applied sciences

MDPI

Article

On-Road Driver Emotion Recognition Using Facial Expression

Huafei Xiao, Wenbo Li, Guanzhong Zeng, Yingzhang Wu, Jiyong Xue, Juncheng Zhang and Chengmou Li and Gang Guo *

College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China; xiaohuafei@cqu.edu.cn (H.X.); liwenbo@cqu.edu.cn (W.L.); guanzhong@cqu.edu.cn (G.Z.); cquwyz@cqu.edu.cn (Y.W.); xuejiyong@cqu.edu.cn (J.X.); zhangjuncheng@cqu.edu.cn (J.Z.); chengmou_li@foxmail.com (C.L.)
* Correspondence: guogang@cqu.edu.cn

Abstract: With the development of intelligent automotive human-machine systems, driver emotion detection and recognition has become an emerging research topic. Facial expression-based emotion recognition approaches have achieved outstanding results on laboratory-controlled data. However, these studies cannot represent the environment of real driving situations. In order to address this, this paper proposes a facial expression-based on-road driver emotion recognition network called FERDERnet. This method divides the on-road driver facial expression recognition task into three modules: a face detection module that detects the driver's face, an augmentation-based resampling module that performs data augmentation and resampling, and an emotion recognition module that adopts a deep convolutional neural network pre-trained on FER and CK+ datasets and then fine-tuned as a backbone for driver emotion recognition. This method adopts five different backbone networks as well as an ensemble method. Furthermore, to evaluate the proposed method, this paper collected an on-road driver facial expression dataset, which contains various road scenarios and the corresponding driver's facial expression during the driving task. Experiments were performed on the on-road driver facial expression dataset that this paper collected. Based on efficiency and accuracy, the proposed FERDERnet with Xception backbone was effective in identifying on-road driver facial expressions and obtained superior performance compared to the baseline networks and some state-of-the-art networks.

Keywords: driving safety; facial expression recognition; human-machine interaction; smart cockpit; affective computing

Citation: Xiao, H.; Li, W.; Zeng, G.; Wu, Y.; Xue, J.; Zhang, J.; Li, C.; Guo, G. On-Road Driver Emotion Recognition Using Facial Expression. *Appl. Sci.* **2022**, *12*, 807. https://doi.org/10.3390/app12020807

Academic Editors: Martin Lauer, Angel Llamazares and Javier Alonso Ruiz

Received: 8 December 2021
Accepted: 11 January 2022
Published: 13 January 2022

1. Introduction

Emotion-related human-machine systems are essential for the intelligent automobile. Driver's emotion affects driving performance and is closely related to traffic accidents. The number of road traffic deaths continues to rise steadily, having reached 1.35 million [1]. Among these incidents, the inability to control emotions has been regarded as one of the critical factors degrading driving safety [2]. Hence, driver emotion detection and recognition are emerging topics for intelligent automotive human-machine systems [3].

Emotion can be divided into internal response, such as electroencephalograph (EEG) and galvanic skin response (GSR); and external response, such as facial expression, gesture, and speech [4]. EEG signals provides excellent time resolution and allow researchers to study emotional stimuli; however, EEG requires many electrodes placed at various places on the head, which is impractical in applications like automotive human-machine systems [5]. GSR monitors emotions and stress due to the change of sweat glands activities; compared with EEG and fMRI, GSR does not require bulky instruments and only needs sensors to be placed on the hands or feet. Nevertheless, even mild exercise can significantly alter the GSR signal and make it unreliable for driver emotion recognition, since drivers frequently move their hands and feet while controlling a vehicle [6,7]. External response requires simple instruments to collect. Gesture is a crucial body language that can deliver emotion states [8];

different body gestures convey various emotions. However, considering that the driving task restricts body movement, it unrealistic to monitor driver emotion through gesture. Speech, as a fundamental means of communication for humans, is also a vital component for affective interaction. Speech-based emotion recognition requires extracting features from raw speech data. It has low accuracy in recognition of highly affective speech [9,10] Considering the implementation of an emotion-based vehicle's human-machine system, speech is not qualified to be the primary pattern due to its insufficient accuracy. Speech is not continuous during the driving task, which means that the system is unable to monitor emotional states when there is no dialogue, which is common during driving.

Among the above-mentioned emotion responses, facial expression is one of the most powerful signals for human beings to convey emotional states [11]. Besides, facial expression is easy to obtain and requires only simple instruments, and many researches have studied facial expression recognition and achieved satisfying accuracy. In addition, the collection of driver facial expression data during driving is less affected by body movement and noise than the EEG or fMRI signals. Hence, facial expression-based emotion recognition is the most appropriate and suitable emotional response recognition for an automotive emotional human-machine system.

Computer vision-based deep learning methods are extensively applied for facial expression recognition and emotion monitoring. Li's [12] research survey of deep facial recognition summarizes facial expression datasets and collection environments such as laboratories or the internet.

However, the existing works are mostly performed on lab-captured datasets due to the shortage of real-scenario databases. There is no on-road driver facial dataset available for the driver facial recognition task, and the driving task may suppress facial expressions. Due to this problem, there is a lack of on-road driver facial expression recognition research, which is vital for automotive human-machine systems.

This paper proposes a novel deep learning-based framework for on-road driver facial expression recognition in an end-to-end manner. To address the above-mentioned dataset limitation, this study collected a driver facial expression dataset. The proposed method identifies drivers' emotions and can further improve driving safety. There are many studies related to the impact of emotion on driving behavior. Anger causes road rage and increases driving risk [13,14], while sadness and nervousness reduce driving concentration [15]. Related researches on driving risk [16,17] also shows that emotion is one of the factors that affect driving risk. Emotions affect the driving behavior and some negative emotions tend to produce dangerous driving behaviors (such as road rage). Therefore, identifying driver emotions is an upstream study for dangerous driving behavior early warning. Based on our research, researchers can further analyze the influence of each emotion on driving behavior and early warning intervention methods. The proposed framework's overall architecture is presented in Figure 1 and further elaborated on in Section 3.

The main contributions of this paper can be described as the following:

- A transfer learning model for on-road driver facial expression recognition, called the facial expression-based on-road driver emotion recognition network (FERDERnet), to classify on-road driver emotion, is proposed. This approach provides a novel method for on-road driver facial expression recognition using insufficient and unbalanced on-road data.
- An on-road driver facial expression dataset was collected. This study designed and conducted the on-road driving experiment to obtain on-road driver facial expression data. The experiment contains various road scenes (traffic lights, pedestrian crossings, urban areas, highways, tunnels, overpasses, bridges, etc.) and road conditions (smooth, traffic jam, congestion, etc.) during various periods (morning, midday, afternoon, night).
- The performance of the proposed FERDERnet was evaluated on the on-road driver facial expression dataset. A comprehensive comparative study of some baseline networks and the corresponding FERDERnet was conducted, and the FERDERnet was

compared with some state-of-the-art deep neural networks. The result demonstrates that the proposed framework improves the recognition accuracy of the on-road driver facial expression dataset.

The remainder of this paper is organized as follows: Section 2 summarizes related preliminary and facial expression emotion recognition works. Section 3 describes the framework proposed called FERDERnet and its components. Section 4 introduces the on-road driver facial expression data collection in detail. Experiment results and discussion are presented in Section 5. Method limitations and future works are presented in Section 6. The conclusion is presented in Section 7.

