Dual-Stream EfficientNet Architecture for the 3D Gaze Estimation Using RGB-D Data

Hafiz Ahmad Qadeer
School of Robot and Smart System Engineering,
Kyungpook National University,
Daegu, South Korea
ahmad.khan2903@knu.ac.kr

Min Young Kim
School of Electronics Engineering,
Research Center for Neurosurgical Robotic System
Kyungpook National University,
Daegu, South Korea
minykim@knu.ac.kr

***Abstract*—Gaze estimation plays a critical role in various applications, including virtual reality, human-computer interaction, and advanced driver monitoring systems. This paper introduces a Gaze Estimation Model that employs a dual-stream architecture, integrating RGB and depth data to achieve precise 3D gaze prediction. The model leverages EfficientNet-B3 as the backbone on both streams and incorporates Long Short-Term Memory (LSTM) layers to capture temporal dependencies, effectively enhancing performance in dynamic environments. Through comprehensive evaluation of the EyeDiap dataset, the proposed model achieves a Mean Angular Error (MAE) of 5.96°, outperforming existing state-of-the-art methods and setting a new benchmark for 3D gaze estimation accuracy. The integration of depth data with RGB imagery enriches the feature set, significantly contributing to the model's accuracy. Despite challenges such as increased computational demands and noise in-depth data, the model demonstrates robust, real-time performance, making it suitable for deployment in real-world applications. Future work will aim to optimize computational efficiency and extend the model's applicability across diverse conditions.**

***Index Terms*—Gaze Estimation, Dual-Stream Architecture, RGB-D Imaging, EfficientNet-B3, EyeDiap Dataset, Real-Time Prediction, Multi-Modal Deep Learning,**

I. Introduction

Gaze estimation, the technique of determining a person's point of focus, is increasingly pivotal in a wide range of applications, including virtual reality, human-computer interaction, and advanced driver monitoring systems. Traditional gaze estimation models have predominantly utilized 2D inputs and simpler neural architectures, often resulting in sub-optimal performance in complex, real-world scenarios [1]. However, recent advances in deep learning and computer vision have paved the way for more sophisticated methodologies. Notably, convolutional neural networks (CNNs) for full-face appearance-based gaze estimation [2] and the integration of RGB-D sensors for precise 3D gaze estimation [3] have emerged as promising approaches.

The introduction of RGB-D sensors has been transformative in the field of gaze estimation, providing depth information that complements traditional RGB data and thereby enabling more accurate 3D gaze predictions [4]. Dual-stream architectures that separately process RGB and depth images have demonstrated significant potential in this domain. For instance, [5] showcased the efficacy of a multi-task CNN leveraging RGB-D data, highlighting the advantages of integrating multiple data streams to enhance gaze estimation accuracy.

Building on these advancements, we propose a Gaze Estimation Model that utilizes a dual-stream architecture, with EfficientNet-B3 as the backbone network for both RGB and depth image processing. EfficientNet, known for its scalable and efficient architecture [6], offers an optimal balance between computational efficiency and model accuracy, making it particularly well-suited for real-time applications. By fusing the features extracted from both streams, our model captures a richer representation of spatial and depth information, which is crucial for precise 3D gaze vector prediction (yaw, pitch, roll).

In addition to the dual-stream architecture, our model incorporates advanced preprocessing techniques, including extensive data augmentation and normalization, to enhance generalization across diverse conditions. We also integrate MTCNN for accurate face detection and alignment [7], ensuring high-quality input data that leads to more reliable gaze estimation. Furthermore, the inclusion of LSTM layers for sequential frame processing allows the model to effectively capture temporal dependencies, thereby improving performance in dynamic environments.

The proposed model is evaluated on the EyeDiap dataset [8], a benchmark dataset known for its comprehensive RGB-D data, making it ideal for evaluating gaze estimation algorithms. Our evaluation demonstrates that the Gaze Estimation Model delivers superior accuracy and robustness, establishing its suitability for real-time applications across various fields.

II. Related Work

Gaze estimation has been a focal point of research in computer vision and human-computer interaction, with significant advancements made over the past decades. Early works, such as the comprehensive survey by [1], laid the groundwork by categorizing various gaze estimation techniques, ranging from traditional model-based methods to emerging appearance-based approaches. These traditional methods often relied on simplistic models, limiting their applicability in complex, real-world environments.

