Spectrum Transformer: An Attention-Based Wideband Spectrum Detector

Weishan Zhang, *Student Member, IEEE*, Yue Wang, *Senior Member, IEEE*, Xiang Chen, *Member, IEEE*, Zhipeng Cai, *Fellow, IEEE*, and Zhi Tian, *Fellow, IEEE*

***Abstract*— Data-driven machine learning techniques have been advocated for signal detection in complex wireless environments. However, when applied to wideband spectrum sensing scenarios, they face practical challenges including very large data dimensionality, insufficient training data, and implicit inter-band dependencies. Current literature focuses on deep convolutional models, whose inherent model structure is not well suited for representing the diverse spectrum occupancy patterns of practical wideband networks, causing inefficient performance-complexity tradeoff and excessive sensing time. To address these issues, this paper develops a novel Spectrum Transformer with multi-task learning for wideband spectrum sensing at high sample efficiency. Empowered by the multi-head self-attention mechanism, the transformer architecture is designed to effectively learn both the inner-band spectral features and the inter-band spectrum occupancy correlations in the wideband regime. Simulations show that the proposed Spectrum Transformer outperforms the existing methods based on convolutional neural networks especially in the small-data case, by achieving higher sensing accuracy with an 89% reduction in model complexity.**

***Index Terms*— Spectrum transformer, cognitive radio, wideband spectrum sensing, deep neural network, multi-head self-attention mechanism.**

I. Introduction

FOR spectrally-efficient wireless communications, dynamic spectrum access has been advocated that allows cognitive radio (CR) users to opportunistically access some spectrum bands that are unoccupied by primary users (PUs) at certain time and space. A key CR technology is spectrum sensing, in which CRs quickly identify access opportunities by detecting PUs' dynamic occupancy over a targeted spectrum pool [2], [3]. Traditional spectrum detectors are designed using signal processing techniques that hinge on the knowledge of some physical models of the system, such as that of the channels, noise, and transmitted signals. These detectors, including energy detector, matched filter and cyclic feature detector [4], [5], [6], achieve good sensing performance under ideal conditions where the assumed models of channels and signals are accurate and known *a priori*. However, this assumption often does not hold in practical CR systems, rendering these methods vulnerable to detrimental model mismatch issues in the presence of channel or noise uncertainty, and unknown signal types [7].

To overcome the limitations of traditional signal processing techniques in complex wireless environments, data-driven spectrum sensing methods using deep neural networks (DNNs) have been developed recently [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. Capitalizing on the powerful representation capability of DNN models, these methods implicitly learn the underlying complex physical models from training data. Existing spectrum learning methods mostly adopt the convolutional neural network (CNN) model structure [18]. Originated for image classification in computer vision, the convolutional filtering mechanism in CNN can efficiently learn the geometric correlation of signal components that are adjacent in either frequency or temporal domain within neurons' small receptive fields [19]. Hence, CNN-based detectors have been used to capture the inner-band spectral features of PU signals for narrowband detection [8], [9], [10], [11], [12], [13].

In order to seek as many spectrum opportunities as possible, CR detectors are usually expected to monitor a large spectrum pool, which leads to the wideband spectrum sensing problem. Apart from a straightforward yet inefficient way of deploying multiple narrowband detectors in parallel, a few works advocate to design CNN-based multi-band detection models [14], [15], [16], [17]. CNN-based classifiers for multi-band occupancy pattern classification are developed [14], [15], but they are not well suited to handle complicated occupancy patterns in the presence of multiple PUs. A wideband CNN framework called DeepSense is developed to process wideband I/Q samples in temporal domain [16], while YOLO, a popular CNN architecture for object detection, is applied on the received spectrograms to estimate the time-frequency range of PU signals [17]. The CNN architecture, by virtue of its built-in convolutional filters, can efficiently learn useful spectral features embedded in small-scale spectral correlations,

Manuscript received 18 September 2023; revised 30 January 2024; accepted 5 April 2024. Date of publication 26 April 2024; date of current version 12 September 2024. This work was supported in part by U.S. NSF under Grant 2003211, Grant 2128596, Grant 2231209, Grant 2146497, Grant 2244219, Grant 2315596, and Grant 2413622. An earlier version of this paper was presented in part at the IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2023), Shanghai, China, September 2023 [DOI: 10.1109/SPAWC53906.2023.10304551]. The associate editor coordinating the review of this article and approving it for publication was M. Caleffi. *(Corresponding author: Yue Wang.)*

Weishan Zhang, Xiang Chen, and Zhi Tian are with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030 USA (e-mail: wzhang23@gmu.edu; xchen26@gmu.edu; ztian1@gmu.edu).

Yue Wang and Zhipeng Cai are with the Department of Computer Science, Georgia State University, Atlanta, GA 30303 USA (e-mail: ywang182@gsu.edu; zcai@gsu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2024.3391515.

Digital Object Identifier 10.1109/TWC.2024.3391515

but may become inefficient for wideband spectrum sensing scenarios featuring in long-range spectral correlations.

To elaborate on the limitations of CNNs for wideband sensing, it is worth highlighting the very large dimensionality of the wideband spectrum learning problem where sensing at high frequency resolution is desired over a very large spectrum pool of wide bandwidth. To process fine-grained spectrum measurements of large dimensionality from all the frequency bands of interest, the computational complexity is a major consideration in deciding a suitable learning model. Meanwhile, the spectrum patterns in modern wideband systems often exhibit long-range, inter-band correlations among frequency components that are non-adjacent or even far apart, e.g, due to power leakage [20], non-contiguous channel allocation for multimedia communications, and carrier aggregation in LTE-A mobile networks [21]. Because of the inherent convolutional filtering mechanism, CNN-based models can capture the correlation among local input entries at each layer within a small receptive field, but have to considerably increase the layer depth and the number of neurons per layer in order to reach adequate model capability for representing long-range correlations. This issue is not conspicuous in narrowband sensing, and thus overlooked in the literature [15], [16]. However, for wideband sensing, CNN cannot rely on a compact model to capture both inner-band and inter-band dependencies over very wide bandwidth, which mandates the use of large and deep CNN models for high-resolution sensing. Training such large models not only incurs expensive computational cost, but also requires a large volume of training data in order to circumvent sensing performance degradation caused by the overfitting issue. As such, existing work on wideband CNNs mainly focus on the situations with relatively large training datasets [14], [15], [16], [17]. However, practical CR devices may operate under constrained resources in terms of energy, computing power and training data availability, rendering the CNN architecture inefficient for wideband spectrum sensing. It is the convolutional filtering mechanism of CNNs that causes the inefficiency in handling long-range dependencies.

The goal of this paper is to develop an effective deep learning architecture for accurate wideband spectrum sensing with high computational efficiency and sample efficiency. The key is to identify a compact model structure with high representation capacity that can efficiently capture both the inner-band dependency within a narrow frequency range and the inter-band long-range correlations among non-contiguous channels, as featured in practical wideband PU networks. We introduce the multi-head self-attention (MSA) mechanism in the Transformer models to wideband spectrum sensing. Originally developed for natural language processing, the MSA-based transformer architecture offers well-appreciated capability in capturing long-distance dependencies among different words in an sentence [22]. For wireless communication applications, MSA also shows superior performance in solving various problems, such as channel estimation in the presence of Doppler frequency shifts experienced by highly dynamic users [23], adaptive live streaming in wireless edge networks [24], and modulation classification based on time-series input [25]. In these works, different Transformer architectures are developed to capture temporal correlations of the target signals. In contrast, we develop a new *Spectrum Transformer* framework that is well suited to capture the various patterns of spectral correlations for wideband spectrum occupancy detection. Main contributions of this work are:

- To the best of our knowledge, this is the first work that introduces the self-attention mechanism to wideband spectrum sensing. The MSA mechanism can effectively learn not only the inner-band spectral features of modulated signals but also the inter-band dependencies across non-contiguous bands in the wideband spectrum pool.
- A wideband *Spectrum Transformer* framework is developed through a full stack of judiciously designed functional modules to facilitate spectrum occupancy detection from wideband PSD data. For efficient implementation, the output layer design adopts a multi-task learning approach to multi-band detection, which considerably improves the sample efficiency in the wideband regime and offers high resistance to the over-fitting issue.
- We evaluate the proposed Spectrum Transformer in terms of sensing accuracy and model/computation complexity, compared with existing CNN-based spectrum sensing approaches. In view of the scarcity of available training data of wideband spectrum measurements, we also present a data labeling approach to generate synthetic wideband spectrum datasets for various modulation types and channel conditions. Simulation results corroborate that our Spectrum Transformer achieves high sensing accuracy at reduced model complexity and computation cost than the benchmarks.

