Facial Features Enhanced Multi-Branch Graph Network for Driver Drowsiness Detection

Songwen Pei[1(✉)], Ying Xie[1], Huichen Zhang[2], and Sen Zhang[2]

[1] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
swpei@usst.edu.cn

[2] Key Laboratory of Ministry of Public Security for Road Traffic Safety, Traffic Management Research Institute of the Ministry of Public Security, Wuxi, China

Abstract. Driver drowsiness detection from driver facial images presents a notable challenge. During driver drowsiness detection, drivers frequently wear masks or sunglasses, which obscure critical facial features like the eyes and mouth. This occlusion significantly reduces detection accuracy. To address this issue, we propose a multi-branch graph network (MBGN) approach designed to enhance facial feature representation. Firstly, the model locates the facial region in the image and extracts facial landmarks. These landmarks are subsequently transformed into graph node structures. Then, three branches of the graph network structure are used to extract the features of the nodes in three regions. Finally, an adaptive learning weight network (ALW-Net) is introduced to learn the weight parameters for each branch's features, reducing the impact of occluded features and improving detection accuracy. We conducted experiments on three public datasets and a synthetic dataset. The results indicate that the proposed method improves accuracy by at least 25.9%, 28.2%, 25%, and 22.1%, respectively, with an average detection time of only 0.05 seconds. This approach significantly enhances facial feature representation, enabling more accurate real-time drowsiness detection even under occlusion.

Keywords: Driver Drowsiness Detection · Multi-Branch · Graph Convolution Networks · Adaptive Learning Weight Network

1 Introduction

Driver drowsiness detection has been widely recognized as a significant cause of traffic accidents [27]. Before a driver enters a drowsiness state, it is difficult for him to detect it himself. Therefore, real-time detection and warning of the driver's condition are crucial. With the rapid development of machine vision technology, real-time drowsiness detection methods based on facial features have received widespread attention and are the most widely applied. These methods typically extract landmark information from the driver's eyes and mouth to infer metrics such as eye closure duration [24], blink frequency [15] or yawning [19, 25], in order to determine the driver drowsiness state [22, 30]. However, in

practical driving scenarios, many drivers tend to wear sunglasses and masks, leading to missing feature representations in the eye or mouth regions, which can significantly impact the accuracy of drowsiness detection. Hence, when facial feature information is partially occluded, it becomes essential to minimize the impact of these occlusions on drowsiness detection and improve the accuracy of real-time detection.

The existing methods of facial drowsiness detection primarily rely on the Percentage of Eye Closure (PERCLOS) [2] to evaluate driver drowsiness. Many researchers [6, 28, 37] have extended the PERCLOS method by incorporating the Eye Aspect Ratio (EAR) to detect eye closure and drooping, providing a basis for determining driver drowsiness. However,this approach faces challenges in maintaining detection accuracy under partial occlusion of facial regions. Deep Neural Network (DNN)-based face detection method [8, 31], which is effective in capturing facial landmark information even under significant occlusion, has demonstrated higher accuracy and stronger resistance to occlusion. This method is often combined with the PERCLOS index to assess driver drowsiness, thereby enhancing the extraction of drowsiness-related features and improving robustness. However, This method still relys on the computation of the PERCLOS value to determine drowsiness, which limits its generalizability.

Some scholars [5, 23] have proposed facial drowsiness detection algorithms based on Dempster-Shafer (DS) [34] Evidence Theory, utilizing an DS fusion decision-making process to determine fatigue levels. This approach effectively addresses the issue of partial facial occlusion in complex environments, it still struggles with low generalization across different facial features. Zhao et al. [35] introduced the MC-Facenet model, which employs a multi-task cascaded convolutional neural network to address partial occlusion detection. They incorporated MobileNetV1 and an attention mechanism to improve face detection accuracy. However, this method primarily enhances facial recognition accuracy and exhibits slower detection times. Xiang et al. [29] used the YOLOv3 model and Faster R-CNN model to study the problem of mouth features loss due to mask occlusion, but the model is only limited to the study of mask occlusion of the mouth, and its generalization is yet to be verified.

To address the issue of improving drowsiness detection accuracy under facial occlusions, this paper introduces a method that enhances facial landmark representation to mitigate the impact of occluded regions on overall detection performance. Our contribution can be summarised in the following three points:

1. We propose MBGN for driver drowsiness detection. This model involves transforming facial landmark information into graph nodes, which serve as the primary input data for the model.

2. We adopt three branches of the graph convolutional network to extract node features associated with the eyes and mouth. This not only reduces the graph size and model complexity but also mitigates potential interference among regional features.

3. We proposed an adaptive learning weight model to learn the feature weights of each branch. we reduce the unreliable feature to the total feature weight, and successfully reduce the influence of occluded regions on the final detection results.

