SRNet: Deep Semantic Segmentation Network for Spectrum Sensing in Wireless Communications

Thien Huynh-The, *Senior Member, IEEE*, Gia-Vuong Nguyen, Thai-Hoc Vu, *Member, IEEE*, Daniel Benevides da Costa, *Senior Member, IEEE*, and Quoc-Viet Pham, *Senior Member, IEEE*

***Abstract*—The evolution towards fifth-generation wireless (5G) and beyond has significantly increased the demand for efficient spectrum management and utilization. Conventional spectrum sensing methods have struggled to accurately characterize spectrum occupancy, particularly when different radio signals share the same frequency band. To address this challenge, we propose a novel spectrum sensing method by exploiting short-time Fourier transform and neural networks for learning spectrogram patterns. Leveraging encoder-decoder architectures, we design a semantic segmentation network, namely SRNet, to precisely detect multiple signals within a spectrum by identifying spectral content based on the frequency and time occupied by the signals. By incorporating an attention mechanism and multi-scale feature extraction, SRNet effectively learns spectral features and improves segmentation efficiency. Extensive simulations demonstrate SRNet's robustness and effectiveness in identifying 5G New Radio and LTE signals, under challenging channel and radio frequency impairments, making it a promising solution for next-generation spectrum sensing.**

***Index Terms*—5G NR, deep learning, LTE, semantic segmentation, signal identification, spectrum sensing.**

I. Introduction

The rapid evolution of fifth generation (5G), sixth generation (6G), and beyond has dramatically increased data transmission rates, driven by the exponentially growing number of mobile users and demand for high-quality services. To meet these needs, innovative spectrum utilization solutions are essential, therein spectrum sensing (SS) plays a key role in dynamic spectrum sharing and interference mitigation, accordingly optimizing frequency use. Reconfigurable intelligent surfaces (RIS) offer potential improvements to communication quality. A recent work [1] explored passive and active RIS by studying one-stage and two-stage optimization algorithms to maximize detection probability while addressing complexity. While increasing reflecting elements can enhance performance, the approach remains heavily reliant on channel state information (CSI). On the other hand, artificial intelligence (AI), including machine learning and deep learning (DL), enhances SS by enabling accurate signal identification without CSI. In 6G networks, AI-driven SS is crucial for dynamic spectrum management [2], thus improving network performance and allowing seamless coexistence of multiple radio components.

Over the last decade, DL has proven effective for intelligent SS in cognitive radio networks, accurately identifying signals based on their modulation type. Recent research, such as the ConvLSTM-based SS method in [3], showed significant improvements in detecting channel occupancy at low signal-to-noise-ratio (SNR) by learning spatio-temporal features from complex envelope data. Another method, described in [4], utilized time-frequency analysis with a simple deep architecture to learn spectral features, offering robust spectrum detection without prior knowledge of the primary signal. This method, though effective, may lack adaptability for dynamic spectrum environments in next-gen networks. The work [5] explored multiple DL architectures, including convolutional neural networks (CNNs), long short-term memory, and their combinations, treating SS as a binary classification problem to detect primary users. This study further underscored the potential of DL for robust spectrum management over radio signal classification. In [6], DeepSense, an unsupervised cooperative sensing approach, combines sparse autoencoders and Gaussian mixture models for representation learning and clustering. Although DeepSense effectively detects PUs using less labeled data and without prior knowledge of channel conditions, it is not suitable for identifying specific signal types like 5G NR and LTE. Recently, the work [7] explored the potential of Transformer models for feature extraction from spectral images. While Transformer models have shown notable success in signal classification, their application to semantic segmentation of spectral images is constrained by their high computational cost and challenges in capturing fine-grained spatial details. In short, while existing methods have made valuable contributions, they have not been able to identify the spectral content (i.e., frequency and time) within a wideband spectrogram when specific signals occupy the same band for the entire frame duration.

