Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9797302, 16 pages
https://doi.org/10.1155/2020/9797302

Research Article

A Deep Convolutional Network for Multitype Signal Detection and Classification in Spectrogram

Weihao Li,[1,2] **Keren Wang**,[2] **and Ling You**[2]

[1]*PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China*
[2]*National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu 610041, China*

Correspondence should be addressed to Weihao Li; liweihao315@gmail.com

Received 13 May 2020; Revised 19 August 2020; Accepted 21 August 2020; Published 12 September 2020

Academic Editor: Paolo Crippa

Wideband signal detection is an important problem in wireless communication. With the rapid development of deep learning (DL) technology, some DL-based methods are applied to wireless communication and have shown great potential. In this paper, we present a novel neural network for detecting signals and classifying signal types in wideband spectrograms. Our network utilizes the key point estimation to locate the rough centerline of the signal region and recognize its class. Then, several regressions are carried out to obtain properties, including the local offset and the border offsets of a bounding box, which are further synthesized for a more fine location. Experimental results demonstrate that our method performs more accurate than other DL-based object detection methods previously employed for the same task. In addition, our method runs obviously faster than existing methods, and it abandons the candidate anchors, which make it more favorable for real-time applications.

1. Introduction

Wireless communication plays an import role in military and civilian life with its flexibility and long-distance transmission capability. The fast development of wireless communication and related technologies makes electromagnetic environment more and more chaotic. Cognitive radio technology that is capable of learning and adapting to environment has attracted a lot of attention [1, 2]. Automatic signal detection (SD) and signal classification (SC) are two of important tasks in cognitive radio which have been researched for decades. Since the novel deep learning (DL) technology performs well in image, speech, and natural language processing, it has also been introduced to wireless communication and brought about great improvements.

SD in this paper specifically refers to detecting signals in the wideband. It is also the basic task of the spectrum sensing [3] and the blind signal separation [4, 5]. For traditional SD methods, the energy detection has once been the most popular technique that can be classified as the threshold-based algorithms [6–11] and the non-threshold-based algorithms [12–16]. The former has lower computational requirements, but it is limited by the signal-to-noise ratio (SNR) with a high false alarm probability. The latter improves accuracy and universality at the cost of computational complexity. Recently, some DL-based methods have been applied in narrowband or wideband environment. In [17, 18], the convolutional neural network (CNN) + long short-term memory network (LSTM) and deep belief network (DBN) were adopted, respectively, to detect signals in the narrowband, with the input of raw data and spectral correlation function (SCF). In [19], researchers selected the narrowband containing signals from a wideband spectrogram by energy detection, and then they utilized CNN to classify the wanted Morse signals. Those methods have achieved excellent detection effects, but their main purposes are mostly to detect the existence of signals, without the time or frequency information.

For SC task, the process is generally to first extract signal features and then design a classifier for recognition. The commonly used features for SC can be classified as instantaneous time features [20], statistical features (cumulants [21] and cyclostationarity [22]), transform features (Fourier transform [23] and wavelet transform [24]), other

features (constellation shape [25] and zero crossing [26]), and feature learning (raw data [27, 28]). The above features have different capabilities in antinoise and complexity, and they are applicable to different signal types. The classifiers mainly include earlier threshold-based algorithms [20, 23], traditional machine learning (ML) methods (support vector machine (SVM) [29, 30], k-nearest neighbor (KNN) [31], and fuzzy classifier [32]), and DL methods (deep neural network (DNN) [25, 33], convolutional neural network (CNN) [24, 27], long short-term memory network (LSTM) [34], and CNN + LSTM [28]). The earlier threshold-based algorithms are fast but depend greatly on the feature design and threshold selection, which need profound expert knowledge. The traditional ML methods reduce the model complexity, but they are sensitive to noise and also need elaborate design of features. DL methods achieve the feature learning on the raw or simply processed data, which obtain a good classification performance, under the precondition of rich training data.

Some research on the joint SD and SC has also been carried out. In [35], an algorithm based on first-order cyclostationarity was proposed for the FSK and AM signals. In [36], the key spectral features of narrowband signals were extracted to train the naive Bayes classifier for the modulation classification of signals and the detection of jamming. Researchers in [37] used the recurrent neural networks (RNN) to process the spectrograms of long-term signals. Recently, some researchers used the single shot multibox detector (SSD) network, which is a classical DL-based object detector, to achieve end-to-end SD and SC in a wideband spectrogram [38, 39]. SSD can find the specific location of the signals, including the start and end time and frequency, which is a promising and valuable approach for a further study.

Inspired by the SSD, in this paper, we convert the spectrogram-based SD and SC tasks to an object detection task, and we exploit the advantages of the DL in computer vision. The SSD and other commonly used object detection methods make predictions at a center point of object. However, since the aspect ratio of signals is usually large and varies dramatically, a signal region is usually beyond the receptive field of one point, which could lead to incomplete predictions. In addition, lots of candidate anchors of SSD reduce the real-time performance. Targeting the above shortcomings, we construct a deep convolutional network and implement end-to-end training in this paper. Since a signal region in spectrogram is usually a horizontally-long rectangle, we propose to model a signal by its centerline, and the signal type and a bounding box (BBox) are regressed from the features at centerline. Compared to most object detection methods, our network abandons the candidate anchors and uses the centerline instead of just a point to make predictions, which are more efficient and task-oriented (we will explain the defects of traditional object detectors and our improvements in detail in Section 2). Experimental results show that our method has a higher detection and classification accuracy, especially for the extremely long instances. Moreover, the simplicity of our network allows it to run at a very fast speed.

To summarize, the main contributions of our works are as follows: (1) we utilize the idea of DL-based object detection for multitype SD and SC in wideband spectrograms, which is capable of detecting time and frequency location and recognizing signal types; (2) different from applying the commonly used object detectors directly, we target the characteristics of signals in spectrograms and propose an improved convolutional network that uses the centerlines to locate the signals and abandons the candidate anchors, which makes our method more accurate and faster.

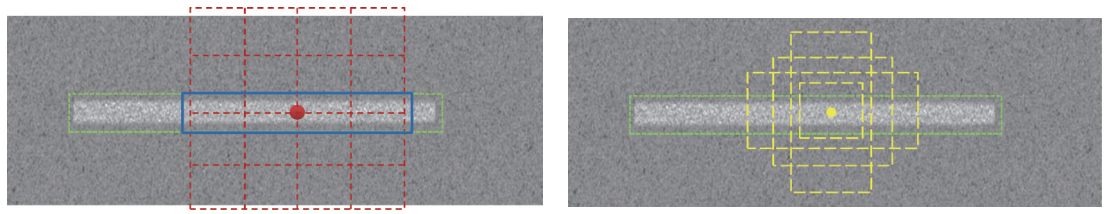
2. Related Work

We want to accomplish the multitype SD and SC in spectrogram by the idea of DL-based object detection. However, we think that the traditional detectors are not suitable for the task solved in this paper. Before us, the researchers in [38, 39] have used the SSD that is a commonly used detector to perform the same task, but the results are not satisfactory, especially for the extremely long instances. In this section, we will explain the defects of traditional detectors for the signals in spectrograms, and then we accordingly propose our centerline-based method.

2.1. Defects of Traditional DL-Based Object Detectors. Most DL-based object detectors [40–45] achieve good performance when the objects have regular shapes and aspect ratios. Nevertheless, the signals in a spectrogram usually have extremely long shapes, and their time duration and frequency band vary dramatically with different signals, which makes them quite different from the general objects. The traditional detectors tend to get frustrated, and two main reasons are considered: (1) Due to the limited receptive field of the CNNs, some one-point based detectors [42, 44, 45] that use only one point or a small area to predict box size cannot get complete BBoxes. (2) Many detectors raise multiple candidate anchors at each pixel in advance [40–44], but since the shapes of signals vary greatly, it is difficult to design a group of common anchors to match the signals, and the regression on those anchors is quite time-consuming.

Figure 1(a) is a detection result of SSD [38]. The green box is the ground truth box of a signal, and an incomplete box proposal as the blue box is predicted. SSD only utilizes the red point to make prediction, whose receptive field (the region in red grids) is smaller than the ground truth box. In addition, in Figure 1(b), the candidate anchors used in [38] are drawn as the yellow grids, and we can see that the shape of anchors differs greatly from that of the ground truth box, which causes the signal to have no suitable anchor to match during the training.

