

Regular Article

DeGS-Net: Interference-Resilient Spectrum Sensing via Geometric-Semantic Consensus

Ngoc-Hoang Mai, Tan Do-Duy, Hai-Trang Phuoc Dang, Hoc Phan, Thien Huynh-The

Department of Electronics and Information Engineering, HCM City University of Technology and Engineering (HCM-UTE), Ho Chi Minh City, Vietnam

Correspondence: Thien Huynh-The, thienht@hcmute.edu.vn

Communication: received 19 March 2026, revised 02 June 2026, accepted 12 July 2026

Online publication: 01 September 2026, Digital Object Identifier: 10.21553/rev-jec.452

Abstract– Accurate spectrum sensing is critical for optimizing the coexistence of radar and communication systems in dynamic wireless environments. While deep learning has emerged as a promising solution existing semantic segmentation models often face a critical trade-off where high precision architectures are too resource intensive for real time applications while lightweight models sacrifice accuracy especially in noisy conditions. To address these challenges this paper introduces DeGS-Net a novel end-to-end framework designed for both high accuracy and computational efficiency. Its core innovation centers on a synergistic multi-task paradigm where an auxiliary denoising task is strategically employed as a powerful regularization mechanism to force the shared encoder to learn noise-invariant features. Through comprehensive experiments on a synthetic dataset containing 5G new radio, long term evolution and radar signals DeGS-Net demonstrates superior performance achieving a state-of-the-art mean intersection over union of 64.29%. Furthermore it proves to be highly efficient delivering this high accuracy with only 1.51M parameters representing a significant reduction in complexity compared to other high performance models. With a compact parameter footprint DeGS-Net offers a highly effective and scalable solution for intelligent spectrum sensing in future wireless systems.

Keywords– 6G wireless networks, deep learning, multi-task learning, semantic segmentation, spectrum sensing.

1 INTRODUCTION

The ongoing evolution towards sixth generation wireless networks is poised to seamlessly fuse the physical and digital worlds enabling a new frontier of intelligent autonomous and immersive applications [1]. Nowhere is this transformation more evident than in the automotive sector where modern vehicles are evolving into intelligent connected nodes [2]. This evolution driven by the convergence of advanced driver assistance systems and vehicle to everything communication (V2X), is creating an unprecedented demand for ultra reliable spectrum access [3]. To address the resulting scarcity of spectral resources the industry is moving towards cognitive radio frameworks [4, 5]. In this specific high mobility context the safety critical automotive Radar systems operate as the incumbent primary users (PUs), holding the highest priority for interference free operation [6]. Conversely the dense fabric of V2X links utilizing technologies like 5G New Radio, denoted as 5G NR, and long term evolution (LTE), functions as the opportunistic secondary users (SUs). This hierarchy creates a stringent operational constraint where the SUs must dynamically access the shared spectrum without disrupting the vital sensing operations of the PUs [7]. Consequently the ability to accurately identify and delineate the bursty high power signature of a Radar signal from continuous communication streams is not merely a matter of spectral efficiency but a direct safety prerequisite. Any failure by the SUs to detect the active PUs jeopardizes the integrity of the

sensing mechanism potentially leading to catastrophic navigational errors [8]. Therefore developing robust and precise spectrum sensing (SS) techniques capable of protecting the PUs while enabling efficient access for SUs is a fundamental prerequisite for realizing the vision of safe and efficient autonomous vehicular networks [9].

The field of SS was initially dominated by classical signal processing techniques, with energy detection, matched filtering, and cyclostationary feature detection being the most prominent [5]. Built on well-established mathematical principles, these methods excel in ideal scenarios, achieving optimal performance when signal characteristics are known beforehand or in high signal-to-noise ratio (SNR) environments. However, these techniques are fundamentally ill-suited for the complexities of modern, dynamic spectral environments [10]. Their primary drawbacks include a sharp performance degradation in low-SNR conditions and an often impractical reliance on prior signal knowledge [11, 12].

These collective limitations created a clear necessity for more adaptive, data-driven approaches, leading subsequent research to shift towards classical machine learning (ML) paradigms. To enhance the adaptability and precision of SS, ML-based models such as Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and K-Means clustering leverage data-driven techniques to learn complex signal patterns and dynamically adapt to spectrum variations [8]. By learning complex decision boundaries in a high-dimensional

feature space, these methods offer higher sensitivity in detecting weak signals in low-SNR conditions where traditional threshold-based methods are constrained. However, the effectiveness of these ML approaches is fundamentally constrained by their critical dependency on handcrafted feature engineering. This process requires significant domain expertise to manually select features, and the resulting feature sets are often not robust enough to generalize across the diverse conditions of modern wireless signals [13].

The dependency on expert-driven, inflexible features was ultimately broken by the advent of deep learning (DL), which fundamentally reshaped SS by enabling automatic and hierarchical feature learning directly from data. Recent deep learning methods have addressed radar-communication waveform classification using CNN-Transformer hybrids [14], while YOLO-based models have been used for signal localization [15]. While these models successfully automated feature engineering, their primary limitation was a lack of precision; classification provides no spatial information, while the coarse bounding boxes of object detection are insufficient for delineating the precise, irregular boundaries of modern wireless signals [16]. This fundamental limitation spurred the adoption of semantic segmentation as the state-of-the-art paradigm for high-precision SS. Unlike object detection, semantic segmentation assigns a class label to every pixel, enabling a detailed understanding of spectrum occupancy and the ability to delineate the exact, non-rectangular boundaries of multiple coexisting signals. Some architectures such as DeepLabV3+ [17] and SRNet [6] have been successfully adapted for this task, demonstrating significant improvements in segmentation accuracy. Despite recent advancements current deep learning solutions remain constrained by two fundamental limitations. First there exists a rigid conflict between efficiency and robustness where high capacity architectures exceed the computational budget of resource limited hardware while lightweight models suffer from severe degradation under interference. Second standard segmentation frameworks lack intrinsic mechanisms to handle spectral corruption often treating noise as valid texture features rather than structural distortion. To address this susceptibility conventional approaches often resort to a disjoint cascaded pipeline where the signal is first denoised by an independent network before being passed to a segmentation model. While this sequential strategy improves performance it incurs prohibitive latency and computational redundancy. Furthermore optimizing the denoising module in isolation often leads to the over smoothing of fine-grained signal boundaries preventing the subsequent segmentation network from recovering precise spatial details. This creates an urgent need for a unified framework that simultaneously achieves low latency operation and high noise resilience through synergistic joint feature learning [9].

To surmount these constraints, we propose DeGS-Net, a unified framework that balances high precision with low computational cost. By integrating an auxiliary denoising task as a regularization mechanism,

the network forces the shared encoder to learn robust noise-invariant features, which are then fused to guide the segmentation decoder. Different from conventional segmentation networks that directly learn from noisy spectrograms, DeGS-Net introduces a geometric-semantic consensus strategy in which the auxiliary denoising task is used as a task-aware regularizer rather than an isolated pre-processing step. The main advantages of this design are threefold: improved robustness under heavy interference, sharper boundary preservation through structural alignment, and low model complexity suitable for efficient spectrum sensing. These advantages are achieved through the coordinated design of the GCS, HLMS, and SGA modules. The main contributions of this paper are summarized as follows:

- We introduce a synergistic architecture featuring a Gradient Context Synergistic (GCS) block for multi scale feature extraction and a Hierarchical Lateral Shift Mixing (HLMS) block for efficient reconstruction. Crucially we propose a Structural Gradient Alignment (SGA) block that explicitly aligns segmentation boundaries with clean structural priors.
- We design a joint optimization strategy that simultaneously minimizes segmentation and reconstruction errors ensuring the auxiliary branch generates high fidelity structural guidance for the primary task.
- Extensive simulations on a synthetic dataset containing 5G NR LTE and radar signals demonstrate that DeGS-Net achieves a state of the art mIoU of 64.29% while requiring only 1.51M parameters significantly outperforming existing heavy architectures in efficiency.

2 RELATED WORKS

2.1 Data-Driven Approaches for SS

The inherent limitations of classical signal processing techniques prompted a shift towards more adaptive, data-driven paradigms, leading to the adoption of classical ML models. This wave of research sought to replace rigid, model-based rules with learned decision boundaries, leveraging the ability of ML to learn complex patterns from data. A wide array of supervised and unsupervised algorithms have been explored for this purpose. For instance, Shi *et al.* [18] successfully employed models such as SVMs, KNN, and K-Means clustering to precisely detect the state of multiple primary users in cooperative non-orthogonal multiple access systems. Furthermore, ML models have been integrated into sophisticated hybrid frameworks. The work in [4] demonstrates a system where a DL model performs the initial SS, while an SVM classifier is subsequently used to perform optimal channel allocation for vehicular networks based on learning features like vehicle density and network capacity. This highlights the continued relevance of classical ML in modern communication systems. However, while these ML models offered a notable improvement, their efficacy is fun-

damentally constrained by their critical dependency on handcrafted feature engineering. This reliance on expert-designed features, such as cyclostationary signatures or wavelet coefficients, is not only labor-intensive but also yields feature sets that are often brittle, failing to generalize across the diverse and dynamic conditions of modern wireless signals [19, 20].

