electronics

MDPI

Article

Intelligent Recognition of Smoking and Calling Behaviors for Safety Surveillance

Jingyuan Zhang [1,2], Lunsheng Wei [1,3], Bin Chen [1,3,*], Heping Chen [1,2] and Wangming Xu [1,2,3,*]

[1] School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China; bada83@163.com (J.Z.); sheng923548906@163.com (L.W.); chenheping@wust.edu.cn (H.C.)
[2] Engineering Research Center for Metallurgical Automation and Detecting Technology of Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China
[3] Institute of Robotics and Intelligent Systems, Wuhan University of Science and Technology, Wuhan 430081, China
* Correspondence: chenbin@wust.edu.cn (B.C.); xuwangming@wust.edu.cn (W.X.)

Abstract: Smoking and calling are two typical behaviors involved in public and industrial safety that usually need to be strictly monitored and even prohibited on many occasions. To resolve the problems of missed detection and false detection in the existing traditional and deep-learning-based behavior-recognition methods, an intelligent recognition method using a multi-task YOLOv4 (MT-YOLOv4) network combined with behavioral priors is proposed. The original YOLOv4 is taken as the baseline network to be improved in the proposed method. Firstly, a K-means++ algorithm is used to re-cluster and optimize the anchor boxes, which are a set of predefined bounding boxes to capture the scale and aspect ratio of specific objects. Then, the network is divided into two branches with the same blocks but independent tasks after the shared feature extraction layer of CSPDarknet-53, i.e., the behavior-detection branch and the object-detection branch, which predict the behaviors and their related objects respectively from the input image or video frame. Finally, according to the preliminary predicted results of the two branches, comprehensive reasoning rules are established to obtain the final behavior-recognition result. A dataset on smoking and calling detection is constructed for training and testing, and the experimental results indicate that the proposed method has a 6.2% improvement in recall and a 2.4% improvement in F1 score at the cost of a slight loss in precision compared to the baseline method; the proposed method achieved the best performance among the compared methods. It can be deployed to related security surveillance systems for unsafe-behavior monitoring and early-warning management in practical scenarios.

Keywords: multi-task learning; object detection; behavior recognition; behavioral prior knowledge; safety surveillance

Citation: Zhang, J.; Wei, L.; Chen, B.; Chen, H.; Xu, W. Intelligent Recognition of Smoking and Calling Behaviors for Safety Surveillance. *Electronics* **2023**, *12*, 3225. https://doi.org/10.3390/electronics12153225

Academic Editors: Junchao Zhang, Moran Ju and Xiangyue Zhang

Received: 30 June 2023
Revised: 21 July 2023
Accepted: 22 July 2023
Published: 26 July 2023

1. Introduction

Smoking and calling are two kinds of human behaviors that are strictly monitored and even prohibited in many occasions. In some public and industrial places such as gas stations, laboratories, libraries, factories, etc., signs prohibiting smoking and calling are usually posted in plain sight. Furthermore, smoking or calling while driving can even cause serious traffic accidents. In recent years, with the development of image processing and computer vision technology, it has become possible to automatically recognize such specific behaviors from surveillance videos [1–6]. The problems of delays, omissions, and the high labor costs of playing back the surveillance video for manual judgments can be overcome by intelligent recognition methods. Traditional behavior-recognition methods are usually realized by extracting hand-crafted low-level image features or combinations of image features. Pan et al. [7] combined a Gaussian mixture model and the frame difference method to detect moving objects from the video, and then used the color feature of smoke

to detect smoking behavior. Zhang et al. [8] proposed a method to detect the human hand region based on the combination of CMOG (co-occurrence matrix of oriented gradients) and HOG (histogram of oriented gradient) features, and then established rules to judge the behavior of holding a mobile phone. Wu et al. [9] introduced a color-based ratio histogram analysis method and fused it with GMM (Gaussian mixture model) to propose an algorithm for the automatic detection of smoking behavior. These traditional methods have achieved a better real-time performance, but they rely on the algorithm design to extract effective image features and it is difficult for them to adapt to complex real-world scenes.

With the popularity of deep-learning-based methods, deep convolutional neural networks have been widely used to extract high-level image features automatically and intelligently with a stronger representation capability, which has significantly improved the performance of object detection and recognition in accuracy and generalization ability. Xiong et al. [10] proposed a deep-learning-based method for recognizing driver's calling behavior, which firstly detected and tracked faces in real time to determine candidate regions, and then detected the phone using a deep convolutional neural network to recognize the calling behavior. Yang et al. [11] proposed a machine-vision-based algorithm for identifying dangerous human behaviors in petrochemical scenes and obtained the final judgment results by fusing the pose-estimation algorithm with the object-detection algorithm of YOLOv3 (You Only Look Once) [12]. Mao et al. [13] also used the YOLOv3 network to recognize the smoking and calling behaviors through dynamic face detection and tracking the cropped face parts. Lu et al. [14] fused deformable and dilated residual blocks with Faster R-CNN (region-based convolutional neural networks) [15] to identify smoking and calling behaviors. Ye et al. [16] introduced an attention mechanism into the Xception network to enhance the feature representation capability and achieved a good performance in the recognition of smoking and calling behaviors. Lu et al. [17] combined the heat maps with the color image of the driver's head and hands and then used a method of keypoint detection to improve the accuracy of driver-behavior recognition. These methods have utilized the advantages of deep convolutional neural networks and partially used the relationship between behavior and behavior-related objects such as the face, head, or hand. However, as single-task learning methods, they fail to fully mine and utilize the prior knowledge between the behaviors and behavior-related objects. Therefore, there is still room for improvement.

In recent years, the idea of MTL (multi-task learning) [18–20] based on deep learning has provided a convenient and effective way to combine information from multiple tasks and achieved a better performance. Xie et al. [21] proposed a recurrent convolution multi-task learning model for text classification, which achieved a good performance in text classification. Zhi et al. [22] used multi-task learning to simultaneously learn similarity metrics and classification tasks and effectively improved the accuracy of video classification. Liu et al. [23] proposed a novel multi-task learning architecture that allows end-to-end learning of task-specific feature-level attention and achieved good results on multiple datasets. Zhang et al. [24] designed a multi-task object detector to detect ships in synthetic aperture radar images and demonstrated that it outperformed a single-task learning method of object detection.

In fact, it is not difficult to conclude that smoking or calling behaviors generally occur with certain prior conditions [25]. For example, when someone is smoking (or calling), there are certain position constraints among the face, the hand, and the cigarette (or phone), and these position constraints also have a certain degree of contribution in judging whether the smoking (or calling) behavior happens. Based on this inspiration and in order to fully utilize the behavioral prior knowledge, we propose an intelligent smoking and calling behavior-recognition method in this paper by combining multi-task learning ideas with behavioral priors to improve the current popular object-detection network of YOLOv4 [26], which is helpful to better prevent safety accidents related to these behaviors and to improve the level of safety management.

