Heritage Science

RESEARCH Open Access

Research on intelligent monitoring technology for roof damage of traditional Chinese residential buildings based on improved YOLOv8: taking ancient villages in southern Fujian as an example

Haochen Qiu[1,2], Jiahao Zhang[1,2*], Lingchen Zhuo[1,2], Qi Xiao[1,2], Zhihong Chen[1,2] and Hua Tian[1,2]

Abstract

In the process of preserving historical buildings in southern Fujian, China, it is crucial to provide timely and accurate statistical data to classify the damage of traditional buildings. In this study, a method based on the improved YOLOv8 neural network is proposed to select aerial photographs of six villages in Xiamen and Quanzhou cities in Fujian Province as the dataset, which contains a total of 3124 photographs. Based on the high-resolution orthophotographs obtained from UAV tilt photography, the YOLOv8 model was used to make predictions. The main task in the first stage is to select the buildings with historical value in the area, and the model's mAP (Mean Accuracy Rate) can reach 97.2% in the first stage task. The second stage uses the YOLOv8 model to segment the images selected in the first stage, detecting possible defects on the roofs, including collapses, missing tiles, unsuitable architectural additions, and vegetation encroachment. In the second stage of the segmentation task, the mAP reaches 89.4%, which is a 1.5% improvement in mAP50 (mean accuracy) compared to the original YOLOv8 model, and the number of parameters and GFLOPs are reduced by 22% and 15%, respectively. This method can effectively improve the disease detection efficiency of historical built heritage in southern Fujian under complex terrain and ground conditions.

Keywords YOLOv8, UAV, Historic buildings, Deep learning, Object detection

Introduction

The southern region of China's Fujian Province (Minnan region) is located on the south-east coast of China and includes the cities of Quanzhou, Xiamen, and Zhangzhou. Historically, it was an important centre of foreign trade, and its handicrafts had a significant impact in the Arab and European regions.1000 Over the years, cultural exchanges have created distinctive urban and architectural styles in the region, leaving a rich cultural heritage for modern China. The area is home to many traditional-style residential buildings, which are an important part of China's cultural heritage. Most of the surviving historical architectural heritage in southern Fujian was built during the Ming and Qing dynasties (fourteenth-nineteenth centuries), along with the development of maritime commerce and trade, as merchants, literati, officials, and overseas Chinese took pride in returning to their hometowns to build properties in southern Fujian [1]. At the 44th session of the UNESCO World Heritage Committee

*Correspondence:
Jiahao Zhang
zhangjh@hqu.edu.cn
[1] School of Architecture, Huaqiao University, No. 668 Jimei Avenue, Jimei District, Xiamen 361021, Fujian, China
[2] Urban and Rural Architectural Heritage Protection Technology Key Laboratory of Fujian Province, No. 668 Jimei Avenue, Jimei District, Xiamen 361021, Fujian, China

in 2021, Quanzhou: Quanzhou: World Centre of Maritime Trade in Song and Yuan China is inscribed on the World Heritage List. However, little research has been done on the architectural heritage of southern Fujian, including Quanzhou. Research on the region's architectural heritage is therefore both urgent and relevant.

The buildings in question contain three different functions: domestic, educational, and ceremonial. They exhibit similarities to buildings in other parts of China in that they are usually built around courtyards surrounded by living rooms, clan temples, bedrooms, and other rooms. Regarding roofing patterns, traditional buildings in other parts of ancient China usually used grey or dark blue tiles, while historical buildings in southern Fujian usually used sintered red clay tiles on sloping roofs. The tiles are unglazed and fixed to the roof with white or grey stucco [2–6] (Fig. 1).

However, there are several important issues that have always been faced in the process of preserving the architectural heritage of southern Fujian:

1. For centuries, under the influence of natural disasters mainly caused by typhoons, the preservation state of architectural heritage is worrying.
2. As one of the fastest growing regions in China after the reform and opening up, the rapid and uncontrolled urbanisation process has led to a very chaotic urban environment around the historic building, as shown in Figure 1, the roof of this historic building not only shows a lot of damage, but also shares the load-bearing walls with the surrounding modern buildings [7, 8].
3. A significant proportion of the owners of historic buildings have relocated from their properties, leaving many of these architectural gems in a state of disrepair for decades. The process of data collection in the field is more challenging [9].

Given the complexity of the environmental situation and the urgency of the conservation of historic built heritage, there is a clear need to develop a robust methodology.

In this instance, the utilisation of drones to ascertain the condition of the edifice is a more efficacious methodology than traditional methods. A significant number of researchers have initiated the use of drone-based tools for the purpose of conducting research on cultural heritage. For example, photogrammetric methods permit the rapid mapping of ancient building complexes and the sampling of their components [10]. In instances where the location of buildings is inaccessible, the deployment of markers manually and the utilisation of drones for regular inspections to monitor the development of cracks in buildings over an extended period is an effective methodology [11].

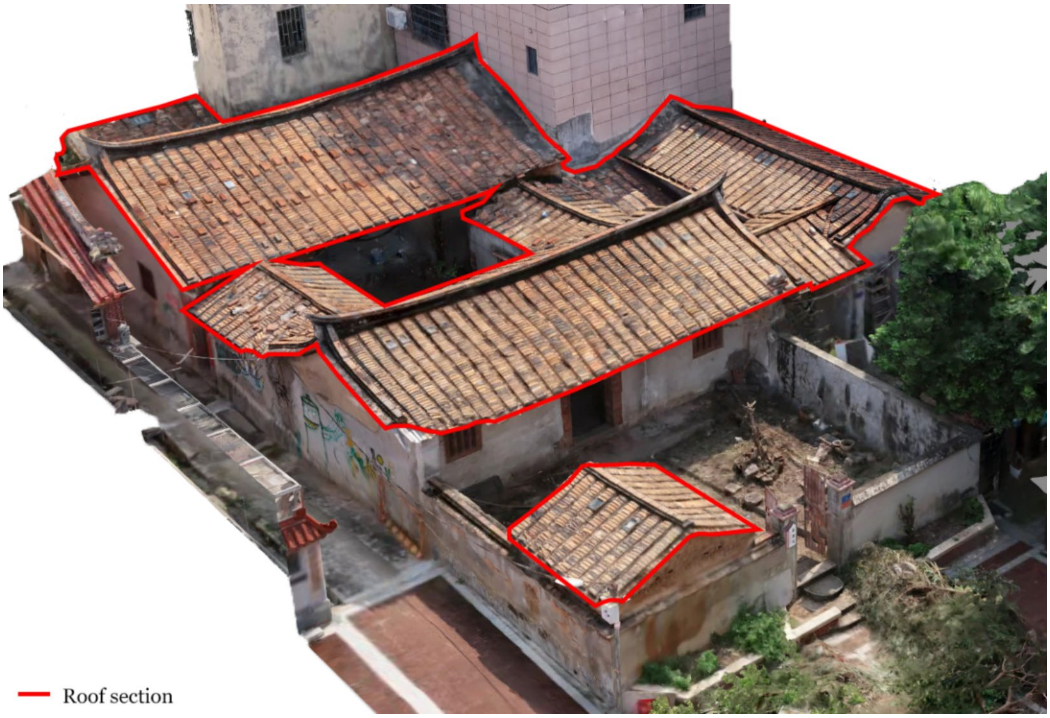


Fig. 1 Typical Architectural Heritage of Southern Fujian Province (from the residence of Su Sui, Xiamen, built in 1727)

Furthermore, the rapid documentation of architectural heritage and the rapid HBIM modelling using drones are also facilitated by this technology [12].

To enhance the efficacy of detecting images captured by UAVs, it is necessary to develop more efficient image processing methods. Based on convolutional neural networks, object detection algorithms can effectively improve the efficiency and accuracy of image analysis. Currently, the mainstream object detection methods mainly include one-stage detectors (such as YOLO, SSD) and two-stage detectors (such as RCNN series). One-stage detectors merge the two processes into one process, achieving classification and regression through a single process, greatly reducing the computational requirements. Two-stage detectors divide the detection process into two stages: feature extraction and proposal, regression, and classification. Finally, the results are outputted (e.g., RCNN series). Although this can provide higher accuracy, it brings high computational requirements, making it inefficient for real-time deployment on resource-constrained edge devices [13, 14].

YOLO, as a mainstream visual detection model and an important representative of one-stage detectors, is widely used in object detection and semantic segmentation tasks. Compared to two-stage detectors, YOLO has faster detection speed and higher accuracy [15, 16]. When deployed on platforms with limited computing power such as drones or other low-computing platforms, YOLO also performs well. When deployed on low-power platforms equipped with Jetson Nano, it also shows good detection performance for underwater natural gas pipeline leaks [17]. By using a microscope mounted on a rail to capture microscopic images of glass and performing instance segmentation on these images, the YOLOACT model demonstrates superior edge delineation performance compared to other models [18]. After typhoons or earthquakes, processing drone remote sensing images using convolutional neural networks can efficiently classify and statistically analyze the damage to buildings in disaster areas based on drone photography images of the roof parts of buildings [19, 20].

Convolutional Neural Networks (CNN) have also played an important role in the protection of architectural heritage. By using transfer learning and adding attention mechanisms, the discerning and contrasting analysis capabilities of architectural heritage images and styles can be effectively improved, enabling rapid retrieval of images of architectural components with similar features [21]. The automatic detection technology for multi-category damages based on CNN models has been applied to style classification of historical buildings, as well as damage detection such as erosion, material loss, stone discoloration, and destruction [22, 23]. It has shown high reliability in identifying facade damage of historical buildings. The CNN model also performs well in classifying Chinese architectural styles [24]. Through CAM visualization technology, the CNN classifier learns what and how to recognize specific architectural styles [24].