This critical dependency on expert-driven, inflexible features ultimately spurred the adoption of DL, which enables automatic and hierarchical feature learning directly from data. A significant early advancement was the application of DL-based object detection, which revolutionized SS by reframing RF signal detection as an image recognition problem. Frameworks such as Faster R-CNN and YOLO variants have been successfully employed for automated signal detection and localization within wideband spectrograms [21, 22]. By treating spectral data as visual information, these models offered a substantial improvement over simple classification by enabling the localization of multiple co-existing signals. Further advancements have focused on enhancing these frameworks for the specific challenges of SS. For instance, to improve localization accuracy for signals of varying durations and bandwidths, some works have incorporated specialized loss functions like Generalized IoU [21]. The versatility of the YOLO paradigm has also been demonstrated in creative applications beyond simple detection, such as for information demodulation in frequency-hopping systems [23]. Other approaches have integrated object detectors into more complex, hybrid architectures; a notable example is the GAN-GRU-YOLO framework proposed by Gao *et al.* [24], which combines generative data augmentation with a YOLOv5 detector. To further improve real-time performance and handle more complex scenarios, the work in [25] leveraged even more modern architectures like YOLOv7 for RF signal classification, enabling the simultaneous localization of multiple, diverse signal types within a single frame [25]. However, a fundamental limitation persists across all these models: they rely on rectangular bounding boxes that assume regular signal boundaries. This assumption is often invalidated by the noise, interference, and channel distortion present in real-world scenarios, thus limiting their ability to precisely capture irregular or overlapping signals, especially in low-SNR environments.

Beyond object detection, the research community has explored a diverse range of advanced DL architectures to address more specific challenges in SS. Recognizing that signals often occupy varying bandwidths and exhibit multi-scale features, models like MDNN were proposed to process multi-resolution spectrograms for more effective open-set signal identification [26]. To further enhance feature extraction, hybrid models like SCBiGNet, introduced by Li *et al.* [27], have combined CNNs with BiGRUs for the complex task of joint channel estimation and signal detection. GANs have also been employed in semi-supervised frameworks to generate synthetic samples near decision boundaries to improve training robustness [28]. Notably, some works have explicitly recognized the critical challenge

of noise. The JDSS framework, for instance, represents a significant step forward by integrating a UNet-based denoising module with a CNN detector, enhanced by self-attention, to improve robustness against fading and interference [29]. However, despite their individual strengths, a common drawback of these diverse approaches is their high computational cost, which often limits their practicality for real-time, embedded processing.

The initial approach involved adapting powerful, general-purpose segmentation architectures from the computer vision domain. Foundational models like U-Net became a cornerstone for this task, and were quickly improved upon by variants such as U-Net++ which utilizes redesigned skip pathways to bridge the semantic gap between the encoder and decoder [30]. Some advanced architectures such as PSPNet and Transformer-based models like SegFormer were explored to capture more extensive contextual information and long-range dependencies [31, 32]. In parallel, recognizing the high computational cost of these models for real-time applications, other research focused on developing architecturally efficient designs. A notable example is BiSeNet, which introduced a novel two-path architecture specifically to address the trade-off between accuracy and inference speed, separating the preservation of high-resolution spatial details from the extraction of deep semantic context [33]. Building on these general-purpose foundations, several specialized networks were subsequently proposed to specifically address the unique challenges of SS. The work by Huynh-The *et al.* [17] pioneered this direction by successfully applying the DeepLabV3+ architecture, while other novel networks like SRNet were introduced to further enhance performance through cutting-edge attention and multi-scale feature extraction [6]. However, while these diverse and advanced segmentation methods significantly improve precision and efficiency, they share two critical limitations for practical deployment. First, the pursuit of higher accuracy often still leads to computationally expensive models that are ill-suited for the low-latency processing required on vehicular edge devices. Second, and more fundamentally, these architectures are not explicitly designed to handle the severe, non-structural noise prevalent in low-SNR spectrograms.

2.2 DL for Spectrogram Denoising

Spectrogram denoising is a critical and extensively studied area, as the quality of the input representation directly impacts the performance of any downstream task like segmentation. Early DL approaches often utilized variants of autoencoders to learn a compressed, latent representation of clean signals. For instance, a convolutional autoencoder network was developed in [34] to remove unstructured noise from radar signals, and this concept was further advanced by models like the Variational Denoising Autoencoder, proposed in [35], to handle the complex characteristics of thrust signals. However, a fundamental limitation of these

standard encoder-decoder structures is the tendency to produce overly smoothed results, losing the fine-grained details critical for precise analysis due to the information bottleneck. This limitation prompted the development of more sophisticated architectures. A significant breakthrough was DnCNN, which introduced residual learning to predict the noise map directly [36]. This foundational concept continues to be enhanced in modern hybrid models, as demonstrated by Li *et al.* [37], who combined a DnCNN with a gated recurrent unit to accurately extract low-frequency signals. Subsequently, UNet-like architectures became highly popular for image restoration, as their characteristic skip connections proved highly effective at preserving the sharp signal boundaries and fine-grained details critical for spectrogram analysis [38].

Building on these foundational architectures, the research community has explored even more advanced strategies to tackle the specific complexities of signal data. Recognizing that signals and noise often exhibit multi-scale features, various sophisticated strategies have been developed. For instance, some works have focused on multilayer decomposition denoising approaches, where techniques like variational mode decomposition are used to separate the signal into different components prior to processing [39]. A more integrated approach is the use of multi-stage networks, which progressively refine the signal through a cascade of denoising blocks, a principle successfully applied to complex distributed acoustic sensing data [40]. Concurrently, attention mechanisms have been integrated into denoising frameworks to allow models to intelligently focus on the most salient features. A notable application is the dual-channel CNN, which has been successfully applied to mitigate interference and improve robustness in carrier frequency offset estimation for communication signals [41]. The recent success of Transformers has also been extended to this domain, with the work in [42] demonstrating a transformer-based denoising network designed to capture the complex dependencies inherent in wireless communication signals for angle-of-arrival estimation in NLoS environments.

In parallel with these architectural enhancements, other advanced learning paradigms have been investigated to enhance data efficiency and the fidelity of the denoising process. To address the critical dependency on large, labeled datasets, self-supervised learning has emerged as a powerful paradigm. Frameworks have been proposed that leverage strategies like Noise2Noise, which trains a network to map from one noisy realization of a signal to another, thereby learning to denoise without ever seeing a clean reference [43]. Furthermore, generative models have become a dominant approach for high-fidelity signal reconstruction, moving beyond traditional pixel-wise losses to enforce perceptual realism. GANs employ an adversarial training process to enforce perceptual realism, moving beyond simple pixel-wise losses. This forces the generator to create sharper, more detailed outputs, a principle successfully applied in wireless communication tasks like high-fidelity channel estima-

tion [44]. A particularly sophisticated application is the TEM-NLnet framework, which uses a GAN not to generate the clean signal directly, but to learn the complex distribution of real-world noise for creating more realistic training data [45]. More recently, DDPMs have achieved state-of-the-art performance. By their very nature, these models are trained to reverse a gradual noising process, making them exceptionally powerful for signal reconstruction and enhancement, particularly in low-SNR wireless channels [46]. Their effectiveness for offline signal enhancement prior to classification has also been demonstrated in [47], and to address their high computational cost, variants such as Latent Diffusion models have been explored for complex tasks like signal recovery in 3D wireless networks [48].

However, while the aforementioned methods demonstrate high efficacy as standalone denoisers, their application as pre-processing tools for a subsequent segmentation task is not inherently optimal. Their architectural sophistication is almost universally geared towards a generic reconstruction objective, such as minimizing pixel-wise error, rather than being explicitly tailored to preserve or even enhance the specific spectral features most critical for a segmentation model. When integrated into a larger pipeline, this results in a decoupled, two-stage approach that is inherently limited. The denoising process, being blind to the final segmentation objective, may inadvertently suppress subtle but critical signal boundaries. This prevents any synergistic, end-to-end learning and highlights the critical research gap that our work aims to fill: a unified framework where the denoising process is not merely a pre-processing step, but an active, regularizing force that is co-optimized with and for the final segmentation task.

