Solid Waste Detection in Cities Using Remote Sensing Imagery Based on a Location-Guided Key Point Network With Multiple Enhancements

Huifang Li, *Member, IEEE*, Chao Hu, Xinrun Zhong, Chao Zeng, and Huanfeng Shen, *Senior Member, IEEE*

***Abstract*—Solid waste is a widespread problem that is having a negative effect on the global environment. Owing to the ability of macroscopic observation, it is reasonable to believe that remote sensing could be an effective way to realize the detection and monitoring of solid waste. Solid waste is usually a mixture of various materials, with a randomly scattered distribution, which brings great difficulty to precise detection. In this article, we propose a deep learning network for solid waste detection in urban areas, aiming to realize the fast and automatic extraction of solid waste from the complicated and large-scale urban background. A novel dataset for solid waste detection was constructed by collecting 3192 images from Google Earth (with a resolution from 0.13 to 0.52 m), and then a location-guided key point network with multiple enhancements (LKN-ME) is proposed to perform the urban solid waste detection task. The LKN-ME method uses corner pooling and central convolution to capture the key points of an object. The location guidance is realized through constraining the key point locations situated of the annotated bounding box of an object. Multiple enhancements, including data mosaicing, an attention enhancement, and path aggregation, are integrated to improve the detection accuracy. The results show that the LKN-ME method can achieve a state-of-the-art $\mathbf{AR_{100}}$(the average recall computed over 100 detections per image) of 71.8% and an average precision of 44.0% for the DSWD dataset, outperforming the classic object detection methods in solving the solid waste detection problem.**

***Index Terms*—Location-guided key point network, multiple enhancements, remote sensing, solid waste detection.**

I. Introduction

With the development of urbanization, solid waste pollution has become a significant environmental problem, and has even been considered as an “icon of the Anthropocene” that cannot be ignored. Solid waste refers to those things that have lost their use value or have been abandoned by humans during the process of industrial production, daily life, or other activities. Unordered and exposed piles of solid waste can lead to material deterioration, methane, and carbon dioxide emission, and leachate land contamination, threatening the living environment of humans [1], [2]. Therefore, the monitoring and management of solid waste is essential for cities, where more than half of the world’s population live [3]. In Europe, the extremely rapid growth of industrial activity in the twentieth century has resulted in a dramatic increase in solid waste volume, and has led to the creation of numerous landfill sites [4], [5]. Along with the progress of human society, many countries and organizations have tried to establish sustainable systems to prevent or reduce the adverse effects of waste processing and disposal on the environment [1]. The United Nations made 17 sustainable development goals to promote prosperity while also protecting the planet. The achievements of many goals, such as clean water and sanitation (goal 6), sustainable cities and communities (goal 11), and responsible consumption and production (goal 12), require the support of rational and tight management of solid waste. Nowadays, both developed and developing countries are experiencing rapid changes in every aspect, including industrial production and social and economic development, during which various amounts of solid waste are being constantly produced. There is therefore an urgent need to search for and locate urban solid waste, to realize appropriate environmentally friendly management [1], [6]. However, the distribution of solid waste is usually random and scattered in cities, causing great difficulties for its detection and location. An efficient and accurate solid waste detection method will be very important for the management of the urban environment.

Remote sensing data are widely used for the monitoring and planning of the urban environment, which can be attributed to their advantages of the large scale and the continuous periodic observation. Remote sensing data have also been used in waste monitoring and management since the 1990s, when aerial remote sensing technology was developed and large-scale photographs of land surfaces were able to be captured. It has been verified that the use of remote sensing data is a very effective and economical way to detect urban waste, as long as the spatial resolution of the data is high enough. A number of studies have been conducted to identify solid waste landfills and the areas around them using aerial remote sensing images [1], [4], [7]. In the beginning, simple visual interpretation was

Manuscript received 8 September 2022; revised 6 November 2022; accepted 22 November 2022. Date of publication 24 November 2022; date of current version 7 December 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB2102901 and in part by the National Natural Science Foundation of China under Grant 41871246. *(Correspongding author: Chao Zeng.)*

Huifang Li and Huanfeng Shen are with the School of Resource and Environmental Sciences and the Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: huifangli@whu.edu.cn; shenhf@whu.edu.cn).

Chao Hu, Xinrun Zhong, and Chao Zeng are with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China (e-mail: chaohu@whu.edu.cn; 2021202050019@whu.edu.cn; zengchao@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3224555

used to identify the solid waste landfills. Subsequently, pattern recognition methods based on spectral signatures were proposed to identify the polluted vegetation areas around the solid waste landfills and evaluate the environmental effects [1], [4], [7]. It can be seen that these previous works have mainly focused on the identification of waste landfills with large areas by the use of visual or simple machine interpretation methods, limited by the data resolution and interpretation ability. Scattered solid waste is rarely discussed, even though this waste occupies available land and poses a threat to the environment. Currently, with the support of high-resolution remote sensing imagery and advanced pattern recognition technology, it should be possible to carry out automatic and intelligent solid waste detection in cities. Therefore, in this article, our aim is to develop an automatic object detection network for solid waste in cities based on deep learning with high-resolution remote sensing images.

Deep learning [8] is a new branch of machine learning, which has shown strong information processing capabilities in computer vision, speech recognition, artificial intelligence and many other fields [9]. It has also been widely used to solve the problems, such as classification [10], [11], [12], segmentation [13], [14], object detection [15], [16], [17], [18], change detection [19], etc., obtaining state-of-the-art results. Solid waste can be considered as a kind of ensemble object in urban areas, with typical spectral and spatial characteristics in remote sensing images, so that it is possible to realize detection by constructing an object detection network. We discovered two distinguishable characteristics of solid waste after interpreting many remote sensing images: solid waste is usually a mixture of multiple substances with varying spectral features; and solid waste is usually dumped in piles with blurry boundaries. In order to fully understand the features of solid waste, we built a novel dataset for solid waste detection (the DSWD dataset) by collecting 3192 high-resolution remote sensing images from Google Earth. Multiple types of solid waste with varying spatial and spectral features are included in the DSWD dataset, based on which a network can be trained to learn the features of solid waste in both shallow and deep layers. A location-guided key point network with multiple enhancements (LKN-ME) is also proposed to perform the task of urban solid waste detection. Multiple enhancements, including data enhancement, feature enhancement, and path enhancement, are also integrated to ensure the detection accuracy of the network.

The main contributions of this article can be summarized as follows.

1) A remote sensing image dataset comprising two categories of solid wastes (i.e., black wastes and white wastes) is constructed, in which the scene is more complex and numbers of targets are included compared with the current dataset.
2) A new key point network with multiple enhancements named LKN-ME is proposed for the solid waste detection, and higher detection accuracy is achieved than six state-of-the-art one-stage and two-stage networks.

