AG-HybridNet: An Attention-Guided Hybrid CNN-Transformer Network for 3D Gaze Estimation

Yue Li
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 327541575@qq.com

Changyuan Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: Cyw901@163.com

***Abstract*—To address the challenge of accurate gaze estimation in unconstrained environments susceptible to various interfering factors, this paper proposes AG-HybridNet, an end-to-end gaze estimation model integrating a dual-branch architecture combining CNN and Transformer components. The model employs Swin Transformer as the backbone for global feature extraction while incorporating an enhanced CNN branch dedicated to local feature capture. We introduce the TDConv-Block, which replaces standard convolution with partial convolution integrated with reparameterization technique, significantly reducing computational load and memory access while forming a T-shaped receptive field focused on central facial regions. Additionally, we design Efficient Additive Attention (ED-Attention) that effectively resolves the computational bottleneck in long-sequence processing for Transformers by reconstructing the computational workflow. Comprehensive experiments on MPIIFaceGaze and Gaze360 datasets validate the model's effectiveness. Experimental results demonstrate that AG-HybridNet achieves mean angular errors of 3.72° and 10.82° on MPIIFaceGaze and Gaze360 datasets respectively. Comparative studies with other mainstream 3D gaze estimation methods confirm that our network model can accurately estimate 3D gaze directions while reducing computational complexity.**

Keywords-Swin Transformer; Hybrid CNN-transformer; 3D Gaze Estimation; Efficient Additive Attention (ED-Attention); TDConv-Block

I. INTRODUCTION

The core objective of 3D gaze estimation, a critical research area within computer vision, is to reliably estimate gaze direction vectors in three-dimensional space. This inference is typically achieved by analyzing two-dimensional image or video sequences captured by standard cameras [1]. By establishing mapping models from visual appearance features to three-dimensional fixation points, this technology demonstrates significant application value across multiple frontier domains including medical diagnostics, psychological research, intelligent driving, and virtual reality. To address this research direction, previous studies [2-5] have proposed various innovative solutions. Based on differences in feature extraction and modeling methodologies, existing approaches can be systematically categorized into two main technical pathways: model-based geometric analysis frameworks and data-driven appearance-based learning frameworks. Model-based approaches [6-8] rely on predefined eye models and image feature points to deduce the 3D gaze direction, typically requiring specialized equipment to capture specific eye information. Appearance-based methods [9-12] have attracted substantial attention. By capitalizing on standard cameras to capture peri-ocular or full-face regions, these techniques learn a direct appearance-to-gaze mapping, demonstrating considerable efficacy in practical applications.

With the widespread adoption of deep learning across various domains, researchers have progressively introduced deep learning methodologies into the field of gaze estimation. Concurrently, the public release of large-scale gaze estimation datasets [13] has significantly propelled the development of numerous appearance-based deep neural network models. In this context, deep learning models built upon CNN and Transformer backbones have become increasingly prevalent, representing the foremost architectural paradigms. However, CNN-based methods [14] inevitably encounter issues such as

DOI: 10.2478/ijanmc-2025-0038

increased network depth and elevated model complexity when pursuing higher accuracy. Furthermore, their limited global modeling capability constrains further advancements in gaze estimation performance. In contrast, Transformer-based approaches, leveraging the self-attention mechanism, demonstrate superior capacity for capturing long-range dependencies within images. Nevertheless, the computational and memory requirements of the self-attention mechanism grow quadratically with spatial dimensions, resulting in substantial computational overhead. The requirements for gaze estimation models in practical applications include high predictive accuracy and simplified architectural design. Hence, investigating approaches to enhance estimation precision while reducing model complexity remains a prominent research direction.

To mitigate the aforementioned problems, the improved network proposed in this paper comprises dual branches: a CNN branch that takes facial images as input, and a Transformer branch that takes randomly initialized learnable tokens as input. The primary contributions of this work can be summarized as follows:

- We employ a dual-branch CNN-Transformer architecture to construct a hierarchical feature representation. This design strategically leverages the CNN branch's proficiency in encoding high-resolution local patterns alongside the Transformer branch's capacity for capturing global spatial relationships. Furthermore, Sigmoid gating mechanisms are embedded to adaptively modulate and integrate feature flows from both pathways. This design ensures that the network prioritizes the most salient features relevant to gaze estimation.
- Building upon the dual-branch CNN and Transformer architecture, we introduce the TDConv-Block and the linear attention mechanism ED-Attention to replace the standard convolution in the CNN branch and the multi-head self-attention mechanism in the Swin Transformer branch, respectively. This enables the network to effectively reduce floating-point operations and enhance feature extraction capability.
- Experimental results on MPIIFaceGaze and Gaze360 show that AG-HybridNet achieves a superior balance between estimation accuracy and computational efficiency when compared to leading 3D gaze estimation methods.

