Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of systematic manipulation of the publication process, including compromised peer review. The Editor and publisher no longer have confidence in the results and conclusions of the article.

CZ, JL, and FZ either did not respond directly or could not be reached.

Transfer learning-based YOLOv3 model for road dense object detection

Chunhua Zhu[a,b,c,*] **Jiarui Liang**[a,b,c] **and Fei Zhou**[a,b,c]

[a]Henan University of Technology, Ministry of Education, Key Laboratory of Grain Information Processing and Control, Zhengzhou, China

[b]Henan University of Technology, Henan Key Laboratory of Grain Photoelectric Detection and Control, Zhengzhou, China

[c]Henan University of Technology, College of Information Science and Engineering, Zhengzhou, China

Abstract. Stemming from the object overlap and undertraining from the few samples, the road dense object detection is confronted with the poor object identification performance and the inability to recognize edge objects. Based on this, one transfer learning-based you only look once, version 3 (YOLOv3) approach for identifying dense objects in the road has been proposed. First, Darknet-53 network structure is adopted to obtain pre-trained YOLOv3 model, then the transfer training is introduced as the output layer for the special dataset of 2000 images containing vehicles; in the proposed model, one random function is adapted to intialize and optimize the weights of the transfer training model, which is seperately designed from the pre-trained YOLOv3; and the object detection classifier replaces the fully connected layer, which further improves the detection effect. The experimental results demonstrate that the object detection accuracy of the presented approach is 87.75% for the Pascal Visual Object Classes (VOC) 2007 dataset, which is superior to the YOLOv2 and the traditional R-CNN by 11.05% and 0.8%, respectively. In addition, the detection speed of the proposed YOLOv3 method reaches 27.3 frames per second (Fps)/s in detecting images, which is 6.4 Fps/s faster than the traditional YOLOv3; the proposed YOLOv3 performs 79.38Bn of floating point operations per second in detecting video, which obviously surpasses the traditional YOLOv3. © *2023 SPIE and IS&T* [DOI: 10.1117/1.JEI.32.6.062505]

Keywords: dense road; object detection; Darknet-53 network; transfer learning.

Paper 221080SS received Oct. 10, 2022; accepted for publication Dec. 13, 2022; published online Jan. 4, 2023; retracted Jul. 8, 2023.

1 Introduction

The dense detection of objects, such as vehicles and pedestrians in the road scenes faces the problems of object overlapping or occlusion, uneven distribution of objects, and difficulty in detecting edge objects. Traditional object detection techniques may be separated into the following two broad categories: region-convolutional neural network (R-CNN), which is based on candidate regions, including R-CNN,[1] Fast R-CNN,[2] Faster R-CNN,[3] and other two-stage networks; and regression-based single-stage networks, such as you only look once (YOLO)[4] and single shot multibox detector (SSD).[5] R-CNN algorithms need extract the characteristics of the candidate regions before feeding them into a pre-trained CNN model to get a characteristic for classification. Due to the large amount of object overlapping in the candidate regions, the features from the overlapping regions will be repeatedly calculated when extracting features, thereby, the real-time performance is poor; comparably, the YOLO model only requires one feed forward neural network to directly predict the classifications and positions of diverse objects for several independent candidate regions, which has the advantages of the simpler training, the better detection efficiency, etc.[6] YOLO has been widely used in object detection,[7] and it has evolved to the YOLOv3.[8] Compared with YOLOv2, the YOLOv3 model has the stronger

*Address all correspondence to Chunhua Zhu, zhuchunhua@haut.edu.cn

1017-9909/2022/$28.00 © 2023 SPIE and IS&T

self-learning ability and larger capacity, which can solve the problem of poor accuracy from the excessive high-resolution dataset and unbalanced positive and negative samples; besides, the YOLOv3 has more candidate frames and can improve the intersection ratio during detection. For matching the special road dense object detection scene, transfer learning[9] will be introduced to the YOLOv3 network to fine-tune and accelerate the original training model, which is embedded in the output layer of pre-trained YOLOv3 thereby one new transfer learning-based YOLOv3 model is shaped.

