Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

C2D-RT-DETR-Based Wideband Signal Detection and Parameter Extraction

Zhibo Shi
Xijing University
Rui Zhu
zhu_r10@163.com

Xijing University
Lulu Liu
Xijing University
Yarong Chen
Xijing University
Le Gao
Xijing University
Huanhuan Gheng
Xijing University

Article

Keywords: Wideband spectrum sensing, Weak signal detection, Real-time object detection, Coordinate attention, Content-aware reassembly, Dynamic head

Posted Date: September 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10869246/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

C2D-RT-DETR-Based Wideband Signal Detection and Parameter Extraction

Zhibo Shi[1,2], Rui Zhu[1,2*],Lulu Liu[1,2], Yarong Chen[1,2], Le Gao[1,2], Huanhuan Gheng[1,2]

[1]School of Electronic Information, Xijing University, Xi'an 710123, China.

[2]Xi'an Key Laboratory of High Precision Industrial Intelligent Vision Measurement Technology, Xijing University, Xi'an, Shaanxi 710123, China.

*Correspondence:zhu_r10@163.com.

Abstract

In the field of wideband spectrum sensing, the accurate detection of weak signals and the extraction of physical parameters in complex electromagnetic environments remain severe challenges. To address the degradation of localization accuracy in traditional detection networks caused by blurred time-frequency features and edge distortion under low signal-to-noise ratio (SNR) conditions, this paper proposes a real-time spectrum sensing model named C2D-RT-DETR. This model achieves high-fidelity reconstruction of corrupted time-frequency features by constructing a synergistic mechanism of "coordinate sensing, content reassembly, and dynamic task alignment": First, the Coordinate Attention (CA) mechanism is introduced into the shallow feature extraction branch to enhance key time-frequency spatial localization and suppress disordered background noise through orthogonal decomposition encoding along the time-delay and frequency axes. Second, the Content-Aware ReAssembly of FEatures (CARAFE) module is deployed in the multi-scale feature fusion path to dynamically generate reconstruction kernels based on local semantics, thereby avoiding the blurring of high-frequency details caused by traditional interpolation. Finally, the Dynamic Head (DyHead) is integrated at the output end to achieve efficient decoupling of classification and bounding box regression feature flows through adaptive regulation in spatial and task dimensions. Building upon this, a mapping mechanism from geometric boundaries to the physical domain is established, enabling the effective extraction of key physical parameters such as center frequency, occupied bandwidth, onset time, and duration. Experimental results demonstrate that C2D-RT-DETR exhibits excellent anti-noise robustness and a high recall rate within the dynamic SNR range of -15 dB to 5 dB. In a 0 dB noise environment, the mAP_{50} of the model reaches 0.687, an improvement of 3.1% over the original baseline network, with a target matching coverage rate up to 92.8%. Furthermore, the Normalized Root

Mean Square Errors (NRMSE) for center frequency and onset time estimation are as low as 0.018% and 0.157%, respectively. In addition, the total parameter count of the model is only 32.76 M, and the single-frame inference latency is 19.46 ms. While balancing real-time performance and low computational overhead, this research provides a reliable technical solution for the intelligent deployment of non-cooperative wideband spectrum monitoring terminals.

Keywords: Wideband spectrum sensing; Weak signal detection; Real-time object detection; Coordinate attention; Content-aware reassembly; Dynamic head.

1. Introduction

Real-time detection, modulation recognition, and parameter estimation of wideband communication signals are core technologies in cognitive radio and modern spectrum sensing. With the rapid evolution of Internet of Things and 5G/6G communication technologies, the electromagnetic spectrum environment is becoming increasingly complex, exhibiting complicated characteristics such as large bandwidth, high overlap, and strong noise interference. In this context, how to achieve robust and high-precision rapid sensing of unauthorized overlapping signals under harsh channel conditions with low signal-to-noise ratios (SNR) is directly related to the dynamic allocation efficiency of spectrum resources and the security of public communications. Therefore, developing an intelligent signal detection architecture with high anti-noise robustness and real-time inference capability has become an urgent key problem to be solved in the fields of wireless communications and intelligent signal processing [1]-[4].

Traditional wideband signal sensing schemes rely heavily on the prior knowledge of domain experts, typically utilizing statistical tools such as higher-order cumulants (HOC) and cyclic spectrum analysis, or refined time-frequency analysis methods like Wigner-Ville Distribution (WVD) and Continuous Wavelet Transform (CWT) to extract signal features for decision-making[5-10]. Although these classical methods possess a solid theoretical foundation and stability under conditions of relatively high SNR and simple channel environments, their detection accuracy and generalization ability often experience significant degradation when facing the pervasive strong noise or dense signals in actual electromagnetic environments, due to the limitations of fixed feature operators and manually set thresholds.

In recent years, deep learning represented by Convolutional Neural Networks (CNN) and Transformers, relying on its excellent data analysis and nonlinear mapping capabilities, has successfully realized the leap from manual feature analysis to automatic feature learning in the signal processing

domain[13-18]. Particularly in the field of multi-object detection, the real-time framework RT-DETR effectively overcomes the local receptive field limitations of traditional CNNs when processing concurrent wideband multi-signals by utilizing global self-attention, providing a highly efficient baseline scheme for the integrated intelligent monitoring of complex spectrum situations[40-42].

However, in practical spectrum monitoring scenarios, signal detection architectures often operate in harsh physical environments with low or even negative SNRs. Under such rigorous conditions, the energy of weak signals is often at the edge of the noise floor, manifesting as discontinuous, diffused, or nonlinear geometric trajectories submerged by Gaussian white noise on the time-frequency spectrogram. The existing RT-DETR architecture exposes two deficiencies when addressing such complex time-frequency sensing tasks: First, in the cross-scale fusion path of the feature pyramid, the Nearest Upsampling operator adopted by RT-DETR lacks the perception ability of local physical features; this static resampling mechanism tends to smooth out the high-frequency details of signals during the feature fusion process, causing irreversible information loss. Second, during the feature extraction process, because RT-DETR lacks a spatial adaptive mechanism for time-frequency energy ridges and struggles to flexibly adjust its receptive field to fit complex and variable frequency-modulated waveforms, it is prone to significant spatial localization deviations and physical quantity estimation drifts when the decoder regresses key geometric parameters such as signal bandwidth and duration.

Aiming at the aforementioned technical bottlenecks, based on the RT-DETR architecture, this paper proposes a multi-dimensional dynamic decoupling feature enhancement detection network, named C2D-RT-DETR, oriented towards complex electromagnetic environments. This model aims to reconstruct the physical time-frequency topology of weak signals with high fidelity under strong noise backgrounds through content-aware reassembly and multi-scale dynamic alignment mechanisms, thereby improving the detection accuracy in high-noise environments. Specifically, first, the Content-Aware ReAssembly of FEatures (CARAFE) module is introduced in the top-down multi-scale feature fusion path to replace the traditional fixed upsampling operation. This module can dynamically generate interpolation kernels based on local semantic features, thereby accurately preserving the time-frequency edge details of fine-grained signals. Second, the Coordinate Attention (CoordAtt) mechanism is embedded in the shallow feature streams P3 and P4, which are rich in high-resolution textures. Through orthogonal encoding along the time and frequency axis directions, it strengthens the coordinate location information and effectively suppresses disordered background noise from the spatial dimension. Finally, before feeding the multi-

scale features into the decoder, the dynamic multi-scale detection head DyHeadBlock is deployed. By integrating scale-aware, spatial-aware, and task-aware mechanisms, it achieves adaptive fine alignment of non-stationary geometrically distorted features, maximizing the recognizability of signals in the time-frequency domain.

The main contributions of this paper are as follows:

(1) Addressing the low SNR scenarios frequently encountered in wideband signal detection, an intelligent detection framework, C2D-RT-DETR, is proposed. This framework integrates the content-aware reassembly of CARAFE, the spatial coordinate refinement of CoordAtt, and the multi-dimensional dynamic feature alignment mechanism of DyHead, breaking through the sensing limitations of traditional visual detection models in environments with low SNR and multi-signal overlap.

(2) A mathematical mapping model from visual bounding box geometric coordinates to physical parameters of wideband signals is refined and validated, providing an engineering solution for automated parameter extraction under harsh channel conditions.

(3) A wideband time-frequency dataset encompassing 9 complex modulation types with an SNR span of -15 dB to 5 dB is created, and comparative and ablation experiments are conducted. The results indicate that the mean Average Precision (mAP_{50}) of the proposed model reaches 0.687 under a 0 dB noise environment, improving by 3.1% over the baseline model, and it maintains robust target recall performance and parameter estimation accuracy even at extreme low SNR boundaries.

The remainder of this paper is organized as follows: Section 2 reviews related research progress in the fields of wideband signal sensing and deep learning object detection. Section 3 details the overall architecture of the C2D-RT-DETR framework and the underlying mechanisms of its core enhancement modules. Section 4 comprehensively analyzes the comparative and ablation results of the proposed model based on the wideband modulated signal dataset, demonstrating its detection and parameter extraction efficacy under extreme interference. Section 5 summarizes the entire paper.

2. RELATED WORKS

Traditional wideband signal analysis relies heavily on classical signal processing theory. Early researchers dedicated themselves to constructing feature libraries based on expert knowledge, utilizing high-resolution time-frequency analysis methods such as the Hilbert-Huang Transform (HHT) and

Wigner-Ville Distribution (WVD), combined with cyclic spectrum features, higher-order cumulants (HOC), and instantaneous envelope statistics[5-12], to accomplish the identification of signal modulation types and parameter extraction. This approach performs well in scenarios with sufficient prior information and controllable SNR. However, this method is highly dependent on manual feature engineering: on the one hand, when confronting unknown non-cooperative signal modulations, preset features often lose their generalization capability; on the other hand, under conditions of low SNR and strong multipath interference, complex environmental noise severely destroys the transient coherence of the signal, leading to a precipitous decline in the robustness of traditional algorithms.

With the rise of data-driven approaches, artificial intelligence models represented by deep Convolutional Neural Networks (CNN) and Transformers have achieved breakthroughs in sensing performance. Through multi-layer nonlinear representation learning, deep networks can automatically mine high-dimensional abstract semantics from original signals[19-23]. However, current deep learning-based modulation recognition schemes are mostly restricted to the ideal assumption of narrowband single signals, ignoring the reality of high-density overlapping of multiple signals in the actual electromagnetic spectrum. When disordered Gaussian white noise textures alias with complex signal energy clusters, traditional deep classification networks struggle to accurately aggregate contextual semantics[24-31], appearing inadequate when handling complex dynamic spectrum monitoring tasks.

To break through the bottleneck of simultaneous multi-object localization and concurrent multi-parameter extraction, researchers in recent years have begun attempting to reframe spectrum sensing tasks as object detection problems in computer vision[32-34]. By projecting one-dimensional time-domain observation sequences into two-dimensional time-frequency spectrograms, the physical parameters of signals-such as center frequency, bandwidth, and duration-are ingeniously transformed into geometric coordinate mappings of image bounding boxes. This perspective thoroughly circumvents the complex pixel-level segmentation and threshold decision logic found in traditional signal processing, significantly elevating the level of automated signal sensing.

Regarding the selection of detection models, although two-stage methods represented by Faster R-CNN possess higher accuracy, their inference latency makes it difficult to meet real-time reconnaissance demands. Meanwhile, single-stage models led by the YOLO series have substantially improved inference efficiency through regression prediction, thus becoming the mainstream solutions[35-39]. As sensing requirements further increase, novel detection models like RT-DETR, leveraging their Non-Maximum

Suppression (NMS)-free design and the global modeling advantages of Transformers, have significantly enhanced sensing robustness and boundary regression accuracy in dense signal environments. Nevertheless, within low SNR regimes, due to the lack of targeted enhancement for the underlying physical noise in time-frequency spectrograms, such generic visual detection models still struggle to evade deep-seated flaws such as feature smoothing and edge localization drift.

To mitigate the contamination of feature representations by underlying noise, the academic community has conducted targeted optimization explorations. Some studies have introduced channel attention mechanisms to recalibrate the contribution of feature maps, but explicit dimensionality reduction operations often degrade the nonlinear correlations among different modulated signals. Looking across existing research, feature enhancement for complex modulated signals mostly focuses on single-dimensional patching, highlighting an urgent need for a multi-dimensional decoupled enhancement mechanism capable of synergistic fusion.

In summary, current spectrum sensing technologies have evolved from manual feature engineering to deep detection paradigms. However, sensing bottlenecks persist when dealing with the nonlinear topological distortion of weak signals in low SNR environments. To further improve the regression accuracy of signal boundaries and classification anti-noise robustness while maintaining real-time sensing efficiency, this paper intends to explore a deeply decoupled spatial and content-aware feature enhancement scheme within the RT-DETR framework.

3.METHODOLOGY

3.1 Overall Framework Overview

To tackle the sensing challenges of the complex electromagnetic spectrum under low SNR conditions, this paper proposes an integrated signal monitoring framework based on visual object detection. The overall operational flow of this scheme is illustrated in Figure 1, which sequentially encompasses three core stages: time-frequency energy mapping, C2D-RT-DETR intelligent detection, and the inverse derivation of physical parameters.

First, in the time-frequency energy mapping stage, the STFT is utilized to project one-dimensional non-stationary time-domain wideband mixed signals into two-dimensional time-frequency spectrograms,explicitly representing the spatial energy distribution of multi-object signals. Subsequently, in the intelligent detection stage, the generated time-frequency feature maps are fed into the novel C2D-

RT-DETR network designed in this paper. To overcome feature fragmentation and topological distortion under low SNR conditions, the network integrates the Coordinate Attention (CA) mechanism to anchor the time-frequency orientation of the targets, introduces the Content-Aware ReAssembly of FEatures (CARAFE) operator to achieve lossless upsampling of the high-frequency edges of weak signals, and combines the Dynamic Head (DyHead) to adaptively fit the distorted energy ridges across multiple feature dimensions. This enables the precise output of bounding boxes and modulation categories for multiple targets. Finally, in the physical parameter derivation stage, relying directly on the geometric coordinates output by the network, a mapping channel between visual bounding boxes and the true time-frequency resolution is established. This allows for the calculation of the center frequency, occupied bandwidth, and time-domain start and end ranges of the wideband signals, thereby achieving complete signal sensing.

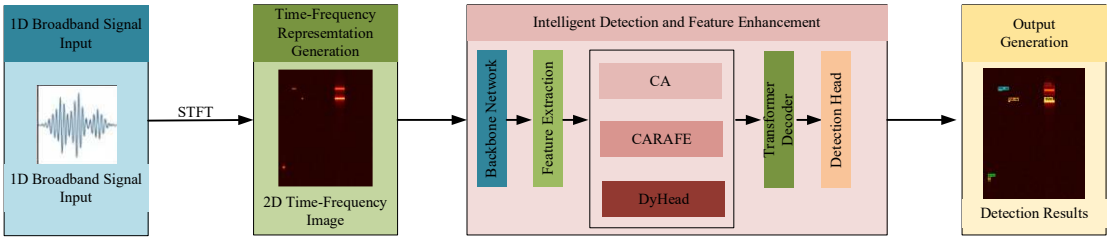


Fig. 1. Overall architecture of the proposed system.

