Latent Diffusion for Spectrum Sensing of Coexisting Radar and Communication Signals

Thien Huynh-The, *Senior Member, IEEE*, Phuoc-Long Huynh, Van-Ca Phan, Thai-Hoc Vu, *Member, IEEE*, and Daniel Benevides da Costa, *Senior Member, IEEE*

***Abstract*—The growing demand for spectrum efficiency in next-generation wireless networks, especially in vehicular environments, necessitates effective spectrum sensing (SS) techniques capable of managing the coexistence of technologies like fifth generation new radio (NR) and radar systems. This letter introduces SpecDiff, an innovative framework based on latent diffusion models for spectrogram segmentation, designed to identify and differentiate these coexisting signals in dynamic, noisy environments. SpecDiff leverages a generative diffusion model in a compact latent space, using an attention-based denoising process to enhance segmentation performance under low signal-to-noise ratios and complex channel conditions. The model achieves state-of-the-art performance, with a mean accuracy of 98.68% and mean intersection-over-union (IoU) of 96.30%, effectively identifying the occupied bandwidth in spectrograms. Furthermore, SpecDiff surpasses existing deep learning models in both accuracy and efficiency, offering a promising solution for spectrum sharing in future wireless networks.**

***Index Terms*—5G NR, deep learning, diffusion models, signal identification, spectrogram segmentation, spectrum sensing.**

I. Introduction

THE EVOLUTION towards fifth-generation (5G) and beyond 5G networks is intensifying spectral congestion, particularly in vehicular environments where vehicle-to-everything communications and automotive radar must coexist within shared spectrum [1]. Connected and autonomous vehicles generate vast amounts of data for safety, navigation, and infotainment, while relying on high-resolution radar for situational awareness. This convergence of technologies drives the need for cellular standards such as long-term evolution (LTE), 5G New Radio (5G NR), and emerging vehicular communication protocols to operate in frequency bands traditionally reserved for primary radar systems [2]. As a result, intelligent spectrum sensing (SS) has become essential for accurately distinguishing diverse signals, enabling efficient spectrum sharing in these dynamic environments. To address these challenges, deep learning (DL) techniques are increasingly being applied to spectrum sensing, offering the potential for adaptive and efficient spectrum management in next-generation wireless networks.

The transition from simple energy detection to fine-grained spectral characterization has exposed key limitations of SS, particularly for reliable primary user (PU) detection under low signal-to-noise ratio (SNR) and dynamic conditions. Recent advancements have moved beyond handcrafted feature-based approaches by leveraging multi-feature fusion with adaptive thresholding for robust PU detection [3], integrating DL with extreme value theory to detect PU emulation attacks without prior attacker knowledge [4], and employing semi-supervised learning to address the scarcity of labeled data and improve SS performance under low-SNR conditions [5]. Additionally, several advanced methods have been introduced by exploiting innovative DL techniques to further enhance SS performance. A hierarchical multi-agent deep reinforcement learning framework was introduced to optimize both sensing and power allocation, using a two-tier deep Q-network and multi-agent deep deterministic policy gradient architecture to treat them as correlated tasks [6]. For wideband spectrum sensing, a spectrum transformer utilizing multi-head self-attention was developed to efficiently capture both local and long-range spectral correlations, overcoming the limitations of traditional convolutional networks [7]. An end-to-end joint denoising and spectrum sensing network employed a U-Net architecture with self-attention to simultaneously enhance signal quality and perform detection [8]. Furthermore, graph convolutional networks were leveraged for cooperative sensing, modeling relationships between users as a graph to adapt to dynamic network topologies and resolve the hidden node problem [9]. However, these SS methods, which focus primarily on PU detection (spectrum hole identification), have yet to address the identification of specific wireless signals like 5G NR and LTE, nor have they resolved the challenges of radar-communication coexistence as issues that are essential for the effective operation of next-generation wireless networks.