The damage automatic detection technology based on the Faster R-CNN model using the ResNet101 framework is used to detect two types of damages (weathering and spalling) in historical masonry structures, effectively reducing the workload of manual inspections [25]. Additionally, it can be deployed online on network cameras for monitoring ancient masonry with smartphones [25]. The improved YOLOv3 algorithm is used for rapid screening of missing ridge beasts in ancient Chinese architecture [26]. Based on the improved YOLOv8 algorithm, a visual protection system for cultural heritage in Suzhou gardens is established [27]. These studies demonstrate the efficiency and stability of convolutional neural network algorithms in the direction of historical heritage preservation.

Convolutional neural networks have been widely used in various fields. In precision agriculture technology, the YOLOv4 algorithm based on improved spatial pyramid pooling (SPP) and path aggregation network (PANet) can detect and prevent various plant diseases at an early stage [28]. The WilDect-YOLO algorithm, which improves YOLOv5, and the DenseSPH-YOLOv5 algorithm, which combines DenseNet and Swin-Transformer, have shown great advantages in wildlife detection and road damage detection [29, 30]. Recent research has also developed more semi-supervised learning-based methods to reduce the hardware cost of supervised learning and improve the learning efficiency [31].

However, in previous research using convolutional neural networks for the preservation of architectural heritage, the focus has often been on analysing and assessing the surface damage of individual buildings or individual components of a complex. For example, identifying individual components of architectural heritage such as ridge beasts, roof tiles, floor tiles or other sculptures. At the same time, disease detection in most studies is still mainly based on image processing from a human perspective. This is due to the limited viewpoints, which make it easy to overlook disease detection in architectural complexes even when image information is processed using convolutional neural network techniques. In a large number of researched contributions to future work, there is a desire to address this problem through the use of automated wayfinding robots or panoramic cameras [25, 32–35].

Research approaches that focus on rapid screening of building damage on a wider and larger scale are more concerned with the overall condition of various types of buildings in an area after an earthquake or other disaster. However, identifying damage to individual building components is challenging due to the limited resolution of data such as satellite imagery. The protection of ancient villages and the daily inspection of ancient building complexes (e.g., city walls or large temples and palaces) require rapid screening of damage to individual building elements.

The present study employs a regional disease study approach and an architectural heritage single-unit disease study approach to develop a methodology for exploring smart monitoring techniques for architectural heritage at the scale of human settlements (e.g., historic districts and traditional villages) and at the scale of historic architectural complexes. The processing capability of convolutional neural networks for high-resolution building complex orthophotos has been enhanced. The analysis of UAV remote sensing images enables the identification of specific damage to monolithic architectural heritage sites in southern Fujian.

The main contributions of this paper are as follows:

1. Construction of a dataset of roof damage to architectural heritage in southern Fujian Province, China.
2. Segmentation prediction of roof orthophotos from ultra-high-resolution UAVs based on the YOLOv8 target detection algorithm, combined with the NMS approach to remove duplicate prediction results.
3. Improve the YOLOv8 backbone using the FasterNet module to simplify the model and improve the detection accuracy, while ensuring its higher detection speed in the UAV dataset, which reduces the number of model parameters by 22% and 15% respectively, resulting in faster detection speed and lower hardware requirements compared to the baseline YOLOv8 algorithm.

Methods

Study area

The main study area of this research is five villages located in different locations in Xiamen and Quanzhou regions of Fujian Province, China. The research object is traditional historical buildings with potential conservation value in the villages, which are called "cuo" in the local dialect. These buildings are characterized by sloping roofs with red tiles. Compared to other regions in China where wooden structures are commonly used, this region has preserved many ancient buildings with brick and stone structures [36, 37] (Fig. 2).

Network structure

YOLOv8 is the latest version of the YOLO series algorithms, which has made multiple modifications based on its predecessor versions. Its structure is illustrated in Fig. 2. This algorithm provides a new SOTA (State-of-The-Art) model, including target detection networks at P5 640 and P6 1280 resolutions, as well as an instance segmentation model based on YOLACT [16].

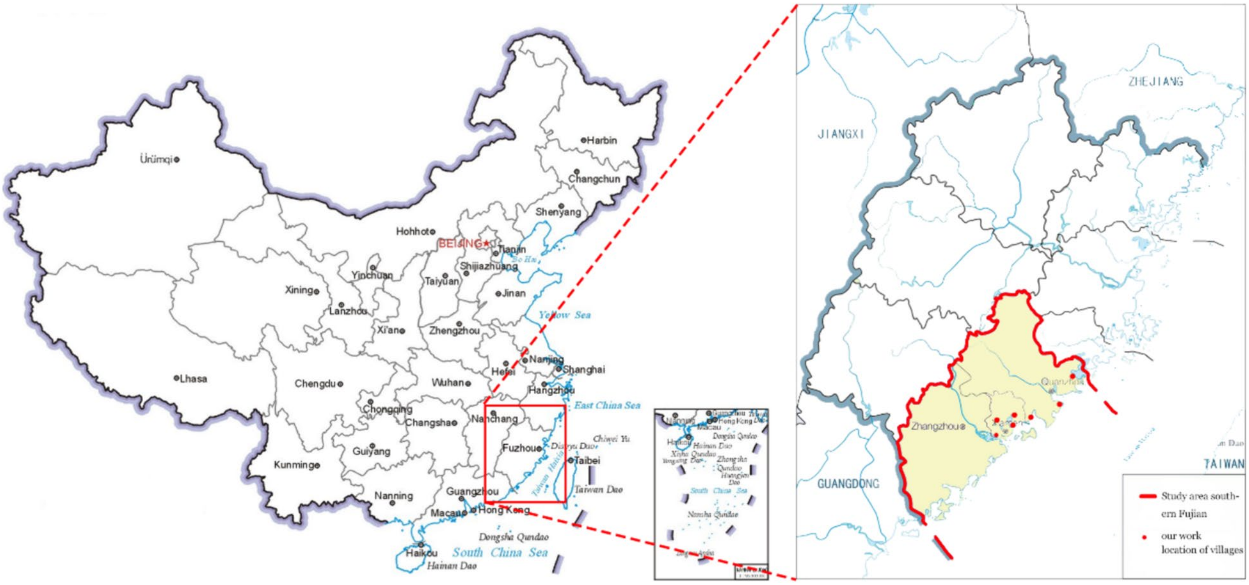


Fig. 2 Study Aera. The main study area of this research is five villages located in different locations in Xiamen and Quanzhou regions of Fujian Province, China. (source of base of map: from Fujian Province Standard Map Open Platform Min(2023) 161 and China National Standard Map Open Platform GS(2019)1673)

In YOLOv8, the backbone part adopts the C2f module, which helps to enhance the model's feature extraction capabilities. The detection head part uses three Decoupled-head structures to distinguish targets of different scales. Each head consists of a classification part and a localization part. The classification part focuses more on the texture information of the target, while the localization part pays more attention to the edge information of the target. This helps the model to locate and identify targets more accurately.

In addition, YOLOv8 has made some adjustments in its training strategy. For example, it disables the Mosaic operation in the last 10 epochs and automatically ends the training process when the model's accuracy no longer improves. These adjustments in training strategies help the model converge faster and more stably, further enhancing its performance.

In our paper, we use YOLOv8 for target detection and instance segmentation tasks. The instance segmentation model is referred to as YOLOv8-seg. Compared to the target detection model, the seg model has only been improved in the YOLO Head part by adding a detection head for instance segmentation to each Decoupled-head structure [38] (Fig. 3).

Orthophoto object detection

Dataset

Southern Fujian in China traditionally includes three cities: Quanzhou, Xiamen, and Zhangzhou. These images were taken in six villages located in different locations in Xiamen and Quanzhou regions of Fujian Province, China. Including data augmentation, there are a total of 3412 images and 67,125 annotated objects.