2 Related Work

2.1 Dlib Face Detection Algorithm

Existing face detection algorithms have good performance in scenes without occlusion. However, when applied to face recognition in partially occluded natural settings, they encounter challenges leading to the loss of facial features and suboptimal recognition results [3, 13]. Addressing this issue is crucial for enhancing face recognition under partial occlusion. The Dlib algorithm [4, 20] is a pre-trained landmark detector which allows the detection of 68 landmarks. Specifically, it detects 6 landmarks for each eye and 20 landmarks for the mouth, divided between the upper and lower eyelids and lips. The focus is on the right eye, left eye, and mouth landmarks. Based on the distance between the upper lip landmark of the eye or mouth and the lower lip landmark of the eye or mouth, it is possible to determine whether the driver is closing his or her eyes or yawning. Zhang et al. [32] proposed a new method of Dlib face recognition based on Ensemble of Regression Trees (ERT) algorithm [9]. This method implements face recognition and feature calibration through Python, utilizes numerous pretrained face model interfaces and demonstrates strong robustness to occlusion. Experimental results show that this method is better than OpenCV method, and can effectively improve the detection sensitivity, recognition accuracy and recognition performance. The Dlib algorithm accurately detects landmarks in unobstructed regions and infers rough positions in occluded regions using structural information. However, these inferred positions lack interpretability and do not closely match actual positions under real occlusion.

2.2 Graph Convolutional Network

Graph neural network (GNN) [10,14] is a deep learning method applied to graph domains, widely recognized for its powerful representation capabilities in graph analysis. Graph convolutional network (GCN) [16] enable effective feature learning in irregular graph structures by extending convolution operations to graph data. The GCN typically includes a graph convolution layer [17], adjacency matrix normalization, a multilayer network, and an output layer. The graph convolution layer is the core of GCN, which aggregates features from neighboring nodes with linear transformations and activations. The introduction of GCN has significantly advanced GNN and is widely applied in graph data processing. Zhang et al. [33] proposed a two-stage regression network by integrating a structural hourglass network (SHN) with the GCN for unconstrained facial landmark detection. Ling Lo et al. [18] who were inspired by the node relationships for constructing GCN proposed an end-to-end AU oriented graph classification

network for classifying facial expressions. The GCN is particularly adept at capturing inter-node relationships in local graph structures. In driver drowsiness detection, the local interactions between facial landmarks, such as around the eyes and mouth, are critical for detecting drowsiness. The GCN effectively extracts these features through graph convolutional layers, improving classification accuracy.

3 Methodology

3.1 Overview of MBGN Model

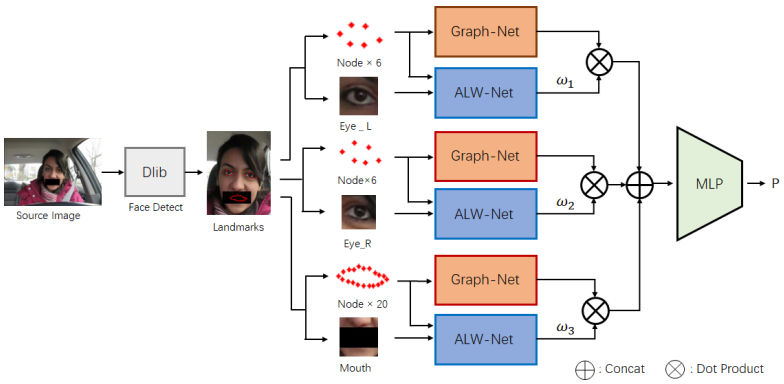

Fig. 1. Framework of multi-branch graph network.

The overall method of this paper is shown in Fig. 1. Firstly, Dlib is used to extract different facial landmarks. These landmarks, specifically from the left eye, right eye, and mouth, are then transformed into graph node structures and input into their respective node feature extraction networks. Next, the GCN extracts node features by considering the node positions and the distance relationships between them, and learns the mapping relationship between the node features and facial drowsiness. Then, the ALW-Net adaptively learns weights for each region and multiplies these with regional features. Finally, the three weighted features are concatenated. The probabilistic detection result P is regressed by a multilayer perceptron, where P ranges from [0,1]. The F1score and other evaluation metrics indicate demonstrate that as the P value approaches 1, the accuracy of drowsiness detection increases. Therefore, the P value greater than 0.5 indicates drowsiness state, while a value less than or equal to 0.5 indicates a normal state.

