scientific reports

Check for updates

OPEN

Adaptive occlusion object detection algorithm based on OL-IoU

Baicang Guo[1,2], Hongyu Zhang[1], Huanhuan Wang[1], Xinwei Li[1] & Lisheng Jin[1,2]✉

The continuous advancement of autonomous driving technology imposes higher demands on the accuracy of target detection in complex environments, particularly when traffic targets are occluded. Existing algorithms still face significant challenges in detection accuracy and real-time performance under such conditions. To address this issue, this paper proposes an improved YOLOX algorithm based on adaptive deformable convolution, named OCC-YOLOX. This algorithm enhances the feature extraction network's ability to focus on occluded targets by incorporating a coordinate attention mechanism. Additionally, it introduces the Overlapping IoU (OL-IoU) loss function to optimize the overlap between predicted and ground truth bounding boxes, thereby improving detection accuracy. Furthermore, the adoption of Fast Spatial Pyramid Pooling (Fast SPP) reduces computational complexity while maintaining real-time performance. Experiments on fused public datasets demonstrate that OCC-YOLOX achieves improvements in accuracy, recall, and average precision by 2.76%, 1.25%, and 1.92%, respectively. In addition to testing on the KITTI, CityPersons, and BDD100K datasets, the effectiveness of the OCC-YOLOX algorithm is further validated through comparisons with self-collected occlusion scene data. The experimental results indicate that OCC-YOLOX outperforms existing mainstream detection algorithms, particularly in handling complex occlusion scenarios, significantly enhancing the accuracy and efficiency of object detection. This study provides new insights for addressing the challenges of occluded target detection in intelligent transportation systems.

Keywords Autonomous driving, Environmental perception, Occlusion detection, YOLOX, Adaptive deformable convolution, Coordinate attention mechanism, Overlapping IoU, Fast spatial pyramid pooling

Environment perception is a key technology for Intelligent Connected Vehicle (ICV), and object detection, as the basis for solving more complex environment sensing tasks, has a direct impact on traffic safety in terms of its accuracy[1]. In recent years, with the continuous development of object detection algorithms based on deep learning, the prediction accuracy has been continuously improved[2]. However, in the face of occluded traffic targets, the truncation phenomenon makes the feature acquisition crippled, often resulting in omission and false alarm phenomenon, which seriously affects the detection effect. There-fore, the study of occluded object detection is of great significance to traffic safety.

Object detection algorithms can be classified into two types: traditional and deep learning-based methods. Compared with the disadvantage of traditional methods[3,4] which rely heavily on manually designed feature sub, deep learning-based methods can overcome the limitation of not being able to adapt to complex scenes. Therefore, deep learning-based object detection algorithms have become the mainstream of application. They are classified into two-stage and single-stage according to their detection logic. The RCNN series of algorithms proposed in the literature[5–8], is a representative of two-stage algorithms, but all the algorithms in this series need to generate a large number of candidate regions first to mark all the possible targets, which has a high detection accuracy, but the redundant anchor frames of repeated calculations will reduce the computing speed. Therefore, the single-stage algorithms that do not need to select candidate regions and directly generate target position coordinates were then born.

Meanwhile, the optimization of object detection algorithms under occlusion conditions mainly follows two technical approaches: one is based on the improvement of overall feature detection algorithms, and the other is based on the enhancement of partial semantic detection algorithms. Although existing methods have made progress in occlusion target detection research through different technical directions, there are still some issues

[1]School of Vehicle and Energy, Yanshan University, Qinhuangdao 066004, China. [2]Hebei Key Laboratory of Special Carrier Equipment, Yanshan University, Qinhuangdao 066004, China. ✉email: jinls@ysu.edu.cn

that need to be addressed. For instance, common occlusion datasets such as the KITTI dataset, Caltech dataset, etc., have single-scene environments, leading to a lack of robustness in the detection results of the algorithms. The complexity of occlusion situations makes it difficult for algorithms to learn an infinite number of occlusion scenarios. Moreover, during occlusions, severe overlaps between prediction boxes may lead to the erroneous suppression of predicted targets, resulting in missed detections. There is an urgent need to propose more specific and robust algorithms based on the characteristics of targets under occlusion conditions, which is of great significance for complex perception systems.

YOLOX has designed four types of networks with different scale sizes for different application scenarios. Among them, YOLOX-s has the best degree of lightness and is of great significance for engineering applications. Therefore, this paper carries out research based on the YOLOX-s algorithm, and the network structure model is shown in Fig. 1.

The YOLOX-s backbone network is CSPDarknet[9], which can deepen the network structure and improve the image feature expression while reducing network redundancy. The SPP structure is used at the end of the backbone network, and three maximum pooling kernels of different sizes are designed for processing to increase the sensory field of the network. The neck network uses three effective feature layers to construct the FPN (Feature Pyramid Network)[10] + PAN (Path Aggregation network)[11] two-way fusion feature pyramid structure, which performs multidimensional feature fusion on the effective features. The prediction end of the YOLOX-s adopts the form of decoupling, and obtains three prediction results for each.

In this paper, YOLOX is improved for the problem of poor detection of occluded targets from vehicle viewpoints, and an adaptive deformable YOLOX occlusion object detection algorithm is proposed. In summary, we make several important contributions in this work:

1. We have applied adaptive deformable convolution to the backbone network of YOLOX, effectively enhancing the algorithm's transformation capability for irregular geometries to adapt to complex and dynamic occlusion scenes, and to strengthen the feature representation ability of occluded targets.

