sensors

MDPI

Article

Automatic Detection Method for Black Smoke Vehicles Considering Motion Shadows

Han Wang [1], Ke Chen [2,*] and Yanfeng Li [2]

[1] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; ms.h.wang@cumt.edu.cn
[2] College of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China; 212204010020@home.hpu.edu.cn
* Correspondence: 212004010025@home.hpu.edu.cn

Abstract: Various statistical data indicate that mobile source pollutants have become a significant contributor to atmospheric environmental pollution, with vehicle tailpipe emissions being the primary contributor to these mobile source pollutants. The motion shadow generated by motor vehicles bears a visual resemblance to emitted black smoke, making this study primarily focused on the interference of motion shadows in the detection of black smoke vehicles. Initially, the YOLOv5s model is used to locate moving objects, including motor vehicles, motion shadows, and black smoke emissions. The extracted images of these moving objects are then processed using simple linear iterative clustering to obtain superpixel images of the three categories for model training. Finally, these superpixel images are fed into a lightweight MobileNetv3 network to build a black smoke vehicle detection model for recognition and classification. This study breaks away from the traditional approach of "detection first, then removal" to overcome shadow interference and instead employs a "segmentation-classification" approach, ingeniously addressing the coexistence of motion shadows and black smoke emissions. Experimental results show that the Y-MobileNetv3 model, which takes motion shadows into account, achieves an accuracy rate of 95.17%, a 4.73% improvement compared with the N-MobileNetv3 model (which does not consider motion shadows). Moreover, the average single-image inference time is only 7.3 ms. The superpixel segmentation algorithm effectively clusters similar pixels, facilitating the detection of trace amounts of black smoke emissions from motor vehicles. The Y-MobileNetv3 model not only improves the accuracy of black smoke vehicle recognition but also meets the real-time detection requirements.

Keywords: intelligent transportation; motion shadows; superpixel segmentation; YOLOv5s localization; MobilNetv3 classification

Citation: Wang, H.; Chen, K.; Li, Y. Automatic Detection Method for Black Smoke Vehicles Considering Motion Shadows. *Sensors* **2023**, *23*, 8281. https://doi.org/10.3390/s23198281

Academic Editor: Daming Shi

Received: 1 September 2023
Revised: 27 September 2023
Accepted: 4 October 2023
Published: 6 October 2023

1. Introduction

Traditional control of motor vehicle exhaust pollution mainly occurs during processes such as vehicle registration and annual inspections rather than effective supervision during vehicle usage. The application of onboard detection technology and road remote sensing monitoring technology can identify motor vehicles emitting black smoke exhaust on roads. However, the size of detection devices is relatively large, making it difficult to deploy them extensively on urban roads. In recent years, with the rapid development of artificial intelligence, methods for automatically detecting black smoke-emitting vehicles based on monitoring videos from road surveillance cameras have become more intelligent and efficient. Cao et al. [1] utilized the Inceptionv3 convolutional neural network to capture spatial information of suspected black smoke frames in monitoring videos, while a long short-term memory network learned the temporal dependencies between video frames. They built a dual-branch black smoke vehicle detection network based on the CenterNet [2] framework, utilizing vehicle feature maps to generate attention mechanisms for guiding the

training of black smoke feature maps. This model achieved a detection speed of 25.46 FPS and mAP@0.5 of 92.5%. Xia et al. [3] proposed using a convolutional neural network model based on LeNet-5 to detect vehicles emitting black smoke. Simultaneously, an Inception module was introduced, and multiple convolutional kernels of different sizes were used to perform convolution operations to extract black smoke features. Zhang et al. [4] proposed a multi-frame classification network based on 2D-3D fusion for detecting black smoke-emitting vehicles. They utilized both 2D and 3D convolutions to extract spatial and spatiotemporal features of black smoke. The model achieved a recognition accuracy of 90.3%, with an average inference time of 45.9 ms per frame. Zhang et al. [5] designed two lightweight networks, YOLOv3-M3-CBAM and YOLOv4-GhostNet, based on the YOLOv3 and YOLOv4 models. After improvement, both models achieved a detection speed of 20 FPS. Liu and others proposed a black smoke vehicle detection model based on a three-dimensional convolutional network and a non-local attention mechanism. This model utilizes three-dimensional convolutional kernels to learn the spatial features and temporal information of black smoke videos. It jointly evaluates the existence of black smoke by considering suspected black smoke regions across multiple consecutive frames [6].

The aforementioned automatic detection methods for vehicles emitting black smoke primarily focus on improving and optimizing model structures based on the target features of black smoke emissions. However, factors that interfere with black smoke vehicle detection in real-world scenarios have not been taken into consideration. For instance, when vehicles are driving under clear weather conditions, they cast dynamic shadows. These dynamic shadows exhibit certain visual similarities to black smoke emissions, which significantly affect the recognition accuracy of black smoke vehicle detection. In areas where shadows are cast and exhibit high brightness and saturation, their color values closely follow a linear relationship with the background image. This principle can be employed for shadow detection, where the brightness in shadow areas is lower than that in non-shadow areas, while chromaticity remains consistent [7]. Khan et al. [8] employed multiple supervised convolutional deep neural networks to learn shadow-related features. However, due to a lack of labeled training data, this approach remains challenging in practical application scenarios. Tian et al. proposed a normalized cross-correlation method based on texture features, which involves calculating the NCC value by comparing the texture similarity between the current frame and the background pixels at the same position and their neighboring pixels for shadow judgment [9]. Shadow removal involves restoring shadow regions in an image while preserving attributes such as texture and color on the object's surface. Shadow binary masks and shadow masks are commonly used for conditional information for generators in generative adversarial networks. Shadow binary masks often utilize alpha matting techniques to label shadow and non-shadow regions, but shadow masks can be easily influenced by human errors [10,11]. The challenge in shadow detection lies in accurately identifying the shadowed areas on object surfaces, while the challenge in shadow removal is to protect object surface information from being altered. However, due to the certain similarity between black smoke emissions and dynamic shadows, the solution of detecting and then removing dynamic shadows is difficult to implement in the task of automatic detection of vehicles emitting black smoke.

The existing intelligent algorithm for detecting smoky vehicles faces several challenges, including difficulties in model deployment, limited model applicability, and the need to improve accuracy in smoky vehicle identification. The large number of model parameters and computational requirements make model deployment challenging, necessitating the development of a more lightweight smoky vehicle detection network. The limited model applicability and low recognition accuracy are due to the fact that existing methods have not adequately considered factors that interfere with the smoky vehicle detection process during optimization and improvement, such as the motion shadows produced by motor vehicles on sunny days. Therefore, this study has designed an automatic smoky vehicle detection solution that takes into account motion shadows, as shown in Figure 1. Based on the "segmentation-classification" concept, it cleverly addresses situations where

motion shadows coexist with smoky exhaust, and it achieves this by using a superpixel segmentation algorithm called simple linear iterative clustering to cluster and re-segment similar pixels in the image [12]. Directly detecting smoky exhaust using YOLO series object detection models faces challenges such as missing small targets, misidentifying motion shadows, and difficulty in associating detected smoky exhaust with motor vehicles in high-traffic areas [13]. However, by locating moving objects that include motor vehicles, smoky exhaust, and motion shadows, the target positioning effect is superior to traditional motion object detection methods. This approach can exclude irrelevant moving objects, such as roadside trees, that are not related to the research being conducted. The images of moving objects are processed using the superpixel segmentation algorithm to obtain superpixel images belonging to three categories: motor vehicles, smoky exhaust, and motion shadows, which serve as training samples. The design of a lightweight network structure, compared with convolutional neural networks, is more suitable for real-time detection tasks. Therefore, the obtained segmented samples of different categories are fed into the smoky vehicle automatic detection model built on the lightweight MobileNetv3 network [14–16] for recognition and classification. In the task of automatic smoky vehicle detection, not only accurate identification of smoky vehicles is required, but also the network inference speed needs to be improved, especially when dealing with a large amount of surveillance video data.

