IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 4, 2020, accepted October 13, 2020, date of publication October 19, 2020, date of current version October 27, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3031990*

# A Lightweight Feature Fusion Single Shot Multibox Detector for Garbage Detection

**WEN MA[1], XIAO WANG[1], AND JIONG YU[2]**
[1]School of Software, Xinjiang University, Ürümqi 830000, China
[2]School of Information Science and Engineering, Xinjiang University, Ürümqi 830000, China

Corresponding author: Jiong Yu (yujiong@xju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China 61862060, in part by the National Natural Science Foundation of China 61462079, in part by the National Natural Science Foundation of China 61562086, and in part by the National Natural Science Foundation of China 61562078.

**ABSTRACT** To solve the problems of a poor manual garbage sorting environment, including heavy tasks and low sorting efficiency, we propose the Lightweight Feature Fusion Single Shot Multibox Detector (L-SSD) algorithm to realize intelligent trash classification and recognition. Since waste has a small volume and the image resolution of garbage is always low, the algorithm that we propose is an enhanced single shot multibox detector (SSD) with a lightweight and novel feature fusion module. This SSD can significantly improve the performance of rubbish detection. In this feature fusion module, features from different layers with different scales are connected in series. A new feature pyramid was generated by using downsampling blocks, which will be fed to appointed multibox detectors to predict the final detection results. Due to the extremely unbalanced ratio of positive samples to negative samples, which leads to a low accuracy of SSD, Focal Loss using balanced cross-entropy is employed, which is provided by easy examples that corresponds to difficult samples with a decline in the loss weight. Thus, the training is biased towards meaningful samples. We have replaced the backbone network of VGG16 with ResNet-101 to achieve more accurate detection. We analyzed the performance of a nonmaximum suppression (NMS) algorithm and discovered that Soft-NMS was more suitable for learning better image representations. The strategy of Soft-NMS is to suppress the undesirable detection box rather than remove it completely. The experimental results show that the L-SSD exceeds a large number of state-of-the-art object detection algorithms in both accuracy and speed.

**INDEX TERMS** Garbage identification, target detection, feature fusion, Focal Loss.

## I. INTRODUCTION

With a rapidly expanding economy, the amount of municipal solid waste has increased rapidly [1]. How to achieve the harmless and resourceful disposal of garbage is a serious problem to be urgently solved. Thus, it is obvious that recycling is significant in modern society. Effective classification of rubbish is the premise of classification processing. By addressing current problems with manual sorting, such as poor waste environment, heavy tasks and low sorting efficiency, intelligent and automated debris sorting can reduce labor costs, improve the reuse ratio of recyclable resources and help to rapidly achieve the goal of ecological construction [2]. This method can be fully utilized in practical applications. We could apply it to the garbage identification and classification of intelligent trash cans or the process of garbage sorting in large garbage dumps to reduce the burden of manual classification. This method also has a role in the process of self-recycling and utilization in manufacturing facilities.

To solve these problems, this paper proposed the Lightweight Feature Fusion Single Shot Multibox Detector (L-SSD) for garbage detection. The L-SSD was improved based on the Single Shot Multibox Detector (SSD) network structure. The main contributions are presented as follows:

1. We redefine the framework of feature fusion and establish a new feature pyramid that has stronger semantics on all types of scales. A simple and compact method was introduced to combine feature maps from different levels. We rely on an architecture that combines low resolution but semantically strong features with high resolution but semantically weak features via top-down pathways and a lateral connection. The features were fully utilized by generating a new feature pyramid.

2. We develop the Focal Loss function to replace the traditional loss function. One-stage methods are encountering a tremendous class imbalance problem during training. These detectors evaluate numerous candidate locations per image, while only a few of them contain objects. Conversely, Focal

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

Loss smoothly handles the class imbalance using a one-stage detector, which enables us to efficiently train all samples without allowing difficult-easy samples to substantially affect loss. During the training time, the function can make this model pay more attention to difficult samples by reducing the weight of easily classified samples.

3. We improve the VGG16 [3] network by replacing several kernels, introduce the idea of the Residual Neural Network (ResNet) [4] algorithm to promote the accuracy of the garbage classification algorithm, and then choose a better network to help us complete the classification.

4. We analyze the impact of optimizing the nonmaximum suppression (NMS) algorithm on L-SSD and discover that this strategy is more suitable for obtaining better image representations. In addition, a sensitivity analysis was carried out on the parameters of Soft-NMS. The experiment determined that the parameters were within the range of 0.4-0.7, which significantly improved the performance of garbage detection.

The remainder of this paper is organized as follows: In Section 2, related work in object detection was divided into one-stage approaches and two-stage approaches. In Section 3, we present the L-SSD framework. The experimental evaluation of the proposed algorithm, which was applied to trash classification, and the comparison with other detectors are shown in Section 4. The conclusions are discussed in Section 5.

## II. RELATED WORK

Object detection is a computer vision task that has attracted the attention of many researchers. In 2012, Alexnet, which is a type of convolutional neural network (CNN), was the ImageNet Challenge winner. Alexnet is on the threshold of a new era in image classification [5]. The architecture used in this contest has a simple structure, whose configuration is not deep but the performance is extremely high. The effective performance of AlexNet in the ImageNet competition with a high degree of difficulty has caused many researchers to contribute to CNN structures in the solution of image classification problems. Deep learning has become particularly popular similar to big data research, and excellent neural network algorithms, such as VGG, Xception and ResNet, have emerged.

