Obstacle detection: improved YOLOX-S based on swin transformer-tiny*

ZHANG Hongying, LU Chengjian, and CHEN Enyao**

College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China

(Received 10 February 2023; Revised 7 April 2023)

Aiming at the accuracy challenge in obstacle detection for autonomous driving, we propose an improved you only look once X-S (YOLOX-S) model based on swin transformer-tiny YOLOX-S (ST-YOLOX-S) for obstacle detection, which could detect multiple targets, including people, cars, bicycles, motorcycles, and buses. Our method mainly comprises two aspects as follows. To improve the capability of local feature extraction and then obtain more accurate detection for obstacles under real-world vehicle conditions, the existing backbone of YOLOX-S is replaced with the swin transformer-tiny backbone. We reduced the number of channels between the swin transformer and path aggregation-feature pyramid network (PA-FPN) from [96, 192, 384, 768] to [192, 384, 768], to decrease the computational cost and then make the swin transformer-tiny more compatible with the PA-FPN. Conclusively, on the popular COCO dataset, the proposed ST-YOLOX-S improves the detection mean average precision (mAP) by 6.1% when compared with YOLOX-S. Among the five types of obstacles that appear in simulated actual vehicle conditions, our ST-YOLOX-S also achieves superior performance compared to YOLOX-S. Furthermore, our method achieves significant performance over the YOLOv3 on obstacle detection, which shows the effectiveness of the proposed algorithm.

Document code: A **Article ID:** 1673-1905(2023)11-0698-7

DOI https://doi.org/10.1007/s11801-023-3018-9

In recent years, with the broad applications of computer vision in various fields, the research and development of autonomous driving have emerged. Autonomous driving systems are mainly implemented by combining vision with radar technology[1]. In the visual aspect, the primary goal is to detect obstacles on the road, such as pedestrians, cars, buses, motorcycles, bicycles, and other objects, utilizing object detection. The efficiency of autonomous driving is dependent upon the speed and accuracy of obstacle detection. A fast detection rate enables the driving system to respond promptly to potential hazards, while a high level of accuracy ensures the appropriate evasion of obstacles. The conundrum of balancing speed and accuracy in object detection has been a persistent challenge within the realm of autonomous driving for an extended period.

Since its introduction, the you only look once (YOLO) detection algorithm has gained widespread attention among the research community benefiting from its remarkable speed in object detection. In 2018, the authors of YOLOv3[2] introduced Darknet-53 as the benchmark backbone network in their study. In 2020, YOLOv4[3] introduced the CSPDarknet-53 backbone network, which was based on Darknet-53 and incorporated modifications to the structure. Specifically, the network was augmented with spatial pyramid pooling (SPP) components[4] at the end of the backbone network. It employed a path aggregation-feature pyramid network (PA-FPN)[5] in the neck region. In 2021, YOLOX[6] adopted the optional configuration feature from YOLOv5[7], enabling the creation of network structures with varying specifications, including YOLOX-S, YOLOX-M, YOLOX-L, and YOLOX-X. Furthermore, the authors innovatively incorporated a decoupled prediction head, the anchor free concept, and a dynamic positive sample matching strategy referred to as SimOTA[6]. Despite the satisfactory detection performance and inference speed demonstrated by YOLOX, the accuracy of its lightweight network, YOLOX-S, for autopilot systems remain suboptimal due to an insufficient number of convolutional layers. In the COCO-2017val set, the mean average precision (mAP) of YOLOX-S was found to be only 40.3%. The backbone network employed by YOLOX-S is the traditional convolutional neural network (CNN) CSPNet[8]. This network primarily focuses on capturing local feature information through its convolution kernel, ignoring the feature map output of the entire CNN architecture, thereby overlooking crucial global feature information.

In terms of the insufficiencies of YOLOX-S in autonomous driving situations, this study concentrates on incorporating an attention mechanism into the backbone

* This work has been supported by the Graduate Research Innovation Project of Civil Aviation University of China (No.2021YJS086).