The rest of this article is organized as follows. Section II reviews the deep learning-based object detection networks and datasets. The details of the DSWD dataset are provided in Section III, and the proposed LKN-ME method is described in Section IV. Section V reports the details of the experiments, as well as the comparison with the existing methods, and the results of the ablation experiments are also presented. Finally, the article is concluded in Section VI.

II. Related Work

A. Deep Learning Based Object Detection Networks

Deep learning was first used by Girshick et al. [20] to solve the object detection problem, and ever since then, many state-of-the-art deep learning object detection networks have been proposed. These networks can be divided into two categories, according to the requirement of the region proposal: one-stage methods and two-stage methods. The one-stage methods can obtain the locations and the categories of the objects simultaneously, without region proposal. Typical examples of the one-stage methods are the single shot multibox detector (SSD) [21], you only look once (YOLO) [22], [23], [24], [25], CornerNet [26], CenterNet [27], and PolarMask [28]. The two-stage methods first obtain numerous region proposals and then identify the precise object locations and categories through fine-tuning and classification. Such methods include the region based convolutional neural network (R-CNN) [20], fast R-CNN [29], faster R-CNN [30], the feature pyramid network (FPN) [31], and mask R-CNN [32]. For both the one-stage and two-stage approaches, the networks for predicting the target boundaries involve two main strategies: anchor-based methods and anchor-free methods. The networks based on anchors, such as faster R-CNN [30], have guidance about the size of the bounding boxes, to achieve better detection effects. The anchor-free methods [20], [25], [26], [27], [28], [29], [30], such as CornerNet [26], omit the design of the size of the bounding boxes. As a result, they are much faster than the anchor-based methods, and are suitable for objects with large scale variations and no definite aspect ratio. In view of these advantages, the anchor-free approach was used in the latest YOLOX [33] method, achieving a first-class detection effect.

B. Object Detection in Remote Sensing Images

Most of the current object detection networks were originally proposed for detecting individual objects in front-viewed natural images covering simple scenes. However, the detection precision is often not satisfactory if these networks are directly used in remote sensing images, as remote sensing images have significantly different characteristics, compared with natural images. Four significant characteristics of objects in remote sensing images can be summarized as follows: various scales; various length-width ratios; blurred boundaries; and complex backgrounds. Solid waste often has a fragmentary appearance, and there are also some objects with characteristics that are similar to those of solid waste, such as groves, parking lots with lots of cars, waves on water surfaces, algae, and windows of tall buildings. These similar objects make the background more complex.

Faster R-CNN, SSD, and YOLO have been shown to be effective in aerial image object detection. However, these methods cannot effectively solve the problems of remote sensing images. Some networks have been proposed to solve the problems of the multiple scales and oriented bounding boxes in aerial images. For example, the rotation-invariant convolutional neural network [34] adds a rotation-invariant layer to the R-CNN architecture to deal with the problem of object rotation variations; the rotation-invariant and Fisher discriminative CNN [35] uses an oriented region proposal network to replace the region proposal network in faster R-CNN; and Fu et al. [36] introduced an oriented region proposal network and orientation region of interest to replace the FPN region proposal network and region of interest. Besides, a shape robust anchor-free network [18] has been proposed for the detection of garbage dumps by generating the targets' bounding boxes, indicating the effects of the anchor-free networks.

C. Datasets for Object Detection

The deep learning object detection task must be driven by datasets. Image datasets composed of large amounts of natural images captured by cameras or cell phones have been constructed for the general object detection tasks in daily life, including the Pascal visual object classes dataset [37] and the MS COCO dataset [38]. Similarly, many datasets composed of remote sensing images have also been constructed to support remote sensing object detection, such as NWPU VHR-10 [39], the vehicle detection in aerial imagery dataset [40], DOTA [41], DIOR [42], and the large-scale dataset for instance segmentation in aerial images [43]. These remote sensing datasets are mainly used for the detection of the typical objects and landscapes in cities, including vehicles, squares, airports, parks, etc. Compared with natural image datasets, remote sensing image datasets can support large object detection in a city from the top view, but the object labels can be very difficult to assign, as the background is complicated and the shapes and scales of the objects vary a lot. For a specific object detection task, a specific dataset is required to drive the learning process. A garbage dumps dataset has been built for the solid waste detection [18], but the data volume is not big enough and most of the scenes are too simple to reflect the true spatial distributions of solid wastes in cities. Therefore, in this article, a new image dataset containing complex scenes and multi-types of garbage for the solid waste detection task was built.

III. Image Dataset for Solid Waste Detection

Solid waste shows very complex characteristics in many aspects, including color, texture, and shape, and the distribution of solid waste in remote sensing images is often very sparse. These characteristics result in the production of solid waste datasets being very difficult. As a result, there are currently no public remote sensing image datasets for solid waste detection, which limits the application of deep learning in this field. In order to solve these problems, we built the novel DSWD dataset.

A. Construction of the DSWD Dataset

The complex features of solid waste brought a lot of problems to the annotation of the DSWD dataset. First, solid waste is usually a mixture of multiple substances which have different colors, textures, and shapes, such as metals, plastics, and rubble. In remote sensing images, solid waste piles are often irregular, and the color of solid waste varies a lot, which leads to huge differences between examples of solid waste. It is therefore a big challenge to make full use of the common features of solid waste and annotate solid waste appropriately. Second, in remote sensing images, solid waste is often dumped in piles with blurry boundaries. There can also be many scattered solid waste objects around the main solid waste pile. It is therefore difficult to judge whether solid waste is one object or more.

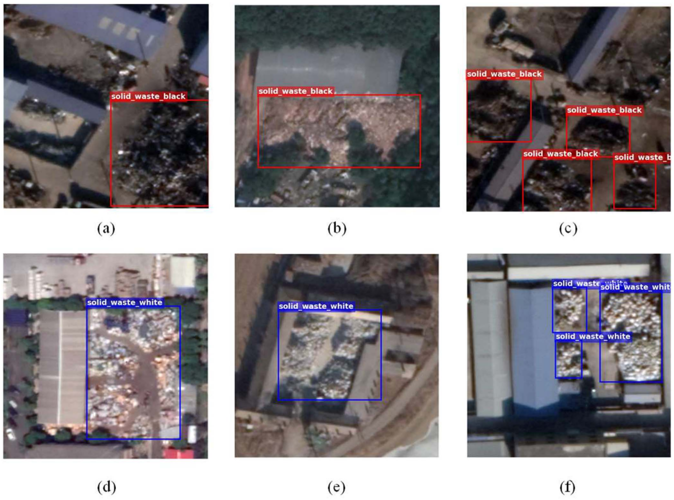


Fig. 1. Some examples of solid waste from the DSWD dataset. The first row shows some examples of black solid waste, and the second row shows some examples of white solid waste.

