Image Processing, Electronics and Computers
L. Trajkovic et al. (Eds.)

doi:10.3233/ATDE240474

POD-YOLO: YOLOX-Based Object Detection Model for Panoramic Image

Mingxuan ZHANG, Hua LI[1], Qi LI, Mingming ZHENG and Iullia DVINIANINA
College of Computer Science and Technology, Changchun University of Science and Technolog, Changchun 130000, Jilin, China

Abstract. Object detection is an important branch of panoramic image scene understanding. Panoramic images possess characteristics such as a wide field of view, significant distortion, and rich content, which leads to constant changes in the convolutional domain of the panoramic image, thus resulting in the fact that using a convolutional kernel of the same shape is not sufficient to perform convolutional feature extraction on the panoramic image. Therefore, traditional perspective-based object detection algorithms frequently exhibit poor performance on panoramic challenges. An enhanced YOLOX-based panoramic image object detection method is suggested as a solution to this problem. To improve the feature extraction capabilities for distorted objects in panoramic photos, an effective feature extraction network is built by integrating deformable convolution v2 and atrous spatial pyramid pooling (ASPP) into the backbone feature extraction network. Furthermore, the accuracy of panoramic image detection is greatly increased by further strengthening the extraction of image channel features by integrating an enhanced attention mechanism between the feature extraction network and the backbone network. Experimental results demonstrate that the proposed panoramic object detection model achieves an average precision (mAP) of 73.35% on a self-built panoramic image dataset, compared with the existing traditional target detection model, achieving significant performance improvement.

Keywords. panoramic image; object detection; YOLOX; deformable convolution; ASPP;POD-YOLOX

1. Introduction

As a direction for the future development of transportation, smart cars possess immense market potential and development space. Globally, smart car technology is rapidly advancing, attracting widespread attention from automotive manufacturers, technology companies, and research institutions. In this wave of development, environmental perception and scene understanding have become important areas of research in smart car technology. Environmental perception enables smart cars to acquire real-time information about the surrounding environment, while scene understanding allows for a deep understanding and inference of road scenes, vehicle behavior, pedestrian behavior, and more. Research on panoramic environmental perception and scene understanding technology is of great significance for enhancing the safety, reliability, and intelligence level of smart cars.

[1] Corresponding Author: Hua LI, lihua@cust.edu.cn.

Panoramic object detection, as a crucial component of panoramic environmental perception and scene understanding technology research, has been a major focus of attention in recent years. Currently, most object detection tasks are based on traditional perspective images, with detection limited to specific regions within the image. In contrast to traditional perspective images, panoramic images provide 360° scene information, encompassing multiple scenes that would be captured in traditional perspective images, thus meeting the requirement for a wide field of view in scene perception.

2. Related Work

Currently, deep learning target detection algorithms based on traditional perspective images mainly rely on convolutional neural networks (CNN [1]) to automatically extract image feature information to complete the target detection task, which is usually divided into two categories: one is the candidate region-based deep learning target detection algorithms represented by the R-CNN [2] series, also known as two-stage algorithms. Generate a priori frames through image features in the candidate regions where target objects may exist, and then classify and positionally correct the generated a priori frames; Secondly, regression-based deep learning target detection algorithms represented by YOLO [3] series and SSD [4] series, also known as One-stage algorithms, which omit the a priori frame generation stage and use the extracted deep features directly for target localization classification.

However, panoramic image object detection faces unique challenges. Firstly, panoramic images typically contain a large amount of background information, resulting in relatively small and dispersed targets, which increases the complexity of detection, as shown in Figure 1. Secondly, due to distortion and projection issues in panoramic images, traditional object detection algorithms may fail in panoramic images, requiring adaptation to the specific geometry and perspective transformations of panoramas. Additionally, due to the large field of view in panoramic images, the labeling process is time-consuming and labor-intensive, leading to a lack of publicly available datasets for panoramic image object detection, making panoramic scene understanding challenging.

Figure 1. Schematic diagram of panoramic image.

Although traditional object detection algorithms are poorly adapted to panoramic images, in recent years, due to the wide range of application areas, researchers have made targeted improvements to traditional algorithms to better fulfill the task of panoramic image detection. Wang et al. [5] proposed an R-SSD network, which filters negative

sample candidate frames by adding an RPN network before the original SSD network, and then adds a transmission conversion module between the two networks to achieve feature fusion. Experimental results show that the algorithm can better achieve the detection of vehicle targets in panoramic video images. Tong et al. [6] proposed a multi-scale feature pyramid network (MS-RPN), which effectively improves the detection accuracy of targets in panoramic images. Peng et al. [7] added an aberration adaptive module to correct the aberration information in the panoramic image before the Faster R-CNN network, the R-CNN network has better input features and the detection effect is improved a lot. Deng et al. [8] proposed a deep learning method based on R-CNN for the detection of objects with strong aberrations on the indoor panoramic image. They first pre-processed the panoramic image for distortion and then used R-CNN algorithm to implement the detection work. Cai et al. [9] established an imaging parameter model based on the characteristics of imaging aberrations in panoramic images, through which the columnar unfolding of the panoramic image was obtained, and then the number of longitudinal grids of the YOLO network was improved to adapt to the object to be detected, which is a good implementation of real-time detection of multi-targets. At present, many scholars have constructed an object detection model for panoramic images based on the available YOLO model [10-12]. Pokuciński et al. [13] demonstrated that publicly available YOLO detectors even without fine-tuning can be successfully used for target detection in ERP.

