Wang *et al.*
EURASIP Journal on Advances in Signal Processing (2023) 2023:58
https://doi.org/10.1186/s13634-023-01021-2

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

A lightweight YOLOX-based model for detection of metallic gaskets on High-speed EMU wheelset

Tengfei Wang[1*], Chongfei Zhu[1], Xiaoning Wang[1], Shuxia Li[2] and Yanchen Hu[1]

*Correspondence:
wtf429@126.com

[1] CRRC QINGDAO SIFANG CO.,LTD., Qingdao 266111, China
[2] Wuxi Xuelang Data and Manufacturing Technology Co., Ltd., Wuxi 214000, China

Abstract

In order to solve the requirements of real-time detection on metallic gaskets of wheelset assembly under the condition of limited hardware resources, here we propose an improved lightweight YOLOX-like network that can be deployed on mobile devices with low computing resource consumption. Firstly, the basic components of ShuffleNetV2 network structure are trimmed to reduce the computation intensity and the number of training parameters. Thus the lightweight ShuffleNetV2 is used as the backbone structure of YOLOX. The prediction layer for 20 × 20 size large targets is also removed from the head of this structure to further reduce the calculation time and improve the inference speed. Besides, the channel-attention enhancement module efficient channel attention is added into the network to improve the capability of feature extraction and the accuracy of target detection. Finally, the verification of inference for the proposed model is carried out on mobile terminal devices.The results show that the improved lightweight algorithm proposed in this paper not only ensures the detection accuracy, but also greatly reduces training parameters and computing resources, and it particularly can be run rapidly on mobile terminal devices with low-cost of computation.

Keywords: High-speed EMU wheelset, Bolt metal gasket, Lightweight neural network, Object detection

1 Introduction

The bolt fastening states on high-speed EMU train wheelset are very important for the security of train running on railways. A metal gasket in use is always attached to a bolt, both of which ensure the condition of whether the gasket has an antiloosening effect on the bolt. It is necessary to configure additional inspectors to check and record the tightness state of antiloose gaskets and bolts. Repeated operations lead to a shocking waste of manual labor and time, and manual inspection probably even leads to the consequences of missing and mischecking. Therefore, deep learning-based computer vision technology is adopted to actualize automatic detection for antiloose gaskets. The traditional target detection algorithm is mainly based on manual feature extraction. The process of the traditional target detection algorithm can generally

be summarized into three parts. First, selecting the region of interest that may contain objects. Second, performing feature extraction on the region that may contain objects. Last, classifying the extracted features. For example, HOG (Histogram of Oriented Gradients) detector is an important improvement of scale invariant feature transform and shape contexts. In order to balance feature invariance (including translation, scale, illumination, etc.) and nonlinear (distinguish different object classes), it needs to improve detection accuracy by computing overlapping local contrast normalization on a dense grid of evenly spaced cells. DPM (Deformable Parts Model) is a SOTA (State of The Art) algorithm in traditional object detection algorithms proposed in 2008. The algorithm consists of a rootfilter and multiple auxiliary partfilters, which are improved detection accuracy by hard negative mining, bounding box regression and context priming techniques. But it cannot adapt to large rotations, so the stability is poor. The traditional target detection algorithm based on manually extracted features has the following shortcomings: the recognition effect is not good enough and the accuracy is low; the huge computation and the operation speed is slow.

With the development of deep learning principles and the improvement of hardware computing power, the large models and big data techniques are extensively utilized to improve the performance of algorithms. Meanwhile, in the feature extraction networks, the numbers of layers are sharply increased and the network structures are becoming more and more complicated. It obviously will cost a large amount of computing resources for the operation of an algorithm deployed onto the terminal device.

Till now, the objection detection algorithms are mainly divided into two categories: the one-stage case and the two-stage case. The representatives of one-stage models are YOLO series [1–5], SSD [6, 7] and RetinaNet [8], whereas the RCNN series models [9–11] based on the analysis of candidate regions are the classic two-stage models [12, 13].

The one-stage algorithms can detect objects at a high rate of speed, but the detection accuracy is usually lower compared with those of the two-stage methods by reason of the complexity of neural network structures. The networks of YOLO series have undergone an iterative evolution from primitive YOLOv1 to the latest YOLOX [14], where the latter one has taken into account both the accuracy and the speed of inference for detection, and it has performed best on varieties of data sets.

