electronics

MDPI

Article

Gaze-Swin: Enhancing Gaze Estimation with a Hybrid CNN-Transformer Network and Dropkey Mechanism

Ruijie Zhao, Yuhuan Wang, Sihui Luo, Suyao Shou and Pinyan Tang *

Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China; 216001944@nbu.edu.cn (R.Z.); 216001847@nbu.edu.cn (Y.W.); luosihui@nbu.edu.cn (S.L.); 216000483@nbu.edu.cn (S.S.)

* Correspondence: tangpinyan@nbu.edu.cn

Abstract: Gaze estimation, which seeks to reveal where a person is looking, provides a crucial clue for understanding human intentions and behaviors. Recently, Visual Transformer has achieved promising results in gaze estimation. However, dividing facial images into patches compromises the integrity of the image structure, which limits the inference performance. To tackle this challenge, we present Gaze-Swin, an end-to-end gaze estimation model formed with a dual-branch CNN-Transformer architecture. In Gaze-Swin, we adopt the Swin Transformer as the backbone network due to its effectiveness in handling long-range dependencies and extracting global features. Additionally, we incorporate a convolutional neural network as an auxiliary branch to capture local facial features and intricate texture details. To further enhance robustness and address overfitting issues in gaze estimation, we replace the original self-attention in the Transformer branch with Dropkey Assisted Attention (DA-Attention). In particular, this DA-Attention treats keys in the Transformer block as Dropout units and employs a decay Dropout rate schedule to preserve crucial gaze representations in deeper layers. Comprehensive experiments on three benchmark datasets demonstrate the superior performance of our method in comparison to the state of the art.

Keywords: gaze estimation; Transformer; convolutional neural networks; Dropkey

check for updates

Citation: Zhao, R.; Wang, Y.; Luo, S.; Shou, S.; Tang, P. Gaze-Swin: Enhancing Gaze Estimation with a Hybrid CNN-Transformer Network and Dropkey Mechanism. *Electronics* **2024**, *13*, 328. https://doi.org/10.3390/electronics13020328

Academic Editor: Chiman Kwan

Received: 11 December 2023
Revised: 6 January 2024
Accepted: 10 January 2024
Published: 12 January 2024

Copyright: © 2024 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Gaze estimation from facial images holds significant importance in understanding human cognition and behavior. Nowadays, gaze estimation serves as an indispensable tool in numerous domains, including psychological research [1], human–computer interaction techniques [2–4], Virtual Reality (VR) [5], and semi-autonomous driving [6,7].

Within the gaze estimation domain, two primary approaches have emerged: model-based and appearance-based methods. Model-based methods [8–15] leverage geometric eye models to predict gaze direction [16]. These methods typically explore anatomical features of the human eyes such as the pupil center, iris edge, and corneal reflections. By analyzing these features, they determine a set of unique eye characteristics, which are essential for accurate gaze prediction [17]. However, these methods often require specialized equipment like high-resolution cameras or infrared sensors and can be sensitive to external factors such as lighting and user head movements. In contrast, appearance-based techniques directly infer gaze orientation by analyzing features in facial or ocular images using machine/deep learning algorithms. This approach does not require specialized equipment, offering a more versatile and accessible solution. Despite potential challenges such as occlusion or varying facial orientations, appearance-based methods have demonstrated high accuracy in gaze estimation [18]. Their adaptability and ease of integration make them particularly beneficial for practical applications in fields like consumer electronics and interactive systems.

In the gaze estimation domain, convolutional neural networks (CNNs) have proven to be highly effective, excelling at extracting hierarchical features from images. This capability enables them to discern intricate details within facial structures, a crucial aspect of accurate

gaze estimation. However, while CNNs make significant contributions to the field, they encounter limitations when analyzing the complex interplay between gaze direction, head pose, and eye orientation. Gaze estimation is a complex task influenced not only by the eyes' orientation but also by the broader context of head pose. CNNs, predominantly designed for local feature extraction, often struggle to capture these broader spatial relationships effectively. This limitation becomes particularly evident when dealing with extreme head poses. Although adept at identifying detailed features of the eyes, CNNs may not fully grasp how these features relate to the overall head pose, potentially compromising the accuracy of gaze estimations.

In the realm of computer vision, the advent of Transformer, originally proposed by Vaswani et al. [19], has been transformative, demonstrating exceptional capabilities in both natural language processing (NLP) and computer vision tasks. The Visual Transformer (ViT) [20], a refined iteration of the Transformer architecture, has proven particularly adept at segmenting images into discrete patches and treating them as tokens, a practice analogous to NLP. This approach allows ViT to effectively manage long-range dependencies, offering a comprehensive understanding of visual information, essential for complex tasks like gaze estimation. However, ViT faces its own set of challenges. Its limited connectivity between adjacent windows may lead to incomplete information extraction from facial images. The patch-based segmentation approach, while innovative, can sometimes distort the structural integrity of the image. This distortion becomes problematic in gaze estimation, a task that requires a complete and accurate representation of the eye and surrounding facial features. Despite ViT's proficiency in object classification through single patches, its application to regression tasks like gaze estimation presents significant challenges due to the need for a more holistic and interconnected understanding of facial features.

The recent advent of the Swin Transformer [21] marks a significant advancement in Transformer-based computer vision. Distinguished by its hierarchical Transformer structure, the Swin Transformer innovatively employs shifted windows for attention computation within local windows. This design not only reduces the number of parameters but also accelerates computation, all while maintaining high levels of accuracy. A key benefit of the shifted windows is their ability to address the issue of image structure distortion, a challenge previously noted in the Visual Transformer (ViT). This makes the Swin Transformer a particularly promising approach for tasks like gaze estimation, where maintaining the integrity of image structure is crucial.

Building on these advancements, we introduce Gaze-Swin, an innovative model that is a confluence of CNNs and the Swin Transformer, encapsulating the strengths of both in a dual-branch CNN-Transformer architecture. Gaze-Swin leverages the hierarchical Transformer framework of the Swin Transformer, incorporating the shifted window mechanism [22]. This integration is pivotal in overcoming the limitations associated with image structure distortion observed in ViT. The shifted window mechanism in Gaze-Swin allows for a more comprehensive and nuanced extraction of information from images, a critical factor in accurately estimating gaze. By capturing intricate details and maintaining the structural integrity of the images, Gaze-Swin represents a significant leap forward in the field of gaze estimation, pushing the boundaries of what is achievable with current computer vision technologies.