Recent advancements in SS have leveraged DL-based semantic segmentation to interpret wideband spectrograms, enabling the identification of the specific time-frequency characteristics that define spectrum occupancy. For instance, DeepLabV3+ [10] explores an atrous spatial pyramid pooling with parallel dilated convolutions to capture multi-scale relevant information, enhancing the localization of coexisting 5G NR and LTE signals in spectrograms under realistic channel conditions. In [11], PRMNet, a resolution-preserving deep network for spectrogram analysis at multiple resolutions, incorporates multi-scale feature extraction modules

Received 15 November 2025; accepted 17 December 2025. Date of publication 22 December 2025; date of current version 31 December 2025. This work was supported by Ho Chi Minh City University of Technology and Education (HCMUTE) under Grant T2025-181. The associate editor coordinating the review of this article and approving it for publication was B. Makki. *(Corresponding author: Thien Huynh-The.)*

Thien Huynh-The, Phuoc-Long Huynh, and Van-Ca Phan are with the Department of Computer and Communications Engineering, Ho Chi Minh City University of Technology and Education, Ho Chi Minh City 71307, Vietnam (e-mail: thienht@hcmute.edu.vn; hphuoclong24@gmail.com; capv@hcmute.edu.vn).

Thai-Hoc Vu is with the Faculty of Electrical Engineering and Computer Science, VSB-Techincal University of Ostrava, 708 00 Ostrava, Czechia (e-mail: thai.hoc.vu@vsb.cz).

Daniel Benevides da Costa is with the Interdisciplinary Research Center for Communication Systems and Sensing (IRC-CSS), Department of Electrical Engineering, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia (e-mail: danielbcosta@ieee.org).

Digital Object Identifier 10.1109/LWC.2025.3646878

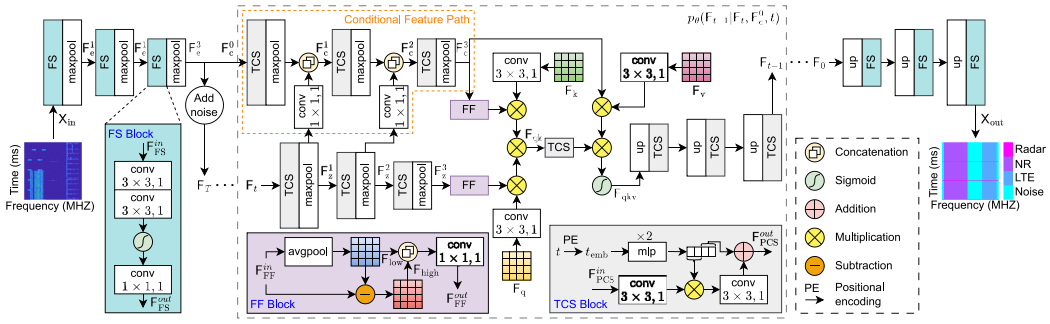


Fig. 1. The SpecDiff framework: an encoder–diffusion–decoder design where diffusion with conditional features and frequency filtering refines spectrogram masks for accurate radar, 5G NR, and LTE signal identification.

to effectively segment complex advanced wireless signals. More recently, with an innovative dual-path architecture, DPSegNetn [12] combines context and spatial paths to capture global spectral information and fine-grained spatial details, thereby enhancing segmentation accuracy and efficiency through a bi-level feature fusion mechanism. However, these architectures often suffer from performance degradation in dynamic low-SNR environments with heterogeneous radar-communication signals, while their inherent complexity exposes a challenging trade-off between segmentation robustness and computational efficiency.

To address the limitations of conventional SS in complex radar–communication coexistence scenarios, we propose SpecDiff, a novel and efficient latent diffusion framework for spectrogram segmentation. SpecDiff performs an iterative denoising process in a compact latent space, guided by an attention-based mechanism to achieve robust identification of heterogeneous signals under challenging channel conditions like low SNR, Doppler, and multipath impairments. The main contributions of this letter are as follows:

- We introduce SpecDiff, a novel framework that adapts a generative diffusion model for the discriminative task of spectrogram segmentation. Its core novelty lies in manipulating an iterative refinement process within a latent space to reconstruct high-fidelity spectral masks, enabling robust signal identification even from heavily corrupted and noisy inputs.
- We design a specialized noise prediction network that serves as the core of SpecDiff. This network integrates three innovative components: a time-conditional squeeze (TCS) block for adapting to varying noise levels across diffusion steps, a conditional path (CP) to provide semantic guidance from spectrogram features, and a frequency filter (FF) to effectively disentangle stable signal structures from noise.
- We demonstrate through extensive evaluation that SpecDiff establishes a new state-of-the-art (SOTA) benchmark, achieving 98.68% mean accuracy and 96.30% mean IoU. It consistently outperforms both general-purpose semantic segmentation models (e.g., HRNet [13], UPerNet [14], SegFormer [15]) and recent SS-driven architectures (e.g., DeepLabV3+ [10], RPMNet [11], DPSegNet [12]), improving upon the next-best competitor by 2.44% in accuracy with a model that is 8× more parameter-efficient.

