Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

RT-DETR-Based Wideband Signal Detection and Modulation Classification

Minghao Cao, Peng Chu *, Pengfei Ma, Bo Fang

School of Electronic Information, XiJing University, Xi'an, Shaanxi, China

*** Corresponding author:** Peng Chu

Abstract: In To address the problems of high computational complexity, low accuracy, and the cumbersome manual feature extraction process in traditional machine learning methods for communication signal modulation recognition, this study proposes a deep learning-based end-to-end recognition model. Built upon the Transformer architecture using the RT-DETR framework, the model directly identifies modulation types from sampled communication signals. It features high recognition accuracy, strong generalization ability, robustness to noise, and a streamlined processing pipeline. Extensive experiments validate the model's effectiveness, demonstrating its superior performance in automatic feature extraction and modulation classification compared to traditional approaches.

Keywords: Feature Extraction Networks; Signal Modulation Recognition; Deep Learning.

1. Introduction

With the rapid advancement of wireless communication and signal processing technologies, signal modulation recognition has become a critical task in various fields, including military communications, radio monitoring, and cognitive radio. Its primary goal is to extract modulation characteristics from received signals and accurately determine their modulation types. Traditional modulation recognition methods often rely on manually crafted feature extraction combined with classical classification algorithms. However, these approaches struggle to maintain high performance in complex electromagnetic environments, resulting in limited recognition accuracy[1].

In recent years, the emergence of deep learning has brought new opportunities to this field. Deep neural networks have shown remarkable capability in automatically learning signal features, offering enhanced robustness and higher accuracy under challenging conditions[2]. Despite these advantages, many existing deep learning-based methods still face trade-offs between model complexity, computational efficiency, and recognition accuracy[3]. Therefore, designing a lightweight and efficient deep learning model that achieves high recognition accuracy remains a pressing research challenge.

In this study, we propose a novel signal modulation recognition method based on the RT-DETR architecture. RT-DETR, originally designed for real-time object detection, provides a strong balance between computational efficiency and generalization ability[4]. To further improve its performance for modulation recognition tasks, we introduce two key modifications:

1.EfficientFormerV2 is employed as the backbone network in place of ResNet-18, enhancing feature extraction while maintaining a lightweight structure [5].

2.The Dynamic-range Histogram Self-Attention (DHS-Attention) module, inspired by HistoFormer, is integrated to replace the original AIFI [6]. This module improves attention distribution and feature fusion through dynamic histogram-based modeling.

These improvements enable the model to better capture modulation features while maintaining low computational cost.

The main contributions of this study are summarized as follows:

We propose a lightweight and efficient signal modulation recognition method based on RT-DETR, integrating EfficientFormerV2 and DHS-Attention to enhance feature extraction and fusion capabilities.

We design a comprehensive set of experiments to evaluate the proposed model across various modulation types and SNR conditions, and perform comparative analyses with traditional methods and baseline models.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 introduces the proposed method in detail, and Section 4 presents the experimental setup, results, and analysis.

2. Related Research

2.1. Traditional Signal Modulation Recognition Methods

Signal modulation recognition plays a vital role in wireless communications. Early research primarily relied on traditional handcrafted feature extraction combined with classical classification algorithms. Common approaches include recognition based on statistical features, instantaneous parameters, and cyclostationary feature extraction. These methods typically derive frequency-domain, time-domain, or instantaneous characteristics of signals—such as amplitude, phase, frequency, and instantaneous power—to distinguish between different modulation types. For example, techniques based on higher-order cumulants (HOC) have shown strong performance in modulation recognition, particularly in high signal-to-noise ratio (SNR) environments.

However, these traditional methods heavily depend on the accuracy and robustness of manual feature extraction. Their performance often degrades significantly in complex and dynamic electromagnetic environments. To address these limitations, machine learning approaches—such as Support Vector Machines (SVMs) and decision trees—have been

widely applied. These models classify signals by feeding manually extracted features into the classifiers. Nevertheless, handcrafted feature extraction remains a critical bottleneck, especially when signals are distorted by noise, interference, or multipath effects, leading to substantial drops in recognition performance under challenging conditions.