The lightweight feature extraction network is able to reduce the amount of model parameters and improve the inference speed of a model by designing a convolutional kernel with a smaller amount of parameters and a more concise network structure, while hardly affecting the accuracy of the model, such as MobileNet [15], MobileNetV2 [16], ShuffleNet [17], ShuffleNetV2 [18], CondenseNet [19], etc. The MobileNet uses separable convolutions instead of traditional convolutions to bring down the amount of model parameters by combining depthwise convolution and pointwise convolution. The ShuffleNet ensures that the input of the next set of group convolutions comes from different groups by "reorganizing" the feature maps obtained by group convolutions, so that the information can be passed through different groups. The CondenseNet proposes a new feature map extraction method, which reduces the amount of model' s calculation through efficient linear operations. But all the existing methods still cannot perform well enough when deployed in the detection environment for railway wheel gaskets.

In this paper, a lightweight YOLOX like object detection network is proposed, and a lightweight ShuffleNetV2-based structure is designed as being the backbone of the YOLOX module. The prediction layer for size 20×20 objects is removed to further reduce the amount of calculation, and the channel attention enhancement module ECA [20, 21] is added into the backbone to improve the feature extraction ability and the detection accuracy. The amount of model parameters is evidently reduced and the speed of model inference is indeed accelerated on the premise of ensuring the precision of detection through the improved lightweight YOLOX like method. Thus, the model can be agilely deployed on mobile devices with low-cost of computational resources.

2 Related work

In the field of SLAM research, the systems of visual simultaneous localization and mapping (SLAM) and visual odometry (VO) have been deeply studied by many scholars. Using vision sensors alone or in combination with inertial sensors, many excellent systems were born with steady improvements in accuracy and robustness. We investigate monocular and binocular visual inertial navigation SLAM systems relying on maximum a posteriori (MAP) estimation, as well as a complete multimap SLAM Atlas system, and find that ORB-SLAM3 as a new visualization reference and a visualized inertial open source SLAM library is undoubtedly excellent.

The baseline of the essential network described in this paper is established on the basis of YOLOX, which is an anchor free model with YOLOv3 being the main structure. The key modules and elements of YOLOX can be introduced as follows.

- Baseline of YOLOv3: YOLOX [14] employs YOLOv3 [3] as the baseline, in which the Darknet53 is applied as the principal structure without FC layer but with an SPP module. Besides, the EMA weight updates, Cosine learning rate mechanism, IoU loss, IoU aware branching are also used along with fine tune training strategy. The cls and obj branches are trained by BCE loss, and the reg branch is trained with IoU loss. In addition, the Random ResizedCrop is removed, while the RandomHorizontalFlip, ColorJitter and multiscale data augmentation are added into use in the preprocessing of training stage.
- Decoupling in head for detection: The detection head of YOLOv3 is in a coupled state and YOLOX uses a decoupling head to replace the head for detection in previous YOLO series networks. The decoupling head is designed like this: First, 1×1 convolution is used to unify the original feature maps with different channel numbers to be 256. Then, two parallel branches are used for classification and regression, respectively, where each branch uses two 3×3 convolutions consecutively. Meanwhile the IoU method is added into the regression branch.
- Data augmentation: At the beginning of training, two data augmentation methods, Mosaic and Mixup are used based on the original YOLOv3. The augmentation for data will be turned off during the final 15 training epochs.
- Anchor free: YOLOX has propelled the development of YOLO series networks from anchor based [11] to anchor free [22, 23]. It is really simple to convert YOLO framework to the anchor free form, where the prediction of anchors changes from 3 sets to

only 1 set with 4 values, which are two coordinate offsets for the upper left corner of the grid and the height and width values of the predicted box.