3 METHODOLOGY

3.1 Signal Model

The received-signal expression in this subsection follows the standard baseband signal model commonly used in wireless communication systems, where the transmitted waveform is distorted by the propagation channel and additive noise. We mathematically model the received baseband signal denoted as $y(t)$ as the convolution of the transmitted sequence $x(t)$ with a time varying channel impulse response $h(t)$ superimposed with additive white Gaussian noise $n(t)$. This fundamental relationship is expressed as

$$y(t) = x(t) * h(t) + n(t). \quad (1)$$

In this framework the transmitted component $x(t)$ encapsulates diverse waveforms including 5G New Radio Long Term Evolution and pulsed radar. The channel response $h(t)$ integrates critical propagation phenomena such as multipath fading and Doppler shifts adhering strictly to the specifications outlined in 3GPP standards [49, 50]. To facilitate deep feature extraction, the continuous signal is first discretized into a sequence $y[n]$ and subsequently transformed into the time-frequency domain via the Short-Time

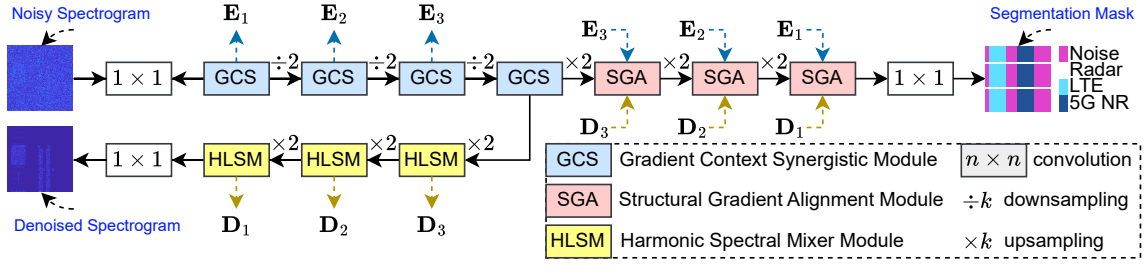


Figure 1. The proposed DeGS-Net architecture. The framework synergizes a shared GCS encoder E_i with an auxiliary HLSM denoising branch D_i to guide the primary SGA segmentation stream via structural alignment, ensuring robust detection under heavy interference.

Fourier Transform. The STFT expression is included here as a standard time-frequency transformation used to construct the spectrogram input for the network. This transformation is governed by the following expression

$$Y[m, k] = \sum_{n=0}^{N-1} y[n + mH] w[n] e^{-j \frac{2\pi}{N} kn}, \quad (2)$$

where $w[n]$ represents the analysis window function of length N while m and k denote the time frame index and discrete frequency bin index respectively with H representing the hop size between adjacent frames. Finally the complex valued coefficients $Y[m, k]$ are converted into a logarithmic power spectrum generating the two dimensional visual input required for the segmentation network.

3.2 Proposed Model

To effectively tackle the challenge of interference-resilient spectrum sensing, we propose a unified framework that synergizes deep feature extraction with structural denoising guidance. The overall architecture, as illustrated in Figure 1, is orchestrated through three strategic stages. First, the framework employs a context-aware encoder powered by the GCS Block to resolve the stability-plasticity dilemma. Complementing this, an auxiliary denoising decoder utilizes the HLSM Block to reconstruct clean geometric priors. Finally, these streams are synthesized by the SGA Module, a geometric-semantic fusion mechanism designed to strictly align the segmentation output with the clean structural reference.

GCS Block: The central component of our encoder is the GCS block, as detailed in Figure 2, meticulously designed to address the stability-plasticity dilemma in spectrogram segmentation. It employs a dual-path architecture that explicitly disentangles feature extraction into two specialized parallel branches, specifically a Micro Gradient path for fine-grained temporal boundary discovery and a Macro Context path for global spectro-temporal aggregation. Formally, let F_{GCS}^{in} denote the input feature map, where the horizontal and vertical axes correspond to the frequency and time domains, respectively. The GCS block processes this input through the following coordinated stages:

Micro-Gradient Path: To capture short-duration, high-bandwidth signal transients such as radar pulses while proactively suppressing the background noise floor, the

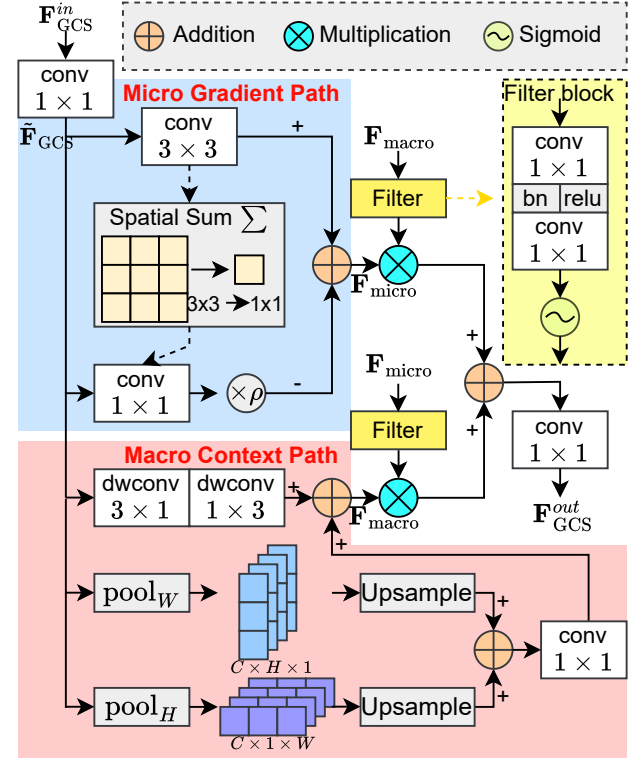


Figure 2. Structure of the proposed GCS block. The module synergizes local transient discovery via the Micro-Gradient Path and global spectral aggregation via the Macro-Context Path. The expanded view illustrates the Filter block, which generates soft attention masks for mutual feature modulation.

first branch functions as a learnable temporal edge detector. Standard convolutions typically process spectral intensity and contrast indiscriminately, often failing to isolate weak transients from a uniform background. To resolve this, we construct a differential operator that explicitly models the local spectral contrast. Specifically, we first project the input feature F_{GCS}^{in} into a calibrated feature space F' via a pointwise convolution to adapt the channel dimensionality to the hierarchical stage requirements. Subsequently, the micro-feature F_{micro} is obtained through a differential operation

$$F_{micro} = C_{3 \times 3}(\tilde{F}) - \rho \cdot C_{1 \times 1}(\tilde{F}), \quad (3)$$

with

$$\tilde{F}_{GCS} = C_{k \times k}(F_{GCS}^{in}), \quad (4)$$

where $C_{k \times k}(\cdot)$ denotes a standard convolution operation with kernel size $k \times k$, and ρ is a learnable scalar parameter. In this specific formulation, the standard

spatial convolution functions as the primary feature extractor, whereas the derived pointwise operator represents the local intensity approximation obtained intrinsically from the primary kernel's weight aggregation. By adaptively subtracting this estimated background intensity via the weighting factor ρ , the operation effectively cancels out low-frequency components to highlight high-frequency signal transients.

Macro-Context Path: Complementing the micro path, the second branch focuses on harvesting long-range dependencies across the time-frequency domain. This is critical because, while radar signals occupy the entire frequency bandwidth instantaneously, 5G and LTE signals typically exhibit continuous transmission over time, requiring a receptive field broader than what a single local convolution can provide. This pathway synergizes directional texture extraction with global context aggregation. Specifically, it employs a cross-shaped receptive field to capture directional spectral textures. Simultaneously, it extends the receptive field to the global scope via orthogonal strip pooling operations to explicitly model the spectral morphology along both the frequency and time axes. The macro-feature $\mathbf{F}_{\text{macro}}$ is derived as

$$\mathbf{F}_{\text{macro}} = \mathcal{A}_{\text{cross}}(\tilde{\mathbf{F}}_{\text{GCS}}) + \mathcal{C}_{1 \times 1}(\mathcal{S}_H(\tilde{\mathbf{F}}_{\text{GCS}}) + \mathcal{S}_W(\tilde{\mathbf{F}}_{\text{GCS}})), \quad (5)$$

where $\mathcal{A}_{\text{cross}}(\cdot)$ denotes the cross-anisotropic convolution stack, mathematically defined as the composition of 1×3 and 3×1 depthwise kernels, designed to capture directional textures with reduced computational cost. The terms $\mathcal{S}_H(\cdot)$ and $\mathcal{S}_W(\cdot)$ represent strip-context operators along the frequency and time axes, respectively, encapsulating a sequence of adaptive average pooling followed by interpolation to broadcast the global context back to the original spatial resolution.

