IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 26, 2019, accepted January 10, 2020, date of publication January 21, 2020, date of current version January 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968464

A Deep-Learning-Based Scheme for Detecting Driver Cell-Phone Use

CHONGCHONG JIN[1], ZHONGJIE ZHU[1], YONGQIANG BAI[1], GANGYI JIANG[2], (Member, IEEE), AND ANQING HE[3]

[1]Ningbo Key Laboratory of DSP, Zhejiang Wanli University, Ningbo 315000, China
[2]Institute of Technology, Ningbo University, Ningbo 315211, China
[3]Zhejiang CRRC Electric Vehicle Company, Ltd., Ningbo 315100, China

Corresponding author: Zhongjie Zhu (zhongjiezhu@yeah.net)

This work was supported in part by the National Natural Science Foundations of China under Grant 61671412, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY19F010002, in part by the Natural Science Foundation of Ningbo, China, under Grant 2018A610053, in part by the Ningbo Municipal Projects for Leading and Top Talents under Grant NBLJ201801006, and in part by the Innovation and Consulting Project from Ninghai Power Supply Company, State Grid Corporation of Zhejiang, China.

ABSTRACT Cell-phone use while driving results in potentially severe safety hazards. In this paper, a scheme for detecting cell-phone use that is based on deep learning is proposed, which can eliminate the potential risk by detecting the driver behavior and issuing an early warning. The proposed scheme consists of two stages: model training and practical testing. In the former, a multi-angle arrangement of cameras is first designed. Then, based on self-established data set, two independent convolutional neural networks (CNNs) are trained by optimizing the size and number of the convolution kernels, which can efficiently recognize cell-phones and hands in real time. In the testing stage, dynamic region extraction and skin color detection are employed as preprocessing to improve the accuracy of target recognition. Then, with the trained CNNs, the detection of cell-phone and hand targets is carried out, and the corresponding early warning is issued based on the distance of the interaction between the cell-phone and the hand. Numerous experiments are conducted and the results demonstrate that the proposed scheme can accurately detect cell-phone use during driving in real time, with a running time of 144 fps and an accuracy of 95.7%.

INDEX TERMS Cell-phone use, deep learning, dynamic region extraction.

I. INTRODUCTION

With the development of technology and the acceleration of the rhythm of life, cell-phones have brought great convenience and have gradually become people's new ''appendages''. However, cell-phone use while driving poses potentially substantial hazards to traffic safety. Studies show that the rate of traffic accidents while drivers are using cell-phones is approximately four times higher than that during normal driving [1], [2]. Thus, most countries have issued regulations that limit drivers' cell-phone use [3]. To implement these regulations, the current practice is to dispatch law enforcement officers on the roadside to visually inspect incoming cars, or to use surveillance cameras to detect the targets, and, subsequently, to impose administrative punishment, which not only consumes a substantial amount of resources but also cannot eradicate this type of behavior fundamentally [4]. Therefore, dynamically detecting cell-phone use while driving and issuing warnings against this behavior via intelligent technology is of substantial importance.

For the behavior of cell-phone use, the first problem to be solved is cell-phone and hand detection and identification. Target detection is an important research area in the modern image processing field and has a broad research foundation and application prospects [5], [6]. Currently, target detection methods can be roughly divided into two categories: artificial-feature-based methods and deep-learning-based methods. In the former, various artificial features are designed according to the target characteristics. Then, target detection and recognition are conducted via regression. Viola *et al.* proposed a rapid object detection method that uses a boosted cascade of simple Haar features [7]. Dalal *et al.* selected the histogram of oriented gradients (HOG) feature and used support vector machine as the classifier to detect targets [8], [9]. Felzenszwalb *et al.* proposed

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

the deformable parts model for further improving the accuracy of target detection [10]. However, these methods focus mostly on single object and relatively intuitive feature. Thus, the robustness and generalization performance of these methods must be further improved in view of the diversity of objectives and characteristics.

Deep-learning-based methods are the mainstream development direction of target detection. Sermanent *et al.* introduced convolutional networks to improve the detection performance [11]. He *et al.* proposed SPP-Net by adding spatial pyramid pooling between the convolutional layer and the fully connected layer to avoid the influence of scaling candidate regions [12]. Girshick *et al.* proposed a target detection model that was based on regional convolutional neural networks (R-CNNs) [13]. Later, Girshick *et al.* proposed the Fast R-CNN detection algorithm, which realized multi-task learning and substantially increased the detection speed [14]. Ren *et al.* designed the candidate region generation network with the Faster R-CNN, and realized end-to-end deep learning with higher speed and accuracy [15]. Lin *et al.* proposed the feature pyramid networks detection algorithm, which is based on Faster R-CNN [16]. In addition, algorithms such as YOLO [17], SSD [18], and Retina-Net [19] also show excellent detection performance.

Behavior detection of cell-phone use has stricter requirements than traditional target detection. After the detection and recognition of the cell-phone and hand, it is necessary to distinguish the interaction between them, namely, whether the driver is using the cell-phone or not. Considering the real-time performance and accuracy of data acquisition, the current research in this field focuses mainly on the hardware design of sensors and the algorithm design of target detection, which will be roughly introduced as follows.

For the hardware design of sensors, signal sensors for detecting audio/network signals from cell-phones are typically installed, which are used to determine whether the cell-phone in the cab is active [20]–[23]. Rodríguez-Ascariz *et al.* proposed an automatic electronic system for capturing the cell-phone voice signal and identifying the time of drivers cell-phone use, which can effectively detect all cell-phone voice signals but cannot distinguish between the signals of pedestrian and passengers [4]. Li *et al.* invented a cell-phone signal shielding device that shields cell-phone signals in a prescribed area of approximately half of a square meter around the driver, which could fundamentally eliminate the possibility of the driver using a cell-phone. However, it also prevented drivers from using cell-phones to seek help in an emergency [24], [25]. Liu *et al.* used embedded sensors to detect whether the driver was using cell-phones by analyzing the associated information, such as data from touching screens [26], [27]. Wang *et al.* developed a cell-phone integrated sensor for detecting the use of cell-phones from changes in cell-phone location information that is obtained by embedded sensors [28]–[30]. Leem *et al.* proposed an impulse radio ultrawideband radar for monitoring driver anomalies, including vital signs and cell phone signals, and detect the use by setting the area between the steering wheel and the operating lever as the region of telephone detection [31]. These hardware designs have yielded satisfactory results by detecting various types of physical information for behavior detection and recognition; however, factors such as cell-phone model, personal habits, and installation methods typically lead to false recognition, privacy disclosure and other problems.

