Accurate Spectrum Sensing with Improved DeepLabV3+ for 5G-LTE Signals Identification

Gia-Vuong Nguyen, Ca Van Phan, and Thien Huynh-The*
vuongng.cce@gmail.com,capv@hcmute.edu.vn,thienht@hcmute.edu.vn
Department of Computer and Communications Engineering
HCM City University of Technology and Education
Ho Chi Minh City, Viet Nam

ABSTRACT

This paper presents a deep learning approach for fifth-generation (5G) and Long-Term Evolution (LTE) signal discrimination, explicitly focusing on identifying modulated signals in next-generation wireless networks. The mixture of modulated signals, which is essentially difficult to discern in the form of a complex envelope, should be converted into a visually informative spectrogram image by applying the Fast Fourier transform (FFT). To segment spectral regions of 5G new radio (NR) and LTE in a spectrogram, we aptly improve DeepLabV3+, a deep encoder-decoder network for semantic segmentation, by incorporating an adaptive Atrous Spatial Pyramid Pooling (ASPP) block and an attention mechanism to accommodate intrinsic signal characteristics and amplify relevant features, respectively. Besides increasing the learning efficiency in the encoder, the improvement enriches the recovery capability of crucial 5G and LTE details, thus resulting in more accurate signal identification in the spectrogram image. Relying on the simulation results benchmarked on a dataset consisting of spectral images containing both LTE and 5G signals, the new network demonstrated effectiveness when compared to the original version by increasing global accuracy, mean intersection-over-union (IoU), and mean boundary-F1-score (BFScore) up to 1.37%, 2.85% and 9.43% in that order. For medium SNR level, it can achieve 98.28% global accuracy and 96.66% mean IoU, while also showing robustness under various practical channel impairments.

CCS CONCEPTS

• **Hardware** → *Digital signal processing*; • **Computing methodologies** → *Neural networks*; *Image segmentation.*

KEYWORDS

5G NR, deep learning, encoder-decoder neural network, intelligent spectrum sensing, signal identification.

*Corresponding author.

SOICT 2023, December 07–08, 2023, Ho Chi Minh, Vietnam

ACM ISBN 979-8-4007-0891-6/23/12…$15.00
https://doi.org/10.1145/3628797.3628798

ACM Reference Format:
Gia-Vuong Nguyen, Ca Van Phan, and Thien Huynh-The. 2023. Accurate Spectrum Sensing with Improved DeepLabV3+ for 5G-LTE Signals Identification. In *The 12th International Symposium on Information and Communication Technology (SOICT 2023), December 07–08, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3628797.3628798

1 INTRODUCTION

Signal recognition is a fundamental task in various domains, including wireless communications, radar systems, and signal intelligence [15]. Accurately identifying and classifying modulated signals is crucial for important tasks such as signal monitoring, interference detection, and network optimization. Over the years, extensive research has been devoted to developing effective techniques and algorithms for modulated signal recognition, ranging from traditional methods based on handcrafted features and machine learning (ML) to more recent approaches exploiting deep learning (DL) algorithms.

Traditional approaches for modulated signal recognition usually rely on designing feature extraction techniques and studying ML algorithms, which require domain expertise and manual engineering knowledge. These methods typically involve extracting statistical features from the time and frequency domains to analyze signals [3, 14]. However, they may need help to capture the intricate patterns and complex relationships inherent in modulated signals, especially in the presence of noise and interference in wireless communication channels.

In recent years, DL has emerged as a powerful tool in signal processing [8, 18], revolutionizing various domains such as natural language processing, computer vision, and robotics. By leveraging the capabilities of deep neural networks (DNN) in learning complex patterns, it is possible to automatically learn intricate patterns and features directly from raw signal data. DL eliminates the need for manual feature extraction and enables end-to-end signal recognition systems that can effectively handle complex and diverse modulated signals in next-generation wireless networks. Several studies have explored the application of DL techniques to the problem of signal recognition and achieved remarkable results in accuracy improvement and complexity reduction compared with traditional approaches. For instance, some authors focus on developing deep neural network architectures to improve the precision of signal detection, such as the three-dimensional convolutional neural network (CNN) in [12], multi-instance multi-label deep convolutional neural network (MIML-DCNN) in [17], RaComNet [13]. Another study tried to find a way to handle raw signals. For example, Zhibo Chen *et al.* has shown that detecting modulated signals

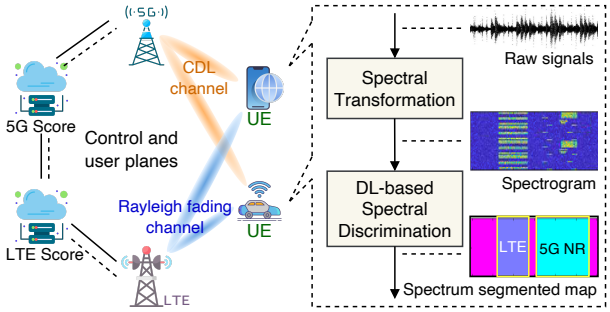


Figure 1: A general wireless network architecture with DL for spectral discrimination of 5G-LTE occupancy.

using Fourier transform produces significantly better results compared to detecting raw signals [6]. Through the Fourier transform (FT), signals in the time domain are converted to a spectral image in the frequency domain, altering the problem from spectrum sensing to image processing. Subsequently, significant advancements in computer vision are applied to achieve a high correct detection rate.