3.2 Branch Graph Network

In current drowsiness detection methods based on facial features, the eye and mouth landmarks primarily serve as the basis for drowsiness classification. Therefore, we transform these points into graph nodes and rank them based on their horizontal and vertical coordinates for analysis. The inputs of GCN are mainly node feature matrix X and node adjacency matrix A. As shown in Eq. (1) and (2), we use the position coordinates of regional landmarks $N_i = (x_i, y_i)$ as input information for each node to construct the node feature matrix X. Based on the Dlib algorithm results, the eye morphology is labeled with 6 landmarks and the mouth morphology with 20 landmarks. Consequently, the model input for the eye region is represented by a 2×6 coordinate matrix constructed from the node positions, while the mouth region is represented by a 2×20 coordinate matrix constructed similarly.

$$X_{eye} = (N_1, N_2, \cdots, N_6) \tag{1}$$

$$X_{mouth} = (N_1, N_2, \cdots, N_{20}) \tag{2}$$

To capture the global structure and characteristics of the graph nodes in each region, as shown in Eq. (3) and (4). The Euclidean distance $D_{i,j}$ between each landmark is computed from the landmark coordinates, thus constructing the corresponding adjacency matrix A. In the extracted facial landmark images, the Euclidean distance reflects the distance between two landmarks in the feature space. A smaller distance indicates greater similarity between the landmarks, while a larger distance indicates a greater difference. Since the distance between its own nodes is zero and the nodes in the graph are undirected, A is a symmetric matrix.

$$D_{i,j} = \sqrt{(x_i - y_i)^2 + (x_j - y_j)^2} \tag{3}$$

$$A = \begin{bmatrix} D_{11} & \cdots & D_{1n} \\ \vdots & \ddots & \vdots \\ D_{n1} & \cdots & D_{nn} \end{bmatrix} \tag{4}$$

In Eq. (3), x_i and y_i represent the horizontal and vertical coordinates of node i, x_j and y_j represent the horizontal and vertical coordinates of node j.

Considering the inconsistency in the numerical scales of coordinate and distance information, and to ensure relatively stable input weights for each node during the graph convolution operation, normalization is applied to the node feature matrix and adjacency matrix before their input. By constructing the graph nodes, the main inputs to the model are three graphs: $G_{eye1} = (X_{eye1}, A_{eye1})$, $G_{eye2} = (X_{eye2}, A_{eye2})$, and $G_{mouth} = (X_{mouth}, A_{mouth})$. As shown in Fig. 2, each branch of the GCN is used for the feature extraction. The specific computation is shown in Eq. (5).

$$H^{l+1} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{5}$$

In Eq. (5), $H^{(l)}$ is the node feature matrix of layer l, $H^{(l+1)}$ is the node feature of layer $l+1$, $W^{(l)}$ is the weight matrix of layer l, $\widetilde{A}$ is the adjacency

matrix with added self-loops, $\widetilde{D}$ is the degree matrix, σ is the activation function. As shown in Fig. 2, the graph network model comprises two GCN blocks. Each

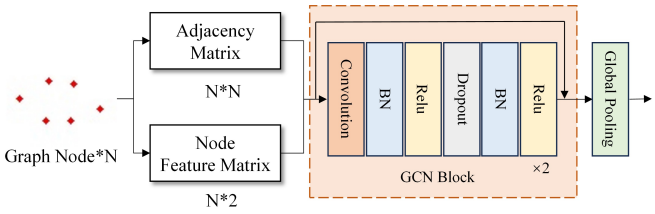

Fig. 2. Structure of branch graph network.

block contains a convolutional layer with a kernel size of 1×1, followed by a Batch Normalization (BN) layer and the ReLU activation function. To improve the model's generalization ability, a DropOut layer with a probability of 0.5 is incorporated. Finally, the input and output of each block are connected to form the final output. After feature extraction, a global pooling layer aggregates node information into global representations, ensuring that the output of the three-branch GCN produces a fixed-size global feature vector.

3.3 Adaptive Learning Weight Network (ALW-Net)

As can be seen from Sect. 2.1, the accuracy of landmarks detected by the Dlib algorithm is significantly reduced in occluded regions. The integration of these less accurate landmarks with other precise features can reduce the reliability of the final detection result and cause contradictions with other features, negatively impacting the model learning process. Therefore, based on the branch graph network from Sect. 3.2, we introduce an ALW-Net to dynamically learn the optimal weights for the features from each branch. This enhances unobstructed facial features and reduces the impact of occluded regions on overall features, improving detection accuracy.

The proposed ALW-Net adaptively detects the degree of alignment between the graph nodes and their corresponding regions in the image, assigning proportional weights ω to branch features within the overall feature representation. As shown in Fig. 3 (a), the outermost coordinates of landmarks in the image are used to extract eye and mouth region images, which serve as inputs to ALW-Net. Eye contour features are initially extracted using a 3×3 convolution with the number of convolution channels C of 32, a stride of 1 and a ReLU activation function. To further enhance feature learning, a Spatial Attention Module (SAM) is incorporated [36], as shown in Fig. 3 (b). Frist, the results of maximum pooling and average pooling of image features are spliced and convolved, Then, a Sigmoid function is used to calculate attention values for target regions in the image. Finally, these attention values are multiplied with the original features to highlight significant areas.