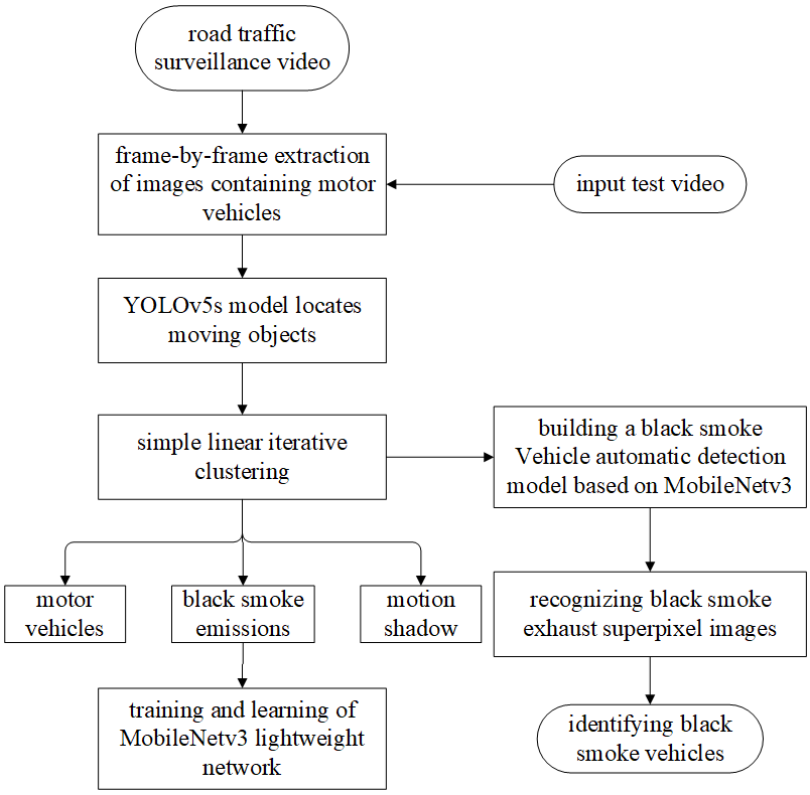

Figure 1. The flowchart for smoky vehicle detection considering motion shadows.

2. Locating Moving Objects

2.1. Object Detection Model

Object detection, as a fundamental problem in computer vision research, involves precisely locating all objects of given classes in an image and predicting the class for each object. The traditional object detection process can be roughly divided into three steps: candidate box generation, feature vector extraction, and region classification. Deep learning-based object detection methods allow for end-to-end learning, eliminating the need for staged training during the process. These methods include two-stage detection algorithms based on candidate windows and single-stage detection algorithms based on regression.

Single-stage detection algorithms do not require generating candidate regions and can directly predict the class probabilities and location information of objects. The YOLO series of algorithms improve accuracy through end-to-end training, and they are compatible and suitable for industrial applications [17,18]. In 2020, YOLOv5 was introduced, followed by the YOLOX model proposed by Megvii in the following year. In 2023, Ultralytics continued to upgrade and optimize the previously introduced YOLOv5 model and released the YOLOv8 model. The performance comparison of these three different models is shown in Table 1. The YOLOXs model and the YOLOv5s model both use Focus and CSPDarknet53 as the backbone networks, and the neck network adopts the FPN + PAN structure. Activation functions include LeakyReLU and Sigmoid, with LeakyReLU used in the hidden layers and Sigmoid used in the detection layers. The YOLOXs model uses a free anchor box strategy for the prediction layer, while the YOLOv5 model learns anchor boxes automatically from the training dataset, reducing the original three anchor box candidates to one and directly predicting the four parameters for each target box [19]. The main feature of YOLOv8 is its scalability, which can be applied not only to YOLO series models but also to non-YOLO models and tasks such as segmentation, classification, and pose estimation. There have been significant improvements in the neck part of the network, where all C3 modules have been replaced with C2f modules, and all CBS modules before upsampling have been removed, with upsampling operations directly performed using C2f modules [20]. YOLOv5 uses a simple convolutional neural network architecture, while YOLOv8 employs multiple residual units and branches and is more complex. Table 1 presents the test results comparison of different object detection models on our custom dataset in this study. YOLOv5 has a smaller parameter count, faster inference speed, and is more suitable for real-time motor vehicle detection.

Table 1. Comparison of different models in the YOLO series.

Model	Batch_Size	mAP (%)	Params (M)	FLOPs (G)
YOLOv5s	256	95.8	7.2	16.5
YOLOXs	128	95.2	9.0	26.8
YOLOv8s	128	94.5	11.2	28.6

The overall structure of the YOLOv5s model consists of an input layer, backbone network, neck network, and prediction layer; as shown in Figure 2. Image preprocessing includes mosaic data augmentation, adaptive image scaling, and adaptive anchor boxes. Mosaic data augmentation involves combining four images through random cropping, flipping, and other methods. This enhances the network's robustness and addresses issues of insufficient dataset samples and uneven size distribution [21]. The backbone feature extraction network is primarily composed of Conv modules, C3 modules, and SPPF modules. In version 6.0, the previous version's focus module has been replaced with a convolutional layer with a kernel size of 6, stride of 2, and padding of 2. For GPUs with limited performance, using a convolutional layer in this context is more efficient than using the focus module. While earlier versions used the CSP module to reduce model computation and achieve cross-layer fusion of local image features, version 6.0 employs the C3 module with a similar role. The difference lies in the removal of the Conv after concatenation, and the standard convolution module after Concat has replaced the Relu activation function with SiLU. In version 6.0, the SPP module is replaced with the SPPF module, both of which aim to fuse output features and enlarge the object receptive field [22,23]. The neck network combines a feature pyramid network with a path aggregation network to reprocess features extracted at different stages. The feature pyramid network transfers strong semantic information from deep feature maps to shallow ones through upsampling, while the path aggregation network transfers positional information from shallow feature maps to deep ones through downsampling. This simultaneous upsampling and downsampling achieves multi-scale feature fusion [24,25]. The prediction layer is responsible for detecting the class and position of target objects. It mainly consists of the loss function and non-maximum sup-

pression. The loss function is the sum of localization loss, confidence loss, and classification loss. Non-maximum suppression is employed to eliminate redundant bounding boxes.

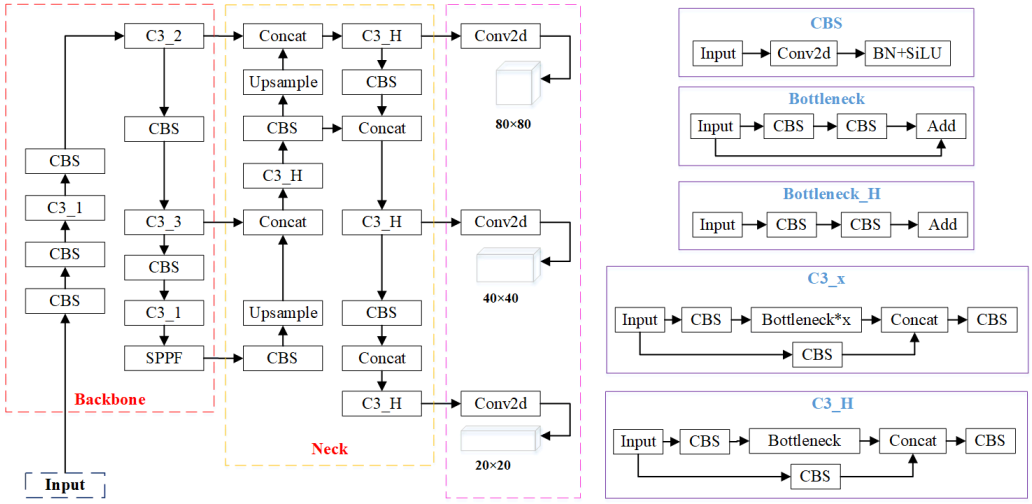

Figure 2. Overall architecture of YOLOv5s model.