The rest of this paper is organized as follows. The signal model and overview of wideband spectrum sensing problem are presented in Section II. The design of Spectrum Transformer architecture and our proposed wideband spectrum detector are described in Section III. Numerical simulation results are presented in Section IV, followed by conclusions in Section V.

II. Signal Model and Problem Statement

This section describes the signal model for the wideband spectrum sensing problem, and highlights some key technical challenges to be addressed in data-driven spectrum learning.

A. Signal Model of Wideband Measurements

Consider a wideband spectrum pool that is uniformly divided into N_f frequency bands, i.e., narrowband channels. Spectrum occupancy by PUs on these bands is indicated by a Boolean vector $\mathbf{y} = [y_1, y_2, \ldots, y_{N_f}] \in \mathbb{B}^{N_f}$, where y_n takes the binary value of either 1 (the n-th band is occupied by a PU) or 0 (unoccupied). A CR detector monitors the entire wideband spectrum, whose received signal is adequately sampled (at or above the Nyquist rate) to form a time sequence as follows:

$$\mathbf{x} = \sum_{n=1}^{N_f} y_n h_n \mathbf{x}_n + \mathbf{w}, \tag{1}$$

where $\mathbf{x}_n$ denotes the modulated signal transmitted on band-n if it is occupied by a certain PU (i.e., $y_n = 1$), $\mathbf{w}$ denotes the additive noise, and h_n represents the channel gain formulated

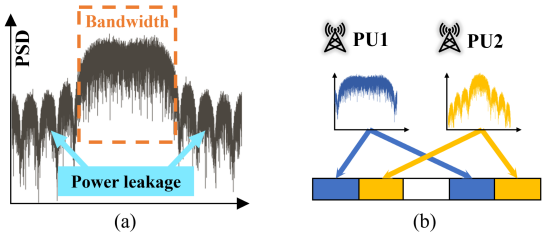


Fig. 1. (a) Power leakage and (b) channel aggregation.

as $h_n = (\beta\,(d_0/d_n)^{\alpha}\,10^{\frac{-\psi_n}{10}})^{0.5}$ [26], where β is a constant related to the antenna characteristics and average attenuation, α is the path-loss exponent, d_n is the distance between the SU and the PU on band-n, d_0 is the reference distance, and ψ_n is a Gaussian-distributed random variable with mean zero and variance $\sigma^2_{\psi_n}$ that measures the shadow fading of the channel over band-n between PU and SU.

To delineate the spectral features of $\mathbf{x}$, we calculate its power spectral density (PSD):

$$\mathbf{s} = \mathbb{FT}(Corr(\mathbf{x})), \tag{2}$$

where $\mathbb{FT}(.)$ denotes Fourier transform, and $Corr(.)$ denotes the autocorrelation of a signal.

For sensing at high spectral resolution, these PSD measurements are collected at fine grain with N_w ($\gg 1$) frequency samples per band, denoted by $\mathbf{s}_n \in \mathbb{R}^{N_w}$ for the n-th band, $\forall n$. Then, the wideband measurement $\mathbf{s} \in \mathbb{R}^{N_w N_f}$ can be regarded as a concatenation of N_f band-wise PSD vectors:

$$\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_n, \ldots, \mathbf{s}_{N_f}].$$

The task of wideband sensing is to estimate the spectrum occupancy vector $\mathbf{y}$ from $\mathbf{x}$ in (1) or equivalently its PSD $\mathbf{s}$.

B. Spectral Correlation Patterns in Wideband Scenarios

Domain knowledge on PUs' spectrum utilization features and correlation patterns can be useful in designing efficient spectrum learning models. To this end, we note that the received wideband PSD $\mathbf{s}$ exhibits both inner-band and inter-band dependencies. The PU signal on each channel exhibits *inner-band* spectrum features that reflects the distinct PSD shape of the adopted modulation scheme. Further, frequency samples across different bands can be correlated as well, giving rise to *inter-band* dependencies.

Fig. 1 illustrates two main sources of inter-band correlations in wideband scenarios. First, practical radio waveforms are not band-limited, but contain PSD sidelobes and even harmonics that spill the signal energy onto adjacent bands (c.f. Fig. 1(a).). Due to such power leakage, PSD samples across adjacent bands are correlated [20]. Second, in practical wireless systems such as LTE-Advanced, non-contiguous channel allocation and carrier aggregation are widely used for enhanced system capacity [21], [27], where a PU may simultaneously transmit waveforms of the same modulation format over multiple non-contiguous bands, forming "aggregated bands" (c.f., Fig. 1(b)). Thus, there are spectral correlations across these non-contiguous channels occupied by the same PU, and the inter-band dependency can even be long range.

As discussed in Section I, existing CNN-based spectrum sensing methods are effective in leveraging the inner-band spectral features [8], [11], but are not efficient in capturing inter-band correlations through convolutional filtering over small receptive fields. The long-range PSD correlations present unique challenges to spectrum sensing in the wideband regime.

C. Basics of Learning-Based Spectrum Occupancy Detection

On a single band, spectrum occupancy detection is a binary hypothesis testing problem with two states representing the band being either occupied (H_1) or vacant (H_0). When a DNN model is adopted for detection, the input layer takes in the PSD samples, say $\mathbf{s}_n \in \mathbb{R}^{N_w}$ on the n-th band, and the output layer only has one neuron that generates the decision statistic $f_n(\mathbf{s}_n; \mathcal{W}_n)$, where $\mathcal{W}_n$ are the DNN model parameters to be trained to reach its optimum $\mathcal{W}_n^*$. Given a prefixed decision threshold λ, the binary detector for the n-th band is given by

$$f_n(\mathbf{s}_n; \mathcal{W}_n^*) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda. \tag{3}$$

Typically, $f_n(\mathbf{s}_n; \mathcal{W}_n) \in [0, 1]$ and $\lambda = 1/2$.

For wideband sensing, a general DNN-based detector can be expressed as $\tilde{\mathbf{y}} = f(\mathbf{s}; \mathcal{W})$, where the input layer has $N_f N_w$ neurons to take in $\mathbf{s} \in \mathbb{R}^{N_f N_w}$, and the output layer has N_f neurons to generate the decision statistic vector $\tilde{\mathbf{y}} \in \mathbb{R}^{N_f}$ for the N_f channels in the spectrum pool. Given a properly chosen loss function $\mathcal{L}(\tilde{\mathbf{y}}, \mathbf{y})$, the DNN model parameters $\mathcal{W}$ can be trained on dataset $\mathcal{D}$ with labeled output $\mathbf{y}$, as follows:

$$\underset{\mathcal{W}}{\operatorname{argmin}} \sum_{\{\mathbf{s},\mathbf{y}\}\in\mathcal{D}} \mathcal{L}(f(\mathbf{s}; \mathcal{W}), \mathbf{y}). \tag{4}$$

It is worth noting that $\mathbf{y} \in \{0, 1\}^{N_f}$, which results in up to $N_p = 2^{N_f}$ spectrum occupancy patterns over the N_f bands. A standard classifer-based DNN design would match each occupancy pattern as one of the 2^{N_f} classes [14], [15], which requires high model complexity and huge dataset for training when N_f is large in the wideband regime. For training efficiency and scalability, it is essential to avoid treating the N_f-band spectrum detector problem naively as a 2^{N_f}-class learning problem with exponential complexity in N_f. Thus, the learning task and loss function need to be designed properly.

III. Spectrum Transformer for Wideband Sensing

To overcome the aforementioned challenges in existing CNN-based wideband sensing methods, this section proposes a novel multi-task Spectrum Transformer architecture. We first introduce the overall design of the proposed Spectrum Transformer. Then, we elaborate on the MSA-based Transformer encoder structure that is a key component of our method.

A. Architecture Design for Wideband Spectrum Transformer

The overall structure of our multi-task Spectrum Transformer is shown in Fig. 2, which is designed to enhance the learning efficiency and scalability for wideband sensing. Overall, wideband PSD data is pre-processed via embedding and position encoding to produce the input to a properly designed Transformer encoder. The encoder output is further processed by a multi-task linear output layer to estimate the multi-band spectrum occupancy at scalable computation.