Fig. 1 shows some examples of solid waste in remote sensing images, where it can be seen that the solid waste is irregularly piled and fuzzy, with uncertain shape and color. There are also many scattered solid waste objects around the main solid waste piles. To deal with the issues mentioned above, we developed corresponding annotation standards.

First, the solid waste piles are usually circular or oval, for which a horizontal bounding box is appropriate. A horizontal bounding box can be described as $(x_c,\ y_c\ ,w,\ h)$, wherex_c ,$y_c$denote the center of the bounding box and w, h denote the width and height of the bounding box. Second, irregular stacking is used as the main basis for judgment, and the surrounding environment is used as an auxiliary basis for judgment, to determine the locations of solid waste. In order to guide the detection results, we classify solid waste into two main classes—white and black—according to its main color and brightness. Third, when annotating the solid waste, we ignore the small scattered solid waste objects and the small intervals between the solid waste, because this can affect the judgment of the bounding boxes. Fig. 1 displays some examples of solid waste annotation. Irregular stacking is the main judgment for solid waste, as shown in Fig. 1(a) and (b). When the features of adjacent solid waste objects are different, we judge that the solid waste is made up of multiple independent objects, as shown in Fig. 1(c). When the features of adjacent solid waste objects are similar, we judge

TABLE I
DETAILS OF THE DSWD DATASET

Details	
Data source	Google Earth
Number of images	3192
Images with solid waste	2690
Images with objects similar to solid waste	502
Image size	512×512
Resolution	0.13–0.52 m
Region	32 major cities in China*

*Note: the 32 major cities are Beijing, Shanghai, Tianjin, Wuhan, Hangzhou, Nanjing, Xi'an, Chongqing, Ningbo, Guangzhou, Shenzhen, Chengdu, Changsha, Hefei, Fuzhou, Xiamen, Datong, Maanshan, Panzhihua, Zhenjiang, Changzhou, Jiaxing, Wuxi, Suzhou, Shijiazhuang, Hong Kong, Macau, Dalian, Shenyang, Harbin, Taiyuan, and Nanning.

that the solid waste is a single object, as shown in Fig. 1(d) and (e). When solid waste has multiple colors, the solid waste is classified according to its main color, as shown in Fig. 1(f). In totally, the first row in Fig. 1 shows images containing black solid waste, and the second row shows images containing white solid waste.

According to the above standards, we built the novel DSWD dataset by collecting high-resolution remote sensing images, for which the details are given in Table I. In total, 3192 images with the size of 512 × 512 are included in the DSWD dataset, in which 2590 images contain solid waste objects and 502 contain objects that are similar to solid waste. We added objects that are similar to solid waste into the DSWD dataset for the reason that some objects in remote sensing images have similar features to solid waste, such as groves, parking lots with lots of cars, waves on water surfaces, algae, and windows of tall buildings. All the images were collected from Google Earth, with resolutions from 0.13 to 0.52 m. The data quality in Google Earth varies, so we mainly chose areas with a good quality to collect the solid waste data. To increase the generalization of the DSWD dataset, we collected images from 32 major cities in China. Fig. 1 shows some examples from the DSWD dataset.

B. Dataset Splits

We randomly selected 70% of the images as the training set, 10% as the validation set, and 20% as the test set. We therefore obtained a training set with 2233 images, a validation set with 320 images, and a test set with 639 images. The training set was used to train the network. The validation set was used to verify the network performance in training. The test set was used to test the network effect after training and to perform unbiased evaluation of a trained model.

IV. METHOD

In this article, we developed the LKN-MEs for solid waste detection. The LKN-ME method uses corner pooling and central convolution to capture the key points of the object bounding boxes. The location guidance is designed to give the network a rough location by taking advantage of the cover area of the annotated bounding boxes of the objects in the dataset. Multiple enhancements, including data mosaicing, an attention enhancement, and path aggregation, are integrated to enhance the data and features in multiple layers and scales. Fig. 2(a) shows the overall architecture of the proposed LKN-ME method.

A. Network Architecture

A flexible bounding box is required to detect solid waste precisely because of the irregular shape of solid waste. The key point network is anchor-free and can provide candidate bounding boxes efficiently and with a high quality. Thus, a key point network is taken as the baseline of the proposed network. Both corner points and center points are detected by the proposed network. The corner pooling module is used to calculate the top-left corners and the bottom-right corners of the bounding boxes, and the central convolution module composed of two convolutional layers and a sigmoid activation function is used to calculate the center points of the bounding boxes. The extreme points of the objects at the top, left, bottom, and right directions are expected to be located by the corner pooling module, which are actually out of the objects. This key point searching mechanism makes the detector pay more attention to the range and outside shape of the object rather than the complicated features of the object itself, which benefits solid waste detection.

An hourglass network [44] is taken as the backbone of the proposed network, from which three heatmaps are produced, i.e., the top-left corner heatmap, the center point heatmap, and the bottom-right corner heatmap. The pixel values of the heatmaps are confidence scores representing the possibilities of the key points of the objects. In addition, the embeddings for the corners, the offsets for the corners, and the center points are also predicted. The embeddings are used to identify whether a top-left corner and a bottom-right corner are from the same object. The offsets are applied to remap the heatmaps to the size of the input image. In order to generate the object bounding boxes from the heatmaps, the top k key points are selected according to their scores. Three standards are then used to match the top-left corners and the corresponding bottom-right corners: the x and $y$coordinates of the bottom-right corner should be larger than those of the top-left corner; the center point exists around the midpoint of the top-left corner and the bottom-right corner; and the distance of the embeddings of the top-left corner and the bottom-right corner is less than a threshold. If the top-left corner and the bottom-right corner meet these three standards, they are matched as a predicted bounding box of an object.

B. Location Guidance

Bounding boxes are used to annotate the location and category of the objects in the object detection task. However, the location information is not fully exploited in most networks. We propose to use the coverage area of the labeled bounding box to achieve the location guidance in the network.

