Domain adaptation for driver's gaze mapping for different drivers and new environments

Ulziibayar Sonom-Ochir [a,b,*], Stephen Karungaru [a,2], Kenji Terada [a,3], Altangerel Ayush [b,4]

[a] Department of Information Science and Intelligent Systems,Tokushima University, Japan
[b] Department of Information Technology, Mongolian University of Science and Technology, Mongolia
[1] ulziibayar.s@gmail.com; [2] karunga@is.tokushima-u.ac.jp; [3] terada@is.tokushima-u.ac.jp; [4] a.altangerel@must.edu.mn
* corresponding author

ARTICLE INFO

Article history
Received June 14, 2023
Revised December 19, 2023
Accepted December 23, 2023
Available online February 29, 2024

Keywords
Gaze mapping
Domain adaptation
Visual attention
Gaze regions

ABSTRACT

Distracted driving is a leading cause of traffic accidents, and often arises from a lack of visual attention on the road. To enhance road safety, monitoring a driver's visual attention is crucial. Appearance-based gaze estimation using deep learning and Convolutional Neural Networks (CNN) has shown promising results, but it faces challenges when applied to different drivers and environments. In this paper, we propose a domain adaptation-based solution for gaze mapping, which aims to accurately estimate a driver's gaze in diverse drivers and new environments. Our method consists of three steps: pre-processing, facial feature extraction, and gaze region classification. We explore two strategies for input feature extraction, one utilizing the full appearance of the driver and environment, and the other focusing on the driver's face. Through unsupervised domain adaptation, we align the feature distributions of the source and target domains using a conditional Generative Adversarial Network (GAN). We conduct experiments on the Driver Gaze Mapping (DGM) dataset and the Columbia Cave-DB dataset to evaluate the performance of our method. The results demonstrate that our proposed method reduces the gaze mapping error, achieves better performance on different drivers and camera positions, and outperforms existing methods. We achieved an average Strictly Correct Estimation Rate (SCER) accuracy of 81.38% and 93.53% and Loosely Correct Estimation Rate (LCER) accuracy of 96.69% and 98.9% for the two strategies, respectively, indicating the effectiveness of our approach in adapting to different domains and camera positions. Our study contributes to the advancement of gaze mapping techniques and provides insights for improving driver safety in various driving scenarios.

1. Introduction

The primary cause of attention deviating from the road is driver distraction, which can be extremely dangerous to the lives of drivers, passengers, and pedestrians [1]. One of the most critical aspects of driver monitoring is observing their visual attention. This assists in keeping track of their status and avoiding traffic accidents caused by distractions. Therefore, it is essential to monitor the driving state, assess the distraction of the driver, and warn the driver. Recently, driver's gaze monitoring has been mentioned in a lot of research [2]. This related research can be categorized using both hardware-based and appearance-based techniques. The studies that are hardware-based often make use of additional equipment to evaluate the driver's visual attention [3]–[5]. Mizuno et al. [5] developed a system that detects visual attention by utilizing a gaze tracker and a device installed in the vehicle. Moreover, some researchers implemented a driver's gaze mapping system that uses an additional camera [6], [7]. While

these approaches have proven to be effective and reliable, they can be intrusive, expensive, and impractical for real-world use. Moreover, such systems can be challenging to operate and may cause driver fatigue. On the other hand, Appearance-based gaze estimation is one of the most effective methods, and it requires the least amount of additional equipment [8]–[12].

Deep learning (DL) technology and Convolutional Neural Networks (CNN) have significantly improved their performance in appearance-based gaze estimation tasks. Over the last few years, numerous versions of different CNN-based gaze estimating techniques have been presented and have achieved remarkable results [13]–[20]. In driver gaze region estimation, also known as gaze mapping, the methods consider eye images, face images, or a combination of the two. Also, some studies used images of the full appearance of the driver [21]–[23]. However, these existing systems in the field of gaze mapping encounter numerous common challenges. The predominant one is that the performance tends to deteriorate for the different drivers and environments. This can be caused by a variety of factors, such as domain disparities, insufficient data for the target driver, environmental influences, and disparate camera positions. In other words, although DL technology and CNN perform well on the learned data, the results are still not satisfactory for different car environments, camera positions, and domains [24]–[27]. To address these challenges, techniques for domain adaptation are employed to mitigate the negative effects of domain shifting, allowing the model to be applicable across different domains and environments. Wang et al. utilize an appearance discriminator and head pose classifier to achieve domain adaptation by adversarial learning [24]. Moreover, Cheng et al. proposed to enhance cross-domain performance without target domain data by eliminating gaze-irrelevant features [25]. Also, Liu et al. designed an outlier-guided collaborative domain adaptation method and tested cross-subjects between several datasets. More recently, Bao et al. have proposed a rotation-enhanced unsupervised domain adaptation technique for the problem of the lack of access to target domain labels in real-world situations [27]. To enhance unsupervised domain adaptation gaze estimation, Guo et.al. developed a novel embedding technique that incorporates prediction consistency loss. This innovative approach allows for the measurement of the variance between the source and target domains [28].

While the aforementioned methodologies have demonstrated notable advancements in enhancing gaze-related tasks, it is imperative to distinguish between gaze mapping and gaze estimation. Gaze estimation primarily involves determining the direction in which a person's gaze is focused, typically relying on technologies like eye-tracking to pinpoint the location of the eyes and infer the point of focus. Essentially, gaze estimation answers the question of 'where' the eyes are directed. On the other hand, gaze mapping extends beyond mere estimation, aiming to provide a comprehensive and spatial representation of the entire gaze behavior. Gaze mapping encompasses not only the predefined regions but also the dynamic patterns, head and body movements, and interactions of the gaze within a given environment. It seeks to create a detailed map or model that reflects how the individual's gaze traverses and engages with different elements in their surroundings. In the context of driver behavior, gaze mapping becomes particularly crucial for understanding not just the instantaneous points of focus but also the broader context of how the driver visually navigates through complex outdoor environments. This includes considerations for factors such as scanning the road, monitoring mirrors, and responding to dynamic stimuli. Despite significant progress in gaze estimation, achieving accurate and robust gaze mapping, especially in the challenging conditions presented by outdoor driving environments and without additional devices, remains a formidable task in the field of research and development.

In this paper, we present a domain adaptation-based solution to gaze mapping for robustness to different drivers and new environments. It's important to find ways to adapt to different domains and environments in a way that is effective and efficient, and this proposal has a lot of potential. We believe that our study has the following contributions.

- Accurate gaze mapping across various drivers is achievable using a simple dashboard camera.
- Self-calibration possibility for different camera positions in the same domains.
- Experimental results demonstrate that the proposed method reduces the gaze mapping error of the pre-trained adapted model and even has better performance on different drivers (cross-subject) and environments (different camera positions).