The key contributions can be described as

1. Based on the pre-trained YOLOv3 model, the transfer training is introduced as the output layer for the special dataset containing detecting objects; moreover, one random function is adapted to initialise and optimize the weights of the transfer training model, which is seperately designed from the pre-trained YOLOv3.
2. In the proposed YOLOv3, the object detection classifier replaces the full convolutional layer of the traditional YOLOv3, aiming to relieve the conflict distinguishing the edge target features and other background features caused by the excessive receptive field of the full connection layer, besides, the introduced classifier can avoid the excessive computation from the fully connected layer in the traditional YOLOv3.

2 Algorithm Principle

Figure 1 shows the architecture of the dense road detection system based on YOLOv3, which include three parts: the YOLOv3 backbone network, transfer training unit, and optimization of network parameters. First, the Visual Object Classes (VOC) 2007, VOC 2012, and Common Objects in Context (COCO) datasets are selected for YOLOv3 network pre-training; then, the images containing vehicles are extracted from the VOC 2007 dataset and re-labeled to form the special vehicle dataset, and the transfer training-based YOLOv3 model is transferred and trained in this dataset; finally, the test dense road pictures or videos are input into the proposed model, and the output is obtained by feature extraction, multi-scale detection, and non-maximum suppression processing. Performance evaluations are performed by confidence, mean precision [mean average precision (mAP)], and precision-recall (P-R) of object detection.

2.1 *Feature Extraction Network*

Compared with the YOLOv2 network, the backbone portion of the YOLOv3 network has evolved from Darknet-19 to Darknet-53, consequently expanding the number of network layers[10] and adding the cross-layer sum operation in the residual network; YOLOv3 network has 53 convolutional layers [residual network (ResNet)]. Darknet-53 is an entirely convolutional

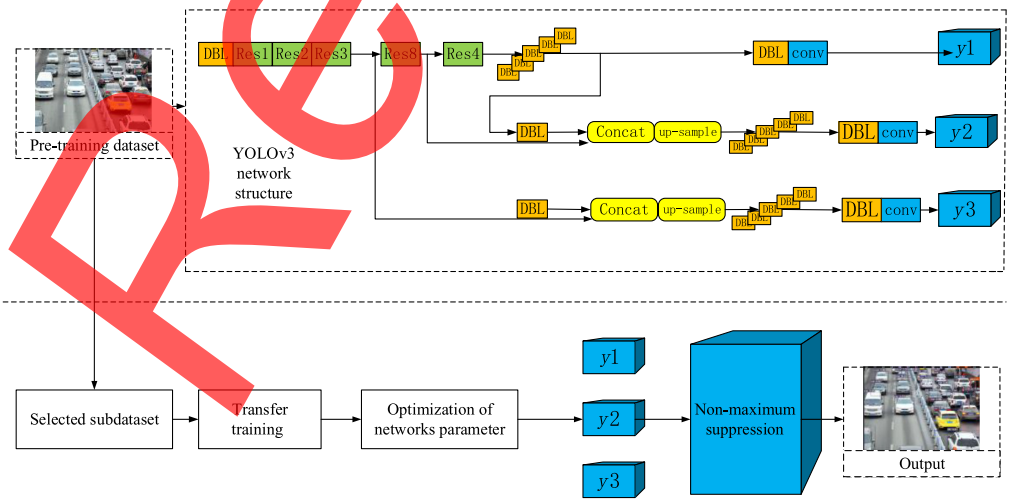

Fig. 1 Architecture of dense road object detection model.