II. Related Work

In the domain of 3D gaze estimation, deep learning-based techniques have assumed a leading position [15]. The year 2015 witnessed a seminal introduction of CNNs into 3D gaze estimation by Zhang et al. [16], who employed a shallow LeNet-derived architecture. Their network processed monocular eye images and integrated auxiliary pose information through feature fusion at the fully-connected layers. While this approach demonstrated superior performance over most traditional appearance-based gaze estimation methods, it required extensive image preprocessing, including separate eye region detection following face localization. Subsequent research witnessed the adoption of various classical convolutional neural networks for 3D gaze estimation. Motivated by the distinct roles of left and right eyes in gaze estimation, Cheng et al. [17] introduced an Asymmetric Regression-Evaluation Network. Their design selectively weights predictions from each ocular input through dedicated performance assessment, achieving accuracy gains in a performance-aware manner. The year 2016 saw Krafka et al. [18] present the iTracker CNN for mobile gaze estimation on Apple devices, along with the accompanying Gaze Capture dataset. This collection quickly became a foundational resource, providing critical infrastructure for subsequent investigations. To address the challenge of the resolution compromise in standard CNNs, Chen et al. [19] developed Dilated-Net in 2019. By utilizing dilated convolutions, the network effectively extracts multi-scale contextual information without sacrificing spatial detail, which consequently refines 3D gaze prediction. However, as investigations deepened, researchers recognized inherent limitations in CNN-based

methods: accuracy improvements inevitably necessitated increased network depth and model complexity, while limited global modeling capabilities constrained further performance enhancements [20].

The Transformer architecture, first presented by Google researchers in 2017, fundamentally reshaped the landscape of natural language processing through its state-of-the-art performance, and was subsequently adapted to computer vision applications [21-23]. A distinctive strength of the Transformer lies in its self-attention mechanism, which captures long-range dependencies through attentional weighting. This capability empowers the model to build comprehensive representations essential for solving complex problems. Consequently, Cheng et al. [24] pioneered GazeTR in 2022 as the first Transformer-based framework for gaze estimation. The architecture schematized in Figure 1 processes visual input through convolutional stages to produce spatially-enriched representations. Transformer modules then decode these refined features by modeling cross-regional dependencies to predict gaze direction.

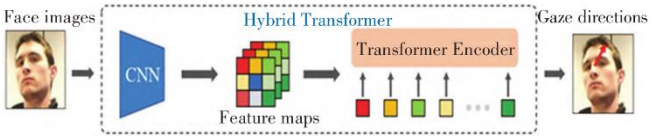

Figure 1. The architecture of GazeTR

Subsequently, addressing the issue of local detail loss potentially caused by patch partitioning in Swin Transformer for visual tasks, Li et al. [25] proposed an innovative convolutional alternative structure. This architecture replaces the original patch splitting and linear embedding mechanisms with hierarchical convolutional layers, enabling the model to capture features from local to global contexts across multiple scales, akin to traditional CNNs, thereby equipping the Transformer with multi-scale feature learning capabilities. This enhancement not only preserves the inherent long-range dependency modeling advantages of the Transformer but also significantly strengthens its perception of local details and spatial structures, demonstrating superior performance in gaze estimation tasks requiring precise localization. In parallel, to address the performance degradation of gaze estimation models in cross-domain scenarios, Cai et al. [26] proposed a framework termed UnReGA. The core concept of this framework lies in its dual uncertainty reduction strategy to enhance the model's domain adaptation capability. Specifically, it mitigates data uncertainty through data augmentation techniques while reducing model uncertainty via consistency regularization. The framework was validated on several cross-domain gaze estimation benchmarks, with results indicating its effectiveness in improving the model's generalization performance [27].

III. The Proposed Methods

A. Overall Architecture of AG-HybridNet

In unconstrained environments, the practical deployment of appearance-based gaze estimation is significantly complicated by the compound effects of continuous head movement, subject-specific anatomical characteristics, and highly variable imaging conditions. These factors substantially influence facial appearance, resulting in images with complex and confounded features. To address these challenges, we propose AG-HybridNet, whose detailed architecture is depicted in Figure 2.

Built upon Swin Transformer [28] as the core, our dual-stream framework implements complementary processing pipelines: a CNN-based limb dedicated to facial region analysis works concurrently with a Transformer limb that operates on tokenized inputs to capture structural relationships. The architecture features a clear division of labor: a branch dedicated to convolutional processing focuses on discerning fine-grained local particulars from specific facial regions, while the other branch, centered on self-attention, focuses on inferring the broader contextual relationships that define a global visual pattern. These local features participate in global feature fusion through two distinct pathways:

- Local features are directly combined with global features from the global feature extraction branch through residual addition, preserving the original details of local

features and providing a detailed foundation for fusion.

- In the fusion process, the local features are first normalized by a sigmoid function to serve as attention coefficients. The global feature set is then calibrated through a point-wise multiplication with these coefficients. This process enables local details to guide the attention allocation of global features, followed by residual addition, allowing global features to focus more on regions strongly correlated with local details.

To further optimize the feature fusion process, we incorporate Sigmoid-based gating units at the output stages of both branches. These gating units employ learnable weight parameters to adaptively calibrate and filter feature responses from different branches, thereby enhancing the preservation of task-critical information while suppressing irrelevant feature interference. Specifically, the gating mechanism performs element-wise weighting operations to achieve differentiated processing of spatial positions and channels within feature maps, consequently improving the discriminative power of feature representations.

Through this design, the locally enhanced features from the CNN branch and the globally contextualized features from the Transformer branch are concatenated at the feature level after gating calibration, forming a comprehensive gaze representation that incorporates multi-level information. The learning objective is completed by passing this integrated feature set through a dedicated fully-connected network, producing the numerical outputs corresponding to the yaw and pitch angles. The entire architecture maintains computational efficiency while ensuring the richness and robustness of feature representations.