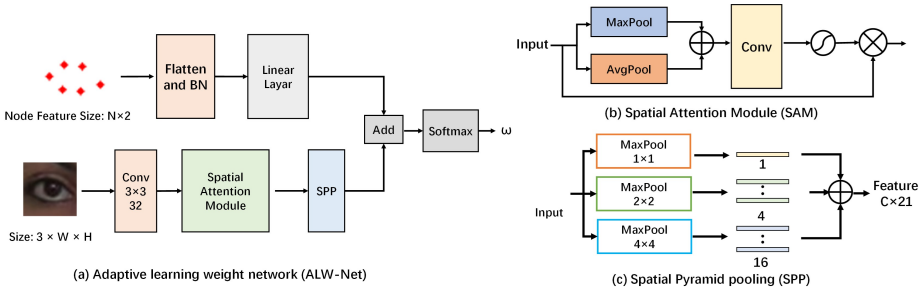

Fig. 3. Structure of adaptive learning weight network.

Considering variations in the sizes of cropped region images, Spatial Pyramid Pooling (SPP) is employed to unify feature map dimensions, as depicted in Fig. 3 (c) [11]. In SPP, three pooling layers with sizes of 1×1, 2×2 and 4×4 are are employed to compute the features, and a fixed-size feature vector (C×21) is outputted after the concat. The node feature matrices constructed for each branch in Sect. 3.2 are also expanded by spreading and normalizing the node feature matrices for linear layer expansion with dimensions consistent with with (C×21). Finally, feature fusion is performed by element-wise summation of the outputs from the two branches, followed by a softmax operation to derive a weight value within (0,1) for each branch. ALW-Net evaluates the degree of alignment between graph node shapes and their corresponding regions in the image to determine occlusion levels, reducing output weights for regions with low alignment.

The proposed model is trained using a cross-entropy loss function, as shown in Eq. (6). The ALW-Net also adapts the learning of the matching degree between the graph nodes and their corresponding regions through backpropagation of the gradients from this loss.

$$L = -\sum_{i} p_i \log(\widehat{p}_i) \tag{6}$$

In Eq. (6), p_i represents the true label, and $\widehat{p}_i$ represents the model prediction result.

4 Experiments

4.1 Experimental Settings

Implementation Details. The method in this paper is implemented on a Linux operating system using an NVIDIA GeForce RTX 3060 GPU, Python 3.9, CUDA 11.7, and the PyTorch deep learning framework. The input image size is 256×256×3, the batchsize is 16, and the model is trained for 100 epochs. The Adam optimizer is used with an initial learning rate of 0.01, a momentum of 0.937. Maintain a constant learning rate for the initial 50 epochs, then linearly reduce it to zero over the following 50 epochs.

Datasets. In this paper, three publicly available datasets are collected: YAWDD [1], NTHU-DDD [12], and Drozy [26]. The YAWDD dataset includes drivers from various racial backgrounds, as well as drivers wearing glasses, sunglasses, and no glasses. The data record driving scenarios such as drivers driving normally, yawning, and slow blinking to sleep under multiple lighting conditions. The NTHU-DDD and Drozy datasets include about 460 videos of individuals driving in various scenarios, from different viewpoints, at different times of the day, and in different countries. These videos contain numerous instances of face-obscuring situations, such as tinted glasses and masks. We randomly extracted discontinuous 100-frame images from each video and recorded the corresponding drowsiness state, totaling about 46,000 pieces of data. To further enhance the model's occlusion resistance, we manually added facial occlusions such as sunglasses and masks by synthesizing 5,000 images selected randomly from the unoccluded images available in the dataset mentioned above, as shown in Fig. 4. In summary the datasets are all divided into training set, validation set, and test set according to 8:1:1 and are uniformly resized to 256×256×3 and normalized before inputting into the model.

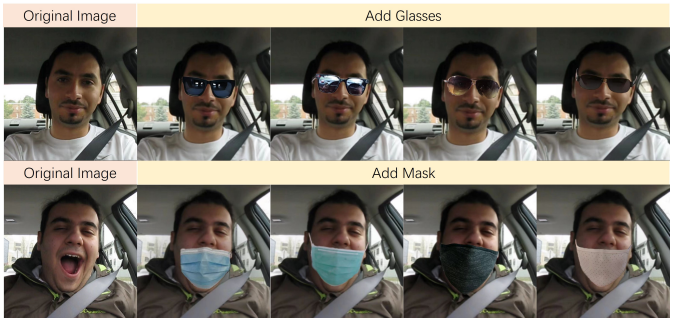

Fig. 4. A synthetic dataset.