A pivotal shift in gaze estimation research occurred with the introduction of appearance-based methods, which leverage machine learning and deep learning techniques to estimate gaze directly from images. [2] explored full-face appearance-based gaze estimation, demonstrating that leveraging the entire facial appearance can significantly enhance estimation accuracy. Similarly, [9] introduced the use of dilated convolutions to refine the appearance-based gaze estimation process further, achieving better performance in challenging scenarios. The integration of depth information has also proven to be a critical advancement in gaze estimation. [3] pioneered the use of RGB-D sensors to estimate gaze in 3D space, showing that depth data significantly improves estimation accuracy compared to RGB-only approaches. [4] corroborated these findings, comparing RGB and RGB-D solutions and concluding that the inclusion of depth information provides a more robust estimation, particularly in dynamic environments.

In recent years, dual-stream architectures have emerged as a promising approach to gaze estimation. These architectures process RGB and depth images separately before combining their features for final estimation. [5] employed a multi-task CNN that utilized both RGB and depth data, demonstrating superior performance compared to single-stream models. This approach aligns with the ongoing trend of utilizing multi-modal inputs to capture more comprehensive features, leading to more accurate and reliable gaze estimation. The advancement in convolutional neural networks (CNNs) has also significantly impacted gaze estimation research. The introduction of EfficientNet by [6] marked a substantial leap forward in CNN design, offering a scalable and efficient architecture that balances accuracy and computational cost. This has been particularly beneficial for real-time gaze estimation applications, where computational efficiency is crucial.

In the domain of real-time gaze estimation, [10] introduced RT-GENE, a system designed for real-time gaze estimation in natural environments. This work underscores the growing importance of developing systems that not only achieve high accuracy but also maintain real-time processing capabilities, a challenge that continues to drive innovation in the field. Face detection and alignment are foundational steps in gaze estimation pipelines, with recent advancements further enhancing overall system performance. [7] Proposed an improved version of the Multi-task Cascaded Convolutional Networks (MTCNN) for face detection, which provides more accurate alignment and contributes to better gaze estimation results.

Finally, the availability of robust datasets has played a critical role in advancing gaze estimation research. The EyeDiap dataset, introduced by [8], is one of the most comprehensive datasets available, offering both RGB and RGB-D data. This dataset has been instrumental in evaluating and benchmarking new gaze estimation algorithms, facilitating the development of more accurate and reliable models. The field of gaze estimation has evolved from simple model-based approaches to sophisticated deep learning models that leverage multi-modal inputs and advanced architectures. The integration of depth information, the development of dual-stream networks, and the focus on real-time capabilities represent significant milestones in this evolution. As research continues, the combination of these advancements promises to push the boundaries of what is possible in gaze estimation, paving the way for more robust and accurate applications.

III. Methodology

This section outlines the methodology underlying our proposed model for 3D gaze estimation, which integrates a dual-stream architecture designed to leverage RGB and depth data. The model architecture is optimized for real-time applications, offering robust performance by capturing complementary information from these two modalities. We detail the architecture, feature fusion, gaze prediction, sequential information processing, and data preprocessing techniques employed in this work.

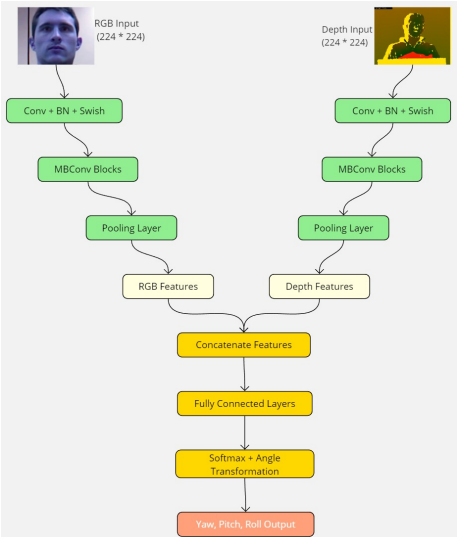

Fig. 1. Schematic Representation of the Dual-Stream Architecture for 3D Gaze Estimation.

