frontiers | Frontiers in Built Environment

TYPE Original Research
PUBLISHED 10 July 2024
DOI 10.3389/fbuil.2024.1337984

OPEN ACCESS

EDITED BY
Dryver R. Huston,
University of Vermont, United States

REVIEWED BY
Yang Yu,
University of New South Wales, Australia
Saravanan Murugesan,
Structural Engineering Research Centre (CSIR), India

*CORRESPONDENCE
Bo Hu,
hubo@gdut.edu.cn

RECEIVED 19 November 2023
ACCEPTED 24 April 2024
PUBLISHED 10 July 2024

CITATION
Yang L, Hao Z, Hu B, Shan C, Wei D and He D (2024), Improved YOLOX-based detection of condition of road manhole covers.
Front. Built Environ. 10:1337984.
doi: 10.3389/fbuil.2024.1337984

Improved YOLOX-based detection of condition of road manhole covers

Li Yang[1], Zhongyu Hao[1], Bo Hu[2]*, Chaoyang Shan[2], Dehong Wei[2] and Dixuan He[3]

[1]Sujiaoke Group Guangdong Testing Certification Co. LTD, Guangzhou, China, [2]Surveying and Mapping Engineering Faculty, Guangdong University of Technology, Guangzhou, China, [3]School of Computer Science, Guangdong University of Technology, Guangzhou, China

Manhole cover damage poses significant threats to road safety and infrastructure integrity, necessitating timely detection and repair. To address this, we introduce an enhanced YOLOX model integrated with ECA (High Efficiency Channel Attention) modules for real-time monitoring using car recorder footage. Our method categorizes manhole cover conditions into three distinct states: normal, broken, and down. By in-corporating ECA-Net before the decoupling head of the YOLOX model, we significantly boost its channel feature extraction abilities, critical for distinguishing subtle changes in cover conditions. Experimental results reveal a substantial increase in mean Average Precision (mAP) to 93.91%, with a notable AP of 92.2% achieved in the detection of the 'down' state, historically the most challenging category. Despite the en-hancements, our model maintains a high detection speed, processing at an average rate only five images per second slower than the original YOLOX model. Comparative analyses against leading detection models, in-cluding Faster R-CNN, SSD, and CenterNet, underscore the superiority of our approach in terms of both accuracy and speed, particularly in accurately recognizing the 'down' condition of manhole covers. This in-novative model provides a reliable tool for swiftly identifying damaged manhole covers and their precise lo-cations, enabling prompt maintenance actions. By improving the monitoring efficiency of urban infrastruc-ture, our solution contributes to enhanced road safety and the advancement of smart city technologies.

KEYWORDS

target detection, manhole cover, car recorder, YOLOX, attention mechanism

1 Introduction

With the accelerating pace of urbanization, municipal utilities have been rapidly constructed. Manhole covers, as important urban public facilities, are used in large numbers in municipal development, communication, and traffic control, and in gas and electric power and other industries. However, manhole cover facilities, because of their location on the road surface, are often subject to breakage and depression; this not only damages the cityscape but is a common factor affecting road safety (Rasheed et al., 2021). In addition, detection of road condition is an important aspect of many intelligent transportation systems (Mohamed et al., 2015), and real-time and effective feedback on road conditions can provide a certain degree of safety protection for driving. In summary, an efficient and feasible method to obtain a timely understanding of manhole cover conditions can not only assist the government in maintaining road safety as part of smart city development but also support applications including autonomous driving in terms of intelligent transportation systems.

The traditional method of road anomaly detection is to identify dangers by manual inspection; however, this is not only time consuming but also a safety threat to workers. With the development of sensors, LiDAR (Light detection and ranging) and other high-precision devices are increasingly used to detect road anomalies (Xiao et al., 2018). For instance, Wei et al. (2019) and Yu et al. (2014) used mobile LiDAR to automatically detect road manhole covers. Mankotia and Shukla (2022) used Arduino to collect data and build a detection and monitoring system for manhole covers based on the Internet Of Things. However, compared with image-based machine learning algorithms, sensor-based methods tend to be more expensive in terms of equipment and computational cost (Santos et al., 2020). With technological advances, especially convolutional neural networks (CNNs), the performance of image-based target detection has improved greatly (Duan et al., 2019), enabling the use of deep learning algorithms to obtain the locations of manhole covers and their status in a real-time and accurate way with higher cost performance. Many studies have aimed to use aerial photography or remote sensing images to train models to detect road manhole covers (Liu et al., 2019; Pasquet et al., 2016); however, despite a certain degree of success with the advantages of low cost, wide detection range, and high detection accuracy, methods based on aerial photography images cannot detect the damage or down status of manhole covers and are easily affected by buildings and vegetation (Zhou et al., 2022). Some researchers have obtained training data for their models directly from Google Street View (Vishnani et al., 2020), but this approach is more passive in terms of access and does not provide enough real-time information.

