Driver Drowsiness Detection Based on Joint Human Face and Facial Landmark Localization With Cheap Operations

Qingtian Wu, *Member, IEEE*, Nannan Li, *Member, IEEE*, Liming Zhang, *Senior Member, IEEE*, and Fei Richard Yu, *Fellow, IEEE*

***Abstract*— Real-time detection of driver drowsiness is critical to reduce the risk of road accidents and fatalities. Current facial landmark-based methods usually use a two-stage paradigm, where faces and facial landmarks are localized separately. Additionally, most methods can be hindered by challenging conditions, such as night driving or eyes closed. To address these challenges, we present a refined YOLO network named YOLOFaceMark that can simultaneously detect faces and their facial landmarks. Furthermore, we introduce a drowsiness detection model based on facial landmarks. This model utilizes extracted eye and mouth information to identify drowsy states. We optimize the original YOLO components through structural re-parameterization, channel shuffling, and the design of a dual-branch detection head with an implicit module. These enhancements are designed to improve the accuracy while maintaining computational efficiency. We validate the real-time performance and accuracy of YOLOFaceMark on public datasets, including 300W and COFW. Additionally, we conduct further validation to demonstrate our ability to achieve effective and robust drowsiness detection solely based on the facial landmarks detected by YOLOFaceMark.**

***Index Terms*— Driver drowsiness detection, face detection, facial landmark detection, end-to-end network, real-time system.**

Manuscript received 20 May 2023; revised 28 August 2023, 5 February 2024, and 6 June 2024; accepted 11 August 2024. Date of publication 8 October 2024; date of current version 27 November 2024. This work was supported in part by the Science and Technology Development Fund of Macau SAR under Grant 0041/2023/RIA1, in part by the Multi-Year Research Grant under Grant MYRG2022-00193-FST and Grant MYRGGRG2023-00056-FST-UMDF, in part by the Hetao Shenzhen-HongKong Science and Technology Innovation Cooperation Zone under Grant HZQB-KCZYZ-2021055, and in part by Shenzhen Deeproute.ai Co., Ltd. The Associate Editor for this article was S. A. Birrell. *(Corresponding author: Liming Zhang.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Shenzhen Polytechnic University Medical Ethics Committee, and performed in line with the principles of the Declaration of Helsinki.

Qingtian Wu is with the Institute of Applied Artificial Intelligence, Guangdong-Hong Kong-Macao Greater Bay Area and the School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen 518055, China (e-mail: wuqingtian@szpu.edu.cn).

Nannan Li is with the International Institute of Next Generation Internet, Macau University of Science and Technology, Macau, China (e-mail: nnli@must.edu.mo).

Liming Zhang is with the Department of Computer and Information Science, Faculty of Sciences and Technology, University of Macau, Macau, China (e-mail: lmzhang@um.edu.mo).

Fei Richard Yu is with the School of Information Technology, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: richard.yu@carleton.ca).

Digital Object Identifier 10.1109/TITS.2024.3443832

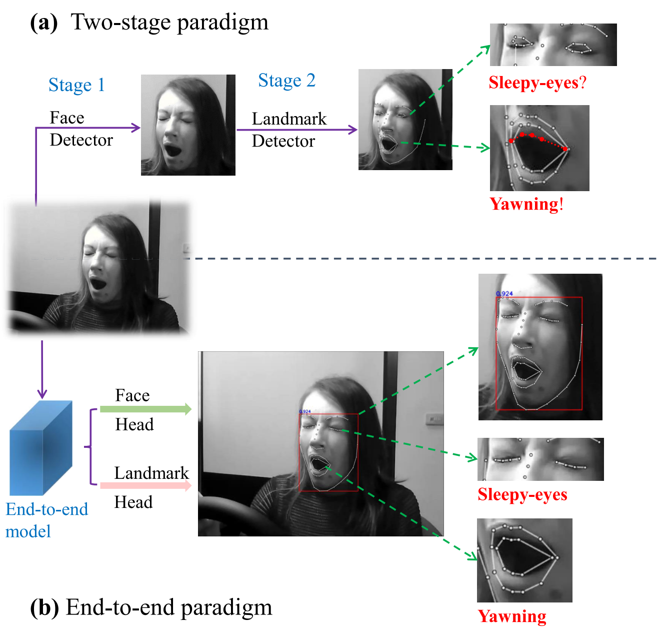
I. Introduction

DRIVER drowsiness detection (DDD) is a critical research area with significant implications for public safety, especially in the fields of transportation [1] and healthcare. Real-time DDD can provide timely warnings to drivers, thereby reducing the risk of accidents.

Various approaches have been developed for DDD, including physiological measures, behavioral measures, and visual solutions. Physiological measures involve monitoring changes in heart rate (Electrocardiogram, ECG) [2], brain activity (Electroencephalogram, EEG) [3], and skin conductance (Electromyography, EMG) [4]. But these methods require the use of some physical contact equipment. Recent research has increasingly focused on developing non-invasive and reliable methods for DDD. Behavioral measures primarily focus on analyzing driving behavior, such as lane deviation [5] and steering wheel movements. Visual approaches that analyze facial landmarks have emerged as promising solutions due to their effectiveness and convenience. These methods leverage the analysis of facial landmarks and eye movements to detect signs of drowsiness in drivers.

Existing facial landmark-based approaches [6], [7], [8] usually follow a two-stage paradigm, as depicted in Fig. 1(a). In this paradigm, they initially employ a sophisticated face detector to locate human faces, and subsequently utilize a refined landmark detector to achieve precise facial landmark detection (FLD). These landmarks, including the eyes and mouth, enable the analysis of momentary changes in facial features and the identification of drowsiness-related patterns such as eye closures or yawning. However, most of the methods do not work well in some complex scenarios, such as occlusions, varying lighting conditions, large head postures. In general, there are two fundamental problems remain challenging:

- **The balance between detection accuracy and inference speed**. Almost the state-of-the-art (SOTA) methods in the literature use the two-stage paradigm as shown in Fig. 1(a). Some methods prioritize accuracy [8] while ignoring real-time constraints. On the other hand, some methods [7] prioritize speed but suffer from accuracy issue. Computing deep features for precise face and landmark detection is a time-consuming process. The features of face and landmark are not shared in the two



(a) Two-stage paradigm

(b) End-to-end paradigm

Fig. 1. Two different paradigms for facial landmark-based DDD. (a) Two-stage based paradigm. (b) Our proposed end-to-end paradigm.

separate stages. Therefore, two-stage methods are not friendly to real-time and accurate industrial applications.

- **Generalization and robustness in the wild**. robustness is crucial for enhancing the safety of assisted driving systems. While certain SOTA methods [6], [9] demonstrate high accuracy in FLD when tested on manually annotated faces, they face challenges when applied in real-world scenarios, e.g., driving in dimmer light, occlusions, and eye closures. In such situations, these methods exhibit missed face detection and false FLD, as depicted in Fig. 2(a). Moreover, it is difficult to detect eye closure accurately, as shown in Fig. 2(c).

To address the above issues, we propose an end-to-end convolutional neural network (CNN) for real-time DDD. Our approach aims to analyze some facial landmarks over time, which can provide insights into the driver's eye closure and yawning patterns, enabling effective drowsiness analysis.

Our proposed method consists of two steps. First, inspired by YOLOPose [10], an end-to-end framework for human pose estimation (HPE), we design our baseline model called YOLOFaceMark for end-to-end FLD. We make modifications to YOLOPose by transforming the task from sparse HPE (17 keypoints) to dense FLD (68 keypoints) and adjusting the label from human body to face. Second, we propose a DDD model based on FLD results. The model can monitor the changes of the corresponding face landmarks when the driver closes his eyes and yawns in real time, and issue DDD warnings in time.

The contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to achieve end-to-end joint detection of eye-closure and yawning based on dense facial landmarks in the literature. Our approach focuses on detecting dense facial landmarks for robust and effective detection of eye-closing and yawning movements.
- To improve the balance between the inference speed and detection accuracy, we adopt some cheap operations such as the structural re-parameterization, channel shuffling, and implicit knowledge modules for accurate FLD.
- Due to the limited number of closed-eye face samples in the public dataset, we create a high-quality training set by employing a "student learns from teacher" strategy. In addition, a fine-tuning model based on the strategy is proposed for the case of closing eyes and yawning under large head turning, which effectively solves the drowsiness detection under occlusions, large posture changes and dimmer light.
- Extensive experiment results show that YOLOFaceMark achieves accurate facial landmark detection with a low normalized mean error (NME), while maintaining real-time inference speed. By leveraging the landmarks of the eyes and mouth estimated by YOLOFaceMark, we successfully detect eye-closing and yawning for driver drowsiness detection under different scenarios, as demonstrated through experiments on the Driver Drowsiness Detection dataset [11].