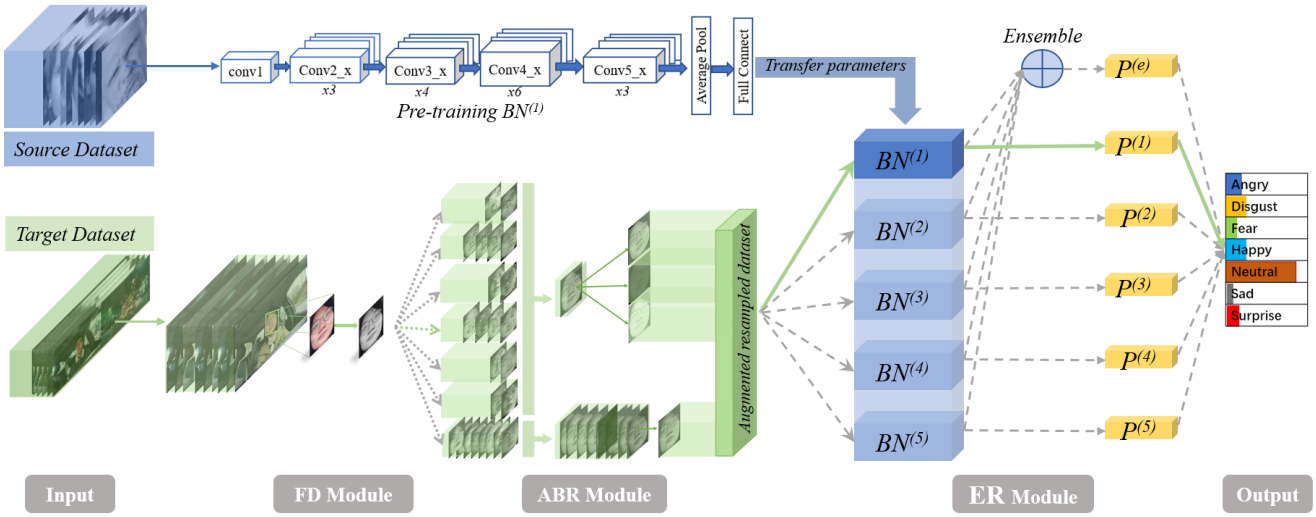


Figure 1. The overall structure of the proposed facial expression-based on-road driver emotion recognition network (FERDERnet). FERDERnet takes two inputs: source dataset (in this study: FER and CK+) and target dataset (in this study: on-road driver facial expression dataset). The flows of the source dataset are indicated by the blue arrows, while the flow of the target dataset is indicated by the green arrows. BN(i) stands for backbone networks; P(i) stands for model predictions.

2. Related Work

2.1. Facial Expression Recognition

There is extensive research related to facial expression recognition based on feature extraction methods, either handcrafted conventional feature extraction or feature extraction by deep neural networks.

The handcrafted features include Gobar wavelets [18], line-based caricatures [19], scale-invariant feature transform [20], histograms of oriented gradients (HOG) [21], local binary patterns (LBP) [22], minutiae points [23], Haar wavelet [24], and histograms of bunched intensity values (HBIV) [25]. The features using machine learning methods include dynamic-Bayesian networks (DBN) [26] and SVM [27–31] for further classification.

More recently, feature extraction and recognition jointly learned by deep learning techniques [32–36] are witnessed to be superior to handcrafted features-based approaches. Zhiding et al. [33] proposed a two-step model: a face detection module and a multiple CNN classification module for facial expression recognition tasks. Karnati et al. [36] proposed a model based on deep convolution neural networks for seven-class facial expression recognition tasks; the model adopted score-level fusion with two branches: a local feature classification branch and a holistic feature classification branch. Pham et al. [37] proposed a masking model to boost the performance of CNNs. Shervin et al. [38] proposed an attentional convolutional network for facial expression recognition. Jiawei et al. [39] proposed an amending representation module (ARM) to be embedded in CNNs to improve network performance.

2.2. Transfer Learning-Based Facial Expression Recognition

Transfer learning is a research problem proposed in machine learning in the 1970s. Extensive research about "learning how to learn" [40], "lifelong learning" [41], "multitask learning" [42], etc. has been conducted. Transfer learning aims to transfer the knowledge learned from one domain to help learning tasks in a new environment [43].

Fine tuning is a commonly used method of transfer learning with the intent of training a deep network with insufficient data. The strategy is usually to first train the network on a large-scale dataset (such as Imagenet, which is commonly used in many studies) and then transfer the network's parameters to training for the target task.

With the lack of large-scale datasets, fine-tuning has been widely investigated for facial expression recognition [44,45]. Networks have been pre-trained on the ImageNet dataset and fine-tuned on other facial expression datasets. Orozco et al. [46] adopted AlexNet, VGG19, and Resnet pre-trained on Imagenet and fine-tuned for recognition tasks on the CK and JAFFE datasets. A. Ravi [47] applied Imagenet pre-trained VGG to extract features on CK+ and JAFFE, then adopted SVM for classification. In [44], AlexNet and VGG were pretrained on ImageNet and twice fine-tuned for small dataset facial expression recognition. Yoursif et al. [48] adopted fine tuning in VGGNet architecture for facial expression recognition.

2.3. Driver Facial Expression Recognition

Driving a vehicle is a complex process involving visual cues, hazard assessment, decision-making, and strategic planning [49] for both the driver and the automotive [50]. Driving style and driving behavior have attracted research interest for driving safety [51,52] and avoiding collision [53]. Wenbo et al. [54] investigated the driver anger regulation in visual attributes. Driver facial expression mirrors driver emotional state [55]; therefore, it is important for emotion recognition and ensuring driving safety.

In [56], a driver facial expression recognition-based emotional stress system was proposed. The data was collected in a static vehicle scenario, then feature extraction applied PCA and implemented a SVM classifier.

Due to the influence of driving tasks, the driver's facial expression may be suppressed or subtle when experiencing emotional states [57]. In that case, driver facial expression recognition is vital for intelligent vehicle human-machine systems.

Vehicle driver facial expression recognition is more difficult compared with a lab-controlled environment. Nonetheless, most of the research mentioned above relating to driver facial expression recognition did not consider the issue of real on-road driver facial expression recognition, which lacks datasets for model training. Furthermore, these methods, regardless of the scenario, mean the datasets are static life scenarios or wild settings. The studies related to driver facial expression recognition also did not collect on-road driving data and did not adopt a transfer learning strategy, which is frequently adopted in ordinary facial recognition tasks.

Emotion perception-based human-computer interaction in a smart cockpit is an important topic. More and more applications of deep learning have been utilized in various fields, yet driver facial expression recognition during the driving task has never been studied. Facial expression recognition during the driving scenario is far more important than the static scenario for emotion recognition in the intelligent vehicle human-machine system. This paper proposes a novel deep learning-based framework for on-road driver facial expression recognition in an end-to-end manner. To address the above-mentioned dataset limitation, this study conducted and collected a driver facial expression dataset. Based on this research, more studies can be migrated to the field of autonomous driving and smart cockpits, such as drone-assisted crowd counting [58], face detection under risk situations [59], and image super-resolution [60].

In this work, a model for on-road driver facial expression recognition based on transfer learning is proposed. To the best of our knowledge, this is a novel work about on-road driver facial expression recognition. This work integrates some excellent technology in computer vision and deep learning with careful design and modification to construct the

proposed FERDERnet and apply it. Furthermore, this research also collected a driver facial expression dataset during on-road driving tasks.

3. Materials and Methods

3.1. Overall Structure

The FERDERnet model proposed in this research is a model for on-road driver facial expression recognition. The entire network consists of three stages. The first stage extracts the faces from the input video frame recorded during the on-road driving process; the second stage employs the image re-sampling algorithm based on the grayscale dataset extracted from the first stage to handle the long-tailed (sample imbalance) issue; the third stage performs emotion recognition on the re-sampling dataset by applying some state-of-the-art deep neural networks for backbone and implements a transfer learning training strategy.

The input of the FERDERnet is the pre-processed video frame (Section 4) and the model outputs the predicted emotion class. In general, the model comprises three modules corresponding to the three stages of the entire network—the face detection module (FD), the augmentation-based re-sampling module (ABR), and the emotion recognition module (ER)—as presented in Figure 1.

3.2. Face Detection Module (FD)

The face detection module (FD) extracts the driver's face area from the input video frames and converts it to grayscale images. Face detection and alignment are essential to many applications in computer vision and many researches have proposed relevant algorithms. Henry et al. [61,62] proposed an algorithm based on template matching and S.Z. Li et al. [63] proposed a Harr feature extraction and Adaboost classifier algorithm. With the development of deep learning, more and more novel algorithms have been proposed [64–67], continuing improvement of face detection accuracy.

The FD module proposed in this work is inspired by the deep cascaded multitask framework proposed by Kaipeng Zhan et al. [67]. The FD module adopts three-stage nets to generate the face window and alignment face landmark positions and extract the face pictures (resolution: 160 × 160, RGB). Then, the extracted face pictures are converted into grayscale images as the output of the FD module. The resolution of the input video influences the performance of the FD module. For the on-road driver facial expression dataset collected in this work, the original dataset was edited and the driver scene (resolution: 1920 × 1080) was taken as the input of FD module to boost the processing speed. The proposed method of this paper is aimed at the driver monitoring system (DMS) [68]. Therefore, the algorithm of the FD module performs largest face detection; when multiple faces appear in one image, only the largest face is considered to be the driver's face. Furthermore, the recorder's placement ensures that the driver's face is largest in the captured images.