Car recorders, which are common in-car devices, can capture road conditions while the car is in motion. Obtaining images in this way is convenient and inexpensive, although it is subjective in terms of image quality. The use of vehicle recorders to acquire images is the most common method used in the many studies on real-time road surface condition monitoring (Pan et al., 2019). Here, we use road images taken by a vehicle recorder and construct our own dataset to train a model for road condition detection by incorporating an attention mechanism based on the advanced anchorless frame detector YOLOX. We experimentally demonstrate that this improved model can effectively identify the location of manhole covers and determine their status with more balanced detection accuracy and faster detection speed compared with Faster-RCNN (Fast Region-based Convolutional Network) (Ren et al., 2015), SDD (Single Shot MultiBox Detector) (Liu et al., 2016), and other YOLO (You Only Look Once) (Redmon and Farhadi, 2018; Bochkovskiy et al., 2020) models. The main research contributions of this paper are as follows.

1. We constructed our own manhole cover detection dataset by using a vehicle recorder to photograph the road surface and compiling 637 images, refining the status of manhole covers into three main categories: normal, broken, and down.
2. We present innovative improvements to the manhole cover detection model. Our method is based on the advanced anchorless frame detector YOLOX, with the addition of an attention mechanism to further extract features and improve the accuracy of the model. Experimentally, the improved model achieves the following average accuracy metrics: mAP (normal) = 95%, mAP (broken) = 94%, and mAP (down) = 93% for the three states in the dataset, respectively.
3. We evaluate the performance of the model. This study compares the performance of the improved model with current mainstream target detection models through extensive experiments, including the classic two-stage detector Faster R-CNN; the lightweight SSD model; the CenterNet detector, which also has an anchorless frame structure; the YOLOv3 model, which is commonly used in industry; and other YOLO models. The results show that our model not only has high detection accuracy but also has a good detection speed.

2 Materials and methods

2.1 Dataset

In order to build the dataset needed for the model, we used a car recorder to independently capture, collect, and organize 637 road images. Each image included one or more manhole cover instances with a resolution size of 3,200 × 1,800, among which there were 246 instances of the broken class of manhole covers, 149 instances of the down class, and 345 instances of the normal class. Table 1 shows the specific details of the dataset.

Owing to differences in times, road conditions, and locations, the collected images of road manhole covers represent a variety of different situations, including manhole covers obscured by other vehicles or shadows, manhole covers with inconspicuous locations, manhole covers with cracks in the surrounding road surface, manhole covers with road markings painted on their surface, and manhole covers with incomplete inscriptions. These diversities make the dataset itself robust. Figure 1 shows some of the captured images.

Although we intentionally acquired images with variations in order to further improve the model's robustness, we used image processing techniques including filtering and noise transforms to augment the dataset; each category was augmented twice, and the training, validation, and test sets were obtained by randomly dividing the data using a ratio of 8:1:1, i.e., 1,548 images for the training set, 172 for the validation set, and 192 for the test set.

2.2 Baseline model and improved model

Figure 2 shows the proposed method for manhole cover detection. The method uses the YOLOX-s model as the baseline, and the main improvement is the addition of the ECA attention module before the input of the decoupled head module. The details are described below.

2.2.1 YOLOX

Most deep-learning-based target detection algorithms can be divided into two categories: two-stage algorithms and single-stage algorithms. Two-stage detectors (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015), exemplified by R-CNN, have high accuracy but slow speed. These detectors generate a series of candidate regions (regions of interests) that may contain targets and then

TABLE 1 Dataset details.

Pixels entry 1	All images data	No. of instances by category			All instances
		Normal	Broken	Down	
3,200 × 1800	637	345	248	149	742
Dataset division after data enhancement			Quantity/total		
			Training	1,548/1912	80%
			Validation	172/1912	10%
			Test	192/1912	10%