In our work, we achieve an accuracy that is at par or better than the state-of-the-art results for tested domain adaptation on gaze mapping from the Driver Gaze Mapping (DGM) dataset (our prepared dataset) to the open dataset Columbia Cave-DB [29]. In this paper, we have structured the content into distinct sections. Section 2 offers insight into the two strategies of our proposed method and the training and domain adaptation process. Section 3 outlines the diverse datasets of camera positions employed for training, the driver's gaze regions that were used, and the experiments conducted and compares them with existing studies. Finally, Section 4 presents our conclusions.

2. Method

2.1. Overview of the Proposed Method

The proposed method has three steps, pre-processing, facial feature extraction, and gaze region classification as shown in Fig. 1, represented by the blue arrow. The process of domain adaptation is represented by the red arrow, while the process of testing the proposed method is represented by the blue arrow. Furthermore, the fine-tuned network parameters are indicated by the dashed line.

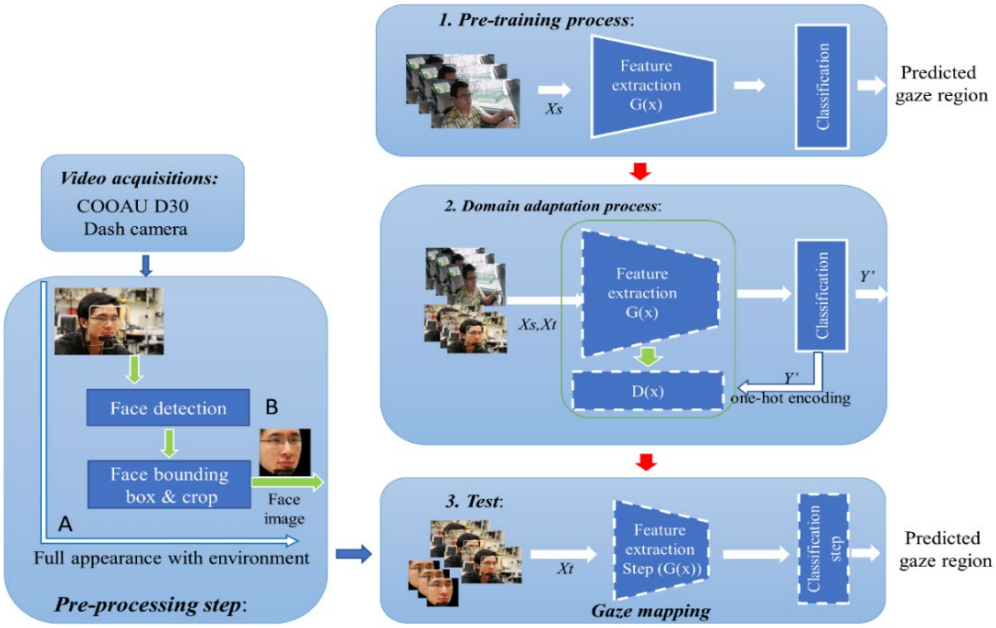

Fig. 1. Structure of the proposed method.

During training, we tried two pre-processing strategies for the input feature extraction step. The first strategy involved using an image of the driver's full appearance and the environment. This allowed us to skip the face detection and face bounding box & crop step and directly train the feature extraction from the input images, shown by line A, in Fig. 1. In the second strategy, we specifically detected the driver's face and used it as input for the feature extraction step, shown by line B, in Fig. 1. The facial feature extraction step involves extracting relevant facial features from images of pre-processing step. Finally, the gaze region classification step predicts one of the 13 predefined gaze regions using these features.

2.2. Domain Adaptation for Gaze Mapping

In this section, we will provide a detailed description of our proposed model, including the principles of the base model, the algorithmic steps, and the mathematical aspects. Our paper's main theoretical underpinning is that the model is designed to address challenges related to domain shift, leveraging adversarial training and transfer learning principles for unsupervised domain adaptation in gaze mapping (DGM dataset to Columbia Cave-DB). To provide a detailed explanation, let's begin by selecting the components of the proposed model structure.

First, we choose a discriminative base model, as we assume that when adapting a model from a source domain to a target domain, the discriminative aspects of the model are more important than the generative aspects. Then, the choice between shared and unshared weights depends on the nature of the adaptation problem. If the domains are expected to have similar characteristics, shared weights might be more appropriate. In our case, the target and source domains are quite different regarding participating drivers' environment and facial appearance. Hence, separate sets of model parameters are used for the source and target domains. This is because unshared weights allow the model to adapt more flexibly to domain-specific characteristics, which is important when there is a significant domain shift. Therefore, we chose an unsupervised domain adaptation method with unshared weights. Moreover, adversarial loss is another important component of our proposed model. It is a crucial component of unsupervised domain adaptation, particularly in methods that leverage domain adversarial training. We used separate sets of model parameters for the source and target domains, and therefore, we chose the GAN loss as the adversarial loss for our case. By combining unshared weight and GAN loss, we assumed that the model can adapt to the specific features present in each domain while minimizing the domain shift through adversarial training.

2.3. Training and Domain Adaptation

In this section, we will provide a detailed description of the adversarial training process of the feature extraction and classification steps, as shown in Fig. 1, represented by the red arrow. The training aims to achieve unsupervised domain adaptation for gaze mapping, specifically from the DGM dataset to Columbia Cave-DB. When images are from different distributions, a feature extractor maps them to different clusters in the feature space. To bring these clusters closer together, a conditional generative adversarial network (CGAN) [30] is used. In detail, we utilized the ResNet18 model [31] as the backbone model. Fig. 2. shows an overview of domain adaptation process.

