Three-Dimensional Line-of-Sight Estimation Based on RTACM-Net and Vision Transformer

Tingjuan Sang
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail: sangtingjuan@163.com

Wuqi Gao
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail: gaowuqi@126.com

Abstract—**Human beings rely primarily on vision to perceive and interact with the external world, with approximately 80% of sensory information input through the visual system. This visual dominance makes the question of "where an individual is looking" not only a key to understanding attention distribution and information processing mechanisms but also a critical factor in optimizing decision-making efficiency and learning outcomes. However, traditional methods for analyzing gaze-related behaviors—such as manual behavioral observation and self-reported evaluation—suffer from inherent limitations: be havioral observation relies on subjective judgment of observers, often missing subtle gaze shifts and failing to achieve real-time tracking; self-evaluation is prone to memory biases and social desirability effects, leading to deviations between reported and actual gaze patterns. These drawbacks highlight the need for a more objective and precise alternative.Gaze estimation, which infers an individual’s visual attention and behavioral intentions by recording and analyzing the spatial position, movement trajectory, and dynamic changes of the eyeball, emerges as an ideal solution. This technology is broadly categorized into model-based (relying on geometric eye models) and appearance-based (using facial/ocular image features) approaches, with appearance-based methods gaining traction due to their non-intrusiveness. Nevertheless, current appearance-based gaze estimation still faces two major challenges: (1) individual differences, such as variations in eye shape, pupil size, eyelid structure, and the presence of glasses, which disrupt consistent feature extraction; (2) environmental interference, including variable lighting, partial facial occlusion, and dynamic head poses, which reduce estimation accuracy. To address these issues, this paper proposes RTACM-Net, a novel gaze estimation network architecture that integrates the strengths of Vision Transformer (ViT) with a multi-scale feature fusion mechanism. Specifically, RTACM-Net employs a lightweight convolutional module to extract local fine-grained features of the ocular region, while leveraging ViT’s multi-head attention mechanism to capture global contextual relationships. This dual-branch design enables the network to balance local feature precision and global context awareness, thereby mitigating the impact of individual differences and environmental noise.Extensive experiments were conducted on two benchmark datasets: MPIIFaceGaze (a large-scale dataset focusing on indoor controlled environments with 21 subjects) and Gaze360 (a challenging dataset covering diverse outdoor/indoor scenes, variable lighting, and large head-pose variations with over 100 subjects). The results show that RTACM-Net : on MPIIFaceGaze, it achieves an average angular error (MAE) of 3.72° ; on Gaze360, it achieves an MAE of 10.46° , Gaze360-Net (11.40°) by 0.94° . These results demonstrate the robustness of RTACM-Net in handling variable individual characteristics and complex environmental conditions. Its practical potential extends to multiple fields: in augmented reality (AR), it can enable adaptive interface rendering; in autonomous driving, it supports dual-task monitorin; in human-robot interaction, it facilitates intuitive service triggering.**

Keywords-Line of Sight Estimation; RTACM-Net; Visi on Transformer; Visual Attention

I. INTRODUCTION

With the rapid popularization of intelligent devices (such as AR glasses, smart cars, and interactive robots), human-computer interaction (HCI) is moving toward "naturalization"—and efficiently understanding and analyzing human visual behavior has become a core research direction to bridge the gap between human intent and machine response. In daily life, vision is the most dominant sensory channel: about 80% of external information (such as text reading, environmental perception, and object recognition) is obtained through the visual system. Therefore,

DOI: 10.2478/ijanmc-2025-0032

exploring the question of "where an individual is looking" is not only academically valuable—it helps reveal the mechanisms of human attention distribution , information processing paths, and decision-making processes —but also has practical application potential across multiple fields: In AR/VR, it enables "gaze-controlled interaction"; in autonomous driving, it monitors driver attention; in assistive technology, it helps people with motor disabilities operate devices via gaze.

As a key task in computer vision, three-dimensional (3D) gaze estimation differs from 2D gaze estimation (which only infers direction within the image plane). It focuses on tracking ocular features in 2D images/videos, then reversely deduces the precise gaze orientation in real 3D physical space — this is critical for scenarios requiring spatial positioning. To solve 3D gaze estimation, researchers have proposed two main types of methods: geometric model-based and appearance-based. Geometric model-based methods rely on pre-established mathematical models of the human eye (incorporating parameters like corneal curvature and pupil-cornea distance) and use specialized equipmentto capture key geometric feature. While they offer high accuracy in controlled environments, their dependence on expensive devices limits deployment in daily scenarios. Appearance-based methods, by contrast, only need a common RGB camera to capture eye or facial images, then learn the mapping from facial appearance to gaze direction via data-driven approaches. They are more flexible and low-cost, and have achieved promising results, but still face major challenges: occlusion, illumination variations, and image noise. These issues make 3D gaze estimation still a difficult problem.