Synergistic Fusion: Finally, to integrate the disentangled features, we employ a mutual gating mechanism where each branch modulates the other. This strategy relies on the intuition that global context filters out local noise, while local gradients sharpen semantic boundaries. The final fused output is computed as

$$\mathbf{F}_{\text{GCS}}^{\text{out}} = \mathcal{C}_{1 \times 1}(\mathbf{F}_{\text{micro}} \odot \mathcal{G}(\mathbf{F}_{\text{macro}}) + \mathbf{F}_{\text{macro}} \odot \mathcal{G}(\mathbf{F}_{\text{micro}})), \quad (6)$$

where $\odot$ denotes the element-wise product. Here, $\mathcal{G}(\cdot)$ represents the gating operator, implemented as a lightweight bottleneck sequence of pointwise convolutions and non-linear activations. This module functions to model inter-channel dependencies and generate a soft sigmoid mask, effectively suppressing irrelevant information in the counterpart branch before the final projection.

HLSM Block: Functioning as the computational keystone of the auxiliary denoising decoder, the HLSM block, presented in Figure 3, is engineered to reconstruct clean signal structures with minimal computational overhead. Unlike the primary segmentation branch which demands deep semantic abstraction, this auxiliary path acts as a structural anchor, guiding the shared encoder to distinguish true signal transients from noise. To fulfill this role without imposing a heavy

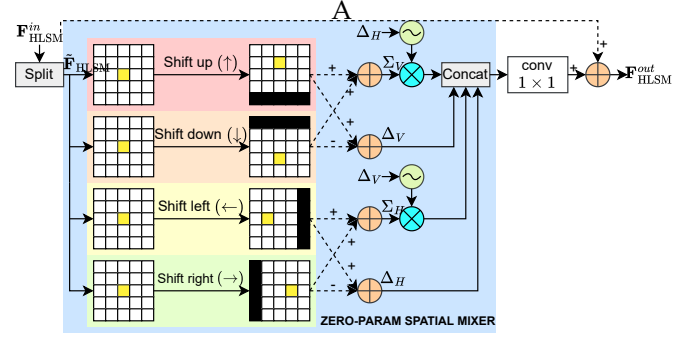


Figure 3. Structure of the proposed HLSM block. The module utilizes a Zero-Param Spatial Mixer with deterministic shifts, where structural difference maps gate the summation maps for efficient, parameter-free reconstruction.

burden on the overall architecture, HLSM substitutes standard convolutions with a paradigm of physical shifting. Specifically, we employ a deterministic spatial tokenization mechanism where the input tensor $\mathbf{F}_{\text{HLSM}}^{\text{in}}$ is first split along the channel dimension into four distinct partitions. Each partition then undergoes a cyclic spatial shift in a cardinal direction (top, bottom, left, right), allowing pixel-level operations to access neighboring contexts with significantly reduced memory bandwidth. Building upon these shifted partitions, we formulate an orthogonal differential interaction strategy to explicitly decouple the feature space via linear combinations of opposing views:

$$\begin{aligned} \Delta_V &= \tilde{\mathbf{F}}_{\text{HLSM}}^{\uparrow} - \tilde{\mathbf{F}}_{\text{HLSM}}^{\downarrow}, & \Sigma_V &= \tilde{\mathbf{F}}_{\text{HLSM}}^{\uparrow} + \tilde{\mathbf{F}}_{\text{HLSM}}^{\downarrow}, \\ \Delta_H &= \tilde{\mathbf{F}}_{\text{HLSM}}^{\leftarrow} - \tilde{\mathbf{F}}_{\text{HLSM}}^{\rightarrow}, & \Sigma_H &= \tilde{\mathbf{F}}_{\text{HLSM}}^{\leftarrow} + \tilde{\mathbf{F}}_{\text{HLSM}}^{\rightarrow}, \end{aligned} \quad (7)$$

where $\tilde{\mathbf{F}}_{\text{HLSM}}^{\uparrow}, \tilde{\mathbf{F}}_{\text{HLSM}}^{\downarrow}, \tilde{\mathbf{F}}_{\text{HLSM}}^{\leftarrow}, \tilde{\mathbf{F}}_{\text{HLSM}}^{\rightarrow}$ denote the channel partitions cyclically shifted in the top, bottom, left, and right directions, respectively. In this context, the difference term $\Delta_{(\cdot)}$ functions as a structural gradient, approximating the first-order derivative to capture high-frequency edges along the vertical V and horizontal H axes. To resolve the potential spatial incoherence where noise suppression in one direction might blur edges in the other, we introduce a harmonic mutual gating mechanism. This strategy utilizes the structural gradients of one axis to spatially modulate the signal context of the orthogonal axis, effectively preventing texture bleeding. We define the harmonized feature groups $\mathbf{F}_V$ and $\mathbf{F}_H$ as:

$$\mathbf{F}_V = [\Sigma_V \odot \sigma(\Delta_H), \Delta_V], \quad (8)$$

$$\mathbf{F}_H = [\Sigma_H \odot \sigma(\Delta_V), \Delta_H], \quad (9)$$

where $[\cdot]$ denotes channel concatenation and $\sigma(\cdot)$ is the sigmoid activation. Finally, the denoised representation is synthesized by fusing these components via a lightweight linear projection

$$\mathbf{F}_{\text{HLSM}}^{\text{out}} = \mathbf{F}_{\text{HLSM}}^{\text{in}} + \mathcal{C}_{1 \times 1}([\mathbf{F}_V, \mathbf{F}_H]). \quad (10)$$

SGA Block: While the primary decoder excels at capturing semantic abstractions, it remains inherently susceptible to noise-induced artifacts, often failing to

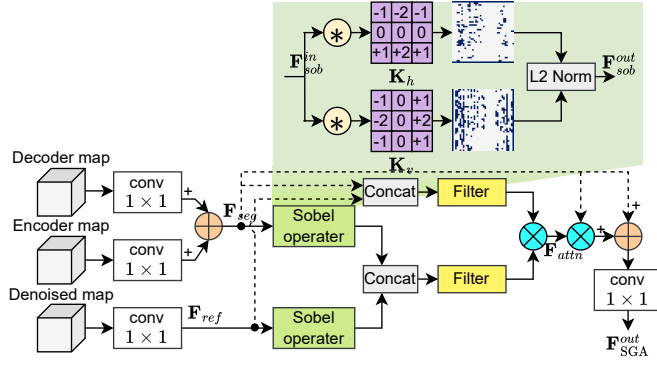


Figure 4. Structure of the SGA module. The design utilizes explicit Sobel-based gradient priors to align the segmentation stream $\mathbf{F}_{seg}$ with the clean reference $\mathbf{F}_{ref}$. A multiplicative gating mechanism fuses these structural and semantic cues to generate the unified attention map $\mathbf{F}_{attn}$ for residual feature rectification.

distinguish true object boundaries from high-frequency interference. This limitation is critical in signal segmentation, where imprecise edges can lead to catastrophic misinterpretation. To overcome this, we introduce the SGA module, visualized in Figure 4, which acts as a geometric validator. Its core motivation is to enforce a rigorous structural consensus: the segmentation stream is allowed to propagate features only if they are geometrically ratified by the clean auxiliary reference $\mathbf{F}_{ref}$. To operationalize this philosophy without introducing learnable biases, we first project the latent feature spaces into an explicit gradient domain. The Sobel-based expression is used as a standard fixed gradient operator to extract structural boundary information from feature maps. This operation is defined as

$$\Psi(\mathbf{F}_{sob}^{in}) = \|\nabla \mathbf{F}_{sob}^{in}\|_2 = \sqrt{(\mathbf{K}_v * \mathbf{F}_{sob}^{in})^2 + (\mathbf{K}_h * \mathbf{F}_{sob}^{in})^2} + \epsilon, \quad (11)$$

where $\mathbf{F}_{sob}^{in}$ denotes an arbitrary input tensor, and $\mathbf{K}_v, \mathbf{K}_h$ represent the fixed vertical and horizontal Sobel kernels, respectively. By applying $\Psi(\cdot)$, we effectively strip away semantic variations to expose the raw structural skeleton of the signal. Building upon this geometric projection, we formulate a dual-gated synergistic mechanism to generate a unified feature attention map $\mathbf{F}_{attn}$. Unlike standard fusion strategies that combine features additively, we employ a multiplicative consensus to strictly enforce that valid features must satisfy consistency in both the structural and intensity domains simultaneously

$$\mathbf{F}_{attn} = \mathcal{G}(\mathbf{F}_{seg}, \mathbf{F}_{ref}) \odot \mathcal{G}(\Psi(\mathbf{F}_{seg}), \Psi(\mathbf{F}_{ref})). \quad (12)$$

In this formulation, the first interaction term evaluates the semantic coherence by comparing raw feature distributions, while the second term assesses the geometric fidelity through the projected gradient representations. Crucially, the integration of these two assessments imposes a strict consensus constraint: acting as a rigorous validity filter, it suppresses the attention weight if either the structural boundary or the intensity context fails to align, thereby minimizing the propagation of noise-induced hallucinations. Finally, the primary segmenta-

tion stream is rectified via a residual modulation

$$\mathbf{F}_{SGA}^{out} = \mathbf{F}_{seg} \odot (1 + \mathbf{F}_{attn}). \quad (13)$$

In this formulation, the term $(1 + \mathbf{F}_{attn})$ serves as a selective amplifier. It preserves the original semantic information while significantly enhancing regions where the multiplicative consensus is high, effectively transforming the auxiliary branch into an active structural supervisor that sharpens object boundaries with high precision.