The rest of the paper is organized as follows. In Section II, we provide a review of related works on FLD and DDD. Section III presents the detailed description of our proposed approach. In Section IV, we compare and analyze the experimental results of our method with other existing methods. Finally, we draw our conclusions in Section V.

II. Related Works

In this section, we introduce the related work of facial landmark detection, followed by driver drowsiness detection based on facial landmark detection.

A. Facial Landmark Detection

FLD is an active research task in computer vision that involves localizing a set of pre-defined landmarks on human faces, such as the corners of the eyes, nose, and mouth. FLD has been studied over the years, with the development of deep learning techniques leading to significant improvements in accuracy and robustness. These methods can be broadly classified into two categories: regression-based methods and heatmap-based methods. Coordinate regression [12], as shown in Fig. 3(a), involves direct mapping of an input image to landmark coordinates. Extracted features are mapped to coordinates through a simple detection head with fully connected (FC) layers. Heatmap generation [8], as shown in Fig. 3(b), entails mapping an image to high-resolution heatmaps. Each heatmap symbolizes the likelihood of a landmark's specific position. Stacked hourglass network are the prevalent structure for generating high-resolution heatmaps. The distinction arises from the inherent simplicity of the detection head within the regression-based paradigm, resulting in diminished accuracy in localizing keypoints in comparison to heatmap-based methodologies. This observation holds true not only in facial landmark detection but also in other similar tasks such as HPE [8].

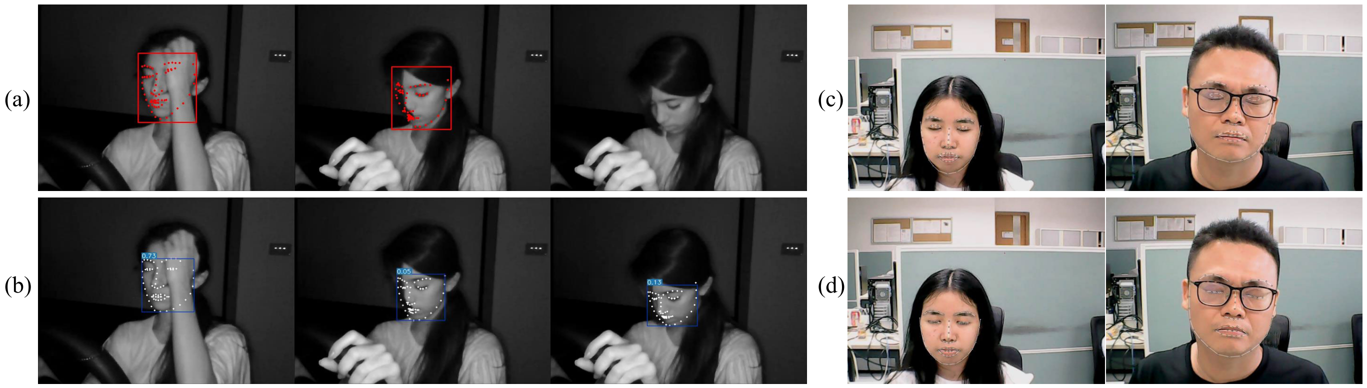


Fig. 2. FLD under some challenging conditions, such as driving at light, occlusion and extreme head pose. (a) False facial landmark detection and missed face detection occur when adopting the SOTA PIPNet [6]. (b) Accurate FLD and robust face detection can be achieved by our YOLOFaceMark. (c) Eye closures can not be recognized accurately by the SOTA 3DDFA2 [7]. (d) Accurate eye closures can be achieved by YOLOFaceMark. Our end-to-end approach offers the advantage of improved accuracy and robustness compared to the two-stage methods [6], [7].

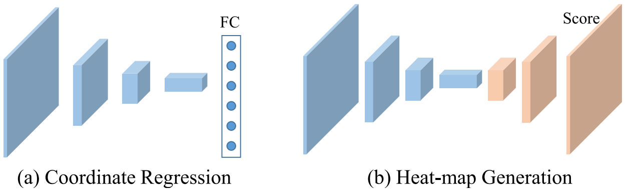


Fig. 3. Two different approaches for landmark localization.

Currently, SOTA performance has been achieved by the two-stage paradigm that first capture face using ground truth (GT) and then focus on designing a refined network for accurate FLD. Among them, high-resolution structures are adopted in [8], [9], and [13] to generate heatmaps for precise landmark localization. For example, ADA [9] relies on an attention-based network to formulate a fully convolutional regional architecture that predicts landmarks in high-resolution facial images without downsampling. In the case of AWing [13], the introduction of adaptive wing loss facilitates the analysis of ideal properties for heatmap regression in face alignment. Unlike the aforementioned methods that restore high-resolution representations from encoded low-resolution counterparts, HRNet [8] maintains high-resolution representations throughout the entire process by sustaining both high-to-low and low-to-high convolutional streams to achieve accurate keypoint estimation. To tackle the computational complexity associated with generating high-resolution heatmaps, a Pixel-in-Pixel Network (PIPNet) [6] is introduced. PIPNet employs a detection head that integrates local constraints from neighboring landmarks and conducts score and offset predictions simultaneously on low-resolution feature maps without using repeated upsampling layers. Some other popular FLD models, including the Convolutional Pose Machines (CPM) [14], Hourglass Networks [15], DLIB [16], 300W [17], and MTCNN [18], have been widely used in various applications.

Recently, there has been a trend towards adopting the single-stage paradigm for FLD. This paradigm treats FLD as an additional task during human face detection, thereby achieving efficient detection with low computational cost. For instance, YOLO5Face [12] accomplishes sparse facial keypoint detection for five points (centers of both eyes, nose tip, and both mouth corners). RetinaFace [19] achieves end-to-end detection of 68 facial landmarks, utilizing a different underlying detection model, RetinaNet [20]. However, detailed information regarding its implementation has not yet been disclosed. Consequently, the task of performing dense 68-point FLD based on the YOLO detection model represents a relatively new task. Moreover, achieving end-to-end dense key point detection is a challenging task due to issues such as non-convergence during training and insufficient localization accuracy.

B. Driver Drowsiness Detection Based on FLD

Facial landmark based DDD is an active research topic in computer vision and has gained significant attention in recent years due to their effectiveness and non-invasiveness. The task involves detecting facial landmarks and analyzing their movements to detect signs of fatigue in drivers.

Researchers in the field of visual features have conducted extensive research on this method. For instance, an ensemble approach is proposed in [21] that adopts four deep CNN models to detect various features such as hand gestures, facial expressions, behavioral features and head movements for DDD. A fuzzy logic algorithm [22] is proposed to determine the degree of fatigue after extracting eye and mouth information, while a real-time DDD system involved in information entropy is proposed in [23] that adopts the YOLOv3-tiny model to capture face and then uses the open DLIB toolkit [16] to capture the landmarks and the coordinates of the facial regions. Considering drowsiness as a continuous process, some methods incorporate temporal information into DDD. For instance, DrowsyNet [24] designs a sequential LSTM mode that considers consecutive 8-frame facial features to learn the spatiotemporal features for embedded systems. A parallel CNN [25] is proposed by using linked time-domain information for analyzing eye states in DDD. However, despite these advancements, several challenges still need to be addressed, such as real-time processing requirements, robustness to facial occlusions, variations in lighting conditions, head poses, and