2.2. Signal as Centerline. In order to solve the above problems, we propose to model the signal as its centerline. Since a signal region is usually a horizontally-long rectangle, we first find the centerlines to locate each signal, and then the features in centerlines are utilized to predict the BBox sizes and signal types. The receptive field of a centerline can easily



(a) (b)

Figure 1: Two typical defects of traditional DL-based object detectors. (a) The limited receptive field of one point. (b) The shape mismatch between the anchors and ground truth boxes.

cover the entire signal, and we abandon the anchor generation, turning to predict the offsets between the centerline and (up/down) border lines directly, which avoids the shape mismatch of anchors and saves a lot of time. The principle of our method is visualized in Figure 2.

3. Data Generation

The amount and richness of dataset are crucial for the training of deep neural networks. Since the signal transmission modality in radio communication has a clear mathematical expression, we simulate the wideband signals by programs, and the data generating process is introduced in this section.

We select 2FSK, 4FSK, (PSK/QAM), Morse, speech, and resident noise (RN) as the signal types of tests that often appear in wideband. These signal types are intuitively distinguishable in the spectrograms except the MPSK and MQAM; thus we merge those two types to the (PSK/QAM). If a further classification on MPSK and MQAM is needed, some subsequent methods such as those in [25, 38, 46, 47] can be introduced.

3.1. Multitype Signal Model. The transmitted digital modulation signal can be modelled as

$$s(t) = \sum_n a_n e^{j(\omega_n t+\phi)} g(t - nT_b), \tag{1}$$

where a_n represents the transmitted symbols, ω_n is the angular frequency, ϕ is the carrier initial phase, $g(t)$ is the shaping filter, and T_b is the symbol period.

For the MFSK signal, it can be presented as

$$a_n = 1,\ \omega_n = \omega_0 + \frac{2\pi}{M} i, \quad i = 0, 1, \ldots, M-1. \tag{2}$$

For the MPSK signal, it can be presented as

$$a_n = e^{j2\pi i/M},\ i = 0, 1, \ldots, M-1, \quad \omega_n = \omega_0. \tag{3}$$

For the MQAM signal, it can be presented as

$$\begin{cases} a_n = I_n + jQ_n, \\ I_n, Q_n = 2i - \dfrac{M}{4} + 1, \quad i = 0, 1, \ldots, \dfrac{M}{4} - 1, \\ \omega_n = \omega_0. \end{cases} \tag{4}$$

For Morse, it can be presented as

$$a_n = 0, 1, \quad \omega_n = \omega_0. \tag{5}$$

For the RN, it is referring to an irrelevant signal with long duration, narrow bandwidth, and random energy changes. Here we present it as the signal with a single frequency and random amplitude change:

$$a_n = 0.5 \sim 1.5, \quad \omega_n = \omega_0, \quad g(t) = 1. \tag{6}$$

For the speech, we modulate the real-world audios to different frequencies by the amplitude modulation.

3.2. Wideband Spectrogram Generation. In the actual communication environment, the received signals of most systems can be expressed as

$$r(t) = e^{j*n_{Lo}(t)} \int_{\tau=0}^{\tau_0} s\left(n_{\mathrm{Clk}}(t-\tau)\right) h(\tau) + n_{\mathrm{Add}}(t). \tag{7}$$

This takes into account the effects of many factors in the real world. $n_{Lo}(t)$ represents the residual carrier random walk, $n_{\mathrm{Clk}}(t)$ represents the time deviation, $h(t)$ is the time-varying channel function, and $n_{\mathrm{Add}}(t)$ is the additive noise.

To make the synthetic data valuable enough, we perform simulation comprehensively in a way identical to the real situation. On the one hand, the pulse shaping and bit rate that are suitable for corresponding modulation modes are set up, and the real voice or text is modulated as transmitted data. On the other hand, a robust channel model is employed including the multipath fading, random frequency walk drifting, and additive white Gaussian noise (AWGN). We pass the synthetic wideband signals through the channel model to obtain the final experimental data.

To calculate the spectrograms of the wideband signals, we utilize the short-time Fourier transform (STFT), which is a common time-frequency analysis method. The calculation of STFT is

$$S_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} s(m) w(n-m) e^{-j\omega m}, \tag{8}$$

$$P_n(\omega) = \left|S_n\left(e^{j\omega}\right)\right|^2, \tag{9}$$

where $s(m)$ is the sampled signal, $w(m)$ is the window function, and $P_n(\omega)$ is the final time-frequency matrix. Figure 3 presents different types of signals in the wideband.

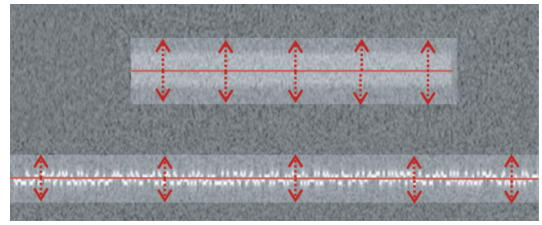


Figure 2: Modeling signal as centerline. The box size and signal type are predicted by the features at the centerline.

We annotate each signal with a ground truth box that is higher than its bandwidth, and the influence of ground truth box height will be discussed in Section 5.2.

4. Approach

The network of our approach is mainly composed of CNNs that perform well in the image recognition. The CNNs learn features via nonlinear transformations as a series of nested layers that introduce several kernels to perform convolution over the input. Generally, the kernels are multidimensional arrays that can be updated by some algorithms [48]. In this section, we first give an overview of our network, and then we elaborate two core modules, and finally we present the details of training and inference.

4.1. Overview. The overall architecture of our method is illustrated in Figure 4, which can be divided into two main parts. First, we extract shared feature maps for subsequent tasks by the backbone network. Our backbone network is a ResNet18 with three up-convolutions, where the features of different levels are effectively merged. Then, we adopt a shape and type expression module (STEM), which utilizes the shared features to predict the BBoxes and signal types. The STEM constructs a shape expression by learning geometry properties including the centerline, local offset, and border offsets. The details of backbone module and STEM are presented as follows.

4.2. Backbone Module. We use a ResNet18 with three up-convolutions as the backbone module to extract shared features, and its architecture is shown in Figure 5. The input image first passes through multiple forward convolutional stages whose structures are detailed in the dotted box on the left. In each convolutional stage, there are two blocks and each block consists of two convolutional layers and a residual structure to connect the input and output of a block. The residual structure is able to solve the gradient transfer problems during the deep network training. We introduce three transposed convolutions to upsample the output of forward convolutions, and each output of transposed convolutions is added with that of the corresponding convolutional stages. By merging the multiscales feature maps, we can make full use of the learned features at different level. The size of output feature map is (1/4) of the input image size. The batch normalization and ReLU activation function are following the convolutional layers, which are not marked in the figure.

4.3. Shape and Type Expression Module. The STEM is a multichannel convolutional network and can be divided into three branches. In each branch, we utilize two 3×3 and 1×1 convolutional layers with different channels to regress the signal property maps including the centerline, local offset, and border offsets.

4.3.1. Centerline. Centerline is a 7-channel (6 signal types + 1 background) map that represents the pixel-wise probabilities of different classes of centerlines. For the generation of ground truth centerline maps, we compute a low-resolution equivalent $\tilde{p} = (p/R)$ for each centerline point p of class c and then splat all $\tilde{p}$ onto a heatmap $Y \in [0,1]^{(W/R)\times(H/R)\times C}$ using a Gaussian kernel $Y_{xyc} = \exp(-((x-\tilde{p}_x)^2 + (y-\tilde{p}_y)^2/2\sigma_p^2))$, where $[W, H]$ are the width and height of input image, R is the downsampling scale, C is the number of signal types, and σ is an object size-adaptive standard deviation. The training objective of centerline is the pixel-wise focal loss:

$$L_{\mathrm{cl}} = \frac{1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right), & \text{if } Y_{xyc}=1,\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)\log\left(1-\hat{Y}_{xyc}\right), & \text{otherwise,}\end{cases} \tag{10}$$

where α and β are hyperparameters of the focal loss and N is the number of centerline points in a spectrogram. Here we chose $\alpha = 2$ and $\beta = 4$ in our all experiments.