A. Model Overview

Our model is designed to estimate 3D gaze directions by leveraging both RGB and depth images. To assess the accuracy of the model, we use the Mean Angular Error (MAE) metric. The MAE is calculated as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \cos^{-1}\left(\frac{\mathbf{g}_{\text{pred}_i} \cdot \mathbf{g}_{\text{gt}_i}}{\|\mathbf{g}_{\text{pred}_i}\| \|\mathbf{g}_{\text{gt}_i}\|}\right) \cdot \frac{180}{\pi} \tag{1}$$

Where:

- $\mathbf{g}_{\text{pred}_i}$ is the predicted gaze vector for the i-th sample.
- $\mathbf{g}_{\text{gt}_i}$ is the ground truth gaze vector for the i-th sample.
- N is the total number of test samples.

This MAE metric provides a reliable measure of model accuracy by averaging the angular discrepancies between predicted and actual gaze vectors across all test samples.

B. Dual-Stream Architecture

The foundation of our model is a dual-stream architecture that processes RGB and depth images concurrently. Each stream utilizes EfficientNet-B3 as its backbone, leveraging its compound scaling approach for an optimal balance between accuracy and computational efficiency [6]. This allows the model to extract detailed spatial features from RGB images and depth cues from depth images, ensuring that both streams contribute effectively to the final prediction.

- RGB Stream: The RGB stream begins with a convolutional layer, followed by batch normalization (BN) and the Swish activation function to extract initial feature maps. These features are further refined through a series of Mobile Inverted Bottleneck Convolution (MBConv) blocks, known for their capacity to capture spatial features efficiently. A pooling layer condenses these features, preparing them for fusion with the depth stream.
- Depth Stream: The depth stream mirrors the RGB stream in its processing steps, ensuring consistent treatment of both modalities. This parallel processing facilitates the eventual fusion of RGB and depth data into a unified feature representation.

The outputs from the RGB and depth streams are then concatenated to form a comprehensive feature representation, which is processed through an LSTM layer to capture temporal dependencies. The final 3D gaze vectors are predicted as follows:

$$\mathbf{g}_{\text{3D}} = \mathbf{W}_{\text{out}} \cdot \mathrm{L}\left(\mathrm{C}\left(\mathrm{E}_{\mathrm{R}}(\mathbf{I}_{\mathrm{R}}),\, \mathrm{E}_{\mathrm{D}}(\mathbf{I}_{\mathrm{D}})\right)\right) + \mathbf{b}_{\text{out}} \tag{2}$$

Where:

- $\mathbf{I}_{\mathrm{R}}$ and $\mathbf{I}_{\mathrm{D}}$ represent the input RGB and depth images, respectively.
- $\mathrm{E}_{\mathrm{R}}(\cdot)$ and $\mathrm{E}_{\mathrm{D}}(\cdot)$ denote the feature extraction processes using EfficientNet-B3 for each modality.
- $\mathrm{C}(\cdot)$ denotes the concatenation (fusion) of RGB and depth features.
- $\mathrm{L}(\cdot)$ captures temporal dependencies in the fused features using LSTM.
- $\mathbf{W}_{\text{out}}$ and $\mathbf{b}_{\text{out}}$ are the learned weights and biases of the output layer, which produce the 3D gaze vectors $\mathbf{g}_{\text{3D}} = [\hat{y}, \hat{p}, \hat{r}]$.

This equation encapsulates the model's core operation, where RGB and depth features are fused, processed through LSTM layers, and mapped to the predicted 3D gaze angles.

C. Feature Fusion and Gaze Prediction

Feature fusion is a critical step in our model, where the outputs from the RGB and depth streams are concatenated to form a comprehensive feature representation. This approach leverages the strengths of both modalities: RGB data provides rich spatial details, while depth data contributes valuable information about the scene's geometry. The fused features are subsequently processed by fully connected layers, where a softmax operation is applied to convert logits to probabilities. Following this, an angle transformation is used to map these probabilities to continuous angles, representing the yaw, pitch, and roll that describe the user's gaze direction in three-dimensional space. The model is trained using a composite loss function, integrating Mean Squared Error (MSE) for the regression of gaze angles with any relevant classification losses to ensure robust performance across tasks.

