EURASIP Journal on Wireless Communications and Networking

RESEARCH

Open Access

Intelligent hyperspectral target detection for reliable IoV applications

Zixu Wang[1*], Lizuo Jin[1] and Kaixiang Yi[2]

*Correspondence: 213191960@seu.edu.cn

[1] Southeast University, Nanjing 210096, China
[2] National University of Defense Technology, Nanjing 210096, China

Abstract

In recent years, hyperspectral imagery has played a significant role in IoV (Internet of Vehicles) vision areas such as target acquisition. Researchers are focusing on integrating detection sensors, detection computing units, and communication units into vehicles to expand the scope of target detection technology with hyperspectral imagery. As imaging spectroscopy technology gradually matures, the spectral resolution of captured hyperspectral images is increasing. At the same time, the volume of data is also increasing. As a result, the reliability of IoV applications is challenged. In this paper, an intelligent hyperspectral target detection method based on deep learning network is proposed. It is based on the residual network structure with the addition of an attention mechanism. The trained network model requires few computational resources and can provide the results in a short time. Our method improves the value of mAP50 by an average of 3.57% for all categories and by up to 5% for a single category on the public dataset.

Keywords: Reliability, Hyperspectral image, Deep learning, Target detection, IoV

1 Introduction

As we all know, the accuracy of the object detection result is very important for the reliability of IoV applications. However, road conditions change rapidly, and the diversity and complexity of pedestrians and obstacles increase the difficulty of target detection. The traditional images have been unable to meet the requirements of target detection tasks. Sometimes it can easily lead to the intelligent driving system getting the wrong environmental information. Therefore, hyperspectral images have been introduced into IoV in recent years. However, object recognition from hyperspectral images is computationally complex. Traditional hyperspectral target recognition algorithms are not only slow and less robust, but also cannot be applied to hyperspectral target recognition systems.

To solve the above problems, one of these research areas is to optimize the communication mechanism and a dynamic task offloading mechanism that can handle the massive amount of data that flows among vehicles and edge servers, to meet the real-time requirements in IoV [1, 2]. Another research area is to optimize the detection algorithms. The feature compression algorithm which is based on the traditional method is proposed to process large dimensionality of hyperspectral image data, such as the projection method. The dimensionality of hyperspectral image data is reduced through feature compression to reduce the computational difficulty. But feature compression method is much more difficult to process, it is complex to solve the feature projection, and the spectral information of the target is reduced after data compression.

Nowadays, with the rapid development of neural network technology, it has been widely used in the field of deep learning [3–5]. Neural networks can simulate the mechanism of feature extraction at the level of the human brain. The more layers the neural network has, the larger its parameters and the stronger its feature extraction capability. Therefore, more and more excellent neural networks have been proposed, achieving better results and less processing time than traditional methods. Meanwhile, their powerful feature extraction capabilities have attracted the attention of scholars in the field of hyperspectral, resulting in the creation of many methods for the classification and detection of targets based on neural networks to extract the spectral properties of hyperspectral images [6, 7].

Therefore, in order to achieve target detection in hyperspectral images, it is of great interest to use neural networks for hyperspectral target detection. So in this paper, a neural network-based model for hyperspectral target detection (Sequeeze and Excitation for Hyperspectral Target Detection: SEHyp) is proposed by analyzing the feature of hyperspectral images. The model adopts the first-order target detection YOLO model structure and proposes a new feature extraction module with an attention mechanism for the spectral features of hyperspectral images. The attention mechanism adaptively weights the feature information to highlight the important feature information in the channel and attenuate the irrelevant feature information, thus improving the network accuracy. In addition, the model output module is modified to adopt a dual output, with separate outputs for target coordinates and categories, to achieve prediction of target coordinates and categories. After the experimental verification, a good detection result is obtained. The trained network model can achieve high accuracy and robustness, it can provide the results in a short time, which is crucial and can meet the reliable requirement of intelligent applications of IoV.

2 Related word

Imaging spectroscopy is an important detection method for acquiring spatial and spectral information from materials, which can be used to obtain hyperspectral images and plays a crucial role in the field of visual perception, such as target identification, classification, and recognition. Imaging spectroscopy is a detection technique that can acquire both spatial and spectral information about a target by extracting the spectral features of a feature at an early stage. The following is the current state of research in hyperspectral target detection technology [8, 9].

JX Yu and others propose a novel workflow performance prediction model (DAG-transformer) that fully exploits the sequential and graphical relationships between workflows to improve the embedding representation and perceptual ability of the deep neural network. Their study provides a new way to facilitate workflow planning [10].

Harsanyi invented the detection method of constrained energy minimization (CEM) [11]. The main idea of CEM method is to extract information in a certain area by reducing the information interference in other areas, and CEM has achieved good results in small target detection and is widely used. However, the CEM method requires prior knowledge of ideal targets for hyperspectral images, so CEM algorithms can only be used to detect ideal targets.

Jimenenz applied genetic algorithms and projection methods to the extraction and categorization of hyperspectral image features [12]. A data processing method of projection tracking proposed by Prof. Friedman [13] was specifically designed for linear dimensionality reduction of high-dimensional data, and the projection method reduces the dimensionality of data to reduce the difficulty of subsequent data processing. However, as mentioned in the background of the study, the feature compression method is difficult to handle and the projection method is complicated to find the eigenprojections, so this method is only suitable for the target detection of pure point image elements.

Li Wang proposed the SSSERN algorithm [14]. The attention mechanism is mainly introduced by adding the SE module to the residual network. The accuracy of the network is improved.

There are also detection algorithms based on sparse representation [15, 16], which represent the image background as a more representative basis vector or spectrum and use the product of spectral prior knowledge and related parameters to represent the original hyperspectral data. Li et al. proposed the BJSR (background joint sparse representation) algorithm [17], an anomaly detection algorithm for hyperspectral images using background joint sparse representation by estimating an adaptive orthogonal background complementary subspace by BJSR, which adaptively selects the most representative background basis vectors for local regions, and then proposed an unsupervised adaptive subspace detection method to suppress the background and highlight the anomalous components at the same time. Although the sparse representation method can detect anomalous pixel points, it may receive the influence of sensor noise and multiple reflections of electromagnetic waves during hyperspectral acquisition, which leads to the spectral variation of the substance, resulting in poor detection of the algorithm, and the method can only identify the anomalous targets and cannot classify the targets.

3 Problem analysis

3.1 Hyperspectral image detection

Hyperspectral images, which are composed of tens to hundreds of wavelength images, are three-dimensional structured images with both spatial and spectral information. The features extracted from hyperspectral images can be roughly divided into two categories: spatial features and spectral features, which are essentially two very different kinds of features. Spatial features are a reflection of the target's position, shape, size, and other information in two-dimensional space, while spectral features are a reflection of the target's ability to reflect light at different wavelengths, which is known as the "spectral fingerprint" and is one of the important optical properties of matter [18–22].

Hyperspectral images are always used as input data for target detection IoV applications. Hyperspectral images are the product of a combination of imaging and spectroscopic techniques. It can store both spatial and spectral information in the range photographed in a kind of data cube. This data cube can detect and identify objects with similar appearance but different materials. Hyperspectral images are not RGB images that only simply integrate the R, G, and B bands. The higher the spectral resolution of the hyperspectral image, the higher the number of spectral channels. For different targets with different representation capabilities, objects with different morphology have rich spatial information, while the spectral information may be weak, when using spatial features to identify and classify the targets can achieve higher accuracy. When the objects are similar in shape, their spatial information is weak, but the spectral information is rich, so the objects can be distinguished by the spectral information of a certain wavelength to obtain higher accuracy.