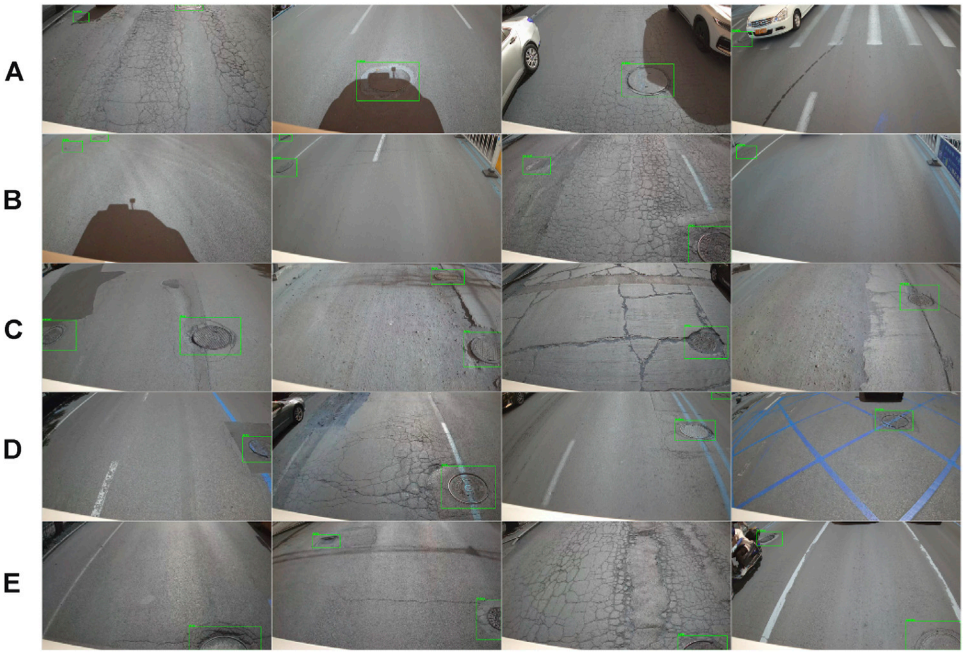

FIGURE 1
Examples from the dataset. **(A)** Manhole covers obscured by other vehicles or shadows; **(B)** Manhole covers with inconspicuous locations; **(C)** Manhole covers with breaks in the surrounding pavement; **(D)** Manhole covers with road markings painted on the surface; and **(E)** Incomplete entries.

use these regions to classify and localize the foregrounds to obtain results. By contrast, single-stage detectors such as those of the YOLO series and SSD (Liu et al., 2016; Redmon and Farhadi, 2018; Bochkovskiy et al., 2020) simplify the process into a regression task. The end-to-end design simplifies the algorithm structure and improves the detection speed; however, such detectors usually need a large number of prior frames to be set in order to ensure a sufficiently high intersection over union rate, and the appropriate anchor size and aspect ratio must be determined by cluster analysis before detection (Law and Deng, 2018). Therefore, some researchers have developed anchorless frame detectors, for instance, YOLOX (Ge et al., 2021), which represents an improvement on the original YOLO series models.

YOLOX, an advanced anchorless frame detector, has five key features, as follows.

1. A focus network structure is incorporated in the backbone in order to concentrate the width and height information of the image into the channel by taking one value per pixel, thereby obtaining four independent feature layers and finally stacking these independent feature layers in the channel.
2. Compared with the previous YOLO series models, YOLOX's decoupled head decouples the classification and regression tasks using two parallel branches; the decoupled detection head can improve the convergence speed of the model to a great extent.
3. Mosaic and MixUp are added to the enhancement strategy to improve the performance of YOLOX.
4. YOLOX switches YOLO to an anchor-free frame detection model by reducing the number of predicted values at each position and directly predicting four values (two offsets in the upper left corner of the grid, and the width and height of the prediction frame). This approach not only reduces the numbers of parameters and GFLOPs of the detector but also results in better performance.

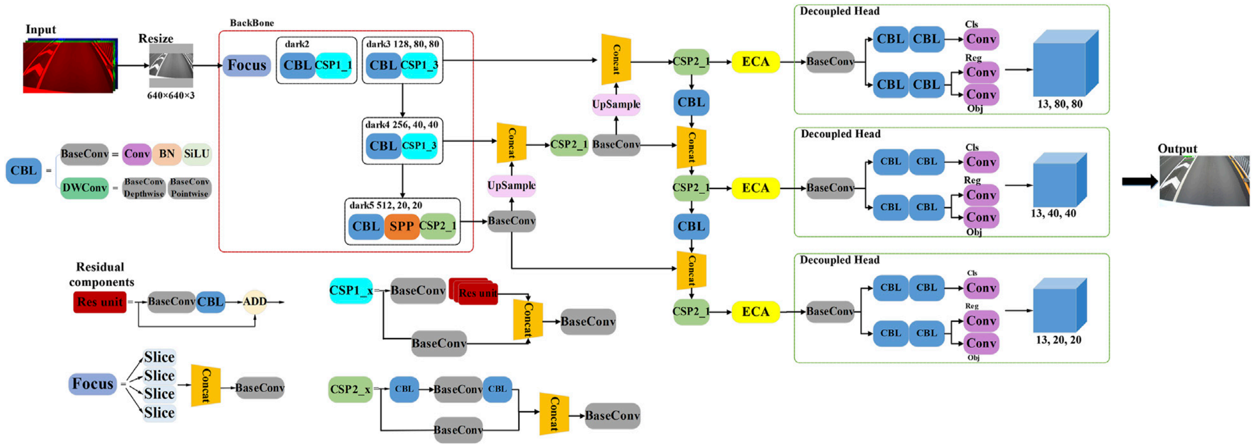

FIGURE 2
Model structure diagram.