- Label matching strategy: The YOLOX chooses the improved OTA (optimal transport assignment) [24] as the label assignment strategy, where OTA analyzes the label assignment from a global perspective and transforms it into an optimal transportation problem [25–27]. Due to the fact that the optimal transportation problem brings additional 25% time cost of training, the authors of YOLOX simplified it to a dynamic topk strategy, named SimOTA [14], to obtain an approximate solution. The SimOTA at first computes the pairwise match, which is represented by the cost or quality of each predicted GT (ground truth) pair. The function of SimOTA can be written as:

$$C_{ij} = L_{ij}^{\text{cls}} + \lambda L_{ij}^{\text{reg}} \tag{1}$$

where λ is the balance coefficient, L_{ij}^{cls} and L_{ij}^{reg} respectively are classification loss and regression loss between GT and the predicted value. The Top-k predicted values with the least cost in the fixed central region are then selected as positive samples of g_i for GT. Finally, the corresponding grids of these positive predictions are designated as positives, while the rest of the grids are designated as negatives. The values of k can be different for different GTs.

3 Improved model with simplifications

3.1 Improved backbone

In position recognition, several similar key frames in the Atlas maps are queried in the DBoW2 database, and the three visible key frames are geometrically verified through DBoW2 candidate key frames and 3D alignment transformation. It also needs to examine whether there are ORB key points whose descriptors match the map points' ORB descriptors. The ratio of distance to the second closest match also needs to be checked for further judgment.

The ShuffleNetV2 [18], proposed by Megvii Research Institute, is a lightweight neural network architecture that can be delicately deployed on mobile and edge terminals. The network structure with high test accuracy is clear and concise, and the memory access cost is considered at the beginning of the design. Therefore, it can be conveniently deployed on mobile devices with very low latency.

In order to further reduce the calculation amount and improve the speed of model inference, two network modules are designed to improve the ShuffleNetV2-based structure. In the standard ShuffleNetV2, one ordinary module and one downsampling module are respectively shown as in Fig. 1, where the 3×3 depthwise convolution (DWConv) and 1×1 convolution (Conv) are employed in the right branch of both of these two modules.

The use of 1×1 convolution before or after DWConv generally has two effects: one is to fuse the information between channels and cover the shortage of information fusion between channels by DWConv; the other is to reduce and increase the dimensions of feature maps. The 1×1 convolution here in ShuffleNetV2 architecture is only to fuse the interchannel information of the DW convolution, while it is somewhat redundant to design two 1×1 convolutions in one branch. Taking into account the weight reduction

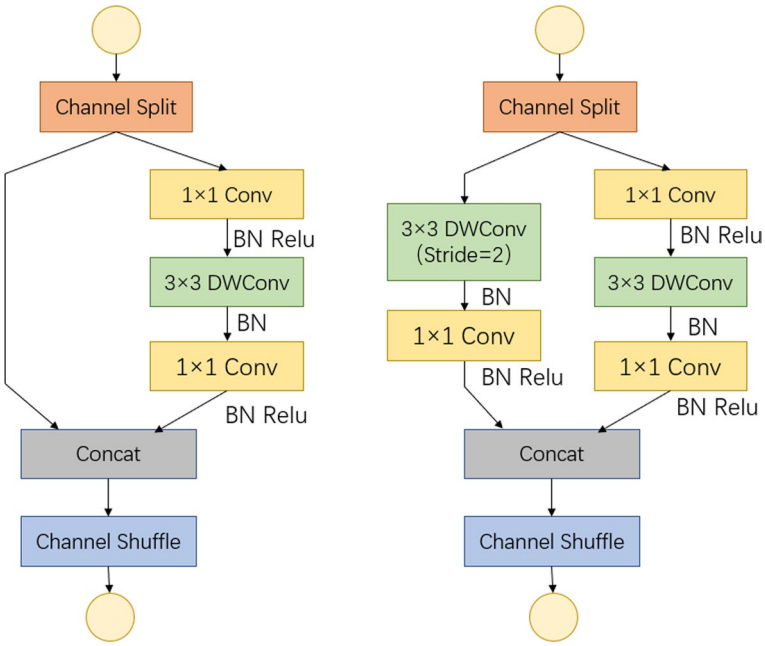

Fig. 1 The 3 × 3 depthwise convolution (DWConv) and 1 × 1 convolution (Conv) in two original modules of ShuffleNetV2.