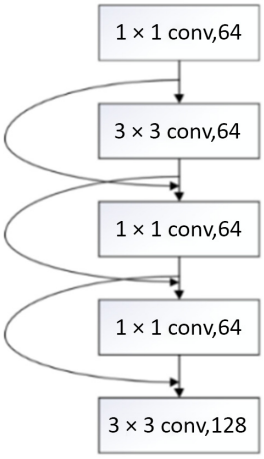

Fig. 2 Schematic diagram of quick link.

network comprised of 3 × 3 and 1 × 1 convolutional layers, including 23 residual modules and layers of detection channels that are completely interconnected. As shown in Fig. 2, the convolutional layers are interconnected by quick link[11] [i.e., shortcut connections (SC)]. This SC structure can greatly enhance the computation performance of the network, enabling the network to obtain faster detection speed in a limited number of network layers. In the detection architecture, YOLOv3 separates three channels for feature detection into distinct grid sizes. These channels include feature maps with grid sizes of 52 × 52, 26 × 26, and 13 × 13, which correspond to the detection of large-scale($y1$), medium-scale($y2$), and small-scale($y3$) picture features, respectively. Thereby, The YOLOv3 can provide a higher detection accuracy with fewer network parameters and fewer superfluous network layers, enabling it to improve both the detection speed and the detection accuracy. By comparison, the conventional R-CNN relies on deepening the network structure to enhance the recognition rate.

2.2 *Training Strategy Design*

The traditional deep-learning network generally improves the recognition accuracy by increasing the training set or deepening the network complexity. In the application of road dense object detection, the real-time detection is an important indicator.[12] Transfer learning is introduced in an effort to improve the training process of the traditional YOLOv3 network, as Fig. 3.

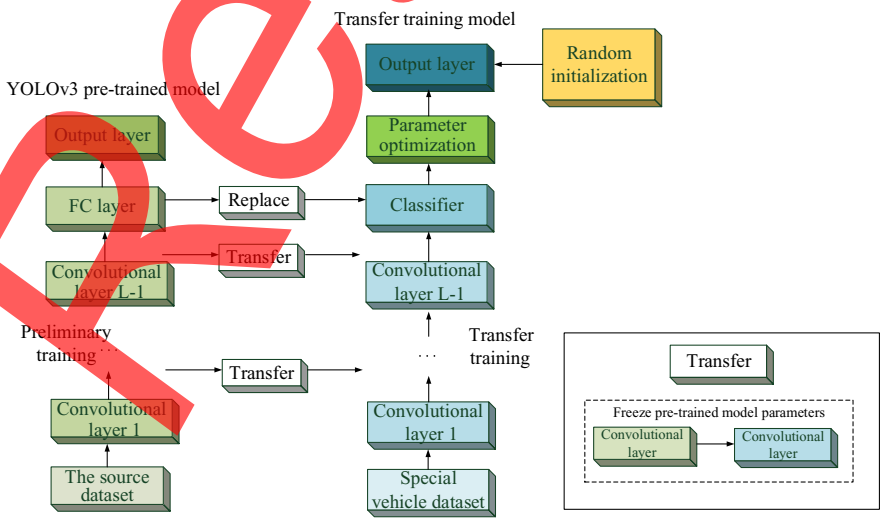

Fig. 3 Transfer training strategy.

Table 1 Performance of YOLOv3 transfer training model.

Classes	AP (%)	TP (%)	FP (%)	Map (%)
Car	81.25	79.65	20.34	83.85
Person	76.43	60	40	
Bicycle	74.12	82.81	17.18	
Motor cycle	73.08	74.14	25.85	
Bus	79.14	81.06	19.02	
Dog	72.44	77.07	22.92	
Cat	73.34	65.41	34.58	

In Fig. 3, the YOLOv3 backbone network is combined with the transfer training model, during pre-training process, the YOLOv3 network is trained in the VOC 2007 and VOC 2012 dataset to obtain the pre-trained model for 20 object types in the VOC dataset, besides, the COCO dataset is trained to obtain 60 object types, totally 80 object types can be obtained; then, the transfer training unit substitutes the convolutional layer of the YOLOv3 backbone network, its parameters are not fixed, which can be randomly initialized and optimized,[13] on the special vehicle dataset. During pre-training, the epoch is set to 300, and the average precision (AP), and mAP of the transfer training model are given in Table 1.