Joint Optimization Objective: To enable end-to-end learning, we formulate a multi-task objective that simultaneously minimizes the segmentation error and the reconstruction discrepancy. The total loss function $\mathcal{L}_{total}$ is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \cdot \mathcal{L}_{rec}, \quad (14)$$

where $\mathcal{L}_{seg}$ denotes the standard multi-class cross-entropy loss applied to the final segmentation mask, and $\mathcal{L}_{rec}$ represents the mean squared error loss enforcing structural fidelity in the auxiliary denoising branch. The term λ serves as a balancing parameter to modulate the contribution of the auxiliary supervision.

4 SIMULATIONS AND RESULTS

4.1 Dataset and Training Configurations

To evaluate the proposed framework, we generated a dataset of 120,000 synthetic spectrograms simulating the coexistence of 5G NR, LTE, and radar signals. Following 3GPP standards [49, 50], the simulation operates at a carrier frequency of 2.8 GHz with a sampling rate of 61.44 MHz. To emulate diverse channel impairments, we use signal-to-noise ratios from -10 to 30 dB and Doppler shifts of $\{0, 5, 70, 500\}$ Hz. The signal configurations include 5G NR bandwidths from 10 to 50 MHz, LTE bandwidths from 5 to 20 MHz, and $1 \mu\text{s}$ radar pulses, all converted into 256×256 RGB time-frequency representations. Although the dataset is synthetically generated, it is designed as a controlled but diverse evaluation environment rather than a single-condition benchmark. This design is motivated by the need for reliable pixel-level ground-truth masks for overlapping 5G NR, LTE, and radar signals, which are difficult to obtain from real over-the-air measurements without synchronized access to all transmitters. To further test generalization, the model is trained only on dual-signal scenarios, such as 5G NR with radar or LTE with radar, and evaluated on unseen triple-signal coexistence. This protocol is intended to evaluate compositional generalization under unseen interference combinations and to examine whether the proposed model can learn noise-invariant features beyond the specific signal mixtures observed during training. The network is optimized using Adam with a learning rate of 10^{-4} for 60 epochs and a batch size of 32 on a P100 GPU.

4.2 Results and Discussions

We first evaluate DeGS-Net under different SNR and Doppler conditions, as shown in Figure 5. As expected,

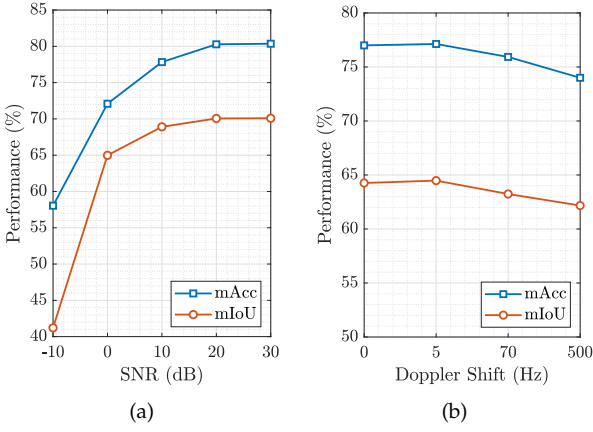


Figure 5. Segmentation performance under varying channel conditions: (a) versus SNR and (b) versus Doppler shift.

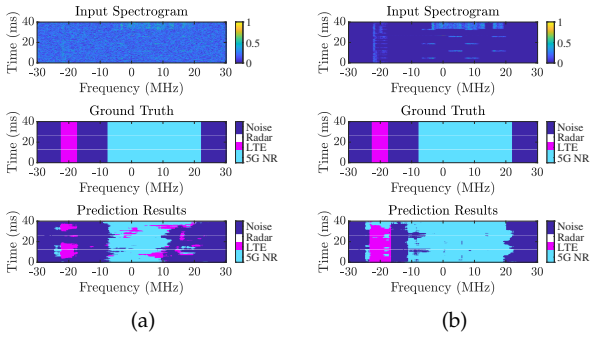


Figure 6. Visualization of segmentation results for the proposed method at different SNR levels: (a) -10 dB and (b) 30 dB.

performance improves as SNR increases. The model maintains a functional mIoU of 41.21% at the challenging -10 dB condition, then rises sharply to 64.98% at 0 dB and saturates at approximately 70.09% under higher SNRs. This trend shows that DeGS-Net can recover reliable segmentation once the signal becomes moderately observable. Under Doppler shifts, the model remains stable, with mIoU decreasing only from 64.48% at 5 Hz to 62.17% at 500 Hz, confirming its robustness in high-mobility scenarios. The qualitative results in Figure 6 further illustrate the boundary condition of the auxiliary denoising task. At 30 dB, DeGS-Net produces sharp and complete signal regions, showing that the denoising branch effectively suppresses non-structural noise while preserving useful spectral geometry. At -10 dB, however, some weak LTE structures become partially discontinuous, indicating that very weak signal traces may be smoothed when they are nearly indistinguishable from the noise floor. Therefore, the auxiliary task is most helpful when corrupted signals still contain recoverable structural patterns, but its benefit becomes limited under extremely low-SNR conditions. Secondly, to validate the contribution of each architectural component and identify the optimal hyperparameters, we conducted ablation studies as shown in Figure 7. We first examine the effect of network depth, controlled by the number of encoder stages n , on both accuracy and complexity. As shown in Figure 7(a), increasing the depth consistently improves

mIoU across all SNR conditions. Moving from a 2-stage to a 3-stage architecture increases the average mIoU by approximately 3.25%, while expanding to four stages provides a larger gain of 5.73%, especially under low-SNR conditions. However, using five stages yields only a marginal 1.09% improvement, while nearly quadrupling the parameter count and increasing inference time. This indicates a point of diminishing returns, where additional depth no longer provides a favorable accuracy-complexity trade-off. We further assess the contribution of each proposed module through the module ablation results in Figure 7(b). The baseline model $\mathcal{M}_1$, built with standard convolutions, achieves an average mIoU of 50.17%. Replacing the standard encoder with the proposed GCS blocks forms $\mathcal{M}_2$ and improves the mIoU to 56.96%, confirming the benefit of local-global spectrogram feature extraction. Adding the auxiliary denoising branch with the HLSM block further increases the mIoU to 59.24%, showing that structural priors from the reconstruction task provide useful guidance for segmentation. Finally, incorporating the SGA module forms the complete DeGS-Net and raises the mIoU to 64.29%, an improvement of over 5% compared with $\mathcal{M}_3$. This confirms that the explicit geometric alignment between denoising and segmentation streams is essential for effectively transferring clean structural priors to the final segmentation task.

Furthermore, we investigated the influence of network width, governed by the number of base channels N , as shown in Figure 7(c). Performance generally improves as N increases. Doubling the base channels from $N = 8$ to $N = 16$ yields an mIoU improvement of 4.08%, and increasing N to 32 provides a further gain of approximately 8.11%. However, this improvement becomes inefficient at higher channel counts. Expanding to $N = 64$ only increases mIoU by 0.88%, while nearly quadrupling the parameter count from 1.51M to 5.96M. This suggests that $N = 32$ offers the best balance between accuracy and complexity. We also evaluate the effect of input resolution, as shown in Figure 7(d). Higher resolutions consistently improve performance by preserving richer time-frequency details. The average mIoU increases from 43.56% with 64×64 inputs to 50.48% with 128×128 inputs, and further reaches 64.29% at 256×256 . Although 512×512 achieves the highest mIoU of 65.38%, the additional gain of only 1.09% is outweighed by the increased computational cost. Therefore, the final configuration with $n = 4$, $N = 32$, and 256×256 input resolution is selected as the best trade-off between segmentation accuracy and computational efficiency. This result also clarifies the trade-off between architectural specialization and performance gain. Although DeGS-Net introduces three task-specific modules, namely GCS, HLSM, and SGA, the complete model remains compact with only 1.51M parameters. The improvement therefore does not mainly come from simply scaling up the network, but from assigning each module a targeted function: GCS extracts local-global spectrogram structures, HLSM provides lightweight denoising guidance, and SGA transfers clean geometric priors to the seg-

