TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 06/21 pp1002–1022
DOI: 10.26599/TST.2023.9010090
Volume 29, Number 4, August 2024

A New High-Precision and Lightweight Detection Model for Illegal Construction Objects Based on Deep Learning

Wenjin Liu, Lijuan Zhou*, Shudong Zhang, Ning Luo, and Min Xu

Abstract: Illegal construction has caused serious harm around the world. However, current methods are difficult to detect illegal construction activities in time, and the calculation complexity and the parameters of them are large. To solve these challenges, a new and unique detection method is proposed, which detects objects related to illegal buildings in time to discover illegal construction activities. Meanwhile, a new dataset and a high-precision and lightweight detector are proposed. The proposed detector is based on the algorithm You Only Look Once (YOLOv4). The use of DenseNet as the backbone of YDHNet enables better feature transfer and reuse, improves detection accuracy, and reduces computational costs. Meanwhile, depthwise separable convolution is employed to lightweight the neck and head to further reduce computational costs. Furthermore, H-swish is utilized to enhance non-linear feature extraction and improve detection accuracy. Experimental results illustrate that YDHNet realizes a mean average precision of 89.60% on the proposed dataset, which is 3.78% higher than YOLOv4. The computational cost and parameter count of YDHNet are 26.22 GFLOPs and 16.18 MB, respectively. Compared to YOLOv4 and other detectors, YDHNet not only has lower computational costs and higher detection accuracy, but also timely identifies illegal construction objects and automatically detects illegal construction activities.

Key words: illegal buildings; object detection; illegal construction objects; high-precision; lightweight

1 Introduction

Illegal construction, also known as illegal building or illegal housing, is a construction project without a valid construction permit. With the rapid development of urbanization and civil engineering, countries around the world are facing an increasing number of illegal buildings. Illegal construction is a significant problem in China. This issue seriously threatens the economic and social stability of China[1]. In Hong Kong of China, illegal buildings often appear on rooftops where villages and towns live[2]. The prevalence of illegal buildings in Argentina is a serious concern that poses significant risks to urban planning, resource conservation, and public safety[3]. The problem of illegal construction in Bulgaria has led to the destruction of ecosystems and the loss of biodiversity[3, 4]. In addition, If the waste of illegal building activities is not managed appropriately, it can lead to environmental pollution, such as air, soil, and water pollution[5]. Illegal constructions have caused significant challenges for social development and economic growth. And it will also lead to the reduction of tax revenue, and the reduction of economic activities

• Wenjin Liu, Lijuan Zhou, Shudong Zhang, and Ning Luo are with School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China. E-mail: 2201002075@cnu.edu.cn; zhoulijuan@hainanu.edu.cn; zsd@hainanu.edu.cn; luoning@hainanu.edu.cn.
• Min Xu is with School of Information Engineering, Capital Normal University, Beijing 100048, China. E-mail: xumin@cnu.edu.cn.
* To whom correspondence should be addressed.

Manuscript received: 2023-06-21 ; revised: 2023-08-05; accepted: 2023-08-27

due to the lack of legal protection and incentives for investment in these areas[6]. Illegal construction has been a significant issue in Greece, with estimates suggesting that up to 25% of buildings in some urban areas are illegal[7]. It has had a significant impact on the country's economy, real estate market, and social infrastructure. In developed countries, such as Britain, New Zealand, and America, illegal construction greatly impacts their economic and social development[8, 9]. Therefore, the automatic and high-precision detection technology for illegal buildings is particularly important.

To realize automatic for illegal building detection, many methods emerge in endlessly. In the early stage, due to the high speed and acceleration of urban change and the expansion of many cities, geographic databases[10, 11] and remote sensing technology[12, 13] were used to reduce the number of illegal buildings. However, these methods must take advantage of land change[14], require manpower to count cadastral data, and spend a considerable amount of labor and resources, with long detection cycle. As machine learning and image processing technology advance, the detection technology of illegal buildings has achieved a certain degree of automation or semi-automation[15]. Building image segmentation[16], Hough transform[17], K-Means[18], and classification based on machine learning[19] are used to detect changes in building areas. Nevertheless, these approaches necessitate manual extraction of illegal building characteristics, and the detection accuracy is inadequate.

Recently, the advancement of deep learning technologies[20–22] has brought new life into the exploration and innovation of identifying unauthorized construction[23–25]. Compared with machine learning methods based on traditional feature engineering[26], deep learning technologies have shown greater precision in identifying illegal construction[27]. However, these methods may lead to substantial inaccuracies in identifying densely located buildings that are occluded or overlapping. To tackle this issue, Convolutional Neural Network (CNN) based object detection models have been devised[28]. These detectors can be categorized into two-stage detectors (e.g., R-CNN[29], Fast R-CNN[30], and Faster R-CNN[31]) or one-stage object detectors[32–36]. However, two-stage object detectors generally suffer from real-time issues.

In recent times, an algorithm known as "You Only Look Once" (YOLO) has appeared[37–40], combining target localization and classification into a regression task. In contrast to two-stage detectors, YOLO does not employ Region Proposal Networks (RPN) and can perform regression directly to detect objects in images, leading to a substantial enhancement in detection speed. This YOLO algorithm has been used in a variety of challenging detection tasks[41–43]. Although these advanced one-stage and two-stage detectors have demonstrated excellent performance in various application scenarios, their performance in illegal building detection is unsatisfactory. They suffer from high computational costs, low detection accuracy, and high demands for hardware computing resources and memory loads, which may not be suitable for mobile and intelligent devices, and therefore difficult to be widely applied in illegal building detection[23]. In addition, deep learning based methods are data-driven. Previous deep learning based illegal building detection methods mostly rely on image data collected by satellites and drones for segmentation or recognition to detect illegal buildings[44]. The constructed datasets are only applicable to specific regions and lack universality, making them difficult to be widely used.

To overcome these above challenges, a new dataset and detector for identifying illegal construction objects is proposed to timely stop and prevent illegal building activities. The illegal construction object dataset constructed comprises 31 categories of 14 046 images. All of the images were captured by a high-definition camera mounted on a tower about 60 m high in various scenarios, such as urban and suburban areas, and under different natural conditions, including rain, snow, and fog. The proposed detector, named YDHNet, is an enhanced variant of the YOLOv4 approach[40], designed for low computational cost and high detection accuracy in identifying illegal construction objects.

In summary, the key contributions of this paper can be stated as follows:

(1) A new method for detecting illegal buildings that is different from previous methods is proposed. Through the use of end-to-end deep learning based detectors identifying illegal construction-related objects, better and faster detection of illegal construction activities is achieved. The proposed method can achieve real-time detection and early prevention of illegal construction activities. Further, a new dataset of illegal building objects has been constructed, overcoming the limitations of previous datasets and making it widely applicable to various

countries and regions.

(2) The proposed detector YDHNet leverages DenseNet121 as the backbone network, effectively enhancing feature extraction, reuse, and transfer. By adopting DenseNet[45], we observed a remarkable improvement in the accuracy of detecting illegal construction objects. Notably, this advancement is accompanied by a substantial reduction in computational cost and parameter count, making our model highly efficient and practical for real-world deployment.

(3) Depthwise Separable Convolution (DSC) is introduced into the neck and head components of the proposed detector. This integration further enhances the model’s efficiency by reducing computation costs and parameter count while optimizing the overall architecture. The use of DSC enables our model to strike an optimal balance between performance and computational resources.

(4) To boost the non-linear feature learning capabilities of the proposed detector, the new high-performance H-Swish activation function is employed. This choice leads to a substantial improvement in the model’s detection precision, enabling it to capture intricate patterns and features associated with illegal construction objects more effectively.

(5) A comprehensive evaluation of our proposed detector on the illegal construction objects dataset with other State-of-the-Art (SOTA) detectors is involved. The results demonstrate that our model achieves an mean Average Precision (mAP) value of 89.60%, surpassing the performance of other state-of-the-art object detectors such as Single Shot MultiBox Detector (SSD)[46], RetinaNet[33], and various versions of the YOLO Series[39, 40, 47]. Notably, the proposed detector exhibits a remarkable 3.78% improvement in mAP compared to the original YOLOv4. Further, the parameter count and computation load of YDHNet and YOLOv4 are merely 22.10% and 25.24%, respectively. The outstanding balance between detection accuracy and computational efficiency achieved by the proposed illegal building object detector YDHNet makes it an attractive solution for real-world deployment.

The remaining sections of this paper are organized as follows. Section 2 discusses the related work in detail. Section 3 presents the proposed methodology. In Section 4, the proposed dataset of illegal construction objects is introduced, along with the experimental settings and evaluation metrics. Section 5 elaborates on the experimental results and provides an in-depth analysis of the findings. Finally, Section 6 presents the conclusions drawn from this work.

2 Related Work

2.1 Early method

In the early stage, many methods for detecting illegal buildings emerged in response to the continuous development of urbanization and the expansion of cities. Geographic Information Systems (GIS) and Global Positioning Systems (GPS) are combined to detect illegal constructions through remote sensing technology using multi-temporal and high-resolution satellite images in space and time[1]. Benedek et al.[48] used rectangular point processing to locate buildings in a single remote-sensing image. Building detection is made possible by a building-detection adaptive technique that Chen et al.[49] developed for seed placement and region growth. Xu et al.[50] proposed an approach for the automatic detection and classification of building changes from airborne laser scanning data over two periods, where changes are confirmed by setting rules on the difference image map. Khalilimoghadam et al.[51] used a city map, municipal property database, and three dual-temporal high-resolution satellite images to detect buildings under construction to improve performance and accuracy. Felice[52] proffered a two-step method for detecting illegal buildings. The first step is to survey the Illegal Buildings (IBs) close to the river. The second step is to calculate the ranking of these buildings, which can be used as the demolition order of the IBs. To achieve automatic and rapid building monitoring, Zhu and Fan[53] presented high-performance computing based digital city illegal building monitoring as a method of quickly identifying all illegal buildings in a city by utilizing high-performance computing to compare official urban planning maps with construction change images or construction images. Because updating the Geographic DataBase (GDB) in urban environments is a challenging and costly endeavor, Bouziani et al.[54] put forward an approach for identifying changes to buildings in urban settings through the use of Very High Spatial Resolution (VHSR) images and pre-existing digital map data. To reduce manual and rapid detection of building changes, Awrangjeb[9] developed a graphical user interface to support the creation of building databases from automatically extracted building data from lidar point cloud data.

These early methods mostly relied on manual statistical analysis of relevant data on buildings, followed by updating the data in the geographic database and comparing it with past historical data to detect changes and detect illegal construction. However, these methods heavily relied on manual labor, had slow detection speeds, required extremely long detection cycles, and were difficult to automate.