To address the aforementioned issues, this paper proposes a panoramic image object detection model based on the YOLOX algorithm. Our contributions mainly include: 1) Improving the network detection capability for distorted objects in panoramic images by introducing deformable convolution v2 and ASPP into the backbone feature extraction network; 2) Adding a multi-attention mechanism on three-dimensional feature layers after feature extraction to enhance the model detection performance; 3) Collecting outdoor panoramic images from the web and annotating them as an experimental dataset, followed by some data augmentation to enhance the model generalization ability. We then evaluate the network performance and compare it with other object detection algorithms.

3. Related Knowledge Content

3.1. Structure of YOLOX

YOLOX [14] (Exceeding YOLO Series in 2021) is YOLOX was proposed by Kuangshi Technology in 2021, and it is the current single-stage target detection algorithm with better detection speed and recognition accuracy. Since panoramic images have a wide field of view, large distortion, and other characteristics, the YOLOX model with more accurate detection accuracy is selected, and the YOLOX algorithm is improved to construct a detection algorithm suitable for panoramic images. The network structure is mainly divided into three parts, namely, the backbone feature extraction network (Backbone), the neck enhancement feature extraction network (Neck), and the YOLOX is different from the YOLO series of the detection head (Decoupled Head), the network structure schematic diagram is shown in Figure 2.

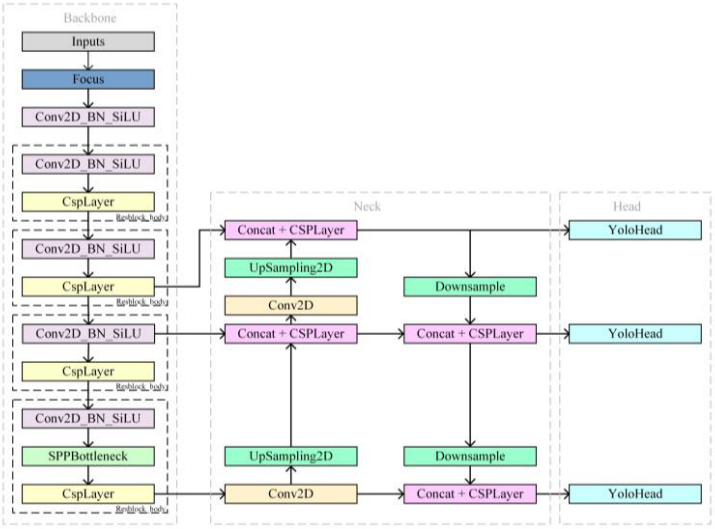


Figure 2. Schematic diagram of the YOLOX network architecture.

The backbone network consists of modules such as Focus, CSPnet, and SPP, which effectively extract image features. The enhanced feature extraction network in the neck adopts the FPN-PAN structure. FPN enables detection at different scales for the same object. By upsampling and merging deep semantic information with low-level target position information, it enhances the network's learning ability for features of objects to be detected. Unlike previous YOLO series, the detection head is divided into two parts in YOLOX. These parts are implemented separately and integrated only during the final prediction stage.

3.2. Deformable Convolution v2

The Deformable Convolution Network (DCN) family of algorithms was proposed to enhance the ability of models to learn complex target invariants. In DCNv1 [15], the authors proposed the modules Deformable Convolution and Deformable Pooling. In DCN v2 [16], the authors added a weighting module to these two deformable modules to enhance the ability of deformable convolutional networks to capture important information.The motivation for the proposal of DCN v2 is to reduce the irrelevant and disturbing information in DCN v1 and to improve the model's adaptability to different geometric changes. In order to solve the problem of introducing some irrelevant regions, in DCN v2, not only the offset of each sampling point is added, but also a weight coefficient is added to distinguish whether the introduced region is a region of interest or not, if the region of this sampling point is not of interest, the weight is simply learnt to be 0:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{1}$$

In DCN v1, the introduced offsets are used to locate regions containing effective information, while in DCN v2, introduced weight coefficients assign weights to the

identified positions. These two aspects ensure accurate extraction of effective information. Introducing deformable convolution in the backbone feature extraction network can improve the detection accuracy of distorted objects in panoramic images. As shown in Figure 3, the first image depicts a standard convolution operation, while the following three images depict deformable convolutions. Such convolution kernels can expand to a large range during training.