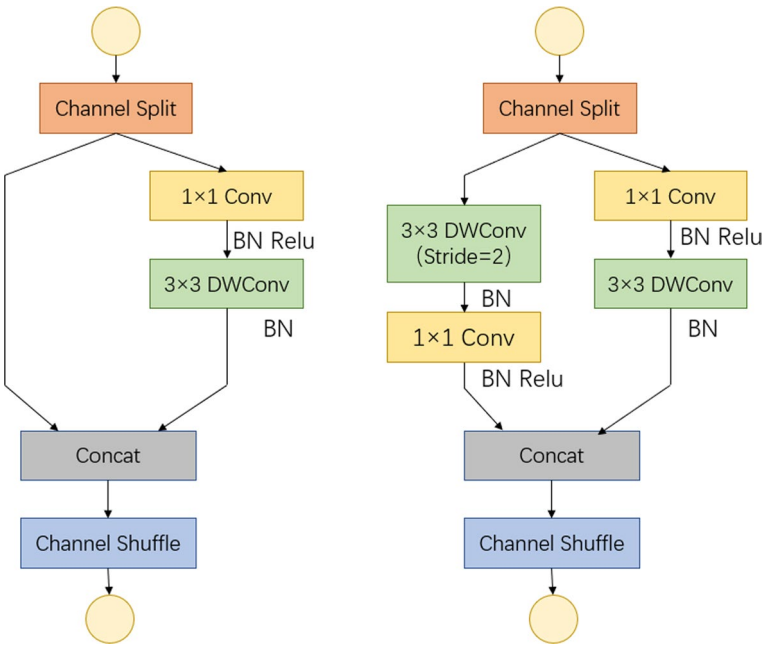

Fig. 2 The improved modules of ShuffleNetV2

effect, the 1 × 1 convolution can be removed after depthwise convolution in the right branches and the improved modules can be shown as in Fig. 2.

3.2 Lightweight attention mechanism

In transformer and CNN, the attention mechanism has achieved good results, while it has been added into the ShuffleNet structure to improve the network performance in this paper. The channel attention mechanism has been shown great potential in improving the performance of deep convolutional neural networks (CNNs). Avoiding dimensionality reduction and proper crosschannel interactions are important for learning high performance and efficient channel attention. The ECA (Efficient Channel Attention) module by considering each channel and its neighbors for local crosschannel exchange of information [20] is an ultralightweight channel attention module. The ECA can be effectively implemented by 1D convolutions of size k, where the convolution kernel size k represents the coverage of local crosschannel interactions, i.e., how many neighbors near this channel participate in the attention prediction of this channel. ECA attention module, which is a channel attention module, is often applied with visual models. It supports plug and play, i.e., it can perform channel feature enhancement on the input feature map and the final ECA module output without changing the size of the input feature map. The flow of the ECA model is as follows: firstly, the input feature map is compressed with spatial features, and a 1×1 feature map is obtained by global average pooling. Secondly, the compressed feature map is subjected to channel feature learning by 1×1 convolution. Finally, the feature map with channel attention is multiplied with the original feature map channel by channel, and finally the feature map with channel attention is output. To avoid manual tuning of k through crossvalidation, this paper proposes a method to adaptively determine k.

The grouped convolutions have shown that high dimensional (low dimensional) channels are proportional to long distance (short distance) convolutions for a fixed number of groups. Similarly, the coverage of crosschannel information interaction (that is, the kernel size k of one dimensional convolution) should also be proportional to the channel dimension C. It is well known that the channel dimension C is usually set to a power of 2. It is hoped that high dimensional channels can have longer interactions and low dimensional channels have relatively short interactions. In summary, and $k = 2$ is set in the experiment. The ECA attention added to the ShuffleNet module can be shown as in Fig. 3.

3.3 Lightweight detection head

The YOLOX network detects large, medium and small size targets based on three different feature layers of $20 \times 20, 40 \times 40$, and 80×80. The total loss of the model is composed of the sum of these three different scale loss functions, thus, poor training results on any scale will affect the whole training effect and the convergence speed. Considering that the gasket is relatively small and there is no large target in the inspected image, the 20×20 object prediction layer and its associated bottleneck connection are removed in order to make the network more suitable for the recognition task of gasket conditions.