2.2 Methods based on image processing and machine learning

To overcome the limitations of manual labor-dependent early detection approaches, various methods have been developed that are based on digital image processing classical machine learning algorithms to detect illegal buildings. An approach for the semi-automatic of illegal constructions detection in satellite images using a pixel-based fuzzy Exclusive OR (XOR) operator was proposed[55]. Konstantinidis et al.[15] employed a region refinement process and an improved Histogram of Oriented Gradients and Local Binary Pattern (HOG-LBP) function to achieve construction detection. To achieve high-accurate and automatic of construction detection from aerial images, Benarchid et al.[56] established an automated framework for extracting buildings by utilizing target-oriented categorization and shadow data in high-resolution multispectral images. Dornaika et al.[57] utilized image segmentation techniques based on matrix covariance descriptors to achieve automatic and accurate detection of buildings from aerial images. Additionally, K-means[58] has also been extensively applied to recognize unauthorized constructions by detecting alterations in pixel values and outlining building boundaries. Moreover, traditional techniques in digital image processing, such as Candy edge detection and Hough transforms, have been employed to identify unlawful buildings. Furthermore, Tan et al.[59] utilized spectral, spatial, texture, and context features to identify buildings. Chen et al.[60] suggested a weight distribution equation and an enhanced adaptive color transfer algorithm to automatically and precisely detect illegal constructions. An and Guo[61] put forward a technique that eliminates points based on multiscale filtering and clustering growth to identify unauthorized constructions. Wang et al.[62] proposed a semi-automated approach for extracting urban boundaries to tackle the problem of unauthorized construction expansion in urban areas. Meanwhile, Zhang and Liu[63] improved the safety of urban traffic and promoted road safety by utilizing techniques such as edge detection and morphological filtering to identify illegal structures situated near roads.

The majority of these approaches exploit satellite remote sensing imagery in digital image processing and classical machine learning methodologies, significantly decreasing manual involvement and time consumption while accomplishing the semi-automated or automated identification of illegal constructions. Nonetheless, the issue of inadequate detection precision still requires attention.

2.3 Methods based on deep learning CNNs

Recently, deep learning based methods have been widely applied, particularly in tasks such as localization, detection, and classification using CNN-based methods, which have shown strong performance. In the detection of illegal constructions, CNN-based deep learning methods have also been used to achieve high-precision and automated detection. Ostankovich and Afanasyev[44] introduced a technique that merges deep learning tools with image processing methods for automating the detection of illegal constructions in national inspections. This approach incorporates four construction detection technologies based on computer vision, utilizes a retrained GoogLeNet to categorize areas, and cross-checks the legality of detected building regions against cadastral maps. Ishii et al.[64] employed CNN for the classification of multispectral satellite images, and adapted the fully connected layers to convolutional layers for processing images of varying resolutions. Xu et al.[19] utilized deep learning based techniques alongside guided filters to isolate constructions from high-resolution images. Prathap and Afanasyev[65] adopted an enhanced version of UNet for detecting constructions. Vakalopoulou et al.[66] utilized a pre-trained AlexNet and supplementary spectral data for discriminating buildings in the course of training. Perez et al.[8] employed CNNs to detect potential construction activities. Martijn et al.[67] used Conditional Generative Adversarial Networks (CGANs) to correctly predict building shapes for detecting illegal buildings. Liu et al.[68] used UNet and morphological analysis to detect roof changes for identifying potential illegal buildings in the roof area. Furthermore, the Faster R-CNN, a conventional target detection algorithm, has also been utilized in the detection of illegal constructions[69]. Li et al.[70] introduced a detection technique that leverages an enhanced Faster R-CNN approach, which employs low-level and high-level features to identify minor foreground constructions in varied environments.

The utilization of deep learning methods has substantially improved the precision of identification, minimized labor and time expenditures, and accomplished the automatic detection of unlawful construction. However, these approaches can only detect illegal buildings that have already been constructed and cannot prevent or stop the occurrence of illegal construction in a timely manner. In addition, current CNN-based methods for illegal building detection still suffer from high computational costs and insufficient detection accuracy. Therefore, this paper proposes a lightweight and high-precision model for illegal construction object detection. The suggested dataset of illegal construction objects can be used to train the proposed detection model, which permits prompt surveillance, deterrence, and intervention of illegal building undertakings.

3 Method

3.1 Basic detection framework YOLOv4

The proposed YDHNet, an object detection method designed for identifying illegal constructions, relies on the YOLOv4 framework and comprises a backbone network, neck, and prediction head. Specifically, the YOLOv4 detection model consists of an input module, CSPDarkNet53 for feature extraction, Spatial Pyramid Pooling (SPP) layer[71], Path Aggregation Network (PANet)[72], and YOLO head for predictions.

The detection process of YDHNet is similar to YOLOv4, as follows:

Firstly, the Mosaic data augmentation technique is used to randomly crop, move, and splice the four original images to generate a new image in the input module. This method can effectively enhance the detection background, solve the challenge of overly homogeneous scale in the dataset, and avoid the need for a large batch size in batch normalization calculations during model training. Then, the new image is resized to 416 × 416 to extract feature in the backbone network.

In the process of extracting features, the image is input into the backbone network, which outputs three effective feature maps of size 52 × 52, 26 × 26, and 13 × 13. The 13 × 13 feature map is then input into the spatial pyramid pooling layer, which is composed of three maximum pooling layers of sizes 5 × 5, 9 × 9, and 13 × 13. The SPP layer effectively addresses issues such as incomplete object clipping and shape distortion caused by R-CNN's clipping and scaling operations, and also solves the problem of redundant feature extraction of CNNs, which greatly improves the speed of candidate box generation and reduces computational costs. The SPP-processed effective feature maps are then input together with other effective feature maps into the PANet. By performing various operations including convolution, upsampling, downsampling, and feature map fusion, PANet produces three improved and useful feature maps.

The YOLO head for prediction mainly consists of 3 × 3 and 1 × 1 standard convolution layers. The 3 × 3 convolutional layer is used for feature integration, and the 1 × 1 convolutional layer is used for channel number adjustment of the feature map, that is, to obtain the prediction result. The YOLO head module is used to determine whether the target is recognizable and whether the category of the target object conforms to the three preset prior boxes in the three improved effective feature layers. Subsequently, Non-Maximum Suppression (NMS) processing and prior box adjustment are employed to produce the final predicted box.

3.2 Improvement of backbone feature extraction network

To enhance identification precision and decrease computational expenses, in the improved detection model YDHNet, the classification layer of DenseNet121[45] was removed, and the remaining layers were used as the backbone network for feature extraction. Table 1 illustrates the complete architecture of DenseNet121, which can be divided into three parts: a 3 × 3 convolutional layer as the head, a stack of repeated dense blocks and transition layers for extracting and reusing features, and a Fully Connected (FC) layer for identification task. The role of dense block is to perform feature extraction and feature reuse, while transition layer is used to perform downsampling between dense blocks to reduce feature map resolution. The structure of transition layer is relatively simple, consisting of three parts: Batch Normalization (BN) layer, 1 × 1 convolutional layer, and 2 × 2 pooling layer. In addition, transition layer can also be used to

Table 1 Network structure of DenseNet121.

Layer	Output size	DenseNet121
Convolution	112×112	7 × 7 conv, stride = 2
Pooling	56×56	3 × 3 max pool, stride = 2
Dense block_1	56×56	6 × (1×1 conv，3×3 conv)
Transition layer_1	56×56	1 × 1 conv
	28×28	2 × 2 average pool, stride = 2
Dense block_2	28×28	12 × (1×1 conv, 3×3 conv)
Transition layer_2	28×28	1 × 1 conv
	14×14	2 × 2 average pool, stride = 2
Dense block_3	14×14	24 × (1×1 conv, 3×3 conv)
Transition layer_3	14×14	1 × 1 conv
	7×7	2 × 2 average pool, stride = 2
Dense block_4	7×7	16 × (1×1conv, 3×3conv)
Classification layer	1×1	7 × 7 global average pool
	—-	1000D fully-connected, softmax

Note: conv is convolutional kernel.

compress the model. In DenseNet, the most important structure is the dense block. Dense block is the basic module that constitutes DenseNet, and its structure is shown in Fig. 1. In the improved detector YDHNet, the improved dense block is stacked with batch normalization, H-Swish activation function[73], and convolutional layers. In dense block, all layers share the identical feature map size and can be linked in the channel dimension.

Compared to the widely used residual structure, which is also used in the backbone network of YOLOv4, DenseNet utilizes a dense connection strategy that aggressively links all layers together. In DenseNet, each layer is connected to all previous layers in the channel dimension and serves as the input for the next layer. The number of connections in DenseNet is calculated as follows:

$$C = \frac{L(L+1)}{2} \tag{1}$$

where C is the total number of connections and L is the number of network layers.

It is precise because of the dense connection structure in the dense block of the YDHNet backbone network, which directly connects feature maps from different layers, that the feature reuse is enabled, and the utilization efficiency of features is improved. This

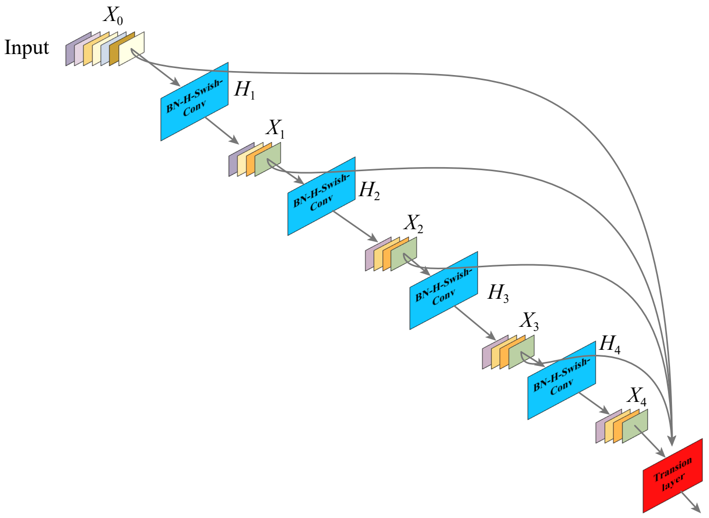

Fig. 1 Structure of dense block in YDHNet.

results in more abundant features extracted by the detection model for illegal building images, which facilitates better recognition. Meanwhile, this dense connection structure also strengthens the flow of gradients, with more skip connections making the gradient easier to propagate forward. Additionally, this structure can obtain more features with fewer parameters, greatly reducing the number of parameters in the detection model.