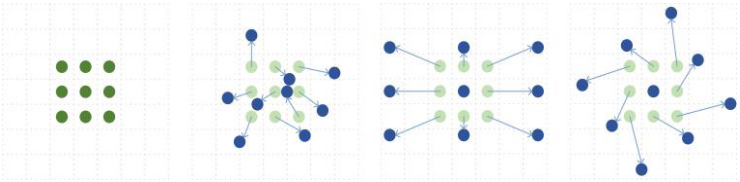


Figure 3. Sampling method of deformable convolution.

4. Improved YOLOX Algorithm

To better adapt to the characteristics of panoramic images, this paper proposes a new panoramic image object detection model called POD-YOLO. The backbone network integrates deformable convolution v2 and replaces SPP with the DCSPDarknet backbone feature extraction network to better extract features of distorted objects in panoramic images. Attention mechanisms are added to three effective feature layers extracted by the backbone network to enhance the network's ability to capture global features, thereby improving detection accuracy. The model prediction network utilizes YOLO layers to predict outputs at large, medium, and small scales, ultimately completing the detection task. The structure of the panoramic image object detection model established in this paper is illustrated in Figure 4.

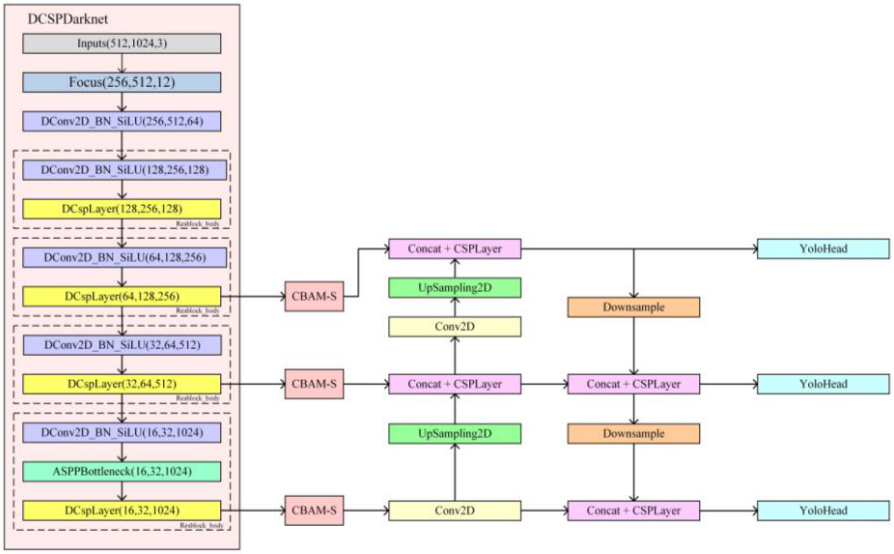


Figure 4. Schematic diagram of the POD-YOLO model structure.

4.1. Deformable Backbone Feature Extraction Network

4.1.1. DCSPDarknet

The primary challenge faced in panoramic image tasks is object distortion caused by projection transformations. Directly applying traditional convolutional neural networks (CNNs) to detect objects in panoramic images often fails to achieve the desired performance. Since panoramic images are typically generated by stitching together multiple perspective images captured at the same location by perspective cameras, the Equirectangular Projection (ERP) can be considered as a transformation from non-Euclidean space to Euclidean space, introducing distortion to objects in panoramic images. As a result, the projections of objects have irregular shapes, and distortion becomes particularly significant for pixels near the poles or the image plane. Therefore, standard convolutions are not suitable for processing panoramic images.

To address the above-mentioned issues, the backbone feature extraction network in this paper is DCSPDarknet, primarily composed of DCSP modules, ASPP modules, and Darknet. DCSPDarknet is based on the CSPDarknet backbone network of YOLOX, drawing from the experience of D-CSP to form the backbone structure. The structure is mainly divided into four parts, each containing DCSP modules, as shown in Figure 5.

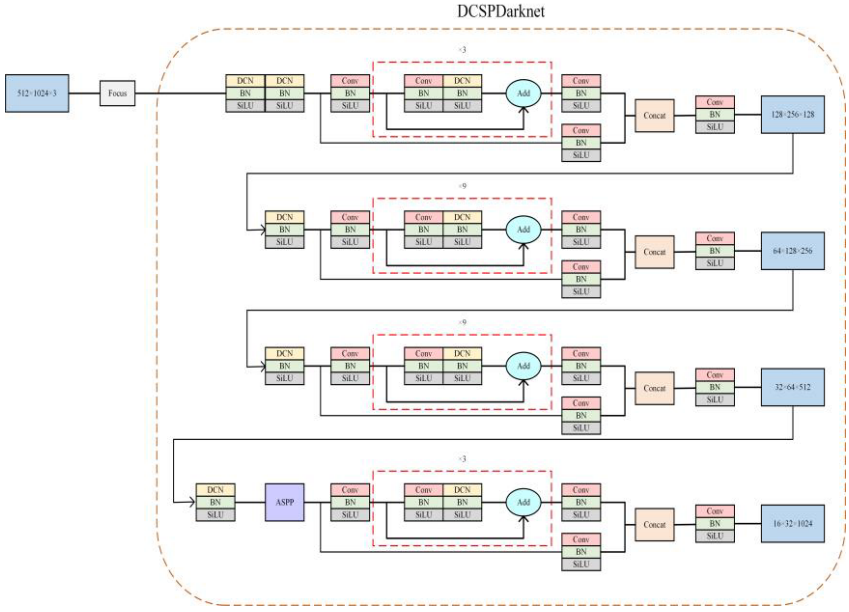


Figure 5. Detailed diagram of the backbone feature extraction network DCSPDarknet.