2. Method

2.1. Behavior-Recognition Procedure

In the proposed intelligent recognition method of smoking and calling behaviors, the input is an image or a frame in a video, and the output is the behavior-recognition result which is a class label: "smoking", "calling", or "normal", indicating whether the smoking or calling behavior happens. The behavior-recognition procedure of the proposed method is shown in Figure 1.

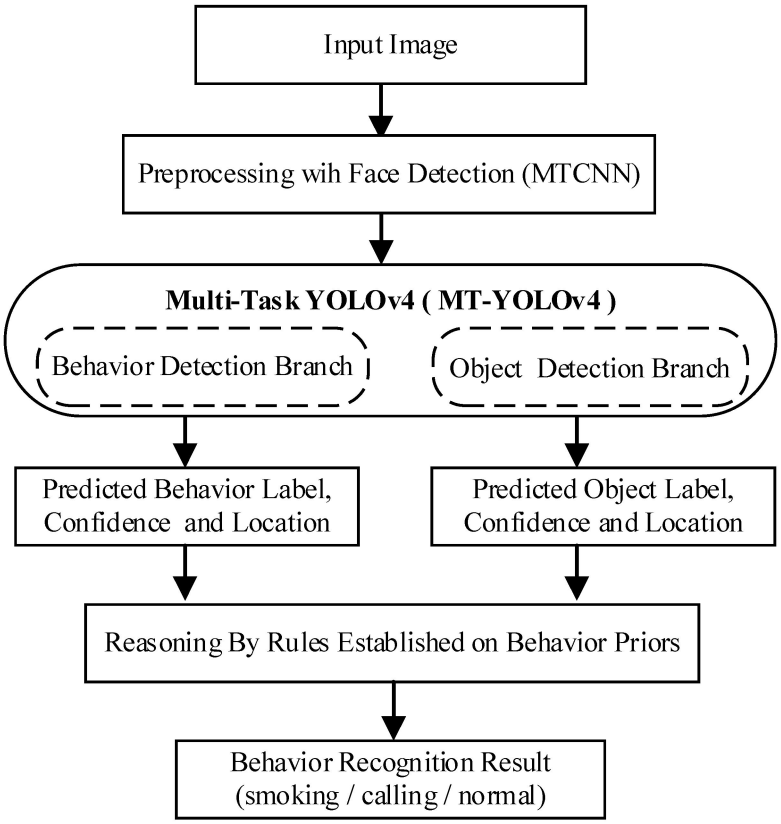

Figure 1. The behavior-recognition procedure of our proposed method.

Considering that the smoking or calling behavior is usually associated with the human face, a face-detection algorithm is used in our proposed method as a preprocessing step, which pre-screens images or video frames containing human faces as the effective image samples and sends them into the subsequent detection network named MT-YOLOv4, i.e., our proposed multi-task YOLOv4 network which is modified from the original YOLOv4 network. If no human face is detected, the output result can be directly set as "normal" which means that no smoking or calling behavior occurs.

For the effective input image preprocessed by face detection in which a human face is detected, our improved detection network MT-YOLOv4 is designed to utilize the relationship between the candidate behavior and the objects related to the behavior. In contrast to the original YOLOv4 network, MT-YOLOv4 has two independent detection branches after the feature-extraction layer: the behavior-detection branch and the object-detection branch. For a simple design, these two branches have the same blocks as the original YOLOv4 network but perform different detection tasks. The behavior-detection branch is used to predict the label, confidence, and location information of the behavior (smoking, calling, or normal behavior). Note that, in order to use the object network like YOLOv4 to detect these behaviors, we treat the behaviors as "objects", which correspond to larger image regions than those of behavior-related objects such as hands, cigarettes, or mobile phones. The object-detection branch is used to predict the label, confidence, and location information of

the behavior-related objects (human hand, cigarette, or mobile phone), which is same as the original YOLOv4 network. Then, with the preliminary prediction results of the two detection branches, the prior knowledge about the smoking or calling behavior is analyzed, and a comprehensive set of reasoning rules based on the behavioral priors is established to further determine whether the behavior happens.

Since the principle of the behavior-detection branch is the same as that of the object-detection branch, the current popular object-detection networks (such as SSD [27], YOLOv3 [12], YOLOv4 [26,28]) can all be used. In this paper, the original YOLOv4 network is adopted as the baseline model, which has the advantages of a small model size, low deployment cost, high flexibility, fast training speed, and short inference time. In addition, MTCNN [29] is used as the face-detection algorithm due to its high speed and good performance. In addition, for the object-detection task using deep-learning neural networks, the network ideally returns valid objects in a timely manner regardless of the scale of the objects, but the use of anchor boxes enables a network to detect multiple objects, objects of different scales, and overlapping objects, and can improve the detection speed and efficiency. Anchor boxes are a set of predefined bounding boxes of a certain height and width, which are defined to capture the scale and aspect ratio of specific object classes that we want to detect and are typically chosen based on object sizes in our training datasets. Therefore, different training datasets should have different anchor boxes suitable to each dataset. In this paper, the anchor boxes are re-generated and optimized by the K-means++ [30] algorithm to improve the performance of behavior and object detection.

2.2. Algorithm Principle of the Proposed MT-YOLOv4

The original YOLOv4 network was proposed in June 2020, representing a single-stage object-detection network like other YOLO networks. Using the ideas of ResNet [31] and CSPNet [32], the YOLOv4 network is composed of CSPDarknet53, SPP (spatial pyramid pooling) [33], and PANet (path aggregation network) [34] in the structure. In the YOLOv4 network, certain shortcut links between layers are set to solve the difficult problem of optimizing the model when the network is deep, as well as to reduce the calculation and memory cost. In addition, the strategy of SAT (self-adversarial training) is used to reduce the risk of overfitting and to improve the generalization ability.

In this paper, a multi-task YOLOv4 (MT-YOLOv4) network is designed based on the idea of MTL (multi-task learning) so as to make effective use of behavioral priors, with different handling of the behavior-detection task (e.g., detecting behavior of smoking or calling) and the behavior-related object detection task (e.g., detecting object of hand, cigarette or mobile phone). The network structure of MT-YOLOv4 is illustrated in Figure 2.

From Figure 2, it can be seen that the MT-YOLOv4 network consists of 3 parts: (1) CSPDarknet-53, for extracting the features of the input image; (2) the behavior-detection branch, for detecting specific behaviors, such as smoking, calling and normal behavior; and (3) the object-detection branch, for detecting objects related to behaviors, such as the human hand, cigarette, and mobile phone. It can also be seen that CSPDarknet-53 and the object-detection branch constitute the original YOLOv4 network. Compared with the original YOLOv4 network, our proposed MT-YOLOv4 network adds a behavior-detection branch which is the same as the object-detection branch in the structure, so that the network can detect behaviors and the related objects simultaneously and independently. The CBL block consists of layers of convolution, batch normalization (BN), and a leaky ReLU activation function, while the CBM block consists of layers of convolution, batch normalization, and a Mish activation function. The RES unit block consists of two CBM blocks and shortcut links, while the CSPX block consists of CBM blocks and RES unit blocks. SPP uses a max-pooling of 1×1, 5×5, 9×9, 13×13 for multi-scale fusion. The input of the MT-YOLOv4 network is a three-channel color image with size of 416×416 pixels, represented as a tensor of $416 \times 416 \times 3$. The outputs are 3 different scales of detection results and each scale is assigned 3 anchor boxes with different sizes. Therefore, every cell in the output will predict 3 bounding boxes and each bounding box corresponds to 5 predicted values, i.e., horizontal

coordinate, vertical coordinate, height, width and confidence. Thus, the dimensions of the output tensor are $S \times S \times 3 \times (5 + C)$, where $S \times S$ is the output scale (i.e., 13×13, 26×26, 52×52) and C is the number of categories which is equal to 3 in this paper.