All the photographs were taken by the DJI Phantom 4 RTK, a small quadrotor drone equipped with a 1-inch image sensor, with an image resolution of up to 2000 w pixels (5472×3648), and an 8.8 mm fixed-focus lens (equivalent focal length of 24 mm). With the help of the DJI Terra software, these aerial images were transformed into orthophoto images that reflect the entire village. A total of 2412 original aerial photos taken at a height of 50 or 100 m were used to construct our benchmark dataset. In this dataset, we divided the sample labels into two categories: "traditional" and "modern", where "traditional" represents traditional buildings that may have conservation value, and "modern" represents modern-style buildings. To facilitate the development of deep learning models, we used the graphical image annotation tool LabelImg [39] to label these aerial photos to generate

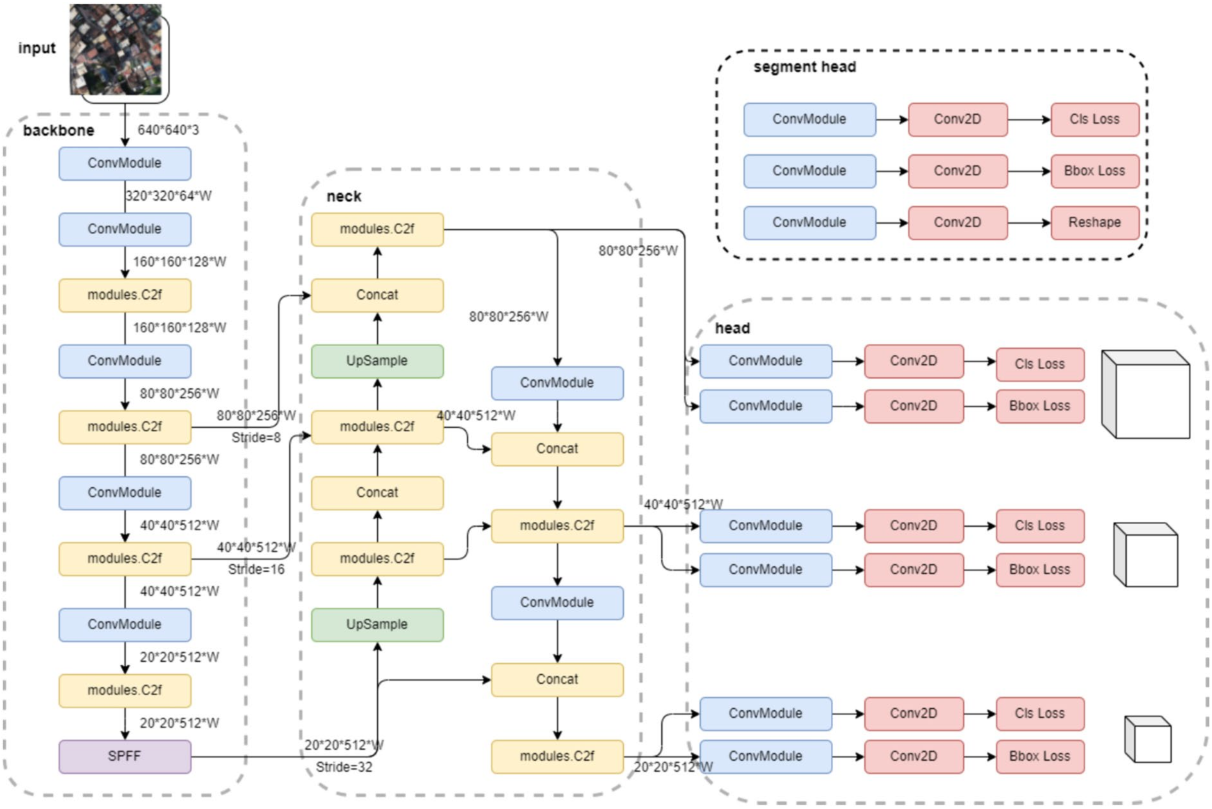


Fig. 3 Detailed structure of YOLOv8

an offline benchmark dataset suitable for deep learning (Fig. 4).

Training

The training was conducted on a high-performance computer with the following specifications: 32 GB RAM, an i5-13400F CPU, and an NVIDIA GeForce RTX 4060Ti 16G GPU. The dataset was divided into two distinct subsets: a training dataset (80%) and a validation dataset (20%). The former was employed to train the model, while the latter was utilised to assess the inference accuracy and speed of the model.

A batch size of 32 was employed, with an initial learning rate of 0.01 and a total of 500 epochs. The number of training observation rounds was specified by YOLOv8, with a value of 90. This implies that if the model's accuracy does not improve following 90 rounds of training, the model will automatically cease training. This method effectively prevents overfitting. Consequently, the actual number of training rounds was 493.

Roof damage detection

In this section, we have used the image prediction results generated by the forward projection object detection task and the YOLOv8-seg model to identify and annotate roof damage on historic buildings of conservation value.

Roof damage dataset

In the preceding step, the images that were predicted to be labelled as "traditional" were selected as the benchmark dataset for the identification of roof damage. A total of 753 roof images were utilised as the dataset, comprising 548 orthographic projection images and 205 oblique photography images. The roof damage was classified according to the situation into the following categories, as shown in Table 1.

The LabelMe [40] image annotation tool was employed to label the dataset, with the Segment Anything tool [41] utilized for assisted labelling to reduce the necessity for manual labour. Initially, the intention was to utilise only orthographic projection images as the basis for the dataset.

In order to enhance the model's generalisation ability and detection accuracy, as well as its capacity to adapt to different lighting conditions, and to address the issue of imbalanced samples, the imgaug library [42] was employed to augment the existing images. A variety of random image augmentation techniques were applied to the photos in the dataset, including blurring and simulating various weather conditions. The final number of images in the dataset was 1500. The quantities for each label are presented in Table 2.

To improve the efficiency of future manual screening, the roof and courtyard areas were labelled in the images (Table 3).

Improvements for YOLOv8-seg to increase speed

To improve the detection speed of YOLOv8 and enable it to be used on platforms with less computing power for outdoor operations, we have made some improvements (Fig. 5). Faster-Net Bottle is derived from the Faster-Net [43] neural network architecture and replaces the Darknet bottleneck in the C2f module with Faster-Net Bottle to improve computational speed without sacrificing accuracy. It reduces redundant computation and memory access through a new technique called PConv (Partial Convolution). PConv's FLOPs are only

$$FLOPs = h \times w \times k^2 \times c_p^2. \tag{1}$$

where cp and c together form the separation ratio: r $= \frac{c_p}{c}$, at r$=\frac{1}{4}$, PConv has only $\frac{1}{16}$ of the FLOPS of traditional convolution, while PConv also has smaller memory access:

$$Memory = h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{2}$$

which is only 1/4 of a regular Conv for r = 1/4. It applies regular conv for spatial feature extraction only to a subset of the input channels, leaving the rest unchanged. For sequential or regular memory access, the first or last sequential channel is considered representative of the entire feature map for computation. Input and output feature maps are assumed to have the same number of channels without loss of generality. This approach allows

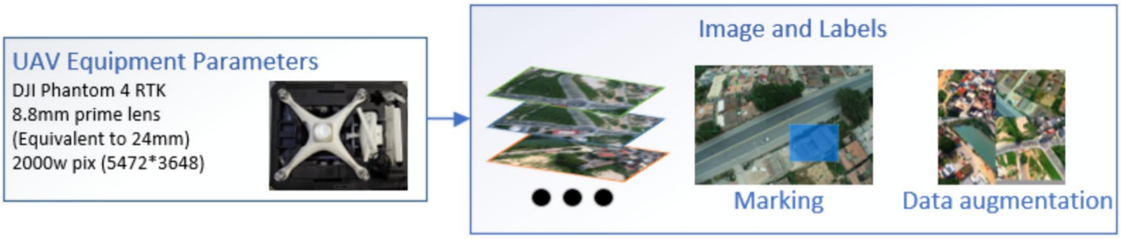
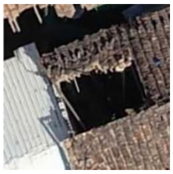
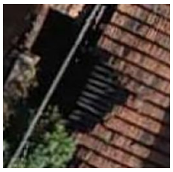
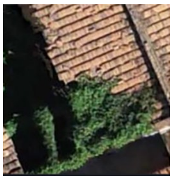
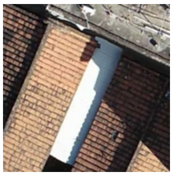


Fig. 4 Data Processing

Table 1 Typical roof damage in Fujian traditional architectural

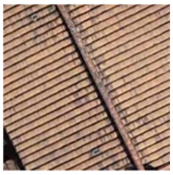
Labels	Introduction	
Collapse	This is the most severe type of roof damage, where a large portion of the building's roof has collapsed, and the load-bearing beams and wooden framework have also collapsed	
Deficiency	Several factors, both physical and chemical, may result in the detachment of the surface layer of roof tiles or porcelain tiles in ancient buildings, exposing the underlying base layer of tiles or the wooden framework of the roof	
Plant	While the presence of vegetation on the surface of a roof may not be immediately detrimental to the roof itself, over time, it can potentially lead to structural damage to the roof of a building. For instance, the roots of plants may erode the tiles, and trees situated in the vicinity of the building may collapse onto the roof during a typhoon	
Addition	Renovating the roof of a building with modern building materials, such as colored corrugated steel plates, plastic sheets, and plastic rainproof cloth, can be effective	

Faster-Net Bottle to run much faster on different devices compared to other networks, while maintaining high accuracy across different visual tasks (Fig. 5).