3.3 Lightweight of neck and head with DSC

To further reduce the computation expense and parameter quantity of the detector, DSC[74] is used in the neck and head of YDHNet. Specifically, the 3 × 3 ordinary convolution in the neck and head is modified to DSC. DSC requires lower computational costs than ordinary convolution. Standard convolution applies the same convolution kernel to all channels of the image for convolution operation, and different convolution kernels are used to extract different features. The convolution kernel of standard convolution is designed for all channels of the input image. Therefore, every time an input image adds a channel, a convolution kernel needs to be added. Finally, standard convolution merges all inputs to obtain a new output. The calculation expression of the output of a standard convolution is shown in Eq. (2), and the computational cost consumption expression is shown in Eq. (3). DSC divides standard convolution into 3 × 3 depthwise convolution and 1 × 1 pointwise convolution, which is shown in Fig. 2. Firstly, 3 × 3 depthwise convolution extracts different features using different convolution kernels on various channels of the image. However, this operation only extracts one aspect of the feature for a specific channel. Therefore, 1 × 1 pointwise convolution is added on this basis to extract different features from the feature map and produce the same output feature map as standard convolution. These two operations greatly reduce the computation expense and parameter quantity. The calculation expression of the output of a depthwise separable convolution is illustrated in Eq. (4), and the computational cost consumption expression is shown in Eq. (5). The expression for the ratio of computational expense between DSC and standard convolution is shown in Eq. (6). Equation (6) shows that the computational cost of DSC is much less than that of ordinary convolution. Therefore, using DSC instead of standard convolution outside the backbone can effectively reduce the parameter count of the detection model. Figure 3 illustrates the structure of DSC in YDHNet. The structures of the neck lightweighted by DSC is shown in Fig. 4.

$$Y_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m} \tag{2}$$

$$S_1 = G_K \cdot G_K \cdot M \cdot N \cdot G_F \cdot G_F \tag{3}$$

$$\hat{Y}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m} \tag{4}$$

$$S_2 = G_K \cdot G_K \cdot M \cdot G_F \cdot G_F + M \cdot N \cdot G_F \cdot G_F \tag{5}$$

$$\frac{G_K \cdot G_K \cdot M \cdot G_F \cdot G_F + M \cdot N \cdot G_F \cdot G_F}{G_K \cdot G_K \cdot M \cdot N \cdot G_F \cdot G_F} = \frac{1}{N} + \frac{1}{G_K^2} \tag{6}$$

where $Y_{k,l,n}$ represents the value in the n-th channel of the output feature map; $K_{i,j,m,n}$ represents the parameters of the convolution kernel, where i and j denote the spatial positions of the kernel, m represents the input channel of the kernel, and n is the output channel of the kernel; $F_{k+i-1,l+j-1,m}$ represents the value in the m-th channel of the input feature map, where k and l represent the spatial positions in the input feature map. $\hat{Y}_{k,l,m}$ represents the value in the m-th channel of the output feature map; $\hat{K}_{i,j,m}$ denotes the parameters of depthwise separable convolution, where i and j represent the spatial positions in the depthwise pointwise convolution kernel, and m represents the

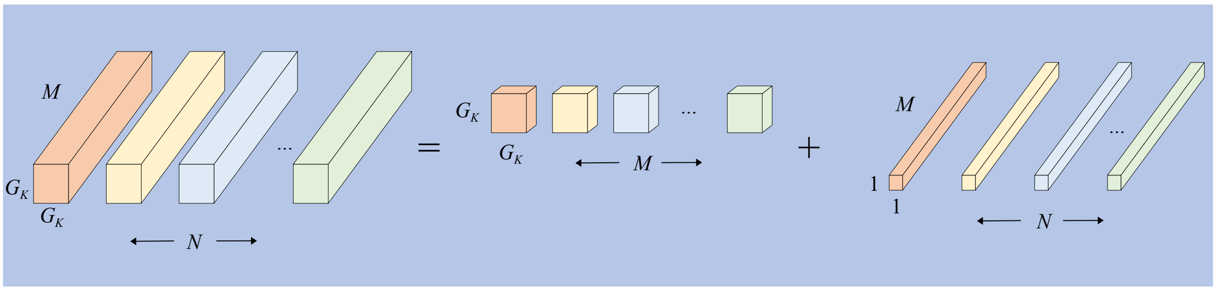

Fig. 2 Schematic diagram of depthwise separable convolution.

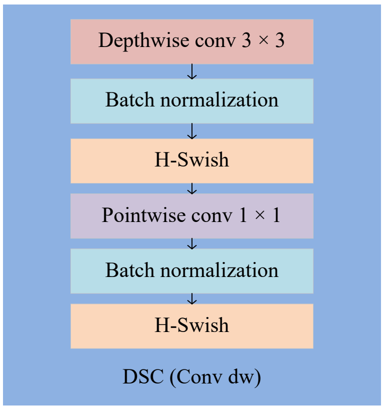

Fig. 3 Structure of DSC in YDHNet.

input channel of the depthwise pointwise convolution; S_1 and S_2 are the computational costs of standard convolution and DSC, respectively; G_F is the width and height of the input feature map; G_K is the spatial dimension of the convolution kernel; M represents the number of input channels; and N represents the number of output channels.

3.4 H-Swish activation function

Choosing an appropriate activation function to improve the identification precision and performance of the object detector is an important aspect of developing such models. In the backbone network of YOLOv4, Leaky ReLU as the main activation function is used outside of the backbone network, and its calculation expression is shown in Eq. (7). The Leaky ReLU activation function solves the problem of neuron death caused by the ReLU function, and its slight positive incline in the negative domain enables backpropagation to be performed on negative input values. Yet, the utilization of the Leaky ReLU across various intervals could result in incongruous outcomes, rendering it incapable of offering dependable forecasts of relationships for both negative and positive input values. Therefore, in the proposed YDHNet model for illegal construction object detection, the H-Swish function[73] is used as the main activation function, and its calculation expression is shown in Eq. (8). The function graphs of Leaky ReLU and H-Swish are shown in Fig. 5.

$$\text{Leaky ReLU}(x) = \begin{cases} x, x \geqslant 0; \\ ax, x < 0 \end{cases} \in \mathbb{R} \tag{7}$$

$$\text{H-Swish}(x) = x \cdot \frac{\text{ReLU6}(x+3)}{6} \tag{8}$$

where x is the input of the activation function, and a is the leakage coefficient.

From the calculation expression and function graph of the H-Swish, it is apparent that this activation

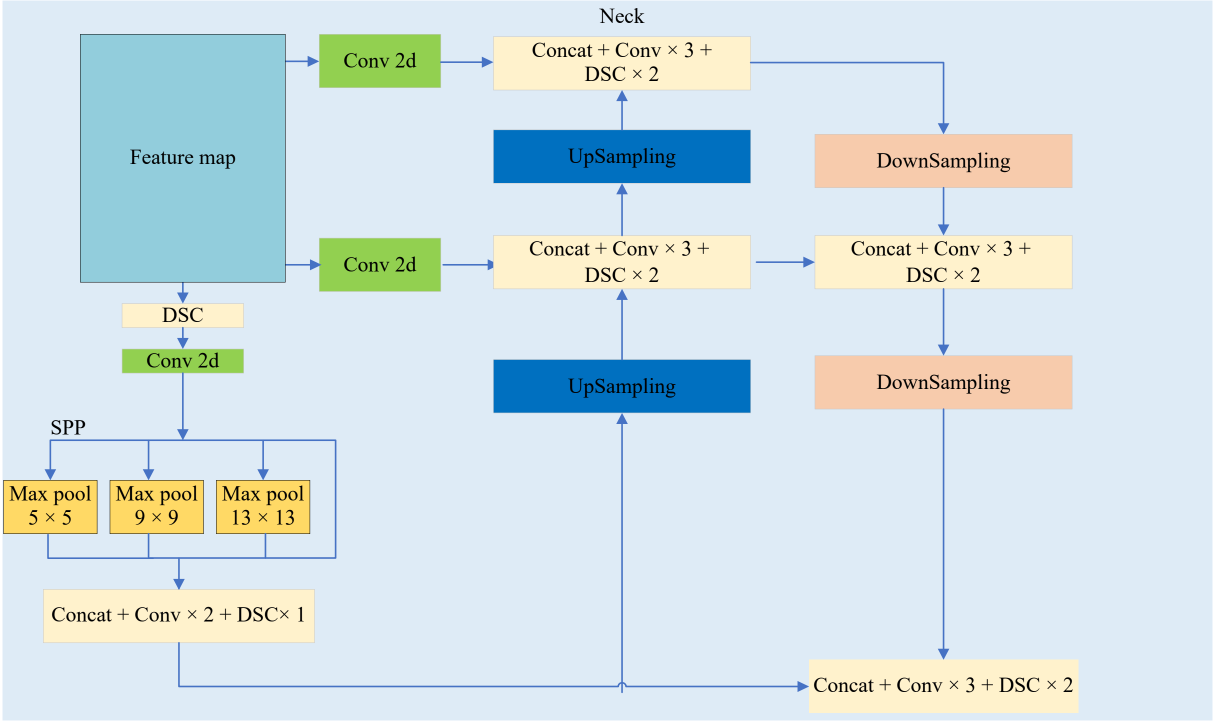
Neck
Feature map
Conv 2d
Conv 2d
Concat + Conv × 3 + DSC × 2
UpSampling
Concat + Conv × 3 + DSC × 2
UpSampling
DownSampling
Concat + Conv × 3 + DSC × 2
DownSampling
DSC
Conv 2d
SPP
Max pool 5 × 5
Max pool 9 × 9
Max pool 13 × 13
Concat + Conv × 2 + DSC× 1
Concat + Conv × 3 + DSC × 2

Fig. 4 Lightweighted neck structure with DSC of YDHNet.

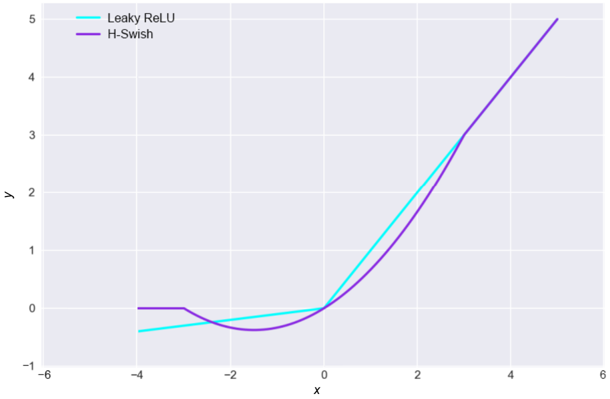

Fig. 5 Function image of H-Swish and Leaky ReLU.

function possesses the following advantages.

(1) Similar to the ReLU function, it has no upper limit. This characteristic is required for any activation function. It can prevent gradient saturation, which results in a significant decline in the speed of training. Therefore, this feature ensures that it does not suffer from gradient saturation problems and can accelerate the training of detection models.

(2) It has a lower bound (the left half-axis of x gradually tends to 0). This can produce stronger regularization effects and effectively prevent overfitting.

(3) Non-monotonic function. This characteristic preserves minor negative values, leading to a stable gradient flow in the network. Most commonly used activation functions cannot maintain negative values, making most neurons unable to be updated.

(4) Everywhere is continuous and differentiable, making it easier to train.

(5) The H-Swish activation function is a differentiable function with robust generalization capability and efficient optimization ability, which can significantly enhance the recognition accuracy of neural networks.

In summary, the unique non-monotonicity of H-Swish enhances the model’s detection precision for damaged illegal construction objects. Due to the absence of an upper limit but the presence of a lower limit, it can effectively address the saturation issue of input neurons and enhance the regularization impact of the model[28]. Furthermore, it offers computational efficiency advantages over the Swish function, which facilitates the training process. Additionally, it considerably minimizes the memory accesses required for the detection model.

