Refining Intelligent Vehicle Driver Gaze Prediction: System Analysis and Improvement

* Pin-Chi Pan
Graduate Institute of Communication Engineering
National Taiwan University
Taipei, Taiwan
r12942103@ntu.edu.tw

***Abstract*—Eye gaze tracking is a longstanding challenge in computer vision, with broad applications across driver assistance, virtual reality, and human-computer interaction. Among these, driver assistance stands out as a critical domain, where gaze tracking models play a pivotal role in predicting a driver's gaze region and assessing their state, thereby significantly enhancing driving safety. Traditional gaze prediction methods rely on estimating gaze direction based on eye states, which are primarily effective with wearable gaze tracking devices but are constrained by their high cost. Non-wearable gaze tracking devices offer a cost-effective alternative; however, they are prone to inaccuracies induced by head movement. In this project, a combined implementation and research approach was undertaken. The 6DRepNet backbone for head pose estimation was modified to enhance model performance, with a meticulous analysis conducted on the impact of various loss functions on estimation accuracy. Despite efforts to modify the L2CS-Net backbone for gaze tracking, significant performance improvements were not achieved. Nonetheless, a viable solution was proposed to address this challenge. Finally, several potential applications for intelligent vehicle driver gaze prediction were explored, highlighting the diverse opportunities for integrating gaze tracking technology into automotive safety systems.**

***Index Terms*—Gaze tracking, Head pose estimation, Eye tracking, Gaze analysis, Decision tree.**

I. Introduction

According to statistics, the primary cause of most traffic accidents is distracted driving, which includes activities such as using mobile phones, holding objects, smoking or eating, and adjusting audio or air conditioning. It is evident that the driver's attention is crucial for safe driving. The current attention of the driver is closely related to their gaze direction. Therefore, studying the driver's gaze direction has been widely applied in driver state and attention detection. Gaze tracking devices are typically divided into wearable and non-wearable devices. Although wearable gaze tracking devices have higher accuracy, they are not commonly used in practical applications due to their high production costs and inconvenience in real-life usage. Conversely, non-wearable gaze tracking systems can significantly reduce manufacturing costs and offer greater flexibility, making them more valuable in practical applications than wearable devices. When predicting gaze using non-wearable devices, since head pose contributes primarily to gaze direction, most methods treat head orientation as an approximation of gaze direction. However, in real driving scenarios, many drivers move both their head and eyes when looking at targets. Tawari et al. [1] compared the predictive performance of gaze when using only head pose versus using both head and eye poses simultaneously. Effective integration of head and eye pose prediction can significantly improve accuracy. Fridman et al. [2] further pointed out that the accuracy of gaze prediction increases more when the driver's head remains stationary compared to when there is significant head movement.

This project investigates the prediction of driver head and eye poses from single images and the effective integration of multiple features to determine their gaze region. For head pose prediction, we enhanced the 6DRepNet backbone to improve model performance, employing a landmark-free method that utilizes a rotation matrix with nine parameters to accurately regress head orientation. This approach ensures full pose regression without encountering Gimbal Lock issues. For eye pose prediction, we adopted a method that utilizes multiple loss functions to estimate 3D gaze angles from images. Although modifying the L2CS-Net backbone for gaze tracking did not yield significant performance improvements, we proposed a viable solution by employing data augmentation and parameter-efficient fine-tuning to enhance prediction accuracy and predict the 3D gaze direction. Finally, by leveraging the predicted head and eye poses, facial coordinates, and distance from the camera, among other features, we use a decision tree algorithm to predict the gaze region the driver is focusing on. Additionally, we explored several potential applications of intelligent vehicle driver gaze prediction. The principal contributions of this study are as follows:

- **Refinement of 6DRepNet backbone:** This research optimized the performance of the 6DRepNet backbone for head pose estimation through systematic adjustments. The model's ability to predict head orientation accurately was significantly enhanced, improving predictive accuracy, robustness, and reliability in real-world scenarios.
- **Analysis of loss function impact:** A comprehensive investigation evaluated the influence of various loss functions on head pose estimation accuracy. This analysis provided insights into their effectiveness in guiding the training process and optimizing model performance, guiding the selection of appropriate loss functions for improved accuracy and efficiency.