To enhance the model's ability to handle dynamic gaze behaviors, we incorporate Long Short-Term Memory (LSTM) layers after the feature fusion stage. LSTM networks are well-suited for sequential data processing, as they maintain and update information across multiple time steps. In the context of gaze estimation, this capability is crucial for capturing temporal dependencies, such as the natural flow of eye movements over time. The integration of LSTM layers allows the model to produce smoother, more consistent gaze predictions, particularly in scenarios where gaze direction changes rapidly.

D. Model Architecture and Training Strategy

The proposed Dual-Stream Gaze Estimation model utilizes EfficientNet-B3 as the backbone for the RGB and depth streams. The backbone networks are initialized with ImageNet-pre-trained weights to leverage robust feature extraction capabilities. Input images are resized to 224x224 pixels, and the Swish activation function is consistently applied throughout the model. EfficientNet-B3 employs Mobile Inverted Bottleneck Convolution (MBConv) blocks to optimize feature extraction for both accuracy and computational efficiency. After feature extraction from the RGB and depth streams, an attention module is introduced to refine the spatial representations, processing 1536 input channels down to 768 intermediate channels. These refined features are then concatenated and fed into two Long Short-Term Memory (LSTM) layers, each with 512 hidden units and a dropout rate of 0.5, to capture temporal dependencies across sequences of frames and enhance performance in dynamic environments. The model is trained using the Adam optimizer with an initial learning rate of 0.0001, which decays by a factor of 0.1 if the validation loss does not improve after 10 epochs. Training is conducted with a batch size of 8 over 20 epochs, optimizing the Mean Squared Error (MSE) as the loss function. L2 regularization with a weight decay of 1×10^{-4} and a dropout rate of 0.5 are employed to prevent overfitting.

IV. Experiments

A. Data Preprocessing

Our preprocessing pipeline systematically processes the EyeDiap dataset [8] for gaze estimation. We organize sessions, load relevant data, and decode camera calibration parameters to transform 3D coordinates. For each selected frame, we calculate head pose and gaze direction, using the 'Norm' class to normalize these relatives to the camera, thereby producing aligned face and eye images [2], [11]. These images are then cropped, equalized, and saved along with metadata such as 3D/2D gaze vectors and head rotations. This process ensures consistent and accurate data preparation for model training.

B. Performance Comparison on the EyeDiap Dataset

The effectiveness of the proposed dual-stream model for 3D gaze estimation was assessed using the EyeDiap dataset, with its performance compared against a baseline method and several state-of-the-art methods. The Mean Angular Error (MAE) metric was utilized as the primary measure of accuracy.

TABLE I
COMPARISON OF MEAN ANGULAR ERROR AND FPS BETWEEN L2CS-NET AND THE PROPOSED MODEL.

Model Architecture	MAE (Degrees)	FPS
L2CS-Net (RGB) Single-Stream CNN [11]	7.56°	25
Proposed Model (RGBD) Dual-Stream CNN	5.96°	20

As shown in Table 1, the baseline L2CS-Net model, which uses a single-stream convolutional neural network (CNN) architecture and processes only RGB images, achieved an MAE of 7.56 degrees. In contrast, our proposed model, which integrates both RGB and depth data through a dual-stream architecture and incorporates Long Short-Term Memory (LSTM) layers for sequential information processing, significantly outperformed the baseline with an MAE of 5.96 degrees.

TABLE II
COMPARISON OF STATE-OF-THE-ART METHODS ON THE EYEDIAP DATASET. THE PROPOSED METHOD ACHIEVED THE LOWEST MAE, ESTABLISHING IT AS THE NEW STATE-OF-THE-ART.

Method	MAE (Degree)
Mnist [12]	7.37°
iTracker [13]	7.13°
Full Face [2]	6.53°
Dilated-Net [9]	6.19°
RT-Gene (4 ensemble) [10]	6.02°
Proposed Method	**5.96°**

Our proposed model achieved a Mean Angular Error (MAE) of 5.96° degrees, the lowest among the methods compared. This result represents a significant improvement over both the baseline and other state-of-the-art methods, including RT-Gene, which previously held the best performance with an MAE of 6.02° degrees. To further illustrate the performance differences between the L2CS-Net model and the proposed model, Figure 1 presents a detailed comparison of the angular errors across different batch indices in the EyeDiap dataset.