In the time-frequency energy mapping stage, suppose the one-dimensional discrete mixed signal sequence acquired by the receiver within a finite observation time window is defined as:

$$y[n]=\sum_{k=1}^{K} s_k[n]+v[n] \tag{1}$$

where $y[n]$ is the comprehensive observation sequence at the n sampling point; K represents the total number of concurrent independent target signals within the current spectrum bandwidth; $s_k[n]$ corresponds to the transmitted signal component injected with the k specific modulation pattern; $v[n]$ denotes the AWGN in the physical channel; and N represents the total length of discrete sampling points in the current frame.

For this one-dimensional non-stationary sequence, this study introduces the STFT to perform two-dimensional time-frequency expansion. Its mathematical operation in discrete form can be expressed as:

$$S(m,f)=\sum_{n=0}^{L-1} y[n]w[n-m\cdot\Delta]e^{-j\frac{2\pi fn}{N_{FFT}}} \tag{2}$$

Here, $w[\cdot]$ is a window function with local truncation and smoothing characteristics, with a

length of L; Δ is the translation step size between adjacent sliding windows; and N_{FFT} represents the number of points for the Fast Fourier Transform. $S(m,f)$ characterizes the complex spectrum response of the signal at the intersection of the m-th time frame and the f-th frequency index.

Considering the enormous dynamic energy difference of the received signals in harsh channels, to highlight the time-frequency geometric textures of weak signals and suppress the overload effect of energy peaks, it is necessary to perform logarithmic scale compression on the magnitude of the aforementioned complex matrix:

$$P(m,f)=10\log_{10}(|S(m,f)|+\delta) \tag{3}$$

Where $|\cdot|$ denotes the modulus operation; the regularization term δ is set to an extremely small positive real constant, aiming to prevent calculation overflow singularities of the logarithmic function in zero-energy noise floor regions.

The single-channel spectrogram matrix obtained through the above processing is transformed via pseudo-color mapping into a three-channel tensor $\mathbf{X} \in \mathbb{R}^{B\times 3\times H\times W}$, which conforms to the standard input format of deep visual networks. Here, B is the Batch Size for network forward propagation, and H and W respectively correspond to the pixel dimensions of the 2D time-frequency spectrogram on the time and frequency axes. This tensor will serve as the direct input to the C2D-RT-DETR network.

In the intelligent detection stage, the time-frequency tensor $\mathbf{X}$ first undergoes multi-scale deep feature extraction by the backbone network. During this process, the CA module encodes the direction-aware information of frequency bands and time delays into the channel dimension, enhancing the precise localization of local signal clusters. Subsequently, in the cross-scale feature fusion path, the traditional nearest neighbor resampling operator is discarded, and the CARAFE module is utilized instead to dynamically generate upsampling kernels based on local content semantics, effectively avoiding the smoothing distortion of high-frequency textures in strong noise environments. Ultimately, the feature stream is fed into the DyHead module, which adaptively aligns the time-frequency edges distorted by strong noise through attention synergy across three dimensions: scale, spatial, and task. The decoder accordingly outputs high-precision detection bounding boxes.

In the physical parameter derivation stage, the system extracts the 2D pixel coordinates of the detection boxes. Relying on the time-pixel and frequency-pixel physical resolution scaling coefficients pre-established in the front-end STFT process, it performs inverse normalization and reverse physical mapping of the geometric coordinates. This method completely eliminates the cumbersome connected-

component post-processing steps in traditional spectrum sensing, realizing an efficient and low-latency conversion from visual detection results to four major physical parameters of the signal: center frequency, bandwidth, onset time, and duration.

3.2 Architecture of the Detection Network

This paper employs the C2D-RT-DETR detection model, the specific architecture of which is illustrated in Fig 2.

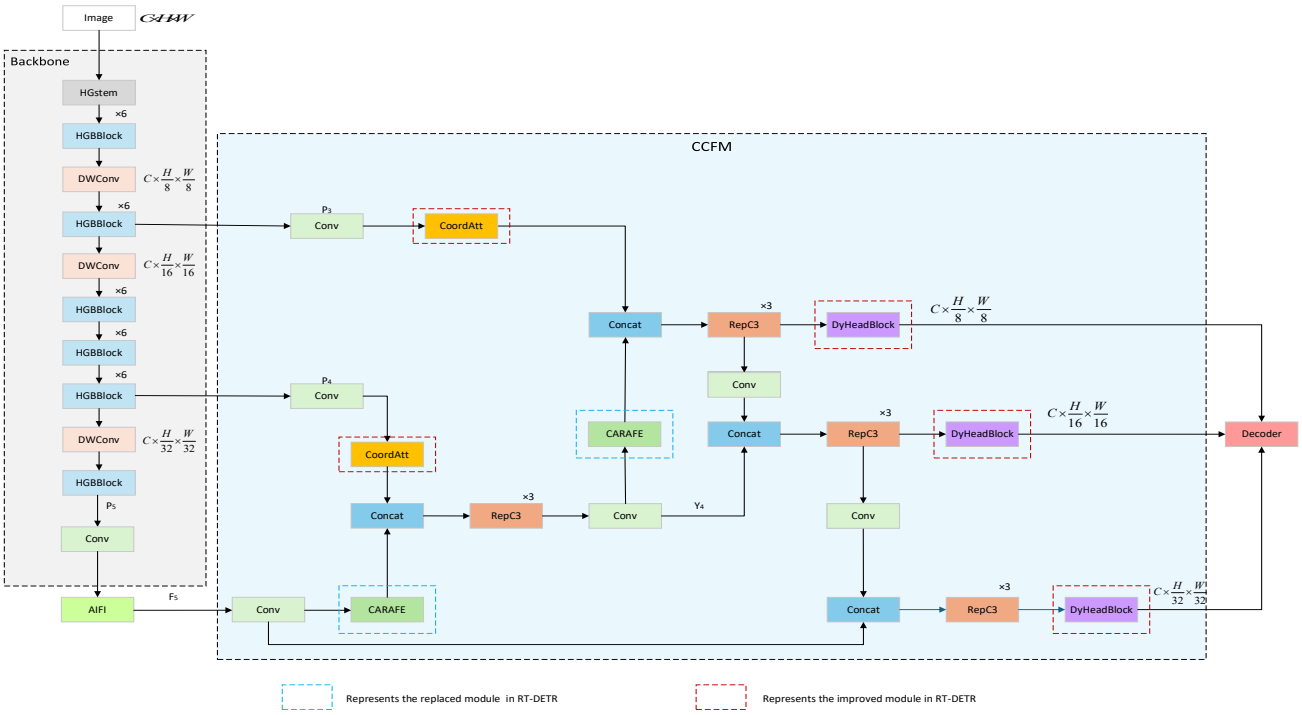


Fig. 2. The network architecture of C2D-RT-DETR.

Compared with the baseline model, C2D-RT-DETR maintains the macroscopic streamlined layout of three core functional components: a deep backbone network responsible for multi-level visual feature extraction; a Cross-Scale Feature Fusion Module (CCFM) aimed at achieving high-level cross-scale feature interaction and anti-noise reassembly; and a Transformer decoder equipped with an Intersection over Union (IoU)-aware query selection mechanism, which is used to directly output the localization and classification results of weak targets.

Specifically, let the input data be the three-channel time-frequency spectrogram tensor $\mathbf{X} \in \mathbb{R}^{B\times 3\times H\times W}$ obtained through front-end STFT processing and conversion. This tensor first enters the backbone network. Through downsampling and deep feature extraction by a series of core modules such as HGStem, HGBlock, and DWConv, the backbone network outputs a set of feature maps

containing three different spatial resolutions and semantic levels at different stages, denoted as $\{P_3, P_4, P_5\}$. Its feature extraction process can be abstractly represented as:

$$P_i = \mathrm{F}^{(i)}_{backbone}(\mathbf{X}), \quad i \in \{3,4,5\} \tag{4}$$

where $\mathcal{F}^{(i)}_{backbone}$ represents the downsampling and feature mapping operations of the i-th stage of the backbone network. The spatial resolution corresponding to the extracted features P_i decreases as the network depth increases, while the semantic abstraction level gradually rises. The feature P_5 located at the deep bottleneck of the network is first fed into the attention-based Intra-scale Feature Interaction module. Through the self-attention mechanism, the aggregation of global context is completed, which is then transformed into high-level semantic features F_5 containing macroscopic time-frequency situations.

Subsequently, this set of multi-scale features is sent into the enhanced CCFM newly designed in this paper. As shown in Fig. 2, targeting the pain points of time-frequency ridge distortion and high-frequency detail loss in harsh environments with low SNR, this paper performs a three-dimensional structural reconstruction of the feature fusion path, characterized by "coordinate sensing, content upsampling, and dynamic task alignment".

First, at the input branches of the shallow high-resolution features P_3 and P_4, the CoordAtt mechanism is innovatively embedded. By capturing long-distance dependencies along the frequency band and time delay directions, CoordAtt can encode precise spatial coordinate information into channel features, thereby firmly anchoring the time-frequency orientation of weak signals at the early stage of cross-scale fusion.

Second, in the bottom-up multi-scale feature upsampling path, the traditional nearest neighbor interpolation operator is completely abandoned and replaced by the CARAFE module. This module can dynamically generate reassembly kernels based on local contextual semantics, achieving lossless anti-noise upsampling of high-frequency edges of signals while improving feature resolution.

Finally, the dynamic detection head module, DyHeadBlock, is additionally deployed at the three multi-scale output ports of the CCFM. This module unifies scale awareness, spatial awareness, and task awareness, and can adaptively adjust the shape of the receptive field to perfectly fit the signal energy edges distorted and deformed by strong background noise.

Ultimately, the multi-scale high-quality features, deeply and synergistically enhanced by CoordAtt, CARAFE, and DyHeadBlock, are fed into the Transformer decoder, completely breaking

through the bottleneck of time-frequency object detection in low SNR environments.

3.2.1 CA Module

The Coordinate Attention (CA) module adopted in this paper is detailed in Fig. 3.

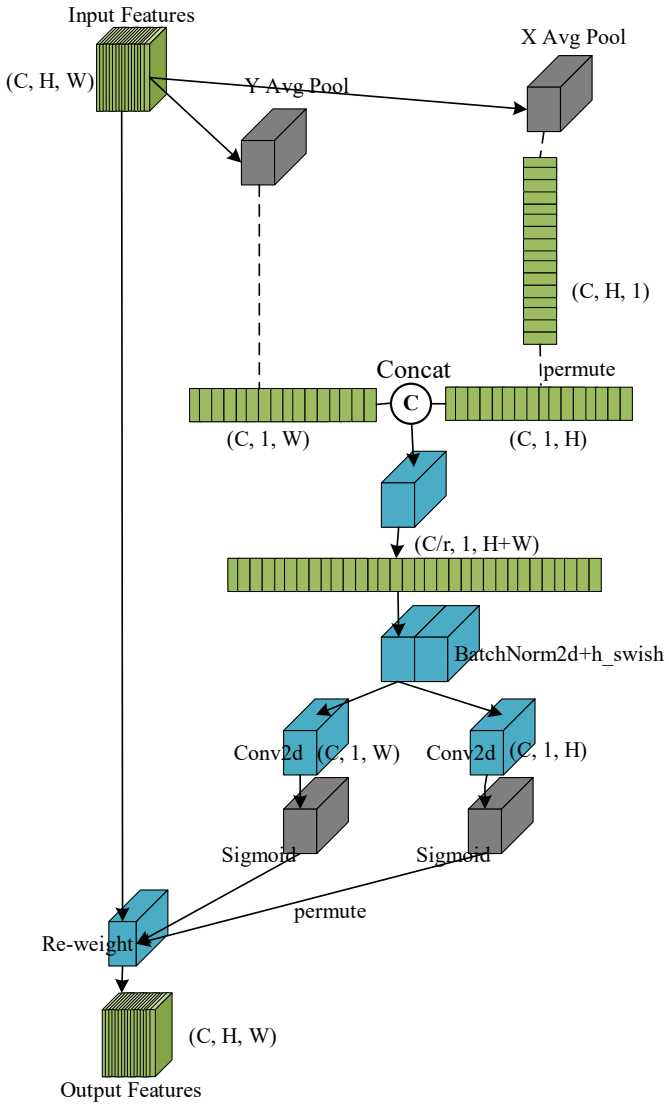


Fig. 3. The network architecture of the CA module.

To more accurately capture the time-frequency boundary features of weak signals in complex electromagnetic environments, this paper introduces the CA module in the P3 and P4 feature layers. Traditional channel attention mechanisms typically compress 2D feature maps directly into 1D via global average pooling. Although this effectively models inter-channel dependencies, it inevitably loses the spatial positional information crucial for object localization. For wideband signal detection tasks, the spatial coordinates on the time-frequency spectrogram directly physically map to the signal's onset time and center frequency. The CA module adopted in this paper not only performs feature recalibration in the channel dimension but also encodes precise positional information into the channel attention[43]. Its specific network architecture is illustrated in Fig. 3.

The core idea of the CA module is to decompose and aggregate positional information through 1D global pooling operations along two spatial directions, thereby strengthening the model's perception of the physical time-frequency boundaries of signals while filtering out high-frequency white noise. Specifically, the forward processing of this module is primarily divided into three steps: coordinate information embedding, coordinate attention generation, and feature re-weighting:

For an input deep feature tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, to encourage the attention module to capture long-range spatial interactions with precise positioning, the CA module first employs two 1D adaptive average pooling kernels of sizes $(H,1)$ and $(1,W)$to encode the input features of each channel along the vertical and horizontal directions, respectively. The output expression for the c-th channel extracted along the vertical direction at height h is formulated as:

$$z_c^h(h)=\frac{1}{W}\sum_{0\le i<W} x_c(h,i) \tag{5}$$

Similarly, the output expression for the c-th channel extracted along the horizontal direction at width w is:

$$z_c^w(w)=\frac{1}{H}\sum_{0\le j<H} x_c(j,w) \tag{6}$$

These two direction-aware pooling operations yield a pair of direction-aware feature maps. This dual aggregation mechanism allows the CA module to capture long-range contextual dependencies along one spatial direction while preserving precise positional information along the other, thereby assisting the network in more accurately localizing the distorted signal energy clusters within the time-frequency spectrogram.