In the transfer training model, the proposed YOLOv3 retains the convolutional layer in the traditional YOLOv3 after pre-training. When performing feature extraction, the pre-trained convolutional layers[14] are selected to be a feature extractor, which structure allows the input image to propagate forward. The convolutional layer parameters[15] retained from the pre-trained model were frozen, and takes the output of convolutional layer L-1 as the proposed YOLOv3 extracted feature.[16] Such a convolutional layer can enable the network to better obtain the semantic information of the target and train it, thereby obtaining higher detection accuracy; then, the proposed YOLOv3 replaces the full convolutional layer of the traditional YOLOv3 with the object detection classifier,[17] which can relieve the conflict distinguishing between the edge target features and other background features caused by the excessive receptive field of the full connection layer, besides, it can reduce the problem of excessive computation caused by the fully connected layer, so that the proposed YOLOv3 can train and detect objects faster; and the corresponding network parameters are accordingly optimized[18] to increase dense detection precision on the road while simultaneously reducing training complexity.

In Table 1, AP represents the mean precision of each object type in the test set. The term correctly classifying a positive example as a positive example (abbreviated "TP") refers to the degree of precision achieved when accurately identifying a positive example, while falsely classifying a negative example as a positive example refers to incorrectly identifying a negative example as a positive example (abbreviated as "FP").

During the transfer training, nearly 2000 images containing vehicles are screened and labeled. This is a special dataset built[19] for dense object detection in road situations. Included among the road objects in the dataset are car, people, bicycles, motorbikes, trucks, cats, and dogs. There are seven types that appear frequently in the road scene, which is given Table 2, and the remaining 13 types that are not common.

From Table 2, a total of 1434 labeled car images in the dataset are selected as special dense road sub-dataset. The images are then divided into a training set and a test set in a 4:1 ratio.[20]

Table 2 Selected special dataset.

Class name	Car	Person	Bicycle	Motor cycle	Bus	Dog	Cat
Total number	1434	1360	860	430	300	220	180

In addition, the classes of the special dense road sub-dataset must be modified to 7 and the file path of "train," "valid," "names," and "backup" must be modified correspondingly for the transfer training model to correspond with the extracted special dense road sub-dataset. This is necessary for the transfer training model to correspond with the special dense road sub-dataset. Meanwhile, parameters batch, LEARN-RATE, and Intersection over Union (IOU) in config.py need to be optimized with the following consideration:

1. The batch function is set to 8. It can make the network complete an epoch in a few iterations, and reach a local optimal state while finding the best gradient descent direction.[21] Increasing this value will prolong the training time, but will better find the gradient descent direction; decreasing this value may cause the training to fall into a local optimum, or not converge.
2. The LEARN-RATE function is set from $1e-4$ to $1e-6$. During the training procedure,[22] the total number of training cycles generally determines a learning rate that continually adapts to new input. About 10 rounds of training are setted. Experiments indicate that the learning rate is initially set at 0.0001 at the beginning of training and then gradually slows down after a certain number of rounds have been completed. As the training nears its conclusion, the learning rate is lowered until it hits 0.000001. The setting of the learning rate, which is based on ten training rounds, not only solves the problems of easy loss value explosion and easy oscillation caused by a learning rate that is too large at the beginning of training; if it is too small, it is easy to over fitting, resulting in slow network convergence.
3. The IOU function is allocated the value 0.65. In computer detection tasks, the IOU value is equal to one and the intersection is the same as the union if the actual bounding box and the predicted bounding box entirely overlap.[23] Generally, a value of 0.5 is utilized as the threshold to determine whether or not the predicted bounding box is correct. It is possible to increase the detection accuracy of tiny items and edge objects while dealing with the detection of dense road objects by setting IOU to 0.65. This will enable the gathering of higher-quality samples and enhance the identification of dense road objects.