4.3.2. Local Offset. Local offset is a 1-channel map that has valid values within the centerlines. To recover the discretization error caused by the downsampling of the backbone network, we predict a vertical local offset $\hat{O} \in R^{(W/R)\times(H/R)\times 1}$ additionally. The local offset will be added to the ordinate of centerline when mapping the shrunken image to the original size. The training objective of local offset is the L1 loss at centerline points:

$$L_{\mathrm{off}} = \frac{1}{\mathrm{N}}\sum_{p}\left|\hat{O}_{\tilde{p}} - \left(\frac{p_y}{R} - \tilde{p}_y\right)\right|. \tag{11}$$

4.3.3. Border Offsets. Border offsets is a 2-channel map that has valid values within the centerlines. The values of two channels $y_u^{(\tilde{p})}$ and $y_d^{(\tilde{p})}$ correspond to the offsets between the centerline and (up/down) border lines. $\hat{S}_{\tilde{p}} = y_u^{(\tilde{p})} - y_d^{(\tilde{p})}$ is

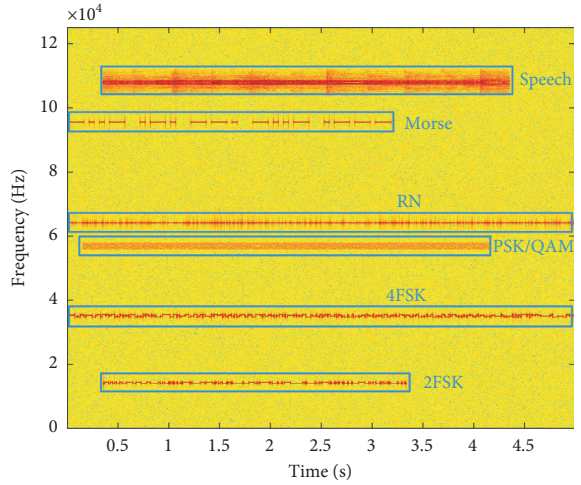


Figure 3: Wideband spectrogram with multitype signals.

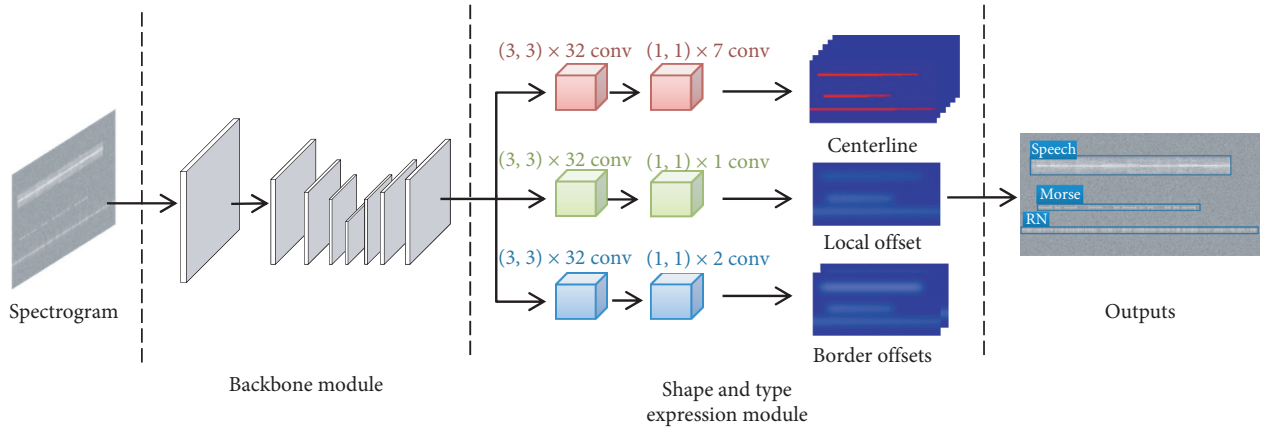


Figure 4: The overall architecture.

the height of the predicted BBox at $\tilde{p}$, and the ground truth height at p is S_p. The same as the local offset training loss, the training objective of border offsets is

$$L_{\text{border}} = \frac{1}{N}\sum_{p}\left|\hat{S}_{\tilde{p}} - S_p\right|. \tag{12}$$

4.3.4. BBox and Signal Type Generation. We have got the predicted centerline, local offset, and border offsets at each pixel and need to regress the final BBoxes and signal types. We set a threshold of 0.5 on the heat map to obtain all of the positive centerline connected domains. Each connected domain corresponds to a signal, and the pixel class that appears most frequently in a domain is the predicted signal type. The horizontal minimum $\hat{x}_{\min}$ and maximum $\hat{x}_{\max}$ of a connected domain are the start time and stop time, respectively. We chose the row with the largest cumulative probability of the predicted class in each connected domain as the centerline of this connected domain. The local offset and border offsets are averaged at the centerline points. So we can obtain the coordinates of the lower left and upper right corners of a BBox as follows:

$$R \times \left(\hat{x}_{\min}, \hat{y} + \hat{O} - \hat{y}_d, \hat{x}_{\min}, \hat{y} + \hat{O} + \hat{y}_u\right). \tag{13}$$

As we can see, all of the BBoxes are regressed directly from the centerlines without the need of deredundant processes such as the intersect over union- (IOU-) based nonmaximum suppression (NMS). The architecture of our model is simple and elegant, compared to most traditional two-stage or one-stage object detection models.

4.4. Training and Inference Details. We train the network end-to-end with the following loss function:

$$L = \lambda_1 L_{\text{cl}} + \lambda_2 L_{\text{off}} + \lambda_3 L_{\text{border}}. \tag{14}$$

The loss is a weighted sum of the three property losses. The weights λ_1, λ_2, and λ_3 that trade off among the three losses are set to 1.0, 0.5, and 0.5 in our experiments.

To make training more efficient and effective, we exploit the data augmentation methods including the crop, scaling, and Gaussian noise. The details of training and validation datasets are presented in Table 1. An Adam optimizer with a learning rate of $2e{-}4$ is used to optimize the overall objective. We train the model with a batch size of 50 for 150 epochs

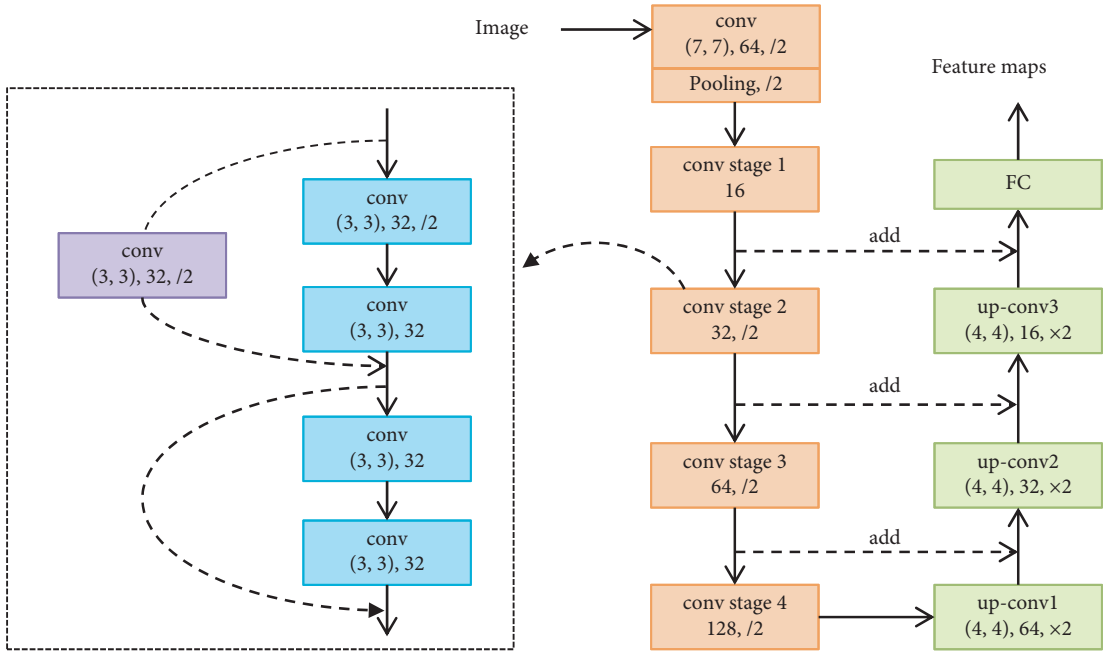


Figure 5: Visualization of the backbone module.

Table 1: Training and validation datasets' details.

	Training dataset	Validation dataset
Amount	8000	2000
Contained signal amount	8–12	8–12
Image size	800 × 4096	800 × 4096
Time range	5 s	5 s
Frequency range	125 kHz	125 kHz
SNR	0–10 dB	0–10 dB

and all of the experiments are performed on a Tesla P40 GPU.

5. Experiments

To the best of our knowledge, it is a relatively new research to implement the multitype SD and SC in wideband spectrograms directly with BBoxes, so there are few related methods. Zha et al. [38] and Yang et al. [39] have used the SSD for this task and made a comparison with other DL-based object detectors; hence we conduct comparative experiments in the same way. Specifically, we present quantitative performance results and analyze the influence of several unstable factors introduced by the channel conditions or manual processing. Moreover, we compare our SC performance separately with two traditional SC methods on narrowband signals. Experiments are conducted on the dataset generated in Section 3, and the implementation details are introduced in Section 4.4. We mark our centerline-based network as CLN in experiments.