** E-mail: carole_zhang@vip.163.com

architecture to focus on both global and local features. In 2017, Google presented a revolutionary model known as the transformer[9], which generated a significant impact on natural language processing and marked a significant turning point in the industry. The initial purpose of the design of this model was to enhance the efficiency of machine translation and this model employed the self-attention mechanism and position encoding as alternatives to conventional convolution blocks. In 2020, Google introduced the concept of the vision transformer (ViT) model[10] which demonstrated the capability of a pure transformer, being applied directly in a sequence of image patches, which exhibits remarkable results in image classification tasks. Compared with prior deep learning networks, the ViT has demonstrated remarkable speed-performance trade-offs across a wide range of computer vision tasks. However, the limitations of ViT are quite evident. The model inputs image patches as vectors into the transformer architecture, thereby flattening the spatial relationships within the image and ignoring its unique visual characteristics. This approach results in the loss of important structural information within the image blocks. In essence, the efficacy of local feature extraction is subpar. To address this issue, the Asian Research Institute introduced the swin transformer[11] in 2021 as a potential solution. The proposed method constructs a hierarchical transformer architecture through the implementation of a hierarchical stacking approach, which is an approach commonly utilized in CNNs and manifests in the form of a pyramid structure. The integration of CNNs and ViT has been demonstrated to be a clever strategy, as it not only enhances the connection between long-distance information, but also elevates the local feature extraction ability of ViT.

To summarize, the most widely used bounding-box-like object detectors in autonomous driving[12] are CNNs composed of stacked Conv blocks, such as YOLO, mask R-CNN[13], RefineNet[14], and faster-RCNN[15]. These detectors aim to reduce computational costs and increase detection speed, but they tend to compromise detection accuracy due to the reduction of kernel maps. On the other hand, ViT, which consists of encoder blocks, excels in linking long-distance information, but its local feature acquisition capability is limited, resulting in poor detection performance on pedestrians and covered vehicles. In light of this, we propose an improved detection algorithm, swin transformer-tiny YOLOX-S (ST-YOLOX-S), which leverages both local and global features from convolution or attention blocks for improving image segmentation. The ST-YOLOX-S model is illustrated in Fig.1.

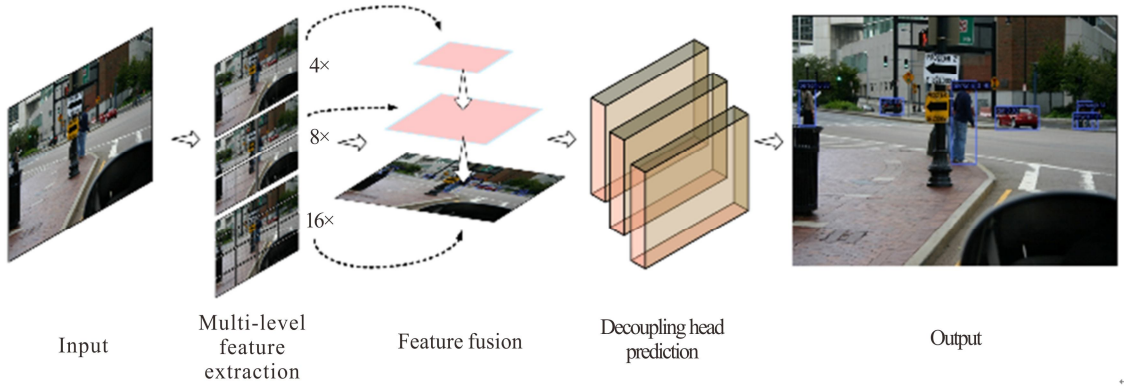

Fig.1 ST-YOLOX-S model diagram

In this paper, we present a novel approach to address the demands for real-time detection and high accuracy in autonomous driving. Our proposed solution, ST-YOLOX-S, integrates the swin transformer-tiny backbone network based on an attention mechanism and YOLOX-S detection algorithm. Furthermore, we reduce the number of subsampling channels between the backbone and neck, making swin transformer-tiny suitable for a single-stage algorithm, the YOLOX series.

In our experiments, the ST-YOLOX-S achieved a remarkable improvement of 6.1% in mAP compared with YOLOX-S on the COCO-2017val dataset[16]. This improvement was particularly evident in obstacle recognition under the simulation of actual vehicle conditions, where specific object categories including people, bicycles, cars, buses, and motorcycles all showed significant improvements.

The architecture diagram for the proposed ST-YOLOX-S model is illustrated in Fig.2.

As shown in Fig.2, the model is comprised of three key components, the swin transformer-tiny backbone, the PA-FPN, and the decoupled head. The swin transformer-tiny backbone with attention mechanism is a compact network that improves the connection between remote information through its window-based hierarchical structure. As a result, the ST-YOLOX-S model adopts a hierarchical design based on the swin transformer-tiny backbone, which comprises four stages in total.