4.2 Result Analysis.

Model Comparison. In order to further validate the performance and advantages of the proposed method, it is compared with the traditional baseline method based on the PERCLOS [2] metric and mainstream facial drowsiness detection methods such as DNN [31], DS [5], MC-Facenet [35], and Yolo-F_RCNN [29]. These comparisons are conducted on three public datasets and a synthetic dataset. This paper uses Accuracy, Detection time, Precision, Recall, and F1score as evaluation metrics. Precision represents the proportion of correctly detected drowsiness states among the predicted positive samples. Recall indicates the proportion of correctly detected drowsiness states among all test samples. F1score is the harmonic mean of precision and recall, used to simulta-

neously consider the number of predicted positive samples and actual positive samples.The formulas for precision and recall are shown in Eqs. (7)-(8):

$$Precision = \frac{T_p}{F_p + T_p} \tag{7}$$

$$Recall = \frac{T_p}{F_n + T_p} \tag{8}$$

In Eq. (7) and Eq. (8), T_p (True positives) represents the number of drowsiness samples correctly predicted as drowsiness. F_p (False positives) represents the number of normal samples incorrectly predicted as drowsiness. F_n (False negatives) represents the number of drowsiness samples incorrectly predicted as normal samples.

Table 1. Comparison of the proposed method with the baseline method and mainstream facial drowsiness detection methods on three public datasets and a synthetic dataset."Improv." denotes the relative improvement in accuracy of each method compared to the baseline.

Dataset	Method	Accuracy (%)	Time (s)	Precision(%)	Recall(%)	F1score	Improv.
YAWDD	PERCLOS [2]	71.2	0.03	72.8	73.6	0.734	-
	DNN [31]	84.2	0.05	84.5	84.8	0.848	13%
	MC-FaceNet [35]	85.6	0.08	84.5	85.2	0.845	14.4%
	DS [5]	93.8	0.09	93.6	94.4	0.937	22.6%
	Yolo-F_RCNN [29]	85.3	0.08	87.2	88.5	0.878	24.1%
	MBGN(Ours)	**97.1**	**0.05**	**97.5**	**97.9**	**0.976**	**25.9%**
NTHU-DDD	PERCLOS [2]	69.4	0.02	72.9	74.3	0.733	-
	DNN [31]	81.3	0.06	84.3	85.4	0.843	11.9%
	MC-FaceNet [35]	85.1	0.07	84.8	85.6	0.851	15.7%
	DS [5]	92.2	0.09	93.6	94.1	0.939	22.8%
	Yolo-F_RCNN [29]	85.8	0.06	87.3	88.8	0.875	16.4%
	MBGN(Ours)	**97.6**	**0.05**	**97.5**	**97.6**	**0.975**	**28.2%**
Drozy	PERCLOS [2]	72.5	0.02	73.0	74.6	0.731	-
	DNN [31]	82.3	0.07	84.2	85.2	0.843	9.8%
	MC-FaceNet [35]	84.7	0.09	84.9	85.5	0.848	12.2%
	DS [5]	93.7	0.10	93.8	94.1	0.939	21.2%
	Yolo-F_RCNN [29]	86.5	0.06	87.1	88.9	0.875	14%
	MBGN(Ours)	**97.5**	**0.06**	**97.3**	**97.8**	**0.973**	**25%**
Synthetic dataset	PERCLOS [2]	75.4	0.02	76.4	77.6	0.761	-
	DNN [31]	83.1	0.05	82.2	84.9	0.833	7.7%
	MC-FaceNet [35]	84.1	0.08	85.9	86.5	0.855	8.7%
	DS [5]	92.5	0.09	93.8	94.3	0.938	17.1%
	Yolo-F_RCNN [29]	87.1	0.08	87.5	88.9	0.877	11.7%
	MBGN(Ours)	**97.5**	**0.05**	**97.4**	**97.6**	**0.974**	**22.1%**

As shown in Table. 1, the baseline model based on PERCLOS relies on the precise landmark positions of the eyes and mouth to determine drowsiness states, which results in poor robustness and generalization. This method is prone to errors due to occlusion. Although it has a high detection efficiency, the accuracy is only around 72%. DNN, and MC-Facenet methods improve detection accuracy

under face occlusion by optimizing the model, so as to indirectly improve the accuracy of the drowsiness classification. In comparison to the baseline method, the average accuracy across the four datasets is improved by approximately 10% and 12%, respectively. However, the two methods still take the PERCLOS index as the final determination criterion, so the accuracy rate is still at a value of less than 86%. The process of DS drowsiness determination avoids the use of the classification method of PERCLOS, and optimizes the influence of occlusion area through the theory of DS evidence. This approach achieving an accuracy rate of over 90%. The Yolo-F_RCNN method excels in accurate eye detection and performs well even when the mouth is occluded. However, its accuracy drops below 90% when the eyes are occluded. The method in this paper still utilizes landmarks as the main data for detection to avoid the influence of other features in the image. At the same time, the drowsiness features corresponding in the graph nodes are learned through the graph network, which greatly improves the generalization. The model adjusts the proportion of the occluded region in the features through adaptive weighting, reducing its impact on classification accuracy. This ensures detection accuracy remains above 97%. A comparison was also made with the aforementioned methods in terms of Precision, Recall, and F1score, where the proposed method still outperforms the comparison methods.