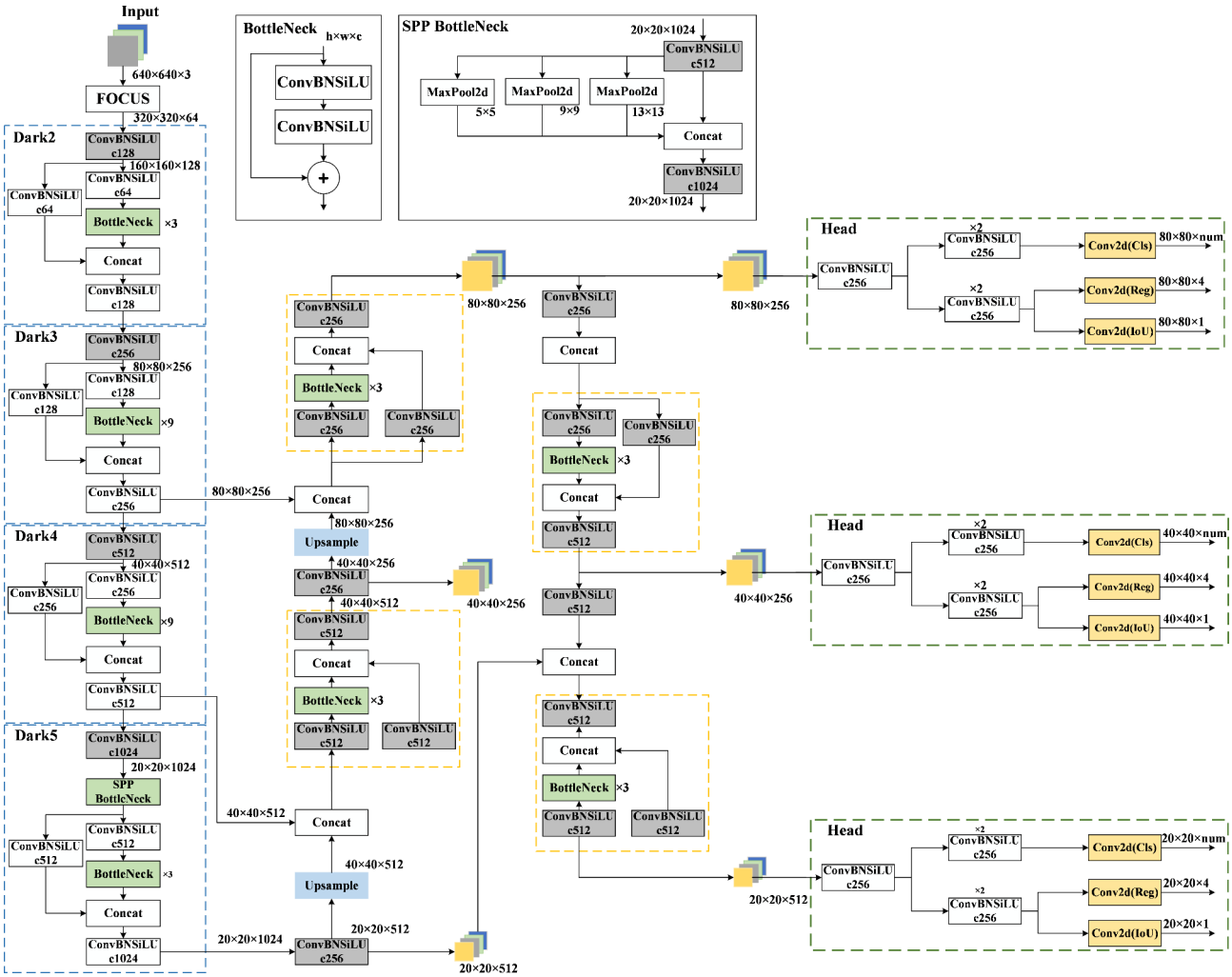

Fig. 1. YOLOX network structure.

2. Coordinate attention mechanism[12] is applied in the Neck part to overcome the restriction of weight application in different channels and spatial domains, so as to improve the fusion ability of residual feature extraction for occluded targets.
3. We propose the Overlapping IoU (OL-IoU) regression loss function based on the intersection and merger ratio (IoU)[13], and a proportional penalty mechanism such as overlapping width and height is added to accelerate the convergence of the bounding box and increase the detection accuracy at the same time.
4. The Spatial Pyramid Pooling (SPP)[14] is replaced by the more efficient Fast SPP[15] to balance the real-time and accuracy of occlusion object detection. Experiments in the occlusion scenario show that the algorithm improves the accuracy and real-time performance of occluded traffic object detection, which verifies the effectiveness of the algorithm.

Related work

Single-stage object detection

Literature[16] proposed the YOLO (You Only Look Once) algorithm, which treats detection as a regression problem and greatly improves the detection speed, but the problem of missed detection is serious in the face of irregular targets and small targets. Literature[17] proposed YOLO9000, which uses a joint dataset to train and design a Darknet19 backbone network, which improves the detection accuracy and speed, but is hindered by the fact that there is only one branch of detection, and it cannot be used for multi-size target detection. Literature[18] proposed SSD (Single Shot MultiBox Detector), which uses CNN to replace the fully connected layer in YOLO detection, and applies feature maps of different scales to detect targets of different sizes, but there is no restriction on the standard size, which makes the miss elected feature maps produce great detection errors. Literature[9] proposed YOLOv3 and designed Darknet53 backbone network to fill the gap of multi-size detection and used binary cross-entropy loss for multi-label classification, but did not consider regression loss, which made the localization results imprecise. Literature[15] proposed YOLOv4, designed CSPDarknet53 backbone network, compared with Darknet53 applying deeper network and more parameters, and added some plug-ins, effectively balancing accuracy and real-time, but the adopted DropBlock random discard does not guarantee the diversity of the feature information, so there is still an indeterminable duplication of operations. Literature[19] proposes YOLOv5 high efficiency detection model, which provides four different depths of network models to adapt to different detection requirements, and also applies Adaptive Anchor Box to train the customized dataset, which improves the operation speed. However, the manual matching rule strategy still requires empirical settings and poorly improves the detection of irregular targets.

YOLOX[20] incorporates the anchor-free framework and decoupled head method, which reduces the tuning pressure of manually setting the Anchor, makes the feature learning and classification regression problems easy to learn, and improves the accuracy and computing speed compared with YOLOv5 models of all sizes. However, there is still the problem of not being able to detect effectively in scenes with a lot of occlusions. In summary, the object detection algorithm still has problems especially in irregular object detection. The irregular targets perceived in the field of intelligent traffic under the vehicle perspective are reflected in the occluded vehicles, pedestrians, cyclists, etc., and the variable traffic scene makes the occlusion relationship of the detection targets extremely complex, thus increasing the difficulty of object detection.

Occlusion object detection

The occlusion object detection algorithm is based on deep learning algorithms and optimized according to its own characteristics, with the aim of training a network model to cope with occlusion. Literature[21] proposes occlusion processing techniques based on various parts of the pedestrian's body, applying the Histogram of Oriented Gradients (HOG) detector to compute the classification scores of the sliding window and applying the sum of its responses to the global detector to reflect the possible partial occlusion of pedestrians. Literature[22] proposes a method to detect partially occluded pedestrians by determining the visible part of the object, which uses a discriminatively trained Deformable Parts Model (DPM) to solve the concave optimization problem to indicate whether the image part belongs to the target object or to the occluder. Literature[23] proposes Soft-NMS non-maximal suppression method to attenuate the scores of high overlapping prediction frames, but its essence is still to consider overlapping suggestion frames as false positives that cannot be accurately classified. Literature[24] integrates deep convolutional neural networks and combinatorial convolutional neural networks, and utilizes a microscopically combinable layer instead of a fully-connected classification layer in order to achieve classification and localization of occluded targets. Literature[25] uses background subtraction to simulate two-wheelers in a crowded scenario. Decision tree is used to evaluate the geometric features of overlapping two-wheelers for classification. All the above detection algorithms for occluded targets are only for a single class of targets and cannot be adapted to the task of occluded object detection in multi-category complex scenes. Literature[26],[27] requires manual annotation at the part level, which is associated with higher data preparation costs. Literature[28–30], are primarily focused on processing video targets, whereas our method is more attentive to static images. Literature[31] employs a combination of classification and proposal methods, which are less efficient and not suitable for complex and dynamic traffic scenarios. Literature[32] concentrates on pedestrian targets, failing to cover various types of traffic participants, and does not address the model lightweight processing required for the computational demands of intelligent vehicles.

Attention mechanism for computer vision

Attention mechanism aims to enhance the ability of neural networks in judging the importance of feature information, and currently Squeeze-and-Excitation (SE)[33], CBAM[34], are mainly used in the field of computer vision. However, SE only considers weighting the importance of each channel by modelling the dependency between channel information, and ignores the importance of positional information. CBAM does not capture

the importance of spatial information through the use of a large-size kernel of convolution to introduce spatial information encoding on top of the channel information, but this approach can only capture the local information and not the information of long range dependencies. Based on the limitations of the above study, Coordinate Attention (CA)[12] is proposed to decompose the 2D global pooling into one-dimensional coding in two spatial directions for obtaining position information and channel relationships, respectively.

OCC-YOLOX Algorithm

Deformable convolution

When conventional convolution is modelling image features, the receptive fields are all regular adjacent rectangles with limited geometric variation capability. However, the location and shape of the occluded targets cannot be predicted, requiring algorithms with the ability to adaptively perceive the target region and adjust the receptive fields. Deformable Convolutional Networks (DCN)[35,36] provides an adaptive learnable offset for increasing spatial sampling points, enabling the network to have adaptive modelling capability. The deformation process can be described by Eq. (1).

$$y\left(p_0\right)=\sum_{n=1}^{N}\omega_n\times x\left(p+p_n+\Delta p_n\right)\times\Delta m_n \tag{1}$$

Where: Δm_n denotes the modulation scalar, and Δm_n is the sigmoid normalized modulation term allowing for more efficient feature level range control. p_0 is the pixel of the output feature, Δp_n denotes the offset position in x, ω_n is the convolution weight, and p_n enumerates the position in x, which can be expressed as Eq. (2). The sampling position of each sample point is affected by summing the offsets of each convolution position.

$$p_n\in R=\{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\} \tag{2}$$

Deformable convolution uses learnable offsets to describe target feature orientations, allowing the network's sense field to be not limited to a fixed range and more flexible to adapt to changes in target geometry. As a result, DCNs are more conducive to adequate detection of complex scenes. Although DCNs do not bring significant extra computation, the application of a large number of DCNs increases the algorithm inference time. Therefore, in order to balance the inference efficiency and detection accuracy, this paper replaces the standard 3×3 convolution in the FEAT3 layer of the backbone network, and adaptively changes the convolution kernel sampling position through the offset, and the offset process is shown in Fig. 2, which effectively adapts to the shape of the occluded target.