To overcome this challenge, we propose an efficient DL-based method capable of identifying spectral content in a wideband spectrogram as shown in Fig. 1. Our method leverages wideband spectrograms generated by short-term Fourier transform (STFT) to capture both the frequency and time characteristics of signals. To learn spectral patterns, we then build an innovative deep network with an encoder-decoder architecture, namely SRNet, to perform semantic segmenta-

This work is supported by Ho Chi Minh City University of Technology and Education (HCMUTE) under Grant No. T2024-114. Thien Huynh-The and Gia-Vuong Nguyen are with the Department of Computer and Communications Engineering, HCMC University of Technology and Education, Ho Chi Minh City 71307, Vietnam (email: thienht@hcmute.edu.vn, vuongng.cce@gmail.com). Thai-Hoc Vu is with the Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Republic of Korea (email:vuthaihoc1995@gmail.com). D. B. da Costa is with the Department of Electrical Engineering, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia (email: danielbcosta@ieee.org). Quoc-Viet Pham is with the School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, D02 PN40, Ireland (email: viet.pham@tcd.ie). Corresponding author: Gia-Vuong Nguyen.

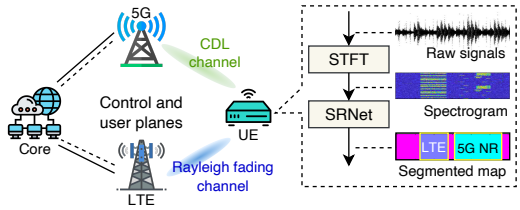

Fig. 1. An example of wireless network architecture for spectrogram-based signals identification.

tion of spectrogram images. This enables the detection and localization of multiple signals of interest in a spectrogram, where their locations correspond to the occupied frequency and time. To enhance spectral feature learning and improve segmentation accuracy, we strategically incorporate two novel components into SRNet's architecture: (i) an attention mechanism integrated with a residual connection to refine important features; and (ii) a multi-scale extraction mechanism with feature subdivision to collectively capture features at different sizes. Extensive simulations demonstrate the robustness and effectiveness of our proposed network in identifying 5G NR and LTE signals in the presence of noise and radio frequency impairments. Our proposed SRNet outperforms several advanced segmentation networks, including U-Net [8], FCN [9], DeepLabV3+ [10], BiSeNet [11], SegFormer [12].

II. Signal Model and Spectrum Representation

A. Signal Model

Within a typical wireless network environment, complex-valued 5G – LTE signals are affected by various noise sources. When the signal passes through the transmission channel, the received signal at the receiver can be generally modeled as

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

where $x(t)$ denotes the noise-free signal, $h(t)$ denotes the channel impulse response, and $n(t)$ denotes the additive noise. This work examines two widely used cellular communication signal types: 5G NR and LTE, along with additive noise at the receiver. Two sets of signal parameters are employed under imperfect channel conditions [13], [14], incorporating real-world effects like Doppler and link-level fading.

B. Spectrum Representation

In wireless communications, STFT is widely used to analyze the time-varying frequency content of signals. Unlike the conventional Fourier transform, which gives a global view of the entire signal's frequency spectrum, STFT segments the signal into short windows and computes the Fourier transform for each, accordingly revealing how the frequency content changes over time. STFT of a signal can be expressed as

$$X(t,f) = \sum_{n=0}^{N-1} x(n)\, w(n-t)\, e^{-i2\pi \frac{fn}{N}}, \tag{2}$$

where $x(n)$ is the discrete time-domain signal, $w(n-t)$ is the window function applied to each segment, $e^{-i2\pi \frac{fn}{N}}$ is the complex exponential factor representing the frequency component, and N is the FFT length (e.g., $N = 4096$ in this work).

TABLE I
Signal Configurations and Channel Conditions.

Channel parameters	**Range of value**	**Unit**
SNR	$\{20, 30, 40, 50, 60\}$	dB
Doppler	$\{0, 70, 750\}$	Hz
Center frequency	$\{1500, 2600, 3500, 4000\}$	MHz
5G parameters [13]	**Range of value**	**Unit**
Subcarrier spacing	$\{15, 30\}$	kHz
Bandwidth	$\{10, 15, 20, 25, 30, 40, 50\}$	MHz
SSB period	$\{5, 10, 20\}$	ms
Modulation	{QPSK, 16-QAM, 64-QAM}	
LTE parameters [14]	**Range of value**	**Unit**
Reference channel	{R.2, R.3, R.5, R.6, R.8, R.9}	
Bandwidth	$\{5, 10, 15, 20\}$	MHz
Modulation	{QPSK, 16-QAM, 64-QAM}	

STFT provides a time-frequency representation that captures the spectral content of wireless signals including 5G NR, LTE, WiFi, and Bluetooth, by transforming complex envelope waveforms into interpretable spectrograms. This representation in combination with advances of DL enables highly accurate signal recognition.