Table 2 Number of labels

Label	Collapse	Deficiency	Plant	Addition
Number	245	480	1800	500

Table 3 Other labels

Labels	Introduction	
Roofing	In traditional buildings in the southern Fujian region, the relatively intact roof is preserved using the traditional Chinese sloping roof form or a combination of sloping and flat roofs. The roof surface is covered with clay tiles or bricks fired with red soil, resulting in a generally brick-red appearance	
Atrium	The courtyard or inner courtyard of a building is usually rectangular in shape and located in the centre of the roof. In an orthographic projection, it is generally shown as the shadow of each roof. Sometimes, modern materials are added to the courtyard to create a roof	

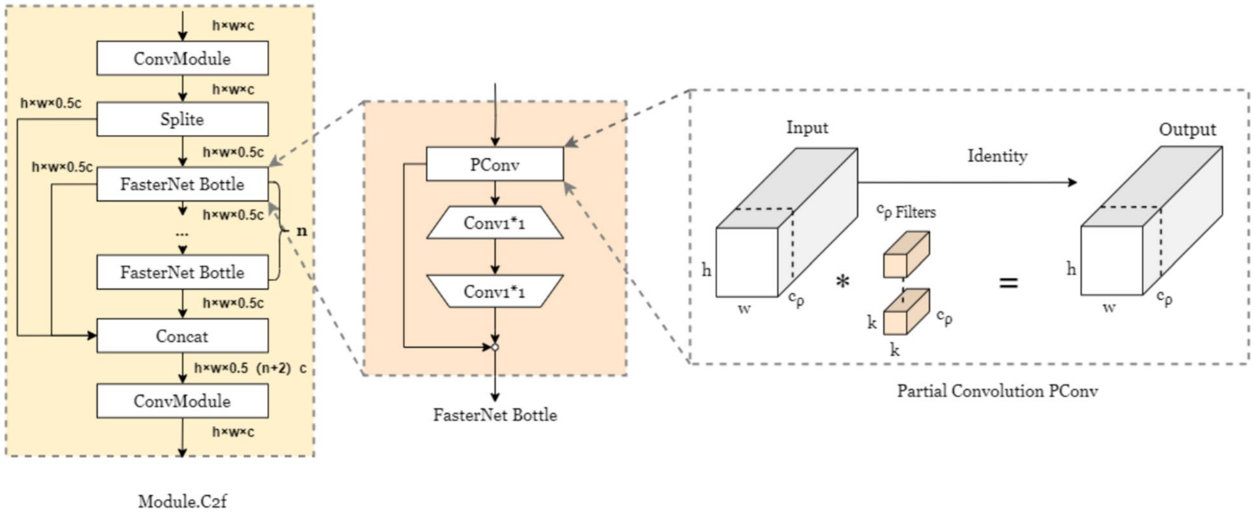


Fig. 5 Detailed structure of Faster-Net Block

Training

The training was also conducted on a high-performance computer with the following specifications: 32 GB RAM, i5-13400F CPU, and NVIDIA GeForce RTX 4060Ti 16G GPU. The dataset was divided into two distinct subsets: a training set (80%) and a validation set (20%). The former was employed to train the model, while the latter was utilised to assess the model's inference accuracy and speed.

A batch size of 16 was employed, with an initial learning rate of 0.01 and a total of 400 epochs. The number of training observation rounds that can be specified by YOLOv8 is 90. Should the model's accuracy not improve following 90 training rounds, the model will automatically cease training.

Result

The result of object detection and contrast experiment

The loss curve during the training process is shown in the figure, and the loss gradually decreases. During training and validation, the loss values converged to 0.26 and 0.55, respectively (Fig. 6).

The number of pixels correctly classified as positive samples is represented by TP; the number of pixels correctly classified as negative samples is represented by FN; the number of error pixels in negative samples is represented by FP; and the number of pixels with errors in positive samples is represented by TN. These values can be calculated using a pixel-based confusion matrix (Fig. 7).

The corresponding precision-recall curve is shown in the figure. As the recall rate increases, the precision value remains stable at around 0.9. (Fig. 8) This also means that the model can make high-quality predictions.

To verify the validity of the models, the YOLOv8 model was compared with the current state-of-the-art models, YOLOv7, YOLOv9, and Faster-CNN, in the comparison experiments. Four metrics, mAP50, mAP50-95, Params (parameters), GFLOPs (Giga Floating Point Operations Per Second), were chosen to compare the accuracy and model size and complexity of different models.

Among them:

(1) mAP50: denotes the average of the mean accuracies across all categories at ioU (intersection and concatenation ratio) = 0.5, which is a key metric for evaluating the overall performance of the object detection algorithm. An increase in the mAP50 reflects an improvement in the performance of the model in all categories.
(2) mAP50-95 indicates that mAP is calculated every 0.05 when ioU is 0.5-0.95 respectively, and the average of the mean accuracy of each category under 10 ioU values is calculated. It can more accurately assess the performance of the model under different ioU thresholds.

(3) GFLOPs (Giga Floating Point Operations Per Second): GFLOPs is a measure of the number of floating-point operations per second in the inference phase of a neural network model. It represents the amount of computation required by the model and is often used to evaluate the computational efficiency

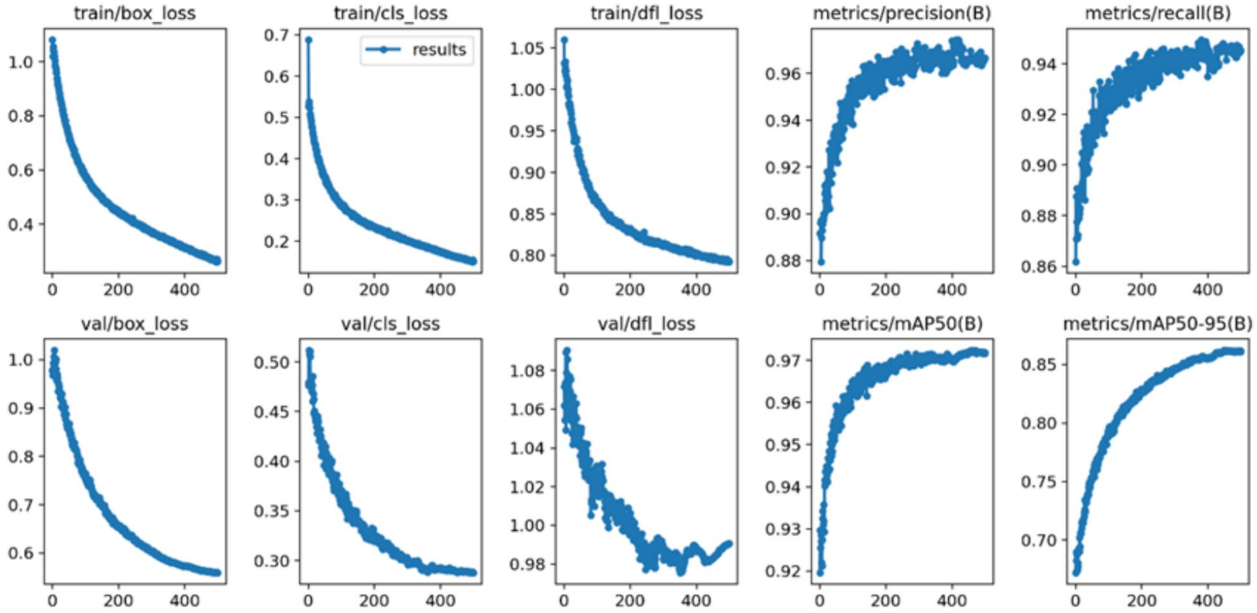


Fig. 6 Performance parameters of model training

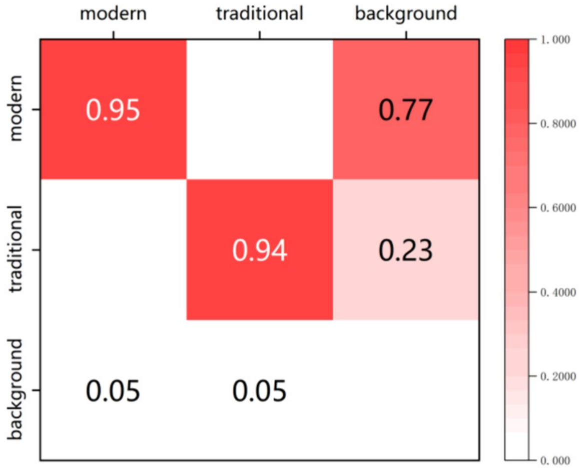


Fig. 7 Confusion matrix

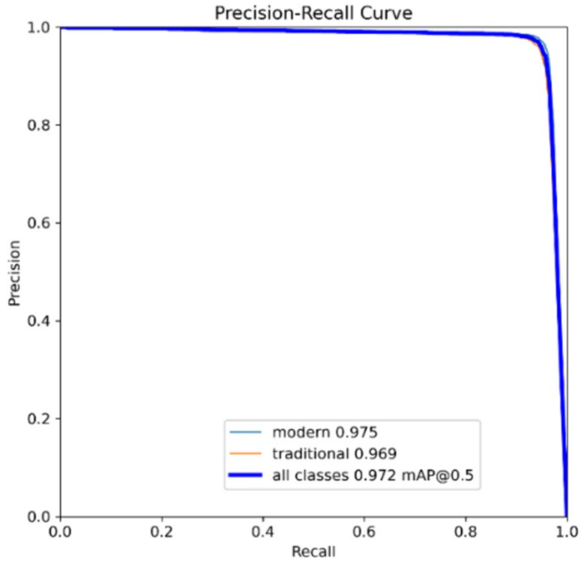


Fig. 8 Precision-recall curve

and speed of the model. Higher GFLOPs mean that the model requires more computational resources to perform the inference task.

(4) Params (parameters): Params are the number of parameters, also known as weights and biases, to be learned and tuned in a neural network model. These parameters determine the complexity and expressiveness of the model and are usually tuned by learning from the training data. A higher number of params means a more complex model, which may have a better fitting ability, but also increases the computational cost and the risk of overfitting.

Overall, YOLOv8 has superior performance in the roof target detection dataset. According to the results of the comparison experiments, on our dataset, compared with the current latest YOLOv9 model, although YOLOv9 has higher accuracy, the model size of YOLOv9 has also reached 238.9, which is almost ten

Table 4 Comparison of experimental results

Backbone	mAP50	mAP50-95	Parameters	GFLOPs
Faster-RCNN	0.855	0.589	41,755,286	251.4
YOLOv6	0.925	0.758	428,342	11.9
YOLOv9	0.981	0.803	51,001,900	238.9
YOLOv8	0.972	0.861	3,011,238	8.2

times that of YOLOv8, which is not a good choice from the point of view of model complexity (Table 4).

