Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230864

YOLO-UB Based Detection and Identification of Illegal Structures in the Ancient City

Kang CI, Gang LI[1], Jiaying SUN, Bao WANG, Ling ZHANG
College of Software, Taiyuan University of Technology, Shanxi, China 030600

Abstract: To address the problem that current target detection algorithms do not work well in the task of detecting illegal building targets in the ancient city, a YOLO-UB-based algorithm for illegal building detection and recognition in the ancient city is proposed based on remote sensing image data taken by UAV. The algorithm incorporates the Coordinate attention (CA) attention mechanism to improve network robustness enhance the model's detection ability for illegal building targets, and make target localization more accurate, and introduces the Swin Transformer V2 structure to use its own self-attention mechanism to deeply mine the target features, which enhances the global information capture capability and enables the network to better integrate multi-scale features. The algorithm is trained iteratively on a custom dataset and compared with other models. The results show that this algorithm achieves an average accuracy (mAP) of 96.8% in detecting illegal building targets, Compared to the algorithm with YOLOv7, the accuracy is improved by 3.1%, and the algorithm has better featureextraction, robustness and generalization than other target detection models.

Keywords: target detection; unauthorized building; YOLOv7; Swin transformerV2; Coordinate attention

1. Introduction

Historic and cultural heritage is not only a witness to the historical development of human beings, but also a material and spiritual wealth acquired by human beings in the practice of transforming the world, which is a non-renewable and irreplaceable valuable resource. Therefore, the protection of historic and cultural heritage is an important issue faced by the government departments of various countries, and has worthy to be researched. Unfortunately, with the accelerated economic development and rapid urbanization last several years, due to the lack of strengthening the protection of cultural heritage, which has caused unprecedented damage to the authenticity and integrity of many precious historical and cultural heritages.

The traditional way of protecting historical and cultural heritage is mainly based on manual daily active detection and inspection, with strong human judgment as the main factor, and cannot establish a scientific and effective monitoring and management mechanism, for which countries have paid a painful price in the protection of world heritage. With the improvement of computer computing power and breakthroughs in

[1] Corresponding Author: Gang LI, College of Software, Taiyuan University of Technology; e-mail: tx2090@126.com

neural network technology, object detection of illegal buildings in ancient cities has been widely applied in the field of deep learning. Zhang Tong [1] from Wuhan University according to remote sensing images proposed automatic building detection algorithm, applying deep learning to the field of unauthorized building identification; Dong Renwei[2] et al. proposed a Faster RCNN-based aerial camera images of UAVs and DCGAN for small-sample enhanced unauthorized building target detection, providing a theoretical and application basis for fast and accurate urban unauthorized building detection; Zheheng Liang, Peng Deng[3] team implement a neural network-based approach unauthorized building detection algorithm for UAV images, with an accuracy and recall rate of 71% and 88% respectively. Although deep learning technology has achieved good results in detecting illegal building targets, there is still much room for improvement in both the accuracy and speed of its detection.

In this paper, we propose a YOLO-UB-based algorithm for detecting and identifying illegal buildings in the ancient city based on the current popular one-stage target detection network algorithm YOLOv7[4], firstly, the illegal building dataset is labeled, then the improved model is used for training, and finally the algorithm is compared with Faster RCNN[5], YOLOv5l[6] and YOLOv7 for the results indicate that the YOLO-UB algorithm achieved better results in the detection performance of illegal building targets in the ancient city, and provides an effective solution for the illegal building detection task, which has greater practical application value and potential.