3.3. Augmentation-Based Re-Sampling Module (ABR)

The primary intent of this work is emotion recognition using the data collected in the on-road driving experiment. In this study, labeled data with seven emotion class labels from the FD module are the original images. The original images of the emotion classes are imbalanced, which means that the image quantity varies from different classes. Therefore, the originally extracted faces required further processing. To settle this problem, related research on imbalanced data [69–71] has proposed methods that deal with long-tailed data, for instance: re-weighting, re-sampling, etc. As shown in Algorithm 1, this work proposed an augmentation-based re-sampling algorithm based on the re-sampling method to alleviate the imbalance of the dataset presented and enhance the model's generalization ability.

In detail, ABR implements different augmentation and sampling methods for the images of each class, including random augmentation, over-sampling, and random under-sampling. In general, the ABR module receives the input of the original driver face grayscale dataset, and outputs the augmented re-sampled dataset.

Algorithm 1 Augmentation-based re-sampling algorithm

Input: images from dataset
Output: augmented and re-sampled dataset
dataset contains cropped faces from FD module
for *directory* in *dataset* **do**
  if *directory* $=$ *Neutral* **then**
    perform undersampling
    sample rate $\leftarrow 0.1$
    random sample(*sample rate*)
  else if *directory* $=$ *Disgust* or *Happy* **then**
    perform augmentation-based oversampling
    sample rate $\leftarrow 0.2$
    random sample(sample rate)
    for *sample* in *random sample* **do**
      one of:
      Random Rotate($-20^\circ \leq angle \leq 20^\circ$)
      Random Bright(*brightness limit* $= 0.3$)
    end for
  else
    perform augmentation-based oversampling
    sample rate $\leftarrow 0.3$
    s1 $\leftarrow$ *random sample*(*sample rate*)
    s2 $\leftarrow$ *random sample*(*sample rate*)
    s3 $\leftarrow$ *random sample*(*sample rate*)
    for *sample* in *s*1 **do**
      Random Rotate($-20^\circ \leq angle \leq 20^\circ$)
    end for
    for *sample* in *s*2 **do**
      Random Bright(*brightness limit* $= 0.3$)
    end for
    for *sample* in *s*3 **do**
      Random Contrast(*contrastlimit* $= 0.1$)
    end for
  end if
end for
return augmented and re-sampled dataset

3.4. Emotion Recognition Module (ER)

The emotion recognition module (ER) in the proposed FERDERnet performs emotion recognition of the augmented re-sampled dataset produced by the ABR module. The ER module utilizes deep neural networks as the backbone, replacing the full connect layer with the seven class emotion recognition task of this work. In this paper, five widely used deep neural networks are selected as the backbones: Googlenet [72], Resnet50 [73], InceptionV3 [74], InceptionV4 [75], and Xception [76]. The ER module's training strategy in this study utilizes the inductive transfer learning method (the target data's labels are available) and adopts the fine-tuning method to transfer the knowledge learned from the source dataset, thereby enhancing module performance on the target dataset. Furthermore, this study adopt the no-weighted sum average ensemble method to the fuse five backbone networks together in order to boost model performance.

In this research, the ER module adopts fine tuning as the transfer learning strategy. The fine tuning method first trains the network on the source dataset and then uses the model weight trained on the source dataset, removes the full connect layer, and constructs a new full connect layer for the target domain training. Hence, knowledge is transferred from the source domain to the target domain by inheriting the model weight. In this work,

the source dataset is the augmented FER [77] and CK+ [78] dataset. The FER and Ck+ were integrated as a whole dataset and data augmentation was conducted (random horizontal flip, random crop, random brightness, random contrast, and random rotation) to increase the quantity of the source dataset. The target dataset is the augmented re-sampled on-road driver facial expression dataset. To perform fine tuning, the model was first trained on the source dataset and then the model weight was transferred to the target dataset training.

As shown in the pipeline of the FERDERnet (Figure 1), the ER module first initializes the deep neural network (in this Figure, Resnet50 is adopted as the backbone) using white cubes to present; then, it pre-trains the initialized network using the source dataset; after that, it fine-tunes the pre-trained network on the target dataset using blue cubes to present; finally, it trains the network using target dataset.

This work adopts cross-entropy loss:

$$Loss = -\sum_{i=1}^{n} y_i \log \widehat{y}_i \tag{1}$$

where n is the emotion class (in this study: seven); $\widehat{y}_i$ represents the probability distribution of the predicted value, and y_i represents the one-hot distribution of the emotion label of one picture:

$$y_i = \begin{cases} 1, & if\ y = G \\ 0, & otherwise \end{cases} \tag{2}$$

where y is the ground-truth of one picture and G represents the Gth emotion class. During the training process, this study adopts the batch training strategy, with gradient update after the iteration of one batch. Therefore, let b represents the batch size and the loss function is:

$$Loss(b) = -\frac{1}{b}\sum_{j=0}^{b}\sum_{i=0}^{n} y_{ij} \log \widehat{y_{ij}} \tag{3}$$

4. Data Collection

To validate the effectiveness of the proposed FERDERnet, an on-road driver facial expression dataset is required. To the best of our knowledge, there is no publicly available on-road driver facial expression dataset. To this end, this study designed and carried out an on-road driver facial expression experiment, which collected driver facial expressions induced by different road scenarios. Compared with other dataset collection methods like lab control environments or static life scenarios that perform facial display, the on-road driving experiment was much more complicated and difficult due to the uncertainty of real road scenarios and the labor-intensive labeling work. Under such circumstances, this study collected 25 subjects' on-road driving facial expression data.

4.1. Ethics Statement

The experimental procedure was approved by Chongqing University Cancer Hospital Ethics Committee, China. Participants and data from participants were treated according to the Declaration of Helsinki. The participants were also informed that they had the right to quit the experiment at any time. The video recordings of the participants were included in the dataset only after they gave written consent for the use of their videos for research purpose. A few participants also agreed to the use of their face images in research articles.

4.2. Participants

Twenty-five participants (20 males and 5 females) with Chinese nationality and from 21 to 31 years old (mean [M] = 24.4 years; standard deviation [SD] = 2.2 years) were recruited to participate in this experiment from Shapingba District, Chongqing, China. Each participant had a valid driving license with at least one year of driving experience (average [M] = 4.1 years; standard deviation [SD] = 2.3 years; range = 1–11 years). All the participants had normal or corrected-to-normal vision (18 participants wore glasses) and

average hearing ability. The presence of occlusions such as lighting conditions and glasses is a significant research challenge for facial expression recognition; hence, experiments during night time and participants wearing glasses were all included to evaluate the robustness of the emotion recognition. Domestic self-driving travel insurance for all participants was purchased. According to the duration of each participant's experiment, all the participants received 60 CNY as financial reimbursement for their participation.

4.3. Experiment Setup

The experiment was carried out in Chongqing, China. Benefiting from Chongqing's unique landform, the route selected included abundant terrain and road scenarios (signal lights at intersections, zebra crossings, pedestrian-intensive sections, downtown sections, highways, tunnels, overpasses, bridges, etc.). Figure 2 shows the experimental setup. The experiment was conducted during various periods (morning, midday, afternoon, night). The experiment route involved four districts of Chongqing and the route varied for different participants.