Using different scales to identify objects was a fundamental challenge in computer vision. Figure 1(a) shows a feature pyramid [6] that is structured with image pyramids forming the basis of a standard solution. All features are independently calculated on each image scale, which is inefficient. The scale of the pyramid is constant, because the scale change of an object is offset by moving its level in the pyramid. For object recognition tasks, manual features have been considerably superseded by ConvNets (deep convolutional networks) [7]. As the modern deep ConvNets becomes active, object detectors show dramatic improvements in accuracy. A deep ConvNet calculates the feature level layer by layer. For the subsampling layers, the feature level has an inherent multiscale and pyramid shape. This intranetwork feature hierarchy generates feature maps with different spatial resolutions

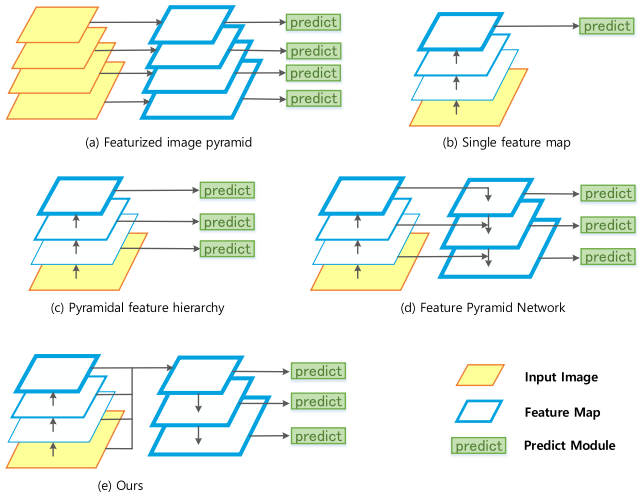


**FIGURE 1.** Different feature pyramid Models.

but introduces a large semantic gap due to different depths. The low-level features of high-resolution maps will damage the representation ability of object recognition [7]. Figure 1(b) shows that a single feature map only selects one scale feature for prediction but introduces anchors with different scales to detect multiscale objects. This feature has been taken into account in recent detection systems to obtain a faster detection speed, which is applied in some two-stage detectors, such as the Faster R-CNN [8], Region-based Fully Convolutional Network (R-FCN) [9], etc.

Target detection algorithms that are based on deep learning are mainly divided into candidate area-based methods and end-to-end regression methods.

### A. TARGET DETECTION BASED ON CANDIDATE REGIONS

Region-based Convolutional Neural Networks (RCNNs) [11], which were proposed by Girshick, initiated the use of convolutional neural networks in target detection. The R-CNN adopts a sliding window strategy extraction feature and fully utilizes the characteristics of the exhaustive method to traverse. Given an input image, 2000 category-independent candidate regions are extracted from the image. A CNN is then applied to extract a fixed-length feature vector for each region. A SVM is used to classify targets in each region. The Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition (SPP-NET) [12] algorithm was proposed by He. The main idea of SPP-NET is to remove the crop/warp operations on the original image and replace it with spatial pyramid pooling (SPP) on the convolutional feature. The OverFeat R-CNN [13] works out the features for all proposals using a CNN, such as ConvNets [7], and classifies each region via a Support Vector Machine (SVM) [14]. Due to the synchronous object detection and localization, the R-CNN is an end-to-end detection method that is applied to object retrieval. However, the R-CNN needs a substantial amount of time to process each object proposal without sharing computing. In addition, a large amount of hard disk space is needed to store these features.

The Fast R-CNN [15] shares features among object proposals to conquer the time-consuming issue of the R-CNN.

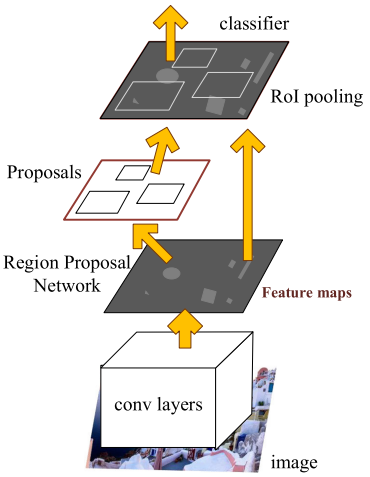


**FIGURE 2.** Algorithm framework of faster R-CNN.

To obtain a faster detection speed, the region of interest (RoI) pooling layer was designed. As a result, features are extracted only once in the process of detection per image. In addition, the SVM procedure is removed; thus, all features are temporarily stored in memory without extra disk space. However, both the R-CNN and the Faster R-CNN contain a common flaw, that is, they extract recommendations outside the training phase and rely on external area suggestion methods. Therefore, these networks are not fully end-to-end object detection systems. To solve this problem, Ren proposed a fully convolutional network named the Regional Proposal Network (RPN) [16], which is connected to the last convolutional layer of the Fast R-CNN to detect object boxes and message scores.

This new combined network was named Faster R-CNN; its structure is shown in Figure 2. The RPN is a fully convolutional network (FCN), whose function generates high-quality regional proposals, each with a confidence score. The RPN predicts both an object boundary and object scores for every location. To produce region proposals, the small network slides over the feature graph from the top transformation layer. By using different anchoring scales, the Faster R-CNN performs more robustly in distorted images than the Fast R-CNN. Owing to the shared convolution feature with the CNN, this network can greatly reduce the computation time.

Typically, the VGG16 model and Fergus model (ZF) [17] of the Fast R-CNN are pretrained in specific image datasets with corresponding annotations, such as Microsoft COCO Dataset [18] and Pascal VOC [19].