III. Methodology

A. Data Preparation

For performance evaluation, we generated a challenging 5G-LTE signal dataset to cover a broader range of frequency bands and signal configurations as described in Table I, including signals with varying center frequencies and modulation schemes, allowing to evaluate the model's robustness across different communication environments. Raw signals with a frame length of 40 ms and a sampling rate of 61.44 MHz were processed using STFT to produce color spectrogram images with a resolution of 256×256. The dataset is divided into two subsets: the first subset contains $32,000$ spectrograms of individual 5G or LTE signals with a split $80/20$ for training and validation; and the second subset consists of $16,000$ synthetic spectrograms representing both 5G and LTE signals used for testing. The spectrograms are distributed across varying SNR levels. The dataset and source code are available on GitHub[1].

B. SRNet: Deep Network for Spectrogram-based Signals Identification

In this section, we propose an efficient semantic segmentation model, named SRNet, for 5G–LTE spectrum recognition with the network architecture shown in Fig. 2.

Encoder with Attention Mechanism (AM): To optimize the model's segmentation performance, a backbone is commonly employed within the encoder to ensure comprehensive feature extraction. Among the available options, ResNet [15] has proven to be particularly effective as a backbone in semantic segmentation models, particularly in the context of SS [10]. While deeper networks such as ResNet50 and ResNet101 can potentially enhance performance by capturing more complex features, they also impose substantially higher computational costs and extended processing times. Consequently, in our work, ResNet18 was selected to maintain model efficiency,

[1]https://github.com/ThienHuynhThe/SpectrumSensing5G

particularly in applications where real-time processing and resource constraints are critical considerations. To enhance the ResNet18 backbone's ability to capture relevant features from spectrograms, we integrate a channel-based attention mechanism called the *squeeze-and-excitation* (SE) [16] block. This mechanism prioritizes important feature maps and suppresses less relevant ones, thereby effectively focusing on the important spectral features in 5G and LTE signals. Given the concentrated spectral information in these signals, the SE block helps capture their essential frequency-domain patterns, accordingly enhancing feature learning for spectrum sensing. The SE block, being lightweight, is particularly well-suited for SS tasks. In our design, the SE block is incorporated at the end of stage 2 in the backbone, where it enhances the interaction of meaningful feature maps before they are passed to the decoder. This allows for more effective recovery of key spectral details, thereby improving segmentation mask accuracy. To do this, a set of attention scores $\mathcal{S} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}}$ (where the feature volume is defined by the height H, the width W, and the depth D) is studied to clarify the relationship across multiple channels in the feature maps. The mathematical formulation of the attention scores can be expressed as follows:

$$\mathcal{S} = \sigma\left(\mathbf{W}_2\left(\theta(\mathbf{W}_1\left(\mathcal{G}\right))\right)\right), \tag{3}$$

where $\mathcal{S}$ represents the vector of attention scores obtained after passing through two fully connected (`fc`) layers, denoted by $\mathbf{W}_1$ and $\mathbf{W}_2$ having the numbers of hidden neurons are 64 and 16, respectively. It is worth noting that each `fc` layer is followed by the corresponding activation functions, where θ symbolizes the rectified linear unit function (`relu`) and σ is the sigmoid (`sigmoid`) function. Moreover, $\mathcal{G} \in \mathbb{R}^{1\times1\times\mathrm{D}}$ denotes the global average pooling (`gap`) operation, subsequently incorporated into the input feature maps $\mathbf{X}_{\mathrm{SE}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}}$. The operation of the `gap` layer can be written as

$$\mathcal{G} = \frac{1}{\mathrm{H}\times\mathrm{W}}\sum_{h=1}^{\mathrm{H}}\sum_{w=1}^{\mathrm{W}}\mathbf{X}_{\mathrm{SE}}(h,w). \tag{4}$$

To manipulate the attention scores, we involve element-wise multiplication, signified by $\odot$, with the input feature maps $\mathbf{X}_{\mathrm{SE}}$, effectively modulating the feature channel importance:

$$\mathbf{Y}_{\mathrm{SE}} = \mathcal{S} \odot \mathbf{X}_{\mathrm{SE}}, \tag{5}$$

where $\mathbf{Y}_{\mathrm{SE}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}}$ is the output feature maps with useful information enhanced by attention scores. As a result, the SE block strengthens relevant features and weakens extraneous features simultaneously to improve learning efficiency.