The input image with dimensions of $H\times W\times 3$ is first processed by the patch partition component, which cuts the RGB image into blocks and embeds them into the embedding. Each patch has a size of 4×4, and the feature

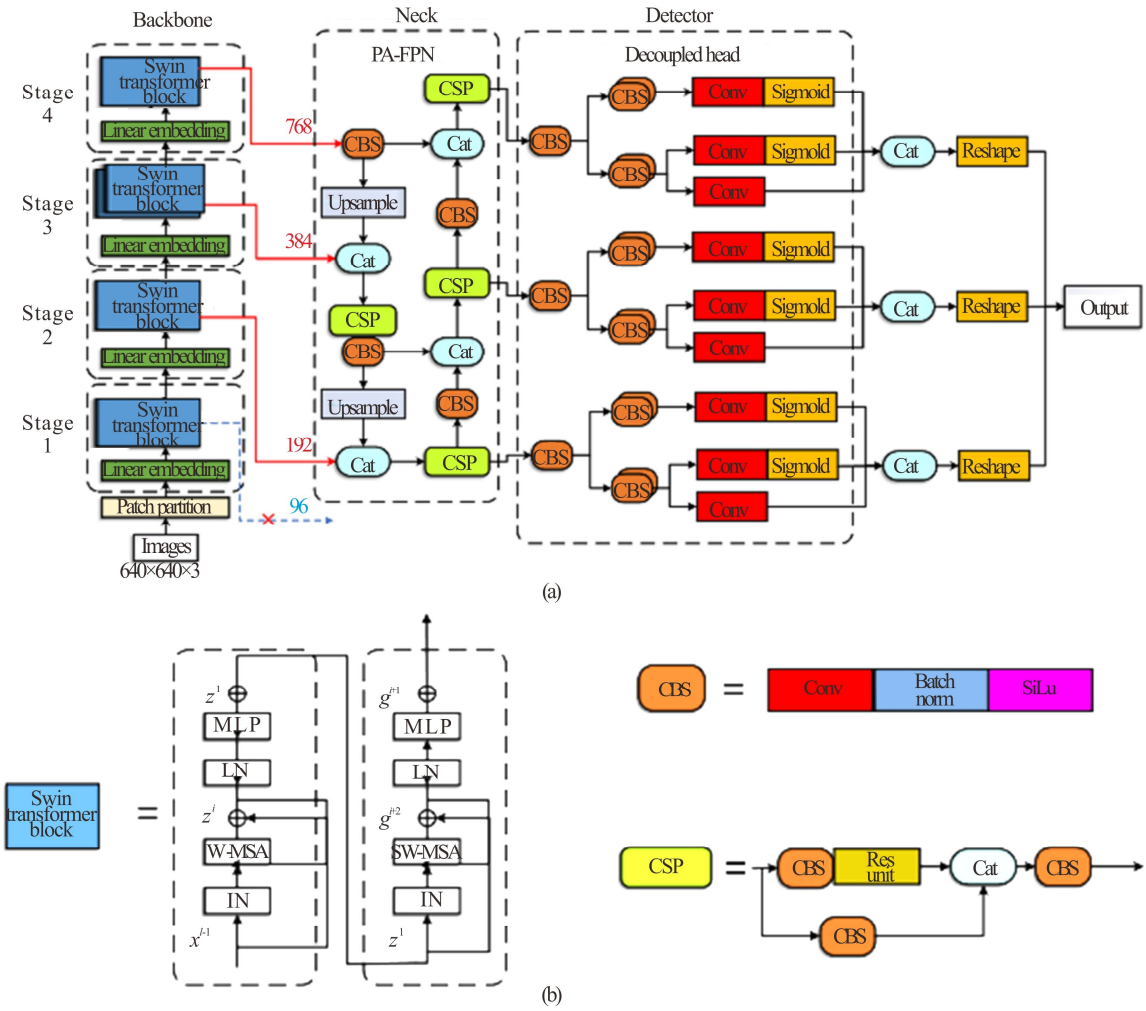

Fig.2 ST-YOLOX-S architecture diagram

dimension of each patch is calculated as 4×4×3=48. Then the feature map is projected to a specific dimension (denoted as C) in the linear embedding of Stage 1. The projected feature map is then fed into the swin transformer block for further calculation and feature extraction.