In practical applications, the limited participants and constrained variations inherent in gaze estimation datasets impede the generalization capabilities of models. To address these challenges, we replace the original self-attention in the Transformer branch with Dropkey Assisted Attention (DA-Attention). The conventional Dropout impacts attention weights by averaging non-Dropout units following Softmax normalization, which fails to penalize localized score peaks. In contrast, the DA-Attention treats keys in the Transformer block as Dropout units. Its unique pre-Softmax Dropout [23] mitigates localized score peaks, effectively counteracting overfitting in Transformer architectures. Furthermore, diverging from traditional Dropout methodologies that maintain a fixed Dropout rate, we employ a progressively reduced Dropout rate as the self-attention layers grow deeper. This

adaptive rate modulation guards against overfitting in the shallow layers while preserving crucial gaze representations in the deeper layers, thereby enhancing the robustness and stability of model training.

To summarize, our main contributions are summarized as follows: first, we pioneer the application of the dual-branch CNN-Transformer architecture to solve the image structure distortion problem commonly associated with ViT in gaze estimation. Second, we introduce DA-Attention within the Gaze-Swin model to enhance its robustness and mitigate the risk of overfitting. Lastly, our proposed Gaze-Swin model achieves state-of-the-art performance on three benchmark datasets, clearly demonstrating its superiority in gaze estimation tasks.

2. Related Work

2.1. Appearance-Based Gaze Estimation Using CNN

In the evolving landscape of gaze estimation, traditional appearance-based methods have made significant strides by harnessing the power of CNNs. These methods initially focused on creating mapping functions that correlate facial or ocular images with gaze directions [24]. A notable advancement in this domain was introduced by Zhang et al. [25], who employed the LeNet architecture to predict gaze direction from grayscale ocular images, demonstrating impressive accuracy. Building upon this, Fischer et al. [26] explored the potential of VGGNets for gaze feature extraction, combining data from two networks for a more comprehensive analysis.

Furthering this progression, Cheng et al. [27] developed a four-channel network that simultaneously processes features from both the left and right eyes, as well as facial characteristics, enhancing the system's capacity to accurately estimate gaze direction. Krafka et al. [28] brought an innovative approach by integrating eye and facial features along with a face grid mask, which aids in pinpointing the head's position, thus significantly boosting the system's overall effectiveness.

Moreover, Zhang et al. [29] introduced a novel spatial weight CNN that facilitates full-face gaze estimation. This method allows for the flexible enhancement or suppression of information in different facial regions, contributing to more refined and accurate gaze prediction. In another notable contribution, Cheng et al. [30] focused on dual-view image processing, utilizing fused CNN features to augment the original feature set, achieving noteworthy improvements in gaze estimation accuracy.

Each of these studies underscores the pivotal role of CNNs in enhancing the feature extraction process, critical for accurate gaze estimation. However, they also highlight the ongoing challenges in the field, particularly in further refining and improving the capabilities of CNNs to meet the diverse and complex requirements of gaze estimation technology.

2.2. Appearance-Based Gaze Estimation Using Transformer

The Transformer architecture, initially introduced by Vaswani et al. [19] for natural language tasks, has garnered remarkable success. Dosovitskiy et al. [20] extended this architecture into the domain of computer vision, resulting in the Visual Transformer (ViT). Further evolution transpired with Liu et al. [22], who introduced the Swin Transformer, incorporating sliding windows to effectively address computational demands. Within the sphere of gaze estimation, Cheng and Lu [24] emerged as pioneers in deploying the Visual Transformer for this task, showcasing substantial performance enhancements. Li et al. [18] utilized a Multi-head Self-Attention mechanism to extract features from eye and facial images, employing a recurrent neural network (RNN) to capture dynamic features from video data. Nagpure et al. [31] employed Neural Architecture Search (NAS) methods to extract feature maps, subsequently utilizing a Transformer to learn gaze direction from these feature maps.

Expanding the horizon of gaze estimation methodologies, the Gaze-Swin framework incorporates a dual-branch CNN-Transformer architecture. This approach is uniquely designed to handle long-range dependencies while extracting local features, substantially enhancing overall performance. The framework is further augmented by a DA-Attention

mechanism, specifically designed to mitigate overfitting risks. This advanced architecture signifies a significant leap in the field, illustrating the continuous evolution and adaptation of Transformer models in addressing complex challenges in gaze estimation technology.

3. Gaze-Swin

3.1. Overall Architecture

The detailed architecture of Gaze-Swin is illustrated in Figure 1 (the images of the human face in Figure 1 were sourced from the public dataset *MPIIFaceGaze*. This dataset is accessible at http://datasets.d2.mpi-inf.mpg.de/MPIIGaze/MPIIFaceGaze.zip, accessed on 26 April 2023). Gaze-Swin employs the Swin Transformer [22] as the foundational framework to efficaciously encapsulate global dependencies inherent in facial images. In addition, an auxiliary CNN branch is incorporated to augment the model's proficiency in extracting local information. After acquiring the local features from the CNN branch and global features from the Transformer branch, the two features are concatenated to formulate the gaze representation. Subsequently, a multilayer perceptron (MLP) is employed to infer the final gaze direction.

