Regular Article

CoDeX-Net: A Coordinate-Gated and Deformable Cross-Scale Network for Efficient Waveform Classification in Integrated Radar-Communication Systems

Thanh-Dat Tran, Minh-Thanh Le, Son Ngoc Truong, Ngoc-Ha Truong, Pham Ngoc Son, Hoc Phan, Tan Do-Duy, Thien Huynh-The

Department of Electronics and Information Engineering
HCM City University of Technology and Engineering (HCM-UTE)

Correspondence: Thien Huynh-The, thienht@hcmute.edu.vn
Communication: received 18 Novemver 2025, revised 18 January 2026, accepted 08 February 2026
Online publication: 03 March 2026, Digital Object Identifier: 10.21553/rev-jec.428

Abstract– **Accurate waveform classification in hybrid radar-communication environments remains challenging, particularly under low-SNR conditions where structured interference and noise severely distort time-frequency signatures. Existing lightweight CNN models often lack the capacity to capture axis-dependent patterns, whereas conventional vision backbones are computationally prohibitive for edge-level deployment and fail to exploit the physical structure of spectrogram data. To address these limitations, this work proposes CoDeX-Net, a compact dual-stream architecture composed of two complementary modules: (i) PAUG, which performs intra-resolution refinement through coordinate-aware spatial gating and channel-adaptive modulation, and (ii) CSDM, a cross-scale deformable mixer that aggregates coarse-resolution context via offset-guided sampling and soft branch routing. Together, these modules enhance both local discriminability and long-range spectral coherence while maintaining extremely low complexity. Extensive experiments on twelve radar and communication waveforms demonstrate that CoDeX-Net achieves 91.01% average accuracy, outperforming state-of-the-art CNN and lightweight radio frequency (RF) classifiers despite operating with only 51K parameters and 0.564 ms inference latency. The results confirm that task-aligned architectural design provides substantive benefits over repurposed vision models and enables practical deployment in real-time embedded RF systems.**

Keywords– **Deep learning, integrated radar-communications systems, time-frequency analysis, waveform classification.**

1 Introduction

The explosive proliferation of next-generation wireless communication systems (fifth-generation/sixth-generation) and the ubiquity of advanced sensor networks have precipitated an increasingly severe scarcity and congestion of the electromagnetic spectrum [1]. The spectral domain, a finite natural resource, has evolved into a highly contested environment populated by a myriad of heterogeneous transmitting devices. This challenge is further exacerbated by the emergence of integrated radar-communication (IRC) systems [2, 3]. The IRC paradigm aims to dissolve traditional frequency boundaries, enabling radar and communication signals to coexist and share the same physical bandwidth to maximize spectral efficiency [4, 5]. This convergence creates a highly complex, hybrid signal environment where receivers must contend not only with background noise but also distinguish between target signals and diverse structured interferers. Consequently, a robust waveform recognition mechanism capable of accurately classifying both signal families under severe noise conditions serves as a critical prerequisite for enabling subsequent signal processing stages, such as interference rejection or adaptive demodulation [6, 7].

Traditionally, waveform recognition has been approached through two primary paradigms: Likelihood-Based (LB) methods and Feature-Based (FB) methods [8–10]. While LB approaches, such as the Average Likelihood Ratio Test, achieve theoretical optimality in the Bayesian sense, they suffer from computational intractability and a heavy reliance on precise prior knowledge of channel parameters—a requirement that is rarely met in blind reconnaissance scenarios. Conversely, FB methods depend on hand-crafted statistical signatures, such as higher-order cumulants, spectral moments, or cyclostationary features [11]. However, these rigid feature extractors often prove brittle, exhibiting severe performance degradation when confronted with non-Gaussian noise or complex multipath fading characteristic of urban or hybrid IRC environments.

The advent of deep learning (DL) has catalyzed a fundamental paradigm shift, replacing laborious manual feature engineering with end-to-end representation learning [12], [13]. Specifically, convolutional neural networks (CNNs) have emerged as the dominant architecture due to their powerful feature extraction capabilities [14]. By treating Time-Frequency Representations (TFR)–such as the Short-Time Fourier Transform [15] or Choi-Williams Distribution [16]–as visual images, CNNs leverage the translation invariance and locality of convolutional kernels to automatically decompose complex signals into hierarchical abstractions. This process enables the network to learn features

ranging from elementary edges and pulses in shallow layers to abstract modulation morphologies in deeper layers, thereby eliminating the need for domain expertise while delivering superior robustness compared to classical statistical classifiers [17].

Initial efforts in DL-based automatic modulation classification primarily applied general vision backbones. Architectures such as ResNet50 [18], MobileNetV2 [19], and EfficientNetB0 [20] were widely used as benchmarks. Although these models possessed strong feature extraction capabilities, they were designed for natural images and often failed to capture the specific time-frequency signatures of radio signals [21]. Their simple isotropic kernels processed all spatial dimensions equally, ignoring the distinct physical meanings of the time and frequency axes in the spectrogram. Furthermore, the massive parameter count rendered them unsuitable for edge deployment [22]. To address this issue, domain-specific architectures with lightweight computation and attention mechanisms have been actively explored [23]. For instance, Huynh-The *el al.* [24] introduced a sophisticated CNN design leveraging the Smoothed Pseudo Wigner-Ville Distribution (SPWVD) to mitigate cross-term interference. The model employed the Rational Spectrum Association module to extract discriminative features. However, RadComNet relied on standard convolutions and dense connections, leading to high memory consumption and computational redundancy. To address the efficiency bottleneck, WaveNet [25] was subsequently proposed as a lighter alternative. WaveNet introduced the cost-efficient feature awareness module, utilizing grouped-of-kernel-wise residual connections and Dual Asymmetric Channel Attention. By decoupling attention along the time and frequency axes, WaveNet achieved a reduction in model size compared to RadComNet while improving accuracy. However, both RadComNet and WaveNet were purely CNN-based models. They suffered from the intrinsic limitation of the convolution operator: a local and fixed receptive field. This prevented them from effectively modeling long-range dependencies. Recently, the frontier of this field shifted toward architectures based on the Self-Attention mechanism, notably the Vision Transformer [26] and hierarchical variants such as the Swin Transformer [27]. These modern models demonstrated that modeling global dependency via the Attention mechanism yielded superior accuracy compared to traditional CNNs, particularly in distinguishing signals with extended temporal structures. However, the standard Self-Attention mechanism incurred a computational complexity that scaled quadratically with the input sequence length ($\mathcal{O}(N^2)$). This resulted in unacceptable inference latency for electronic warfare or real-time radar applications. Nevertheless, the success of Transformers in computer vision inspired hybrid architectures aimed at mitigating this drawback. Tran *el al.* [28] proposed a core innovation with the Decoupled Attention (DTA) mechanism, which split attention into temporal, frequency, and channel streams to reduce the quadratic complexity of standard Self-Attention. Despite its high accuracy (91.11%), it still imposed a significant computational burden compared to pure CNNs. Another direction was CMNet [29], which leveraged State Space Models (Mamba) to achieve ultra-fast inference speeds (0.059 ms) thanks to linear complexity. However, the accuracy of CMNet (90.7%) was slightly lower than that of Transformer-based hybrid models, suggesting that although SSMs were efficient, they struggled to filter heavy noise as effectively as attention-based mechanisms in low signal-to- noise ratio (SNR) regimes.

It is evident that none of the aforementioned architectures possess the capability to dynamically align features across different scales. The absence of such mechanisms limits the ability to extract robust features in noisy environments, varying frequencies, or complex waveform structures. To effectively address these challenges and bridge the gap between performance and efficiency, this paper proposes CoDeX-Net, a compact yet robust time-frequency classifier. Our architecture is built upon the synergistic interaction of two complementary mechanisms: intra-resolution refinement via coordinate gating and cross-scale context aggregation via deformable sampling. This hybrid design allows the model to flexibly adapt to complex signal variations and dynamic noise without incurring unnecessary computational costs. The main contributions of this study are summarized as follows:

- We propose CoDeX-Net, a lightweight architecture for waveform classification in integrated radar-communication systems. By leveraging SPWVD to obtain high-resolution, interference-mitigated time-frequency representations, the model achieves superior trade-off between accuracy and efficiency.
- We develop two novel complementary modules: the Pan-Axis Unified Gating (PAUG) module for intra-resolution refinement via dynamic routing and coordinate attention; and the Cross-Scale Deformable Mixer (CSDM) module for cross-resolution context aggregation via content-adaptive deformable sampling. This combination enables the network to effectively capture both fine-grained local details and global signal structures.
- Comprehensive experiments on a dataset comprising 12 waveforms demonstrate the superior performance of CoDeX-Net. The model achieves an average accuracy of 91.01% and impressive noise robustness of 66.66% at -5 dB SNR , outperforming heavy vision backbones and specialized lightweight models such as CMNet and DTANet, while maintaining a compact size of 51K parameters and a low inference latency of 0.564 ms.