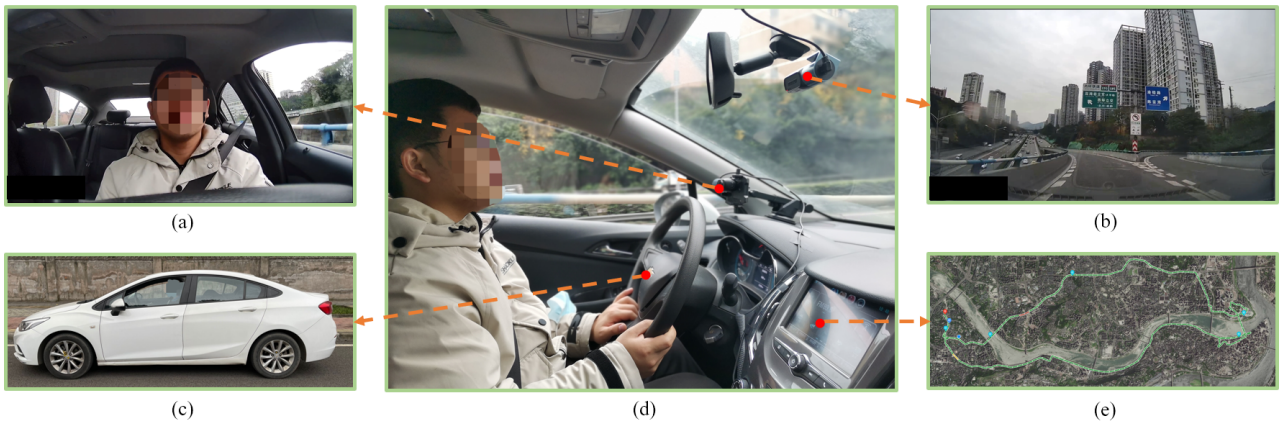


Figure 2. Experimental setup of on-road driver facial expression dataset collection: (**a**) driver facial expression recording, (**b**) road scenario recording, (**c**) experimental vehicle, (**d**) experiment setup, (**e**) navigation of the test road.

The experiment's intent was to collect on-road driver facial expressions induced by the road scenario; hence both the driver's face and the road scenario needed to be collected. To record the driver's face as well as the road scenario synchronously, two driving recorders were required. Based on the vehicle condition of this study, the driving recorder needed to meet the following requirements:

- Mainstream driving recorder power supplies require 5 volts DC. The on-board cigarette lighter of the test vehicle meets the requirement, meaning that the driving recorder needs to support an on-board cigarette lighter port.
- The video recorded by the recorder must be unencrypted (videos recorded by some driving recorders are encrypted) so that the original recorded driving data can be exported for further processing.
- The recorder needs to be small and easy to install to reduce interference with driving.

Based on the above factors, this study selected Hikvision's D1 Driving Recorder, one for the road scenario recording (Resolution: 1920 × 1080, TS) and the other for driver facial expression recording (Resolution: 1920 × 1080, TS), with a frame rate of 30 frames per second (fps). This experiment used a TF memory card to store and read the experimental data collected in the recorder.

In order to reduce interference and maintain the real driving environment, the recorder equipment was installed slightly higher than the steering wheel. Each driver was also required to adjust the driving seat. Furthermore, the cab's interior ceiling lights (except for the dashboard light) and the sunroof were closed to reduce reflection.

4.4. Experiment Protocol

At the beginning of the experiment, the participants were informed about the driving route before driving the vehicle. During the driving process, two driving recorders recorded each scene simultaneously. For each driving experiment participant, an experimenter accompanied them in the vehicle (for safety reasons).

During the driving process, emotions can be induced by road scenarios, communication with passengers, and other none-driving-related tasks such as music and phones. However, non-driving-related tasks such as communication, occur not only in the vehicle, in but many situations. This research aimed at the special situation that can induce emotion only when driving a vehicle, which is the diverse road scenario. In this case, the driver was not allowed to communicate with the co-pilot except for safety reasons, to ensure that the collected driver emotions were induced by the road scenarios.

The recorded driving time for participants was around 90 min. After the driving process, the experimenter exported the collected video to the computer through the TF memory card, then performed data pre-processing and labeling.

4.5. Data Pre-Processing and Labeling

Data pre-processing included video format conversion, merging, and alignment of the original recorded facial video and the scenario video. The, each participant's video clips were spliced and edited into a 3 s short video clip sequence for the labeling work.

The initially recorded video clips for each participant include facial records and scenario records, both containing a quantity of one-minute video clips in TS format. After removing those video clips that involve communication between the driver and the experimenter, the short video clips were combined into a long video clip and converted to MP4 format. After that, we obtained two time-aligned videos of equal length (a long video clip of the driver's face and a long video clip of the scenario; resolution: 1920 $\times$ 1080, MP4); Then, the two long video clips were spliced into one clip with the layout of the face on the left and scenario on the right (resolution: 3840 $\times$ 1080, MP4). After that, this long clip was edited into a series of 3 s short video clips for the participant to label.

To obtain the emotion label of the pre-processed dataset, this work adopted a manual labeling method. The annotation tool is used for manual labeling in many researches [79,80]. This study developed an annotation software called the "Driver Emotion Label Tool" for the labeling process; this tool ensures the annotation quality and the reliability of the collected dataset. Figure 3 demonstrates the annotation tool's GUI.

In this study, the labeling work adopted a discrete emotion method (emotion categories), with each participant required to label their emotion in each 3 s short video clip edited from the pre-processing procedure (0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral). To eliminate the individual differences among different facial expressions of participants, each driver was employed to label their own clips. The road scenario also helped the drivers in labeling the dataset. Because the facial expressions may be subtle in some situations, the road scenarios enhanced the driver's judgment of his/her current emotional state. An experimenter accompanied the driver to assist in labeling and avoid miss-labeling.

After that, video frames were extracted from each participant's labeled short video clip series. With all participants' video extracted, the entire on-road driver facial expression dataset was obtained (effective images: 69,923; resolution: 3840 $\times$ 1080, RGB). The dataset contains various road scenarios during day and night, and the corresponding driver facial expression. Figure 4 demonstrates part of the dataset.

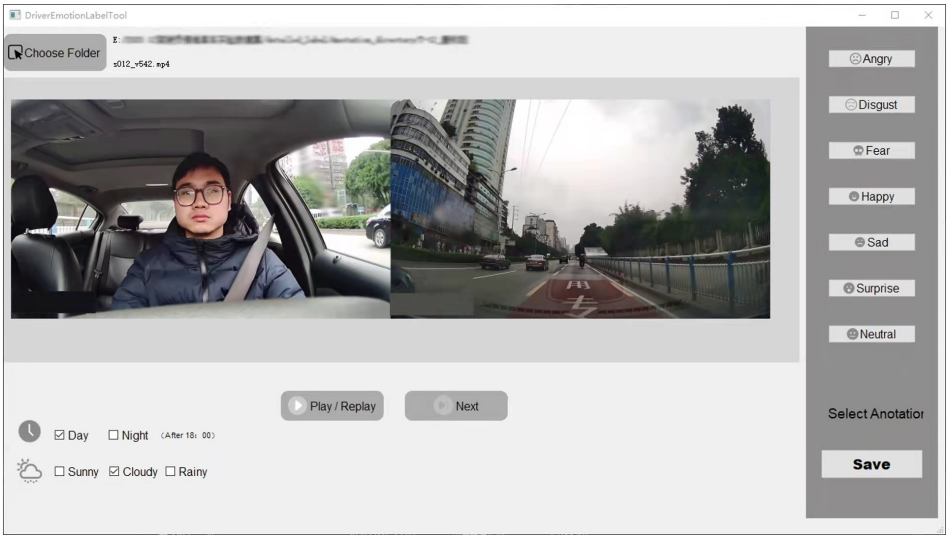


Figure 3. Screenshot of the developed annotation tool called "DriverEmotionLabelTool", used to annotate the video sequences of the on-road driver facial expression dataset.

Figure 4. Collected on-road driver facial expression dataset. Each clip contains the driver facial expression and road scenario synchronously; the driver facial expression is induced by various scenarios during day (the first three columns) and night (the fourth column).

5. Results and Discussion

5.1. Training Details

The datasets adopted for the source domain training were FER and CK+, and dataset adopted for the target domain training was the on-road driver facial expression dataset collected in this study. FERDERnet adopts the data augmentation re-sampling module (ABR) for the target original long-tailed dataset to alleviate data imbalance. Hence, the original long-tailed dataset was input into the face detection module (FD) of FERDERnet to generate grayscale faces as shown in Figure 5. It was then input into the ABR module, which outputs the augmented re-sampled dataset (effective images: 15,523; resolution: 160 $\times$ 160, GRAY).

The augmented and re-sampled dataset produced by the ABR module containing 15,523 images was fed into the ER module for model training. The target dataset was divided into a training set with 12,418 images and a test set with 3105 images. The performance in the target dataset was the result of performing stratified cross-validation, which has better performance in small, imbalanced datasets.

Figure 5. Driver facial expressions in gray images processed by the FD module.

