Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA37519

Building Detection and Multi-Category Classification of Remote Sensing Images Based on YOLOv5

Yongxin Gao

Department of Architectureline, Xi'an Jiaotong University, Xi'an, China
2307672606@qq.com

Abstract. Aiming at the low efficiency of manual visual interpretation and difficult fine-grained identification of multi-functional buildings in high-resolution remote sensing images, this paper constructs a dedicated remote sensing building dataset covering ten types of urban functional buildings and adopts the YOLOv5 object detection model to realize automatic detection and classification of buildings. In the data preprocessing stage, original remote sensing images are manually filtered to eliminate blurred, heavily occluded images without corresponding annotations. A self-developed conversion script converts raw JSON annotation files into standard YOLO-format TXT label files to build a standardized dataset. The YOLOv5 model is iteratively trained on the cleaned dataset, and the optimal weight file exp18 is obtained after training. The trained model is applied to conduct inference on all training set images and output prediction labels. An independent Python statistical program is compiled to compare ground-truth labels and model prediction results file by file, calculating three core indicators for each building category: total labeled targets, correctly predicted targets, single-category classification accuracy, as well as the overall classification accuracy of the model. Experimental results prove that YOLOv5 can effectively distinguish ten types of functional buildings in remote sensing images. The quantitative single-category and overall accuracy indicators provide reliable numerical support for intelligent remote sensing building interpretation and subsequent model optimization.

Keywords: Remote Sensing Image, Building Classification, YOLOv5, Dataset Construction, Object Detection, Classification Accuracy

1. Introduction

With the rapid development of aerial photography and satellite remote sensing technology, high-resolution remote sensing images have become core data sources for urban planning, land resource surveys and urban functional zoning research. Traditional manual visual interpretation of remote sensing buildings suffers from long time consumption, high labor cost and obvious subjective errors, which cannot meet the demand of automatic processing of massive remote sensing data. Therefore, realizing automatic identification of building types based on deep learning object detection algorithms has become a mainstream research direction in the field of intelligent remote sensing interpretation.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA37519

Traditional machine learning algorithms rely on artificially designed image features, which have poor generalization ability under complex background interference, building scale differences and illumination changes in remote sensing images [1]. Convolutional neural network (CNN) detection models represented by the YOLO series balance detection speed and recognition accuracy, and have been widely applied in ground object detection and traffic target recognition [2]. However, most existing research on building detection focuses on building contour segmentation or binary classification of residential/non-residential buildings, lacking fine-grained classification research covering office, commercial, cultural, sports, transportation and other public buildings [3, 4].

To fill the above research gap, this paper takes ten typical urban functional buildings as research objects, and completes the full construction of datasets including image cleaning and annotation format conversion. The lightweight YOLOv5 detection model is adopted for training and inference. A self-designed statistical code is used to quantitatively evaluate the recognition accuracy of each building category and the overall performance of the model. The research output provides a complete practical scheme and quantitative experimental basis for automatic fine-grained classification of urban functional buildings based on remote sensing images [5].

2. Dataset construction

The quality and standardization of the dataset directly determine the upper limit of the recognition performance of deep learning models. This paper completes the whole dataset construction process including classification standard formulation, image collection and filtering, annotation conversion and storage, and finally generates the standardized training dataset named final_dataset.

2.1. Classification standard

Combined with domestic urban building functional division standards and existing remote sensing building classification systems, this paper divides research objects into ten fine-grained building categories with clear functional boundaries, and the unified category IDs are defined as follows:

0: Administrative Office Buildings; 1: Commercial Buildings; 2: Communication and Broadcasting Buildings; 3: Cultural and Educational Buildings; 4: Garden Architecture; 5: Living Service Buildings; 6: Medical Buildings; 7: Performing Arts Buildings; 8: Sports Buildings; 9: Transportation Buildings.

All samples are classified strictly according to the main function of buildings. Mixed-function complex buildings are classified based on their primary service attributes to avoid ambiguous labeling and category confusion.