II. Methodology

A. Signal Model and Spectrogram Representation

In shared spectrum environments, the received discrete-time signal $y[n]$, affected by time-varying channel impairments, is modeled as

$$y[n] = \sum_{\ell} h[n,\ell]\, x[n-\ell] + v[n], \tag{1}$$

where $x[n]$ is the transmitted waveform (e.g., 5G NR, LTE, radar), $h[n,\ell]$ is the time-varying channel impulse response capturing multipath fading and Doppler effects, and $v[n]$ denotes additive white Gaussian noise (AWGN). To reveal time–frequency structure for signal characteristics identification, we apply the short-time Fourier transform (STFT) to $y[n]$:

$$Y(m,k) = \sum_{n=-\infty}^{\infty} y[n]\, w[n-mH]\, e^{-j\frac{2\pi kn}{N}}, \tag{2}$$

where $w[\,\cdot\,]$ is the analysis window, H is the hop size, N is the DFT length, and $\mathrm{k} \in \{0,\ldots,N-1\}$. From these complex STFT coefficients, we construct the spectrogram as

$$\mathrm{Spectrogram}(m,k) = |\,Y(m,k)|^2, \tag{3}$$

which captures temporal and spectral dynamics, yielding distinct visual patterns for different signal types. Consequently, a DL network trained for semantic segmentation can analyze these spectrograms to identify the exact time-frequency boundaries of each signal and thus determine its occupied bandwidth.

B. SpecDiff: Diffusion for Spectrogram Segmentation

As illustrated in Fig. 1, SpecDiff formulates spectrogram segmentation as a conditional generative task within a compact latent space. The model's operation is divided into two primary stages. First, a convolutional encoder-decoder handles the latent space mapping and mask reconstruction. The encoder distills the input spectrogram into a multi-scale feature representation that serves as a semantic guide. Symmetrically, the decoder's role is to upscale the final latent representation into a full-resolution segmentation mask. Second, the core segmentation is performed by the noise prediction network, which operates entirely within this latent space. Conditioned on the features from the encoder, this network iteratively refines a random noise variable into a clean representation of the segmentation mask, which is then passed to the decoder for final reconstruction.

1) Latent Space Mapping With Mask Reconstruction: Spectrograms from congested environments are often noisy and contain complex, multi-scale structures. Performing generative tasks directly on such inputs is inefficient. SpecDiff addresses this by first using an encoder to compress the spectrogram into a more compact and manageable latent representation, thus simplifying the subsequent generative process. Given the output $\mathbf{F}_{\mathrm{FS}}^{out}$, the feature selection (FS) module (as shown in Fig. 1) extracts features as follows:

$$\mathbf{F}_{\mathrm{FS}}^{out} = \mathcal{C}_{1\times1}^{1}\Big(\sigma\Big(\mathcal{C}_{3\times3}^{1}\Big(\mathcal{C}_{3\times3}^{1}\Big(\mathbf{F}_{\mathrm{FS}}^{in}\Big)\Big)\Big)\Big), \tag{4}$$

where $\mathbf{F}_{\mathrm{FS}}^{in}$ is the input feature map, $\mathcal{C}_{k\times m}^{s}(\cdot)$ denotes a standard convolution with kernel size $k \times m$ and stride s, which is immediately followed by batch normalization and a rectified linear unit (ReLU) activation function. The term $\sigma(\cdot)$ is the sigmoid activation function, which acts as a self-attention gate to modulate the feature map, thereby allowing the module to selectively emphasize salient features and suppress noise. The i-th stage of the encoder is defined as:

$$\mathbf{F}_{\mathrm{e}}^{i} = \begin{cases} \mathcal{M}_{2\times2}^{2}(\mathrm{FS}(\mathbf{X}_{\mathrm{in}})) & \text{if } i = 1, \\ \mathcal{M}_{2\times2}^{2}\Big(\mathrm{FS}\Big(\mathbf{F}_{\mathrm{e}}^{i-1}\Big)\Big) & \text{if } i \in \{2,3\}, \end{cases} \tag{5}$$