B. CNN-Based Local Feature Extractor

As depicted in Figure 3, a ResNet-18 architecture serves as the computational foundation for the local feature branch, which is specialized for deriving regional characteristics from input facial imagery. The input to this branch is propagated through a computational graph of four stacked Basic Blocks to progressively transform the facial representations. Each module contains 3×3 convolutional kernels and ReLU activation functions, enabling hierarchical local feature learning. Crucially, each residual block incorporates skip connections between its two convolutional layers, effectively mitigating gradient vanishing and explosion issues, thereby facilitating the learning of complex hierarchical representations. This residual design not only stabilizes the optimization process but also promotes the learning of more sophisticated feature hierarchies. The architecture ensures that the convolutional stages are immediately succeeded by a global average pooling operation, which transforms the expansive feature maps into a singular 512-dimensional representation.

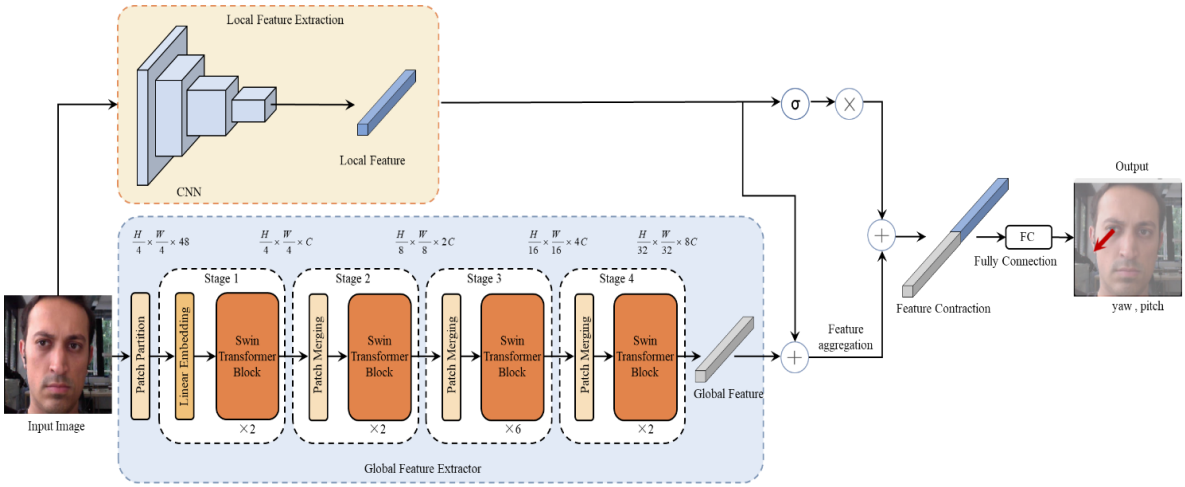

Figure 2. The detailed architecture of AG-HybridNet.

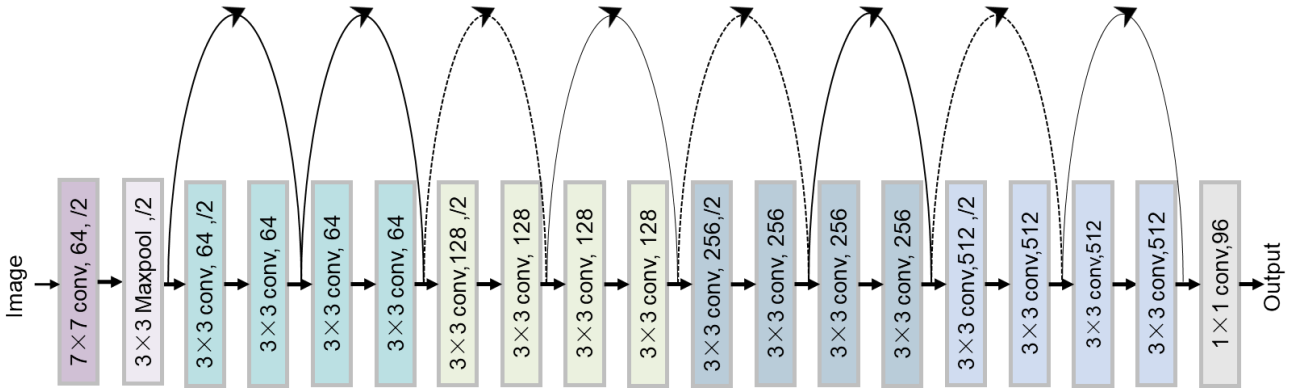

Figure 3. The detailed architecture of the CNN branch.

Achieving accurate 3D gaze estimation primarily demands a faithful representation of critical ocular details—namely pupil position, iris boundaries, and eyelid morphology—coupled with the precise spatial configuration of local facial constructs including eyebrow alignment and nasal bridge topography. These critical features are predominantly concentrated in the central image regions, while features from background or non-critical areas contribute minimally to gaze direction prediction. Standard convolution operations, which perform undifferentiated computations across all input channels and spatial positions, inevitably introduce abundant redundant features from non-critical regions, thereby increasing noise interference.

To enhance feature extraction capabilities, this paper introduces a crucial modification to the fundamental convolutional block within the CNN branch by replacing certain standard convolutional modules with TDConv Blocks incorporating Reparametrized Partial Convolution (RPConv), thereby constructing a more efficient feature extraction unit. A detailed comparison of these three convolutional structures is presented in Figure 4. Specifically, we first replace traditional standard convolution with partial convolution followed by pointwise convolution, then optimize the partial convolution through reparameterization techniques to construct a feature extraction module that balances both efficiency and accuracy.