As building upon prior works, this work presents an advanced approach to discriminate 5G and LTE signals in a wideband spectrogram using deep learning. First, our method leverages the fast Fourier transform (FFT), a specific algorithm of FT, to convert complex waveform received signals into spectrogram images, which capture the frequency and temporal characteristics of the signals. Finally, we use these spectrograms as input data for DeepLabV3+ deep convolutional neural network (DCNN) [4, 5], which is state-of-the-art architecture in computer vision. DeepLabV3+ has proven its capability in semantic segmentation tasks by using atrous convolution and ASPP block to capture contextual information at multiple scales and enhance feature representation. In addition to utilizing DeepLabV3+, we incorporate an adaptive ASPP block that adapts to the characteristics of the signal data. This allows the network to effectively identify the boundary information between signals in spectrum images and increase performance. Furthermore, we integrate an attention mechanism into the network to emphasize relevant features and enhance the discriminative power of the model. To evaluate the effectiveness of our proposed method, we conduct simulations on a dataset comprising a wide range of modulated signals and compare the performance of our approach with existing backbones, like ResNet18, ResNet50 [10], MobileNetV2 [19]. The performance is evaluated based on some metrics such as accuracy, IoU, and BFScore. Our results demonstrate the advancements achieved in recognizing modulated 5G-LTE signals, as evidenced by the significant improvement in metric values compared to the baseline.

2 METHODOLOGY

This work aims to develop an intelligent DL-based spectrum sensing method to discriminate 5G NR and LTE signals. Taking inspiration from the advancements of deep networks in image processing or computer vision, especially semantic segmentation, we adopt a spectral representation with FFT for incoming signals at the receiver as the input to a semantic segmentation model. We fine-tune existing DL architectures with some sophisticated mechanisms to enhance the overall accuracy of signal segmentation. Particularly, inspired by the advancements of DeepLabV3+ in semantic segmentation, we aptly leverage the deep encoder-decoder architecture of DeepLabV3+ with an adaptive ASPP block and an attention mechanism to improve the segmentation efficiency of 5G and LTE regions in spectrogram images. As shown in Figure 1, our proposed method consists of two modules: spectrogram representation with FFT and signal identification with improved DeepLabV3+.

2.1 Spectrogram representation with FFT

We simulate 5G NR and LTE signals similar to reality. The signals are generated by following their separate sets of variable signal parameters. With LTE, the parameters include the reference channel, bandwidth, and duplex mode, wherein the reference channel refers to the specific channel used as a reference for signal transmission, bandwidth represents the frequency range allocated to the LTE signal, and the duplex mode determines whether the LTE signal uses frequency division duplex (FDD) or time division duplex (TDD) for uplink and downlink communications [1]. With NR, the signal is specialized by more parameters like bandwidth, sub-carrier spacing (SCS), synchronization signal block (SSB) block pattern, and SSB Period, wherein bandwidth denotes the frequency range assigned to the NR signal, SCS refers to the spacing between adjacent subcarriers in the 5G NR signal, SSB block pattern indicates the pattern of SSBs within the NR signal, and SSB period denotes the time interval between consecutive SSBs [2].

Additionally, over-the-air signals are transmitted through channels from a transmitting device (base station or cell tower) to a receiving device (smartphones, tablets, or other devices). The transmission and reception of signals through the wireless channel can be mathematically represented by the following equation:

$$y(t) = x(t) \times h(t) + n(t), \tag{1}$$

where t represents time, x denotes the initial signals generated, y denotes received signals, h denotes the channel coefficient, and n denotes the additive noise introduced when the signal passes through the transmission channel. The additive noise can include various types, which may be mentioned as additive white Gaussian noise (AWGN), thermal noise (TN), multipath fading noise (MFN), impulse noise (IN), etc. Moreover, some real-world phenomena in wireless communications, such as multipath scattering effects, time dispersion, and Doppler shifts (i.e., arise from relative motion between the transmitter and receiver), may cause the degradation of signal quality.

Fast Fourier transform (FFT): Several studies have shown that modulated signal recognition yields more accurate results than raw signal recognition. A recent work [6] demonstrated a remarkable effectiveness of FT in signal classification, in which an FT-based method outperformed several ones that leveraged raw data decomposed from complex envelopes of signals. Therefore, we represent our signal from the time domain to the frequency domain. However, we propose substituting FT with FFT to improve computational efficiency and increase speed when working with large signal data. This substitution will enable the model to process real-time signals

more responsively and effectively. For each received signal waveform y, we apply a 4096-point FFT to convert the signal y from the time domain to the signal in the frequency domain $\boldsymbol{Y}$ as follows:

$$Y(f) = \sum_{t=0}^{T-1} y(t) \times e^{-i2\pi \frac{ft}{T}}, \tag{2}$$

where $e^{-i2\pi \frac{ft}{T}}$ is the complex factor that forms the frequency component of the FFT, i is the imaginary unit, f is the frequency, T is the number of samples in the cycle of the input signal, here is 4096. $\boldsymbol{Y}$ represents the spectrum of the input signal at frequency k. To enhance intuitiveness, the spectrum images are normalized to the range of $[0, 1]$ before being transformed into color values and resized to 256×256.