After obtaining the aggregated feature maps, the module concatenates these two feature maps containing specific directional information along the spatial dimension, and subsequently utilizes a shared 1×1 convolutional layer for channel dimensionality reduction to decrease the model's computational overhead. Following processing by BatchNorm2d and an h-swish activation function, an intermediate feature map $\mathbf{f} \in \mathbb{R}^{\frac{C}{r}\times(H+W)\times 1}$ containing joint spatial information of both horizontal and vertical directions is generated:

$$f=\delta(B(Conv_{1\times 1}([z^h,z^w]))) \tag{7}$$

where $[z^h,z^w]$ denotes the concatenation operation along the spatial dimension, r is the reduction ratio controlling the compression proportion, $\mathcal{B}$ represents the batch normalization operation, and δ indicates the non-linear activation function h-swish.

Subsequently, the intermediate feature map $\mathbf{f}$ is split again along the spatial dimension into two independent directional tensors, $\mathbf{f}^h \in \mathbb{R}^{\frac{C}{r}\times H\times 1}$ and $\mathbf{f}^w \in \mathbb{R}^{\frac{C}{r}\times 1\times W}$. Two additional 1×1convolutional layers are employed to respectively restore the channel counts of these two tensors to the same dimension C as the input feature tensor, and a Sigmoid activation function is applied to generate the final vertical and horizontal attention weight matrices:

$$g^h=\sigma(Conv_{1\times1}(f^h)) \tag{8}$$

$$g^w=\sigma(Conv_{1\times1}(f^w)) \tag{9}$$

Finally, the generated horizontal and vertical attention weights undergo element-wise multiplication with the original input features $\mathbf{X}$ to obtain the ultimate high-quality output feature $\mathbf{Y}$ of the coordinate attention module:

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \tag{10}$$

Through the aforementioned mathematical and logical mechanisms, the CA module achieves a joint purification of the channel domain and the spatial coordinate domain in the extremely deep feature space. The weight regions approaching 0 adaptively suppress the disordered Gaussian white noise responses, while the weight regions approaching 1 faithfully let through the time-frequency ridge features of the wideband signals. More importantly, compared to traditional global pooling, this operation completely preserves the spatial coordinate distribution capable of inversely deriving physical parameters, providing a clean and precisely localized data foundation for the global context aggregation and bounding box regression in the subsequent decoder.

To intuitively reveal the spatial orientation and anti-noise enhancement mechanisms of the CA module on weak time-frequency signals during the shallow feature extraction stage, Fig. 4 illustrates a set of channel response comparisons of feature maps before and after integrating the CA module under a 0 dB noise background.

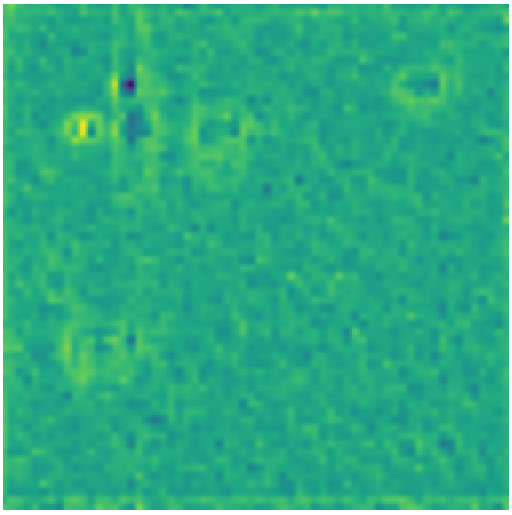


Fig. 4(a)

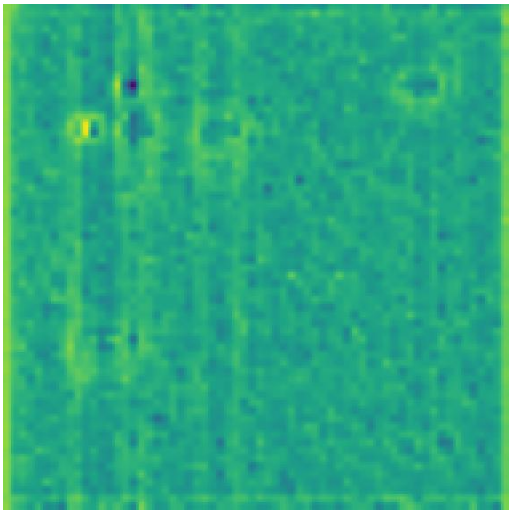


Fig. 4(b)

As shown in Fig. 4(a), in the baseline feature space without CA processing, the feature plane is filled with high-frequency, disordered, and discrete noise points due to the influence of strong Gaussian white noise. This causes severe diffusion and fragmentation phenomena at the boundaries of weak real signals, making it difficult to form a continuous topological representation. In contrast, as shown in Fig. 4(b), after introducing the horizontal and vertical 1D adaptive pooling and direction-aware recalibration of the CA module, the feature responses exhibit significant orthogonal time-frequency axial focusing characteristics. Specifically, the strip-like energy responses along the time and frequency axes are significantly strengthened, enabling the originally blurred and discrete signal energies to be precisely anchored at the orthogonal coordinate intersections. Simultaneously, the signal-free background regions lacking geometric energy relevance are assigned extremely low weights, and the background noise floor of the entire feature map tends towards uniform silence. This visual evidence directly corroborates that the CA module can effectively extract long-range time-frequency coordinate dependencies through orthogonal spatial decomposition. While suppressing disordered electromagnetic background noise, it provides a feature representation foundation with high SNR and clear localization for subsequent cross-scale fusion and parameter regression.

3.2.2 CARAFE Module

The specific architecture of the CARAFE module adopted in this paper is illustrated in Fig. 5.

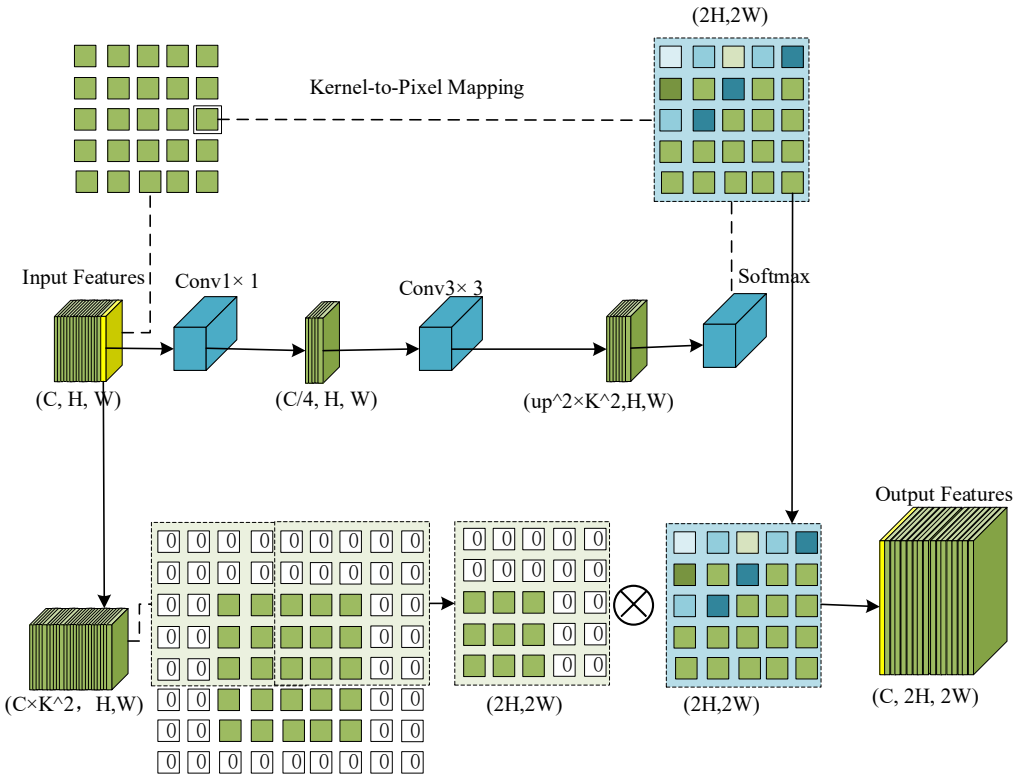


Fig. 5. Architecture of the CARAFE module.

In the top-down feature traceback path of the hybrid encoder, cross-scale feature fusion relies on spatially upsampling the deep low-resolution feature maps (F_5 or Y_4) to achieve spatial dimension alignment with the shallow high-resolution feature maps. The standard RT-DETR architecture typically employs Nearest Neighbor Interpolation to accomplish this process. However, this traditional upsampling method relies solely on the relative geometric positions between sampling points for pixel completion, falling into a fixed interpolation paradigm.

Under low SNR electromagnetic interference, traditional interpolation methods can trigger severe negative effects: isolated high-energy artifacts generated by strong background noise are smoothly amplified by the interpolation operation, forming large areas of false background noise patches. Meanwhile, the inherently distorted and fragmented edges of weak signals are highly susceptible to semantic diffusion during pixel replication or weighted interpolation, causing the boundary between time-frequency ridges and the background to become further blurred. To achieve content-aware adaptive edge reconstruction, this paper introduces the CARAFE module[44] into the top-down feature recovery path. Its specific structure is shown in Fig. 5.

As illustrated in Fig. 5, the CARAFE module decouples the upsampling process into two core sub-stages: the Kernel Prediction Module and the Content-aware Reassembly Module.

Let the input deep low-resolution feature tensor be $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, and the upsampling ratio be $\sigma = 2$. First, to reduce the computational complexity when generating reassembly kernels, the module performs channel compression on the input tensor via a 1×1 convolutional layer, reducing the number of channels from C to $C_m = C/4$, generating compressed features $\mathbf{X}_{compress} \in \mathbb{R}^{\frac{C}{4}\times H\times W}$.

Subsequently, the module utilizes a 3×3 content encoding convolutional layer to adaptively predict the reassembly weights within the spatial receptive field based on local contextual semantics. For each input pixel, the network needs to generate a prediction kernel of size $K_{up} \times K_{up}$ and corresponding weights for σ^2 high-resolution output grids. Therefore, the output channel number of the content encoding convolution is mapped to $\sigma^2 \cdot K_{up}^2$. Then, the spatial dimension is expanded to the high-resolution grid $(2H, 2W)$ via a PixelShuffle operation.

Finally, a Softmax function is applied along the channel dimension to normalize the predicted reassembly kernel weights, ensuring the energy conservation of the reassembly weights. The mathematical expression for this process is:

$$W_{kernel}=Softmax\left(PixelShuffle\left(Conv_{3\times3}(Conv_{1\times1}(X))\right)\right) \tag{11}$$

where $\mathbf{W}_{kernel} \in \mathbb{R}^{K_{up}^2\times 2H\times 2W}$ represents the $K_{up} \times K_{up}$ local receptive field reassembly weight corresponding to each target pixel point on the high-resolution grid.

In the reassembly stage, for any target pixel point $l' = (i', j')$ in the high-resolution output feature map $\mathbf{Y} \in \mathbb{R}^{C\times 2H\times 2W}$, its corresponding center point mapped back to the original low-resolution input map is $l = (\lfloor i'/2 \rfloor, \lfloor j'/2 \rfloor)$.

The module first extracts a local feature block $\mathcal{N}(\mathbf{X}_l, K_{up})$ of size $K_{up} \times K_{up}$ on the input feature tensor $\mathbf{X}$ centered at l. Subsequently, this local feature block undergoes an element-wise product with the adaptive weight $\mathbf{W}_{kernel}(l')$ generated by the kernel prediction module, and the results are accumulated to ultimately aggregate and generate the high-resolution target pixel feature:

$$Y_{l'}=\sum_{k=-r}^{r}\sum_{m=-r}^{r} W_{kernel}^{(k,m)}(l')\cdot X_{l+(k,m)} \tag{12}$$

where $r = \lfloor K_{up}/2 \rfloor$ represents the radius of the local receptive field.

At the programmatic implementation level, this process first performs constant padding on the channel-aligned input features $\mathbf{X}_{align}$, and utilizes the F.unfold operator to efficiently extract local $K_{up} \times K_{up}$ feature blocks; then, it initially aligns the high-resolution spatial grid through nearest interpolation, and finally performs point-wise multiplication and addition with the reassembly kernel

matrix to achieve intelligent edge reshaping in the high-resolution space.

To intuitively reveal the repair and anti-noise advantages of the CARAFE module on the time-frequency edges of weak signals during the feature upsampling process, Fig. 6 visualizes the comparative responses of time-frequency spectrogram features under a 0 dB noise background processed by traditional nearest neighbor upsampling and the CARAFE module.

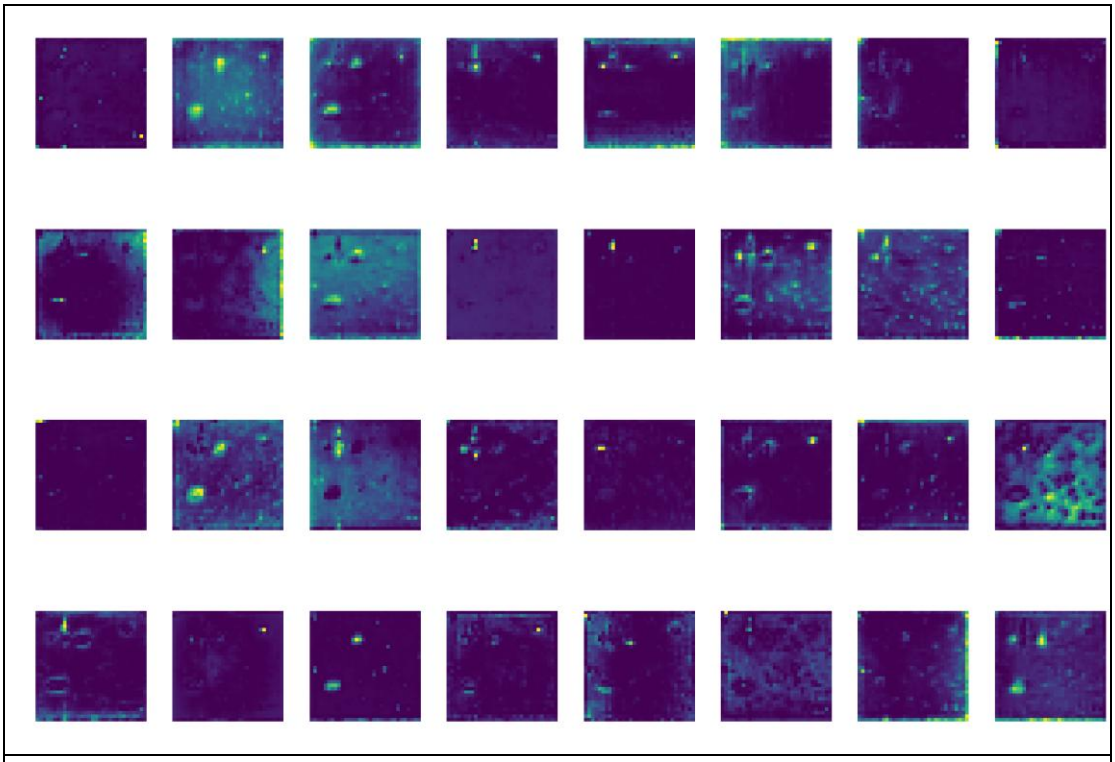


Fig. 6(a)