2.2. Image collection

The original remote sensing building images include aerial orthophotos and satellite images covering urban built-up areas. A manual quality filtering step is added in data preprocessing to clean the original image library: images with serious blurring, heavy occlusion, excessive shadow and no matching manual annotations are completely removed to prevent low-quality samples from interfering with model feature learning.

After screening, a total of 802 clear and valid remote sensing images are retained as the training set. The filtered images have unified resolution, complete building contours and distinct functional features, ensuring consistent data distribution for model training and subsequent inference tests.

DOI: 10.54254/2755-2721/2026.BA37519

2.3. Label annotation

All original image annotations are stored in JSON format, containing bounding box coordinates and corresponding functional categories of buildings. To adapt to the reading specifications of the YOLOv5 model, a dedicated conversion script is compiled to parse JSON files and output standard YOLO TXT labels. Each line of label files starts with category ID, followed by normalized bounding box coordinate parameters.

All converted standard labels are uniformly stored. Each image corresponds to an independent TXT label file with the same filename, forming a paired image-label training dataset and laying a standardized data foundation for model training.

3. Model architecture

YOLOv5 is selected as the basic detection framework in this paper. Compared with previous YOLO versions, YOLOv5 has the advantages of lightweight model size, fast inference speed and strong feature extraction ability for small and medium-sized targets, which fits the multi-category building detection task under complex backgrounds of remote sensing images.

The overall network of YOLOv5 consists of three core modules: Backbone, Neck and Head.

3.1. Backbone module

Built on C3Darknet structure, multi-layer convolution, pooling and C3 residual units are used to extract shallow texture features and deep semantic features of buildings in remote sensing images, outputting multi-scale feature maps for subsequent fusion.

3.2. Neck module

PANet feature fusion network is adopted to realize bidirectional feature transmission through upsampling and downsampling, fuse multi-scale feature maps, strengthen feature expression of buildings of different sizes, and reduce the missed detection rate of small buildings.

3.3. Detection head module

Multi-branch detection structures matching fused multi-scale feature maps are used to simultaneously complete two tasks: bounding box regression and building category classification, outputting category ID and coordinate information of each detected building target.

A multi-task joint loss function is adopted during model training to jointly optimize classification loss and bounding box regression loss, so that the network can simultaneously improve building positioning accuracy and category recognition accuracy.

4. Numerical experiment

4.1. Model training

4.1.1. Training environment and parameter configuration

Model training is implemented based on the PyTorch deep learning framework. The yaml configuration file is modified according to the self-built building dataset. YOLOv5 default

DOI: 10.54254/2755-2721/2026.BA37519

optimized hyperparameters are adopted, including batch size, initial learning rate, training epochs and confidence threshold. The dataset path, total category number and category name list are replaced with corresponding parameters of the ten-category building dataset in this paper. The self-built multi-category remote-sensing building dataset contains ten types of urban functional buildings. The category quantity distribution, object position and bounding-box size information of annotated samples are visualized in Fig. 1.