- **Proposal of viable gaze tracking enhancement:** Despite unsuccessful attempts to improve gaze tracking performance through L2CS-Net backbone modifications, a novel solution was proposed. Through innovative methodologies and strategic adjustments, this solution addresses limitations and offers a promising pathway for enhancing gaze tracking accuracy and effectiveness in intelligent vehicle systems.
- **Exploration of potential applications:** Beyond technical advancements, this study explores practical implications of intelligent vehicle driver gaze prediction. Discussions on potential applications, from driver distraction detection to emotion recognition, underscore the broader relevance and significance of findings. These insights facilitate the integration of gaze prediction technology into automotive safety and driver assistance systems, advancing intelligent transportation systems.

The following Chapter 2 will introduce relevant research on facial and gaze prediction. Chapter 3 will describe our experimental framework, improvement methods, and experimental results. Finally, Chapter 4 will conclude and discuss future prospects of this study.

II. Related Work

A. *Rotation Representation*

The key approach in angle prediction lies in using an appropriate rotation representation. Euler angles are the most commonly used and convenient rotation representation method, as shown in Figure 1. However, this representation is not optimal because it is susceptible to gimbal lock, wherein the same head pose appearance can result in multiple rotation parameters. Another method of rotation representation is quaternions, as depicted in Figure 2. Although quaternions are not affected by gimbal lock, their mirroring symmetry may lead to decreased predictive performance when learning full-range head poses. The best rotation representation method for predicting full-range head poses is using rotation matrices, as illustrated in Figure 3. Rotation matrices provide a continuous representation, and each rotation has unique parameters.

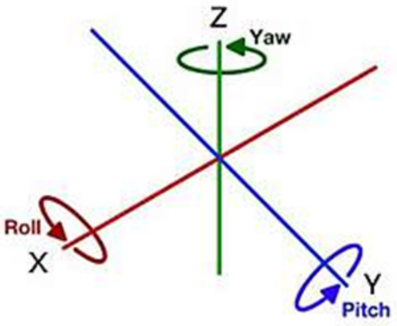


Fig. 1. Euler angle representation

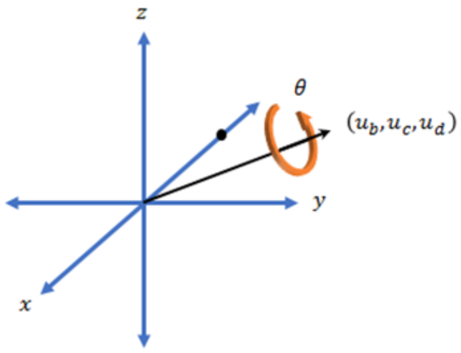


Fig. 2. Quaternion representation

$$\mathcal{R}_x(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix}$$

$$\mathcal{R}_y(\theta_y) = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix}$$

$$\mathcal{R}_z(\theta_z) = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Fig. 3. Rotation matrix representation

B. *Head Pose Predictions*

The current commonly used methods for head pose prediction are typically categorized into landmark-based and landmark-free approaches. Landmark-based methods [3] initially detect facial keypoints and subsequently establish correspondences between these keypoints and a 3D head model to recover the 3D head pose. While this method can yield highly accurate results, it heavily relies on the correct prediction of keypoints. Therefore, poor-quality keypoints caused by occlusion and extreme rotations can compromise accurate head pose estimation. On the other hand, landmark-free methods like HopeNet [4] overcome this issue by directly estimating head pose, which often aids deep neural networks in formulating orientation prediction as an appearance-based task.

In the context of appearance-based head pose prediction, 6DRepNet [5] proposes a landmark-free end-to-end head pose prediction method. This approach addresses the issue of ambiguous rotation labels by introducing rotation matrices, where a nine-parameter matrix facilitates full-pose prediction. Additionally, the paper introduces a continuous 6D matrix representation, which can be transformed into rotation matrices in subsequent tasks to achieve efficient and stable prediction methods, as shown in Figure 4.