### B. TARGET DETECTION BASED ON REGRESSION MODELS

Target detection based on regression is generally referred to as a one-stage method. These methods detect targets mainly by sampling regularly with dense frames of various positions, sizes, and aspect ratios on an image. The network directly processes the input image to generate the category probability and location coordinate values of each object. In 2016, You Only Look Once (YOLO) [20] was proposed by Joseph Redmon to improve the time consumption. The YOLO algorithm removes the network layer that generates the candidate area and distributes one image into S×S grids, where each grid gives B bounding boxes. To affirm the position of one object, YOLO uses NMS to select the highest score in a bounding box, merge the overlapping regions, and output the target bounding box and category. Compared with the Faster R-CNN, the target positioning accuracy of YOLO is extraordinarily low, and each grid can only predict one target. In addition, when the size of the targets is small and the arrangement is similar, this algorithm frequently encounters missed detection. Redmon proposed the YOLOv3 [21] algorithm to adjust the network structure, carried out object detection using multiscale features, and applied logistic regression instead of soft-max regression to carry out object classifications, which improved the mean average precision (mAP) and small object detection.

Nothing is more important than balancing the relationship between speed and accuracy. SSD enabled the advantages of Faster R-CNN and YOLO to be combined. Taking account of the detection accuracy, VGG16 is adopted as the trunk network, and the final full connection layer was replaced by a convolutional layer. In this model, the anchor box in Faster R-CNN is used as an auxiliary method to abandon the RPN. To improve the accuracy of small objects, the multilayer feature map of SSD is used to predict the target categories and directly show the boundary boxes. NMS is used to postprocess the final detection results. Because SSD detects objects directly from the plane ConvNet feature maps, it can achieve complete real-time object detection, which is faster than most of the other advanced target detectors. An alternative is to reuse the pyramidal feature hierarchy that is computed by a ConvNet as if it were a characterized image pyramid. As illustrated in Figure 1(c), the pyramid of the conventional SSD would reuse the multiscale feature maps from different layers calculated in the forward transfer. To avoid using low-level features, however, the SSD refuses to reuse previously computed layers. At the top of the network, the network builds the pyramid instead of adding several new layers. Thus, the network misses the opportunity to reuse the higher-resolution maps of the feature hierarchy, which was important for detecting small objects.

To improve the accuracy of the SSD, the Deconvolutional Single Shot Detector (DSSD) [22] uses deconvolution layers to augment the SSD with an additional large-scale context. However, an unduly complex model operates with a slow speed. The RSSD (enhancement of SSD by concatenating feature maps for object detection) [23] uses rainbow concatenation, which passes through both pooling and joining to fully utilize the relationship between the feature pyramid layers for enhancing the accuracy with a slight loss of speed. Learning Deeply Supervised Object Detectors from Scratch (DSOD) [24] investigates how to train an object detector from scratch, and DenseNet [25] architecture was designed to increase the efficiencies of parameters. Using feature fusion algorithms in ConvNet, which utilize multiple layer features, can improve the vision task performance. Before predicting

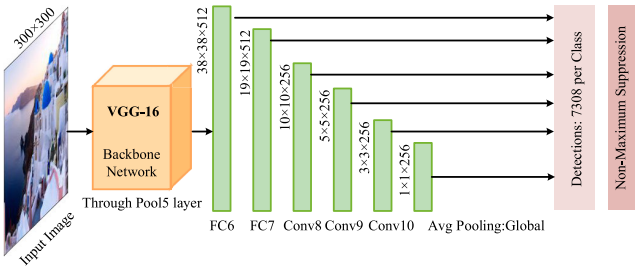


**FIGURE 3.** Framework of traditional SSD.

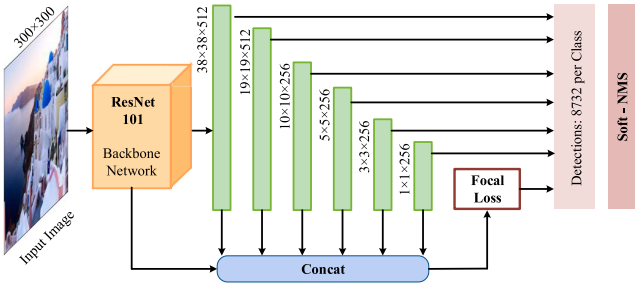


**FIGURE 4.** Framework of L-SSD.

the results, HyperNet [26] and Parsenet [27] concatenate features from multiple layers. Fully convolutional networks (FCN) [28], U-Net (Convolutional Networks for Biomedical Image Segmentation) [29], and Stacked Hourglass networks [30] also use skip connections to associate low-level feature maps with high-level feature maps to fully utilize synthetic information. To enhance the performance, SharpeMask [31] and Feature Pyramid Networks for Object Detection (FPNs) [10] introduce a top-down structure to combine the different levels of features (Figure 1(d)). Although they have made considerable efforts to balance the accuracy and speed of target detection, the effect was not obvious. There were still some problems, such as missing small objects or confusing the foreground and background.

## III. METHODOLOGY

Although the accuracy of the SSD algorithm was greatly improved compared with YOLO, there are still problems of missing small targets and simultaneously detecting the same object with different sizes of boxes. The framework of the traditional SSD algorithm is shown in Figure 3.