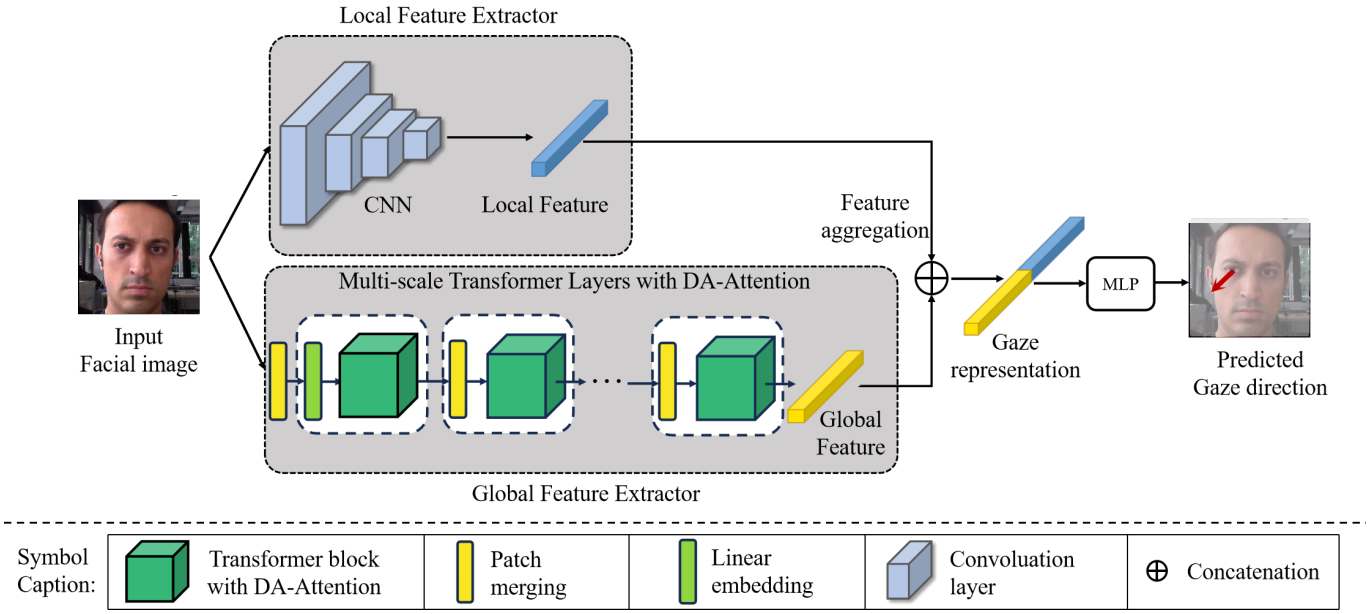

Figure 1. The detailed architecture of Gaze-Swin. The red arrow represents the estimated gaze direction.

3.2. Transformer-Based Global Feature Extractor

In the context of the Swin Transformer branch (as illustrated in Figure 2), the model takes a facial image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ as input and partitions it into non-overlapping patches $\boldsymbol{I}_i \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 48}$ using a patch size of 4×4. A linear embedding layer maps the patches to C dimensions in the initial stage. This is followed by several Swin Transformer blocks that employ a modified self-attention calculation while maintaining the number of tokens. The subsequent three stages comprise a patch merging layer and Swin Transformer blocks. As the network deepens, the number of tokens decreases through patch merging. In each patch merging layer, groups of 2×2 patches are concatenated into $4C$-dimensional concatenated features, then downsampled to $2C$ dimensions using a linear layer. Additionally, the shifted window approach [22] is used to learn global relationships in Swin Transformer blocks. This mechanism allows consecutive Swin Transformer blocks to compute as follows:

$$\hat{\mathbf{z}}_1^l = \text{LN}(\text{W-MSA}(\mathbf{z}_1^{l-1})) + \mathbf{z}_1^{l-1} \quad (1)$$

$$\mathbf{z}_1^l = \mathrm{LN}(\mathrm{MLP}(\hat{\mathbf{z}}_1^l)) + \hat{\mathbf{z}}_1^l \tag{2}$$

$$\hat{\mathbf{z}}_1^{l+1} = \mathrm{LN}(\text{SW-MSA}(\mathbf{z}_1^l)) + \mathbf{z}_1^l \tag{3}$$

$$\mathbf{z}_1^{l+1} = \mathrm{LN}(\mathrm{MLP}(\hat{\mathbf{z}}_1^{l+1})) + \hat{\mathbf{z}}_1^{l+1} \tag{4}$$

Here, $\mathbf{z}_1^{l-1}$ represents the features from the previous layer entering the Window Multi-head Self-Attention (W-MSA) layer module before the Layer Norm (LN) layer. Residual connections exist between each module. Subsequently, the feature similarly passes through the Shifted Window Multi-head Self-Attention (SW-MSA) layer. As a result, we obtain the feature map $\boldsymbol{O} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 8C}$. We employ average pooling to obtain a $8C$-dimensional gaze representation $\mathbf{z}_1^{l+1}$.

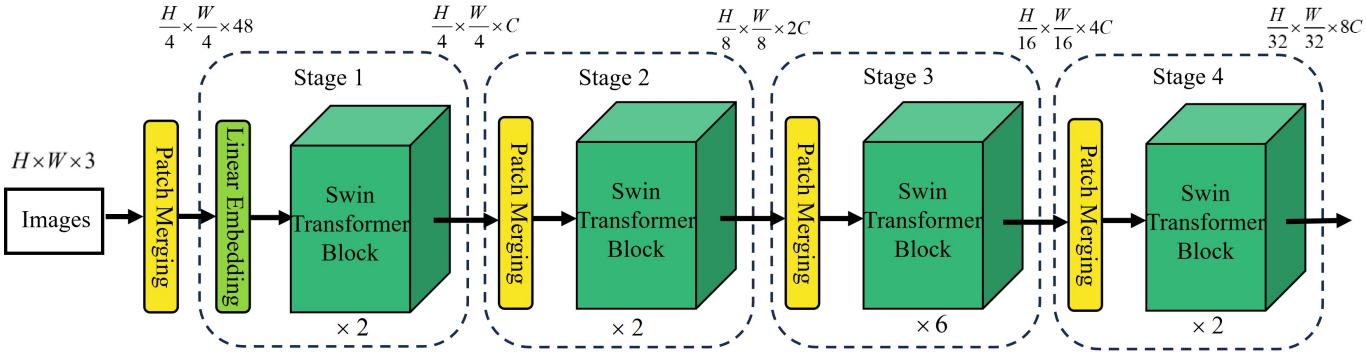

Figure 2. The architecture of the Transformer branch.