With the development of deep learning, many deep neural network models have been designed to address these challenges. Early models mainly used convolutional neural networks (CNNs), which excel at extracting local eye features but struggle to capture global contextual relationships. Recently, the combination of CNNs and Transformer architectures has opened new avenues: CNNs are used to extract fine-grained local features of the ocular region, while Transformers leverage self-attention to model long-range dependencies. This fusion effectively balances local precision and global context, providing a new way to improve 3D gaze estimation accuracy.

Zhang et al. [14] proposed a shallow network similar to LeNet, taking the image of a single eye as input and adding a head Angle vector to the output. The performance of this method surpasses many traditional appearance-based gaze estimation methods, but the steps in the image preprocessing part of this method are too complicated. Chen et al. proposed the Dilated convolutional network (Dilated Net), which uses dilated convolution to extract advanced features without reducing spatial resolution, thereby enhancing the accuracy of three-dimensional line of sight estimation. Krafka et al. constructed the dataset Gaze Capture for line of sight estimation. Although these methods are constantly improving the accuracy of gaze estimation, gaze estimation models are also becoming increasingly complex as a result.

II. METHOD

To achieve high-precision gaze estimation—addressing the bottleneck of single-branch models that struggle to balance local detail and global context—this paper designs an eye-tracking model based on deep learning networks: Res Telu Attention Convolution Mix Network (RTACM-Net). This module is specifically optimized for extracting local facial features, with a structured design that includes three core components: first, a 3-layer convolutional block (using 3×3 kernels with stride=1) to capture primary texture information of facial regions; second, embedded Telu activation functions to avoid gradient vanishing during deep-layer training, ensuring subtle feature signals (like faint pupil-edge transitions) are not lost; third, a spatial attention module (CBAM-based) that dynamically assigns higher weight to the ocular region (64×64 pixel cropped area, the core of gaze inference) by calculating channel and spatial importance. This targeted design focuses on capturing fine-grained details such as the irregular shape of the pupil under varying light, the texture contrast between iris and sclera, and subtle movements of the eyelid

— all of which are critical for accurate gaze direction estimation. Notably, when the face is partially obscured or lighting is suboptimal, the spatial attention module can still lock onto the ocular region, ensuring valid local features are extracted. The importance of this precise local feature extraction cannot be overstated: in real-world scenarios where faces are rarely perfectly aligned with the camera, local feature integrity directly determines whether the model can distinguish between small gaze shifts , and experiments show that omitting the RTACM-Net's attention module leads to a 1.8 ° increase in average angular error on the MPIIFaceGaze dataset.

On this basis, this paper also introduces Vision Transformer (ViT) to complement local feature learning with global context modeling. ViT is an advanced deep learning architecture that has shown remarkable performance in capturing long-range spatial dependencies in image data — an advantage traditional convolutional neural networks (CNNs) lack, as CNNs rely on small receptive fields and can only model local pattern correlations within limited ranges. For gaze estimation, this limitation means CNNs struggle to associate head pose changes with corresponding gaze direction adjustments, often leading to misjudgment. ViT addresses this by treating the entire 224×224 facial image as a sequence of 16 ×16 pixel patches, converting each patch into a 768-dimensional embedding vector, and processing these vectors through 12 encoder layers. The self-attention mechanism enables ViT to calculate the correlation between any two patches — for example, learning the spatial relationship between the outer corner of the eye and the tip of the nose or between the eyebrow position and iris orientation (to correct gaze bias caused by head tilting). In the context of gaze estimation, ViT is used to extract global features of human faces and gaze, including the overall facial contour, relative positions of facial landmarks, and even subtle correlations between head movement trends and gaze shifts. This capability is especially valuable in handling large variations in head pose and gaze direction, as it allows the model to adapt to different user orientations and camera perspectives without requiring strict face alignment.

In this paper, ViT and RTACM-Net are designed as two parallel branches that process the input facial image simultaneously — ensuring no time lag in feature extraction and avoiding information loss from sequential processing. Before entering the branches, the input image undergoes standard preprocessing: first, MTCNN is used for face detection to crop the 224×224 facial region from the original image; then, facial landmarks are detected to align the eyes horizontally, reducing interference from initial face misalignment. After preprocessing, the image is fed into both branches: the RTACM-Net branch first crops the 64×64 ocular region and extracts local features, outputting a 256-dimensional feature map that focuses on ocular details; the ViT branch processes the full 224×224 facial image, outputting a 512-dimensional global feature vector that contains contextual information like head pose and facial contour. For feature aggregation, the two feature outputs are first normalized to ensure consistent data distribution, then concatenated along the channel dimension to form a 768-dimensional unified feature vector — this concatenation method retains both the fine-grained spatial information from RTACM-Net and the long-range contextual information from ViT, avoiding the loss of either feature type that occurs with weighted summation or element-wise addition. The aggregated feature vector is then passed through a fully connected layer module: the first FC layer performs feature dimensionality reduction and semantic fusion, the second FC layer further refines the feature representation, and the final output layer predicts the gaze pitch angle, representing upward/downward gaze and yaw angle, representing left/right gaze — these two angles fully define the individual's gaze direction in three-dimensional physical space, with each angle's prediction error directly mapped to the model's core evaluation metric.