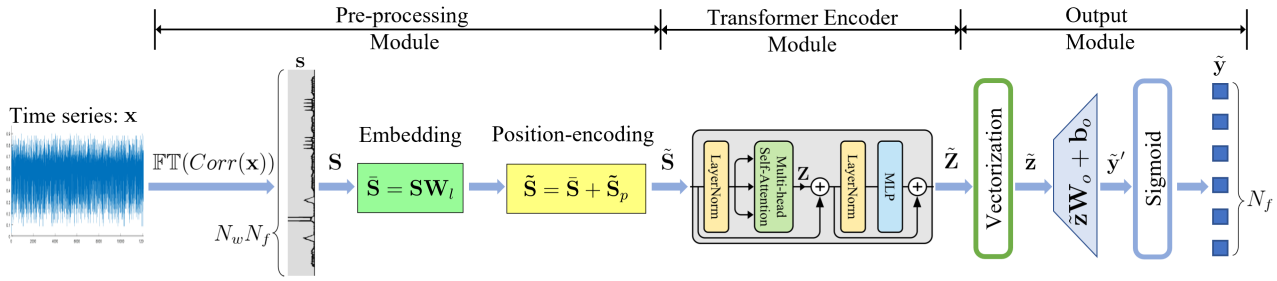


Fig. 2. Structure of Spectrum Transformer for wideband sensing.

1) Wideband PSD Data Embedding Module: The raw data collected by a CR device is the temporal $\mathbf{x}$ in (1), which can be processed to yield the cross-correlation data $Corr(\mathbf{x})$ and the high-resolution PSD data $\mathbf{s}$ in (2). Current Transformers for wireless applications build on the time series $\mathbf{x}$ [23], [24], [25], which are not well suited to capture spectral correlations. Instead, our method explicitly works on the wideband PSD data $\mathbf{s}$, which contain the semantic information of both the inner-band features of modulated signals and the inter-band dependencies due to power leakage and/or cross-band channel aggregation by PUs. This semantic information can be effectively captured by the MSA mechanism, which will be described in Section III-B.

For ease of computation, we rearrange the input PSD vector $\mathbf{s} \in \mathbb{R}^{N_f N_w}$ into a matrix $\mathbf{S} = [\mathbf{s}_1; \ldots; \mathbf{s}_{N_f}] \in \mathbb{R}^{N_f \times N_w}$, where each row of $\mathbf{S}$ is the PSD segment $\mathbf{s}_n$ on a single band. To reduce the model complexity and computation cost, we further apply a learnable linear embeddding matrix $\mathbf{W}_l \in \mathbb{R}^{N_w \times \bar{N}_w}$, with $\bar{N}_w < N_w$, on all the single-band fine-grained PSD segments, so as to compress them into $\bar{\mathbf{S}} \in \mathbb{R}^{N_f \times \bar{N}_w}$:

$$\bar{\mathbf{S}} = \mathbf{S}\mathbf{W}_l \tag{5}$$

The ensuing modules operates on the compressive $\bar{\mathbf{S}}$, which reduces the computation cost and model complexity for MSA.

2) Position Encoding Module: Standard Transformer encoders may not need to identify the positions of segments in the input sequence [22]. However, the positions of input PSD segments are essential in determining the locations of the occupied spectrum in the multi-band spectrum sensing problem of interest. Hence, we employ a learnable position-encoding module, by adding a trainable matrix $\tilde{\mathbf{S}}_p$ to the compressed PSD $\bar{\mathbf{S}}$ in (5):

$$\tilde{\mathbf{S}} = \bar{\mathbf{S}} + \tilde{\mathbf{S}}_p, \tag{6}$$

where $\tilde{\mathbf{S}}_p \in \mathbb{R}^{N_f \times \bar{N}_w}$ denotes the matrix for additive position-encoding. After these pre-processing operations in (5) and (6), the compressive position-cognizant matrix $\tilde{\mathbf{S}}$ is fed into the next MSA-based Transformer Encoder module to extract useful inner-band and inter-band spectral correlation features.

3) Transformer Encoder Module: Based on MSA [22], we customize the design of a Transformer encoder module to fulfill the feature extraction for wideband spectrum occupancy detection. The block diagram of our Transformer encoder is shown in Fig. 2. The MSA block provides the key operations to linearly extract inner-band and inter-band features, which will be discussed in Section III-B. To represent the non-linearity of the spectrum occupancy state as a function of the extracted spectral correlation features, the output of the MSA block needs to be processed by a multi-layer perceptron (MLP) unit [28], which acts as a feed-forward network with one hidden layer plus a Gaussian-error linear unit activation function [29]. Both the MSA and MLP units are processed by a LayerNorm operation, to ensure stability of the model training [30]. Also, residual connections [31] are employed for both MSA and MLP to avoid losing the input information. The output of the Transformer encoder is a matrix $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1; \ldots; \tilde{\mathbf{z}}_{N_f}] \in \mathbb{R}^{N_f \times \bar{N}_w}$ of the same size as that of the input PSD matrix $\tilde{\mathbf{S}}$, while each of the N_f rows of $\tilde{\mathbf{Z}}$ captures rich spectral features for occupancy detection of the corresponding band.

4) Mulit-Task Output Module: The next module makes multi-band spectrum occupancy decisions based on the encoder output $\tilde{\mathbf{Z}}$. As explained in Section II-C, in the wideband regime with large N_f, it is essential to avoid treating the N_f-band spectrum detector problem naively as a 2^{N_f}-class spectrum pattern classification problem. Existing classifier-based multi-band CNN models incur such exponential complexity in N_f [14], [15], because they treat the N_f-band detection problem as a single task, but there are 2^{N_f} possible outcomes of this task to be examined by the detector. To circumvent such high model complexity and computation cost, we design a multi-task output module extended from our previous work [32]. We treat the spectrum detection on each band as a single task, resulting in N_f parallel tasks to be fulfilled by jointly training the common MSA-based transformer encoder.

Specifically, our multi-task output module is the last neural layer of the overall network model. First, the encoder output $\tilde{\mathbf{Z}}$ is vectorized into $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_{N_f}] \in \mathbb{R}^{N_f \bar{N}_w}$. Then, affine transformation is applied on $\tilde{\mathbf{z}}$ with a learnable weight matrix $\mathbf{W}_o \in \mathbb{R}^{N_f \bar{N}_w \times N_f}$ and bias $\mathbf{b}_o \in \mathbb{R}^{N_f}$, yielding

$$\tilde{\mathbf{y}}' = [\tilde{y}'_1, \ldots, \tilde{y}'_{N_f}] = \tilde{\mathbf{z}}\mathbf{W}_o + \mathbf{b}_o, \tag{7}$$

Here, each $\tilde{y}'_n$ is treated as the decision statistic for the binary detection task on the n-th band, $\forall n$. Thus, multi-task learning over N_f bands can be implemented through parallel nonlinear activations on all the entries of $\tilde{\mathbf{y}}'$, using the sigmoid function to map each $\tilde{y}'_n$ into the spectrum occupancy estimate $\tilde{y}_n$:

$$\tilde{y}_n = \text{Sigmoid}(\tilde{y}'_n) \in (0, 1), \quad \text{for } n = 1, \ldots, N_f. \tag{8}$$

Given the labeled output $\mathbf{y} \in \{0,1\}^{N_f}$ in (1) for the true spectrum occupancy states, all the N_f sigmoid estimators are trained simultaneously, along with the common model parameters in preceding modules. To this end, we specify the training objective function in (4) by the total binary cross-entropy (BCE) loss, defined as:

$$\mathcal{L}^{\text{BCE}}(\tilde{\mathbf{y}}, \mathbf{y}) = \sum_{n=1}^{N_f} \tilde{y}_n \log y_n + (1-\tilde{y}_n)\log(1-y_n). \quad (9)$$

In our multi-task design, the number of classes to be distinguished from data-driven learning is $2N_f$ for the N_f binary states, which is greatly reduced from 2^{N_f} in the conventional single-task case [14], [15]. Accordingly, the model parameter size is reduced, which alleviates the sample size requirements and the overfitting issue. Meanwhile, all the tasks share the same Transformer encoder, which allows to effectively extract and utilize the inter-band dependencies in the wideband PSD data for enhanced sensing accuracy across all bands.

B. The MSA Mechanism for Wideband Sensing

The core of our wideband Spectrum Transformer is the adopted MSA mechanism in the encoder module. We now delve into its working principle and explain how it efficiently addresses the unique challenges imposed by long-range PSD correlations in the wideband regime (c.f. Section II-B).