The rest of this paper is structured as follows. Section 2 details the proposed CoDeX-Net architecture and its core components. Section 3 presents the experimental setup, comprehensive results analysis, and a comparative study against state-of-the-art methods. Finally, Section 4 concludes the paper and outlines future research directions.

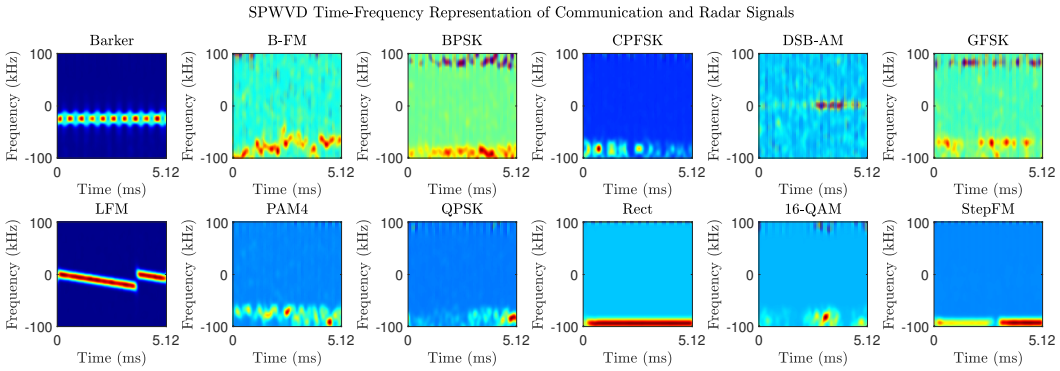


Figure 1. Representative SPWVD spectrograms illustrating the distinct time-frequency signatures of the 12 considered radar-communication waveform categories.

2 Methodology

2.1 Signal Modeling and TFR via SPWVD

We consider an IRC receiver tasked with identifying radar and communication signals. The received discrete-time complex signal $y[k]$ is modeled as the convolution of the transmitted waveform $x[k]$ with an unknown channel response $h[k]$, corrupted by additive white Gaussian noise (AWGN) $n[k]$

$$y[k] = x[k] * h[k] + n[k], \tag{1}$$

where k represents the time index and $*$ denotes the convolution operator. Our objective is to classify twelve distinct waveform types, comprising four radar signatures (LFM, Rect, Barker, StepFM) and eight communication modulations (B-FM, BPSK, CPFSK, DSB-AM, GFSK, PAM4, QPSK, 16-QAM), without prior channel knowledge.

To facilitate CNN-based classification, we utilize TFRs. The Wigner-Ville Distribution (WVD) is chosen for its high resolution, defined for a signal $s(t)$ as

$$\mathrm{WVD}_s(t,\omega) = \int_{-\infty}^{\infty} s\left(t+\frac{\tau}{2}\right) s^*\left(t-\frac{\tau}{2}\right) e^{-j\omega\tau} d\tau. \tag{2}$$

However, for multi-component received signals $y(t) = s(t) + n(t)$, WVD suffers from cross-term interference

$$\begin{aligned}\mathrm{WVD}_y(t,\omega) = \mathrm{WVD}_s(t,\omega) + \mathrm{WVD}_n(t,\omega) \\ + 2\mathcal{R}\left[\mathrm{WVD}_{s,n}(t,\omega)\right],\end{aligned} \tag{3}$$

where the cross-term $2\mathcal{R}\left[\mathrm{WVD}_{s,n}\right]$ can obscure genuine signal features. To mitigate this, we employ the SPWVD, as analyzed in recent deep learning frameworks [30], which suppresses interference via independent temporal ($g(t)$) and spectral ($H(\omega)$) smoothing windows

$$\begin{aligned}\mathrm{SPWVD}_x(t,\omega;\Phi) = \int_{-\infty}^{\infty} g(t)\, H(\omega)\, x\left(t+\frac{\tau}{2}\right) \\ x^*\left(t-\frac{\tau}{2}\right) e^{-j\omega\tau} d\tau.\end{aligned} \tag{4}$$

For discrete implementation with sequence length N, this is computed as

$$\begin{aligned}\mathrm{SPWVD}[n,k] = \sum_{m=-N}^{N} g[n]\, H[k]\, x\left[n+\frac{m}{2}\right] \\ \times x^*\left[n-\frac{m}{2}\right] e^{\frac{-j2\pi km}{N}}.\end{aligned} \tag{5}$$

We utilize Kaiser windows for both smoothing domains to balance resolution and sidelobe suppression. The coefficients are empirically set following [25] and are given by

$$w[n] = \frac{\mathrm{I}_0\left(\gamma\sqrt{1-\left(\frac{n-N/2}{N/2}\right)^2}\right)}{\mathrm{I}_0(\gamma)},\ 0 \le n \le N, \tag{6}$$

where $\mathrm{I}_0(\cdot)$ is the zeroth-order modified Bessel function of the first kind, and γ is the shape parameter controlling spectral leakage. In this study, we generate square SPWVD images of size 224×224 by aligning the number of frequency bins with the time samples. As illustrated in Figure 1, using Kaiser windows configured with $N = 40$ and $\gamma = 0.5$ provides robust interference suppression while preserving essential local time-frequency details for the twelve waveform classes. This selection of input representation is a deliberate design choice. Although direct learning from raw I/Q data is feasible, it mandates that the network implicitly approximate time-frequency transformations, thereby imposing significant computational overhead. By integrating SPWVD with Kaiser windows, the proposed framework incorporates a strong inductive bias. The deterministic suppression of cross-term interference functions as a mathematical prior rather than heuristic tuning, yielding a disentangled feature manifold. This physics-aware decoupling effectively offloads the signal transformation task, enabling CoDeX-Net to maintain minimal parameter complexity while achieving high classification accuracy.

2.2 CoDeX-Net: Coordinate-Gated and Deformable Cross-Scale Network for Waveform Classification

In this paper, we introduce CoDeX-Net, a novel hybrid neural network architecture designed to address the unique challenges of waveform classification in IRC systems. Departing from standard computer vision backbones that treat inputs as spatially isotropic images, CoDeX-Net adopts a physics-aware design to explicitly leverage the anisotropic information inherent in time-frequency domains. The architecture optimizes feature learning via the synergy of two core mechanisms, as illustrated in Figure 2, the PAUG for dynamic intra-resolution refinement that mimics adaptive signal

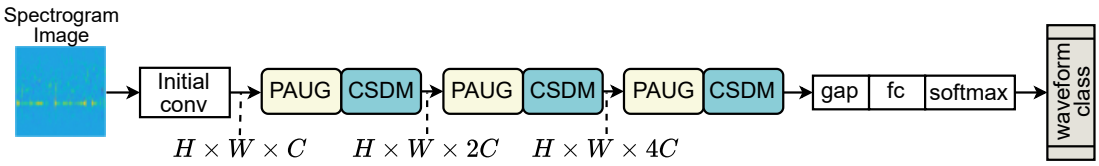


Figure 2. The overall architecture of the proposed CoDeX-Net. The network processes the input spectrogram through an initial stem, followed by three cascaded stages. Each stage strategically alternates between a PAUG module for intra-resolution refinement and a CSDM module for cross-resolution context aggregation, culminating in a compact classification head.