3.3. CNN-Based Local Feature Extractor

Turning attention to the CNN branch, we employ ResNet-18 [32] for local information extraction (as illustrated in Figure 3). Gaze-Swin feeds the facial image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ into the CNN branch. The CNN branch comprises four BasicBlocks, each employing 3×3 convolutional kernels, and Rectified Linear Unit (ReLU) activation functions to extract local features from the input images. The residual blocks, each composed of two convolutional layers and a skip connection, effectively address challenges related to vanishing and exploding gradients in deep convolutional neural networks. These residual blocks empower the network to adeptly learn intricate hierarchical representations. After the convolutional layers, a global average pooling layer is implemented to convert the feature maps into a 512-dimensional feature vector, which serves as the gaze representation $\mathbf{z}_2^{l+1}$.

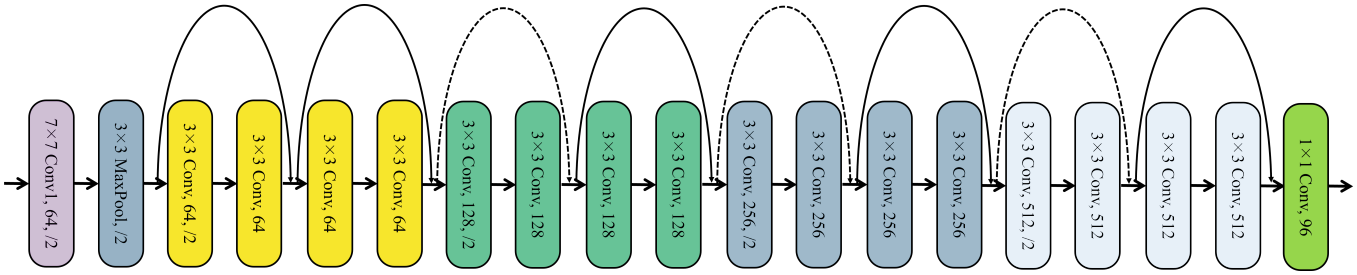

Figure 3. The architecture of the CNN branch.

3.4. Unified Prediction Module

The culmination of the Gaze-Swin architecture lies in the Unified Prediction Module, where the ultimate output $\mathbf{z}^{l+1}$ is obtained by concatenating $\mathbf{z}_1^{l+1}$ and $\mathbf{z}_2^{l+1}$. This concatenated representation is then processed through a multilayer perceptron (MLP) to regress the estimated gaze $\mathbf{g}$ from the gaze representation $\mathbf{z}^{l+1}$. The mathematical formulation of this process is represented as

$$\mathbf{z}^{l+1} = \text{Concat}(\mathbf{z}_1^{l+1}, \mathbf{z}_2^{l+1}) \tag{5}$$

$$\mathbf{g} = \text{MLP}(\mathbf{z}^{l+1}) \tag{6}$$

To train the model effectively, the L1 distance between the predicted gaze direction and the corresponding ground truth is employed as the loss function. This approach ensures its effectiveness in handling long-range dependencies and extracting global features.

3.5. DA-Attention for Gaze-Swin

In this section, we introduce the DA-Attention employed in Gaze-Swin. Figure 4 illustrates the architecture of the Attention layer in the Transformer branch.

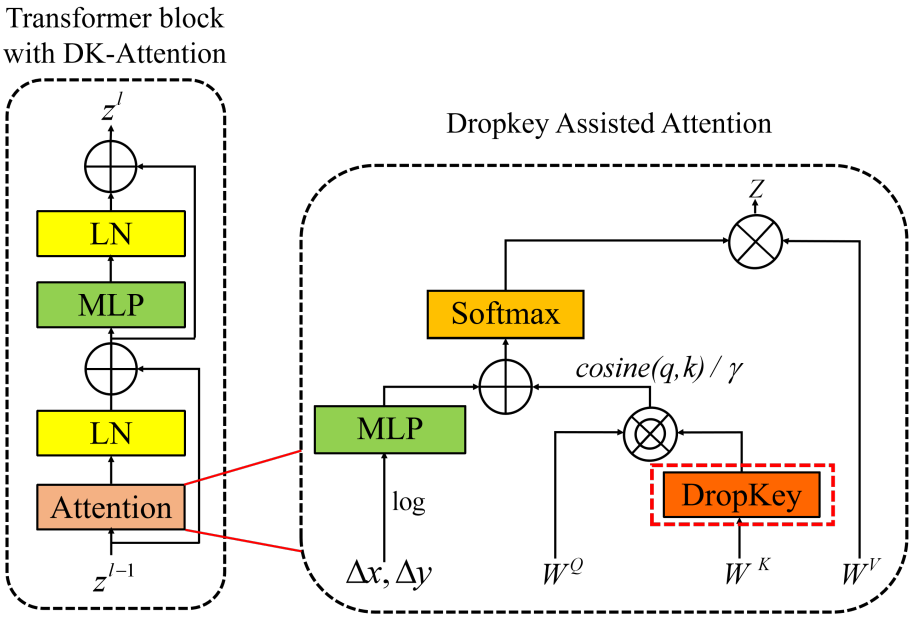

Figure 4. The architecture of DA-Attention in Transformer branch.

Initially, we utilize the technique of continuous relative position bias along with log-spaced coordinates to enhance the network's overall accuracy. The formulation for the relative position bias is given by:

$$B(\Delta x, \Delta y) = \mathcal{G}(\Delta x, \Delta y), \tag{7}$$

where Δx and Δy denote the linear-spaced coordinates. The transformed log-spaced coordinates are defined as $\widehat{\Delta x} = \text{sign}(x) \cdot \log(1 + |\Delta x|)$ and $\widehat{\Delta y} = \text{sign}(y) \cdot \log(1 + |\Delta y|)$. Here, $\mathcal{G}$ represents a small meta-network comprising a 2-layer MLP with ReLU activation, and B indicates the relative position bias in the feature map.

Subsequently, we employ a scaled cosine attention method to prevent attention values from reaching extreme levels. We denote the feature map as $\boldsymbol{Q} \in \mathbb{R}^{N^2 \times C}$, $\boldsymbol{K} \in \mathbb{R}^{N^2 \times C}$, and $\boldsymbol{V} \in \mathbb{R}^{N^2 \times C}$, where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ present the *query*, *key*, and *value*, respectively. Here, C indicates the number of channels in the feature map, and N^2 denotes the number of patches in a window. The attention is calculated as follows:

$$\text{Sim}(\boldsymbol{Q}, \boldsymbol{K}) = \frac{\cos(\boldsymbol{Q}, \boldsymbol{K})}{\gamma} + B \tag{8}$$

where γ is a learnable scalar; unlike heads and layers, γ is not shared and is set to a value greater than 0.01.