The core mechanism of partial convolution confines its processing to a contiguous subset of input channels. Our implementation specifically applies convolutional operations to 50% of these consecutive channels and leaves the remainder unmodified. This strategic design alone substantially curtails both redundant computations and memory access. Building upon this, RPConv further enhances performance through a reparameterization strategy of multi-branch learning during training and parameter merging during inference: during the training phase, RPConv incorporates parallel partial convolution branches alongside identity mapping branches, capturing richer local features through differentiated processing of different channel subsets. During inference, these branch parameters are merged into a single convolutional kernel, eliminating multi-branch redundant computations and thereby improving inference speed on embedded devices.

When RPConv is combined with subsequent pointwise convolution, the resulting effective receptive field on the feature map exhibits a more precise T-shaped structure. Figure 4 clearly demonstrates the receptive field differences compared to standard convolution and ordinary partial convolution: compared to ordinary partial convolution, RPConv T-shaped focus concentrates more intensively on central regions such as the ocular area and surrounding key regions, while simultaneously preserving perception capability for important features in non-central areas through the multi-branch information fused via reparameterization. This characteristic perfectly aligns with the functional positioning of the CNN branch in our dual-branch architecture for extracting fine-grained local features to support 3D gaze estimation, not only enhancing the

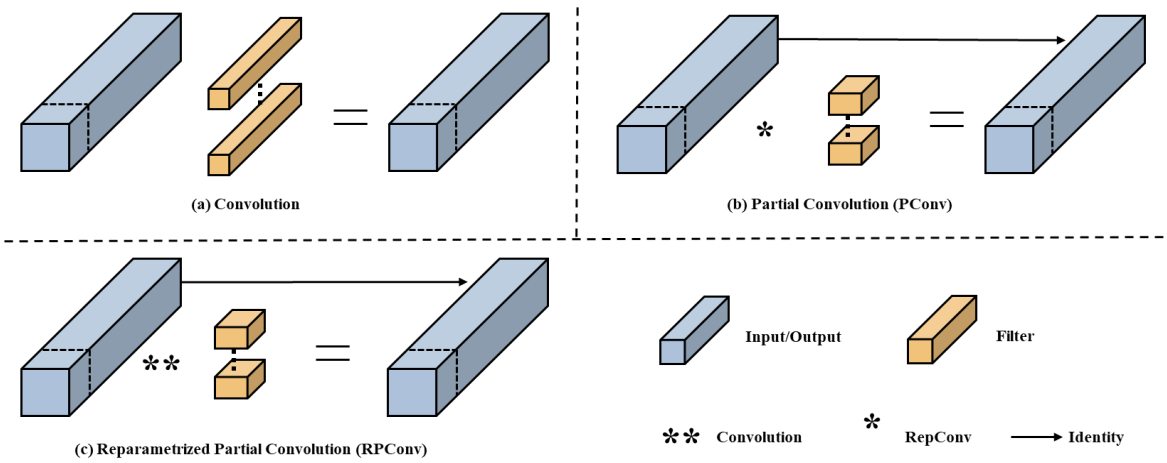

Figure 4. Schematic diagram of the Reparametrized Partial Convolution (RPConv) structure.

capture of core cues such as pupil position and iris boundaries but also improving robustness to occlusions and illumination variations through multi-branch fused features.

The output features of this branch undergo global average pooling and are flattened into a 1×1152-dimensional vector. After being filtered by a Sigmoid gating unit to select high-information-content features, they are concatenated with global contextual features from the Transformer branch. Finally, the combined features are projected to a 128-channel representation through fully-connected layers, providing comprehensive feature support for subsequent regression of 3D gaze direction (yaw and pitch angles).

Partial convolution is instrumental in building an inherently robust feature representation for partially occluded facial inputs, a frequent and challenging condition in practical gaze estimation applications. Unlike standard convolution operations that treat all input regions equally, partial convolution selectively updates only the visible regions of the feature map. Through this gated computation, the model sustains the robustness of its local feature extraction pipeline, thereby maintaining performance integrity when processing partially corrupted inputs.

Partial convolution is characterized by its selective processing mechanism, which optimizes computational resource utilization in neural networks by mitigating computational redundancy. This operation partitions input feature maps into multiple sub-regions and selectively performs convolutional operations on a subset of channels while preserving the remaining channels unchanged. The approach achieves a synergistic reduction in both the computational burden and the demand on memory access bandwidth. The computational complexity is primarily quantified by Floating Point Operations (FLOPs), which serves as a fundamental metric for evaluating operational efficiency. The FLOPs can be expressed as:

$$FLOPs_{Conv} = h \times w \times k^2 \times c^2 \tag{1}$$

Where c indicate the dimensionality of the input and output channels, respectively. k denotes the convolutional kernel size, h and w define the spatial dimensions of the feature maps.

Partial convolution reduces computational complexity by performing convolutional operations on only a subset of input channels. Given a total of c input channels with c_p channels selected for processing, the FLOPs can be expressed as:

$$FLOPs_{PConv} = h \times w \times k^2 \times \left(\frac{c_p^{\,2}}{c} \right) \tag{2}$$

When the ratio of processed channels to total channels $R = c_p / c = 1/4$, the computational cost of partial convolution reduces to merely one-quarter of standard convolution.

Furthermore, since partial convolution only requires reading data from a subset of channels, its memory access frequency is significantly lower than standard convolution. The memory access volume can be quantified as:

$$h \times w \times 2c_p \tag{3}$$

At the ratio of R = 1/4, the memory access volume decreases to one-quarter of standard convolution, demonstrating substantial improvement in memory efficiency.