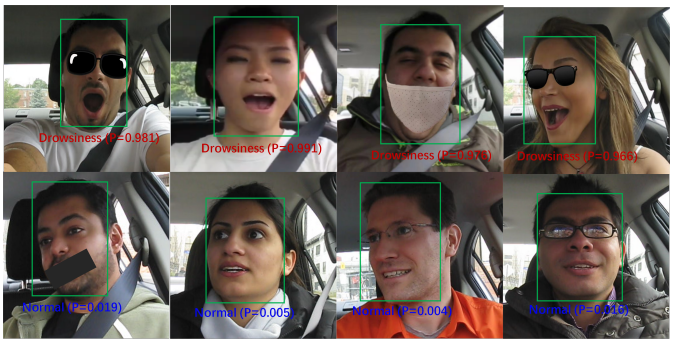

Fig. 5. Recognition results of normal and drowsiness states.

We visualized the drowsiness classification performance of the proposed method, as shown in Fig. 5. It displays the classification results of drowsiness and normal states, where the green checkboxes represent Dlib tracking boxes and indicate the probability values of normal and drowsy states. It can be observed that the model accurately detects the driver's drowsiness state even when facial features are partially occluded.

Ablation Study. In order to demonstrate the effectiveness of GCN introduced in this paper, we performed ablation experiments on popular network models

Table 2. Accuracy of different model combinations.

Model	YAWDD	NTHU-DDD	Drozy	Synthesis
CNN [26]	73.6	71.6	72.5	76.2
ResNet [7]	81.5	79.2	80.4	82.5
LSTM [21]	89.1	88.9	89.3	89.4
MBGN(Ours)	**96.5**	**97.6**	**97.7**	**97.5**

(CNN, ResNet, LSTM) and GCN for image detection and classification, comparing their accuracies on a common dataset.

As shown in Table. 2, models like CNN [26], ResNet [7], and LSTM [21] use image pixels to extract features. This process is affected by background and other organ features in facial images, resulting in an average accuracy of less than 90%. We used GCN for feature extraction, directly extracting landmark features to avoid interference from other image information, achieving an average detection accuracy of over 97%.

Occluded Area Comparison. To verify the effectiveness of the proposed model in detecting driver fatigue under different occlusion scenarios, we synthesized four types of occluded data images (left eye, right eye, both eyes and mouth) for experimental evaluation. To assess the overall performance of the model on the occlusion datasets, we conducted tests on the synthetic dataset. The classification results of the model are evaluated using the most common metrics in binary classification tasks, namely the receiver operating characteristic (ROC) curve and the area under the curve (AUC).

Table 3. Accuracy under occlusion in different facial regions.

Occluded areas	Accuracy (%)	F1score (%)	AUC (%)
Left eye	97.5	97.2	97.6
Right eye	97.1	96.8	96.7
Both eyes	96.5	96.1	95.8
Mouth	96.5	96.1	95.8

Table. 3 shows that occluding the left or right eye separately has minimal impact on detection accuracy, indicating good model stability. However, when both eyes are occluded and the landmarks of both eyes are missing, the detection accuracy of the model decreases. Since the mouth region contains more landmarks and drowsiness features are more prominent, occlusion of the mouth area leads to a relatively lower accuracy when processing the occlusion of the

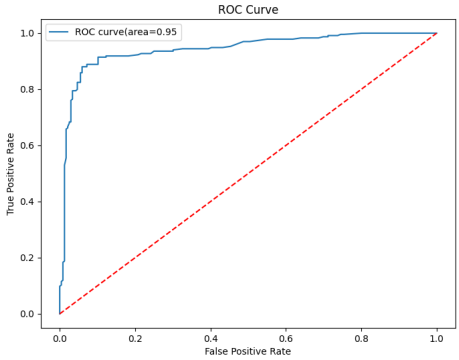

Fig. 6. ROC curve for MBGN on the synthetic dataset.

mouth. Nevertheless, whether both eyes or the mouth are occluded, the model's detection accuracy remains above 96%, which is within an acceptable range. Moreover, the experiments show that different types of occlusion do not have a significant effect on the detection performance of the proposed model.

The model's performance is evaluated using the ROC curve, as shown in Fig. 6. It can be observed that the AUC value of MBGN on the synthetic dataset is higher than 0.9, approaching 1. The overall performance of MBGN demonstrates a good classification effect, effectively completing the driver state classification task under various occlusion conditions. The optimal value of the ROC curve corresponds to a high level of sensitivity, indicating stronger generalization ability of the model.