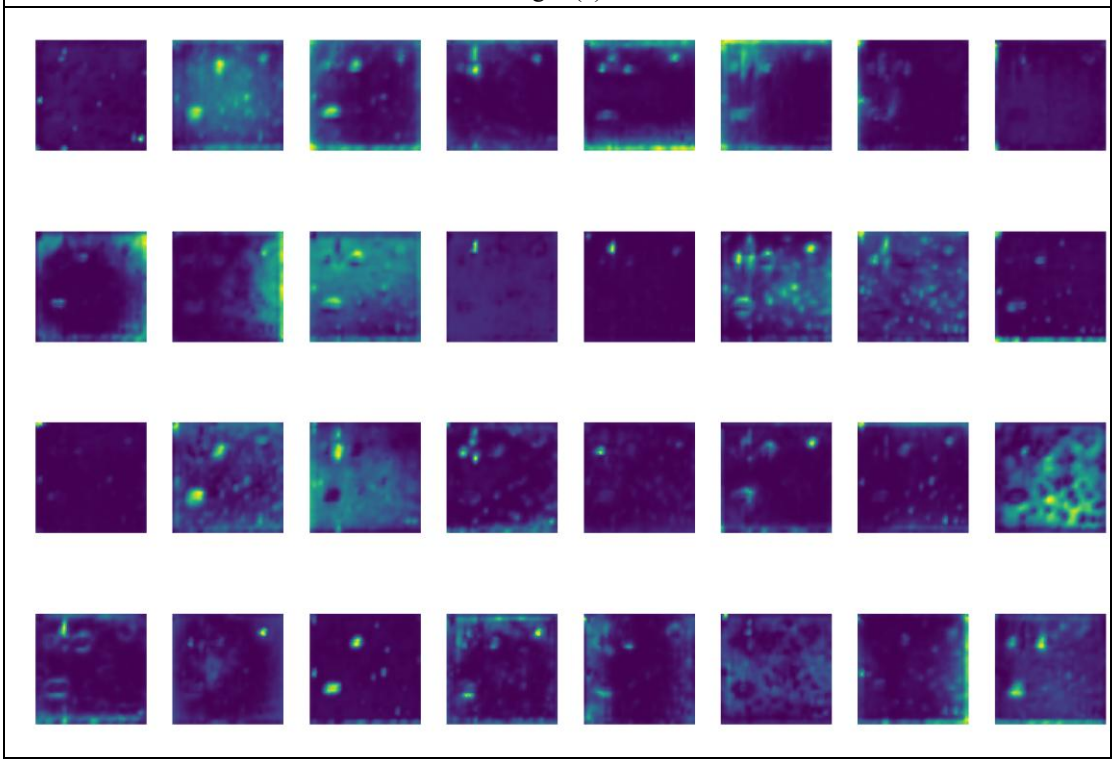


Fig. 6(b)

As shown in Fig. 6(a), in the deep low-resolution feature space before upsampling, due to the small spatial grid size and the limitations of discrete sampling, the feature distribution exhibits obvious pixelation and high-contrast granular effects. Simultaneously, the residual white noise manifests as isolated high-frequency noise points, which can easily trigger spatial aliasing during direct transmission.

In contrast, as shown in Fig. 6(b), after spatial dimension enlargement by the CARAFE module, the spatial resolution of the feature map is expanded from the low-dimensional grid to twice its original size. During this scale transformation process, CARAFE dynamically predicts reassembly kernels based on local contextual semantics, executing content-adaptive weighted aggregation on neighborhood pixels rather than simple geometric pixel replication. This mechanism breaks the original discrete step boundaries of the low-resolution grid, effectively suppressing the high-frequency noise points that originally exhibited a blocky discrete distribution, and transforming the geometric envelopes of weak signals into spatially continuous, smooth energy clusters. Consequently, the features visually present a softer and more continuous manifold distribution, effectively circumventing the jagged distortion and noise amplification effects brought about by traditional fixed interpolation, thereby laying a structural foundation for the subsequent high-fidelity fusion of multi-scale features.

3.2.3 DyHead Module

After being purified by the deep CA module and reconstructed by the CARAFE content-aware upsampling, the multi-scale feature maps still face two core challenges before being fed into the Transformer decoder for final classification and localization regression: First, residual spatial white noise can forge false high-frequency edges in the spatial dimension. Second, the classification and bounding box regression tasks in object detection systems have distinctly different preferences for feature representations-classification focuses on the global semantic invariance of the response region, whereas regression is extremely sensitive to the local abrupt edges of the signal's time-frequency boundaries. Traditional detection heads often forcibly share the same feature stream, which easily triggers task misalignment between classification and regression.

To break through the aforementioned limitations, this paper designs and deploys a dynamic head block, the DyHead module[45], at the front end of the detection head, with its specific architecture illustrated in Fig. 7.

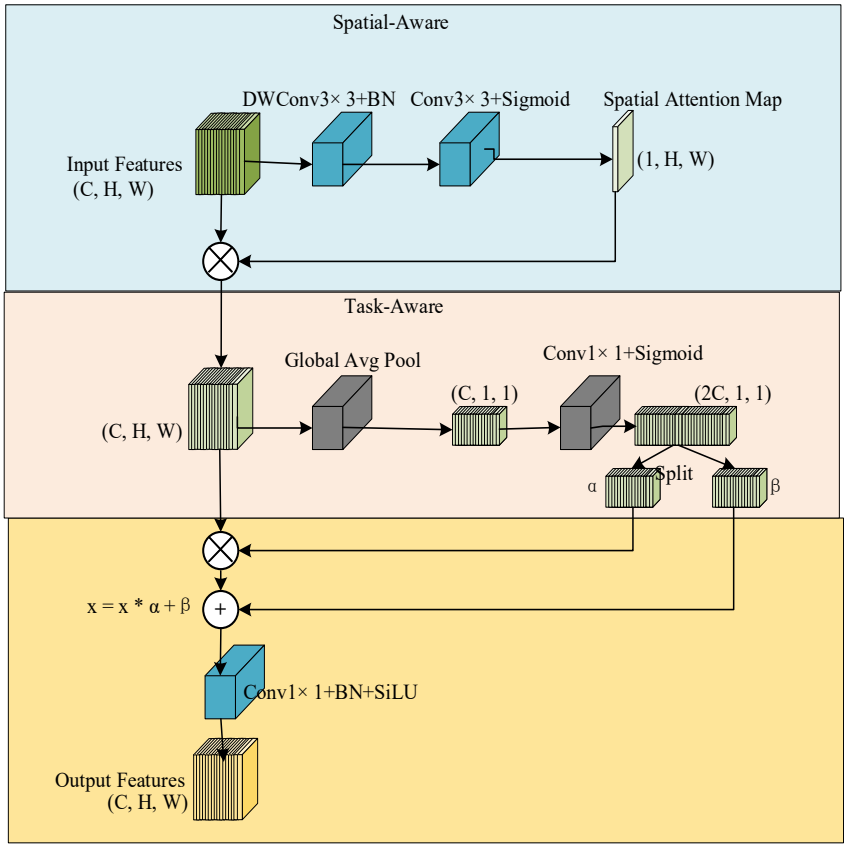


Fig. 7. Architecture of the DyHead module.

As illustrated in Fig. 7, the DyHead module decouples the feature regulation process into two serial core sub-stages: the spatial-aware stage and the task-aware stage.

Let the multi-scale feature tensor input to this module be $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$. To precisely highlight true signals and suppress disordered electromagnetic white noise in the spatial dimension, the network first extracts local spatial geometric features through a depthwise separable convolution, maintaining channel independence with an extremely low parameter count; subsequently, it stabilizes the feature distribution through batch normalization, and compresses the number of channels to 1 via a 1×1 convolution and a Sigmoid activation function, generating a 2D spatial attention weight map $\mathbf{S}_{attn} \in \mathbb{R}^{1\times H\times W}$:

$$S_{attn}=\sigma\left(Conv_{1\times 1}\left(BN(DWConv_{3\times 3}(X))\right)\right) \tag{13}$$

where $\sigma(\cdot)$ represents the Sigmoid function. Subsequently, the generated spatial attention map undergoes element-wise multiplication with the input tensor $\mathbf{X}$ to achieve adaptive highlighting and noise suppression in the spatial dimension:

$$\mathbf{X}_{spatial} = \mathbf{X} \otimes \mathbf{S}_{attn} \tag{14}$$

Through this mechanism, the diffused white noise in the background regions is assigned a weight

approaching 0, while the time-frequency ridge energy of the true signal is significantly focused.

After obtaining the spatially purified feature $\mathbf{X}_{spatial}$, to dynamically allocate and decouple the feature stream for the subsequent classification and regression branches, the module introduces a task-aware attention mechanism based on dynamic channel scaling and shifting.

Specifically, the network first compresses the spatial dimension to 1×1 via global average pooling to acquire a global channel descriptor. Following this, a 1×1 convolution is utilized to increase the channel dimensionality to $2C$, and a Sigmoid function is applied for mapping to generate a task regulation vector $\mathbf{T}_{attn}$ with dimensions of $2C \times 1 \times 1$:

$$\mathbf{T}_{attn} = \sigma(Conv_{1\times1}(GAP(\mathbf{X}_{spatial}))) \tag{15}$$

Next, $\mathbf{T}_{attn}$ is evenly split along the channel dimension into two independent factor vectors $\alpha, \beta \in \mathbb{R}^{C\times1\times1}$:

$$\alpha, \beta = Split(\mathbf{T}_{attn}, dim = 1) \tag{16}$$

where α represents the dynamic scaling coefficient, and β represents the dynamic shifting bias. The module utilizes these two adaptive parameters to perform affine transformation and reassembly on the feature map:

$$\mathbf{X}_{task} = \mathbf{X}_{spatial} \otimes \alpha + \beta \tag{17}$$

Through this adaptive scaling and shifting, the network can flexibly offset and adjust the feature distribution according to different classification and localization needs. Finally, the feature is processed by a 1×1 output convolution, batch normalization, and a SiLU activation function to generate the ultimate high-quality feature representation:

$$\mathbf{Y}_{out} = SiLU(\mathcal{B}(Conv_{1\times1}(\mathbf{X}_{task}))) \tag{18}$$

To intuitively reveal the physical advantages of the DyHead module in eliminating background white noise interference and resolving task misalignment, Fig. 8 compares the feature response distributions of a time-frequency spectrogram under a 0 dB noise background processed by a conventional convolutional layer and the DyHead module.

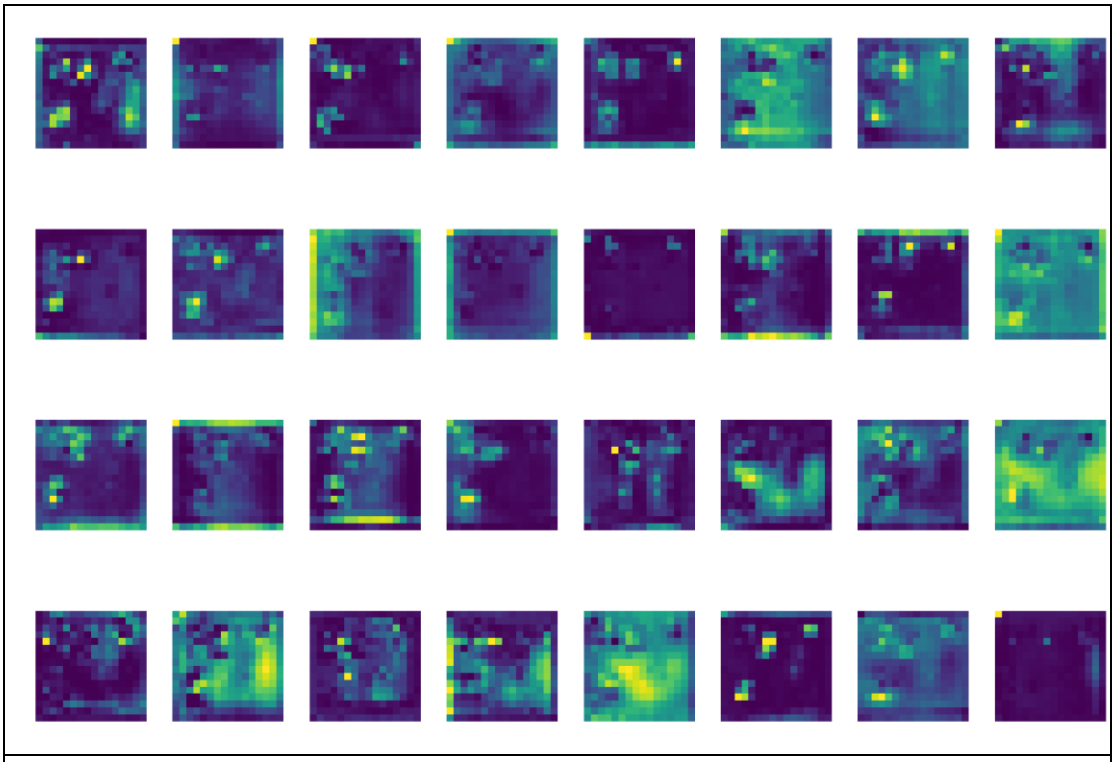


Fig. 8(a)

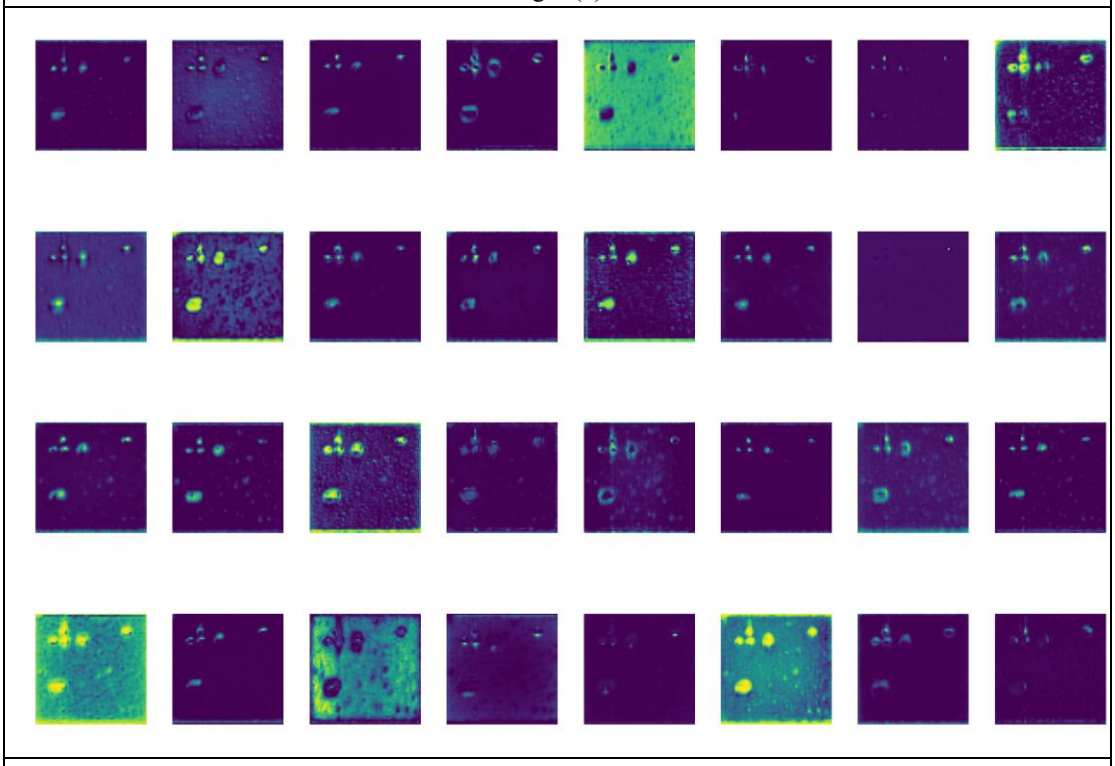


Fig. 8(b)

As illustrated in Fig. 8(a), in the feature space without DyHead regulation, strong AWGN causes the signal-free background regions to be flooded with large areas of scattered false activation responses. Simultaneously, the geometric edges of weak signals exhibit relatively obvious diffusion effects, resulting in limited contrast between the effective signal energy and the background noise

floor.