As demonstrated in Figure 5, The enhanced CNN branch adopts a ResNet-18 backbone architecture, establishing a hierarchical feature extraction pipeline. Input facial images initially pass through a series of convolution-batch normalization layers to accomplish preliminary feature extraction and data normalization. Subsequently, the feature tensor sequentially undergoes processing through one basic residual block and multiple meticulously designed TDConv blocks for deep feature transformation, ultimately producing detail-rich local features for subsequent cross-branch feature fusion.

The core component of this architecture, the TDConv block, employs a multi-level residual connection structure that accomplishes feature refinement and enhancement through four consecutive stages. The initial phase utilizes standard 3×3 convolutional kernels for preliminary spatial feature extraction. The subsequent phase introduces an innovative reparametrized partial convolution mechanism, which incorporates a dual-path architecture combining identity mapping with partial convolution. This design enables parallel learning of complementary feature patterns during the training phase. Notably, the partial convolution path performs computations on only fifty percent of consecutive channels, significantly reducing computational redundancy. During inference, this dual-path structure is consolidated into a single efficient computational unit through parameter reorganization techniques. The third phase employs pointwise convolution to achieve inter-channel feature fusion and dimensional adjustment. The final phase reapplies 3×3 convolution for feature refinement. The entire processing flow effectively mitigates gradient vanishing issues through cross-layer residual connections while preserving multi-scale feature information.

This modular design not only inherits the representational capacity of traditional residual structures for complex hierarchical features but also enhances the detail-focused capability for key facial regions such as the ocular area through reparametrized partial convolution. The ultimate design provides more discriminative local representations for feature fusion with the Transformer branch, achieving an optimal balance between computational efficiency and feature representational capacity, thereby establishing a solid foundation for precise gaze estimation.

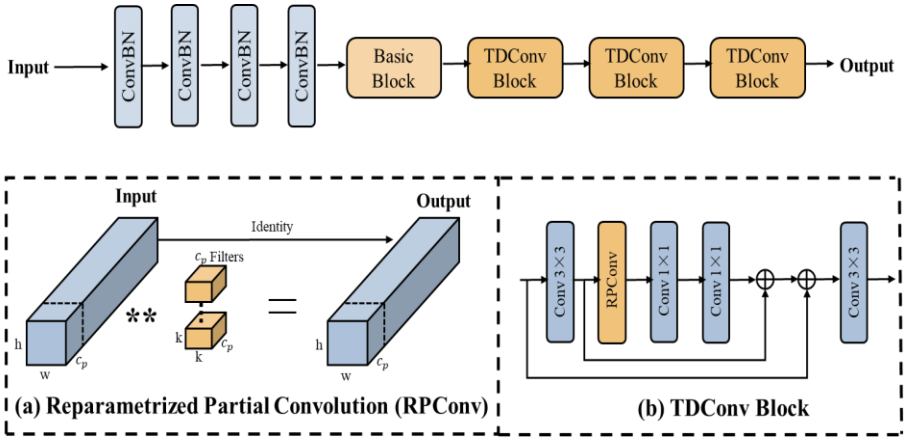

Figure 5. Network architecture diagram showing the RPConv and TDConv Block structure.

C. Improvement of feature extraction network

As demonstrated in Figure 6, the self-attention mechanism is employed within Transformer blocks to extract gaze-related features from images and capture long-range dependencies across image sequences. The operation of self-attention adheres to the following standard mathematical formulation:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Where Q, K and V represent query, key, and value matrices obtained through linear transformations of the input sequence, d_k denotes the dimensionality of the key vectors, and $\sqrt{d_k}$ is applied to the dot product to prevent gradient vanishing issues. The input sequence is first projected into three distinct representations through learned linear mappings. A compatibility matrix is then formed by computing the pairwise similarities between each query and all keys. Following this, the SoftMax function is applied to normalize these affinities into a set of probabilistic attention weights. The final, contextualized output is generated by performing a weighted aggregation of the value vectors based on these weights.

The conventional multi-head attention mechanism further replicates this process multiple times. Specifically, for a sequence of length n, each attention head requires pairwise comparisons between all sequence elements, resulting in the construction of an attention matrix of size n × n where each element represents the similarity between a pair of elements. This matrix exhibits both temporal and spatial complexity of O ($n^2 d_k$), consuming substantial memory resources and significantly prolonging computation time. As the sequence length increases, this quadratic growth rapidly becomes computationally intractable. To mitigate computational complexity in the Transformer branch while enhancing key feature modeling, we replace the standard self-attention mechanism with Efficient Dot-product Attention (ED-Attention). The computation of ED-attention can be formulated as:

$$ED - Attention(Q,K,V) = \frac{Q(K^TV)}{\sqrt{d_k}} \quad (5)$$

The algorithm of ED-Attention is implemented by adjusting the computational logic based on the standard self-attention. Its core lies in transforming the process of "first calculating Query-Key similarity and then aggregating Value" into "first aggregating the global information of Value and then interacting with Query", thereby reducing the complexity. This adjustment decreases the complexity from O ($n^2 d_k$) to O ($n d_k$). When d_k are much smaller than n (e.g., in 3D gaze estimation, the dimensionality of facial features is generally much smaller than the sequence length), the complexity is approximately linear with respect to the sequence length, significantly improving efficiency.