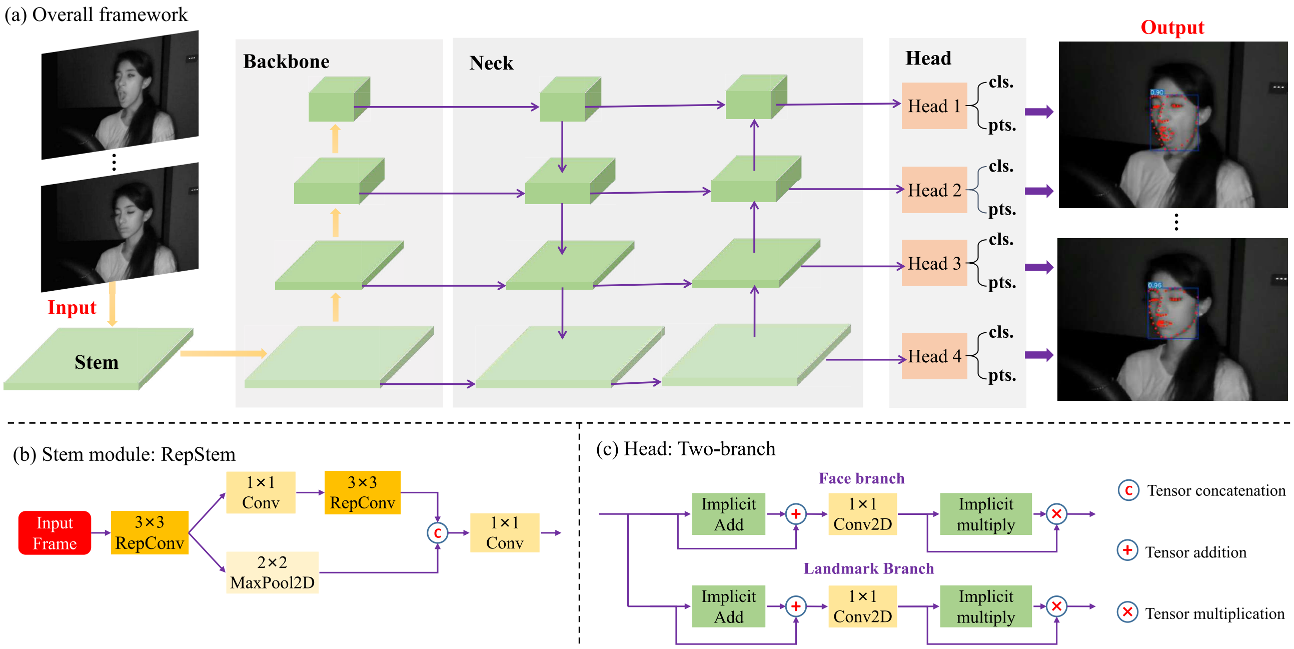


Fig. 4. The overall framework for FLD. (a) The structure of our proposed YOLOFaceMark. It contains four components (i.e., Stem, Backbone, Neck and Head). (b) The stem is designed to generate low-semantic features, which are fed into the backbone to generate the hierarchical pyramid features. The neck is designed with a bidirectional network to fuse these features. (c) The head is designed with two branches, one is for face detection and the other for FLD.

facial expressions. In this work, we propose YOLOFaceMark, a solution designed to tackle these challenges.

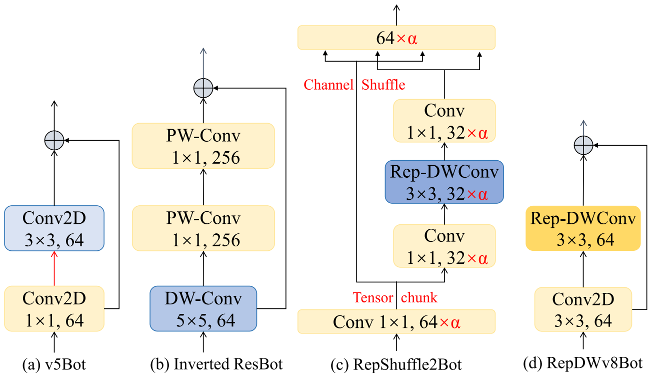
III. Methodology

In this section, we introduce our proposed YOLOFaceMark, which leverages the naive YOLOv5 architecture and incorporates a dual-branch detection head to jointly detect faces and their dense landmarks. By utilizing the estimated key facial landmarks, specifically those of the eyes and mouth, we analyze the driver's eye closure and yawning movements to detect signs of drowsiness.

A. Network Architecture

The overall architecture for FLD is illustrated in Fig. 4(a), which contains four components (i.e., Stem, Backbone, Neck and Head).

1) Stem: The stem receives the raw input image(s) and outputs low-level semantic feature maps. We try to design efficient stem that can learn abundant features with low computational cost. Our proposed stem shown in Fig. 4(b) is designed with two branches: the max-pooling branch to reduce computational cost, and the small-kernel (i.e., $K = 1, 3$) convolutional branch to increase the visual receptive fields of features. Different from the naive stem of YOLOv5 or YOLOv7, we adopt the RepConv to replace the original Conv3 $\times$ 3. The idea of RepConv is to use small-kernel (k $=$ 0, 1, 3) Conv in a multi-branch structure to learn rich features in the training stage and then re-parameterize their kernels to construct a single 3 $\times$ 3 kernel in the inference stage. Thus, it can learn rich features with different visual receptive fields in the training stage and can save computational cost by integrate multi-branches into a single stream in the inference stage. Both Conv and RepConv adopt the Sigmoid Linear Unit (SiLU) as the activation function.



(a) v5Bot (b) Inverted ResBot (c) RepShuffle2Bot (d) RepDWv8Bot

Fig. 5. Various bottlenecks. (a) The naive YOLO v5Bot. (b) The inverted ResBot [26]. (c) The RepShuffle2Bot (here, $\alpha = 2$). (d) The RepDWv8Bot. It adopts Rep-DWConv to replace the DWConv in residual bottleneck.

2) BottleNeck: Bottleneck as the basic block of a backbone plays a key role to learn effective features. Due to the success of ResNet [26], current mainstream backbones adopt residual connection to design the bottleneck. The naive YOLOv5 bottleneck, depicted in Fig. 5(a), simplifies the residual bottleneck with two Convs. Fig. 5(b) shows an inverted ResBot, where the DW-Conv is placed ahead of the PW-Conv. To further improve feature representation while minimizing computational costs, we present RepShuffle2Bot, illustrated in Fig. 5(c). In this design, the structural re-parameterization and channel shuffling techniques are employed, where $\alpha = 2$ signifies a doubling of the input channels. We also introduce a tiny version of YOLOFaceMark (i.e, YOLOFaceMark-t) that keep the channel number (here, $\alpha = 1$) thus saving lots of computation cost with slight performance degradation. This variant is well-suited for deployment on embedded systems with limited computational resources. Furthermore, we integrate the latest YOLOv8 as the basic framework, adapting it for our specific joint tasks. We employee re-parameterization and modify the original v8Bot to construct the RepDWv8Bot, as shown in Fig. 5(d).

3) Neck: Neck is designed to aggregate the hierarchical features generated by the backbone. Efficient neck can enable more information interaction among features of different granularity. It provides the multiscale features, which are fed into the detection head for robust detection especially for the multi-scale target detection. Our neck adopts the naive path aggregation network (PAN). The structure is bidirectional: the top-down path continuously fuses high-level semantic features with those of low-level semantics through up-sampling, while the bottom-up path continuously fuses low-level semantic features with these high-level semantic features through convolutional operations. We adopt the proposed RepShuffleBot to design the Cross-Stage-Partial-connections (CSP) module, which is a widely-adopted lightweight block in naive YOLOv5. We call it RepShuffleCSP. This leads to obtaining multiscale and rich features through two reversible processes.

4) Head: The detection head is designed as an adapter to generate results adapted to the target tasks. Beyond the naive detection head of YOLOv5 that regresses the target bounding-boxes (BBox), we add an extra branch to regress the facial landmarks. Fig. 4(c) shows our proposed detection head with two branches, the face branch and the landmark branch. Each branch contains two extra modules, i.e., the implicit addition module (called ImplicitA) and the implicit multiplication module (called ImplicitM). Both of them contain a vector to learn implicit knowledge for accurate FLD. The vector has the same dimension with the input features. ImplicitA is added with the input vector while ImplicitM is multiplied with the input vector. They act the roles of performing shift and scale transformation on the input features, respectively.

B. Face Detection

Based on YOLOPose [10], we modify the human labels to the face labels in our YOLOFaceMark. For each anchor, its detection head outputs three kinds of information: the BBox location, box confidence, and face confidence. The BBox location can be presented by four elements, i.e., the box center (C_x, C_y) and its width and height (w, h). Above three kinds of information can be summarized by a vector with 6 elements, which is generated by the detection head for each anchor. The vector for face detection can be presented as:

$$O_f = (C_x, C_y, w, h, box_{conf}, face_{conf}). \tag{1}$$

C. Facial Landmark Detection

We further extend the original YOLOPose and change the task of the 17 human keypoint regression to the 68 facial landmark regression. Whether it is a human keypoint or a facial landmark, each point information contains its coordinate and confidence: $(L_x, L_y, conf)$. All 68 facial landmarks correspond to 204 element outputs. The landmark head of our model outputs a 204-element vector in each anchor, which is defined as:

$$O_l = (L_x^1, L_y^1, L_{conf}^1, \ldots, L_x^{68}, L_y^{68}, L_{conf}^{68}). \tag{2}$$

D. Loss Function

Various loss functions have been proposed to solve the task of point regression, such as the widely-used L1, smooth-L1, and L2. MTCNN [18] adopts the L2 loss function to jointly detect faces and their five facial landmarks. YOLO5Face [12] adopts Wing-Loss [13] as the loss function for facial keypoint regression to overcome the problem that L-norm functions are not sensitive to small errors approaching 0. However, these loss functions do not consider object scales and keypoint types.