In contrast, as shown in Fig. 8(b), following the joint spatial and task-aware regulation by the DyHead module, the feature maps exhibit an extremely pure and highly focused response representation. On the one hand, the spatial attention mechanism significantly suppresses the widely diffused electromagnetic background noise through adaptive gating and weighting of the feature plane, rendering the non-target regions virtually completely silent. On the other hand, the task-aware mechanism performs high-contrast purification of the true signal's time-frequency energy skeleton through dynamic affine transformations in the channel dimension. This causes different channels to present distinct task-differentiation characteristics: specific channels retain the global envelope energy beneficial for modulation recognition, while the majority of channels precisely converge to the local abrupt edges advantageous for bounding box regression. This visual evidence demonstrates that the DyHead module effectively eliminates background noise interference and mitigates task conflicts between classification and localization features, providing highly discriminative feature inputs for the final decoder's high-precision target capture and physical parameter inversion.

3.3 Bounding Box Regression and Parameter Mapping

To achieve the quantitative conversion from detection results in the image space to electromagnetic physical quantities, this section establishes a physical parameter inversion mechanism based on 2D bounding box coordinates.

Suppose the set of predicted target bounding boxes output by the C2D-RT-DETR decoder is denoted as $\mathcal{B} = \{b_k\}_{k=1}^{K}$. For any individual predicted bounding box b_k, its geometric location can be uniquely determined by $\mathbf{p}=[x_{min},y_{min},x_{max},y_{max}]^T \in [0,1]^4$, where (x_{min},y_{min}) and (x_{max},y_{max}) correspond to the relative normalized coordinates of the top-left and bottom-right corners of the predicted box in the image coordinate system, respectively.

Let T_{obs} denote the continuous observation time window of the front-end signal receiving system, and let f_s denote the system sampling frequency. Based on the linear mapping characteristics of the 2D time-frequency spectrogram, the physical parameters of the signal in both the time and frequency domains can be derived as follows:

In the time dimension, the horizontal pixel axis of the image linearly corresponds to the temporal evolution of the signal. The signal start time t_{start} and its temporal duration Δt are expressed as:

$$t_{start}=x_{min}\cdot T_{obs} \tag{19}$$

$$\Delta t=(x_{max}-x_{min})\cdot T_{obs} \tag{20}$$

In the frequency dimension, the vertical pixel axis of the image maps to the baseband spectral distribution of the system. Considering the geometric convention that the vertical axis increases downwards in the digital image coordinate system, the occupied bandwidth B and baseband center frequency f_c of the signal can be calculated by:

$$B=(y_{max}-y_{min})\cdot f_s \tag{21}$$

$$f_c=\left(\frac{1}{2}-\frac{y_{min}+y_{max}}{2}\right)\cdot f_s \tag{22}$$

The above equations establish a deterministic mapping relationship from the geometric boundaries to the physical-domain parameters. This mapping method eliminates the need for pixel-level semantic segmentation or energy-threshold-based binarization searches, directly calculating the core physical attributes of the signal via linear operations on the extreme coordinates of the bounding boxes. Combined with the high-precision boundary localization capability of the C2D-RT-DETR model under low SNR conditions, this mechanism achieves high-accuracy reconstruction of physical parameters in strong-interference scenarios while maintaining the low-latency advantage of single-stage detection networks.

4 Experimental Results

4.1 Dataset Description

To evaluate the detection and parameter extraction performance of the proposed C2D-RT-DETR model in harsh electromagnetic environments, a wideband multi-signal dataset was generated via simulation. The detailed simulation parameters are listed in Table 1. During data generation, the receiver sampling rate was set to $2.4MHz$, and the observation time window for each sampling period was $2s$. To emulate real-world communication scenarios where multiple signals overlap, all signals were down-converted to baseband for processing. For each generated signal, key parameters-including the start time, duration, carrier frequency, and symbol rate-were randomly selected within predefined intervals. This stochastic combination strategy enriches the diversity of the synthesized time-frequency spectrograms, enabling a comprehensive and objective evaluation of the model's generalization capability under complex conditions.

Table 1. Parameters for dataset generation

Parameter	Value / Set
Sampling rate (f_s)	2.4 MHz

Baseband center frequency(f_c)	{-1.0, -0.8, -0.4, 0.4, 0.6, 0.8} MHz
Symbol rate (R_s)	{1.2, 2.4, 4.8, 9.6, 19.2} kHz
Signal duration (T_d)	{0.01, 0.03, 0.05, 0.08, 0.10, 0.12, 0.20} s
Signal start time (T_s)	{0.1, 0.3, 0.5, 0.8, 1.2, 1.5, 1.6} s
Observation time window	2 s
Signal-to-noise ratio (SNR)	-15 dB ~ 5 dB
Number of samples per SNR	10,000

As shown in Table 2, nine representative wideband signal modulation types are selected for this study. The entire dataset contains 10,000 independent samples for each SNR gradient, which are uniformly split into training, validation, and test subsets at a ratio of 7:2:1.

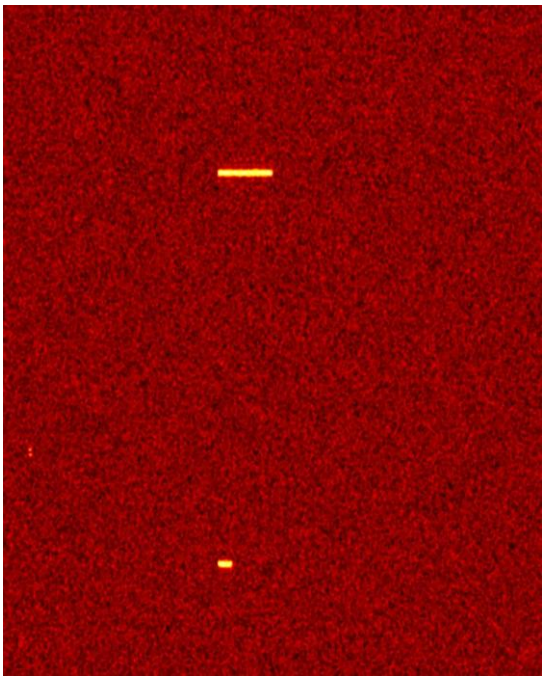
Regarding the sample generation pipeline, spectral overlapping in real-world scenarios is replicated by randomly superimposing one or multiple modulated signals within a single sampling window. Subsequently, AWGN of varying intensities is injected into the signals to simulate fading and interference introduced by harsh channels. To facilitate direct processing by visual models, the mixed one-way time-domain waveforms are uniformly mapped into two-dimensional spectrograms via the STFT. Their dynamic ranges are then compressed using logarithmic transformation, followed by pseudo-color encoding, ultimately yielding standard three-channel RGB time-frequency samples.

Table 2. Signal Modulation Types

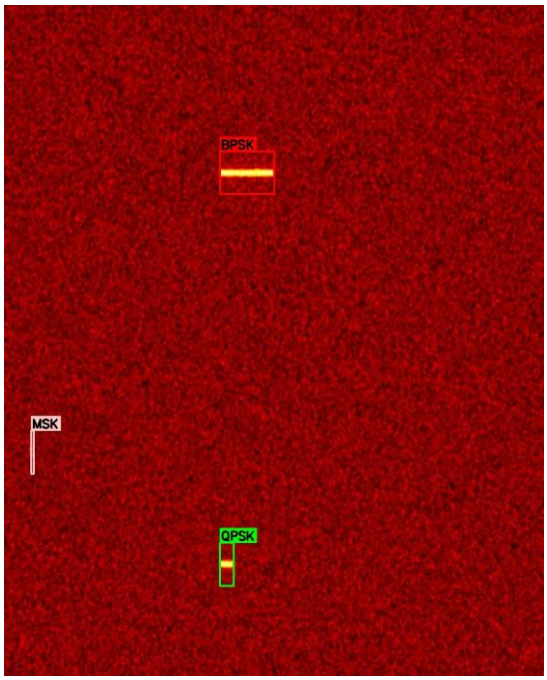
Modulation Type	Modulation Method
Digital Modulation	Binary Phase Shift Keying, BPSK
	Quadrature Phase Shift Keying, QPSK
	8 Phase Shift Keying, 8PSK
	16 Quadrature Amplitude Modulation,16QAM
	64 Quadrature Amplitude Modulation,64QAM
	Minimum Shift Keying, MSK
Analog Modulation	Double-Side Band Amplitude Modulation, DSB-AM
	Single Side Band Amplitude Modulation, SSB-AM
	Frequency Modulation, FM

Fig. 9 presents the time-frequency spectrograms along with their corresponding ground-truth bounding boxes for several representative signals under various SNR conditions. The visual degradation of time-frequency features induced by noise can be clearly observed: at a high SNR of 5 dB, the signals exhibit well-defined contours and concentrated energy; however, as the SNR

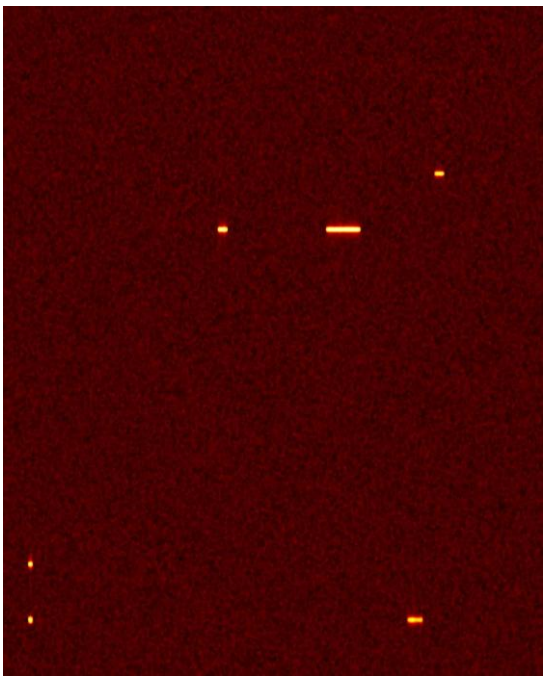
deteriorates to 0 dB and down to an extremely low level of -15 dB, the random background noise intensifies substantially. Consequently, the originally regular signal stripes are severely corrupted by noise, leading to fragmentation, blurring, and edge dispersion, to the point where they almost merge with the background. Such feature submergence and geometric distortion represent the most challenging bottlenecks in practical spectrum reconnaissance, thereby serving as a highly persuasive benchmark to validate the proposed C2D-RT-DETR framework.