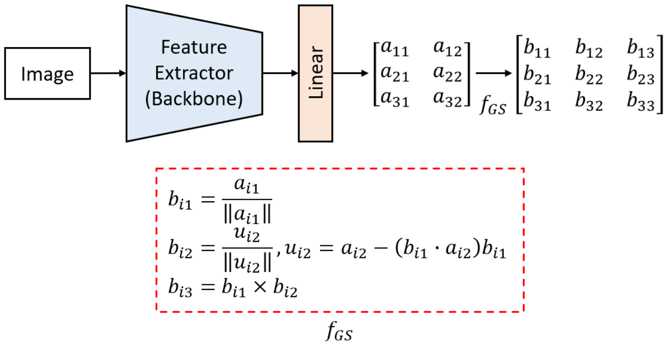


Image
Feature Extractor (Backbone)
Linear
$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \xrightarrow[f_{GS}]{} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$
$b_{i1} = \frac{a_{i1}}{\|a_{i1}\|}$
$b_{i2} = \frac{u_{i2}}{\|u_{i2}\|}, u_{i2} = a_{i2} - (b_{i1} \cdot a_{i2}) b_{i1}$
$b_{i3} = b_{i1} \times b_{i2}$
f_{GS}

Fig. 4. Overview of 6DRepNet head pose prediction method

The method proposed by 6DRepNet allows for learning the complete rotational appearance, contrasting with previous approaches that restricted pose prediction to narrow angles to achieve satisfactory results. Additionally, the loss function used in this method is the proposed Geodesic Distance-Based Loss rather than the commonly used mean square error loss function, as shown in Equation (1).

$$\mathcal{L}_g = \cos^{-1}\left(\frac{tr\left(R_p R_{gt}^T\right) - 1}{2}\right) \tag{1}$$

Although the L2-Norm is commonly used as the loss function for tasks related to head pose, using the Frobenius norm to measure the distance between two matrices disrupts the geometric structure of the SO(3) manifold. Instead, the shortest path between two 3D rotation matrices is interpreted geometrically as the Geodesic Distance.

C. *Gaze Tracking*

The direction of human eye gaze has always been a crucial cue utilized in various applications such as human-computer interaction and virtual reality. Although significant progress has been made in predicting gaze direction using deep learning through convolutional neural network methods, predicting gaze on non-wearable devices remains a challenging problem due to the uniqueness of eye appearance, lighting conditions, and the diversity of head poses and gaze directions.

Most CNN-based gaze estimation models predict 3D gaze as the gaze direction in spherical coordinates (Yaw, Pitch). Training loss functions typically employ mean square error (MSE) loss to penalize the network. L2CS-Net [6] proposes a CNN-based gaze tracking model for predicting gaze direction in unconstrained environments. The network utilizes ResNet50 as the main network architecture, regressing Yaw and Pitch angles separately during training to enhance the accuracy of each angle prediction, thereby improving the overall gaze prediction performance. Additionally, the network employs two identical losses, each a combination of Cross Entropy Loss and Mean-Squared Error, to improve network learning and increase its generalization.

In terms of gaze prediction, L2CS-Net does not directly predict continuous gaze angles; instead, it uses a Softmax Layer with Cross Entropy to predict discrete gaze classifications. It then transforms the discrete gaze prediction results into continuous gaze angles and incorporates mean square error into the output to improve gaze prediction accuracy. The overall gaze tracking approach is illustrated in Figure 5.

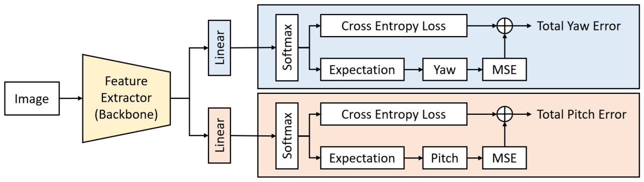


Fig. 5. Overview of L2CS-Net gaze tracking prediction method

D. *Decision Tree*

Decision trees generate a rule tree based on training data and use the learned rules to predict new samples. Decision tree algorithms can evaluate the quality of branches using various methods such as Information Gain, Gain Ratio, Gini Index, etc. By identifying suitable rules from the training data, a rule tree is ultimately generated to make decisions, with the aim of maximizing information gain for each decision, as depicted in Figure 6. In decision-making at higher levels of the tree, features with greater influence on the final decision are first considered. Subsequently, as the tree descends, the most suitable decision factors are identified from these features until the maximum depth is reached, at which point tree growth ceases.