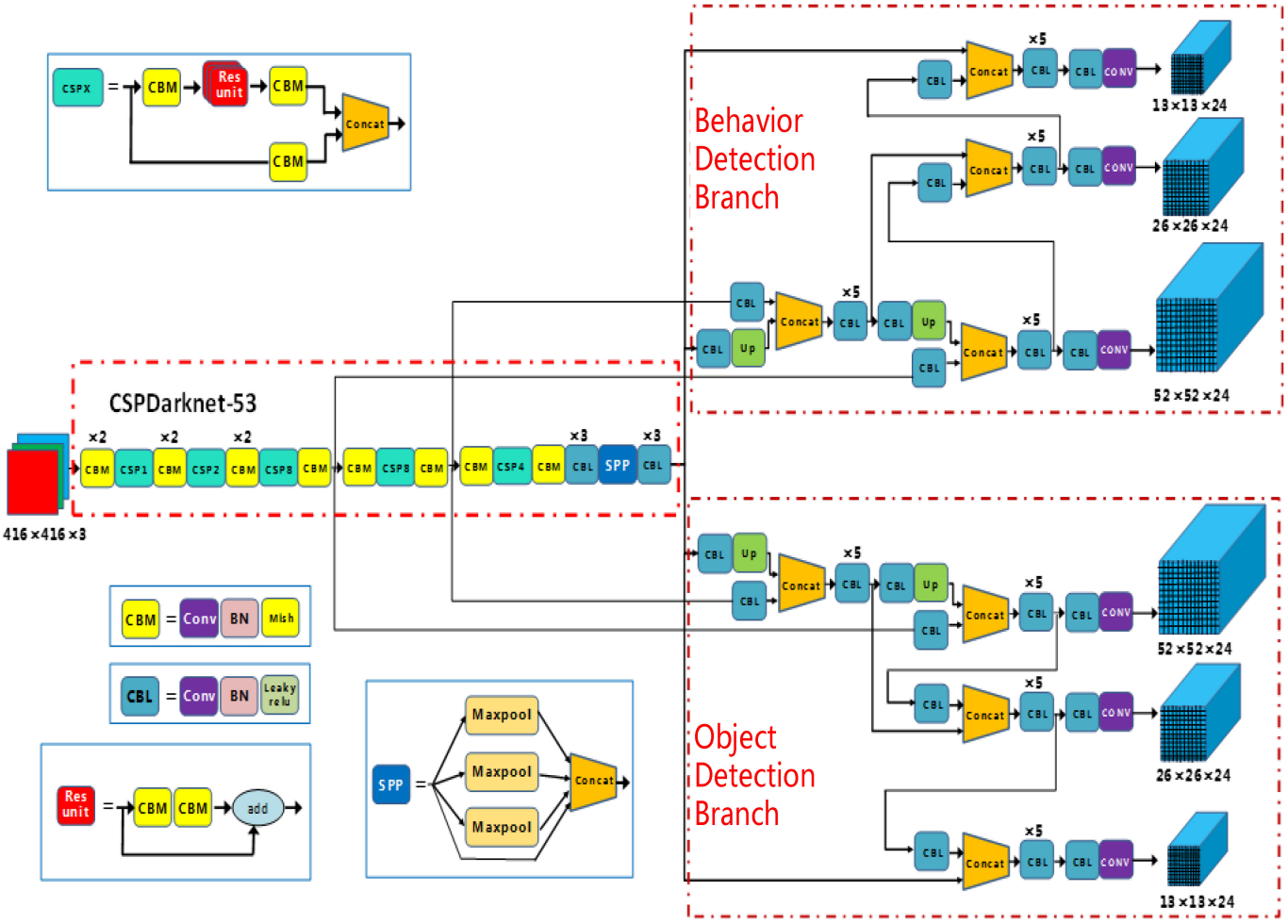

Figure 2. The network structure of MT-YOLOv4.

3. Model Training and Resulting Reasoning

3.1. Dataset Construction

No public dataset specifically for recognizing smoking and calling behaviors is available; thus, we construct a new dataset for training the MT-YOLOv4 network. Firstly, we collect many videos containing smoking or calling behaviors of different people, indoors and outdoors at different places and times, which increases the diversity of the data. In addition, the changes in viewpoint and human body posture are taken into consideration. Then, the MTCNN face-detection algorithm is used for preprocessing every several video frame to screen for the effective sample images containing human faces, and the coordinates of the face regions are recorded at the same time. Finally, we label all the collected effective sample images using LabelMe, a free graphical annotation tool. When labeling a behavior in the effective sample image, the region that covers the face and behavior-related objects is selected, and the coordinates of the bounding box and the behavior label, namely "smoking", "calling", or "normal", are recorded. When labeling the behavior-related objects, we only need to mark the regions of the hands, the cigarettes, or the mobile phones and record their class labels and bounding box coordinates. Since the face region is already detected and recorded by the MTCNN during the preprocessing step, it is unnecessary to label the faces.

As a result, a total of 10,461 frames are pre-screened from the collected videos as the training set. The numbers of the labeled behavior samples are 4516 (smoking), 3968 (calling), and 1977 (normal), respectively; the numbers of the labeled behavior-related-object samples are 9861 (hand), 6197 (cigarette), and 4873 (phone), respectively. In addition, data augmentation strategies are adopted to further increase the number and complexity of the samples, such as image scaling, image flipping, random adjustment of brightness, adding noise, and mosaic, etc.

For performance evaluation, another 1167 testing images are collected from different videos and webpages, which are also labeled and used as the test set.

It should be noted that for safety considerations, it is difficult to collect actual sample images from public places such as gas stations and libraries, as well as industrial environments such as petrochemical production and coal mining. However, experimental results have shown that the model trained from the dataset constructed by our method can accurately recognize smoking and calling behaviors in images collected in various real-world scenarios without using the strategy of transfer learning.