For the algorithm design of target detection, available methods mainly acquire image information through cameras to recognize targets and determine behaviors [32]. The research in this field remains in its infancy for practical applications. Artan *et al.* acquired images through a near-infrared camera and designed an elastic deformation model to locate the facial area of the driver. Subsequently, they used machine-learning-based image classifier technology to detect the use of cell-phone [33]. Based on the machine learning, Xu *et al.* designed a deformable part model to locate the frontal windshield region and utilized the Fisher vector representation in one side of windshield to classify the violation behavior of cell-phone use [34]. All these methods typically acquired drivers' photos by mounting the camera on road poles to make law enforcement officers obtain evidence easier. However, due to factors like the camera angle, occlusion between objects and body postures, the accuracy of these methods must be improved further. Therefore, Wang *et al.* utilized a vehicle camera to capture driver's pictures and determined driver's irregular behavior based on the skin color around face indirectly [35]. The accuracy of this method is also affected by above factors inevitably.

As discussed above, the existing techniques of driving behavior detection of cell-phone use are not satisfactory, the development of more efficient techniques are expected. Hence, in this paper, a new deep-learning based scheme is proposed. In our scheme, A multi-angle arrangement of cameras is used to improve the integrity of image acquisition and ensure the detection accuracy of target recognition. Two independent CNNs are trained by optimizing the size and number of convolution kernels, which can recognize cell-phones and hands efficiently in real time. With the trained CNNs, the target relationship between cell-phones and hands are carried out. Then, the corresponding early warning and recording are issued according to the distance of the interaction between the cell-phone and hand. The main contributions of this paper are as follows:

1. Multi-angle arrangement of cameras. Due to weather variation, object occlusion, human posture and other factors, the camera mounting position has a substantial influence on the detection accuracy when the camera is installed outside the vehicle for image acquisition. In this paper, multi-angle arrangements of cameras are installed in the vehicle, which can not only more clearly capture the complete internal area to effectively improve the accuracy of the follow-up target detection but also provide a large amount of photographic evidence for law enforcement departments as a new technology in the vehicle assistant system.

2. Optimization of image acquisition. Vehicle driving is a continuous behavior, which will produce many interrelated image sequences for image acquisition. Based on the regularity of human behavior and the characteristics of the human visual system, we consider the interaction of the multi-angle camera system and design a reasonable frame rate for image acquisition to ensure the efficiency of target detection and the real-time performance of the scheme.

3. Design of parallel processing of dual networks and interactive algorithms. For the training of the neural network, we adopt the structure of parallel processing of dual networks, namely, two independent CNNs are trained with the database to recognize cell-phones and hands efficiently and in real time. Meanwhile, we calculate the Euclidean distance between the cell-phone and the hand to distinguish their interaction and realize the recognition and early warning of cell-phone use behavior.

The remainder of the paper is organized as follows: In Section II, the proposed scheme, which includes the establishment of the database, the training stage and the testing stage are described in detail. The experimental results and analysis are presented in Section III. Finally, we present the conclusions of our work in Section IV.

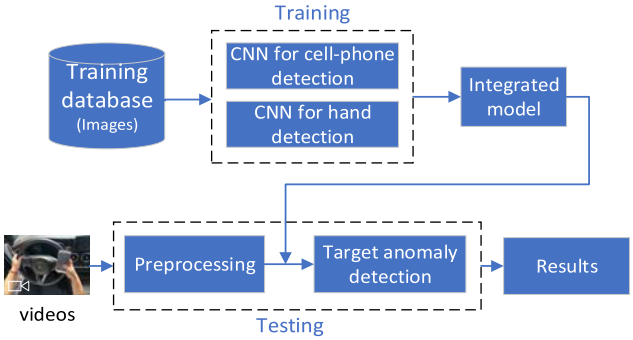


FIGURE 1. Diagram of the proposed scheme.

II. PROPOSED SCHEME

The proposed scheme consists of two parts, camera setting and network design. First, multi-angle cameras are installed around the driver for sample collection to improve the robustness of behavior detection of cell-phone use. Then, collected images are labeled manually and two CNNs are optimized and trained for object detection. Finally, the two networks are integrated together for testing the accuracy and real-time performance of target detection. In order to more in line with actual traffic condition, we used videos as the input during testing stage and preprocessed videos to reduce the interference of background. The preprocessed video frames are input into trained CNNs for testing, and abnormality detection is performed on targets to obtain the ultimate experimental results. A diagram of the proposed scheme is shown in Fig. 1.

A. ESTABLISHMENT OF THE DATABASE

1) NUMBER AND LOCATIONS OF VEHICLE CAMERAS

In the traditional camera arrangement for target detection, some cameras are mounted on road poles outside the vehicle

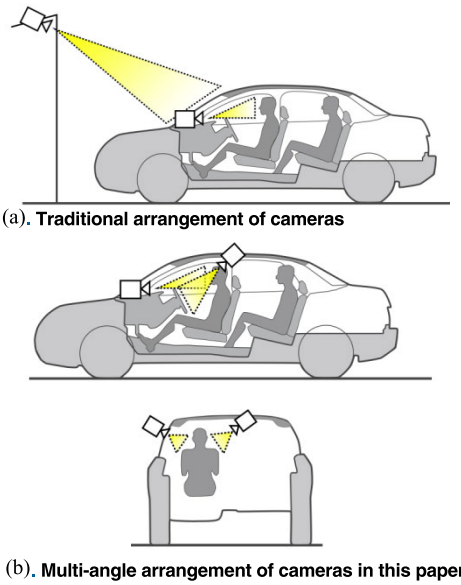


FIGURE 2. (A) Traditional arrangement of cameras. (B) Multi-angle arrangement of cameras in this paper.

to photograph the windshield area then detect driver's face further, the others are installed inside the vehicle in front of the driver to detect face area directly, as illustrated in Fig. 2(A). All captured images are the front face of the driver [33], [34], and the behaviors that are occurring in steering wheel area and joystick area cannot be captured.