2.2. Deep Learning-Based Signal Modulation Recognition

With the rapid development of deep learning, researchers have begun exploring the use of deep neural networks to automatically extract features from signals, addressing the shortcomings of traditional handcrafted feature extraction. Convolutional Neural Networks (CNNs), known for their excellent performance in image processing, were introduced into signal modulation recognition. O'Shea et al. first proposed using CNNs to directly learn features from the I/Q data of signals and perform modulation classification. This method achieved remarkable results on multiple public datasets.

Subsequently, Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory networks (LSTMs), were introduced to capture temporal information in signals [9]. By modeling the temporal dependencies of the signals, these methods have demonstrated strong robustness in low SNR environments. However, RNN-based models are time-consuming to train and are prone to gradient vanishing or exploding issues.

In recent years, hybrid architectures combining CNNs and RNNs, such as CNN-RNN and CNN-LSTM models, have gained attention. These models improve modulation recognition accuracy by fusing spatial and temporal features. Although these deep learning methods outperform traditional methods, they often require substantial computational resources and training time, posing challenges for real-time applications.

2.3. Lightweight Networks and Transformer Applications

To address the high computational complexity of deep learning models, lightweight network structures have been widely researched. Lightweight CNN models, such as MobileNet and SqueezeNet, reduce the number of parameters and computation, enabling efficient performance on resource-constrained platforms like mobile devices. While these models perform well in specific application scenarios, they still struggle with accuracy when handling complex signal modulation recognition tasks.

The Transformer structure, known for its success in natural language processing, has garnered attention and has been gradually introduced into visual tasks. Its self-attention mechanism can capture global features without relying on temporal sequences, providing new approaches to feature extraction. RT-DETR, a lightweight object detection model that integrates the advantages of the Transformer, can extract rich feature information while maintaining computational efficiency, making it highly promising for signal modulation recognition tasks.

2.4. Baseline Model – RT-DETR

RT-DETR (Real-Time DEtection TRansformer) is a recent end-to-end object detection model that eliminates the need for post-processing steps such as Non-Maximum Suppression (NMS). Built upon the DETR framework, RT-DETR introduces several architectural optimizations to achieve a better balance between accuracy and inference speed, making it the first transformer-based object detector to operate effectively in real-time. The model consists of four main components: the input module, backbone, neck encoder, and head decoder, as illustrated in Figure 1.

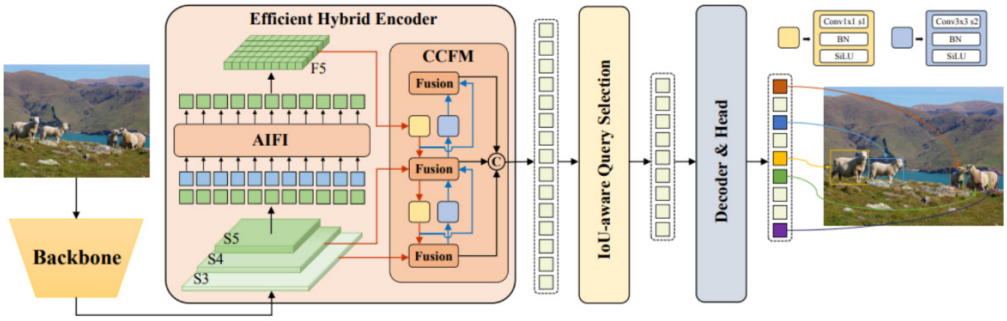

Figure 1. Overview of RT-DETR

2.4.1. Input Module

The input module is responsible for preprocessing raw images into a unified format suitable for the model. Typically, images are resized to a fixed resolution (e.g., 640 × 640) and normalized. To enhance generalization performance, a variety of data augmentation techniques are applied during training

2.4.2. Backbone

The backbone of RT-DETR is a convolutional neural network used to extract multi-scale semantic features from the input image. Common backbone options include ResNet-50 and ResNet-101. The model utilizes three feature levels (denoted as S3, S4, and S5), each corresponding to a different receptive field and semantic level.