Due to the distortion present in both the background and objects in panoramic images, the convolutional field in panoramic images constantly changes. Convolutional kernels with the same shape are insufficient to extract complete features of objects in the image. To enhance the feature extraction capability of objects in panoramic images, the model introduces deformable convolution (DCNv2) into the CSPDarknet53 backbone feature extraction network. Additionally, to balance detection accuracy and speed, we choose to replace some convolutions in the network with deformable convolutions

(DCNv2). This is because deformable convolutions require more computational resources than traditional convolutions, as they need to compute offsets and weights for each region.

Specifically, in a standard convolutional network, 1×1 convolutions compress and rearrange features to reduce computational complexity and maintain model training stability, while 3×3 convolutions are typically used for feature extraction. Therefore, this paper only modifies the smallest component, the CBS module, in the backbone network structure. The 3×3 standard convolution (Conv) in the CBS module is replaced with deformable convolution (DCNv2) to create a new module called DBS. In addition, the 3×3 CBS module in the Res Unit residual module of CSP is replaced with the DBS module, while retaining the original 1×1 CBS module. The DCSP module is illustrated in Figure 6.

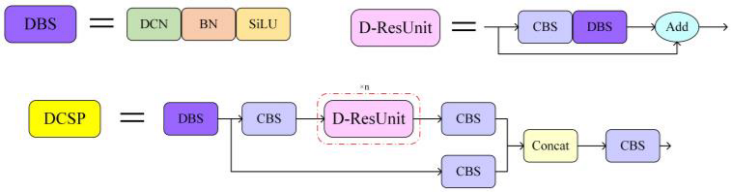


Figure 6. DCSP module.

Each DCSP module has a convolutional kernel size of 3×3, which can play the role of up-sampling and improve the computational efficiency, and has a good feature extraction capability. Since there are four DCSP modules in the backbone network and the size of the input image is 512×1024, the feature map is varied as 512×1024→128×256→64×128→32×64→16×32, and finally a 16×32 feature map is obtained. The SiLU activation function is used in the backbone network, which is a smooth activation function that better allows the image information to penetrate the neural network, thus improving the detection accuracy and the generalization ability of the model.

4.1.2. ASPP

To address the challenges posed by the large resolution and wide field of view of panoramic images, the ASPP (Atrous Spatial Pyramid Pooling) module is introduced into the backbone network. This module employs parallel dilated convolutional layers with different dilation rates to capture multiscale information from the feature map. The output results are then fused to obtain the detection results of the image. The goal is to have a large receptive field for the extracted features while maintaining the resolution of the feature map without significant reduction (too much reduction in resolution can lead to loss of detailed information at the image boundaries). However, these two objectives are contradictory: to achieve a larger receptive field, larger convolutional kernels or larger strides in pooling are needed, which would result in either excessive computational complexity or loss of resolution. Atrous convolution is used to resolve this contradiction by allowing the model to obtain a larger receptive field while preserving resolution, effectively aggregating contextual information, and obtaining multiscale information.

Specifically, the ASPP module has five branches, as shown in Figure 7, and each branch provides different sizes of sensory fields respectively, the convolutional layer with a large void rate provides more global contextual feature information for the

network, and the convolutional layer with a small void rate supplements the detailed information for the network. The feature maps obtained by processing under different null rates are the multi-scale information of the same layer feature maps. In Figure 8, the convolution operation of the first four branches does not change the spatial dimension of the feature map and only changes the channel dimension size to obtain X_1, X_2, X_3, X_4; the feature map X_5 after the global average pooling operation reduces the spatial dimensions to the same size as that of the input feature map $X_5^{'}$, through bilinear interpolation up-sampling. The feature maps after five branch processing are spliced in the channel dimension, and after splicing, the feature maps are compressed to the specified dimensions by 1×1 convolution at the same time, and the multi-scale information interactions in the same hierarchical feature maps are realized in the channel dimension.

The ASPP module can extract high-level features from deep feature maps of the network, obtaining feature maps with rich semantic information. This enhances the model's capability to extract features of wide-field objects in panoramic images, providing more abstract and advanced features for the next stage of feature enhancement in the network.