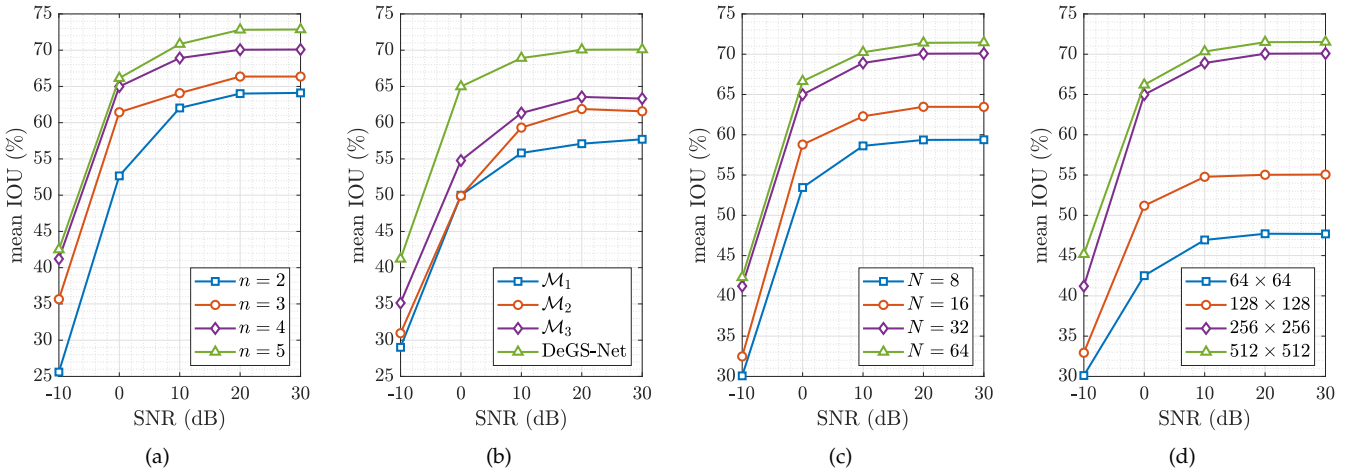


Figure 7. Ablation study of DeGS-Net's key components and hyperparameters. The impact of varying (a) number of encoder stages, (b) module ablation configurations, (c) base channels, and (d) input image size.

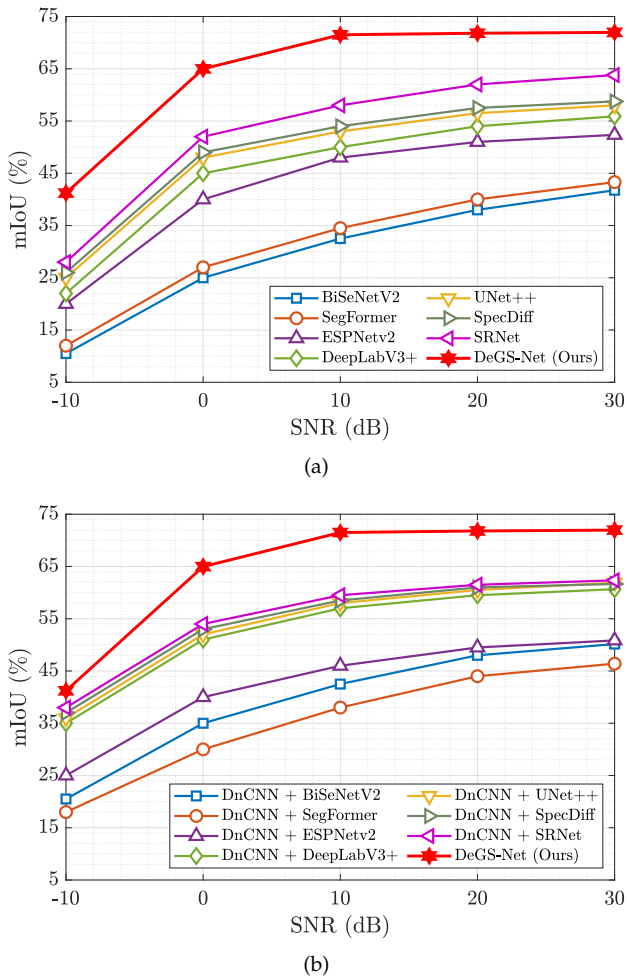


Figure 8. Performance comparison of segmentation models, detailing: (a) results on raw noisy spectrograms and (b) results on spectrograms pre-denoised by a DnCNN.

mentation stream. This leads to a favorable accuracy-efficiency balance for resource-constrained spectrum sensing.

Finally we benchmark the proposed DeGS Net against seven state of the art semantic segmentation models as illustrated in Figure 8(a). The analysis reveals

a distinct performance hierarchy driven largely by architectural suitability for stochastic signal processing. First distinct observation is that models relying on extensive global context aggregation like BiSeNetV2 and SegFormer proved fundamentally ill suited for spectrogram analysis in low Signal to Noise Ratio regimes. Both architectures are optimized for natural scenes where integrating background context aids object recognition. However in the domain of wireless signal processing the background consists primarily of high power stochastic interference. Consequently BiSeNetV2 via its global pooling layers and SegFormer through its global self attention mechanism inadvertently assimilate this background noise into the feature representation resulting in the lowest average mIoU scores of 29.55% and 31.36% respectively. Intermediate improvements are observed with lightweight architectures such as ESPNetv2 and DeepLabV3+ which achieve average mIoU scores of 42.27% and 45.38% respectively. While DeepLabV3+ leverages atrous spatial pyramid pooling to capture multi scale features its reliance on a lightweight MobileNet backbone limits its capacity to disentangle weak signals from heavy interference. Moving to the top tier baselines UNet++, SpecDiff and SRNet demonstrate significantly better resilience reaching average mIoU scores ranging from 48.10% to 52.76%. Their superiority stems from the utilization of dense nested skip pathways in UNet++ and specialized attention mechanisms in SRNet and SpecDiff which facilitate the retention of fine grained features. Nevertheless this precision comes at the cost of prohibitive computational complexity and high parameter counts rendering them impractical for resource constrained edge deployment. Furthermore, to validate the benefit of the proposed synergistic multi-task learning strategy, we also evaluate the baselines using a cascaded pipeline where the input spectrograms are first pre-denoised by a standalone DnCNN, as shown in Figure 8(b). While pre-denoising provides noticeable recovery for some lightweight architectures, it yields only marginal gains for stronger models such as SRNet and UNet++ at higher SNR levels, where their

Table I
PERFORMANCE BENCHMARK OF DeGS-Net AND COMPARISON WITH EXISTING METHODS

Model	n	GCS	HLSM	SGA	N	Image Size	mAcc (%)	mIoU (%)	Params (M)	Speed (ms)
DeGS-Net (Figure 7(a))	2	✓	✓	✓	32	256×256	71.63	55.31	0.12	10.5
	3	✓	✓	✓	32	256×256	75.79	58.56	0.47	17.4
	4	✓	✓	✓	32	256×256	77.11	64.29	1.51	20.3
	5	✓	✓	✓	32	256×256	78.23	65.38	6.01	26.8
$\mathcal{M}_1$ $\mathcal{M}_3$ $\mathcal{M}_4$ DeGS-Net } (Figure 7(b))	4	✗	✗	✗	32	256×256	65.08	50.17	0.52	12.1
	4	✓	✗	✗	32	256×256	71.22	56.96	1.12	15.3
	4	✓	✓	✗	32	256×256	72.95	59.24	1.14	16.7
	4	✓	✓	✓	32	256×256	77.11	64.29	1.51	20.3
DeGS-Net (Figure 7(c))	4	✓	✓	✓	8	256×256	67.44	52.10	0.11	17.9
	4	✓	✓	✓	16	256×256	72.40	56.18	0.38	19.1
	4	✓	✓	✓	32	256×256	77.11	64.29	1.51	20.3
	4	✓	✓	✓	64	256×256	77.67	65.17	5.96	24.8
DeGS-Net (Figure 7(d))	4	✓	✓	✓	32	64×64	61.56	43.56	1.51	19.0
	4	✓	✓	✓	32	128×128	66.98	50.48	1.51	19.5
	4	✓	✓	✓	32	256×256	77.11	64.29	1.51	20.3
	4	✓	✓	✓	32	512×512	78.30	65.38	1.51	25.1
BiSeNet [33] – MobileNetV2 SegFormer [32] – MiT_B1 ESPNetv2 [51] DeeplabV3+ [17] – MobileNetV2 UNet++ [30] – MobileNetV2 SpecDiff [52] SRNet [6] – MobileNetV2 DeGS-Net (proposed)	} (Figure 8(a))					256×256	46.19	29.55	4.31	18.5
						256×256	48.79	31.36	13.63	21.9
						256×256	57.80	42.27	1.49	20.2
						256×256	59.52	45.38	4.37	18.9
						256×256	62.35	48.10	6.82	19.7
						256×256	63.19	49.05	9.72	22.6
						256×256	66.79	52.76	10.72	19.4
DnCNN [36] + BiSeNet [33] – MobileNetV2 DnCNN [36] + SegFormer [32] – MiT_B1 DnCNN [36] + ESPNetv2 [51] DnCNN [36] + DeeplabV3+ [17] – MobileNetV2 DnCNN [36] + UNet++ [30] – MobileNetV2 DnCNN [36] + SpecDiff [52] DnCNN [36] + SRNet [6] – MobileNetV2 DeGS-Net (proposed)	} (Figure 8(b))					256×256	54.21	39.23	4.81	20.5
						256×256	50.28	35.28	14.13	23.9
						256×256	57.80	42.27	1.49	20.2
						256×256	66.16	52.63	4.87	20.9
						256×256	63.55	53.68	7.32	21.7
						256×256	64.73	54.23	10.22	24.6
						256×256	67.13	55.07	11.22	21.4
						256×256	77.11	64.29	1.51	20.3