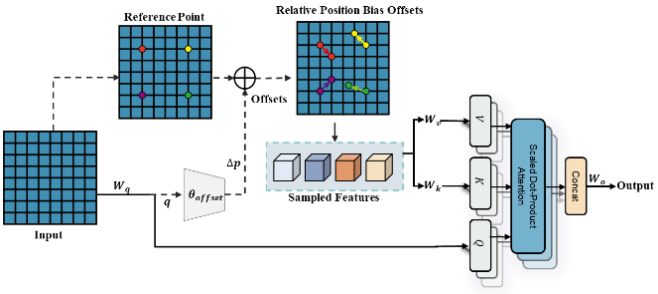

Figure 6. The detailed architecture of the attention mechanism.

Furthermore, the ED-Attention mechanism substantially reduces computational overhead by streamlining the normalization process, while simultaneously enhancing focus on critical features through an implicit feature selection scheme. In 3D gaze estimation tasks, this design effectively captures long-range dependencies between ocular regions and global facial pose configurations, significantly improving the precision of feature representation. Through its linear complexity architecture, the module optimally adapts to the stringent resource constraints of embedded systems, achieving an exceptional balance between global feature modeling capacity and computational efficiency.

Consequently, ED-Attention serves as the pivotal component for lightweight optimization in the Transformer branch, enabling robust performance in resource-constrained environments while maintaining competitive estimation accuracy.

IV. Experiments

A. Experimental Environment and Configuratio

The experiments were conducted on an Ubuntu server equipped with 60 GB RAM and an RTX 3080 GPU, using PyTorch 1.10 and Python 3.9.

On the MPIIFaceGaze dataset, the model was trained with a batch size of 64, an initial learning rate of 2e-4, and a weight decay parameter of 0.5. The learning rate scheduling commenced with a linear warm-up for the initial 2 epochs, followed by a decay phase initiating at the 15th epoch. For the Gaze360 dataset, the training configuration employed a batch size of 100 and a starting learning rate of 1e-4, which underwent a decay with a factor of 0.97 from the second epoch, continuing for a total of 80 epochs.

The AdamW optimizer was used to train the model on both datasets, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The loss function is calculated as follows:

$$L = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

The ResNet-18 network for the head pose branch utilizes publicly available pre-trained weights from the 300W-LP dataset. It was trained using the Adam optimizer with a learning rate of 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, for 25 epochs.

B. Experimental design

MPIIFaceGaze [29] and Gaze360 [30] are commonly used datasets for 3D gaze estimation methods. This paper normalizes the MPIIFaceGaze and Gaze360 datasets using the same approach as other gaze estimation methods. Specifically, we apply translation and rotation to a camera, removing head roll, and maintaining a fixed distance from the face center, then train and evaluate on these two datasets.

Figure 7 shows the loss convergence curves for subject p00 from the MPIIFaceGaze dataset and for training on the Gaze360 dataset.

The MPIIFaceGaze dataset contains data from 15 subjects, with each subject contributing 3,000 facial images. The evaluation follows a leave-one-subject-out cross-validation strategy. Specifically, the images from a single folder serve as the test set, with the images from all other folders comprising the training set. This procedure was repeated for every folder within the dataset, and the outcomes were averaged to form the final evaluation of the model's performance.

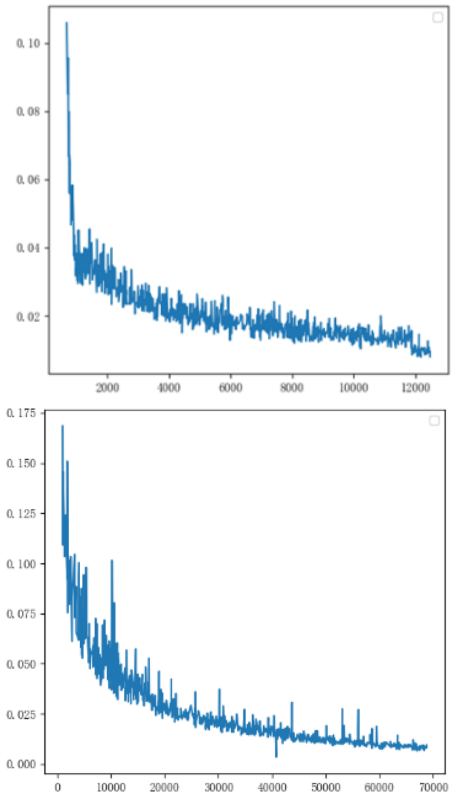

Figure 7. MPIIFaceGaze and Gaze360 Loss Convergence Curves.

The applicability of appearance-based gaze estimation on the Gaze360 dataset is inherently limited. This is primarily because portions of the dataset consist solely of occipital views, which do not contain the requisite facial information. The dataset partitioning for our experiments nevertheless followed the original training, validation, and test splits. The exclusion of images with unsuccessful face detection, consistent with the Phi-ai Lab methodology, was implemented to

guarantee the accuracy of the appearance-based gaze estimation framework.

The proposed AG-HybridNet directly yields estimates for the pitch and yaw angles. This result enables the immediate computation of the corresponding 3D gaze vector, which fully defines the direction of gaze. The calculation method is as follows:

$$\begin{aligned} x &= \cos(\phi) \cdot \cos(\theta) \\ y &= \cos(\phi) \cdot \sin(\theta) \\ z &= \sin(\phi) \end{aligned} \tag{7}$$

Where ϕ is the yaw angle and θ is the pitch angle, both in radians.

In the 3D gaze estimation task, performance is evaluated using the Mean Angular Error (MAE) in degrees. This metric is derived from the angular difference between the ground-truth and predicted gaze vectors, given by the formula:

$$L_{angular} = \cos^{-1}\left(\frac{g \cdot g'}{\|g\|\|g'\|}\right) \tag{8}$$

Here, g represents the ground-truth gaze direction, and g' corresponds to the predicted one.