2.2. Motion Object Extraction

Motion object detection is a crucial component of intelligent video surveillance systems. Currently, mainstream methods for motion object detection include optical flow, frame differencing, and background subtraction [26–28]. Background subtraction involves comparing the current image with a background image. This method can adapt to changes in application scenarios and handle noise disturbances to some extent [29,30]. Frame differencing is simple to implement, has low computational requirements, and exhibits strong adaptability and robustness in dynamic environments. However, in the presence of large areas of similar grayscale values on the surface of the moving object, frame differencing may result in holes in the image [31,32]. In recent years, deep learning technology has shown its remarkable feature extraction capabilities. Object detection algorithms can locate motion objects, thereby predefining the scope of study and reducing the interference of influencing factors. Two-stage object detection algorithms have slow processing speeds, making them inadequate for real-time detection tasks. On the other hand, the YOLO series of one-stage object detection algorithms can significantly improve detection speed while sacrificing only a slight decrease in accuracy. Thus, this study chooses the YOLOv5 model, which excels in object detection performance, to locate the regions of moving objects in road traffic surveillance videos. Based on network depth and width, the model is available in four sizes: small, medium, large, and extra-large. In practical applications, there is a need to balance the relationship between model accuracy, speed, and volume. Considering the relatively small dataset and the requirement for real-time detection, the YOLOv5s model with the smallest volume is selected to locate motion objects. The extracted motion object regions include moving vehicles, black smoke emissions from the tailpipes, and the dynamic shadows generated by vehicles under clear weather conditions. Figure 3 demonstrates the motion object regions with both black smoke emissions and dynamic shadows extracted by the YOLOv5s model from road traffic surveillance videos.

Figure 3. YOLOv5s model to extract motion target regions.

3. Motion Target Segmentation

3.1. Optimal Segmentation Parameters

Image segmentation involves dividing an image into different regions with specific semantic meanings based on certain similarity criteria. In the early days, image segmentation was mostly performed at the pixel level, using a two-dimensional matrix to represent an image, without considering the spatial relationships between pixels [33]. Simple linear iterative clustering uses the similarity of features between pixels to group pixels and classify pixels of the same type. This is advantageous for reducing data dimensions and computational complexity, thus enhancing the efficiency of image processing [34]. The objective of this research is to automatically detect vehicles emitting black smoke emissions. However, the presence of dynamic shadows generated by vehicles under clear weather conditions can impact the accuracy of black smoke vehicle detection. Therefore, a superpixel segmentation algorithm is employed to process the images of the regions, with moving objects extracted by the YOLOv5s model. This process aims to obtain superpixel images belonging to three categories: vehicles, black smoke emissions, and dynamic shadows. These superpixel images are then used as training samples.

The implementation process of the SLIC involves converting a color image into a five-dimensional feature vector V = [L, a, b, x, y] in the CIELAB color space and XY coordinates. Each pixel's color vector (LL, aa, bb) and position vector (xx, yy) together form a five-dimensional feature vector, enabling the local clustering of image pixels [35]. Firstly, the color space conversion is performed, and a nonlinear tone mapping of the image is achieved using the gamma function. The initial set of k superpixel seed points is evenly distributed over the image containing N pixels [36]. The generated seed points might fall on the edges of superpixels with significant gradients or noisy pixel locations. Therefore, the initial seed points are generally chosen as the positions with the smallest gradient values within a 3×3 neighborhood. The similarity between pixel points and seed points is measured using a distance metric that combines color distance and spatial distance. The parameter m represents a weight factor that gauges the relative importance between color and spatial distances, while S denotes the distance between adjacent seed points. The value of D indicates the similarity between two pixels, with higher values implying greater similarity [37].

$$d_{Lab} = \sqrt{(L_i - L_j)^2 + (a_i - a_j)^2 + (b_i - b_j)^2} \quad (1)$$

$$d_{xy} = \sqrt{(x_{\mathrm{i}} - x_j)^2 + (y_i - y_j)^2} \tag{2}$$

$$D = \sqrt{(d_{Lab})^2 + \left(\frac{d_{xy}}{S}\right)^2 m^2} \tag{3}$$

In the equation, L_i, a_i, and b_i represent the three channel components of a pixel in the CIELAB color space, while x_i and y_i, respectively, denote the horizontal and vertical coordinates of pixel i.

To enhance the computational efficiency of the SLIC, a search for similar pixels is conducted within a 2S × 2S region centered around the seed point. Clustering involves calculating the distance metric between all pixels within this region and the seed point. Through repetitive iterations and assignments, similar feature pixels are grouped to form super pixel blocks. The initial number of seed points, k, and the weight factor, m (which determines the relative importance between color distance and spatial distance), both influence the generation of the superpixel image [38,39]. Therefore, in this experiment, a controlled variable method is employed to analyze and compare the effects of different parameter combinations on the segmentation of motion object regions. This analysis aims to determine the optimal parameter values for the SLIC.

In the first set of comparative experiments, the balancing parameter m of the SLIC was set to 10 and the number of seed points, k, was set to 500, 1000, 1500, and 2000, respectively. The segmentation results of motion object regions are shown in Figure 4. In Figure 4, the red rectangular boxes highlight the segmentation outcomes at the junctions between vehicle tail, dynamic shadow, and road surface. As the number of seed points increases, the under-segmentation phenomenon at the junctions of different objects gradually diminishes, resulting in more consistent content within the generated superpixel blocks. When the segmentation accurately captures the junctions between different objects, increasing the number of seed points will lead to a higher number of superpixel blocks generated during motion object region segmentation. Consequently, this can amplify the computational workload during model classification. Considering the segmentation outcomes from the four different parameter settings, the best segmentation results were achieved when the number of seed points, k, was set to 1500.

The second set of comparative analysis experiments involved setting the number of seed points in the SLIC to 1500. The balancing parameter was varied as 5, 10, 15, and 20, respectively. The segmentation results of motion object regions are shown in Figure 5. In Figure 5, the red rectangular boxes highlight the segmentation details at the junction between black smoke emissions and the road surface. When the balancing parameter is set too small, the boundaries of the object's contours appear blurry. Conversely, when the balancing parameter is set too large, the boundary segmentation of the object's contours becomes imprecise. Considering the segmentation outcomes from the four different parameter settings, the best segmentation results were achieved when the balancing parameter m was set to 10. Consequently, the optimal parameters for the SLIC in this application scenario are selected as k = 1500 and m = 10.

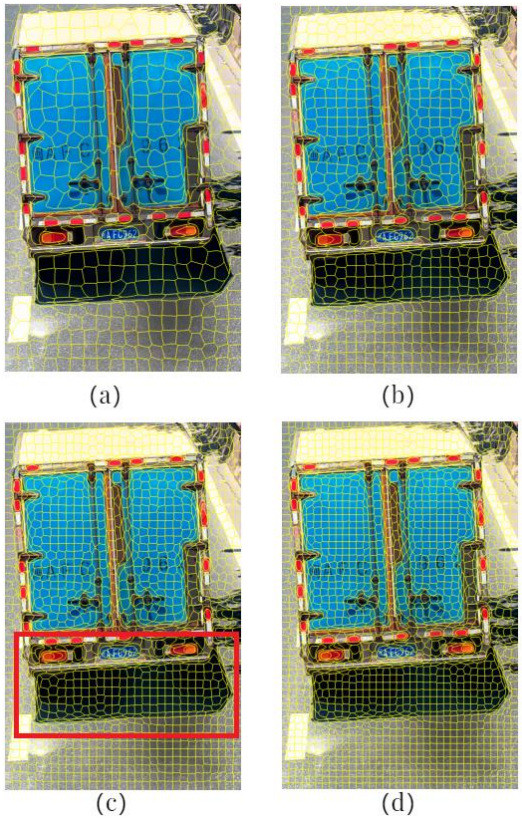

Figure 4. Segmentation results with different numbers of seed points when m = 10: (**a**) k = 500; (**b**) k = 1000; (**c**) k = 1500; (**d**) k = 2000.

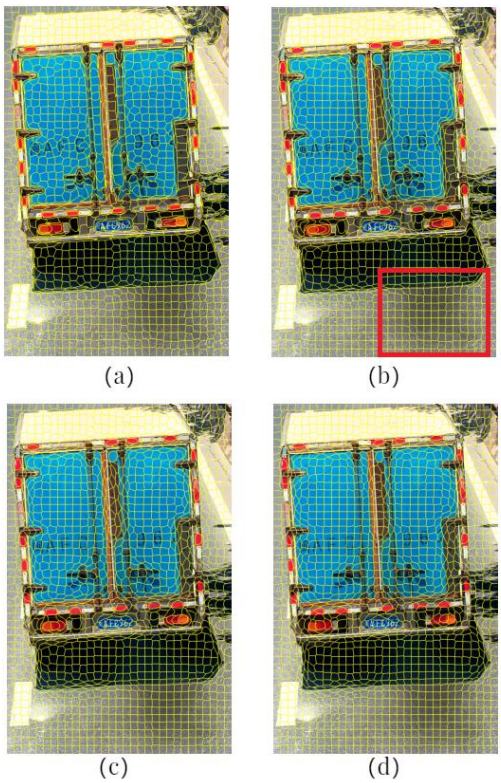

Figure 5. Segmentation results with different balance parameters when k = 1500: (**a**) m = 5; (**b**) m = 10; (**c**) m = 15; (**d**) m = 20.