On our dataset, YOLOv8 has a huge advantage in terms of accuracy and model complexity over the Faster-RCNN and YOLOv6 models. Compared to YOLOv6, mAP50 improves by 0.05 and mAP50-95 improves by 0.103, while the number of covariates and GFLOPs in the model is reduced by 25%. Compared to the Faster RCNN model (based on the Resnet50 backbone), there is a 90% reduction in the number of GFLOPs and model parameters, and a 0.117 improvement in mAP50.Overall, YOLOv8 has superior performance on the roof target detection dataset.

Predict result

In this study, a DJI Terra was used to synthesise an orthophoto of a village from an aerial triangulation mission and aerial photographs. Orthophotos typically have a resolution of up to 26,000 × 26,000 pixels, which is approximately one billion pixels. As a representative of a first-order detector, YOLOv8 adjusts the resolution of the photographs to 640 × 640 pixels during object prediction and training. Consequently, YOLOv8 is deficient in its ability to detect small targets in high-resolution images and is essentially incapable of performing the target detection inference task on the synthetic orthophoto of the input model. To enhance the detection accuracy, the original high-resolution orthophoto (96 dpi, with a typical resolution of approximately 26,000 × 26,000 pixels) is divided into square images of 2560 × 2560 or 5120 × 5120 pixels. To ensure that objects between the cropped images are fully recognised, there is an overlap region of 512 pixels between each neighbouring image. The pixel position of the upper left corner of the segmented image in the original image is used as the filename of the segmented image in the following format:

patch_i_j.jpg,

where i and j are the height and width pixel numbers of the upper left corner of the segmented image in the original image. The process begins with the prediction of each segmented image, which is then stitched back into the original image. The results of the individual images are then combined and redrawn on the original image.

Table 5 Dataset format of YOLOv8

Classes	X	Y	W	H	Confidence
0	0.5	0.5	0.24	0.48	0.8961

Coordinate conversion

In the context of YOLO for object detection, the output of a static image is a text file with the same file name as the original image. The text file contains all the bounding box information, with each line representing a set of bounding boxes. Each bounding box corresponds to a predicted target object, its category, and confidence score. In particular, the dataset generated by YOLO typically comprises the following elements:

1. Category label: Each bounding box is associated with the category label of a target object, which is typically represented by an integer value.
2. Bounding box coordinates: Each bounding box is represented by four floating-point numbers, which are the x and y coordinates of the centre point of the predicted box, as well as the width and height of the box. The values range from 0 to 1, representing the ratio between the pixel value and the predicted image size.
3. The confidence score is defined as: Additionally, each bounding box is accompanied by a confidence score, which is employed to indicate the probability of the corresponding target object existing within the bounding box. In general, a higher confidence score indicates a greater probability that the corresponding target object exists in the bounding box.

To illustrate, consider an image with a width and height of 416 pixels. If a target object with a category of “person” is present in the image, with a centre point coordinate of (208, 208), a width of 100 pixels, a height of 200 pixels, and a confidence score of 0.961, the content of the corresponding TXT file is as follows: [16] (Table 5):

In order to stitch the predicted results of the segmented images back into the original image, it is necessary to calculate the content of the bounding box coordinates. Assuming the width and height of the original image are W_0 and H_0, and the width and height of the segmented image are both a, the calculation of converting the predicted results coordinates of the segmented image to the coordinates in the original image is as follows:

$$X = j + (X * a) / H_0 \tag{3}$$

$$Y = i + (Y * a) / W_0 \tag{4}$$

$$W = W * a / W_0 \tag{5}$$

$$H = H * a / H_0 \tag{6}$$

Non-maximum suppress (NMS) and other post-processing

In order to ensure the comprehensive capture of target information during prediction, a 512-pixel overlap area is established between adjacent images during the segmentation process. During the process of image merging, the aforementioned overlapping areas will result in the generation of overlapping detection boxes. The non-maximum suppression (NMS) method is employed to eliminate redundant detection boxes. The fundamental principle of the NMS algorithm is that there may be multiple overlapping detection boxes for a single target in the detection results. In such instances, the detection box with the highest confidence score is selected as the final result. In particular, the NMS process typically comprises the following steps:

1. For each target, calculate its detection box and corresponding confidence score in the image.
2. Sort all detection boxes based on their confidence scores and select the detection box with the highest confidence score as the optimal detection box for the current target.
3. For the remaining detection boxes, calculate their IoU (Intersection over Union) values with the current optimal detection box. If the IoU value is greater than a certain threshold (usually 0.5), the detection box is discarded, otherwise it is retained.
4. Repeat steps 2 and 3 until all detection boxes have been processed.

Through the NMS algorithm, the phenomenon of repeated detection and excessive detection can be effectively reduced, thereby improving the accuracy and efficiency of target detection (Fig. 9).

In the context of object detection tasks based on orthophoto images, it is often the case that non-maximum suppression (NMS) is unable to completely remove all duplicate prediction boxes. In the majority of cases,

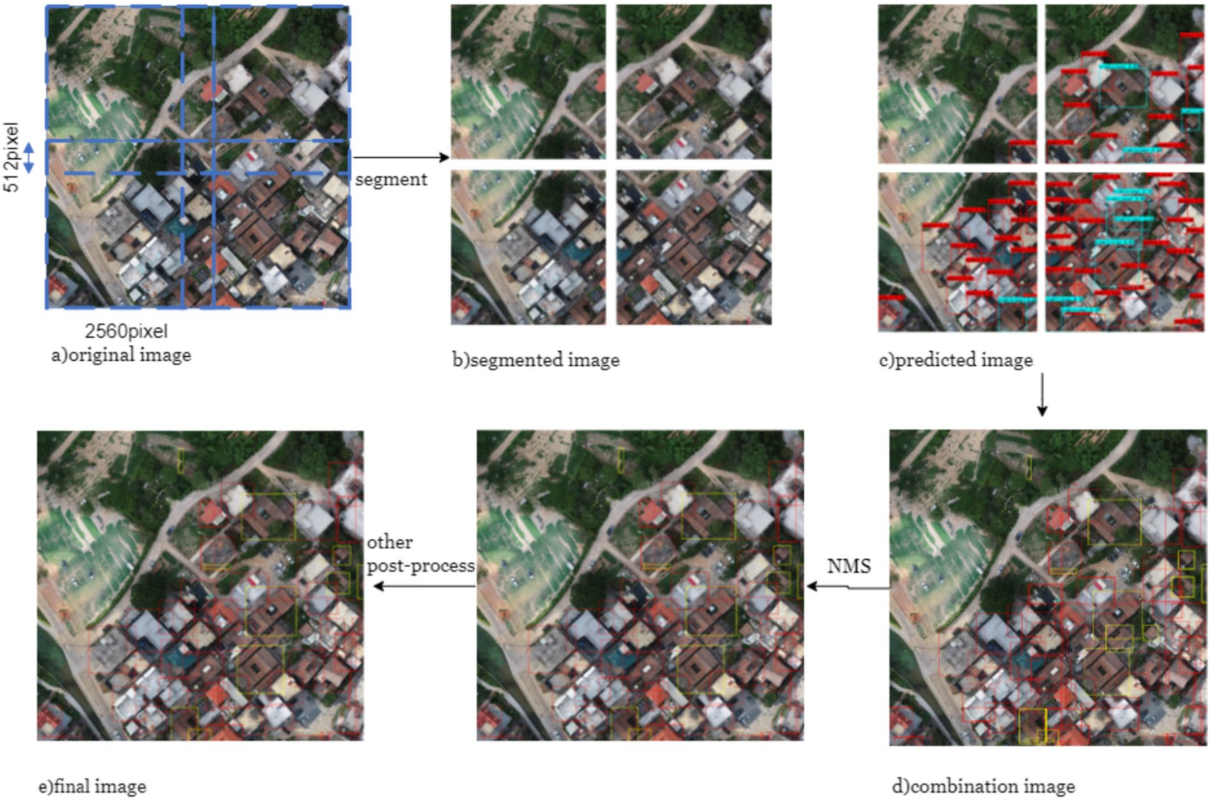


Fig. 9 Result of post-process. (Modern buildings are represented by red boxes and architectural heritage by yellow boxes)

buildings in orthophoto images do not overlap, and there is often a certain distance between two buildings. Consequently, in the majority of cases, there is a minimal overlap between any two prediction boxes generated by YOLO, rendering a method such as NMS an appropriate solution. In such instances, the prediction box with the larger area is retained, provided that the intersection between two objects of the same category is significant (or even contains each other). A series of experiments were conducted to ascertain the impact of varying values between 0.5 and 0.9 on the efficacy of retention. (Figs. 9 and 10).

Following the post-processing procedures outlined in Sect. "Non-Maximum suppress (NMS) and other post-processing.", there were 35 discrepancies between the target identification results and the manual counting and classification. These discrepancies included 5 instances of missed detections, 2 instances of classification errors, and 28 instances of other objects being identified as targets. The accuracy rate was found to be 95.1%.