Multi-scale Extraction with Feature Subdivision: The primary concept behind multi-scale feature extraction is the parallel execution of convolutional layers with varying receptive fields. This approach enhances the extracted features across different scales but significantly increases complexity due to the large resolution of feature maps. To overcome this drawback, the information extracted from the backbone is processed through an innovative module, called feature subdivision (FS) with the structure illustrated in Fig. 2, to reduce complexity and facilitate multi-scale feature extraction.

By leveraging the periodic characteristics of 5G NR and LTE signals after applying STFT, it is possible to observe distinct patterns in the spectrogram. Within a frame length, the

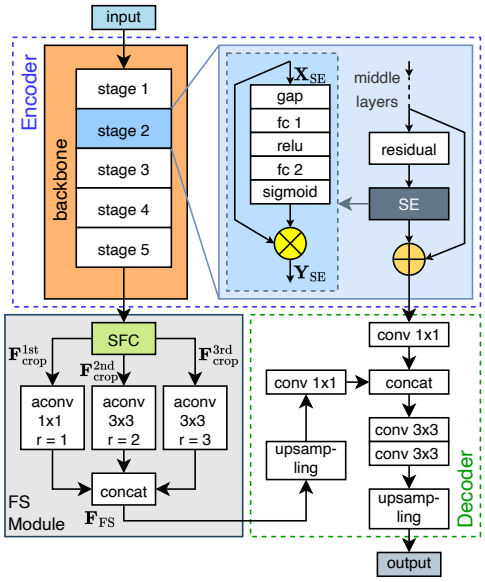

Fig. 2. SRNet for segmenting the signal spectrum: the orange area is the backbone (i.e., ResNet18) with five stages, the light blue area provides a detailed explanation of our enhancement with attention mechanism in the stage 2, and the gray region indicates the structure of the FS module.

amplitude values at each frequency exhibit periodic similarities, thus reflecting the inherent structure of the signals. These periodic characteristics are preserved during feature extraction using convolutional layers. When scanning through the spectrogram's feature map, convolutional kernels detect similar features across time and frequency, consequently resulting in consistent feature maps that highlight the periodic patterns of the signals. Inspired by this observation, we introduce a subframe cropping (SFC) mechanism to divide and reduce the resolution of feature maps along the vertical axis. This approach maintains the signal's distribution characteristics in the frequency domain while reducing the complexity of the feature maps by subdividing the frame length into smaller subframes, thereby facilitating subsequent multi-scale extractions. In detail, based on the input feature map $\mathbf{F} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}}$, we apply the SFC to generate three sub-parts with the same size $\mathbf{F}_{\mathrm{SFC}} \in \mathbb{R}^{\frac{\mathrm{H}}{2}\times\frac{\mathrm{W}}{2}\times\mathrm{D}}$, enabling to minimize computational costs and ensure the amount of information retained about the signal characteristics. The operation of SFC can be expressed as

$$\mathbf{F}_{\mathrm{SFC}} = \{\mathbf{F}[i,j] \mid i \in [u, u+\frac{\mathrm{H}}{2}], j \in [v, v+\mathrm{W}]\}, \tag{6}$$

where $[i,j]$ denotes the position of a pixel within the feature map and $\{u,v\}$ is the position of the pixel located in the top left corner of the cropped area. In the implementation, as the input feature map $\mathbf{F}$ with the size of 16×16, the first part $\mathbf{F}^{1\mathrm{st}}_{\mathrm{SFC}}$ comprises the upper half of the original map with $\{u,v\} = \{0,0\}$, the second part $\mathbf{F}^{2\mathrm{nd}}_{\mathrm{SFC}}$ consists of the lower half with $\{u,v\} = \{4,0\}$, and the third part $\mathbf{F}^{3\mathrm{rd}}_{\mathrm{SFC}}$ focuses on the central area with $\{u,v\} = \{8,0\}$. This configuration provides 50% overlap between these parts and ensures the effectiveness of multi-scale feature extraction.