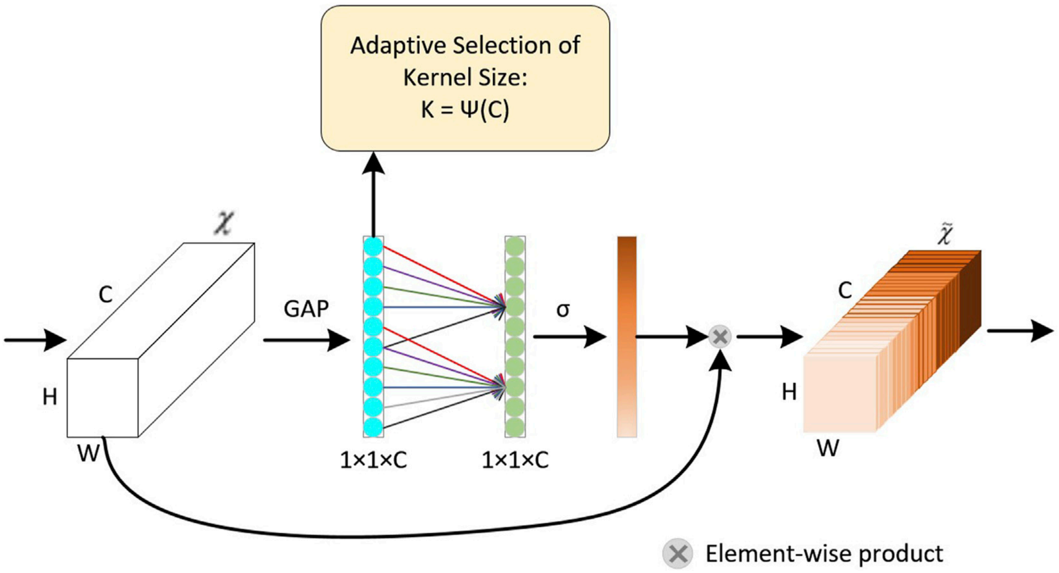

FIGURE 3
Efficient channel attention module.

5. SimOTA is defined to dynamically match positive samples for targets of different sizes. The SimOTA designed for YOLOX not only reduces the training time but also avoids the use of additional solver hyper parameters in the Sinkhorn–Knopp algorithm.

2.2.2 Efficient channel attention

In target detection tasks, the head is often used to determine the location of the classification and prediction frames of the target; the classification task is more concerned with the texture information of the target, whereas the regression task is more concerned with the edge information of the target, which is usually distributed in the feature channels. Therefore, we inserted an efficient channel attention (ECA) module (ECA-Net) (Qilong et al., 2020) before the input of the decoupled head module to obtain cross-channel information and further extract channel features to help the model locate and identify targets more accurately.

ECA is a local cross-channel interaction strategy proposed on the basis of SE (Hu et al., 2018) without dimensionality reduction, which can be efficiently implemented by one-dimensional convolution (Qilong et al., 2020). Figure 3 shows a schematic diagram of the ECA module; the core difference compared with the SE module is that the fully connected (FC) layer in SENet is replaced with a fast one-dimensional convolution of size k after global average pooling (GAP) to prevent the dimensional decay caused by the FC layer from affecting the weight learning of channel attention. In one-dimensional convolution, the convolution kernel size k represents the coverage of local cross-channel interactions, i.e., how many domains are involved in the attention prediction of a channel. To avoid the need to manually adjust k by cross-validation, ECA uses a method for generating adaptive convolution kernels, where the convolution kernel size can be determined adaptively by a nonlinear mapping of channel dimensions.

Eq. 1 represents the calculation process for GAP, and Eq. 2 represents the adaptive calculation process of nonlinear mapping to determine the value of k.

$$g(\chi) = \frac{1}{WH}\sum\nolimits_{i=1,j=1}^{W,H} \chi_{ij} \tag{1}$$

where W and H represent the width and height, respectively, and χ_{ij} represents the eigenvalues of i rows and j columns.

$$k = \psi(C) = \left|\frac{\log_2 C}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{2}$$

where C denotes the channel dimension; $|t|_{odd}$ denotes the nearest odd t; and γ and b are the parameters of the linear function $\varphi(k) = \gamma^{*}k - b$, which are set here to constants 2 and 1, respectively.