The result of segmentation and contract experiment

To investigate the effectiveness of the method described in this paper in the roof debris dataset, the model described in this paper is compared with advanced semantic segmentation models that have emerged in recent years with lightweight feature extraction networks

Table 6 Result of contrast test

Model (backbone)	mAP50	mAP50-95	Parameters	GFLOPs
Resnet18 [44]	0.884	0.717	13,327,394	35.2
Swin transformer [45]	0.863	0.620	30,259,516	83.0
Mobilenet_v3 [46]	0.859	0.625	2,434,322	9.3
Efficientnet [49]	0.883	0.680	5,796,350	14.9
Ghostnet [48]	0.860	0.637	2,419,476	10.1
Mobilenet_v2 [47]	0.850	0.624	2,699,712	11.0
YOLOv8_seg(baseline)	0.879	0.659	3,259,243	12.0
Our work(YOLOv8-seg-fasster)	0.894	0.639	2,554,034	10.2

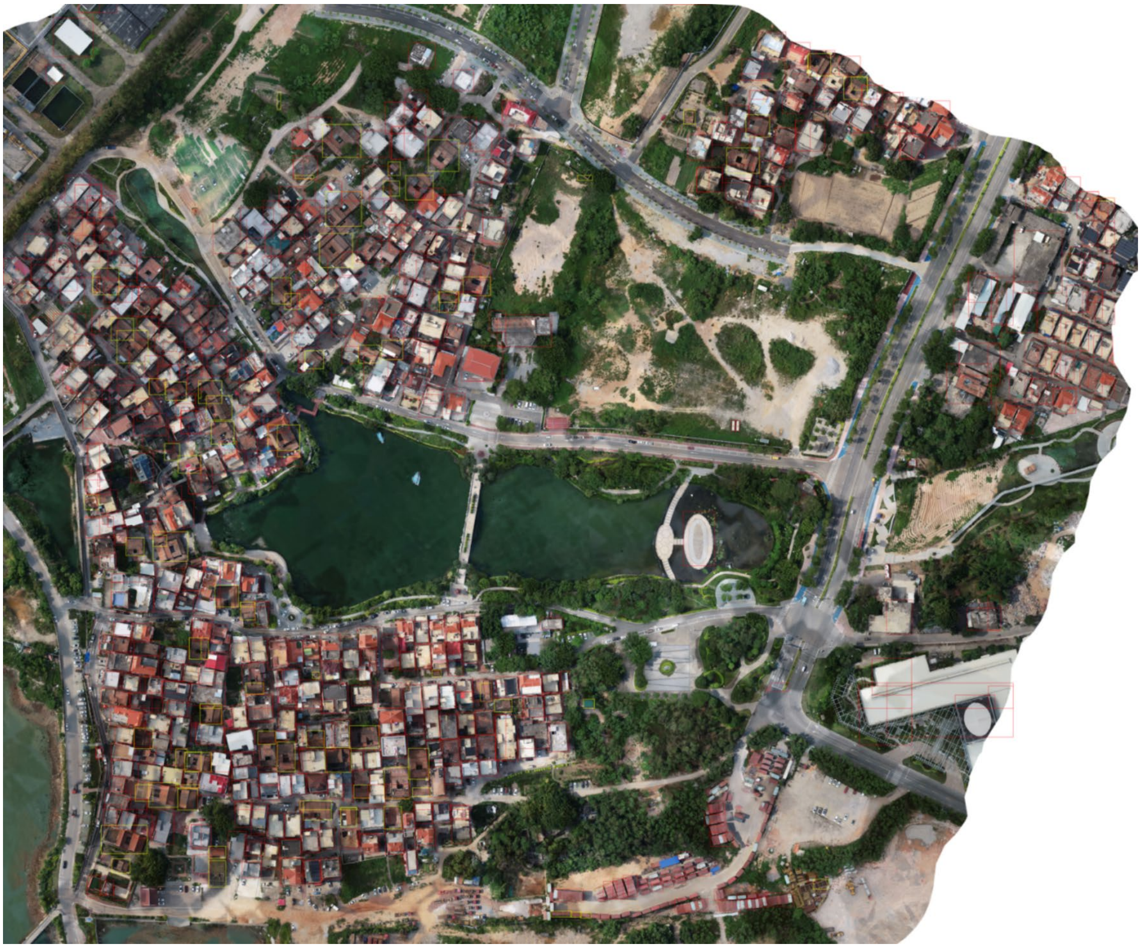


Fig. 10 Detection results in a village area (Modern buildings are represented by red boxes and architectural heritage by yellow boxes)

as the backbone. Including: resnet [44], swin_transformer [45], Mobilenet_v3 [46], Mobilenet_v2 [47], Ghostnet [48] for the roof debris dataset for comparison. And using quantitative analysis to compare the experimental results and model complexity, the experimental results are shown in Table 6.

The confusion matrix indicates that the accuracy of the categories "roofing," "atrium," "collapse," "deficiency," "plant," and "addition" in YOLOv8-seg-faster is 0.97, 0.89, 0.93, 0.78, 0.76, and 0.85, respectively. In comparison to YOLOv8-seg, the accuracies of roofing and atrium remain consistent in YOLOv8-seg-faster. Conversely, the accuracies of collapse and addition have increased by 0.11 and 0.01, respectively, while the accuracies of plant and addition have decreased by 0.03 and 0.02, respectively. It is evident that YOLOv8-seg-faster is capable of maintaining consistent segmentation capabilities for different target categories in comparison to YOLOv8-seg. The training process and segmentation results are shown in Figs. 11, 12 and 13

Conclusion

Research findings and contributions

The full paper presents a method for identifying architectural heritage disease damage based on drone images and convolutional neural networks. The proposed method addresses the challenge of rapid identification and detection of historical architectural heritage diseases within the context of a large-scale historic district or traditional village.

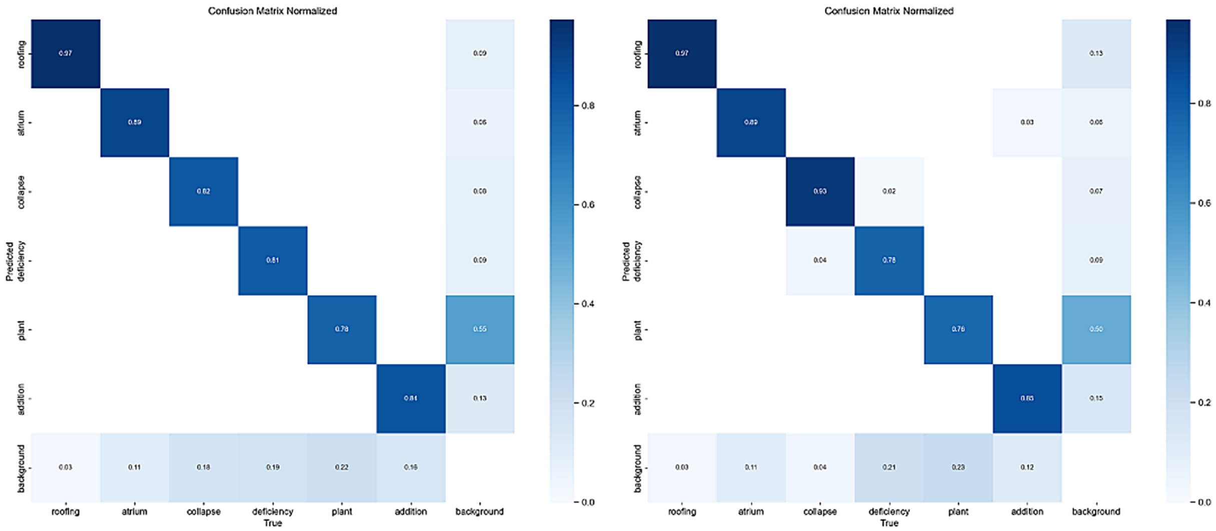
To achieve this goal, two phases of the task were carried out. The first phase involved target detection of the historic built heritage from historic countryside orthophotos. The second phase entailed instance segmentation of the historic buildings for roofing diseases.

In the initial phase of the task, historical landscape orthophotos were detected using the YOLOv8 target detection model with a mAP of 0.971, which is an acceptable level of accuracy. Comparison experiments also showed that the method still outperforms current state-of-the-art models in terms of accuracy and complexity. Subsequently, the recognition results were post-processed, thus effectively improving the target detection capability of ultra-high resolution orthophotos synthesised from UAV work. In the second phase of the task, we employed a lightweight YOLOv8-seg segmentation algorithm comprising the Faster-Net module, with a mAP50 of 0.894. The results of the comparison experiments indicated that the model exhibited a similar level of accuracy to that of the state-of-the-art algorithm. In comparison to the benchmark model, the number of model parameters and GFLOPs were reduced by 21.64% and 15%, respectively, to 2554034. This reduction in complexity effectively maintains high accuracy while reducing the model's complexity. In summary, our method exhibits clear advantages over other state-of-the-art models.