The combination of RTACM-Net and Vision Transformer creates a synergistic effect between local feature extraction and global context understanding, addressing two key pain points of existing gaze estimation models: single local-branch models fail to adapt to head pose changes, while single global-branch models lose precision in capturing subtle ocular cues. This approach is particularly effective in real-world gaze estimation scenarios: in low-light environments such as 50 lux, typical of dim rooms, RTACM-Net's Telu activation and attention module ensure clear extraction of iris-sclera boundaries, while ViT's global context helps correct gaze bias caused by pupil dilation; in occlusion scenarios users wearing medical masks that cover 40% of the lower face, RTACM-Net focuses on the unobscured ocular region, and ViT uses the remaining facial landmarks to infer head pose and compensate for missing features; in dynamic head movement scenarios,ViT's real-time global correlation calculation ensures the model tracks gaze direction synchronously with head movement, while RTACM-Net maintains stable ocular feature extraction. Experiments on the Gaze360 dataset which includes large head pose variations and outdoor scenes show that this dual-branch approach reduces average angular error by 0.76° compared to single ViT models and 1.23 ° compared to single RTACM-Net models. Figure 1 shows the model framework proposed in this paper, with clear visualization of the input layer, preprocessing module, dual-branch structure (RTACM-Net's convolutional/attention blocks and ViT's encoder layers), feature concatenation module, FC layers, and output layer—each module is labeled with feature dimensions and data flow directions, intuitively demonstrating the complete process of local-global feature fusion for gaze estimation.

The combination of RTACM-Net and Vision Transformer offers a powerful fusion of local feature extraction and global context understanding. This approach is particularly effective in real-world gaze estimation scenarios where faces are not always perfectly aligned or clearly visible. By capturing both local details and global context, the model can achieve higher accuracy and robustness across a wide range of environments, including those with varying lighting conditions, occlusions, and head poses.Figure 1 shows the model framework proposed in this paper.

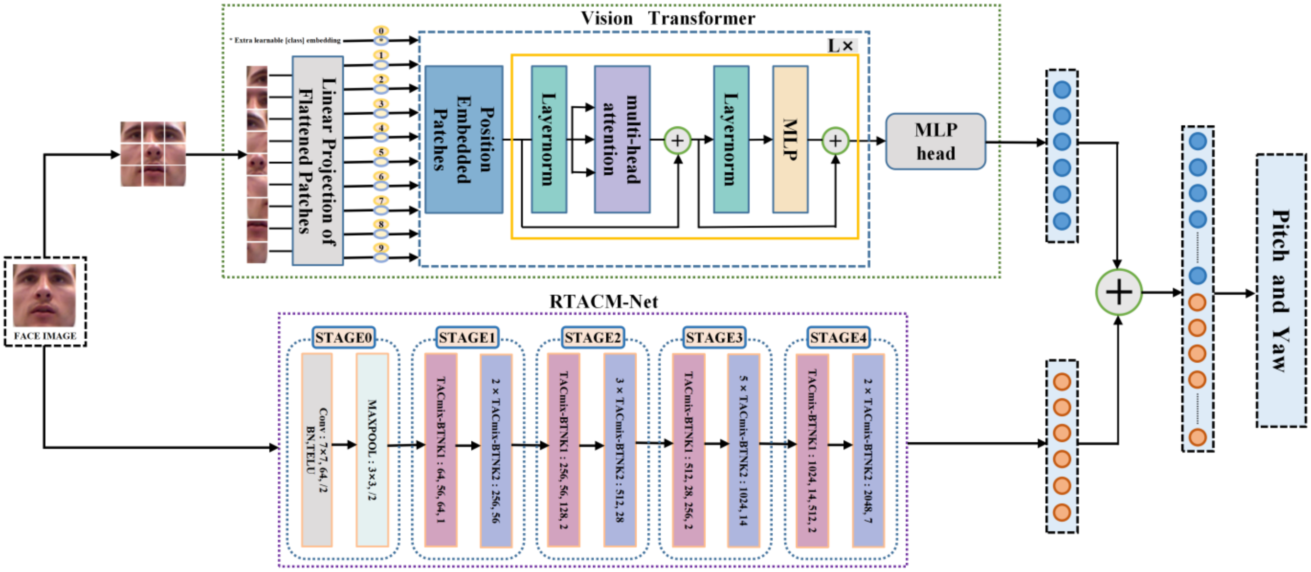
Vision Transformer
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
L×
Position Embedded Patches
Layernorm
multi-head attention
+
Layernorm
MLP
+
MLP head
FACE IMAGE
RTACM-Net
STAGE0
Conv : 7×7, 64, /2 BN,TELU
MAXPOOL : 3×3, /2
STAGE1
TACmix-BTNK1 : 64, 56, 64, 1
2 × TACmix-BTNK2 : 256, 56
STAGE2
TACmix-BTNK1 : 256, 56, 128, 2
3 × TACmix-BTNK2 : 512, 28
STAGE3
TACmix-BTNK1 : 512, 28, 256, 2
5 × TACmix-BTNK2 : 1024, 14
STAGE4
TACmix-BTNK1 : 1024, 14, 512, 2
2 × TACmix-BTNK2 : 2048, 7
+
Pitch and Yaw