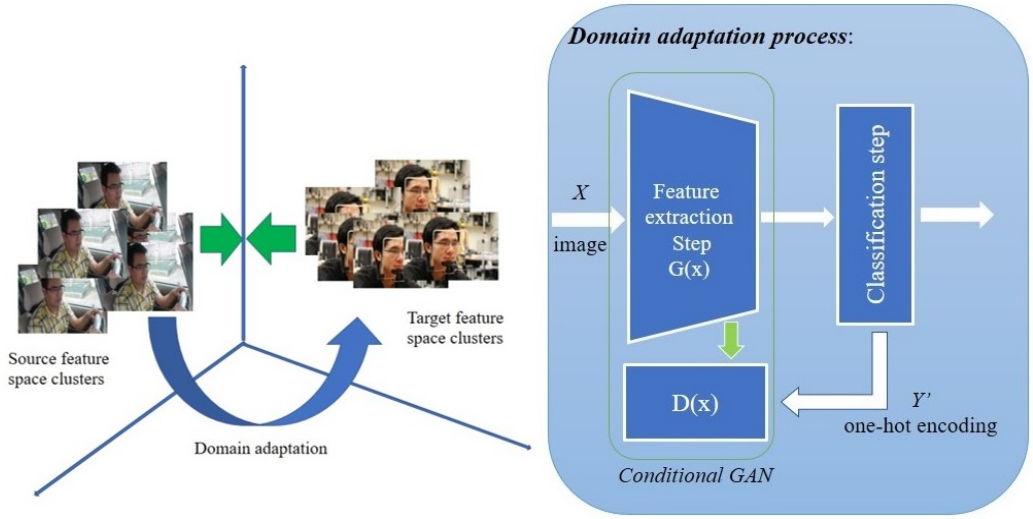

Fig. 2. An overview of domain adaptation

To modify the network, we replaced its final layer with a new fully connected layer consisting of 13 neurons. Additionally, a softmax layer was included on top of it. The model uses a feature extractor as a generator G(x), where x represents the input image, and an external multi-layer perceptron acts as a discriminator D(x), which determines whether the extracted feature is from the source or target domain. This classification is represented through one-hot encoding, Y(x). During each epoch, the discriminator is optimized first, to minimize the difference between D(G(x)) and Y(x) for all x in both domains. The generator is then optimized to confuse the discriminator, to minimize the difference between D(G(x)) and Y', where Y' represents the one-hot encoding for the source domain and x is from the target domain, as shown in Fig. 2. This process maps images from the target domain to a cluster that is closer to the cluster in the source domain's feature space. We provide a step-by-step algorithm that gives a mathematical overview of the key components and processes involved in unsupervised domain adaptation for gaze mapping (Fig. 3).

Algorithm 1: Training procedure

Inputs: Source domain images X_s with labels Y_s ; Target domain images X_t

Outputs: Feature extractor network *G;* Classifier network *C*

1. Initialize and pre-training:

 # This involves training the feature extractor network G and the classifier network C on the source domain data (X_s) with corresponding labels (Y_s). This is a standard pre-training step.

- ResNet18 network *G* with modified final layers (13 neurons + softmax), *G = ResNet18 (num_classes=13)*
- Multi-layer perceptron (MLP) discriminator *D, D = MLP(input_dim=G.output_dim, hidden_dim=128, output_dim=1)*
- Classifier network *C, C = MLP(input_dim=G.output_dim, hidden_dim=64, output_dim=13)*
- For each epoch:
  - For each image x in X_s
    - Compute features *f = G(x)*
    - Encode domain label *y = Y(x)*
    - Calculate loss L_s based on X_s and ground truth labels Y_s
    - Back-propagate L_s to update *G* parameters *θ*

***2. Feature Extraction and Adversarial Training*:**

 # After pre-training, the algorithm freezes the pre-trained feature extractor (G) and introduces a multi-layer perceptron (MLP) discriminator (D) for adversarial training.

- Freeze pre-trained feature extractor parameters *θ*
- For each epoch:
  - For each image x in X_s and X_t:
    - Compute features *f = G(x)*
    - Encode domain label *y = Y(x)*
  - Train discriminator *D* to minimize loss:
  - L_D *= -1/n * sum(y * log(D(f)) + (1-y) * log(1-D(f)))*

    # n: total number of images (including both source and target.
  - Train generator *G* to minimize loss:

    # n_t: number of target domain images.

    # Y': the one-hot encoded vector for the source domain.
  - L_G *= -1/n_t * sum(Y' * log(D(G(X_t))))*

3. Joint Fine-tuning and Classifier Training:

 # The feature extractor parameters (θ) are unfrozen for joint fine-tuning.

 # The algorithm then samples batches from both source ($X_s$and target (X_t)domains, extracts features using the fine-tuned G, and computes both classification loss (L_{cls}) and domain adaptation loss (L_{DA}) using Maximum Mean Discrepancy (MMD).

 # The combined loss ($L_{combined}$) includes both classification and domain adaptation components. The feature extractor (G) and classifier (C) parameters are updated based on this combined loss.

- Un-Freeze feature extractor parameters *θ*
- For fine_tune_epochs:
  - Sample a batch from source and target domains (X_s_batch, X_t_batch, Y_s_batch)
  - Extract features: $f_{X_s\_batch}$ *= G(X_s_batch)*, $f_{X_t\_batch}$ *= G(X_t_batch)*
  - Compute classification loss: L_{cls} *= C($f_{X_s\_batch}$).loss(Y_s_batch)*
  - Compute domain adaptation loss: L_{DA} *= MMD(f_{X_s}, f_{X_t})*
  - Combined loss: $L_{combined} = L_{cls} + \lambda * L_{DA}$
  - Back-propagate $L_{combined}$ to update both *G* and *C* parameters

4. Output:

 # The output includes the fine-tuned feature extractor (G) with domain adaptation and the classifier (C) trained on the source domain and fine-tuned by the target domain.

- Feature extractor network G(fine-tuned with domain adaptation)
- Classifier network C (trained on source domain and fine-tuned by target domain)

Fig. 3. Training procedure for for gaze mapping

The feature extractor parameters are frozen, and the classifier is trained on the source domain. Since the feature extractor is generalized, training on the source domain can enhance performance on the target domain. Furthermore, we will explain the proposed model in terms of process. One important aspect of our method is the adversarial training procedure.

3. Results and Discussion

3.1. Implementation details

We experimented with gaze mapping using the domain adaptation method and trained the model with specific parameters in both the source and target domains. For the feature extractor, the learning rates were set to 0.001 in the source domain and 0.0005 in the target domain. The classifier's learning rate was set to 0.001 in both domains. We set the adversarial loss weight and domain classifier weight to 0.1. The training was done with a batch size of 64 and 30 epochs. We initialized the feature extractor with a pre-trained model and used the Adam optimizer. We provided a step-by-step algorithm adversarial training in the following section. For more details on the training process, please refer to Algorithm 1.

3.2. Evaluation metrics

Strictly correct estimation rate (SCER) and Loosely correct estimation rate (LCER) are mostly used in our field research. In our study, the accuracy of gaze mapping also was measured based on the Strictly correct estimation rate (SCER) and the Loosely correct estimation rate (LCER).

$$SCER = \frac{Number\ of\ Strictly\ Correct\ Frames}{Total\ Nubmber\ of\ Frame} \tag{1}$$

SCER measures the ratio of the number of frames where the estimated gaze region is strictly correct (equivalent to the ground truth gaze region) to the total number of frames.

Strictly Correct Frame: A frame is considered strictly correct if the estimated gaze region is precisely equal to the ground truth gaze region.

$$LCER = \frac{Number\ of\ Frame\ with\ Estimated\ Region\ in\ (GT \cup N)}{Total\ Number\ of\ Frames} \tag{2}$$

LCER measures the ratio of the number of frames where the estimated gaze region is loosely correct (within the ground truth gaze region and neighboring regions) to the total number of frames.

Loosely Correct Frame: A frame is considered loosely correct if the estimated gaze region is placed within the ground truth gaze region or in one of the neighboring regions.

The numerator now represents the count of frames where the estimated gaze region is in the union of the ground truth gaze region (GT) and the set of neighboring regions (N).