This study employs the Adam optimizer with a learning rate of 10^{-4}. To confirm the training optimizer for the proposed FERDERnet, an experiment with different training algorithms was conducted. Five different optimizers were chosen, namely Adam, SGD, RMSprop, Adagrad, and Adamax. FERDERnet_R (FERDERnet with resnet50 backbone) was trained using the above five optimizers. As can be seen from Table 1 and Figure 6, the Adam optimizer obtained the best accuracy as well as the best F1-score. As for the training time, each optimizer required a similar time. Therefore, the Adam optimizer was confirmed to be the FERDERnet optimizer. With the Adam optimizer and cross-entropy loss, training FERDERnet executed 150 epochs for the source dataset and 50 epochs for the target dataset. For backbone details, in each backbone this research adopted, the batch normal layer was placed between convolution layer and ReLU layer, and dropout was only adopted for the full connect layer, with the dropout rate set to 0.5 for all five backbone networks. The model implementation was done using Pytorch. The model was trained and tested on a server with an Intel Xeon E5-2678 v3@2.5GHz CPU and a NVIDIA GeForce RTX 2080Ti GPU.

Table 1. Comparison of different optimization algorithms for FERDERnet.

Algorithm	Precision	Recall	Specificity	F1-Score	ACC	Training Time
Adam	0.941	0.96	0.991	0.951	0.953	2303.3 s
SGD	0.471	0.469	0.917	0.4656	0.575	2284.8 s
RMSprop	0.941	0.949	0.991	0.945	0.952	2297.8 s
Adagrad	0.706	0.685	0.951	0.695	0.748	2296.2 s
Adamax	0.944	0.943	0.988	0.943	0.948	2320.8 s

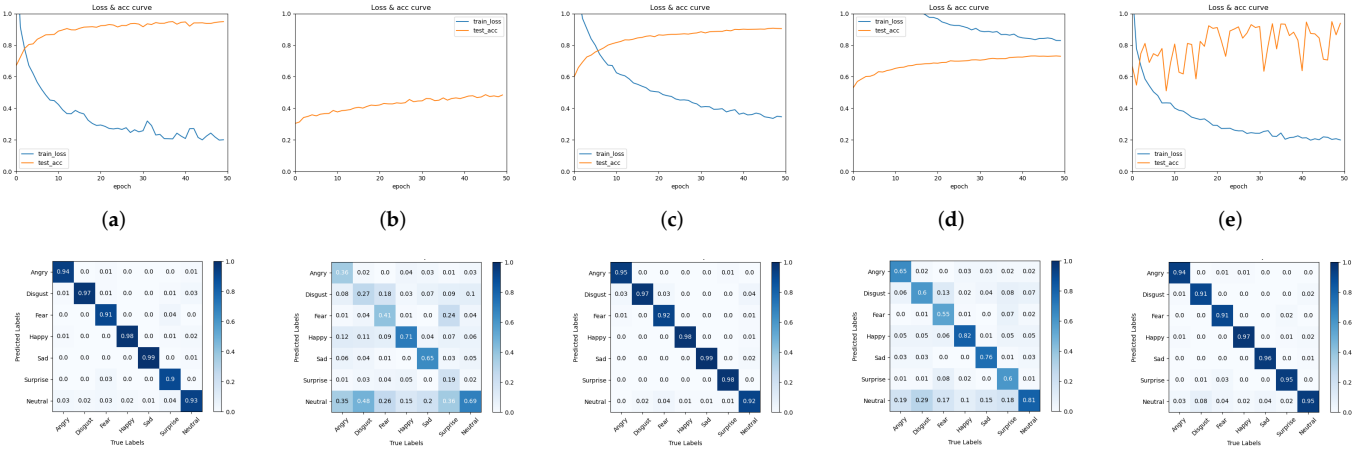
5.2. The Baseline Methods

Several of the most common networks such as Googlenet, Resnet50, InceptionV3, InceptionV4, and Xception were employed as baseline methods in this study. For these methods, the training strategy was the same as in FERDERnet's target domain.

5.3. Performances

The dataset this study adopted to validate the performance for all baseline methods and FERDERnet with different backbones was the on-road driver facial expression dataset.

To evaluate the performance of the network, the results obtained were reported using accuracy, precision, recall, and F1-score.



(**a**) (**b**) (**c**) (**d**) (**e**)

(**f**) (**g**) (**h**) (**i**) (**j**)

Figure 6. Results of different training algorithms. The above charts present training loss (blue line) and test accuracy (orange line); the bottom charts present the corresponding confusion matrix. (**a**) ACC and Loss Adam, (**b**) ACC and Loss SGD, (**c**) ACC and Loss RMSprop, (**d**) ACC and Loss Adagrad, (**e**) ACC and Loss Adamax, (**f**) matrix Adam, (**g**) matrix SGD, (**h**) matrix RMSprop, (**i**) matrix Adagrad, (**j**) matrix Adamax.

Table 2 demonstrates the performance between each baseline network and the corresponding FERDERnet that applies the same network for its backbone. Among the five backbones employed in this experiment, FERDERnet_G (FERDERnet with Googlenet backbone) reached a classification top 1 accuracy of 88.8%, higher than Googlenet (top 1 accuracy: 80.3%) by 8.5%. Furthermore, as can be seen, the F1-score of FERDERnet_G is also significantly higher than the Googlenet, by 9.9%. For each backbone adopted, the table reveals that the proposed FERDERnet achieved remarkably high emotion classification accuracy, as each FERDERnet exceeded its baseline for the top 1 accuracy by 3.3% to 8.3%.

It is clear that the samples in the collected on-road driver facial expression dataset are all of the same nationality so that all methods meet a reasonably high classification performance. Despite that, the FERDERnet still surpasses the baseline networks.

Another comparative study can be seen in Table 2. FERDERnet with different backbone networks achieves a diverse top 1 accuracy range from 88.8% to 96.6%. The FERDERnet_X (FERDERnet with Xception backbone) achieved the best classification top 1 accuracy of 96.6%. For further discussion, the F1-score plays an important role when evaluating multi-class classification tasks on imbalanced data. Hence, the performance of the methods proposed was also assessed on F1-score. The Table shows that FERDERnet_X achieved the highest F1-score of 0.962. The results indicate that FERDERnet_X performs best among the backbones applied.

Table 2. Model performance of on-road driver facial expression dataset.

Models	Precision	Recall	F1-Score	ACC
FERDERnet_G VS. baseline				
FERDERnet_G	0.881	0.879	0.88	0.888
Googlenet	0.793	0.78	0.781	0.803
FERDERnet_R VS. baseline				
FERDERnet_R	0.939	0.956	0.947	0.956
Resnet50	0.898	0.876	0.886	0.907
FERDERnet_I3 VS. baseline				
FERDERnet_I3	0.964	0.952	0.957	0.959
InceptionV3	0.81	0.913	0.884	0.897
FERDERnet_I4 VS. baseline				
FERDERnet_I4	0.956	0.959	0.958	0.961
InceptionV4	0.898	0.93	0.91	0.928
FERDERnet_X VS. baseline				
FERDERnet_X	0.952	0.972	0.962	0.966
Xception	0.926	0.939	0.932	0.93

Note: G represents Googlenet backbone; R represents Resnet50 backbone; I3 represents InceptionV3 backbone; I4 represents InceptionV4 backbone; X represents Xception backbone.

As shown in Figure 7, the confusion matrix of FERDERnet_X performed significantly well in seven-class emotion classification. The model performed pretty well in classes such as Angry, Happy, and Sad, while performing slightly worse in Fear and Surprise. This may be caused by class imbalance because the samples are still not fully balanced after the ABR module (Fear contained 464 images, Surprise contained 642 images, Neutral contained 6265 images). Overall, the proposed FERDERnet with Xception backbone obtained excellent classification accuracy for the on-road driver facial expression recognition task.

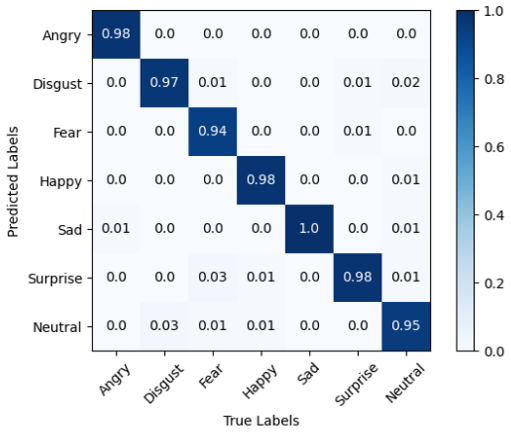


Figure 7. The confusion matrix of FERDERnet_X when trained on the on-road driver facial expression dataset.