C. Comparative Experiments and Analysis

The performance of the 3D gaze estimation model was assessed by benchmarking it against leading contemporary methods using the mean angular error.

Detailed per-subject results of the mean angular error are reported for the MPIIFaceGaze dataset. A comparative analysis is then conducted with Dilated-Net and GazeTR. Among the 15 subjects, compared to Dilated-Net, the proposed method achieves higher 3D gaze accuracy for 13 subjects, and compared to GazeTR, it achieves higher accuracy for 9 subjects. The results are demonstrated in Figure 8.

In our experiments, the proposed method was benchmarked on the MPIIFaceGaze dataset, achieving a mean angular error of 3.72 degrees. The models compared in Table 1 include pure CNN, pure Transformer, and hybrid models, demonstrating that the approach adopted in this study surpasses other methods in 3D gaze estimation accuracy.

TABLE I. COMPARATIVE EXPERIMENTAL RESULTS ON THE MPIIFACEGAZE DATASET

Model	Mean Angular Error (°)
MPIIGaze	5.40
Dilated-Net	4.80
CA-Net	4.10
AGE-Net	4.09
GazeTR	4.00
L2CS-Net	3.92
Res-Swin-Ge	3.75
Ours	3.72

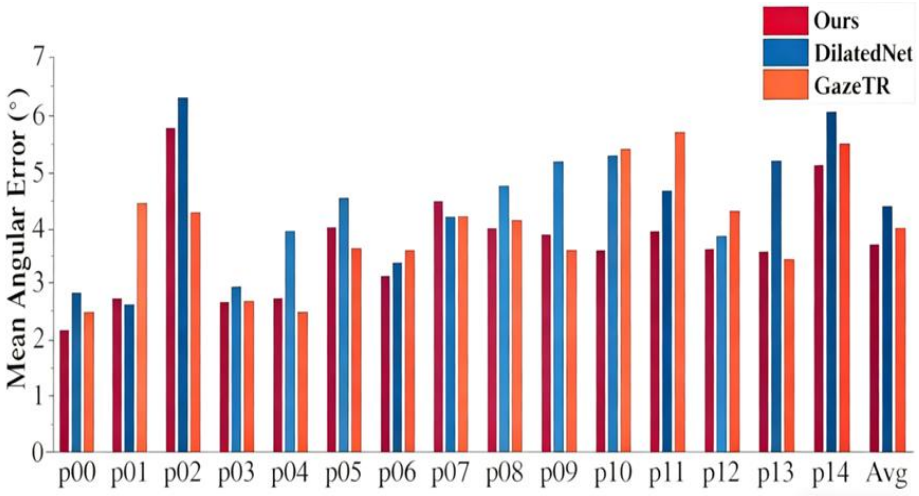

Figure 8. Comparison of Mean Angular Error on the MPIIFaceGaze Dataset.

Table 2 details the comparative results obtained from the Gaze360 test set. The proposed method achieves a mean angular error of 10.85 °within the front 180 °range, demonstrating its superiority over the existing gaze estimation approaches evaluated.

TABLE II. COMPARATIVE EXPERIMENTAL RESULTS ON THE GAZE360 DATASET

Model	Mean Angular Error (°)
Full-Face	14.99
Dilated-Net	13.73
RT-Gene	12.26
Gaze360	11.40
Bot2L-Net	11.53
Ours	10.82

The computational complexity of the proposed AG-HybridNet was assessed to evaluate its efficacy for 3D gaze estimation. Using the THOP PyTorch library, we conducted a comprehensive analysis of the model's complexity, evaluating both its parameter quantity and computational footprint (FLOPs). The detailed results are presented in Table 3.

TABLE III. COMPARISON OF PARAMETERS AND FLOPS FOR DIFFERENT MODELS

Model	Mean Angular Error (°)	Parameters	FLOPs
Dilated-Net	4.80	3.920	3.153
GazeTR	4.00	11.394	1.834
Ours	3.72	21.201	1.505

To assess the performance of the proposed model in terms of mean angular error and FLOPs, it was compared with Dilated-Net and GazeTR. The experimental results are presented in Figure 9. Data points closer to the bottom-left corner indicate fewer FLOPs and higher gaze estimation accuracy. The proposed method significantly outperforms both Dilated-Net and GazeTR in these two aspects. Although the number of parameters increased slightly, it is noteworthy that the proposed method was trained for only 18 epochs on the MPIIFaceGaze dataset, whereas Dilated-Net and GazeTR were trained for 100 and 80 epochs, respectively. This fully demonstrates that the proposed improvements enhance performance while maintaining strong fitting capability.