3.3. Experimental datasets

training process, we utilized the DGM dataset as the source domain. We trained on this dataset and subsequently adapted and tested it to the Cave-DB dataset as the target domain. As a result of the training described in Section 2, our proposed method shown in Fig. 1 is prepared for testing on the target domain. In the preprocessing step, there are two modes available - full appearance image and face image, mentioned in Section 2.1. So, in this section, we will present and analyze the experimental results of strategies of the proposed method. Additionally, we conduct experiments on the DGM dataset, which includes different camera positions. It explores the possibility of adapting to different camera positions in the same domain for self-calibration tasks. Furthermore, we provided an analysis of the results obtained from the proposed model. This includes a discussion of the implications of these results, a comparison with existing methods, and the limitations of our study.

3.3.1. IDGM Dataset

The Driver Gaze Mapping (DGM) dataset was used for the gaze mapping task. This dataset features 13 distinct gaze regions and data collected from two different camera positions, as described in Fig. 4.

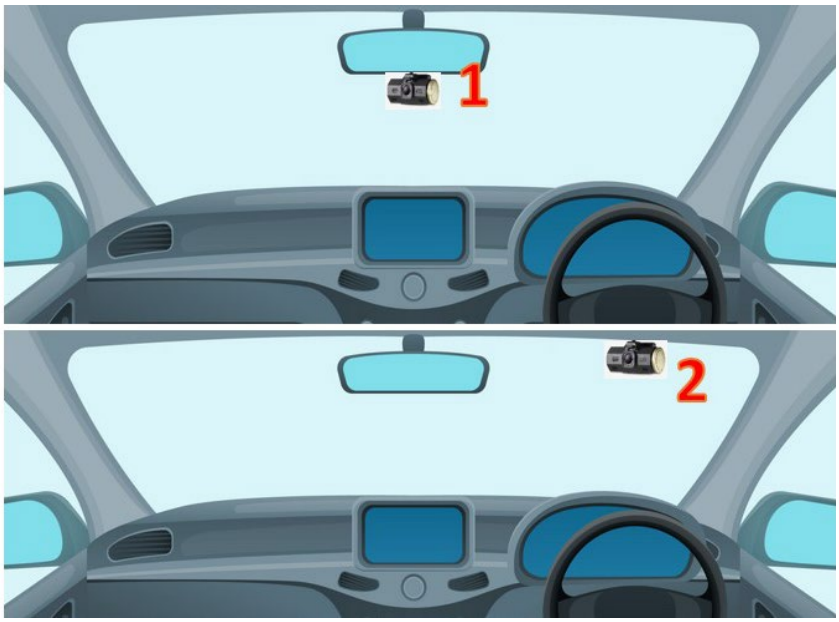

Fig. 4. Camera positions (1) bottom of the rear mirror (2) top-front of windshield

The dataset comprises the driver's gaze and information about the driving environment. The 13 predefined gaze regions, illustrated in Fig. 5, include the gaze region on the windshield, left and right-side mirrors, and left and right-side windows (regions 1-9). To avoid the risk of the driver causing an accident while observing the designated areas, we opted to capture images of various locations - including university campus roads and parking lots - while the vehicle was in motion. We used a COOAU-D30-1080P dual dash camera to take pictures in the morning, afternoon, and night, ensuring that we had a diverse set of images from different times of the day. The drivers gazed at 13 predefined gaze regions, and they were allowed to move their heads and bodies freely, simulating the naturalistic movements of a driver. Additionally, we have accounted for the neighboring regions adjacent to each gaze region.

Fig. 5. Predefined 13 gaze region

The dataset includes 12,285 images with 13 labels using camera position 1. Additionally, we collected 3900 images with the same labels using camera position 2, for the self-calibration task experiment.

3.3.2. Open dataset Cave-DB

We created a new dataset using the open dataset Columbia gaze dataset CAVE-DB for fair comparison. This enabled us to apply unsupervised classification to different domain shifts. This was done as previous studies [17], [19], [20] also evaluated their methods using SCER and LCER through CAVE-DB. The CAVE-DB contains a large gaze database of 56 individuals with 5880 images that vary in head poses and gaze directions. There are 105 gaze directions as 5 head poses with 21 gaze directions per head pose. From the database, we chose 13 gaze direction images considering the DGM dataset in Fig. 5. The examples of images with gaze regions are shown in Fig. 6.

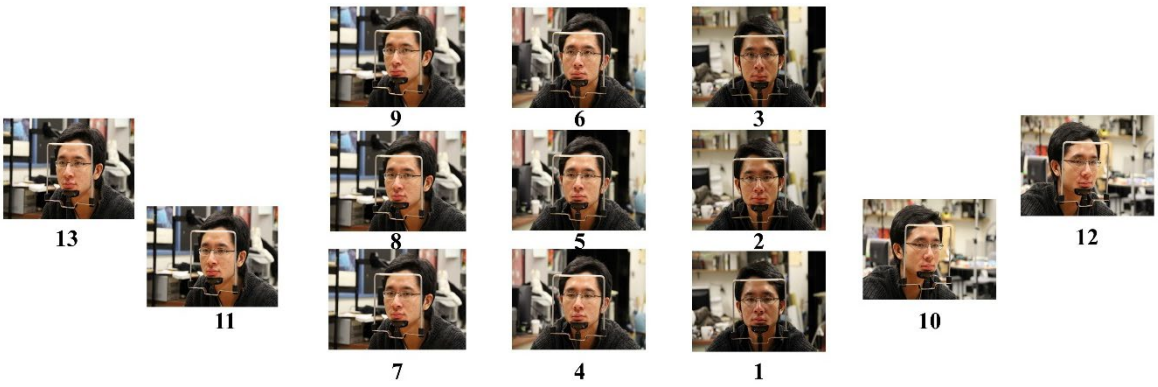

Fig. 6. Sample images selected from CAVE-DB

3.4. Experimental result

In this section, we have provided a detailed analysis of experimental results obtained from the proposed model. This includes a discussion of the implications of these results, a comparison with existing methods, and the limitations of our study. In this, we prepared the source and target datasets in the following ways: on different drivers in the same environment, on the same driver in different environments, and on different drivers in different environments, as shown in Fig. 7. As a result, domain adversarial training was performed on the above differentially trained datasets. As a result, we determined how different drivers, different environments, and different environments and different drivers affect the results of gaze estimation methods using domain adaptation. Also, during domain adaptation, we determined which of the driver's full appearance with environment images and face images were effective for adaptive training.