2.3 *Loss Function Selection*

In the proposed YOLOv3, the error loss mainly comes from the misjudgment of prediction frame, confidence and category. The error in the prediction frame is determined by the rate of coincidence between the a priori frame and the prediction frame. If the rate at which the prediction frame and the a priori frame agree is large, this implies that the prediction is accurate and error margins are small; confidence error refers to the mistake induced by random sampling during testing in the test set, sampling the test set, and then estimating all test sets; category misjudgment error is the error caused by detecting one type into another. These three types of losses are expressed as lbox, lobj, and lcls, respectively,

$$\text{lbox} = \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{\text{obj}} (2 - w_i \times h_i)[(x_i - \hat{x}_i) + (y_i - \hat{y}_i) + (w_i - \hat{w}_i) + (h_i - \hat{h}_i)], \quad (1)$$

$$\text{lcls} = \lambda_{\text{class}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{\text{obj}} \sum_{c \in \text{classes}} p_i(c) \log(\hat{p}_i(c)), \quad (2)$$

$$\text{lobj} = \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{\text{noobj}} (c_i - \hat{c}_i)^2 + \lambda_{\text{obj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{i,j}^{\text{noobj}} (c_i - \hat{c}_i)^2, \quad (3)$$

$$\text{loss} = \text{lbox} + \text{lobj} + \text{lclcs}. \quad (4)$$

Here, s is the grid size; B is amount of prediction frames; $I_{i,j}^{\text{obj}}$ is the indicator function, which stand for if the prediction frame at i, j has a object, its value is 1, otherwise it is 0; $I_{i,j}^{\text{obj}}$ is the indicator function, which stands for if the prediction box at i, j has no object and its value is 1, otherwise it is 0.

Table 3 Configuration environment.

Operating system	CPU	Memory	GPU	CUDA	CUDNN
Windows 10	Intel i5	8 GB	NVIDIA GEFORCE RTX 745	CUDA 10.04	CUDNN 7.04

3 Experimental Testing and Evaluation

3.1 *Required Environment*

The experimental configuration is given in Table 3, using Tensorflow framework in deep learning and Opencv Python framework in computer vision.

3.2 *Detection Performance Evaluation*

In the existing object detection algorithms, map and *P*-*R* value are the common evaluation indicators of recognition accuracy. The *P*-*R* curve is composed of precision and recall[24]

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{5}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{6}$$

Here, "true positive" (TP) refers to the genuine example, "false positive" (FP) refers to a FP example, and "false negative" (FN) refers to a FN example. In addition, the index of mAP[25] is also adapted to show the mAP.

Figure 4 shows the *P*-*R* curve generated by applying the proposed YOLOv3 to the data for the car object detection in road scene. If the classification results of all test samples are positive, the recall rate of the model will be one and the accuracy rate will be extremely low; if the classification results of almost all test samples are negative, the accuracy rate will be high and the recall rate will be very low. Precision and recall are two rather contradicting metrics. Figure 4 shows that the proposed YOLOv3 is capable of reaching a high accuracy rate, that the recall rate slope drops in a moderate and steady fashion, and that the accuracy rate and recall rate achieve a more balanced state.

In the special dense road sub-dataset, the training set and test set are divided into two sections with 4:1 ratio. AP statistics are performed on the seven types as Table 2, and mAP is calculated.

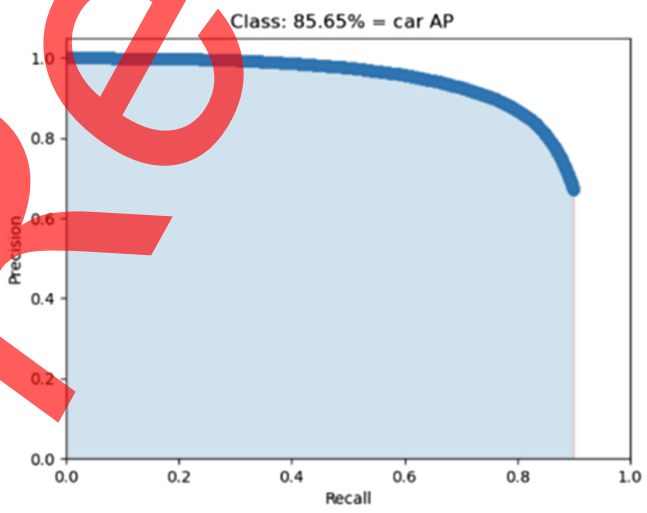

Fig. 4 *P*-*R* curve of vehicle object detection.