2.2 Signal identification with improved DeepLabV3+

Applying the success in computer vision, we intended to build a DNN to segment modulated signals. Nowadays, various DNN architectures are proven efficient in semantic segmentation, such as U-Net [16], Mask R-CNN [9] and especially the DeepLab family [4, 5], which has achieved notable advancements and has been the new state-of-the-art method in semantic segmentation tasks. To learn features from a larger receptive field while still maintaining computational efficiency, DeepLab networks were introduced by incorporating multiple atrous (or dilated) convolutional layers. These layers enable the extraction of spatial features in an enlarged region. Atrous convolutional layers compute features to produce the output using the following formula:

$$\mathbf{Y}_a[i, j] = \sum_{u,v} \mathbf{X}_a[i + r_H u, j + r_W v] \times \mathbf{W}_a[u, v], \tag{3}$$

where the filter $\mathbf{W}_a$ has the size of $u \times v$ to extract the local features of the input $\mathbf{X}_a$; r_H and r_W indicates the dilation factors (a.k.a., dilation rates) in the height and width dimension, respectively. Interestingly, the regular convolution layers are defined with $r_H = r_W = 1$. Alongside the atrous convolutional layers, the ASPP module is designed to facilitate feature extraction at multiple different identical dilation rates r, which means, $r_H = r_W = r$ in DeepLabV3+ [5]. As a result, ASPP enriches the spatial contextual information of extracted features thanks to enlarging the receptive field (the region of the input that the layer can perceive) of the layer without increasing the number of parameters or computation, thus enabling the model to facilitate segmentation tasks. The output of ASPP is the summary of multiple feature maps of different atrous convolutional layers with different dilation rates via a depthwise concatenation layer as follows:

$$\mathbf{F}_{ASPP} = concat(\mathbf{A}_{1,1}^{1\times1}(\mathbf{X}), \mathbf{A}_{1,6}^{3\times3}(\mathbf{X}), \mathbf{A}_{1,12}^{3\times3}(\mathbf{X}), \mathbf{A}_{1,18}^{3\times3}(\mathbf{X})), \tag{4}$$

where $\mathbf{A}_{s,r}^{n\times n}$ denotes a sequential operation, including atrous convolution (with the filter size $n\times n$, the stride s, and the identical dilation rate r), batch normalization (BN), and ReLU (rectified linear unit) activation, and $\boldsymbol{X}$ is the input of ASPP. Although DeepLabV3+ with ASPP has presented some advancements to yield remarkable performance in semantic segmentation, it has remained two drawbacks: first is the weak spatial features extracted by the backbones, such as MobileNetV2 [19] and ResNet [10], in the encoder; second is the uncertain dilation rate selection of atrous convolutional layers in ASPP. To overcome these drawbacks, we improve DeepLabV3+ with an adaptive dilation rate (ADR) in ASPP and an attention mechanism (ATM) to improve multi-scaling feature fusion and strengthen relevant features, respectively.

ASPP with ADR: We realize that the dilation rates in the original ASPP block are not really identical to our dataset, wherein the input spectrogram images are generated and normalized to the size of 256×256. With some standard backbones (such as ResNet18, ResNet50, and MobileNetV2) deployed in the encoder, the final feature maps of these backbones finalized with the size of 16×16, which serves as the input to ASPP. In DeepLabV3+, ASPP consists of one atrous convolution with 1×1 kernels and three atrous convolutions with 3×3 kernels, where the dilation rates are 6, 12 and 18. Observably, the atrous convolutional layer with a 3×3 kernel and a dilation rate 12 can extract spectral features with a receptive field 25×25 and up to 37×37 with a dilation rate 18. These receptive fields are significantly larger than 16×16 feature maps, thus leading to inefficient feature learning. To address this issue, dilation rates should be adjusted to be smaller and in a range of $[1, 7]$ to achieve a receptive field smaller than 15×15, which is suitable for the given input size to ASPP. On the other hand, our deep segmentation network is trained on a dataset consisting of individual spectral signals (i.e., either 5G NR or LTE with noise). Still, it is benchmarked on another synthetic dataset containing multiple spectral-occupied signals. Concretely, atrous convolutional layers with a large receptive field perform effectively in seizing the spectral features of individual signals in the training set, but many unforeseen features, such as the 5G-LTE correlations captured by a larger receptive field, may cause the degradation of learning efficiency. As a result, this phenomenon decreases the overall segmentation accuracy of our segmentation network. In this work, we reduce the dilation rates of the atrous convolutional layers to smaller (respectively $[1, 2, 4, 6]$) in ASPP to increase the model accuracy. At this point, the output of ASPP should be written as follows:

$$\mathbf{F}_{ASPP} = concat(\mathbf{A}_{1,1}^{1\times1}(\mathbf{X}), \mathbf{A}_{1,2}^{3\times3}(\mathbf{X}), \mathbf{A}_{1,4}^{3\times3}(\mathbf{X}), \mathbf{A}_{1,6}^{3\times3}(\mathbf{X})). \tag{5}$$