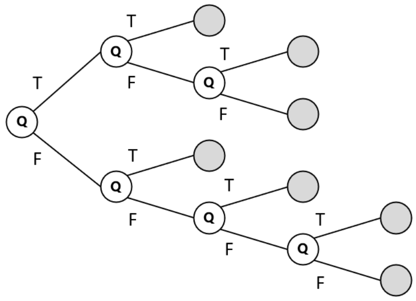


Fig. 6. Decision tree diagram

The generation of decision trees follows a greedy approach to determine each layer's question, aiming to make each branch more distinctly represent its corresponding category after classification. However, assessing the quality of each decision requires relying on measures of impurity. Objective criteria to determine each branch of the decision tree are crucial, necessitating an evaluative metric to assist in decision-making. Decision tree algorithms can employ various metrics to evaluate the quality of branches, with common measures of decision impurity including Information Gain, Gain Ratio, and Gini Index, among others. The objective of this evaluation method is to derive a set of decision rules from the training data that maximize information gain for each decision.

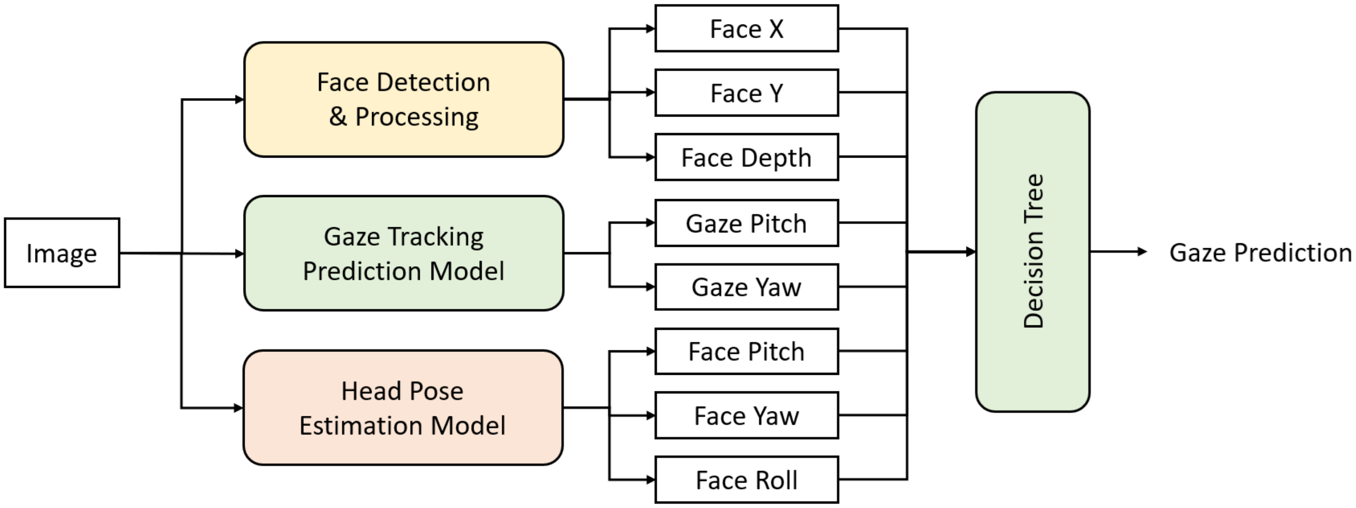


Fig. 7. Overview of the Intelligent Vehicle Driver Gaze Prediction System

The training process of decision trees involves continually searching for features to make decisions, attempting to group data into the same category as much as possible while minimizing disorder. Although increasing the depth of decision trees can enhance accuracy, it may also lead to overfitting issues. A well-trained decision tree model can visualize its structure, offering relatively high interpretability. Additionally, compared to other machine learning models, decision trees exhibit clear decision stages due to their tree-like structure, resulting in rapid execution speed, making them suitable for real-time applications.

E. Intelligent Vehicle Driver Gaze Prediction System

The Intelligent Vehicle Driver Gaze Prediction Auxiliary System is a screen-based gaze prediction model developed using 6DRepNet and L2CS-Net for facial pose prediction and human eye gaze prediction, respectively. This model primarily utilizes the outputs of 6DRepNet and L2CS-Net as features, incorporating user facial coordinates and the distance between the user's face and the non-wearable camera as additional features. The model employs decision tree algorithms to learn to predict the screen area the user is viewing, replacing the prediction of the gaze area for drivers. The overall architecture is illustrated in Figure 7.