Table 4 Object detection performance of the proposed YOLOv3.

Classes	AP (%)	TP (%)	FP (%)	AP (%)
Car	89.25	86.5	13.4	87.85
Person	83.43	73.2	26.8	
Bicycle	82.12	88.9	11.0	
Motor cycle	82.08	78.6	21.3	
Bus	84.14	89.8	10.2	
Dog	79.44	79.6	20.4	
Cat	86.34	69.6	30.3	

Table 5 Object detection performance of the YOLOv2.

Classes	AP (%)	TP (%)	FP (%)	mAP (%)
Car	78.25	79.65	20.34	76.8
Person	75.43	60	40	
Bicycle	72.12	82.81	17.18	
Motor cycle	70.08	74.14	25.85	
Bus	73.14	81.06	19.02	
Dog	69.44	77.07	22.92	
Cat	78.34	65.41	34.58	

Tables 4 and 5 show the object detection performance of the proposed YOLOv3 and the YOLOv2, respectively.

From Tables 4 and 5, compared to the YOLOv2, the proposed YOLOv3 exhibits a obvious improvement in the detection accuracy of each category and average detection accuracy.

3.3 *Comparative Analysis of Different Algorithms*

For detecting static road pictures, Fig. 5 shows the original object image in the top row, the detection result of YOLOv2 in the second row, and the proposed YOLOv3 detection result in the third row. It is evident that the proposed YOLOv3 is capable of accurately differentiating between different types of vehicles and pedestrians, as well as accurately detecting pedestrians at the edge of the image; it is also capable of using a confidence box to indicate the object type division probability, and to mark the object type division probability; besides, the confidence will not decrease as the density of road objects increasing.

When detecting the real-time collected videos, the test results of YOLOv2 and the proposed YOLOv3 utilizing of road dense objects are shown as Figs. 6 and 7, respectively.

Comparing Figs. 6 and 7, the proposed YOLOv3 can detect and divide vehicles and pedestrians in real-time, even the detection confidence for different types of objects can also be provided; while the YOLOv2 can only roughly divide the object types, and cannot provide confidence information.

Considering the the difference of object image resolutions, and for the YOLOv2, traditional YOLOv3 and the proposed YOLOv3, the detection time is shown in Table 6. Table 7 gives the corresponding detection time of selecting the length 42.6 s of video as the input.

Table 6 gives that the proposed YOLOv3 has the faster detection time, which changes proportionally with image resolution. The accordingly result can be taken from Table 7, which

Fig. 5 Object detection results of a single picture: (a) few object image, (b) medium object image, (c) dense object image, (d) few object image (YOLOv2), (e) medium object image (YOLOv2), (f) dense object image (YOLOv2), (g) few object image (YOLOv3), (h) medium object image (YOLOv3), (i) dense object image (YOLOv3).

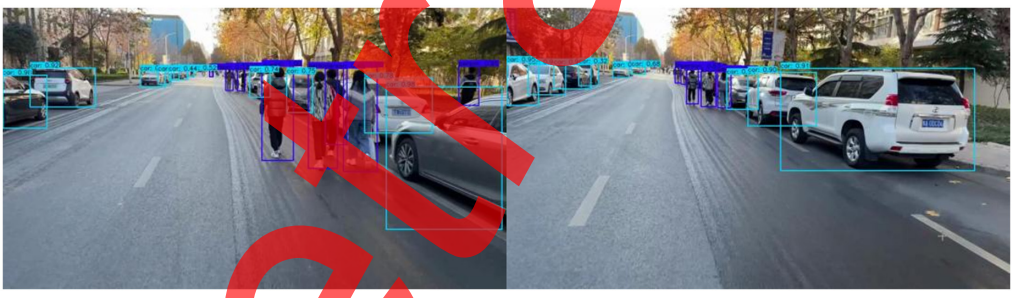

Fig. 6 Real-time video detection based on YOLOv2: (a) close-range object detection (YOLOv2) and (b) long-range object detection (YOLOv2).