Since both keypoint detection and object detection tasks consider localization, MS COCO [27] proposes the object key similarity (OKS) metric to evaluate keypoint regression. OKS acts a similar role as the intersection over union (IoU) in the object detection task. By following the practice of YOLOPose, we adopt OKS strategy as the loss function. We refer to the OKS definition in MS COCO, which is formulated as:

$$OKS = \textstyle\sum_i [exp(-d_i^2/2s^2k_i^2)\delta(v_i > \theta)]/\sum_i[\delta(v_i > \theta)], \tag{3}$$

where d_i denotes the Euclidean distance between the predicted keypoint and GT; v_i represents the visibility flag of GT; s is the object scale, and k_i is a per-keypoint constant that controls the falloff. OKS ranges from 0 to 1.

E. Driver Drowsiness Detection

Facial landmark based DDD gain attention in recent years due to their effectiveness and non-invasiveness. The task involves detecting facial landmarks and analyzing signs of fatigue in drivers, such as eye closure and yawning by calculating parameters such as distance, angle between key points of the eyes and mouth. Eye Aspect Ratio (EAR) is a measure of the ratio of the height to the width of an eye. The common EAR [28] is defined as:

$$EAR_{com} = \frac{d_1 + d_2}{2 \times d_0}, \tag{4}$$

where d_i represents the euclidean distances between the corresponding two points. EAR is a widely-used measure to analyze the sleepy eyes (shown in Fig. 6(b)) for DDD.

By referring to the definition of EAR, we define Mouth Aspect Ratio (MAR) to measure the ratio of the height to the width of a mouth (shown in Fig. 6(c)). It is formulated as:

$$MAR_{com} = \frac{d_5}{d_6}. \tag{5}$$

We adopt MAR to determine if the driver is yawning and further evaluate the driver's fatigue status. The combination of EAR and MAR can provide more accurate DDD analysis.

When extreme head postures (i.e., tilting left and right or nodding) occur, the common EAR or MAR cannot clearly reflect their occurrence. Since the changes in the global face height and width are more sensitive to head postures than local variations in eye or mouth, we define an improved EAR that

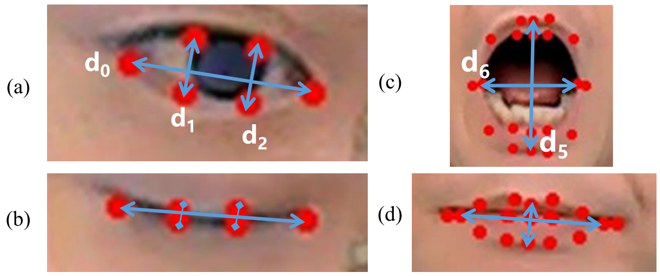


Fig. 6. Illustration of the open and closed states of eyes and mouth. Geometric distances of corresponding points enable the calculation of EAR and MAR.

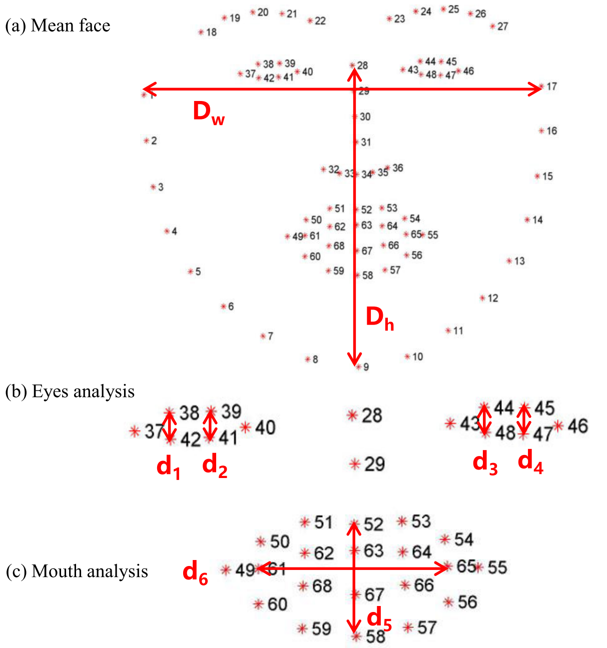


Fig. 7. DDD analysis. (a) Mean face [17]. (b) Eye analysis. (c) Mouth analysis.

employs the global face size D_{ref} instead of the local eye width d_0. The formulation is as follows:

$$D_{ref} = \frac{D_h + D_w}{2},$$
$$EAR_{ours} = \alpha \times (\frac{d_1 + d_2}{2 \times D_{ref}} + \beta), \tag{6}$$

where D_h, D_w represent the height and width of a face in Fig. 7(a). α represents the scaling ratio between the reference size D_{ref} and the actual eye width d_0. β denotes the bias associated with the eye-closed state. As illustrated in Fig. 6(a), we can see a discernible gap between the upper and lower eyelids persists even when the eyes are closed. Fig. 7(a) displays the specifications for facial annotations on the 300W dataset, representing the average model of facial landmark on that dataset. By utilizing the coordinates from the mean face, we can calculate D_{ref} and d_0, resulting in $\alpha = 6$. Additionally, by calculating the gap between the upper and lower eyelids, we calculate $\beta = -0.02$. This calibration ensures that the EAR value attains zero when eye closure occurs.

Similarly, we leverage the global face height and width to enhance the definition of MAR. Our proposed MAR definition is expressed as:

$$d_{5\_ref} = \gamma \times (\frac{d_5}{D_{ref}} + \delta),$$
$$d_{6\_ref} = \frac{d_6}{D_{ref}},$$
$$MAR_{ours} = \frac{d_{5\_ref}}{d_{6\_ref}} = \frac{\gamma \times d_5 + \delta \times D_{ref}}{d_6}, \tag{7}$$

where γ represents the scaling factor between the mouth height d_5 and the mouth width d_6. Fig. 6(c) shows a yawning movement in which we can get a maximum γ as 1.68 by calculating the geometric distances of corresponding points in the mean face. δ represents the bias related to the mouth-closed state, as indicated in Fig. 6(d). Moreover, by calculating the gap between the upper and lower lips, we calculate $\delta = -0.13$. This adjustment ensures that the MAR value reaches zero when mouth closure happens.

IV. Experiments and Analysis

In this section, we provide a comprehensive analysis of the experimental design and its results, including datasets, evaluation metrics, ablation studies, and performance comparisons.

A. Datasets and Evaluation Metrics

We evaluate the performance of our YOLOFaceMark for FLD on two benchmark datasets: 300W [17] and COFW [29]. Additionally, we assess the effectiveness of our method for DDD on the Driver Drowsiness Detection dataset [11].

1) 300 Faces in-the-Wild (300W): 300W is widely used for FLD. It is a collection of existing datasets that includes LFPW, AFW, Helen, XM2VTS, and Ibug. It comprises two parts: the training and testing set. The training set contains 3148 images derived from LFPW, Helen, and the full AFW. The testing set includes 689 images derived from LFPW, Helen, and the full Ibug. The testing set can be further divided into two parts based on their degree of difficulty: the common set and the challenging set. The common set is from LFPW and Helen while the challenge set is from Ibug.

2) Caltech Occluded Faces in the Wild (COFW): COFW is a challenging benchmark for FLD. It contains 1345 training images and 507 test images that are collected from various backgrounds, including multiscale changes, occlusions and invisible parts. Each face has 29 landmarks initially annotated. The test set is re-annotated with 68 landmarks to form the COFW-68 dataset [29]. We use the COFW-68 dataset as another test set to verify the robustness and generalization of cross-dataset detection.

3) National Tsuing-Hua University Driver Drowsiness Detection (NTHU-DDD): NTHU-DDD [11] is widely used among the image-based DDD systems. It covers various scenarios (day and night illuminations) and drowsiness features, including: normal driving; yawning; slow blink rate; falling asleep. The dataset includes training, evaluation, and test datasets and contains recorded videos for 36 subjects from different ethnicities. Here, we take the evaluation dataset that contains videos with drowsy and non-drowsy features, mixed under multiple scenarios.