3.5 Proposed detection model YDHNet

Figure 6 shows the architecture of the suggested detector YDHNet. Compared with YOLOv4, YDHNet has significant advantages in both detection accuracy and parameter quantity. Feature extraction of the input illegal building images is performed using DenseNet121, which has dense connections, instead of the original CSPDarkNet53 as the backbone network. In DenseNet, the fully connected layer used for classification is removed but other parts are retained, which includes the 3 × 3 convolutional layer of its head, pooling layers, and dense blocks structure, to maintain consistency in the network. The improved YOLOv4 algorithm YDHNet uses DenseNet121 as the feature extraction network, which not only alleviates gradient vanishing but also strengthens the feature extraction and transfer of illegal construction objects in the network, more effectively reusing features, greatly enhancing feature utilization efficiency, and improving the recognition precision of the detection model for illegal construction objects. Therefore, only the number of feature layer channels is changed in the detection network, while the size of the feature layers in the network remains the same as before. Additionally, due to the dense connection mechanism of DenseNet, the number of model parameters is greatly reduced. Furthermore, in addition to improving the backbone feature extraction network of YDHNet, the 3 × 3 standard convolution in neck and head was also optimized by modifying it to DSC, further decreasing the parameter quantity and computational expense of the detection model. In addition, the H-Swish[73] is utilized as the primary activation function of the detector. Compared with the Leaky ReLU activation function mainly used in YOLOv4, H-Swish can not only perform better nonlinear feature extraction but also has better generalization ability and result optimization ability, enabling the model to better recognize illegal construction objects and further enhance the identification precision of the detector for illegal construction objects. The experiments show that after these modifications, the model can effectively solve the problems of low recognition precision and high computational cost in the current illegal building detection methods, and its performance in illegal construction object detection exceeds that of other SOTA detectors.

Furthermore, the loss function of YDHNet primarily

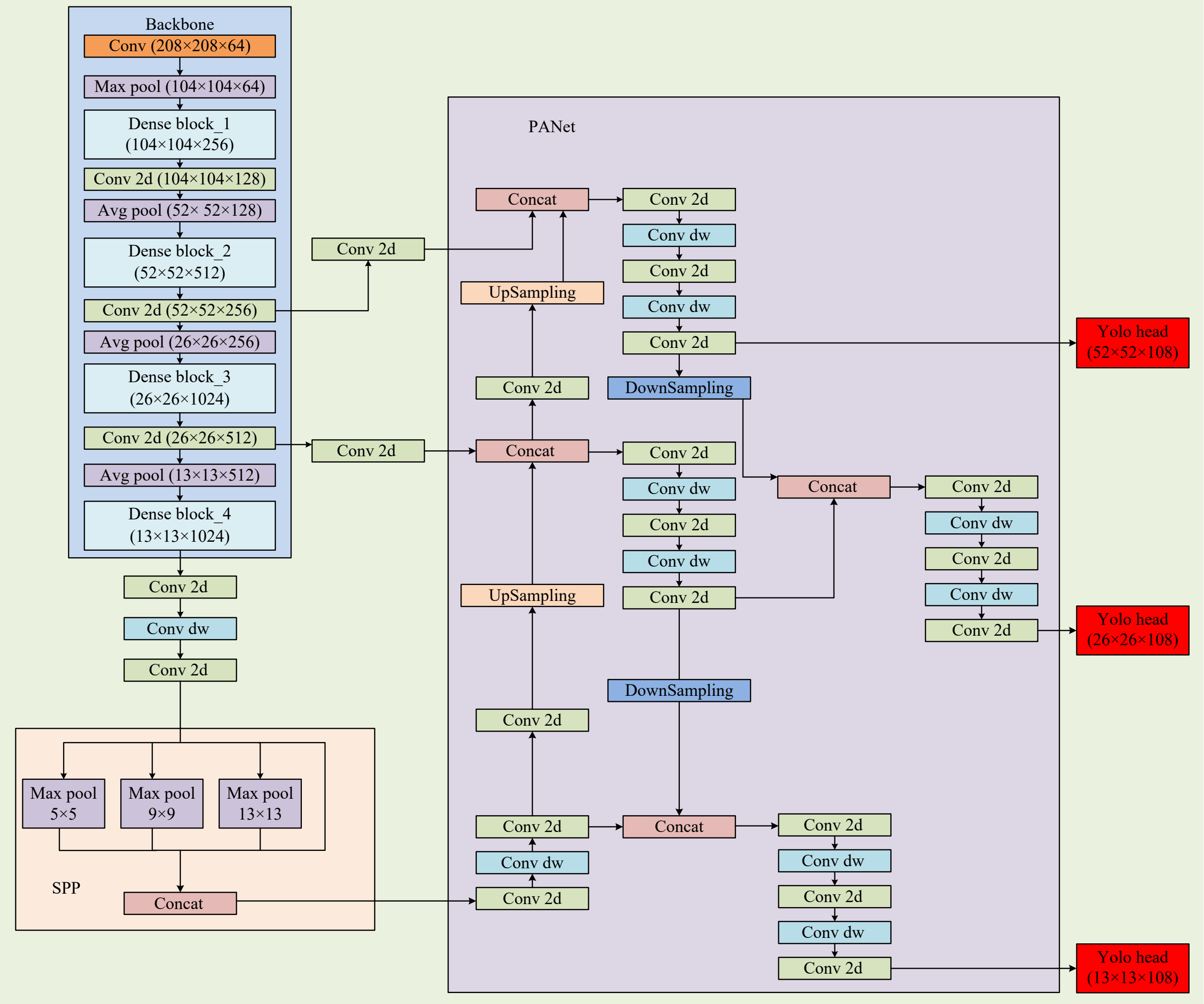

Fig. 6 Network structure of proposed detection model YDHNet.

consists of the subsequent components: object position loss function (L_l), classification loss function (L_c), and confidence loss function (L_cf). The loss function is calculated as follows:

$$\mathrm{Loss} = L_\mathrm{l} + L_\mathrm{c} + L_\mathrm{cf} \tag{9}$$

The object position loss function (L_l) is computed using the following equation:

$$L_\mathrm{l} = 1 - \mathrm{IoU} + \frac{\alpha^2(b_\mathrm{p}, b_\mathrm{t})}{\gamma^2} + \beta v \tag{10}$$

$$\mathrm{IoU} = \frac{b_\mathrm{p} \cap b_\mathrm{t}}{b_\mathrm{p} \cup b_\mathrm{t}} \tag{11}$$

$$\beta = \frac{v}{(1 - \mathrm{IoU}) + v} \tag{12}$$

$$v = \frac{4}{\pi}(\arctan\frac{w_\mathrm{t}}{h_\mathrm{t}} - \arctan\frac{w_\mathrm{p}}{h_\mathrm{p}}) \tag{13}$$

where b_p represents the prediction box, b_t is the real box, $\alpha^2(b_\mathrm{p}, b_\mathrm{t})$ is the Euclidean distance between the center point of the prediction box and the center point of the real box, γ represents the diagonal distance of the smallest closed rectangular area that can contain both the prediction box and the real box, and IoU represents the intersection ratio of the prediction box and the ground truth. w_t and h_t represent the width and height of the real boundary box, respectively. w_p and h_p represent the width and height of the predicted boundary box, respectively.

The computation formula for the classification loss function (L_c) is as follows:

$$L_\mathrm{c} = \sum\nolimits_{i=0}^{s^2} l_{i,j}^{o} \sum\nolimits_{c\in m} \hat{p}_i^j(c)\log(p_i^j(c)) + (1 - \hat{p}_i^j(c)\log(p_i^j(c))) \tag{14}$$

where s^2 is the number of grids to divide the image. $l_{i,j}^{o}$ represents whether the i-th grid contains the target in the j-th prediction box, if it contains, its value is 1;

otherwise, it is 0. $\hat{p}_i^j$ represents the probability of the target object category in the prediction box, p_i^j is the probability of the target object category in the real box, c is the object category, and m is the total number of target object categories.

The computation equation for the confidence loss function (L_{cf}) is as follows:

$$L_{\text{cf}} = \sum_{i=0}^{s^2} \sum_{j=0}^{m} l_{i,j}^{o} (\hat{c}_i^j \log(c_i^j) + (1-\hat{c}_i^j)\log(1-c_i^j)) - \\ \lambda_{\text{no}} \sum_{i=0}^{s^2} \sum_{j=0}^{m} l_{i,j}^{\text{no}} (\hat{c}_i^j \log(c_i^j) + (1-\hat{c}_i^j)\log(1-c_i^j)) \quad (15)$$

where λ_{no} represents the error weight parameter. $l_{i,j}^{\text{no}}$ is whether the i-th grid contains the target object in the j-th prediction box, if it does not, its value is 1; otherwise, it is 0. c_i^j is the confidence of the ground truth, and $\hat{c}_i^j$ represents the confidence of the predicted box.

4 Experiment

4.1 Dataset

The dataset of illegal building structures constructed by our research team was captured by using high-definition cameras. The high-definition camera model used was DS-2TD6236-75C2L. The images of illegal building structures collected were obtained from various complex real-world scenarios. Table 2 shows the specifications of these High-Definition (HD) cameras. The images in this dataset were taken in natural conditions, involving various scenes, weather conditions, occlusions, and overlaps, and have a size of 1080 pixel × 1920 pixel. This dataset includes 14 046 images containing 31 classes of illegal construction objects, including construction vehicles (such as concrete transport trucks, bulldozers, excavators, cranes, etc.), construction materials (such as bricks, piles of soil, wood, etc.), and construction tools (such as concrete mixers, etc.). For the collected image data, all images were first cut into pictures with a size of 416 pixel × 416 pixel. Then, the images were manually annotated by LableImg to identify the category and position of the illegal construction objects in the collected images. Finally, all annotated image data were formatted as the Pascal VOC dataset to train the detection model. Table 3 shows the category names and descriptive information of the illegal construction objects in the dataset, and Fig. 7 displays some sample images of the illegal construction objects. The number and proportion of each illegal construction object category in the dataset are illustrated in Figs. 8 and 9, respectively.

Table 2 Parameter values of our camera.

Camera parameter	Value
Protect model number	iDS-2DY9437IX-A/SP
Manufacturer	Hikvision
Horizontal range	360°
Vertical range	(−20°, 90°)
Horizontal preset point speed	0.1°/s−210°/s
Vertical preset point speed	0.1°/s−150°/s
Operating temperature	(−40 ℃, 70 ℃)
Working humidity	< 95%
Weight	8 kg
Protection level	IP67

4.2 Experimental setting

All experiments in this study were performed on a GPU server with a graphics card, NVIDIA GeForce GTX Titan X, having 12 GB memory, and Ubuntu 18.04 as the operating system. The server configuration details are provided in Table 4. The proposed detection model, YDHNet, was trained using both pre-training and fine-tuning approaches. Initially, the detector was trained on the Pascal VOC dataset, and then it was fine-tuned using a specialized dataset consisting of illegal construction objects. The illegal construction object dataset was split into a training set, a validation set, and a test set, with 11 376, 1265, and 1405 instances, respectively. The YDHNet detection model was trained for 100 epochs, utilizing the Adam optimizer. For the first 50 epochs, the backbone network was frozen, DenseNet121, and only trained the non-backbone parts of the model. In the subsequent 50 epochs, the entire model was trained with the unfrozen backbone network. The learning rate was set to 0.001 for the first 50 epochs and 0.0001 for the last 50 epochs, with the batch size being 16 for the first 50 epochs and 8 for the remaining 50 epochs.