2. Yolov7 Network Structure

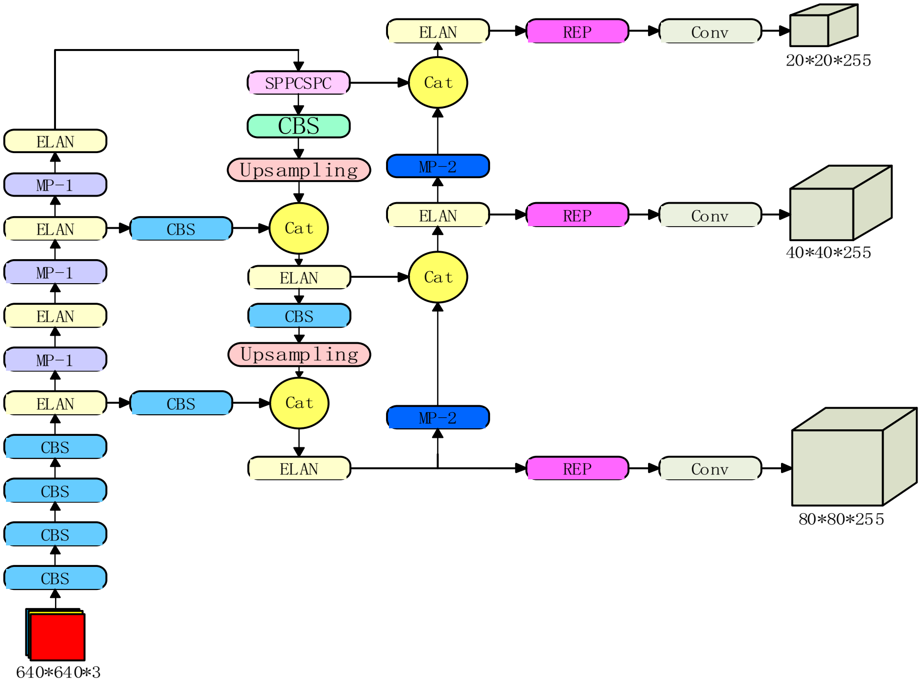


Figure 1. YOLOv7 network structure

YOLOv7, as an excellent target detection and recognition network, mainly consists Input, Backbone, Neck and Prediction. First, Input preprocesses the input image with data enhancement and data compression; then, the processed image is sent to the backbone network for feature extraction; then, the fusion of features is performed in the Neck module and three different sizes of features are obtained; finally, the detection head further processes the fused features and outputs the detection results, the network structure is shown in Figure 1. the YOLOv7's main research components are:

(1) Parameterization in the model: applying the idea of reparameterization to the network, so that the model can use the structure of multiple branches to extract the image feature information during training, and merge multiple branches into one branch during testing to reduce the computation and optimizing model parameters and improving network detection performance;

(2) Label assignment strategy: incorporating YOLOv5's cross-grid search and YOLOX's matching strategy. Accelerate the convergence speed of the model and select the highest quality pre-selection boxes through matching strategies, reducing the costs incurred during the matching process and improving the speed of the network;

(3) ELAN (Efficient Layer Aggregation Network, ELAN), as a new scalable network structure proposed by YOLOv7, ELAN uses two shortest and longest convolutional branches to strengthen the feature extraction ability of the model and make the model more robust;

(4) Training method with auxiliary head: By increasing the training cost, Increase the learning ability of the network on images to improve the accuracy of model detection.

3. Network Structure Improvement

Although the YOLOv7 algorithm performs well in natural scene target detection tasks in terms of accuracy and results, there are still serious misdetection and omission when facing the remote sensing image target detection task with the characteristics of light, cloud noise pollution and complex background. Therefore, in this paper, based on the original YOLOv7, CA attention mechanism and Swin transformer V2 algorithm structure, enhance the detection accuracy of the model and improve detection effect. The improved network structure is shown in Figure 2.

The MPCA structure in the figure consists of MP-1 module plus CA attention mechanism, and STV2 is Swin transformer V2 structure.