According to the labels of a bounding box in the dataset, a binary mask is generated as the guidance, in which the pixels within the bounding box are annotated as 1, while those outside

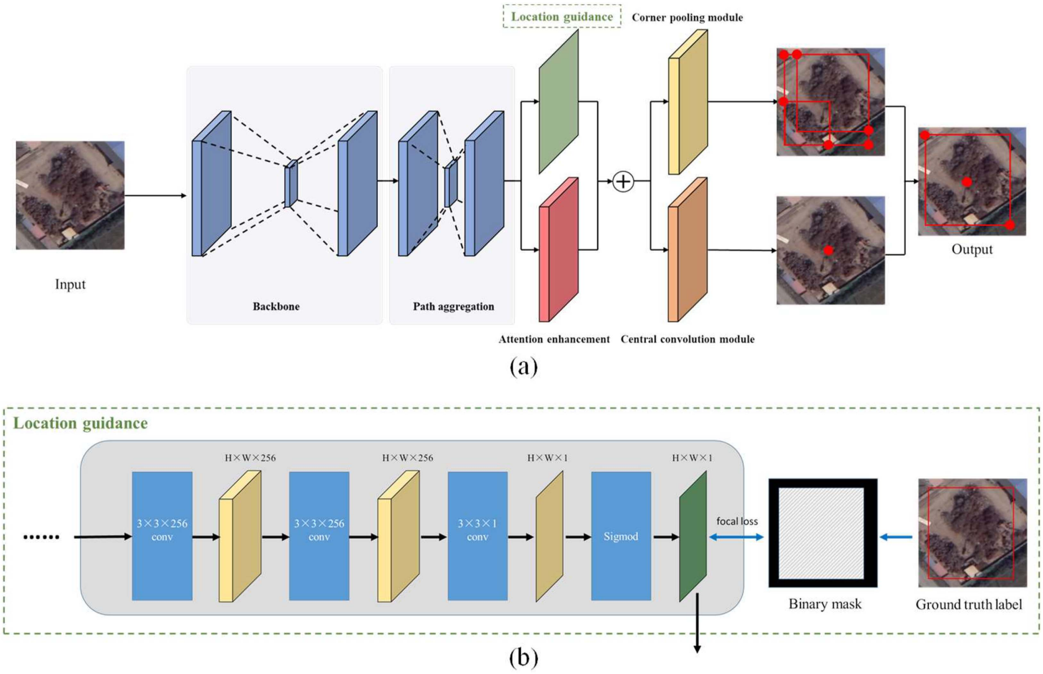


Fig. 2. Proposed LKN-ME. (a) Architecture of the proposed LKN-ME method. (b) Structure of the location guidance module.

of the bounding box are annotated as 0. In this way, a coarse localization criterion is provided by the binary mask guidance, to guide the training of the LKN-ME method. Fig. 2(b) shows the structure of the location guidance, which is an independent branch in the proposed network. Three convolutional layers and a sigmoid function are connected in the location guidance to obtain the coarse localization heatmap of the objects. The kernel sizes of all convolution layers are 3×3. Focal loss is used to calculate the difference between the heatmap and the ground-truth map. Since the heatmap depicts the coarse localization of the objects, element-wise addition is used to add the heatmap back to the LKN-ME network, to make full use of the location information.

C. Multiple Enhancements

1) Mosaic Data Augmentation: It is difficult to manually select a large number of images containing solid waste from high-resolution urban remote sensing images, which greatly limits the amount of data in the DSWD dataset and leads to low robustness of the LKN-ME method. Therefore, mosaic data augmentation [25] was used to expand the solid waste dataset.

Data mosaicing is a method that can be used to mix four images to generate an image that contains the scenes of all four images. Before an $n \times n$image is input into the network, three images are randomly selected from the training set. The four images can then be spliced into an $2n \times 2n$image. We then randomly pick a point from the central $n \times n$ area of the composite image and take this point as the center point to crop an $n \times n$ image, as illustrated in Fig. 3(a). The $n \times n$ image is then input into the network for training. The robustness of the LKN-ME method can be significantly improved because of the data mosaicing. There are two reasons for this: there are different combinations of scenes in an image, which enriches the scenario of the training data; and since an object can be sliced by the mosaicing, a part of the object rather than the whole object is input into the network.

2) Path Aggregation: The hourglass network can obtain features $[P_1, P_2 ,P_3, P_4 ,P_5]$ with five scales. However, only feature P_5 is used to obtain the final result, while the features with other scales are not fully used, Remote sensing images are often of multiple resolutions and the object scales are usually varied. Thus, a network for solid waste detection should be capable of dealing with multiresolution remote sensing data and mining multiscale features. Therefore, a path aggregation module composed of a top-down path and a bottom-up path is proposed in the LKN-ME method, to explore more multi-scale features and support solid waste detection.

Fig. 3(b) displays the structure of the path aggregation. For the top-down path, each feature $P_i (i = 1, 2, 3, 4, 5)$ goes through a 3×3 convolutional layer to obtain feature $T_i (i = 1, 2, 3, 4, 5)$. Each feature $T_j (j = 1, 2, 3, 4, 5)$ then goes through a 3×3 convolutional layer with stride 2 to reduce the spatial scale. Finally, each feature map ${T'}_{j+1}$ and the down-sampled map are added using element-wise addition to obtain the features $[{P'}_1,\ {P'}_2,\ {P'}_3,\ {P'}_4,\ {P'}_5]$, where ${P'}_i$ denotes the feature generated by the top-down path. For the bottom-up path, each feature ${P'}_i (i = 1, 2, 3, 4, 5)$ goes through a 3×3 convolutional layer to obtain feature ${T'}_i (i = 1, 2, 3, 4, 5)$. Each feature ${T'}_j (i = 1, 2, 3, 4, 5)$ then goes through an up-sampling layer to increase the spatial scale. Finally, each feature map ${T'}_{j+1}$ and the up-sampled map are added using element-wise addition to obtain the features $[{P''}_1, {P''}_2,\ {P''}_3,\ {P''}_4,\ {P''}_5]$, where ${P''}_i$ denotes the

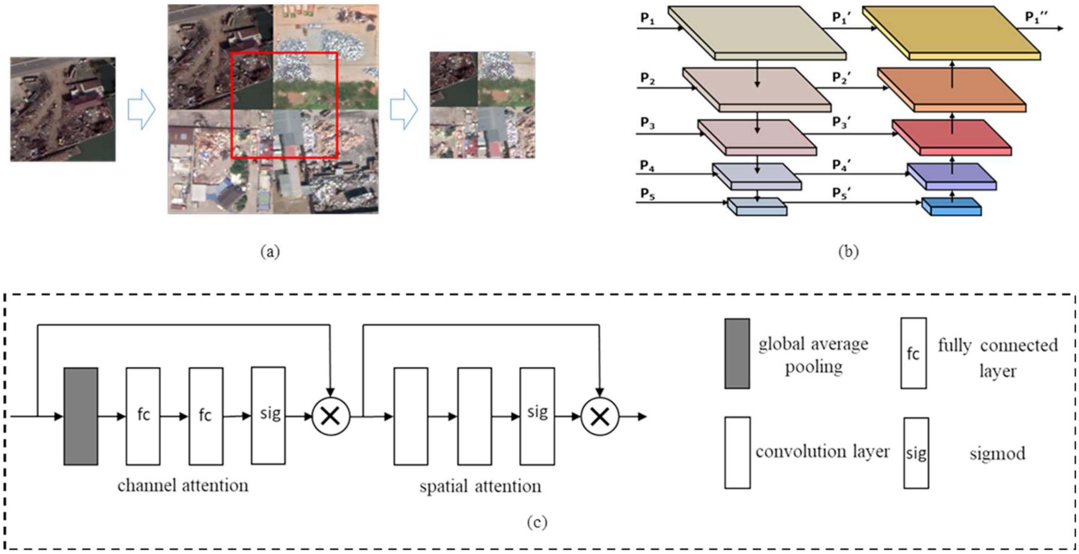


Fig. 3. Multiple enhancements. (a) Process of mosaic data augmentation. (b) Structure of the path aggregation module. (c) Structure of the attention enhancement module.