4.3 Evaluation metrics

To evaluate the efficacy of the proposed detection model, this study chose to use commonly employed metrics in the field of object detection. These metrics are mAP, Average Precision (AP), precision, recall, and $F1$-score, and their respective calculation formulas are illustrated as follows:

Table 3 Category and description of illegal construction objects.

Index	Category name	Description
1	Bar deposits	Tubular building materials.
2	Big building	A tall continuously building having multiple floors.
3	Big truck	Large trucks transporting construction materials.
4	Black cover	A black material is used to prevent dust.
5	Blue enclosure	A temporary color steel fence is constructed around the perimeter of a construction project.
6	Boxcar	Container trunk.
7	Brick	Tile rotation.
8	Bricks	Board bricks.
9	Building frame	A building that is on hold or under construction at a certain stage.
10	Building	Sample building.
11	Bulldozer	Bulldozing construction vehicles.
12	Concrete simple house	A simple house made of concrete.
13	Concrete truck	Drum type cement mixing device is usually installed.
14	Crane closed	Crane not working.
15	Crane	There are many types of machines for lifting or moving heavy objects, used in mines, construction sites, etc.
16	Digger	A machine for excavating.
17	Drill	A tool used for making round holes or driving fasteners.
18	Earth vehicles	A truck is used to transport building materials such as sand and stone.
19	Fuel tank	Vehicles carrying large quantities of combustible liquids.
20	Grave mound	Burial heads and graves for the dead.
21	Green cover	A green covering that serves to conceal or shelter something.
22	Greenhouse	Facilities that can transmit light and keep warm are used to cultivate plants.
23	Grey enclosure	The utility model takes light steel as the skeleton and sandwich plate as the enclosure material.
24	Mixer	Simple concrete mixer.
25	Piece deposits	Piled up by particles, soil, and rocks.
26	Scaffold	A temporary arrangement is erected around a building for the convenience of workers.
27	Simple house	Board room.
28	Small truck	Small trucks transporting construction materials.
29	Tower	Tower crane.
30	Vehicle	A vehicle is used to transport people or building materials.
31	Woodpile	A pile of wood.

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{16}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{17}$$

$$F1 = \frac{2\times\mathrm{TP}}{(2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN})} \tag{18}$$

$$\mathrm{AP} = \int_0^1 P(R)\mathrm{d}R \tag{19}$$

$$\mathrm{mAP} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{AP}_i \tag{20}$$

where P represents precision; R represents recall, AP represents average precision; mAP represents mean average precision; $F1$ represents $F1$-score; TP represents the number of correctly detected illegal construction objects; FP is the number of non-illegal construction objects that are mistakenly detected as illegal; and FN is the number of illegal construction objects that are not detected. Detecting illegal construction objects involves two sub-tasks: classification and localization. Illegal construction objects not only need to be correctly classified, but also must be accurately localized to be detected.

5 Result and Analysis

5.1 Comparison with YOLOv4

To test the efficacy of the proposed detector, YDHNet

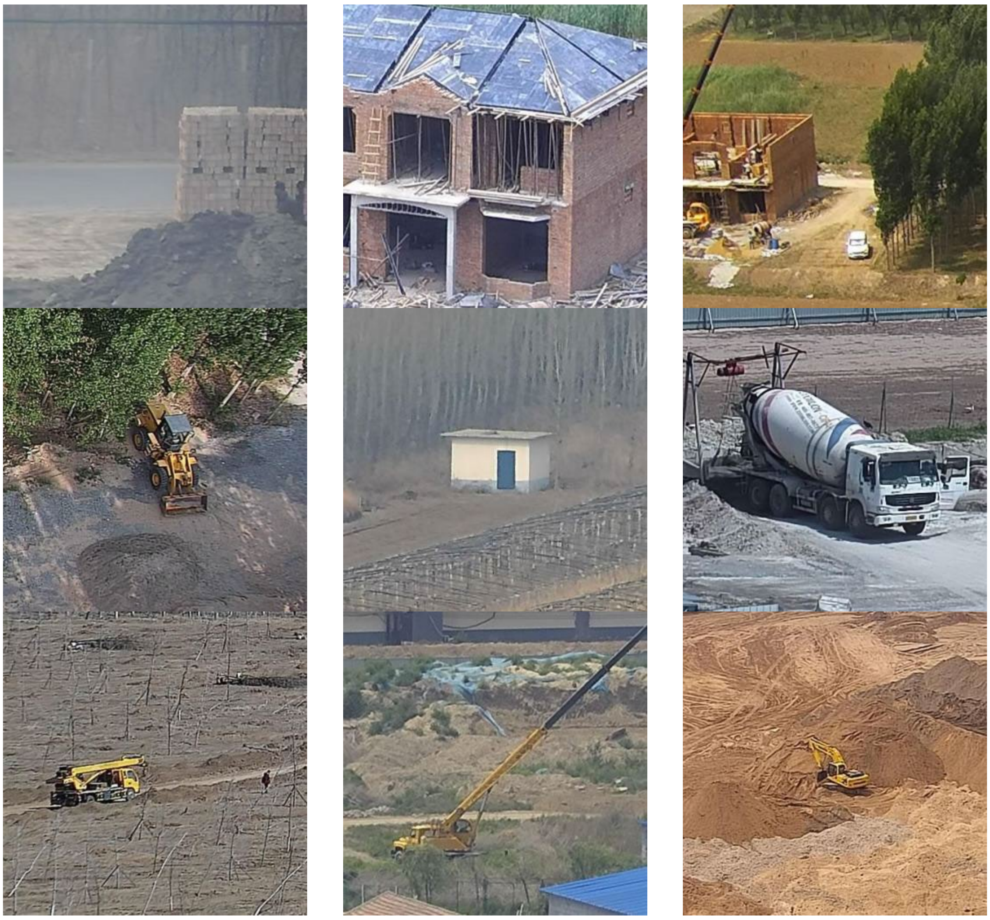

Fig. 7 Samples images of the illegal construction objects dataset.

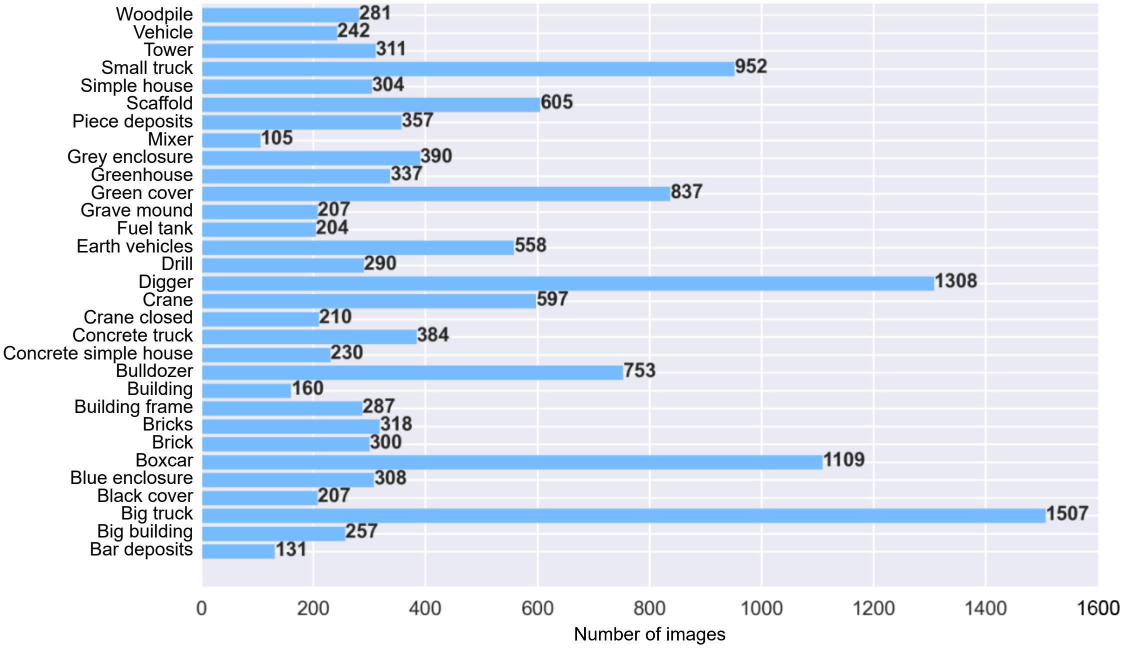

Fig. 8 Number of images for each category of illegal construction object in the dataset.

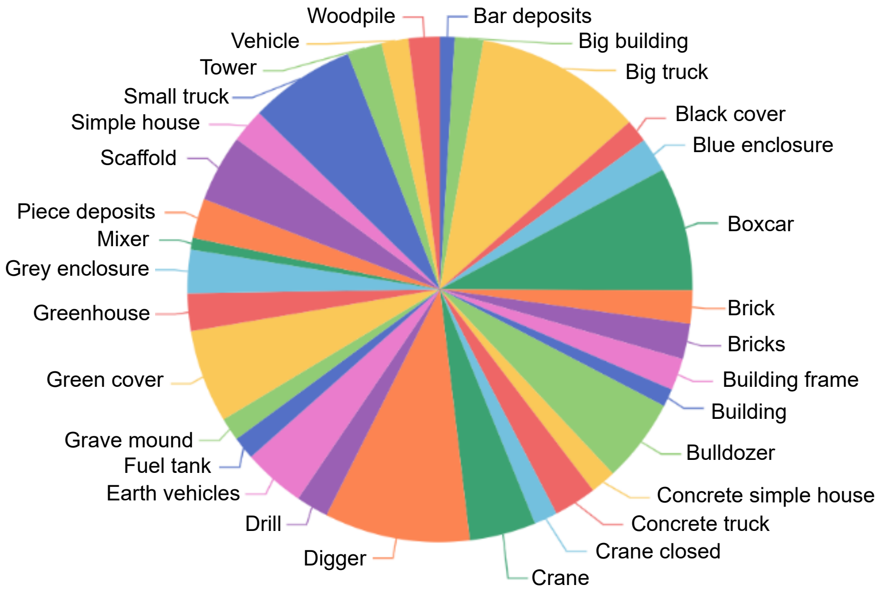

Fig. 9 Ratio of the number of images for each category of illegal construction objects in the dataset.

Table 4 Configuration information of the GPU server.