The canonical self-attention mechanism was developed primarily for language understanding tasks [33], in which each input word vector is partitioned into contiguous small segments, and correlations of these segments are captured and weighted to build global dependencies between inputs and outputs [22]. In our formulation for wideband spectrum sensing, an input "word" becomes a normalized version of the position-embedded PSD matrix $\tilde{\mathbf{S}} \in \mathbb{R}^{N_f \times \bar{N}_w}$ in (6). For simplicity, we view each row of $\tilde{\mathbf{S}}$ as a "word segment" of length $\bar{N}_w$, while other segment partitioning strategy is allowable as well. For each segment, the self-attention mechanism employs linear projections to extract three vectors, namely "Query", "Key" and "Value". For row-i $\tilde{\mathbf{s}}_i \in \mathbb{R}^{\bar{N}_w}$, these three vectors are denoted by $\mathbf{q}_i \in \mathbb{R}^{d_q}, \mathbf{k}_i \in \mathbb{R}^{d_k}$, and $\mathbf{v}_i \in \mathbb{R}^{d_v}$, respectively, where d_q, d_k, and d_v are pre-defined dimensionalities. Let the matrices $\mathbf{Q} \in \mathbb{R}^{N_f \times d_q}, \mathbf{K} \in \mathbb{R}^{N_f \times d_k}$, and $\mathbf{V} \in \mathbb{R}^{N_f \times d_v}$ store the "Queries", "Keys" and "Values" of the entire wideband PSD sequence, respectively, and let $\mathbf{W}_q \in \mathbb{R}^{\bar{N}_w \times d_q}, \mathbf{W}_k \in \mathbb{R}^{\bar{N}_w \times d_k}$, and $\mathbf{W}_v \in \mathbb{R}^{\bar{N}_w \times d_v}$ denote the corresponding linear projection matrices, respectively. It holds that

$$\begin{aligned}
\mathbf{Q} &= \tilde{\mathbf{S}}\mathbf{W}_q = [\tilde{\mathbf{s}}_1\mathbf{W}_q; \ldots; \tilde{\mathbf{s}}_{N_f}\mathbf{W}_q] = [\mathbf{q}_1; \ldots; \mathbf{q}_{N_f}], \\
\mathbf{K} &= \tilde{\mathbf{S}}\mathbf{W}_k = [\tilde{\mathbf{s}}_1\mathbf{W}_k; \ldots; \tilde{\mathbf{s}}_{N_f}\mathbf{W}_k] = [\mathbf{k}_1; \ldots; \mathbf{k}_{N_f}], \\
\mathbf{V} &= \tilde{\mathbf{S}}\mathbf{W}_v = [\tilde{\mathbf{s}}_1\mathbf{W}_v; \ldots; \tilde{\mathbf{s}}_{N_f}\mathbf{W}_v] = [\mathbf{v}_1; \ldots; \mathbf{v}_{N_f}]. \quad (10)
\end{aligned}$$

In (10), the matrix $\mathbf{W}_v$ extracts the *inner-band* features of each $\tilde{\mathbf{s}}_i$, which is stored in the form of the vector "Value" $\mathbf{v}_i$. Given the extracted N_f $\mathbf{v}_i$'s as the rows of $\mathbf{V}$, the self-attention process works as a learnable fusion of such inner-band features of all different segments. This process is implemented in the form of a weighted summation of all "Values" by incorporating the "Queries" $\mathbf{Q}$ and "Keys" $\mathbf{K}$ in (10) as well:

$$\begin{aligned}
\mathbf{Z} &= \text{Att}(\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \tilde{\mathbf{S}}) \\
&= \text{Softmax}\left(\frac{\tilde{\mathbf{S}}\mathbf{W}_q\mathbf{W}_k^T\tilde{\mathbf{S}}^T}{\sqrt{d_k}}\right)\tilde{\mathbf{S}}\mathbf{W}_v \\
&= \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \quad (11)
\end{aligned}$$

Now, the output attention matrix $\mathbf{Z} \in \mathbb{R}^{N_f \times d_v}$ contains the learned *inter-band* dependencies in terms of the correlation features across all N_f $\mathbf{v}_i$'s. In (11), the weights are calculated through the product of $\mathbf{Q}\mathbf{K}^T$. Specifically, to compute the attention $\mathbf{z}_i$, "Query" $\mathbf{q}_i$ multiplies with the "Keys" $\mathbf{K}$ corresponding to all bands:

$$\begin{aligned}
a_i^j &= \frac{1}{\sqrt{d_k}}\mathbf{q}_i\mathbf{k}_j^T = \frac{1}{\sqrt{d_k}}\tilde{\mathbf{s}}_i\mathbf{W}_q\mathbf{W}_k^T\tilde{\mathbf{s}}_j^T, \\
\mathbf{a}_i &= [a_i^1, \ldots, a_i^{N_f}] = \frac{1}{\sqrt{d_k}}\mathbf{q}_i\mathbf{K}^T, \quad (12)
\end{aligned}$$

where a_i^j indicates the correlation of $\tilde{\mathbf{s}}_j$ to $\tilde{\mathbf{s}}_i$. In this way, $\mathbf{W}_q$ and $\mathbf{W}_k$ can be trained to extract the co-existing features between highly correlated bands. Afterward, for band-i, a softmax operation is conducted to calculate the weights as:

$$\alpha_i^j = \text{Softmax}(\mathbf{a}_i)_j = \frac{e^{a_i^j}}{\sum_{j'=1}^{N_f} e^{a_i^{j'}}} \in (0,1). \quad (13)$$

The attention vector of band-i is calculated as:

$$\mathbf{z}_i = \sum_{j=1}^{N_f} \alpha_i^j \mathbf{v}_j. \quad (14)$$

Intuitively, the input $\tilde{\mathbf{s}}_j$'s that are more relevant to $\tilde{\mathbf{s}}_i$ lead to larger weights α_i^j's and vice versa. To capture the bidirectional inter-band dependencies, the self-attention mechanism needs to repeat the calculations of $\mathbf{z}_i$'s on all N_f rows of $\tilde{\mathbf{S}}$ and stack them altogether into $\mathbf{Z} \in \mathbb{R}^{N_f \times d_v}$ as the self-attention output given the wideband PSD input.

So far, self-attention is performed through one head, that is, one set of trainable parameter matrices $\{\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v\}$. To improve the performance in extracting diverse inner-band features and handling complicated inter-band dependencies, self-attention is extended to MSA which has multiple "Heads" to boost learning capability [22]. The structure of MSA is illustrated in Fig. 3. By using H sets of distinct linear projection matrices $\{\mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h\}_{h=1}^H$ and getting H parallel output matrices $\{\mathbf{Z}^1, \ldots, \mathbf{Z}^H\}$, MSA is not only capable of extracting the inner-band spectrum features from H subspaces spanned by $\mathbf{W}_v^h$'s, but also good at learning the potential inter-band correlations captured by different sets of $\{\mathbf{W}_q^h, \mathbf{W}_k^h\}$'s. Then, a learnable linear projection $\mathbf{W}_m \in \mathbb{R}^{Hd_v \times \bar{N}_w}$ is applied to merge the H outputs into matrix $\mathbf{Z}$ that contains the combined wideband spectrum features:

$$\mathbf{Z} = [\mathbf{Z}^1, \ldots, \mathbf{Z}^H]\mathbf{W}_m. \quad (15)$$

Constructed based on the MSA mechanism, the Transformer encoder has high capability in capturing inner-band features and inter-band dependencies jointly, through a compact model.

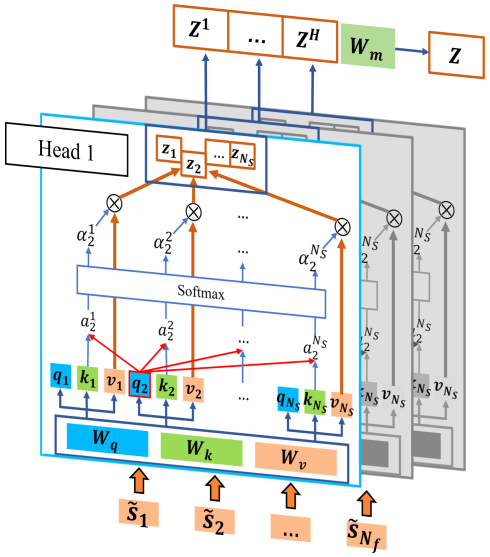


Fig. 3. Structure of MSA.