5.1. End-to-End SD and SC Performance

5.1.1. Baselines. Zha et al. [38] have exploited the SSD to detect and classify signals in spectrograms and compared it with the Fast-RCNN [40], which suggests that the Fast-RCNN has a high accuracy, while the SSD is superior in speed. Here we compare our method with two traditional DL-based object detectors, SSD and Faster-RCNN (we can expect that the Faster-RCNN could do better than the Fast-RCNN in both accuracy and speed):

(1) SSD [42]: SSD is a representative of the one-stage object detection methods. The main idea of SSD is to first raise candidate anchors over different aspect ratios and scales at each pixel of extracted feature map, and then it predicts class scores and size adjustments for each anchor. The SSD used in our experiment is the same as that in [38], where the feature extraction CNN is a VGG-16 [49].

(2) Faster-RCNN [41]: Faster-RCNN is a representative of the two-stage object detection methods. Faster-RCNN also raises candidate anchors in advance, and then the region proposal network (RPN) predicts positive scores and adjustments for each anchor to propose regions; finally, class predictions and rere-gressions of regions are carried out by the Fast-RCNN. The feature extraction CNN of Faster-RCNN in our experiments is VGG-16.

5.1.2. Metrics. For the end-to-end SD and SC performance evaluation, we compare different methods in terms of the precision, speed, and model size. The precision metric is the mean Average Precision (mAP) [50], which is the mean of different classes' AP. AP summarizes the predicted precision and recall of one signal class at a given IOU threshold, as in equations (15)–(17), where TP, FP, and FN denote the True Positive, False Positive, and False Negative, respectively, and $P(r)$ is the Precision-Recall curve.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{TP} = \int_0^1 P(r)dr, \tag{16}$$

$$\text{mAP} = \frac{\sum_{\text{num\_classes}} \text{AP}_i}{\text{num\_classes}}. \tag{17}$$

The speed metric is the frames per second (FPS), representing the number of spectrograms that a model can process per second, as in the following equation:

$$\text{FPS} = \frac{\text{num\_process}}{\text{time\_process}}. \tag{18}$$

The model size metric is the memory usage of model parameters.

For the SSD and Faster-RCNN, we use the same training methods as in our network. All of the models are trained to converge. Figure 6 shows the quantitative comparison on end-to-end SD and SC performance. CLN has the highest mAP at different IOU thresholds, while the Faster-RCNN is the second, and the SSD is a little bit worse. The results demonstrate that our centerline-based method is more suitable for the signals in spectrograms than the one-point-based methods.

Figures 7 and 8 show the speed and model size, respectively, of different methods, for the evaluation of computational complexity. The results show that CLN is significantly faster and smaller than the other two models. Our method abandons the candidate anchors and regresses the BBox properties directly from the centerlines, thus greatly reducing the number of parameters. The simple architecture of our method ensures that it runs at a very fast speed, even compared with the fast SSD method.

To further visualize and analyze the performance, in Figure 9 we randomly show some detection results of the three methods. Figure 9(a) is the result of CLN, which is able to trace out the precise BBoxes containing the whole signals and successfully classify types with high confidence scores. Benefitting from the twice boundary regressions, the Faster-RCNN also has a nice detection and recognition performance in Figure 9(b), but it occasionally confuses two signals that are very close to each other (such as the two speeches in the bottom spectrogram). In Figure 9(c), although the SSD has found the existence of signals, it fails to regress complete BBoxes, especially for the extremely long instances.

We need to emphasize that, for the SSD and Faster-RCNN models, the default aspect ratios (height/width) of their candidate anchors are too large for the signals in spectrograms. We adjust their aspect ratios to [(1/2), (1/4), (1/6), (1/8)] to make the ground truth boxes match more candidate anchors during input encoding. The above process makes the models more task-specific, but the performance of SSD is still not satisfactory. We can expect that there is still room for improvement through further adjustments of anchors, but it could be a cumbersome and patient process compared to our method that does not need anchors.

5.2. Sensitivity Analysis. To evaluate the robustness of methods, we conduct the sensitivity tests on several unstable factors introduced by the channel conditions or manual processing. All of the factors are able to influence the signal presentation in the spectrograms that are the inputs of networks.

5.2.1. SNR. Figure 10(a) shows the mAP 50 curves of different methods versus SNR. It can be seen that all of the performances drop quickly when the SNR is lower than −2 dB. The CLN and Faster-RCNN always have better performances than the SSD, and they can hold a mAP 50 greater than 0.5 at a low SNR. In Figures 10(b)–10(d), we draw the classification confusion matrixes of the CLN at different SNR. The results show that there are few classification errors at a high SNR, and more errors happen as the SNR goes down. The confusions are often related to the similarity of the bandwidth or the shape, such as the 2FSK and 4FSK, RN and Morse, and speech and (PSK/QAM).

5.2.2. Rayleigh Fading. Rayleigh fading mainly describes that signals transmit through multiple paths of different directions to the receiver. We simulate a Rayleigh fading channel to test the robustness of CLN. We assume that the received signals are combination of two path reflections, the gains of two paths are 0 dB and −10 dB, and the time delay between them is $1e{-}7$ s. In addition, the maximum Doppler frequency shift (MDFS), introduced by the relative motion between the transmitter and receiver, is set to 0 Hz, 50 Hz, 100 Hz, and 500 Hz. The test results are shown in Figure 11. Although the "MDFS: 0 Hz" has no MDFS, a multiple path fading in test data leads to a slight performance drop compared to the "AWGN." When at a very low SNR, the mAP 50 of different MDFS are very similar, indicating that SNR is the main constraint in this situation. With the increase of SNR, the performances show a difference and a larger MDFS gets a lower mAP 50, especially after −2 dB SNR. When the SNR exceeds 2 dB, our method can hold an approximate 0.8 mAP at 500 Hz MDFS, which shows a good robustness, and we can also expect a performance improvement by enriching training data.

5.2.3. Frequency Resolution. The frequency resolution is an important and necessary parameter for spectrogram expression. To test the robustness on frequency resolution, we vary it from 20 Hz to 40 Hz and plot the mAP 50 curves in Figure 12. Generally, all of the methods perform best in 30 Hz, since our training frequency resolution is approximately 30.5 Hz. When the frequency resolution changes, the method's performances do not fluctuate obviously.

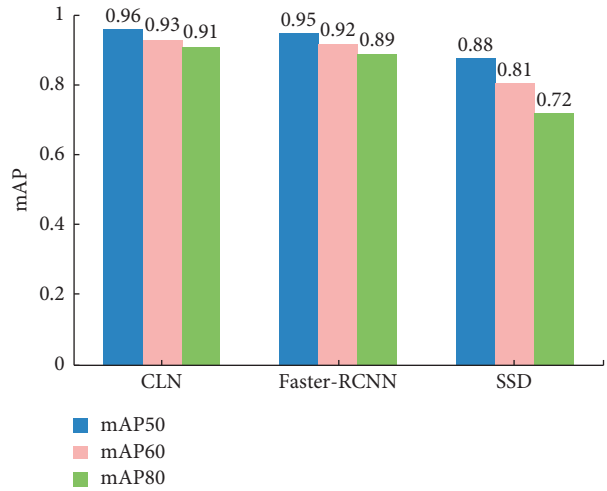


Figure 6: mAP comparison results of different methods. mAP 50, mAP 60, and mAP 80 represent the mAP at IOU thresholds of 0.5, 0.6, and 0.8, respectively.

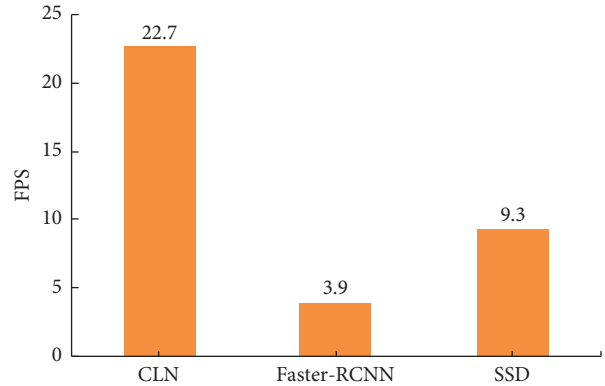


Figure 7: FPS comparison results of different methods. Results are tested on a Tesla P40 GPU.