To solve the previously mentioned problems of SSD, this paper proposed the L-SSD algorithm. which adds a lightweight but efficient feature fusion module to the conventional SSD. Different from previous methods of adjusting the training strategy to improve the performance, our strategy takes the whole network and fully utilizes different layers in the feature pyramid by changing its backbone network. First, considering the relationship between layers in the feature pyramid, the complete network structure of the L-SSD was implemented for target detection. Changing the backbone network VGG16 with ResNet-101 is an effective way to improve the performance of feature extraction. Second, as shown in Figure 4, a feature fusion module that is used for object detection in the L-SSD was designed to realize effective fusion of feature information between feature layers. The structure of the feature pyramids was replaced by top-down pathways and lateral connections. Thus, the fused feature map can contain richer details and semantic information. In addition, a more balanced and lightweight Focal Loss function was proposed to solve the problem of imbalance between easy-hard samples and the tasks involved in the conventional SSD algorithm. The function used a more balanced cross-entropy to reduce the loss weight. The principle of this function is that easy samples vary with negative samples, so that the training is biased towards meaningful samples. Eventually, to improve the shortcomings of the traditional NMS algorithm, we developed the Soft-NMS algorithm to attenuate the detection box scores with obvious overlap instead of removing them completely.

### A. FEATURE PYRAMID NETWORKS

The goal of this paper is to create a feature pyramid with strong semantics for all scales on the pyramid shape of the ConvNets' feature hierarchy. To achieve this goal, we rely on an architecture that combines low-resolution but semantically strong features with high-resolution but semantically weak features via top-down pathways and a lateral connection. In the feature fusion and feature pyramid generation method that we proposed in this paper, as shown in Figure 1(e), features from different layers with different scales are concurrently concatenated and subsequently generate a series of pyramid features. The method consists of a feature pyramid that has rich semantics on all levels. A single input image will be quickly built according to its proportion to the feature pyramid. Therefore, a significant solution for a ConvNet object detector to improve the accuracy is to synthesize the features with a slight structure.

As mentioned in Section 2, many algorithms attempt to observe and fully utilize the pyramidal features. The most prevalent method is shown in Figure 1(d). FPN [10] and DSSD [22] employed this type of feature fusion to substantially improve the performance, which was verified in their papers. However, this design adds a process of merging multiple features. The new features on the right side can only blend the features from the left side that are higher than the same levels. Addressing potential features and multifeature elements also consumes a considerably amount of time. These tasks would be improved with a lightweight but efficient feature fusion module that we propose. Our motivation is to merge different levels of features at one time in an appropriate way and generate a feature pyramid from the fused features. The largest difference between Figure 1(d) and (e) is that instead of horizontally connecting the feature semantics of each layer, our algorithm aggregates them to the top of the left pyramid and then horizontally connects them to the feature pyramid on the right, which will solve the time-consuming problem and integrate feature semantics of all scales to obtain stronger semantic information.

The feature fusion module consists of top-down pathways and the lateral connection [32]. In our approach, this kind

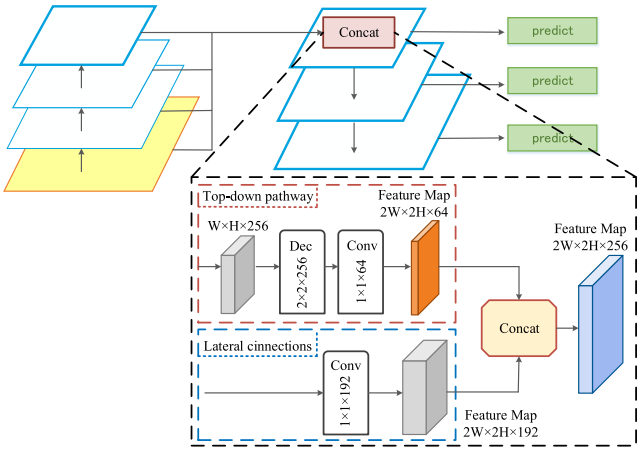


**FIGURE 5.** Building block that illustrates the lateral connection and top-down pathway, merged by the Concat method.

of path utilizes deconvolution to recognize higher levels of features. Figure 5 shows the details of the feature fusion module. We adopt the following strategies to concatenate the features with different scales using a simple and efficient way. Spatial feature maps are mapped twice using convolution layers, and then sampled feature maps are processed by 1 $*$ 1 convolution kernels to reduce the number of high-level feature channels. To satisfy the output characteristic mapping of the same channel numbers, we take 1 $*$ 1 convolution operations to convolution operations that are generated from bottom-up paths. Horizontal connections merge feature mappings that are generated from bottom-up paths with feature mappings that are based on Concat operations. The process of Concat can be expressed as follows (1):

$$Z_{\text{concat}} = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_{i+c} \tag{1}$$

Xi represents the input of the bottom-up pathway, while Yi is the input of the lateral connection. Here, $*$ denotes the convolution operation. Concat is an increase in the number of channels [33], while the information under each feature is not increased. We set the number of channels of output feature maps to 256 in our experiments, so all convolutional layers have 256 channel outputs.

### B. FOCAL LOSS

The dominant paradigm in modern object detection is based on two-stage approaches. As pioneered in the Selective Search work, the first stage generates a sparse set of candidate proposals, which should contain all objects while filtering out the majority of negative positions. The second stage divides these proposals into foreground classes or backgrounds. However, the SSD has renewed interest in one-stage methods. The speed of the detector has been reduced, and its accuracy is improved compared with two-stage methods. Accordingly, one-stage methods are experiencing a tremendous class imbalance problem during training. These detectors evaluate numerous candidate locations per image, while only a few of them contain objects. Conversely, Focal Loss smoothly handles the class imbalance using a one-stage detector and enables us to efficiently train all examples without allowing difficult-easy samples to substantially affect loss.