C. The Complexity of Spectrum Transformer

In this part, we analyze the parameter complexity of our Spectrum Transformer. The trainable parameters of the linear embedding module are saved in $\mathbf{W}_l$, which has $N_w\bar{N}_w$ weight values. The parameter volume of the position encoding module is the same as the embedded wideband PSD $\tilde{S}$, which is $N_f\bar{N}_w$. For the Transformer encoder module, the majority of trainable parameters are saved in the MSA block and the MLP unit. For each "head" of the MSA block, i.e., a single self-attention structure, the trainable parameters are contained in $\{\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v\}$, which makes up $\bar{N}_w(d_q + d_k + d_v)$ weight values. For the H "heads" in the MSA block, the total number of parameters is $H\bar{N}_w(d_q+d_k+d_v)$. The operation to project the output attention vectors of H "heads" into $\mathbf{Z}$, as defined in (15), requires $Hd_v\bar{N}_w$ weight values. The MLP unit in the Transformer encoder is supposed to process each row of $\mathbf{Z}$, which is a single attention vector, into the corresponding row in $\tilde{\mathbf{Z}}$. Parameters of the MLP unit are contained in a hidden layer and an output layer, which has weight values and biases. Assuming that the hidden layer has N_h neurons, the trainable parameter volumes of the hidden layer and the output layer are $(\bar{N}_wN_h + N_h)$ and $(\bar{N}_wN_h + \bar{N}_w)$, respectively. The operation of the multi-task output module is formulated in (7), which involves $(\bar{N}_wN_fN_f + N_f)$ trainable parameters in $\mathbf{W}_o$ and $\mathbf{b}_o$. In total, the Spectrum Transformer contains $\bar{N}_w[N_w+2N_h+N_f+N_f^2+H(d_q+d_k+2d_v)+1]$ parameters. This indicates that, for a specific wideband sensing problem with fixed N_w and N_f, the dimension of input embedding $\bar{N}_w$, the size of the hidden layer in the MLP unit, the number of "heads", and the dimensions of $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$.

IV. Simulation Results

In this section, we testify the proposed Spectrum Transformer for wideband sensing,[1] and compare its performance with DeepSense, a CNN-based solution with demonstrated superiority in the existing spectrum sensing literature [16].

[1]For implementation details, please refer to our online repository available at https://github.com/FrancisZWS/SpectrumTransformerDraft.

A. Data Generation for Wideband Sensing

Considering the scarcity of wideband sensing datasets, we first explain how to generate the synthetic datasets with labels, which shall reflect not only the inner-band features of single-band signals but also the different inter-band dependencies caused by PUs' spectrum occupancy patterns.

Per the signal model in (1), we generate the narrowband transmitted signal $\mathbf{x}_n$ on the n-band by modulating a random message sequence through a predefined modulation scheme used by the PU on this band. It results in the modulation-specific inner-band features of the PSD samples.

To depict the power leakage issue, we suppose that the signal on each band n is leaked into the two adjacent bands only. That is, we compute the PSD of $\mathbf{x}_n$ and retain its mainlobe sample vector $\mathbf{s}'_{n,\text{main}} \in \mathbb{R}^{N_w}$ on the n-th band, and two sidelobe vectors $\mathbf{s}'_{n,\text{left}} \in \mathbb{R}^{N_w}$ and $\mathbf{s}'_{n,\text{right}} \in \mathbb{R}^{N_w}$ that are leaked to the $(n-1)$-th band on the left and the $(n+1)$-th band on the right, respectively. Per (1), the received PSD samples $\mathbf{s}_n$ on the n-th band is given by

$$\begin{aligned}\mathbf{s}_n = {} & y_n|h_n|^2\mathbf{s}'_{n,\text{main}} + y_{n-1}|h_{n-1}|^2\mathbf{s}'_{n-1,\text{right}} \\ & + y_{n+1}|h_{n+1}|^2\mathbf{s}'_{n+1,\text{left}} + \tilde{\mathbf{w}}_n,\end{aligned} \tag{16}$$

where $\tilde{\mathbf{w}}_n \in \mathbb{R}^{N_w}$ represents the noise PSD, and y_n and h_n indicate the ground-truth occupancy condition and channel gain on the n-th band, respectively. The wideband PSD $\mathbf{s}$ is the concatenation of $\{\mathbf{s}_n\}_{n=1}^{N_f}$ over all N_f bands, with inter-band correlations induced by power leakage.

To account for the channel aggregation effect, we let $\mathcal{B}_n$ contain the indices of channels that aggregate with channel-n, $\forall n$. We set $y_{n'} = y_n$ for all $n' \in \mathcal{B}_n$, so that they share the same occupancy state. Long-range inter-band correlations arise when some aggregated channels are far apart in frequency.

B. Simulation Settings

1) CR Environments: Consider a general wideband sensing scenario where a CR user monitors $N_f = 10$ bands. There are $N_w = 64$ frequency points per band, which amounts to a dimension of $N_wN_f = 640$ for the input PSD vector $\mathbf{s}$. For signals transmitted by PUs, a random message sequence is modulated through a predefined modulation scheme used by the PU over allocated bands. Then, the received PSD samples on each band is given by (16). For 10 bands, we generate Boolean vectors $\mathbf{y} = [y_1, y_2, \ldots, y_{10}] \in \mathbb{B}^{10}$ and save them as the library of all ground truth wideband occupancy labels, where each has a matching wideband occupancy pattern.

Our Spectrum Transformer is evaluated in four different cases. Case 1 considers power leakage only, Case 2 features channel aggregation alone, Case 3 considers both power leakage and channel aggregation effects, and Case 4 is a highly dynamic scenario with random channel aggregation patterns. To address the practical issue of limited training data in wireless applications, the dataset of each case in our simulations contains $[6000, 8000]$ training samples, if not particularly mentioned. To reflect the generalization capability

TABLE I
MODULATION SCHEMES

Band index	1, 5	2, 4	3	6, 10	7, 9	8
Modulation	BPSK	MSK	2FSK	16PSK	4FSK	16QAM

TABLE II
CHANNEL AGGREGATION ASSIGNMENT

PU index	1	2	3	4	5	6
Band index	1, 5	2, 4	3	6, 10	7, 9	8
Modulation	BPSK	MSK	2FSK	16PSK	4FSK	16QAM

of our proposed Spectrum Transformer, we run simulations in dynamic scenarios, where the patterns of spectrum occupancy by PUs are different between training and testing stages.

For Case 1, there are $N_{\text{PU}} = 10$ PUs, each of which is assigned to one channel. The modulation scheme for each band is shown in Table I. For each of the total $2^{N_{\text{PU}}} = 1024$ occupancy patterns in this case, we generate 7 wideband PSD samples $\mathbf{s}$ for training and the training dataset volume is 7168.

For Case 2, there are $N_{\text{PU}} = 6$ PUs, and 4 of them use channel aggregation over non-contiguous bands when they occupy their assigned bands. Each PU uses the same modulation scheme over the aggregated bands, as shown in Table II. For each of the $2^{N_{\text{PU}}} = 64$ occupancy patterns, we generate 100 wideband PSDs for training, such that the training dataset volume is 6400.

The settings for Case 3 are similar to that of Case 2. The difference is that the per-band PSD $\{\mathbf{s}_n\}$'s are generated from (16) to account for power leakage. Thus, the datasets for Case 3 reflect the impact of inter-band dependencies caused by both power leakage and channel aggregation.

For Case 4, we keep the same configuration of training data volume and the modulation types of the 6 PUs as in Case 3. Different from Cases 1-3, the training and testing in Case 4 exhibit different patterns of channel aggregation. Specifically, in the training phase, each of the 6 PUs is allocated the same number of bands as in the previous cases. However, these bands are randomly chosen from the spectrum pool. Then, the testing of the trained model is conducted to simulate a more dynamic scenario, where both the channel aggregation pattern and the number of channels utilized by each PU are both randomly varying.

2) Model Configuration: The learnable input embedding module of the proposed Spectrum Transformer compresses the dimension of per-band PSD from $N_w = 64$ to $\bar{N}_w = 16$. The MSA module has $H = 4$ "heads", while the dimensionalities d_q, d_k, and d_v are all set to 8. The hidden layer of the MLP unit has 32 neruons.

The benchmark method of DeepSense CNN has two serial building blocks before the output layer, each of which consists of two convolutional layers followed by one maximal pooling layer [16]. While the original DeepSense model has a predefined model size, we also adjust its structure to implement a compact version called "DeepSense Mini", which has similar model complexity as our Spectrum Transformer to facilitate fair comparison. The detailed structures and model configurations of DeepSense and DeepSense Mini are specified in the format of layer-wise output dimensions, as shown in Table III. Correspondingly, these models entail different model sizes and computational costs, which can be evaluated in terms of the number of trainable parameters and the number of multiply-accumulate operations (MACs) during testing, respectively. As summarized in Table IV, the model size of our Spectrum Transformer is only 11% of that of the full-size DeepSense, while the computation is only about 1.5% of it. Compared with DeepSense Mini, our Transformer has a similar model size, at a reduced computational cost of 77%.