The proposed network FERDERnet_X performed best among the backbones applied. To further evaluate the method compared to other state-of-the-art networks in the facial expression recognition task, some robust networks—DeepEmotion [38], ARM [39], VGGNet [48], ResMaskingNet [37], Resnet [73], Inception [74]—and FERDERnet with an ensemble of five backbone networks were trained on our on-road driver facial expression dataset under the same training strategy. Table 3 shows the comparison between the proposed method and other facial expression recognition SOTA work.

As shown in Table 3, the ensemble method reaches the best top 1 accuracy and F1-score, outperforming other networks. The training time reflects the network parameters and depth. It can be observed that DeepEmotion has the most straightforward network; hence it finished training in the shortest time, but obtained a much lower accuracy. The ensemble of five backbone networks required much longer training time; therefore, it is not acceptable for engineering application. On the other hand, FERDERnet_X used a relatively short training time to reach high accuracy. Considering demands on time and computation resources, FERDERnet_X is most suitable for practical application.

Table 3. Comparison with the state-of-the-art methods on the on-road driver facial expression dataset.

Method	Dataset	ACC	F1-Score	Training Time
DeepEmotion	*ours*	0.667	0.553	856 s
ARM	*ours*	0.925	0.913	3143 s
VGGNet	*ours*	0.937	0.925	3982 s
ResMaskingNet	*ours*	0.94	0.937	5252 s
ResNet	*ours*	0.956	0.947	3741 s
Inception	*ours*	0.959	0.957	5937 s
FERDERnet_X	*ours*	0.966	0.962	2503 s
FERDERnet_E5	*ours*	0.972	0.967	23,686 s

Note: *ours* stands for on-road driver facial expression dataset. FERDERnet_X stands for FERDERnet with Xception backbone, FERDERnet_E5 stands for the ensemble method that fuses 5 backbone networks (Googlenet, Resnet50, InceptionV3, InceptionV4, Xception) together.

Light conditions are various in real driving situations, hence, the model's performance under difficult conditions is important. An experiment to compare FERDERnet performance under different light conditions was conducted. Daytime driver facial images and nighttime driver facial images were randomly selected from the original dataset. Figure 8 shows some of the test samples. The test data included 140 day images and 140 night images (20 images for each emotion category) to compare the model's performance. Furthermore, the influence of a different training set was considered; the FERDERnet_X (FERDERnet with Xception backbone) model was trained on different datasets (day-images-only on-road driver dataset vs. the full on-road driver dataset (containing both day and night data)) to compare recognition accuracy. As illustrated in Table 4, the night images lower the model's recognition accuracy. However, by involving nighttime data in the training process, our model gains significantly improved nighttime recognition performance. Furthermore, comparing recognition accuracy of the model trained on different datasets, the nighttime recognition accuracy was much higher when the model's training set involved night data.

Figure 8. Facial samples from the collected on-road driver facial expression dataset; the top row shows day samples and the bottom row shows night samples.

Table 4. Comparison of model performance under day and night images.

Method	Training Set	Test Set	ACC	F1-Score
FERDERnet_X	ours	day-images	0.966	0.962
FERDERnet_X	ours	night-images	0.812	0.795
FERDERnet_X(D/O)	ours(day-images)	day-images	0.958	0.951
FERDERnet_X(D/O)	ours(day-images)	night-images	0.624	0.554

Note: FERDERnet_X stands for FERDERnet with Xception backbone, FERDERnet_X(D/O) stands for FERDERnet_X trained under the day-images-only on-road driver facial expression dataset.

5.4. Ablative Analysis

The FERDERnet model this study proposes contains of three modules: the FD module, which performs the processing of face detection, crop, and image format transform; the ABR module, which performs facial image augmentation-based re-sampling; the ER module, which employs the deep network backbone and adopts the fine-tuning strategy to perform emotion classification.

Hence, to validate if and how much the ABR module impacts the emotion recognition task and compare it with the augmentations that torch transforms provide, ablative studies were conducted. The proposed FERDERnet was modified and evaluated by including or removing the ABR module in the architecture. The ablative analysis adopts precision, recall, and F1-score as metrics.

The ablative analysis result is presented in Table 5. As can be observed, the model suffers significant performance loss when the ABR module is removed; the F1-score shows a decline by 4.6%. Furthermore, the confusion matrix of FERDERnet_X without the ABR module is demonstrated in Figure 9, and the classification results significantly drop in classes like Disgust, Fear, and Sad compared with Figure 7, which also shows the importance of the ABR module. The probable reason is that the ABR module performs not only adjustment of imbalanced data, but also enhances the images' diversity of lightness, angle, and contrast, which improves the model's recognition performance of night images and images under hard conditions in the dataset. This demonstrates the importance of data re-balance and diversity that the ABR module performs in the on-road driver emotion classification task.

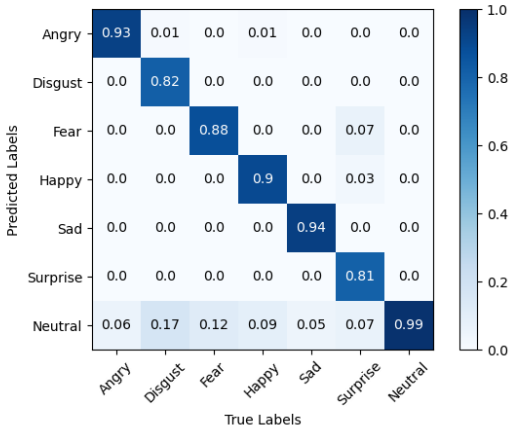


Figure 9. The confusion matrix of FERDERnet_X without ABR module when trained on the on-road driver facial expression dataset.

Table 5. Ablative evaluation with and without (w/o) the ABR module of FERDERnet.

Model	Dataset	Precision	Recall	F1-Score
FERDERnet	*ours*	0.952	0.972	0.962
FERDERnet(w/o)	*ours*	0.934	0.9	0.916

Note: *ours* stands for on-road driver facial expression dataset.

5.5. Discussion

The proposed FERDERnet applies transfer learning by fine-tune training the backbone network and combines face detection, crop, and image format transform image augmentation-based re-sampling to classify on-road driver emotions. The whole model is a novel method that achieves excellent recognition accuracy with insufficient and imbalanced data.

As shown in Table 2, the experimental result of the proposed model obtained high accuracy with all five backbones (FERDERnet with Googlenet backbone achieved 89.8%

top accuracy, which was the lowest among the five backbones). Apart from the model's effectiveness, the reason for the high top 1 accuracy of seven emotion classification is the lack of samples during data collection. Each subject did not necessarily show rich emotions in their facial expression; as a matter of fact, quite a few drivers rarely showed non-neutral emotions during the driving experiment. Due to this, the dataset did not contain a diversity of samples, which lowered the classification difficulty for the networks.

Despite this fact, the proposed model still outperformed the baseline models (note that the dataset collected in this study contains driver facial expressions in night-driving) for 3.4% to 8.4% top 1 accuracy. In addition, the expression of different emotions was less identical than the lab controlled facial display, because drivers are highly focused on the road in the reality of driving.

As the proposed network aims to apply a fast and accurate method for on-road driver facial expression recognition, recognition speed is essential. The best-trained model FERDERnet_X performed recognition at the speed of 12.8 FPS. However, image size significantly affected the system's speed, since the original image size is 1920×1080. In practical applications, lower resolution images can improve the recognition speed. In future practical applications, the network needs to recognize the driver's emotion in different situations, including nighttime. The ABR module can be modified to perform image augmentation, enhance night image brightness, or adjust contrast in order to improve the emotion recognition accuracy.

Under such circumstances, the result and ablative analysis demonstrate that the FERDERnet proposed in this study obtains significant improvements compare to the baseline networks. To the best of our knowledge, this is novel research in addressing on-road driver expression recognition, as this research also conducted experiments to collect on-road driver facial expressions induced by various scenarios.