where $\mathcal{M}_{2\times2}^{2}(\cdot)$ is a 2×2 max pooling operation with a stride of 2, and $\mathbf{X}_{\mathrm{in}} \in \mathbb{R}^{3\times H\times W}$ is the input color spectrogram of height H and width W; $\mathbf{F}_{\mathrm{e}}^{i} \in \mathbb{R}^{C_i\times H_i\times W_i}$ represents the latent feature map at stage i, where H_i and W_i denote the height and width of the feature map, which change due to downsampling. C_i is the number of output channels at stage i, where $C_i \in \{32, 48, 96\}$. The decoder symmetrically reconstructs the high-resolution segmentation mask from the denoised latent representation $\mathbf{F}_0$ by progressively applying $2\times$ upsampling and FS modules at each stage, matching the encoder's channel dimensions at the corresponding stage and ending with a convolution to produce the output logits $\mathbf{X}_{\mathrm{out}}$.

2) Noise Prediction Network: The noise prediction network is the central component of SpecDiff, responsible for estimating the noise $\epsilon_\theta(\mathbf{F}_t, t, \mathbf{F}_{\mathrm{c}}^{0})$ within the noisy latent mask $\mathbf{F}_t$, conditioned on features $\mathbf{F}_{\mathrm{c}}^{0}$ and timestep t. Its multi-scale, symmetrical architecture captures global contextual information and fine-grained details, while an attention-like mechanism enables it to focus on the most salient features for robust prediction. These capabilities are realized through three key mechanisms: the TCS block forms the multi-scale backbone; the CP module provides semantic guidance; and the FF module enhances features for the attention mechanism.

Time-conditional squeeze: The TCS block conditions the network on the current noise level by converting the diffusion timestep t into a time embedding vector t_{emb} of size 128. This vector parameterizes a scale-and-shift operation that modulates the feature maps. While increasing the size of t_{emb} generally improves accuracy by enabling finer-grained feature adaptation, it also increases model complexity. The operation of a TCS block is defined as:

$$\mathbf{F}_{\mathrm{TCS}}^{out} = \mathrm{TCS}\Big(\mathbf{F}_{\mathrm{TCS}}^{in}, t\Big) = \mathcal{C}_{3\times3}^{1}(\mathbf{H} \odot \mathrm{Scale}(t)) + \mathrm{Shift}(t), \tag{6}$$

with

$$\begin{aligned} \mathbf{M}_{\mathrm{mlp}}^{out} &= \mathrm{MLP}_2(\mathrm{MLP}_1(\mathrm{PE}(t))), \\ \mathrm{Scale}(t) &= \mathbf{M}_{\mathrm{mlp}}^{out}[:, 0{:}C_{\mathrm{M}}/2], \\ \mathrm{Shift}(t) &= \mathbf{M}_{\mathrm{mlp}}^{out}[:, C_{\mathrm{M}}/2{:}C_{\mathrm{M}}], \end{aligned}$$

where C_{M} is the output channel dimension of $\mathbf{M}_{\mathrm{mlp}}^{out}$; and $\mathbf{H} = \mathcal{C}_{3\times3}^{1}(\mathbf{F}_{\mathrm{TCS}}^{in})$. $\mathrm{PE}(t)$ represents the sinusoidal positional embedding of timestep t; MLP_1 and MLP_2 are multi-layer perceptrons, and $\odot$ denotes element-wise multiplication.

The downsampling path of the noise prediction architecture is designed to process the noisy latent mask $\mathbf{F}_t$. During training, this mask is obtained via the forward diffusion process by adding a controlled amount of Gaussian noise ϵ to the clean latent mask $\mathbf{F}_0$ according to a random timestep t:

$$\mathbf{F}_t = \sqrt{\bar{\alpha}_t}\mathbf{F}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{7}$$

where $\bar{\alpha}_t$ is a predefined value from a fixed noise schedule. The path then progressively reduces the spatial dimensions of $\mathbf{F}_t$ to expand the network's receptive field and build a multi-scale representation for capturing broad, contextual features. Its j-th stage is defined as:

$$\mathbf{F}_{\mathrm{z}}^{j} = \begin{cases} \mathcal{M}_{2\times2}^{2}(\mathrm{TCS}(\mathbf{F}_t, t)) & \text{if } j = 1, \\ \mathcal{M}_{2\times2}^{2}\Big(\mathrm{TCS}\Big(\mathbf{F}_{\mathrm{z}}^{j-1}, t\Big)\Big) & \text{if } j \in \{2,3\}, \end{cases} \tag{8}$$

where $\mathbf{F}_{\mathrm{z}}^{j} \in \mathbb{R}^{C_j\times H_j\times W_j}$ denotes the latent feature map characterized by height H_j, width W_j, and C_j output channels at stage j, where $C_j \in \{32, 48, 96\}$. Moreover, to guide the denoising process towards a semantically meaningful output, our network also incorporates the CP.

Conditional feature path: The CP processes the conditional features $\mathbf{F}_{\mathrm{c}}^{0}$ from the spectrogram encoder and integrates them at multiple scales, thereby providing the network with explicit information about the current state of the latent mask, which is essential for generating a context-aware and time-step-specific noise prediction. The j-th stage of processing in the CP, denoted as $\mathbf{F}_{\mathrm{c}}^{j} \in \mathbb{R}^{2C_j\times H_j\times W_j}$, is formulated as follows:

$$\mathbf{F}_{\mathrm{c}}^{j} = \begin{cases} \Big\langle \mathcal{M}_{2\times2}^{2}\Big(\mathrm{TCS}\Big(\mathbf{F}_{\mathrm{e}}^{j-1}, t\Big)\Big), \mathcal{C}_{1\times1}^{1}\Big(\mathbf{F}_{\mathrm{z}}^{j}\Big)\Big\rangle & \text{if } j \in \{1,2\}, \\ \mathcal{M}_{2\times2}^{2}\Big(\mathrm{TCS}\Big(\mathbf{F}_{\mathrm{e}}^{j-1}, t\Big)\Big) & \text{if } j = 3, \end{cases} \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes the concatenation operation responsible for fusing conditional features from the encoder with features from the noisy latent path at the same resolution.

Frequency filter: The noise prediction process is enhanced by the FF module, which decomposes the latent representation $\mathbf{F}_{\mathrm{FF}}^{in}$ into two components to better separate the underlying signal structure from the superimposed noise. A low-frequency component $\mathbf{F}_{\mathrm{low}}$, representing the stable signal structure, is obtained via average pooling, while a high-frequency component $\mathbf{F}_{\mathrm{high}}$, containing fine details and potential noise, is isolated through subtraction. This decomposition allows the subsequent attention mechanism to more effectively differentiate between the signal's true details and the noise that needs to be denoised. The operation of the FF block is defined as:

$$\mathbf{F}_{\mathrm{FF}}^{out} = \mathcal{C}_{1\times1}^{1}\big(\langle\mathbf{F}_{\mathrm{high}}, \mathbf{F}_{\mathrm{low}}\rangle\big), \tag{10}$$

with $\mathbf{F}_{\text{low}} = \mathcal{A}^1_{3\times3}(\mathbf{F}^{in}_{\text{FF}})$ and $\mathbf{F}_{\text{high}} = \mathbf{F}^{in}_{\text{FF}} - \mathbf{F}_{\text{low}}$, where $\mathbf{F}^{out}_{\text{FF}} \in \mathbb{R}^{C\times\frac{H}{64}\times\frac{W}{64}}$ with $C = 128$ is the number of output channels, and $\mathcal{A}^1_{3\times3}(\cdot)$ is an average pooling layer with a kernel size of 3 and a stride of 1.