TABLE III
DEEPSENSE CNN ARCHITECTURE

Layer(Activation)	DeepSense (channels, _)	DeepSense Mini (channels, _)
Input	640	640
Conv1(LeakyRelu)	(16, 640)	(2, 640)
Conv2(LeakyRelu)	(16, 640)	(2, 640)
Maxpool1	(16, 320)	(2, 320)
Conv3(LeakyRelu)	(32, 320)	(4, 320)
Conv4(LeakyRelu)	(32, 320)	(4, 320)
MaxPool2	(32, 160)	(4, 160)
FC(LeakyRelu)	10	10
Output(Sigmoid)	10	10

TABLE IV
MODEL COMPARISON: PARAMETER SIZE AND COMPUTATIONAL COST

Complexity	Spectrum Transformer	DeepSense	DeepSense Mini
Parameter	6362	60346	6572
MACs	49312	3379200	64000

3) Model Training Method and Setting: The proposed Spectrum Transformer and the CNN baselines are implemented and evaluated on Pytorch. For the scenarios with sufficient training samples (say more than 100 training samples per occupancy pattern), we select a training batch size of 500. For the scenarios with fewer than 100 training samples per occupancy pattern, we reduce the batch size to the number of samples per occupancy pattern, which is set to be either 20 or 50 depending on the dataset. We train all deep models by using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ and the base learning rate is commonly set to 5×10^{-4}. Given the data insufficiency in the most of our experimental configuration that leads to the notorious overfitting issues, we apply 100 training epochs for all models. For fair comparison among different models, we retain the model parameters that have attained the highest validation accuracy during the training stage. For scenarios with less than 100 training samples per occupancy pattern, we apply a small weight decay of 5×10^{-4} in training for all models. In the Transformer encoder module, we apply a dropout rate of 0.1 to the MLP unit and the embedding unit.

C. Simulation Results and Discussions

We conduct simulations for various system settings to evaluate the spectrum sensing performance of the Spectrum Transformer with reference to the benchmark models in wideband environments. Performance metrics include the average probability of detection (PD) under a fixed average probability of false alarm (PFA) of 5%, and the receiver operating

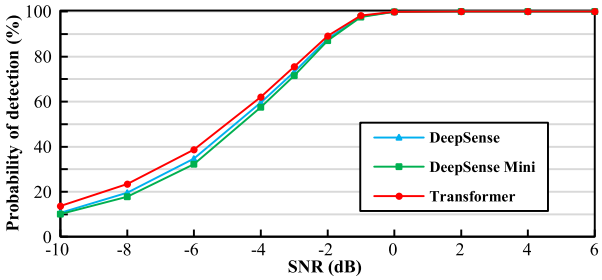


Fig. 4. PD of different models in Case 1.

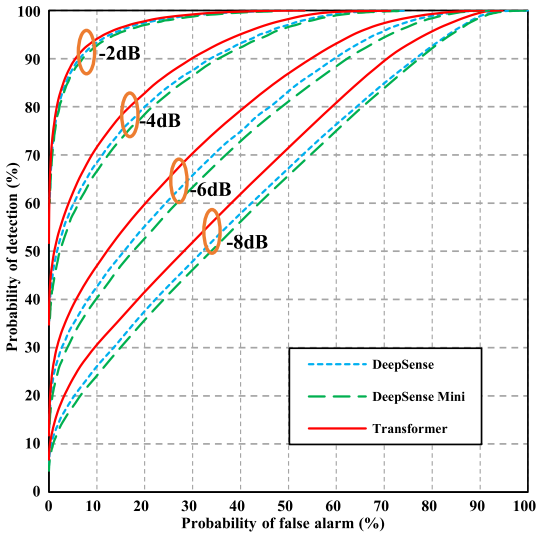


Fig. 5. ROC of different models in Case 1.

characteristic (ROC) in terms of PD versus PFA for various signal to noise ratio (SNRs) over the entire wide bandwidth.

1) Case 1. Power Leakage: When inter-band correlation arises among adjacent bands due to power leakage alone, the sensing accuracy in terms of PD is tested for a range of SNRs between $[-10, 6]$ dB. As shown in Fig. 4, our Spectrum Transformer outperforms the other two CNN-based solutions at any SNR values. Meanwhile, the full-size DeepSense offers better sensing accuracy than DeepSense Mini. This demonstrates that improvement in the learning capability of CNN-based models comes at the cost of increased model complexity. The superiority of the Spectrum Transformer over the others is also corroborated by the comparative ROC performances depicted in Fig. 5 for various SNRs.

2) Case 2. Channel Aggregation: Channel aggregation, as shown in Table II, gives rise to long-range inter-band correlations, which is evaluated in Case 2. Fig. 6 depicts the PD performance, and Fig. 7 shows the ROC curves for various SNR values. The proposed Spectrum Transformer achieves the best performances among the three methods.

3) Case 3. Both Power Leakage and Channel Aggregation: For this case of both inter-band correlation patterns, the sensing accuracy of our Spectrum Transformer consistently outperforms the other two CNN-based methods as well, as shown in the comparative performance of PD in Fig. 8 and that of ROC in Fig. 9.

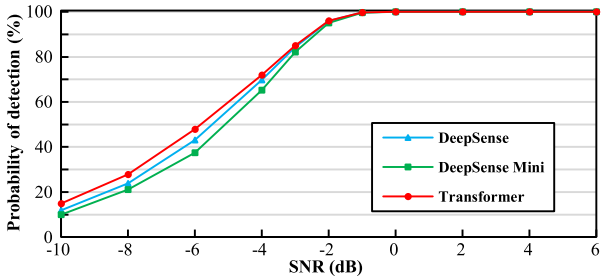


Fig. 6. PD of different models in Case 2.

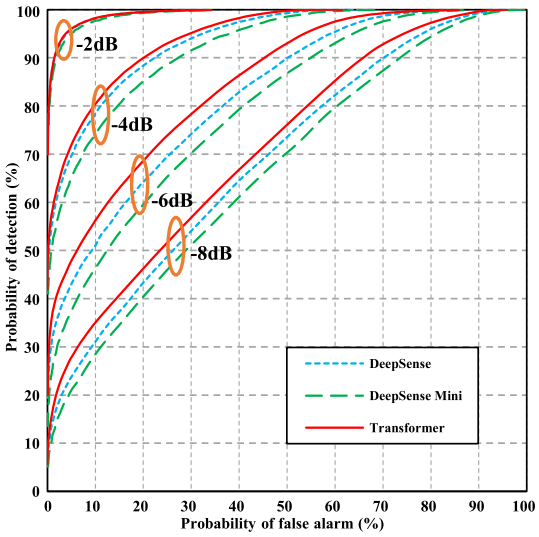


Fig. 7. ROC of different models in Case 2.

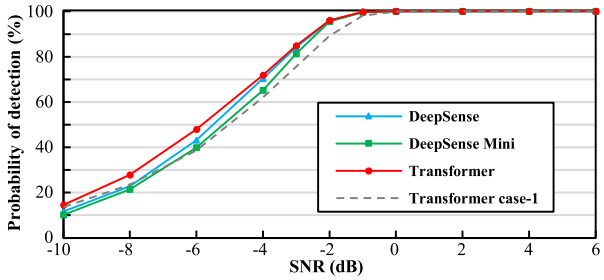


Fig. 8. PD of different models in Case 3.

Our learning model design has been motivated by the domain knowledge that long-range spectral correlations are present in the wideband regime. To illustrate the importance of this knowledge, we train the Spectrum Transformer model on a dataset without channel aggregation and then test its sensing performance in Case 3, which is labeled as "Transformer Case 1" in Fig. 8. Apparently, this trained model does not utilize the long-range correlation features, and thus yields worse sensing performance than all the other models trained in Case 3. Such performance gap indicates the usefulness of the domain knowledge in improving the accuracy of wideband spectrum occupancy detection, which is efficiently utilized by our Spectrum Transformer model.