In contrast to Stage 1, Stages 2, 3, and 4 will utilize a patch merging process before feature extraction. The purpose of the patch merging is to reduce the resolution of the input feature map by down-sampling, adjusting the number of channels, and creating a hierarchical design that allows the model to expand its receptive field layer by layer, which is similar to a CNN. This also results in a reduction of computation. The down-sampling process is doubled, so elements are selected at intervals of 2 in both the row and column directions, which makes computation a reduction.

Fig.2(b) depicts a swin transformer block that comprises two blocks. Initially, the input feature map processes through a block that includes windows multi-head self-attention (W-MSA), followed by computation with a block that consists of W-MSA. Each of these blocks is equipped with a two-layer multilayer perceptron (MLP). Additionally, a layer-norm (LN) layer is implemented before each MSA module and each MLP module, and residual connections are established after each module.

The W-MSA module was designed for the purpose of decreasing the computational load and enhancing the real-time processing speed. The conventional MSA module performs the self-attention calculation globally. On the other hand, the W-MSA module divides the feature map into smaller windows of size $M\times M$ before conducting the self-attention calculation within these windows. The computational complexities of both MSA and W-MSA are listed below:

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C, \tag{1}$$

$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC, \tag{2}$$

where h represents the height of the feature map, w represents the width of the feature map, C represents the depth of the feature map, and M represents the size of each window.

However, the implementation of the W-MSA module will result in a communication breakdown between the windows. To address this issue and facilitate the exchange of information between windows, we integrate a shifted windows multi-head self-attention (SW-MSA) module, which forms a truly effective self-attention mechanism

module in combination with the W-MSA. Fig.3 shows the difference between W-MSA and SW-MSA.

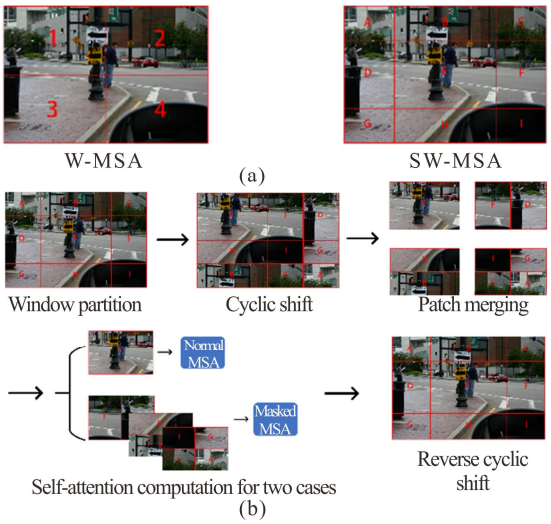

Fig.3 (a) W-MSA and SW-MSA; (b) Computational methods for realizing efficient information interaction between windows

As shown in Fig.3(a), the conventional W-MSA divides the entire window into four equal-sized sub-windows and calculates their self-attention. To enhance the information exchange between sub-windows, SW-MSA divides the window into nine uneven sub-windows marked A to I by moving two pixels to the right and down. When calculating self-attention in the SW-MSA sub-window, windows such as B, D, H, F and E are distinct from the 1, 2, 3, and 4 windows in the W-MSA method. For example, when calculating self-attention in the B window in SW-MSA, it is equivalent to computing partial self-attention in the 1 and 2 windows of the W-MSA method. In this way, SW-MSA connects independent windows 1 and 2, thus resolving the communication break between windows. However, the sub-windows partitioned by SW-MSA are irregular in shape and they require padding for ease of calculation, which increases the computational amount. To address this issue, we adopt the window shifting method to achieve efficient calculation.

Fig.3(b) demonstrates that four sub-windows can be obtained after multiple window displacements. However, a batch window may consist of several sub-windows which are not nearby in the feature map, which prohibits the combination of these sub-windows for self-attention computation. To address this issue, we use masked MSA to calculate non-adjacent windows. Masked MSA acts to mask the connection between the pixels of a window and its non-adjacent sub-windows. The final step in the process is the restoration of the image through a reverse cyclic shift.

In this discussion, we will delve into the neck and prediction components of the ST-YOLOX-S framework, which have been derived from the YOLOX-S architecture. As shown in Fig.2(a), the neck component implements the PA-FPN architecture, effectively integrating feature maps extracted from multiple levels of the backbone network through down-sampling and up-sampling. Furthermore, to realize real-time target detection on mobile carriers in the future, this research reduces the number of channels between the backbone and neck in the model. The proposed ST-YOLOX-S in this paper reduces the original swin transformer-tiny backbone output channel number of [96, 192, 384, 768] to [192, 384, 768] to better fit the PA-FPN. Although it may result in a slight decrease in accuracy, deleting the 96 channels output from Stage 1 to the neck part could improve the detection frame rate. Keeping 30 fps is crucial for identifying obstacles in real-life vehicular conditions.