Limitations and future work

The following limitations exist in this study:

(1) The variety of problems that occur with conventional building roofs is very complex. Building roofs may have minor leakage, cracks and many other



a)YOLOv8-seg

b)YOLOv8-seg-faster

Fig. 11 Confusion matrix

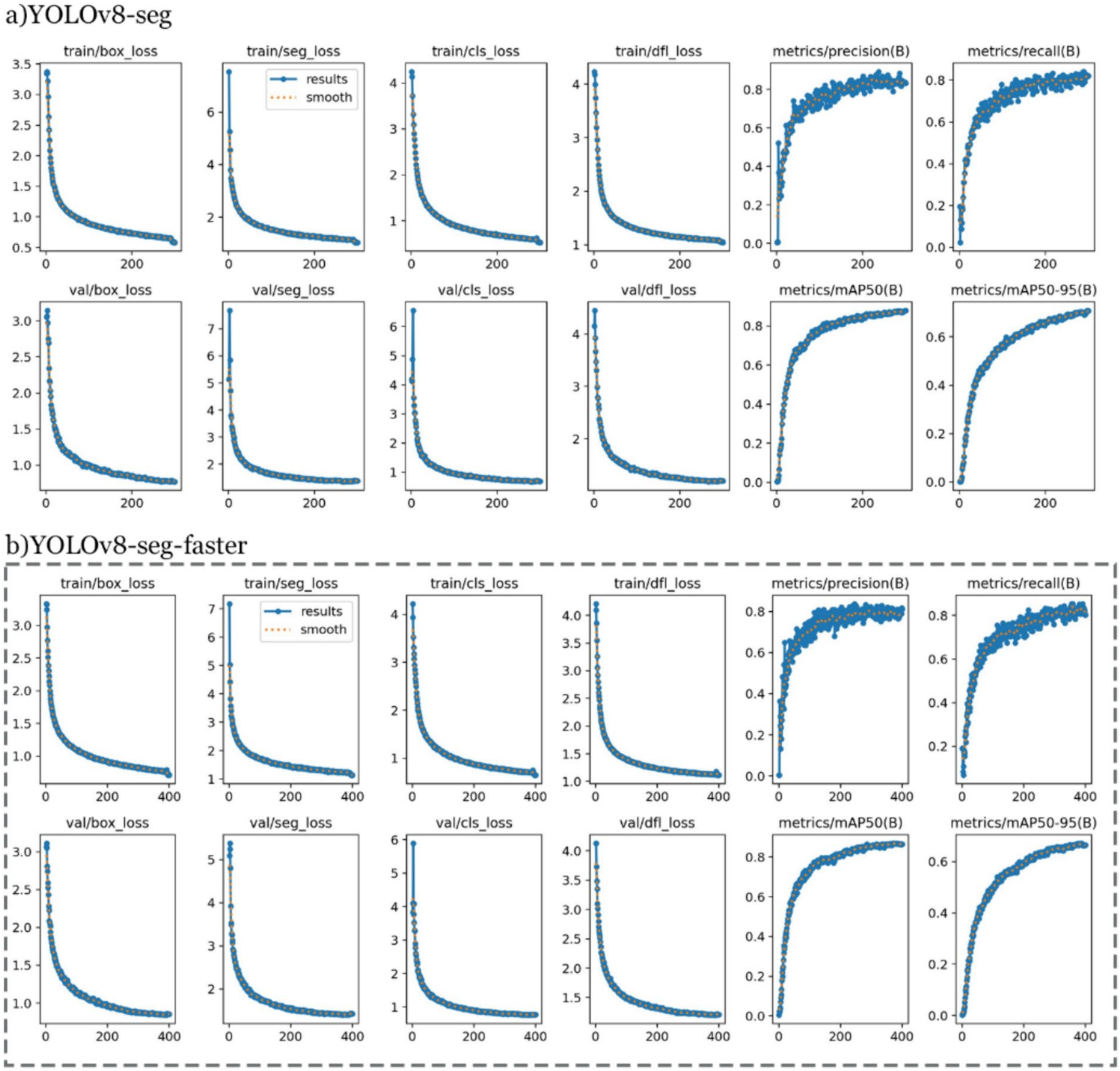


Fig. 12 Performance parameters of model training

problems which cannot be fully covered by the current dataset.

(2) Current research still focuses on the historical architectural heritage of southern Fujian, and there is a lack of data on similar buildings in other parts of China and even in regions such as Southeast Asia. Similarly, although our model shows satisfactory performance for all types of architectural heritage in southern Fujian, the historical value of public architectural heritage such as clan temples and Buddhist temples is still not adequately represented in our study.

(3) Although this study has made slight progress in the non-destructive detection of building diseases, the problem of building diseases is a comprehensive problem that needs to be combined with related disciplines such as climatic conditions, image information, and material properties.

To address the above issues, we will make the following improvements in our future research.

(1) Increase the content of the dataset to improve the robustness of the model and its ability to handle

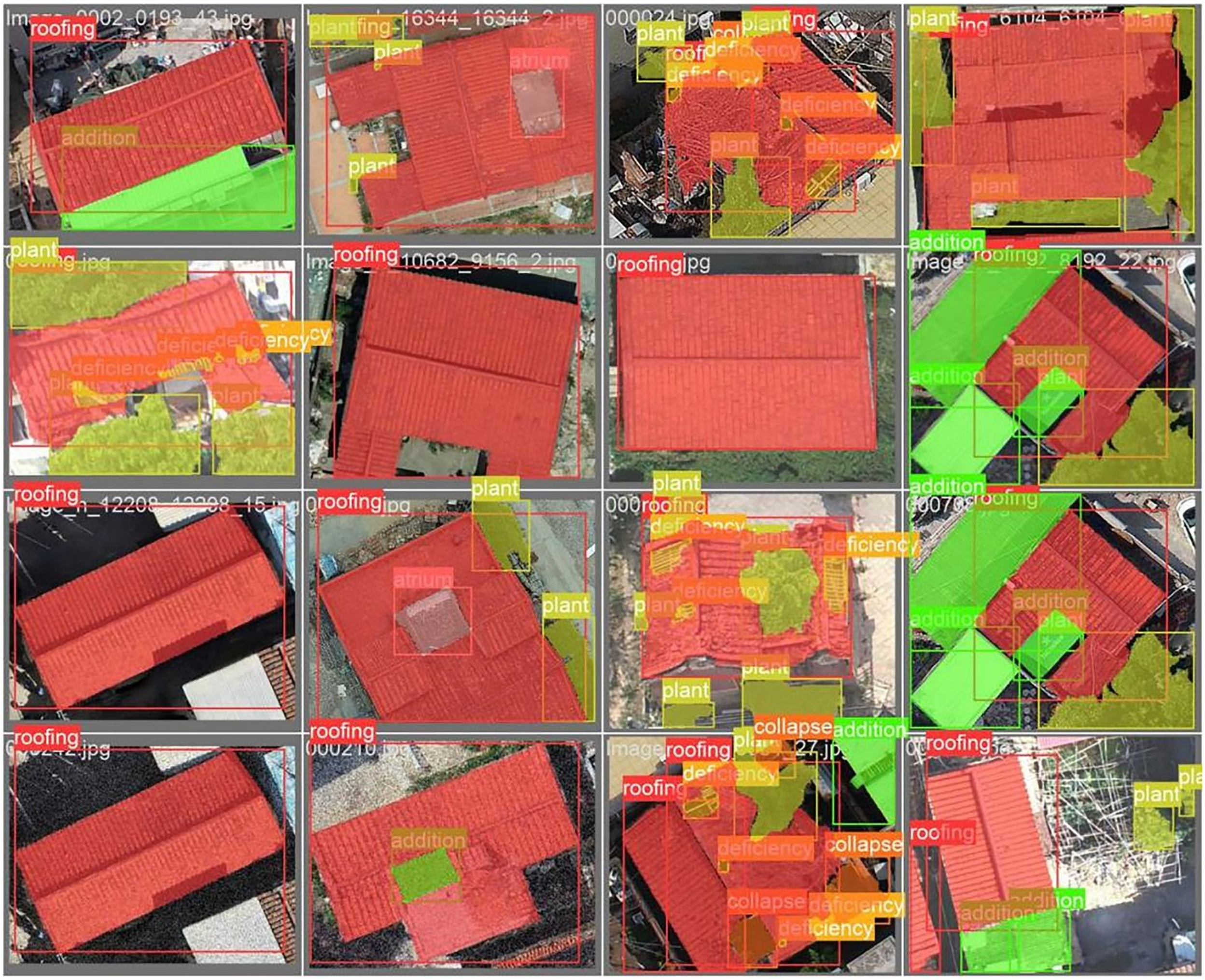


Fig. 13 Roof damage inspection results

complex weather conditions, complex environmental contexts and more damage data. Further data on complex weather conditions, urban built-up environment and more building damage will be added.

(2) The inclusion of image data of built heritage in the environment will extend the research direction and content of the methodology. The image data of architectural heritage of different functional types will be improved, and the historical value of public architectural heritage such as Buddhist temples and ancestral halls will be explored.

(3) The YOLO model will be further improved to incorporate multi-modal information such as LiDAR imaging results, multi-spectral images, etc. for efficient heritage damage detection.

Acknowledgements
Not applicable.

Author contributions
Author 1 (HQ) (the first author): Conceptualization, Methodology, Software, Formal Analysis Investigation, Formal Analysis, Writing-Original Draft. Author 2 (JZ) (Corresponding Author): Funding Acquisition, Project administration, Supervision. Author 3 (LZ): Investigation, Data Curation. Author 4 (QX): Investigation, Data Curation. Author 5 (ZC): Supervision. Author 6 (HT): Investigation and Data Curation. All authors have read and agreed to the published version of the manuscript. All the authors have read and approved the final manuscript.

Funding
The project is supported by the National Natural Science Foundation of China (52008175), the National Natural Science Foundation of China (52078223) and the Fundamental Research Funds for the Central Universities of Huaqiao University (ZQN-919).

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable, the study does not include any human or animal experimentation.

Competing interests

The authors declare that they have no competing interests.