feature generated by the bottom-up path. P''_1 is the output of the path aggregation.

3) Attention Enhancement: An attention enhancement is used to make the network focus on the important features of the objects of interest and suppress unnecessary ones. In the proposed approach, the attention enhancement is divided into spatial attention and channel attention enhancement. The structure of the attention enhancement is shown in Fig. 3(c).

To compute the channel attention efficiently, we first use global average pooling to obtain a $1 \times 1 \times C$ feature map. We then apply two fully connected layers and a sigmoid function after the last fully connected layer to produce a channel attention map $M_C(F) \in R^{1\times 1\times C}$. The channel attention enhancement focuses on which channel contains meaningful information about the objects. The spatial attention enhancement is used to obtain the inter-spatial relationship of the feature map, and it focuses on the location of the objects. To compute the spatial attention, two convolutional layers and a sigmoid function after the final convolutional layer are used to generate a two-dimensional spatial attention map $M_s(F) \in R^{W\times H}$.

D. Loss Function

According to the above description, the heatmaps of the key points are produced by the proposed network, as well as the offsets of the key points, the embeddings of the top-left corners and bottom-right corners, and the heatmaps of the coarse localization. Thus, the loss function of the proposed network consists of five parts, expressed as

$$L = L_{\rm det} + aL_{\rm pull} + bL_{\rm push} + cL_{\rm off} + dL_{\rm localization}$$

where L is the loss function of LKN-ME. $L_{\rm det}$ and $L_{\rm localization}$ are both a variant of focal loss, and $L_{\rm det}$ was introduced in CornerNet [26]. $L_{\rm pull}$ is used to minimize the distance between the top-left and bottom-right corner embeddings which belong to the same objects, and $L_{\rm push}$ is used to maximize the distance between the top-left and bottom-right corner embeddings which belong to different objects. $L_{\rm off}$ is the smooth L1 loss, which is applied at the ground-truth corner locations. a, b ,c, and d denote the weights of the corresponding losses, which are set to 0.1, 0.1, 0.1, and 0.1 in the proposed network.

V. Experiments

The experiments were conducted in PyTorch. An RTX 2080Ti (11 GB) GPU was used to accelerate the calculation. In addition to adopting standard data augmentation techniques, including random horizontal flipping, random cropping, random coloring, and random scaling, mosaicing was used for further data augmentation. The Adam optimizer [45] was used to optimize the full training loss, and a rectified linear unit was used as the activation function. The batch size of the network was set to 5 and the maximum number of epochs was set to 180. The learning rate was 2.5×10^{-4} for the first 150 epochs and then 2.5×10^{-5} for the last 30 epochs.

In the testing, we selected the top 70 top-left corners, center points, and bottom-right corners from the heatmaps to detect the bounding boxes. The score of each bounding box was the average of the key points of this bounding box. Soft-NMS [46] was used to remove the redundant bounding boxes. The threshold of soft-NMS was set to 0.5. We finally selected the top 100 bounding boxes, according to the scores of these boxes, as the final detection results. The annotation of each bounding box was solid waste rather than white solid waste or black solid waste. Multiscale testing with 0.6, 1, 1.2, 1.5, and 1.8 times the resolution of the input image was used to detect the objects.

A. Evaluation Metrics

In this article, the average recall (AR) computed over 100 detections per image (AR_{100}), the average precision (AP), and

TABLE II
PERFORMANCE OF THE PROPOSED LKN-ME METHOD AND THE OTHER STATE-OF-THE-ART METHODS

Models	Backbone	AR_{100}	AR_{small}	AR_{medium}	AR_{large}	AP	AP_{50}	AP_{small}	AP_{medium}	AP_{large}
Faster R-CNN	ResNet-50	60.6	8.0	53.3	64.8	36.2	66.3	7.9	27.6	41.0
FPN	ResNet-50	58.8	10.0	52.2	62.6	37.8	69.4	4.6	28.9	42.6
PANet	ResNet-50	59.4	20.0	53.2	63.0	38.3	69.7	7.4	30.1	42.7
YOLOv4	Darknet-53	53.5	52.0	54.2	53.2	35.3	63.6	**38.8**	36.2	35.4
CornerNet	Hourglass-52	65.1	48.0	51.9	71.9	26.5	43.6	1.5	18.5	32.6
CenterNet	Hourglass-52	70.2	47.0	63.5	73.8	40.7	65.7	27.2	30.8	45.9
LKN-ME	Hourglass-52	**71.8**	**53.0**	**64.8**	**75.5**	**44.0**	**70.3**	17.6	**37.0**	**48.2**

AP_{50} are used as the metrics for the solid waste detection, where AP is computed over the maximum 100 detections per image and the average of 10 different intersection over union (IoU) thresholds from 0.5 to 0.95 in 0.05 intervals. AP_{50} is computed over the maximum 100 detections per image with an IoU threshold of 0.5. AR_{100} is the AR computed over the maximum 100 detections per image and the 10 different IoU thresholds. As the purpose of solid waste detection is to identify as much solid waste as possible, AR_{100} is used as the main evaluation metric. In addition, different object scales of small (area $< 32^2$), medium ($32^2 <$ area $< 96^2$), and large (area $> 96^2$) are used to calculate the AP and AR, to evaluate the performance of the networks. All the AP and AR metrics are computed over the maximum 100 detections per image and the 10 different IoU thresholds.

B. Detection Results

The last row in Table II lists the results of the LKN-ME method obtained with the DSWD dataset. The LKN-ME method achieves 71.8% AR, 44.0% AP, and 70.3% AP_{50} on the DSWD test dataset, which is the best result among all seven detectors. The AP and AR values for the large solid waste objects are better than those for the medium and small solid waste objects. The reason is that there is more information in the large solid waste objects, so it is easier to obtain the centers of the large solid waste objects.