To solve the occlusion problem, four vehicle cameras are installed at the front, top, left and right sides of the driver, as illustrated in Fig. 2(B). Multi-angle cameras capture the posture of holding cell-phone at various angles to increase the diversity of samples and reduce body position occlusion of targets which can ensure the robustness of proposed algorithm.

2) TARGET TYPES IN THE SAMPLES

In this paper, two forms of data, images and videos, are obtained for training and testing, respectively. The regions of hands are labeled as hand positive samples and both the regions of hands and the hand-backgrounds constitute the hand-training database. Similarly, the phone positive samples and the phone-backgrounds constitute the phone-training database. Videos are used for testing and can be divided into three types to simulate different relationships between the hand and the cell-phone. The first type has only the hand. Obviously, the phone is not used at the moment; In the second type, both the hand and the phone exist but do not overlap, which is used to simulate the presence of the cell-phone in the cab but the driver does not touch it; In the third type, both the hand and the phone exist and overlap, which is used to simulate the behavior of cell-phone use while driving.

3) LABELING

To improve the accuracy of target detection, we label hands and cell-phones separately. During labeling, the backgrounds

FIGURE 3. Labeling of target and background.

are carefully selected to enhance detection robustness. For example, due to its color similarity to hand, the arm area is easily mis-detected as a hand. Thus, when labeling the background of the hand, we select more areas similar to the hand to improve the flexibility of samples, as shown in Fig. 3. Meanwhile, when labeling the hand and the cell-phone, we require that the labeling frame of positive samples does not contain their overlapping areas.

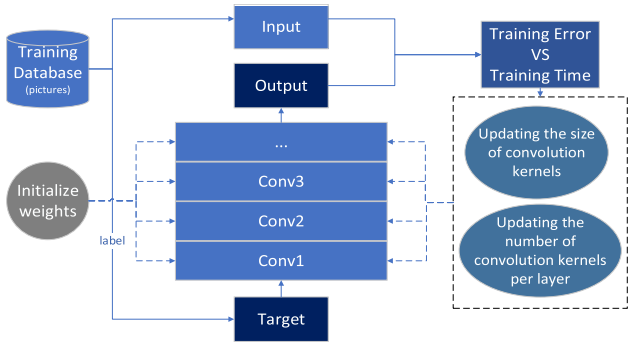


FIGURE 4. Diagram of the training stage.

B. TRAINING STAGE

The strategies of CNN training for hands and cell-phones are similar. Here, we take the CNN training of cell-phone as example to demonstrate the optimization process. The flow chart of training stage is presented in Fig. 4. When cell-phone images are input into the initial CNN for training, the sizes of the convolution kernels and the number of convolution kernels per layer must be continuously optimized to improve detection accuracy and the real-time performance.

1) INITIAL NETWORK MODEL SELECTION

CNN structures are widely used in target detection. In 2012, Hinton *et al.* [36] proposed the Alex-Net model, which is mainly used in the field of image classification, and the error rate of results on the ImageNet data set is reduced to 15.3%. Later, VGGNet [37], which was proposed by Oxford University, became the mainstream CNN structure due to its narrow and deep convolution structure. Experiments demonstrated that its migration learning performance is very strong and is used by many models. There are also two mainstream CNNs, namely, GoogLeNet [38] and ResNet [39], for the target detection model. The GoogLeNet complex inception structure utilizes multi convolution kernels. the results demonstrate that it can effectively improve the utilization of computing resources. ResNet has a higher level of precision due to its deep hierarchy and its use of residual nodes. To improve the performance of the model on special scenes, other cnn structures such as densenet [40] have emerged.

With the continuous development of neural networks, various networks are constantly improving in terms of accuracy. However, CNN-based devices have not been effectively popularized in view of the real-time performance. Hence, considering accuracy, real-time and copyright issues, we draw on the classic Alexnet model and set up a target detection scheme with dual channel network, which can share training data on two GPU computers, and finally combine the results of two GPUs to take advantage of limited hardware resources.

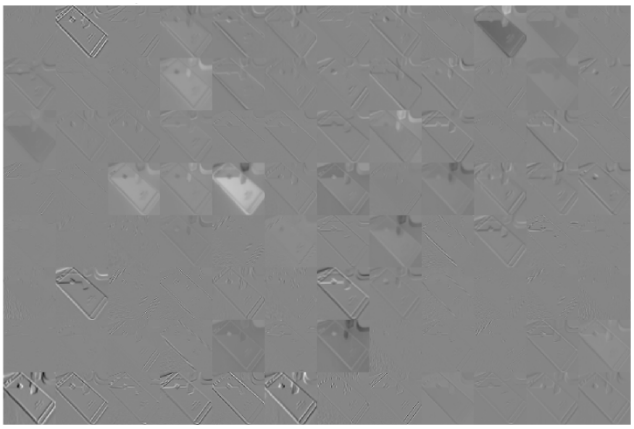


FIGURE 5. Feature map.

2) NETWORK STRUCTURE OPTIMIZATION

The initial model has 5 convolutional layers and 3 fully connected layers [36], and the average time for the convolutional layers occupies 87% of the total network time. Therefore, the time of image detection mostly depends on the complexity of the convolutional layers, and the complexity of the convolutional layers depends on the sizes of the convolution kernels and the number of convolution kernels in each layer. These two factors are determined via training to reduce the time of detection and realize real-time detection while maintaining the detection accuracy.

The parameter of each layer is computed as follows:

$$resolutions = \left(\frac{m-n}{s}+1\right) \times \left(\frac{m-n}{s}+1\right) \times k \quad (1)$$

where m is the side length of the normalized square image, n is the side length of the square convolution kernels, s is the step length that the convolution kernels traveled in the whole normalized picture, and k is the number of convolution kernels per layer.