As depicted in Figure 2, the proposed model consistently exhibits lower angular errors than the L2CS-Net model across the entire dataset. The mean angular error of the proposed model is significantly lower, represented by the green dashed line, indicating its improved accuracy in predicting 3D gaze directions. The flagged errors in the figure further emphasize instances where the L2CS-Net model's performance notably deteriorates, while the proposed model maintains a more stable error profile.

To illustrate the processing stages of the facial images utilized in our model, Figure 3 presents a visualization organized into three columns, each representing a different subject. The

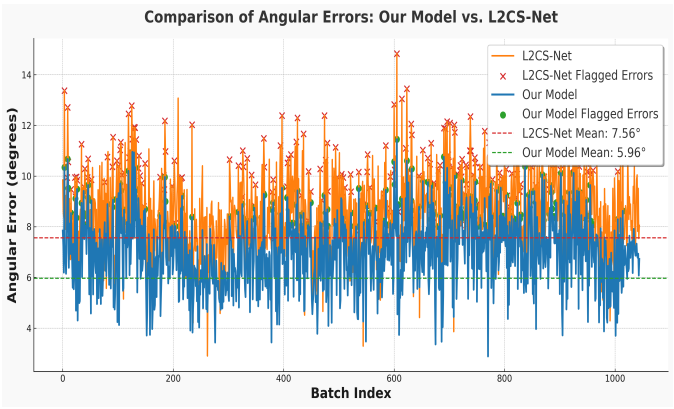

Fig. 2. Comparison of Angular Errors: Proposed Model vs. L2CS-Net. The figure shows the angular errors of both models across different batch indices. The mean angular errors are indicated by dashed lines: 7.56° for L2CS-Net (red) and 5.96° for the proposed model (green). Flagged errors highlight specific instances where the models exhibited the most significant deviations.

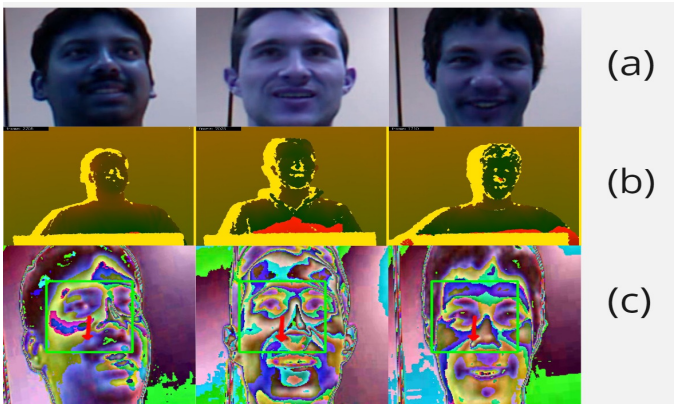

Fig. 3. Visualization of Facial Image Processing: (a) Original RGB images, (b) Corresponding depth images, and (c) Fused RGB-D images. These rows illustrate the progression from raw data to a fused representation, combining color and depth information for facial feature analysis.

figure is structured into three rows, with each row highlighting a distinct type of image data:

- **RGB Images(a)**: Displays the original RGB images of three subjects, captured using a standard RGB camera. These images provide the color information of the subjects' faces, serving as the foundational input for subsequent processing.
- **Depth Images(b)**: Shows the corresponding depth images for the same subjects depicted in Row (a). Depth images convey information about the distance of various facial points from the camera, represented by a color gradient where different colors correspond to varying depth levels. This depiction captures the three-dimensional structure of the face, offering valuable spatial context.
- **Fused RGB-D Images(c)**: Presents each subject's concatenated RGB and depth images. This fusion of color and depth data enhances the model's ability to analyze facial features by integrating both texture and spatial information. The colorful patterns in these images high-

light regions where significant facial landmarks have been detected, making this step crucial for facial recognition and emotion detection applications.