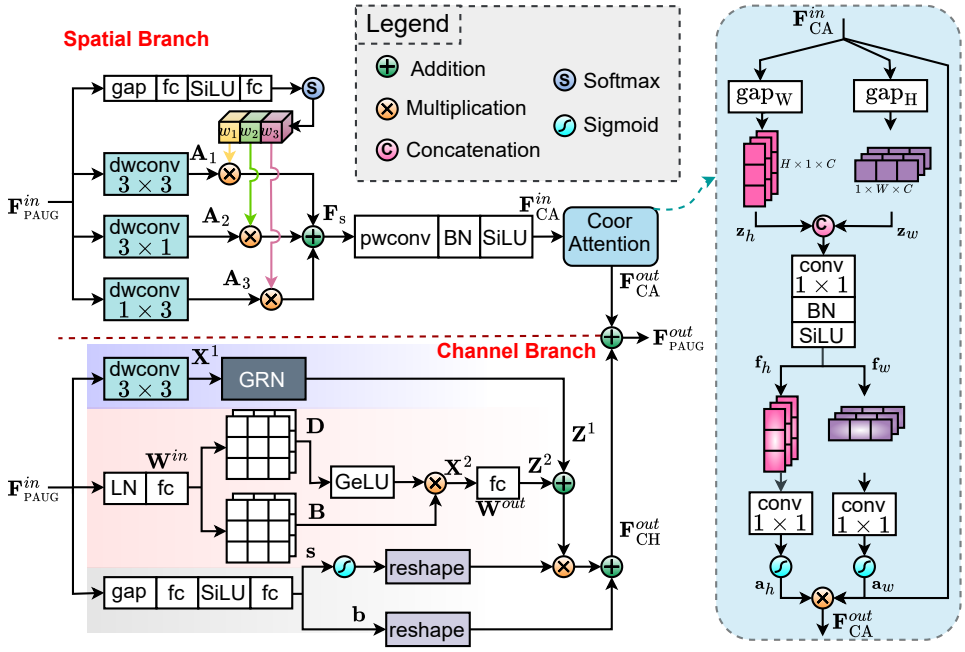


Figure 3. Detailed architecture of the proposed PAUG module.

processing windows, and the CSDM for deformable cross-scale mixing. This design enables the model to simultaneously capture local details, axial long-range relationships, and dynamic multi-scale context, thereby overcoming the geometric rigidity of generic CNNs.

The backbone begins with an efficient stem for aggressive spatial downsampling. It stacks two 3×3 convolutional (conv) layers, each with stride 2, yielding a $4\times$ reduction in spatial resolution. Each convolution is followed by Batch Normalization (BN) and a Sigmoid Linear Unit (SiLU) activation to stabilize training and provide robust nonlinearity. This early reduction substantially lowers the compute of all subsequent blocks while preserving essential edge and texture cues.

PAUG module: This is a sophisticated intra-resolution feature refiner. Its primary function is to process and enhance feature maps within a single scale. It operates by executing two parallel processing streams–one spatial and one channel–which are subsequently fused intelligently, as detailed in Figure 3. The process begins with the input feature map $\mathbf{F}_{\text{PAUG}}^{in} \in \mathbb{R}^{H\times W\times C}$, obtained from the initial conv block. The extracted features are then forwarded into two parallel branches. The first branch is dedicated to spatial feature refinement. This process commences by applying three parallel depthwise convs (dwconv) to the input $\mathbf{F}_{\text{PAUG}}^{in}$, chosen to capture diverse, multi-orientation spatial structures

$$\mathbf{A}_i = \begin{cases} \mathcal{DW}_{3\times3}\left(\mathbf{F}_{\text{PAUG}}^{in}\right), & \text{if } i=1 \\ \mathcal{DW}_{3\times1}\left(\mathbf{F}_{\text{PAUG}}^{in}\right), & \text{if } i=2 \\ \mathcal{DW}_{1\times3}\left(\mathbf{F}_{\text{PAUG}}^{in}\right), & \text{if } i=3, \end{cases} \tag{7}$$

where $\mathcal{DW}$ represents an $m \times n$ depthwise convolution. This operator is utilized for its high parameter efficiency and significantly lower computational cost relative to a standard conv layer. To dynamically synthesize these features, we introduce a Softmax Router. This router is a lightweight MLP designed to first extract a global context vector from the input $\mathbf{F}_{\text{PAUG}}^{in}$ via global average pooling (gap). This vector is subsequently processed by an MLP and a softmax function to predict three dynamic weights. The entire weight computation process is summarized as

$$\mathbf{w} = \mathcal{S}\left(\text{fc}\left(\text{SiLU}\left(\text{fc}\left(\text{gap}\left(\mathbf{F}_{\text{PAUG}}^{in}\right)\right)\right)\right)\right), \tag{8}$$

where $\mathcal{S}$ denotes the softmax function applied along the feature dimension, ensuring the weights sum to unity, and fc represents a fully connected layer. This mechanism allows the model to dynamically learn to select and balance the kernel combination–prioritizing horizontal, vertical, or symmetric features–that is most suitable for each input sample, rather than using a static aggregation. The spatial feature map, $\mathbf{F}_s$, is then computed as a weighted summation of the

parallel branches, where each weight $\mathbf{w}_i \in \mathbf{w}$ is spatially broadcast

$$\mathbf{F}_s = \mathbf{w}_1\mathbf{A}_1 + \mathbf{w}_2\mathbf{A}_2 + \mathbf{w}_3\mathbf{A}_3. \quad (9)$$

By utilizing dynamic weights derived from the global context, the network can dynamically prioritize and balance the combination of spatial features (e.g., horizontal, vertical, or symmetric) that is most suitable for each input sample, creating a more flexible receptive field compared to static aggregation. Subsequently, the feature map $\mathbf{F}_s$ is passed through a pointwise conv (pwconv), BN, and SiLU sequence. This processing block serves two purposes: it performs channel fusion to mix information and learn cross-channel correlations (which were kept separate by the preceding dwconvs), and it applies a non-linear transformation to refine the features before they are fed into the Coord Attention block. Standard convolutions often struggle to capture long-range, position-sensitive dependencies. To explicitly address this limitation, the Coord Attention module is employed to capture these relationships. Instead of computing an computationally expensive full 2D attention map, Coord Attention factorizes spatial attention into two 1D vectors that independently encode information along the two spatial axes. This process commences by applying 1D gap along the horizontal and vertical dimensions to compress the input $\mathbf{F}_{\text{CA}}^{in}$ into two separate axis descriptors: a vertical-axis descriptor $\mathbf{z}_h \in \mathbb{R}^{H\times 1\times C}$ and a horizontal-axis descriptor $\mathbf{z}_w \in \mathbb{R}^{1\times W\times C}$. These two vectors, which now embed position information along their respective axes, are then concatenated and passed through a shared encoding block. This step is crucial, as it enables the model to learn interactions between the positional information of the two axes, producing a unified context representation $\mathbf{f}$

$$\mathbf{f} = \text{SiLU}\left(\text{BN}\left(\mathcal{C}_{1\times 1}\left(\left[\mathbf{z}_h, \mathbf{z}_w^{\top}\right]\right)\right)\right), \quad (10)$$

where $[\cdot]$ denotes the concatenation operation, $\mathbf{z}_w^{\top}$ indicates that $\mathbf{z}_w$ is permuted to match the spatial dimensions of $\mathbf{z}_h$ prior to concatenation, and $\mathcal{C}$ denotes a standard conv operator. The tensor $\mathbf{f}$ is then split back into its horizontal and vertical components, $\mathbf{f}_h$ and $\mathbf{f}_w$. These components are passed through two separate 1×1 conv layers and a sigmoid function (σ) to reconstruct the final attention masks

$$\mathbf{a}_h = \sigma\left(\mathcal{C}_{1\times 1}\left(\mathbf{f}_h\right)\right), \qquad \mathbf{a}_w = \sigma\left(\mathcal{C}_{1\times 1}(\mathbf{f}_w)^{\top}\right). \quad (11)$$

The final output of the spatial branch, $\mathbf{F}_{\text{CA}}^{out}$, is obtained by applying both attention masks element-wise to the feature map $\mathbf{F}_{\text{CA}}^{in}$, expressed as

$$\mathbf{F}_{\text{CA}}^{out} = \mathbf{F}_{\text{CA}}^{in} \odot \mathbf{a}_h \odot \mathbf{a}_w, \quad (12)$$

where $\odot$ denotes element-wise multiplication. The effect of this mechanism is that $\mathbf{a}_h$ functions as a vertical gating mechanism, re-weighting the importance of each row. Similarly, $\mathbf{a}_w$ re-weights the importance of each column. Consequently, a feature at a given location (h, w) is strongly emphasized only if both its corresponding row h and its column w are deemed contextually significant by the model, thereby achieving a sophisticated, global-aware feature refinement with minimal computational overhead.