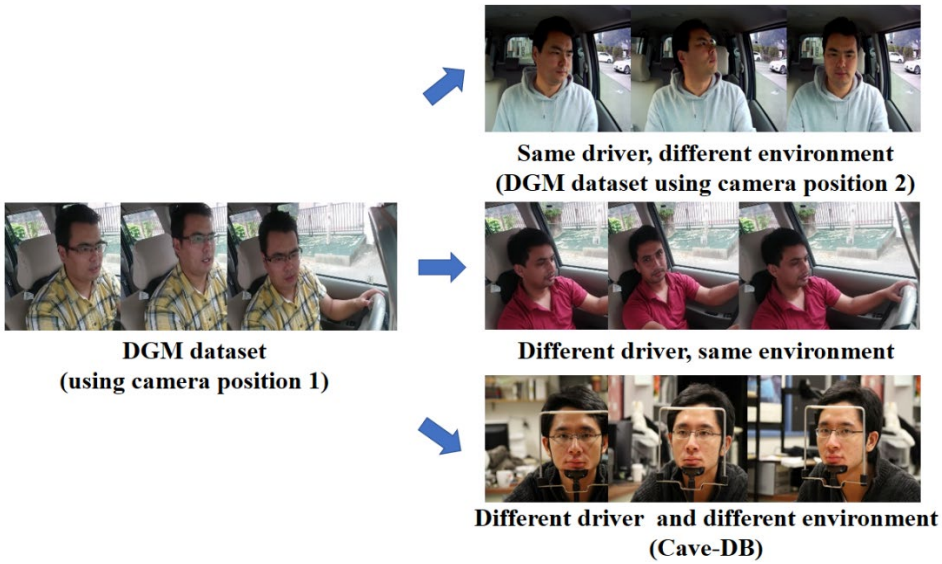

Fig. 7. Prepared datasets and domain adaptation versions

3.4.1. Experiment of Strategy A

As shown in Fig. 7, we organized domain adaptation training in 3 different versions using the driver's full appearance with environment images (Strategy A). First, we trained the DGM dataset using camera position 1 as the source domain, and camera position 2 as the target domain, using sets of datasets as shown in Table 1.

Table 1. Amount of datasets used in domain adaptation versions

Domain adaptation versions	Source	Target	Test
DGM-1 to DGM-2	12285 images with 13 labels	3900 images with 13 labels	1300 images with 13 labels
DGM-1 to DGM (diff.driver)	12285 images with 13 labels	3900 images with 13 labels	1300 images with 13 labels)
DGM-1 to Cave-DB	12285 images with 13 labels	3900 images with 13 labels	1300 images with 13 labels

In this experiment, we explored the possibility of learning from each other between datasets with the same driver or facial features but different camera positions. Based on the results, the average accuracy was 85%. In the experiment, it is evident from Fig. 8 that there is significant confusion between gaze regions 6 and 9, as well as between gaze regions 7 and 8. Furthermore, it can be observed that there is some confusion in regions with low head movement. Also, a small of confusion was formed between gaze regions 1 and 2, and gaze regions 8 and 11, which are regions that can be moved by the movement of the eyeball. This suggests a risk of confusion between gaze regions that require minimal head movement and require small changes in gaze direction. Although the confusion was between the aforementioned gaze regions, the feasibility of self-calibration was demonstrated using the domain adaptation method across different camera positions within the same domain.

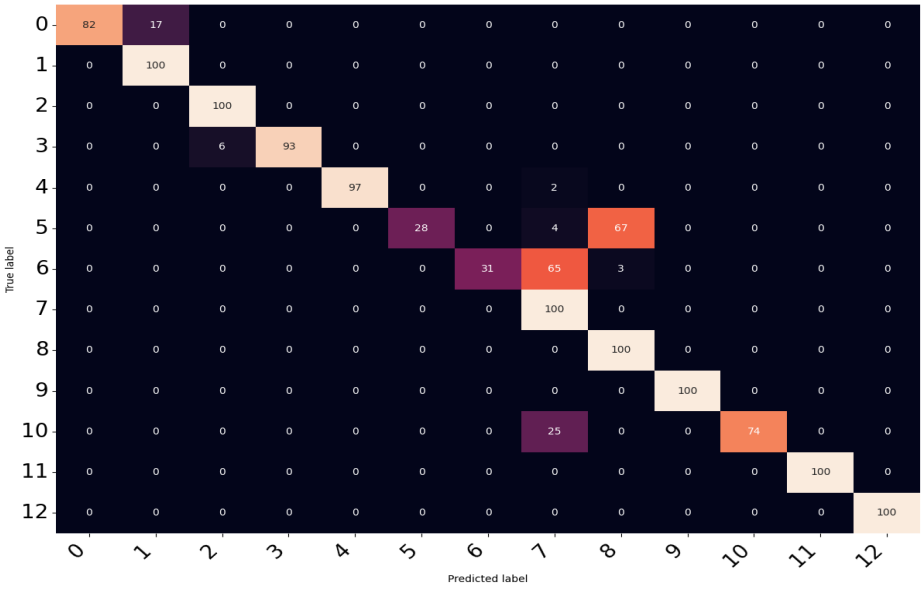

Fig. 8. Confusion matrix of strategy A on the same driver, different environment

Second, we trained the DGM dataset using camera position 1 as the source domain, and a different driver with the same environment as the target domain, using sets of datasets as shown in Table 1. In this experiment, we aimed to determine the adaptive performance of different domains in the same environment. According to the results of the experiment, the performance of each gaze region demonstrated that the minimum accuracy was 76% or more, and the average accuracy was 88.76%. This indicates that serious confusion has not occurred in each region. Also, it can be seen from Fig. 9 that the resulting confusion is usually observed with the neighboring gaze region.

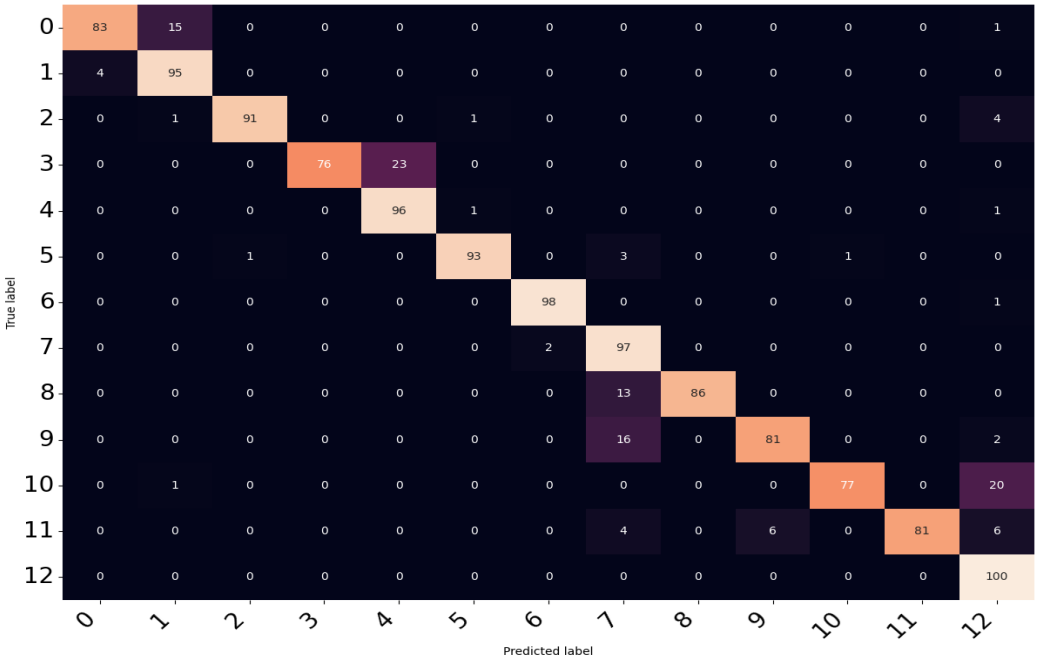

Fig. 9. Confusion matrix of strategy A on different drivers, same environment

Finally, we conducted an experiment where we used the DGM dataset as the source domain and the Cave-DB dataset as the target domain, using sets of datasets as shown in Table 1. The purpose of this experiment was to demonstrate how our proposed model can adapt to different domains and environments. The results showed that the target domain was classified with reasonable accuracy, except for gaze region 5 which was misclassified as neighboring gaze region 4. Apart from this, the results were reasonable, with an average accuracy of 81.38%, as shown in Fig. 10.