It is noted that adjusting the dilation rates does not change the complexity of ASPP, including the number of trainable parameters and computation cost.

Attention mechanism: Segmenting spectral signals based solely on spatial features that are extracted by convolutional layers in a regular manner is insufficient. In fact, the success of semantic segmentation extends beyond enhancing the representational capability of CNN through the improvement of the encoding quality of spatial features. It also diligently explores inter-channel relationships, enabling recalibrating channel-wise feature responses by explicitly modeling inter-dependencies between channels [11]. Furthermore, several existing works have demonstrated that some significant accuracy improvements are yielded by cooperatively leveraging spatial and channel-wise features [20]. In this work, besides the adaptive ASPP block, we develop an attention mechanism to reinforce the learning capability of backbones adopted in the encoder of DeepLabV3+ by studying the squeeze-and-excitation (SE) block [11] to seize the channel-wise correlations. The SE block is constructed by integrating two distinct processing steps: *squeeze* and *excitation*. The *squeeze* step aggregates comprehensive global

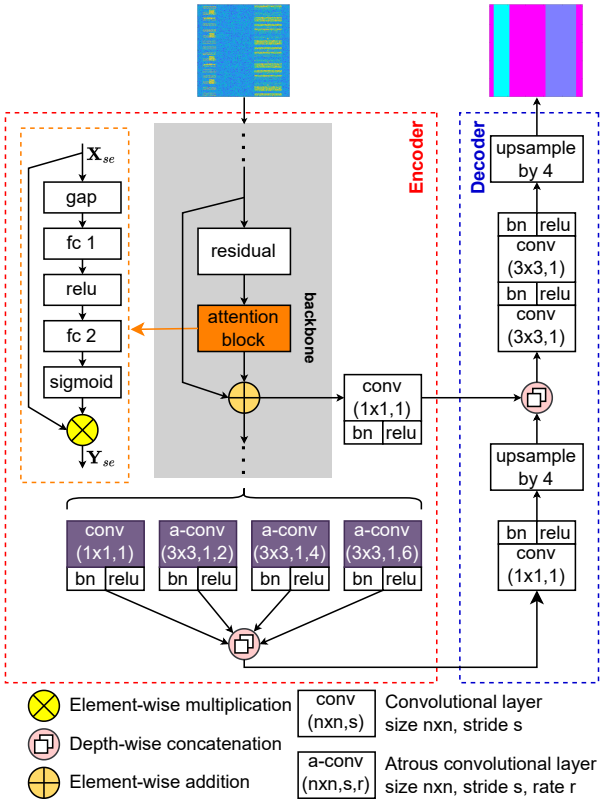


Figure 2: DeepLabV3+ architecture with adaptive ASPP block combined with an attention mechanism.

information from input feature maps (i.e., encompassing all spatial locations for each channel). To this end, it first constructs a global average pooling layer (gap), which computes the average for each channel across the entire spatial dimensions. The *squeeze* step can be represented by the following equation:

$$\mathbf{z}(c) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{X}_{se}(h, w, c), \tag{6}$$

where $\mathbf{X}_{se}$ denotes an input feature map of SE with the volume of $H \times W \times C$, H and W are the height and width of the feature map, respectively, and C is the number of channels; and $\mathbf{z}$ is the channel descriptor vector containing the average value of channel c. Next, the *excitation* step captures channel-wise dependencies by learning channel-specific important weights. It involves two fully connected (fc) layers (a.k.a., dense connected layers) interfered by an activation ReLU layer. The first fc layer is for dimensionality reduction of $\mathbf{z}$ and is followed by an activation layer to eliminate negative values. As stated in [11], the reduction ratio is 16 (i.e., the number of fc neurons). The second fc layer increases the dimensionality back to C with a sigmoid activation layer to obtain the non-excluding channel-wise attention scores $\mathbf{s} \in \mathbb{R}^C$. The mathematical representation of the *excitation* stage can be written as follows:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})), \tag{7}$$

Table 1: Channel and signal configurations.