performance tends to saturate. More importantly, under the severe -10 dB condition, the decoupled denoiser still fails to prevent performance degradation across the baselines. This suggests that generic denoising, when optimized independently from the segmentation objective, may remove not only noise but also weak signal details that are important for pixel-level prediction. In contrast, DeGS-Net treats denoising as a task-aware regularizer rather than an isolated pre-processing step. The auxiliary branch encourages the encoder to learn noise-invariant structural features, while the SGA module transfers such guidance only when the denoised reference and segmentation features are geometrically consistent. As a result, DeGS-Net maintains an mIoU of 41.21% even at -10 dB and consistently outperforms all compared models. Nevertheless, this result also reveals a boundary condition of the auxiliary denoising strategy: when weak signals are almost submerged by noise, some subtle structures may still be partially smoothed or missed.

Table I summarizes the comprehensive performance of DeGS-Net across various configurations and against other state of the art methods. Our ablation studies confirm that the final configuration is a carefully optimized architecture with each proposed module incrementally improving performance. While the optimized DeGS-Net achieves a competitive inference time of approximately 20.3 ms comparable to top tier models like SRNet a crucial distinction lies in its efficiency. DeGS-Net maintains a remarkably small footprint of

only 1.51M parameters a significant reduction compared to other high performance models like SRNet with 11.22M parameters and SegFormer with 14.13M parameters. More specifically, Table I can be interpreted from three complementary perspectives. First, the module ablation results show the progressive contribution of the proposed components, where GCS improves feature extraction, HLSM introduces denoising guidance, and SGA further enhances segmentation by aligning semantic features with clean geometric priors. Second, the hyperparameter analysis indicates that the setting $n = 4$, $N = 32$, and 256×256 input resolution provides a practical balance between accuracy and computational cost. Third, the comparison with existing segmentation models confirms that DeGS-Net achieves the highest mIoU of 64.29% while maintaining only 1.51M parameters, demonstrating a favorable accuracy-efficiency trade-off.

5 CONCLUSION

In this paper, we proposed DeGS-Net, a novel end-to-end framework for robust and efficient spectrogram segmentation. The core innovation lies in a synergistic multi-task design, where an auxiliary denoising task acts as a regularization mechanism through specialized modules for multi-scale analysis and feature fusion. Experiments on a synthetic dataset of 5G NR, LTE, and radar signals show that DeGS-Net achieves a state-

of-the-art mIoU of 64.29% with only 1.51 million parameters, demonstrating strong robustness and efficiency compared with other high-performance models. The results also confirm that the final configuration is the outcome of deliberate design choices, with each proposed component contributing incrementally to the overall performance. Nevertheless, we acknowledge that the current validation is still based on simulated RF spectrograms. Although the simulation follows standardized waveform and channel configurations and includes diverse SNR, Doppler, bandwidth, and interference conditions, real-world RF environments may introduce additional impairments, such as hardware nonlinearity, phase noise, I/Q imbalance, synchronization mismatch, calibration errors, and non-Gaussian interference. These factors may lead to a synthetic-to-real gap that cannot be fully resolved by simulation alone. Although DeGS-Net is developed for 5G NR, LTE, and radar spectrogram segmentation, its core design is not limited to these specific signal types. Since the proposed modules operate on general time-frequency structures, the framework can potentially be adapted to related tasks such as RF interference localization, anomaly segmentation, and spectrogram-based sensing for other wireless protocols. Nevertheless, applying DeGS-Net to substantially different signal modalities may require task-specific retraining, calibration, and hyperparameter adjustment. Future work will focus on leveraging unlabeled real-world RF data through semi-supervised learning, optimizing the model for edge deployment using quantization and pruning, and extending the framework to additional RF signals such as WiFi and emerging IoT protocols.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.02-2024.21

REFERENCES

- [1] R. Liu, L. Zhang, R. Y.-N. Li, and M. D. Renzo, "The ITU vision and framework for 6G: Scenarios, capabilities, and enablers," *IEEE Vehicular Technology Magazine*, vol. 20, no. 2, pp. 114–122, Jun. 2025.
- [2] M. Elloumi *et al.*, "Digital-twin-empowered interference management for multihop internet of vehicles networks over millimeter wave bands," *IEEE Internet of Things Journal*, vol. 12, no. 11, pp. 17 807–17 827, Jun. 2025.
- [3] Y. Xu *et al.*, "Joint resource allocation for UAV-assisted V2X communication with mean field multi-agent reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 1, pp. 1209–1223, Sep. 2025.
- [4] R. Ahmed *et al.*, "Hybrid machine-learning-based spectrum sensing and allocation with adaptive congestion-aware modeling in CR-Assisted IoV networks," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25 100–25 116, Aug. 2022.
- [5] X. Cao *et al.*, "AI-empowered multiple access for 6G: A survey of spectrum sensing, protocol designs, and optimizations," *Proceedings of the IEEE*, vol. 112, no. 9, pp. 1264–1302, Sep. 2024.
- [6] T. Huynh-The *et al.*, "SRNet: Deep semantic segmentation network for spectrum sensing in wireless communications," *IEEE Wireless Communications Letters*, vol. 14, no. 2, pp. 355–359, Feb. 2025.
- [7] L. Du, P. Fan, Z. Ma, and Q. Liang, "RIS assisted radar-communication coexistence system with discrete reflection coefficients," *IEEE Transactions on Wireless Communications*, vol. 24, no. 4, pp. 3014–3028, Jan. 2025.
- [8] N. C. Luong *et al.*, "Advanced learning algorithms for integrated sensing and communication (ISAC) systems in 6G and beyond: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, pp. 1–1, Jul. 2025.
- [9] M. Temiz *et al.*, "Deep-learning-based techniques for integrated sensing and communication systems: State-of-the-art, challenges, and opportunities," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 5940–5968, Jul. 2025.
- [10] T.-D. Tran, M.-T. Le, S. N. Truong, N.-H. Truong, P. N. Son, P. Hoc, T. Do-Duy, and T. Huynh-The, "CoDeX-Net: A coordinate-gated and deformable cross-scale network for efficient waveform classification in integrated radar-communication systems," *REV Journal on Electronics and Communications*, vol. 16, no. 1, pp. 12–23, 2026.
- [11] T. Huynh-The, N.-H. Mai, Q.-V. Pham, and H.-T. P. Dang, "Denoise2seg: A deep denoising-segmentation model for interference-resilient spectrum sensing," *IEEE Wireless Communications Letters*, vol. 15, pp. 3119–3123, 2026.
- [12] N.-H. Mai, S. N. Truong, D. B. da Costa, Q.-V. Pham, and T. Huynh-The, "Learning spectral structures: Efficient segmentation via hierarchical kernel attention," *IEEE Wireless Communications Letters*, vol. 15, pp. 3866–3870, 2026.
- [13] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using RF data: A review," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 77–100, 2023.
- [14] T.-D. Tran and T. Huynh-The, "DTANet: High-performance network with decoupled three-dimensional attention for radar-communications waveform classification," *IEEE Communications Letters*, vol. 29, no. 11, pp. 2741–2745, 2025.
- [15] S. Sarkar, D. Guo, and D. Cabric, "RadYOLOLet: Radar detection and parameter estimation using YOLO and wavelet," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, Dec. 2024.
- [16] H.-T. Nguyen, H.-T. Phuoc Dang, Q.-V. Pham, and T. Huynh-The, "Resolution-preserving multi-scale network for 5G-LTE spectrogram-based spectrum sensing," *IEEE Wireless Communications Letters*, vol. 14, no. 6, pp. 1673–1677, Jun. 2025.
- [17] T. Huynh-The, Q.-V. Pham, T.-H. Vu, D. B. da Costa, and V.-P. Hoang, "Intelligent spectrum sensing with convnet for 5G and LTE signals identification," in *Proceedings of the IEEE Statistical Signal Processing Workshop*, Hanoi, Vietnam, Jul. 2023, pp. 140–144.
- [18] Z. Shi *et al.*, "Machine learning-enabled cooperative spectrum sensing for non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 19, no. 9, pp. 5692–5702, May 2020.
- [19] Y. Li *et al.*, "Enhanced support vector machine for cooperative spectrum sensing against byzantine attack in cognitive wireless sensor networks," *IEEE Sensors Journal*, vol. 24, no. 23, pp. 39 835–39 844, Dec. 2024.
- [20] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using RF data: A review," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 77–100, Sep. 2023.
- [21] K. N. R. S. V. Prasad *et al.*, "A downscaled Faster-RCNN framework for signal detection and time-frequency localization in wideband RF systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4847–4862, Jul. 2020.
- [22] A. Vagollari, V. Schram, W. Wicke, M. Hirschbeck, and