The smaller the convolution kernels are, the fewer parameters and lower computational complexity. The resolution of the normalized image is $227 \times 227 \times 3$, the resolution of the convolution kernels is 11×11, the step length is 4 and there are 96 convolution kernels initially. Hence, the resolution after the first convolution operation is $[(227-11) \div 4+1] \times [(227-11) \div 4+1] \times 96 = 55 \times 55 \times 96$. The resolution of the pooling layer is 3×3, thus, the resolution after the first pooling operation is $[(55-3) \div 2+1] \times [(55-3) \div 2+1] \times 96 = 27 \times 27 \times 96$. We reset the size of the first-layer convolution kernels to 3×3, 7×7, 11×11, and 15×15 (n=3, 7, 11, and 15, respectively) and keep the step length s and the number of kernels k unchanged. $m-n$ should be divided by the step length and we fine tune the size of the pooling kernels to ensure that the results after pooling are unchanged. For example, we set the resolutions of the convolution kernels from 11×11 to 7×7 in the first layer; Therefore, the resolution after the first convolution operation is $[(227-7) \div 4+1] \times [(227-7) \div 4+1] \times 96 = 56 \times 56 \times 96$. To divide by the pooling kernels, we will fine-tune the resolution of the first pooling layer from 3×3 to 4×4, hence, the resolution after the pooling operation remains $[(56-4) \div 2+1] \times [(56-4) \div 2+1] \times 96 = 27 \times 27 \times 96$.

In addition, we optimize the number of convolution kernels per layer. For the CNN, many feature maps are generated through each convolutional layer, and each feature map reflects the quality of the target feature extraction by each convolution kernel. Therefore, we propose a method of using feature maps to optimize the convolution kernels in each layer. Considering the detection of cell-phones as an example. The first layer of the convolutional layer contains 96 convolution kernels, which will produce 96 feature maps, as shown in Fig. 5. The entropy of each feature graph is calculated:

$$H = \sum_{i=0}^{255} p_i \log p_i \tag{2}$$

where p_i is the probability of a specified gray level.

The calculated entropy values are normalized, and the 96 values are sorted, as plotted in Fig. 6. A visual map of the target quality of the convolution kernel detection is obtained.

We set a threshold for deleting the convolution kernels:

$$\frac{D_{\mathrm{E}}}{N_{\mathrm{E}}} < Threshold \tag{3}$$

where D_E is the number of convolution kernels that must be deleted and N_E is the number of convolution kernels that the current layer contains.

We set the threshold in percentage format. If the threshold is set as 10%, the convolution kernels with the lowest entropy values that are sorted from 1 to 9 are deleted. When setting the threshold for deleting the convolution kernels, the detection accuracy and the real time calculation results are inversely proportional. Thus, setting a suitable threshold for the targets is critical.

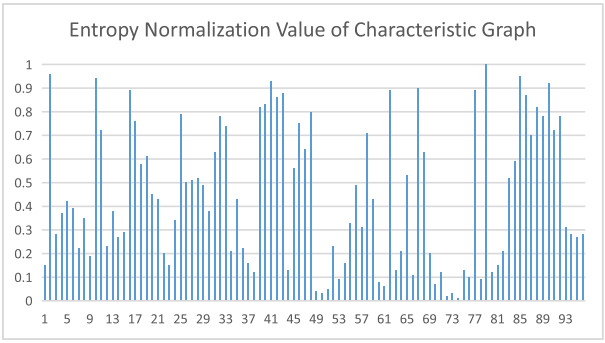


FIGURE 6. Sorted normalized entropy values.

C. TESTING STAGE

The testing flow chart is shown in Fig. 7. The test database that must be detected is initially preprocessed. A preprocessed image is input into the trained CNN to locate the target. The threshold of the Euclidean distance between the hand and the cell-phone is set for detecting anomalies between targets.

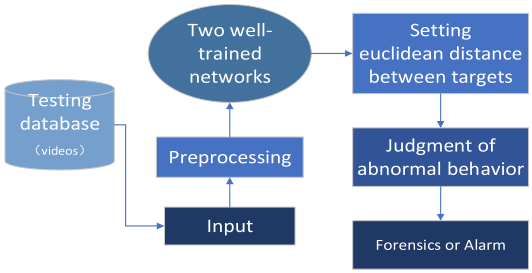


FIGURE 7. Diagram of the network testing phase.

1) PREPROCESSING

Traditionally, target detection involves capturing an image and inputting the entire image into the trained CNN [13]. This method is computationally intensive and has a high false-positive rate. Our experiments must issue the warning before the driver uses the cell-phone and must capture the photographic forensic evidence while the driver is using the cell-phone. Therefore, the algorithm must have the ability to detect in real time. Hence, we propose using the video as the sample and preprocessing the sample before it enters the trained CNN.

In our scheme, dynamic region detection [41] is conducted to reduce the number of possible candidate regions. The classic three-frame difference method is employed, and its basic steps can be described as follows:

a) Let $I_{t-1}(x, y)$, $I_t(x, y)$, and $I_{t+1}(x, y)$ be three consecutive frames of a video. Two difference images are calculated:

$$D_{t-1,t}(x, y) = |I_{t-1}(x, y) - I_t(x, y)| \tag{4}$$

$$D_{t,t+1}(x, y) = |I_t(x, y) - I_{t+1}(x, y)| \tag{5}$$

b) Binarize the difference images by choosing an appropriate threshold:

$$B_{t-1,t}(x,y) = \begin{cases} 1 & D_{t-1,t}(x,y) \geq threshold \\ 0 & D_{t-1,t}(x,y) < threshold \end{cases} \quad (6)$$

$$B_{t,t+1}(x,y) = \begin{cases} 1 & D_{t,t+1}(x,y) \geq threshold \\ 0 & D_{t,t+1}(x,y) < threshold \end{cases} \quad (7)$$

where the threshold is determined by considering driving environment to distinguish background dither and moving amplitude of targets. Based on experiments, we find that the threshold can be set between 45 and 60.

c) The logic operation with the "and" operator is applied to the two binary images to obtain the initial dynamic area:

$$R_t(x,y) = \begin{cases} 1 & B_{t-1,t}(x,y) \&\& B_{t,t+1}(x,y) = 1 \\ 0 & B_{t-1,t}(x,y) \&\& B_{t,t+1}(x,y) = 0 \end{cases} \quad (8)$$

d) Eliminate the noise via median filtering with a window size of 3×3.