After completing the spatial branch, the PAUG module employs a channel refinement branch that focuses on content-dependent modulation along the channel dimension. While the spatial branch models axis-aligned spatial relationships in the 2D plane, the channel branch is designed to enhance semantic discrimination and stabilize per-channel responses. It captures complex nonlinear interactions across channels while leveraging global contextual information to modulate its output adaptively. This process consists of three sub-mechanisms that operate in parallel and are later merged. To first collect local spatial context, the input feature map $\mathbf{F}_{\text{PAUG}}^{in}$ is passed through a dwconv 3×3 convolution, producing an intermediate output $\mathbf{X}^1$. Immediately afterward, we apply Global Response Normalization (GRN) [31]. GRN is an advanced normalization technique that enhances contrast and feature selectivity by normalizing each channel's response relative to its own global spatial energy. This creates a competitive dynamic among channels by normalizing the feature map $\mathbf{X}^1$ according to its global response magnitude $\mathbf{g}$. As a result, weak channels are amplified while overly dominant channels are suppressed. The entire operation is expressed as a gated residual update, summarized by

$$\mathbf{Z}^1 = \mathbf{X}^1 + \gamma\hat{\mathbf{X}}^1 + \beta, \quad (13)$$

with

$$\hat{\mathbf{X}}^1 = \frac{\mathbf{X}^1}{\mathbf{g}} = \frac{\mathbf{X}^1}{\sqrt{\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{X}_{i,j}^1)^2 + \epsilon}}, \quad (14)$$

where ϵ is a small constant added for numerical stability, and $\gamma, \beta \in \mathbb{R}^{1\times 1\times C}$ are learnable per-channel scale and bias parameters. This stage diffuses local information within each channel while simultaneously regulating activation magnitudes, which is crucial before the subsequent gating mechanisms that can further amplify responses. In parallel, the channel branch employs a pixel-wise MLP utilizing the GE-GLU (GELU-Gated Linear Unit) mechanism [32] to model complex nonlinear inter-channel interactions. This is a variant of the GLU, where the GELU (Gaussian Error Linear Unit) activation is used as the non-linear gating function. The process first normalizes the input using layer norm (LN), followed by a linear projection ($\mathbf{W}^{in}$) that is decomposed into two components: $\mathbf{D}$ (the data component) and $\mathbf{B}$ (the gating component). The intermediate feature $\mathbf{X}^2$ is then computed by the element-wise multiplication of the GELU-activated data component and the gating component

$$\mathbf{X}^2 = \text{GeLU}\left(\mathbf{D}\right) \odot \mathbf{B}, \quad \mathbf{D}, \mathbf{B} = \text{LN}\left(\mathbf{F}_{\text{PAUG}}^{in}\right)\mathbf{W}^{in}. \quad (15)$$

The resulting feature $\mathbf{X}^2$ is then projected back to the original channel dimension via $\mathbf{W}^{out}$ to obtain the final output $\mathbf{Z}^2$. This mechanism enables the model to learn complex inter-channel relationships with high computational efficiency, as the entire transformation is applied independently at each spatial location. After the

two main feature branches ($\mathbf{Z}^1$ and $\mathbf{Z}^2$) are computed, they are summed together. However, rather than utilizing this result directly, we apply a final dynamic modulation mechanism.This mechanism, known as FiLM (Feature-wise Linear Modulation) [33], functions as a control branch. It utilizes the global context of the image to refine the output of the Channel Branch. First, a global context vector $\mathbf{v}$ is extracted from the input $\mathbf{F}^{in}_{\text{PAUG}}$ via gap. This vector is then passed through a 2-layer MLP (with SiLU) to predict two control vectors: a scale vector $\mathbf{s}$ and a bias vector $\mathbf{b}$, which are applied as follows

$$\mathbf{s}, \mathbf{b} = \text{fc}\left(\text{SiLU}\left(\text{fc}\left(\text{gap}\left(\mathbf{F}^{in}_{\text{PAUG}}\right)\right)\right)\right). \tag{16}$$

To ensure stability, the scale vector $\mathbf{s}$ is normalized using the sigmoid function. A critical reshape step is then performed to enable the tensor broadcasting mechanism. This ensures that the same scale value s_c and bias value b_c are applied to all pixels (H, W) within the same corresponding channel c. These vectors $\mathbf{s}$ and $\mathbf{b}$ are not features themselves, but rather supplementary coefficients used to modulate the fused result of the other two branches. The final output of the entire Channel Branch, $\mathbf{F}^{out}_{\text{CH}}$, is the result of this FiLM modulation

$$\mathbf{F}^{out}_{\text{CH}} = (\mathbf{Z}^1 + \mathbf{Z}^2) \odot \sigma(\mathbf{s}) + \mathbf{b}. \tag{17}$$

This mechanism allows the global context of the image (via $\mathbf{s}$ and $\mathbf{b}$) to determine how to refine (e.g., amplify or suppress) the mixed channel features. Finally, the PAUG module completes its task with a fusion step, merging the feature outputs from the Spatial Branch ($\mathbf{F}^{out}_{\text{CA}}$) and the Channel Branch ($\mathbf{F}^{out}_{\text{CH}}$) to create a unified feature map

$$\mathbf{F}^{out}_{\text{PAUG}} = \mathbf{F}^{out}_{\text{CA}} + \mathbf{F}^{out}_{\text{CH}}. \tag{18}$$

This feature map now combining both dynamically routed spatial information and contextually modulated channel information, completes the intra-resolution feature refinement process of the PAUG module.

CSDM module: While the PAUG module focuses on intra-resolution refinement, the CSDM is designed to flexibly integrate cross-resolution context in a content-adaptive manner. Instead of operating at a single spatial resolution, CSDM constructs three feature maps that share the same channel dimensionality but differ in spatial scale, and subsequently learns deformable sampling operators that propagate information from coarse resolutions back to the full-resolution grid. The overall architecture of this module is illustrated in Figure 4. Given the input feature map $\mathbf{F}^{in}_{\text{CSDM}} \in \mathbb{R}^{H\times W\times C}$, the module generates three spatial scales

$$\begin{cases} full-resolution : \mathbf{X}_0 = \mathbf{F}^{in}_{\text{CSDM}} \in \mathbb{R}^{H\times W\times C} \\ half-resolution : \mathbf{X}_1 = \mathcal{D}\left\langle \mathbf{F}^{in}_{\text{CSDM}} \right\rangle \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C} \\ quater-resolution : \mathbf{X}_2 = \mathcal{D}\left\langle \mathbf{F}^{in}_{\text{CSDM}} \right\rangle \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}, \end{cases} \tag{19}$$

where $\mathcal{D}\langle\cdot\rangle$ denotes the adaptive average pooling downsampling operator. Each level is then lightly refined by a DS block (dwconv + pwconv + BN + SiLU) to yield $\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2$. Among them, $\mathbf{P}_0$ preserves the full spatial resolution and serves as the Query feature map used to determine how information should be gathered from the coarse scales. To establish a geometric correspondence between the full-resolution grid and the coarse scales, CSDM constructs two fixed base grids, $\mathbf{G}_1$ and $\mathbf{G}_2$. Each base grid $\mathbf{G}_k(h, w) \in [-1, 1]^2$ maps the position (h, w) on the full-resolution feature map $\mathbf{P}_0$ to its normalized coordinates on the corresponding coarse feature map $\mathbf{P}_k$, for $k \in \{1, 2\}$. These grids are non-learnable and encode the default sampling coordinates that would be used if no deformation were applied. They form the geometric foundation required for stable deformable warping. From the full-resolution query feature $\mathbf{Q} = \mathcal{C}_{1\times1}(\mathbf{P}_0)$, two prediction heads (one for the half-resolution and one for the quarter-resolution branch) jointly estimate offsets and attention weights. A lightweight convolutional head processes $\mathbf{Q}$ and produces

$$\Delta\mathbf{o}_k, \bar{\mathbf{w}}_k = \phi_k(\mathbf{Q}), \tag{20}$$

where $\phi_k(\cdot)$ denotes the lightweight convolutional prediction head associated with scale k, $\Delta\mathbf{o}_k \in \mathbb{R}^{H\times W\times K\times 2}$ contains K 2-D offsets per spatial position, and $\bar{\mathbf{w}}_k \in \mathbb{R}^{H\times W\times K}$ holds attention weights normalized by a softmax over the K samples. The raw offsets are then normalized through a Δ-Norm step to constrain their magnitude within a predefined radius r_{max}. This normalization is applied independently to the horizontal offset component Δx_k and the vertical offset component Δx_k, each predicted at every spatial location (h, w). The normalized offsets are computed as

$$\Delta x_k = \tanh\left(\Delta x_k^{raw}\right) r_{max}, \quad \Delta y_k = \tanh\left(\Delta y_k^{raw}\right) r_{max}, \tag{21}$$

where Δx_k^{raw} and Δy_k^{raw} denote the unbounded offset predictions produced by the convolutional head. This bounded mapping prevents excessively large spatial shifts that could destabilize the subsequent deformable sampling operation.