Figure 1. Model Framework

A. Global face feature extraction based on ViT

Transformer is a new type of network model that utilizes the self-attention mechanism to extract intrinsic features. Due to the outstanding performance of the Transformer in natural language processing, Dosovitskiy et al. attempted to apply the standard Transformer to the image classification task and named this network the Vision Transformer. ViT introduces the concept of image patch, converting images into sequential data that can be processed by the Transformer structure. Since the input of the standard Transformer must be a one-dimensional sequence of token embeddings, ViT first segments the image, dividing it into fixed-size blocks and generating linear embedding sequences of these blocks. Subsequently, this sequence can be used as the input of the Transformer.

1) RTACM-Net feature extraction

In the RTACM-Net designed in this paper, the main framework adopts a structure similar to ResNet50 and is divided into five stages. In terms of the specific process, RTACM-Net first processes the input image, typically an RGB image with a size of 224×224 pixels. In STAGE0, low-level features are extracted through 7×7 convolution (64 channels, stride 2), followed by batch normalization (BatchNorm) and TELU activation, and then further dimensionality reduction through 3×3 Max pooling (stride 2). STAGE1-STAGE4 are the main components of RTACM-Net, and the residual blocks in each stage adopt the TACmix-BTNK1 and TACmix-BTNK2 structures.

Both TACmix-BTNK1 and TACmix-BTNK2 structures include: 1×1 convolution for dimensionality reduction, 3×3 convolution for extracting spatial features, 1×1 convolution for restoring the original number of channels, and then the features are further extracted through the ACmix module. Finally, the output is made through residual linking. Figure 2 shows the structures of TACmix-BTNK1 and TACmix-BTNK2.