In each training iteration, the Dropkey method randomly masks a specific rate of keys in the input key map for each query. Additionally, we introduce variation in the drop ratio for the lower layers of the Swin Transformer blocks. The drop ratio d_i is calculated as

$$d_i = \max(d - (i-1) \cdot \alpha, 0) \tag{9}$$

Here, d_i represents the drop ratio in the i-th Transformer block, and α represents the coefficient for the linear decay of the mask ratio.

Next, we generate a mask matrix $\boldsymbol{M} \in \mathbb{R}^{N^2 \times N^2}$ with a drop ratio d_i, which leads specific elements of the similarity matrix to be $-\infty$. The formulation for $\boldsymbol{M}$ is as follows:

$$m_j = \begin{cases} 0 & \textit{with probability } 1 - d_i \\ -\infty & \textit{with probability } d_i \end{cases} \tag{10}$$

Upon applying the Softmax function, the *key* will be masked at the elements where $m_j = -\infty$. Consequently, the self-attention is calculated as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{SoftMax}(\boldsymbol{M} + \text{Sim}(\boldsymbol{Q}, \boldsymbol{K})) \cdot \boldsymbol{V} \tag{11}$$

4. Experiment

4.1. Datasets

We employ four datasets to demonstrate the efficacy of the Gaze-Swin: ETH-XGaze [33], Gaze360 [34], RT-Gene [26], and EyeDiap [35]. We follow the processing method described by Cheng and Lu [24] for these datasets. The ETH-XGaze dataset comprises a total of 1.1 M images from 110 subjects. For pre-training purposes, we utilize the training set of ETH-XGaze, which consists of 765 K images from 80 subjects. The Gaze360 dataset consists of 129 K training and 26 K test images with gaze annotations under various lighting conditions. Notably, the Gaze360 dataset includes images that feature the back side of subjects. These particular images are marked and were excluded from our analysis, following the approach used in GazeTR [24]. The EyeDiap dataset, on the other hand, comprises a series of video clips from 16 subjects. For our study, we treat the images of each individual as a separate subject. It is important to note that 14 of these subjects are involved in screen target sessions, which formed the primary focus of our analysis. We evaluate the screen target sessions and sample one image every 15 frames from the RGB Kinect videos. To ensure robustness, a 4-fold cross-validation is conducted on the EyeDiap dataset. The RT-Gene dataset comprises 92 K images from 13 subjects and we adhere to the protocol and conduct a 3-fold cross-validation. In our implementation, we utilize the ETH-XGaze dataset for pre-training and the rest of the datasets for evaluation.

4.2. Setup

4.2.1. Training

In our experiments, we use facial images with the size of 224×224 as input, along with a 2D gaze vector comprising the yaw and pitch angles as output. The L1Loss is employed as the loss function for training.

For the Transformer branch, we configure the window size as 7 and the layer numbers as $\{2, 2, 6, 2\}$. The first stage sets the hidden layers' channel count to $C = 96$. The number of heads in the Multi-head Self-Attention for the four stages is $\{3, 6, 6, 24\}$. The drop ratio and linear decay are set to $d = 0.1$, $\alpha = 0.005$. In the CNN branch, Gaze-Swin uses ResNet-18 to attain local information.

In our experimentation, we execute the pre-training phase of Gaze-Swin on the ETH-Gaze dataset, utilizing a batch size of 128 and conducting training for a duration of 10 epochs. The initial learning rate is established at 0.0005, accompanied by a linear warmup of the learning rate for 5 epochs. The parameters for the Adam optimizer, denoted as (β_1, β_2), are configured to (0.9, 0.999). For the subsequent evaluation of three test datasets,

we maintain the majority of the aforementioned training parameters. Furthermore, we adjust the batch size to 32, extend the training duration to 80 epochs, and set the decay step as 60 epochs with 0.5 weight decay.

4.2.2. Evaluation Metric

We adopt the angular error as the primary metric for performance evaluation, a widely accepted measure in the assessment of gaze direction accuracy. A lower angular error signifies heightened model proficiency. The angular error is computed using the following formula:

$$L_{\text{angular}} = \arccos\left(\frac{\mathbf{g} \cdot \hat{\mathbf{g}}}{\|\mathbf{g}\|\|\hat{\mathbf{g}}\|}\right) \tag{12}$$

Here, $\mathbf{g}$ denotes the veritable gaze direction, and $\hat{\mathbf{g}}$ represents the estimated gaze direction. To complement the quantitative assessment, visual representations of Gaze-Swin's performance are depicted in Figure 5 (the human face pictures in Figure 5 were sourced from the public datasets *Gaze360*, *RT-Gene*, and *EyeDiap*. Access to these datasets is available at the following webpages: http://gaze360.csail.mit.edu/, https://zenodo.org/records/2529036, and https://www.idiap.ch/en/scientific-research/data/eyediap, accessed on 26 April 2023).



Figure 5. The proposed Gaze-Swin method estimates gaze direction on three datasets of face images. The green arrows represent the true direction of gaze, while the red arrows indicate the estimated gaze direction produced by Gaze-Swin.

4.3. Comparison with State of the Art

In this study, we conduct a comprehensive evaluation of Gaze-Swin's performance relative to state-of-the-art methodologies on three prominent datasets, namely Gaze360 [34], RT-Gene [26], and EyeDiap [35]. As detailed in Table 1, the top six comparative models are primarily CNN- or RNN-based, pre-trained on ImageNet [36]. The GazeTR [24] leverages the Visual Transformer as its primary structure, while the GazeNAS-ETH [31] employs Neural Architecture Search (NAS) methodologies to extract feature maps and Transformer is utilized to learn gaze direction from these feature maps. The synthesized findings affirm the remarkable efficacy of Gaze-Swin as it attains the lowest overall angular error across the evaluated datasets. Specifically, in comparison to Gaze-NAS, Gaze-Swin exhibits a noteworthy enhancement of 3.6% on Gaze360, a modest yet discernible improvement of 0.4% on RT-Gene, and a substantiated advancement of 2% on EyeDiap. The superior per-

formance of Gaze-Swin on the Gaze360 dataset can be attributed to the inherent challenges associated with significant head pose variability within the dataset. This particular domain, characterized by diverse head poses, aligns with the strengths of the Transformer architecture integral to Gaze-Swin. This observation underscores the adaptability and robustness of Gaze-Swin in scenarios characterized by diverse head poses.