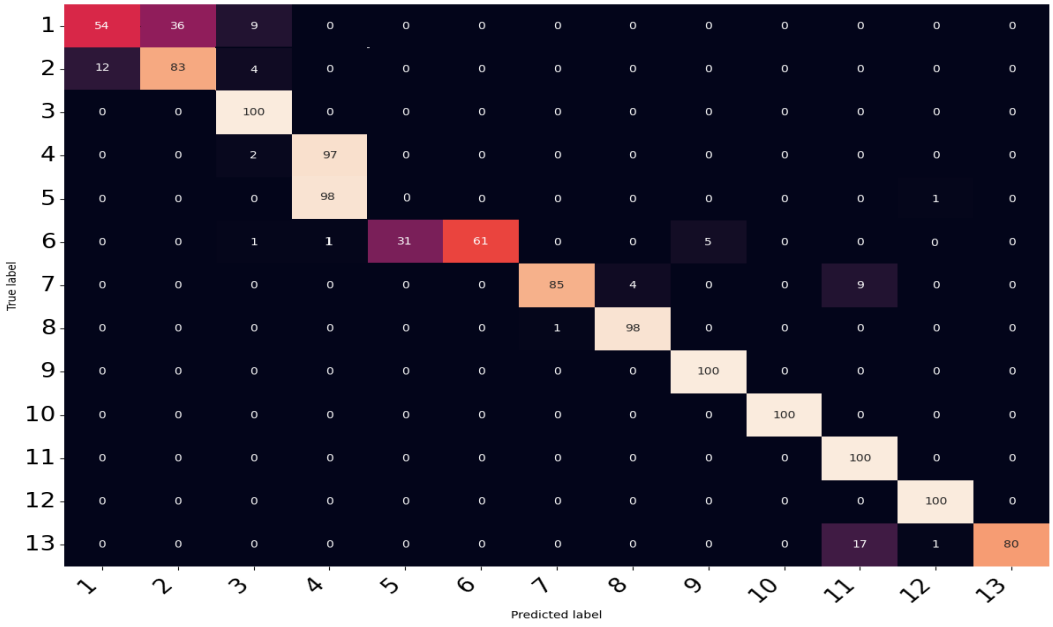

Fig. 10. Confusion matrix of strategy A on different drivers, different environments

3.4.2. Experiment of Strategy B

In this experiment, we trained the DGM dataset as the source domain and the Cave-DB dataset as the target domain by strategy B of pre-processing which uses a face image. The average SCER accuracy was 93.53% and the LCER rate was 98.9%, as shown in Fig. 11.

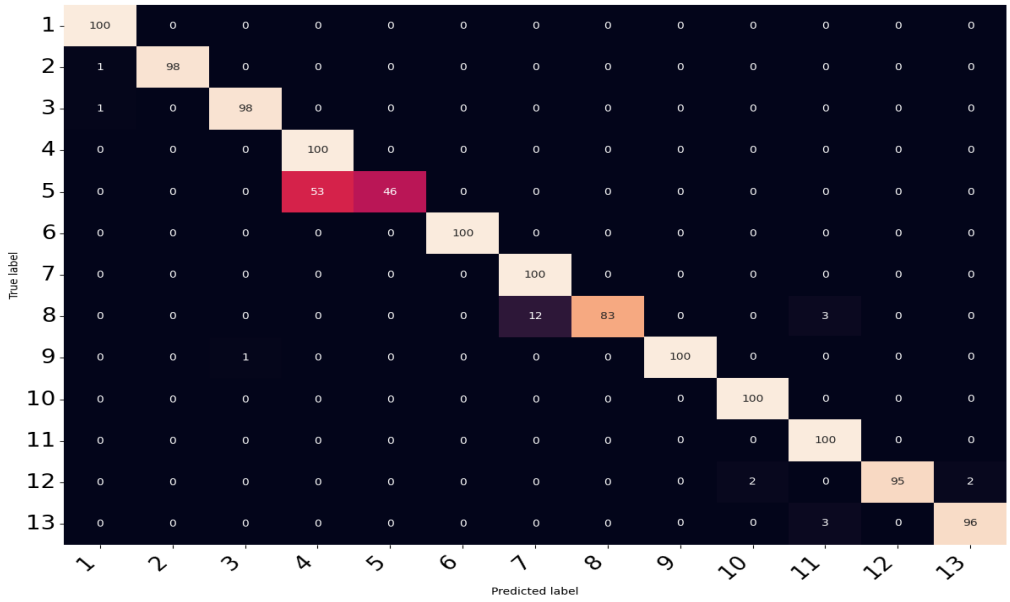

Fig. 11. Confusion matrix of strategy B on different drivers, and different environments

Based on the results, Strategy B proves to be more effective than Strategy A. The average SCER accuracy rate of Strategy B is 12.15% higher compared to Strategy A which uses the driver's full appearance image. Moreover, the experiment's findings indicated that there is more confusion when transitioning between gaze regions that require only slight head and eye movements, such as gaze regions 1 and 2. However, there seems to be less confusion when transitioning between gaze regions that require more significant head and eye movements. For example, the gaze regions of side mirrors can be mentioned.

Then, we tested on the DGM dataset, where camera position 1 was the source domain and camera position 2 was the target domain. We have achieved the following results in this experiment. The accuracy of strategy B of pre-processing which uses a face image was reasonable. On average, the accuracy of the SCER was 94.85%, as illustrated in Fig. 12. This indicates that Strategy B is also more efficient than Strategy A, with an average SCER accuracy rate that is 9.8% higher. As a result, strategy B proved to be more effective on the above two tasks.