GPU server	Configuration information
Architecture	x86_64
CPU op-mode(s)	32-bit, 64-bit
Model name	Intel (R) Xeon (R) CPU E5-2667 v4 @ 3.20 GHz
GPU	NVIDIA GeForce GTX TiTan X
Cuda memory	12 GB
System	Ubuntu18.04

was compared with YOLOv4. The study was conducted using a dataset of illegal construction objects, which was proposed in this research. Table 5 presents the experimental results obtained by running both models on the proposed dataset. The precision, recall, and $F1$-score values of YOLOv4 are 85.03%, 82.76%, and 83.56%, respectively. On the other hand, YDHNet demonstrated improved performance with precision, recall, and $F1$-score values of 90.08%, 86.06%, and 87.80%, respectively. When compared to YOLOv4, YDHNet showed a significant increase of 5.05%, 3.30%, and 4.24% in precision, recall, and $F1$-score, respectively. In terms of the most important evaluation metric, mAP, YDHNet achieved a value of 89.60%, which is 3.78% higher than YOLOv4. The proposed model outperformed YOLOv4 in 25 out of 31 classes of illegal construction objects, while performing slightly worse in only 6 classes, including blue enclosure, earth vehicles, grave mound, green cover, scaffold, and woodpile. However, this indicates that the proposed method is effective and can enhance the detection accuracy of most classes of illegal construction objects, with lower AP values observed only in the detection of green cover, blue enclosure, and woodpile. This could be attributed to the imbalanced distribution of these classes in the dataset, or the complexity of detecting these objects due to their unclear characteristics and complex surroundings. During the experiment, the majority of the models failed to distinguish between grave mounds and piece deposits, despite their distinct differences. Therefore, these two classes were labeled separately in the annotated dataset, leading to improved detection accuracy. Furthermore, Fig. 10 demonstrates the detection results for several sample images.

5.2 Comparison with SOTA detectors

To better access the performance of YDHNet, 18 popular and advanced object detection models were selected and compared with the proposed YDHNet on multiple evaluation metrics, including mAP, precision, recall, and $F1$-score, which are illustrated in Table 6. SSD[46], RetinaNet[33], Faster R-CNN[70], YOLOv5[75], YOLOX[34], EfficientDet[36], YOLO Series[39, 40, 47], and the proposed detector YDHNet were trained on the proposed dataset until the models converged. Table 6 illustrates the experimental results, demonstrating that the proposed detector YDHNet achieved the highest

Table 5 Results of comparative experiment with YOLOv4.

Category	YDHNet				YOLOv4			
	AP (%)	Precision (%)	Recall (%)	*F*1-score (%)	AP (%)	Precision (%)	Recall (%)	*F*1-score (%)
Bar deposits	**89.10**	100.00	76.92	86.95	79.60	81.82	69.23	75.00
Big building	**84.13**	84.00	84.00	84.00	62.94	72.73	64.00	68.09
Big truck	**96.66**	94.23	96.71	95.45	96.14	91.08	94.08	92.56
Black cover	**84.84**	88.89	69.57	78.05	80.62	79.17	82.61	80.85
Blue enclosure	65.70	65.22	53.57	58.82	**71.46**	80.95	60.71	69.38
Boxcar	**99.70**	98.11	100.00	99.05	99.47	98.11	100.00	99.05
Brick	**77.01**	73.53	75.76	74.63	71.67	77.42	72.73	75.00
Bricks	**86.40**	83.33	88.24	85.71	85.43	81.25	76.47	78.79
Building frame	**88.67**	86.67	86.67	86.67	84.23	80.65	83.33	81.97
Building	**74.51**	92.31	75.00	82.76	60.67	68.75	68.75	68.75
Bulldozer	**99.89**	97.87	97.87	97.87	97.83	98.91	96.81	97.85
Concrete simple house	**90.04**	95.24	90.91	93.02	89.18	79.17	86.36	82.61
Concrete truck	**99.78**	97.87	97.87	97.87	99.31	97.78	93.62	95.65
Crane closed	**100.00**	100.00	100.00	100.00	94.74	100.00	94.74	97.30
Crane	**96.87**	96.30	94.55	95.42	96.71	92.73	92.73	92.73
Digger	**99.90**	100.00	96.92	98.44	99.09	99.21	96.92	98.05
Drill	**98.96**	96.67	96.67	96.67	96.44	96.43	90.00	93.10
Earth vehicles	94.08	87.50	88.89	88.19	**95.36**	85.29	92.06	88.55
Fuel tank	**100.00**	100.00	100.00	100.00	99.41	95.65	100.00	97.78
Grave mound	90.72	88.89	84.21	86.49	**93.39**	70.83	89.47	79.07
Green cover	68.21	77.59	54.88	64.29	**72.86**	86.54	54.88	67.17
Greenhouse	**91.69**	93.55	87.88	90.63	82.82	87.50	84.85	86.15
Grey enclosure	**94.02**	94.29	86.84	90.41	93.94	88.24	78.95	83.34
Mixer	**97.46**	85.71	92.31	88.89	82.29	68.75	84.62	75.86
Piece deposits	**87.03**	82.35	82.35	82.35	86.16	82.86	85.29	84.06
Scaffold	82.23	80.33	77.78	79.03	**83.67**	81.25	82.54	81.89
Simple house	**95.69**	92.59	89.29	90.91	87.25	92.00	82.14	86.79
Small truck	**94.27**	86.67	95.12	90.70	89.61	90.91	85.37	88.05
Tower	**99.83**	95.83	95.83	95.83	84.04	70.37	79.17	74.51
Vehicle	**80.42**	95.00	73.08	82.61	68.47	85.71	69.23	76.59
Woodpile	69.00	81.82	78.26	80.00	**75.47**	73.91	73.91	73.91
Average	**89.60**	**90.08**	**86.06**	**87.80**	**85.82**	**85.03**	**82.76**	**83.56**

mAP value of 89.60%. The mAP values of YOLOv3 and SSD closely follow, at 86.37% and 86.19%, respectively. The mAP values of these two detection models are very similar. Next are YOLOv4 and RetinaNet, with mAP values of 85.82% and 85.22%, respectively. The mAP values of other state-of-the-art detectors are much lower than these detectors. In addition, the mAP value and *F*1-score of YOLOv7 are 81.54% and 79.75%, respectively, which are much lower than the proposed illegal building detector YDHNet. Furthermore, the *F*1-score of the proposed detection model is 87.80%, which is much higher than other popular and SOTA detectors. Meanwhile, the proposed YDHNet model has parameter numbers and computational costs of 16.18 MB and 26.22 GFLOPs, respectively, which are less than those of popular detectors such as SSD, YOLOv4, YOLOv3, RetinaNet, and Faster R-CNN. This indicates that the proposed detection model requires low memory overhead and computational cost, making it suitable for mobile and intelligent devices. Therefore, the proposed model can accurately detect illegal construction objects in real-time and effectively prevent illegal construction activities.

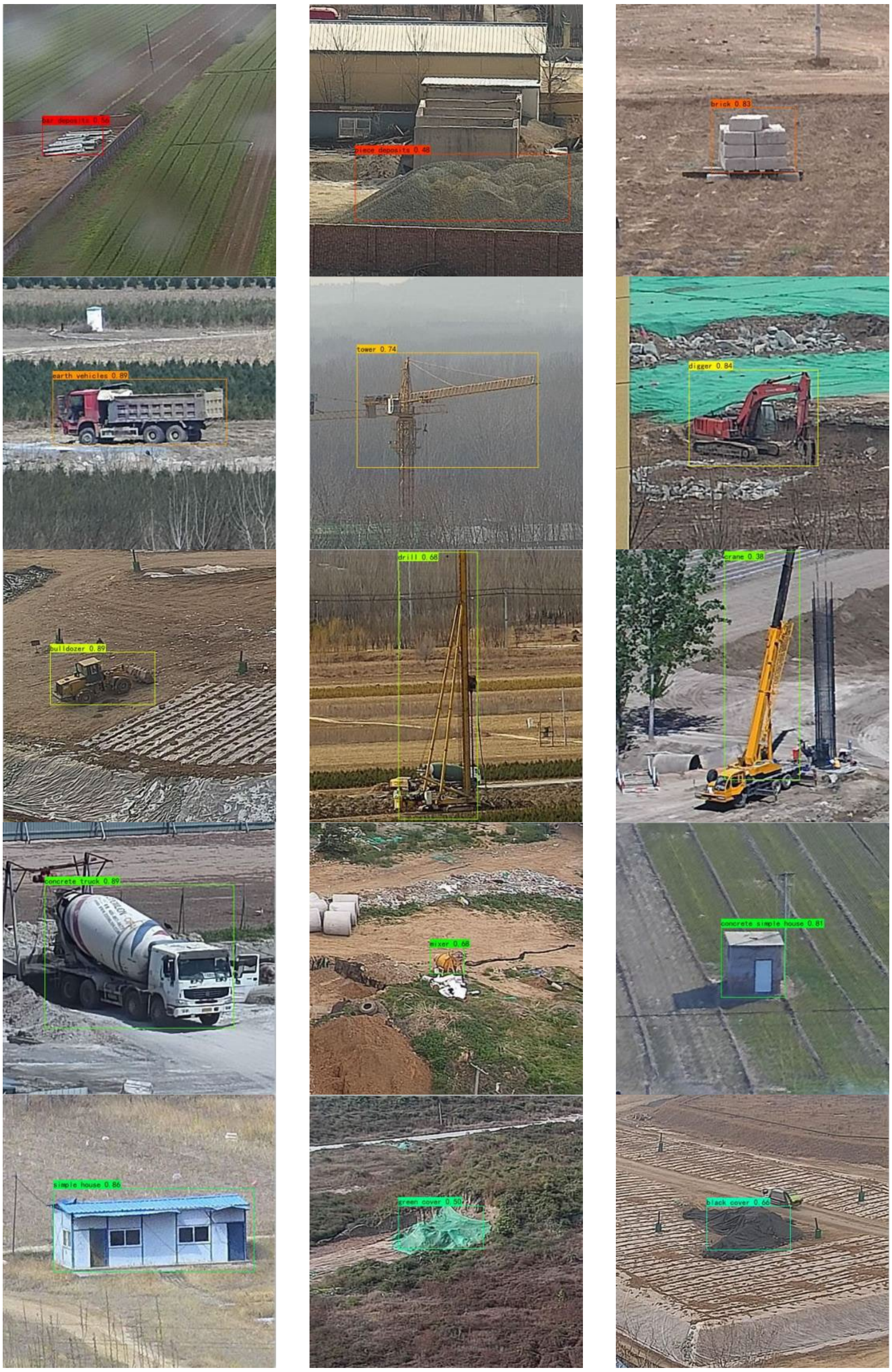

Fig. 10 Detection results of sample images of YDHNet.

5.3 Ablation study

Table 7 illustrates the results obtained using YOLOv4 as the baseline model on the proposed dataset, as well as the results of the improved detectors. YD represents the utilization of DenseNet121 as the backbone network to enhance the utilization efficiency of features, thus significantly improving the accuracy of illegal construction object recognition while reducing

Table 6 Results of comparative experiment with other SOTA detectors.