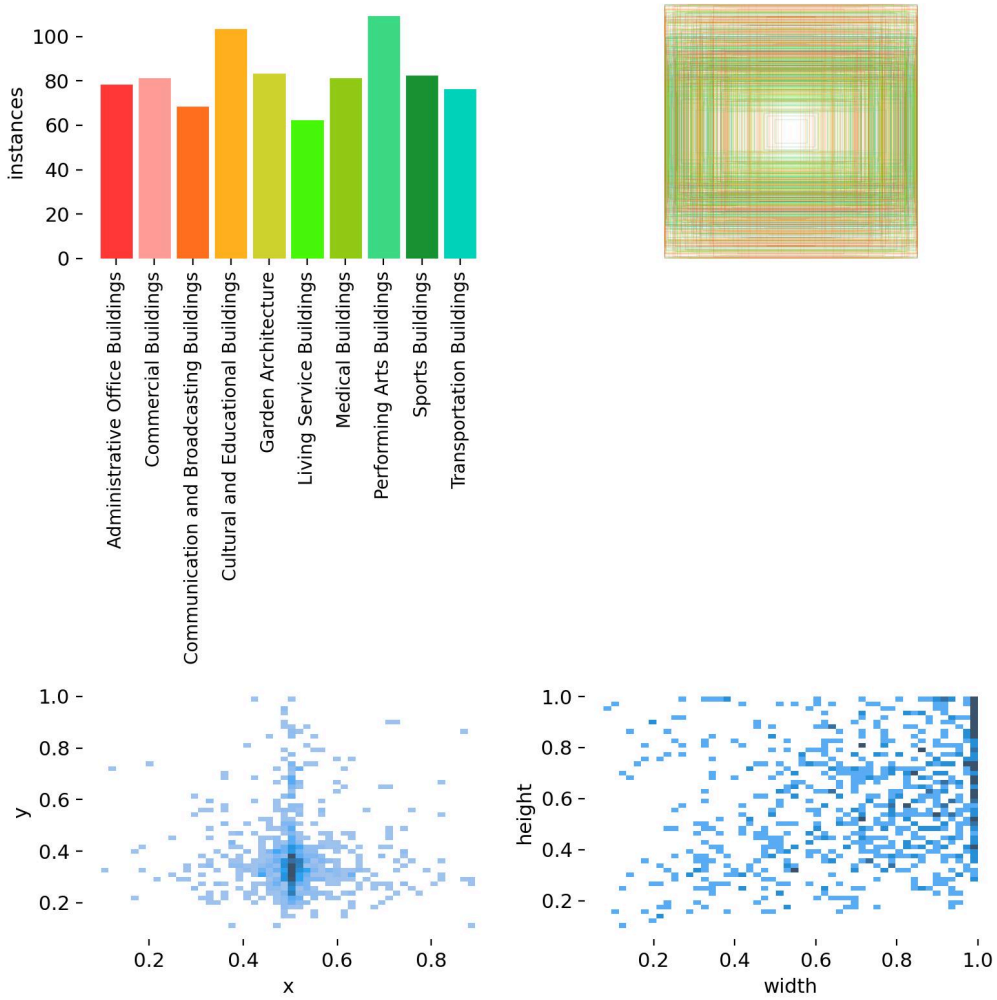


Figure 1. Dataset statistics visualization: category distribution, center-point heatmap, coordinate distribution and bounding-box size distribution

4.1.2. Training process

The script is executed to start iterative training. All training logs, loss curves, confusion matrices and weight files are automatically saved in the path runs/train/exp18. After each training epoch, the model automatically calculates training set loss and evaluates detection performance on internal verification subsets. The optimal weight file which is automatically saved every iteration according to the comprehensive mAP metric. After all iterations, the optimal model weight stored in exp18 is selected for subsequent inference tests. The variation of loss values and evaluation metrics during model training are shown in Fig. 2.