Focal Loss [34] is an amendment to standard cross-entropy loss that can make the model focus more on difficult samples in training by reducing the weights of easily classified samples. In the model, y represents the real label of the sample, and p represents the predicted value of the sample generated by the classifier. We introduce Focus Loss in binary classification based on cross-entropy (CE) loss: (2)

$$CE(\mathrm{p}, \mathrm{y}) = \begin{cases} -\log(p) & \text{if y} = 1 \\ -\log(1-p) & \text{otherwise .} \end{cases} \tag{2}$$

For convenience, we redefine pt: (3)

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases} \tag{3}$$

Therefore, we can rewrite CE(p,y): (4)

$$CE(\mathrm{p}, \mathrm{y}) = CE\ (\mathrm{p}_t) = -\log\ (\mathrm{p}_t) \tag{4}$$

Focal loss adds two factors to the binary cross-entropy loss function. There is a large gap between the number of positive samples and the number of negative samples in one-stage detector training. Thus, a common way is to add the weight of $\alpha$ to adjust positive and negative samples. In practice, $\alpha$ may be set to the inverse class frequency or treated as a hyper-parameter set by cross-validation. If the frequencies of the negative samples are large, then the weights of the negative samples will be reduced. If the number of positive samples is small, the weights of the positive samples will be relatively increased. Therefore, the shared weights of the positive and negative samples to the total loss can be controlled by setting the value of $\alpha$. In most cases, $\alpha$ takes a relatively small value to reduce the weights of negative samples. We suggest a restoring loss function to reduce the weights of simple samples and focus training on difficult cases. Focal Loss was ultimately defined as follows: (5)

$$FL\ (\mathrm{p}_t) = -\alpha\ (1-\mathrm{p}_t)^{\gamma}\ \log\ (\mathrm{p}_t) \tag{5}$$

Although $\alpha$ has an important role in balancing positive and negative examples, it does not distinguish easy and hard examples. To solve this problem, another factor named the focusing parameter $\gamma$ was proposed. This parameter can regulate the rate of weight reduction of simple samples. In this experiment, we use the loss function form of formula (5). In this way, we can not only adjust the weights of positive and negative samples but also control the weights of difficult and easy classification samples.

### C. BACKBONE NETWORK STRUCTURE

SSD algorithm uses the traditional VGG16 network. As shown in Figure 6, VGG16 contains 16 hidden layers, including 13 convolutional layers and 3 fully connected layers.

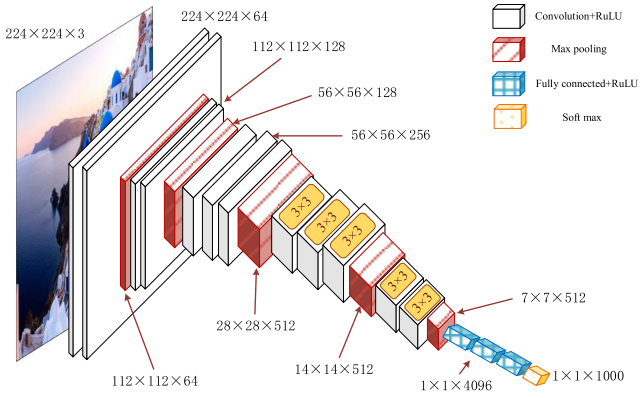


**FIGURE 6.** Network architecture of VGG16.

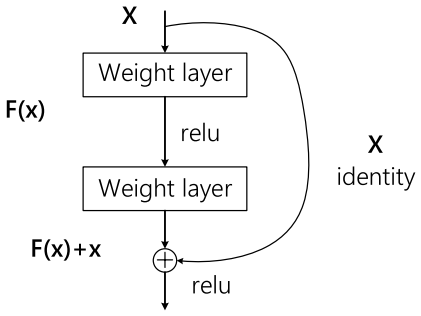


**FIGURE 7.** Residual block structure of ResNet. X is the input, F(x) is the residual map, ReLu is the activation function.

In this study, we made improvements based on the conventional SSD target detection model. Aimed at the characteristics of small garbage volume and low image resolution, ResNet-101 with a deeper network depth and less computation was used to replace the original VGG16, and richer features were extracted. ResNet-101 is a residual module that enables us to train deeper networks. The structure of the module mainly uses 3 × 3 convolution, and its internal residual blocks adopt jumping connections, which alleviate the problem of gradient disappearance by increasing the depth in a deep neural network [35]. Figure 7 shows the residual network model. For a given receptive field, it is better to use a small convolutional kernel with accumulation than a large convolutional kernel, because multiple nonlinear layers can increase the network depth to ensure that more complex patterns are learned with less cost and fewer parameters. After optimization, we adopt this method to replace the 7 × 7 convolution kernel with three 3 × 3 convolution kernels and the 5 × 5 convolution kernel with two 3 × 3 convolution kernels. The main purpose of this approach is to improve the depth of the network, and the neural network has better performance with the same perceptual field.