3.2. Creating Dataset

The three essential elements of deep learning are data, algorithms, and computing power. Data hold a crucial position in deep learning, as a high-quality dataset often improves the accuracy of model predictions. When data are scarce, it is also crucial to utilize existing data resources to create high-quality datasets. A high-quality dataset not only considers the quantity and quality of the raw data but also takes into account the factors that can interfere with experiments during the data preprocessing process. In this study, the data are sourced from road traffic monitoring videos, and the research goal is to automatically detect motor vehicles emitting black smoke on the road. First, the original images containing motor vehicles are obtained through video frame-by-frame processing and selection, as shown in Figure 6.

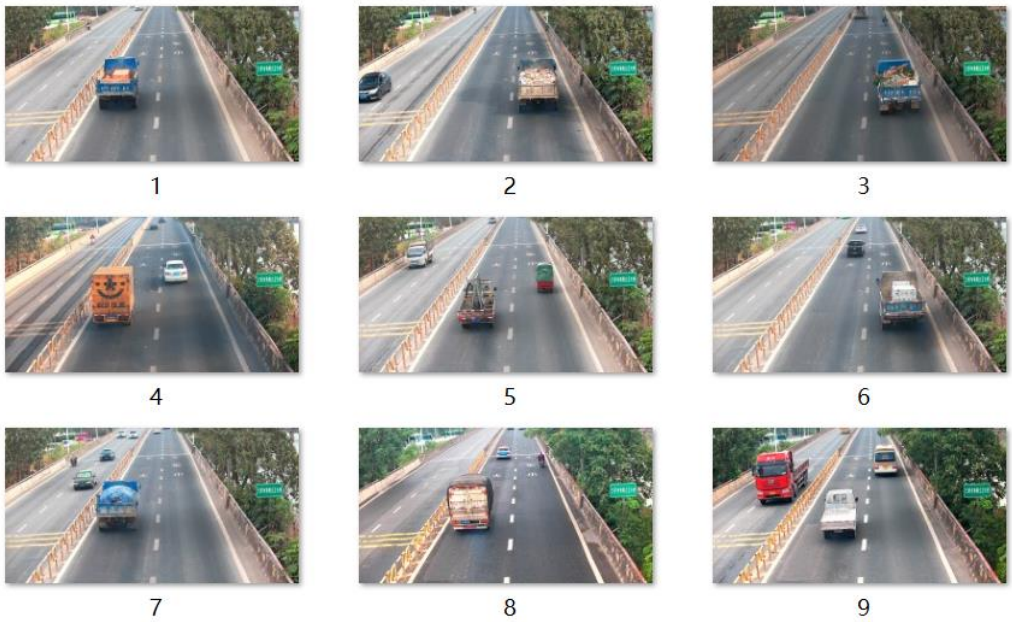

Figure 6. Video frame-by-frame processing to obtain original images containing motor vehicles: 1–3 and 6–7 are heavy truck; 4 and 8 are medium truck; 5 and 9 are light truck.

Based on the YOLOv5s model, we located moving targets and obtained a total of 2900 images containing motor vehicles. Next, based on the 2900 images of located moving targets, two sets of experimental plans were designed to obtain training samples for different models. The automatic detection model for black smoke vehicles considering motion shadows is referred to as "Y-MobileNetv3", while the model not considering motion shadows is referred to as "N-MobileNetv3". The extracted images of moving targets were processed using a superpixel segmentation algorithm, resulting in 1082 images of black smoke emissions, 1035 images of motion shadows, and 1118 images of motor vehicles as training samples for the Y-MobileNetv3 model. The extracted images of moving targets include heavy-duty trucks, medium-sized vans, and light sedans. Adaptive thresholds were designed based on the aspect ratios of the extracted images of moving targets. The last third of the images was selected as the suspected black smoke region, resulting in a total of 2320 non-black smoke emissions and 580 black smoke emissions used as input for training the N-MobileNetv3 model. The process for creating training samples with and without considering motion shadows is shown in Figure 7. The experimental process ensures the consistency of YOLOv5s in locating images of moving targets, with the difference being that the training samples for the model considering motion shadows undergo superpixel segmentation to classify non-black smoke emissions into motor vehicles and motion shadows as two separate categories.

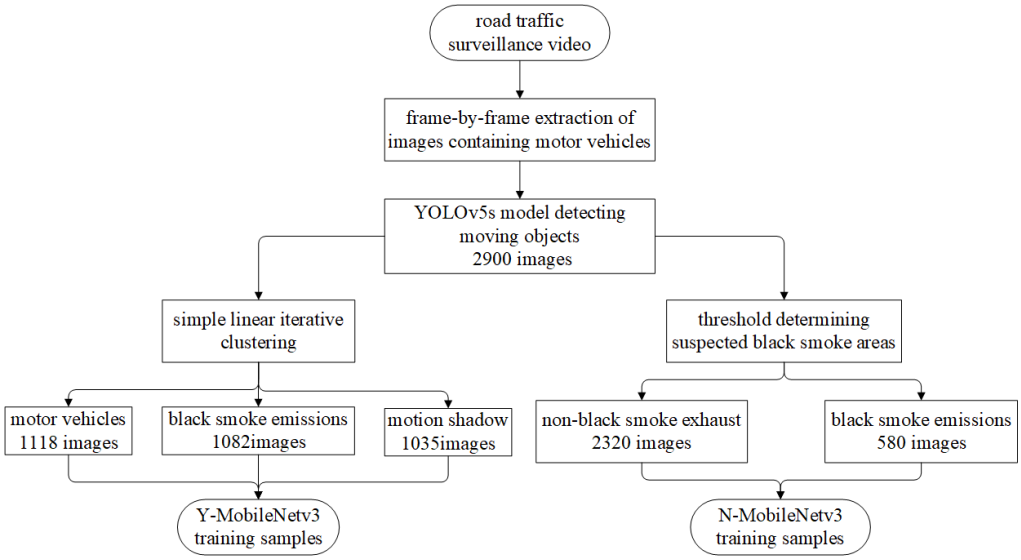

Figure 7. The flowchart for creating training samples for Y-MobileNetv3 and N-MobileNetv3 model.

The settings of two key parameters in the superpixel segmentation algorithm need to be adjusted according to the specific application scenarios. When selecting training samples from different categories after motion target segmentation, it is important to ensure that superpixel images taken from the center of each category region are preserved. This approach helps avoid issues related to excessive segmentation of neighboring objects from different categories, which can negatively impact the quality of training samples. Superpixel images with a resolution of 100 × 100 are saved, as shown in Figure 8, for training samples of some motor vehicles, black smoke emissions, and motion shadows. For motor vehicles, key features that are easy to identify, such as vehicle taillights, rear bumpers, and vehicle body colors, are selected for the superpixel images. The dataset covers various types of motor vehicles, including heavy-duty trucks, medium-sized vans, and light sedans. Superpixel images of black smoke emissions exhibit a hazy and blurry appearance with no distinct texture features, while superpixel images of motion shadows have clearer texture features. These visual differences help distinguish between the two categories.

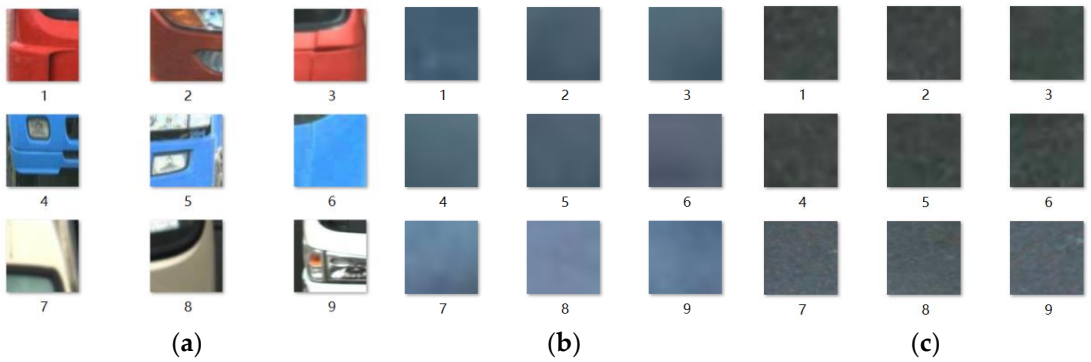

Figure 8. Training samples obtained by the SLIC: (**a**) motor vehicles' black smoke; (**b**) black smoke emissions; (**c**) motion shadows.