To predict the gaze region a user is focusing on, the Intelligent Vehicle Driver Gaze Prediction Auxiliary System uses the Yaw, Pitch, and Roll values obtained from facial pose prediction, along with the Yaw and Pitch values from eye gaze prediction. These values, combined with the facial x and y coordinates and the distance from the camera, are used as input to train a decision tree model. The "Face Detection and Processing" module detects five facial points using a face detection model: the left eye, right eye, nose, left side of the lips, and right side of the lips. The nose coordinates represent the facial x and y coordinates. The distance between the face and the camera is calculated using the method shown in Figure 8.

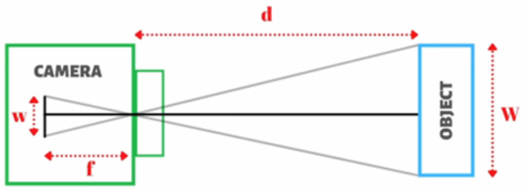


Fig. 8. Illustration of distance between object and lens

In the formula, f represents the focal length, w denotes the size of the object on the screen, d indicates the distance between the object and the camera, and W stands for the size of the object in the real world. W is set as the distance between the two eyes, with the average distance between human eyes statistically recorded as 6.3 centimeters. The value of w represents the distance between the two eyes on the screen, measured in pixels. To calculate the distance between the object and the camera, it is necessary to first determine the camera's focal length f. In this study, we can obtain the camera's focal length f by fixing the distance between the object and the camera d and using equations (2) and (3).

$$f = (w \cdot d)/W \tag{2}$$

$$d = (W \cdot f)/w \tag{3}$$

Since the camera's focal length remains constant, there is no need to calculate it in real-time. After obtaining the camera's focal length f, we can estimate the distance between the face and the camera based on the distance between the two eyes. It is worth noting that because the interocular distance W used to estimate the camera's focal length is an average distance obtained through statistics, there may be some error in calculating the distance between the face and the camera.

III. Experiments and Results

A. *Datasets*

In the training of head pose prediction, our primary dataset is 300W-LP [7], consisting of 66,225 facial samples, further augmented to 122,415 images through image flipping. For testing, we utilize the AFLW2000 dataset [8], which comprises the first 2000 images from the AFLW dataset. These images are annotated with real 3D facial data and corresponding 68 landmarks, exhibiting significant variations, diverse lighting, and occlusion conditions. In our study, we conduct model training using the 300W-LP dataset and evaluate it on 1,969 images from AFLW2000.

For training in predicting human eye gaze, we employ two environment-independent datasets to train and evaluate our model: Gaze360 and MPIIGaze. Gaze360 [9] offers the most extensive 3D gaze annotations, covering a maximum range of 360 degrees. It includes 238 subjects of various ages, genders, and races, captured in different indoor and outdoor environments using a multi-camera system. MPIIGaze [10] provides 213,659 images captured over several months from 15 subjects during their daily activities. Hence, it encompasses images with diverse backgrounds, times, and lighting conditions, making it suitable for unrestricted gaze prediction. Regarding image collection, software is used to gather images, requiring participants to look at randomly moving points on a laptop screen.

In the training for predicting the screen area users are viewing, we divide the computer screen into nine regions using a 3x3 grid. We collect our custom dataset through the laptop screen camera, capturing 200 images for each region for training, totaling 1800 training images. Additionally, we use 50 images per region for testing, amounting to 450 test images.

B. *Data Preprocessing*

For the 300W-LP and AFLW2000 datasets, we followed the preprocessing strategy of other methods [11], [12], retaining only the images with Euler angles between -99° and 99°. For the Gaze360 and MPIIGaze datasets, we normalized the images in both datasets using the same method as described in [13]. This process involved applying rotations and translations to simulate a virtual camera, eliminating the roll angle of the head while keeping the distance between the virtual camera and the center of the face constant. Additionally, we discretized the continuous gaze directions (Yaw, Pitch) into a binary-labeled discrete representation in each dataset, allowing classification based on the range of gaze annotations. Therefore, both datasets have two different target annotations: continuous labels and discrete labels, making them suitable for combined regression and classification losses.

For our custom screen gaze area dataset, we simply divided it into nine folders corresponding to the nine regions on the computer screen. We estimated facial angles and gaze directions using the trained facial and gaze prediction models and added eight features, including facial coordinates and distance from the camera. The eight features of each image, along with their corresponding folder names, served as the training input and output data for the decision tree model.