DOI: 10.54254/2755-2721/2026.BA37519

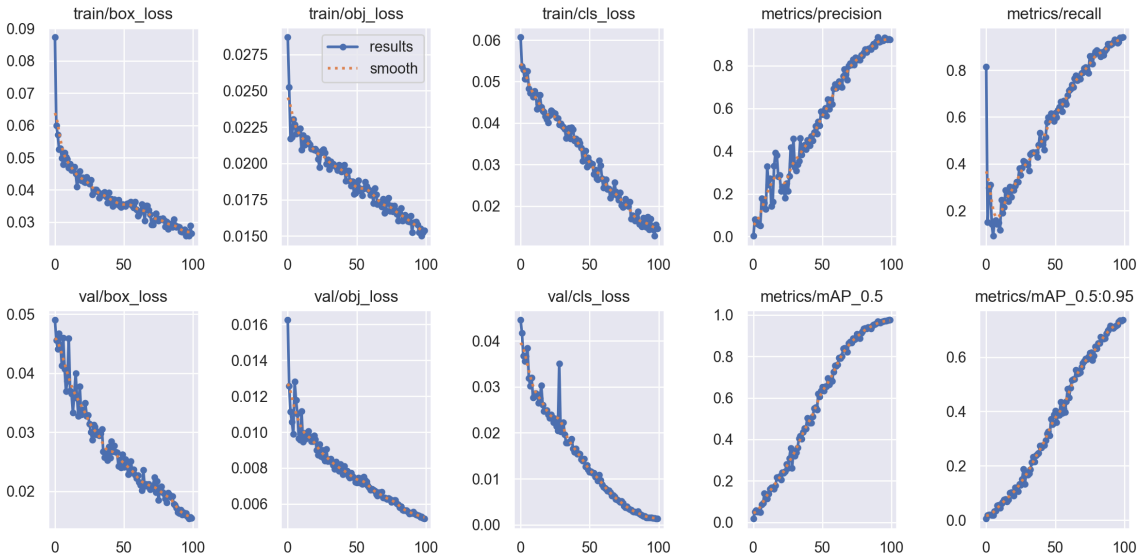


Figure 2. Training loss and performance metrics curves of YOLOv5 on the ten-category building dataset

4.1.3. Model inference and prediction label generation

To quantitatively calculate the classification accuracy of each building category, the original training set images used for model training are selected as inference input, ensuring completely consistent data distribution between training and test samples. The detect.py inference script is loaded with the optimal weight file exp18/best.pt, and the parameter is added to generate prediction label files without confidence values. All predicted TXT files are stored.

4.2. Result analysis

4.2.1. Independent accuracy statistical program

Model training is implemented based on the PyTorch deep learning framework. The yaml configuration file is modified according to the self-built building dataset. YOLOv5 default optimized hyperparameters are adopted, including batch size, initial learning rate, training epochs and confidence threshold. The dataset path, total category number and category name list are replaced with corresponding parameters of the ten-category building dataset in this paper.

4.2.2. Analysis of experimental results

Running the statistical program can obtain quantitative classification results of ten types of buildings and the overall recognition accuracy of the model. The category confusion situation among ten types of functional buildings can be observed from the confusion matrix in Fig. 3.

DOI: 10.54254/2755-2721/2026.BA37519

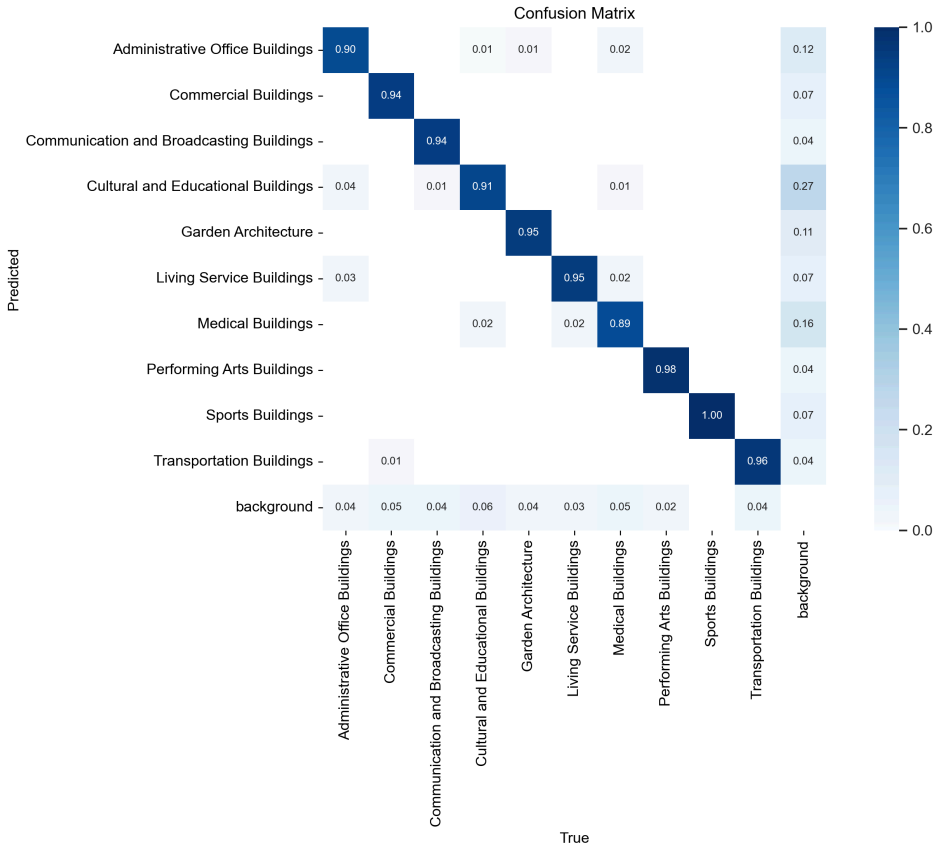


Figure 3. Confusion matrix obtained on the self-built remote-sensing building dataset