3.2. Re-Cluster Anchor Boxes by K-Means++

To adapt our proposed model to the new dataset, the K-means++ algorithm is adopted to re-cluster and optimize all of the annotated boxes in the training set and assign three anchor boxes for each scale of the output. Since the network has outputs of three scales, the number of clustering centers is set to nine. Unlike the original K-means algorithm that randomly selects K points as the initial clustering centers, the optimization strategy of the K-means++ algorithm is to make the distances between the K initial clustering centers as far as possible. Therefore, the resultant clustering centers are not affected by the random initialization of the cluster centers, which ensures that the obtained anchor boxes are more relevant to the data themselves. The re-clustering steps are as follows:

Step 1: Randomly select a sample from the dataset as the first initial clustering center.

Step 2: Calculate the distance between each sample and the current clustering center and calculate the probability of the sample becoming the next initial clustering center. Select the sample with the highest probability as the next initial clustering center.

Step 3: Repeat the previous step until K initial clustering centers are selected, and then execute the standard K-means algorithm.

For comparison, Table 1 lists the sizes of the original anchor boxes of YOLOv4 and the optimized anchor boxes obtained by using the K-means++ re-clustering algorithm.

Table 1. Comparison of the size of the original and the re-clustered anchor boxes.

Output Scale	Receptive Field	Original Anchor Box Size	Re-Clustered Anchor Box Size
13 × 13	Large	459 × 401	338 × 217
		192 × 243	314 × 180
		142 × 110	271 × 141
26 × 26	Medium	72 × 146	128 × 65
		76 × 55	91 × 61
		36 × 75	64 × 60
52 × 52	Small	40 × 28	47 × 14
		19 × 36	33 × 52
		12 × 16	20 × 13

As shown in Table 1, the differences between the re-clustered anchor boxes and the original anchor boxes are significant, no matter the output scale or receptive field. The re-clustered anchor boxes are more suitable to capture the scale and aspect ratio of the specific object classes we want to detect because they are trained using the samples in our dataset. The performance comparison will be further analyzed in the experimental section.

3.3. Loss Function for Network Training

Similar to the original YOLOv4 algorithm, the total loss function (*Loss*) used for training the MT-YOLOv4 model includes the bounding box position loss (L_{CIoU}), the confidence loss ($L_{\text{confidence}}$), and the classification loss (L_{class}). It can be represented by the following formula:

$$Loss = L_{CIoU} + L_{\text{confidence}} + L_{\text{class}} \tag{1}$$

$$L_{\text{CIou}} = 1 - IoU + \frac{d^2}{c^2} + \alpha\nu \tag{2}$$

$$L_{\text{confidence}} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} K[-\log(p) + BCE(\hat{n}, n)] \tag{3}$$

$$L_{\text{class}} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{\text{noobj}} \left[-\log(1 - p_c)\right] \tag{4}$$

In above formula, IoU is the ratio of the intersection over the union between the predicted bounding box and the ground-truth bounding box, c and d are the center distance and the diagonal distance of the two bounding boxes, respectively. $\nu = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$, $\alpha = \frac{\nu}{(1-IoU)+\nu}$. w^{gt} and h^{gt} are the width and height of the ground-truth bounding box, and w and h are the width and the height of the predicted bounding box. S is the number of grids, B is the number of anchor boxes for each grid, and $K = 1_{i,j}^{obj}$ represents the weight. If there is an object in the j-th anchor box of the i-th grid, its value is 1; otherwise, it is 0. The cross-entropy loss is represented as $BCE(\hat{n}, n) = -\hat{n}\log(n) - (1 - \hat{n})\log(1 - n)$, where $\hat{n}$ and n are the ground-truth class label and the predicted class label for the j-th anchor box in the i-th grid, respectively. p represents the probability that the detection result corresponds to the correct category label.

3.4. Reasoning by Combining the Predicted Results of MT-YOLOv4 and the Behavioral Priors

The variability and complexity of behavioral performance make it difficult to design a unified standard method for labeling the training sample images of behaviors, because behavioral performance is susceptible to many subjective factors. In addition, there are certain types of interference in complex image backgrounds. Therefore, the preliminary predicted results of the behavior-detection branch of the MT-YOLOv4 network will have the problems of missed detection or false detection.

Therefore, the proposed method in this paper combines the behavior-branch prediction results with the prior knowledge of the behavior and establishes the reasoning rules to obtain the final result. Firstly, the face region is detected using the MTCNN face-detection algorithm, and the behavior bounding box predicted by the MT-YOLOv4 network should cover the face region. Secondly, there are constraints on the positions of the face, hand, and object (cigarette or mobile phone) when someone is smoking or calling.

We use $Dist(face, cigarette)$, $Dist(face, phone)$, and $Dist(face, hand)$ to represent the distance between the face and the object of the cigarette, mobile phone, and human hand, where the distance is defined as the distance between the center points of the bounding boxes. At the same time, we take the length of the face bounding box, $Len(face)$, as the reference distance to adapt to the scale change in the actual images. For a candidate behavior of smoking or calling, the following three rules about confidence in the occurrence of the behavior are established:

(1) When $Dist(face, cigarette) \leq a' \cdot Len(face)$, the confidence increases by p_1;
(2) When $Dist(face, phone) \leq b' \cdot Len(face)$, the confidence increases by p_2;
(3) When $Dist(face, hand) \leq c' \cdot Len(face)$, the confidence increases by p_3.

In above rules, p_1, p_2, p_3 can be manually set to a value ranging from 0 to 1 according to the contribution of the rule to the judgment of the behavior, a', b', c' can be obtained through a statistical analysis as follows.

For each training sample image, we first calculate the following three ratios based on its labeling information and detected face information:

$$a = \frac{Dist(face, cigarette)}{Len(face)}, b = \frac{Dist(face, phone)}{Len(face)}, c = \frac{\text{Dist}(face, hand)}{Len(face)} \tag{5}$$

The original ratio values of a, b, c cannot be directly used to calculate their means and variances because certain outliers need to be filtered out. However, the outliers are inherently irregular and difficult to find by statistical methods such as means and variances. Therefore, the distribution map of these ratio values is taken as an effective and intuitive tool to discriminate the outliers and the inliers. The distribution of the ratio values of all samples is shown in Figure 3.