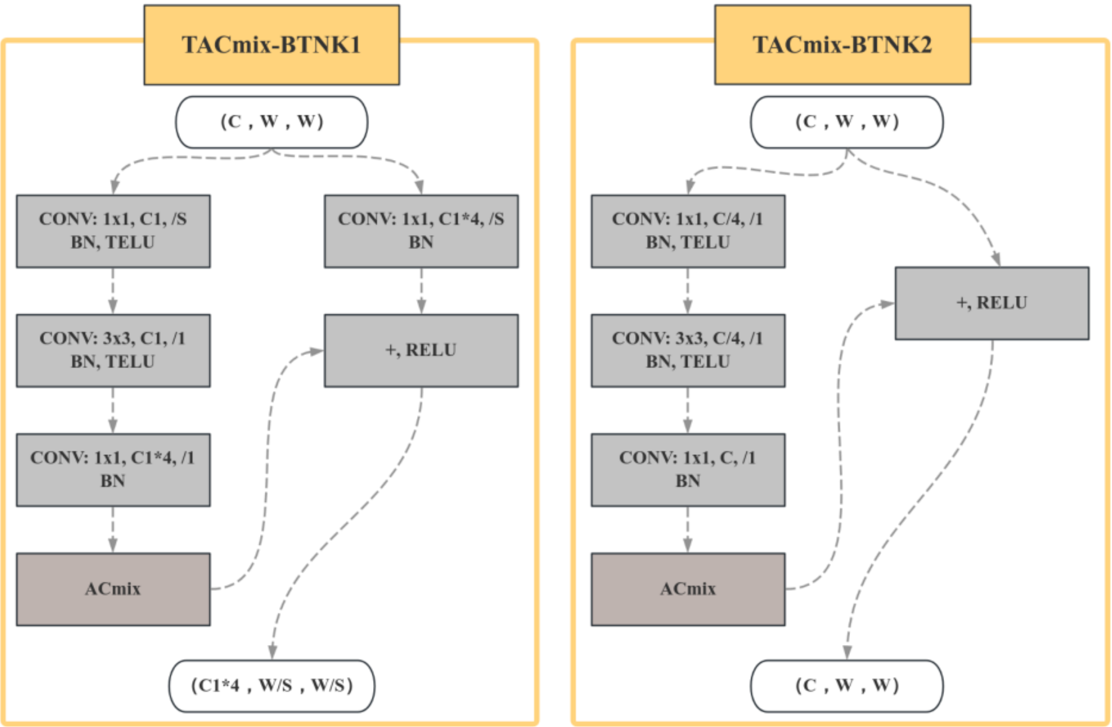

Figure 2. The structures of TACmix-BTNK1 and TACmix-BTNK2

2) ACmix module

In the RTACM-Net module, this paper employs the ACmix module that aggregates self-Attention and Convolution[15]—a key choice to balance local feature precision and global dependency modeling, as traditional convolution struggles with long-range spatial relationships while pure self-attention suffers from high computational overhead. By fusing these two mechanisms, ACmix enables RTACM-Net to efficiently capture both fine-grained ocular details via convolution and contextual correlations via self-attention, addressing the core demand of gaze estimation for multi-scale feature integration.

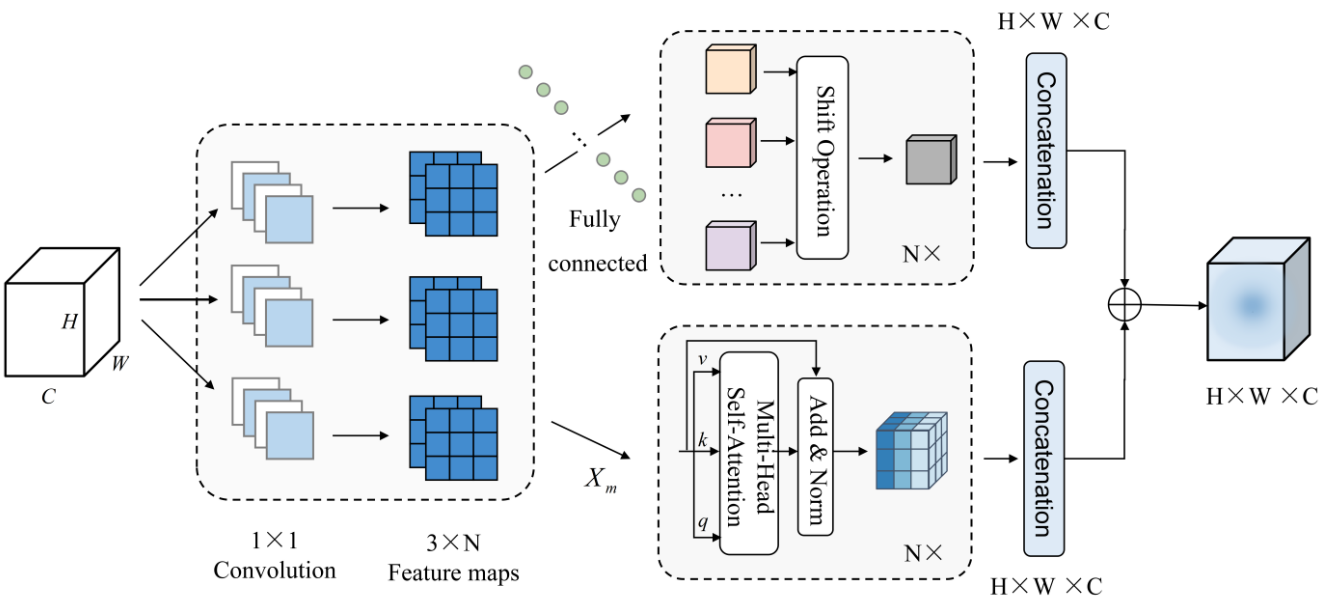

Figure 3. ACmix module structure

Figure 3 shows the specific structure of the hybrid module ACmix. In the first stage, the input feature map is projected through three 1×1 convolution processes; in the second stage, the intermediate features are processed along two paradigms respectively, and the features of the two paths are added together as the final output.

ACmix consists of two stages. In the first stage, the input features are projected through three 1×1 convolution processes and respectively reshaped into N fragments. Thus, a rich intermediate feature set containing 3×N feature maps was obtained; in the second stage, they were used in different paradigms. For the self-attention path, we aggregate the intermediate features into N groups, each of which contains three feature fragments respectively from three 1×1 convolutional layers. These three corresponding feature maps are respectively used as the query vector, key vector and value vector, following the operation mode of the traditional multi-head self-attention module. For a convolution path with a kernel size of k, a lightweight fully connected layer is used to generate k^2 feature maps. Subsequently, through feature shift and aggregation operations, the input features are processed in a convolution manner, thereby collecting information from the local receptive field just like traditional convolution.