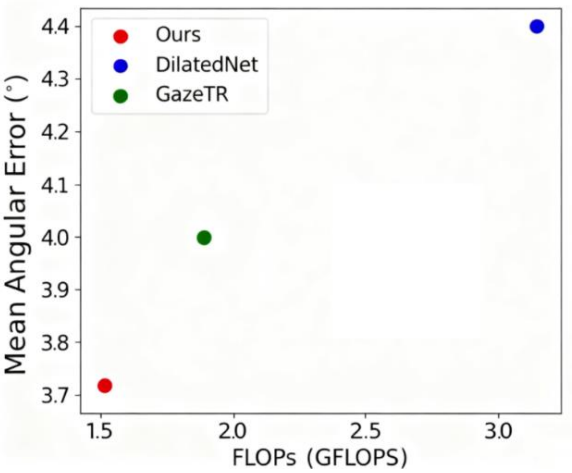

Figure 9. Comparison of Mean Angular Error and FLOPs

V. CONCLUSIONS

With the growing demand for accurate and efficient gaze estimation in practical applications, designing models that achieve an optimal balance between precision and computational efficiency has become increasingly critical. Inspired by the complementary strengths of Convolutional Neural Networks and Swin Transformers, this paper proposes AG-HybridNet—a novel dual-branch architecture that leverages parallel CNN-Transformer pathways and introduces lightweight convolutional modules along with an efficient dot-product attention mechanism to reduce computational complexity. The proposed model not only enhances spatial feature extraction and suppresses redundant information but also adaptively captures relationships among Region of Interest (ROI) features, thereby improving generalization capability and robustness. Experimental results demonstrate that our method significantly improves the accuracy of gaze estimation while effectively reducing computational complexity.

REFERENCES

[1] [Strazdas D, Hintz J, Al-Hamadi A. Robo-hud: Interaction concept for contactless operation of industrial cobotic systems [J]. Applied Sciences, 2021, 11(12): 5366.

[2] Zhang X C, Sugano Y, Fritz M, et al. It's written all over your face: Full-face appearance-based gaze estimation[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 51-60.

[3] [Kellnhofer P, Recasens A, Stent S, et al. Gaze360: Physically unconstrained gaze estimation in the wild[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6912-6921.

[4] Abdelrahman A A, Hempel T, Khalifa A, et al. L2CS-Net: Fine-grained gaze estimation in unconstrained environments[C]// Proceedings of the 8th International Conference on Frontiers of Signal Processing. Corfu: IEEE, 2023: 98-102.

[5] Liu J H, Chi J N, Hu W X, et al. 3D model-based gaze tracking via iris features with a single camera and a single light source[J]. IEEE Transactions on Human-Machine Systems, 2020, 51(2): 75-86.

[6] Liu S, Liu D P, Wu H Y. Gaze estimation with multi-scale channel and spatial attention[C]// Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition, 2020: 303-309.

[7] Liu S L, Li F, Zhang H, et al. DAB-DETR: Dynamic anchor boxes are better queries for DETR[J]. arXiv preprint arXiv:2201.12329, 2022.

[8] Liu Y F, Wang T C, Zhang X Y, et al. PETR: Position embedding transformation for multi-view 3D object detection[C]// European Conference on Computer Vision. Springer, 2022: 531-548.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000-6010.

[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]// International Conference on Learning Representations. 2021.

[11] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 568-578.

[12] WU H, XIAO B, CODella N, et al. CvT: Introducing convolutions to vision transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 22-31.

[13] CAO J, PANG Y, ANWER R M, et al. PSTR: End-to-end one-step person search with transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9448-9457.

[14] XUE F, WANG Q, GUO G. Transfer: Learning relation-aware facial expression representations with transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3581-3590.

[15] Onyemauche U C, Nkwo S M, Mbanusi C E, et al. Towards the Use of Eye Gaze Tracking Technology: Human Computer Interaction (HCI) Research[C]// AfriCHI 2021: 3rd African Human-Computer Interaction Conference. 2021.

[16] Zhang X C, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild[C]// Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2021: 4511-4520.

[17] Cheng Y H, Lu F, Zhang X C. Appearance-based gaze estimation via evaluation-guided asymmetric regression[C]// Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 100-115.

[18] Krafka K, Khosla A, Kellnhofer P, et al. Eye tracking for everyone[C]// Proceedings of the 2023 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2023: 2176-2184.

[19] Chen Z K, Shi B E. Appearance-based gaze estimation using dilated-convolutions[C]// Proceedings of the 14th Asian Conference on Computer Vision. Perth: Springer, 2019: 309-324.

[20] Zhang X C, Sugano Y, Fritz M, et al. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(1): 162-175.

[21] Cheng Y H, Huang S Y, Wang F, et al. A coarse-to-fine adaptive network for appearance-based gaze estimation[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 10623-10630.

[22] Murthy L R D, Biswas P. Appearance-based gaze estimation using attention and difference mechanism[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3143-3152.

[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]// Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

[24] Cheng Y H, Bao Y W, Lu F. PureGaze: Purifying gaze feature for generalizable gaze estimation[C]// Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 436-443.

[25] Li Y J, Chen J H, Ma J X, et al. Gaze estimation based on convolutional structure and sliding window-based attention mechanism[J]. Sensors, 2023, 23(13): 6226.

[26] Wang X H, Zhou J, Wang L, et al. BoT2L-Net: Appearance-based gaze estimation using bottleneck Transformer block and two identical losses in unconstrained environments[J]. Electronics, 2023, 12(7): 1704.

[27] Fischer T, Chang H J, Demiris Y. RT-GENE: Real-time eye gaze estimation in natural environments[C]// Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 334-352. Conference on Computer Vision. Seoul: IEEE, 2019. 6912–6921.

[28] Li, Y. J.; Chen, J. H.; Ma, J. X.; et al. Gaze Estimation Based on Convolutional Structure and Sliding Window-Based Attention Mechanism. Sensors 2023, 23(13), 6226.

[29] Zhang Xucong, Sugano Y, Fritz M, et al. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 41(1): 162-175.

[30] Kellnhofer P, Recasens A, Stent S, et al. Gaze360: Physically unconstrained gaze estimation in the wild[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6912-6921.