Table 1. Comparison with state-of-the-art methods. The top six are CNN- or RNN-based gaze estimation models that are pre-trained on ImageNet [36]. The remaining models use the Transformer as the backbone and pre-train on ETH-Xgaze. (Bold numbers denote the best performance in each dataset. This notation is consistent throughout all tables).

Method	Gaze360	RT-Gene	EyeDiap
FullFace [29]	14.99°	10.00°	6.53°
RT-Gene [26]	12.26°	8.60°	6.02°
Dilated-Net [37]	13.73°	8.38°	6.19°
CA-Net [38]	11.20°	8.27°	5.27°
RCNN [39]	11.23°	10.30°	5.31°
Gaze360 [34]	11.04°	7.06°	5.36°
CADSE [40]	10.70°	7.00°	5.25°
GazeTR-Pure [24]	13.58°	8.06°	5.72°
GazeTR-Hybrid [24]	10.62°	6.55°	5.17°
GazeNAS-ETH [31]	10.52°	6.40°	5.00°
Gaze-Swin	**10.14°**	**6.38°**	**4.90°**

4.4. Hyper-Parameters in Gaze-Swin

In this section, we delve into the analysis of four crucial hyper-parameters in Gaze-Swin, presenting our findings in Table 2. The impact of varying Transformer layers L_1, input channels C, heads N, and CNN layers L_2 on the model's performance is systematically explored.

Table 2. Hyper-parameter analysis in Gaze-Swin.

Hyper-Parameter		Gaze360	RT-Gene	EyeDiap
Transformer layers L_1	3	10.22°	6.69°	5.16°
	6	10.14°	**6.38°**	**4.90°**
	12	**10.04°**	6.83°	5.27°
Input channel C	48	10.17°	6.87°	5.46°
	96	**10.14°**	**6.38°**	**4.90°**
	192	10.29°	6.63°	5.14°
Heads N	6	10.14°	**6.38°**	**4.90°**
	12	10.16°	6.67°	5.12°
	24	**10.05°**	7.01°	5.30°
CNN layers L_2	18	10.14°	**6.38°**	**4.90°**
	34	10.04°	6.73°	4.98°
	50	**10.01°**	6.97°	5.06°

4.4.1. Transformer Layers L_1

Performance deterioration is observed after the threshold of $L_1 = 6$. An exception to this trend is noted in the Gaze360 dataset, where the best performance is achieved at $L_1 = 12$. In contrast, the remaining datasets demonstrate a decline in performance when exceeding $L_1 = 6$.

4.4.2. Input Channel C

The examination of input channels, denoted as C, discloses that optimal performance across all datasets is achieved when $C = 96$. This configuration signifies an equilib-

rium, wherein the model capitalizes on an optimal tradeoff between complexity and information representation.

4.4.3. Heads N

The influence of heads N on performance is distinct, with only the Gaze360 dataset exhibiting performance improvements with incremental increments. In contrast, the remaining two datasets experience a discernible decrement in performance as the number of heads increases.

4.4.4. CNN Layers L_2

The analysis of CNN layers L_2 elucidates a nuanced relationship with performance. For a comprehensive evaluation, we employ ResNet18, ResNet34, and ResNet50. While performance on the Gaze360 dataset improves incrementally with the increasing complexity of CNN layers, our observations indicate a decrement in performance on the other two datasets beyond a certain layer depth. This suggests that, while larger networks like ResNet34 and ResNet50 may offer benefits, the efficiency–performance balance offered by ResNet18 makes it a preferable choice for our application.

4.5. Dropkey Parameter Analysis and Comparison with Dropout

We change the drop ratio d and linear decay α to analyze the parameter setting of Dropkey, and we compare Dropkey with Dropout to prove the effectiveness of Dropkey. The datasets mentioned above are employed for executing these experiments. The results are presented in Table 3, demonstrating that both Dropkey and Dropout techniques positively contribute to improving Gaze-Swin across all three datasets. Notably, the impact of Dropkey is more pronounced. Additionally, linear decay for drop ratio is useful for Gaze-Swin.

Table 3. Parameter analysis of Dropkey and comparison with Dropout. d means Dropkey ratio and α means linear decay.

Method	Hyper-Parameter		Gaze360	RT-Gene	EyeDiap
	d	α			
Dropkey	0.1	0	10.19°	6.52°	5.01°
		0.005	10.14°	**6.38°**	4.90°
		0.01	**10.12°**	6.50°	5.06°
	0.2	0	10.18°	6.49°	4.99°
		0.01	10.18°	6.42°	**4.89°**
		0.02	10.15°	6.51°	4.97°
	0.3	0	10.36°	6.57°	5.28°
		0.01	10.31°	6.48°	5.23°
		0.02	10.27°	6.51°	5.08°
Dropout	0	0	10.26°	6.49°	5.13°
	0.1	0	10.20°	6.51°	5.06°
	0.2	0	10.22°	6.43°	5.11°
	0.3	0	10.41°	6.54°	5.44°

4.6. Ablation Study

In this section, we perform an ablation study by removing two essential design elements from our model: (1) ablating the CNN branch and (2) ablating the Transformer branch.

4.6.1. Gaze-Swin Without CNN Branch

The CNN branch is designed to capture local features and intricate texture detail. To evaluate the impact of the CNN branch on our model, we remove the CNN branch from Gaze-Swin. As previously stated, we conducted comparative analyses on three datasets. The results are presented in Table 4. They demonstrate that the CNN branch is necessary for Gaze-Swin.

Table 4. Ablation study. Both the CNN branch and the Transformer branch are necessary for Gaze-Swin.