4) Case 4. Dynamic Channel Aggregation: In this case, all deep models are tested under a dynamic channel aggregation pattern in testing that is different from training. The

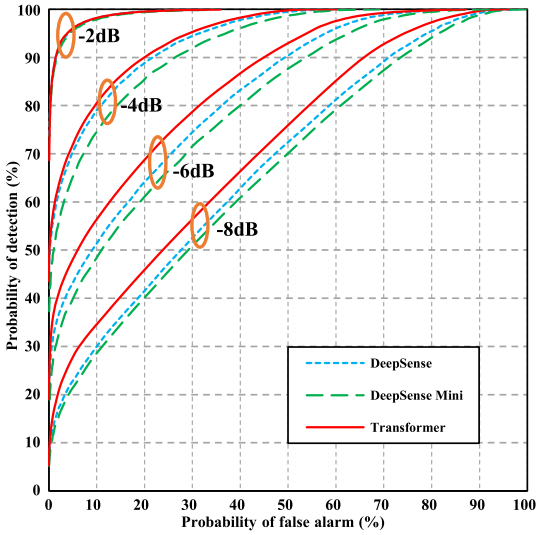


Fig. 9. ROC of different models in Case 3.

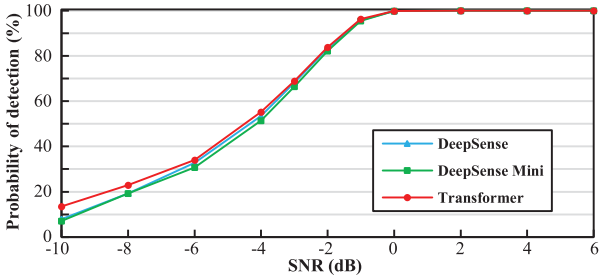


Fig. 10. PD of different models in Case 4.

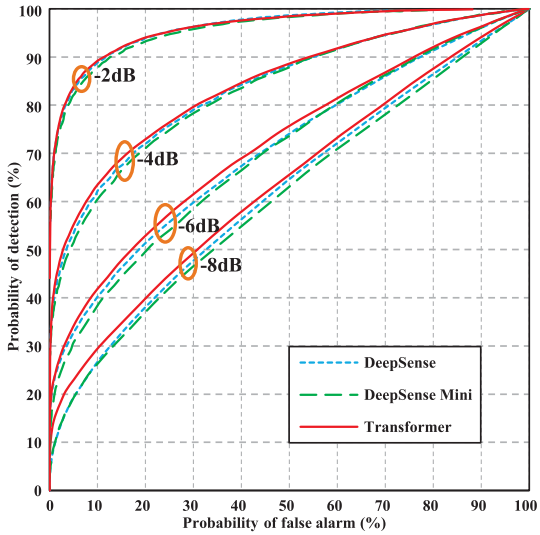


Fig. 11. ROC of different models in Case 4.

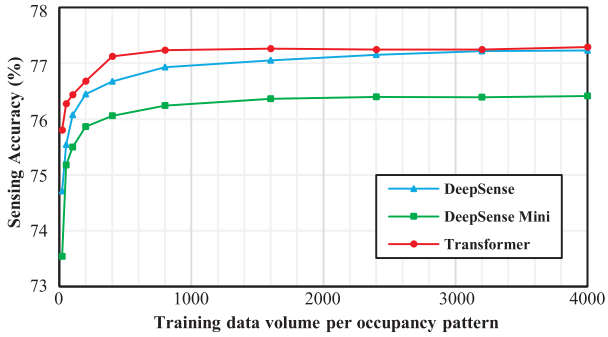


Fig. 12. Sensing accuracy of different models in Case 4 with SNR=-4dB.

performance of different models are tested and compared in terms of PD and ROC, as shown in Fig 10 and Fig 11, respectively. We can see that our Spectrum Transformer still outperforms other CNN-based methods for any given SNR. These results demonstrate that our Spectrum Transformer works effectively when applied to a diverse wireless scenario that is different from the one used for training. Meanwhile, we also observe that the performance is degraded a little compared with Case 3, and the gap between our method and the CNN models becomes smaller. This motivates us to further enhance the transferability of our Spectrum Transformer in future work.

In all cases above, the proposed Spectrum Transformer delivers superior sensing performance, because its built-in self-attention mechanism can efficiently capture inter-band correlation features using a compact model given small data. This capability is particularly beneficial to practical CR systems, which require short sensing time over limited training samples in order to respond to dynamic operating environments.

5) Impact of Training Data Volume: While previous test cases examine the small data scenario with 6400 or 7168 wideband training data samples, we now evaluate these learning models over a wide range of training data volumes in Case 3, for a fixed SNR$=-4$dB. As we mentioned, there are 6 PUs over the 10 channels, resulting in $2^6 = 64$ channel occupancy patterns. The data volume is measured by the number of training samples per occupancy pattern, which is evaluated at $\{20, 50, 100, 200, 400, 800, 1600, 2400, 3200, 4000\}$. The adopted metric for sensing accuracy is $\frac{\text{PD}+(1-\text{PFA})}{2}$, which is an average of the PD and PFA performances.

As shown in Fig. 12, the Spectrum Transformer achieves the highest sensing accuracy under various data volumes, and finally reaches 77.3% for large datasets. When the data volume is small with 20 per occupancy pattern, our method is 1.1% more accurate than the full-size DeepSense CNN and 4% more accurate than DeepSense Mini. As the data volume reduces from 4000 to 100, the accuracy of our method only drops by less than 0.86%, while that of DeepSense and DeepSense Mini degrades by 1.2% and 0.9%, respectively. The sensing performance by the full-size DeepSense model becomes better as the data volume increases, but saturates at a small performance gap from ours when the dataset is large. The compact DeepSense Mini, on the other hand, reaches its steady-state sensing performance much faster at lower data volume than DeepSense, but the accuracy is not as good. These phenomena reflects the inefficient bias-variance tradeoff of CNN-based methods in the wideband regime. Essentially, DeepSense adopts a large model size with deep layers for small bias, but needs a large data volume to reduce the variance caused by overfitting. The compact CNN, on the other hand, circumvents the overfitting issue at small data volume, but does not appear to have adequate model representation

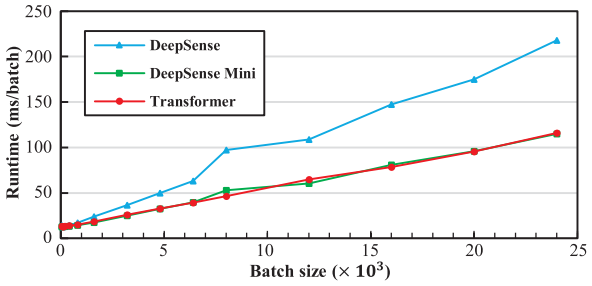


Fig. 13. Runtime of different models per batch for variant batch size.

power, causing large bias. These experimental results, along with the comparison of model complexity and computation costs in Table IV, corroborate that our Spectrum Transformer, can efficiently capture various spectral correlation patterns through a moderate model size, by virtue of its self-attention mechanism. Hence it offers an appealing model architecture with efficient bias-variance tradeoff for the wideband spectrum sensing problem of interest.

6) Runtime Evaluation: We also evaluate the computation complexity of different methods by comparing their execution time on our desktop computer with a Intel Core i9-12900KF CPU, Nvidia RTX3090 GPU, and 128GB-RAM. Specifically, we count the runtime per iteration to train each model on a data batch whose volume ranges from 20 to 24000. As shown in Fig. 13, under most batch sizes, the time consumption to train our Spectrum Transformer is similar to that spent for DeepSense Mini, which is nearly half of that to train DeepSense. These results align with the model size comparison presented in Table IV. Therefore, our proposed model enjoys the low complexity, while achieving the desired sensing accuracy as has been verified by the other figures of PD and ROC curves.

V. Conclusion

This paper develops a novel learning model for wideband sensing, called Spectrum Transformer, which is empowered by the multi-head self-attention structure. The Spectrum Transformer captures the pairwise correlations of all spectrum segments in a parallel manner, which allows it to effectively leverage both the inner-band features and long-range inter-band dependencies for high-resolution spectrum occupancy detection in the wideband regime. Extensive simulations verify that the Spectrum Transformer model outperforms the CNN architecture in the wideband regime, in terms of model capacity, computation efficiency and resistance to over-fitting. It is particularly attractive for wideband spectrum sensing with small-volume datasets.

References

[1] W. Zhang, Y. Wang, X. Chen, and Z. Tian, "Spectrum transformer: Wideband spectrum sensing using multi-head self-attention," in *Proc. IEEE 24th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2023, pp. 101–105.

