Machinery, 2022, p. 38–41.

[34] M. L.R.D, A. Mukhopadhyay, K. Anand, S. Aggarwal, P. Biswas, PARKS-Gaze - a precision-focused gaze estimation dataset in the wild under extreme head poses, in: Companion Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22 Companion, Association for Computing Machinery, 2022, p. 81–84.

[35] Y. Cheng, H. Wang, Y. Bao, F. Lu, Appearance-based gaze estimation with deep learning: A review and benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[36] I. Loshchilov, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[37] L. Murthy, P. Biswas, Appearance-based gaze estimation using attention and difference mechanism, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3143–3152.

[38] H. Wang, J. O. Oh, H. J. Chang, J. H. Na, M. Tae, Z. Zhang, et al., GazeCaps: Gaze estimation with self-attention-routed capsules, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2669–2677.

[39] Z. Cheng, Y. Wang, G. Xia, EM-Net: Gaze estimation with expectation maximization algorithm, arXiv preprint arXiv:2412.08074 (2024).

[40] L. Wu, B. E. Shi, Merging multiple datasets for improved appearance-based gaze estimation, in: Proceedings of the International Conference on Pattern Recognition, Springer Nature Switzerland, 2024, pp. 77–90.

[41] Y. Liu, Z. Shao, N. Hoffmann, Global attention mechanism: Retain information to enhance channel-spatial interactions, arXiv preprint arXiv:2112.05561 (2021).

[42] Y. Si, H. Xu, X. Zhu, W. Zhang, Y. Dong, Y. Chen, H. Li, SCSA: Exploring the synergistic effects between spatial and channel attention, Neurocomputing 634 (2025) 129866.

[43] D. Han, T. Ye, Y. Han, Z. Xia, S. Pan, P. Wan, S. Song, G. Huang, Agent Attention: On the integration of softmax and linear attention, in: Computer Vision–ECCV 2024, Vol. 15108 of Lecture Notes in Computer Science, Springer, Cham, 2025.