As each part is processed by a corresponding atrous convolution (`aconv`), the three parts can operate in parallel to extract multi-scale information from the respective feature maps. As

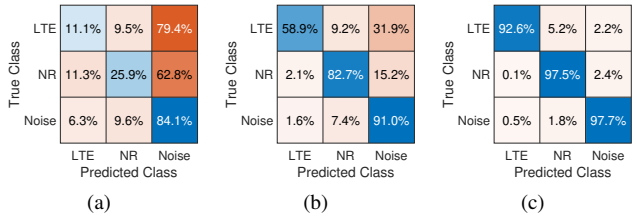

Fig. 3. Confusion matrices of spectral content segmentation at various SNR regimes: (a) 20 dB, (b) 40 dB, (c) and 60 dB.

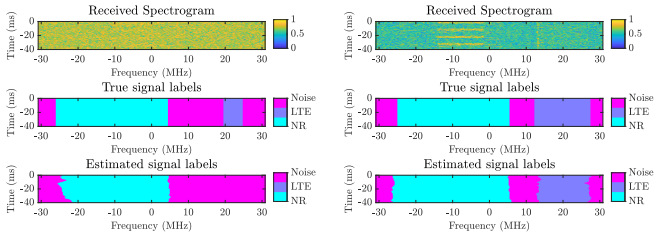

Fig. 4. Visualization of the spectrogram images, ground truth masks, and the results of SRNet at different SNR regimes: (a) 20 dB, and (b) 60 dB.

a result, the output of an `aconv` layer with the kernel $\mathbf{W}_a$ and the input $\mathbf{X}_a$ at a coordinate $[i, j]$ can be computed as

$$\mathbf{Y}_a[i,j] = \sum_{u=1}^{\mathrm{U}} \sum_{v=1}^{\mathrm{V}} \mathbf{X}_a[i+ru, j+rv] \times \mathbf{W}_a[u,v] + b, \quad (7)$$

where U and V are the height and width of the kernel, respectively, and b is the arbitrary bias value. It is known that atrous convolution is characterized by an additional hyper-parameter, called the dilation rate r. This parameter controls the spacing between elements within the kernel to expand the receptive field effectively. By increasing r, the kernel can capture information from a wider spatial area in the input feature map without increasing the number of learnable parameters. In our proposed mechanism, where the input spatial size of `aconv` is 8×16, the dilation rate should be lower than 3 to prevent the receptive field from exceeding the feature map size. The computation of FS can be mathematically represented as

$$\mathbf{F}_{\mathrm{FS}} = \langle \mathcal{A}_{1,1}^{1\times1}(\mathbf{F}_{\mathrm{SFC}}^{\mathrm{1st}}), \mathcal{A}_{1,2}^{3\times3}(\mathbf{F}_{\mathrm{SFC}}^{\mathrm{2nd}}), \mathcal{A}_{1,3}^{3\times3}(\mathbf{F}_{\mathrm{SFC}}^{\mathrm{3rd}}) \rangle, \quad (8)$$

where $\langle . \rangle$ denotes the depth-wise concatenation (`concat`) layers, $\mathbf{F}_{\mathrm{FS}}$ is the output of FS module, $\mathcal{A}_{s,r}^{n\times n}$ is the atrous convolution with kernel size $n \times n$, stride s, and rate r.

Decoder: The decoder in SRNet is structured by a sequence of regular `conv` layers, upsampling layers, and `concat` layers. As shown in Fig. 2, two upsampling layers increase the feature maps' spatial dimensions by 4 times, thus allowing the restoration of the segmented mask to the same spatial size of an input spectrogram image. Notably, attention-enhanced feature maps from stage 2 are passed to the first upsampling layer via a skip connection and a `concat` layer to ensure that essential signal information is preserved. This structural connection improves the accuracy of spectral content identification in SRNet's segmentation mask, accordingly enhancing the model's performance.

C. Training Configurations

SRNet leverages the cross-entropy loss function and utilizes pre-trained weights of ResNet18 in the encoder. During the training stage, the stochastic gradient descent with momentum algorithm optimizes the network with the momentum of 0.9 and an initial learning rate of 0.02 that decays by 0.1 every 20 epochs. Training is conducted for 100 epochs with a mini-batch size of 20 on an RTX 2080 GPU. Model performance is evaluated using the global accuracy, the mean intersection-over-union (IoU), and mean F1 score metrics besides the number of learnable parameters (params) and execution time (time) for efficiency measurement.