Received: 22 May 2024 Accepted: 23 June 2024
Published online: 04 July 2024

References:

1. Lin L, Qiu H. The changes and driving factors of coastal areas land use in time and space: a case study in Fujian Province Southeast China. Pol J Environ Stud. 2022;31(3):2695–707.
2. Li W. *The Research of the Design of Minana Traditional Roofs(in Chinese)*. Master's Thesis. Xiamen University; 2014.
3. Chen Z. *A Study of Modern Regional Architecture in Overseas Chinese's Hometown in South of Fujian(in Chinese)* . PhD Thesis. Tianjin University; 2005.
4. Wang Q, Witte A. Hokkien Chinese diaspora visitors' image construction of their ancestral hometown: the role of the tourist gaze. J Herit Tour. 2023;18(6):768–84.
5. Ding Y, Cai Y, Liu J. Reconstruction of single-bay buddhist architecture based on stylistic comparisons in Northeast Fujian, the core hinterland of the Changxi river basin-using Gonghoulong temple as an example. Religions. 2024;15(4):474.
6. Wu S, Huang X, Tan WH. The construction of smart city community environment under the application of animation immersion technology integrated into southern Fujian folk culture. Int J Urban Sci. 2024. https://doi.org/10.1080/12265934.2024.2317180.
7. Li Y, Jia L, Wu W, Yan J, Liu Y. Urbanization for rural sustainability—rethinking China's urbanization strategy. J Clean Prod. 2018;178:580–6. https://doi.org/10.1016/j.jclepro.2017.12.273.
8. Rong W, Bahauddin A. A bibliometric review of the development and challenges of vernacular architecture within the urbanisation context. Buildings. 2023;13(8):2043.
9. Chen Y, Wu Y, Sun X, Ali N, Zhou Q. Digital documentation and conservation of architectural heritage information: an application in modern Chinese architecture. Sustainability. 2023;15(9):7276.
10. Liu Z. Extraction and mapping of component information of ancient buildings in Huizhou based on UAV technology. Ecol Inform. 2021;66:101437. https://doi.org/10.1016/j.ecoinf.2021.101437.
11. Germanese D, Leone GR, Moroni D, Pascali MA, Tampucci M. Long-term monitoring of crack patterns in historic structures using UAVs and planar markers: a preliminary study. J Imaging. 2018. https://doi.org/10.3390/jimaging4080099.
12. Lin G, Giordano A, Stendard L, Xu L, Yang X. Using UAV for the digital, documentation, monitoring of culture ar-chitectural heritage: a case study of Sisheng ancestral hall, Shenzhen China. Disegnarecon. 2022;15(29):14–14. https://doi.org/10.20365/disegnarecon.29.2022.13.
13. Jiao LC, Zhang F, Liu F, et al. A survey of deep learning-based object detection. IEEE Access. 2019;7:128837–68. https://doi.org/10.1109/ACCESS.2019.2939201.
14. Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines. 2023;11(7):677. https://doi.org/10.3390/machines11070677.
15. Ultralytics YOLOv8 Docs. 2023; 2023. https://docs.ultralytics.com/modes/
16. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLO. 2023.https://github.com/ultralytics/ultralytics. Accessed 28 Jun 2024.
17. Hu S, Feng A, Shi J, et al. Underwater gas leak detection using an autonomous underwater vehicle (robotic fish). Process Saf Environ Prot. 2022;167:89–96. https://doi.org/10.1016/j.psep.2022.09.002.
18. Pan Z, Yang J, Wang X, Wang F, Azim I, Wang C. Image-based surface scratch detection on architectural glass panels using deep learning approach. Constr Build Mater. 2021. https://doi.org/10.1016/j.conbuildmat.2021.122717.
19. Yeom J, Han Y, Chang A, Jung J. Hurricane Building Damage Assessment using Post-Disaster UAV Data. In: *IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE; 2019:9867–9870. https://doi.org/10.1109/IGARSS.2019.8900477
20. Lei T, Zhang Y, Lu J, et al. The application of UAV remote sensing in mapping of damaged buildings after earthquakes. In: Jiang X, Hwang JN (eds). *Tenth international conference on digital image processing (ICDIP 2018)*. SPIE; 2018:265. https://doi.org/10.1117/12.2503163
21. Gao L, Wu YQ, Yang T, et al. Research on image classification and retrieval using deep learning with attention mechanism on diaspora Chinese architectural heritage in Jiangmen, China. Buildings. 2023. https://doi.org/10.3390/buildings13020275.
22. Samhouri M, Al-Arabiat L, Al-Atrash F. Prediction and measurement of damage to architectural heritages facades using convolutional neural networks. Neural Comput Appl. 2022;34(20):18125–41. https://doi.org/10.1007/s00521-022-07461-5.
23. Dornaika F, Moujahid A, El Merabet Y, Ruichek Y. Building detection detection from orthophotos using a machine learning approach: an empirical study on image segmentation and descriptors. Expert Syst Appl. 2016;58:130–42. https://doi.org/10.1016/j.eswa.2016.03.024.
24. Han Q, Yin C, Deng YY, Liu PL. Towards classification of architectural styles of Chinese traditional settlements using deep learning: a dataset, a new framework, and its interpretability. Remote Sens. 2022. https://doi.org/10.3390/rs14205250.
25. Wang NN, Zhao XF, Zhao P, Zhang Y, Zou Z, Ou JP. Automatic damage detection of historic masonry buildings based on mobile deep learning. Autom Constr. 2019;103:53–66. https://doi.org/10.1016/j.autcon.2019.03.003.
26. Hou M, Hao W, Dong Y, Ji Y. A detection method for the ridge beast based on improved YOLOv3 algorithm. Herit Sci. 2023;11(1):167. https://doi.org/10.1186/s40494-023-00995-4.
27. Gao C, Zhang Q, Tan Z, et al. Applying optimized YOLOv8 for heritage conservation: enhanced object detection in Jiangnan traditional private gardens. Herit Sci. 2024;12(1):1–20. https://doi.org/10.1186/s40494-024-01144-1.
28. Roy AM, Bose R, Bhaduri J. A fast accurate fine-grain object detection model based on YOLOv4 deep neural network. Neural Comput Appl. 2022;34(5):3895–921.
29. Roy AM, Bhaduri J, Kumar T, Raj K. WilDect-YOLO: an efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. Ecol Inform. 2023;75:101919.
30. Roy AM, Bhaduri J. DenseSPH-YOLOv5: an automated damage detection model based on DenseNet and Swin-transformer prediction head-enabled YOLOv5 with attention mechanism. Adv Eng Inform. 2023;56:102007.
31. Jiang B, Chen S, Wang B, Luo B. MGLNN: semi-supervised learning via multiple graph cooperative learning neural networks. Neural Netw. 2022;153:204–14.
32. Valero E, Forster A, Bosché F, Hyslop E, Wilson L, Turmel A. Automated defect detection and classification in ashlar masonry walls using machine learning. Autom Constr. 2019. https://doi.org/10.1016/j.autcon.2019.102846.
33. Zhang J, Zhang Y, Liu J, Lan Y, Zhang T. Human figure detection in Han portrait stone images via enhanced YOLO-v5. Herit Sci. 2024;12(1):119. https://doi.org/10.1186/s40494-024-01232-2.
34. Gao C, Zhang Q, Tan Z, et al. Applying optimized YOLOv8 for heritage conservation: enhanced object detection in Jiangnan traditional private gardens. Herit Sci. 2024;12(1):31. https://doi.org/10.1186/s40494-024-01144-1.
35. Li Y, Zhao M, Mao J, Chen Y, Zheng L, Yan L. Detection and recognition of Chinese porcelain inlay images of traditional Lingnan architectural decoration based on YOLOv4 technology. Herit Sci. 2024;12(1):137. https://doi.org/10.1186/s40494-024-01227-z.

36. Xie X, Xu W, Lian X, Fu YL. Sustainable restoration of ancient architectural patterns in fujian using improved algorithms based on criminisi. Sustainability. 2022;14(21):13800.
37. Huang H, Xie Y, Chen J, Liang S, Chen Z. Bioclimatic design strategy of vernacular architecture in the south-east of China: a case study in Fujian. China Int J Low Carbon Technol. 2024;19:ctad079.
38. Bolya D, Zhou C, Xiao F, Lee YJ. YOLACT plus plus better real-time instance segmentation. IEEE Trans Pattern Anal Mach Intell. 2022;44(2):1108–21.
39. HumanSignal/labelImg. 2024. https://github.com/HumanSignal/labelImg. Accessed 28 June 2024.
40. labelmeai/labelme. 2024. https://github.com/labelmeai/labelme. Accessed 28 June 2024.
41. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. In: *2023 IEEE/CVF international conference on computer vision, ICCV*. IEEE International conference on computer vision. IEEE Computer Soc, 2023:3992–4003.
42. Jung A. aleju/imgaug. Published online March 8, 2024. https://github.com/aleju/imgaug. Accessed March 8, 2024
43. Chen J, Kao SH, He H, et al. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In: *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE conference on computer vision and pattern recognition. IEEE computer Soc; 2023:12021–12031.
44. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2015. https://doi.org/10.48550/arXiv.1512.03385
45. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021. https://doi.org/10.48550/arXiv.2103.14030
46. Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. 2019. https://doi.org/10.48550/arXiv.1905.02244
47. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2019. https://doi.org/10.48550/arXiv.1801.04381
48. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. GhostNet: More Features from Cheap Operations. 2020. https://doi.org/10.48550/arXiv.1911.11907
49. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2020. https://doi.org/10.48550/arXiv.1905.11946

Publisher's Note

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