### D. SOFT-NMS

NMS is a fundamental problem in object detection. The principle of NMS is to generate a detection box according to the target detection score [36]. NMS selects the detection box with the highest score and suppresses other detection boxes that overlap with the selected detection box. This process

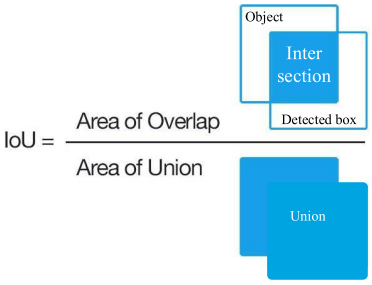


**FIGURE 8.** The meaning of IoU.

applies recursively to the remaining detection boxes. According to the design of this algorithm, if the target is located within the preset overlap threshold range, the object may not be detected. Therefore, the Soft-NMS [37] was proposed. A continuous function attenuates the detection score of the nonmaximal detection box instead of removing it completely. Soft-NMS is a greedy algorithm that does not guarantee the globally optimal solution for the scores of detection boxes. However, the Soft-NMS algorithm is a more general nonmaximal suppression algorithm, and the traditional NMS algorithm can be regarded as a special case that uses a discontinuous binary weight function. The traditional NMS algorithm was defined as (6)

$$S_{\mathrm{i}} = \begin{cases} S_i, & \mathrm{IoU}\left(\mathrm{M}, \mathrm{b}_i\right) < \mathrm{N}_t \\ 0, & \mathrm{IoU}\left(\mathrm{M}, \mathrm{b}_i\right) \geq \mathrm{N}_t \end{cases} \tag{6}$$

where Si is the score calculated by the classifier for each test box, M is the candidate box with the highest prediction score, bi is the candidate box that waits for processing, and Nt is the NMS threshold. As shown in Figure 8, IoU is the ratio of the intersection between the candidate detection box and the real target to the union of the candidate detection box and the real target. If the value of the IoU exceeds predetermined threshold values, set its value to zero. Otherwise, it still retains the original value.

The improved Soft-NMS algorithm attenuates the box rather than removing the detection score of the part where the IoU value of M and bi exceeds the Nt threshold. The Soft-NMS algorithm was defined as follows: (7)

$$S_i = \begin{cases} S_i, & \mathrm{IoU}\left(\mathrm{M}, \mathrm{b}_i\right) < \mathrm{N} \\ S_i\left(1 - \mathrm{IoU}\left(\mathrm{M}, \mathrm{b}_i\right)\right), & \mathrm{IoU}\left(\mathrm{M}, \mathrm{b}_i\right) \geq \mathrm{N}_t \end{cases} \tag{7}$$

We analyze the impact of optimizing NMS on the SSD and discover that this strategy is more suitable for learning better image representations. In addition, sensitivity analysis was carried out on the parameters of Soft-NMS, and experiments determined that the parameters were within the range of 0.4-0.7, which were significantly higher than the average accuracy of traditional NMS algorithm.

## IV. EXPERIMENTS

### A. EXPERIMENTAL SETUP

#### 1) DATASETS

Since there are no public datasets in garbage classification studies, the data used in this study were obtained by network

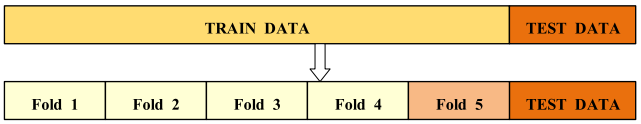


**FIGURE 9.** Dataset division.

collection and camera shooting. Ensuring image quality and clarity, network acquisition accounts for 20% of the total dataset, and the remainder of the dataset is captured by the Sony SLR camera. There are 9000 pictures in the dataset, including cardboard, glass, paper, plastic and metal. Each category accounts for one fifth of the total with 1800 images. The size of the picture is 1891 $\times$ 1263, and the format is JPG. To ensure the reliability of category tags, all trash categories have been identified manually in advance.

#### 2) DIVISION OF DATASET

The dataset is differentiated into the training set, verification set and test set. The training set refers to the data images that are employed in the process of model training; It has a significant role in the initial model fitting and parameter optimization. In this paper, a 5-fold cross-validation method was employed to test the accuracy of the optimization algorithm. A total of 9000 waste pictures were taken as input images in this model. We separate them into six equal parts on average and then rotate among five of them as garbage detection training datasets. The remaining part is a test dataset to test the models. Through training and testing experiments, each model will obtain the corresponding correct rate. The correct rate results were calculated for the average value, which were used to estimate the accuracy of the target detection algorithm. As a result, the actual training set accounts for 4/5 of the total dataset. The verification set refers to checking the state and convergence of their model after each epoch is completed. This set does not participate in the process of gradient descent but adjusts the super parameters, such as the number of iterations and learning rate.

The verification set determines which group of hyperparameters has a brilliant performance and adopts them according to the performance of the five groups in the models. In the process of training, the verification set can also be used to monitor whether the model has been fitted to judge the time when training stops. As shown in Figure 9, one of the training datasets is randomly selected as a verification set, and the remaining four copies are used as a training set to record the accuracy. Cycle in turn until each copy has a validation set, which marks the end of cross-validation. After calculating the average value of the five times precision, the highest accuracy model will be chosen, and the superparameter of the model will be determined in the final model. The test set tests the model after completing the training to obtain the optimal model. Testing is performed to determine the accuracy of the model detection and classification and further verify its generalization ability. Testing ensures the correctness and validity of this model in practical application in the future.


**FIGURE 10.** Detection examples with L-SSD.

#### 3) EXPERIMENTAL ENVIRONMENT

We performed the experiments using an Intel Xeon Gold 5117@2.0 GHz processor with a 32 Gb RAM and a Nvidia Tesla V100-PCIE-16Gb. All experiments were conducted with TensorFlow 1.3, Python 3.5, and OpenCV 3.1.