In order to eliminate other moving interference, we added a skin color model to exclude non-skin dynamic regions. Firstly, a Gaussian function is used to establish a similar function. Then, binarization is conducted via the optimal threshold method. At the end, the skin region that is identified via morphological processing must occupy 1/4 of the total area of the pixels.

Finally, the obtained skin-moving region by projecting horizontally and vertically to obtain the bounding box and extended double distance of the bounding box to surround the cell-phone. The screened dynamic skin region is normalized to a fixed size to facilitate the subsequent CNN-based target segmentation. The normalization does not affect the accuracy of the subsequent object classification. In this paper, each dynamic region is normalized to a resolution of 227×227.

2) SETTING THE EUCLIDEAN DISTANCE THRESHOLD

In the case of the simultaneous presence of the hand and the cell-phone, it is necessary to determine whether the two regions overlap or are far away from each other. Due to the variability of hand postures and cell-phone sizes, the size of the area that is obtained by the bounding box is not fixed. Therefore, we use the method of setting the distance threshold of two regions to determine whether there is any overlap between the hand area and the cell-phone area to judge whether the driver is using his or her cell-phone or not. The main steps are as follows:

First, the four vertex coordinates of the hand bounding box are $a1, a2, a3, a4$; the four vertex coordinates of the cell-phone bounding box are $b1, b2, b3, b4$; and a diagram of the target models is shown in Fig. 8.

Second, the coordinates of the central points of the two regions are calculated as follows:

$$c1 = (\frac{a1+a2}{2}, \frac{a1+a3}{2}), \quad c2 = (\frac{b1+b2}{2}, \frac{b1+b3}{2}) \quad (9)$$

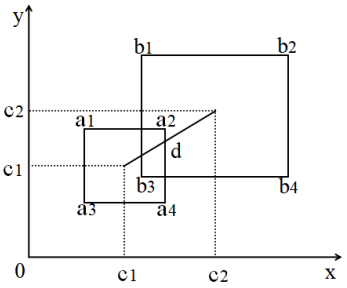


FIGURE 8. Model map of the target area.

Third, we calculate the distance between the two regions as follows:

$$d = \sqrt{(y_{c2} - y_{c1})^2 + (x_{c2} - x_{c1})^2} \quad (10)$$

Finally, the radii of the two rectangular bounding boxes are calculated as follows:

$$r1 = \frac{a2 - a1}{2}, \quad r2 = \frac{b2 - b1}{2} \quad (11)$$

If $d \leq r1 + r2$, there is an overlap between the regions of the hand and the cell-phone. We can prove that the driver has violated the law regarding cell-phone use, and we can use the picture as evidence. In contrast, if $d > r1 + r2$, the driver is not using his cell-phone; however, since the hand and the cell-phone are in the picture at the same time, we can issue early warnings to the driver.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In the simulation experiment, two PCs are used, which have a 3.20 GHZ CPU, 16 GB of memory and a GTX 1060 6 G GPU, and the test software is MATLAB R2017b.

A. TYPE AND QUANTITY OF THE DATABASE

the original data samples are collected by four mini-adhesive cameras with Huawei HiSilie chip, the charging method is USB, camera size is $3.5 \times 3.5 \times 4$ cm, memory card is 64 G, the photo resolution is 4032×3024, video resolution is 1920×1080, and video frame is 35 FPS.

TABLE 1. Type and quantity of the database.

Type	Format	Notes	Quantity
Training	Image	Hand	11000
		Hand-background	11000
		phone	11000
		Phone-background	11000
Testing	Video (35fps, 5s)	Only-hand existed	8
		Both hand and phone existed and overlapped	15
		Both hand and phone existed and no-overlapped	15

As mentioned above, the database captured by cameras has two forms: images and videos, which are shown in Table 1. All testing videos took five seconds and can be classified into

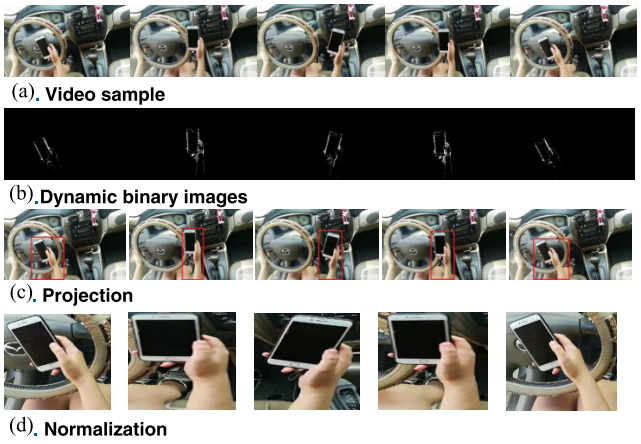


FIGURE 9. Results of preprocessing.

three categories: only-hand (8 videos), both hand and phone existed and overlapped (15 videos), both hand and phone existed but no-overlapped (15 videos).

B. TEST RESULTS AND ANALYSIS

1) PREPROCESSING RESULTS AND COMPARISON

The preprocessing results are shown in Fig. 9. Fig. 9(A) shows sample frames of cell-phone behavior, Fig. 9(B) shows the binary results of dynamic targets, Fig. 9(C) shows the bounding boxes that are generated via the projection of dynamic targets and the skin color model, and Fig. 9(D) shows the normalized results.

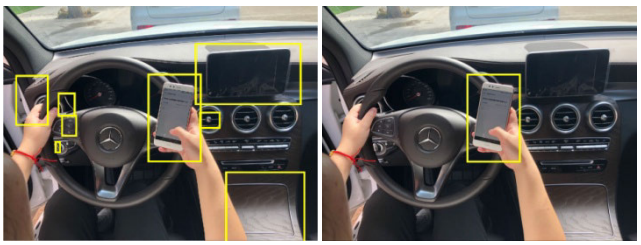


FIGURE 10. Results of candidate area optimization.