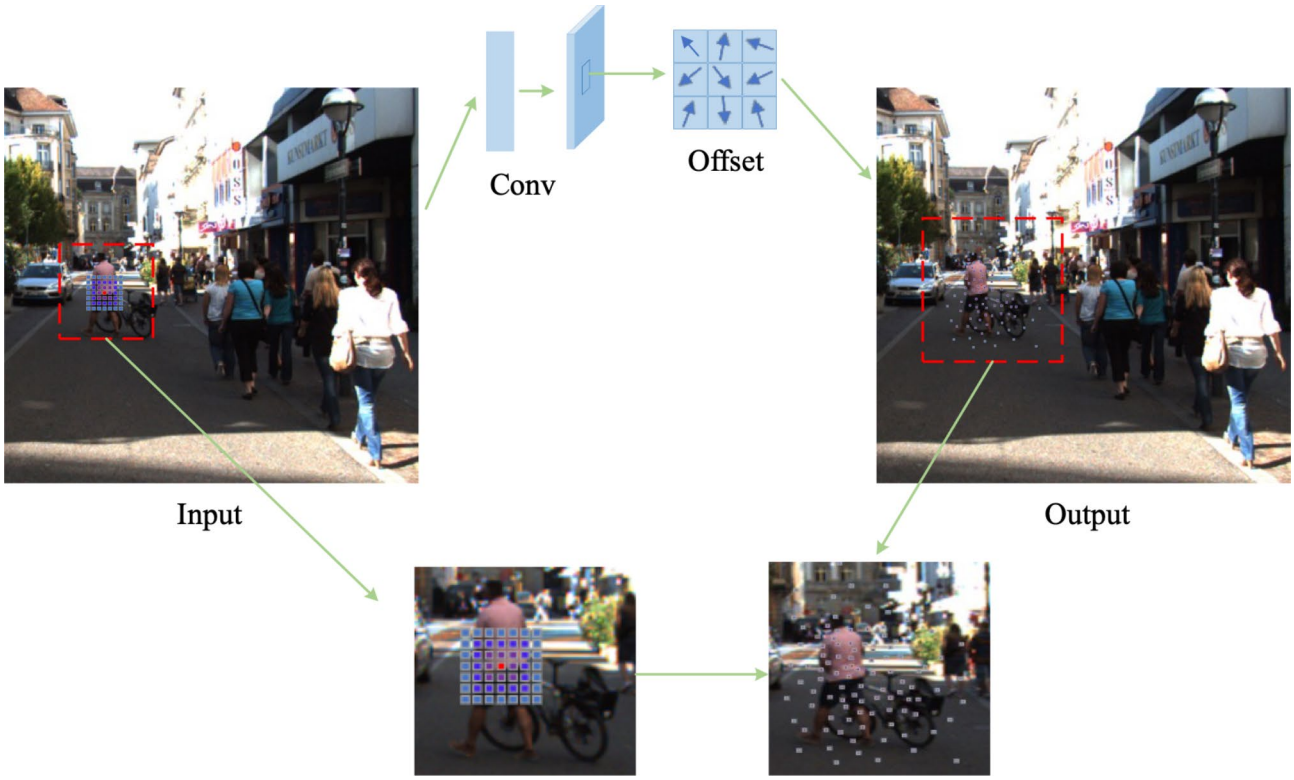

Fig. 2. Deformable convolution offset effect.

Fast SPP spatial pyramid pooling

YOLOX-s applies the Spatial Pyramid Pooling Structure (SPP) in the backbone network. The SPP takes the feature map through three different sizes of maximum pooling kernels of 5×5, 9×9, and 13×13 for feature extraction and then aggregation, which efficiently enlarges the sensory field of the network.

However, the use of multilayer concatenated convolutional kernels of different sizes would greatly enhance the computational difficulty and affect the detection speed of the algorithm. Therefore, the FAST SPP structure is utilized for replacement in the backbone network of OCC-YOLOX, and the FAST SPP structure model is shown in Fig. 3. Compared with the SPP structure, FAST SPP calculates three convolutional kernels as 5×5 pooling layers in series, so that the output of each pooling will become the input of the next pooling, and aggregates the features under different scales of the same feature map, which improves the utilization rate of the network parameters and reduces the difficulty of the operation.

Coordinate attention network

The CA attention module is shown in Fig. 4, where given the inputs two spatial ranges of the pooling and are used to encode each channel along the horizontal and vertical coordinates, respectively. encoding. Therefore, the output of the cth channel with height of h can be formulated as Eq. (4):

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \leqslant i \leqslant W} x_c(h, i) \tag{3}$$

The output of the cth channel with a width of w can be written as Eq. (5):

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \leqslant i \leqslant H} x_c(j, w) \tag{4}$$

And then the features are aggregated along the two spatial directions respectively, as in Eq. (6), to produce a pair of intermediate feature maps f that encode spatial information in the horizontal and vertical directions.

$$f = \delta\left(F_1\left(\left[Z^h, Z^w\right]\right)\right) \tag{5}$$

where, F_1 denotes the 1×1 convolutional transform function and δ is a nonlinear activation function, which performs cascade operations on the two spatial dimensions. Then use two 1×1 convolution F_h and F_w encoded as tensor with the same channel respectively and calculate the attention weights of -two spatial directions, the operation process is as in Eqs. (7) and (8), and finally the coordinate attention output can be obtained as in Eq. (9).

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{6}$$

$$g^w = \sigma\left(F_w\left(f^W\right)\right) \tag{7}$$

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

The CA attention module can satisfy the need to capture long-range dependencies along one direction while still ensuring accurate position information in the other direction. After adding the attention module to the up-sampling and down-sampling of the bidirectional FPN feature pyramid, it decides which part needs to be concerned, assigns the weights of different feature maps and activates them by the swish activation function, so that the part with smaller weights will be less concerned, and allocates the processing resources more reasonably. The location of the CA attention module is shown in Fig. 5.

YOLOX uses dynamic matching of positive samples to get the feature point corresponding to each real frame, and then takes out the predicted frame of that feature point, and uses the real frame and the predicted frame to calculate the IoU regression loss. IoU (Intersection over Union) is the intersection and union ratio, which is used to reflect the detection effect of the real frame and the predicted frame. However, it cannot accurately represent the way the two frames overlap, especially when the IoU is the same, the different two frames position indicates the regression effect is not the same, the calculation formula is Eq. (10), Such a representation is therefore flawed.

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

DIoU[37] introduces the minimum outer rectangular diagonal distance between the prediction box and the real box to accelerate the regression to the Euclidean distance between the centroids of the two boxes, which avoids the phenomenon that Loss is too large to be optimised when the two boxes are far away from each other. The formula is expressed as Eq. (11):

$$L_{DIoU} = 1 - \mathrm{IoU}(A, B) + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} \tag{10}$$

where: c denotes the minimum outer rectangular diagonal distance between the prediction frame and the real frame, and b、 b^{gt} denotes the center point of the prediction frame and the real frame, respectively.