6. Limitations and Future Works

As with any study, the present research has limitations. The limitations of this work can be summarized as follows:

- The manual individual labeling process was conducted after the on-road driving experiment; the time passed between the driving experiment and labeling may have caused some labeling accuracy problems.
- The verification on other datasets: the proposed method is a three-stage model designed specifically for driver facial expression recognition. The FD module detected the driver's face and the ABR module handled the data augmentation as well as re-sampling, followed by the ER module that predicted the emotion category. Verifying the proposed method on other datasets requires other publicly available on-road driver facial expression datasets. However, there are currently no publicly available on-road driver facial expression datasets; most facial expression datasets are lab controlled or from the internet. More verification needs to be done on available on-road driver facial expression datasets when future researches is published.

7. Conclusions

In this paper, a novel deep learning model, namely FERDERnet for on-road driver facial expression recognition that is robust against insufficient and imbalanced data, is proposed. The model is transfer learning-based, using fine tuning and employing augmentation-based re-sampling to enhance recognition performance.

To validate the effectiveness of the proposed model, extensive experiments were conducted against other state-of-the-art deep networks. The FERDERnet model delivers exceedingly high emotion classification accuracy improvements of around 3% to 8%. This is because of the fine-tuning strategy and the augmentation-based data re-sampling that assisted the FERDERnet to learn from insufficient and imbalanced on-road driving data. This can be observed from the ablative analysis, where removing the augmentation-based re-sampling module (ABR) caused the classification accuracy to decline significantly.

To the best of our knowledge, the proposed FERDERnet is a novel study aimed at recognizing on-road driver facial expressions through a transfer learning approach, which collected the on-road driving dataset for model training. Work is in progress to extend this model and improve the dataset in sample quantity and label quality in order to make the dataset publicly available for related research.

The collected dataset contains driver facial expressions in various road scenarios. The dataset contains 25 subjects, which is a relatively small number for the network training process. However, the dataset collection was also difficult. The on-road driving experiment was much more labor-intensive than static expression collection experiments due to both the uncertainty of the road scenario and the labeling work, which is reflected in the scarcity of on-road driver facial expression datasets. Future work should carefully design experiments and contain more subjects. Furthermore, there are researches aimed at deploying deep network models to some embedded devices, for instance NVIDIA Jetson devices [81,82], for practical applications, even though they may lower performance. However, due to a device shortage, this research was unable to test performance on the NVIDIA Jetson; a performance test on a common laptop (graphics card with 4G memory) was conducted and the trained model was deployed and running with 13.2 FPS when the resolution of the input video was 1280×720. Thus, more work to deploy this model on embedded devices is essential for practical applications.

Furthermore, the proposed model can be applied to the smart cockpit of intelligent automobiles for driver emotion recognition, in order to improve the human-machine system, reduce driving risk, and improve driving safety.

Author Contributions: Conceptualization, W.L.; methodology, H.X., G.Z., Y.W., J.Z., J.X., and C.L.; writing—original draft preparation, H.X.; writing—review and editing, H.X.; supervision, G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key Research and Development Program of China (Grant Number: 2017YFB1401702).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Chongqing University Cancer Hospital Ethics Committee, China (CZLS2020261-A, 24 December 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: Data sharing is not applicable to this article.

Acknowledgments: The authors would like to thank Lei Shi, Qiuyang Tang, Mengjin Zeng, Ying Lin, Renjie Ma, Dehua Xia, Yujing Liu and Lei Wu for their assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. World Health Organization. *Global Status Report on Road Safety 2018: Summary*; Technical Report; World Health Organization: Geneva, Switzerland, 2018.
2. Li, G.; Lai, W.; Sui, X.; Li, X.; Qu, X.; Zhang, T.; Li, Y. Influence of traffic congestion on driver behavior in post-congestion driving. *Accid. Anal. Prev.* **2020**, *141*, 105508. [CrossRef] [PubMed]
3. Braun, M.; Chadowitz, R.; Alt, F. User Experience of Driver State Visualizations: A Look at Demographics and Personalities. In Proceedings of the IFIP Conference on Human-Computer Interaction, Paphos, Cyprus, 2–6 September 2019; Springer: Cham, Switzerland, 2019; pp. 158–176.
4. Mühl, C.; Allison, B.; Nijholt, A.; Chanel, G. A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges. *Brain-Comput. Interfaces* **2014**, *1*, 66–84. [CrossRef]
5. Alarcao, S.M.; Fonseca, M.J. Emotions recognition using EEG signals: A survey. *IEEE Trans. Affect. Comput.* **2017**, *10*, 374–393. [CrossRef]

6. Nisa'Minhad, K.; Ali, S.H.M.; Khai, J.O.S.; Ahmad, S.A. Human emotion classifications for automotive driver using skin conductance response signal. In Proceedings of the 2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES), Putrajaya, Malaysia, 14–16 November 2016; pp. 371–375.
7. Eesee, A.K. The suitability of the Galvanic Skin Response (GSR) as a measure of emotions and the possibility of using the scapula as an alternative recording site of GSR. In Proceedings of the 2019 2nd International Conference on Electrical, Communication, Computer, Power and Control Engineering (ICECCPCE), Mosul, Iraq, 13–14 February 2019; pp. 80–84.
8. Shen, Z.; Cheng, J.; Hu, X.; Dong, Q. Emotion Recognition Based on Multi-View Body Gestures. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3317–3321.
9. Wu, C.H.; Liang, W.B. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Trans. Affect. Comput.* **2010**, *2*, 10–21.
10. Wang, K.; An, N.; Li, B.N.; Zhang, Y.; Li, L. Speech emotion recognition using Fourier parameters. *IEEE Trans. Affect. Comput.* **2015**, *6*, 69–75. [CrossRef]
11. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115. [CrossRef] [PubMed]
12. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
13. Jeon, M. Towards affect-integrated driving behaviour research. *Theor. Issues Ergon. Sci.* **2015**, *16*, 553–585. [CrossRef]
14. Wells-Parker, E.; Ceminsky, J.; Hallberg, V.; Snow, R.W.; Dunaway, G.; Guiling, S.; Williams, M.; Anderson, B. An exploratory study of the relationship between road rage and crash experience in a representative sample of US drivers. *Accid. Anal. Prev.* **2002**, *34*, 271–278. [CrossRef]
15. Lee, Y.C. Measuring drivers' frustration in a driving simulator. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; Sage Publications: Los Angeles, CA, USA, 2010; Volume 54, pp. 1531–1535.
16. Tao, D.; Zhang, R.; Qu, X. The role of personality traits and driving experience in self-reported risky driving behaviors and accident risk among Chinese drivers. *Accid. Anal. Prev.* **2017**, *99*, 228–235. [CrossRef]
17. Sun, C.; Li, B.; Li, Y.; Lu, Z. Driving risk classification methodology for intelligent drive in real traffic event. *Int. J. Pattern Recognit. Artif. Intell.* **2019**, *33*, 1950014. [CrossRef]
18. Liu, C.; Wechsler, H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Process.* **2002**, *11*, 467–476.
19. Gao, Y.; Leung, M.K.; Hui, S.C.; Tananda, M.W. Facial expression recognition from line-based caricatures. *IEEE Trans. Syst. Man -Cybern.-Part A Syst. Hum.* **2003**, *33*, 407–412.
20. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
21. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
22. Bhattacharjee, D.; Seal, A.; Ganguly, S.; Nasipuri, M.; Basu, D.K. A comparative study of human thermal face recognition based on Haar wavelet transform and local binary pattern. *Comput. Intell. Neurosci.* **2012**, *2012*, 6. [CrossRef]
23. Seal, A.; Ganguly, S.; Bhattacharjee, D.; Nasipuri, M.; Basu, D.K. Automated thermal face recognition based on minutiae extraction. *Int. J. Comput. Intell. Stud.* **2013**, *2*, 133–156. [CrossRef]
24. Seal, A.; Ganguly, S.; Bhattacharjee, D.; Nasipuri, M.; Basu, D.K. Thermal human face recognition based on haar wavelet transform and series matching technique. In *Multimedia Processing, Communication and Computing Applications*; Springer: New Delhi, 2013; pp. 155–167.
25. Seal, A.; Bhattacharjee, D.; Nasipuri, M.; Gonzalo-Martin, C.; Menasalvas, E. Histogram of bunched intensity values based thermal face recognition. In *Rough Sets and Intelligent Systems Paradigms*; Springer: Cham, Switzerland, 2014; pp. 367–374.
26. Ontañón, S.; Montaña, J.L.; Gonzalez, A.J. A Dynamic-Bayesian Network framework for modeling and evaluating learning from observation. *Expert Syst. Appl.* **2014**, *41*, 5212–5226. [CrossRef]
27. Trujillo, L.; Olague, G.; Hammoud, R.; Hernandez, B. Automatic feature localization in thermal images for facial expression recognition. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, San Diego, CA, USA, 21–23 September 2005; p. 14.
28. Littlewort, G.; Bartlett, M.S.; Fasel, I.; Susskind, J.; Movellan, J. Dynamics of facial expression extracted automatically from video. In Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004; p. 80.
29. Kotsia, I.; Pitas, I. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. Image Process.* **2006**, *16*, 172–187. [CrossRef]
30. Berretti, S.; Amor, B.B.; Daoudi, M.; Del Bimbo, A. 3D facial expression recognition using SIFT descriptors of automatically detected keypoints. *Vis. Comput.* **2011**, *27*, 1021–1036. [CrossRef]
31. Lemaire, P.; Ardabilian, M.; Chen, L.; Daoudi, M. Fully automatic 3D facial expression recognition using differential mean curvature maps and histograms of oriented gradients. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–7.
32. Liu, P.; Han, S.; Meng, Z.; Tong, Y. Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1805–1812.

33. Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 435–442.
34. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-piloted deep network for facial expression recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 425–442.
35. Villanueva, M.G.; Zavala, S.R. Deep neural network architecture: application for facial expression recognition. *IEEE Lat. Am. Trans.* **2020**, *18*, 1311–1319. [CrossRef]
36. Mohan, K.; Seal, A.; Krejcar, O.; Yazidi, A. Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5003512 . [CrossRef]
37. Luan, P.; Huynh, V.; Tuan Anh, T. Facial Expression Recognition using Residual Masking Network. In Proceedings of the IEEE 25th International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2020; pp. 4513–4519.
38. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef] [PubMed]
39. Shi, J.; Zhu, S. Learning to amend facial expression representation via de-albino and affinity. *arXiv* **2021**, arXiv:2103.10189.
40. Schmidhuber, J. *On Learning How to Learn Learning Strategies*; Technical Report FKI-198-94; Fakultat fur Informatik: Leuven, Belgium, 1995 .
41. Thrun, S. Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, 1996; pp. 640–646.
42. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
43. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
44. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 443–449.
45. Kaya, H.; Gürpınar, F.; Salah, A.A. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image Vis. Comput.* **2017**, *65*, 66–75. [CrossRef]
46. Orozco, D.; Lee, C.; Arabadzhi, Y.; Gupta, D. *Transfer Learning for Facial Expression Recognition*; Florida State Univ.: Tallahassee, FL, USA, 2018.
47. Ravi, A. Pre-Trained Convolutional Neural Network Features for Facial Expression Recognition. *arXiv* **2018**, arXiv:1812.06387.
48. Khaireddin, Y.; Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. *arXiv* **2021**, arXiv:2105.03588.
49. Li, G.; Wang, Y.; Zhu, F.; Sui, X.; Wang, N.; Qu, X.; Green, P. Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China. *J. Saf. Res.* **2019**, *71*, 219–229. [CrossRef] [PubMed]
50. Li, G.; Yang, Y.; Qu, X.; Cao, D.; Li, K. A deep learning based image enhancement approach for autonomous driving at night. *Knowl.-Based Syst.* **2020**, *213*, 106617. [CrossRef]
51. Li, G.; Li, S.E.; Cheng, B.; Green, P. Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. *Transp. Res. Part Emerg. Technol.* **2017**, *74*, 113–125. [CrossRef]
52. Li, G.; Chen, Y.; Cao, D.; Qu, X.; Cheng, B.; K Li. Automatic segmentation and understanding on driving behavioral signals using unsupervised Bayesian methods. *Mech. Syst. Signal Process.* **2021**, 156, 107589. [CrossRef]
53. Li, G.; Yang, Y.; Zhang, T.; Qu, X.; Cao, D.; Cheng, B.; Li, K. Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios. *Transp. Res. Part C Emerg. Technol.* **2021**, *122*, 102820. [CrossRef]
54. Li, W.; Zhang, B.; Wang, P.; Sun, C.; Zeng, G.; Tang, Q.; Guo, G.; Cao, D. Visual-Attribute-Based Emotion Regulation of Angry Driving Behaviours. *IEEE Intell. Transp. Syst. Mag.* **2021**. [CrossRef]
55. Li, W.; Zeng, G.; Zhang, J.; Xu, Y.; Xing, Y.; Zhou, R.; Guo, G.; Shen, Y.; Cao, D.; Fei-Yue, W. CogEmoNet: A Cognitive-Feature-Augmented Driver Emotion Recognition Model for Smart Cockpit. *IEEE Trans. Comput. Soc. Syst.* **2021**, 1–12. [CrossRef]
56. Gao, H.; Yüce, A.; Thiran, J.P. Detecting emotional stress from facial expressions for driving safety. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 5961–5965.
57. Li, W.; Cui, Y.; Ma, Y.; Chen, X.; Li, G.; Zeng, G.; Guo, G.; Cao, D. A Spontaneous Driver Emotion Facial Expression (DEFE) Dataset for Intelligent Vehicles: Emotions Triggered by Video-audio Clips in Driving Scenarios. *IEEE Trans. Affect. Comput.* **2021**. [CrossRef]
58. Woźniak, M.; Siłka, J.; Wieczorek, M. Deep learning based crowd counting model for drone assisted systems. In Proceedings of the 4th ACM MobiCom Workshop on Drone Assisted Wireless Communications for 5G and Beyond, Virtual Event, 29 October 2021; pp. 31–36.
59. Wieczorek, M.; Sika, J.; Wozniak, M.; Garg, S.; Hassan, M. Lightweight CNN model for human face detection in risk situations. *IEEE Trans. Ind. Inform.* **2021**. [CrossRef]
60. Liu, X.; Chen, S.; Song, L.; Woźniak, M.; Liu, S. Self-attention negative feedback network for real-time image super-resolution. *J. King Saud-Univ.-Comput. Inf. Sci.* **2021**. [CrossRef]
61. Rowley, H.A.; Baluja, S.; Kanade, T. Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 23–38. [CrossRef]
62. Rowley, H.A.; Baluja, S.; Kanade, T. Rotation invariant neural network-based face detection. In Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231), Santa Barbara, CA, USA, 25 June 1998; pp. 38–44.

63. Li, S.Z.; Zhu, L.; Zhang, Z.; Blake, A.; Zhang, H.; Shum, H. Statistical learning of multi-view face detection. In Proceedings of the European Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; Springer: Berlin/Heidelberg, Germany 2002; pp. 67–81.
64. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
65. Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. Densebox: Unifying landmark localization with end to end object detection. *arXiv* **2015**, arXiv:1509.04874.
66. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. Faceness-net: Face detection through deep facial part responses. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1845–1859. [CrossRef]
67. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
68. Baldwin, K.C.; Duncan, D.D.; West, S.K. The driver monitor system: A means of assessing driver performance. *Johns Hopkins APL Tech. Dig.* **2004**, *25*, 269–277.
69. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
70. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
71. Zhou, B.; Cui, Q.; Wei, X.S.; Chen, Z.M. BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9719–9728.
72. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
73. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
74. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
75. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
76. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
77. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 279–283.
78. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
79. Ma, Z.; Mahmoud, M.; Robinson, P.; Dias, E.; Skrypchuk, L. Automatic detection of a driver's complex mental states. In Proceedings of the International Conference on Computational Science and Its Applications, Trieste, Italy, 3–6 July 2017; Springer: Cham, Switzerland, 2017; pp. 678–691.
80. Yan, Y.; Lu, K.; Xue, J.; Gao, P.; Lyu, J. Feafa: A well-annotated dataset for facial expression analysis and 3d facial animation. In Proceedings of the 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 8–12 July 2019; pp. 96–101.
81. Inthanon, P.; Mungsing, S. Detection of drowsiness from facial images in real-time video media using nvidia Jetson Nano. In Proceedings of the 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Phuket, Thailand, 24–27 June 2020; pp. 246–249.
82. Xun, D.T.W.; Lim, Y.L.; Srigrarom, S. Drone detection using YOLOv3 with transfer learning on NVIDIA Jetson TX2. In Proceedings of the 2021 Second International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), Bangkok, Thailand, 20–22 January 2021; pp. 1–6.