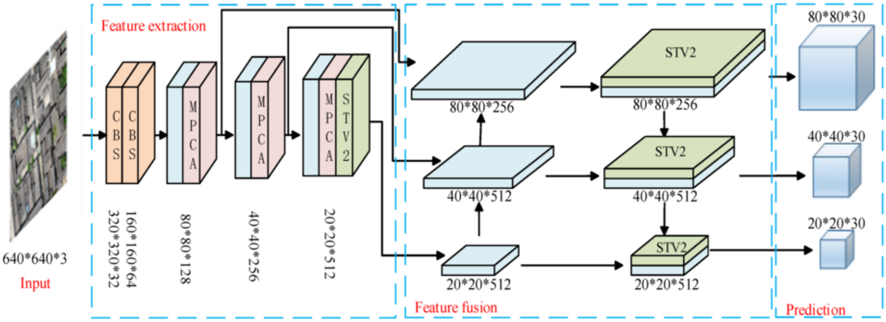


Figure 2. YOLO-UB network structure

3.1 Add CA Attention Mechanism Module

The CA attention mechanism was published and proposed by Qibin Hou[7],the structure is shown in Figure 3. Is a lightweight network structure that aims to enhance network feature extraction capabilities while reducing the computation of parameters, as a way to balance the accuracy and speed of the network in detecting targets. The CA attention mechanism can be divided into two main steps: channel relationship computation and feature fusion. Channel relationship calculation: For the input feature map, it is first partitioned into multiple channels, then a self-attentive mechanism is applied to each channel and the weight coefficients of each position in each channel are calculated, and finally the calculated result indicates the degree of contribution of the position to the other positions in the channel. Feature fusion: The features of each channel are weighted and averaged according to their corresponding weight coefficients to obtain the optimized feature information.

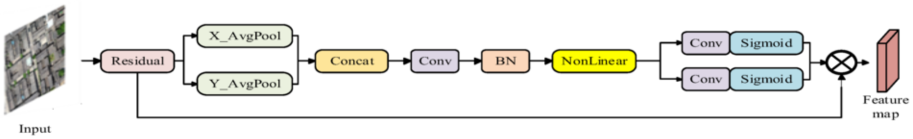


Figure 3. Coordinate attention Module

Compared with single-channel SE[8] attention mechanism and two-channel CBAM[9] attention mechanism, the CA attention mechanism performs feature extraction both in communication information and directional location information, which improves the shortcomings of the above two attention mechanisms and enables the model to pay more attention to important channel and location information, thus improving the expressiveness and generalization ability of the model.

3.2 Introduction of Swin Transformer V2

To enhance the detection performance of the network model for unauthorized building targets, introducing the Swin TransformerV2[10] multi-headed attention structure, the network structure is shown in Figure 4. To enhance the network's ability to capture image global information and rich contextual information, while using the self-attention mechanism to exploit the feature potential, so as to improve the accuracy of the network.

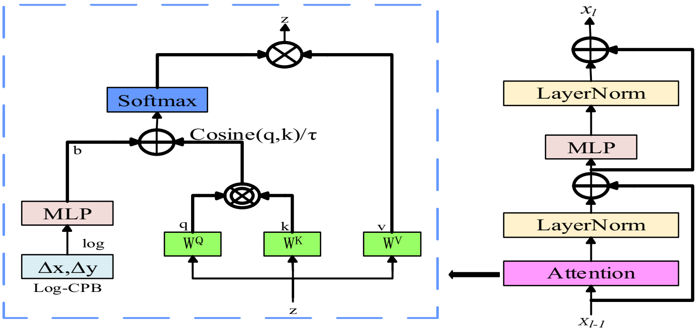


Figure 4. Swin TransformerV2 Structure

Swin-TransformerV2 addresses three main problems caused by the large model size in vision training:

(1) Model scale instability problem. Post Normalization technique and Scaled cosine attention are used to improve the stability of large visual models. The Post Normalization technique changes the traditional Attention + Layer Norm combination into Layer Norm + Attention, and normalizes the output of each residual module before merging it with the main branch, which solves the problem that the model is difficult to converge because the model becomes larger and the output layer is scaled up layer by layer. The Scaled cosine attention formula is shown in Eq. (1) which solves the problem of model instability caused by extreme values during model training;

$$Sim(q_i, k_i) = \cos(q_i, k_i) / \tau + B_{ij} \tag{1}$$

where q denotes the query vector, k denotes the vector of correlations between the queried information and other information, τ denotes the scaling factor, and B_{ij} denotes the relative position deviation between pixel pairs i and j;

(2) The problem of needing to extrapolate most of the relative coordinate range when the training model migrates from low to high resolution. The logarithmically spaced continuous position bias (Log-CPB) method is used, which allows the attention window size to be variable. When the small window migrates to a large window, the window size can be changed naturally to fit the current resolution, making the required extrapolation ratio smaller and the perceptual field can acquire features more accurately;

(3) When inputting large resolution images and high volume model training, GPU video memory occupancy is a serious problem. It reduces the GPU video memory occupation by techniques such as zero-optimizer optimizer and checkpoint, which can extract features of small targets to a greater extent while reducing the training speed, so that the detection performance is greatly improved.