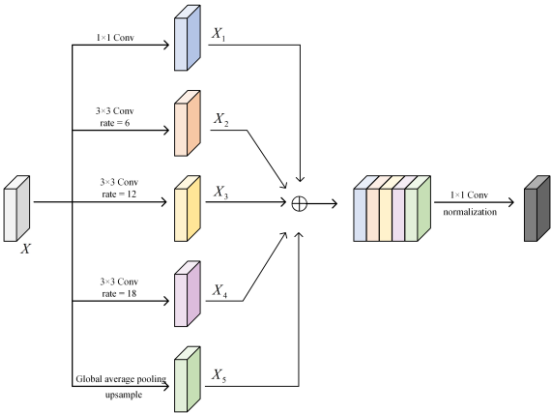


Figure 7. ASPP module.

4.2. Multi-attention Mechanism

People can perceive a variety of things through their eyes, capturing a wealth of information about the world. However, human consciousness can selectively focus on what it considers important while ignoring unimportant information, thus shielding oneself from the overwhelming volume of information. The attention mechanism works similarly, weighting the inputs of the model before outputting them. This allows the model to weight features based on learned attention weights, assigning lower weights to less relevant features and higher weights to more relevant ones. As a result, irrelevant background information in the image is weakened, and important information is separated.

CBAM is a lightweight attention mechanism module that can be easily embedded into existing convolutional neural network architectures without the need for any additional computations. To achieve feature recalibration of the original inputs, it utilizes both max pooling and average pooling. It adopts a serial approach to generate the final

output weights from information in both channel and spatial dimensions. The structure of the CBAM module is illustrated in Figure 8.

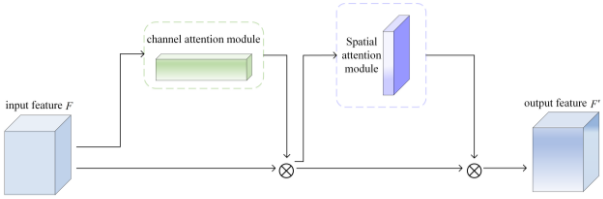


Figure 8. Schematic diagram of the CBAM module structure.

The CBAM module adopts a 'serial' structure, where the input feature F first undergoes the one-dimensional convolution M_c of channel attention to obtain a weighted feature map, and the output F_1 of the channel attention module is used as the input of the spatial attention module, and then undergoes the spatial attention M_s to obtain the final output feature F'. The formula of the CBAM module Tabulation:

$$F_1 = M_c(F) \otimes F \tag{2}$$

$$F' = M_s(F_1) \otimes F_1 \tag{3}$$

However, due to the wide field of view of panoramic images, which encompass a 360° range of scenes from a single viewpoint, the background tends to be complex. Typically, rectangular panoramic images are selected as detection data for object detection. When a panoramic image is unfolded along a certain axis, distortions occur, leading to poor performance of conventional object detection models. To enable the model to pay more attention to the characteristics of panoramic images, this paper introduces the CBAM-S mixed attention mechanism.

Unlike CBAM, where either channel attention or spatial attention is applied first, in CBAM-S, the order of attention mechanisms affects the features input to subsequent modules to some extent, leading to unstable model training. The CBAM-S module in this paper adjusts the original "serial" structure to a "parallel" structure, where both attention modules learn simultaneously on the initial input features. It no longer considers the sequence of channel and spatial attention modules. Moreover, it not only considers spatial relationships but also emphasizes position information in channels. For the feature maps extracted by the backbone network, CBAM-S can effectively locate and recognize objects of interest, thereby improving detection efficiency and accuracy. The schematic diagram of the CBAM-S structure is shown in Figure 9, and the formulation of CBAM-S is as follows:

$$F' = M_c(F) \otimes M_s(F) \otimes F \tag{4}$$

The channel attention module mainly consists of two stages: compression and excitation. Compression involves processing the input features through global average

pooling, which compresses the features along the spatial dimension to reduce computational complexity. The calculation process is as follows:

$$Z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i, j) \tag{5}$$

To effectively utilize the aggregated information obtained during the compression process, excitation is performed to comprehensively capture channel dependencies. The calculation process is as follows:

$$A_c = F_{ex}(Z_c, W) = \sigma(W_2 \delta(W_1 Z_c)) \tag{6}$$

Finally, the attention weights obtained earlier are applied to each channel's features by weighting. Specifically, the normalized weights obtained after compression and excitation operations are assigned to each channel's features. The calculation process is as follows:

$$F_1' = F(U_c, A_c) = A_c U_c \tag{7}$$

The spatial attention module first performs average pooling and max pooling along the channel dimension and then concatenates the feature maps generated by these operations. On the concatenated feature map, convolutional operations are used to produce the final spatial attention feature map. The calculation process is as follows:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}(F^s_{avg}; F^s_{\max})) \end{aligned} \tag{8}$$

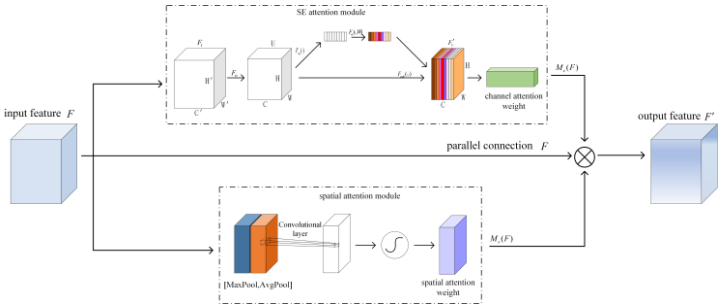


Figure 9. Schematic diagram of the CBAM-S module structure.