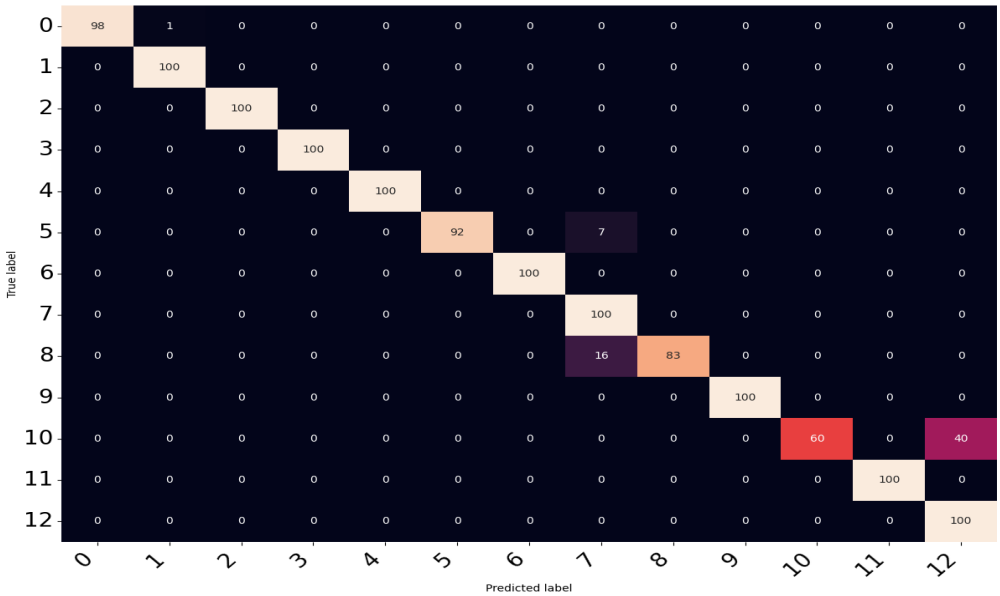

Fig. 12. Confusion matrix of strategy B on the same driver, different environment

3.4.3. Findings and discussion

During the above three domain adaptation experiments, shown in Table 2, various findings were observed. Firstly, it was discovered that accurate gaze mapping on different drivers can be performed using domain adaptation. Secondly, the position of different cameras in the same domain can self-calibrate. Additionally, the experimental results show that the proposed method can reduce gaze mapping errors. The findings also demonstrate that the proposed method can reduce the gaze mapping error of the pre-trained adapted model, and even perform better on different drivers (cross-subject) and environments (different camera positions). In addition, when analyzing the results of the above three confusion matrices, it was seen that the model is very stable only in domain transition without environmental change. These results underscore the effectiveness of our method in adapting to different domains. On the other hand, it was observed that it is comparable weak for the same domain and different environments (using different camera positions). In other words, we noticed that our domain adaptation model for gaze mapping, while robust for different domains, is affected by significant camera changes.

Table 2. Performance results on domain adaptation versions

Training versions	Full-appearance image		Face image	
	SCER	*LCER*	*SCER*	*LCER*
DGM-1 to DGM-2	85.00%	98.80%	94.85%	99.23%
DGM-1 to DGM (diff.driver)	88.76%	96.23%	-	-
DGM-1 to Cave-DB	81.38%	96.69%	93.53%	98.90%

This highlights the adaptability of our approach to diverse environments and even different camera positions for the same driver, indicating potential self-calibration capabilities. We also discovered that strategy B was more effective than strategy A in both of the given tasks. This indicates that strategy B is more successful in domain adaptive learning. In other words, we observed that the feature extraction step produces cleaner output as the environment's influence decreases.

3.5. Comparative experiments with existing studies

In this section, we conducted a comparison between the two strategies of our proposed methods and the other existing studies on the Cave-DB dataset. All of these studies were conducted using the same database and used the same number of gaze regions. The results are shown in Table 3. We found that our results were slightly better than those of the previous studies.

Table 3. Comparison of the Existing Studies on Cave-DB

Methods	Accuracy /%/	
	SCER	*LCER*
Choi et al. [19]	53.09%	88.66%
Naqvi et al. [17]	77.70%	96.31%
Lee et al. [20]	44.00%	85.07%
Strategy A of our proposed method	81.38%	96.69%
Strategy B of our proposed method	93.53%	98.90%

In addition, it should be noted that Vora et al. [22] and Shah et al. [32] have conducted state-of-the-art studies in addition to the ones mentioned earlier. However, they only used 6 and 7 gaze regions respectively, which were Forward, Right, Left, Center Stack, Rearview mirror, and speedometer. It is difficult to compare the results of these studies with the ones that have 13 gaze regions as there are fewer gaze regions and less possibility of confusion. In the study by Vora et al., SqueezeNet was used, which is the best method for Face Embedded FoV, with an accuracy of 89.37% [22]. Also, Shah et al. used a Driver gaze estimation method based on Deep Learning, which has an accuracy of 91% [32]. However,

our proposed method's strategy B obtained matching results for more gaze regions, which can be considered decent results comparable to state-of-the-art studies.

4. Conclusion

This paper introduces a novel gaze mapping solution designed to enhance robustness across diverse drivers and environmental conditions. By integrating pre-processing, facial feature extraction, and gaze region classification, our method explores two feature extraction strategies, leveraging both full appearance and facial focus. Through unsupervised domain adaptation employing GAN loss and unshared weight, we successfully align feature distributions between source and target domains. Experimental evaluations on the Driver Gaze Mapping (DGM) dataset and the Columbia Cave-DB dataset demonstrate a notable reduction in gaze mapping error and superior performance compared to existing methods. Our approach achieves an average Strictly Correct Estimation Rate (SCER) accuracy of 81.38% and 93.53%, and a Loosely Correct Estimation Rate (LCER) accuracy of 96.69% and 98.9% for the two strategies, respectively. These results underscore the effectiveness of our method in adapting to different domains. Furthermore, we attain an average SCER accuracy of 85.00% and 94.84%, and LCER accuracy of 98.80% and 99.23% for the two strategies, respectively. This highlights the adaptability of our approach to diverse environments and even different camera positions for the same driver, indicating potential self-calibration capabilities. Our study significantly contributes to the evolution of gaze mapping techniques, providing valuable insights for enhancing driver safety in a variety of driving scenarios. The achieved accuracies demonstrate the practical effectiveness of our proposed solution, positioning it as a promising advancement in the field. Looking forward, this work lays the groundwork for future research in gaze mapping and its applications.

Declarations

Author contribution. Ulziibayar Sonom-Ochir implemented the methods and conducted the experiments under the supervision of Stephen Karungaru, Kenji Terada, and Altangerel Ayush. Ulziibayar Sonom-Ochir and Stephen Karungaru were involved in the writing of the paper. All the authors read and approved the final manuscrip.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] "Pedestrians, cyclists among main road traffic crash victims," *World Health Organization*, 2010. [Online]. Available at: https://www.who.int/news/item/11-12-2010-pedestrians-cyclists-among-main-road-traffic-crash-victims.

[2] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver Inattention Monitoring System for Intelligent Vehicles: A Review," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 596–614, Jun. 2011, doi: 10.1109/TITS.2010.2092770.