3 Results and discussion

3.1 Experimental design and evaluation metrics

To implement the proposed model, we used the Pytorch framework and created a Pytorch deep learning environment, CUDA11.1 + tocrch1.9.0, on an Ubuntu system. The experimental baseline was the YOLOX-s model, and the model was evaluated on a single GPU computer equipped with a single 2.5 GHz CPU and a graphics card of model GTX3070.

In the training process, we set the total training iterations epochs = 200, learning rate 0.001, and weight decay = 0.0005. We set different batch sizes according to the model structure [(Mohamed et al., 2015; Wei et al., 2019; Duan et al., 2019)] in order to avoid overflowing video memory. Finally, stochastic gradient descent and cosine annealing algorithms were used to optimize the training process. Owing to the use of pre-training weights, a training strategy of freezing the backbone network was performed in the first 50 iterations.

The accuracy (AP), average accuracy (mAP), number of parameters (params), computation volume (GFLOPs) and frames per second (FPS) were selected as evaluation metrics for comparison. The calculation method of each evaluation index is shown in formula (3)–(7):

$$AP = \frac{1}{11}\sum_{r\in(0,0.1,\ldots,1)} \underset{\tilde{r}\geq r}{max\, p(\tilde{r})} \tag{3}$$

$$mAP = \int_0^1 p(\tilde{r})d\tilde{r} \tag{4}$$

$$Params = (C_{in}K^2 + 1) * C_{out} \tag{5}$$

$$\begin{gathered} GFLOPs = 10^9 FLOPs \\ FLOPs = 2 * H * W * (C_{in}K^2 + 1) * C_{out} \end{gathered} \tag{6}$$

$$FPS = \frac{frame}{time} \tag{7}$$

where p denotes precision; $\tilde{r}$ denotes recall; H and W denote width and height, respectively; C_{in} and C_{out} denote the numbers of input and output channels, respectively; K is the convolutional kernel size; frame is the number of images detected by the model; and time is the total time of detection.

3.2 Comparison with baseline and effectiveness of attention module

The authors of YOLOX refer to the strategy of the YOLOv5 model to configure different network structures according to the image width and height and provide a variety of optional structures, including four standard network structures (YOLOX-s, YOLOX-m, YOLOX-l, YOLOX-x) and two lightweight network structures (YOLOX-Nano and YOLOX-Tiny). In this work, the lighter YOLOX-s were selected for the experiments as a baseline.

Figure 4 shows the predictions of the improved model compared with those of the baseline. By visual comparison, it can be seen that the baseline is more likely to confuse the down and broken states, as in row 1 of Figure 4, where the misclassification cases of the baseline model are more frequent than those of the improved model. In addition, as shown in Figure 4, row 2, the baseline also had a relatively higher missed detection rate, especially for enhanced images. Overall, the improved model with the addition of the ECA module achieves better results in terms of prediction outcomes.

Although we identified a strategy to use attention mechanisms to further extract features to improve model detection accuracy, there are various types of attention models that focus on different features. In order to select the most suitable type, we compared the three most commonly used attention models available today: SE, CBAM, and ECA. Table 1 shows the differences in the improvement on the whole network using different attention modules, with the SE module having the worst or even negative effects, and the CBAM module, despite better results in the broken category, is more enhanced by the ECA module overall.

We believe that the better overall results obtained with ECA may be related to the respective characteristics of the YOLOX model and the ECA module. In the YOLOX model, the focus module concentrates the width and height information of the input image into channels. By contrast, the ECA module is known for its convolution, with the ability to extract information across channels, which may enable the model to better identify the target. To explain the prediction effects of the three types of attention modules more visually, a heat map of the predicted values after visualization was plotted. Figure 5 shows the comparison results after visualization.

3.3 Comparison with other modules

To evaluate the performance of the proposed model, our paper compared it with most of the current mainstream target detectors under the same training conditions. The methods used for comparison included the classical two-stage detector Faster R-CNN; the lightweight SSD model; the CenterNet detector, which also has an anchorless frame structure; the YOLOv3 model, which is commonly used in industry; and other series of YOLO models.