5. Experiment

5.1. Dataset Preparation

Currently, there is no benchmark public dataset specifically designed for panoramic image object detection research. Most existing studies are based on self-built datasets or

utilize datasets proposed in previous literature for experimental purposes. In this paper, we utilize the omnidirectional street-view (OSV) dataset provided by Wuhan University [17], which contains panoramic images of streets. We annotate objects in the panoramic images using the labelimg image annotation tool. The annotated dataset includes seven classes of objects: light, car, traffic_sign, crosswalk, warning_line, building, and person, totaling 14,394 objects. This is an extension of the original OSV dataset, with the addition of three new classes: traffic_sign, person, and building. Overall, the number of annotated objects has doubled compared to the original OSV dataset. We refer to the annotated dataset as OSV-B. The specific number of objects for each class is shown in Table 1.

Table 1. Dataset used in the experiment

	light	car	traffic_sign	crosswalk	warning_line	building	person	total
OSV	1777	2059	-	867	355	-	-	5058
OSV-B	1811	2760	1053	959	516	7114	181	14394

5.2. Experimental Environment

The detailed configuration information used in this paper's experiments is shown in Table 2.

Table 2. Experimental platform configuration

Configuration	**Release parameter**
Operating system	Windows11
CPU	Intel Core i5-13400F
RAM	32G
GPU	NVIDIA GeForce RTX 4070
VRAM	12G
Deep Learning Framework	Pytorch

5.3. Analysis of Experimental Results

To confirm the effectiveness of the proposed model, five sets of experiments were designed to investigate the overall impact of deformable convolution, ASPP, CBAM-S, and data augmentation on panoramic image detection. The results obtained are shown in Table 3.This paper utilizes common metrics for object detection to analyze the performance of the panoramic image object detection model. Evaluation metrics include Average Precision (AP) and mean Average Precision (mAP).

For the re-annotated panoramic dataset OSV-B, to enhance the model's generalization ability for detecting distorted objects, a data augmentation method tailored for panoramic images is proposed. Firstly, images are horizontally flipped and vertically flipped with a probability of 50%, augmenting the diversity of training data for panoramic images. Secondly, an affine transformation method is employed to scale, translate, and rotate images.

The accuracy of the different classifications is shown in Table 4. In the dataset used in this paper, distorted objects mainly include building and crosswalk located at the poles. From the experimental results, it is evident that our improved model compared to the original network has increased the accuracy for building by 5.86% and for crosswalk by 13.27%. However, there is a slight decrease in accuracy for caution lines, mainly due to the limited number of samples in the caution line class in the dataset, uneven distribution in images, and most of them being concentrated near the panoramic image acquisition

device, causing some occlusion to caution lines. Hence, there is a slight decrease in accuracy for warning_line detection. Nevertheless, the model shows varying degrees of improvement in detecting the other five classes of objects, leading to an overall accuracy improvement of 8.03%. In summary, the model proposed in this paper demonstrates effective performance in panoramic image object detection tasks.

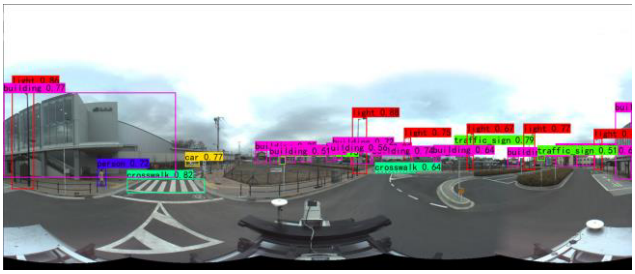
Table 3. Results of ablation experiments.

No.	CBAM-S	DCN v2	ASPP	data aug	mAP/%
1					65.32
2	√				66.08
3	√	√			67.55
4	√	√	√		67.18
5	√	√	√	√	73.35

Table 4. Results of ablation experiments by class

No.	AP/%						
	building	car	crosswalk	light	person	traffic_sign	warning_line
1	81.94	75.81	60.33	86.72	19.4	50.73	82.32
2	83.19	75.18	66.54	87.51	21.89	43.63	84.63
3	83.88	74.65	72.37	88.11	27.33	43.41	83.08
4	84.99	76.5	70.1	88.47	25.71	44.28	80.2
5	87.8	76.56	73.6	88.56	55.9	53.74	77.27

Figure. 10 illustrates the visual detection results of the proposed model in different panoramic scenes, demonstrating that POD-YOLO can accurately detect distorted objects. This indicates that the introduction of deformable convolution and ASPP allows the model to better fit the shapes of objects in panoramic images during feature extraction. Furthermore, the experimental results show that the proposed model exhibits good detection performance even for small-scale objects in dense scenes, validating the effectiveness of the hybrid attention mechanism.