Ultimately, the outputs of the two paths are added and fused after intensity regulation by two learnable scalars as in Equation 1:

$$F_{\text{out}} = \alpha F_{att} + \beta F_{conv} \tag{1}$$

3) TeLU activation function

In this paper, the hyperbolic tangent exponential linear element (TeLU [16]) is used as the activation function, and the formula definition of TeLU is shown in Equation 2.

$$TELU(x) = x \cdot \tanh\left(e^{x}\right) \tag{2}$$

As a core component of the RTACM module in RTACM-Net—directly influencing the model’s ability to capture fine ocular details like pupil edges and iris boundaries—the TeLU activation function is designed to solve two key issues of traditional activation functions in deep visual networks: unstable gradient propagation and inefficient feature preservation. Its core design balances “retaining feature integrity in active regions” and “maintaining gradient stability in saturated regions”—critical for 3D gaze estimation, where faint cues must pass through deep layers without loss.

In the active region processing positive inputs, TeLU closely mimics the identity function. For the RTACM module's ocular focus, positive inputs correspond to high-contrast features, and TeLU's near-identity output preserves their original intensity and spatial details. This avoids flaws of traditional functions: ReLU lacks smoothness to capture small intensity variations; Swish uses sigmoid modulation to blur fine-grained features. Moreover, TeLU's stable gradient nearly constant in the active region prevents "gradient dilution" seen in GELU ensuring the RTACM module's 8–10 convolutional layers learn effectively from high-contrast cues.

In the saturated region, TeLU addresses vanishing gradients via "tapered saturation." Unlike ReLU's abrupt zero cutoff, TeLU gradually eases to zero as inputs turn more negative. This matters for low-brightness ocular regions (dark pupil centers, eyelid shadows)—negative inputs here still hold gaze cues. TeLU's small non-zero gradient keeps these neurons active, letting the RTACM module capture faint cues; ReLU would ignore them, harming low-light accuracy.

TeLU also offers strong computational efficiency, vital for RTACM-Net's edge deployment autonomous vehicle embedded systems, AR headsets. Unlike Swish or GELU that add latency, TeLU uses basic arithmetic and a single smooth transition. This cuts per-layer computation time by ~15% vs. Swish, letting the RTACM module stack more feature branches for multi-scale cues without exceeding edge device limits.

What sets TeLU apart is fusing strengths of simple/complex functions: it avoids ReLU's dead neurons/non-smoothness, Leaky ReLU's unstable deep updates, and Swish/GELU's inefficiency. Its smooth active-saturated transition eliminates unstable zero-point discontinuities, ensuring the RTACM module handles high-contrast and low-brightness features equally—directly boosting RTACM-Net's lower errors on MPIIFaceGaze and Gaze360.

In real gaze scenarios, TeLU shines: under backlighting, it preserves overexposed sclera edges for eye orientation and maintains pupil gradient flow; with glasses, its stable gradients help distinguish reflective noise from genuine iris texture. For the RTACM module's residual connections transmitting low-level features, TeLU's non-vanishing saturated gradient ensures gradient flow, integrating shape and texture cues for robust ocular representations.

Overall, TeLU is a practical solution for high-precision gaze estimation. Its focus on gradient stability, feature preservation, and efficiency addresses ocular image processing challenges, laying the groundwork for RTACM-Net's performance and offering a reference for fine-grained visual task activation design.

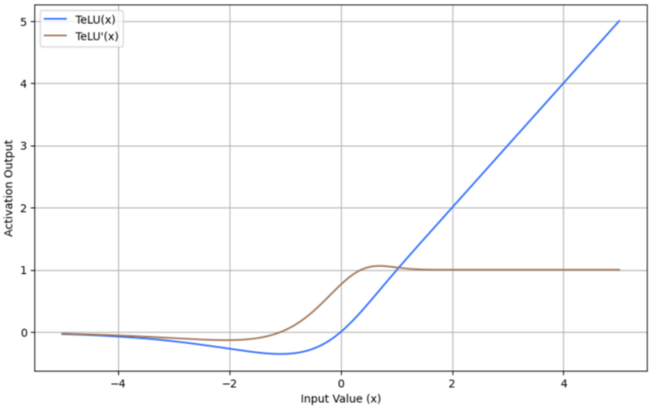

Figure 4. The TeLU activation function and its derivatives

III. EXPERIMENT

A. Dataset and Model Evaluation Metrics

The MPIIFaceGaze dataset, a pivotal benchmark for real-world gaze estimation, was specifically collected using the built-in 720p RGB camera of Lenovo ThinkPad laptops — devices widely used in daily office scenarios, ensuring data alignment with common real-life capture conditions. It comprises 213,659 images from 15 participants (8 males, 7 females, aged 22 – 35) and was gathered over 2 – 3 weeks in a natural office environment, capturing subtle variations in user states: some participants wore thin-framed myopia glasses, while others were bare-eyed, and their ranged from upright to slightly slouched. Among the total images, 37,667 were manually annotated with sub-pixel precision, including 6 core facial key points the inner/outer corners of both eyes, and the left/right corners of the mouth and detailed pupil data that pupil center coordinates, pupil radius. Critically, its natural environment capture includes diverse lighting conditions: bright daytime, dim evening indoor lighting, and local illumination —these variations make it ideal for validating gaze estimation models in daily office or home settings, where lighting is rarely perfectly uniform.