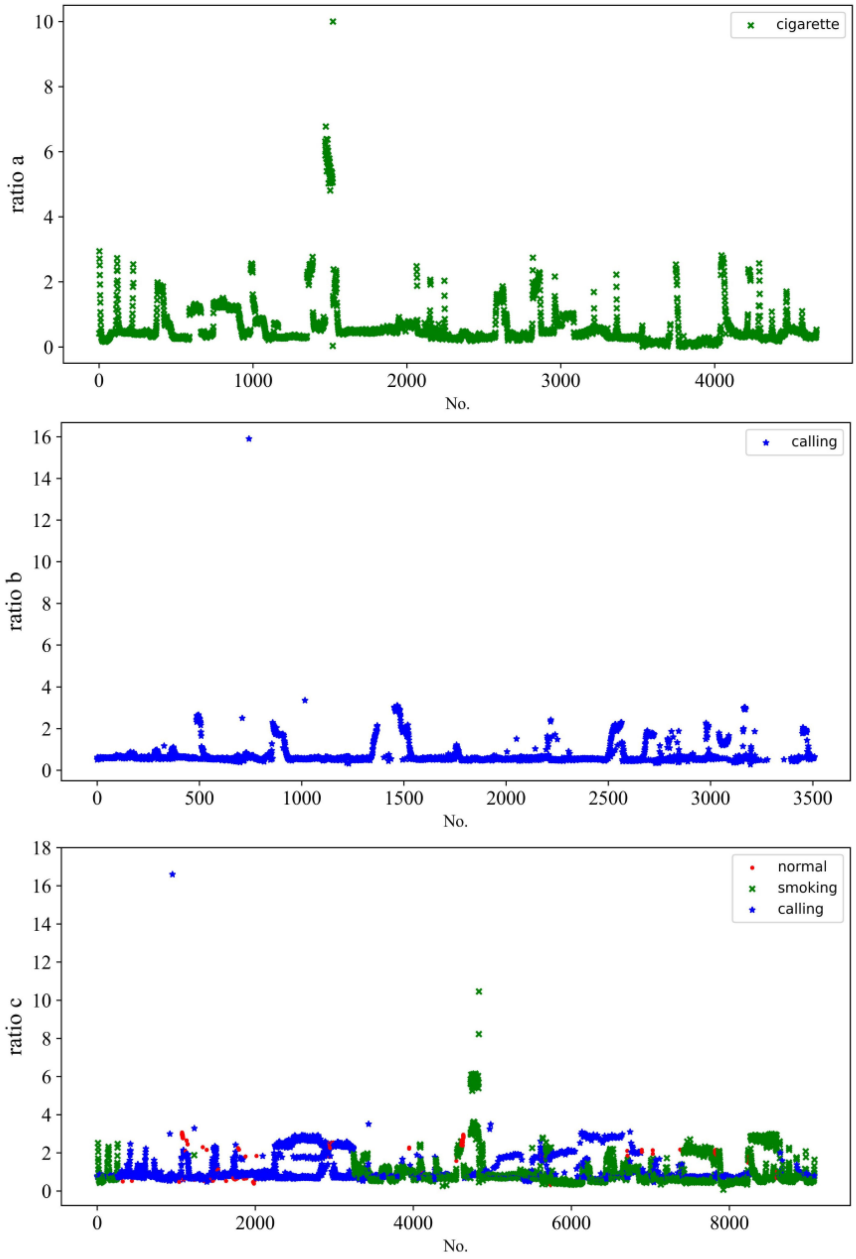

Figure 3. The distribution of the ratio values a, b and c.

From Figure 3, it can be seen that the ratio value corresponding to a certain behavior has shown a certain regularity in its distribution. A few points farther from the majority can be regarded as outliers due to the diversity of the training sample images. It is obvious that we can eliminate the effects of these outliers by excluding such points with a ratio value greater than four. The choice of four as the threshold is based on what is shown in Figure 3. Note that, this threshold does not need to be too precise, and it can be seen that a smaller one (e.g., 3.7) or a larger one (e.g., 4.3) has little effect according to the point distribution in Figure 3. Thus, the means and variances of the inliers can be calculated by the following equations:

$$\overline{x_a} = \frac{\sum_{i=1}^{n_a} x_{ai}}{n_a}, \overline{x_b} = \frac{\sum_{i=1}^{n_b} x_{bi}}{n_b}, \overline{x_c} = \frac{\sum_{i=1}^{n_c} x_{ci}}{n_c} (x_{ai}, x_{bi}, x_{ci} \leq 4) \tag{6}$$

$$\sigma_a{}^2 = \frac{\sum_{i=1}^{n_a} (\overline{x_a} - x_{ai})^2}{n_a}, \sigma_b{}^2 = \frac{\sum_{i=1}^{n_b} (\overline{x_b} - x_{bi})^2}{n_b}, \sigma_c{}^2 = \frac{\sum_{i=1}^{n_c} (\overline{x_c} - x_{ci})^2}{n_c} (x_{ai}, x_{bi}, x_{ci} \leq 4) \tag{7}$$

According to the central limit theorem, suppose that random variables $\{X_n\}$ are independently equally distributed, and the mathematical expectation and variance are finite values: $E(X) = \mu, D(X) = \sigma^2 > 0$. When n is large enough, its mean value approximately follows a normal distribution:

$$\lim_{n\to\infty} P(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < a) = \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{8}$$

According to the frequency distribution law in the normal distribution, a', b', c' can be set as:

$$a' = \overline{x_a} + 3\sigma_a, b' = \overline{x_b} + 3\sigma_b, c' = \overline{x_c} + 3\sigma_c \tag{9}$$

Thus, we can guarantee the validity of the distance priori information in the labeled training samples with a high probability. Calculated with the specific data, we obtain $a' = 1.3$, $b' = 1.4$, and $c' = 2.4$. If the training samples are changed, these parameters can be re-calculated in the same way.

When judging whether a certain behavior happens in an input image, we produce a comprehensive reasoning and analysis by combining with the preliminary predicted results of the MT-YOLOv4 network and the prior knowledge of the behavior. First, the confidence of a behavior p_0 is obtained according to the results predicted by the behavior-detection branch of MT-YOLOv4. Then, with the results predicted by the object-detection branch of MT-YOLOv4 and according to the above three rules, the increase in the confidence of the behavior corresponding to one rule can be determined:

$$\Delta p_i = \begin{cases} p_i, \ if \ Rule \ i \ is \ satisfied \\ 0, \ otherwise \end{cases} \tag{10}$$

Therefore, the summary increase in confidence corresponding to the three rules is $\sum_{i=1}^{3} \Delta p_i$, and the final confidence of the behavior is:

$$p = \min\left(p_0 + \sum_{i=1}^{3} \Delta p_i, 1\right) \tag{11}$$

In Equation (11), $\min(\cdot)$ represents the operation to find the minimum which ensures that the value of the confidence cannot be larger than one. According to the degree of

contribution for the above-mentioned three rules to judge whether a smoking or calling behavior happens, we set the parameters as $p_1 = 0.5$, $p_2 = 0.4$ and $p_3 = 0.1$.

Letting T be the confidence threshold that indicates the occurrence of the behavior, the rule for exactly determining whether the behavior occurs are as follows:

$$\begin{cases} behavior\ occurs,\ if\ p > T \\ behavior\ doesnotoccur,\ otherwise \end{cases} \tag{12}$$

Note that T is set to 0.8 in this paper to ensure the reliability of the final recognition result.

4. Experiments

4.1. Experimental Setting and Performance Metrics

Experiments are conducted with Keras deep learning framework on the Windows platform. The network models are trained with NVIDIA Quadro GP100 16 GB GPU using CUDA 10.0 and cuDNN 7.6.5. Adam is adopted as the optimizer, and the learning rate is initialized to 0.001 and adjusted dynamically by the method of cosine annealing, which is beneficial to achieve the global optimal solution during the training process.