During deformable warping, the module simultaneously consumes two streams of data. The first stream consists of the coarse feature map $\mathbf{P}_k$, which is reshaped or replicated along the sample dimension to produce $\mathbf{P}_k^{re}$; this tensor serves as the source data from which the warped features will be drawn. The second stream is the sampling grid, constructed by adding the normalized offsets to the base grid and reshaping the result along the K dimension to match the form expected by the sampling operator. For each spatial position (h, w) and each sample $m \in \{1, ..., K\}$, the actual sampling coordinate on the coarse map is given by

$$\mathbf{G}_k^{(m)}(h, w) = \mathbf{G}_k(h, w) + \Delta\mathbf{o}_k(h, w, m), \tag{22}$$

where bilinear sampling is performed to extract

$$\mathbf{S}_k^{(m)}(h, w, :) = \text{Sample}\left(\mathbf{P}_k^{re}, \mathbf{G}_k^{(m)}(h, w)\right), \tag{23}$$

This mechanism enables content-dependent, spatially adaptive warping, allowing each full-resolution position to gather non-local information from appropriate coarse-scale neighborhoods. Instead of relying on a single sampled point, the CSDM employs a Mix-K aggregation step to merge the K warped

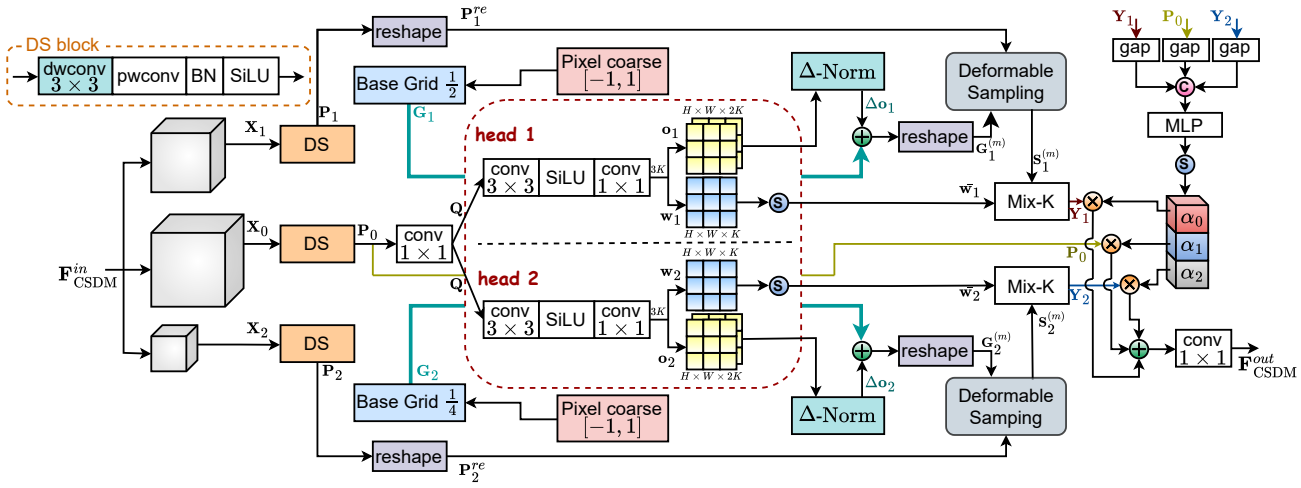


Figure 4. Detailed architecture of the proposed CSDM module.

Table I
Performance benchmark of CoDeX-Net and method comparison

Waveform configurations

Type	Param.	Range of value	Type	Param.	Range of value
LFM	N	1024	Barker	N	1024
	N_{pw}	$[512, 1920]$		f_c	$[f_s/6, f_s/5]$
	B	$[f_s/20, f_s/16]$		c_w	$(\lfloor k \cdot f_s/f_c \rfloor - 1)/f_s, k \in \{1,5\}$
	f_c	$[f_s/6, f_s/5]$		c_{pp}	$\{3,4,5,7,11\}$
	PRF	f_s/N_{pw}		PRF	$f_s/((c_w \cdot f_s) \cdot c_{pp} + 1)$
	Sweep Dir.	{Up, Down}			
Rect	N	1024	StepFM	N	1024
	N_{pw}	$[512, 1920]$		N_{cc}	$[512, 1024]$
	p_w	N_{pw}/f_s		N_{steps}	$\{3,4,5,7,11\}$
	f_c	$[f_s/6, f_s/5]$		f_{step}	$k \times$ PRF, $k \in \{1,2,3,4,5\}$
	PRF	f_s/N_{pw}		PRF	$f_s/(N_{cc} + \Delta N)$, $\Delta N \in [50, 100]$

Notations

Param.	Description
N	Total number of samples
N_{pw}	Samples for pulse width
B	Bandwidth
f_c	Center frequency
f_s	Sampling frequency
p_w	Pulse width
c_w	Chip width
c_{pp}	Cycles per phase code
PRF	Pulse repetition frequency
N_{cc}	Samples per pulse
N_{steps}	Number of frequency steps
f_{step}	Frequency step size
ΔN	Random sample offset for PRF

Table II
Dataset information of 12 waveform types

Category	Waveform type	Abbr	No. signals
Radar	Rectangular	Rect	384,000 signals per type
	Linear frequency modulation	LFM	
	Barker Code	Barker	
	Step frequency modulation	StepFM	
Communication	16-Quadrature amplitude modulation	16-QAM	
	Binary phase-shift keying	BPSK	
	Quadrature phase-shift keying	QPSK	
	Pulse amplitude modulation 4-level	PAM4	
	Gaussian frequency shift keying	GFSK	
	Continuous phase frequency shift keying	CPFSK	
	Broadcast frequency modulation	B-FM	
	Double sideband amplitude modulation	DSB-AM	

samples according to the learned attention weights

$$\mathbf{Y}_k(h,w,:) = \sum_{m=1}^{K} \bar{\mathbf{w}}_k(h,w,m)\, \mathbf{S}_k^{(m)}(h,w,:), \qquad (24)$$

where the weights satisfy $\sum_{m=1}^{K} \bar{\mathbf{w}}_k(h,w,m) = 1$. This produces two full-resolution representations $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}^{H\times W\times C}$, each carrying context propagated from the half and quarter-resolution features, respectively, but aligned to the full-resolution lattice. To complete the fusion, CSDM aggregates the three pathways $\mathbf{P}_0, \mathbf{Y}_1, \mathbf{Y}_2$ via a global softmax router. The global average pooling vectors of the three branches are concatenated and passed through a small MLP to generate mixing coefficients $\alpha_0, \alpha_1, \alpha_2$. The fused representation is obtained by first forming a convex combination of the three branches and then applying a 1×1 convolution