C. *The Head Pose Estimation Performance With Different Backbone Architectures*

For head pose estimation, we selected the Vision Transformer (ViT) and Swin Transformer architectures to compare with the original RepVGG backbone used in 6DRepNet. The results of our implementation are depicted in Figure 9.

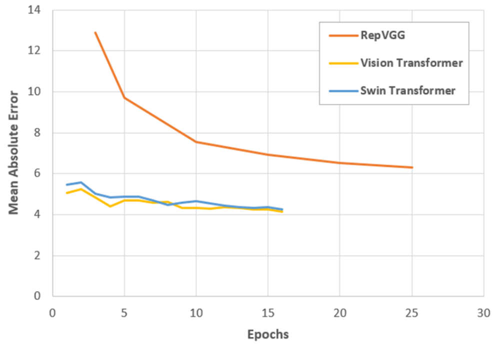


Fig. 9. The head pose estimation performance with different backbone architectures

As illustrated in Figure 9, although the original 6DRepNet paper reports an average absolute error (MAE) of 3.97, our training yielded an MAE of 6.3, suggesting that the original model is insufficient for real-world applications. Consequently, we replaced the RepVGG backbone with Vision Transformer and Swin Transformer architectures. For a robust comparison, we employed the base variants of both models, pretrained on the ImageNet dataset. Our findings demonstrate that the Transformer-based 6DRepNet models reduced the MAE to below 6 after a single training iteration, significantly outperforming the RepVGG-based 6DRepNet. Ultimately, we trained the Vision Transformer and Swin Transformer models to achieve MAEs of 4.1563 and 4.2513, respectively. Given that the Vision Transformer model exhibited superior performance in our study, we adopted the Vision Transformer-based 6DRepNet for subsequent research and applications.

D. *The Head Pose Estimation Performance With Different Loss Functions*

Similarly, in the context of head pose estimation, we compared the performance of Mean Squared Error (MSE) Loss and L1 Loss with the original Geodesic Distance-Based Loss employed in 6DRepNet. The implementation results are presented in Figure 10.

The original 6DRepNet paper highlights that the Geodesic Distance-Based Loss provides a more accurate measurement of the distance between two rotation matrices. Our experimental results indicate that the differences in performance between MSE Loss and L1 Loss when training the 6DRepNet model are negligible. As illustrated in Figure 10, the Geodesic Distance-Based Loss significantly outperforms the commonly used MSE Loss and L1 Loss, achieving markedly superior results.

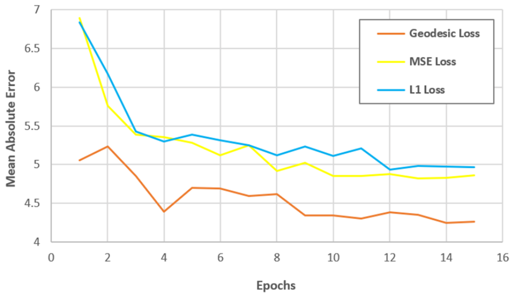


Fig. 10. The head pose estimation performance with different loss functions

E. *The Gaze Tracking Performance With Different Backbone Architectures*

In the realm of eye gaze prediction, we opted to substitute the primary ResNet50 backbone of L2CS-Net with Vision Transformer (ViT) and Swin Transformer models for a comparative analysis. The outcomes of our implementations are delineated in Table 1.

TABLE I
THE GAZE TRACKING PERFORMANCE WITH DIFFERENT BACKBONE ARCHITECTURES

Methods	MAE
L2CS-Net (Paper Result)	10.41
L2CS-Net (ResNet50)	10.98
L2CS-Net (Vision Transformer)	11.77
L2CS-Net (Swin Transformer)	11.65

As discerned from Table 1, although the original research on L2CS-Net delineated an average absolute error (MAE) of 10.41 on the Gaze360 dataset, our training regimen yielded an MAE of 10.98, closely resembling the original findings. In the pursuit of heightened performance, we also explored Transformer-based architectures. Employing Vision Transformer and Swin Transformer models akin to those utilized in head pose estimation, our findings reveal that Transformer-based models exhibit inferior performance compared to their CNN-based counterparts in gaze prediction tasks, albeit necessitating prolonged training intervals.