Figure 1 shows the spectral image of the hyperspectral image. Each spectral image is represented as a grayscale image.

Although hyperspectral stores a large amount of information, it also brings problems such as large amount of data and redundancy. The target detection application of the IoV needs to get target feature information from these data. Then, the specified target and its coordinates and categories are detected from the feature information. Most of the traditional hyperspectral target detection algorithms only consider the difference of spectra between different objects and utilize the spectral information of hyperspectral images, while the spatial information of objects is less used. And their spectral information extraction methods generally use feature compression methods to reduce the computational effort, or manually extract the feature spectral bands for difference analysis. The design of feature compression methods is difficult, such as the complicated feature projection in the projection method. The manual extraction of the feature spectrum is designed only for a single target, which is less adaptable to other targets. With the rapid development of computer hardware, deep learning technology has been improved unprecedentedly. Neural networks, as the most widely used theory in deep learning, has demonstrated powerful high-level (more abstract and semantic) feature representation and learning capabilities. It is able to extract nonlinear correlation features between data. This makes a qualitative leap in the detection accuracy and speed of target detection technology [23].

However, models for target detection on hyperspectral images are relatively rare, and most of them are color images or grayscale images, which lack spectral information

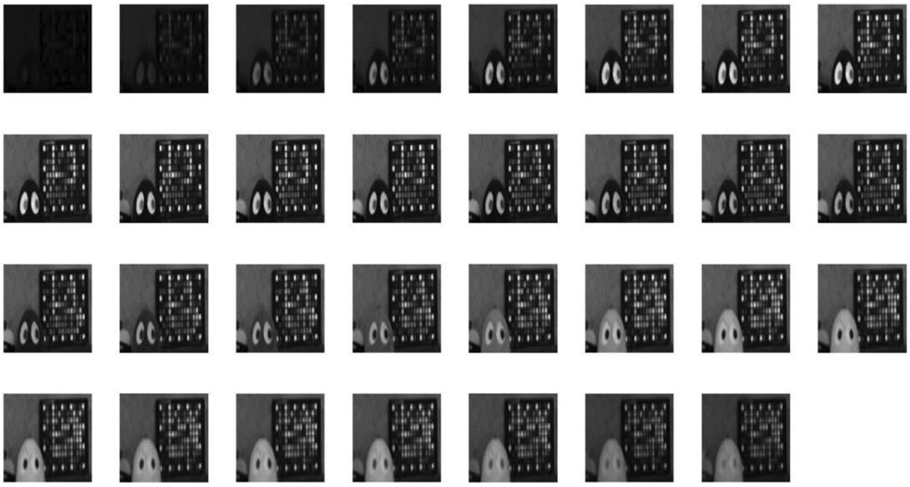


Fig. 1 Hyperspectral image display map

compared to hyperspectral images. Therefore, the feature extraction module of the traditional neural network target detection model pays less attention to the spectral information and lacks feature extraction for feature extraction between channels.

3.2 Target detection network analysis

The target detection task can be divided into two operations, which are target localization and target classification. Target localization is to detect the location of the target in the image and output the coordinates of the target box, which are continuous data and belong to the regression task. Target classification is to classify the targets in the target frame, and the predicted values belong to discrete data, and only a specified number of categories are predicted for the targets. However, in the neural network target detection model, the target coordinate frame and the target category prediction values are often output only in the last convolution operation at the same time, and the two different task operations increase the difficulty of convolution prediction, so it is also necessary to improve the output module of the target detection model by using two convolution operations, respectively.

Deep learning-based algorithms for target detection can be broadly divided into two-stage and single-stage methods. They have similar frameworks, and both the two-stage and single-stage algorithms can be divided into the structure shown in Fig. 2. Only for different types of detection models, the designed improvements focus on different aspects. The first-order model is an end-to-end model, where the input data is passed through the neural network and the prediction results are directly output.

The backbone module is used to extract features from the input data. Common backbone networks include VGG16, ResNet50, CSPResNeXt50, and CSPDarknet53-[24–27]. Usually, some functional layers are inserted in the middle of the backbone and head modules, which are used to collect feature mappings from different stages for fusion. These functional layers are called neck modules. For single-stage algorithms, the DensePrediction module completes the regression prediction of the prediction frame and categories, while for two-stage algorithms, further regression operations are required on the preselected frame [28]. As Sparse Prediction is shown in Fig. 2, the detection time of the second-order model therefore increases. However, all these models only process target detection on color images, which have only three channels, while hyperspectral images have tens or even hundreds of channels, and a complete and continuous spectral curve can be extracted from each pixel point of hyperspectral images. In order to extract

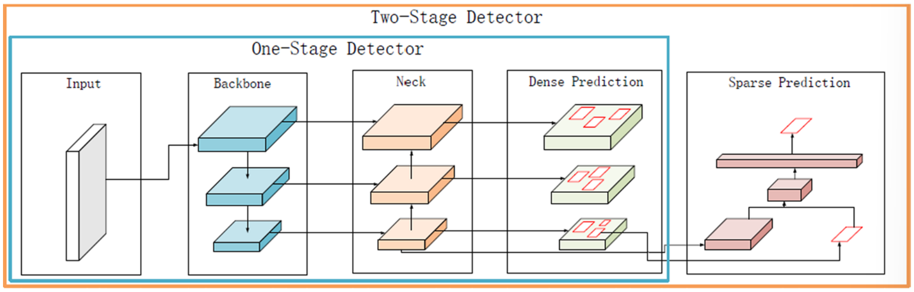


Fig. 2 Target detection architecture

hyperspectral image features effectively, it is necessary to improve the structure of traditional target detection models.

In the target detection model structure, the backbone module is used to obtain the feature map by convolution operation on the input data with convolution kernels. However, the convolution operation only performs feature extraction on the image space dimension. For hyperspectral images, the spectral information of matter is also an important basis for judging the target, so it is necessary to operate on the data channel dimension in the feature extraction module of the model.

In the hybrid-CNN network model [29], a joint null-spectrum operation is used for feature extraction of the spatial dimension and spectral segment dimension in the classification operation of the target. It employs an attention mechanism that allows the network to adaptively weight the feature information to highlight a certain important channel feature information. Meanwhile, attenuating irrelevant channel characteristic information. The network accuracy improves a lot. Although what we do in this paper is target detection, its method can also be borrowed by introducing the attention mechanism in the feature extraction module of target detection to improve the feature extraction capability for the channel dimension of the input data.

In summary, in order to solve the problems of complex feature compression and poor adaptability of traditional target detection, this paper will research on how to adopt neural network methods for target detection of hyperspectral images. For the neural network method, most of the target detection models are for color images. In order to be able to utilize the unique spectral features of hyperspectral images, we consider introducing an attention mechanism in the feature extraction stage to improve the feature extraction ability of the network between channels. We also decouple the target localization task and the target classification task in target detection and try to use two convolutional operations in the output module to predict the two tasks separately.