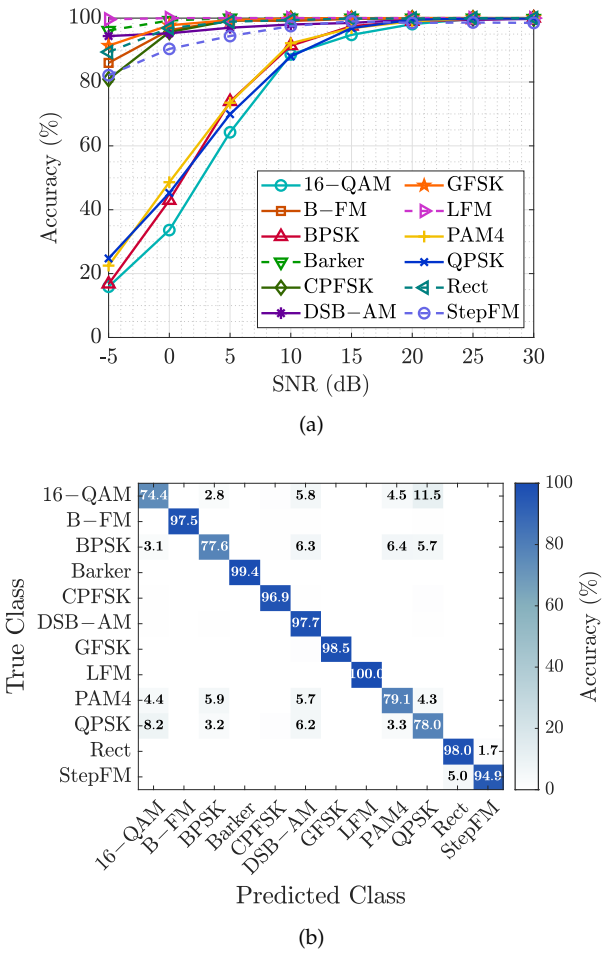
$$\mathbf{F}_{\text{CSDM}}^{out} = \mathcal{C}_{1\times1}(\alpha_0 \mathbf{P}_0 + \alpha_1 \mathbf{Y}_1 + \alpha_2 \mathbf{Y}_2), \qquad (25)$$

which adjusts the channel statistics and prepares the features for subsequent stages of the network. Overall, the CSDM module is a lightweight, content-adaptive cross-scale mixer that injects coarse-level context into full-resolution features via deformable sampling and soft routing, enriching high-resolution representations with multi-scale information at modest computational cost. At the end of the model, a compact classification head, consisting of a gap layer, a fc layer, and a final softmax operator, is employed to produce the output predictions.

3 Simulations and Results

3.1 Dataset and Training Configurations

To conduct a controlled and reproducible evaluation, we construct a large-scale synthetic dataset comprising 12 waveform categories (detailed information is reported in Table II), covering both radar and communication modalities. Based on the radar waveform



(a)

(b)

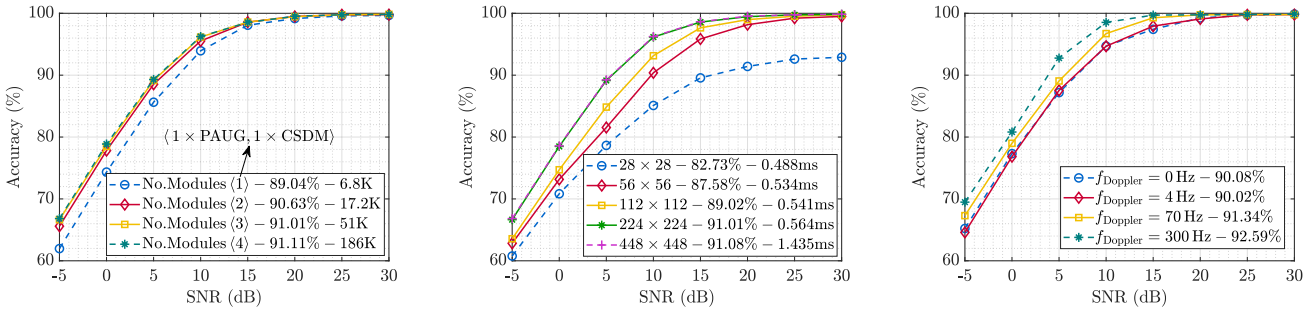
Figure 5. Performance evaluation: (a) classification accuracy versus SNR for each waveform type, and (b) confusion matrix aggregated across all SNR levels.

configurations in Table I, the radar subset includes LFM, Rect, Barker, and StepFM signals, whose defining parameters–such as pulse duration, sweep bandwidth, chip count, and sweep direction–are randomly drawn within operationally realistic bounds. The communication signals span eight representative modulation formats, including digital schemes (BPSK, QPSK, 16-QAM, PAM4), continuous-phase modulations (GFSK, CPFSK), and analog waveforms (B-FM, DSB-AM). All digitally modulated signals are pulse-shaped with a root-raised-cosine filter of roll-off 0.35. Each waveform propagates through a simulated Rician fading channel characterized by a K-factor of 4, multipath delays of $[0, 30, 150, 310, 370, 710, 1090]$ ns, and associated average path gains of $[0, -1.5, -1.4, -3.6, -0.6, -9.1, -7.0]$ dB. Mobility-induced distortions are modeled via Doppler shifts selected from $\{0, 4, 70, 300\}$ Hz, while communication signals additionally incorporate up to 5 ppm clock offset. The entire corpus spans SNR levels from -5 dB to 30 dB in steps of 5 dB, resulting in a total of 384,000 waveform instances. Each received signal is converted into a 224×224 SPWVD time-frequency representation prior to network input. CoDeX-Net is trained for 60 epochs using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 32. All experiments are performed strictly on the test partitions to ensure a fair comparison across ablation settings and backbone baselines. Training and inference are conducted on a workstation equipped with a 3.80 GHz CPU, 64 GB RAM, and an NVIDIA GeForce RTX 3060Ti GPU.

3.2 Results and Discussions

Classification robustness: In the first experiment, we present the classification accuracy for 12 waveform types across a range of SNRs, as illustrated in Figure 5(a). The results indicate that the recognition accuracy for all 12 waveform types steadily improves as the SNR increases, reaching near-perfect levels (above 99%) at high SNRs (20 dB and above). This confirms that CoDeX-Net exhibits superior feature discrimination capabilities. However, the performance at low SNRs reveals a clear distinction in signal robustness. Waveforms with unique time-frequency structures, such as LFM, DSB-AM, and Barker, demonstrate impressive resilience, maintaining accuracies above 94% even at -5 dB. Conversely, digital modulations, particularly those with complex constellations, prove more sensitive to noise, making them prone to confusion as the noise level increases. For example, at -5 dB, the accuracies for 16-QAM, QPSK, and BPSK drop sharply below 25% before rapidly recovering to near 100% as the SNR improves. The confusion matrix in Figure 5(b) provides further justification for this performance drop: confusion at low SNRs occurs primarily between modulations with similar constellation patterns (e.g., 16-QAM and QPSK; BPSK and PAM4), whereas waveforms with unique frequency sweep or coding structures (like LFM and Barker) remain clearly separable. Overall, these results demonstrate that CoDeX-Net achieves robust performance across diverse waveform families, effectively modeling both modulation-dependent features and specific time-frequency signatures. This makes the model particularly well-suited for practical radio environments where SNR conditions can vary widely.

Hyper-parameter investigation: To validate the design of CoDeX-Net and to determine the optimal configuration, in the second experiment we carried out a series of tests on the main hyperparameters: model depth, input resolution, and robustness to Doppler shift. First, Figure 6(a) illustrates the relationship between network depth, represented by the number of PAUG-CSDM block pairs. When the number of module pairs increases from 1 to 2, a clear jump in performance is observed. Specifically, the accuracy rises from 89.04% to 90.63%, corresponding to an increase of 1.59%, while the number of parameters increases from 6.8K to 17.2K. Adding a third module pair, raising the total number of parameters to 51K, further improves the performance to 91.01%, an additional increase of 0.38%, which shows that a deeper model is able to learn more complex feature representations. However, a clear diminishing return appears when a fourth module pair is added. The accuracy gain from the fourth module pair is negligible at only 0.10%, yet it incurs a disproportionate parameter cost, growing 3.6× to 186K. This



(a) (b) (c)

Figure 6. Hyperparameter sensitivity and robustness analysis of CoDeX-Net: (a) impact of network depth (number of module pairs) on accuracy and model size, (b) effect of input spectrogram resolution on recognition performance and inference latency, and (c) robustness assessment under varying Doppler shift conditions.

demonstrates that the three module pair configuration, with 91.01% accuracy and 51K parameters, achieves the optimal trade-off between performance and complexity.