We compare the preprocessing results with those of the traditional candidate region extraction method. The traditional candidate region extraction method typically traverses the whole picture and sets a confidence score threshold to reduce the effects of invalid bounding boxes [42]. However, due to the interference from background colors and textures, the candidate regions will have high false detection rates. The experimental results of detecting the cell-phone are shown in Fig. 10(A). After the dynamic target extraction method and the skin color model are improved, the background interference of the candidate regions is reduced, the bounding boxes are more precise and the efficiency of extracting the candidate regions is substantially improved. The experimental results are shown in Fig. 10(B). Moreover, the dynamic target detection and skin color model do not take much time in the entire CNN model.

2) RESULTS OF OPTIMIZING THE SIZE AND NUMBER OF CONVOLUTION KERNELS

First, we vary the convolution kernel size in the first convolutional layer among 3×3, 5×5, 7×7, 11×11 and the training times are all approximately 20 hours. However, the training error rates differ substantially. The error rate is the lowest when the 7×7 convolution kernel is used. Hence, we set the size of the convolution kernel to 7×7. The results are plotted in Fig. 11.

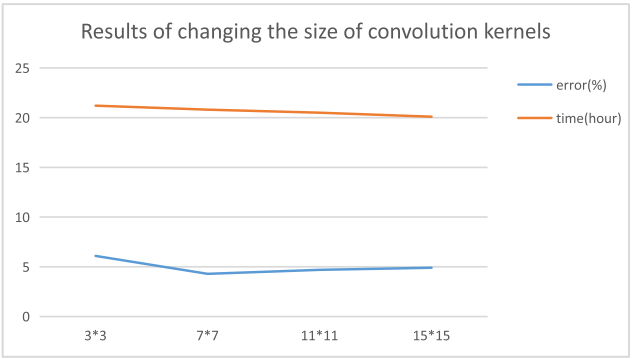


FIGURE 11. Results of optimizing the size of the convolution kernels.

Second, we set the threshold for removing weak convolution kernels from each layer, the data are presented in Table 2. From the experimental results, we conclude that when 0% of the convolution kernels are deleted, the accuracy is the highest, but real-time detection of frames cannot be realized. When 30% of the convolution kernels are deleted, the model has high real-time performance and the number of convolution kernels in the model parameters can be reduced from 4 million to 720,000, which reduces the calculation time by approximately five times compared to the 0% model. However, the accuracy is low.

Therefore, to balance the accuracy and real-time performance in the experiment, we set the threshold to 10%. After removing 10% of the convolution kernels, the accuracy of the experiment is still 95.7%. Compared with the original model, the accuracy decreases by 2%, but the real-time performance is improved by approximately four times. When we removed 12% of the convolution kernels, the accuracy began to drop dramatically. Hence, it is not effective to continue to remove convolution kernels.

In cell-phone detection experiments, deleting 10% of the convolution kernels is an appropriate threshold. It guarantees an accuracy of 95.7%, and the average test time of each frame is 144 ms. Hence, approximately 6-7 frames can be detected in one second, which ensures that all four cameras can detect at least one frame per second. It maintains a satisfactory balance between accuracy and real-time performance, which is of important social significance in practical applications.

TABLE 2. Quantity and parameters of convolution kernels.

Threshold for Deleting Convolution Kernels（%）	Accuracy Rate（%）	Test Time （ms/frame）
0	97.7	548
5	96.1	276
8	95.9	238
10	95.7	144
12	86.8	98
15	73.3	75
20	62.5	54
30	51.6	32

3) RESULT COMPARISON BETWEEN SINGLE CNN AND DUAL CNN

In order to evaluate the two CNNs in terms of classification accuracy, the confusion matrix is used which is an error matrix that is often used to visually evaluate the performance of supervised learning algorithms. In this paper the size of the confusion matrix is n_classes × n_classes, where n_classes represents the number of classes. We compare the confusion matrices of the hand, cell-phone and background classification, and visualize the main issues of the three categories.

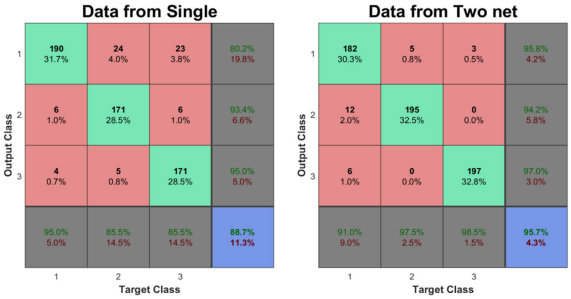


FIGURE 12. Classification results of single CNN and dual CNN.

The numbers of optimal classifications in the CNN are compared in Fig. 12. We randomly select 200 pictures as test samples, and we compare the accuracies of single network training and two networks training, where "1" represents background, "2" represents hands and "3" represents cell-phones. Among the four colors in the graph, green represents the correct number of classifications, pink represents the number of incorrect classifications for one target, gray represents the recognition accuracy for each part, and gray-purple represents the overall accuracy. For example, entry (2, 2) of the single network matrix indicates that there are 171 hand pictures that are correctly classified, and entry (2, 1) indicates that 24 hand pictures are incorrectly classified as background.