4 Proposed methodology

In this paper, we propose a neural network-based hyperspectral target detection model, whose network structure is shown in Fig. 3. The overall network structure of SEHyp consists of a convolutional neural network with a first-order target detection framework,

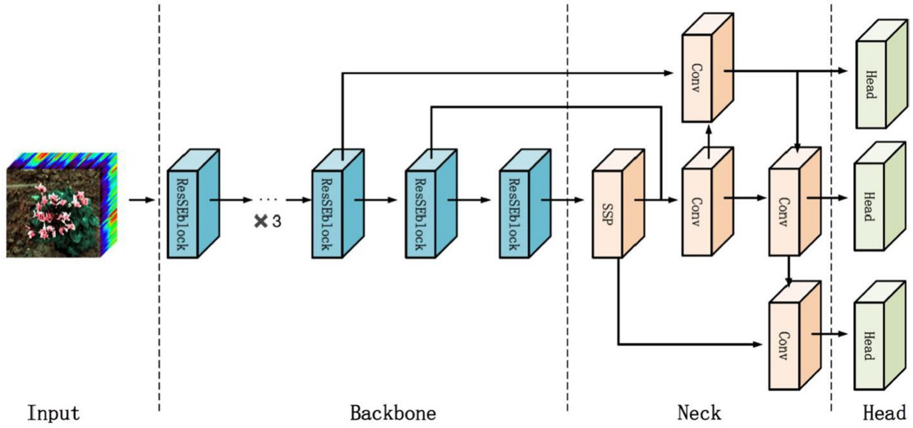


Fig. 3 SEHyp network structure

which contains three major network modules: backbone module, neck module, and head module. The backbone module consists of six ResSEblock blocks to extract the feature information from the input data. The neck module consists of SPP blocks and a pyramidal convolutional structure, which uses different stages of feature information output from the feature extraction module as input, then performs fusion of different feature information to achieve detection of targets of different sizes. Finally, the head module is the output module of the network model, which is used to output the prediction value, including the coordinates of the target frame, the target category and the confidence level, which is used to determine whether there is a target in the target frame.

4.1 SEHyp backbone

As mentioned in the introduction part for the spectral characteristics of hyperspectral images, modifications are needed in the feature extraction module, so the attention residual module (ReSEblock) is proposed in this paper, and its network structure is shown in Fig. 4. The overall structure of the network consists of a residual network, and an attention mechanism is added to the final output link in order to be able to extract feature information between channels.

ResSEblock contains N Resblocks. N is the number multiplied by Resblock in Fig. 4. The number of channels is halved by the convolution of the input feature map, and then fused on both sides. This not only reduces parameters, but also reduces computation, while the number of channels remains the same. Finally, the output of the ResSEblock also goes through the SE block. It selectively emphasizes informative features through an attention mechanism. Nonlinear features between spectral bands are effectively extracted by selectively emphasizing informative features.

Resblock mainly consists of 1×1 and 3×3 convolution kernels. It uses 1×1 convolution kernels to compress the number of channels. This not only reduces the number of parameters of the model, but also reduces the number of computations of the model. The residual structure can effectively prevent gradient explosion and gradient disappearance as the number of network layers is added. It operates by jumping connections, adding the input feature map data to the output feature map data, and then transferring the

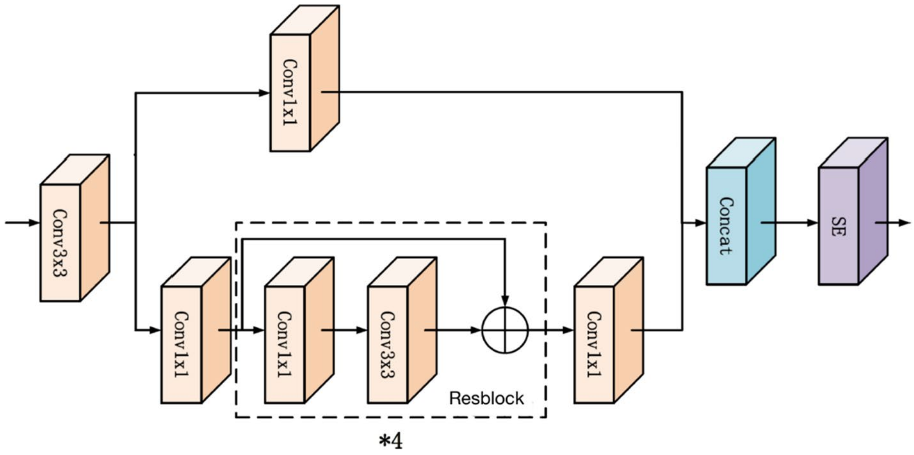


Fig. 4 ResSEblock module structure

result to the next layer. Finally, nonlinearity is introduced with the Mish activation function. The Mish formula is shown in formula (1)

$$Mish = x * tanh\left(ln\left(1 + e^{x}\right)\right) \tag{1}$$

The SE block (Sequeeze-and-Excitation Block) was proposed by Jiehu et al [30]. It is an implementation of the attention mechanism which can improve the response of channel features. The SE module adaptively recalibrates the representation of feature channels. It learns to use global information to selectively enhance channel feature representations and suppress useless parts.

The structure of the SE module is shown in Fig. 5. The whole SE module can be divided into three steps. First, global average pooling is performed on U, outputs channel eigenvalues of 1 × 1 × C size. The data were then subjected to two 1 × 1 convolution operations. The first convolution compressed C channels into C/r channels and used the ReLU activation function to add nonlinearity to the data, where *r* is the compression ratio; the second convolution uses the sigmoid activation function to restore the channel to the C channel again. The obtained 1 × 1 × C weight data are multiplied with the input feature map of the corresponding channels as the next level input feature map. The mathematical formula is as follows:

$$Out = X * Sigmoid\left(F_2\left(ReLU\left(F_1\left(F_{sq}(X), W_1\right)\right), W_2\right)\right) \tag{2}$$

W_1 and W_2 are parameters in the convolution operation, and $F1(\cdot, \cdot)$ and $F2(\cdot, \cdot)$ are convolution operations. The SE module selectively emphasizes informative features to enhance important channels and weakens non-important channels to improve the representation of features by learning and adaptively weighted features. Therefore, adding an attention mechanism to the feature extraction network allows the network to adaptively weight the feature information to highlight important features. The accuracy of the network can then be improved.

4.2 SEHyp neck

The neck of SEHyp model consists of SPP module and pyramid structured convolutional layers. The feature information is fused by different pooling operations and upsampling or downsampling and finally outputs the fused feature *Si*. The SSP module uses 1 × 1, 3

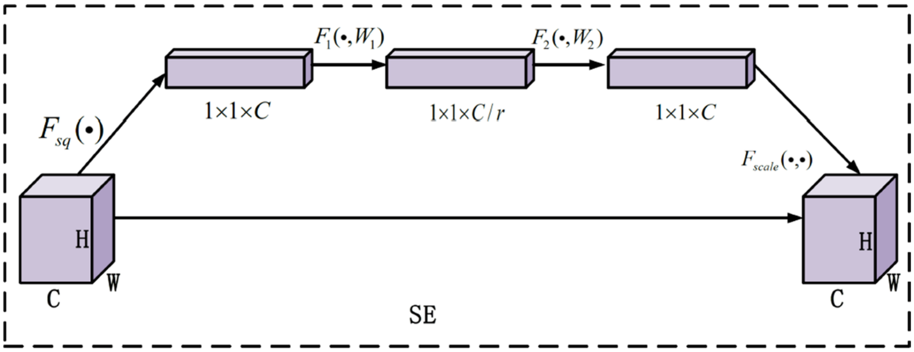


Fig. 5 SE module structure

× 3, 5 × 5 and 13 × 13 pooling kernels for max pooling. The feature values get different perceptual field by pooling of different sizes. This allows different feature information to be obtained, therefore improving the detection capability of the network for small targets and the localization accuracy.