### *B. EVALUATION METRIC*

The mean average accuracy (mAP) is the most commonly employed metric for evaluating the target detection accuracy. mAP is defined as the average of the average precision (AP) of all object categories, which is the area under the precision and recall rate (P-R) curves. Thus, we use mAP as an authoritative metric to evaluate the performance of our method.

### *C. RESULTS AND DISCUSSION*

#### 1) ABLATION STUDY ON GARBAGE DATASET

Figure 10 shows the detection results of the L-SSD on the garbage dataset. We have displayed each of these categories.

We note that garbage has irregular folding characteristics, such as cardboard, paper and metal. Sometimes, cans have different degrees of extrusion phenomenon. This problem adds to the difficulty of identification. Some wastes are similar in appearance but have different materials, which also affects the accuracy of identification. The difference in the appearance between cardboard and paper is not obvious. Bottles that are composed of glass or plastic hinder the detection. In addition, the picture taken by the image in the case of insufficient light will produce poor pixels, which will also hinder processing the blurred image. However, the algorithm

**TABLE 1.** Ablation study: effects of various improvement methods on garbage dataset.

| Method | mAP |
|---|---|
| SSD (VGG 16) | 74.02% |
| SSD (ResNet-101) | 79.63% |
| SSD (VGG 16) + FPN | 75.90% |
| SSD (ResNet-101) + FPN | 76.87% |
| SSD (VGG 16) + Feature Pyramid (ours) | 78.31% |
| SSD (ResNet-101) + Feature Pyramid (ours) | 80.16% |
| SSD (ResNet-101) + Focal Loss | 77.64% |
| SSD (ResNet-101) + Soft-NMS | 76.54% |
| L-SSD (ours) | 83.48% |

that we proposed can reasonably classify and detect the materials. Our algorithm can accurately distinguish rubbish from different materials, although these materials sometimes appear similar in their outlines. In addition, the algorithm can also avoid being disturbed by wrinkles and low pixels.

To reflect the effect of the series of actions that we added to the traditional SSD, we ran models with different settings and recorded their evaluations in Table 1, where the Feature Pyramid indicates the Feature Fusion module that we proposed. The mAP of the conventional SSD with VGG16 is 74.02%. After changing the backbone structure, which is replaced with ResNet-101, the mAP improved to 79.63%. We added the pyramid model of FPN to VGG16 and ResNet-101, and the mAP rose to 75.90% and 76.87%, respectively. After using our lightweight Feature Pyramid, which can fully mine context information, the mAP of VGG16 rose to 78.31% and the ResNet-101 increased to 80.16%, because fused feature layers contain rich details and semantic information. By adding Focal Loss, the mAP increased to 77.64%, mainly because our more balanced Focal Loss function can solve the imbalance problems by reducing the weight of the easily classified samples to achieve better convergence. In addition, the NMS algorithm was substituted by Soft-NMS in the SSD framework, and the mAP climbed to 76.54%. Thus, it can be seen that the improved methods that we have proposed are effective. As shown in Table 1, when we combined our proposed improvements and adopted a better network framework, the optimal performance of our L-SSD algorithm was 83.48%.

### 2) PERFORMANCE COMPARISON

To verify and evaluate the performance of our proposed method, this part makes a quantitative and qualitative comparison between our method and some advanced target detection algorithms. As shown in Table 2, For Faster R-CNN (VGG16) or Faster R-CNN (ResNet-101), the mAP values are 72.34% and 76.48% respectively, and the FPS values are 7 and 9, respectively. As an improvement in Faster R-CNN, the two-stage detection algorithm achieves excellent detection accuracy, but due to the region proposals, it is much slower than one-stage detection algorithms. Although the detection speed of YOLOv3 reaches 65 FPS, its mAP is

**TABLE 2.** Mean average precision and time results comparison between different models.

| Method | Back bone | AP (%) | | | | | mAP (%) | FPS |
|---|---|---|---|---|---|---|---|---|
| | | Card board | Paper | Plastic | Metal | Glass | | |
| Faster R-CNN | VGG16 | 75.4 | 62.7 | 68.2 | 79.3 | 76.1 | **72.34** | **7** |
| Faster R-CNN | ResNet-101 | 91.3 | 75.3 | 57.7 | 78.3 | 79.8 | **76.48** | **9** |
| YOLO v3 | Darknet | 69.4 | 68.6 | 68.6 | 52.9 | 62.1 | **64.32** | **65** |
| SSD | VGG16 | 81.4 | 74.7 | 74.4 | 70.1 | 69.5 | **74.02** | **46** |
| L-SSD | ResNet-101 | 87.5 | 78.6 | 86.2 | 82.4 | 82.7 | **83.48** | **40** |

9.7% lower than that of the traditional SSD. The conventional SSD algorithm is one of the most popular target detection algorithms with high accuracy and speed. For SSD, the mAP and FPS are 74.02% and 46, respectively. Compared with it, L-SSD has a 9.46% accuracy gain. However, due to the complexity of the model structure, the feature fusion method drops the speed to 40 frames per second. Compared with SSD and Faster R-CNN, our L-SSD shows higher accuracy and faster speed because the structure of our feature fusion module is simpler, and a more balanced Focal Loss function will not increase the additional detection time.