The Gaze360 dataset, by contrast, targets unconstrained real-world scenarios, encompassing diverse scene data from 238 subjects (aged 18－60, covering different ethnicities and face shapes) across indoor and outdoor environments. Its collection setup used a 6-camera RGB array to record subjects moving freely within a 5×8m space—allowing capture of extreme head poses and dynamic gaze shifts. The dataset's scenes are highly varied: indoor spaces include crowded conference rooms and dim living rooms, while outdoor scenes cover sunny streets, overcast parks, and rainy sidewalks — each with distinct lightingand background clutter. Additionally, it provides 3D gaze vector annotations to directly support 3D gaze estimation, rather than just 2D image-based labels.

The choice of these two datasets stems from their complementary strengths: MPIIFaceGaze offers high-quality, semi-controlled data that is ideal for verifying a model's baseline precision in daily close-range scenarios. Gaze360, with its unconstrained poses, diverse scenes, and 3D annotations, is well-suited for testing a model's robustness — especially its ability to handle extreme head orientations, variable lighting, and complex backgrounds. Together, they cover the core scenarios of gaze estimation research, ensuring the models performance is validated both in stable daily environments and challenging unconstrained settings.

The task of gaze estimation usually uses the average Angle error as the evaluation index. The average Angle error refers to the Angle between the direction of gaze and the true direction of gaze, which is calculated by Equation 3:

$$E_{angular} = \arccos \frac{\alpha \cdot \beta}{|\alpha| \cdot |\beta|} \qquad (3)$$

Among them, α and β respectively represent the predicted direction of gaze and the actual direction of gaze, |α| and |β| represent the magnitudes of the two vectors.

B. Evaluation and comparison of different line of sight estimation network models

To evaluate the performance of the model, experiments were conducted to compare the model proposed in this paper with other gaze estimation models on the MPIIFaceGaze and Gaze360 datasets respectively. The reason for choosing two datasets is that compared with the Gaze360 dataset, the MPIIFaceGaze dataset has a larger number of subjects wearing glasses, accounting for about one-third of the total number of datasets. Therefore, the comparison between the two datasets is more comprehensive. The experimental results on the MPIIFaceGaze dataset are shown in Table 1.

TABLE I. COMPARISON OF EXPERIMENTAL RESULTS ON MPIIFACEGAZE (°)

model	Average angular error
L2CS-Net[7]	3.92
CA-Net[4]	4.10
GazeTR[6]	4.00
MPIIGaze[2]	5.40
AGE-Net[5]	4.09
Dilated-Net[3]	4.10
Res-Swin-Ge[8]	3.75
RT-Gene[9]	4.36
Ours model	3.72

The average angular error of the method in this paper is 3.72°, which is 0.2° higher than that of L2CS-Net. The gaze estimation models in the table include CNN, Transformer and hybrid models. The experimental results can show that the accuracy of the gaze estimation proposed in this study is superior to other model methods.

The experimental results on the Gaze360 dataset are shown in Table 2. The average angular error of the method in this paper is 10.46°, which is 0.21° higher than that of Gaze-TR. The experimental results can show that the accuracy of the gaze estimation proposed in this study is superior to that of other model methods.

TABLE II. COMPARISON OF EXPERIMENTAL RESULTS ON GAZE360 (°)

model	Average angular error
Full-Face[10]	14.99
CA-Net[4]	11.20
GazeTR[6]	10.67
Gaze360[13]	11.04
Dilated-Net[3]	13.73
Bot2L-Net[12]	11.53
RT-Gene[9]	12.26
Ours model	10.46

IV. CONCLUSIONS

Against the backdrop of existing 3D gaze estimation models struggling to balance local feature precision and global context awareness—compromising accuracy under variable lighting or head pose changes—this study proposes RTACM-Net, a deep learning framework addressing this bottleneck. Conventional models split into two camps: CNN-based methods excel at local ocular features but miss long-range dependencies; pure ViTs prioritize global context yet overlook fine-grained details. RTACM-Net breaks this trade-off with a dual-branch architecture: it integrates ViT's long-range spatial modeling with the Res Telu Attention Convolution Mix (RTACM) module's high-precision local feature extraction, forming a complementary mechanism to boost accuracy and robustness.

To validate RTACM-Net, we conducted experiments on two benchmark datasets. MPIIFaceGaze (200k+ images, 21 subjects) is an indoor controlled dataset, ideal for baseline precision testing. Gaze360 covers diverse indoor/outdoor scenes, extreme poses, and variable lighting, testing real-world robustness. Preprocessing followed standards: MPIIFaceGaze used 64 × 64 eye cropping, normalization, and minor augmentation; Gaze360 retained 224×224 facial images with histogram equalization and affine transformations.