The role of the neck module of the SEHyp model is to fuse different feature information, enabling the network to improve the detection of targets of different sizes. The main work of this module is to fuse the feature information extracted from the backbone module. As shown in Fig. 6, the ith ReSEblock block in the backbone module is represented by $\mu i(xi|wi)$, where xi denotes the input data of the ith block and wi denotes the network parameters of the block. Use $Q(a|wQ)$ for the neck module, where a is the input data of the neck module and wQ is its network parameters. The output feature values $\mu i(x5|w5)$, $\mu i(x6|w6)$ and $\mu i(x7|w7)$ of the backbone module are used as the input of the neck module parameter a, the different stages of the convolution layer obtain different feature information, the perceptual field of each pixel point is different, and the perceptual field obtained at different depth network structures will increase accordingly, so inputting feature information with different perceptual fields to improve the feature fusion efficiency can enhance the accuracy of different size target detection.

4.3 SEHyp head

The SEHyp head module is an output module of the model. For the model output module coupling problem, as shown in Fig. 7, the two branches are used to predict the target class and coordinates.

Object classification determination is a classification problem, while object location prediction is a regression problem. If both types of predictions use a convolution operation, the information is combined. This can make regression more difficult. Moreover, the spectral information unique to the hyperspectral map plays a crucial role in the detection, so two convolutional branches are used here for classification and coordinate prediction respectively. Two parallel branches are used to do the prediction of two tasks separately, so that different head modules can do their respective tasks and reduce the difficulty of prediction regression.

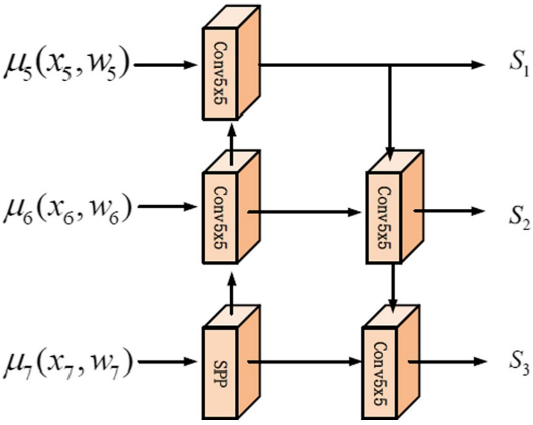


Fig. 6 SEHyp neck module structure

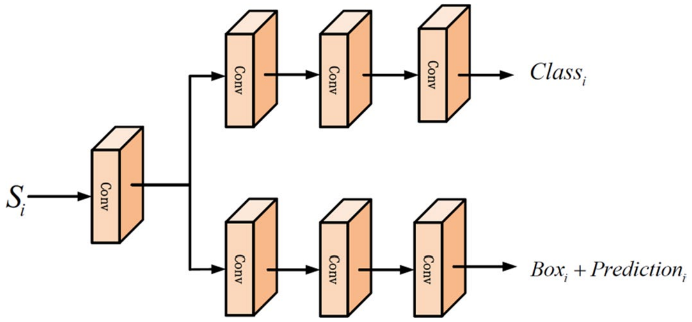


Fig. 7 SEHyp head module structure

The head module is denoted by $H(Si|wHi)$, Si is input data of the head module, which is the feature information output by the neck module. The results *Classi*, *Boxi* and *Predicioni* output finally, where *Classi* is the probability of each category, *Boxi* is the coordinate of the center point of the target box with the box length and width values, *Predicioni* is the confidence level, which is used to determine whether the target exists in the box. The output of the head module, the target classification prediction output sizes are (52, 52, ClassNum × 3), (26, 26, ClassNum × 3) and (13, 13, ClassNum×3). Here, ClassNum is the number of models that can be classified, and 3 is the three prediction frames that the model predicts for each pixel. The target coordinates lose the predicted output sizes for (52, 52, 15), (26, 26, 15) and (13, 13, 15). 15 is calculated from the coordinate point and confidence from the three boxes, whereas the confidence level is used to determine if there is a target in the box.

4.4 Loss function

In order for the model to perform the inverse process, a loss function is also required to calculate the difference between the predicted and true values. The size of the final predicted output value is $K \times K \times ((\text{ClassNum} + 5) \times 3)$, where the side length of the grid is K. Therefore, the grid has a total of $K \times K$ grids. Each grid predicts three prediction frames. And each predicted frame requires the predicted class, the vertex coordinates of the frame, and the confidence. The loss function is shown in formula (3).

$$\begin{aligned}
\text{loss}\left(\text{object}\right) = {} & \lambda_{\text{coord}} \sum_{i=0}^{K\times K} \sum_{j=0}^{M} I_{ij}^{\text{obj}} (2 - w_i \times h_i)[1 - CIOU] \\
& - \sum_{i=0}^{K\times K} \sum_{j=0}^{M} I_{ij}^{\text{obj}} \left[\hat{C}_i \log\left(C_i\right) + (1 - C_i)\log(1 - C_i) \right] \\
& - \lambda_{\text{noobj}} \sum_{i=0}^{K\times K} \sum_{j=0}^{M} I_{ij}^{\text{obj}} \left[\hat{C}_i \log\left(C_i\right) + \left(1 - \hat{C}_i\right)\log(1 - C_i) \right] \\
& - \sum_{i=0}^{K\times K} \sum_{j=0}^{M} I_{ij}^{\text{obj}} \sum_{c\in\text{classes}} \left[\hat{p}_i(\text{c})\log(p_i(\text{c})) + \left(1 - \hat{p}_i(c)\right)\log(1 - p_i(c)) \right]
\end{aligned} \tag{3}$$

The loss function I_{ij}^{obj} is used to determine whether there is a target in the *j*th prediction box of the ith grid. When it is 1, there is a target. When it is 0, there is not. Therefore, when there is no target in the grid, only the fourth row of the formula is calculated that means only the confidence loss is calculated. The first two lines of the loss function are the loss function for the predicted frame. The CIOU algorithm [15, 31] was used. For the traditional IOU loss function [32], if the two boxes do not intersect, the distance between the two boxes cannot be reflected, which means the loss is 0. It does not accurately reflect the size of the overlap of the two boxes. The CIOU loss function solves the problem that the loss is 0 when the two frames do not overlap by calculating the Euclidean distance between the centroids of the two frames. It also increases the scale loss of the predicted frame. Thus, the regression accuracy is improved. This improves the accuracy of the regression. λ_{coord} is the weight coefficient, K is the side length of the grid, M is the number of predicted frames per grid, w and h are the width and height of the predicted frame, and x and y are the coordinates of the grid center point of the predicted frame. The formula for calculating confidence loss is in the third and fourth lines. Confidence loss is calculated using cross-entropy. The loss value is still calculated when there are no objects in the grid. But its share in loss is controlled by λ_{noobj} weights. The last line of the equation is the loss function for the class. The cross-entropy loss function is used. But the class loss is only calculated if there are targets in the grid.