## V. CONCLUSION

This paper is aimed at the urgent need for effective refuse classification to address the problems of a poor manual litter sorting environment, heavy tasks and low sorting efficiency. Research on the use of deep learning for rubbish classification is lacking. Thus, we proposed L-SSD to assist us in effectually solving these problems. L-SSD is an enhanced SSD, in which a lightweight but efficient feature fusion module is applied to its framework. First, considering the relationship between layers of the feature pyramid, a new feature fusion module is adopted to change the traditional SSD network structure, which can effectively integrate the feature layers that are used for object detection in the SSD network structure. In the previous qualitative analysis, after adding the feature pyramid module, the accuracy of garbage identification was greatly improved. The effectiveness of this module is also affirmed. This consequence was supported by two reasons. The first reason is that all the features are fused once in the topmost feature map in this feature pyramid module. The feature maps of different scales can be fused to obtain more rich features. The second reason is that we used only one horizontal connection to reduce the amount of repetitive computation, which shortens the detection time. Second, by utilizing a more balanced Focal Loss function, the gradient contributions of easy and hard samples were effectively balanced to the whole localization loss function, so that the

imbalances of samples and multitasks in the conventional SSD algorithm can be partially solved. Third, we explore the impact of the VGG16 and ResNet-101 architecture on the performance of the garbage classification algorithm and choose a better network to help us complete the classification tasks. In addition, we analyze the impact of optimizing NMS on L-SSD and determine that Soft-NMS is more suitable for learning better image representations. The results demonstrate that our L-SSD model outperforms the classical SSD framework, especially for small targets, while still maintaining a comparable detection speed for other reasonable detectors. In the future, for better performance, the use of more powerful backbone networks, such as DenseNet [13], to enhance our L-SSD is worthwhile. The addition of the attention module to our algorithm is a meaningful innovation of this research field.

## REFERENCES


[1] G. J. G. Mollica and J. A. P. Balestieri, "Is it worth generating energy with garbage? Defining a carbon tax to encourage waste-to-energy cycles," *Appl. Thermal Eng.*, vol. 173, Jun. 2020, Art. no. 115195.

[2] R. F. Service, "Electricity turns garbage into high-quality graphene," *Science*, vol. 367, no. 6477, p. 496, Jan. 2020.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: IEEE Computer Society, Jun. 2016, pp. 770–778.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[6] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. ECCV*, 2014, pp. 346–361.

[8] 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.

[9] J. Dai, Y. Li, K. He, and J. Sun R-fcn:, "Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems 29*, vol. 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 379–387.

[10] T. Lin, P. Dollar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[11] 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.*, Jun. 2014, pp. 580–587.

[12] 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, Sep. 2015.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*. [Online]. Available: https://arxiv.org/abs/1312.6229

[14] T. Joachims, "Making large-scale SVM learning practical," Komplexitätsreduktion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Germany, Tech. Rep. SFB 475, 1998.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*. [Online]. Available: http://arxiv.org/abs/1506.02640

[21] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[22] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659

[23] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, *arXiv:1705.09587*. [Online]. Available: https://arxiv.org/abs/1705.09587

[24] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1919–1927.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[26] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 845–853.

[27] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: http://arxiv.org/abs/1506.04579

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[29] O. Ronneberger, P. Fischer, and T. Brox "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[30] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision—ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 483–499.

[31] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollar, "Learning to refine object segments," in *Proc. ECCV*, 2016, pp. 75–91.

[32] H. Wu, K. Zhang, and G. Tian, "Simultaneous face detection and pose estimation using convolutional neural network cascade," *IEEE Access*, vol. 6, pp. 49563–49575, 2018.

[33] W. Lu, L. Liang, X. Wu, X. Wang, and J. Cai, "An adaptive multiscale fusion network based on regional attention for remote sensing images," *IEEE Access*, vol. 8, pp. 107802–107813, 2020.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[35] S. Fekri-Ershad and F. Tajeripour, "Multi-resolution and noise-resistant surface defect detection approach using new version of local binary patterns," *Appl. Artif. Intell.*, vol. 31, nos. 5–6, pp. 395–410, 2017.

[36] F.-E. Shervan, "Bark texture classification using improved local ternary patterns and multilayer neural network," *Expert Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113509.

[37] 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. (ICCV)*. Venice, Italy: IEEE, Oct. 2017, pp. 5562–5570.


**WEN MA** was born in Ürümqi, Xinjiang, in 1993. She received the B.S. degree in network engineering from the Tianjin University of Science and Technology, China, in 2016. She is currently pursuing the master's degree with the School of Software, Xinjiang University, China. Her research interests include image recognition, image enhancement, image classification, and image processing. She is a member of China Computer Federation (CCF).

**XIAO WANG** received the B.S. degree in software engineering from Xinjiang University, China, in 2018, where he is currently pursuing the master's degree with the School of Software. His research interests include image enhancement, image recognition, and under-warder image processing.

**JIONG YU** was born in Beijing, China, in 1964. He received the M.S. degree from the Key Laboratory of Materials Modification by Laser, Ion and Electron Beams, Ministry of Education, Dalian University of Technology, China, in 1995, and the Ph.D. degree from the School of Computer Science and Technology, Beijing Institute of Technology, China, in 2009. He is currently working as a Full Professor with the School of Information Science and Engineering, Xinjiang University, and the Dean of the Graduate School of Xinjiang University, China. He was a Visiting Professor with the National Research Council of Canada, in 2003. He served as the Vice-Dean of the School of Computer Science, Beijing University of Technology, in 2005. He has hosted several funded projects from the National Science Foundation of China (NSFC) and published articles in several international journals, including eight articles in SCI and six articles in ISTP. His research fields include image processing, deep learning, cloud/grid/cluster computing, parallel and distributed computing, high-performance computing, network security, and computer networks. He is a Senior Member of China Computer Federation (CCF).