According to the experimental data, the main reason for the low accuracy of a single network is that hands and cell-phones are misclassified as background. In practice, there is typically an overlap between a hand and cell-phone. To improve this classification, we train and test the target networks separately,

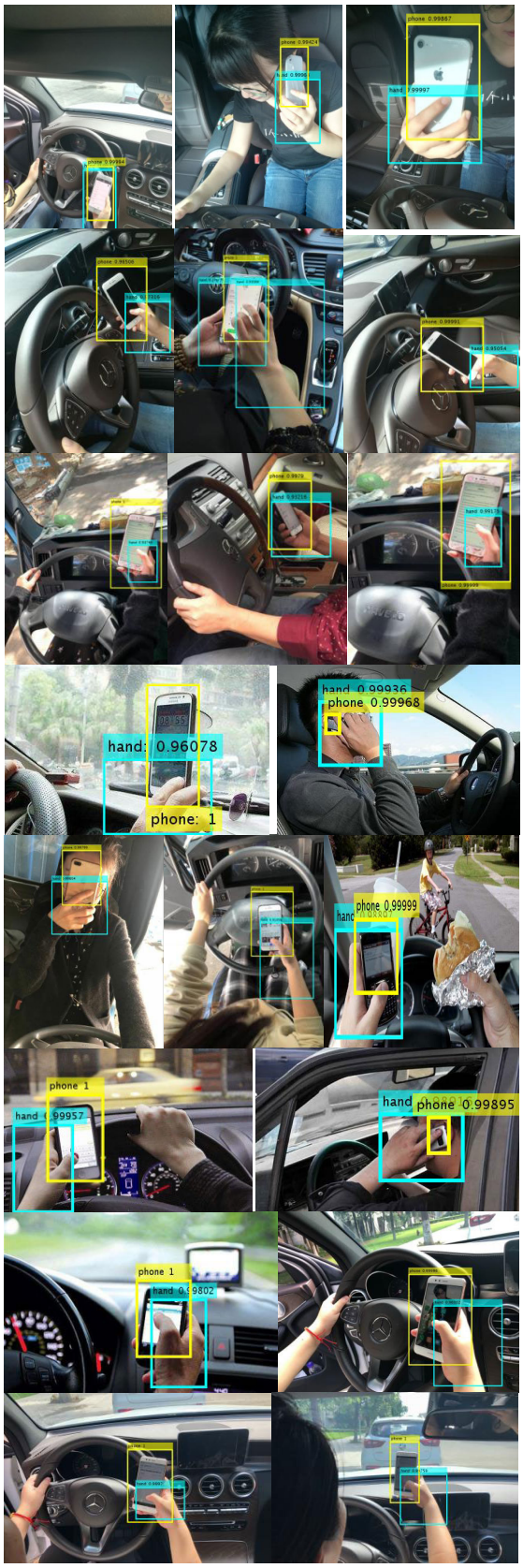


FIGURE 13. Example results of image detection.

and the experimental results that are obtained are shown in Fig. 12, which have been modified. After the targets are trained on their own networks, the numbers of hands and cell-phones that are misclassified as background are substantially reduced. For example, as specified in entry (1, 2), the number of hands that are classified as background is 5, compared with 24 in the single network, hence, the error rate is reduced by approximately 5 times. The problem of targets overlapping is substantially reduced, and the overall recognition rate is increased by approximately 7%, which substantially increases the applicability of the algorithm.

4) VISUALIZATION OF THE EXPERIMENTAL RESULTS

The final classification results are divided into three categories. If the classification result is only hand exist, there is no cell-phone use. If the hand and cell-phone both exist but no-overlap, we alarm the driver. If both the hand and cell-phone exist and overlap, we confirm that there exists cell-phone use behavior. A sample of the experimental results for cell-phone use is shown in Fig. 13.

IV. CONCLUSION

This paper proposed a cell-phone-use behavior detection scheme that is based on deep learning, which can eliminate the potential risk by detecting the driver behavior and issuing an early warning efficiently and in real time. A multi-angle arrangement of cameras is used to improve the integrity of image acquisition and to ensure the detection accuracy of target recognition for the scheme design. Two independent CNNs are trained by optimizing the size and number of the convolution kernels, which can efficiently recognize cell-phones and hands in real time. Then, with the trained CNNs, the corresponding early warning or forensics is issued based on the distance of the interaction between the cell-phone and the hand. Numerous experiments are conducted, and the results demonstrate that the proposed scheme can accurately detect cell-phone use behavior while driving in real time, with running time of 144 fps and the accuracy of 95.7%.

Although this scheme has yielded satisfactory test results and is efficient and practicable, as a preliminary attempt at behavior detection, many aspects, such as the camera arrangement, video data acquisition and network structure optimization, must be further studied. The design of end-to-end network structure will be the main direction of future research.

REFERENCES

[1] V. Žuraulis, S. Nagurnas, and R. Pečeliūnas, "The analysis of drivers' reaction time using cell phone in the case of vehicle stabilization task," *Int. J. Occupational Med. Environ. Health*, vol. 31, no. 5, pp. 633–648, Oct. 2018.

[2] J. Wahlstrom, I. Skog, and P. Handel, "Smartphone-based vehicle telematics: A ten-year anniversary," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 10, pp. 2802–2825, Oct. 2017.

[3] A. T. Mccartt, D. G. Kidd, and E. R. Teoh, "Driver cellphone and texting bans in the United States: Evidence of effectiveness," *Adv. Automot. Med.*, vol. 58, pp. 99–114, Dec. 2014.

[4] J. M. Rodríguez-Ascariz, L. Boquete, J. Cantos, and S. Ortega, "Automatic system for detecting driver use of mobile phones," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 673–681, Aug. 2011.

[5] Á. F. García-Fernández, L. Svensson, and M. R. Morelande, "Multiple target tracking based on sets of trajectories," *IEEE Trans. Aerosp. Electron. Syst.*, to be published.

[6] T. Intharah, D. Turmukhambetov, and G. J. Brostow, "HILC: Domain-independent PbD system via computer vision and follow-up questions," *ACM Trans. Interact. Intell. Syst.*, vol. 9, nos. 2–3, p. 16, Mar. 2019.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, vol. 1, Dec. 2001, pp. 511–518.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[9] S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 211–223, Feb. 2016.

[10] P. Felzenszwalb, D. Mcallester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[11] P. Sermanet, D. Eigen, M. Mathieu, R. Fergus, Y. LeCun, and X. Zhang, "OverFeat: Integrated recognition, localization and detection using convolutional networks," Dec. 2013, *arXiv:1312.6229*. [Online]. Available: https://arxiv.org/abs/1312.6229

[12] K. He, X. Zhang, J. Sun, and S. Ren, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jan. 2015.