TABLE I
DATA AUGMENTATIONS IN TWO DIFFERENT TRAINING STAGES

Stage	rotate	translate	scale	fliplr	mosaic	mixup
Training	0	0.1	0.5	0.5	1.0	0.0
Fine-tune	0.373	0.245	0.898	0.5	1.0	0.243

4) Evaluation Metrics: We achieve the joint detection of faces and facial landmarks. They have different evaluation metrics. To evaluate the face detection accuracy, we use the common-used strategy of mean average precision (mAP) [27] over the IoU thresholds. Following the widely-used indicator in most detecting tasks, the precision is also adopted in face detection task. The indicator is defined as:

$$P = \frac{TP}{TP + FN}, \tag{8}$$

where P delegates the precision; TP, FN represent the true positive and false negative respectively. To evaluate the FLD accuracy, we adopt the strategy of normalized mean error (NME) by normalizing the interocular distances. The drowsiness state can be directly indicated by the EAR and MAR.

B. Experimental Settings

1) Training: The initial YOLOv5 is pre-trained on the MS COCO dataset [27]. For training our specific YOLOFaceMark, we utilize a GPU server equipped with a GeForce RTX2080. Each GPU has a maximum batch size of 16. We employ the SGD algorithm as our optimizer, with an initial learning rate of 1e-2. The training process is terminated at the 300-th epoch. The input resolution for training is fixed at 640 × 640. To preprocess the input frames, we follow the common practice in naive YOLOv5. The long side of the input frame is resized to the desired size, and the short side is padded to achieve a square image. Further elaboration of data augmentations employed in both stages can be found in Table I.

2) Testing: Our proposed YOLOFaceMark is designed as an end-to-end solution for FLD. For testing, we use a batch size of 16 and ensure that the input frames are consistently adjusted to a fixed resolution of 640 × 640. The inference speed of our model is specifically evaluated at this resolution.

C. Hyperparameter Learning

Our task involves the joint detection of faces and their corresponding keypoints. In this pursuit, two hyperparameters influence the face detection performance: the threshold for face-label confidence and the IoU threshold between anchor box and ground-truth bounding-box. Through dedicated experimentation using the full 300W dataset and NTHU-DDD, we aim to attain a nuanced balance.

For the face-label confidence threshold, we explore three typical values: 0.002, 0.02, and 0.1. Regarding the IoU threshold, we investigate two values: 0.50 and 0.65. As indicated in Table II, the most promising results surface when the face confidence threshold is set to 0.002, and either an IoU threshold of 0.65 or 0.50 is applied. Considering the sensitivity of the system, directed towards preserving accuracy while curbing false positives, our final choice leans towards adopting the higher IoU threshold of 0.65. This deliberate selection guarantees that candidate boxes closely align with the ground-truth bounding boxes, fostering a strong correspondence between the detected and actual objects.

TABLE II
THRESHOLD HYPERPARAMETER LEARNING

Face Conf.	IoU	mAP@0.5	mAP
0.002	0.50	**97.9**	**92.6**
0.02	0.50	97.6	92.5
0.1	0.50	96.7	92.0
0.002	0.65	**97.9**	**92.6**
0.02	0.65	97.6	92.5
0.1	0.65	96.7	92.0

Bold represents the best result.

D. Ablation Study of BottleNeck

The bottleneck plays a crucial role in extracting effective and efficient features. Our goal is to design an optimal bottleneck that can enhance the accuracy of FLD while also saving computational cost. In Fig. 5, we illustrate three types of bottlenecks: the Darknet bottleneck (DarkBot) [36], the Inverted ResNet bottleneck (Inverted ResBot) [26], and our proposed RepShuffleαBot that incorporates structural re-parameterization and channel shuffling modules.

In the end-to-end paradigm of Table III, we compare the performance of different bottlenecks employed by our YOLOFaceMark. It is evident that YOLOFaceMark utilizing the RepShuffle2Bot achieves the highest mAP of 92.1% and a low NME of 3.91 on the full 300W test set. Furthermore, it has a computational cost of 16.9 GFlops, which is lower than that of the DarkBot-based model. The tiny model with RepShuffleBot (here, $\alpha = 1$) achieves the fewest parameters and GFlops while still maintaining a satisfactory NME of 4.21.

E. FLD Performance Comparisons With SOTAs

1) Tests on 300W: Table III provides a comprehensive analysis of the performance of state-of-the-art methods for facial landmark detection on the 300W dataset. These methods follow the two-stage paradigm, while our approach adopts an end-to-end paradigm. It should be noted that all the listed two-stage methods utilize ground-truth annotations to locate the human face, with some methods [6], [13] resizing the face bounding box to a higher resolution for improved performance. In contrast, our end-to-end paradigm employs the YOLOFaceMark framework to jointly detect the face and its key landmarks without additional computational cost.

Examining Table III, we observe that STAR [35] achieves the best NME of 2.87 on the 300W-Full subset with resizing the face GT to a relatively large resolution. ADNet [34] attains the second-best NME of 2.93 through the generation of three distinct heatmaps, namely landmark, edge and point heatmaps. In contrast to these two-stage SOTA methods that utilize GT to extract faces, our YOLOFaceMark can simultaneously detect faces and their facial landmarks in an end-to-end

TABLE III

FLD PERFORMANCE COMPARISON WITH OTHER ADVANCED METHODS.GT INDICATES THE GROUND TRUTH OF FACE DETECTION

Method	Backbone	Input Size	Params	Flops	mAP	NME (300W)			NME
			(M)	(G)		Full	Common	Challenging	COFW
Two-stage paradigm									
DAC-CSR [30]	-	100×100	-	-	GT	-	-	-	6.03
Lab [31]	ResNet-18	-	52.4	29.1	GT	3.49	2.98	5.19	5.58
Wing [32]	ResNet-50	-	91.0	5.5	GT	-	-	-	5.07
HRNet [8]	HRNetV2-W18	256×256	9.7	**4.8**	GT	3.32	2.87	5.15	3.45
ADA [9]	Hourglass	256×256	-	-	GT	3.50	**2.41**	5.68	-
PIPNet [6]	ResNet-50	256×256	26.7	5.6	GT	3.24	2.80	5.03	3.18
AWing [13]	Hourglass	256×256	24.1	26.7	GT	3.07	2.72	4.52	-
RepFormer [33]	ResNet-50	256×256	-	-	GT	3.03	-	-	**3.01**
ADNet [34]	Hourglass	256×256	-	-	GT	2.93	2.53	4.58	-
STAR [35]	Hourglass	256×256	13.4	-	GT	**2.87**	2.52	**4.32**	-
End-to-end paradigm									
Ours	Inverted ResBot	640×640	11.8	16.2	91.2	4.15	3.62	6.30	4.62
	RepShuffleBot	640×640	9.9	11.8	91.5	4.21	3.43	6.65	4.58
	RepShuffle2Bot	640×640	13.3	16.9	92.1	3.91	3.31	6.43	4.52
	RepShuffle2Bot*	640×640	13.3	16.9	91.9	3.88	3.29	6.08	4.46
	RepDWv8Bot-t	640×640	**5.1**	17.7	**92.5**	3.85	3.29	6.18	4.50

These two-stage methods use GT to extract face and do not consider the computational cost in the face-detection stage. ‘-’ indicates null or not given. * means our fine-tuned model. ‘t’ means our tiny model.

paradigm. Despite YOLOFaceMark exhibiting a slightly lower accuracy compared to these two-stage SOTA methods, it still achieves a commendable NME of 3.88 on the complete 300W. When using the Tiny YOLOv8 with RepDWv8Bot, it demonstrates a significant reduction in parameter count, standing at merely 5.1 million, nearly a 60% decrease compared to utilizing YOLOv5 with RepShuffle2Bot. Despite the reduction in parameters, this configuration achieves an improved NME of 3.85.”

2) Cross Tests on COFW: To assess the generalization ability of YOLOFaceMark, we conduct experiments on the cross-dataset COFW [29]. This involves training YOLOFaceMark on the 300W dataset and then testing it on the challenging COFW dataset. Table III also presents the NME results on COFW for YOLOFaceMarks and several SOTA methods. It shows that our fine-tuned YOLOFaceMark with RepShuffle2Bot achieves an NME of 4.46, which falls behind these SOTAs, such as RepFormer [33], PIPNet [6] and HRNet [8]. It is important to note that these two-stage methods utilize GT to locate faces and scale up the face BBox to a higher resolution (e.g., 256 × 256) for enhanced performance. In contrast, our approach involves detecting the face BBox in the entire large image (with a resolution of 648 × 648) and subsequently detecting the corresponding facial landmarks.