Channel Parameters	Value	Unit
SNR	[0 20 40 60 80 100]	dB
Doppler	[0 10 500]	Hz
5G Parameters [2]	**Value**	**Unit**
SCS	[15 30]	kHz
Bandwidth	[10 15 20 25 30 40 50]	Mhz
SSB Period	20	ms
LTE Parameters [1]	**Value**	**Unit**
Reference Channel	[R.2, R.6, R.8, R.9]	
Bandwidth	[10 5 15 20]	MHz
Duplex Mode	FFD	

where $\mathbf{s}$ is the channel-wise attention scores; σ and δ are sigmoid and ReLU activation functions, respectively. $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively, are the weight matrices of two fully connected layers. Finally, the refined features $\mathbf{Y}_{se}$ are obtained by re-weighting the input feature maps $\mathbf{X}_{se}$ of the SE block with attention scores as follows:

$$\mathbf{Y}_{se} = \mathbf{s} \odot \mathbf{X}_{se}, \tag{8}$$

where $\odot$ represents the element-wise multiplication. To effectively exploit the SE block, it is crucial to identify suitable locations in the encoder to enhance the weighting information of the channels (i.e., emphasize relevant features and attenuate irrelevant or less meaningful features). Particularly, placing the attention block too close to the network's input can improve the correct classification rate of high-SNR signals. However, learning complex patterns of low and medium-SNR signals is inefficient. This outcome could be better when considering the imperfect nature of signal transmission environments. On the other hand, it is advisable to avoid locating the SE block too deep in the network architecture to prevent the increment of computational cost and parameter overhead as well as the dimension reduction of feature maps. Therefore, this block is recommended to place in the middle of an encoder before the feature map, which is transferred from the encoder to the decoder (as shown in Figure 2). This strategic placement enhances the information about the relationships between different channels, which is then transmitted to the decoder to facilitate the upsampling procedure.

3 PERFORMANCE EVALUATION AND DISCUSSIONS

3.1 Simulation Setups

Dataset generation: For performance evaluation, we generate a synthetic dataset of 5G and LTE signals, where the signals suffer from different channel impairments to simulate the real-world transmission phenomena [1, 2], by using 5G Toolbox and LTE Toolbox functions in MATLAB and making some changes to create signals that are similar which signals over-the-air. For more details, the specific channel transmission parameters and signal configurations are presented in Table 1.

We generate two separated datasets for model training and validation and performance measurement: (i) one set contains only 5G

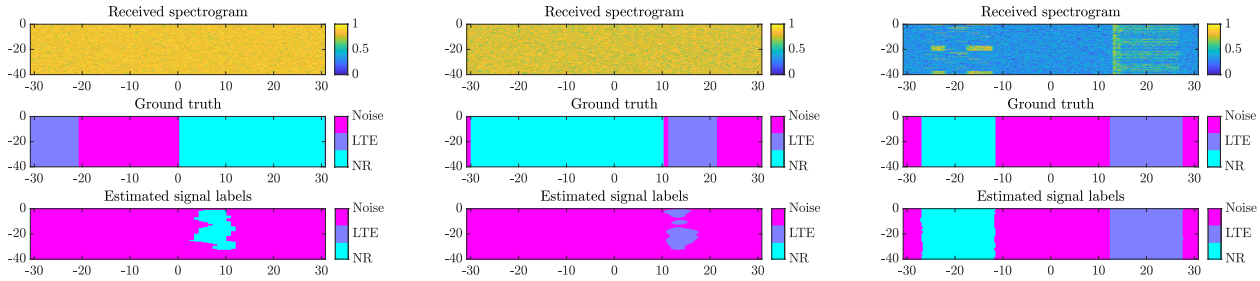


Figure 3: Visualization of the received spectrum, ground truth, and predicted labels at different SNR levels: (a) 0 dB, (b) 20 dB, and (c) 60 dB. The horizontal axis represents the frequency values of the signal spectrum, normalized in the range of $[-30, 30]$ megahertz (MHz), and the vertical axis represents time measured in milliseconds (ms), normalized from -40 to 0 ms.

NR or LTE signals, which are subsequently shifted to the frequency within the band of interest randomly, and (ii) another set contains both 5G NR and LTE signals occupying the same band of interest. Each signal frame has a length of 40 ms corresponding to the duration of 40 subframes. To support the most of current standard signals and facilitate compatibility with affordable software-defined radio (SDR) systems, the sampling rate is set to 61.44 MHz, providing approximately 50 MHz of valuable bandwidth. With the FFT length of 4096, the color spectrogram image has the size of 256×256 pixels. The first dataset, which consists of 20,000 spectrogram images equally distributed to 5G and LTE classes, is split into the training and validation sets with the ratio of 8/2 (i.e., 16,000 images for training and the remainder is for validation). For performance measurement in the testing phase, the second dataset has 10,000 images, where our proposed deep network should identify the regions of three classes (i.e., 5G, LTE, and noise) in the occupied spectrum images effectively.