Fig. 4 shows some results of the LKN-ME method for the DSWD test dataset. The first row shows the detection of single solid waste objects, where the location of the solid waste detection is very accurate and the candidate bounding boxes contain the solid waste completely. In the second and third rows, there are multiple solid waste objects, and the solid waste in the images is basically all detected. However, the solid waste in the fourth-row scenes is fuzzy and mixed with various substances. The detection effect declines in these scenes, and some solid waste is not detected in the fourth-row images. The fifth row displays objects with similar properties to solid waste. The objects from left to right are algae, waves on the surface of water, a grove, a parking lot with lots of cars, and windows of tall buildings. None of these objects are mistakenly detected as solid waste. Generally speaking, the overall test results for the LKN-ME method are good for both the solid waste objects and the objects similar to solid waste, but the detection effect does decrease when the features of the solid waste are complex.

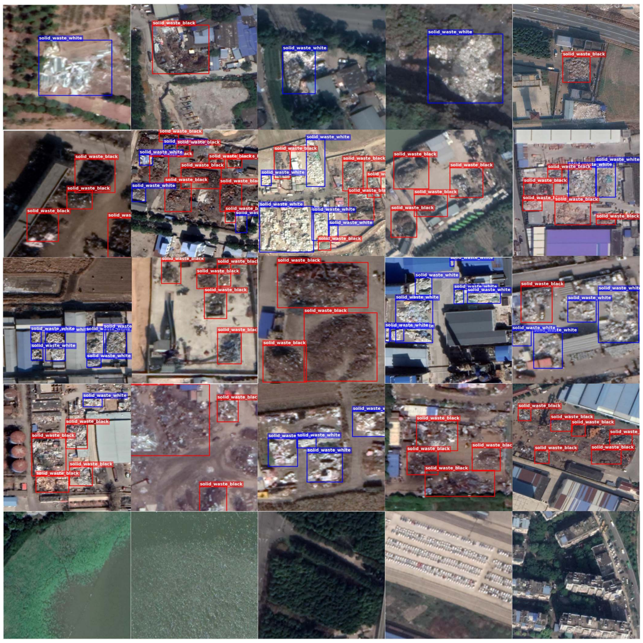


Fig. 4. Results of the proposed LKN-ME method with the DSWD test dataset. The first row shows images with single solid waste objects. The second to fourth row show images containing multiple solid waste objects. The fifth row shows images with objects that are similar to solid wastes.

From Table IV, it can be seen that the AP and AP_{50} of the results based on the dataset with two categories are almost

TABLE III
COMPARISONS OF THE TRAINING SPEEDS AND ACCURACIES WITH THE EXISTING KEY POINT NETWORKS

Models	AR$_{100}$ (%)	AP (%)	AP$_{50}$ (%)	Speed (s)
CornerNet	65.1	26.5	43.6	**1.5**
CenterNet	70.2	40.7	65.7	3.6
Key point network	69.9	40.9	66.7	**1.5**
LKN-ME	**71.8**	**44.0**	**70.3**	1.7

TABLE IV
COMPARISON OF DATASET WITH ONE CATEGORY AND TWO CATEGORIES FOR SOLID WASTE DETECTION

Dataset	**AR$_{100}$**	**AP**	**AP$_{50}$**
One category	66.7%	43.9%	70.3%
Two categories	**71.8%**	**44.0%**	**70.3%**

consistent with that with one category. However, the AR$_{100}$ is improved by 5.1%, changing from 66.7% to 71.8% in the results based on two categories. It strongly demonstrates that the dataset with two categories promotes the performance of the solid waste detection by using deep learning networks.

C. Comparison Experiments

1) Accuracy Comparison With the State-of-the-Art Detectors: Table II is a comparison of the proposed network with six state-of-the-art detectors on the DSWD test dataset, i.e., faster R-CNN, FPN, PANet, YOLO, CornerNet, and CenterNet, where faster R-CNN, FPN, and PANet are two-stage networks, and the other three detectors are one-stage networks.

In order to conduct peer-to-peer experiments between different networks, the backbones of all the networks were set to around 50 layers. Nine metrics were calculated for each method. LKN-ME obtains the best results in eight of the nine metrics, with an AP of 44.0% and an AR$_{100}$ of 71.8%, which outperforms the other detectors. Faster R-CNN, FPN, and PANet use anchors with three scales and three sizes to detect objects. These two-stage methods show a poor performance on the small targets, resulting in a low overall detection accuracy.YOLOv4 has advantages in small-object detection and obtains the best AP$_{\mathrm{small}}$ score of 38.8%. Because multi-scale fusion is adopted in YOLOv4, the accuracy of the small-scale prediction is higher. CornerNet obtains a high average maximum recall but a low AP on the DSWD dataset. The reason for this is that the complex boundaries of the solid waste objects mislead CornerNet to detect many false bounding boxes. CenterNet adds center point detection to CornerNet to strengthen the constraints on the key point matching. CenterNet achieves remarkable improvements, from 26.5% to 40.7% in AP, from 43.6% to 65.7% in AP$_{50}$, and from 65.1% to 70.2% in AR$_{100}$, which means that using deep learning to detect the key points of bounding boxes to achieve solid waste detection is effective. This method pays more attention to the boundaries of the objects, which weaken the feature differences inside the objects, and it more easily extract the unified features of solid waste.

Compared with CenterNet, the proposed LMN-KE method shows certain improvements in eight metrics, except for AP$_{\mathrm{small}}$. The AR$_{100}$ is increased from 70.2% to 71.8%, the AP is increased from 40.7% to 44.0%, and the AP$_{50}$ is increased from 65.7% to 70.3%. In addition, AR$_{\mathrm{small}}$ is increased by 6%, AR$_{\mathrm{medium}}$ is increased by 1.3%, AR$_{\mathrm{large}}$ is increased by 1.7%, AP$_{\mathrm{medium}}$ is increased by 6.2%, and AP$_{\mathrm{large}}$ is increased by 2.3%. However, AP$_{\mathrm{small}}$ is decreased by 9.6%. The detection accuracy of the LKN-ME method is reduced on small objects. This may be because the path aggregation fuses low-scale features, which blur the features of the small objects. As a result, LKN-ME detects too many small objects, resulting in a decrease in AP$_{\mathrm{small}}$.

2) Training Speed Comparisons: In the key point network, the corner pooling module and the central convolution module are used to calculate the corners and the center points of the bounding boxes. Table III gives a comparison between the proposed key point network and other key point networks. For solid waste detection, the training speed of CornerNet is fast but the detection accuracy is poor. Solid waste has a lot of edge information, which makes CornerNet detect many false bounding boxes. CenterNet is better able to detect solid waste but the training speed is slow. CenterNet detects the center points of the bounding boxes to decrease the false bounding boxes, which results in a huge improvement in detection accuracy. However, CenterNet uses corner pooling repeatedly, which is a slow calculation method when backpropagating. As a result, the training speed of CenterNet is slow.