5 Conclusion

To address the feature loss caused by facial occlusion, which impacts drowsiness detection accuracy, we propose a novel multi-branch network structure. This model enhances facial features to improve detection accuracy and utilizes a multi-branch graph network to handle partially occluded faces. The model includes a graph node module for transforming feature points, with eye and mouth landmarks converted into a graph node feature matrix. These transformed matrices are then input into the corresponding branches of the graph network to learn the mapping between node features and facial drowsiness. The adaptive learning weight network accurately learns the weight information of each branch feature, reducing the influence of occluded regions on overall feature representations. Experimental results demonstrate the effectiveness of this model in drowsiness detection under partial facial occlusion, achieving an average classification accuracy of over 97% and an average single-frame detection time of 0.05 seconds, thus meeting the real-time requirements for driving detection. Future work will focus

on detecting multi-frame video information and improving the model's ability to capture changes in drowsiness through graph networks.

Acknowledgements. The authors would like to thank the anonymous reviewers for their invaluable comments.This work was partially funded by the National Natural Science Foundation of China under Grant No.61975124, State Key Laboratory of Computer Architecture (ICT, CAS) under Grant No.CARCHA202111, the Key R&D Program Projects in Sichuan Province under Grant No.24ZDYF1640, Engineering Research Center of Software/Hardware Co-design Technology and Application, Ministry of Education, East China Normal University under Grant No.OP202202, and Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety under Grant No.2023ZDSYSKFKT04. Any opinions, findings and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

References

1. Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., Hariri, B.: Yawdd: A yawning detection dataset. In: Proceedings of the 5th ACM multimedia systems conference. pp. 24–28 (2014)
2. Acıoğlu, A., Ercelebi, E.: Real time eye detection algorithm for perclos calculation. In: 2016 24th Signal Processing and Communication Application Conference (SIU). pp. 1641–1644. IEEE (2016)
3. Aydın, M.T., Menemencioğlu, O., Orak, İ.M.: Face recognition approach by using dlib and k-nn. Current Trends in Computing **1**(2), 93–103 (2024)
4. Boussaki, H.E., Latif, R., Saddik, A.: Drowsiness detection using dlib: an overview. In: 7th IEEE Congress on Information Science and Technology, CiSt 2023, Agadir - Essaouira, Morocco, December 16-22, 2023. pp. 150–154. IEEE (2023)
5. Cai, J., Liao, X., Bai, J., Luo, Z., Liu, L., Bai, J.: Face fatigue feature detection based on improved D-S model in complex scenes. IEEE Access **11**, 101790–101798 (2023)
6. Chen, W., Zhang, X., Chen, S.: Fatigue detection system for extracting driver's eye features. In: 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE). pp. 891–894. IEEE (2024)
7. Choi, S., Chung, S., Lee, S., Han, S., Kang, T., Seo, J., Kwak, I.Y., Oh, S.: Tb-resnet: Bridging the gap from tdnn to resnet in automatic speaker verification with temporal-bottleneck enhancement. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10291–10295. IEEE (2024)
8. Dharmadhikari, S., Raut, R., Ray, A., Basak, A.: A unified mixed deep neural network for fatigue damage detection in components with different stress concentrations. Applied Sciences **13**(3), 1542 (2023)
9. Enriquez, M.L., Ducut, J.D., Baun, J.J., Concepcion, R., Relano, R.J., Francisco, K., Vicerra, R.R., Bandala, A., Dungca, J., Co, H., et al.: Spartacus: Sampling precision and rate transformation algorithm for continuous uniform sampling of an underground imaging antenna. In: 2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). pp. 1–6. IEEE (2023)