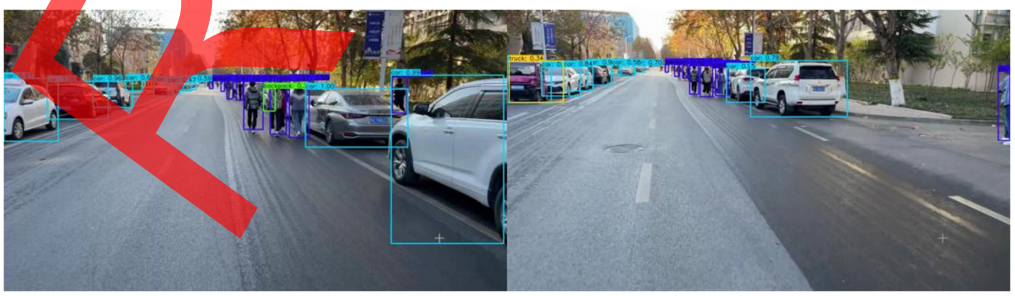

Fig. 7 Real-time video detection based on the proposed algorithm: (a) close-range object detection (YOLOv3) and (b) long-range object detection (YOLOv3).

Table 6 Detection time for the images with different resolutions.

Algorithm	320 × 320 detection time/ms	416 × 416 detection time/ms	608 × 608 detection time/ms
R-CNN	85	85	85
YOLOv2	31	37	64
YOLOv3	22	29	51
The proposed YOLOv3	19	24	39

Table 7 Detection time for the video.

Algorithm	Detection time/s	FPS	Floating point operations per second (Bn)
R-CNN	183.2	12	23.62
YOLOv2	76.8	67	59.36
YOLOv3	66.9	78	65.86
The proposed YOLOv3	61.8	87	79.38

demonstrates that the proposed YOLOv3 can identify 87 frames per second (FPS), which is clearly superior to the 12 FPS of the traditional R-CNN and superior to the 9 FPS of the traditional YOLOv3. In addition, the proposed YOLOv3 does 79.38 billion floating-point operations per second, which is three times more than the traditional R-CNN, and obviously surpasses the YOLOv2 and the traditional YOLOv3.

4 Conclusion

The application of YOLOv3 network in road dense object detection is studied in this paper, mainly including the deepening of backbone network layers in YOLOv3 network structure and the cross-layer addition and operation in residual network. Different convolutional layers can realize image detection of small, medium and large scale features, respectively. Thus, the traditional idea of deepening and improving the recognition rate by relying on network structure is fundamentally improved. It can provide higher recognition accuracy with fewer network parameters and network layers, and the detection speed is also taken into account. The proposed algorithm can accurately distinguish different types of vehicles and pedestrians, even pedestrians at the edge of the detection area. A confidence box can be used to mark the probability of object classification, and the confidence does not decrease with the increasing of road object density. By contrast, YOLOv2 model can only roughly divide the object type and cannot provide confidence information. In addition, the proposed training strategy of transfer training in a special dataset can also be extended to other dense object detection scenarios, which can achieve high target detection precision and is simple to train.

Acknowledgments

This research is financially supported by National Natural Science Foundation of China (Grant No. 61871176): Research of Abnormal Grain Conditions Detection using Radio Tomographic Imaging based on Channel State Information; Applied research plan of key scientific research projects in Henan colleges and Universities (Grant No. 22A510013): Research of Abnormal

Grain Conditions Detection using Cone-beam CT based on Convolutional Neural Network; Open subject of Scientific research platform in Grain Information Processing Center (Grant No. KFJJ2022011): Grain condition detection and classification-oriented semantic communication method in artificial intelligence of things; The Innovative Funds Plan of Henan University of Technology Plan (Grant No. 2020ZKCJ02): Data-Driven Intelligent Monitoring and Traceability Technique for Grain Reserves.