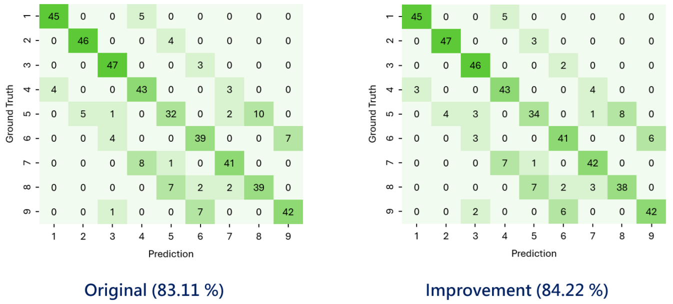
Despite extensive efforts to enhance the L2CS-Net backbone for gaze tracking, observed improvements in model performance were not achieved. Nonetheless, I proposed a pragmatic solution to address these challenges. In the context of limited datasets, particularly relevant to Vision Transformers (ViTs) and Swin Transformers, it is evident that robust feature learning necessitates ample data volumes. Insufficient data may lead to suboptimal model performance, underscoring the importance of dataset augmentation techniques.

To mitigate the constraints posed by small datasets, I recommended augmenting the dataset using established methodologies such as data augmentation. This strategy aims to enrich the training data, thus enhancing the model's ability to generalize and perform effectively across diverse scenarios. Furthermore, the phenomenon of catastrophic forgetting, a potential consequence of fine-tuning neural networks with additional data, poses a significant challenge. This issue can lead to overfitting and a deterioration in performance, thereby undermining the efficacy of the model. In response, I proposed the implementation of Parameter-Efficient Fine-Tuning (PEFT) methods, leveraging techniques such as Adapters or LoRA. These approaches selectively adjust model parameters to mitigate forgetting while adapting to new data, ensuring robust performance without compromising generalization capabilities. Despite encountering these challenges, the original L2CS-Net architecture remains the preferred choice for real-time gaze tracking applications, ensuring continuity and stability in the pursuit of optimal performance.

Additionally, beyond Gaze360, we conducted training on the MPIIGaze dataset. It is noteworthy that although our models attained results proximate to the original paper on MPIIGaze, they proved ineffectual in real-time eye gaze tracking endeavors. We conjecture that this challenge stems from the Gaze360 dataset employing entire face images for training to prognosticate gaze direction, while MPIIGaze exclusively employs eye images. Given the prevalence of smaller-angle eye images in MPIIGaze, the models exhibited diminished performance when forecasting large-angle gaze directions. Consequently, for subsequent real-time gaze prediction tasks, we exclusively relied on the eye gaze prediction model trained on the Gaze360 dataset.

F. *Quantitative Comparison*

Figure 11 presents the prediction outcomes of the Intelligent Vehicle Driver Gaze Prediction System, depicted through a confusion matrix. The graph indicates that the most prevalent misclassifications in each region of our model's predictions tend to occur above and below. Our research reveals that regions closer to the camera exhibit an accuracy rate exceeding that of other areas by 10%. Moreover, the central position of the screen is notably more susceptible to misclassifications compared to other regions. Despite focusing solely on enhancing the model performance of head pose estimation in this project, we observed a 1-percentage point improvement in the overall success rate of prediction across the entire gaze prediction auxiliary system. This underscores the value of individually refining the models for head pose estimation and gaze tracking.



Ground Truth \ Prediction	1	2	3	4	5	6	7	8	9
1	45	0	0	5	0	0	0	0	0
2	0	46	0	0	4	0	0	0	0
3	0	0	47	0	0	3	0	0	0
4	4	0	0	43	0	0	3	0	0
5	0	5	1	0	32	0	2	10	0
6	0	0	4	0	0	39	0	0	7
7	0	0	0	8	1	0	41	0	0
8	0	0	0	0	7	2	2	39	0
9	0	0	1	0	0	7	0	0	42

Original (83.11 %)

Ground Truth \ Prediction	1	2	3	4	5	6	7	8	9
1	45	0	0	5	0	0	0	0	0
2	0	47	0	0	3	0	0	0	0
3	0	0	46	0	0	2	0	0	0
4	3	0	0	43	0	0	4	0	0
5	0	4	3	0	34	0	1	8	0
6	0	0	3	0	0	41	0	0	6
7	0	0	0	7	1	0	42	0	0
8	0	0	0	0	7	2	3	38	0
9	0	0	2	0	0	6	0	0	42

Improvement (84.22 %)

Fig. 11. Improvements in Gaze Region Prediction

IV. Conclusion

A. *Future Work*

Addressing the intricacies of Level-3 autonomous driving, a pivotal focus area is the seamless transfer of control authority between automated systems and human drivers. Integrating gaze prediction with segmentation and depth estimation offers potential for enhanced hazard perception and informed control handover decisions.