Next, the sensitivity of the model to the time-frequency input resolution is analyzed in Figure 6(b). The results show that very low resolutions (28×28 or 56×56) cause severe degradation due to the loss of fine time-frequency structures, yielding only 82.73% and 87.58% accuracy, respectively. When the resolution increases to 112×112 and 224×224, the accuracy improves significantly, reaching approximately $89 - 91\%$, while the inference time increases only slightly (from 0.541 ms to 0.564 ms). This shows that radio waveforms contain characteristic details at both local and global scales, and that maintaining sufficient resolution allows the CSDM and PAUG modules to exploit these structures more effectively. It is noteworthy that increasing the resolution to 448×448 does not yield any substantial improvement in accuracy, remaining at only 91.08%, but increases the latency to almost three times (1.435 ms). Therefore, the 224×224 resolution represents the optimal point between information resolution and computational cost, suitable both for real-time scenarios and for resource-constrained devices. Finally, in Figure 6(c), we investigate Doppler robustness, an important factor in communication environments with motion. Over the Doppler range from 0 Hz to 300 Hz, the model performance remains stable and even improves slightly, from 90.08% (0 Hz) to 92.59% (300 Hz). These results show that the network does not rely solely on absolute frequency positions but instead learns time-frequency shape characteristics that remain stable under frequency shifts. This provides important evidence of the generalization capability of CoDeX-Net in dynamic scenarios such as mobile radar, time-selective fading channels, or communication systems in which the transmitter and receiver experience motion.

Ablation study: In the third experiment, we perform an ablation study to quantify the individual and combined contributions of the PAUG and CSDM modules. Table III summarizes the performance of four model configurations under different SNR levels and also reports the inference speed and the number of parameters. The baseline configuration $\mathcal{M}_1$, which does not use PAUG or CSDM, yields the lowest results across all SNR levels, achieving only 50.49% accuracy at -5 dB and an average accuracy of 80.03%. This confirms that a plain convolutional backbone is not sufficient for reliable waveform discrimination, especially in highly noisy environments. When only CSDM is added ($\mathcal{M}_2$), the performance improves significantly: the accuracy at -5 dB increases from 50.49% to 65.25%, and the average accuracy reaches 89.94%. This shows that the cross-scale deformable mixing mechanism in CSDM can effectively capture coarse-to-fine spectral structures that remain stable even under strong noise. In contrast, when only PAUG is integrated ($\mathcal{M}_3$), the results also improve compared with the baseline model, especially at -5 dB (64.49%) and 10 dB (95.20%), which shows that intra-resolution refinement and coordinate-based gating enhance the separability of features at the native resolution. The full CoDeX-Net model, which combines both PAUG and CSDM, achieves the highest performance at all SNR levels, including 66.66% accuracy at -5 dB and 99.83% at 30 dB, corresponding to an average accuracy of 91.01%. This demonstrates that the two modules provide complementary benefits: PAUG improves local and axis-aware feature representations, whereas CSDM supplies long-range cross-scale context. Notably, this combined design still maintains a compact model size (51K parameters) and fast inference speed (0.564 ms), which shows that the performance gain does not come at the expense of a large computational cost.

Table III
Performance comparison of different model configurations in CoDeX-Net

Model	Configurations		Accuracy (%)				Speed	Size
			SNR (dB)			Avg.	(ms)	(params)
	PAUG	CSDM	-5	10	30			
$\mathcal{M}_1$	✗	✗	50.49	84.38	91.73	80.03	0.161	7.2K
$\mathcal{M}_2$	✗	✓	65.25	94.14	99.57	89.94	0.557	33K
$\mathcal{M}_3$	✓	✗	64.49	95.20	99.78	90.28	0.474	25.4K
CoDeX-Net	✓	✓	**66.66**	**96.16**	**99.83**	**91.01**	**0.564**	**51K**

Method comparison: In the final simulation, we construct a state-of-the-art (SOTA) comparison table to evaluate the performance of CoDeX-Net against many popular CNN architectures as well as modern

Table IV
Performance comparison with state-of-the-art models

Model	Accuracy (%)				Speed (ms)	Size (params)
	SNR (dB)			Avg.		
	−5	10	30			
ResNet50 [18]	63.73	93.42	99.67	89.35	2.073	23.5M
InceptionV3 [34]	65.13	93.94	99.56	89.84	1.351	21.8M
EfficientNetB0 [20]	64.47	93.59	99.73	89.49	1.137	4M
MobileNetV2 [19]	65.31	94.64	99.69	89.91	0.825	2.2M
WaveNet [25]	65.75	94.63	99.64	90.25	0.611	140K
RadComNet [24]	65.71	94.67	99.77	90.28	0.957	178K
CMNet [29]	65.81	94.95	99.60	90.43	0.582	**38K**
DTANet [28]	65.75	95.49	99.69	90.42	0.585	75K
CoDeX-Net (Ours)	**66.66**	**96.16**	**99.83**	**91.01**	**0.564**	51K

lightweight models. Table IV shows that CoDeX-Net achieves an average accuracy of 91.01%, which is the highest among all compared models. It can be seen that common vision backbones such as ResNet50, InceptionV3, EfficientNetB0, and MobileNetV2 do not fully exploit the nature of time-frequency signals. These models all remain around 89-90% average accuracy, whereas CoDeX-Net surpasses them by about 1-2 percentage points with a much more compact architecture. The advantage of CoDeX-Net is particularly evident in the low-SNR region. At -5 dB, the model reaches 66.66%, higher than ResNet50 (63.73%) and InceptionV3 (65.13%), and also better than specialized lightweight designs such as CMNet and DTANet (around 65%). This is the most challenging operating region, where most of the signal energy is masked by noise. The fact that CoDeX-Net still maintains a 2-3% margin over the baselines shows that the PAUG and CSDM modules truly enable the model to capture more robust time-frequency structural cues, rather than relying too heavily on instantaneous amplitude patterns as in conventional CNNs. In other words, the model is not only strong in the clean region (high SNR) but also scores clearly in the challenging region (low SNR), where practicality in radar-communication systems is most important. At medium and high SNR levels (10 dB and 30 dB), CoDeX-Net continues to match or slightly exceed the other models, achieving 96.16% and 99.83%, respectively. Although some models also reach near-saturation at 30 dB, the consistently high performance across all SNR levels shows that CoDeX-Net not only has a high performance ceiling but also maintains its advantage under challenging noise conditions. With respect to computational cost, CoDeX-Net exhibits clearly superior efficiency. With only 51K parameters, the model is smaller than ResNet50 by nearly three orders of magnitude (23.5M parameters) and still significantly smaller than EfficientNetB0, or MobileNetV2 (2.2-4M parameters). Nevertheless, CoDeX-Net surpasses all of these models in both average accuracy and inference speed, with 0.564 ms compared with 0.825-2.073 ms. When compared with lightweight architectures tailored for waveform classification such as CMNet and DTANet, CoDeX-Net still retains its advantage. Although CMNet has fewer parameters (38K) and DTANet is of similar scale (75K), both remain around 90.4% average accuracy. CoDeX-Net improves this by about 0.6 percentage points while maintaining a comparable latency (0.564 ms versus 0.582-0.585 ms). In summary, Table IV shows that CoDeX-Net not only outperforms the other models numerically, but also provides an architecturally optimal configuration from a system perspective: high average accuracy, compact model size, and very low latency. This confirms that designing the architecture to match the characteristics of radio frequency signals yields real benefits compared with directly reusing vision backbones, and makes CoDeX-Net a viable candidate for practical deployment on embedded platforms or edge devices.

4 Conclusion

In this paper, we introduce CoDeX-Net, a lightweight deep learning framework addressing the challenge of waveform classification in IRC systems. By leveraging high-resolution SPWVD representations, our method bridges the gap between accuracy and computational efficiency. The core of the architecture lies in the synergy between two novel modules: the PAUG for dynamic intra-resolution refinement, and the CSDM for adaptive cross-resolution context aggregation. This design effectively captures both local details and global signal structures, overcoming the geometric rigidity of standard CNNs. Extensive experimental results on a diverse 12-class dataset demonstrate that CoDeX-Net achieves SOTA performance. The model achieves an average accuracy of 91.01% and exhibits superior robustness under severe noise conditions. Notably, CoDeX-Net outperforms both heavy vision backbones and specialized lightweight models while maintaining a compact size of only 51K parameters and a low inference latency of 0.564 ms. These computational advantages indicate CoDeX-Net's potential suitability for deployment on resource-constrained edge devices in real-time spectrum surveillance. However, acknowledging that current simulations do not fully capture hardware impairments and complex environmental interference, our future work will prioritize validating the model on real-world testbeds to bridge this sim-to-real gap.

Acknowledgment

This research is funded by Ministry of Education and Training, and hosted by Ho Chi Minh City University of Technology and Engineering under grant number B2025-SPK-03.

References

[1] D. Hou, L. Li, W. Lin, J. Liang, and Z. Han, "ClST: A convolutional transformer framework for automatic modulation recognition by knowledge distillation," *IEEE Transactions on Wireless Communications*, vol. 23, no. 7, pp. 8013–8028, 2024.