The proposed key point network detects the center points of the bounding boxes to improve the detection accuracy, and decreases the use of corner pooling to improve the training speed. The proposed key point network achieves the detection accuracy of CenterNet and the training speed of CornerNet, which results in an improved overall performance.

The LKN-ME method adds the mosaic data augmentation, the attention enhancement, path aggregation, and location guidance to the key point network. As a result, the training speed increases by 0.2 s/iter, from 1.5 a/iter to 1.7 s/iter, the AR$_{100}$ is improved by 1.9%, from 69.9% to 71.8%, the AP is improved by 3.1%, from 40.9% to 44.0%, and the AP$_{50}$ is improved by 3.6%, from 66.7% to 70.3%. Therefore, the LKN-ME method obtains a superior detection accuracy, with only a slight increase in the training cost.

D. Ablation Study

The proposed LMN-KE method consists of four components, i.e., mosaic data augmentation, the attention enhancement, path aggregation, and location guidance. An ablation study was conducted to analyze the contribution of each individual component.

TABLE V
ABLATION STUDY (%) FOR THE MAJOR COMPONENTS OF THE LKN-ME METHOD

MDA	AM	PA	LG	AR$_{100}$	AP	AP$_{50}$
				69.9	40.9	66.7
√				70.8	42.9	67.8
√			√	**72.0**	43.1	69.0
	√			71.5	40.3	65.6
√	√			71.1	43.2	69.0
√	√	√		71.7	43.5	69.0
√	√		√	71.4	43.3	69.1
√	√	√	√	71.8	**44.0**	**70.3**

MDA denotes mosaic data augmentation, AM denotes attention enhancement, PA denotes path aggregation, and LG denotes location guidance.

The backbone of each experiment was Hourglass-52. We conducted the ablation study with a variety of combinations of the four components. The results are listed in Table V, where the first row is the result of the proposed key point network.

To demonstrate the importance of the mosaic data augmentation, the results of the network with and without mosaicing were compared. As shown in the first two rows in Table V, the mosaicing results in an improvement in AR$_{100}$ of 0.9%, from 69.9% to 70.8%, an improvement in AP of 2%, from 40.9% to 42.9%, and an improvement in AP$_{50}$ of 1.1%, from 66.7% to 67.8%. These results demonstrate that the mosaic data augmentation is an effective way to improve the detection effect.

The third row shows the result of adding location guidance to the key point network. The location guidance results in an improvement in AR$_{100}$ of 1.2%, from 70.8% to 72.0%, an improvement in AP of 0.2%, from 42.9% to 43.1%, and an improvement in AP$_{50}$ of 1.2%, from 67.8% to 69.0%. Therefore, it is confirmed that the location guidance is an effective way to improve the detection effect for solid waste, in both AP and AR.

To verify the effectiveness of the attention enhancement, the attention enhancement was added to the key point network. The fourth row in Table V shows that the attention enhancement can improve the AR$_{100}$ greatly, from 69.9% to 71.5%, but it decreases the AP, from 40.9% to 40.3%, and also the AP$_{50}$, from 66.7% to 65.6%. Because the task is to detect solid waste as much as possible, the attention enhancement is a suitable way to improve the AR of the proposed network.

However, the results listed in the fifth row show that using both mosaic data augmentation and the attention enhancement can improve all three-evaluation metrics, compared with the second row, but AR$_{100}$ decreases a little, from 71.5% to 71.1%, compared with the fourth row. The reason for this is that the attention enhancement can enhance the detection effect for some unobvious objects, such as small objects, and the mosaic data augmentation generates more small objects to input into the training network. Therefore, using the combination of both mosaic data augmentation and the attention enhancement improves the solid waste detection effect. However, the mosaic data augmentation splits up the bounding boxes so that some large objects are not all input into the network. This reduces the detection effect for large objects, which leads to the decrease in AR.

The sixth row in Table V shows the effect of path aggregation being added into the network. Compared with the fourth row, the path aggregation results in an improvement in AR$_{100}$ of 0.6%, from 71.1% to 71.7% and in improvement in AP of 0.3%, from 43.2% to 43.5%. This suggests that the path aggregation, which merges features at different scales, can improve the detection effect of the network slightly. The reason for this is that the path aggregation merges the features of a small scale, which blurs the location information.

The seventh row in Table V shows that the location guidance results in an increase in AR$_{100}$ of 0.3%, an increase in AP of 0.1%, and an increase in AP$_{50}$ of 0.1%, compared with the fifth row in Table V. We believe that the effect of the location guidance is similar to that of the attention enhancement, but the difference is that the location guidance is a supervised method and the attention enhancement is an unsupervised method, which leads to little improvement when this method is used directly.

However, when using both path aggregation and location guidance, as shown in the eighth row in Table V, compared with the fifth row, the AR$_{100}$ increases by 0.1%, from 71.7 to 71.8%, the AP increases by 0.5%, from 43.5% to 44.0%, and the AP$_{50}$ increases by 1.3%, from 69.0% to 70.3%. These results indicate that using both location guidance and path aggregation increases the AR slightly while increasing the AP significantly. This is not surprising because the path aggregation merges feature at different scales, but the fuzzy location information and the location guidance provide more accurate location information to the network. Therefore, the location guidance can compensate for the problem caused by the path aggregation.

VI. Conclusion

In this article, we have described how we built the DSWD in remote sensing images, which contains both a large number of solid waste scenes and some negative samples. An LKN-MEs is proposed for the urban solid waste detection task in remote sensing imagery. The LKN-ME method is a key point network integrating location guidance and multiple enhancements, including mosaic data augmentation, path aggregation, and an attention enhancement. The experimental results showed that the LKN-ME method can achieve state-of-the-art results of 71.8% in AR_{100} and 44.0% in AP for the DSWD dataset, and it outperformed six other classical detectors. The ablation study verified the effect of each module of the proposed network. The mosaic data augmentation has the most obvious effect in enhancing the network performance. The attention enhancement allows the network to focus more on the regions of interest. The use of the path aggregation module and the location guidance can further improve the performance of the proposed network. Although data with different resolutions are included in the DSWD dataset, whether network training based on the DSWD dataset is appropriate for other data will need further verification in the future.

ACKNOWLEDGMENT

The authors appreciate the editors and anonymous reviewers for their valuable suggestions.

REFERENCES

[1] G. Ottavianelli, S. Hobbs, R. Smith, and D. Bruno, "Assessment of hyperspectral and SAR remote sensing for solid waste landfill management," *Environment*, vol. 1, p. 593, 2005.