Method	Gaze360	RT-Gene	EyeDiap
Gaze-Swin	**10.14°**	**6.38°**	**4.90°**
Without CNN branch	10.40°	6.61°	5.19°
Without Transformer branch	10.77°	6.64°	5.23°

4.6.2. Gaze-Swin without Transformer Branch

The Transformer branch is designed to handle long-range dependencies to extract global features. To investigate the impact of the Transformer branch, we delete the Transformer branch while keeping other components unchanged. We compare three datasets and present the results in Table 4. Our results demonstrate the importance of the Transformer branch.

4.7. Comparison with Visual Transformer

As aforementioned, ViT methods encounter difficulties in extracting comprehensive information from facial images due to the limited connectivity between adjacent windows. To address this limitation, we employ a 14 $\times$ 14 patch partitioning strategy for facial image analysis. In configuring the ViT architecture, we use 64 heads in Multihead Self-Attention (MSA), 12 Transformer layers, and a hidden size of 4096 for the two-layer MLP. The outcomes, as delineated in Table 5, demonstrate the superior performance of our Gaze-Swin in comparison to ViT. This superiority is evident not only in accuracy but also in terms of model efficiency, as reflected in the parameters (Params) and floating point operations (FLOPs). Moreover, we categorize samples based on their gaze directions and divide them into clusters of 1° $\times$ 1° regions. We calculate the average improvement in each cluster, as depicted in Figure 6. The brighter points on the graph indicate greater improvement. Notably, the edge regions are brighter, demonstrating that the most significant improvement occurred for extreme gaze directions.

Table 5. Comparison with Visual Transformer. Gaze-Swin surpasses ViT in terms of performance metrics. Gaze-ViT is formally characterized as the transformation of the Swin Transformer into a Visual Transformer within the Gaze-Swin framework, as outlined in this table.

Method	Params	FLOPs	Gaze360	RT-Gene	EyeDiap
Gaze-Swin	**32.28 M**	**5.17 G**	**10.14°**	**6.38°**	**4.90°**
Visual Transformer	227.19 M	44.84 G	13.58°	8.06°	5.72°
Gaze-ViT	238.37 M	46.66 G	10.36°	6.60°	5.06°

Moreover, we transition the Swin Transformer to a Visual Transformer, thereby redefining it as Gaze-ViT, to substantiate the efficacy of hierarchical Transformers in the gaze estimation task. The outcomes of this transformation are elucidated in Table 5. These results serve as evidence that multiscale Transformer layers effectively address challenges arising from patch division, mitigating potential disruptions to facial and ocular image integrity in the gaze estimation task.

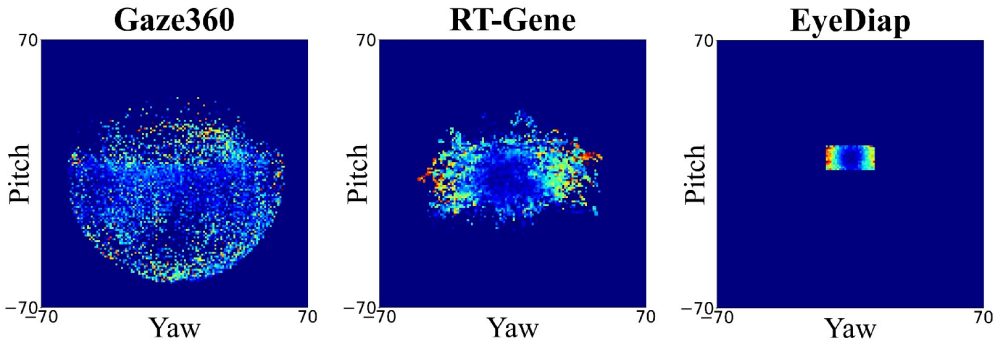

Figure 6. The performance improvement of Gaze-Swin in contrast to ViT. The x-axis and y-axis in each figure mean the yaw and pitch of the gaze direction. Brighter points indicate a larger improvement.

4.8. Impact of Pre-Training

Pre-training, a methodology entailing the initial training of a Transformer model on an extensive dataset followed by subsequent fine-tuning on a task-specific dataset, is designed to augment model efficacy and curtail training duration. In the pursuit of assessing the influence of pre-training on our Gaze-Swin model, we conduct comprehensive experiments. Our evaluations involve a comparative analysis of the model's performance when subjected to both pre-training and non-pre-training scenarios, leveraging the ETH-XGaze dataset. A comparative benchmark, GazeTR-Hybrid [24], served as the reference point. The empirical findings, presented in Figure 7, underscore the nuanced impact of pre-training on gaze estimation.

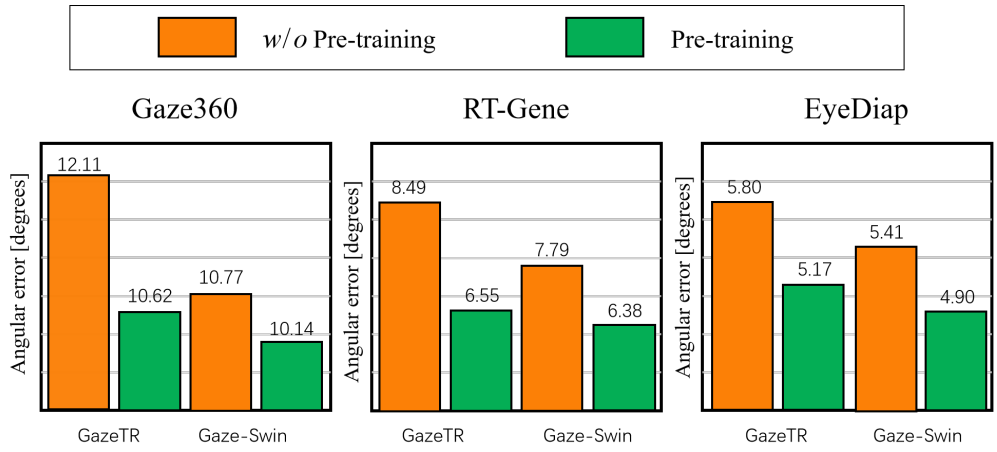

Figure 7. The impact of pre-training on Gaze-Swin's performance is illustrated by comparing its pre-trained and non-pre-trained versions against both conditions of GazeTR-Hybrid. The results confirm that Gaze-Swin consistently surpasses GazeTR-Hybrid, regardless of pre-training status.