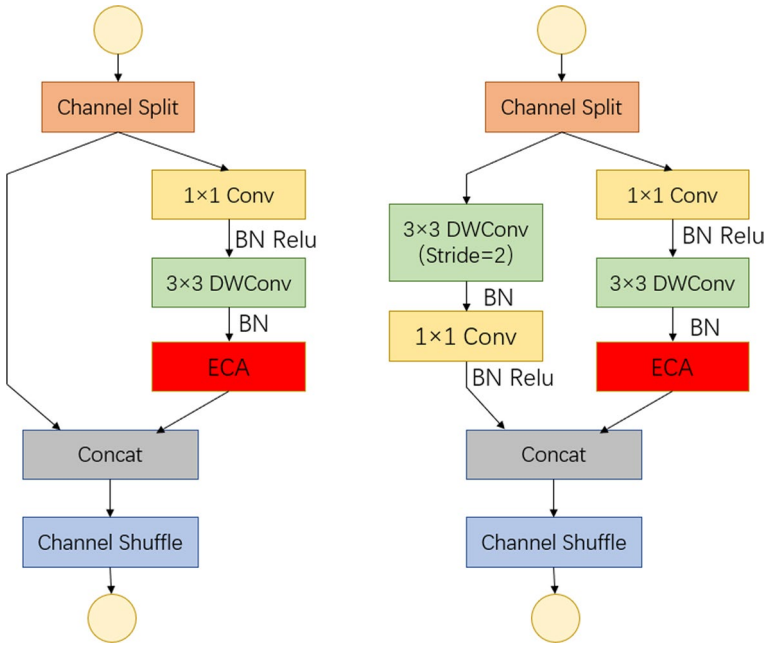

Fig. 3 The ECA added ShuffleNetV2 module

4 Experiments and analysis

The model inference hardware was configured with Intel® Core™ i5-8350U CPU@1.70 GHz.

Experimental data A total of 800 gasket images were collected in the workshop, and the images were marked by LabelImg software with labels "ok-dp" (attached tightly) and "ng-dp" (not attached tightly). There are about 2100 "ok-dp" labels and 2000 "ng-dp" labels, and the input image size for model training and testing is 416×416.

Experimental parameters The training learning rate is set to 0.01, the optimizer selects Momentum, the moving average parameter momentum is given as 0.9, only the weight parameter of the convolutional layer in the network uses L2 decay, the factor is 0.0005, the activation function chooses Relu, the input image size, the batch size and epoch are (416, 416), 128 and 300, respectively. We use 10-fold cross-validation to prevent the algorithm from overfitting and to improve the accuracy of the algorithm.

Evaluation index In this study, mAP (mean Average Precision) is selected as the evaluation index of algorithms. The precision indicates the proportion of correct predictions in the prediction results, the formula is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

The recall rate represents the proportion of the target that is correctly predicted by the model.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

where TP is the number of samples that are actually positive and determined to be positive, FP is the number of actually negative samples determined to be positive samples, and FN is the number of actually positive samples determined to be negative samples. Each category is integrated by the PR (precision recall) curve:

$$\mathrm{AP} = \int_0^1 p(r)\mathrm{d}\mathbf{r} \tag{4}$$

where the precision (p) denotes the proportion of true positive samples in the prediction results. Recall (r) represents the proportion of correctly predicted values among all positive samples. The average of APs of all categories is mAP.

$$\mathrm{mAP} = \frac{\Sigma_{i=1}^{N} \mathrm{AP}_i}{N} \tag{5}$$

The network after cutting Conv1*1 layer, described in Sect. 3.1, is named ShuffleNetLite, based on which the lightweight network with ECA attention module added, described in Sect. 3.2, is named ShuffleNetLite-ECA.

The comparison for ablation experiments are shown in Table 1, where different network structures are compared considering four dimensions of mean average precision (mAP), inference latency, calculation amount and parameter amount.

As for deployment on the mobile devices, the average precision and inference latency are two most important indicators to be valued. For the ShuffleNetLite, the mAP decreased from 36.28 to 35.80% with a drop of 0.48%, and the inference latency decreased from 19.06 to 13.52 ms with a speed increase of 5.53 ms, comparing with ShuffleNetV2. It can be seen that the model size and the amount of parameters and calculation amount are significantly reduced after the 1×1 convolution is correspondingly removed, while it has only a slight impact on the precision but can boost the inference speed evidently. Compared with ShuffleNetLite, the mAP of ShuffleNetLite ECA has increased from 35.80 to 38.40% with a gap of 2.60%, and the inference latency has slightly increased by 0.19 ms. From the experimental results, it effectively illustrated that the ShuffleNetLite-ECA model is able to improve the detection precision.