C. Discussion

The proposed dual-stream architecture for 3D gaze estimation demonstrates significant advancements, reducing the Mean Angular Error (MAE) from 7.56° with L2CS-Net to 5.96°. This improvement is achieved through the integration of RGB and depth data, where RGB data captures detailed spatial features and depth data provides essential geometric context. Our model surpasses state-of-the-art methods, achieving a lower MAE than RT-Gene (6.02°) and Dilated-Net (6.19°), particularly in dynamic environments where rapid gaze changes occur. The incorporation of LSTM layers after feature fusion has proven effective in capturing temporal dependencies, resulting in smoother and more consistent gaze predictions, which are crucial for real-time applications.

However, several challenges were encountered during development. One notable issue is the inherent noise in depth data, especially when the subject is positioned farther from the depth sensor, which can degrade the accuracy of the depth stream and, by extension, the overall model performance. Future work could explore advanced noise reduction techniques or improved sensor calibration methods to mitigate these issues. Additionally, the simultaneous processing of RGB and depth data increases the model's computational demands, achieving an average processing rate of 20 frames per second (FPS). Future research could focus on optimizing the model to improve FPS without sacrificing accuracy, potentially through techniques such as model pruning or quantization.

A limitation of this study is the reliance on the EyeDiap dataset for evaluation, which may not capture the full range of real-world variability. Most available gaze estimation datasets, such as GazeCapture, MPIIGaze, and ETH-XGaze, lack depth data, making them unsuitable for direct evaluation of RGBD models. To address this, future work could focus on generating synthetic RGBD data or using multi-camera setups to create depth information for existing RGB datasets. Alternatively, adapting the model for RGB-only data using domain adaptation or transfer learning could extend its applicability to scenarios without depth data.

Further research should also evaluate the model in real-time gaze tracking using RGB-D cameras in dynamic environments. Such efforts would help validate the model's robustness and effectiveness across various applications, including virtual reality, driver monitoring, and human-computer interaction.

V. Conclusion

This paper presented a dual-stream gaze estimation model that leverages both RGB and depth data for 3D gaze prediction. When evaluated on the EyeDiap dataset, the proposed model achieved a Mean Angular Error (MAE) of 5.96°, outperforming existing state-of-the-art methods. The combination of EfficientNet-B3 for feature extraction and LSTM layers for capturing temporal dependencies proved effective in enhancing the model's robustness in dynamic environments.

Future work will focus on broadening the model's evaluation to include synthetic RGB-D datasets and exploring domain adaptation techniques to extend its applicability to RGB-only datasets. Additionally, efforts will be directed toward optimizing the model's computational efficiency to enable its deployment in real-time applications such as virtual reality, driver monitoring, and human-computer interaction.

VI. Acknowledgement

This work was supported by the Basic Science Research Program through the National Research Institute Foundation of Korea (NRF) funded by the Ministry of Education(2021R1A6A1A03043144) and the BK21 Four project funded by the Ministry of Education, Korea (4199990113966). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2C2008133).

References

[1] D. W. Hansen and Q. Ji, "In the eyes of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[2] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 2299–2308.

[3] K. A. Funes Mora and J.-M. Odobez, "Gaze estimation in the 3d space using rgb-d sensors," *International Journal of Computer Vision*, vol. 113, no. 3, pp. 267–283, 2015.

[4] X. Xiong, Z. Liu, Q. Cai, and Z. Zhang, "Eye gaze tracking using an rgbd camera: A comparison with a rgb solution," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. New York, NY, USA: ACM, 2014, pp. 1113–1121.

[5] D. Lian, Z. Zhang, W. Luo, L. Hu, M. Wu, Z. Li, J. Yu, and S. Gao, "Rgbd based gaze estimation via multi-task cnn," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2488–2495.

[6] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[7] M. Gu, X. Liu, and J. Feng, "Classroom face detection algorithm based on improved mtcnn," *Signal, Image and Video Processing*, vol. 16, pp. 1355–1362, 2022.

[8] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "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*. New York, NY, USA: ACM, 2014, pp. 255–258.

[9] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," 2019. [Online]. Available: https://arxiv.org/abs/1903.07296

[10] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, 2018.

[11] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges, "L2cs-net: Fine-grained gaze estimation in unconstrained environments," in *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, Corfu, Greece, 2023, pp. 98–102.

[12] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4511–4520.

[13] K. Krafka, A. Khosla, P. Kellnhofer, and et al., "Eye tracking for everyone," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176–2184.