Firstly, only the last layer of the network is unfrozen and trained for 50 epochs in a method of transfer learning for fine-tuning. Then, all of the layers are unfrozen and trained for another 50 epochs to obtain the final network model. The loss curves of the network training are illustrated in Figure 4, with the training epoch number in the horizontal coordinate and the loss value in the vertical coordinate. The blue solid line and the orange dotted line represent the loss curves for the object-detection branch and the behavior-detection branch, respectively. It can be seen that both of them are converged within 100 epochs, and the best model among those in the 100 epochs is adopted for inference.

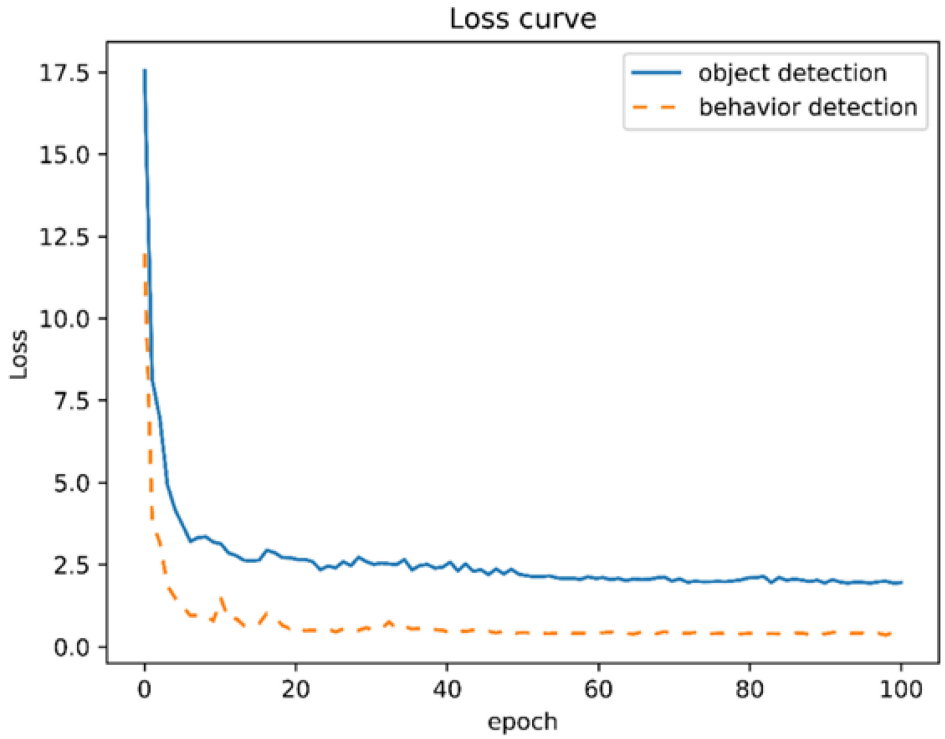

Figure 4. The loss curves of network training.

P (precision), R (recall) and $F1$ (F1 score) are always used as the evaluation criteria for object detection and recognition tasks, which are defined as:

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{15}$$

where TP is the number of positive samples to be correctly predicted, FP is the number of negative samples to be predicted as positive, and FN is the number of positive samples predicted as negative. For the task of behavior recognition, more attention should be paid to recall than precision. In addition, F1 score is the harmonic mean of precision and recall, reflecting the balance between precision and recall. Generally speaking, a higher recall may come at the expense of a degree of precision loss. Therefore, we choose recall and F1 score as the performance metrics.

4.2. Ablation Experiment

This ablation experiment is carried out to verify the effectiveness of the strategies of re-clustering anchor boxes and using behavioral priors for behavior recognition. We take an additional 1167 images as the test set which are labeled in advance and used as a reference, i.e., ground truth for calculating related performance metrics. The results of the ablation experiments are shown in Table 2.

Table 2. The results of ablation experiments.

Using Re-Clustered Anchor Boxes	Using Behavioral Priors	P (%)	R (%)	F1 (%)
No	No	86.9	83.4	85.1
Yes	No	87.9	84.7	86.3
No	Yes	86.5	87.9	87.2
Yes	Yes	85.5	**89.6**	**87.5**

As shown in Table 2, the method of using re-clustered anchor boxes results in an improvement of 1.3% in the recall rate and an improvement of 1.2% in the F1 score compared with the original YOLOv4 method. The reason is that the new anchor boxes re-clustered from our new dataset are more suitable for the behavior-detection task. In addition, the method of using behavioral priors results in an improvement of 4.5% in the recall rate and an improvement of 2.1% in the F1 score compared with the original YOLOv4 method. When using re-clustered anchor boxes, there is a 4.9% improvement in recall and a 1.2% improvement in the F1 score by using behavioral priors. The main reason is that the reasoning rules established by behavioral priors comprehensively utilize the information generated by the two branches of behavior detection and object detection in the MT-YOLOv4 network, rather than relying on the behavior-detection result alone. In total, the proposed method in this paper uses the strategies of re-clustering anchor boxes and using behavioral priors and leads to a great improvement in performance, that is, an improvement of 6.2% in recall and 2.4% in the F1 score.

4.3. Comparative Experiments with Other Deep Networks

In order to show the advantages of the method proposed in this paper, MT-YOLOv4 is compared with other three typical deep-learning-based detection networks: SSD, YOLOv3, and YOLOv4. These detection networks have been applied in some of the existing behavior-recognition methods as described in Refs. [12,13]. The experimental results are illustrated in Table 3.

Table 3. The comparison of behavior-recognition results using different detection networks.

Detection Network	P (%)	R (%)	F1 (%)
SSD	92.1	75.2	82.8
YOLOv3	93.5	76.1	83.9
YOLOv4	86.9	83.4	85.1
MT-YOLOv4 (the proposed method)	85.5	**89.6**	**87.5**

As can be seen from Table 3, the original YOLOv4 network outperforms the SSD and YOLOv3 network, and the proposed method has achieved the highest recall and F1 score. Although the precision rate is slightly reduced compared with the original YOLOv4 network, the recall rate increases from 83.4% to 89.6%, and the F1 score increases from 85.1% to 87.5%. For the behavior-recognition task in the field of security surveillance, recall is a more crucial assessment criterion. Therefore, the proposed method is the most effective among the compared methods. The main reason lies in the fact that the different detection tasks carried out by our MT-YOLOv4 network make beneficial contributions to the behavior-recognition performance by combining the advantages of deep-learning methods and behavioral priors.

4.4. Examples of Behavior-Recognition Results

To intuitively demonstrate the effectiveness of the proposed method, Figure 5 shows the recognition results of two groups of test samples for behavior recognition before and after using behavioral priors for reasoning. Figure 5a,c show the original recognition results without using behavioral priors, which are only based on the predicted results of the behavior-detection branch of the MT-YOLOv4 network. It can be seen that the smoking behavior in Figure 5a is not recognized, and the smoking behavior in Figure 5c is misrecognized as a calling behavior. Correspondingly, Figure 5b,d show the recognition results after using behavioral priors. By utilizing the predicted results of the behavior-detection branch and the object-detection branch of the MT-YOLOv4 network together and combining them with behavioral priors for reasoning, the smoking behaviors in both figures are accurately recognized.