IV. Experiments and Discussions

This section evaluates the performance of SRNet for wireless signal identification on the 5G–LTE spectrum dataset under different frequency bands and signal configurations via three comprehensive experiments.

The first experiment comprehensively evaluates the spectrum segmentation-based signal identification achieved by SRNet on synthetic spectrogram images containing both 5G NR and LTE signals. Fig. 3 presents the confusion matrices of multi-class pixel-level image segmentation at different SNR regimes. The accuracy of 5G NR and LTE at 20 dB (see Fig. 3(a)) is notably poor due to significant signal quality distortion caused by additive noise. A high misclassification rate of LTE is likely due to the relatively small number of pixels representing LTE in the spectrum, thus resulting in insufficient features for effective learning. The confusion matrices for 40 dB in Fig. 3(b) and 60 dB in Fig. 3(c) show some remarkable accuracy improvements compared to 20 dB. Specifically, the accuracy for LTE signals increases significantly by 47.8% and 81.5% at 40 dB and 60 dB, respectively. For 5G NR signals, the improvements are 56.8% and 71.6% the same SNR regimes. Notably, insufficient features extracted in spectrogram images of LTE compared to 5G NR across different SNR levels may lead to higher misclassification rates. Fig. 4 visualizes two examples of spectrum segmentation at different SNR levels, wherein SRNet is unable to segment the LTE spectrum region at 20 dB due to various channel impairments.

The second experiment investigates the impact of AM and FS modules on the overall segmentation performance of SRNet. As reported in Table II, these modules significantly improve the model's performance across all metrics with a slight trade-off in model complexity. The baseline model, consisting only of the encoder and decoder parts, achieves a global accuracy of 71.82%, a mean IoU of 52.65%, and a mean F1 score of 49.07%. Compared to the baseline, the AM configuration achieves a slight accuracy increase 0.68%, a noticeable IoU improvement of 3.11%, and a higher F1 score of 3.32%. Similarly, incorporating an FS module for multi-scale information integration from the encoder leads to further gains in performance. Remarkably, the best-performing model, including both AM and FS modules in the architecture to selectively capture and gather the important features at different resolutions for a more effective spectrum pattern learning, achieves the highest results of all metrics. While improving segmentation performance often increases parameters

TABLE II
ABLATION STUDY OF SRNET WITH DIFFERENT MODULES. THE BASELINE CONSISTS SOLELY OF AN ENCODER AND A DECODER.

Method	Global Acc	Mean IoU	Mean F1 Score	Params (M)	Time (ms)
Baseline	71.82	52.65	49.07	16.82	5.04
Baseline + AM	72.31	54.29	50.70	16.84	5.10
Baseline + FS	72.69	54.33	51.08	19.30	5.33
Baseline + AM + FS	72.96	55.05	53.85	19.32	5.41

TABLE III
METHOD COMPARISON WITH DIFFERENT STATE-OF-THE-ART SEGMENTATION NETWORKS FOR SPECTRUM SENSING.

Method	Global Acc	Mean IoU	Mean F1 Score	Params (M)	Time (ms)
U-Net [8]	69.57	51.45	49.02	31.34	8.33
FCN [9]	70.13	51.97	50.19	42.72	9.45
DeepLabV3+ [10]	71.69	53.21	51.84	20.61	6.74
BiSeNet [11]	72.74	53.73	52.29	49.12	7.63
SegFormer [12]	72.92	54.23	53.41	24.98	9.88
SRNet (Proposed)	72.96	55.05	53.85	19.32	5.41

and processing time, our proposed model with AM and FS maintains efficiency. Thanks to primarily focusing on learning critical weights, AM conduct to a very slight increase in parameters. In contrast, FS requires additional layers but the increase in parameters has a negligible impact on processing speed. As a result, our model executes in around 5.41 ms, thus making it suitable for real-time applications.