[3] I. Dua, A. U. Nambi, C. V. Jawahar, and V. N. Padmanabhan, "Evaluation and Visualization of Driver Inattention Rating From Facial Features," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 2, no. 2, pp. 98–108, Apr. 2020, doi: 10.1109/TBIOM.2019.2962132.

[4] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi, "Driver Gaze Tracking and Eyes Off the Road Detection System," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2014–2027, Aug. 2015, doi: 10.1109/TITS.2015.2396031.

[5] N. Mizuno, A. Yoshizawa, A. Hayashi, and T. Ishikawa, "Detecting driver's visual attention area by using vehicle-mounted device," in *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, Jul. 2017, pp. 346–352, doi: 10.1109/ICCI-CC.2017.8109772.

[6] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4318–4327, Oct. 2020, doi: 10.1109/TITS.2019.2939676.

[7] Y. Wang, X. Ding, G. Yuan, and X. Fu, "Dual-Cameras-Based Driver's Eye Gaze Tracking System with Non-Linear Gaze Point Refinement," *Sensors*, vol. 22, no. 6, p. 2326, Mar. 2022, doi: 10.3390/s22062326.

[8] P. Smith, M. Shah, and N. da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 4, pp. 205–218, Dec. 2003, doi: 10.1109/TITS.2003.821342.

[9] J. Jo, "Vision-based method for detecting driver drowsiness and distraction in driver monitoring system," *Opt. Eng.*, vol. 50, no. 12, p. 127202, Dec. 2011, doi: 10.1117/1.3657506.

[10] S. Guasconi, M. Porta, C. Resta, and C. Rottenbacher, "A low-cost implementation of an eye tracking system for driver's gaze analysis," in *2017 10th International Conference on Human System Interactions (HSI)*, Jul. 2017, pp. 264–269, doi: 10.1109/HSI.2017.8005043.

[11] Z. Guo, Q. Zhou, and Z. Liu, "Appearance-based gaze estimation under slight head motion," *Multimed. Tools Appl.*, vol. 76, no. 2, pp. 2203–2222, Jan. 2017, doi: 10.1007/s11042-015-3182-4.

[12] X. Zhang, Y. Sugano, and A. Bulling, "Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, May 2019, pp. 1–13, doi: 10.1145/3290605.3300646.

[13] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically Unconstrained Gaze Estimation in the Wild," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, vol. 2019-Octob, pp. 6911–6920, doi: 10.1109/ICCV.2019.00701.

[14] J. Araluce *et al.*, "Gaze Focalization System for Driving Applications Using OpenFace 2.0 Toolkit with NARMAX Algorithm in Accidental Scenarios," *Sensors 2021, Vol. 21, Page 6262*, vol. 21, no. 18, p. 6262, Sep. 2021, doi: 10.3390/S21186262.

[15] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018, pp. 59–66, doi: 10.1109/FG.2018.00019.

[16] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha, "DGaze: CNN-Based Gaze Prediction in Dynamic Scenes," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 5, pp. 1902–1911, May 2020, doi: 10.1109/TVCG.2020.2973473.

[17] R. Naqvi, M. Arsalan, G. Batchuluun, H. Yoon, and K. Park, "Deep Learning-Based Gaze Detection System for Automobile Drivers Using a NIR Camera Sensor," *Sensors*, vol. 18, no. 2, p. 456, Feb. 2018, doi: 10.3390/s18020456.

[18] L. Fridman, J. Lee, B. Reimer, and T. Victor, "'Owl' and 'Lizard': patterns of head pose and eye pose in driver gaze classification," *IET Comput. Vis.*, vol. 10, no. 4, pp. 308–314, Jun. 2016, doi: 10.1049/iet-cvi.2015.0296.

[19] In-Ho Choi, Sung Kyung Hong, and Yong-Guk Kim, "Real-time categorization of driver's gaze zone using the deep learning techniques," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, Jan. 2016, pp. 143–148, doi: 10.1109/BIGCOMP.2016.7425813.

[20] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim, "Real-Time Gaze Estimator Based on Driver's Head Orientation for Forward Collision Warning System," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 254–267, Mar. 2011, doi: 10.1109/TITS.2010.2091503.

[21] U. Sonom-Ochir, S. Karungaru, K. Terada, and A. Ayush, "Detection of Driver'S Visual Distraction Using Dual Cameras," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 5, pp. 1445–1461, 2022. [Online]. Available at: https://repo.lib.tokushima-u.ac.jp/en/118053.

[22] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver Gaze Zone Estimation Using Convolutional Neural Networks: A General Framework and Ablative Analysis," *IEEE Trans. Intell. Veh.*, vol. 3, no. 3, pp. 254–265, Sep. 2018, doi: 10.1109/TIV.2018.2843120.

[23] U. Sonom-Ochir, S. Karungaru, K. Terada, and A. Ayush, "Appearance-based Driver's Gaze Mapping Using a Dash Camera," in *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and*

23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Nov. 2022, pp. 1–5, doi: 10.1109/SCISISIS55246.2022.10001875.

[24] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing Eye Tracking With Bayesian Adversarial Learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, vol. 2019-June, pp. 11899–11908, doi: 10.1109/CVPR.2019.01218.

[25] Y. Cheng, Y. Bao, and F. Lu, "PureGaze: Purifying Gaze Feature for Generalizable Gaze Estimation," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, pp. 436–443, Jun. 2022, doi: 10.1609/aaai.v36i1.19921.

[26] Y. Liu, R. Liu, H. Wang, and F. Lu, "Generalizing Gaze Estimation with Outlier-guided Collaborative Adaptation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 3815–3824, doi: 10.1109/ICCV48922.2021.00381.

[27] Y. Bao, Y. Liu, H. Wang, and F. Lu, "Generalizing Gaze Estimation with Rotation Consistency," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, vol. 2022-June, pp. 4197–4206, doi: 10.1109/CVPR52688.2022.00417.

[28] Z. Guo, Z. Yuan, C. Zhang, W. Chi, Y. Ling, and S. Zhang, "Domain Adaptation Gaze Estimation by Embedding with Prediction Consistency," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12626 LNCS, Springer Science and Business Media Deutschland GmbH, 2021, pp. 292–307, doi: 10.1007/978-3-030-69541-5_18.

[29] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, Oct. 2013, pp. 271–280, doi: 10.1145/2501988.2501994.

[30] J. Luo, J. Huang, and H. Li, "A case study of conditional deep convolutional generative adversarial networks in machine fault diagnosis," *J. Intell. Manuf.*, vol. 32, no. 2, pp. 407–425, Feb. 2021, doi: 10.1007/s10845-020-01579-w.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[32] S. M. Shah, Z. Sun, K. Zaman, A. Hussain, M. Shoaib, and L. Pei, "A Driver Gaze Estimation Method Based on Deep Learning," *Sensors*, vol. 22, no. 10, p. 3959, May 2022, doi: 10.3390/s22103959.