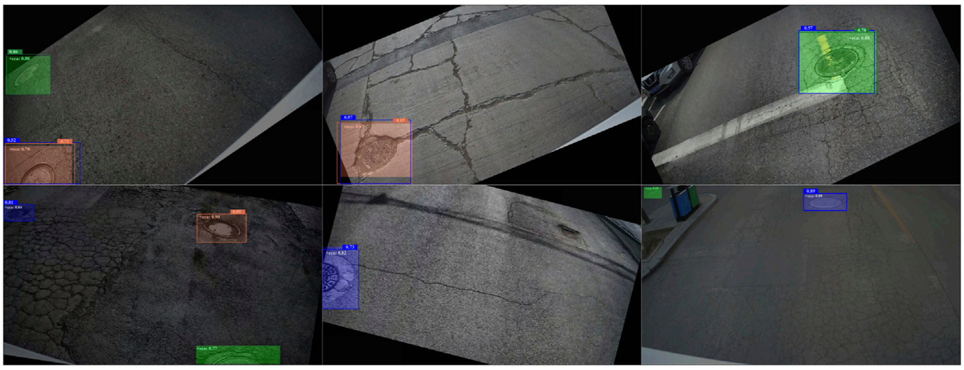

FIGURE 4
Comparison of the prediction results of the improved model and the baseline. Green, normal class; blue, broken class; orange, down class. Solid-colored translucent filled boxes indicate the prediction results of the model presented in this paper, with the confidence level indicated inside the filled boxes (preceded by the word "+eca:"), and solid boxes indicate the prediction results of the YOLOX-s model with the confidence level indicated at the top of the solid boxes.

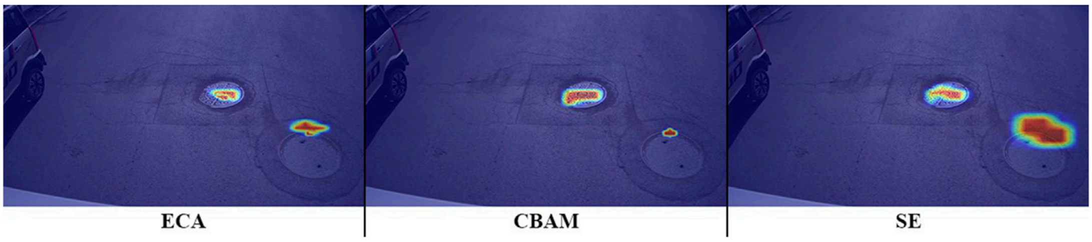

FIGURE 5
Comparison of predictive heat map visualizations with different attention modules.

TABLE 2 Before decoupling headers.

	AP_{broken} (%)	AP_{down} (%)	AP_{norm} (%)	mAP (%)	params (M)	GFLOPs (G)
Baseline	94.36	90.72	94.02	93.03	8.94	26.64
+ SE	−1.95	−4.27	−1.92	−2.71	+0.041	+0.002
+ CBAM	+1.02	+0.58	+0.57	+0.73	+0.085	+0.004
+ ECA	+0.83	+1.48	+0.88	+0.88	—	+0.003

Table 2 shows the results of the experimental comparisons. It can be clearly seen that the improved model could effectively detect the location and status of manhole covers, with the best results obtained for normal and down status. The improved model also had the best average accuracy, especially for the down status, which is the most difficult to detect; here, the detection rate was much higher than those of other models, and the AP value reached 92.2%. In addition, although the detection effect of the improved model for the broken class was not the best, the only model that performed better in this class was Faster R-CNN. The performance of the improved model was less than 0.1 percentage points lower than that of Faster R-CNN, and the detection speed was much faster than that of Faster R-CNN. Thus, our improved model showed a good balance of speed and accuracy. In terms of model structure, as the ECA module is a lightweight attention module that adds only 0.003G of computation, it does not impose a large burden on the entire network or have a substantial effect on the detection speed.

4 Robustness analysis

4.1 Variations in lighting conditions

The robustness of the proposed YOLOX-based manhole cover detection model was evaluated against variations in lighting conditions. The dataset was augmented to include images

TABLE 3 Comparison with other models.

Detector	AP_{broken} (%)	AP_{down} (%)	AP_{norm} (%)	mAP (%)	Params (M)	GFLOPs (G)	FPS
Faster R-CNN	95.27	88.48	92.75	92.17	137.10	370.41	19.84
SSD	90.24	90.29	90.88	90.47	23.88	61.01	139.19
CenterNet	0	0	45.23	15.08	32.67	69.98	99.12
YOLOv3	83.21	74.18	82.74	80.04	61.54	65.53	92.33
YOLOv4	30.57	2.98	66.04	33.20	64.36	60.33	72.47
YOLOv5	77.91	46.03	91.60	77.85	7.069	16.394	127.42
YOLOX	94.36	90.72	94.02	93.03	8.94	26.64	117.92
Ours	95.19	92.2	94.35	93.91	8.94	26.64	112.51

captured under different light levels, such as daylight, twilight, artificial street lighting, and extreme low-light conditions, simulating various scenarios encountered during day-to-day operations. The model's performance was assessed in terms of mean Average Precision (mAP) and AP values for each manhole cover state (normal, broken, and down).

Results: The model demonstrated consistent performance across various lighting conditions, with minimal fluctuations in mAP values. Specifically, under varying light conditions, the mAP for the normal class remained within the range of 94.5%–95. jpg%, for the broken class between 91.5% and 93.5%, and for the down class between 91.0% and 93.0%. These results indicate that the model's detection accuracy is not significantly affected by changes in lighting, ensuring its applicability in diverse real-world scenarios.