Detector	mAP (%)	Precision (%)	Recall (%)	$F1$-score (%)	Size of parameters (MB)	GFLOPs
SSD	86.19	87.49	79.85	81.88	27.62	63.74
RetinaNet	85.22	78.46	80.84	79.01	36.95	154.74
Faster R-CNN	78.64	56.66	86.38	67.44	137.30	370.46
YOLOv5-L	78.14	79.03	61.78	66.36	46.79	48.62
YOLOv5-M	77.30	78.10	62.32	65.37	21.18	21.54
YOLOv5-S	72.73	72.69	65.71	64.70	7.14	7.07
YOLOv5-X	81.27	84.66	59.37	64.90	87.45	92.29
YOLOX-L	83.29	83.34	70.85	74.44	54.17	65.83
YOLOX-M	83.70	83.52	75.37	78.02	25.30	31.19
YOLOX-S	83.44	79.86	79.86	79.86	8.95	11.33
YOLOX-X	62.45	79.52	79.52	79.52	99.02	119.23
EfficientDet-D0	79.47	81.51	70.97	73.06	3.85	4.93
EfficientDet-D1	77.45	75.97	71.20	71.11	6.58	11.90
EfficientDet-D2	79.58	74.18	75.58	71.75	8.08	21.28
EfficientDet-D3	77.41	68.19	75.35	69.51	11.98	48.34
YOLOv3	86.37	87.82	81.55	84.29	61.69	65.82
YOLOv4	85.82	85.03	82.76	83.56	64.10	60.17
YOLOv7	81.54	81.87	78.69	79.75	71.02	189.53
YDHNet	89.60	90.08	86.06	87.80	16.18	26.22

Table 7 Results of ablation experiment.

Detector	DenseNet121	DSC	H-Swish	mAP (%)	Precision (%)	Recall (%)	$F1$-score (%)
YOLOv4	×	×	×	85.82	85.03	82.76	83.56
YD	√	×	×	88.60	87.72	84.50	85.87
YM	×	√	×	86.46	86.63	80.58	82.42
YA	×	×	√	87.09	87.12	80.09	82.58
YDM	√	√	×	88.93	89.23	86.06	87.39
YDA	√	×	√	89.02	87.74	86.39	86.82
YMA	×	√	√	87.45	87.75	85.61	86.49
YDHNet	√	√	√	89.60	90.08	86.06	87.80

computational costs and parameters. YM represents the modification of the 3 × 3 standard convolution outside the backbone network to DSC, further reducing the parameters of the detection model. YA represents the use of the H-Swish activation function instead of Leaky ReLU to improve the detection accuracy of the model. YDM represents the utilization of DenseNet121 as the backbone network, with DSC applied outside the backbone network. YDA represents the use of DenseNet121 as the backbone network and the adoption of H-Swish instead of Leaky ReLU as the activation function. YMA represents the application of DSC outside the backbone network and the usage of H-Swish instead of Leaky ReLU as the activation function. YDHNet is the detector that integrates all the improvement methods mentioned above.

The baseline model YOLOv4 achieved an mAP value of 85.82%. After modifying the original backbone network CSPDarkNet53 to DenseNet121, the mAP value improved to 88.60%, representing a 2.78% increase compared to the unimproved version. When only incorporating DSC into the detection model structure, the mAP value reached 86.46%, showing a 0.64% improvement. By using the H-Swish activation function outside the backbone network, the mAP value is 87.09%, which is a 1.27% improvement over YOLOv4. By replacing CSPDarkNet53 with DenseNet121 for feature extraction and further incorporating the DSC modification, the mAP value reached 88.93%. Compared to YOLOv4, this is a 3.11% improvement. When using DenseNet121 for feature extraction and employing the H-Swish activation function, the mAP value of the improved detector reached 89.02%, indicating a 3.20% improvement over YOLOv4. With the optimization of both DSC and H-Swish activation function, the mAP value reached 87.45%, a 1.63% improvement over YOLOv4. When all the optimization methods were applied to the baseline YOLOv4 model, the mAP value reached 89.60%, reflecting a 3.78% improvement over the original YOLOv4. Furthermore, the proposed detector YDHNet achieved an $F1$-score value that was 4.24% higher than YOLOv4, indicating significant improvements in detection accuracy and avoiding false positives. This is crucial for practical applications of illegal construction object detection, as a higher $F1$-score means the model can accurately identify and locate illegal construction objects, providing more reliable results.

6 Conclusion

This paper presents a new method for detecting illegal constructions to prevent and address illegal building activities effectively. Key contributions include the construction of a specialized dataset, the development of a high-precision, and lightweight illegal construction object detector. Compared to previous methods, the proposed detector YDHNet achieves higher detection accuracy with lower computational costs. The optimization of YOLOv4 involves replacing CSPDarkNet53 with DenseNet121 as the backbone network, enhancing feature extraction effectiveness, and reducing model parameters. In addition, the introduction of DSC further reduces computational costs while improving the model structure. H-Swish activation is employed for enhanced feature learning, leading to improved detection accuracy. YDHNet outperforms existing state-of-the-art detectors, achieving the highest mAP, precision, recall, and $F1$-score for illegal construction object detection. Parameters and computational complexity are significantly reduced compared to YOLOv4. YDHNet has demonstrated excellent accuracy and efficiency in detecting illegal building objects. The potential for its application in illegal building detection in the real world is promising. In the future, we plan to further expand the dataset by increasing the diversity and scale of the illegal construction object dataset. In addition, we will continue to optimize the model's structure and algorithms to enhance real-time performance and efficiency, thereby promoting the advancement of intelligent urban development.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Nos. 62177034 and 61972046).

References

[1] L. Yang, T. H. Chi, L. Peng, and X. Sun, Research of illegal building monitoring system construction with 3S integration technology, in *Proc. 2nd Int. Conf. Information Science and Engineering*, Hangzhou, China, 2010, pp. 3908–3911.

[2] D. Liu, Photogrammetric approach for detection of rooftop illegal structures, https://theses.lib.polyu.edu.hk/handle/200/7044, 2013.

[3] P. F. Lezaun and G. Olivieri, Undeclared constructions: A government's support deep learning solution for automatic change detection, in *Proc. 2018 ITU Kaleidoscope*: *Machine Learning for a 5G Future* (*ITU K*), Santa Fe, Argentina, 2018, pp. 1–7.

[4] N. Seror and B. A. Portnov, Identifying areas under potential risk of illegal construction and demolition waste dumping using GIS tools, *Waste Manag*., vol. 75, pp. 22–29, 2018.

[5] F. da Conceição Leite, R. dos Santos Motta, K. L. Vasconcelos, and L. Bernucci, Laboratory evaluation of recycled construction and demolition waste for pavements, *Constr. Build. Mater.*, vol. 25, no. 6, pp. 2972–2979, 2011.

[6] B. Varol, E. Ö. Yılmaz, D. Maktav, S. Bayburt, and S. Gürdal, Detection of illegal constructions in urban cities: Comparing LIDAR data and stereo KOMPSAT-3 images with development plans, *Eur. J. Remote. Sens.*, vol. 52, no. 1, pp. 335–344, 2019.

[7] C. Ioannidis, C. Psaltis, and C. Potsiou, Towards a strategy for control of suburban informal buildings through automatic change detection, *Comput. Environ. Urban Syst.*, vol. 33, no. 1, pp. 64–74, 2009.

[8] D. Perez, D. Banerjee, C. Kwan, M. Dao, Y. Shen, K. Koperski, G. Marchisio, and J. Li, Deep learning for effective detection of excavated soil related to illegal tunnel activities, in *Proc. 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference* (*UEMCON*), New York, NY, USA, 2017, pp. 626–632.

[9] M. Awrangjeb, Effective generation and update of a building map database through automatic building change detection from LiDAR point cloud data, *Remote. Sens.*, vol. 7, no. 10, pp. 14119–14150, 2015.

[10] T. Baylor, Active contour and prior information for change analysis: Applied to updating urban building digital maps based on high-resolution remote sensing optical images, (in French), PhD dissertation, National Institute of Technology, Toulouse, France, 2005.

[11] E. P. Baltsavias, Object extraction and revision by image analysis using existing geodata and knowledge: Current status and steps towards operational systems, *ISPRS J. Photogramm. Remote. Sens.*, vol. 58, nos. 3&4, pp. 129–

151, 2004.

[12] Y. Tang, X. Huang, and L. Zhang, Fault-tolerant building change detection from urban high-resolution remote sensing imagery, *IEEE Geosci. Remote. Sens. Lett.*, vol. 10, no. 5, pp. 1060–1064, 2013.

[13] X. Huang, L. Zhang, and T. Zhu, Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index, *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 7, no. 1, pp. 105–115, 2014.

[14] V. Walter, Object-based classification of remote sensing data for change detection, *ISPRS J. Photogramm. Remote. Sens.*, vol. 58, nos. 3&4, pp. 225–238, 2004.

[15] D. Konstantinidis, T. Stathaki, V. Argyriou, and N. Grammalidis, Building detection using enhanced HOG–LBP features and region refinement processes, *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 10, no. 3, pp. 888–905, 2017.

[16] K. Karantzalos and D. Argialas, A region-based level set segmentation for automatic detection of man-made objects from aerial and satellite images, *Photogramm. Eng. Remote. Sens.*, vol. 75, no. 6, pp. 667–677, 2009.

[17] P. V. C. Hough, Method and means for recognizing complex patterns, US Patent 3069654, 1962-12-18.

[18] N. K. Moghadam, M. R. Delavar, and P. Hanachee, Automatic urban illegal building detection using multi-temporal satellite images and geospatial information systems, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XL-1/W5, pp. 387–393, 2015.

[19] Y. Xu, L. Wu, Z. Xie, and Z. Chen, Building extraction in very high resolution remote sensing imagery using deep learning and guided filters, *Remote. Sens.*, vol. 10, no. 1, p. 144, 2018.

[20] W. Liu, S. Zhang, and L. Zhou, High-precision snore detection method based on deep learning, in *Proc. 5th Int. Conf. Mechatronics and Computer Technology Engineering* (*MCTE 2022*), Chongqing, China, 2022, pp. 1492–1497.

[21] W. Liu, S. Zhang, and L. Zhou, Snoring detection method in sleep based on MBAM-ResNet, in *Proc. 5th Int. Conf. Computer Information Science and Application Technology* (*CISAT 2022*). Chongqing, China, 2022, pp. 547–551.

[22] L. Wang, Y. Liu, P. Du, Z. Ding, Y. Liao, Q. Qi, B. Chen, and S. Liu, Object-aware distillation pyramid for open-vocabulary object detection, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 11186–11196.

[23] W. Liu, S. Zhang, L. Zhou, M. Xu, and Z. Ren, High-precision automatic detection method of illegal construction object images in complex scenes, *J. Electron. Imag.*, vol. 32, no. 3, p. 031803, 2022.

[24] S. Zhang, X. Wang, J. Wang, J. Pang, C. Lyu, W. Zhang, P. Luo, and K. Chen, Dense distinct query for end-to-end object detection, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 7329–7338.

[25] L. Zhou, W. Liu, S. Zhang, N. Luo, and M. Xu, CRMNet: Development of a deep-learning-based anchor-free detection method for illegal building objects, *Int. J. Patt. Recogn. Artif. Intell.*, vol. 37, no. 6, p. 2352007, 2023.

[26] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: A survey, *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, 2018.

[27] Z. Wang, Y. Li, X. Chen, S. N. Lim, A. Torralba, H. Zhao, and S. Wang, Detecting everything in the open world: Towards universal object detection, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 11433–11443.