- W. Gerstacker, "Joint detection and classification of RF signals using deep learning," in *Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, Helsinki, Finland, May 2021, pp. 1–7.
- [23] L. M. Hoang et al., "Frequency hopping joint radar-communications with hybrid sub-pulse frequency and duration modulation," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2300–2304, Nov. 2022.
- [24] Z. Gao et al., "Intelligent spectrum sensing of consumer IoT based on GAN-GRU-YOLO," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 6140–6148, Aug. 2024.
- [25] T. Anna and R. D. Reese, "YOLO for radio frequency signal classification," in *Proceedings of the 2024 IEEE International Symposium on Dynamic Spectrum Access Networks*, USA, May 2024, pp. 177–178.
- [26] H. Lin, X. Dai, and C. Wang, "A deep learning-based method for open-set wireless signal detection in HF band," in *Proceedings of the International Conference on Wireless Communications and Signal Processing*, Hefei, China, Oct. 2024, pp. 163–168.
- [27] J. Li, Z. Zhang, Y. Wang, B. He, W. Zheng, and M. Li, "Deep learning-assisted OFDM channel estimation and signal detection technology," *IEEE Communications Letters*, vol. 27, no. 5, pp. 1347–1351, May 2023.
- [28] G. Xu, Y. Wang, B. Zheng, and J. Li, "Deep semi-supervised learning-based spectrum sensing at low SNR," *IEEE Communications Letters*, vol. 28, no. 11, pp. 2558–2562, Nov. 2024.
- [29] Z. Su et al., "Signal enhancement aided end-to-end deep learning approach for joint denoising and spectrum sensing," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 3, pp. 4424–4428, Mar. 2024.
- [30] Z. Zhou et al., "UNet++: A nested U-Net architecture for medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Granada, Spain, Sep. 2018, pp. 3–11.
- [31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, Jul. 2017.
- [32] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proceedings of the Conference on Neural Information Processing Systems*, vol. 34, Virtual, Oct. 2021, pp. 12 077–12 090.
- [33] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, Munich, Germany, Sep. 2018.
- [34] Y. Chen et al., "Physical layer authentication for industrial control based on convolutional denoising autoencoder," *IEEE Internet of Things Journal*, vol. 11, no. 9, pp. 15 633–15 641, May 2024.
- [35] X. Chen et al., "A high-generalization variational denoising autoencoder for microneutron thrust signal noise removal and step reconstruction," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–15, Feb. 2025.
- [36] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
- [37] G. Li et al., "Low-frequency magnetotelluric data denoising using improved denoising convolutional neural network and gated recurrent unit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.
- [38] J. Chen, G. Chen, J. Li, R. Du, Y. Qi, C. Li, and N. Wang, "Efficient seismic data denoising via deep learning with improved MCA-SCUNet," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, Jan. 2024.
- [39] C. Wang et al., "The denoising method with variational mode decomposition for signal-dependent counting noise in molecular communications," *IEEE Sensors Journal*, vol. 24, no. 15, pp. 24 337–24 343, Aug. 2024.
- [40] Y. Li et al., "Distributed acoustic sensing vertical seismic profile data denoising based on multistage denoising network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, Jul. 2022.
- [41] Y. Zhang et al., "Carrier frequency offset estimation based on signal denoising and dual-channel convolutional attention network under low signal-to-noise ratio," *IEEE Communications Letters*, 2025.
- [42] J. Liu et al., "A transformer-based signal denoising network for AoA estimation in nlos environments," *IEEE Communications Letters*, vol. 26, no. 10, pp. 2336–2339, Jun. 2022.
- [43] X. Wang, S. Fan, C. Zhao, D. Liu, and W. Chen, "A self-supervised method using noise2noise strategy for denoising crp gathers," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, Jun. 2023.
- [44] M. Ye, H. Zhang, and J.-B. Wang, "Channel estimation for intelligent reflecting surface aided wireless communications using conditional GAN," *IEEE Communications Letters*, vol. 26, no. 10, pp. 2340–2344, Apr. 2022.
- [45] M. Wang et al., "TEM-NLnet: A deep denoising network for transient electromagnetic signal with noise learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, Jan. 2022.
- [46] M. Letafati et al., "Conditional denoising diffusion probabilistic models for data reconstruction enhancement in wireless communications," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 3, pp. 133–146, Dec. 2025.
- [47] Y. Zhang et al., "Offline real-world wireless interference signal classification algorithm utilizing denoising diffusion probability model," *IEEE Signal Processing Letters*, vol. 30, pp. 1132–1136, Aug. 2023.
- [48] A. B. M. Adam et al., "GS3D: Signal recovery in 3-D wireless networks with gaussian splatting-synergized stable diffusion," *IEEE Wireless Communications Letters*, vol. 14, no. 7, pp. 2019–2023, Jul. 2025.
- [49] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project; Technical Specification Group Radio Access Network, Technical Report TR 38.901, Jan. 2023.
- [50] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception," 3GPP, Technical Specification TS 36.101, Mar. 2023.
- [51] S. Mehta et al., "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, Jun. 2019.
- [52] T. Huynh-The, P.-L. Huynh, V.-C. Phan, T.-H. Vu, and D. Benevides da Costa, "Latent diffusion for spectrum sensing of coexisting radar and communication signals," *IEEE Wireless Communications Letters*, vol. 15, pp. 1025–1029, Dec. 2025.



Ngoc-Hoang Mai is currently pursuing a B.S. degree in Embedded System and IoT Technology with the Department of Electronics and Information Engineering, Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam. His current research interests include digital image processing, radio signal processing, computer vision, machine learning, and deep learning.



Tan Do-Duy received the B.S. degree in Electronics and Telecommunications from the Ho Chi Minh City University of Technology (BK-HCM), Vietnam, in 2010, the M.S. degree in Wireless Communications from the Kumoh National Institute of Technology, South Korea, in 2013, and the Ph.D. degree in Electronic and Telecommunication Engineering from the Autonomous University of Barcelona, Spain, in 2019. He is currently with the Department of Electronics and Information Engineering, HCMC University of Technology and Engineering (HCM-UTE), Vietnam, as an associate professor. His main research interests include resource allocation optimization for wireless networks and advanced coding for reliable systems. He has served as a TPC Member for many IEEE conferences, such as GLOBECOM, ICC, and WCNC. He is also currently serving as a regular reviewer for many IEEE journals and international conferences.



Hai-Trang Phuoc Dang is a lecturer at Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam, since 2007. He received his B.Eng and M.Eng. degrees in Electronic Engineering from HCMC University of Technology and Education in 2007 and 2011. His research focuses on Internet of Things, signal processing, wireless communications, and artificial intelligence.



Hoc Phan received the B.S. degree in electronics and computer engineering from Da Nang University of Technology, Da Nang, Vietnam, in 2001; the M.S. degrees in geographic information systems and electrical and electronics engineering from Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2005 and 2006, respectively; and the Ph.D. degree in telecommunication systems from Blekinge Institute of Technology, Karlskrona, Sweden, in March 2013. He is currently a Lecturer of Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam. His current research interests include cooperative communications, relay networks, cognitive radio networks, and network coding.



Thien Huynh-The (IEEE Senior Member) received the Ph.D. degree in Computer Science and Engineering from Kyung Hee University (KHU), South Korea, in 2018. He was a recipient of the Superior Thesis Prize awarded by KHU. From March 2018 to August 2018, he was a Postdoctoral Researcher with Ubiquitous Computing Laboratory, KHU. From September 2018 to May 2022, he was a Postdoctoral Researcher with ICT Convergence Research Center, Kumoh National Institute of Technology, South Korea. He is currently a Lecturer of Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam. He was a recipient of Golden Globe Award 2020 for Vietnamese Young Scientist and IEEE ATC Best Paper Award in 2023, 2024, and 2025. His research interests include digital image processing, radio signal processing, computer vision, wireless communications, IoT applications, machine learning, and DL. He is currently serving as an Editor of IEEE COMML.