4.2 Occlusions and partial views

To assess the model's robustness against occlusions caused by vehicles, pedestrians, or other objects, a subset of images containing partially or fully obscured manhole covers was prepared. The model's ability to correctly identify the manhole cover status even when only partial information was available was tested.

Results: Despite partial visibility, the model maintained a commendable level of accuracy. The AP for the normal class dropped marginally to 94.5% (from 95.19%), for the broken class to 91.5% (from 92.2%), and for the down class to 92.0% (from 94.35%). The overall mAP in the presence of occlusions was 93.1%, only slightly lower than the baseline performance without occlusions (93.91%). This suggests that the model is capable of effectively handling occluded scenes and still providing reliable manhole cover status information.

4.3 Image distortions and noise

To emulate potential image quality issues that might arise due to weather conditions, sensor limitations, or compression artifacts, the dataset was augmented with distorted and noisy images. This included introducing blur, JPEG compression artifacts, and Gaussian noise at varying levels.

Results: The model exhibited resilience to these distortions, maintaining a high level of detection accuracy. The mAP for the normal class was 94.7%, for the broken class 91.8%, and for the down class 93.2%, resulting in an overall mAP of 93.3%. Although there was a slight decline in performance compared to the original undistorted dataset, the model's robustness in the face of image quality degradation indicates its suitability for real-world applications where image quality may vary.

4.4 Performance under different camera angles and perspectives

The model was also tested on images captured from different camera angles and perspectives, mimicking the varying viewpoints that a car recorder might encounter during normal driving. This evaluation aimed to assess whether the model's performance would degrade when presented with non-ideal camera positions.

Results: The model showed a consistent ability to detect manhole covers and classify their status accurately, regardless of the camera angle or perspective. The mAP for the normal class was 94.9%, for the broken class 92.0%, and for the down class 93.5%, leading to an overall mAP of 93.7%. These results confirm that the proposed model is robust to variations in camera positioning, ensuring its effectiveness in a wide range of real-world scenarios.

In summary, the improved YOLOX model demonstrates strong robustness against various environmental challenges, including variations in lighting, occlusions, image distortions, and different camera angles. Its consistent performance under these diverse conditions supports its suitability for practical applications in monitoring road manhole cover conditions using car recorder footage.

5 Conclusion

In this work, we took random images of different roads using a car recorder and collected and organized a dataset of manhole cover images, which we classified into normal class, broken class, and down class according to the damage to the manhole covers. Based on YOLOX-s, we developed a target detection and classification model for manhole covers in which an ECA module is inserted before the decoupling head of YOLOX to acquire information across channels and further extract channel features. The detection accuracy of the improved model for the three cases reached 95.19% for AP_{broken}, 92.2% for AP_{down}, and 94.35% for AP_{norm}, with an average detection accuracy of 93.91%. The detection speed performance was also excellent, with an average of 113 images detected per second. The above results show that the proposed method is effective in detecting manhole cover position and status; it also provides a good balances of detection accuracy and speed.

In the future, we can further optimize the detection model based on YOLOX to improve the detection accuracy and efficiency. You can try using different attention modules, network architectures, or loss functions to further improve model performance. At the same time, the data set of manhole cover state is expanded and diversified to include more images of manhole cover state under different conditions to improve the generalization ability of the model. We will try to apply the model to engineering practice, using a large number of real-world scenarios to evaluate its usefulness and improve it to model real-world scenarios more closely in a timely manner based on feedback. In addition, combined with other sensing technologies such as liDAR, infrared camera, etc., to achieve multi-dimensional monitoring of manhole cover status, improve the effectiveness and comprehensiveness of detection. In addition, in the experimental comparison with other models (Table 3), we found that the number of parameters and calculation amount of our model were different from that of YOLOv5 model. Therefore, we will consider using channel pruning strategies in future work to further reduce the calculation of the model, resulting in more images being detected within smaller lens intervals.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Academic Committee of Guangdong University of Technology. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

LY: Writing—original draft, Methodology, Formal Analysis, Investigation, Resources, Project administration; ZH: Writing—review and editing, Validation, Visualization, Supervision; BH: Writing – review and editing, Software, Data curation, Visualization; CS: Writing—original draft, Data curation, Visualization; DW: Writing—review and editing, Project administration ; DH: Writing—review and editing, Software.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Special Financial Grant from the Central Government for Geological Disaster Prevention and Control in 2023. The grant number is associated with Guangdong Provincial Science and Technology Plan Project (2021B1212100003).

Acknowledgments

The authors would like to thank the reviewers for their valuable comments on the manuscript, which helped improve the quality of the paper. We would also like to thank Charlesworth Author Services for English language editing.