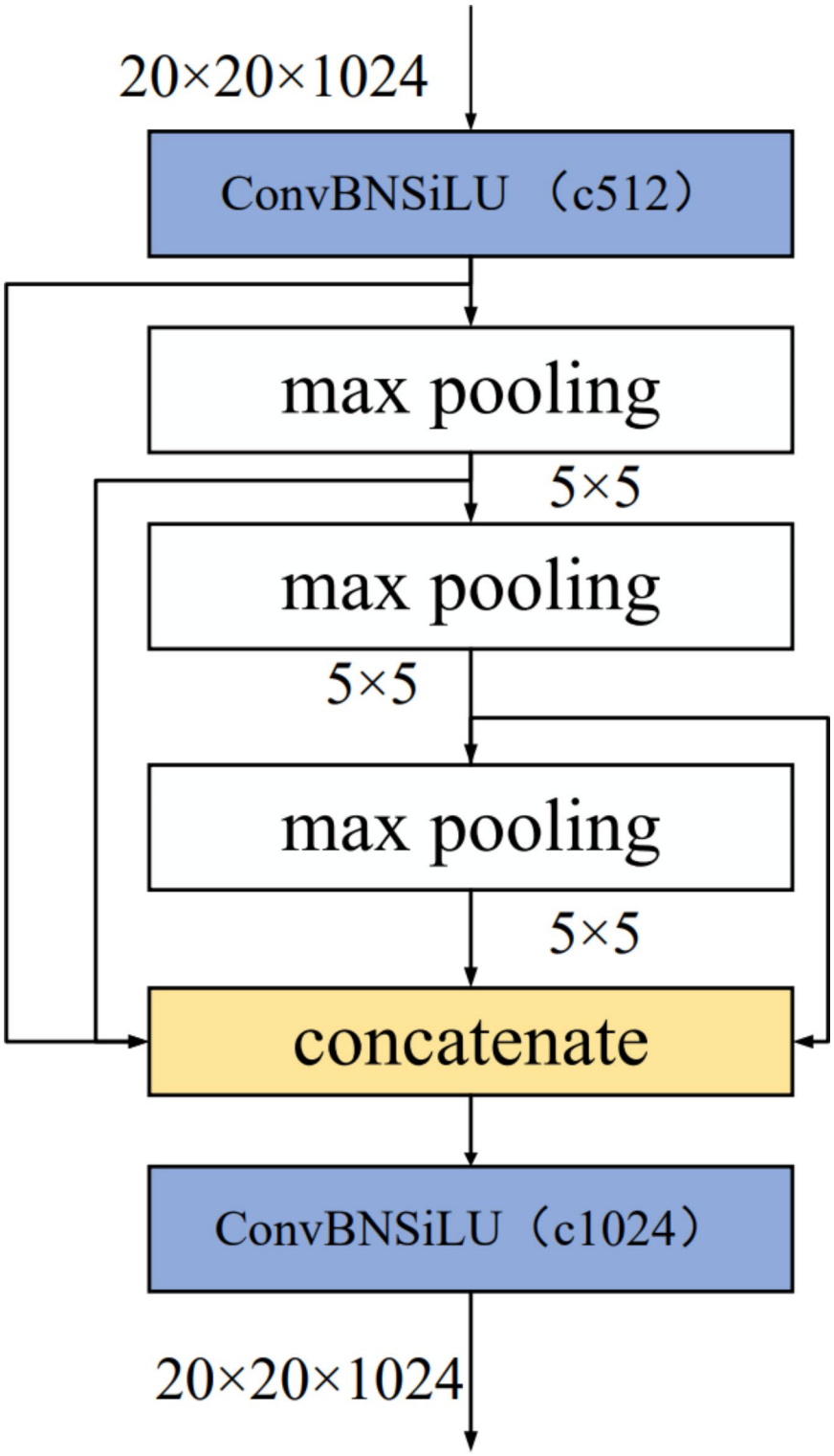

Fig. 3. FAST SPP structural model.

CIoU[37] adds the factor of aspect ratio consistency between predicted and real frames on the basis of DIoU, and the formula is as in (12). Where, α is the weight factor, which indicates the proportion of the consistency case in joining the consistency loss, and ν_1 indicates the metric aspect ratio consistency, and the formula is calculated as Eq. (13):

$$L_{CIoU} = 1 - \mathrm{IoU}(A, B) + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v_1 \tag{11}$$

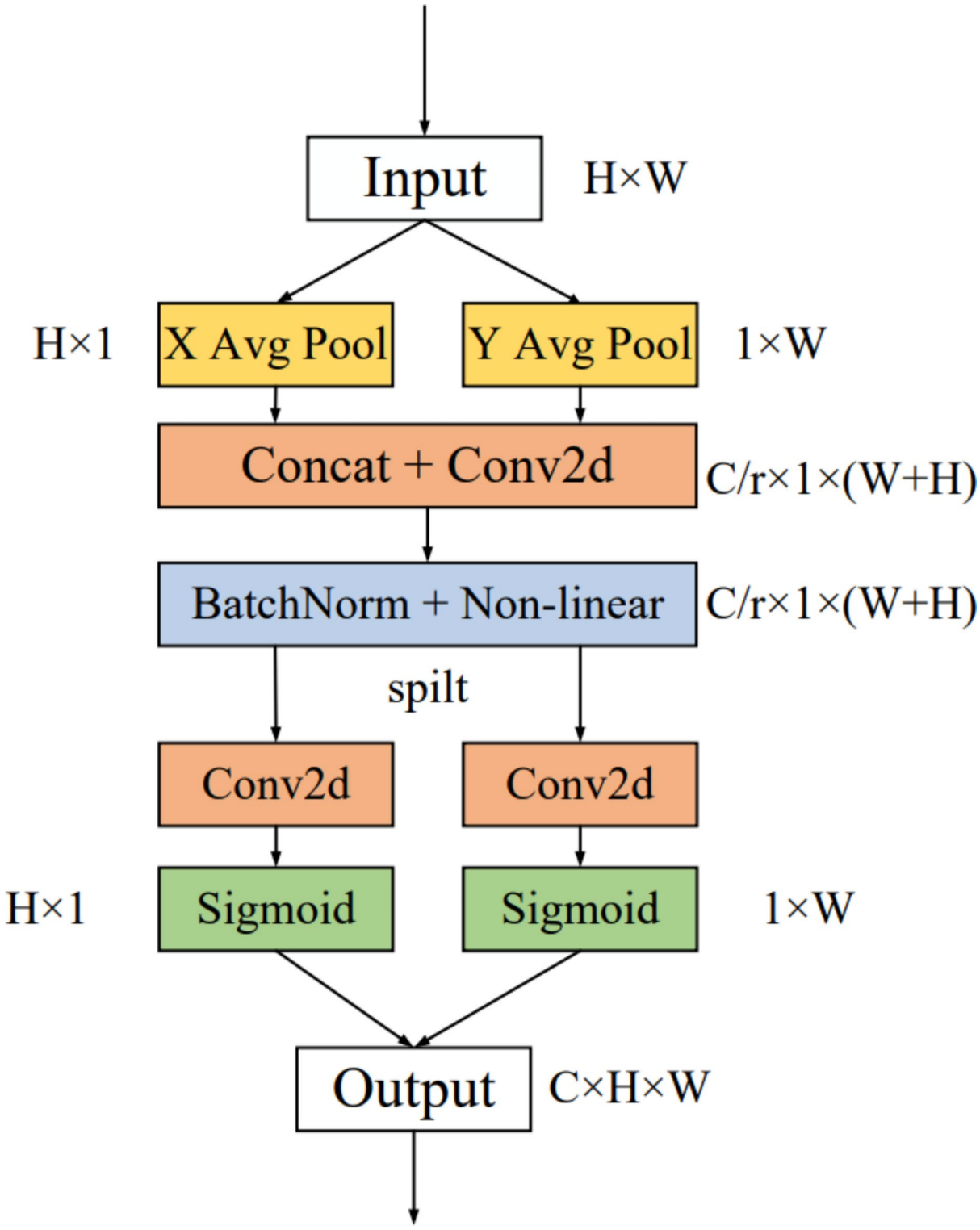

Fig. 4. CA Attention Module Diagram.