(a) visual diagram of different categories

(b) visual illustration of distorted buildings

Figure 10. Selected visualizations of model predictions.

5.4. Comparative Experiment

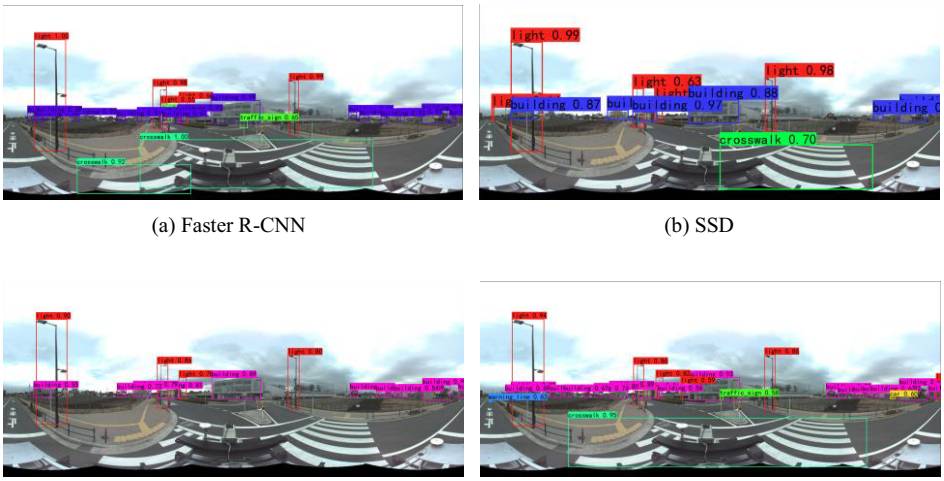
As there is currently no standard public dataset for panoramic image object detection tasks, most works rely on self-collected panoramic data or publicly available datasets from existing literature for experimental comparisons. This study tests the proposed panoramic object detection model on the self-annotated panoramic image dataset OSV-B and compares it with other network models, including Faster R-CNN, SSD, YOLOv3, and YOLOv5, which are common object detection models. The comparative experimental results on the OSV-B dataset are shown in Table 5.

Table 5. Comparative experimental results on OSV-B dataset

	class	method				
		SSD	Faster R-CNN	YoloV3	YoloV5	POD-YOLO（Ours）
AP/%	building	77.74	62.65	76.97	78.49	87.8
	car	56.96	27.04	68.39	63.52	76.56
	crosswalk	53.64	26.67	51.1	44.62	73.6
	light	72.96	54.84	75.16	75.09	88.56
	person	25.32	11.02	25.68	21.36	55.9
	Traffic_sign	17.71	5.96	40.43	26.7	53.74
	warning_line	48.82	27.06	68.76	66.86	77.27
mAP/%		50.45	30.75	58.07	53.81	73.35

From Table 5, it can be observed that Faster R-CNN, as a representative of two-stage object detection algorithms, has limitations due to the structure of the two-stage network and constraints on the number of positive and negative samples during training, leading to significant limitations in prediction accuracy. Since the baseline model used in this paper is a one-stage object detection algorithm, we compared it with some representative one-stage object detection algorithms. SSD has relatively fewer convolutional layers for low-level features, resulting in insufficient extraction of feature information and a higher rate of missed detections for the target objects. YOLOv5, as a representative method in the YOLO series, adapts anchor boxes by calculating different optimal anchor box values during training. However, for distorted objects in the panorama, the calculated anchor boxes may not be optimal, resulting in lower detection accuracy. The experimental data comparison results in Table 5 show that POD-YOLO outperforms all the comparison models in terms of accuracy, especially demonstrating good detection performance for highly distorted objects in the panorama.

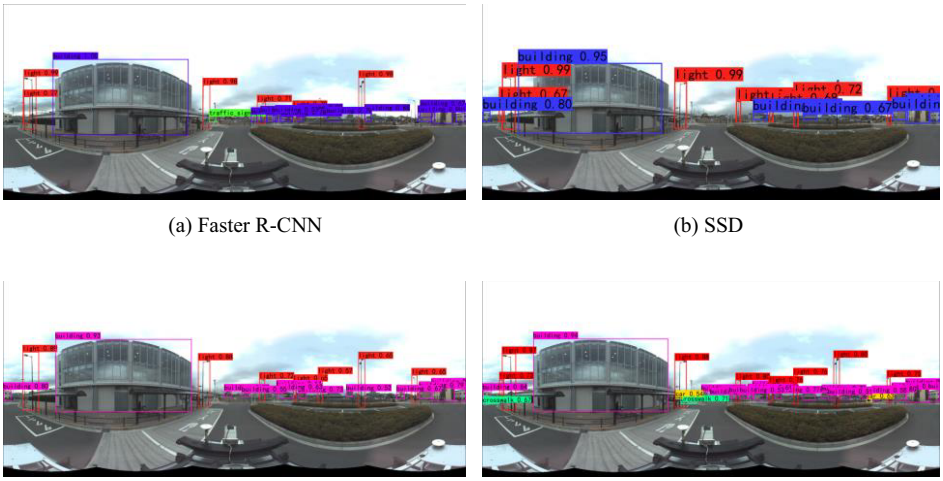
The visualization results of the comparative experiments are shown in Figure 11 and Figure 12, from left to right: Faster R-CNN, SSD, YOLOv5, and POD-YOLO. From the first horizontally compared image, it can be observed that Faster R-CNN detects two crosswalks at the bottom of the image, SSD only detects half of a crosswalk, YOLOv5 fails to detect the crosswalk, while POD-YOLO detects the distorted crosswalk completely with a confidence of 0.95. In the other three horizontally compared images, although the first three algorithms have high confidence for individual categories, they suffer from higher overall missed detection rates, whereas our model exhibits fewer missed detections for small objects across the entire scene range. In summary, the proposed model in this paper demonstrates good detection capabilities in panoramic images, achieving higher accuracy in object detection. However, there are still missed detection issues in the model, which can be addressed through further research.