4. MobileNetv3 Classification

In 2017, the Google team introduced the lightweight MobileNetv1 model. While ensuring model accuracy, this model significantly reduced the computational load of network model parameters, making it suitable for running applications on mobile terminal devices. Compared with the traditional convolutional neural network VGG16 model, the

MobileNetv1 model had 1/32 of the parameters, while only sacrificing 0.9% of classification accuracy [40,41]. The MobileNetv2 model is an optimized and upgraded version of the MobileNetv1 model by the Google team. It boasts higher accuracy and a smaller model size. This model dramatically reduces the computational load of parameters, making it highly efficient for deployment on mobile devices and suitable for real-world applications. Similar to MobileNetv1, the design of the MobileNetv2 model's architecture also incorporates depthwise separable convolutions instead of standard convolutions. A pointwise convolution is added before the depthwise convolution to increase the dimensionality, allowing the network model to extract features in a higher-dimensional space [42]. Drawing inspiration from the design philosophy of the ResNet network architecture, the input and output are added together in the model, facilitating the flow of information between layers; this aids in feature reuse during forward propagation and mitigates the vanishing gradient problem during backward propagation. The most innovative aspect of the MobileNetv2 model's architecture design is the inverted residual structure. A shortcut connection is only established when the stride is 1 and the input and output feature matrices have the same shape.

The inverted residual structure shown in Figure 9 utilizes a 1×1 pointwise convolution before the depthwise separable convolution to increase the channel dimension of the feature map, followed by a 1×1 convolution for dimension reduction. The classic order of residual blocks is reversed to form the inverted residual structure. The ReLU6 activation function is employed within the inverted residual structure, while the linear activation function is used in the final 1×1 convolution layer. In this context, using the ReLU6 activation function would lead to significant loss of low-dimensional feature information. The overall design of the inverted residual structure is characterized by narrower channels at the two ends and a wider middle section. Applying a linear activation function helps mitigate information loss in the output. The MobileNetv3 model, proposed by Howard and his team in 2019, continues to utilize depthwise separable convolutions from the v1 version and the inverted residual structure from the v2 version [43]. The MobileNetv3 model introduces a new SE (squeeze and excitation) attention mechanism and replaces the swish activation function with the $h-swish$ activation function. The SE attention mechanism comprises compression and excitation parts, involving two fully connected layers with Relu6 and $h-swish$ activation functions, respectively, after global average pooling of features [44,45]. The original authors approximated the $swish$ activation function with ReLU6 to create the $h-swish$ activation function, which effectively addresses the issue of complex gradient calculation [46,47]. The computation formula for the h-swish activation function is as follows:

$$swish(x) = x \cdot sigmoid(\beta x) \tag{4}$$

$$Relu = \max(0, x) \tag{5}$$

$$h-swish(x) = x\frac{Relu(x+3)}{6} \tag{6}$$

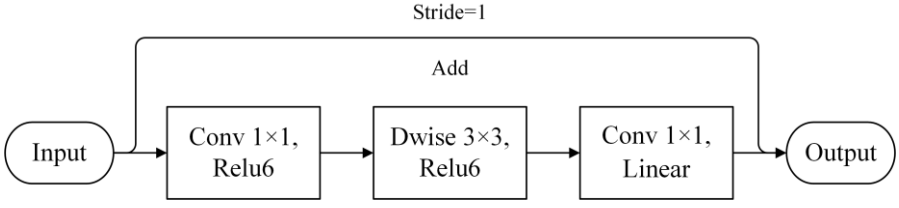

Figure 9. Inverted residual module composition structure.

In the equation, x represents the input and β is a constant or a training parameter.

The MobileNetv3 model strengthens feature extraction through a combination of 3 × 3 standard convolutions and the neck structure. It further enhances the model by incorporating a max pooling layer, substituting 1 × 1 convolution blocks for fully connected layers, and implementing a series of operations to reduce network parameters and complexity [48]. The MobileNetv3 model comes in two scale sizes: "large" and "small". In the ImageNet classification competition, the MobileNetv3-large network achieved a 4.6% increase in accuracy and a 5% improvement in detection speed compared with the v2 version [49]. Similarly, the MobileNetv3-small network demonstrated a 3.2% accuracy improvement and a 15% increase in detection speed over the v2 version.

Taking into account the small size of the experimental dataset and the real-time detection requirements, the MobileNetv3-small model, which has a smaller volume, was chosen for identifying black smoke-emitting vehicles in this study. The training process of the Y-MobileNetv3 model for automatic detection of black smoke-emitting vehicles with consideration of motion shadows is depicted by the loss function variation curve in Figure 10. As the training epochs reach 120 rounds, the loss function fluctuates between 0.1 and 0.2, indicating that the model training is effective and stable.

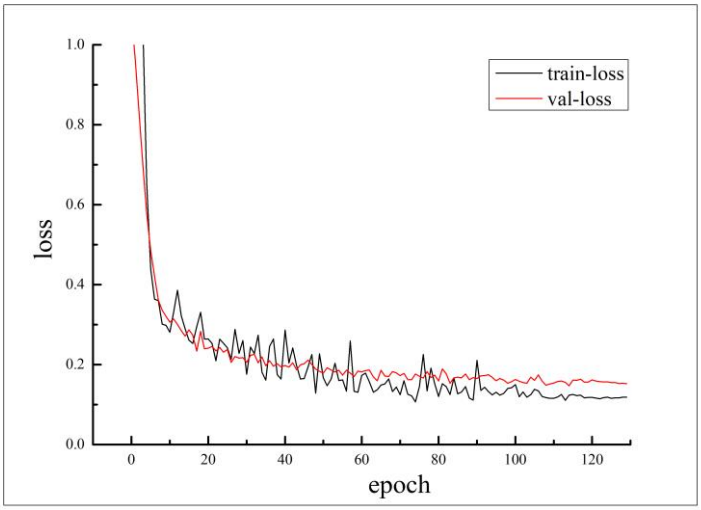

Figure 10. Loss function variation curve of Y-MobileNetv3 model.

5. Experimental Results and Analysis

5.1. Experimental Environment Configuration

The experimental hardware and software environment configuration parameters are shown in Table 2. The hyperparameters of the YOLOv5s model were determined based on previous relevant research and comparative experiments, with input image resolution uniformly scaled to 640 × 640. Prior to training, the initial anchor boxes were clustered using the k-means algorithm, resulting in (10, 13, 16, 30, 33, 23), (30, 61, 62, 45, 59, 119), and (116, 90, 156, 198, 373, 326) The YOLOv5s model was trained for a total of 200 epochs, with a batch size of 8. The Adam optimizer was selected, and the initial learning rate was set to 1×10^{-3} with an initial decay rate of 1×10^{-5}. The learning rate reduction was performed using the cosine annealing strategy. For the MobileNetv3 model, the initial learning rate was set to 0.0001, and the batch size was set to 16 for a total of 130 epochs. The training process utilized the mosaic data augmentation method to enhance the model's robustness, and the SGD optimizer was employed for gradient updates during training.

Table 2. Experimental environment configuration.

Name	Version Model
Operating system	Windows 10
CPU	Intel(R) Core (TM) i5-11400F @2.60 GHz
GPU	NVIDIA GeForce GTX 1650
Programming language	Python 3.8.13
Deep learning framework	Pytorch 1.13.0, CUDA 11.7

5.2. Comparative Experimental Analysis

The test results for automatic detection of black smoke vehicles based on the MobileNetv3 model are shown in Table 3. Y-MobileNetv3 represents the automatic detection model for black smoke vehicles considering motion shadows. The training samples input for Y-MobileNetv3 are superpixel images obtained through segmentation of motion target regions extracted by YOLOv5s. N-MobileNetv3 represents the automatic detection model for black smoke vehicles without considering motion shadows. The training samples input for N-MobileNetv3 are motion target images extracted by YOLOv5s. The confusion matrix, also known as an error matrix, is capable of determining the quality of the model's classification. Predicted values and actual values for all classes are placed in the same table, providing a clear view of the number of correct and incorrect recognitions for each class. Each column of the confusion matrix represents the predicted class of images, with the values indicating the number of images predicted for each class. Each row of the confusion matrix represents the actual class of images, with the values indicating the number of images belonging to each actual class. The results of the confusion matrix can be used to calculate more advanced classification evaluation metrics such as average accuracy, precision, and recall. Average accuracy is the most commonly used classification evaluation metric, calculated by dividing the number of correctly classified instances by the total number of samples. A higher value indicates better classification performance of the model.