As depicted in Fig.2(a), the prediction components decoupled head provides a unique approach to object detection compared to traditional coupled detection heads. Instead of having all classification tasks share parameters, the decoupled head separates each task, resulting in separate predictions, which have been shown to have a faster convergence rate and improved detection accuracy, as demonstrated in the YOLOX study[6].

In our experiment, the COCO-2017 dataset is utilized, which comprises 118 000 training and 5 000 validation images. Our models were tested using PyTorch 1.9.0, Python 3.8, Cuda 11.1, and NVIDIA GeForce RTX 3090 (24 GB) with a consistent configuration. The batch size was set at 13 and the optimizer used was SGD with an initial learning rate of 0.002 5 and a weight decay of 0.000 5. The training schedule consisted of 300 epochs and FP-16 precision was employed. To ensure fairness, the same training strategy as YOLOX-S was implemented, including data augmentation techniques such as Mosaic[17] and MixUp[18] (during the last 15 epochs of training, these augmentations were turned off), as well as label assignment SimOTA. We introduce them in the following.

Data augmentation is a technique used to generate additional data from a limited set, thus increasing the number and diversity of training samples and making the model more robust. Conventional data augmentation methods include rotation, cropping, and flipping. In the ST-YOLOX-S system, the Mosaic data augmentation method proposed by Ultralytics-YOLOv3[17] is utilized. This method splices the data through random scaling, cropping, and arrangement, which effectively improves the detection of remote obstacles in real-world vehicle conditions. Additionally, MixUp, an enhancement strategy originally developed for image classification, has been added to the Mosaic method. It has been proven to increase classification accuracy by approximately 1% with minimal computational effort. By combining MixUp with Mosaic, the accuracy of obstacle recognition is further improved.

Label assignment refers to the crucial stage of object detection where positive and negative samples are distinguished and assigned appropriate learning targets in the

training phase. The advanced label assignment process is influenced by four key insights outlined in the optimal transport assignment (OTA)[19] study. To further optimize the process, the SimOTA approach improves upon the classical Sinkhorn-Knopp[20] algorithm by solving the approximate optimal solution, thereby avoiding the need for additional solver hyperparameters and reducing training time. This leads to improved performance of the object detector.

On COCO-2017val, we analyze the importance of our proposed component. The impact of components is listed in Tab.1. As the YOLOX algorithm is a one-stage algorithm, our experiments can only be run under three-channel conditions. This is one of the reasons why reducing the number of channels to [192, 384, 768] makes swin transformer-tiny more adaptable to YOLOX-S. With swin transformer-tiny backbone, the proposed method ST-YOLOX-S is 6.1% higher than YOLOX-S in mAP and 7.5% higher than YOLOX-S in AP_{50} accuracy.

Tab.1 Ablation study on COCO-2017val

Method	mAP	AP_{50}
YOLOX-S	40.3	59.1
YOLOX-S+swin transformer-tiny+channels [192, 384, 768]	46.4(+6.1)	66.6(+7.5)

Our proposed method ST-YOLOX-S achieves an average accuracy of 46.4% in the COCO-2017val set. As shown in Tab.2, we have compared the accuracy and speed of a few popular target detection algorithms. Our algorithm ST-YOLOX-S performs well in terms of both accuracy and speed, achieving our original aim of improving accuracy. Relative to the YOLOX-S, the real-time inference speed (means FPS) of our proposed method is 19.1 frames lower than the original YOLOX-S. The reason for this is that there are many parameters in the swin transformer module. Although the real-time detection speed of our proposed ST-YOLOX-S is not as good as that of YOLOX-S, our method mAP improves more than YOLOX-S and also meets the needs of real-time detection (FPS≥30).