4. Experiment and Analysis

4.1 Data Set

This article uses a DJI Genie 4Pro UAV to collect data with 20 million effective pixels and has 2870 images with a resolution size of 7952*5034. Since the resolution of the collected images is too large, directly inputting them into the model for training will easily cause the memory to "explode" and the detection of small targets will not be satisfactory. Therefore, this paper first crops the dataset and obtains a total of 9780 valid images with the size of 1024*1024; then uses the RoLabelImg tool to label the three types of targets in the dataset, namely, advertisement board, color steel tile and glass roof, and converts the labeled dataset to the format required by YOLOv7; finally, the dataset is randomly divided into training sets according to the ratio of 8:2.

4.2 Experimental Configuration

All experiments in this paper were conducted under Windows 10 Professional operating system environment, with GeForce RTX 3090 graphics card selected for GUP, starting learning rate of 0.01, input image resolution of 1024*1024, and training times of 300.

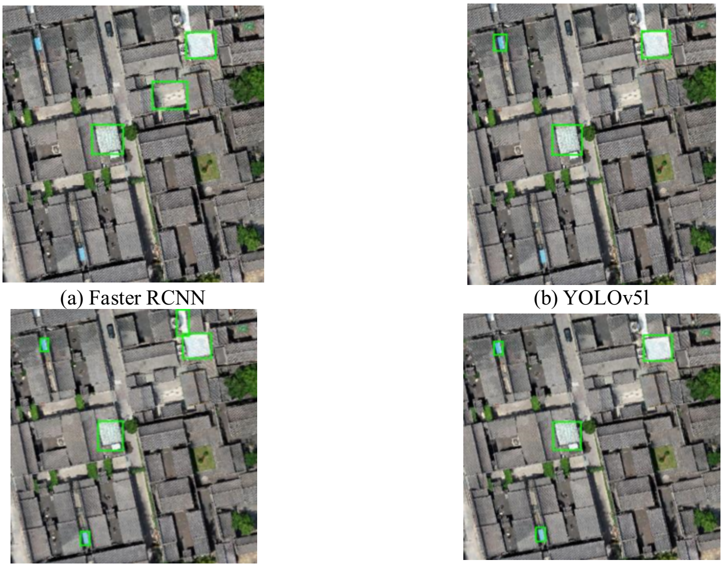
4.3 Experimental Results and Analysis

With the above experimental configuration and experimental environment unchanged, the models in this paper are compared with the two-stage model Faster RCNN and the one-stage models YOLOv5l and YOLOv7 for experiments to derive the effects of different models on the recognition performance of illegal buildings in the ancient city. The experimental results are shown in Table 1.

Table 1. Comparative Experiment

	Models	mAP50	Recall	FPS
A	SSD	60.6%	65.8%	73
B	Faster RCNN	79.4%	77.3%	78
C	YOLOv4	83.6%	85.1%	86
D	YOLOv5l	87.2%	89.6%	92
E	YOLOv7	93.7%	72.4%	107
F	ours	96.8%	97.2%	97

The results in the table show that after 300 rounds of model training, detection accuracy aspect, the mAP of model E in this paper is the highest, reaching 96.8%, compared with models A, B, C, D, and E, which are improved by 36.2%, 17.4%, 13.8%, 9.6%, and 3.1%, respectively, indicating that the model proposed in this paper significantly improves the detection of illegal buildings in the ancient city. Detection speed aspect, because of the fusion of other network structures, the model parameters are increased and the computation volume is increased, resulting in a decrease in frame processing rate compared to the YOLOv7 algorithm 10 , but compared to the SSD, Faster RCNN, YOLOv4 and YOLOv5l models, the frame processing rate is improved by 24, 19, 11 and 5, respectively, explaining that the model of ours also has good performance in detection speed also has good performance and can be applied in real-time detection scenarios.