Table 3. Test results based on MobileNetv3 modeling.

Confusion Matrix		Y-MobileNetv3		N-MobileNetv3	
		Smoke	No Smoke	Smoke	No Smoke
True value	smoke	145	8	138	15
	no smoke	6	131	13	124

The average accuracy variation curves based on the MobileNetv3 model are presented in Figure 11. The red curve represents the Y-MobileNetv3 model for automatic detection of black smoke vehicles considering motion shadows, while the black curve represents the N-MobileNetv3 model for automatic detection of black smoke vehicles without considering motion shadows. Observing the average accuracy variation curves reveals that the trends of average accuracy for both models change similarly with the epochs, and their learning efficiency is comparable. When the training epochs reach around 80, the average accuracy of the Y-MobileNetv3 model fluctuates around 95%, while the average accuracy of the N-MobileNetv3 model fluctuates around 90%.

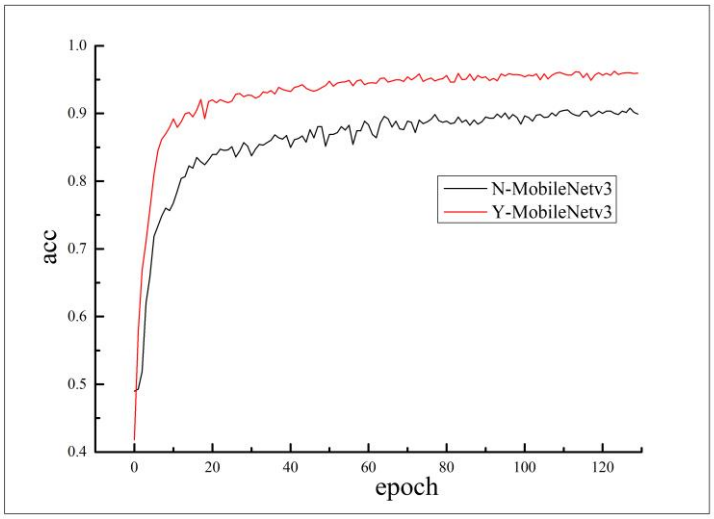

Figure 11. Average accuracy change curve based on MobileNetv3 models.

Through the confusion matrix in Table 3, we can compute the model evaluation metrics, as shown in Table 4. The average accuracy of the Y-MobileNetv3 model is 95.17%, while the average accuracy of the N-MobileNetv3 model is only 90.34%. Average accuracy is an evaluation metric for the entire classification model, but for evaluating each category, we primarily use precision and recall. Precision refers to the proportion of samples identified by the model as black smoke exhaust that are actually black smoke exhaust. Recall is the proportion of actual black smoke exhaust samples that the model correctly predicts as black smoke exhaust. The Y-MobileNetv3 model has a precision of 96.03% and a recall of 94.77%, both of which are 4.64% and 4.58% higher than the N-MobileNetv3 model, respectively. The Y-MobileNetv3 model has a single-image inference speed of 7.3 ms, slightly faster than the N-MobileNetv3 model. This improvement is due to the superpixel segmentation algorithm that groups and classifies similar pixels, enhancing the efficiency of model recognition and classification computations. Compared with existing research on black smoke vehicle detection algorithms, the algorithm proposed in this study, which takes into account motion shadows, has advantages in both detection speed and accuracy, as shown in Table 5. The most important contribution of this research is that it goes beyond previous detection algorithms that solely rely on improving the model network structure to enhance detection performance. Instead, it considers the mutual influence between the research objectives and interfering factors, thereby improving both recognition accuracy and model generality. Under the same test dataset, the Y-MobileNetv3 model's average accuracy improves by 4.73%, clearly demonstrating that using the superpixel segmentation algorithm in the data preprocessing phase to process motion target images and classify motion shadows as a separate category can effectively enhance the recognition accuracy and computational efficiency of the automatic black smoke vehicle detection model.

Table 4. Evaluation metrics based on MobileNetv3 model.

Motion Shadow	Average Accuracy	Precision	Recall	Inference Speed per Image
N-MobileNetv3	90.34%	91.39%	90.19%	8.7 ms
Y-MobileNetv3	95.17%	96.03%	94.77%	7.3 ms

Table 5. Test results of different algorithms for detecting black smoke vehicles.

Model	P (%)	mAP (%)	FPS (ms)
Improved LeNet-5 [3]	87.34	86.75	8.2 ms
CenterNet-ResNet18 [2]	89.67	90.54	22.2 ms
2D-3D Fusion Network [4]	88.93	87.45	45.9 ms
YOLOv3-M3-CBAM [5]	92.57	93.80	49.2 ms
Ringelman-3D CNN [6]	88.56	86.74	5.9 ms
Ours	96.03	95.17	7.3 ms

The results of the Y-MobileNetv3 model are illustrated in Figure 12. Figure 12a, depicts an example where the moving object consists solely of black smoke exhaust. In Figure 12b, an example shows a moving object consisting exclusively of motion shadows. In Figure 12c, an instance demonstrates the coexistence of black smoke exhaust and motion shadows. On clear days, motor vehicles generate motion shadows, and the Y-MobileNetv3 model is capable of excluding the interference of motion shadows and accurately identifying black smoke exhaust. The left side of Figure 9 displays the motion object regions extracted by the YOLOv5s model, while the right side showcases the visualized images of the Y-MobileNetv3 model's test results. Superpixels marked in green represent black smoke exhaust, while those in red denote motion shadows. The motion object regions are recognized and classified by the Y-MobileNetv3 model. The presence of superpixel images indicating black smoke exhaust in the classification results serves as the basis for determining whether a motor vehicle emits black smoke. When black smoke exhaust and motion shadows coexist within the same superpixel block, the model's classification will identify it as black smoke exhaust.

(**a**)