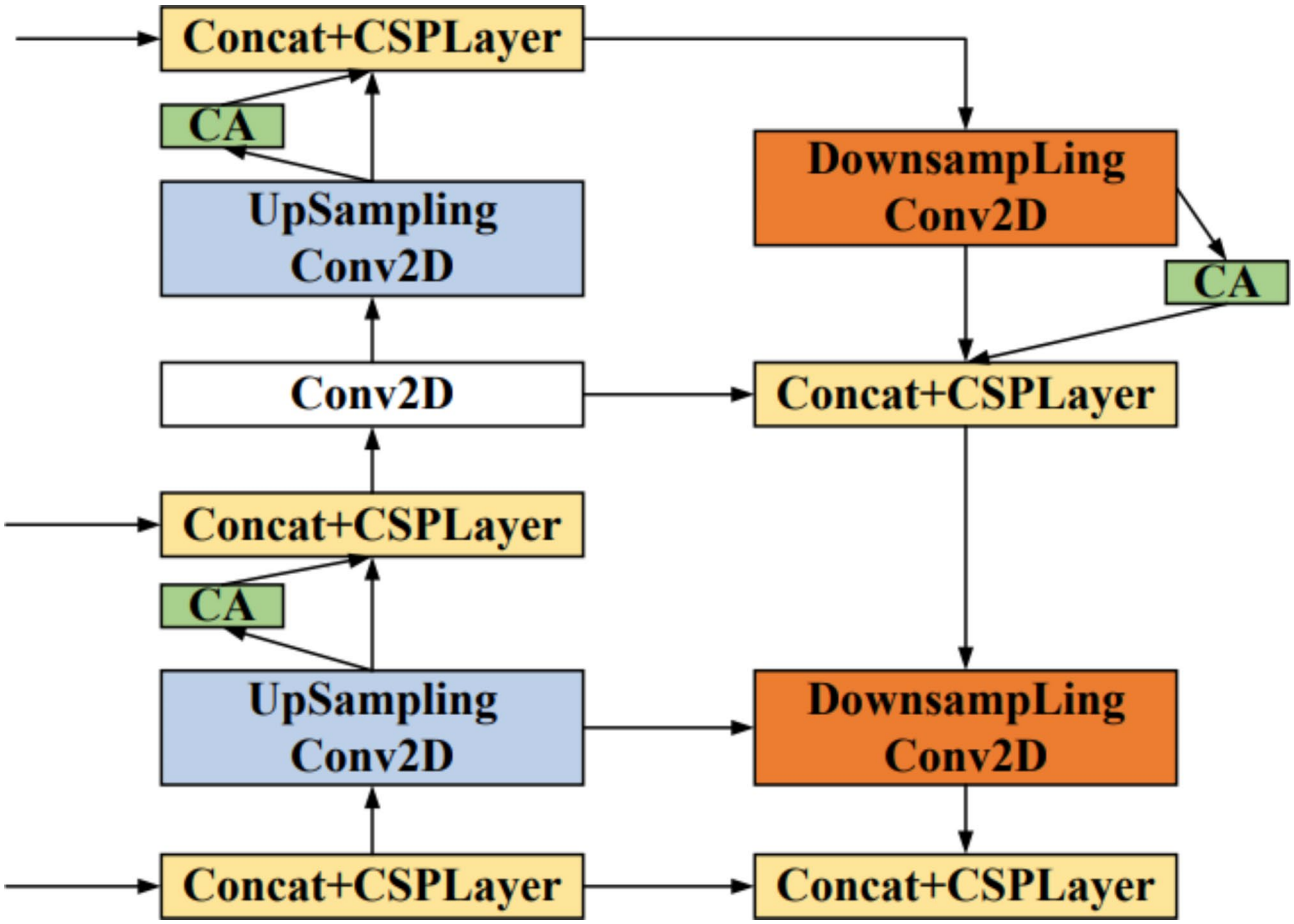

Fig. 5. Enhancement of the structure of the feature extraction section.

$$v_1 = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{12}$$

$$\alpha = \frac{v_1}{1 - \mathrm{IoU}(A,B) + v_1} \tag{13}$$

EIoU[38] splits the aspect ratio on the basis of CIoU, and considers the consistency of the width and height of the overlapped part and the width and height of the smallest external rectangle respectively, and the formula is as (15). Where c_w denotes the width of the minimum outer rectangle, c_h denotes the width of the minimum outer rectangle, I_w denotes the width of the overlapping part, and I_h denotes the height of the overlapping part. The EIoU makes it clearer that the width and height are respectively the real differences with their confidence levels, which promotes the effective optimisation of the model.

$$L_{EIoU} = 1 - \mathrm{IoU}(A,B) + \frac{\rho^2(b,b^{gt})}{c^2} + \frac{(c_W - I_w)^2}{c_W^2} + \frac{(c_h - I_h)^2}{c_h^2} \tag{14}$$

In this paper, we propose OL-IoU, which introduces the consistency factor of the aspect ratio of the predicted frame to the real frame ν_1 and the consistency factor of the overlapping aspect-accelerated regression ν_2 on the basis of EIoU. Among them, ν_1 is the same as the representation in CIoU, and ν_2 is calculated as Eq. (16):

$$v_2 = \frac{4}{\pi^2}\left(\arctan\frac{I_w}{c_w} - \arctan\frac{I_h}{c_h}\right)^2 \tag{15}$$

where, β is the weight coefficient, which indicates the proportion of the accelerated regression consistency case in the joining consistency loss, and the larger value indicates that the detection frame width and height can be closer to the real frame at the same time. the OL-IoU calculation formula is expressed as Eq. (17):

$$L_{oL-IoU} = 1 - \mathrm{IoU}(A, B) + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{(c_w - I_w)^2}{c_W^2} + \frac{(c_h - I_h)^2}{c_h^2} + \alpha v_1 + \beta v_2 \tag{16}$$

Design $\beta\,\nu_2$ Overlapping width and height accelerated regression consistency penalties are used to promote the approach of predicted frames in both directions of width and height to the real frames with equal speed, avoiding that a single direction cannot achieve the purpose of maximising the optimization. At the same time, this design more encourages that the approach of the target frame in a single direction can promote the equivalent operation in the other direction, even if the overlap area reaches an increase of 2D square degree, accelerating the improvement of IoU. it is schematically shown in Fig. 6 below, where (a) is the intersection-parallel ratio representation without adding the overlap width-height accelerated regression consistency penalty, and (b) indicates that the prediction frames are more active after adding the overlap width-height accelerated regression consistency penalty closer to the real frame. By improving the YOLOX model above, the OCC-YOLOX[1] as shown in Fig. 7 is finally obtained.

Experiments

Experimental environment

In order to verify the effectiveness of the proposed algorithm in this paper, the experimental environment uses Windows 10 platform and PyTorch deep learning framework, RTX 3060Ti graphics card, i7-11700 F processor. Python programming language was applied for programming development. ANACONDA was used for environment management and VSCODE was used for IDE.

Datasets preprocessing

Due to the adaptability of detecting occluded objects scenes, a large number of multi-target occlusion scenes need to be selected. Secondly, in order to meet the needs of in-vehicle viewpoint detection, it is necessary to select in-vehicle viewpoint traffic target public datasets, such as KITTI[39], CityPersons[40], and BDD100K[41]. The KITTI dataset is based on the in-vehicle camera captured more than 10,000 real image data of multiple scenes, labelled with eight types of traffic targets, which include various degrees of occlusion and truncation. CityPersons The Cityscapes in-vehicle dataset contains 5,000 images of pedestrians, including a large number of pedestrian occlusion scenes, and the BDD100K dataset collects more than 100,000 images based on the in-vehicle viewpoint, including different lighting, weather, and other scenarios, and annotates 10 types of traffic targets. The above three datasets meet the requirements of the experimental scenarios, so this paper integrates the public datasets KITTI 7481, CityPersons 1599, and BDD100K 2500 images, a total of 11,580 images, as the data source for training and testing, and the selection of images pays special attention to the occlusion scenarios. As shown in Fig. 8, the three selected dataset image samples are shown.

Firstly, we choose KITTI dataset label form as a sample, intercept the category, degree of occlusion and 2D bounding box coordinates in the labels of the three datasets, and convert the TXT label format to VOC format uniformly and reconstruct the picture names. Secondly, we merge similar targets with different category names in different datasets, such as merging Human, Person, and Pedestrian into Pedestrian, and finally establish the vehicle view occlusion object detection dataset with five categories of targets (Car, Pedestrian, Truck, Van, and Cyclist). The occlusion object detection dataset is divided into training set, test set and validation set according to 8:1:1.