[2] W. Zhou, R. Zhang, G. Chen, and W. Wu, "Integrated sensing and communication waveform design: A survey," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1930–1949, 2022.

[3] Z. Liu, X. Li, H. Ji, H. Zhang, and V. C. M. Leung, "Toward STAR-RIS-empowered integrated sensing and communications: Joint active and passive beamforming design," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 12, pp. 15 991–16 005, 2023.

[4] S. Ning, J. Li, J. Gao, and Q. Wu, "RA-CNN: An effective automatic modulation recognition method for joint communication and radar system," in *Proceedings of the 2024 IEEE/CIC International Conference on Communications in China*, Hangzhou, China, 2024, pp. 1217–1221.

[5] X. Zhang, H. Zhao, H. Zhu, B. Adebisi, G. Gui, H. Gacanin, and F. Adachi, "NAS-AMR: Neural architecture search-based automatic modulation recognition for integrated sensing and communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 3, pp. 1374–1386, 2022.

[6] Y. Guo, H. Sun, H. Liu, and Z. Deng, "Radar signal recognition based on CNN with a hybrid attention mechanism and skip feature aggregation," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–13, 2024.

[7] H. Milczarek, I. Djurović, C. Leśnik, and J. Jakubowski, "Automatic classification of frequency-modulated radar waveforms under multipath conditions," *IEEE Sensors Journal*, vol. 23, no. 16, pp. 18 349–18 361, 2023.

[8] T. Huynh-The, Q.-V. Pham, T.-V. Nguyen, T. T. Nguyen, R. Ruby, M. Zeng, and D.-S. Kim, "Automatic modulation classification: A deep architecture survey," *IEEE Access*, vol. 9, pp. 142 950–142 971, 2021.

[9] G. Vanhoy, T. Schucker, and T. Bose, "Classification of LPI radar signals using spectral correlation and support vector machines," *Analog Integrated Circuits and Signal Processing*, vol. 91, no. 2, pp. 305–313, 2017.

[10] G. Kong, M. Jung, and V. Koivunen, "Waveform classification in radar-communications coexistence scenarios," in *Proceedings of the 2020 IEEE Global Communications Conference*, Taipei, Taiwan, 2020, pp. 1–6.

[11] A. Pavy and B. Rigling, "SV-Means: A fast SVM-based level set estimator for phase-modulated radar waveform classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 191–201, 2018.

[12] J. Pan, N. Ye, H. Yu, T. Hong, S. Al-Rubaye, S. Mumtaz, A. Al-Dulaimi, and I. Chih-Lin, "AI-driven blind signature classification for IoT connectivity: A deep learning approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6033–6047, 2022.

[13] X. Zhang, Z. Luo, W. Xiao, and L. Feng, "Deep learning-based modulation recognition for MIMO systems: Fundamental, methods, challenges," *IEEE Access*, vol. 12, pp. 112 558–112 575, 2024.

[14] L. Zhang, S. Xu, and J. Li, "CNN based target classification in vehicular networks with millimeter-wave radar," in *Proceedings of the 2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, Helsinki, Finland, 2022, pp. 1–6.

[15] B. Wang, J. Xie, and F. Wang, "Specific emitter identification based on ACGAN and STFT," in *Proceedings of the 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, Shanghai, China, 2024, pp. 400–403.

[16] T. Huynh-The, V.-S. Doan, C.-H. Hua, Q.-V. Pham, T.-V. Nguyen, and D.-S. Kim, "Accurate LPI radar waveform recognition with CWD-TFA for deep convolutional network," *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1638–1642, 2021.

[17] T. Huynh-The, Q.-V. Pham, T.-V. Nguyen, D. B. da Costa, and D.-S. Kim, "RaComNet: High-performance deep network for waveform recognition in coexistence radar-communication systems," in *Proceedings of the IEEE International Conference on Communications (ICC 2022)*, Seoul, Korea, 2022, pp. 1–6.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, NV, USA, Jun. 2016, pp. 770–778.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, UT, USA, Jun. 2018, pp. 4510–4520.

[20] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, vol. 97, California, USA, Jun. 2019, pp. 6105–6114.

[21] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[22] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.

[23] S. Xu, D. Zhang, Y. Lu, Z. Xing, and W. Ma, "MCCSAN: Automatic modulation classification via multiscale complex convolution and spatiotemporal attention network," *Electronics*, vol. 14, no. 16, 2025.

[24] T. Huynh-The, N. C. Luong, H. Phan, D. B. d. Costa, and Q.-V. Pham, "Improved waveform classification for integrated radar-communication 6G systems via convolutional neural networks," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 9, pp. 13 921–13 925, Sep. 2024.

[25] T. Huynh-The, V.-P. Hoang, J.-W. Kim, M.-T. Le, and M. Zeng, "WaveNet: Toward waveform classification in integrated radar-communication systems with improved accuracy and reduced complexity," *IEEE Internet of Things Journal*, vol. 11, no. 14, pp. 25 111–25 123, 2024.

[26] A. Dosovitskiy *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the 9th International Conference on Learning Representations*, May 2021, pp. 1–21.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 10 012–10 022.

[28] 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.

[29] T. Huynh-The, T.-T. Le, T.-H. Vu, and D. Benevides da Costa, "CMNet: Radar-communication waveform recognition via convolution and mamba networks," *IEEE Wireless Communications Letters*, vol. 14, no. 9, pp. 2997–3001, Sep. 2025.

[30] F. Ayaz, B. Alhumaily, S. Hussain, M. A. Imran, K. Arshad, K. Assaleh, and A. Zoha, "Radar signal processing and its impact on deep learning-driven human activity recognition," *Sensors*, vol. 25, no. 3, 2025.

[31] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 2023, pp. 16 133–16 142.

[32] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 5728–5739.

[33] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, no. 483, New Orleans, Louisiana, USA, 2018, pp. 3942–3951.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

Thanh-Dat Tran is currently pursuing a B.S. degree in Computer Engineering Technology with the Department of Computer and Communication 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.

Minh-Thanh Le received the B.S and M.Sc degrees in Electronics and Telecommunication Engineering from Ho Chi Minh City University of Technology (HCMUT), Vietnam, in 2000 and 2003, respectively. He is currently a Lecturer in 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, wireless communications, IoT applications, and artificial intelligence.

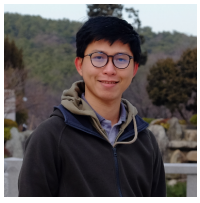
Son Ngoc Truong received his B.S. and M.S. degrees in electronic engineering from the Ho Chi Minh City University of Technology and Education, Vietnam, in 2006 and 2011, respectively. He completed his Ph.D. degree in electronic engineering at Kookmin University, Seoul, South Korea, in 2016. He was a postdoctoral researcher at Kookmin University from 2016 to 2017. He is currently an Associate Professor in Department of Electronics and Information Engineering, Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam. His research interests include neuromorphic computing systems, brain-inspired circuits, VLSI design, hardware acceleration of deep neural networks, and related areas of electronic systems design. He can be contacted at email: sontn@hcmute.edu.vn

Ngoc-Ha Truong received the B.E. degree in Electronics and Telecommunications Engineering from Ho Chi Minh City University of Technology (BKU), Vietnam, in 2004, and the M.S. degree in Electronics Engineering from Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam. He is currently a PhD Candidate in Electronics Engineering with the Faculty of Electrical and Electronics Engineering at HCM-UTE. His current research interests include artificial intelligence and deep learning applications in wireless communications, 4G/5G/6G networks, antenna design, and robust channel estimation.

Pham Ngoc Son received the B.E. degree (2005) and M.Eng. degree (2009) in Electronics and Telecommunications Engineering from Post and Telecommunication Institute of Technology, Ho Chi Minh City and Ho Chi Minh City University of Technology, Vietnam, respectively. In 2015, he received the Ph.D. degree in Electrical Engineering from the University of Ulsan, South Korea. He is currently an Associate Professor in the Faculty of Electrical and Electronics Engineering (FEEE) of Ho Chi Minh City University of Technology and Engineering (HCM-UTE). His major research interests are cooperative communication, cognitive radio, physical layer security, energy harvesting, intelligent reflecting surfaces, short packet communications, and deep learning.

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.

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 Computer and Communication Engineering, HCMC University of Technology and Engineering (HCMUTE), 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.



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.