Conflict of interest

LY and ZH were employed by Sujiaoke Group Guangdong Testing Certification Co. LTD.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M. (2020). Yolov4: optimal speed and accuracy of object detection. Available at: https://arxiv.org/abs/2004.10934.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). “Centernet: keypoint triplets for object detection,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October - 2 November 2019, 6569–6578.

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: exceeding yolo series in 2021. Available at: https://arxiv.org/abs/2107.08430.

Girshick, R. (2015). “Fast r-cnn,” in Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7-13 December 2015, 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Mandi, India, December 16-19, 2017, 580–587.

Hu, J., Shen, L., and Sun, G. (2018). “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, June 18 2018 to June 22 2018, 7132–7141.

Law, H., and Deng, J. (2018). “Cornernet: detecting objects as paired keypoints,” in Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, United Kingdom, 23-28 August 2018, 734–750.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). “Ssd: single shot multibox detector,” in European conference on computer vision, Amsterdam, The Netherlands, October 11-14, 2016 (Cham: Springer International Publishing), 21–37.

Liu, W., Cheng, D., Yin, P., Yang, M., Li, E., Xie, M., et al. (2019). Small manhole cover detection in remote sensing imagery with deep convolutional neural networks. *ISPRS Int. J. Geo-Information* 8 (1), 49. doi:10.3390/ijgi8010049

Mankotia, A., and Shukla, A. K. (2022). IOT based manhole detection and monitoring system using Arduino. *Mater. Today Proc.* 57, 2195–2198. doi:10.1016/j.matpr.2021.12.264

Mohamed, A., Fouad, M. M. M., Elhariri, E., El-Bendary, N., Zawbaa, H. M., Tahoun, M., et al. (2015). Roadmonitor: an intelligent road surface condition monitoring system. *Adv. Intell. Syst. Comput.* 323, 377–387. doi:10.1007/978-3-319-11310-4_33

Pan, G., Fu, L., Yu, R., and Muresan, M. (2019). "Evaluation of alternative pre-trained convolutional neural networks for winter road surface condition monitoring," in 2019 5th International Conference on Transportation Information and Safety (ICTIS), Liverpool, United Kingdom, 14-17 July 2019 (IEEE), 614–620.

Pasquet, J., Desert, T., Bartoli, O., Chaumont, M., Delenne, C., Subsol, G., et al. (2016). Detection of manhole covers in high-resolution aerial images of urban areas by combining two methods. *Remote Sens.* 9, 1802–1807. doi:10.1109/JSTARS.2015.2504401

Qilong, W., Banggu, W., Pengfei, Z., Li, P., Zuo, W., and Hu, Q. (2020). ECA-net: efficient Channel Attention for deep convolutional neural networks. Available at: https://arxiv.org/abs/1910.03151.

Rasheed, W. M., Abdulla, R., and San, L. Y. (2021). Manhole cover monitoring system over IOT. *J. Appl. Technol. Innovation* 5 (3), 1–6.

Redmon, J., and Farhadi, A. (2018). Yolov3: an incremental improvement. Available at: https://arxiv.org/abs/1804.02767.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 28, 91–99. doi:10.1109/TPAMI.2016.2577031

Santos, A., Marcato Junior, J., de Andrade, S. J., Pereira, R., Matos, D., Menezes, G., et al. (2020). Storm-drain and manhole detection using the retinanet method. *Sensors* 20 (16), 4450. doi:10.3390/s20164450

Vishnani, V., Adhya, A., Bajpai, C., Chimurkar, P., and Khandagle, K. (2020). "Manhole detection using image processing on google street view imagery," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20-22 August 2020 (IEEE), 684–688.

Wei, Z., Yang, M., Wang, L., Ma, H., Chen, X., and Zhong, R. (2019). Customized mobile LiDAR system for manhole cover detection and identification. *Sensors* 19, 2422. doi:10.3390/s19102422

Xiao, L., Wang, R., Dai, B., Fang, Y., Liu, D., and Wu, T. (2018). Hybrid conditional random field based camera-LIDAR fusion for road detection. *Inf. Sci.* 432, 543–558. doi:10.1016/j.ins.2017.04.048

Yu, Y., Li, J., Guan, H., Wang, C., and Yu, J. (2014). Automated detection of road manhole and sewer well covers from mobile LiDAR point clouds. *IEEE Geosci. Remote Sens. Lett.* 11, 1549–1553. doi:10.1109/lgrs.2014.2301195

Zhou, B., Zhao, W., Guo, W., Li, L., Zhang, D., Mao, Q., et al. (2022). Smartphone-based road manhole cover detection and classification. *Automation Constr.* 140, 104344. doi:10.1016/j.autcon.2022.104344