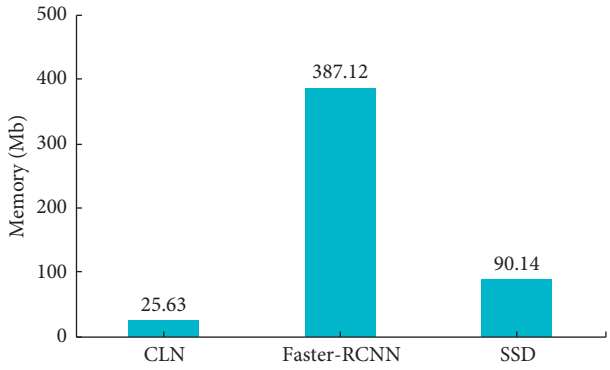
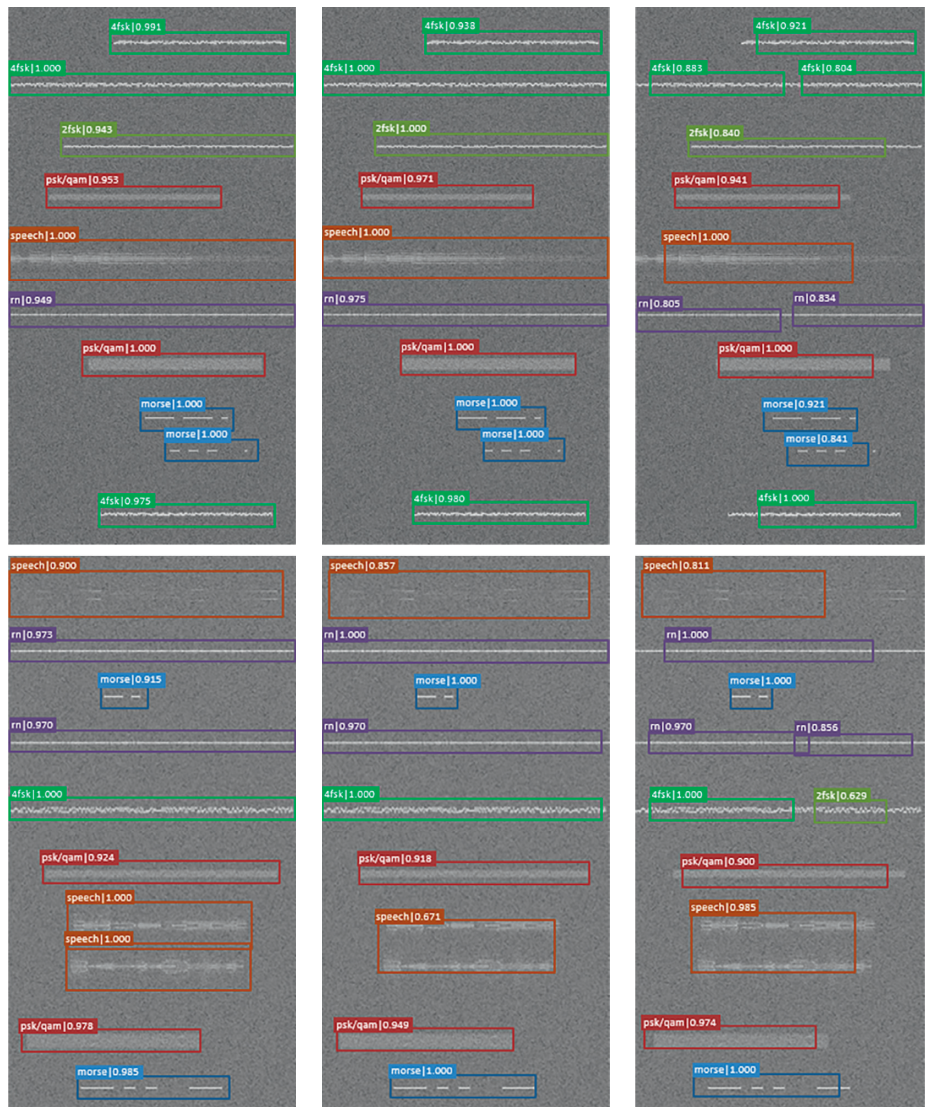


Figure 8: Model size comparison results of different methods.

Therefore, the change of resolution in a certain range has limited impact on detection and classification effects. The reason may be that, at the test frequency resolutions, the target signal types can still be intuitively distinguished in spectrograms.

5.2.4. Height of Ground Truth Box. During the dataset generation in section 3.2, we annotate the signals with BBoxes whose height is larger than the bandwidth of signals. In this setup, the height of ground truth boxes is annotated as the bandwidth of the signals, which leads to the decrease of the aspect ratios of ground truth boxes. To adapt to those adjustments, we also have increased the aspect ratios of anchors for Faster-RCNN and SSD to [(1/2), (1/10), (1/15), (1/20)]. The detection and classification results are presented in Figure 13, and we can see that our method is still able to predict the BBoxes closed to the ground truth, while the performances of the SSD and Faster-RCNN are greatly reduced, especially the missing and redundant predictions.

For our method, it focuses on the centerlines of signals that do not change with the height of ground truth boxes; thus it only needs to predict different border offsets. For the SSD and Faster-RCNN, since the sizes and aspect ratios of BBoxes vary greatly for different signals, it is difficult to design a group of common anchors, resulting in that some



(a) (b) (c)

FIGURE 9: The detection and classification results of different methods. (a) Results of the CLN. (b) Results of the Faster-RCNN. (c) Results of the SSD.

ground truth boxes could not match the candidate anchors during input encoding. In addition, the signals with a narrow bandwidth or long duration may get repeated but not overlapping predictions, which cannot be filtered by the IOU-based NMS. So if you want to use the anchor-based methods to detect the signals in spectrograms, you had better decrease the aspect ratios of anchors and increase those of the ground truth boxes.

5.3. Separate SC Performance. The proposed method is based on the spectrogram, which can classify the types of signals when they are detected. By adjusting the types of training data, our network can classify some modulation modes and even the specific signal types such as Morse and speech. To further evaluate the classification capability of our method, we conduct a series of separate SC experiments in this subsection. We first conduct classification on several common signal types to analyze which types can and cannot be classified by our method. Then, targeting the types that can be classified, we compare our method with two classic SC methods to evaluate the performance.

We select the signal types of 2ASK, 4ASK, 2FSK, 4FSK, BPSK, QPSK, and 16QAM, which are all the basic modulation modes used in practical communication, to test the classification effect of our method. We simulate wideband data that randomly contain the above signals using the same approach as in section 3.2, and the confusion matrix is shown in Figure 14. The results show that 2ASK, 4ASK, 2FSK, and 4FSK can be well classified by our method, while BPSK, QPSK, and 16QAM are seriously confused. The results are in line with our expectations, since the shapes of MPSK and MQAM in spectrograms are quite similar, while those of the other signals are distinguishable.