Figure 12. *Cont.*

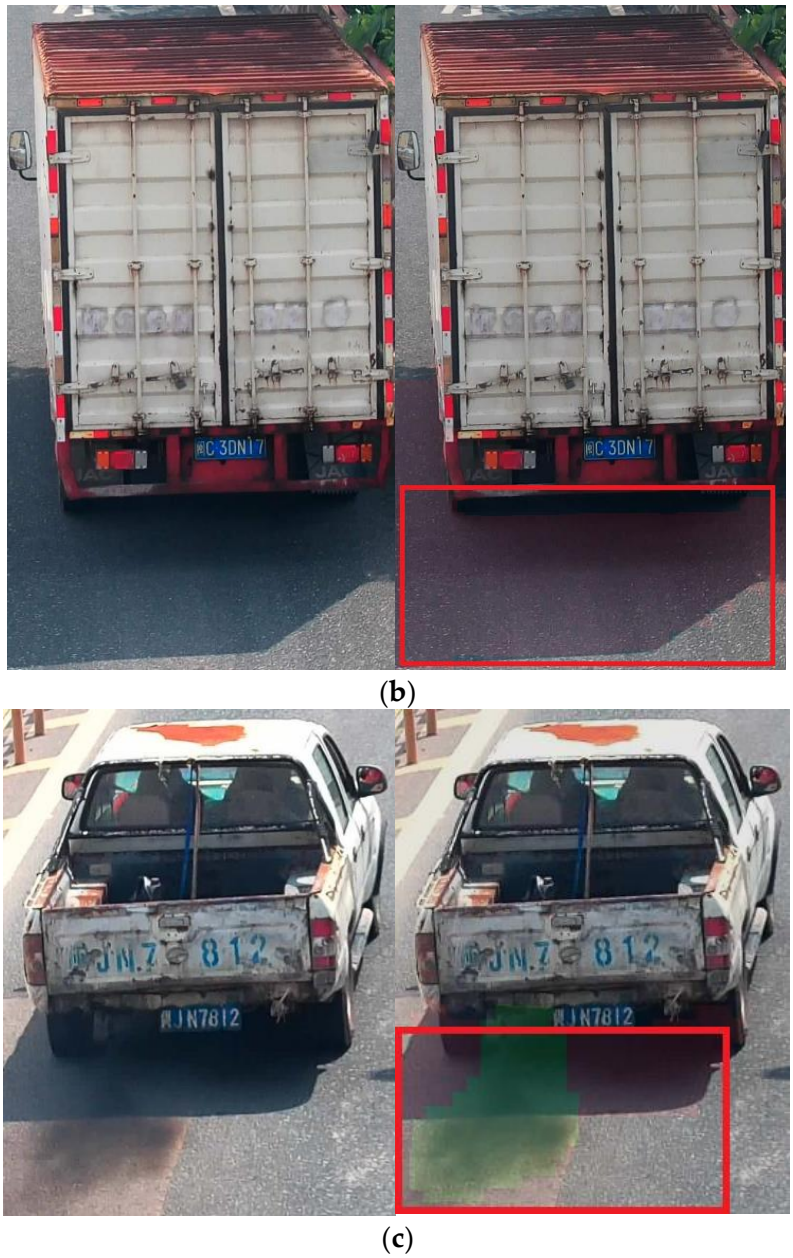

Figure 12. Test results of the Y-MobileNetv3 model: (**a**) black smoke exhaust; (**b**) motion shadows; and (**c**) coexistence of both.

The YOLOv5s model locates the moving target regions of motor vehicles, effectively avoiding interference from other irrelevant moving objects in the research. Experimental results indicate that the N-MobileNetv3 model exhibits false positives and false negatives when detecting motor vehicles emitting trace amounts of black smoke exhaust. In contrast, the Y-MobileNetv3 model can accurately identify them. As shown in Figure 13, the primary reason for false positives and false negatives in the N-MobileNetv3 model is the imprecise identification of suspected black smoke regions. However, the Y-MobileNetv3 model identifies the entire motion target region obtained through the superpixel segmentation algorithm, allowing for accurate recognition of motor vehicles emitting trace amounts of black smoke exhaust. The superpixel segmentation algorithm groups pixels based on the similarity of their features. This characteristic not only aids in distinguishing between black smoke exhaust and motion shadows but also assists the model in identifying motor vehicles emitting trace amounts of black smoke exhaust. By processing the extracted motion target regions using the superpixel segmentation algorithm and classifying motion shadows as

a separate category, it effectively improves the recognition accuracy of automatic black smoke vehicle detection.

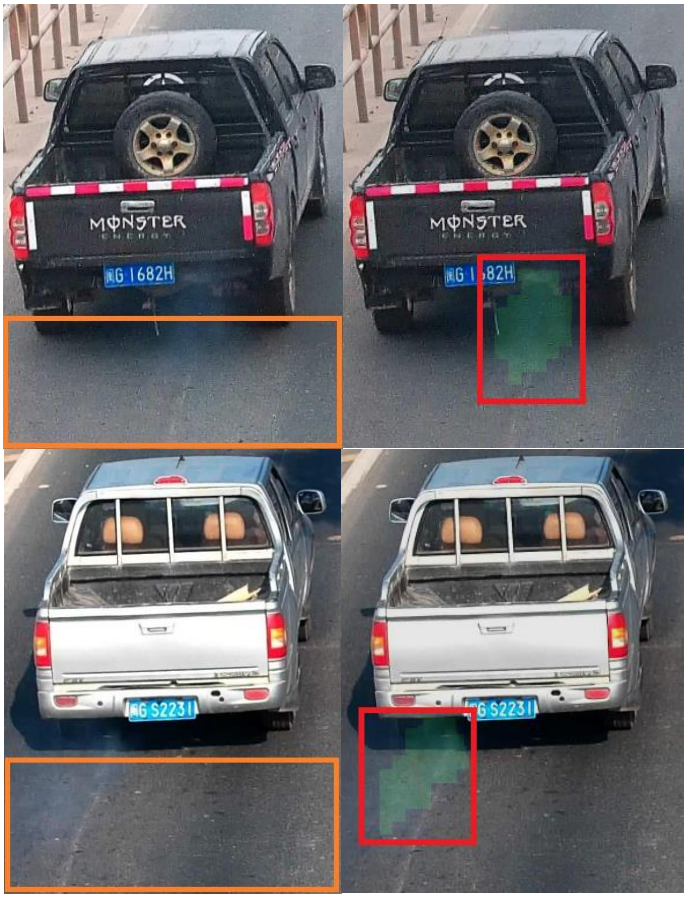

Figure 13. The N-MobileNetv3 (**left**) and Y-MobileNetv3 (**right**) models recognize motor vehicles emitting trace amounts of black smoke exhaust.

6. Conclusions

In the context of road traffic surveillance videos, deep learning-based methods can be employed for automatic detection of black smoke-emitting vehicles. However, these methods often suffer from challenges such as lower recognition accuracy and limited model generalization. The "segmentation-classification" approach effectively distinguishes between black smoke exhaust and motion shadows, reducing instances where motion shadows are misclassified as black smoke exhaust. This approach breaks away from the conventional technique of detecting first and then removing shadows, enhancing both the accuracy of identifying black smoke-emitting vehicles and the general applicability of the automatic detection model. Using the same test dataset, the Y-MobileNetv3 model for black smoke vehicle automatic detection, which considers motion shadows, achieves an average accuracy of 95.17%, precision of 96.03%, and recall of 94.77%. In comparison with the N-MobileNetv3 model, which does not consider motion shadows, all evaluation metrics show significant improvement in results, and the Y-MobileNetv3 model also demonstrates faster inference speeds. The recognition computation time for the Y-MobileNetv3 model is 7.3 ms per image, ensuring real-time detection of black smoke-emitting vehicles while maintaining accuracy.

The model's recognition and classification results are visually displayed through color-coded superpixel images, effectively illustrating the model's successful differentiation between black smoke exhaust and motion shadows. The SLIC aggregates and classifies neighboring pixels with similar features, not only distinguishing between black smoke exhaust and motion shadows but also significantly enhancing the model's deployment

applicability. The superpixel images generated during image segmentation are beneficial for detecting vehicles emitting small amounts of black smoke exhaust, thereby reducing the false negative rate of the automatic detection model. Currently, quantitative calculation of black smoke exhaust concentration from road surveillance video data using computer vision technology remains challenging. However, a color-coded approach can roughly depict the outline of black smoke exhaust. Further research will be to conduct in-depth research on image segmentation of moving targets, to further explore the differences between black smoke exhaust and moving shadows, with the aim of more accurately depicting the black smoke exhaust outline. It realizes the hierarchical classification management of smoky vehicles and helps relevant law enforcement departments to efficiently monitor smoky vehicles.

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

Funding: This study was supported by the Fundamental Research Funds for the Central Universities (No. 2023KYJD1003), National Natural Science Foundation of China (No. 42075132; No. 41975036), Natural Science Foundation of Jiangsu Province Basic Research Program (No. BK20231502).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are not publicly available due to privacy restrictions.