F. Robust Face and Eye-Closure Detection

The detection of eye closure is a crucial aspect of facial landmark-based DDD. However, publicly available datasets like 300W and COFW have a limited number of face samples with closed eyes. In 7(a), we present the mean facial keypoint model derived from the 300W dataset, which illustrates that the common state of human eyes is open. Since the YOLOFaceMark deep learning model relies on data-driven supervised learning, it learns the knowledge and patterns associated with open-eye states from the dataset. As a result, the detection performance for closed eyes is relatively poor, as depicted in Fig. 8(a) and (c).

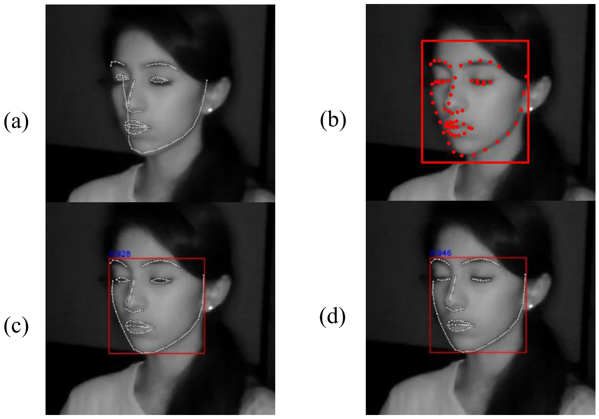


Fig. 8. Eye closure detected by various methods. (a) 3DDFA2 [7]. (b) PIPNet [6]. (c) Our model. (d) Our fine-tuned model.

YOLOFaceMark is initially pretrained on the 300W dataset, which includes fewer eye closure samples. In order to achieve accurate eye closure detection, it is necessary to gather more positive samples of eye closures to retrain our model. However, annotating dense facial keypoints for these samples is a labor-intensive and time-consuming task. To address this, we employ a “student learns from teacher” strategy to learn the eye closure pattern. We create a high-quality training set by capturing a video that simulates driver drowsiness in a controlled laboratory environment with good lighting

TABLE IV
FACE DETECTION PERFORMANCE COMPARED WITH OTHER METHODS

Method	VJ [37]	DLIB [16]	MTCNN [18]	Ours
FPS	**125**	45	40	98
P(%)	87.0	90.7	97.3	**99.1**

P(%) is calculated on NTHU-DDD [11].

and a simple background. We use PIPNet [6] to detect eye closures in the video and use the results to form a new training set. We then fine-tune our model using this newly collected set, which contains a significant number of eye closure samples.

During the fine-tuning stage, we employ a more robust data augmentation strategy compared to the original training stage. The aim is to enhance the robustness of face detection and the accuracy of facial landmark detection. Fig. 8(d) illustrates that our fine-tuned model can effectively detect faces with closed eyes in various challenging conditions, such as driving at night. Face performance is also compared in Table IV, from which we can see that our fine-tune model can achieve the highest precision of 99.1% on NTHU-DDD [11] with an inference speed of 98 FPS. It should be stated that our methods contains the task of FLD beyond face detection. Furthermore, we compare the performance of the original and fine-tuned models in the end-to-end paradigm of Table III, which clearly demonstrates that the fine-tuned model achieves improved accuracy with the lowest NME of 3.88 among our proposed models on 300W testset.

G. Real-Time Inference Speed

In practical applications, high-speed performance is crucial especially in facial landmark-based DDD,. YOLOFaceMark meets the real-time demand. Inference experiments are conducted on a workstation equipped with a CPU (Intel(R) Xeon(R) Gold6134@3.20GHz) and a GeForce RTX2080 GPU. We calculate the inference speed by testing the validation sample at 640 × 640 resolution with a batch size of 1.

The time consumption mainly consists of two processes: inference and non-maximum suppression (NMS), which take 9.6 and 0.6 milliseconds respectively. As a result, we achieve a real-time test speed of 98 frames per second (FPS). When utilizing the tiny model, it achieves a speed of 102 FPS with less computational cost. Table V compares the speed performance with other SOTA methods. Notably, while our method exhibits lower accuracy compared to SOTAs like RepFormer [33] and PIPNet, it achieves nearly 100 FPS, making it approximately 4-5 times faster than them.

It should be noted that when applying these two-stage SOTA methods to practical applications, the time consumed by the face detector should be considered, as ground-truth is not provided in wild images. Fig. 9 illustrates that YOLOFaceMark performs well in challenging scenarios.

TABLE V
INFERENCE SPEED COMPARISON WITH OTHER METHODS

Method	Resolution	GPU	NME	FPS
Two-stage paradigm				
DAC-CSR [30]	100×100	-	6.03	10
LAB-18 [31]	256×256	-	5.58	17
Wing-50 [32]	256×256	-	5.07	30
HRNet [8]	256×256	-	3.45	12
PIPNet-50+FB [6]	256×256	RTX2080	3.18	20
STAR [35]	256×256	RTX2070	-	14
RepFormer-50+FB [33]	256×256	RTX2080	**3.01**	20
End-to-end paradigm				
Ours-t	640×640	RTX2080	4.58	**102**
Ours	640×640	RTX2080	4.46	98

NME is calculated on COFW-68. '-t' means our tiny model.

H. Driver Drowsiness Detection

To assess the effectiveness of our approach, we conduct experiments on three different scenes, ranging from easy to difficult. In this context, we employ our proposed MAR and EAR metrics, as defined in Equation (6) and (7), to analyze the driver's drowsiness status.

Firstly, we conduct experiments using short videos capturing diverse individuals displaying drowsiness within controlled laboratory conditions. As illustrated in Fig. 10, an eye-closure event is indicated by the EAR curve when the EAR value falls below the threshold of 0.2. Similarly, a yawning event is indicated by the MAR curve when the MAR value exceeds the threshold of 0.5. Furthermore, MAR values below the threshold of -0.1 correspond to the dozing state. Notably, we observe that the two troughs in the curve correspond to instances where a person is sleeping with their head down. This demonstrates that our MAR design effectively considers extreme head postures, such as dozing. Fig. 11 shows more results of fatigue driving detection in simulated driving scenarios.

Secondly, we evaluate our method in driving scenarios, both during daytime and nighttime conditions, using the DDD evaluation set. Fig. 12 shows the MAR and EAR curves under daytime conditions. Significant fluctuations in the EAR curve indicate intermittent and drowsy driving, especially after the MAR curve showed yawning and the value of the EAR curve dropped below the threshold several times.

Finally, we verify the viability of the system in dimmer light. FLD detector faces robustness challenges in pitch-dark nighttime scenes due to limited illumination, which may lead to missed or false detections. However, YOLOFaceMark consistently detects faces and actions such as eye closure and yawning. Fig. 13 show sharp fluctuations in EAR, with thresholds below 0.2 indicating eye closure. In the MAR curve, a threshold above 0.5 indicates yawning. Given the gradual nature of driver fatigue, we supplement our approach with PERCLOS [38] to evaluate the drowsiness state. It is essential to note that PERCLOS operates within the frame-by-frame detection paradigm, setting it apart from the approach adopted

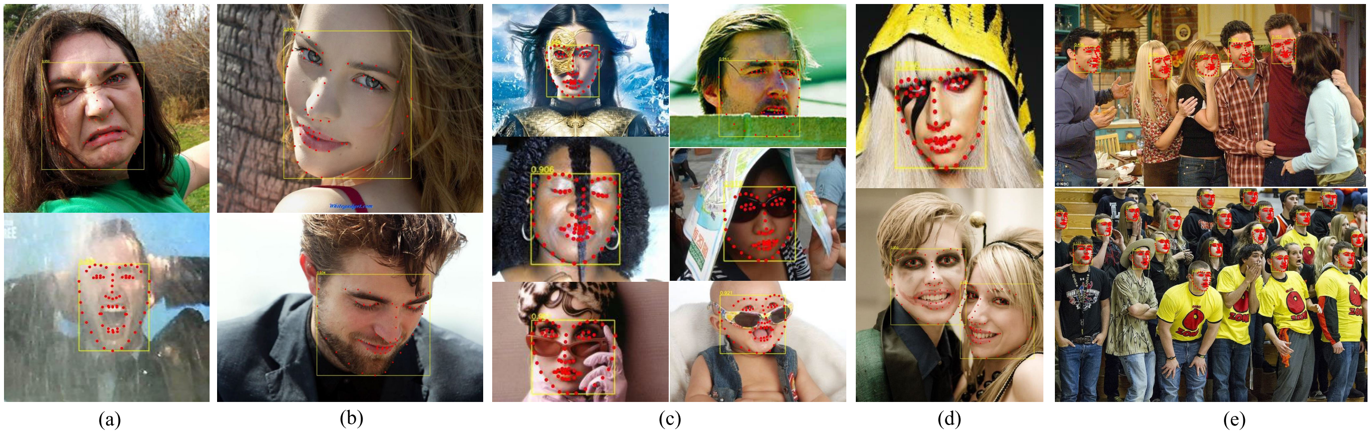


Fig. 9. More FLD results under challenging conditions, i.e., (a) expressions, (b) viewpoints, (c) occlusions, (d) makeups and (e) crowds. These challenging scenes are derived from 300W, primarily used to illustrate the robustness of our FLD method in handling complex situations.