[13] R. Girshick, J. Donahue, J. Malik, and T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[14] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[16] T. Y. Lin, P. Dollár, K. He, B. Hariharan, S. Belongie, and R. Girshick, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

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

[18] W. Liu, D. Anguelov, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, and D. Erhan, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[20] J. Yang, S. Sidhom, T. Vu, H. Liu, N. Cecan, Y. Chen, M. Gruteser, R. P. Martin, and G. Chandrasekaran, "Detecting driver phone use leveraging car speakers," in *Proc. ACM 17th Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2011, pp. 97–108.

[21] J. Yang, S. Sidhom, G. Chandrasekaran, T. Vu, H. Liu, N. Cecan, Y. Chen, M. Gruteser, and R. P. Martin, "Sensing driver phone use with acoustic ranging through car speakers," *IEEE Trans. Mobile Comput.*, vol. 11, no. 9, pp. 1426–1440, Sep. 2012.

[22] M. Rosen, "Method and system for automated detection of mobile phone usage," U.S. Patent 8 384 555, Feb. 26, 2013.

[23] M. L. Zeinstra and P. J. Vanderwall, "In-vehicle electronic device usage blocker," U.S. Patent 13 978 540, Jul. 11, 2013.

[24] L. Nian-Feng, G. He, Z. Tong, and Z. Meng, "Study on mobile phone call monitoring and positioning and shielding algorithms," in *Proc. Int. Conf. Transp., Mech., Elect. Eng. (TMEE)*, Changchun, China, Dec. 2011, pp. 1771–1774.

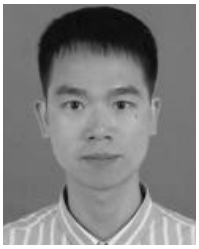
[25] N. F. Li, M. Zhang, H. Gu, and Y. J. Yang, "Study on a kind of mobile phone signals monitoring and shielding system," in *Advances in Intelligent Systems* (Advances in Intelligent and Soft Computing), G. Lee, Ed. Berlin, Germany: Springer, 2012, pp. 321–326.

[26] X. Liu, J. Cao, S. Tang, Z. He, and J. Wen, "Drive now, text later: Nonintrusive texting-while-driving detection using smartphones," *IEEE Trans. Mobile Comput.*, vol. 16, no. 1, pp. 73–86, Jan. 2017.

[27] C. Bo, X. Jian, X. Mao, Y. Wang, F. Li, and X. Y. Li, "You're driving and texting: Detecting drivers using personal smart phones by leveraging inertial sensors," in *Proc. Int. Conf. Mobile Comput. Netw.*, vol. 7, Dec. 2013, pp. 199–202.
[28] Y. Wang, Y. J. Chen, J. Yang, M. Gruteser, R. P. Martin, H. Liu, L. Liu, and C. Karatas, "Determining driver phone use by exploiting smartphone integrated sensors," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 1965–1981, Aug. 2016.
[29] Y. Wang, J. Yang, Y. Chen, M. Gruteser, R. P. Martin, and H. Liu, "Sensing vehicle dynamics for determining driver phone use," in *Proc. ACM 11th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2013, pp. 41–54.
[30] J. Lindqvist and J. Hong, "Undistracted driving: A mobile phone that doesn't distract," in *Proc. ACM 12th Workshop Mobile Comput. Syst. Appl.*, Mar. 2011, pp. 70–75.
[31] S. Leem, F. Khan, and S. Cho, "Vital Sign monitoring and mobile phone usage detection using IR-UWB radar for intended use in car crash prevention," *Sensors*, vol. 17, no. 6, p. 1240, May 2017.
[32] M. J. Smith and D. R. Stephens, "Detecting use of a mobile device by a driver of a vehicle, such as an automobile," U.S. Patent 13 290 126, Aug. 23, 2012.
[33] Y. Artan, O. Bulan, P. Paul, and R. P. Loce, "Driver cell phone usage detection from HOV/HOT NIR images," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, Columbus, OH, USA, 2014, pp. 225–230.
[34] B. Xu and R. P. Loce, "A machine learning approach for detecting cell phone usage," *Proc. SPIE*, vol. 9407, Mar. 2015, Art. no. 94070A.
[35] D. Wang, M. Pei, and L. Zhu, "Detecting driver use of mobile phone based on in-car camera," in *Proc. 10th Int. Conf. Comput. Intell. Secur.*, Kunming, China, 2014, pp. 148–151.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[39] 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.
[40] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.
[41] P.-Y. Lv, S.-L. Sun, C.-Q. Lin, and G.-R. Liu, "Space moving target detection and tracking method in complex background," *Infr. Phys. Technol.*, vol. 91, pp. 107–118, Jun. 2018.
[42] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

CHONGCHONG JIN received the M.E. degree from Zhejiang Wanli University, China, in January 2019, under the supervision of Prof. Z. J. Zhu. She is currently pursuing the Ph.D. degree in information and communication engineering with Ningbo University, China. Her research interests include image detection and location, neural networks, and deep learning.

ZHONGJIE ZHU received the Ph.D. degree in electronics science and technology from Zhejiang University, China, in 2004. He is currently a Professor with the Faculty of Electronics and Information Engineering, Zhejiang Wanli University, China. His research interests mainly include video compression and communication, image analysis and understanding, watermarking and information hiding, and 3D image signal processing.



YONGQIANG BAI received the B.S. and M.S. degrees from Zhengzhou University, China, in 2006 and 2009, respectively, and the Ph.D. degree from Ningbo University, China, in 2019. He is currently a Researcher with Zhejiang Wanli University, China. His research interests mainly include data hiding and image processing.

GANGYI JIANG (Member, IEEE) received the M.S. degree from Hangzhou University, China, in 1992, and the Ph.D. degree from Ajou University, South Korea, in 2000. He is currently a Professor with the Faculty of Information Science and Engineering, Ningbo University, China. His research interests mainly include digital video compression and communication, multiview video coding, image-based rendering, and image processing.

ANQING HE received the B.Sc. degree from Huadong JiaoTong University, in 1990. He is currently a Professor of engineering with CRRC and the Chief Engineer of Zhejiang CRRC Electric Vehicle Company, Ltd. His research interests mainly include electric vehicle, drive motor, and motor controller.