In summary, the learning process of the hyperspectral target detection network model is as follows.

1. Data processing is performed on the training data.
2. Input the data into the network model.
3. Perform data feature extraction by the backbone module.
4. The feature values extracted from the previous module are fused with the features by the neck module.
5. Input the fused features into the head module to make the final output data prediction.
6. Calculate the loss function from the corresponding image labels and predicted values, and update the network parameters.
7. Repeat steps (2)–(6) until the network converges or the training count is completed.

5 Experimental results and analysis

This section describes the training regime and experiments for our models.

5.1 Experimental setting

In this section, we conduct simulation experiments to verify the deep learning-based hyperspectral target detection algorithm. The data categories used in the experiments are six different categories of shoes. The object coordinates and category labels in the images are stored in an XML file which is in the same format as the coco dataset labels. The data source is divided into two parts. One is the real hyperspectral image acquired by the hyperspectral image acquisition system I built in the optical laboratory, and the other is the hyperspectral image generated by the AWAN algorithm.

The hyperspectral acquisition system is an image acquisition system based on the principle of single-slit push-broom spectral imaging. The initial model of the system is shown in Fig. 8. The AWAN algorithm was proposed by Li Yunsong et al. [33]. It performs spectral reconstruction from RGB images based on deep learning. The paper proposes the AWCA module. This is an adaptive weighted attention mechanism. It can improve the reconstruction accuracy of the network. Thus, the simulation accuracy of RGB image to hyperspectral image is improved. The AWAN algorithm is used to convert the acquired RGB images into hyperspectral images. This expands the training dataset.

There are 10236 hyperspectral images which are divided into 6 categories. The spectral bands range from 400 to 700 nm, separated by 10 nm. There are a total of 31 spectral bands, which are stored in mat file format. Eighty percentage of the dataset is used as the training set, 10% as the validation set, and 10% as the test set. The training set is used to train the neural network model. The validation set is used to verify the effect of network training. The test set is used to test the actual learning ability of the network.

5.2 Experimental evaluation index

The evaluation metrics used in this experiment mainly include: Precision, Recall, Average Accuracy (AP), and Mean Average Precision (mAP).

(1) IOU: The IOU formula is used to determine the similarity of two rectangular boxes, as shown in formula (4). When calculating AP, it is usually necessary to state at what IOU value the average correct rate is. For example, AP50 means that when the IOU is greater than or equal to 0.5, the prediction box has selected the target.

$$\mathrm{IOU} = \frac{\text{Area of intersection of two rectangular boxes}}{\text{Area of two rectangular boxes merged}} \tag{4}$$

(2) Precision, Recall, and F1: As shown in Fig. 9, the True class is the true value, and the Hypothesized class is the predicted value. Y is the positive class and N is the negative class. TP represents the probability that the model predicts that the target exists and the true target also exists. TN represents the probability that the model predicts that the target does not exist and the true target does. FP represents the probability that the model predicts that the target exists and the real target does not. TN represents the probability that the model predicts that the target exists and the true target does not. Precision is

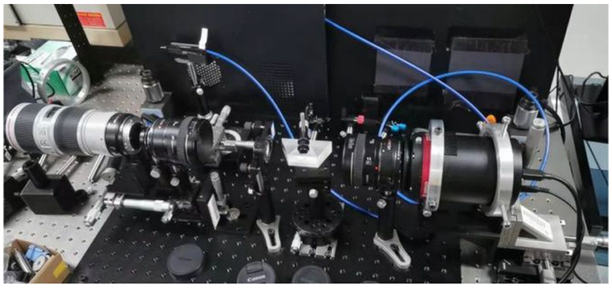
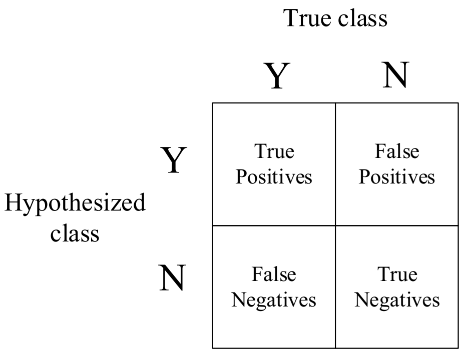


Fig. 8 The process of building a hyperspectral image acquisition system



		True class	
		Y	N
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives

Fig. 9 Schematic diagram of evaluation indicators

the probability of the positive class, which is the percentage of true classes among the positive classes predicted by the model shown in Eq. (5). Recall is the ratio of the positive class detected by the model to the true class. It is the percentage of true classes detected by the model which is shown in Eq. (6). F1 score can combine the balance of recall and precision, which can well distinguish the advantages and disadvantages of algorithms. It is shown in Eq. (7).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

(3) AP and mAP: AP (Average Precision) is the average precision. Precision and Recall tend to be mutually exclusive. When Precision was little, Recall was large. Precision was large when Recall was little. AP balances the two well using the surface area under the Precision and Recall curves. The area under the curve is the AP value for that category. The larger the area, the more accurate the model is for that class. mAP (mean Average Precision) is the mean average precision. It is used to measure the performance of the model in all categories. For object detection systems, usually multiple objects are detected. AP is an evaluation index for a single category. Therefore, mAP obtained by calculating the average of APs of all categories can effectively measure the quality of the model for all categories.

5.3 Experimental environment and procedure

This experiment was performed on Ubuntu 20.4. Pytorch is used to build deep learning models. The main software components used are: torch version 1.2, torchvision version 0.4, tqdm version 4.60, opencv_python version 4.1.2.30, h5py version 2.10, etc. The server hardware configuration is: Intel Xeon 6226R processor, 256G RAM, 11TB hard drive capacity, two GeForce RTX3090.

For model training, the backbone network uses pretrained weights. First, a backbone network is used for classification training on RGB images. The weights of the backbone network will not be too random. It is helpful for the training weights of the later target detection model to converge better. The input size of the hyperspectral object detection model is 416 × 416 × 31. The training method is freezing training. It is divided into two stages: freezing period and thawing period. Freeze period freeze the backbone network. That is, the parameters of the feature extraction network do not change during freezing. Only the parameters of neck and head are fine-tuned. The epoch number is 50 and the learning rate is 0.001. The learning rate is changed by the cosine annealing method. During the thawing period, the backbone network is thawed. The entire model parameters were trained with 50 epochs and a learning rate of 0.0001. It will decrease as the number of training cycles increases. The optimizer used for model training is adaptive moment estimation (Adam) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight_decay $= 0.0005$.

5.4 Experimental results

In this experiment, the detection performance of three models, YHyp network with only modified inputs, YSE network with added attention mechanism, and SEHyp network designed in this chapter is tested.

Figure 10 illustrates the change of loss function values between the training set and the test set during the training process. It can be seen that the loss function decreases faster initially, which is due to the fact that during the training process, migration learning is used. The method first trains the feature extraction module in a classification task and then migrates it to the model. So the freeze training is performed first to freeze the parameters of the feature extraction module and train only the parameters of the neck and head modules, Therefore, after several training sessions, the network can quickly adapt to the target detection task and the loss function decreases faster. When the training reaches 50 times, the training will be unfrozen and the parameters of the feature extraction module will be changed, so it can be found from the figure that the loss has a large decline after the 50th training.