In the YOLOX-based experiments, similarly as YOLO-S and YOLO-Nano presented in paper [1], the YOLOX-S and YOLOX-Nano are considered as being the baselines for comparison and analysis. Besides, in this paper, YOLOX-LH means the lightweight YOLOX model with 20×20 prediction layer removed and bottleneck layer removed, which is mentioned in Sect. 3.3. The YOLOX-Shuf2 stands for the YOLOX network using ShuffleNetV2 as the backbone. The YOLOXLite represents our proposed lightweight network.

Table 1 Comparison of ShuffleNetV2 ablation experiments

Model	mAP	Latency (ms)	FLOPs (GB)	Param (MB)
ShuffleNetv2	36.28	19.06	1.47	1.28
ShuffleNetLite	35.80	13.53	1.03	0.915
ShuffleNetLite-ECA	38.40	13.72	1.03	0.916

Table 2 Experimental results of improved-YOLOX networks

Model	mAP	Latency (ms)	FLOPs (GB)	Param (MB)
YOLOX-S	38.81	386.14	26.80	9.0
YOLOX-Nano	35.70	23.35	1.08	0.91
YOLOX-LH	39.16	383.65	26.71	9.0
YOLOX-Shuf2	36.28	36.28	19.06	1.28
YOLOXLite	38.40	1.03	1.03	0.916

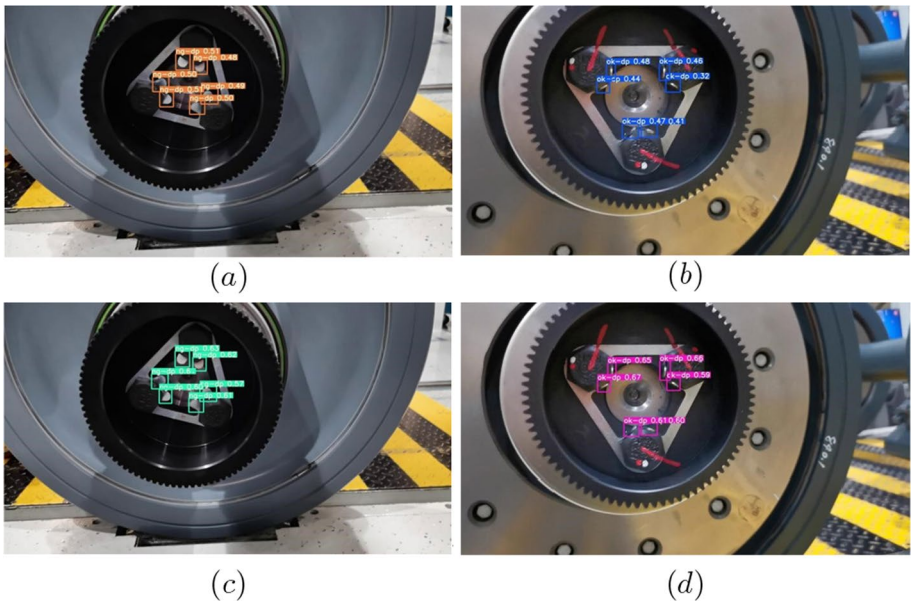

Fig. 4 The detection results of gasket states by using YOLOX and YOLOXLite. **a** Status "Not-attached" detected by YOLOX model; **b** status "Attached" detected by YOLOX model; **c** status "Not-attached" detected by YOLOXLite model; **d** status "Attached" detected by YOLOXLite model

The experimental results for network ablation can be shown in Table 2. Compared with YOLOX-Nano and YOLOX-Shuf2, the YOLOXLite has obviously improved the detection precision and inference speed. On the other hand, although the YOLOX-S has slightly higher detection accuracy, the amount of parameters and computation is too large and the inference time takes too long. The YOLOX-LH brings tiny increase in detection precision and inference speed compared with YOLOX-S, which means that the detection quality can be improved via removing 20×20 prediction layer and its related bottleneck layer.