To accurately simulate the obstacle recognition of real-life vehicle conditions, we selected results from COCO-2017val for comparison. The results selected were for the detection of obstacles commonly encountered by vehicles such as people, bicycles, cars, buses, and motorcycles. As shown in Tab.3, our method ST-YOLOX-S has excellent performance in the detection of each of the above-mentioned obstacles. Comparing with YOLOX-S, our method showed an improvement of 3.1% for person recognition, 6.6% for bicycle recognition, 6% for car recognition, 4.5% for bus recognition, and 5.4% for motorcycle recognition. This improvement in obstacle recognition can be attributed to the effective use of the attention mechanism in the swin transformer. This mechanism improves the global understanding of the detector, thus enhancing the ability to identify larger obstacles by connecting the relationships between local parts.

Fig.4 presents the loss of the ST-YOLOX-S model, which was trained for 300 iterations. As evident from the figure, the loss experiences a rapid decrease during the initial stages of training. However, as the number of epochs increases, the loss gradually stabilizes. In the final 15 epochs, the use of Mosaic and Mixup is discontinued and the loss function is altered from Eq.(3) to Eq.(4). While this change results in a break in the loss, it does not impact the convergence of the loss towards stability.

$$Loss = Loss_{\text{cls}} + Loss_{\text{bbox}}, \tag{3}$$

$$Loss = Loss_{\text{cls}} + Loss_{\text{bbox}} + Loss_{\text{l1}}. \tag{4}$$

Tab.2 Comparison of the speed and accuracy of different object detectors on COCO-2017val dataset

Method	AP	AP_{50}	AP_{75}	Parameter	FPS
YOLOv3[2]	33.7	56.6	35.3	61.95M	63.8
YOLOX-S[6]	40.3	59.1	43.4	8.97M	54.8
SSD512[21]	29.5	49.3	30.9	36.04M	34.6
Retinanet-R101[22]	38.5	57.6	41.0	56.74M	20.8
Cascade R-CNN-R101[23]	42.0	60.4	45.7	88.16M	18.2
CornerNet[24]	41.2	57.2	44.0	201.04M	3.0
Gird R-CNN-R101[25]	41.5	59.8	44.9	83.31M	11.0
Sparse-R-CNN-R101[26]	44.2	63.1	47.8	125.06M	17.5
ST-YOLOX-S	**46.4**	**66.6**	**50.2**	35.87M	**35.7**

Tab.3 The results of all models simulating actual vehicle conditions on COCO-2017val (including person, bicycle, car, bus and motorcycle)

Method	AP (%) (IoU=0.5)				
	Person	Bicycle	Motorcycle	Bus	Car
YOLOv3[2]	46.9	27.4	37.4	57.9	36.0
YOLOX-S[6]	54.8	29.9	42.9	64.2	40.7
SSD512[21]	40.4	21.3	31.8	57.3	30.1
Retinanet-R101[22]	51.7	29.0	40.5	64.6	40.3
Cascade R-CNN-R101[23]	56.4	31.9	42.8	66.3	45.5
CornerNet[24]	50.6	29.3	42.7	66.4	42.1
Gird R-CNN-R101[25]	55.7	30.4	42.2	66.1	44.9
Sparse-R-CNN-R101[26]	55.7	30.7	45.1	67.2	45.2
ST-YOLOX-S	**57.9**	**36.5**	**48.3**	**68.7**	**46.7**

Fig.5 displays the qualitative results of the object detection performed by ST-YOLOX-S on random images

from the COCO-2017 test set. These results demonstrate that our method, ST-YOLOX-S, exhibits a commendable level of recognition accuracy even in real-world vehicle conditions.