Evaluation indicators

The evaluation metrics reflect the model performance by applying the precision-recall (P-R) curve and the detection speed FPS (Frames Per Second). The IoU threshold takes the value of 0.5, in which the average precision of a category is, and the mean average precision (mAP) is the average precision of each category.

[1] [Online]. Available at https://github.com/Bryceder/OCC-YOLOX.

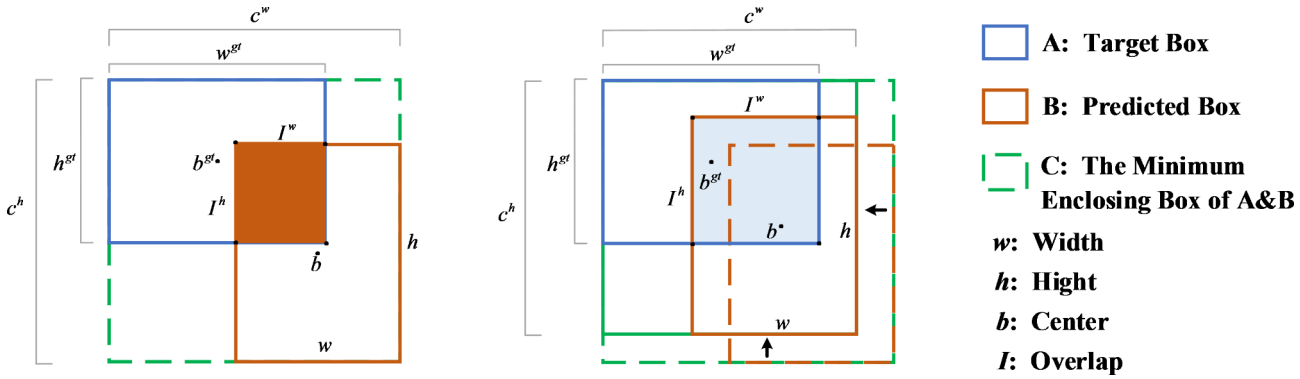

Fig. 6. Schematic diagram of overlapping IoU regression.

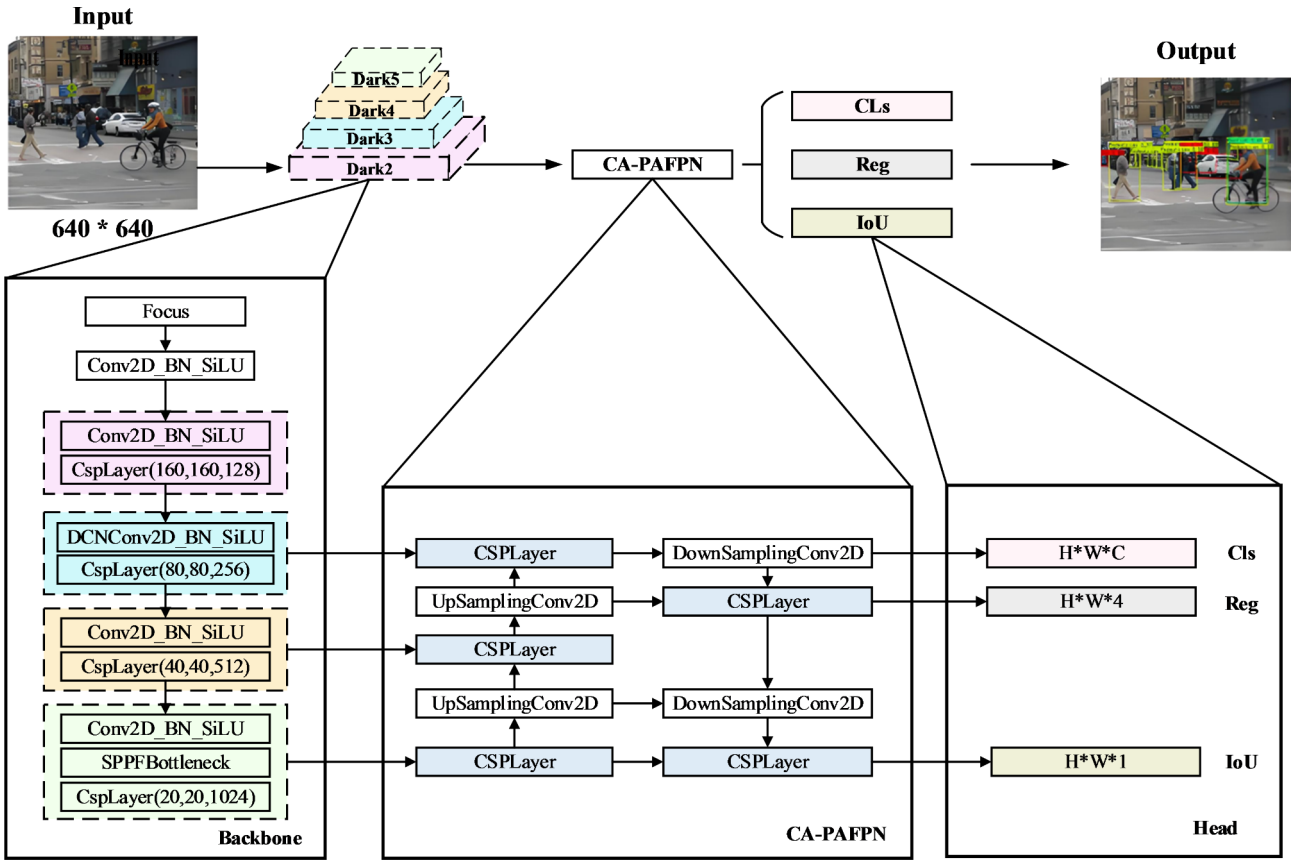

Fig. 7. OCC-YOLOX structural model.

Fig. 8. Sample plot of data set selection.

Detection speed FPS indicates how many frames per second can be detected, the larger the FPS, the higher the detection speed and the higher the real-time performance.

Model training

A single GPU is trained with 300epoch, batch size is 8, adam optimizer is selected, momentum is 0.937, learning rate is set to stochastic gradient descent, initial learning rate is 0.001, and minimum learning rate is 0.00001.

The model training results are shown in Fig. 9. The left side of the figure represents the change in training loss over time, The right side of the figure shows the change in training accuracy for each round. It can be observed that the OCC-YOLOX algorithm reaches the peak accuracy at the 210th round.

Results and discussion

Model testing

The model performance is tested in the test set of the fusion dataset and compared with the YOLOX model, and the comparison results are shown in Table 1 below. In addition, the model test visual image comparison results are shown in Figs. 10, 11, 12, 13, 14 and 15, where Figs. 10, 11 and 12 are the occlusion scene images selected in the test set of the public dataset, and Figs. 13, 14 and 15 are the test images of the occlusion scene collected by ourselves. The left side (a) shows the YOLOX model test results, and the right side (b) shows the OCC-YOLOX model test results. As can be observed from the visual comparison, the OCC-YOLOX algorithm demonstrates

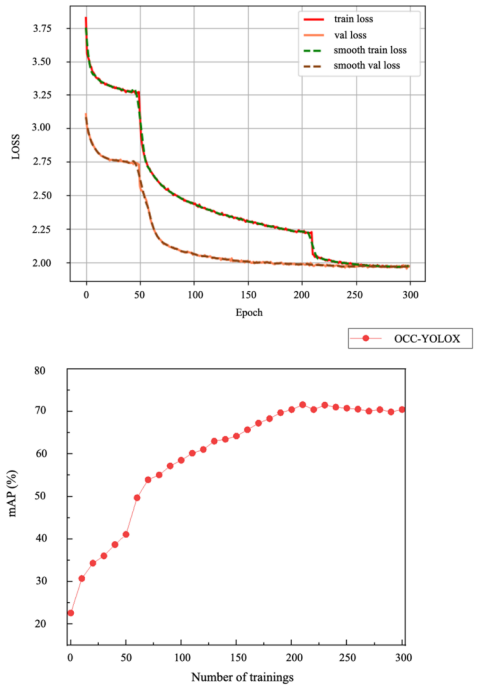

Fig. 9. Comparison of average accuracy.