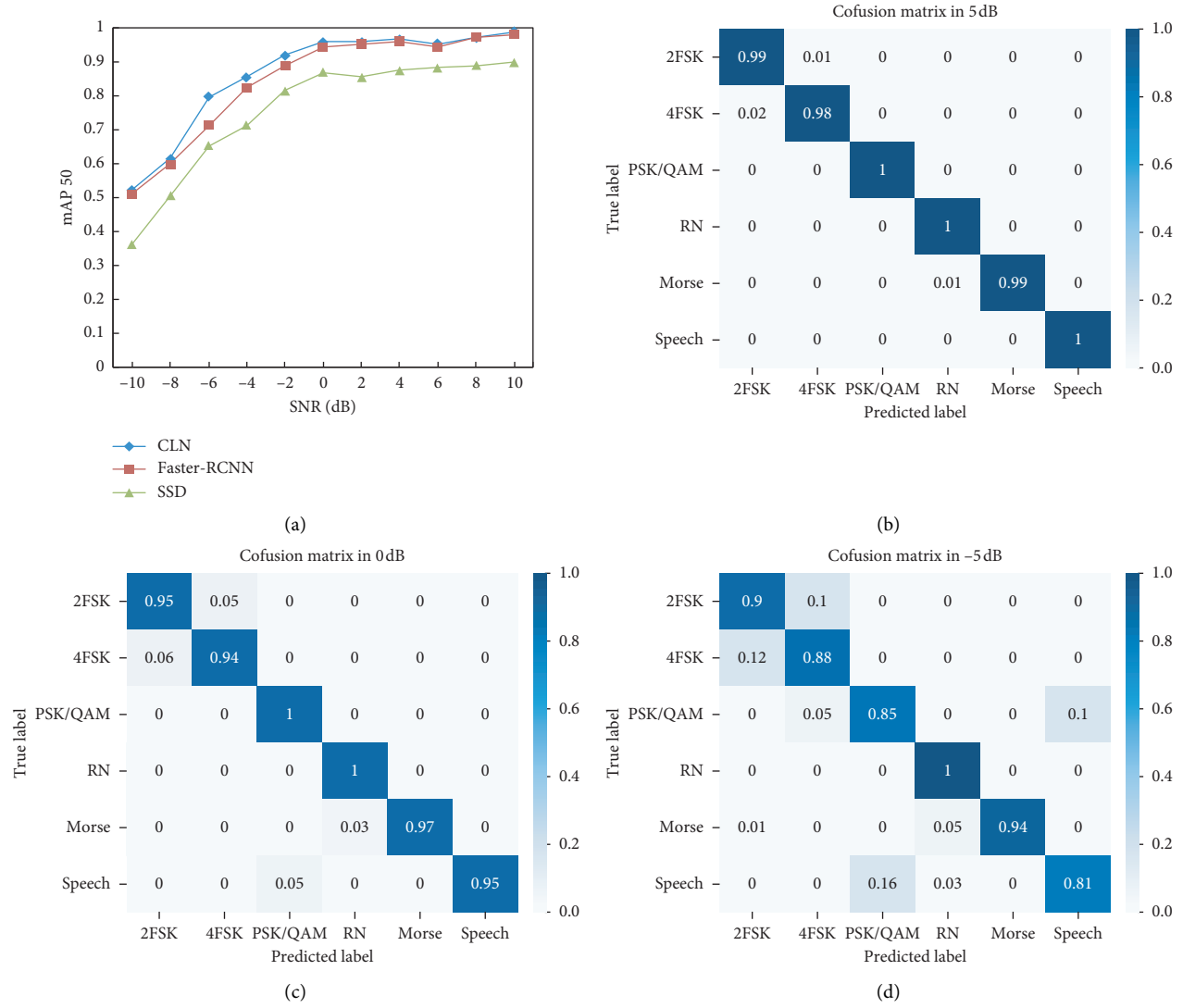


Figure 10: The performance at different SNR. (a) Detection of mAP 50 of different methods versus SNR. ((b)–(d)) The confusion matrixes of CLN at 5 dB, 0 dB, and −5 dB, respectively.

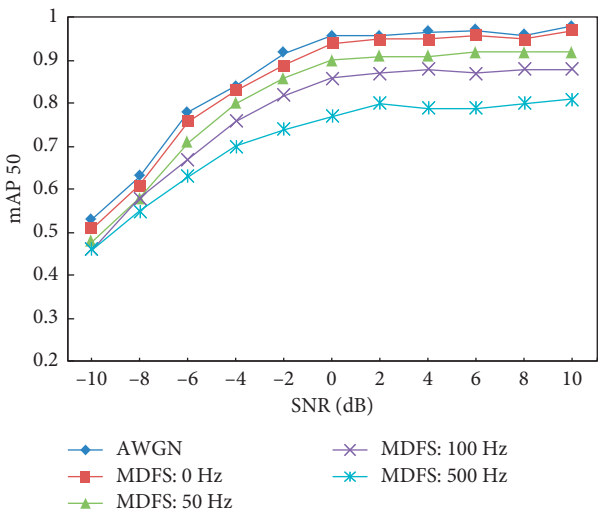


Figure 11: The performance of CLN on different MDFS.

In consideration of the above results, we merge the signal types of BPSK, QPSK, and 16QAM to the "(PSK/QAM)." To further evaluate the classification performance on 2ASK, 4ASK, 2FSK, 4FSK, and (PSK/QAM), we compare our method with two classic SC methods:

(1) High order cumulant (HOC) + SVM [51]: An ML-based algorithm using the SVM with seven HOC as the feature vectors: C_{20}, C_{21}, C_{40}, C_{41}, C_{42}, C_{60}, and C_{63}.

(2) IQ waveform + CNN [47]: A DL-based algorithm using a 4-layer network of two CNN layers and two fully connected layers, with the signal IQ waveform as features.

The inputs of methods are baseband signals. Figure 15 shows the average classification accuracy of three methods versus SNR. The results show that CLN is more robust at a low SNR, partly because the spectrogram features can effectively present the characteristics of tested signal types and

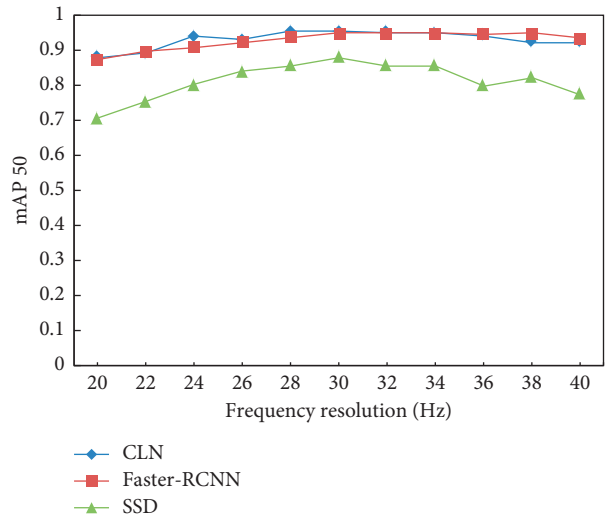


Figure 12: The performance of different methods versus frequency resolution.

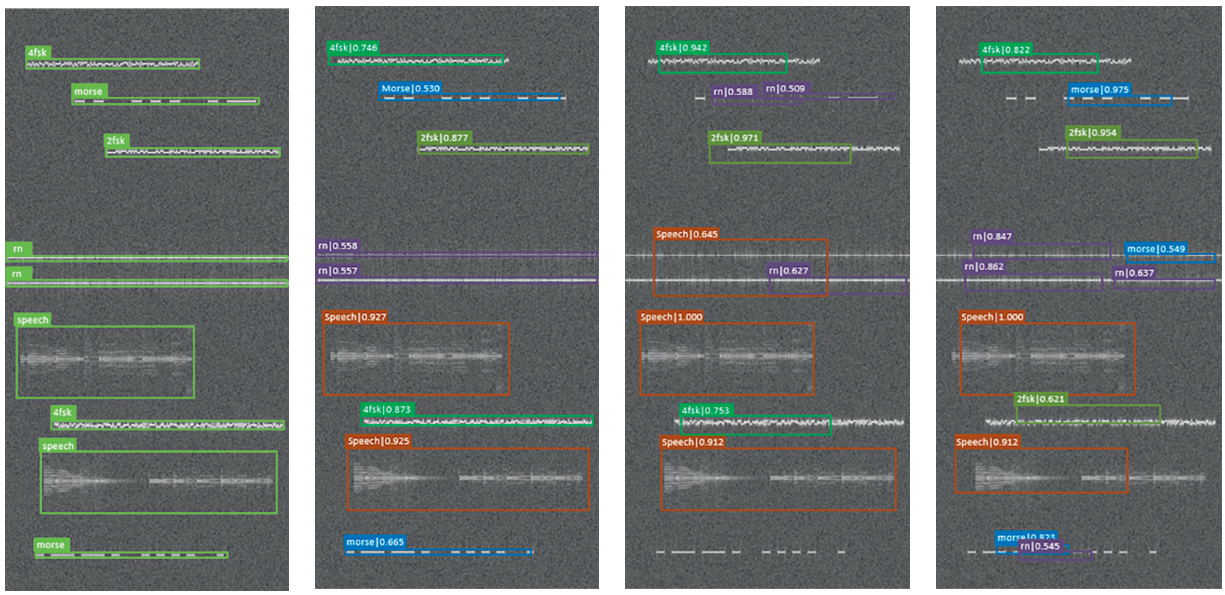


Figure 13: The detection and classification results with the ground truth box height close to the signal bandwidth. (a) Ground truth boxes. (b) Results of the CLN. (c) Results of the Faster-RCNN. (d) Results of the SSD.

partly because the DL algorithm could learn deeper and richer features.

Considering the deviation of carrier frequency estimation by CLN or Fourier Transform in practice, we evaluate the influence of the frequency offset on classification performance. We calculate the average accuracy at 5 dB SNR versus the frequency offset normalized by the symbol sampling frequency, and the results are shown in Figure 16. It can be seen that the increase of frequency offset could have an obvious effect on method performances, and when at a large offset, all of the methods are no longer suitable. However, our method still shows a stronger robustness.

5.4. Parameter Tuning. We implement parameter tunings on several important hyperparameters of our neural network, including the layers of up-convolutions, the channels of convolutional stages and up-convolutions, and the channels of the first CNN layer in STEM. Specifically, we tune the object parameters while keeping the others fixed and plot the mAP 50 curves versus SNR.