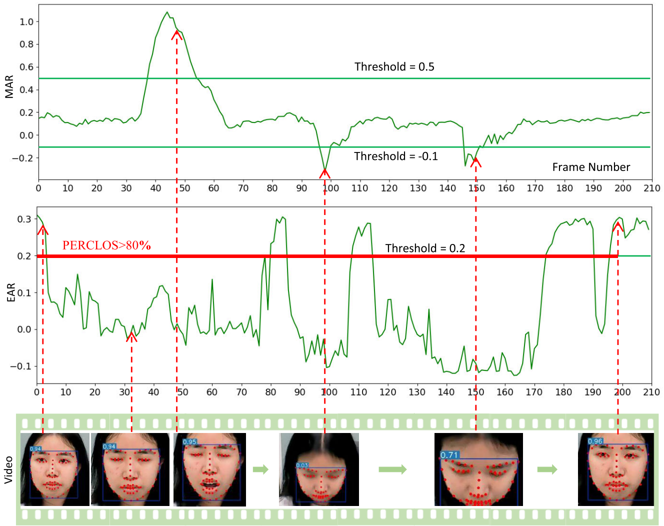


Fig. 10. EAR and MAR analysis in laboratory condition. The bold red lines on the timeline (i.e., frame number) donates the intervals corresponding to sleep processes identified by PERCLOS.

by Drowsynet [24], which employs contiguous-frame state blocks. The additional utilization of PERCLOS introduces minimal computational overhead, requiring computation solely for intervals surpassing the 80% threshold based on our refined EAR within each video frame. These experiments demonstrate the successful detection of eye closure and yawning by YOLOFaceMark across different scenarios, enabling the determination of drowsiness levels.

I. Analysis of Advantages and Disadvantages

We conduct an analysis of our single-stage paradigm for FLD. Our end-to-end model demonstrates real-time and robust detection of faces and their corresponding facial landmarks. It offers several key advantages: First, it combines face detection and FLD into a unified framework. This approach improves computational efficiency as the two tasks can share deep feature maps, reducing unnecessary redundant computation. Second, the robustness of face detection can be enhanced through data augmentations during the preprocessing stage. These augmentations have no impact on the inference speed, ensuring resilient face detection even in challenging scenarios such as nighttime driving. Third, our approach achieves accurate FLD even during instances of eye-closure, a pivotal component for DDD. We employ the approach of "learning from teacher" to gather a substantial dataset of closed-eye instances. Leveraging this dataset, we fine-tune our model to accurately localize landmarks within closed-eye scenarios, facilitating effective analysis of yawning states.

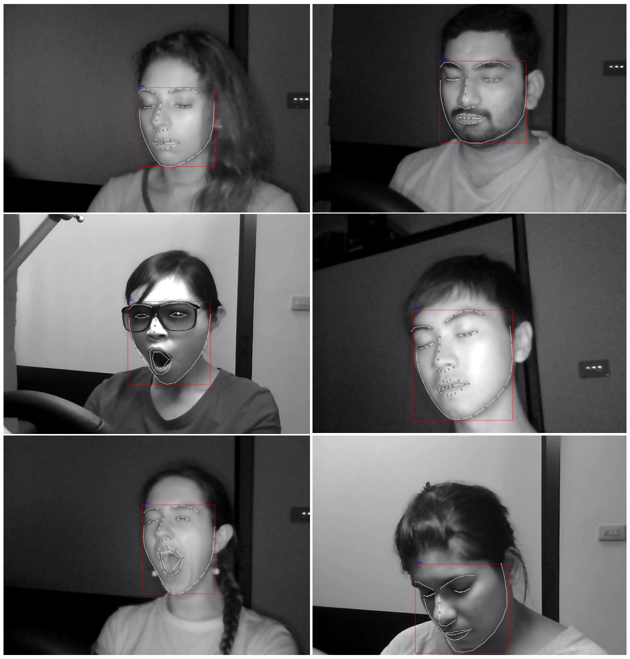


Fig. 11. More fatigue driving detection results in simulated driving scenarios.

While YOLOFaceMark demonstrates impressive capabilities, certain limitations also need to be addressed. (1) Challenges with occlusion and extreme poses. When the human eyes are partially occluded or the face is rotated at an extreme angle, the system encounters difficulties in detecting eye closure status. This leads to an inability to accurately judge driver drowsiness. (2) Limited precision in comparison to other SOTA approaches (i.e., PIPNet [6] and HRNet [8]). Despite

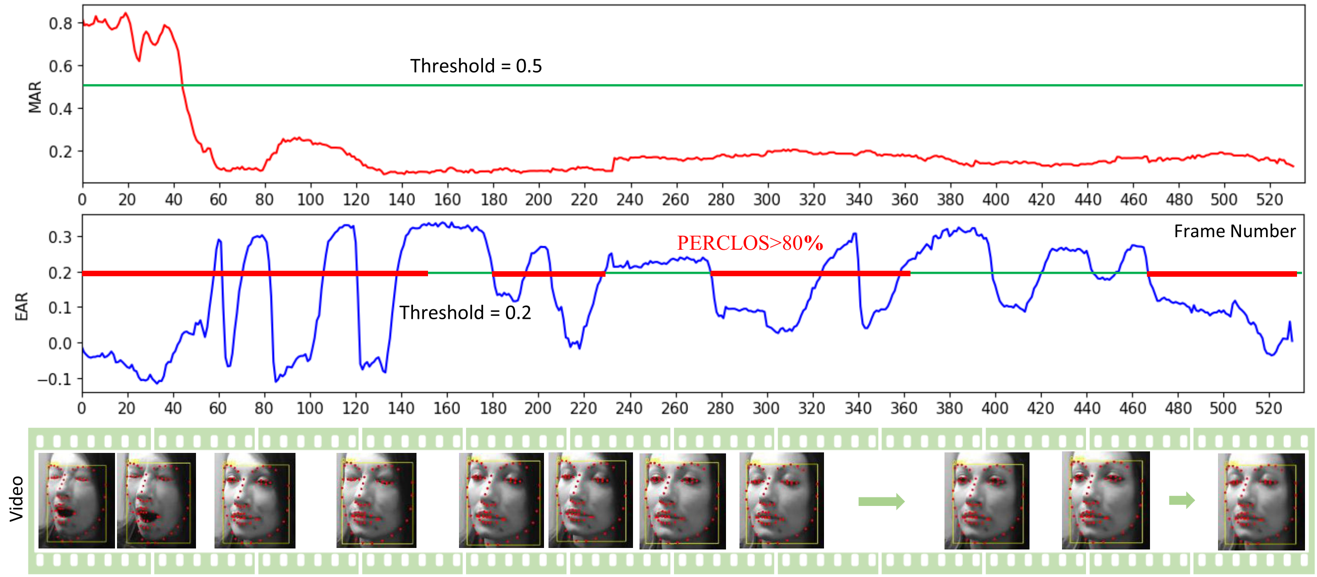


Fig. 12. EAR and MAR analysis under daytime condition on DDD evaluation dataset. The bold red lines on the timeline (i.e., frame number) donates the intervals corresponding to sleep processes identified by PERCLOS.