References

1. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 580–587 (2014).
2. G. Oliveira et al., "Automatic graphic logo detection via fast region-based convolutional networks," in *Int. Joint Conf. Neural Networks (IJCNN)*, pp. 985–991 (2016).
3. Y. Chen et al., "Domain adaptive faster R-CNN for object detection in the wild," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 3339–3348 (2018).
4. J. Redmon et al., "You only look once: unified, real-time object detection," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 779–788 (2016).
5. W. Liu et al., "SSD: single shot multibox detector," in *Eur. Conf. Comput. Vision*, pp. 21–37 (2016).
6. S. Bell et al., "Inside-outside Net: detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2874–2883 (2016).
7. Y. H. Peng, W. H. Zheng, and J. F. Zhang, "Deep learning-based on-road obstacle detection method," *J. Comput. Appl.* **40**(8), 2428–2433 (2020).
8. J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," in *Comput. Vision Pattern Recognit.*, pp. 01–06 (2018).
9. L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: a survey," *IEEE Trans. Neural Networks Learn. Syst.* **26**(5), 1019–1034 (2014).
10. K. Huang and W. Chang, "A neural network method for prediction of 2006 World Cup Football Game," in *Int. Joint Conf. Neural Networks (IJCNN)*, pp. 1–8 (2010).
11. O. Oyedotun et al., "Training very deep networks via residual learning with stochastic input shortcut connections," *Lect. Notes Comput. Sci.* **10635**, 23–33 (2017).
12. B. Zhu, M. F. Huang, and D. K. Tan, "Pedestrian detection method based on neural network and data fusion," *Autom. Eng.* **42**(11), 1482–1489 (2020).
13. B. Thomee et al., "The new data and new challenges in multimedia research," *Commun. ACM* **59**(2), 64–73 (2015).
14. S. Wang, M. Huang, and Z. Deng, "Densely connected CNN with multi-scale feature attention for text classification," in *IJCAI'18: Proc. 27th Int. Joint Conf. Artif. Intell.*, pp. 4468–4474 (2018).
15. S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. Twenty-Third AAAI Conf. Artif. Intell.*, vol. 8, pp. 677–682 (2008).
16. E. Rezende et al., "Malicious software classification using transfer learning of ResNet-50 deep neural network," in *16th IEEE Int. Conf. Mach. Learn. and Appl. (ICMLA)*, IEEE, pp. 1011–1014 (2017).
17. M. Wang et al., "Intelligent classification of ground-based visible cloud images using a transfer convolutional neural network and fine-tuning," *Opt. Express* **29**(25), 41176–41190 (2021).
18. E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning convolutional neural networks for fine art classification," *Expert Syst. Appl.* **114**, 107–118 (2018).
19. A. Majee, K. Agrawal, and A. Subramanian, "Few-shot learning for road object detection," in *AAAI Workshop Meta-Learn. and MetaDL Challenge*, PMLR, pp. 115–126 (2021).
20. R. Xu et al., "Opv2v: an open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Int. Conf. Rob. and Autom. (ICRA)*, IEEE, pp. 2583–2589 (2022).

21. E. Dogo et al., "A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks," in *Int. Conf. Comput. Tech., Electron. and Mech. Syst. (CTEMS)*, IEEE, pp. 92–99 (2018).
22. K. Zhang et al., "Instance transfer subject-dependent strategy for motor imagery signal classification using deep convolutional neural networks," *Comput. Math. Methods Med.* **2020**, 1683013 (2020).
23. Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 6154–6162 (2018).
24. H. Wang et al., "A comparative study of state-of-the-art deep learning algorithms for vehicle detection," *IEEE Intell. Transp. Syst. Mag.* **11**(2), 82–95 (2019).
25. H. Mao, X. Yang, and W. J. Dally, "A delay metric for video object detection: what average precision fails to tell," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 573–582 (2019).

Chunhua Zhu is a professor, doctoral supervisor. Her research interest include intelligent signal and information processing, advanced detection technology and abnormal recognition, etc.

Jiarui Liang is a master's student. His research focuses on deep learning-based object detection.

Fei Zhou is a lecturer, PhD. His research interests are signal processing, target detection and image processing.