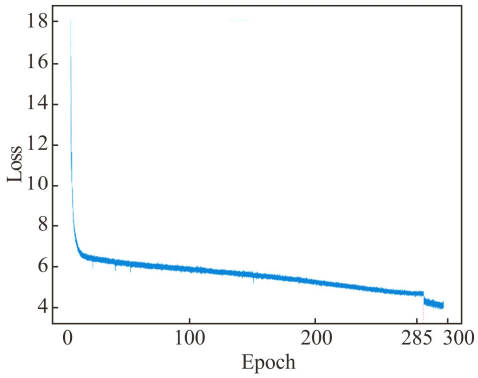

Fig.4 The loss curve of ST-YOLOX-S on COCO-2017 train at 300 epoch

Fig.5 The rendering of ST-YOLOX-S

In this research, an improved obstacle detection model, ST-YOLOX-S, is proposed to address the issue of low accuracy in recognizing obstacles under actual vehicle conditions. This model aims to improve accuracy based on the YOLOX-S model. To achieve this, the model utilizes swin transformer-tiny in the backbone section, which enhances the collection of global information through the incorporation of six swin transformer blocks with an attention mechanism. To make it compatible with YOLOX-S, ST-YOLOX-S decreases the number of channels in the backbone and neck from [96, 192, 384, 768] to [192, 384, 768]. Tab.2 and Tab.3 demonstrate that the proposed method has significantly improved the accuracy of obstacle categorization in real-world vehicle conditions compared with the benchmark model YOLOX-S. Despite a slight decrease in real-time reasoning speed, the method still satisfies the requirement for real-time detection with a minimum of 30 fps. In future endeavors, the objective should enhance the real-time inference speed without sacrificing the level of accuracy.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] ZHANG X, ZHOU M, QIU P, et al. Radar and vision fusion for real-time obstacle detection and identification[J]. Industrial robot：the international journal of robotics research and application, 2019, 46(3)：391-395.

[2] REDMON J, FARHADI A. Yolov3：an incremental improvement[EB/OL]. (2018-04-08) [2023-01-22]. https：//arxiv.org/abs/1804.02767.

[3] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4：optimal speed and accuracy of object detection[EB/OL]. (2020-06-05) [2023-01-22]. https：//github.com/kiccho1101/paper/issues/27.

[4] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9)：1904-1916.

[5] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, USA. IEEE：New York, 2018：8759-8768.

[6] GE Z, LIU S, WANG F, et al. YOLOX：exceeding YOLO series in 2021[EB/OL]. (2021-08-06) [2023-01-22]. https：//arxiv.org/abs/2107.08430.

[7] JOCHER G, STOKEN A, BOROVEC J, et al. Ultralytics/YOLOv5：v5.0-YOLOv5-P6 1280 models AWS supervisely and youtube integrations[J]. Zenodo, 2021, 11.

[8] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet：a new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, June 14-19, 2020, Seattle, WA, USA. IEEE：New York, 2020：390-391.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words：transformers for image recognition at scale[EB/OL]. (2010-11-09) [2023-01-22]. https：//arxiv.org/pdf/2010.11929.pdf.

[11] LIU Z, LIN Y, CAO Y, et al. Swin transformer：hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, Canada. IEEE：New York, 2021：10012-10022.

[12] GRIGORESCU S, TRASNEA B, COCIAS T, et al. A survey of deep learning techniques for autonomous driving[J]. Journal of field robotics, 2020, 37(3)：362-386.

[13] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, October 24-27, 2017, Italy. IEEE：New York, 2017：2961-2969.

[14] LIN G, MILAN A, SHEN C, et al. Refinenet：multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, USA. IEEE：New York, 2017：1925-1934.

[15] REN S, HE K, GIRSHICK R, et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco：common objects in context[C]//European Conference on Computer Vision, September 6-12, 2014, Zurich, Switzerland. Berlin：Springer, Cham, 2014：740-755.

[17] JOCHER G, KWON Y, VEITCH-MICHAELIS J, et al. Ultralytics/YOLOv3 ： v9.5.0-YOLOv5 v5.0 release compatibility update for yolov3[J]. Zenodo, 2021.

[18] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup：beyond empirical risk minimization[EB/OL]. (2017-10-25) [2023-01-22]. https：//arxiv.org/abs/1710.09412.

[19] GE Z, LIU S T, LI Z M, et al. OTA：optimal transport assignment for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, virtual. IEEE：New York, 2021.

[20] KNIGHT P A. The sinkhorn-knopp algorithm ：convergence and applications[J]. SIAM journal on matrix analysis and applications, 2008, 30(1) ： 261-275.

[21] LIU W, ANGUELOV D, ERHAN D, et al. SSD：single shot multibox detector[C]//14th European Conference on Computer Vision, October 11-14, 2016, Amsterdam, the Netherlands. Berlin： Springer International Publishing, 2016：21-37.

[22] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, October 24-27, 2017, Italy. IEEE：New York, 2017：2980-2988.

[23] CAI Z, VASCONCELOS N. Cascade R-CNN：high quality object detection and instance segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(5) ： 1483-1498.

[24] LAW H, DENG J. Cornernet：detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin：Springer International Publishing, 2018：734-750.

[25] LU X, LI B, YUE Y, et al. Grid R-CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, USA. IEEE：New York, 2019：7363-7372.

[26] SUN P, ZHANG R, JIANG Y, et al. Sparse R-CNN: end-to-end object detection with learnable proposals[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, virtual. IEEE：New York, 2021：14454-14463.