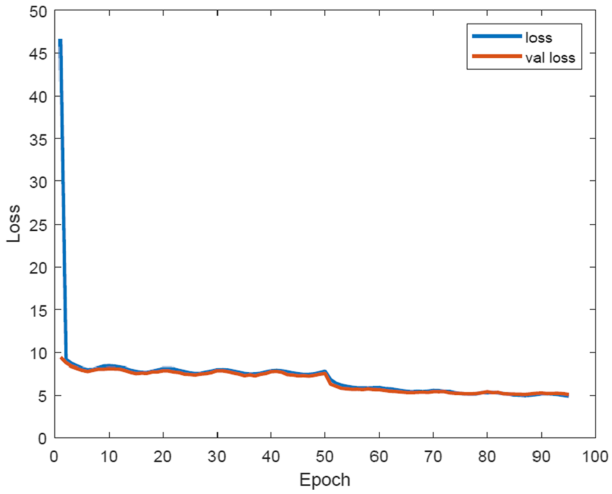


Fig. 10 Network model training process

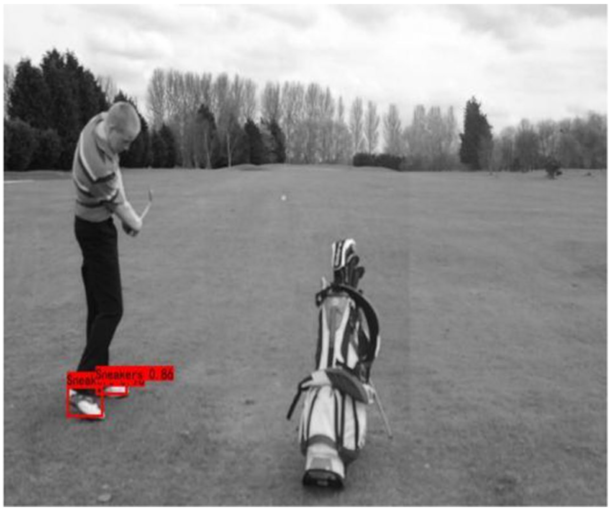


Fig. 11 SEHyptarget detection results (19th spectrum)

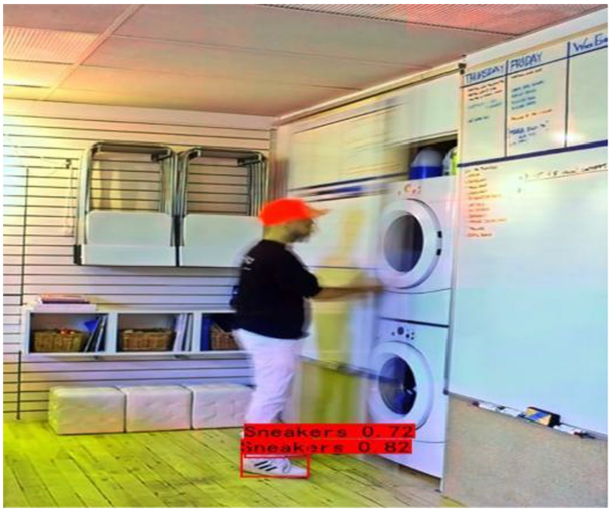


Fig. 12 Pseudo-RGB images 1(29th, 19th, and 9th spectrum)

The detection results are shown in Figs. 11, 12, 13, and 14, which is part of the results detected by the SEHyp model. Figures 12, 13, and 14 show the pseudo-color image, because the hyperspectral image has 31 spectral bands containing visible light from violet to red, which cannot be displayed directly, so the 9th spectral band, the 19th spectral band, and the 29th spectral band are extracted from the hyperspectral image to form the RGB triple channel of the color image, so the color image is displayed. In Fig. 11, the separate results are also presented in the grayscale image of the 19th spectral band. The detection results are displayed in the 19th spectrum, which has a better light intensity in the middle, because the appearance of the target is not well displayed in the front of the visible spectrum in hyperspectral images.

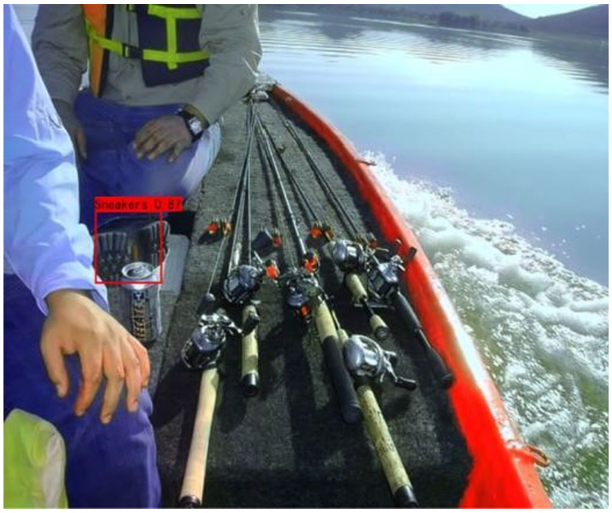


Fig. 13 Pseudo-RGB images 2(29th, 19th, and 9th spectrum)

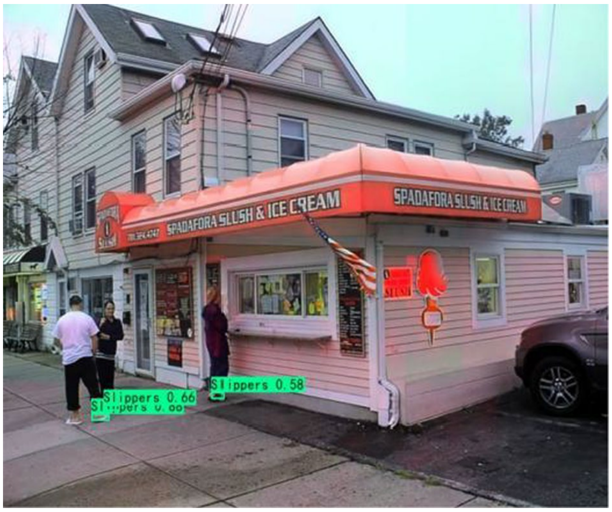


Fig. 14 Pseudo-RGB images 3(29th, 19th, and 9th spectrum)

As shown in Figs. 15, 16, and 17, three metrics of target detection effectiveness are illustrated in the figure, Fig. 15 shows the value of F1 for target detection, Fig. 16 shows the value of Precision for target detection, and Fig. 17 shows the value of recall rate for target detection. When the IOU threshold value is taken as 0.5, F1 and recall rate can be found by the figure that the present method has been improved considerably, which indicates that the overall effectiveness of the model with the ability to detect the presence of targets has been increased. For the precision value, the method is lower than the detection precision of the YSE model, but its value still reaches 86.71%, and the accuracy of the detected targets is high.