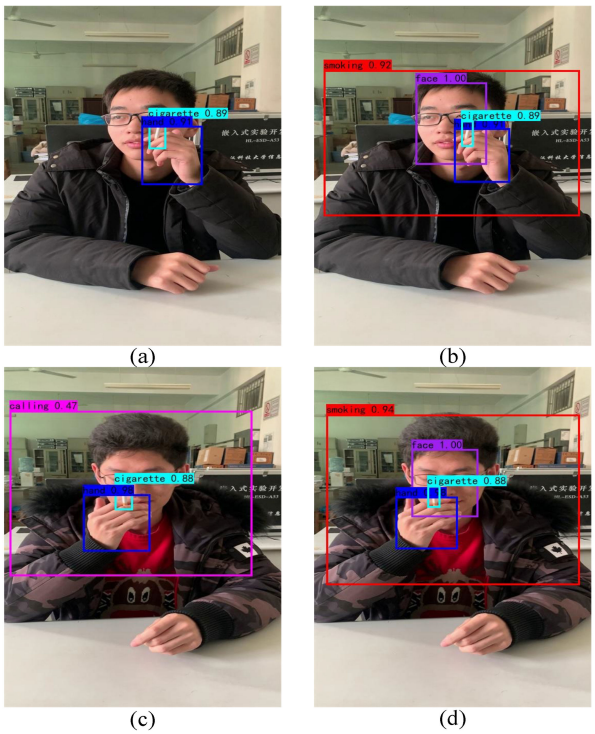

Figure 5. Examples of the comparison of the recognition results before and after using behavioral priors.

To further verify the practicability of the proposed method, certain typical real-world images are collected from webpages for testing. Related real scenarios include gas stations, libraries, coal mines, airplanes, and cars. For the sake of fairness, the collected images are not in our self-built dataset; that is, they have not been used in the model training and the above testing experiments. As shown in Figure 6, our proposed method has accurately recognized the smoking or calling behavior in each scene image, indicating that the proposed method has a good generalizability. Therefore, it can be deployed to safety surveillance systems for monitoring and preventing unsafe human behaviors.

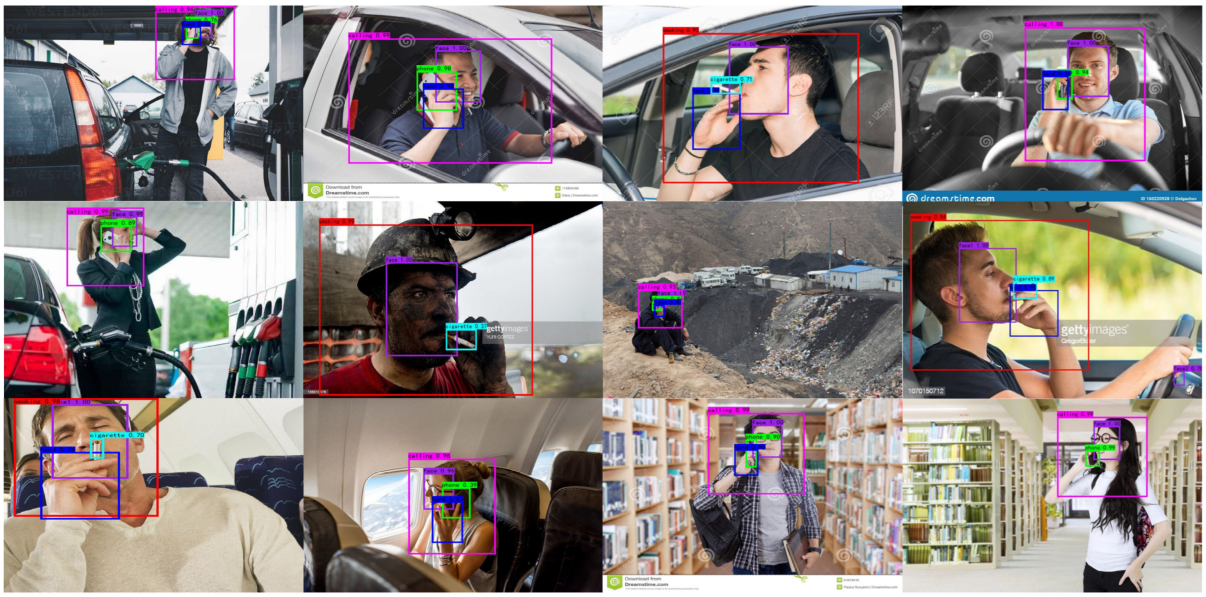

Figure 6. Examples of smoking and calling recognition in practical scenarios. The source images are collected from webpages.

5. Conclusions

To meet the needs of safety surveillance, we propose a new intelligent method to simultaneously recognize smoking and calling behaviors. The existing traditional and deep-learning-based behavior-recognition methods cannot make full use of prior knowledge, resulting in the problems of missed detection and false detection. By combining deep learning and behavioral priors, our method significantly improves the effectiveness of behavior recognition and can be used for safety monitoring and accident prevention in practical scenarios. In order to make full use of the behavioral priors, the original single-task YOLOv4 network is improved into multi-task YOLOv4 (MT-YOLOv4) network. After the CSPDarknet-53 layer, the network is divided into two branches: behavior detection and object detection, with the same blocks but independent tasks. K-means++ is used to re-cluster the anchor boxes, which further improves the performance. In order to make full use of behavioral priors, the reasoning rules suitable for smoking and calling behavior-recognition tasks are established, and a method based on statistical theory for solving required parameters is presented. A new dataset for smoking and calling behavior recognition is constructed which consists of tens of thousands of images for model training and testing. The trained model also has an outstanding performance when applied to images in actual scenes, which shows the practical value of the proposed method.

One limitation of the proposed method is that it depends on a face-detection algorithm as a preprocessing step, so it is not applicable to cases where the face-detection algorithm cannot detect the face due to the large variation in imaging conditions, such as illumination and viewpoint. For example, the face is difficult to detect when photographed from the side or behind. Another limitation is that there are several empirical parameters when

reasoning with the behavioral priors. These parameters are generally set according to the contribution of behavioral priors for the occurrence of smoking or calling.

Therefore, one future research direction is to replace face detection with head detection, because the head-detection task is not affected by the direction of imaging. Image-enhancement methods [35,36] can be used to improve the quality of the input image to achieve a better performance in head detection. In addition, recent state-of-the-art convolutional neural networks such as YOLOv7, YOLOv7v8, and NAS architectures can be considered as an alternative to YOLOv4. Furthermore, we can try to put the prior knowledge of behaviors into the design of loss functions and train a new end-to-end behavior-recognition model which directly combines convolutional neural networks and behavioral priors and can avoid the problem of setting empirical parameters.