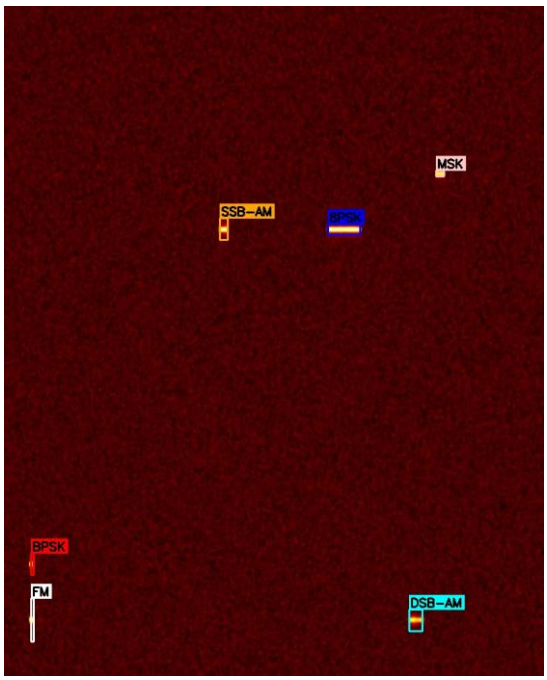


-15dB



Annotated signal spectrograms at -15 dB





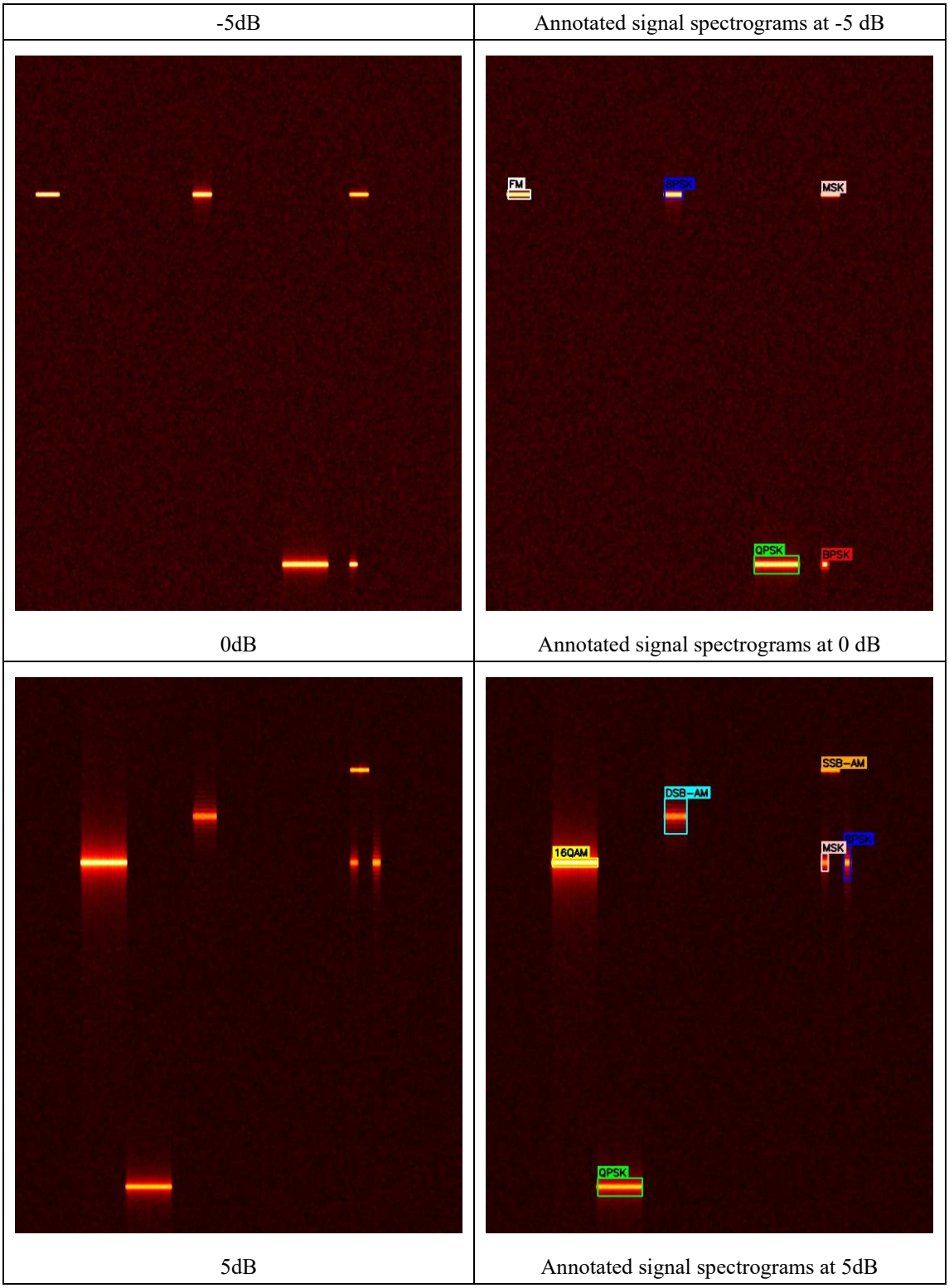


Fig. 9. Original and annotated spectrograms under different SNRs.

4.2 Experimental Setup

To ensure the objectivity of the performance evaluation and the rigorousness of the comparative experiments, the training, validation, and inference testing of all models were independently conducted on the same hardware and software platform. The experiments were implemented using the PyTorch

deep learning framework, with GPU acceleration enabled via the CUDA architecture.

Regarding the training and optimization strategies, customized optimizer configurations were adopted to account for the varying sensitivities of different network architectures to gradient updates. For the benchmark models in the YOLO series, which are built upon the conventional convolutional paradigm, Stochastic Gradient Descent (SGD) with momentum was uniformly employed to ensure smooth convergence of the deep convolutional weights. Conversely, for the proposed C2D-RT-DETR and the official RT-DETR-L baseline, the AdamW optimizer was selected for parameter updates, given the sensitivity of Transformer modules to parameter initialization and weight decay. By decoupling weight decay from gradient updates, AdamW better stabilizes the optimization trajectory of the self-attention mechanism, accelerates convergence, and reduces the risk of overfitting.

Furthermore, to rigorously control experimental variables and eliminate performance discrepancies induced by hyperparameter differences, all evaluated networks were trained to convergence under strictly consistent training schedules, input resolutions, and batch sizes. The specific hardware specifications, operating system version, and underlying software dependencies are detailed in Table 3.

Table 3. Hardware configuration and experimental environment.

Name	Model
CPU	Intel(R) Core(TM) i9-13900K
System	Windows 11
GPU	NVIDIA GeForce RTX 4090 D 24GB
RAM	128GB
Python	3.10.13
CUDA	12.1
PyTorch	2.2.1
Torchvision	0.17.0
Epochs	100

4.3 Evaluation Metrics

To comprehensively and multi-dimensionally assess the sensing efficacy of the proposed model for complex wideband signals, the evaluation framework is divided into two distinct dimensions: object

detection performance and physical parameter estimation accuracy. In the dimension of object detection and modulation classification, the standard metrics widely adopted in computer vision-namely Precision (P), Recall (R), and mean Average Precision (mAP)-are utilized as baseline evaluation criteria.

To accurately evaluate the fidelity of the reconstructed core physical quantities derived via inverse mapping from the 2D geometric bounding boxes, the Mean Squared Error (MSE) and Normalized Root Mean Square Error (NRMSE) are introduced as quantitative benchmarks. Let $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$ denote the set of ground-truth annotations for a specific physical parameter in the test set, and let $\hat{\mathcal{Y}} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N\}$ denote the set of estimated values inverted by the model, where N represents the total number of valid matched samples. The MSE between the predicted and ground-truth values is formulated as follows:

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{23}$$

Because physical parameters such as frequency and time have inconsistent physical dimensions and significantly disparate numerical scales, direct evaluation using absolute error precludes fair cross-dimensional comparisons. Therefore, this study adopts the range of a given physical quantity across the entire test set,$y_{\max} - y_{\min}$, as a normalization factor to non-dimensionalize the deviation, yielding the expression for NRMSE:

$$NRMSE=\frac{\sqrt{MSE}}{\mathrm{y}_{\max}-\mathrm{y}_{\min}} \tag{24}$$

where $y_{\max}$ and $y_{\min}$ denote the maximum and minimum values of the corresponding physical parameter observed across the test samples, respectively. An NRMSE value closer to 0 indicates a higher degree of overlap between the geometric envelope inferred by the model and the actual time-frequency distribution of the electromagnetic signal, reflecting superior robustness and accuracy in parameter estimation, thereby demonstrating the practical utility of the mapping scheme under strong noise interference.

4.4 Comparative Experiments

To evaluate the sensing efficacy of the C2D-RT-DETR model in complex electromagnetic environments, comparative experiments were conducted against the original RT-DETR-L baseline and mainstream YOLO detection networks across an SNR range of -15 dB to 5 dB.

Setting the evaluation scope to -15 dB ~ 5 dB holds explicit practical engineering significance. In

practical non-cooperative spectrum reconnaissance scenarios, when the SNR exceeds 5 dB, signals maintain a substantial energy margin above the noise floor, allowing conventional models or traditional energy detection methods to achieve reliable perception with marginal performance divergence among algorithms. Conversely, when the SNR falls below -15 dB, the time-frequency signatures of signals are almost completely submerged by the background noise, approaching the physical limits of detectability. Therefore, -15 dB ~ 5 dB represents the most realistic and challenging operating regime. Within this interval, although the time-frequency boundaries of weak signals undergo severe distortion due to heavy noise interference, they still preserve localized structural cues viable for feature extraction.

Fig. 10 illustrates the mAP_{50} curves of the primary comparative models across various SNRs. The experimental results demonstrate that C2D-RT-DETR exhibits superior detection performance across the entire SNR range from -15 dB to 5 dB. Specifically, at the critical operating points of 0 dB and 5 dB, the mAP_{50} of C2D-RT-DETR reaches 0.687 and 0.701, achieving improvements of 3.1% and 2.9% over the RT-DETR-L baseline (0.656 and 0.672), respectively. Even under the extreme low-SNR condition of -15 dB, where the baseline mAP_{50} degrades to 0.597, C2D-RT-DETR still maintains a detection accuracy of 0.612, demonstrating robust noise immunity.

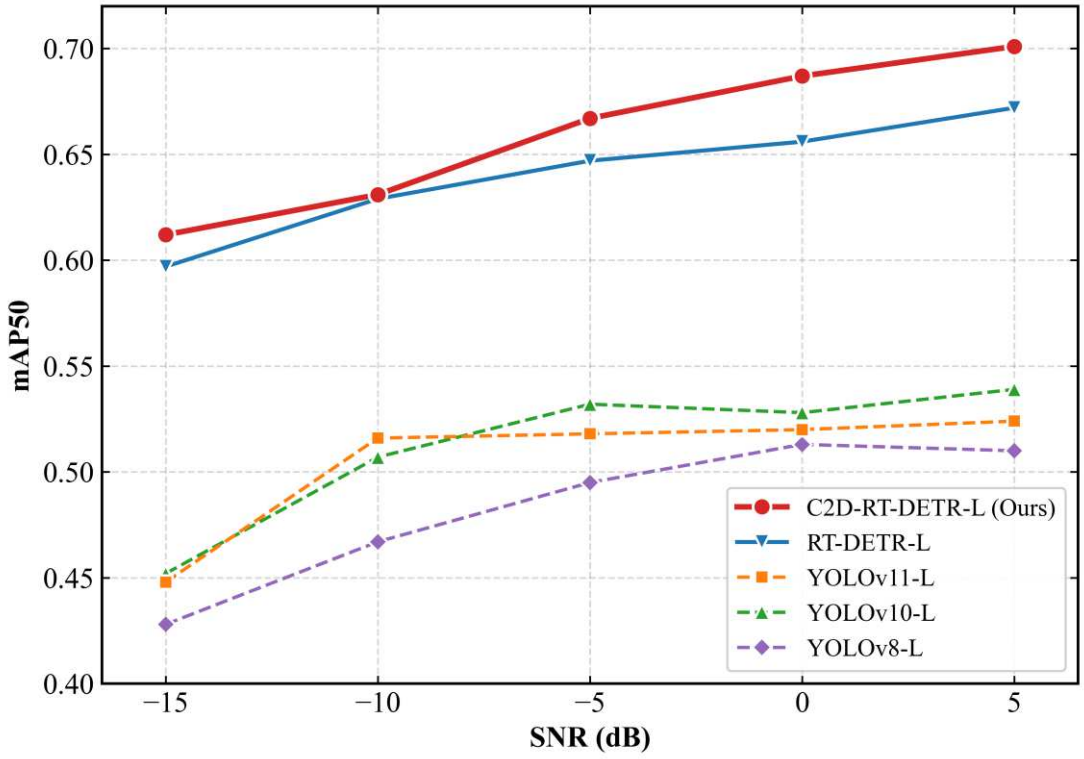


Fig. 10. mAP_{50} curves of the primary comparative models under different SNRs

The comparative experimental data objectively demonstrate the adaptability bottlenecks of the classic YOLO series architectures under low SNR conditions. Across the entire test range from -15 dB to 5 dB, the detection accuracies of YOLOv8-L, YOLOv10-L, and YOLOv11-L remain at relatively low levels. This is primarily because the time-frequency boundaries of weak signals are prone to severe blurring and fragmentation against strong background noise; conventional convolutional operators with fixed rectangular receptive fields struggle to adaptively capture irregular time-frequency structures, readily leading to feature blurring and prediction failures.

In contrast, the superior performance exhibited by C2D-RT-DETR validates the efficacy of the proposed joint enhancement mechanism. Specifically, the CA module incorporates direction-aware encoding at the front end of feature extraction, accurately pinpointing the 2D spatial energy distribution of the signal and suppressing widespread disordered noise. The CARAFE module adaptively reconstructs shallow spatial details via content-aware reassembly kernels, effectively circumventing the aliasing and artifact amplification inherent in conventional interpolation-based upsampling. Furthermore, the DyHead module deployed at the output stage achieves dynamic decoupling of classification and localization feature streams through the dual modulation of spatial refinement and task awareness. This collaborative paradigm enables the model to efficiently extract discriminative features even in harsh environments at or below 0 dB, attaining detection accuracy superior to all benchmark models.

Furthermore, based on the aforementioned geometric-to-physical domain inverse mapping mechanism, this study quantifies the number of effectively recalled targets and the NRMSE of key physical parameters for each model across three representative SNR gradients: -15 dB, 0 dB, and 5 dB. The detailed results are presented in Tables 4 through 6.

Table 4. Multi-class Classification Metrics and Parameter Estimation Accuracy at -15 dB（4424 targets）

Model	Avg. mAP50	Avg. Precision	Avg. Recall	Onset Time NRMSE	Duration NRMSE	Center Frequency NRMSE	Bandwidth NRMSE	Matched Targets
RT-DETR-L	0.573	0.605	0.605	0.00952	0.01274	0.000526	0.03693	3813
YOLOv11-L	0.571	0.589	0.599	0.001004	0.01577	0.000216	0.0346	2518
YOLOv10-L	0.564	0.611	0.597	0.00898	0.01828	0.000181	0.03372	2405

Model	Avg. mAP50	Avg. Precision	Avg. Recall	Onset Time NRMSE	Duration NRMSE	Center Frequency NRMSE	Bandwidth NRMSE	Matched Targets
YOLOv8-L	0.561	0.581	0.591	0.00779	0.01484	0.000225	0.03952	2507
C2D-RT-DETR	**0.578**	**0.619**	**0.598**	**0.00181**	**0.02022**	**0.00023**	**0.04217**	**4118**

Table 5. Multi-class Classification Metrics and Parameter Estimation Accuracy at 0 dB（4424 targets）

Model	Avg. mAP50	Avg. Precision	Avg. Recall	Onset Time NRMSE	Duration NRMSE	Center Frequency NRMSE	Bandwidth NRMSE	Matched Targets
RT-DETR-L	0.656	0.653	0.656	0.00137	0.02055	0.00022	0.03153	4179
YOLOv11-L	0.613	0.705	0.628	0.00084	0.01425	0.00021	0.03122	2798
YOLOv10-L	0.606	0.706	0.614	0.00142	0.01688	0.00018	0.02731	2812
YOLOv8-L	0.629	0.679	0.649	0.00132	0.01997	0.00035	0.03983	2807
C2D-RT-DETR	**0.687**	**0.735**	**0.711**	**0.00157**	**0.01956**	**0.00018**	**0.02687**	**4107**

Table 6. Multi-class Classification Metrics and Parameter Estimation Accuracy at 5 dB（4539 targets）

Model	Avg. mAP50	Avg. Precision	Avg. Recall	Onset Time NRMSE	Duration NRMSE	Center Frequency NRMSE	Bandwidth NRMSE	Matched Targets
RT-DETR-L	0.672	0.718	0.685	0.00113	0.01330	0.00030	0.04655	3977
YOLOv11-L	0.618	0.688	0.635	0.00108	0.01206	0.00028	0.02503	2923
YOLOv10-L	0.611	0.722	0.596	0.00158	0.01580	0.00022	0.03408	2897
YOLOv8-L	0.623	0.693	0.63	0.00074	0.00988	0.00028	0.03766	2912
C2D-RT-DETR	**0.701**	**0.742**	**0.72**	**0.0019**	**0.01688**	**0.00019**	**0.03211**	**4062**

An analysis of the data in Tables 4 through 6 reveals that the primary breakthrough of C2D-RT-DETR lies in its exceptional target capture capability under extreme operating conditions. Taking the 0 dB SNR scenario as an example, where the test set contains a total of 4,424 valid signals, the number of matched targets for YOLOv10-L drops sharply to 2,812, resulting in a miss rate approaching 36.4%. In contrast, C2D-RT-DETR successfully detects 4,107 targets, achieving a matching coverage rate as high as 92.8%. This pronounced disparity indicates that conventional YOLO models-relying on fixed

receptive fields and NMS post-processing-are prone to an overall collapse in prediction confidence when faced with signal fragmentation caused by heavy noise corruption. It is worth noting that the marginally lower errors occasionally exhibited by the YOLO series on specific parameters are fundamentally attributed to their limitation of detecting only easy samples characterized by locally high SNRs and sharp boundaries. In comparison, by leveraging the global context modeling and bipartite graph matching mechanism of the Transformer architecture, C2D-RT-DETR overcomes the bottleneck of missed detections, achieving high recall in tandem with high precision.