The results of the three network tests are shown in Table 1. APss, mAP50, and mAP75 are also shown here as evaluation metrics. APss is the value of AP50 in the

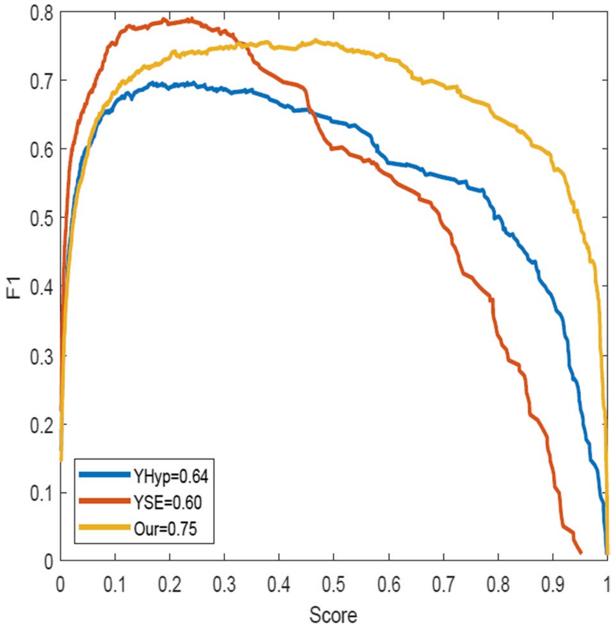


Fig. 15 Value of F1 in each model

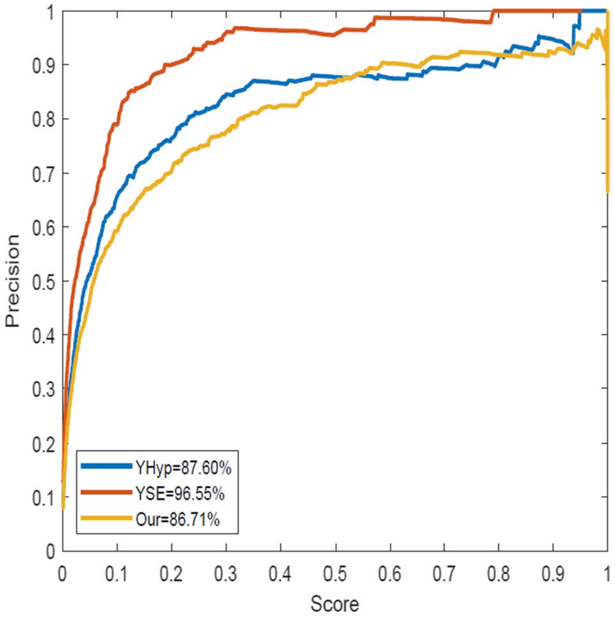


Fig. 16 Value of Precision in each model

Skating and Skiing category of the detection dataset. mAP50 and mAP75 are the average value of AP when IOU is set to 0.5 and 0.75. We also count the detection time on the GPU to judge the model performance. As the attention mechanism SE is added to the backbone network, the value of AP for skating and skiing category is shown in Fig 18. Compared with YHyp, the overall method has improved, especially in the Recall value greater than 0.8, the Precision has improved more. Compared with YSE,

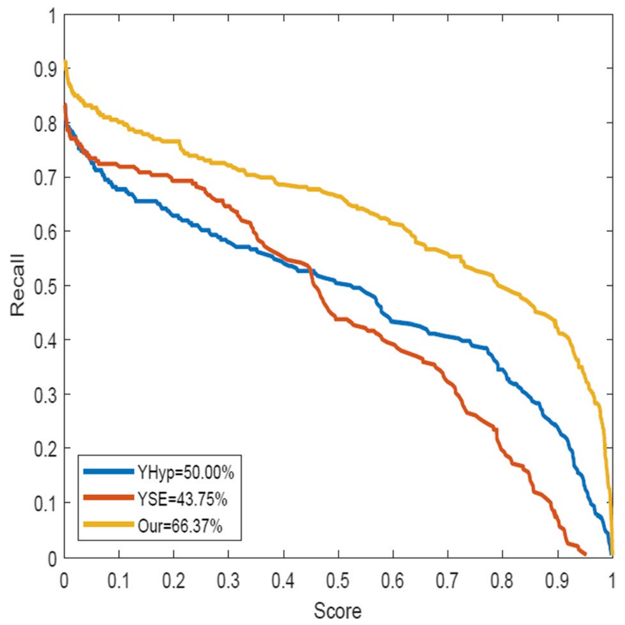


Fig. 17 Value of Recall in each model

Table 1 Model test results

Method	F1$_{50}$	Pre$_{50}$	Recall$_{50}$	AP$_{ss}$	mAP50	mAP75	detection times (s)
YHyp	0.64	87.60%	50.00%	70.64	46.64	34.52	0.86
YSE	0.60	96.55%	43.75%	75.45	50.21	40.28	0.92
Our	0.75	86.71%	66.37%	76.25	51.10	40.84	0.93

the area of AP decreases when Recall is less than 0.7, but the area of AP is improved. Overall, the value of APss is increased by 5%, the mAP value is significantly improved. mAP50 value is increased by 3.57%. But the detection time increases with the increase of backbone network parameters. The overall time is within 1 second, which can meet the requirements of the IoV system. In the case of the improved head module for binary branch prediction, the value of mAP50 is improved. But the time did not add much. Because the number of channels is reduced by using a 1 × 1 convolution operation before entering branch prediction. And the subsequent computations are also reduced. It can be seen that the hyperspectral target detection model designed in this paper can effectively perform the hyperspectral target detection task of IoV applications. Although the time-consuming has increased, it is still within the acceptable range for the IoV system.

6 Conclusion

It can be seen that the hyperspectral target recognition model based on deep learning developed in this paper can effectively perform the task of hyperspectral target recognition. Compared with other traditional detection algorithms, the algorithm not only can be processed for the collected hyperspectral images, thus greatly improving the detection

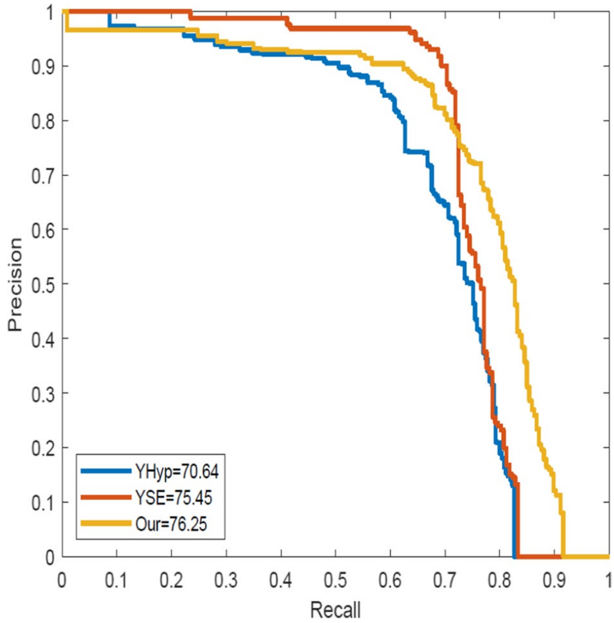


Fig. 18 The value of AP in the Skating and Skiing shoes class

accuracy in the process of IoV applications. The introduction of the attention mechanism makes the cost of time small, which is crucial and can meet the reliable requirement of intelligent applications of IoV.

A two-order target detection architecture model can be added in the future. Although the two-order target detection architecture model is computationally time-consuming, the detection accuracy is generally higher than that of the first-order target detection model.

Abbreviations

IoV	Internet of vehicle
RGB	Red, green, and blue
SE	Sequeeze-and-excitation
SEHyp	Sequeeze and excitation for hyperspectral target detection
mAP	Mean average precision
AP	Average accuracy

Acknowledgements

The authors acknowledged the anonymous reviewers and editors for their efforts in valuable comments and suggestions.