At the lowest resolution stage, features from the noisy latent path and the CP are fused with a lightweight attention mechanism. Similar to standard attention, it employs three learnable tensors ($\mathbf{F}_{\text{q}}$, $\mathbf{F}_{\text{k}}$, $\mathbf{F}_{\text{v}}$) to compute interaction scores and selectively integrate the most relevant conditioning information for denoising. The integration is performed as follows:

$$\mathbf{F}_{\text{qkv}} = \text{TCS}\Big(\sigma\Big(\mathbf{F}_{\text{qk}} \odot \mathcal{C}^1_{3\times3}(\mathbf{F}_{\text{v}})\Big)\Big), \tag{11}$$

with

$$\mathbf{F}_{\text{qk}} = \Big(\text{FF}\Big(\mathbf{F}^3_{\text{c}}\Big) \odot \mathcal{C}^1_{3\times3}(\mathbf{F}_{\text{k}})\Big) \odot \Big(\text{FF}\Big(\mathbf{F}^3_{\text{z}}\Big) \odot \mathcal{C}^1_{3\times3}(\mathbf{F}_{\text{q}})\Big),$$

where $\mathbf{F}_{\text{qk}}$ represents the correlation of the CP and the noisy latent mask, which enables the network to selectively focus on relevant conditional features when predicting noise from the latent mask. Finally, the upsampling path takes the integrated feature $\mathbf{F}_{\text{qkv}}$ and progressively reconstructs the final noise prediction. This path consists of 3 upsampling stages, each using a $2\times$ upsampling operation followed by a TCS block.

3) Model Learning: SpecDiff is trained with a hybrid loss, $\mathcal{L}_{\text{total}}$, defined as a weighted sum of a diffusion and a segmentation term $\mathcal{L}_{\text{total}} = \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}$. The diffusion loss, $\mathcal{L}_{\text{diff}}$, adapts the denoising diffusion probabilistic model objective to the latent space, minimizing the error between the true and predicted noise: $\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\mathbf{F}_0,\epsilon}[||\epsilon - \epsilon_\theta(\mathbf{F}_t, t, \mathbf{F}^0_{\text{c}})||^2]$. The segmentation loss, $\mathcal{L}_{\text{seg}}$, is a standard cross-entropy loss that ensures the final decoded mask is pixel-wise accurate. Furthermore, we use a standard cross-entropy function as the segmentation loss, $\mathcal{L}_{\text{seg}}$. To accelerate the high-latency reverse process, we use a DPM-solver++ [16] that refines the output $\mathbf{F}_0$ from a noise sample $\mathbf{F}_t$ using far fewer steps, enabling efficient, high-fidelity sampling.

III. Simulations and Discussions

A. Dataset Generation and Training Setup

Since collecting real-world 5G–LTE-radar spectrograms is prohibitively costly and complex, we constructed a realistic synthetic dataset, which serves for performance benchmark with all parameters adhering to 3GPP standards and practical radar specifications. The channel model captures urban and airport environments with up to 30 scatterers, Doppler shifts of 0–500 Hz, and SNRs from –10 to 30 dB. Radar signals emulate S-band Airport Surveillance Radar at 2.8 GHz with 25 kW peak power and high-gain antennas. 5G NR and LTE waveforms strictly follow 3GPP settings, including synchronization signal block (SSB) periodicity, subcarrier spacing, bandwidths, and Reference Measurement Channels. The dataset contains 5,000 core instances with diverse combinations (e.g., Radar+5G NR, Radar+LTE, all types), each converted into 256 × 256 spectrograms from 40 ms frames. This yields 120,000 labeled samples, with 96,000 for training/validation and 14,000 for testing. All models are trained for 60 epochs with a batch size of 32 using the Adam optimizer, starting with a 10^{-4} learning rate. Training and validation are conducted on dual T4 GPUs. Model performance is evaluated based on mean accuracy, mean IoU, learnable parameters (params), and inference execution time.

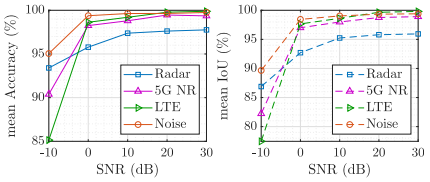
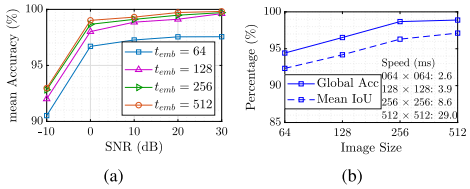


Fig. 2. Accuracy of 5G NR-LTE and radar signals versus SNR levels.



(a) (b)

Fig. 3. SpecDiff performance analysis: (a) Accuracy vs. number of DPM-solver++ steps across SNRs; (b) Accuracy vs. time embedding size across SNR; (c) Speed-accuracy trade-off at different spectrogram resolutions.