	Accuracy (%)		Recall rate (%)		AP (%)	
	YOLOX	OCC-YOLOX	YOLOX	OCC-YOLOX	YOLOX	OCC-YOLOX
Car	91.62	**92.12**	68.89	**69.31**	79.77	**81.37**
Van	86.1	**87.17**	80.46	**81.23**	85.7	**86.52**
Truck	74.55	**76.95**	55.56	**55.56**	63.56	**63.18**
Cyclist	85.86	**86.54**	53.35	**52.93**	63.02	**66.19**
Pedestrian	86.14	**86.82**	45.22	**45.7**	58.22	**60.70**

Table 1. Comparison of YOLOX, OCC-YOLOX accuracy, recall, and AP values. The metrics of the OCC-YOLOX are presented in bold.

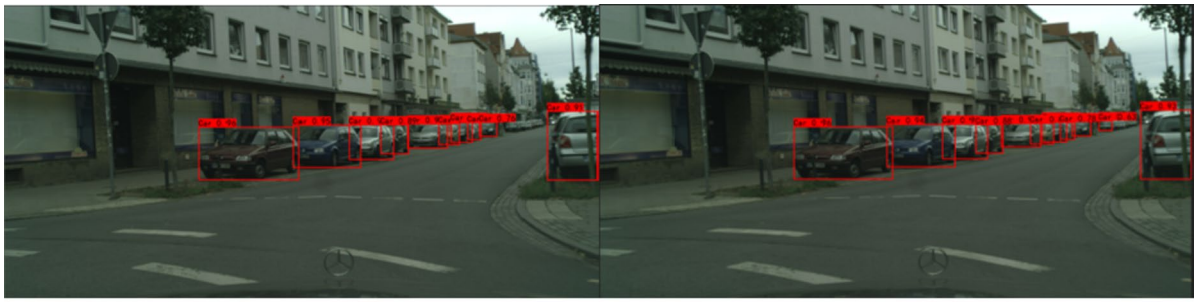

Fig. 10. Comparison of test results on CityPersons dataset.

significant improvements in detecting occluded objects compared to YOLOX. As shown in Fig. 10, OCC-YOLOX effectively detects small and occluded objects at a distance, demonstrating its ability to focus on distant targets with occlusion. Figure 11 illustrates the superior performance of OCC-YOLOX in detecting closely packed pedestrians in the near field within the KITTI dataset. Figure 12 highlights the improved attention given by OCC-YOLOX to occluded vehicles on the opposite lane in nighttime scenarios. Figures 13 and 14 showcase the

(a)¤ (b)¤ ¤

Fig. 11. Comparison of test effects on KITTI dataset.

(a)¤ (b)¤

Fig. 12. Comparison of test effects on BDD100K dataset.

(a)¤ (b)¤

Fig. 13. Comparison of Car and Pedestrian test results.

higher detection rate and reduced false positives of OCC-YOLOX in complex and congested traffic conditions within the campus environment. Lastly, Fig. 15 reveals the robustness of OCC-YOLOX in detecting occluded targets under challenging conditions of snowy weather and low visibility at night. These results collectively demonstrate the effectiveness of OCC-YOLOX in addressing the challenges of occlusion in diverse scenarios.

Ablation experiments

The results of applying the validation set to compare the average accuracy and detection speed of each improvement scheme are shown in Table 2. Replacing the adaptive deformable convolution led to a 0.77%

Fig. 14. Comparison of Car and Cyclists test results.

Fig. 15. Comparison of Car test results at night and in snowy weather.

Modelling	Adaptive deformable convolution	Fast SPP	CA Attention Module	OL-IoU	mAP (%)	Detection speed/FPS
YOLOX					69.86	70.80
OCC-YOLOX	√				70.82	57.60
	√	√			70.74	68.02
	√	√	√		71.60	65.23
	√	√	√	√	71.78	68.55

Table 2. Ablation study of OCC-YOLOX modules.

increase in accuracy while still meeting real-time performance requirements. The theoretical rationale behind this improvement lies in the ability of adaptive deformable convolution to dynamically adjust the convolutional kernel based on the shape and position of the target, effectively extracting target features in occlusion scenarios. Implementing Fast SPP resulted in a 10.42 FPS increase in detection speed without any significant degradation in accuracy. This improvement is attributed to the simplification of the spatial pyramid pooling structure, reducing computational burden while maintaining the capability for multi-scale feature fusion, enabling the model to efficiently handle targets of various sizes. The addition of the CA module led to a 0.86% increase in accuracy. This module enhances the model's ability to focus on occluded targets by emphasizing their positional information, thereby improving detection accuracy in complex occlusion scenarios. Replacing the IoU loss with OL-IoU loss resulted in a 0.18% increase in accuracy. The OL-IoU loss functions more accurately in evaluating the overlap between the predicted and ground truth bounding boxes, especially in cases of partial occlusion, aiding in more precise regression of the bounding boxes. Overall, the refinements implemented in the OCC-YOLOX algorithm have concurrently preserved real-time responsiveness and markedly enhanced its accuracy in detecting objects amid occlusions. These enhancements have not only been corroborated through experimental validation but also enjoy a robust theoretical underpinning.

Modelling	mAP (%) ↑	Detection speed (FPS) ↑	Model weight size (MB) ↓
YOLOv3	30.26	32.56	235.78
YOLOv4-s	54.1	49.55	244.78
YOLOv5s	63.95	53.22	178.64
YOLOX-s	69.86	70.80	34.36
YOLOv6-s	70.58	72.80	60.23
YOLOv7	74.17	35.07	142.32
OCC-YOLOX	**71.78**	**68.55**	**34.64**

Table 3. Comparison of the effectiveness of single-stage object detection algorithms. The metrics for OCC-YOLOX are in bold.

Comparison of models

Table 3 shows the comparison between the method proposed in this paper and the current mainstream single-stage object detection algorithms in terms of average accuracy, detection speed, model size, and model complexity. The algorithm's metrics of mAP, Detection speed, and Model weight size were obtained after training on the occluded object detection dataset. Through the comparison, it can be concluded that the accuracy of the OCC-YOLOX algorithm has achieved competitive results with YOLOv7, while also surpassing other mainstream algorithms. At the same time, the advantage of our work lies in achieving a balance between detection precision and speed. Although the accuracy of OCC-YOLOX is 2.39% lower than that of YOLOv7, its detection frame rate is 93.7% higher than YOLOv7, and the model complexity and size are less than a quarter of YOLOv7, with the added benefit of being lightweight compared to other mainstream algorithms. Therefore, the OCC-YOLOX algorithm has more advantages in applications with an onboard perspective.

Conclusions

In this paper, for the occluded target in vehicle view object detection, based on the single-stage algorithm YOLOX, integrating the backbone network, neck network, and multiple improvement methods of the loss function, we propose the occluded object detection algorithm OCC-YOLOX. fusion of three public datasets and self-picked data to validate this paper's algorithm, which proves the effectiveness of the algorithm. The conclusions are as follows:

(1) The OL-IoU is proposed, and overlapping width-height accelerated regression consistency penalties are designed to facilitate the regression of prediction frames with equal speeds in both the width-height directions, which improves the real-time performance of the algorithm.
(2) Adaptive deformable convolution is used to replace the traditional backbone convolution and change the spatial distribution of feature points to make it more flexible to adapt to the change of target geometry, which effectively improves the object detection accuracy of the occlusion scene. The number, size and connection of spatial pyramid pooling kernels are simplified using FAST SPP, which effectively improves the efficiency of detection. Adding coordinate attention in enhancing the feature extraction part, using two-dimensional information improves the ability of the convolutional network to judge the importance of feature information, which in turn improves the algorithm's feature extraction ability.