Author contributions

Zixu Wang proposes the theoretical analysis and carries out experiments. Lizuo Jin conceived of the study and participated in its design and coordination and helped to draft the manuscript. Kaixiang Yi is responsible for the experimental part of the update and processing. All authors read and approved the final manuscript.

Funding

Our work is supported by the'Research and Development of Coal Mine Big Data Platform with Artificial Intelligence Applications' program of China Coal Energy Research Institute Co. Ltd.

Availability of data and materials

The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 25 May 2022 Accepted: 15 August 2022
Published online: 05 September 2022

References

1. L. Liu, M. Zhao, M. Yu, M.A. Jan, D. Lan, A. Taherkordi, Mobility-aware multi-hop task offloading for autonomous driving in vehicular edge computing and networks. IEEE Trans. Intell. Transp. Syst. (2022). https://doi.org/10.1109/TITS.2022.3142566
2. S. Mao, L. Liu, N. Zhang, M. Dong, J. Zhao, J. Wu, V.C. Leung, Reconfigurable intelligent surface-assisted secure mobile edge computing networks. IEEE Trans. Veh. Technol. (2022). https://doi.org/10.1109/TVT.2022.3162044
3. M. Li, J. He, R. Zhou, L. Ning, Y. Liang, Research on prediction model of mixed gas concentration based on CNN-LSTM network, In: 2021 3rd International Conference on Advanced Information Science and System (AISS 2021). Association for Computing Machinery, New York (2021)
4. T. Zhu, H. Gu, Z. Chen, A median filtering forensics CNN approach based on local binary pattern. In: Proceedings of the 11th International Conference on Computer Engineering and Networks. Springer, Singapore, pp. 258–266 (2022)
5. Y. Gu, J. Li, A novel WiFi gesture recognition method based on CNN-LSTM and channel attention, in: 2021 3rd International Conference on Advanced Information Science and System (AISS 2021). Association for Computing Machinery, New York (2021)
6. W. Wang, S. Dou, Z. Jiang et al., A fast dense spectral–spatial convolution network framework for hyperspectral images classification. Remote Sens. **10**(7), 1068 (2018)
7. W. Ma, Q. Yang, Y. Wu et al., Double-branch multi-attention mechanism network for hyperspectralimage classification. Remote Sens. **11**(11), 1307 (2019)
8. L. Dong, Q. Ni, W. Wu, C. Huang, T. Znati, D.Z. Du, A proactive reliable mechanism-based vehicular fog computing network. IEEE Int. Things J. **7**(12), 11895–11907 (2020)
9. L. Dong, W. Wu, Q. Guo, M.N. Satpute, T. Znati, D.Z. Du, Reliability-aware offloading and allocation in multilevel edge computing system. IEEE Trans. Reliab. **70**(1), 200–211 (2021)
10. J. Yu, M. Gao, Y. Li, Z. Zhang, W.H. Ip, K.L. Yung, Workflow performance prediction based on graph structure aware deep attention neural network. J. Ind. Inf. Integr. **27**, 100337 (2022)
11. J.C. Harsanyi, Detection and classification of subpixel spectral signatures in hyperspectral image sequences, Baltimore County (1993)
12. S.S. Chiang, C.I. Chang, I.W. Ginsberg, Unsupervised target detection in hyperspectral images using projection pursuit. IEEE Trans. Geosci. Remote Sens. **39**(7), 1380–1391 (2001)
13. J.H. Friedman, W. Stuetzle, Projection pursuit regression. J. Am. Stat. Assoc. **76**(376), 817–823 (1981)
14. L. Wang, J. Peng, W. Sun, Spatial-spectral squeeze-and-excitation residual network for hyperspectral image classification. Remote Sens. **11**, 884 (2019)
15. J. Liu, W. Zhang, Z. Wu, A distributed and parallel anomaly detection in hyperspectral images based on low-rank and sparse representation, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium(2018)
16. X. Ou, Y. Zhang, H. Wang, Hyperspectral image target detection via weighted joint k-nearest neighbor and multitask learning sparse representation. IEEE Access **8**, 11503–11511 (2020)
17. J. Li, H. Zhang, L. Zhang, L. Ma, Hyperspectral anomaly detection by the use of background joint sparse representation. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **8**(6), 2523–2533 (2015)
18. W. Lan, V. Baeten, B. Jaillais, Comparison of near-infrared, mid-infrared, raman spectroscopy and near- infrared hyperspectral imaging to determine chemical, structural and rheological properties of apple purees. J. Food Eng. **323**, 111002–111002 (2022)
19. P. Shi, Q. Ling, J. Wu, Z. Lin, Collaborative representation based on the constraint of spectral characteristics for hyperspectral anomaly detection, in 2021 3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC), pp. 199–204 (2021)
20. G. Squeo, D. De Angelis, C. Summo, Assessment of macronutrients and alpha-galactosides of texturized vegetable proteins by near infrared hyperspectral imaging. J. Food Compos. Anal. **108**, 104459–104459 (2022)
21. T. Mu, R. Nie, C. Ma, J. Liu, Hyperspectral and panchromatic image fusion based on CNMF, in 2021 3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC), pp. 293–297 (2021)
22. L. Yan, M. Zhao, X. Wang, Object detection in hyperspectral images. IEEE Signal Process. Lett. **28**, 508–512 (2021)
23. L. Dong, M.N. Satpute, W. Wu, Two-phase multi-document summarization through content-attention-based subtopic detection. IEEE Trans. Comput. Soc. Syst. **8**(6), 1379–1392 (2021)
24. J. Redmon, A. Farhadi, Yolov3: an incremental improvement (2018)
25. K. He, X. Zhang, S. Ren, Deep residual learning for image recognition, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
26. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014)
27. S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp. 5987–5995 (2017)
28. L. Dong, Q. Guo, W. Wu, Speech corpora subset selection based on time-continuous utterances features. J Comb Optim **37**, 1237–1248 (2019)
29. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
30. J. Hu, L. Shen, S. G, Squeeze-and-excitation networks, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

31. Z. Zheng, P. Wang, W. Liu, Distance-iou loss: faster and better learning for bounding boxregression. Proc. AAAI Conf. Artif. Intell. **34**, 12993–13000 (2020)
32. J. Yu, Y. Jiang, Z. Wang, Unitbox: An advanced object detection network, in Proceedings of the 24th ACM International Conference on Multimedia, pp. 516–520 (2016)
33. J. Li, C. Wu, R. Song, Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 2020 (2020)

Publisher's Note

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

Zixu Wang is currently pursuing for a bachelor's degree with the school of Automation, Southeast University. His research interests include cooperative communication, big data processing, and mobile intelligent computing.

Lizuo Jin received his Ph.D. degree from Southeast University, China, in 2000. From 2002 to 2004, he was a post-doctoral fellow at Institute of Industrial Technology, Tokyo University, Japan. He is now an associate professor at School of Automation, Southeast University, and serves as a member of IEEE, SPIE, and a committee member of Information Fusion Technical Committee of CSAA, China. His research interests include theories and methods for machine learning, pattern recognition, computer vision, and embedded computation.

Kaixiang Yi is going to graduate from Undergraduated Military Educational Academy of National University of Defense Technology in 2022. His specialty is navigation engineering.