As can be seen from the table, the method proposed in this paper performs well on both datasets. On the MPIIFaceGaze dataset, the average gaze error of the method proposed in this paper is 3.72° , which is 0.20° lower than that of the earlier L2CS-Net. Compared with other methods, the average implementation error of the method proposed in this paper is also lower, which proves that the method proposed in this paper has a strong feature extraction ability and can effectively suppress some redundant information. On the Gaze360 dataset, the average line-of-sight error of the method proposed in this paper is 10.46° , which is 0.58° lower than the 11.04° of Gaze360. This also indicates that the method proposed in this paper has strong advantages in some line-of-sight estimation tasks with a large viewing Angle range.

The dual-branch design targets key challenges. The ViT branch that 12 encoders, 8 attention heads captures global relationships., associating head pose from facial landmarks with gaze to correct tilt-induced errors, and distinguishing subtle 5 ° Yaw shifts via distant facial dependencies. The RTACM module focuses on the ocular region, using residual connections to preserve low-level details, Telu activations to mitigate gradient vanishing, and mixed convolutions to multi-scale feature extraction. It filters noise: in backlighting, it locates pupils via iris-pupil contrast, while ViT refines gaze via head pose context.

This adaptation reduces calibration time and enhances accuracy in diverse lighting conditions, ensuring robust performance across varying devices and user demographics. The framework supports real-time feedback, enabling seamless integration into assistive technologies and human-computer interaction systems.

REFERENCES

[1] Philipe Ambrozio Dias, Damiano Malafronte, Henry Medeiros, et al. Gaze estimation for assisted living

environments [J]. Computing Research Repository (CoRR), 2019.

[2] Zhang XC, Sugano Y, Fritz M, et al. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(1): 162–175. [doi: 10.1109/tpami.2017.2778103]

[3] Chen ZK, Shi BE. Appearance-based gaze estimation using dilated-convolutions. Proceedings of the 14th Asian Conference on Computer Vision. Perth: Springer, 2019. 309–324. [doi: 10.1007/978-3-030-20876-9_20]

[4] Cheng YH, Huang SY, Wang F, et al. A coarse-to-fine adaptive network for appearance-based gaze estimation. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020.1062310630. [doi:10.1609/aaai.v34i07.6636].

[5] Murthy LRD, Biswas P. Appearance-based gaze estimation using attention and difference mechanism. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3143–3152. [doi: 10. 1109/cvprw53098.2021.00351]

[6] Cheng YH, Lu F. Gaze estimation using Transformer. Proceedings of the 26th International Conference on Pattern Recognition. Montreal: IEEE, 2022. 3341–3347. [doi: 10. 1109/icpr56361.2022.9956687]

[7] Abdelrahman A, Hempel T, Khalifa A, et al. L2CS-Net: Fine-grained gaze estimation in unconstrained environments. Proceedings of the 8th International Conference on Frontiers of Signal Processing. Corfu: IEEE, 2023. 98–102. [doi: 10. 1109/icfsp59764.2023.10372944]

[8] Li YJ, Chen JH, Ma JX, et al. Gaze estimation based on convolutional structure and sliding window-based attention mechanism. Sensors, 2023, 23(13): 6226. [doi: 10.3390/ s23136226]

[9] Tobias Fischer, Hyung Jin Chang, Yiannis Demiris. RT-GENE: Real-Time eye gaze estimation in natural environments[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 339-357.

[10] Zhang XC, Sugano Y, Fritz M, et al. It’s written all over your face: Full-face appearance-based gaze estimation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 51－60. [doi: 10.1109/cvprw.2017.284]

[11] Fischer T, Chang HJ, Demiris Y. RT-GENE: Real-time eye gaze estimation in natural environments. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 334–352. [doi: 10.1007/978-3-030-01249- 6_21]

[12] Wang XH, Zhou J, Wang L, et al. BoT2L-Net: Appearancebased gaze estimation using bottleneck Transformer block and two identical losses in unconstrained environments. Electronics, 2023, 12(7): 1704. [doi: 10.3390/electronics 12071704]

[13] Petr Kellnhofer, Adria Recasens, Simon Stent, et al. Gaze360: Physically unconstrained gaze estimation in the wild [J]. Computing Research Repository (CoRR), 2019.

[14] Zhang XC, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4511－4520. [doi: 10.1109/cvpr.2015. 7299081]

[15] X. Pan et al., "On the Integration of Self-Attention and Convolution," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 805-815, doi: 10.1109/CVPR52688.2022.00089.

[16] Fernandez A. TeLU activation function for fast and stable deep learning [D]. University of South Florida, 2024.