Additionally, for heavily occluded objects, the algorithm can only detect their approximate location and shape. It is unable to recover the complete object information, and the algorithm currently does not address the understanding of occlusion relationships. Therefore, it cannot yet be applied to scenarios requiring scene understanding. In future work, we plan to collect more data from extreme occlusion scenarios. We will enhance the algorithm's robustness in these scenarios through techniques such as data augmentation and model refinement. Additionally, we will explore the modeling of occlusion relationships to achieve a better understanding of the occlusion relationships between occluded objects.

Data availability

The data that support the findings of this study are openly available in KITTI dataset at https://www.cvlibs.net/datasets/kitti/raw_data.php; BDD100K datasat at https://www.bdd100k.com/ ; CityPersons dataset at https://www.cityscapes-dataset.com/.

Received: 28 December 2023; Accepted: 5 November 2024
Published online: 12 November 2024

References

1. Li, K., Dai, Y., Li, S. & Bian, M. Development Status and trends of Intelligent Connected Vehicle (ICV) Technology. *J. Automot. Saf. Energy Conserv.* **8**, 1–14 (2017).
2. Zhang, B., Qin, H., Jiang, S., Zheng, J. & Wu, Z. A method of vehicle detection at night based on RetinaNet and optimized loss functions. *Automot. Eng.* **43**, 1195–1202 (2021).
3. Yang, S., Wang, J., Hu, L., Liu, B. & Zhao, H. Research on occluded object detection by Improved RetinaNet. *Comput. Eng. Appl.* **58**, 209–214 (2022).
4. Dalal, N. & Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 886–893. (2005).

5. Felzenszwalb, P., Mcallester, D. & Ramanan, D. A discriminatively trained, multiscale, deformable part model In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 1–8. (2005).
6. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 580–587. (2014).
7. Girshick, R. & Fast, R-C-N-N. June,. In Proceedings of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 1440–1448. (2015).
8. Ren, S., He, K., Girshick, R., Sun, J. & Faster, R-C-N-N. Towards real-time object detection with region proposal Networks[J]. *IEEE Trans. Pattern Anal. Mach. Intell.* **39** (6), 1137–1149 (2017).
9. Redmon, J. & Farhadi, A. YOLOv3: an incremental improvement. *arXiv*. arXiv:1804.02767 (2018).
10. Lin, T. Y. et al. Feature Pyramid Networks for Object De-tection. In Proceedings of the IEEE Conference on Computer Vision and Pattern R ecognition, Honolulu, HI, USA, 21–26, 936–944. (2017).
11. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23, 8759–8768. (2018).
12. Hou, Q., Zhou, D. & Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25, 13708–13717.
13. Yu, J., Jiang, Y., Wang, Z., Cao, Z. & Huang, T. Unitbox: an advanced object detection network. *arXiv*. 1608.01471 (2016).
14. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1904–1916 (2015).
15. Bochkovskiy, A., Wang, C. Y. & Mark Liao, H. Y. YOLOv4: Optimal Speed and Precision of Object Detection. arXiv:2004.10934. (2020).
16. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 779–788. (2016).
17. Redmon, J. & Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 6517–6525. (2017).
18. Liu, W. et al. SSD: Single Shot MultiBox Detector. in *Computer Vision – ECCV 2016,Lecture Notes in Computer Science* 21–37, (2016). https://doi.org/10.1007/978-3-319-46448-0_2
19. https://github.com/ultralytics/yolov5, 2021.
20. Ge, Z., Liu, S., Wang, F. & Sun, J. YOLOX: Exceeding Yolo series in 2021. *arXiv*. arXiv:2107.08430 (2021).
21. Wang, X., Han, T. X., Yan, S. & An HOG-LBP human detector with partial occlusion handling. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 32–39. (2009).
22. Chan, K. C., Ayvaci, A. & Heisele, B. Partially occluded object detection by finding the visible features and parts. In Proceedings of IEEE International Conference on Image Processing, Quebec City, QC, Canada, 2130–2134. (2015).
23. Bodla, N., Singh, B., Chellappa, R., Davis, L. S. & Soft -NMS -- Improving Object Detection With One Line of Code. In Proceedings of IEEE International Conference on Computer Vision, Venice, Italy, 5562–5570. (2017).
24. Kortylewski, A., He, J., Liu, Q. & Yuille, A. L. June,. Compositional convolutional neural networks: a deep architecture with innate robustness to partial occlusion, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 8940–8949. (2020).
25. Phan, H. N., Pham, L., Tran, L. H. & Ha, D. N. S.V. Occlusion vehicle detection algorithm in crowded scene for traffic surveillance system. In Proceedings of International Conference on System Science and Engineering, Ho Chi Minh City, Vietnam, 215–220. (2017).
26. Zhang, S., Wen, L., Bian, X., Lei, Z. & Li, S. Z. Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd. in Computer Vision – ECCV 2018,Lecture Notes in Computer Science 657–674, https://doi.org/10.1007/978-3-030-01219-9_39 (2018).
27. Reddy, N. D., Vo, M., Narasimhan, S. G. & Occlusion-Net 2D/3D Occluded Keypoint Localization Using Graph Networks. in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) doi: (2019). https://doi.org/10.1109/cvpr.2019.00750
28. Li, A., Yuan, Z. & SymmNet: A Symmetric Convolutional Neural Network for Occlusion Detection. (2018).
29. Cui, Y., Yan, L., Cao, Z. & Liu, D. TF-Blender: Temporal Feature Blender for Video Object Detection. in 2021 IEEE/CVF International Conference on Computer Vision (ICCV) https://doi.org/10.1109/iccv48922.2021.00803 (2021).
30. Liu, D., Cui, Y., Tan, W. & Chen, Y. S. G. N. Spatial Granularity Network for One-Stage Video Instance Segmentation. in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/cvpr46437.2021.00969 (2021).
31. Wang, A., Sun, Y., Kortylewski, A. & Yuille, A. Robust Object Detection Under Occlusion With Context-Aware CompositionalNets. in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/cvpr42600.2020.01266 (2020).
32. Xie, H., Zheng, W. & Shin, H. Occluded pedestrian detection techniques by deformable attention-guided network (DAGN). *Appl. Sci.* **6025** https://doi.org/10.3390/app11136025 (2021).
33. Hu, J., Shen, L. & Sun, G. Squeeze-and-Excitation Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23, 7132–7141. (2018).
34. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. & Cbam Convolutional block attention module. In Proceedings of the 2018 IEEE European conference on computer vision (ECCV), Munich, Germany, 8–14, 3–19. (2018).
35. Dai, J. et al. Deformable convolutional networks. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26, 764–773. (2017).
36. Zhu, X., Hu, H., Lin, S. & Dai, J. Deformable ConvNets V2: More Deformable, Better Results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20, 9300–9308. (2019).
37. Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI conference on artificial intelligence. **34**(07): 12993–13000.(2020)
38. Zhang Y F, Ren W, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing. **506**: 146–157.(2022).
39. Geiger, A., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2012.6248074 (2012).
40. Cordts, M. et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 3213–3223. (2016).
41. Yu, F. et al. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2633–2642. (2020).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (52202503), Hebei Natural Science Foundation (F2024203112, F2022203054), Major Scientific and Technological-Special Projects in Jilin Province and Changchun City (20220301008GX), Science and Technology Project of Hebei Education Department (BJK2023026) and Hebei Higher Education Society's "14th Five-Year Plan" 2024 Annual Higher Education Research Project Achievement (GJXHZ2024-05).

Author contributions

Conceptualization, L.J.; Methodology, X.L.; Software, H.Z.; Validation, X.L., H.Z.; Formal Analysis, X.L.; Investigation, H.Z.; Resources, B.G.; Data Curation, B.G.; Writing—original draft preparation, X.L.; Writing—review and editing, H.Z.; Visualization, H.Z.; Supervision, B.G.; Project administration, B.G.; Funding acquisition, L.J. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.