[2] K. Glanville and H.-C. Chang, "Remote sensing analysis techniques and sensor requirements to support the mapping of illegal domestic waste disposal sites in Queensland, Australia," *Int. J. Remote Sens.*, vol. 7, pp. 13053–13069, 2015.

[3] M. A. Nwachukwu, M. Ronald, and H. Feng, "Global capacity, potentials and trends of solid waste research and management," *Waste Manage. Res.*, vol. 35, pp. 923–934, 2017.

[4] S. Silvestri and M. Omri, "A method for the remote sensing identification of uncontrolled landfills: Formulation and validation," *Int. J. Remote Sens.*, vol. 29, pp. 975–989, 2007.

[5] J. Krook, N. Svensson, and M. Eklund, "Landfill mining: A critical review of two decades of research," *Waste Manage.*, vol. 32, pp. 513–520, 2012.

[6] B. Esmaeilian, B. Wang, K. Lewis, F. Duarte, C. Ratti, and S. Behdad, "The future of waste management in smart and sustainable cities: A review and concept paper," *Waste Manage.*, vol. 81, pp. 177–195, 2018.

[7] G.-S. Kim, Y.-J. Chang, and D. Kelleher, "Unit pricing of municipal solid waste and illegal dumping: An empirical analysis of Korean experience," *Environ. Econ. Policy Stud.*, vol. 9, pp. 167–176, 2008.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[9] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[10] T. He and S. Wang, "Multi-spectral remote sensing land-cover classification based on deep learning methods," *J. Supercomput.*, vol. 77, pp. 2829–2843, 2020.

[11] W. S. Lu and J. J. Chen, "Computer vision for solid waste sorting: A critical review of academic research," *Waste Manage.*, vol. 142, pp. 29–43, 2022.

[12] Y. Ding, X. Zhao, Z. Zhang, W. Cai, and N. Yang, "Multiscale graph sample and aggregate network with context-aware learning for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4561–4572, Jun. 2021.

[13] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Int. J. Remote Sens.*, vol. 12, p. 207, 2020.

[14] Y. X. Li, B. Peng, L. L. He, K. L. Fan, and L. Tong, "Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2279–2287, Jul. 2019.

[15] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[16] H. Zhou, X. Feng, Z. Dong, C. Liu, and W. Liang, "Multiparameter adaptive target classification using full-polarimetric GPR: A novel approach to landmine detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2592–2606, Apr. 2022.

[17] J. Su, J. Liao, D. Gu, Z. Wang, and G. Cai, "Object detection in aerial images using a multiscale keypoint detection network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1389–1398, Jun. 2021.

[18] X. Sun, Y. F. Liu, Z. Y. Yan, P. J. Wang, W. H. Diao, and K. Fu, "SRAF-Net: Shape robust anchor-free network for garbage dumps in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6154–6168, Nov. 2021.

[19] M. Gong, H. Yang, and P. Zhang, "Feature learning and change feature classification based on deep learning for ternary change detection in SAR images," *ISPRS J. Photogramm. Remote Sens.*, vol. 129, pp. 212–225, 2017.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014. pp. 580–587.

[21] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016. pp. 779–788.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[24] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

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

[26] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[27] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6569–6578.

[28] E. Xie, P. Sun, X. Song, W. Wang, and P. Luo, "PolarMask: Single shot instance segmentation with polar representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12193–12202.

[29] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015. pp. 1440–1448.

[30] 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, Jun. 2017.

[31] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016. pp. 2117–2125.

[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017. pp. 2980–2988.

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

[34] C. Gong, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[35] C. Gong, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2018.

[36] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 294–308, 2020.

[37] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.

[38] T.-Y. Lin et al., "Microsoft coco: Common objects context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[39] G. Cheng, J. Han, P. Zhou, and G. Lei, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, 2014.

[40] F. Jurie and S. Razakarivony, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, 2016.

[41] G. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018. pp. 3974–3983.

[42] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.

[43] S. W. Zamir, A. Arora, A. Gupta, S. Khan, and X. Bai, "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019. pp. 28–37.

[44] A. Newell, K. Yang, and D. Jia, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[45] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[46] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS - Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017. pp. 5561–5569.

Huifang Li (Member, IEEE) received the B.S. degree in geographical information science from China University of Mining and Technology, Xuzhou, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013.

She is currently a Professor with the School of Resources and Environmental Science, Wuhan University. She focuses on the study of radiometric correction of remote sensing images, including cloud correction, shadow correction, and urban thermal environment analysis and alleviation.

Chao Hu received the M.S. degree in geomatics engineering from Wuhan University, Wuhan, China, in 2021.

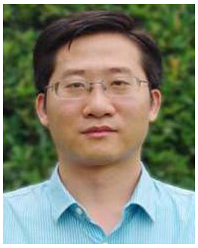
His research interests include remote sensing, deep learning, and object detection.

Xinrun Zhong received the B.S. degree in geographic information science from Lanzhou university, Lanzhou, China, in 2021. She is currently working toward the M.S. degree in human geography with the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China.

Her research interests include deep learning, computer vision and local climate zone classification

Chao Zeng received the B.S. degree in resources environment and urban–rural planning management, the M.S. degree in surveying and mapping engineering, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009, 2011, and 2014, respectively.

He was a Post-Doctoral Researcher with the Department of Hydraulic Engineering, Tsinghua University, Beijing, China. He is with the School of Resources and Environmental Science, Wuhan University. His research interests focus on remote sensing image processing and hydrological remote sensing applications.



Huanfeng Shen (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively. He is currently a Distinguished Professor with Wuhan University, where he is currently the Dean with the School of Resource and Environmental Sciences. He was or is the Principal Investigator of two projects supported by the National Key Research and Development Program of China and six projects supported by the National Natural Science Foundation of China. He has authored or coauthored more than 150 peer-reviewed international journal articles, where over 60 appeared in IEEE journals, and published four books as a Chief Editor. His research interests include remote-sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is a Fellow of the Institution of Engineering and Technology, an Education Committee Member of the Chinese Society for Geodesy Photogrammetry and Cartography, and a Theory Committee Member of the Chinese Society for Geospatial Information Society. He was a recipient of the First Prize in Natural Science Award of Hubei Province in 2011, the First Prize in Nature Scientific Award of China's Ministry of Education in 2015, and the First Prize in Scientific and Technological Progress Award of Chinese Society for Geodesy Photogrammetry and Cartography in 2017. He is also a Senior Regional Editor for *Journal of Applied Remote Sensing* and an Associate Editor for *Geography* and *Geo-Information Science and Journal of Remote Sensing*.