B. Simulation Results and Discussions

Signal-wise Robustness: As plotted in Fig. 2, the per-class accuracy and IoU across varying SNRs show that SpecDiff performs reliably even in extremely noisy conditions. At –10 dB, it maintains strong recognition of 5G NR and radar, with 82.23%–93.40% mean accuracies and IoU, while LTE shows greater sensitivity to low SNR with a 77.47% IoU. This degradation is reasonable, since LTE spectrograms exhibit denser subcarrier structures that are more easily distorted under heavy noise. As SNR increases, all signal types benefit from clearer spectral features, with accuracies exceeding 97.30% for radar and surpassing 98.00% for 5G NR and LTE. These results highlight two aspects: (i) SpecDiff remains robust under severe channel impairments, important for low-SNR scenarios in practical spectrum environments; and (ii) its strong radar recovery at higher SNR confirms the ability to separate transient, wideband radar emissions from the more structured 5G NR and LTE signals. This fine-grained separation demonstrates SpecDiff's strength in multi-signal characterization beyond traditional PU detection.

Hyper-parameter Investigation: In Fig. 3, we analyze the sensitivity of SpecDiff to key hyperparameters, highlighting the trade-off between accuracy, complexity, and inference speed. As shown in Fig. 3(a), increasing the time-embedding dimension improves performance up to $t_{emb} = 128$, beyond which the gains become marginal. This suggests that $t_{emb} = 128$ offers the best balance between performance and complexity. Additionally, Fig. 3(b) demonstrates that 256 × 256 resolution strikes the optimal balance, achieving 98.68% global accuracy and 96.30% mean IoU, with an efficient inference speed of 8.6 ms. Based on these observations, we adopt $t_{emb} = 128$ and 256 × 256 resolution as the default configuration for all experiments.

Ablation Study: Table I highlights the contribution of each SpecDiff component. The baseline model achieves 95.01%

TABLE I
ABLATION STUDY OF SPECDIFF COMPONENTS

Method	Components FF	Components CP	Mean Accuracy	Mean IoU	Params (M)
Baseline	✗	✗	95.01	92.55	8.2
SpecDiff	✓	✗	97.22	94.66	9.1
	✗	✓	97.80	95.48	9.5
	✓	✓	98.68	96.30	9.7

TABLE II
PERFORMANCE COMPARISON WITH VARIOUS SOTA MODELS

Method	Mean Accuracy	Mean IoU	Params (M)	Speed (ms)
DeepLabV3+ [10]	93.60	87.43	26.8	9.14
RPMNet [11]	95.52	90.47	14.5	4.63
DPSegNet [12]	92.19	86.56	7.0	7.09
HRNet [13]	91.65	86.51	29.5	9.25
SegFormer [15]	89.11	84.98	14.0	6.91
UPerNet [14]	96.24	93.10	80.3	10.99
SpecDiff	**98.68**	**96.30**	**9.7**	**8.6**

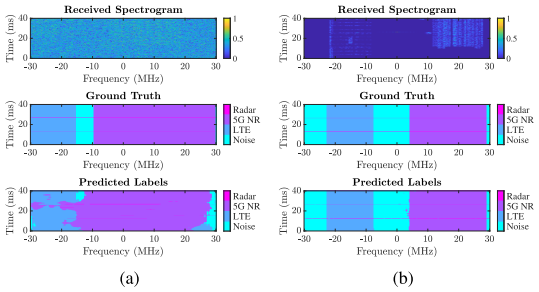


Fig. 4. Visualization of input spectrograms, ground truth masks, and the segmentation output under two SNR conditions: (a) –10 dB and (b) 20 dB.

accuracy and 92.55% IoU. Introducing the FF module improves accuracy by 2.21% and IoU by 2.11%, while the CP module provides even greater gains with a 2.79% accuracy increase and a 2.93% IoU improvement. Combining both FF and CP in the full SpecDiff model results in the highest performance, demonstrating the complementary benefits of these components in enhancing feature representation and providing semantic context for denoising and segmentation.