Under typical electromagnetic interference at 0 dB, the model exhibits a fine-grained capability in reconstructing physical parameters. Specifically, the NRMSE of C2D-RT-DETR for baseband center frequency and start time is confined to 0.018% and 0.157%, respectively, demonstrating its ability to penetrate Gaussian noise cover and accurately capture both the geometric energy centers and abrupt onset transitions of the signals. For boundary-sensitive metrics that are vulnerable to frequency-domain dispersion and aliasing, the model maintains bandwidth and duration NRMSEs of only 2.687% and 1.956%, respectively. This outstanding performance is credited to the concentrated spatial direction-aware sensing of the CA module, the adaptive reconstruction of shallow high-frequency details by CARAFE, and the dynamic decoupling of classification and localization feature streams by the DyHead module. Their collaborative synergy allows the predicted bounding boxes to tightly envelope the true time-frequency distribution.

When the channel condition deteriorates to the extreme low-SNR threshold of -15 dB, C2D-RT-DETR still exhibits noise resilience that conventional pure convolutional networks cannot achieve. Under this condition, although the topological structure of the signals is heavily fragmented by intense noise-causing the target match counts of the YOLO series to plummet to a trough of 2,400~2,500,C2D-RT-DETR consistently matches 4,118 targets, maintaining a high match rate of 93.1%. Simultaneously, its center frequency error remains at an exceptionally low level of 0.023%, while the bandwidth error increases only marginally to 4.217%. This highlights that the integration of CA coordinate attention and CARAFE reassembly kernels empowers the network to perceive true physical boundaries even within highly discrete and distorted residual energy distributions.

Overall, across the entire SNR progression from -15 dB to 5 dB, the estimation errors of C2D-RT-DETR for core physical parameters-including center frequency, bandwidth, start time, and duration-consistently remain within a minimal range. The superior recall and stability demonstrated across this

broad SNR span not only validate the feasibility of the parameter mapping mechanism, but also underscore the engineering deployment potential of this model for field-ready spectrum monitoring equipment in complex electromagnetic environments.

4.5 Ablation Experiments

To quantitatively analyze the individual contributions and synergistic effects of each enhancement module in the C2D-RT-DETR architecture toward wideband signal detection and parameter estimation performance, multi-dimensional ablation experiments were conducted using the standard RT-DETR-L as the baseline model, under strictly consistent training strategies and hyperparameters. The experimental evaluation is divided into two phases: first, taking the 0 dB SNR scenario as a benchmark, progressive module ablation and combination tests were performed to evaluate the enhancement effects of CA, content-aware upsampling, and the dynamic feature head in the dimensions of feature extraction and reconstruction; second, the evaluation scope was extended across the dynamic SNR range from -15 dB to 5 dB to plot performance curves, thereby verifying the robustness and anti-noise ceiling of the proposed C2D network under severe noise corruption.

4.5.1 Validation of Enhanced Modules and Computational Complexity Evaluation

The comparative performance and computational metrics for various network configurations under the 0 dB SNR condition are detailed in Table 7.

Table 7. Ablation Experiment Results(0dB)

Experiment	CA	CARAFE	DyHead	Parameters	GFLOPs	P	R	mAP50	mAP50-95
1	-	-	-	32,002,235	103.5	0.717	0.653	0.656	0.517
2	√	-	-	32,015,595	103.5	0.708	0.712	0.676	0.542
3	-	√	-	32,150,531	103.8	0.709	0.684	0.661	0.537
4	-	-	√	32,425,147	103.2	0.693	0.667	0.663	0.546
5	√	√	-	32,163,891	103.8	0.718	0.717	0.684	0.57
6	√	-	√	32,616,171	104.7	0.716	0.678	0.676	0.562
7	-	√	√	32,751,107	104.9	0.716	0.671	0.661	0.538
8	√	√	√	**32,764,467**	**105.0**	**0.735**	**0.711**	**0.687**	**0.588**

As evidenced by the data in Table 7, the independent introduction of each single enhancement module yields varying degrees of performance gains over the baseline, while exhibiting distinct operational mechanisms and resource overheads:

Experiment 2 demonstrates the most prominent individual performance increment. With the sole incorporation of CA, the model's mAP_{50} increases significantly from 0.656 to 0.676, and the recall rate (R) jumps to 0.712, representing a 5.9% improvement over the baseline's 0.653. Crucially, the CA module introduces a negligible parameter overhead of only about 0.013 M, with the computational cost

remaining unchanged at 103.5 GFLOPs. This indicates that positional feature attention can effectively suppress the unconstrained spread of electromagnetic background noise at a virtually negligible computational cost by accurately pinpointing the 2D coordinate information of signals in the time-frequency domain.

Experiment 3 integrates only the CARAFE content-aware adaptive upsampling module, boosting mAP_{50} and mAP_{50-95} to 0.661 and 0.537, respectively. Because CARAFE replaces conventional nearest-neighbor interpolation by dynamically generating reconstruction kernels based on local feature semantics, the parameter count increases slightly to 32.150 M. This mechanism circumvents the aliasing distortion of high-frequency signal details during the sampling process, providing superior continuous representations for restoring weak edge features.

Experiment 4 introduces only DyHead, achieving the highest mAP_{50-95} among single-module configurations at 0.546, a 2.9% improvement over the baseline (0.517). However, because DyHead integrates multi-scale and multi-channel dynamic attention mechanisms in the detection head, the model parameter count expands to 32.425 M. This suggests that without front-end feature purification, deploying a heavy dynamic perception head at the back end enhances localization precision but offers limited improvements to the overall recall rate.

Building upon the validated efficacy of each individual module, Experiments 5 through 7 further investigate the combinatorial synergies between dual-module configurations. The results reveal that complementary fusion across multiple dimensions breaks the performance bottlenecks of single-dimension improvements. For example, Experiment 5, combining CA and CARAFE, pushes mAP_{50} further to 0.684, with mAP_{50-95} leaping substantially to 0.570; meanwhile, Experiment 6, integrating CA and DyHead, solidifies mAP_{50} at 0.676 at a computational cost of 104.7 GFLOPs.

Ultimately, the fully integrated C2D-RT-DETR architecture-combining CA, CARAFE, and DyHead-achieves comprehensively superior performance across all benchmarks, with precision (P), mAP_{50}, and mAP_{50-95} reaching the highest levels among all tested groups. From a systems engineering perspective, this optimal performance stems from the organic synergy among the three cascaded stages: time-frequency localization, feature reconstruction, and dynamic head modulation. The complete architecture entails a final parameter count of 32.764 M and a computational complexity of 105.0 GFLOPs, successfully delivering a substantial leap in signal sensing capability within complex electromagnetic environments at minimal cost.

To thoroughly assess the engineering feasibility of the model for real-world deployment and satisfy the stringent real-time requirements of wideband spectrum sensing tasks, this section conducts comprehensive benchmark evaluations of computational cost and inference latency across a unified hardware platform. The proposed C2D-RT-DETR framework is compared against mainstream models, including YOLOv8-L, YOLOv10-L, YOLOv11-L, and the official RT-DETR-L baseline. The evaluation metrics encompass model parameters, computational complexity (GFLOPs), and single-frame GPU inference latency. The input resolution for all models was uniformly set to 640×640 with a batch size of 1. The detailed benchmark results are presented in Table 8.

Table8.Model complexity and inference time.

Model	Parameters	GFLOPs	GPU Time (ms)
YOLOv11-L	25,286,251	86.6	10.83
YOLOv10-L	25,779,206	127.3	11.06
YOLOv8-L	43,613,547	164.9	6.86
RT-DETR-L	32,002,235	103.5	17.81
C2D-RT-DETR	**32,764,467**	**105.0**	**19.46**

Table 8 reports the quantitative evaluation results of different models regarding hardware resource consumption and inference speed. Comparative analysis reveals that, compared with the baseline model RT-DETR-L, the proposed C2D-RT-DETR architecture successfully maintains hardware overhead at a low level through the lightweight integration of the CA, CARAFE, and DyHead modules, while substantially enhancing detection efficacy under low SNR conditions. Specifically, the model parameters increase marginally from 32.00 M to 32.76 M, and the computational complexity increases from 103.5 GFLOPs to 105.0 GFLOPs, demonstrating superior computational cost-effectiveness. In terms of inference latency, due to the matrix multiplications in the multi-dimensional dynamic perception head of DyHead and the GPU memory access overhead of the dynamic reconstruction kernels in CARAFE, the single-frame inference time of C2D-RT-DETR slightly rises from 17.81 ms to 19.46 ms, with the corresponding throughput remaining at 51.4 FPS. Although this latency is slightly higher than that of the pure convolutional YOLO series, it well satisfies the engineering threshold of modern wideband spectrum sensing systems for real-time edge detection (sub-30 ms). Furthermore, a horizontal comparison with state-of-the-art YOLO architectures demonstrates that the parameters and computational complexity of C2D-RT-DETR are reduced by approximately 24.88% and 36.33%, respectively, compared to the classical large-scale model YOLOv8-L, thereby alleviating the memory and bandwidth burdens on edge devices. Concurrently, its computational complexity of 105.0 GFLOPs is markedly lower than the 127.3 GFLOPs of YOLOv10-L, exhibiting

superior computational efficiency within the Transformer paradigm.

In summary, at the expense of negligible computational and parameter increments, the C2D-RT-DETR architecture breaks through the perception bottlenecks of conventional detectors in strong noise environments. It achieves a finely tuned balance between anti-noise robustness in complex electromagnetic scenarios and real-time inference efficiency on edge hardware, demonstrating deployment and application potential for non-cooperative spectrum monitoring terminals.

4.5.2 Analysis of Effective Operating Regimes for Enhanced Modules

To further reveal the robustness boundaries and perceptual limits of the proposed C2D-RT-DETR architecture under low SNR conditions, this section systematically tracks the dynamic trajectory of mAP_{50} across a broad SNR span ranging from -15 dB to 5 dB under different ablation configurations. The comparative performance curves and degradation trends are illustrated in Fig. 11.

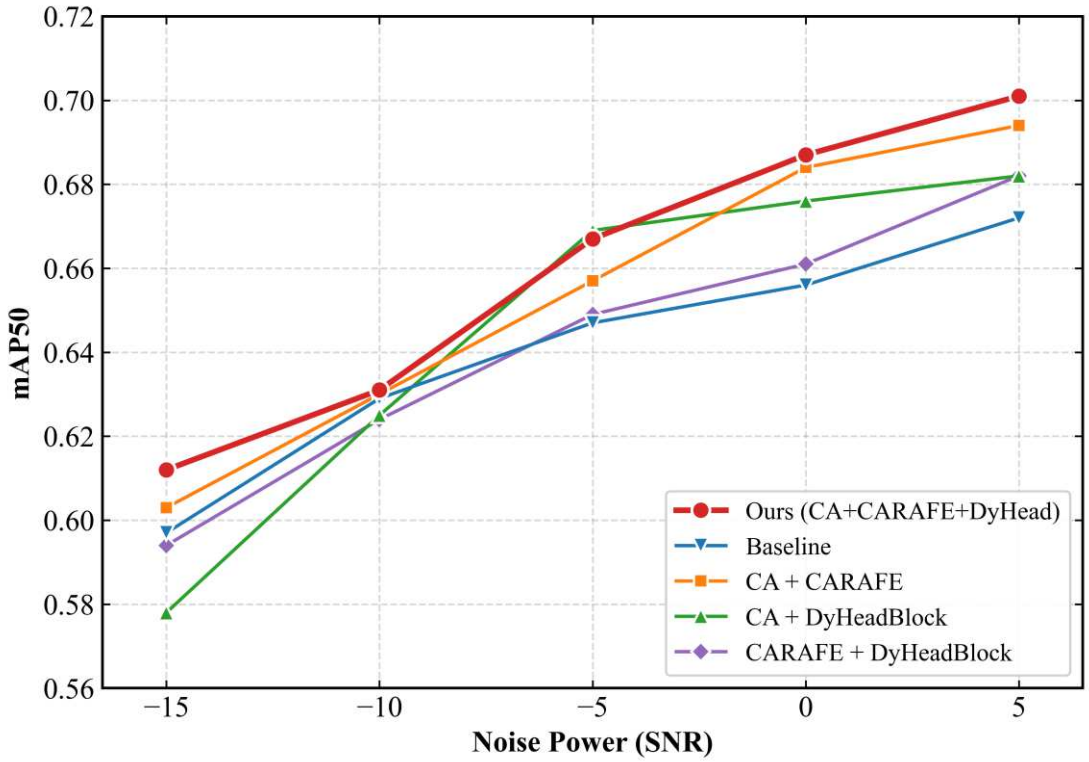


Fig. 11. Ablation trajectory curves and performance gain profiles across different SNRs.

Overall, the proposed C2D-RT-DETR consistently maintains superior detection accuracy across the entire SNR span, validating the strong adaptability of the multi-module collaborative strategy across diverse noise scenarios. Within the higher SNR range of 0 dB to 5 dB, C2D-RT-DETR exhibits substantial performance advantages. Taking the 0 dB condition as an example, the complete architecture achieves an mAP_{50} of 0.687, representing a 3.1 percentage point gain over the baseline

model.

Within the core operational reconnaissance regime of -10 dB to 5 dB, C2D-RT-DETR displays a notably steeper performance growth trajectory. As noise interference gradually diminishes, the architecture captures emerging weak signal features from the background noise with enhanced sensitivity, rapidly widening the accuracy margin compared to both the baseline and dual-module combinations. This indicates that the composite architecture enables highly efficient semantic aggregation when handling wideband time-frequency spectrograms corrupted by strong interference, breaking through the performance bottlenecks inherent in single- or dual-module setups.

Notably, when the electromagnetic environment deteriorates to the critical low-SNR limit of -15 dB, where the noise power significantly exceeds the signal energy threshold and true time-frequency structures are largely submerged within the random noise field, the performance margins of all ablation variants undergo varying degrees of compression. Nevertheless, C2D-RT-DETR still attains the highest mAP_{50} of 0.612, preserving an approximate 1.5 percentage point advantage over the baseline and markedly outperforming configurations lacking shallow-level denoising capability or effective feature reconstruction. This physical phenomenon confirms that the foundational representations established by CA and CARAFE at the front end effectively prevent DyHead from falling into disordered sampling when energy gradients are absent, ensuring that the model retains robust anti-noise resilience and a solid representation baseline even near the physical limits of detectability.