Table III presents a comparison of SRNet with several state-of-the-art segmentation networks for SS, where the proposed model outperforms them across all performance metrics. Concretely, SRNet achieves the highest global accuracy of 72.96%, mean IoU of 55.05%, and mean F1 score of 53.85%, while maintaining a relatively low model complexity. Compared to U-Net, SRNet offers significantly better performance with fewer parameters. FCN, while offering slight improvements over U-Net, suffers a substantial increase in model size with over 42M parameters. Despite being more lightweight than FCN, DeepLabV3+ performs segmentation more precisely by 2.22% global accuracy, 2.38% mean IoU, and 3.29% mean F1 score. While BiSeNet achieves comparable performance metrics, it is the heaviest model with over 49M parameters. Notably, SegFormer presents a strong competitor with competitive performance across all metrics. Compared to SegFormer, SRNet outperforms with a small margin in both mean IoU and mean F1 score, while maintaining a lower number of trainable parameters (24.98M of SegFormer versus 19.32M of SRNet). SRNet outperforms all comparison models in processing time, achieving 5.41 ms, compared to DeepLabV3+ (6.74 ms) and BiSeNet (7.63 ms), both of which are optimized for real-time tasks. This demonstrates SRNet's superior efficiency and segmentation performance, thus making it highly suitable for real-time applications. Additionally, SRNet's ability to deliver better results with fewer parameters enhances its practicality. Simulations using a realistic dataset under real-world conditions confirm SRNet's potential to accurately identify spectral bands while maintaining real-time responsiveness, making it an ideal solution for spectrum sensing in next-generation wireless networks.

V. CONCLUSION

In this study, we present SRNet, a high-performance CNN model tailored for spectrum occupancy monitoring by extracting spectral features from synthetic 5G and LTE spectrograms. SRNet leverages an attention mechanism with residual connections and multi-scale feature extraction with feature subdivision to refine and selectively capture crucial spectrum features, significantly enhancing segmentation accuracy. The model achieves outstanding results, with an average global accuracy of 72.96% across various channel impairments and a processing time of 5.41 ms, outperforming state-of-the-art networks such as DeepLabV3+, BiSeNet, and SegFormer. Future work will focus on integrating noise removal techniques and advanced DL methods, extending the model to low SNR conditions and exploring its potential in diverse wireless communication scenarios, including mmWave and urban multipath environments.

REFERENCES

[1] H. Xie, D. Li, and B. Gu, "Enhancing spectrum sensing via reconfigurable intelligent surfaces: Passive or active sensing and how many reflecting elements are needed?" *IEEE Trans. Wireless Commun.*, pp. 1–1, Jan. 2024.

[2] Z. Wei *et al.*, "Integrated sensing and communication signals toward 5G-A and 6G: A survey," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11 068–11 092, Jul. 2023.

[3] Q. Wang, B. Su, C. Wang, L. P. Qian, Y. Wu, and X. Yang, "ConvLSTM-based spectrum sensing at very low SNR," *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 967–971, Jun. 2023.

[4] Z. Chen, Y.-Q. Xu, H. Wang, and D. Guo, "Deep STFT-CNN for spectrum sensing in cognitive radio," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 864–868, Nov. 2021.

[5] S. N. Syed *et al.*, "Deep learning approaches for spectrum sensing in cognitive radio networks," in *Proc. WPMC*, Herning, Denmark, Nov. 2022, pp. 480–485.

[6] N. A. Khalek and W. Hamouda, "DeepSense: An unsupervised deep clustering approach for cooperative spectrum sensing," in *Proc. ICC*, Rome, Italy, Jun. 2023, pp. 1868–1873.

[7] W. Zhang, Y. Wang, X. Chen, Z. Cai, and Z. Tian, "Spectrum transformer: An attention-based wideband spectrum detector," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 12 343–12 353, Jan. 2024.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234–241.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[10] G.-V. Nguyen, C. V. Phan, and T. Huynh-The, "Accurate spectrum sensing with improved DeepLabV3+ for 5G-LTE signals identification," in *Proc. SOICT*, Ho Chi Minh, Dec. 2023, p. 221–227.

[11] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 325–341.

[12] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12 077–12 090, Dec. 2021.

[13] 3GPP TS 38.901, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project; Technical Specification Group Radio Access Network, Technical Specification, Sep. 2019, version 15.1.0.

[14] 3GPP TS 36.101, "Evolved universal terrestrial radio access; user equipment radio transmission and reception," 3rd Generation Partnership Project; Technical Specification Group Radio Access Network, Technical Specification, Sep. 2019, version 15.8.0.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, USA, Jun. 2016, pp. 770–778.

[16] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.