Method Comparison: In Table II, SpecDiff achieves SOTA performance, outperforming all listed segmentation models in mean accuracy and IoU. Concretely, it delivers a significant performance gap over established models, achieving an accuracy improvement of 5.08% to 7.03% compared to models like DeepLabV3+, HRNet, and DPSegNet, which highlights the superiority of its diffusion-based design. The model's balance of performance and efficiency is particularly compelling, that means, SpecDiff is not only more accurate than the heavyweight UPerNet but is also vastly more efficient, using over 8× fewer parameters with a faster inference speed. While RPMNet is the fastest model, SpecDiff surpasses it in accuracy by a significant margin of 3.16%, yeilding a more effective trade-off between segmentation quality and latency. pecDiff's performance arises from its core architecture, which leverages a latent diffusion process to denoise features and an attention mechanism to focus on the most salient information. As shown in Fig. 4, while the predicted signal boundaries at a low SNR of –10 dB are not perfectly aligned, SpecDiff still accurately identifies the overall occupied regions.

IV. CONCLUSION

This letter has introduced SpecDiff, an efficient latent diffusion framework designed for the semantic segmentation of coexisting 5G NR, LTE, and radar signals. Despite its small size of only 9.7M parameters, SpecDiff delivers excellent performance with 98.68% mean accuracy and 96.30% mean IoU. Its effectiveness arises from diffusion-based denoising in a latent space, guided by attention to reconstruct precise signal masks. These results demonstrate the potential of SpecDiff for multi-signal coexistence in shared spectrum environments. SpecDiff's primary limitations are potential signal confusion under extreme noise and the added computational overhead from its auxiliary modules. Future work will focus on improving robustness in harsh noise conditions and extending the framework to recognize a broader range of signals, such as satellite and next-generation radar communications, for enhanced spectrum management in heterogeneous networks.

REFERENCES

[1] H. Zhang et al., "Coexistence designs of radar and communication systems in a multi-path scenario," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 3733–3749, Mar. 2024.
[2] Y. Zhong et al., "Empowering the V2X network by integrated sensing and communications: Background, design, advances, and opportunities," *IEEE Netw.*, vol. 36, no. 4, pp. 54–60, Jul./Aug. 2022.
[3] F. Xu et al., "Deep learning-based spectrum sensing for TV white space in 5G-MBMS networks," *IEEE Trans. Broadcast.*, vol. 71, no. 3, pp. 706–716, Sep. 2025.
[4] M. Xu et al., "A reliable spectrum sensing method based on deep learning for primary user emulation attack detection in cognitive radio network," *IEEE Commun. Lett.*, vol. 28, no. 3, pp. 547–551, Mar. 2024.
[5] G. Xu et al., "Deep semi-supervised learning-based spectrum sensing at low SNR," *IEEE Commun. Lett.*, vol. 28, no. 11, pp. 2558–2562, Nov. 2024.
[6] X. Li et al., "Intelligent spectrum sensing and access with partial observation based on hierarchical multi-agent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3131–3145, Apr. 2024.
[7] W. Zhang et al., "Spectrum transformer: An attention-based wideband spectrum detector," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 12343–12353, Sep. 2024.
[8] Z. Su et al., "Signal enhancement aided end-to-end deep learning approach for joint denoising and spectrum sensing," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 4424–4428, Mar. 2024.
[9] D. Janu et al., "A graph convolution network based adaptive cooperative spectrum sensing in cognitive radio network," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2269–2279, Feb. 2023.
[10] T. Huynh-The et al., "Intelligent spectrum sensing with ConvNet for 5G and LTE signals identification," in *Proc. IEEE SSP*, Hanoi, Vietnam, Jul. 2023, pp. 140–144.
[11] H.-T. Nguyen et al., "Resolution-preserving multi-scale network for 5G-LTE spectrogram-based spectrum sensing," *IEEE Wireless Commun. Lett.*, vol. 14, no. 6, pp. 1673–1677, Jun. 2025.
[12] T.-T. Le et al., "Efficient spectrum sensing via a multi-scale dual-path segmentation network," *IEEE Wireless Commun. Lett.*, vol. 14, no. 7, pp. 2134–2138, Jul. 2025.
[13] J. Wang et al., "Deep high–resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[14] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF ICCV*, Montreal, QC, Canada, Oct. 2021, pp. 10012–10022.
[15] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NeurIPS*, vol. 34, Dec. 2021, pp. 12077–12090.
[16] C. Lu et al., "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," *Mach. Intell. Res.*, vol. 22, no. 4, pp. 730–751, Jun. 2025.