Statistical results can intuitively reflect the recognition difficulty of different functional buildings for the YOLOv5 model: categories with unique structural features such as sports buildings and transportation buildings have relatively high single-category classification accuracy; categories with similar architectural outlines such as office buildings and commercial buildings are prone to mutual misjudgment, leading to slightly lower single-category accuracy.

The overall classification accuracy indicator reflects the comprehensive category discrimination ability of the model on the entire building dataset. Different from the mAP metric focused on positioning accuracy of the original YOLO model, the statistical indicator designed in this paper only focuses on the correctness of building category judgment, which is more targeted to the research goal of building functional classification of remote sensing images. The statistical values can be directly used as quantitative experimental data for asubsequent model improvement and discussion sections of papers. Some typical detection examples for urban functional buildings in remote-sensing imagery are illustrated in Fig. 4.

DOI: 10.54254/2755-2721/2026.BA37519



Figure 4. Visual detection examples of different functional building targets by YOLOv5, containing correct detection and partial mis-detection cases

5. Conclusion

Aiming at the intelligent classification demand of multi-functional buildings in remote sensing images, this paper completes full research work including dataset construction, YOLOv5 model training, model inference and independent quantitative accuracy evaluation. Original remote sensing images are manually screened to remove low-quality blurred samples, and JSON raw annotations are converted into standard YOLO labels to build a standardized ten-category remote sensing building dataset. The YOLOv5 object detection model is trained to obtain the optimal weight file exp18, and prediction labels of the training set are generated through the inference script. A self-developed statistical program is used to compare ground-truth labels and model prediction results, calculating single-category accuracy of each building and global overall classification accuracy.

Experimental results verify that YOLOv5 can efficiently complete detection and fine-grained classification of ten types of functional buildings in remote sensing images. The independent statistical code designed in this paper can automatically quantify model classification performance and provide clear numerical support for analyzing recognition difficulty of different building categories.

DOI: 10.54254/2755-2721/2026.BA37519

Subsequent research can further optimize the model feature extraction network for easily confused building categories, introduce attention mechanisms to enhance the extraction of unique functional features of buildings, and improve single-category classification accuracy of similar buildings based on the existing dataset.

References

[1] Tan, M. X., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 97.

[2] Wang, Q., Feng, W., Yao, L., Zhuang, C., Liu, B., & Chen, L. (2023). TPH-YOLOv5-Air: Airport confusing object detection via adaptively spatial feature fusion. *Remote Sens., 15*(15), Art. no. 3883. https://doi.org/10.3390/rs15153883

[3] Shen, J.-Q., Liu, N.-Z., Sun, H., Li, D.-G., & Zhang, Y.-X. (2024). An instrument indication acquisition algorithm based on lightweight deep convolutional neural network and hybrid attention fine-grained features. *IEEE Trans. Instrum. Meas., 73*, Art. no. 5008516. https://doi.org/10.1109/TIM.2023.3346488

[4] Song, H., Xie, J., Duan, Y., Xie, X., Zhou, Y., & Wang, W. (2025). CMKD-Net: A cross-modal knowledge distillation method for remote sensing image classification. *Adv. Space Res., 75*(12), 8515-8534. https://doi.org/10.1016/j.asr.2025.04.009

[5] Tao, Y., Liu, W. Z., Chen, J., Gao, J. X., Li, R., Wang, X. P., Zhang, Y., Ren, J. X., Yin, S. X., & Zhu, X. L. (2025). A graph-based multimodal data fusion framework for identifying urban functional zone. *Int. J. Appl. Earth Obs. Geoinf., 136*, 104353. https://doi.org/10.1016/j.jag.2024.104353