10. Fan, J., Liang, J., Liu, H., Huan, Z., Hou, Z.: Robust face alignment via adaptive attention-based graph convolutional network. Neural Computing and Applications **35**(20), 15129–15142 (2023)
11. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III. Lecture Notes in Computer Science, vol. 8691, pp. 346–361. Springer (2014)
12. Huang, Y., Liu, C., Chang, F., Lu, Y.: Self-supervised multi-granularity graph attention network for vision-based driver fatigue detection. IEEE Transactions on Emerging Topics in Computational Intelligence pp. 1–14 (2024)
13. Khan, S.S., Sengupta, D., Ghosh, A., Chaudhuri, A.: Mtcnn++: A cnn-based face detection algorithm inspired by mtcnn. The Visual Computer **40**(2), 899–917 (2024)
14. Kim, H., Lee, J.H., Ko, B.C.: Facial expression recognition in the wild using face graph and attention. IEEE Access (2023)
15. Li, X., Lin, H., Du, J., Yang, Y.: Computer vision-based driver fatigue detection framework with personalization threshold and multi-feature fusion. Signal Image Video Process. **18**(1), 505–514 (2024)
16. Liu, N., Zhang, F., Chang, L., Duan, F.: Facial attribute classification by deep mining inter-attribute correlations. IET Computer Vision **17**(3), 352–365 (2023)
17. Liu, Z., Zhang, C., Wu, Y., Zhang, C.: Joint face completion and super-resolution using multi-scale feature relation learning. Journal of Visual Communication and Image Representation **93**, 103806 (2023)
18. Lo, L., Xie, H.X., Shuai, H.H., Cheng, W.H.: Mer-gcn: Micro-expression recognition based on relation modeling with graph convolutional networks. In: 2020 IEEE conference on multimedia information processing and retrieval (MIPR). pp. 79–84. IEEE (2020)
19. Massoz, Q., Langohr, T., François, C., Verly, J.G.: The ulg multimodality drowsiness database (called drozy) and examples of use. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–7. IEEE (2016)
20. Poeloemgam, A.A., Mulyana, E., Hermawan, W., et al.: Web-based face detection and recognition using yolo and dlib. In: 2023 17th International Conference on Telecommunication Systems, Services, and Applications (TSSA). pp. 1–6. IEEE (2023)
21. Prakash, S., Jalal, A.S., Pathak, P.: Forecasting covid-19 pandemic using prophet, lstm, hybrid gru-lstm, cnn-lstm, bi-lstm and stacked-lstm for india. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON). pp. 1–6. IEEE (2023)
22. Qian, Y., Pei, S., Qin, W., Tan, J.: Video arbitrary style transfer via style attention and contrastive learning. In: 2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS) (2023)
23. Sekihara, M., Sakurai, S.: Creep-fatigue behavior in a ni-based ds alloy during high temperature oxidation. In: Creep and Fracture of Engineering Materials and Structures: Proceedings of the 9th International Conference: Proceedings of the 9th International Conference. pp. 479–486. CRC Press (2024)
24. Shvets, O., Smakanov, B., Györök, G., Kovács, L.: A driver fatigue recognition system, based on an artificial neural network. Acta Polytechnica Hungarica **21**(8), 211–226 (2024)
25. Tang, X., Guo, P.: Fatigue driving detection methods based on drivers wearing sunglasses. IEEE Access **12**, 70946–70962 (2024)

26. Wang, J., Zhang, X., Gao, G., Lv, Y.: Op mask r-cnn: An advanced mask r-cnn network for cattle individual recognition on large farms. In: 2023 International Conference on Networking and Network Applications (NaNA). pp. 601–606. IEEE (2023)
27. Wang, L., Wang, H., Liu, J.: Discrimination of driver fatigue based on distortion energy density theory and multiple physiological signals. IEEE Access **9**, 151824–151833 (2021). https://doi.org/10.1109/ACCESS.2021.3125052
28. Wu, D.: Improving automatic detection of driver fatigue and distraction using machine learning. arXiv preprint arXiv:2401.10213 (2024)
29. Xiang, Y., Yang, H., Hu, R., Hsu, C.Y.: Comparison of the deep learning methods applied on human eye detection. In: 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA). pp. 314–318 (2021)
30. Xu, K., Li, F., Chen, D., Zhu, L., Wang, Q.: Fusion of lightweight networks and deepsort for fatigue driving detection tracking algorithm. IEEE Access **12**, 56991–57003 (2024). https://doi.org/10.1109/ACCESS.2024.3386858
31. Yu, Z., Li, L., Xu, L., Chen, K.: Fatigue detection for public transport drivers under the normalization of epidemic prevention. In: 2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES). pp. 225–228 (2022)
32. Zhang, D., Li, J., Shan, Z.: Implementation of dlib deep learning face recognition technology. In: 2020 International Conference on Robots & Intelligent System (ICRIS). pp. 88–91 (2020)
33. Zhang, J., Hu, H., Feng, S.: Robust facial landmark detection via heatmap-offset regression. IEEE Transactions on Image Processing **29**, 5050–5064 (2020)
34. Zhao, K., Li, L., Chen, Z., Sun, R., Yuan, G., Li, J.: A survey: Optimization and applications of evidence fusion algorithm based on dempster–shafer theory. Applied Soft Computing **124**, 109075 (2022)
35. Zhao, Y., Wang, L., Tan, M., Yan, X., Zhang, X., Feng, H.: Face recognition with partial occlusion based on attention mechanism. In: 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS). pp. 562–566 (2021)
36. Zhu, X., Cheng, D., Zhang, Z., Lin, S., Dai, J.: An empirical study of spatial attention mechanisms in deep networks. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 6687–6696. IEEE (2019)
37. Zulkarnanie, M.A., Shanmugam, K.S., Badruddin, N., Saad, M.N.M.: Enhancements to perclos algorithm for determining eye closures. In: 2022 International Conference on Future Trends in Smart Communities (ICFTSC). pp. 76–81. IEEE (2022)