Author Contributions: Conceptualization, J.Z., L.W., B.C., H.C. and W.X.; methodology, J.Z., L.W., B.C., H.C. and W.X.; software, L.W.; validation, J.Z., L.W. and W.X.; formal analysis, H.C. and W.X.; investigation, J.Z. and B.C.; resources, J.Z.; data curation, L.W.; writing—original draft preparation, J.Z. and L.W.; writing—review and editing, B.C. and W.X.; visualization, L.W.; supervision, H.C. and W.X.; project administration, B.C. and W.X.; funding acquisition, B.C. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 51805386; the Open Project of Metallurgical Automation and Testing Technology Engineering Research Center of the Ministry of Education, grant number MADTOF2021B02; and the Scientific Research Program of the Hubei Provincial Department of Education, grant number D20191104.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank all anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions, which helped us to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chu, J.; Zhang, S.; Lu, W. A Driving Behavior Analysis Algorithm Based on Convolutional Neural Network. *Prog. Laser Optoelectron.* **2020**, *57*, 180–189.
2. Ji, X.; Teng, B. Detection of Abnormal Escalator Behavior Based on Deep Neural Network. *Prog. Laser Optoelectron.* **2020**, *57*, 140–149.
3. Jebur, S.A.; Hussein, K.A.; Hoomod, H.K.; Alzubaidi, L.; Santamaría, J. Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. *Electronics* **2023**, *12*, 29. [CrossRef]
4. Shi, Y.; Guo, B.; Xu, Y.; Xu, Z.; Huang, J.; Lu, J.; Yao, D. Recognition of Abnormal Human Behavior in Elevators Based on CNN. In Proceedings of the 2021 26th International Conference on Automation and Computing (ICAC), Portsmouth, UK, 2–4 September 2021; pp. 1–6.
5. Vrskova, R.; Hudec, R.; Kamencay, P.; Sykora, P. A New Approach for Abnormal Human Activities Recognition Based on ConvLSTM Architecture. *Sensors* **2022**, *22*, 2946. [CrossRef]
6. Ali, M.A.; Hussain, A.J.; Sadiq, A.T. Deep Learning Algorithms for Human Fighting Action Recognition. *Int. J. Online Biomed. Eng.* **2022**, *18*, 71–87.
7. Pan, G.; Yuan, Q.; Fan, C.; Qiao, H.; Wang, Z. Cigarette-smoke Detection Based on Gaussian Mixture Model and Frame Difference Method. *Comput. Eng. Des.* **2015**, *36*, 1290–1294+1336.
8. Zhang, B.; Wang, W.; Wei, M.; Cheng, B. Detection Handheld Phone Use by Driver Based on Machine Vision. *J. Jilin Univ. (Eng. Technol. Ed.)* **2015**, *45*, 1688–1695.
9. Wu, P.; Hsieh, J.W.; Cheng, J.C.; Cheng, S.C.; Tseng, S.Y. Human smoking event detection using visual interaction clues. In Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; IEEE: Piscataway, NJ, USA; pp. 4344–4347.
10. Xiong, Q.; Lin, J.; Yue, W.; Liu, S.; Luo, X.; Ding, C. A Driver's Call Behavior Detection Method Based on Deep Learning. *Control Inf. Technol.* **2019**, *6*, 53–56+62.
11. Yang, B.; Yun, X.; Dong, K.; Liu, X.; Huang, H. Personnel Dangerous Behavior Recognition in Petrochemical Scene Based on Machine Vision. *Laser Optoelectron. Prog.* **2021**, *58*, 355–365.
12. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.

13. Mao, P.; Zhang, K.; Liang, D. Driver Distraction Behavior Detection Method Based on Deep Learning. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2020; Volume 782, p. 022012.
14. Lu, M.; Hu, Y.; Lu, X. Driver Action Recognition Using Deformable and Dilated Faster R-CNN with Optimized Region Proposals. *Appl. Intell.* **2020**, *50*, 1100–1111. [CrossRef]
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef] [PubMed]
16. Ye, L.; Chen, C.; Wu, M.; Nwobodo, S.; Antwi, A.A.; Muponda, C.N.; Ernest, K.D.; Vedaste, R.S. Using CNN and Channel Attention Mechanism to Identify Driver's Distracted Behavior. In *Transactions on Edutainment XVI*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 175–183.
17. Lu, M.; Hu, Y.; Lu, X. Pose-guided Model for Driving Behavior Recognition Using Keypoint Action Learning. *Signal Process. Image Commun.* **2022**, *100*, 116513. [CrossRef]
18. Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv* **2017**, arXiv:1706.05098.
19. Nadeem, M.I.; Ahmed, K.; Li, D.; Zheng, Z.; Naheed, H.; Muaad, A.Y.; Alqarafi, A.; Abdel Hameed, H. SHO-CNN: A Metaheuristic Optimization of a Convolutional Neural Network for Multi-Label News Classification. *Electronics* **2023**, *12*, 113. [CrossRef]
20. Zhang, W.; Miao, Z.; Xu, W. A Video Anomalous Behavior Detection Method Based on Multi-Task Learning. In Proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 15–17 April 2022; pp. 396–400.
21. Xie, J.; Li, J.; Kang, S.; Wang, Q.; Wang, Y. Multi-Domain Text Classification Method Based on Recurrent Convolution Multi-Task Learning. *J. Electron. Inf.* **2021**, *43*, 2395–2403.
22. Zhi, H.; Yu, H.; Li, S.; Gao, C.; Wang, Y. A Video Classification Method Based on Deep Metric Learning. *J. Electron. Inf.* **2018**, *40*, 2562–2569.
23. Liu, S.; Johns, E.; Davison, A.J. End-to-end Multi-Task Learning with Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1871–1880.
24. Zhang, X.; Huo, C.; Xu, N.; Jiang, H.; Cao, Y.; Ni, L.; Pan, C. Multitask Learning for Ship Detection from Synthetic Aperture Radar Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8048–8062. [CrossRef]
25. Xu, W.; Xu, T.; Li, C.; Wu, S. A Smoking and Calling Detection Method Based on Deep Learning and Behavior Prior. *Comput. Appl. Softw.* **2022**, *39*, 199–204.
26. Bochkovskiy, A.; Wang, C.; Liao, H. Yolov4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
28. Wang, Y.; Hua, C.; Ding, W.; Wu, R. Real-time Detection of Flame and Smoke Using an Improved YOLOv4 Network. *SIViP* **2022**, *16*, 1109–1116. [CrossRef]
29. 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]
30. Arthur, D.; Vassilvitskii, S. K-means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007.
31. 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.
32. Wang, C.; Liao, H.; Wu, Y.; Chen, P.; Hsieh, J.; Yeh, I. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
34. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
35. Zheng, C.; Li, Z.; Yang, Y.; Wu, S. Single image brightening via multi-scale exposure fusion with hybrid learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1425–1435. [CrossRef]
36. Zheng, C.; Jia, W.; Wu, S.; Li, Z. Neural Augmented Exposure Interpolation for Two Large-Exposure-Ratio Images. *IEEE Trans. Consum. Electron.* **2023**, *69*, 87–97. [CrossRef]