Figure 17 shows the performance on different layers of up-convolutions. The layers of up-convolutions determine the size and channels of extracted feature map; for example, one layer up-convolution outputs feature map of (1/16) input size and 64 channels. In backbone module, the up-convolutions merge the high-level features that present more

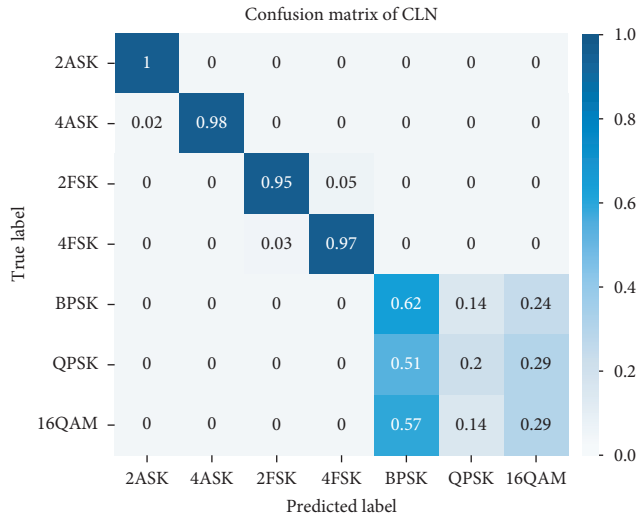


Figure 14: Confusion matrix of CLN on several modulation modes.

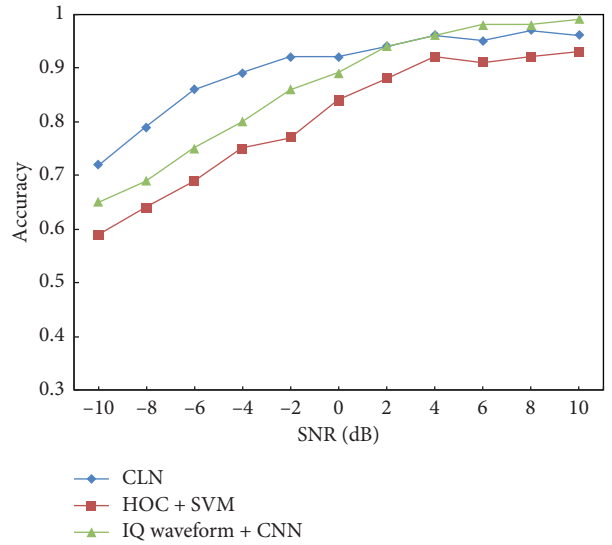


Figure 15: Classification accuracy of different methods versus SNR.

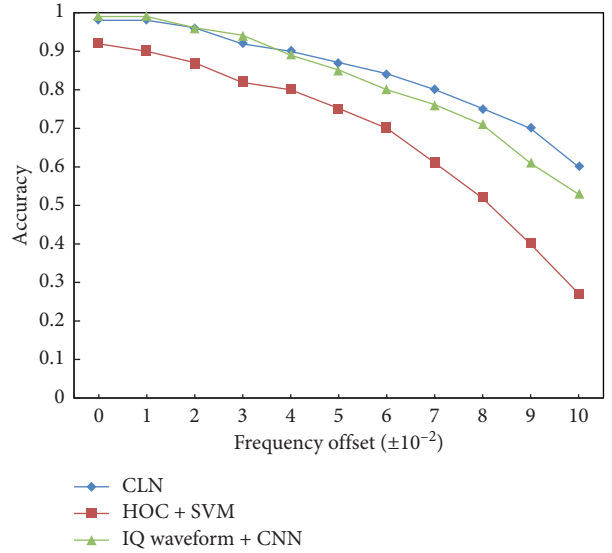


Figure 16: Classification accuracy of different methods versus frequency offset.

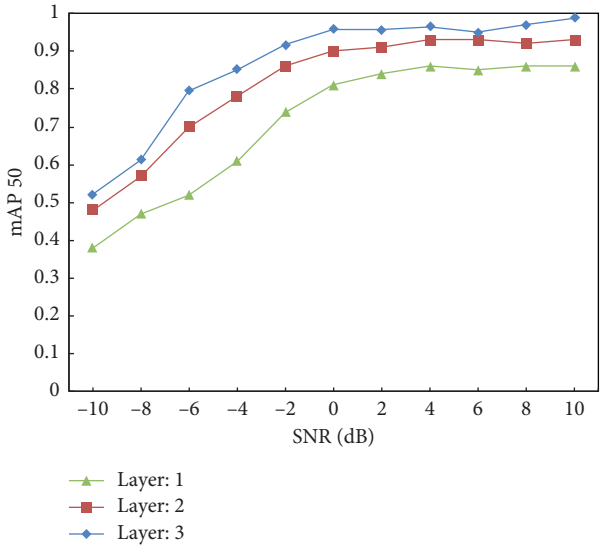


Figure 17: The performance of CLN on different layers of up-convolutions.

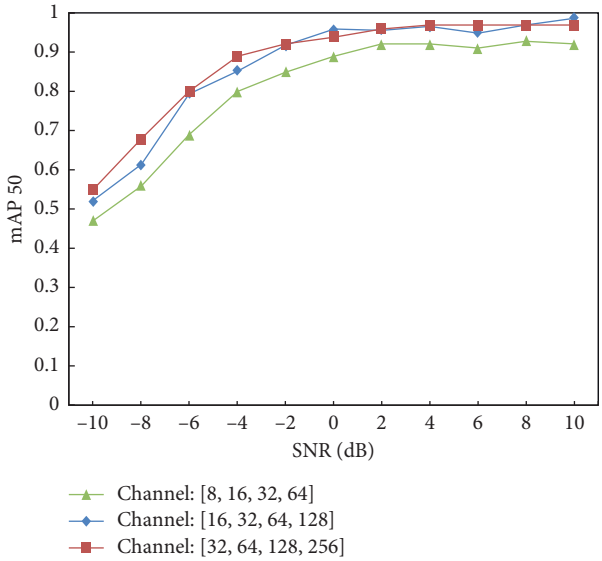


Figure 18: The performance of CLN on different channels of convolutional stages and up-convolutions. "Channel: [8, 16, 32, 64]" refers to the channels of conv stage 1/up-conv 3, conv stage 2/up-conv 2, conv stage 3/up-conv 1, and conv stage 4 which are 8, 16, 32, and 64, respectively.

general information with the low-level features. Thus, increasing up-convolution layers could consider more detailed information but more discrete noise. The results in Figure 17 show the three layer up-convolutions obtain the best effect.

Figure 18 shows the performance on different channels of convolutional stages and up-convolutions. In backbone module, four forward convolutional stages and three up-convolutions achieve the extraction and merge of the different level features. Increasing channels could let the CNN learn features of more dimensions, but it also increases the space and time usage of a model. We set three channel groups for test. Following the principle of ensuring accuracy and keeping the model as small as possible, we prefer the setting of "Channel: [16, 32, 64, 128]."

Figure 19 shows the performance on different channels of the first CNN layer in STEM. Following the principle of ensuring accuracy and keeping the model as small as possible, we prefer the setting of "Channel: 32."

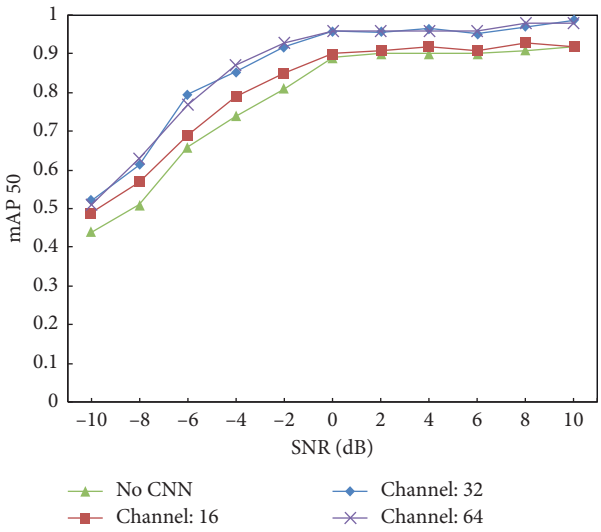


Figure 19: The performance of CLN on different channels of the first CNN layer in STEM.

6. Conclusions

In this paper, we exploit the idea of object detection technology and present a deep convolutional network for multitype SD and SC in the wideband spectrograms. We have analyzed the defects of traditional DL-based object detectors for the tasks solved here and proposed a centerline-based method. Targeting the characteristics of the signals, our method first finds the centerlines of signal regions, and then it regresses to the complete BBoxes and classifies the signal types. In experiments, we have conducted comparisons with other object detection methods in terms of accuracy, speed, and model size, and we have also implemented sensitivity tests on some channel conditions and manual processing. The results indicate that our method has a higher detection mAP with an obvious speed advantage, and it is also more robust in different conditions. In addition, a series of separate classification experiments have shown the good classification capability of our method. As a consequence, the proposed method achieves outstanding SD and SC performances in spectrograms while keeping a real-time capability, which makes it valuable for the engineering applications.