Expanding gaze tracking integration into existing in-vehicle driver assistance systems, such as distraction and emotion detection, holds promise for developing more context-aware driver assistance frameworks. Furthermore, extending gaze tracking to exterior driving assistance systems, particularly in road segmentation and depth prediction, could significantly bolster environmental awareness and driving safety.

Lastly, advancing condition assessment methodologies for Level-3 autonomous driving transitions is crucial. This involves refining algorithms for assessing driver readiness and environmental conditions to ensure safe and efficient transitions between automated and manual driving modes.

B. *Contribution*

In conclusion, this project has significantly advanced the field of intelligent vehicle driver gaze prediction. By refining the 6DRepNet backbone for head pose estimation, we have achieved notable improvements in model performance, enabling more accurate determination of head orientation. This enhancement enhances the overall robustness and reliability of gaze prediction systems, crucial for ensuring driver safety in autonomous and semi-autonomous vehicles. Furthermore, our comprehensive analysis of various loss functions' impact on head pose estimation accuracy has yielded valuable insights into the intricacies of model training and optimization. This knowledge contributes to the ongoing refinement of predictive models, enhancing their effectiveness in real-world applications.

Despite encountering challenges in improving gaze tracking performance through modifications to the L2CS-Net backbone, our proposal of a viable solution underscores the study's commitment to innovation and problem-solving. This resilience is essential for driving progress in complex research domains such as intelligent transportation systems. Finally, our exploration of potential applications for intelligent vehicle driver gaze prediction highlights the broader implications of our research. From driver distraction detection to emotion recognition and beyond, the integration of gaze tracking technology holds promise for revolutionizing various aspects of automotive safety and human-machine interaction.

In summary, this study not only deepens our understanding of driver gaze prediction but also offers practical solutions and insights that have the potential to significantly impact the future of automotive technology and driver assistance systems.

V. Declaration of Originality

In accordance with academic standards and ethical guidelines, this section attests to the originality of the contents presented in this report. It confirms that the materials herein have not been previously utilized for any other assignment, including but not limited to a thesis or coursework from another academic endeavour completed prior to the current semester.

References

[1] A. Tawari, K. H. Chen and M. M. Trivedi, ”Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation,” 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2014, pp. 988-994.

[2] L. Fridman, J. Lee, B. Reimer, T. Victor. ””Owl” and ”Lizard”: Patterns of Head Pose and Eye Pose in Driver Gaze Classification.” arXiv preprint arXiv:1508.04028, 2015.

[3] P. Werner, F. Saxen, and A. Al-Hamadi, ”Landmark based head pose estimation benchmark and method,” in 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 3909–3913.

[4] N. Ruiz, E. Chong and J. M. Rehg, ”Fine-Grained Head Pose Estimation Without Keypoints,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 2155-215509.

[5] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, ”6D Rotation Representation For Unconstrained Head Pose Estimation,” arXiv preprint arXiv:2202.12555.

[6] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, ”L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments,” arXiv preprint arXiv:2203.03339.

[7] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li, ”Face Alignment Across Large Poses: A 3D Solution,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 146–155, 2016.

[8] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, ”High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 787–796.

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

[10] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, ”Mpiigaze: Real-world dataset and deep appearance-based gaze estimation,” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 1, pp. 162–175, 2017.

[11] N. Ruiz, E. Chong, and J. M. Rehg, ”Fine-Grained Head Pose Estimation Without Keypoints,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2155–215509, 2018.

[12] Z. Cao, Z. Chu, D. Liu, and Yingjie Chen, ” A Vector-based Representation to Enhance Head Pose Estimation,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2021, pp. 1188–1197.

[13] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, ”It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation,” in Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017, pp. 2299–2308.