Training configuration and evaluation metrics: In the training phase, we configure some following options to train our improved DeepLabV3+: the optimizer is stochastic gradient descent with the momentum (SGDM) algorithm with momentum factor 0.9 and L_2 regularization factor 0.0001, the maximum number of epochs is 40, the mini-batch size is 20, and the initial learning rate is 0.02 with a learning rate schedule of dropping rate 0.1 every 20-epochs period. Notably, the deep model with trained weights is returned to the epoch that the model achieves the lowest validation loss. To address the training set's imbalance, where 5G NR signals typically occupy a wider bandwidth than LTE signals and background noise fills the samples, we employ class weighting to alleviate the bias in learning caused by the unequal distribution of observations. The results are evaluated based on five common and widely-used metrics for semantic image segmentation [7]:

- *Global accuracy:* the ratio of correctly classified pixels to the total number of pixels.
- *Mean accuracy*: the average accuracy of all classes in all images, where accuracy is the ratio of correctly classified pixels to the total number of pixels in a class.
- *Mean IoU*: the average IoU score of all classes in all images, where IoU score is the ratio of correctly classified pixels to the total number of ground truth and predicted pixels in a class.
- *Weighted IoU*: the average IoU of each class, weighted by the number of pixels in a class.
- *Mean BFScores*: the average BF score of that class over all images, where the BF score indicates how well the predicted boundary of each class aligns with the true boundary.

Table 2: The performance of our improved DeepLabV3+ using ResNet50 as the backbone with ADR and ATM.

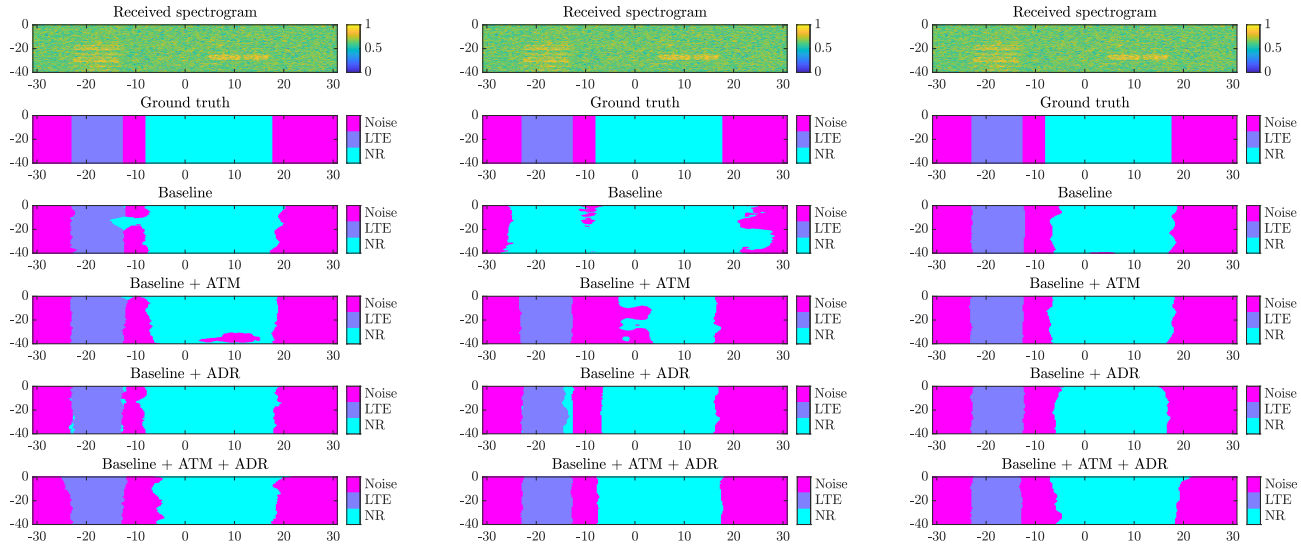
SNR (dB)	Global Accuracy	Mean Accuracy	Mean IoU	Weighted IoU	Mean BFScore
0	0.3900	0.3341	0.1330	0.1554	0.4353
20	0.3799	0.3336	0.1370	0.1547	0.3088
40	0.7853	0.7494	0.6237	0.6452	0.5296
60	0.9776	0.9750	0.9555	0.9562	0.9082
80	0.9855	0.9855	0.9721	0.9714	0.9378
100	0.9855	0.9855	0.9724	0.9714	0.9341
> 40	0.9828	0.9820	0.9666	0.9663	0.9267
Overall	0.7501	0.7261	0.5926	0.5988	0.6749

3.2 Results and Discussions

In the first simulation, we evaluate our improved DeepLabV3+ network with ResNet50 as the backbone in the encoder with ADR and ATM for spectrum segmentation on the generated dataset. Table 2 presents the results with different SNR levels, in which all the metrics improve along with the increase of SNR. It is evident that the model performs poorly at low SNR (e.g., 0 and 20 dB). The IoU scores reach around 0.13, indicating that the architecture just predicts a small region of low-SNR spectral signal correctly. The global accuracy of 5G-LTE-noise discrimination is low, approximately 0.37 to 0.39, to reveal the limitation of the discrimination capability of our model at low SNRs. Besides, the model performs poorly in terms of spectral signal boundary recognition as BFScore is less than 0.5 when SNR $\leq$ 20 dB. These metrics are significantly improved when the spectral signals become more obvious as corresponding to the increase of SNR. For example, the overall accuracy doubles at 40

Table 3: Performance comparison of our improved DeepLabV3+ with different backbones.