Our findings demonstrate that the Gaze-Swin model consistently outperforms GazeTR, irrespective of pre-training. This consistent edge reveals our model's strong generalization capability in gaze estimation. Remarkably, Gaze-Swin excels in both pre-trained and non-pre-trained conditions, underscoring its adaptability and transfer learning capabilities. These qualities not only affirm its robustness but also its flexibility for various tasks, particularly those involving diverse and smaller datasets. This adaptability enhances its scalability and practicality for real-world eye-tracking applications, suggesting its potential for wider applications where data variability is a significant factor.

5. Conclusions

In this work, we introduce Gaze-Swin, an innovative end-to-end gaze estimation model that utilizes a synergistic dual CNN-Transformer architecture. This model is distinguished by its Dropkey Assisted Attention mechanism within the transformer component,

strategically employing drop units for keys to effectively combat overfitting. Rigorous testing across three benchmark datasets has underscored Gaze-Swin's exceptional performance, outperforming existing models. The integration of a dual architecture coupled with advanced anti-overfitting attention techniques facilitates robust and reliable feature learning. This enhances the practical application and real-world effectiveness of gaze estimation technology, marking a significant advancement in the field.

Author Contributions: Conceptualization, R.Z.; methodology, R.Z. and S.L.; software, Y.W. and R.Z.; writing—original draft preparation, R.Z.; writing—review and editing, P.T. and S.L.; visualization, S.S.; supervision, P.T.; funding acquisition, P.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Zhejiang Province, China, under Grant LQ22H090011 and Grant LQ22F020020.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372. [CrossRef] [PubMed]
2. Jacob, R.J.; Karn, K.S. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The Mind's Eye*; Elsevier: Amsterdam, The Netherlands, 2003; pp. 573–605.
3. Mutlu, B.; Shiwa, T.; Kanda, T.; Ishiguro, H.; Hagita, N. Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, La Jolla, CA, USA, 9–13 March 2009; pp. 61–68.
4. Morimoto, C.H.; Mimica, M.R. Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Underst.* **2005**, *98*, 4–24. [CrossRef]
5. Patney, A.; Kim, J.; Salvi, M.; Kaplanyan, A.; Wyman, C.; Benty, N.; Lefohn, A.; Luebke, D. Perceptually-based foveated virtual reality. In Proceedings of the SIGGRAPH '16: ACM SIGGRAPH 2016 Emerging Technologies, Anaheim, CA, USA, 24–28 July 2016; pp. 1–2.
6. Demiris, Y. Prediction of intent in robotics and multi-agent systems. *Cogn. Process.* **2007**, *8*, 151–158. [CrossRef] [PubMed]
7. Park, H.S.; Jain, E.; Sheikh, Y. Predicting primary gaze behavior using social saliency fields. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 3503–3510.
8. Yoo, D.H.; Chung, M.J. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Comput. Vis. Image Underst.* **2005**, *98*, 25–51. [CrossRef]
9. Zhu, Z.; Ji, Q. Eye gaze tracking under natural head movements. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 918–923.
10. Zhu, Z.; Ji, Q.; Bennett, K.P. Nonlinear eye gaze mapping function estimation via support vector regression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 1, pp. 1132–1135.
11. Hennessey, C.; Noureddin, B.; Lawrence, P. A single camera eye-gaze tracking system with free head motion. In Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, San Diego, CA, USA, 27–29 March 2006; pp. 87–94.
12. Ishikawa, T.; Baker, S.; Matthews, I.; Kanade, T. Passive Driver Gaze Tracking with Active Appearance Models. In Proceedings of the 11th World Congress on Intelligent Transportation Systems, Nagoya, Japan, 18–22 October 2004.
13. Chen, J.; Ji, Q. 3D gaze estimation with a single camera without IR illumination. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–4.
14. Valenti, R.; Sebe, N.; Gevers, T. Combining head pose and eye location information for gaze estimation. *IEEE Trans. Image Process.* **2011**, *21*, 802–815. [CrossRef] [PubMed]
15. Hansen, D.W.; Pece, A.E. Eye tracking in the wild. *Comput. Vis. Image Underst.* **2005**, *98*, 155–181. [CrossRef]
16. Huang, M.X.; Li, J.; Ngai, G.; Leong, H.V. Screenglint: Practical, in-situ gaze estimation on smartphones. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 2546–2557.
17. Ansari, M.F.; Kasprowski, P.; Obetkal, M. Gaze tracking using an unmodified web camera and convolutional neural network. *Appl. Sci.* **2021**, *11*, 9068. [CrossRef]
18. Li, Y.; Huang, L.; Chen, J.; Wang, X.; Tan, B. Appearance-Based Gaze Estimation Method Using Static Transformer Temporal Differential Network. *Mathematics* **2023**, *11*, 686. [CrossRef]

19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
22. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.
23. Li, B.; Hu, Y.; Nie, X.; Han, C.; Jiang, X.; Guo, T.; Liu, L. DropKey for Vision Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22700–22709.
24. 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: Piscataway, NJ, USA, 2022; pp. 3341–3347.
25. 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.
26. Fischer, T.; Chang, H.J.; Demiris, Y. Rt-gene: Real-time eye gaze estimation in natural environments. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 334–352.
27. 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.
28. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye tracking for everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2176–2184.
29. 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.
30. Cheng, Y.; Lu, F. DVGaze: Dual-View Gaze Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 20632–20641.
31. Nagpure, V.; Okuma, K. Searching Efficient Neural Architecture with Multi-resolution Fusion Transformer for Appearance-based Gaze Estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 890–899.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. 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 Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part V 16; Springer: Cham, Switzerland, 2020; pp. 365–381.
34. 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.
35. 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.
36. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
37. Chen, Z.; Shi, B.E. Appearance-based gaze estimation using dilated-convolutions. In Proceedings of the Asian Conference on Computer Vision, Perth, WA, Australia, 2–6 December 2018; Springer: Cham, Switzerland, 2018; pp. 309–324.
38. 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.
39. Palmero, C.; Selva, J.; Bagheri, M.; Escalera, S. Recurrent cnn for 3d gaze estimation using appearance and shape cues. *arXiv* **2018**, arXiv:1805.03064.
40. Oh, J.O.; Chang, H.J.; Choi, S.I. Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4992–5000.

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.