[2] Y. Wang, Z. Tian, and C. Feng, "Sparsity order estimation and its application in compressive spectrum sensing for cognitive radios," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2116–2125, Jun. 2012.

[3] Y. Wang, Z. Tian, and C. Feng, "Collecting detection diversity and complexity gains in cooperative spectrum sensing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2876–2883, Aug. 2012.

[4] C. Jiang, Y. Li, W. Bai, Y. Yang, and J. Hu, "Statistical matched filter based robust spectrum sensing in noise uncertainty environment," in *Proc. IEEE 14th Int. Conf. Commun. Technol.*, Nov. 2012, pp. 1209–1213.

[5] J. Lunden, V. Koivunen, A. Huttunen, and H. V. Poor, "Spectrum sensing in cognitive radios based on multiple cyclic frequencies," in *Proc. 2nd Int. Conf. Cognit. Radio Oriented Wireless Netw. Commun.*, Aug. 2007, pp. 37–43.

[6] Z. Tian, Y. Tafesse, and B. M. Sadler, "Cyclic feature detection with sub-Nyquist sampling for wideband spectrum sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 1, pp. 58–69, Feb. 2012.

[7] 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, Jan. 2019.

[8] W. M. Lees, A. Wunderlich, P. J. Jeavons, P. D. Hale, and M. R. Souryal, "Deep learning classification of 3.5-GHz band spectrograms with applications to spectrum sensing," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 2, pp. 224–236, Jun. 2019.

[9] P. Shachi, K. R. Sudhindra, and M. N. Suma, "Convolutional neural network for cooperative spectrum sensing with spatio-temporal dataset," in *Proc. Int. Conf. Artif. Intell. Signal Process. (AISP)*, Jan. 2020, pp. 1–5.

[10] W. Lee, M. Kim, and D.-H. Cho, "Deep cooperative sensing: Cooperative spectrum sensing based on convolutional neural networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3005–3009, Mar. 2019.

[11] J. Gao, X. Yi, C. Zhong, X. Chen, and Z. Zhang, "Deep learning for spectrum sensing," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1727–1730, Dec. 2019.

[12] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Proc. Int. Conf. Eng. Appl. Neural Netw.* Cham, Switzerland: Springer, 2016, pp. 213–226.

[13] S. Zheng, S. Chen, P. Qi, H. Zhou, and X. Yang, "Spectrum sensing based on deep learning classification for cognitive radios," *China Commun.*, vol. 17, no. 2, pp. 138–148, Feb. 2020.

[14] N. Ambika, K. Muthumeenakshi, and S. Radha, "Classification of primary user occupancy using deep learning technique in cognitive radio," in *Advances in Automation, Signal Processing, Instrumentation, and Control*. Cham, Switzerland: Springer, 2021, pp. 1795–1804.

[15] Y. Zhang, B. Shen, J. Wang, and T. Yan, "CNN based wideband spectrum occupancy status identification for cognitive radios," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2020, pp. 569–574.

[16] D. Uvaydov, S. D'Oro, F. Restuccia, and T. Melodia, "DeepSense: Fast wideband spectrum sensing through real-time in-the-loop deep learning," in *Proc. IEEE Conf. Comput. Commun.*, May 2021, pp. 1–10.

[17] A. Vagollari, V. Schram, W. Wicke, M. Hirschbeck, and W. Gerstacker, "Joint detection and classification of RF signals using deep learning," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–7.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[19] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–12.

[20] T.-H. Yu, S. Rodriguez-Parera, D. Markovic, and D. Cabric, "Cognitive radio wideband spectrum sensing using multitap windowing and power detection with threshold adaptation," in *Proc. IEEE Int. Conf. Commun.*, May 2010, pp. 1–6.

[21] Nidhi, A. Mihovska, and R. Prasad, "Overview of 5G new radio and carrier aggregation: 5G and beyond networks," in *Proc. 23rd Int. Symp. Wireless Pers. Multimedia Commun. (WPMC)*, Oct. 2020, pp. 1–6.

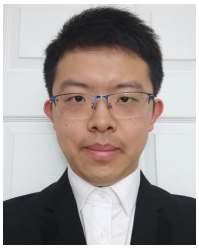
[22] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[23] Z. Chen, F. Gu, and R. Jiang, "Channel estimation method based on transformer in high dynamic environment," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2020, pp. 817–822.

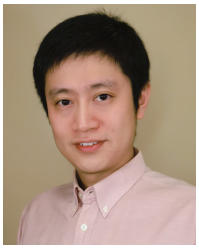
[24] S. Wang, S. Bi, and Y. A. Zhang, "Deep reinforcement learning with communication transformer for adaptive live streaming in wireless edge networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 308–322, Jan. 2022.

[25] W. Kong, Q. Yang, X. Jiao, Y. Niu, and G. Ji, "A transformer-based CTDNN structure for automatic modulation recognition," in *Proc. 7th Int. Conf. Comput. Commun.*, Chengdu, China, 2021, pp. 159–163.

[26] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[27] S. H. R. Bukhari, M. H. Rehmani, and S. Siraj, "A survey of channel bonding for wireless networks and guidelines of channel bonding for futuristic cognitive radio sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 924–948, 2nd Quart., 2016.
[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[29] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[30] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[32] W. Zhang, Y. Wang, F. Yu, Z. Qin, X. Chen, and Z. Tian, "Wideband spectrum sensing based on collaborative multi-task learning," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2022, pp. 01–06.
[33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–8.

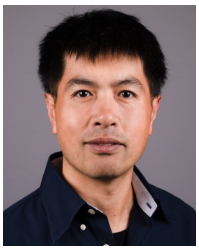


Weishan Zhang (Student Member, IEEE) received the B.S. degree in electronic science and technology from the University of Science and Technology of China, Hefei, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, George Mason University. His current research interests include deep learning for wideband spectrum sensing.



Yue Wang (Senior Member, IEEE) received the Ph.D. degree in communication and information system from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, in 2011. He has been an Assistant Professor with the Department of Computer Science, Georgia State University, Atlanta, GA, USA, since August 2023. Prior to that, he was a Research Assistant Professor with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA. His research interests include the interdisciplinary areas of machine learning, signal processing, wireless communications, and their applications in cyber physical systems. His specific research focuses on distributed optimization and machine learning, sparse signal processing, massive MIMO, mmWave communications, cognitive radios, spectrum sensing, the Internet of Things, direction of arrival estimation, and high-dimensional data analysis.

Xiang Chen (Member, IEEE) received the Ph.D. degree in computer engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 2016. After that, he joined George Mason University, Fairfax, VA, USA, and founded the Intelligence Fusion Laboratory. With close collaboration with industrial labs and vast universities, he is currently leading multiple research projects funded by NSF and Air Force Research Laboratory (AFRL). His research interests include high-performance computing, artificial intelligence, large-scale systems, and various mobile and edge applications. He received the NSF CAREER Award, the Best Paper Award in DATE, and the several other competition awards.



Zhipeng Cai (Fellow, IEEE) is currently a Professor with the Department of Computer Science and the Director of the INSPIRE Center, Georgia State University. His research expertise lies in the areas of resource management and scheduling, high performance computing, cyber-security, privacy, networking, and big data. He was a recipient of the NSF CAREER Award. He is the Editor-in-Chief of *Wireless Communications and Mobile Computing*, the Associate Editor-in-Chief of *High-Confidence Computing* (Elsevier), and an Editor of various prestigious journals, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, and IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS.

Zhi Tian (Fellow, IEEE) has been a Professor with the ECE Department, George Mason University, since 2015. Prior to that, she was a Faculty Member with Michigan Technological University, from 2000 to 2014. She was the Program Director of U.S. National Science Foundation from 2012 to 2014. Her general interests are in the areas of signal processing, communications, detection, and estimation. Her current research interests include decentralized optimization and learning over networks, statistical inference from distributed data, compressed sensing for random processes, cognitive radios, and millimeter-wave MIMO communications. She received the IEEE Communications Society TCCN Publication Award in 2018. She served on the Board of Governors for the IEEE Signal Processing Society from 2019 to 2021. She was the Chair of the IEEE Signal Processing Society Big Data Special Interest Group and a member of the IEEE Signal Processing for Communications and Networking Technical Committee. She was a Distinguished Lecturer of both the IEEE Communications Society and the IEEE Vehicular Technology Society. She is the Editor-in-Chief of IEEE TRANSACTIONS ON SIGNAL PROCESSING and served as an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE TRANSACTIONS ON SIGNAL PROCESSING.