(a) Faster R-CNN (b) SSD

(c) YOLO v5 (d) ours

Figure 11. Visualization of the distortion of crosswalk.



(a) Faster R-CNN (b) SSD

(c) YOLO v5 (d) ours

Figure 12. Visualization of the distortion of building.

6. Conclusions

In this paper, for the task of panoramic image target detection, for the panoramic image that exists in high resolution, distortion, large field of view, and other issues, proposed a panoramic image target detection method POD-YOLO based on the improvement of YOLOX. Through enhancing the feature extraction network of the backbone network and subsequent dataset processing, an object detection model capable of perceiving distortions in panoramic images is developed, thereby improving the performance of

panoramic image object detection algorithms. Testing on a self-constructed panoramic image dataset, comparing the detection results of this model with other models, and conducting quantitative analysis using AP (Average Precision) values and mAP (mean Average Precision) values, experimental results indicate that the improved YOLOX detection algorithm demonstrates the best detection performance both in visual observation and numerical comparison aspects. Therefore, applying this model to panoramic image object detection tasks has significant advantages.

Model parameters have increased as a result of the addition of deformable convolution, even though the enhanced object detection algorithm with its distortion perception capabilities has produced good detection results. Therefore, further research will be conducted on the model lightweight.

Acknowledgments

This project was supported by Science and technology research planning project of Education Department of Jilin Province (No. JJKH20240951KJ)

References

[1] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Communications of the ACM 60, 2017, 6: 84-90.
[2] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. IEEE Computer Society, 2014: 580-587.
[3] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection. Computer Vision & Pattern Recognition. IEEE, 2016: 779-788.
[4] Liu Wei, Dragomir Anguelov, Dumitru Erhan, et al. Ssd: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 2016, Part I 14: 21-37.
[5] Wang D, Zhao M, Liu Y, et al. Improved R-SSD Panoramic Video Image Vehicle Detection Algorithm. Computer Engineering and Applications, 2021, 57(3): 189-195.
[6] Li Y, Tong G, Chen H, et al. Object Detection for Panoramic Images Based on MS-RPN Structure in Traffic Road Scenes. IET Computer Vision, 2019, 13(5): 500-506.
[7] Peng J L. Research on object detection algorithm based on panoramic image. Guangdong University of Technology, 2022.
[8] Deng F, Zhu X, Ren J. Object detection on panoramic images based on deep learning. International Conference on Control. IEEE, 2017, 375-380.
[9] Cai C, Wu K, Liu Q, et al. Panoramic multi-object real-time detection based on improved YOLO algorithm. Computer Engineering and Design, 2018, 39(10): 3259-3264+3271.
[10] Jia P, Tie Y, Qi L, et al. PV-YOLO: An Object Detection Model for Panoramic Video based on YOLOv4. 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML). IEEE, 2022: 56-61.
[11] Jia P F. Research and Application of Object Detection Algorithm Based on Panoramic Stereo Video. Zhengzhou University, 2022.
[12] Tie Y, Jia P, Lu Y, et al. Object Detection Algorithm Based on Panoramic Stereo Video. SSRN Electronic Journal, 2023.
[13] Pokuciński S, Mrozek D. YOLO-based Object Detection in Panoramic Images of Smart Buildings. 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2023: 1-9.
[14] Ge Z, Liu S, Wang F, et al. YOLOX: Exceeding YOLO Series in 2021. arXiv e-prints, 2021.
[15] Dai J, Qi H, Xiong Y, et al. Deformable Convolutional Networks. IEEE, 2017 : 764-773.
[16] Zhu X, Hu H, Lin S, et al. Deformable ConvNets V2: More Deformable, Better Results. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 9308-9316.
[17] Yu D, Ji S. Grid Based Spherical CNN for Object Detection from Panoramic Images. Sensors, 2019, 19(11): 2622.