(a) Faster RCNN

(b) YOLOv5l

(c) YOLOv7

(d)ours

Figure 5. Comparison of detection effects in small target scenes

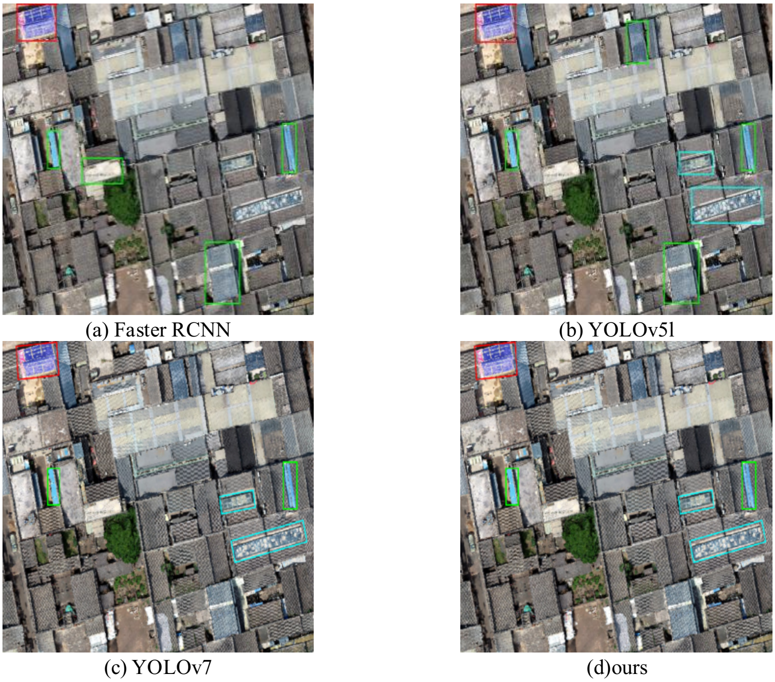

(a) Faster RCNN
(b) YOLOv5l
(c) YOLOv7
(d)ours

Figure 6. Comparison of detection effects in complex background scenes

To better validate the detection performance of the above algorithm, two scenarios are selected to compare the effect of target recognition of illegal buildings in the old city. One is the small target scene, as shown in Figure 5, compared with other algorithms, the improved YOLO-UB algorithm can better avoid the phenomenon of wrong detection and missed detection, and also has a significant improvement for the detection of small targets in images.

Another is the background complex scene, as shown in Figure 6. The YOLO-UB algorithm is the best among the four models in terms of detection accuracy and effect, although there are cases of missed detection. At the same time, the algorithm YOLO-UB has higher localization accuracy when the target object and background color are not clearly distinguished, which again proves the effectiveness of the YOLO-UB model for the identification of illegal buildings in the ancient city.

5. Conclusion

In this paper, we propose a YOLO-UB-based algorithm for detecting and identifying illegal buildings in the ancient city based on the complex detection environment, low detection accuracy, and easy to miss and mis-detect. Firstly, the CA attention mechanism is added to the original YOLOv7 algorithm to improve the feature extraction ability of the network and the detection accuracy of the model; secondly, the Swin transformer V2 structure is introduced to enhance the network's ability to capture global and contextual information, and it improves the stability of the network and solves the problem that the network does not detect well in large resolution scenes. Through the comparison test results of Faster RCNN, YOLOv5l and YOLOv7 algorithms, it is concluded that the algorithm YOLO-UB has good performance in both detection accuracy and detection speed, and through visualization analysis, it is proved that the algorithm YOLO-UB has accurate localization of illegal buildings, low error detection and leakage detection rate, which indicates that the model YOLO-UB has high accuracy, fast speed and The algorithm can also provide ideas and references for national and local government departments to implement historical and cultural heritage protection measures.

References

[1] Zhang Tong,Pan Li. An automatic detection method for buildings with high-resolution remote sensing images[J]. Mapping and Geographic Information,2020,45(2):101-105.

[2] Dong Renwei. Research on urban unauthorized building detection method based on deep learning [D]. Shandong University, 2020. DOI:10.27272/d.cnki.gshdu.2020.000839.

[3] Liang Zheheng, Deng Peng, Jiang Fuquan, Sheng Sen, Wei Rulan, Xie Gansheng. Application of convolutional neural network based UAV image for unauthorized building detection[J]. Survey and Mapping Bulletin,2021,(04):111-115.

[4] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.

[5] Ren Shaoqing, He Kaiming, Grishick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proc of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 21-37.

[6] SONG Q,LI S,BAI Q,et al. Object detection method for grasping robot based on improved YOLOv5[J].Micromachines,2021,12(11):1273.

[7] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate Attention for Efficient Mobile Network Design. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13708-13717.

[8] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[9] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[10] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[J/OL]. arXiv Preprint arXiv, 2103.14030.