The detection results of gaskets' states on High-speed EMU wheelset are demonstrated as in Fig. 4 by utilizing YOLOX and YOLOXLite, respectively. It can be evidently proven that our proposed method YOLOXLite is able to detect and identify the tightness of the antiloose gaskets and the bolt much better with higher confidence values. As shown in Fig. 4, due to the very small proportion of the detected object to the whole detected object, and the presence of shadow occlusion resulting in insufficient light, etc. our detection results also have missed and false detections, but compared to YOLOX, our YOLOXLite achieved significant detection results.

5 Conclusions

This paper proposes an improved lightweight YOLOX-wise model that can be deployed on mobile devices with low computing resource consumption. The original ShuffleNetV2 backbone has been cut to be a new lightweight structure applied in YOLOX. The 20×20 prediction layer for large objects detection are removed to further reduce the calculation amount and increase the detection precision, and the channel attention enhancement module has also been added into the model structure to increase the capabilities of feature extraction and object detection. All the experimental results illustrate that, the lightweight network proposed in this paper not only ensures the detection accuracy, but also greatly reduces the amount of parameters and the resource of algorithmic calculations. The proposed algorithm improves the model inference speed with low computing necessities and can even run fast on low power small mobile devices.

Acknowledgments
Not applicable.

Author contributions
Conceptualization, TW and CZ; methodology, TW; software, SL; validation, YH; formal analysis, TW; investigation, CZ; resources, TW; data curation, XW; writing-original draft preparation, TW; writing-review and editing, TW. All authors have read and agreed to the published version of the manuscript.

Funding
Not applicable.

Availability of data and materials
The data presented in this study are available on request from the corresponding author.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Approved.

Competing interests
The authors declare that they have no competing interests.

Received: 30 December 2022 Accepted: 16 May 2023
Published online: 27 May 2023

References

1. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 779–788
2. J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 7263–7271
3. J. Redmon, A. Farhadi, *Yolov3: an incremental improvement*. arXiv preprint arXiv:1804.02767 (2018)
4. A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, *Yolov4: optimal speed and accuracy of object detection*. arXiv preprint arXiv:2004.10934 (2020)
5. F. Zhou, H. Zhao, Z. Nie, Safety helmet detection based on yolov5, in *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*(IEEE, 2021), pp. 6–11
6. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in *European Conference on Computer Vision* (Springer, 2016), pp. 21–37
7. C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, *Dssd: deconvolutional single shot detector*. arXiv preprint arXiv:1701.06659 (2017)
8. T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2980–2988
9. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 580–587
10. R. Girshick, Fast r-cnn, in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1440–1448

11. S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28**, 66 (2015)
12. Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, Y. Fu, Rethinking classification and localization for object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 10186–10195
13. A. Li, X. Yang, C. Zhang, *Rethinking classification and localization for cascade r-cnn*. arXiv preprint arXiv:1907.11914 (2019)
14. Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, *Yolox: Exceeding yolo series in 2021*. arXiv preprint arXiv:2107.08430 (2021)
15. A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, *Mobilenets: efficient convolutional neural networks for mobile vision applications*. arXiv preprint arXiv:1704.04861 (2017)
16. A. Howard, A. Zhmoginov, L.-C. Chen, M. Sandler, M. Zhu, *Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation* (2018)
17. X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 6848–6856
18. N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: practical guidelines for efficient cnn architecture design, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 116–131
19. G. Huang, S. Liu, L. Van der Maaten, K.Q. Weinberger, Condensenet: an efficient densenet using learned group convolutions, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2752–2761
20. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Supplementary material for 'eca-net: efficient channel attention for deep convolutional neural networks, in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA* (2020), pp. 13–19
21. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141
22. H. Law, J. Deng, Cornernet: detecting objects as paired keypoints, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 734–750
23. Z. Tian, C. Shen, H. Chen, T. He, Fcos: fully convolutional one-stage object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 9627–9636
24. Z. Ge, S. Liu, Z. Li, O. Yoshie, J. Sun, Ota: optimal transport assignment for object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 303–312
25. Y. Ma, S. Liu, Z. Li, J. Sun, Iqdet: instance-wise quality distribution sampling for object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1717–1725
26. S. Zhang, C. Chi, Y. Yao, Z. Lei, S.Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9759–9768
27. X. Zhang, F. Wan, C. Liu, R. Ji, Q. Ye, Freeanchor: learning to match anchors for visual object detection. Adv. Neural Inf. Process. Syst. **32**, 66 (2019)

Publisher's Note