[28] A. M. Roy, R. Bose, and J. Bhaduri, A fast accurate fine-grain object detection model based on YOLOv4 deep neural network, *Neural Comput. Appl.*, vol. 34, no. 5, pp. 3895–3921, 2022.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580–587.

[30] R. Girshick, Fast R-CNN, in *Proc. 2015 IEEE Int. Conf. Computer Vision* (*ICCV*), Santiago, Chile, 2015, pp. 1440–1448.

[31] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[32] A. Rahimzadeganasl and E. Sertel, Automatic building detection based on CIE LUV color space using very high resolution Pleiades images, in *Proc. 2017 25th Signal Processing and Communications Applications Conference* (*SIU*), Antalya, Turkey, 2017, pp. 1–4.

[33] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, Focal loss for dense object detection, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 2999–3007.

[34] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, YOLOX: Exceeding YOLO series in 2021, arXiv preprint arXiv: 2107.08430, 2021.

[35] X. Zhou, D. Wang, and P. Krähenbühl, Objects as points, arXiv preprint arXiv: 1904.07850, 2019.

[36] M. Tan, R. Pang, and Q. V. Le, EfficientDet: Scalable and efficient object detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 10778–10787.

[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 779–788.

[38] J. Redmon and A. Farhadi, YOLO9000: Better, faster, stronger, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 6517–6525.

[39] J. Redmon and A. Farhadi, YOLOv3: An incremental improvement, arXiv preprint arXiv: 1804.02767, 2018.

[40] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, YOLOv4: Optimal speed and accuracy of object detection, arXiv preprint arXiv: 2004.10934, 2020.

[41] J. Wang, N. Wang, L. Li, and Z. Ren, Real-time behavior detection and judgment of egg breeders based on YOLO

v3, *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5471–5481, 2020.

[42] I. Martinez-Alpiste, G. Golcarenarenji, Q. Wang, and J. M. Alcaraz-Calero, Search and rescue operation using UAVs: A case study, *Expert Syst. Appl.*, vol. 178, p. 114937, 2021.

[43] M. Choudhary, V. Tiwari, and V. Uduthalapally, Iris presentation attack detection based on best-*k* feature selection from YOLO inspired RoI, *Neural Comput. Appl.*, vol. 33, no. 11, pp. 5609–5629, 2021.

[44] V. Ostankovich and I. Afanasyev, Illegal buildings detection from satellite images using GoogLeNet and cadastral map, in *Proc. 2018 Int. Conf. Intelligent Systems* (*IS*), Funchal, Portugal, 2018, pp. 616–623.

[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 2261–2269.

[46] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, SSD: Single shot multibox detector, in *Proc. 14th European Conf. Computer Vision* (*CVPR*), Amsterdam, the Netherlands, 2016, pp. 21–37.

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

[48] C. Benedek, X. Descombes, and J. Zerubia, Building detection in a single remotely sensed image with a point process of rectangles, in *Proc. 2010 20th Int. Conf. Pattern Recognition*, Istanbul, Turkey, 2010, pp. 1417–1420.

[49] D. Chen, S. Shang, and C. Wu, Shadow-based building detection and segmentation in high-resolution remote sensing image, *J. Multimed.*, vol. 9, no. 1, pp. 181–188, 2014.

[50] S. Xu, G. Vosselman, and S. O. Elberink, Detection and classification of changes in buildings from airborne laser scanning data, *Remote. Sens.*, vol. 7, no. 12, pp. 17051–17076, 2015.

[51] N. Khalilimoghadama, M. R. Delavar, and P. Hanachi, Performance evaluation of three different high resolution satellite images in semi-automatic urban illegal building detection, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W7, pp. 505–514, 2017.

[52] P. D. Felice, Ranking of illegal buildings close to rivers: A proposal, its implementation and preliminary validation, *ISPRS Int. J. Geo Inf.*, vol. 8, no. 11, p. 510, 2019.

[53] D. Zhu and J. Fan, IBMDCH: Illegal building monitoring in digital city based on HPC, in *Proc. Geoinformatics 2008 and Joint Conf. GIS and Built Environment*: *Monitoring and Assessment of Natural Resources and Environments*, Guangzhou, China, 2008, pp. 425–436.

[54] M. Bouziani, K. Goïta, and D. C. He, Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge, *ISPRS J. Photogramm. Remote. Sens.*, vol. 65, no. 1, pp. 143–153, 2010.

[55] N. Moghadam, M. Delavar, and P. Hanachi, Semi-automatic illegal building detection in urban areas using satellite images and a pixel-based fuzzy XOR operator method, *Geospatial Engineering Journal*, vol. 7, no. 3, pp. 105–116, 2016.

[56] O. Benarchid, N. Raissouni, E. A. Samir, A. El Abbous, A. Azyat, N. B. Achhab, M. Lahraoua, and C. Asaad, Building extraction using object-based classification and shadow information in very high resolution multispectral images, a case study: Tetuan, Morocco, *Canadian Journal on Image Processing and Computer Vision*, vol. 4, no. 1, pp. 1–8, 2013.

[57] F. Dornaika, A. Moujahid, Y. El Merabet, and Y. Ruichek, A comparative study of image segmentation algorithms and descriptors for building detection, in *Handbook of Neural Computation*, P. Samui, S. Sekhar, and V. E. Balas, eds. New York, NY, USA: Academic Press, 2017, pp. 591–606.

[58] D. Comaniciu and P. Meer, Mean shift: A robust approach toward feature space analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[59] Q. Tan, Q. Wei, and F. Liang, Building extraction from VHR multi-spectral images using rule-based object-oriented method: A case study, in *Proc. 2010 IEEE Int. Geoscience and Remote Sensing Symp.*, Honolulu, HI, USA, 2010, pp. 2754–2756.

[60] X. F. Chen, X. G. Zhang, R. K. He, and Y. Wang, An algorithm to detect illegal buildings using color transferring and texture difference, in *Proc. 2018 IEEE 9th Int. Conf. Software Engineering and Service Science* (*ICSESS*), Beijing, China, 2018, pp. 545–550.

[61] L. An and B. Guo, Outdoor illegal construction identification algorithm based on 3D point cloud segmentation, *IOP Conf. Ser.*: *Mater. Sci. Eng.*, vol. 322, p. 052013, 2018.

[62] H. Wang, X. Ning, H. Zhang, Y. Liu, and F. Yu, Urban boundary extraction and urban sprawl measurement using high-resolution remote sensing images: A case study of China's provincial, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLII-3, pp. 1713–1719, 2018.

[63] C. Zhang and N. Liu, Illegal construction detection of road based on unmanned aerial vehicle vision, http://en.cnki.com.cn/Article_en/CJFDTotal-WJFZ201807030.htm, 2018.

[64] T. Ishii, E. Simo-Serra, S. Iizuka, Y. Mochizuki, A. Sugimoto, H. Ishikawa, and R. Nakamura, Detection by classification of buildings in multispectral satellite imagery, in *Proc. 2016 23rd Int. Conf. Pattern Recognition* (*ICPR*), Cancun, Mexico, 2016, pp. 3344–3349.

[65] G. Prathap and I. Afanasyev, Deep learning approach for building detection in satellite multispectral imagery, in *Proc. 2018 Int. Conf. Intelligent Systems* (*IS*), Funchal, Portugal, 2018, pp. 461–465.

[66] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, Building detection in very high resolution multispectral data with deep learning features, in *Proc. 2015 IEEE Int. Geoscience and Remote Sensing Symp.* (*IGARSS*), Milan, Italy, 2015, pp. 1873–1876.

[67] L. Martijn, S. Meijerink, and J. Ezeiza, Detect illegal buildings based on LiDAR point cloud data, https://event.cwi.nl/lsde/2019/results/group02.pdf, 2023.

[68] Y. Liu, Y. Sun, S. Tao, M. Wang, Q. Shen, and J. Huang, Discovering potential illegal construction within building roofs from UAV images using semantic segmentation and object-based change detection, *Photogramm. Eng. Remote. Sens.*, vol. 87, no. 4, pp. 263–271, 2021.

[69] Z. Liang, P. Deng, F. Jiang, S. Sheng, R. Wei, and G. Xie, The application of illegal building detection from VHR UAV remote sensing images based on convolutional neural network, *Bulletin of Surveying and Mapping*, doi: 10.13474/J.CNKI.11-2246.2021.0120

[70] X. Li, L. Fu, Y. Fan, and C. Dong, Building recognition based on improved faster R-CNN in high point monitoring image, in *Proc. 2021 33rd Chinese Control and Decision Conference* (*CCDC*), Kunming, China, 2021, pp. 1803–1807.

[71] K. He, X. Zhang, S. Ren, and J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[72] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, Path aggregation network for instance segmentation, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8759–8768.

[73] A. Howard, M. Sandler, B. Chen, W. Wang, L. C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, et al., Searching for MobileNetV3, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision* (ICCV), Seoul, Republic of Korea, 2019, pp. 1314–1324.

[74] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv: 1704.04861, 2017.

[75] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision Workshops* (*ICCVW*), Montreal, Canada, 2021, pp. 2778–2788.

Wenjin Liu received the BS degree from Wenzhou University, Zhejiang, China in 2020, and the MS degree from Capital Normal University, Beijing, China in 2023. He is currently pursuing the PhD degree at School of Cyberspace Security (School of Cryptology), Hainan University, China. His research interests include computer vision, image processing, artificial intelligence, LLM, and deep learning.

Lijuan Zhou received the BS degree in computer software from Heilongjiang University, China in 1991, the MS degree in computer science from Harbin University of Science and Technology, China in 1998, and the PhD degree in computer science from Harbin Engineering University, China in 2004. She is a professor at School of Cyberspace Security (School of Cryptology), Hainan University, China. Her main research interests include intelligence analysis, data mining and data analysis, business intelligence, and artificial neural network.

Min Xu received the PhD degree from Renmin University of China, Beijing, China in 2012. She is currently an associate professor at School of Information Engineering, Capital Normal University, Beijing, China. She has authored more than 30 scientific papers in peer-reviewed journals and conferences. Her research interests include computer vision, pattern recognition, and machine learning.

Shudong Zhang received the BS degree in computer science from Beijing Institute of Technology, Beijing, China in 1993, the MS degree from China Academy of Engineering Physics, Beijing, China in 1996, and the PhD degree in computer science from Beijing Institute of Technology, Beijing, China in 2005. He worked as a postdoctoral researcher at the Institute of Software, Chinese Academy of Science from 2005 to 2007. He was a professor at the Information Engineering College, Capital Normal University, China from 2008 to 2022. He is a professor at School of Cyberspace Security (School of Cryptology), Hainan University, China. His main research interests include computer architecture, computer system software, high performance computing, virtualization, and cloud computing.

Ning Luo received the BS and MS degrees in computer science from Tsinghua University, Beijing, China in 1995 and 1997, respectively, and the PhD degree from University of International Business and Economics, Beijing, China in 2015. He was an associate researcher at the Institute of Software, Chinese Academy of Science, China. Now he is an associate professor at School of Cyberspace Security (School of Cryptology), Hainan University, China. His main research interests include service computing, high performance computing, cloud computing, and computer system software.