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

References

1. Cao, Y.; Lu, X. Learning spatial-temporal representation for smoke vehicle detection. *Multimed. Tools Appl.* **2019**, *78*, 27871–27889. [CrossRef]
2. Guo, T.; Ren, M. Dual branch network for black smoke and vehicle detection based on attention mechanism. *Comput. Digit. Eng.* **2022**, *50*, 147–151.
3. Xia, X. *Research on Smoke Vehicle Detection Technology Based on Video Image*; Southeast University: Nanjing, China, 2019.
4. Zhang, G.; Zhang, D.; LU, X.; Cao, Y. Smoky Vehicle Detection Algorithm Based on Improved Transfer Learning. In Proceedings of the 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; pp. 155–159.
5. Zhang, Q. *Research on Smoky Vehicle Detection Technology Based on Computer Vision*; Hebei University of Science and Technology: Shijiazhuang, China, 2021.
6. Liu, R. *Research on Detection Algorithm of Vehicle Black Smoke Based on Video*; Dalian University of Technology: Dalian, China, 2022.
7. Kumar, A. SEAT-YOLO: A squeeze-excite and spatial attentive you only look once architecture for shadow detection. *Opt.-Int. J. Light Elect. Opt.* **2023**, *273*, 170513. [CrossRef]
8. Khan, S.; Bennamoun, M.; Sohel, F.; Togneri, R. Automatic Feature Learning for Robust Shadow Detection. In Proceedings of the 2014 Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1939–1946.
9. Tian, J.; Tang, Y. Linearity of Each Channel Pixel Values from a Surface in and out of Shadows and Its Applications. In Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 985–992.
10. Hu, X.; Jiang, Y.; Fu, C.; Heng, P. Mask-Shadow GAN: Learning to Remove Shadows from Unpaired Data. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 20–26 October 2019; pp. 2472–2481.
11. Choi, S.; Jeong, G. Shadow Compensation Using Fourier Analysis with Application to Face Recognition. *IEEE Signal Process. Lett.* **2011**, *18*, 23–26. [CrossRef]
12. Wu, X.; Liu, X.; Chen, Y.; Shen, J.; Zhao, W. A graph based superpixel generation algorithm. *Appl. Intell.* **2018**, *48*, 4485–4496. [CrossRef]
13. Chen, J.; Bao, E.; Pan, J. Classification and Positioning of Circuit Board Components Based on Improved YOLOv5. *Procedia Comput. Sci.* **2022**, *208*, 613–626. [CrossRef]
14. Dong, J.; Chen, S.; Miralinaghi, M.; Chen, T.; Labi, S. Development and testing of an image transformer for explainable atoomous driving systems. *J. Intell. Connect. Veh.* **2022**, *5*, 235–249. [CrossRef]

15. Tong, Z.; Wu, Y.; Liu, Y. Single-stage Multi-scale Receptive Field Improvement Lightweight Object Detection Network Based on MobileNetV3. In Proceedings of the 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), Chi Zhou, China, 20–26 October 2022; pp. 103–106.
16. Chen, J.; Wang, W.; Zhang, D.; Zeb, A.; Nanehkaran, Y. Attention embedded lightweight network for maize disease recognition. *Plant Pathol.* **2020**, *70*, 630–642. [CrossRef]
17. Liao, X.; Zeng, X. Review of target detection algorithm based on deep learning. In Proceedings of the 2020 International Conference on Artificial Intelligence and Communication Technology (AICT 2020), Chongqing, China, 28–29 March 2020; pp. 55–59.
18. Li, W.; Sheng, F.; Zha, K.; Li, S.; Zhu, H. Summary of target detection algorithms. *J. Phys. Conf. Ser.* **2021**, *1757*, 012003. [CrossRef]
19. He, Q.; Xu, A.; Ye, Z.; Zhou, W.; Cai, T. Object Detection Based on Lightweight YOLOX for Autonomous Driving. *Sensors* **2023**, *23*, 7596. [CrossRef]
20. Oh, G.; Lim, S. One-Stage Brake Light Status Detection Based on YOLOv8. *Sensors* **2023**, *23*, 7436. [CrossRef] [PubMed]
21. Alex, K.; Ilya, S.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90.
22. Hu, T.; Wang, W.; Gu, J.; Xia, Z.; Zhang, J.; Wang, B. Research on Apple Object Detection and Localization Method Based on Improved YOLOX and RGB-D Images. *Agronomy* **2023**, *13*, 1816. [CrossRef]
23. Tang, H.; Liang, S.; Yao, D.; Qiao, Y. A visual defect detection for optics lens based on the YOLOv5-C3CA-SPPF network model. *Opt. Express* **2023**, *31*, 2628–2643. [CrossRef] [PubMed]
24. 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]
25. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A deep network architecture for pan-sharpening. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 21–26 July 2017; pp. 1753–1761.
26. Yin, L.; Wang, L.; Li, J.; Lu, S.; Tian, J.; Yin, Z.; Liu, S.; Zheng, W. YOLOV4_CSPBi: Enhanced land target detection model. *Land* **2023**, *12*, 1813–1829. [CrossRef]
27. Lin, T.; Doollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
28. Liu, L.; Fan, J. Multi-Scale Motion Attention Fusion Algorithm for Video Moving Target Detection. *J. Phys. Conf. Ser.* **2018**, *1098*, 012030.
29. Ding, F.; Mi, G.; Tong, E.; Zhang, N.; Bao, J.; Zhang, D. Multi-channel high-resolution network and attention mechanism fusion for vehicle detection model. *J. Automot. Saf. Energy* **2022**, *13*, 122–130.
30. Tom, A.; George, S. Video Completion and Simultaneous Moving Object Detection for Extreme Surveillance Environments. *IEEE Signal Process. Lett.* **2019**, *26*, 577–581. [CrossRef]
31. Feng, Y.; Wu, Q.; He, G. Motion Target Detection Algorithm Based on Monocular Vision. In Proceedings of the Sixth International Conference on Software and Computer Applications (ICSCA), Bangkok, Thailand, 26–28 February 2017; pp. 107–111.
32. Tian, J.; Ma, B.; Lu, S.; Yang, B.; Liu, S.; Yin, Z. Three-Dimensional point cloud reconstruction method of cardiac soft tissue based on binocular endoscopic images. *Electronics* **2023**, *12*, 3799–3817. [CrossRef]
33. Shang, L.; You, F.; Han, C.; Wang, X.; Zhao, S. Optimization of Three-Frame Difference Method and Improvement of Pedestrian Detection Code Book. *J. Phys. Conf. Ser.* **2019**, *1302*, 022014. [CrossRef]
34. Ng, T.; Choy, S.; Lam, S.; Yu, K. Fuzzy Superpixel-based Image Segmentation. *Pattern Recognit.* **2023**, *134*, 109045. [CrossRef]
35. Maame, G.; Anh, H.; Salman, A.; Zaher, A.; Andrzej, C. Image reconstruction using superpixel clustering and tensor completion. *Signal Process.* **2023**, *212*, 109158.
36. Pouria, S.; Nicolas, V.; Eva, m.; Piet, R.; Malcolm, J. DeepCount: In-Field Automatic Quantification of Wheat Spikes Using Simple Linear Iterative Clustering and Deep Convolutional Neural Networks. *Front. Plant Sci.* **2019**, *10*, 1176.
37. Zhu, Y.; Luo, K.; Ma, C.; Liu, Q.; Jin, B. Superpixel Segmentation Based Synthetic Classifications with Clear Boundary Information for a Legged Robot. *Sensors* **2018**, *18*, 2808. [CrossRef] [PubMed]
38. Nur, A.; Mohd, A.; Wan, M.; Aini, H. An automated glaucoma screening system using cup-to-disc ratio via Simple Linear Iterative Clustering superpixel approach. *Biomed. Signal Process. Control* **2019**, *53*, 101454.
39. Chang, C.; Ding, J.; Lin, H. Learning Based SLIC Superpixel Generation and Image Segmentation. In Proceedings of the 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Madison, WI, USA, 24–26 March 2019.
40. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
41. Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
42. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the2018 Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–21 June 2018; pp. 4510–4520.
43. Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudenvan, V.; et al. Searching for MobileNetV3. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2019**, *1905*, 02244.
44. Bi, C.; Xu, S.; Hu, N.; Zhang, S.; Zhu, Z.; Yu, H. Identification Method of Corn Leaf Disease Based on Improved Mobilenetv3 Model. *Agronomy* **2023**, *13*, 300. [CrossRef]

45. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2011–2023. [CrossRef]
46. Zhao, Y.; Huang, H.; Li, Z.; Huang, Y.; Lu, M. Intelligent garbage classification system based on improve MobileNetV3-Large. *Connect. Sci.* **2022**, *34*, 1299–1321. [CrossRef]
47. Liu, K.; Wang, J.; Zhang, K.; Chen, M.; Zhao, H.; Liao, J. A Lightweight Recognition Method for Rice Growth Period Based on Improved YOLOv5s. *Sensors* **2023**, *23*, 6738. [CrossRef] [PubMed]
48. Zheng, H.; Duan, J.; Dong, Y.; Liu, Y. Real-time fire detection algorithms running on small embedded devices based on MobileNetV3 and YOLOv4. *Fire Ecol.* **2023**, *19*, 31. [CrossRef]
49. Mohamed, A.; Abdelghani, D.; Naser, A.; Ammar, H.; Amal, I.; Mahmoud, A. Boosting COVID-19 Image Classification Using MobileNetV3 and Aquila Optimizer Algorithm. *Entropy* **2021**, *23*, 1383.