Backbone	ATM	ADR	Global Accuracy	Mean Accuracy	Mean IoU	Weighted IoU	Mean BFScore
MobileNetV2	✗	✗	0.7350	0.7116	0.5671	0.5788	0.5928
	✓	✗	0.7398	0.7139	0.5730	0.5846	0.6036
	✗	✓	0.7423	0.7141	0.5794	0.5877	0.6371
	✓	✓	0.7444	0.7166	0.5833	0.5909	0.6456
ResNet18	✗	✗	0.7284	0.6989	0.5588	0.5685	0.5824
	✓	✗	0.7184	0.6969	0.5524	0.5547	0.6154
	✗	✓	0.7338	0.7052	0.5672	0.5753	0.6011
	✓	✓	0.7384	0.7158	0.5697	0.5833	0.6031
ResNet50	✗	✗	0.7427	0.7181	0.5816	0.5889	0.6167
	✓	✗	0.7440	0.7168	0.5812	0.5898	0.6595
	✗	✓	0.7435	0.7150	0.5811	0.5891	**0.6771**
	✓	✓	**0.7501**	**0.7261**	**0.5926**	**0.5988**	0.6749



(a) **(b)** **(c)**

Figure 4: Results visualization with different backbones: (a) MobileNetV2, (b) ResNet18, and (c) ResNet50 at 40 dB (from top to bottom: the spectrogram of the received signal, ground truth of spectrum segmentation, and segmentation results obtained by the baseline (a.k.a., just DeepLabV3+ without any improvements), baseline+ATM, baseline+ADR, and baseline+ATM+ADR).

dB; however, the boundary detection, as indicated by the BRScore of 0.5296, remains vague with no any significant improvement. Our model achieves an astonishing performance growth with SNR > 40 dB, reaching the average global accuracy of 0.9828. The accurate segmentation regions take over 96% of the spectral image and the boundaries are detected more precisely with the average BFScore of 0.9267. To summarize, we provide an overall validation of the entire dataset to conduct a more comprehensive analysis of the semantic segmentation capability. Our improved DeepLabV3+ model yields a global accuracy of 0.7501 that is primarily caused by misclassified pixels at low SNR regimes. The mean IoU is around 0.5926 because the capability to detect the presence of signals in noisy conditions was humble. Additionally, we show three examples of received spectrogram images along with their actual signal labeling mask (a.k.a., ground truth) and the estimated signal labeling mask obtained by our improved deep model at different SNR regimes in Figure 3. The spectrogram images are represented in the frequency domain with horizontal and vertical axes denoting the signal frequency in megahertz (MHz) and the time in milliseconds (ms).

In the second simulation, we investigate the performance of our improved DeepLabV3+ with different backbones (including MobileNetV2, ResNet18, and ResNet50) in the encoder and further

analyze the effectiveness of two proposed components, i.e., ADR and ATM. As the quantitative results showed in Table 3, all the evaluation metrics improved when we applied ADR and ATM concurrently in DeepLabV3+ with different backbones. Compared with MobilenetV2 and ResNet18, ResNet50 with ADR and ATM yields the highest performance in all segmentation evaluation metrics except mean BFScore (slightly lower than ResNet50+ADR around 0.0022). With ResNet18 as the backbone, the global accuracy, mean IoU, and BFScore increase by 1.37%, 1.95%, and 3.56%, respectively, when compared to the baseline (i.e., DeepLabV3+ without ADR and ATM). In the case of setting MobileNetV2 as the backbone, these metrics show an improvement of approximately 1.28%, 2.85%, and 8.91%, respectively. Interestingly, the adaptive dilation rate mechanism is more important than the attention mechanism in the DeepLabV3+ with different backbones, wherein many evaluation metrics of ADR-enhanced models are higher than those of ATM-enhanced models. Integrating the ADR and ATM mechanisms into the encoder of DeepLabV3+ with different backbones significantly enhances the learning meaningful feature capacity, in which ADR and ATM allow to capture of spatial correlations in spectral images at multiple appropriate regions and comprehensively refine features (i.e., intensify relevant features and decline irrelevant features simultaneously). As a result, the precision and clarity of segmented boundaries for each region are improved to identify the spectrum regions of 5G NR and LTE more accurately. To provide a visual insight, we present the qualitative results of 5G-LTE spectrum segmentation at 40 dB, a commonly encountered SNR level in practical scenarios. In Figure 4, we show the spectrum segmentation masks obtained by DeepLabV3+ with various backbones, along with the adoption of ADR and ATM mechanisms. The sequential adoption of these mechanisms demonstrates a gradual enhancement in segmentation performance; particularly, more accurate signal recognition is realized by reducing the confusion in identifying 5G-LTE signals against noise in the frequency domain.