5 Conclusion

To tackle the vulnerabilities of wideband signals to severe noise interference and the challenges of boundary alignment in complex electromagnetic environments, this paper proposes a real-time signal detection framework based on content-aware reassembly and multi-dimensional coordinate alignment, termed C2D-RT-DETR. Built upon RT-DETR-L, the proposed architecture introduces the CA mechanism at the front end of feature extraction, deploys the CARAFE module along the cross-scale fusion path, and integrates the DyHead at the output stage. This establishes a synergistically optimized network spanning spatial coordinate localization, lossless shallow topological reconstruction, and dynamic feature decoupling, thereby significantly enhancing the model's capabilities in time-frequency edge regression and modulation classification under low-SNR conditions.

Comparative and ablation experiments demonstrate that the proposed C2D-RT-DETR architecture achieves superior perceptual identification and target recall coverage across the SNR range from -15

dB to 5 dB. The investigation further confirms that the CA module accurately locks the time-frequency orientation of signals via orthogonal directional encoding, the CARAFE module prevents the blurring of high-frequency textures caused by conventional interpolation, and the DyHead module effectively alleviates task misalignment between classification and localization. Combined with the inverse mapping mechanism from geometric coordinates to physical quantities, the model achieves high-precision reconstruction of essential physical parameters, including center frequency, occupied bandwidth, start time, and duration.

In terms of engineering deployment overhead, the complete C2D-RT-DETR architecture entails 32.76 M parameters and a computational complexity of 105.0 GFLOPs, representing marginal increases of only 2.38% and 1.45% over the baseline model, respectively. Its single-frame GPU inference latency of 19.46 ms well satisfies the strict latency requirements for real-time spectrum reconnaissance. Notably, when the operating environment deteriorates to the extreme low-SNR regime of -15 dB and below, the marginal gains of the feature enhancement modules slightly contract due to the physical sensing limits of short-time observation windows and the heavy submergence by Gaussian white noise. Nevertheless, leveraging the global context modeling capability of the Transformer, the architecture consistently maintains high target recall coverage.

In summary, the technical solution presented in this work establishes a reliable mapping from visual geometric boundaries to wideband signal physical parameters and achieves a well-balanced trade-off between recognition robustness and edge computing overhead, providing a practical engineering reference for the intelligent deployment of non-cooperative wideband spectrum monitoring terminals.

Research Funding
This research did not receive funding.
Author Contributions
Z.S. conceived the idea, designed the network architecture, conducted the experiments, and wrote the original draft. R.Z. supervised the project, provided resources, and critically revised the manuscript. L.L. and Y.C. contributed to data generation and signal preprocessing. L.G. and H.G. performed the ablation studies and helped analyze the experimental results. All authors reviewed and approved the final version of the manuscript.
Competing interests
The authors declare no competing interests.
Data Availability Statement
The dataset generation scripts used in this study are openly available in the GitHub repository at https://github.com/Chrislo09-netizen/C2D-RT-DETR-Data-Generation. The full simulated datasets generated during the current study are available from the corresponding author upon reasonable request.

References
[1] T. Yucek and H. Arslan, 'A survey of spectrum sensing algorithms for cognitive radio applications', IEEE Commun. Surv. Tutor., vol. 11, no. 1, pp. 116–130, 2009, doi: 10.1109/SURV.2009.090109.
[2] S. Haykin, 'Cognitive radio: brain-empowered wireless communications', IEEE J. Sel. Areas Commun., vol. 23, no. 2, pp. 201–220, Feb. 2005, doi: 10.1109/JSAC.2004.839380.
[3] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, 'NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey', Comput. Netw., vol. 50, no. 13, pp. 2127–2159, Sep. 2006, doi: 10.1016/j.comnet.2006.05.001.
[4] H. Sun, A. Nallanathan, C.-X. Wang, and Y. Chen, 'Wideband spectrum sensing for cognitive radio networks: a survey', IEEE Wirel. Commun., vol. 20, no. 2, pp. 74–81, Apr. 2013, doi: 10.1109/MWC.2013.6507397.
[5] A. K. Nandi and E. E. Azzouz, 'Algorithms for automatic modulation recognition of communication signals', IEEE Trans. Commun., vol. 46, no. 4, pp. 431–436, Apr. 1998, doi: 10.1109/26.664294.
[6] W. Wei and J. M. Mendel, 'Maximum-likelihood classification for digital amplitude-phase modulations', IEEE Trans. Commun., vol. 48, no. 2, pp. 189–193, Feb. 2000, doi: 10.1109/26.823550.
[7] A. Swami and B. M. Sadler, 'Hierarchical digital modulation classification using cumulants', IEEE Trans. Commun., vol. 48, no. 3, pp. 416–429, Mar. 2000, doi: 10.1109/26.837045.
[8] V. D. Orlic and M. L. Dukic, 'Automatic modulation classification algorithm using higher-order cumulants under real-world channel conditions', IEEE Commun. Lett., vol. 13, no. 12, pp. 917–919, Dec. 2009, doi: 10.1109/LCOMM.2009.12.091711.

[9] J. Liu and Q. Luo, ‘A novel modulation classification algorithm based on daubechies5 wavelet and fractional fourier transform in cognitive radio’, in 2012 IEEE 14th International Conference on Communication Technology, Nov. 2012, pp. 115–120, doi: 10.1109/ICCT.2012.6511199.
[10] M. W. Aslam, Z. Zhu, and A. K. Nandi, ‘Automatic Modulation Classification Using Combination of Genetic Programming and KNN’, IEEE Trans. Wirel. Commun., vol. 11, no. 8, pp. 2742–2750, Aug. 2012, doi: 10.1109/TWC.2012.060412.110460.
[11] B. Boashash, ‘Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals’, Proc. IEEE, vol. 80, no. 4, pp. 520–538, Apr. 1992, doi: 10.1109/5.135376.
[12] A. Akan and O. Karabiber Cura, ‘Time–frequency signal processing: Today and future’, Digit. Signal Process., vol. 119, p. 103216, Dec. 2021, doi: 10.1016/j.dsp.2021.103216.
[13] T. O’Shea and J. Hoydis, ‘An Introduction to Deep Learning for the Physical Layer’, IEEE Trans. Cogn. Commun. Netw., vol. 3, no. 4, pp. 563–575, Dec. 2017, doi: 10.1109/TCCN.2017.2758370.
[14] T. J. O’Shea, J. Corgan, and T. C. Clancy, ‘Convolutional Radio Modulation Recognition Networks’, in Engineering Applications of Neural Networks, C. Jayne and L. Iliadis, Eds., Cham: Springer International Publishing, 2016, pp. 213–226, doi: 10.1007/978-3-319-44188-7_16.
[15] T. J. O’Shea, T. Roy, and T. C. Clancy, ‘Over the Air Deep Learning Based Radio Signal Classification’, IEEE J. Sel. Top. Signal Process., vol. 12, no. 1, pp. 168–179, Feb. 2018, doi: 10.1109/JSTSP.2018.2797022.
[16] F. Meng, P. Chen, L. Wu, and X. Wang, ‘Automatic Modulation Classification: A Deep Learning Enabled Approach’, IEEE Trans. Veh. Technol., vol. 67, no. 11, pp. 10760–10772, Nov. 2018, doi: 10.1109/TVT.2018.2868698.
[17] S. Peng et al., ‘Modulation Classification Based on Signal Constellation Diagrams and Deep Learning’, IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 3, pp. 718–727, Mar. 2019, doi: 10.1109/TNNLS.2018.2850703.
[18] T. Huynh-The et al., ‘Automatic Modulation Classification: A Deep Architecture Survey’, IEEE Access, vol. 9, pp. 142950–142971, 2021, doi: 10.1109/ACCESS.2021.3120419.
[19] C. Zhang, P. Patras, and H. Haddadi, ‘Deep Learning in Mobile and Wireless Networking: A Survey’, IEEE Commun. Surv. Tutor., vol. 21, no. 3, pp. 2224–2287, 2019, doi: 10.1109/COMST.2019.2904897.
[20] N. C. Luong et al., ‘Applications of Deep Reinforcement Learning in Communications and Networking: A Survey’, IEEE Commun. Surv. Tutor., vol. 21, no. 4, pp. 3133–3174, 2019, doi: 10.1109/COMST.2019.2916583.
[21] W. Lee, M. Kim, and D.-H. Cho, ‘Deep Cooperative Sensing: Cooperative Spectrum Sensing Based on Convolutional Neural Networks’, IEEE Trans. Veh. Technol., vol. 68, no. 3, pp. 3005–3009, Mar. 2019, doi: 10.1109/TVT.2019.2891291.
[22] R. Mei and Z. Wang, ‘Deep Learning-Based Wideband Spectrum Sensing: A Low Computational Complexity Approach’, IEEE Commun. Lett., vol. 27, no. 10, pp. 2633–2637, Oct. 2023, doi: 10.1109/LCOMM.2023.3310715.

[23] S. Guo, Z. Hu, X. Li, X. Liang, Y. Zhao, and Z. Huang, ‘Robust Wideband Spectrum Sensing in Low-SNR Environments With a Dynamic Convolution-Augmented Transformer’, IEEE Trans. Instrum. Meas., vol. 75, pp. 6510515–6510515, 2026, doi: 10.1109/TIM.2026.3699718.
[24] A. Vaswani et al., ‘Attention Is All You Need’, Aug. 02, 2023, arXiv: arXiv:1706.03762, doi: 10.48550/arXiv.1706.03762.
[25] W. Ma, Z. Cai, and C. Wang, ‘A Transformer and Convolution-Based Learning Framework for Automatic Modulation Classification’, IEEE Commun. Lett., vol. 28, no. 6, pp. 1392–1396, Jun. 2024, doi: 10.1109/LCOMM.2024.3380623.
[26] W. Kong, X. Jiao, Y. Xu, and Q. Yang, ‘An Effective Masked Transformer Model for Automatic Modulation Recognition’, IEEE Trans. Cogn. Commun. Netw., vol. 12, pp. 128–143, 2026, doi: 10.1109/TCCN.2025.3550729.
[27] S. Ansari et al., ‘Attention-Enhanced Hybrid Automatic Modulation Classification for Advanced Wireless Communication Systems: A Deep Learning-Transformer Framework’, IEEE Access, vol. 13, pp. 105463–105491, 2025, doi: 10.1109/ACCESS.2025.3580574.
[28] Q. Zheng, P. Zhao, H. Wang, A. Elhanashi, and S. Saponara, ‘Fine-Grained Modulation Classification Using Multi-Scale Radio Transformer With Dual-Channel Representation’, IEEE Commun. Lett., vol. 26, no. 6, pp. 1298–1302, Jun. 2022, doi: 10.1109/LCOMM.2022.3145647.
[29] Z. Chen, X. Zhang, and K. He, ‘Multi-Channel Convolutional Distilled Transformer for Automatic Modulation Classification’, in 2024 International Joint Conference on Neural Networks (IJCNN), Jun. 2024, pp. 1–8, doi: 10.1109/IJCNN60899.2024.10650112.
[30] R. Luo, T. Hu, Z. Tang, C. Wang, X. Gong, and H. Tu, ‘A Radio Signal Modulation Recognition Algorithm Based on Residual Networks and Attention Mechanisms’, Sep. 27, 2019, arXiv: arXiv:1909.12472, doi: 10.48550/arXiv.1909.12472.
[31] X. Zhao, X. Zhou, J. Xiong, F. Li, and L. Wang, ‘Automatic Modulation Recognition Based on Multi-Dimensional Feature Extraction’, in 2020 International Conference on Wireless Communications and Signal Processing (WCSP), Oct. 2020, pp. 823–828, doi: 10.1109/WCSP49889.2020.9299797.
[32] A. Vagollari, V. Schram, W. Wicke, M. Hirschbeck, and W. Gerstacker, ‘Joint Detection and Classification of RF Signals Using Deep Learning’, in 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Apr. 2021, pp. 1–7, doi: 10.1109/VTC2021-Spring51267.2021.9449073.
[33] X. Kang, H. Chen, G. Chen, K.-C. Chang, and T. M. Clemons, ‘Optimal Preprocessing for Joint Detection and Classification of Wireless Communication Signals in Congested Spectrum Using Computer Vision Methods’, Aug. 13, 2024, arXiv: arXiv:2408.06545, doi: 10.48550/arXiv.2408.06545.
[34] T. Huynh-The, P.-T. Huynh, and Q.-V. Pham, ‘Spec-YOLO: An Efficient Deep Network for Spectrogram-Based Signals Identification’, IEEE Wirel. Commun. Lett., vol. 14, no. 11, pp. 3550–3554, Nov. 2025, doi: 10.1109/LWC.2025.3598052.
[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ‘You Only Look Once: Unified, Real-Time Object Detection’, in 2016 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR), Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
[36] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, ‘YOLOv4: Optimal Speed and Accuracy of Object Detection’, Apr. 23, 2020, arXiv: arXiv:2004.10934, doi: 10.48550/arXiv.2004.10934.
[37] J. Terven and D. Cordova-Esparza, ‘A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS’, Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.
[38] A. Wang et al., ‘YOLOv10: Real-Time End-to-End Object Detection’, Oct. 30, 2024, arXiv: arXiv:2405.14458, doi: 10.48550/arXiv.2405.14458.
[39] R. Sapkota and M. Karkee, ‘Ultralytics YOLO Evolution: An Overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 Object Detectors for Computer Vision and Pattern Recognition’, Mar. 16, 2026, arXiv: arXiv:2510.09653, doi: 10.48550/arXiv.2510.09653.
[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, ‘End-to-End Object Detection with Transformers’, May 28, 2020, arXiv: arXiv:2005.12872, doi: 10.48550/arXiv.2005.12872.
[41] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, ‘Deformable DETR: Deformable Transformers for End-to-End Object Detection’, Mar. 18, 2021, arXiv: arXiv:2010.04159, doi: 10.48550/arXiv.2010.04159.
[42] Y. Zhao et al., ‘DETRs Beat YOLOs on Real-time Object Detection’, in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2024, pp. 16965–16974, doi: 10.48550/arXiv.2304.08069.
[43] Q. Hou, D. Zhou, and J. Feng, ‘Coordinate Attention for Efficient Mobile Network Design’, in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 13708–13717, doi: 10.1109/CVPR46437.2021.01350.
[44] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, ‘CARAFE: Content-Aware ReAssembly of FEatures’, Oct. 29, 2019, arXiv: arXiv:1905.02188, doi: 10.48550/arXiv.1905.02188.
[45] X. Dai et al., ‘Dynamic Head: Unifying Object Detection Heads with Attentions’, in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 7369–7378, doi: 10.1109/CVPR46437.2021.00729.