In the future, we will enrich our dataset, including the real-received signals, and we will explore more comprehensive features for the signal detection and classification.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key Laboratory of Science and Technology on Blind Signal Processing.

References

[1] A. A. Khan, M. H. Rehmani, and M. Reisslein, "Cognitive radio for smart grids: survey of architectures, spectrum sensing mechanisms, and networking protocols," *IEEE Communications Surveys & Tutorials*, vol. 18, p. 1, 2015.

[2] G. Joshi, S. Nam, and S. Kim, "Cognitive radio wireless sensor networks: applications, challenges and research trends," *Sensors*, vol. 13, no. 9, pp. 11196–11228, 2013.

[3] Y. Arjoune and N. Kaabouch, "A comprehensive survey on spectrum sensing in cognitive radio networks: recent advances, new challenges, and future research directions," *Sensors*, vol. 19, no. 1, p. 126, 2019.

[4] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1622–1637, 2016.

[5] A. Sadhu, S. Narasimhan, and J. Antoni, "A review of output-only structural mode identification literature employing blind source separation methods," *Mechanical Systems and Signal Processing*, vol. 94, pp. 415–431, 2017.

[6] J. E. Salt and H. H. Nguyen, "Performance prediction for energy detection of unknown signals," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3900–3904, 2008.

[7] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

[8] J. J. Lehtomaki, J. Vartiainen, M. Juntti, and H. Saarnisaari, "Analysis of the LAD methods," *IEEE Signal Processing Letters*, vol. 15, pp. 237–240, 2008.

[9] J. J. Lehtomaki, J. Vartiainen, M. Juntti, and H. Saarnisaari, "CFAR outlier detection with forward methods," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4702–4706, 2007.

[10] A. A. Tadaion, M. Derakhtian, S. Gazor, M. M. Nayebi, and M. R. Aref, "Signal activity detection of phase-shift keying signals," *IEEE Transactions on Communications*, vol. 54, no. 6, p. 1143, 2006.

[11] P. Salembier, S. Liesegang, and C. López-Martínez, "Ship detection in SAR images based on maxtree representation and graph signal processing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 2709–2724, 2018.

[12] D. Bao, L. De Vito, and S. Rapuano, "A histogram-based segmentation method for wideband spectrum sensing in cognitive radios," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 7, pp. 1900–1908, 2013.

[13] D. Bao, L. De Vito, and S. Rapuano, "Spectrum segmentation for wideband sensing of radio signals," in *Proceedings of the 2011 IEEE International Workshop on Measurements and Networking Proceedings (M & N)*, pp. 47–52, Anacapri, Italy, October 2011.

[14] S. Koley, V. Mirza, S. Islam, and D. Mitra, "Gradient-based real-time spectrum sensing at low SNR," *IEEE Communications Letters*, vol. 19, pp. 391–394, 2014.

[15] S. Men, P. Chargé, Y. Wang, and J. Li, "Wideband signal detection for cognitive radio applications with limited resources," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, no. 1, p. 2, 2019.

[16] P. Y. Dibal, E. N. Onwuka, J. Agajo, and C. O. Alenoghena, "Wideband spectrum sensing in cognitive radio using discrete wavelet packet transform and principal component analysis," *Physical Communication*, vol. 38, Article ID 100918, 2020.

[17] D. Ke, Z. Huang, X. Wang, and X. Li, "Blind detection techniques for non-cooperative communication signals based on deep learning," *IEEE Access*, vol. 7, pp. 89218–89225, 2019.

[18] G. J. Mendis, J. Wei, and A. Madanayake, "Deep learning based radio-signal identification with hardware design," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 5, pp. 2516–2531, 2019.

[19] Y. Yuan, Z. Sun, Z. Wei, and K. Jia, "DeepMorse: a deep convolutional learning method for blind Morse signal detection in wideband wireless spectrum," *IEEE Access*, vol. 7, pp. 80577–80587, 2019.

[20] E. E. Azzouz and A. K. Nandi, "Procedure for automatic recognition of analogue and digital modulations," *IEE Proceedings-Communications*, vol. 143, no. 5, pp. 259–250, 1996.

[21] A. Abdelmutalab, K. Assaleh, and M. El-Tarhuni, "Automatic modulation classification based on high order cumulants and hierarchical polynomial classifiers," *Physical Communication*, vol. 21, pp. 10–18, 2016.

[22] P. Sutton, K. Nolan, and L. Doyle, "Cyclostationary signatures in practical cognitive radio applications," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 13–24, 2008.

[23] P. R. U. Lallo, "Signal classification by discrete Fourier transform," in *Proceedings of the Military Communications Conference*, pp. 197–201, Atlanta, GA, USA, March 1999.

[24] W. Yongshi, G. Jie, L. Hao, L. Li, W. Zhigang, and W. Houjun, "CNN-based modulation classification in the complicated communication channel," in *Proceedings of the IEEE International Conference on Electronic Measurement Instruments*, pp. 512–516, Yangzhou, China, October 2017.

[25] A. Ali and F. Yangyu, "Unsupervised feature learning and automatic modulation classification using deep learning model," *Physical Communication*, vol. 25, pp. 75–84, 2017.

[26] S.-Z. Hsue and S. S. Soliman, "Automatic modulation classification using zero crossing," *IEE Proceedings F Radar and Signal Processing*, vol. 137, no. 6, pp. 459–464, 1990.

[27] T. J. Oshea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Proceedings of the International Conference on Engineering Applications of Neural Networks*, pp. 213–226, Aberdeen, UK, September 2016.

[28] N. West and T. J. Oshea, "Deep architectures for modulation recognition," in *Proceedings of the IEEE International Symposium on Dynamic Spectrum Access Networks*, pp. 1–6, Baltimore, MD, USA, March 2017.

[29] X. Teng, P. Tian, and H. Yu, "Modulation classification based on spectral correlation and SVM," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1–4, Dalian, China, October 2008.

[30] H. Hu, J. Song, and Y. Wang, "Signal classification based on spectral correlation analysis and SVM in cognitive radio," in *Proceedings of the Advanced Information Networking and Applications*, pp. 883–887, GinoWan, Japan, March 2008.

[31] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Automatic modulation classification using combination of genetic programming and KNN," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2742–2750, 2012.

[32] J. Lopatka and M. Pedzisz, "Automatic modulation classification using statistical moments and a fuzzy classifier," in *Proceedings of the International Conference on Signal Processing*, pp. 1500–1506, Istanbul, Turkey, June 2000.

[33] L. Zhao, C. Su, Z. Dai et al., "Indoor device-free passive localization with DCNN for location-based services," *The Journal of Supercomputing*, pp. 1–18, 2019.

[34] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.

[35] O. A. Dobre, S. Rajan, and R. Inkol, "Joint signal detection and classification based on first-order cyclostationarity for cognitive radios," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, Article ID 656719, 2009.

[36] M. O. Mughal and S. Kim, "Signal classification and jamming detection in wide-band radios using Naïve Bayes classifier," *IEEE Communications Letters*, vol. 22, no. 7, pp. 1398–1401, 2018.

[37] M. Titos, A. Bueno, L. Garcia, M. C. Benitez, and J. Ibanez, "Detection and classification of continuous volcano-seismic signals with recurrent neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 1936–1948, 2019.

[38] X. Zha, H. Peng, X. Qin, G. Li, and S. Yang, "A deep learning framework for signal detection and modulation classification," *Sensors*, vol. 19, no. 18, p. 4042, 2019.

[39] S. Yang, S. Jin, H. Peng, X. Hou, and J. Fu, "Ultra-Short wave specific signal detection and recognition based on spectrogram and deep convolution neural network," *Journal of Information Engineering University*, vol. 20, 2019.

[40] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 91–99, Montreal, Canada, December 2015.

[42] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *Proceedings of the 14th European Conference ECCV 2016*, pp. 21–37, Amsterdam, The Netherlands, October 2016.

[43] J. Redmon and A. Farhadi, "YOLOV3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, Venice, Italy, October 2017.

[45] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, https://arxiv.org/abs/1904.07850.

[46] B. Kim, J. Kim, H. Chae, D. Yoon, and J. W. Choi, "Deep neural network-based automatic modulation classification technique," in *Proceedings of the 2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 579–582, Jeju, South Korea, October 2016.

[47] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Proceedings of the International Conference on Engineering Applications of Neural Networks*, pp. 213–226, Cham, Switzerland, August 2016.

[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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

[50] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: a retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[51] S. Li, F. Chen, and W. Long, "Modulation recognition algorithm of digital signal based on support vector machine," in *Proceedings of the Control and Decision Conference*, pp. 3326–3330, Maui, HI, USA, December 2012.