4 CONCLUSION

In this paper, we have improved DeepLabV3+, an encoder-decoder neural network architecture, to discriminate 5G NR and LTE signals in the scenario of spectrum occupancy. To this end, we first converted a received signal from a complex envelope form to a visually informative spectrogram image using FFT. Subsequently, an encoder-decoder network was developed based on DeepLabV3+ to segment the spectral regions of 5G and LTE against noise on spectrogram images. Remarkably, to boost the performance of spectrum segmentation, we contributed two improvements in the architecture of DeepLabV3+: (i) an adaptive dilation rate mechanism deployed ASPP to comprehensively compute spatially correlated features at different spectrum resolutions via atrous convolution and (ii) an attention mechanism consolidated in the encoder to intensively sophisticate the meaningful features. As a result, our improved DeepLabV3+ model increased the spectral signal segmentation capability, leading to more accurate signal identification based on segmented spectral regions on the spectrogram images. Based on simulated datasets with reasonable and reliable parameter settings, we evaluated the segmentation performance at various SNR levels and further investigated with different backbones to demonstrate the robustness and effectiveness of two improvement mechanisms.

REFERENCES

[1] 3GPP TS 36.101. 2019. *Evolved Universal Terrestrial Radio Access (E-UTRA);User Equipment (UE) radio transmission and reception.* Technical Specification (TS). 3rd Generation Partnership Project; Technical Specification Group Radio Access Network. Version 15.8.0.

[2] 3GPP TS 38.901. 2019. *Study on channel model for frequencies from 0.5 to 100 GHz.* Technical Specification (TS). 3rd Generation Partnership Project; Technical Specification Group Radio Access Network. Version 15.1.0.

[3] Anas Alarabi and Osama A. S. Alkishriwo. 2021. Modulation Classification Based on Statistical Features and Artificial Neural Network. In *2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA.* IEEE, Tripoli, Libya, 748–751.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2018), 834–848.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV).* Springer, Munich, Germany, 833—-851.

[6] Zhibo Chen, Yi-Qun Xu, Hongbin Wang, and Daoxing Guo. 2021. Deep STFT-CNN for Spectrum Sensing in Cognitive Radio. *IEEE Communications Letters* 25, 3 (2021), 864–868.

[7] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. 2013. What is a good evaluation measure for semantic segmentation?.. In *Proc. 24th British Mach. Vision Conf. (BMVC)*, Vol. 27. BMVA Press, University of Bristol, 10–5244.

[8] Jiabao Gao, Xuemei Yi, Caijun Zhong, Xiaoming Chen, and Zhaoyang Zhang. 2019. Deep Learning for Spectrum Sensing. *IEEE Wireless Communications Letters* 8, 6 (2019), 1727–1730.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* IEEE, Venice, Italy, 2980–2988.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Las Vegas, NV, USA, 770–778.

[11] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 8 (2020), 2011–2023.

[12] Thien Huynh-The, Toan-Van Nguyen, Quoc-Viet Pham, Daniel Benevides da Costa, and Dong-Seong Kim. 2022. MIMO-OFDM Modulation Classification Using Three-Dimensional Convolutional Network. *IEEE Transactions on Vehicular Technology* 71, 6 (2022), 6738–6743.

[13] Thien Huynh-The, Quoc-Viet Pham, Toan-Van Nguyen, Daniel Benevides da Costa, and Dong-Seong Kim. 2022. RaComNet: High-Performance Deep Network for Waveform Recognition in Coexistence Radar-Communication Systems. In *ICC 2022 - IEEE Int. Conf. Commun.* IEEE, Seoul, Korea, Republic of, 1–6.

[14] Sahan Damith Liyanaarachchi, Taneli Riihonen, Carlos Baquero Barneto, and Mikko Valkama. 2021. Optimized Waveforms for 5G–6G Communication With Sensing: Theory, Simulations and Experiments. *IEEE Transactions on Wireless Communications* 20, 12 (2021), 8301–8315.

[15] Naseer A. Mousa and Sattar B. Sadkhan. 2021. Identification of digitally modulated signal used in cognitive radio network -A survey. In *Proc. 1st Babylon Int. Conf. Inf. Technol. Sci. (BICITS).* IEEE, Babil, Iraq, 311–314.

[16] Thomas Brox Olaf Ronneberger, Philipp Fischer. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* Springer International Publishing, Cham, 234–241.

[17] Zesi Pan, Shafei Wang, Mengtao Zhu, and Yunjie Li. 2020. Automatic Waveform Recognition of Overlapping LPI Radar Signals Based on Multi-Instance Multi-Label Learning. *IEEE Signal Processing Letters* 27 (2020), 1275–1279.

[18] Quoc-Viet Pham, Nhan Thanh Nguyen, Thien Huynh-The, Long Bao Le, Kyungchun Lee, and Won-Joo Hwang. 2021. Intelligent Radio Signal Processing: A Survey. *IEEE Access* 9 (2021), 83818–83850.

[19] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).* IEEE, Salt Lake City, UT, USA, 4510–4520.

[20] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European conference on computer vision (ECCV).* Springer-Verlag, Munich, Germany, 3–19.