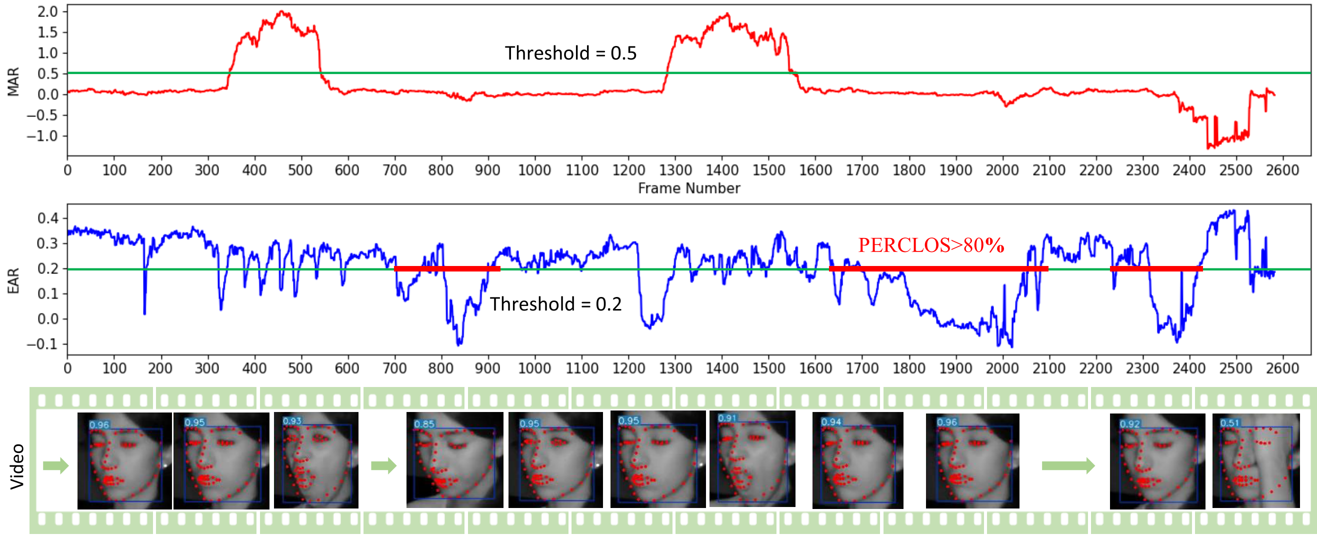


Fig. 13. EAR and MAR analysis under nighttime condition on DDD evaluation dataset.

achieving real-time performance, YOLOFaceMark still falls short in terms of accuracy when juxtaposed with two-stage heatmap-based methods. Since YOLOFaceMark adopts the coordinate regression paradigm, its fixed connections to specific feature map locations result in inaccuracies and biases. Further improvements are required to address the identified limitations and continue refining our framework, potentially by incorporating additional contextual information.

V. Conclusion

Leveraging facial landmark-based driver drowsiness detection presents a promising avenue to enhance road safety by promptly identifying drowsy drivers. This study introduces an end-to-end model named YOLOFaceMark, facilitating simultaneous detection of faces and their dense landmarks. Through the implementation of cost-effective operations, including re-parameterizing the stem, incorporating re-parameterization and channel shuffling in the bottleneck, and designing a dual-branch detection head with an implicit module, we achieve accurate and robust FLD. Various experiments demonstrate that YOLOFaceMark can effectively detect precise face and facial landmarks in real-time. We validate its generalization across datasets using COFW. Moreover, by relying on the key landmarks of the eyes and mouth estimated by YOLOFaceMark, we can analyze driver drowsiness by detecting eye-closing and yawning movements.

References

[1] M. N. Azadani and A. Boukerche, "Driving behavior analysis guidelines for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6027–6045, Jul. 2022.

[2] M. Gromer, D. Salb, T. Walzer, N. M. Madrid, and R. Seepold, "ECG sensor for detection of driver's drowsiness," *Proc. Comput. Sci.*, vol. 159, pp. 1938–1946, Jan. 2019.

[3] Y. Jiang, Y. Zhang, C. Lin, D. Wu, and C.-T. Lin, "EEG-based driver drowsiness estimation using an online multi-view and transfer TSK fuzzy system," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1752–1764, Mar. 2021.

[4] Y. Fan, F. Gu, J. Wang, J. Wang, K. Lu, and J. Niu, "SafeDriving: An effective abnormal driving behavior detection system based on EMG signals," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12338–12350, Jul. 2022.

[5] Z. Qiu, J. Zhao, and S. Sun, "MFIALane: Multiscale feature information aggregator network for lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24263–24275, Dec. 2022.

[6] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel Net: Towards efficient facial landmark detection in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3174–3194, Sep. 2021.

[7] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 152–168.

[8] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[9] P. Chandran, D. Bradley, M. Gross, and T. Beeler, "Attention-driven cropping for very high resolution facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5860–5869.

[10] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2637–2646.

[11] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Proc. Int. Workshops Comput. Vis.*, vol. 13. Cham, Switzerland: Springer, 2017, pp. 117–133.

[12] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why reinventing a face detector," 2021, *arXiv:2105.12931*.

[13] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6971–6981.

[14] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[15] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 483–499.

[16] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.

[17] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016.

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[19] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5203–5212.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[21] M. Dua, R. Singla, S. Raj, and A. Jangra, "Deep CNN models-based ensemble approach to driver drowsiness detection," *Neural Comput. Appl.*, vol. 33, no. 8, pp. 3155–3168, Apr. 2021.

[22] M. S. Devi and P. R. Bajaj, "Fuzzy based driver fatigue detection," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Istanbul, Turkey, Oct. 2010, pp. 3139–3144.

[23] F. You, Y. Gong, H. Tu, J. Liang, and H. Wang, "A fatigue driving detection algorithm based on facial motion information entropy," *J. Adv. Transp.*, vol. 2020, pp. 1–17, Jun. 2020.

[24] S. Zu, Y. Jin, D. Yang, and H. Xu, "DrowsyNet: Multivariate time series classification for embedded driver drowsiness detection," in *Proc. 8th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2022, pp. 442–451.

[25] J.-Y. Shiau, K. Nishiyuki, S. Nagae, T. Yabuuchi, K. Kinoshita, and Y. Hasegawa, "Driver drowsiness estimation by parallel linked time-domain CNN with novel temporal measures on eye states," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 937–942.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[28] T. Soukupova and J. Cech, "Eye blink detection using facial landmarks," in *Proc. 21st Comput. Vis. Winter Workshop*, Rimske Toplice, Slovenia, 2016, pp. 1–8.

[29] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1899–1906.

[30] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3681–3690.

[31] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2129–2138.

[32] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2235–2245.

[33] J. Li, H. Jin, S. Liao, L. Shao, and P.-A. Heng, "RePFormer: Refinement pyramid transformer for robust facial landmark detection," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1088–1094.

[34] Y. Huang, H. Yang, C. Li, J. Kim, and F. Wei, "ADNet: Leveraging error-bias towards normal direction in face alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3060–3070.

[35] Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji, "STAR loss: Reducing semantic ambiguity in facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15475–15484.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[37] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2001, pp. 511–518.

[38] U. Trutschel, B. Sirois, D. Sommer, M. Golz, and D. Edwards, "PERCLOS: An alertness measure of the past," in *Proc. Driving Assesment Conf.*, vol. 6. Iowa, IA, USA: University of Iowa, 2011, pp. 172–179.

Qingtian Wu (Member, IEEE) received the B.S. degree in information engineering from Shenzhen University in 2014, the M.S. degree in pattern recognition from the University of Chinese Academy of Sciences in 2017, and the Ph.D. degree in computer science from the University of Macau in 2024.

He was a Research Associate with the National University of Singapore in 2019. He is currently a Lecturer with the School of Artificial Intelligence, Shenzhen Polytechnic University. His research interests include computer vision and stable diffusion-based content generation.

Nannan Li (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, in 2015. Then, he was a Post-Doctoral Researcher with the School of Electronic and Computer Engineering, Peking University. Since 2021, he has been joining Macau University of Science and Technology as an Assistant Professor. His research interests include computer vision, artificial intelligence, and machine learning.

Liming Zhang (Senior Member, IEEE) received the B.S. degree in computer software from Nankai University, China, the M.S. degree in signal processing from Nanjing University of Science and Technology, China, and the Ph.D. degree in image processing from the University of New England, Australia.

She is currently an Assistant Professor with the Faculty of Science and Technology, University of Macau. Her research interests include signal processing, computer vision, and multimedia computing.

Fei Richard Yu (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of British Columbia (UBC) in 2003. His research interests include connected/autonomous vehicles, artificial intelligence, blockchain, and wireless systems. He has been named in the Clarivate's list of "Highly Cited Researchers" in computer science since 2019, Standford's Top 2% Most Highly Cited Scientist since 2020. He received several Best Paper Awards from some first-tier conferences, Carleton Research Achievement Awards in 2012 and 2021, and the Ontario Early Researcher Award (formerly Premiers Research Excellence Award) in 2011. He is a Board Member the IEEE VTS and the Editor-in-Chief for IEEE VTS Mobile World newsletter. He is a Member of the Academia Europaea (MAE), Canadian Academy of Engineering (CAE), Engineering Institute of Canada (EIC), and IET. He is a Distinguished Lecturer of IEEE in both VTS and ComSoc.