2.4.3. Neck Encoder

To address the inefficiency of the original DETR encoder, RT-DETR proposes an Efficient Hybrid Encoder consisting of two components:

AIFI (Attention-based Intra-scale Feature Interaction): Applies self-attention only to high-level features (S5), modeling long-range dependencies while reducing sequence length and computation cost;

CCFF (CNN-based Cross-scale Feature Fusion): A lightweight convolutional module that fuses S3, S4, and the transformed S5 features to integrate fine-grained and semantic information across scales.

This decoupled design of intra-scale and cross-scale operations ensures efficient computation without sacrificing detection performance.

2.4.4. Head Decoder

The head decoder predicts object categories and bounding boxes based on the features processed by the encoder. It retains the standard Transformer decoder structure and introduces a novel Uncertainty-Minimal Query Selection mechanism, characterized by:

Evaluating both classification and localization confidence to quantify the uncertainty of each encoder feature;

Selecting the most reliable features as object queries;

Iteratively refining these queries through multiple decoder layers to produce final predictions.

Unlike traditional detectors, RT-DETR eliminates the need for NMS by employing Hungarian Matching to establish a one-to-one correspondence between predictions and ground-truth objects. This enables fully end-to-end training and simplifies the detection pipeline.

2.5. Model Enhancement and Feature Fusion Techniques

To further enhance the performance of deep learning models in signal modulation recognition, researchers have proposed various enhancement and optimization strategies. Attention mechanisms have been widely applied in deep learning, particularly when dealing with signals with complex patterns, as they effectively enhance the expression of important features. AIFI, a dynamic attention mechanism, can adaptively adjust attention weights according to the characteristics of the input signal, improving the model's feature extraction capabilities.

Moreover, feature fusion techniques have been applied in multimodal signal processing and complex scene recognition. The AIFI-EfficientAdditive module improves traditional additive feature fusion methods, enhancing the synergy between features while maintaining computational efficiency, thereby improving modulation recognition accuracy.

2.6. Contributions of This Research

Building upon previous research, this study presents a novel signal modulation recognition method based on the RT-DETR architecture. By incorporating EfficientFormerV2 as the backbone network and integrating the Dynamic-range Histogram Self-Attention (DHS-Attention) mechanism, the proposed model achieves significant improvements in both recognition accuracy and computational efficiency, while maintaining a lightweight structure. Extensive experiments conducted under various complex channel conditions demonstrate that the proposed approach consistently outperforms traditional methods in terms of both classification performance and processing speed.

3. Wideband Modulated Signal Dataset

3.1. Data Preprocessing

Before performing signal modulation recognition, we preprocess the input signal data using the following methods:

Normalization: All input signal data is normalized to the [0,1] range to eliminate amplitude differences between signals, reducing their impact on the model[10].

Signal Framing and Overlapping: To capture richer temporal information, the input signal is divided into multiple frames, with a certain overlap between frames. Each frame of the signal is treated as an independent input sample.

Data Augmentation: To increase the diversity of the training data, we applied data augmentation techniques such as random noise addition and frequency shifting[11]. These augmentations help the model generalize better to different channel environments.

3.2. Network Architecture Design

The overall structure of the RT-DETR model is composed of the following parts:

Backbone Network: The original ResNet-18 backbone in RT-DETR was replaced with. EfficientFormerV2 provides a better trade-off between performance and efficiency, offering stronger feature representation while maintaining a lightweight structure. This allows the model to extract more discriminative time-frequency features, improving recognition accuracy for complex modulation types.

Transformer Encoder: After extracting primary features, the feature maps are input into the Transformer encoder. The self-attention mechanism of the Transformer allows it to effectively capture global information in the signal, enhancing recognition performance [13].

AIFI Module: To further refine the attention mechanism following the Transformer encoder, we replaced the AIFI module with the Dynamic-range Histogram Self-Attention (DHS-Attention) module, inspired by HistoFormer. DHS-Attention introduces histogram-based dynamic range modeling, which enables the network to capture long-range dependencies more effectively, enhancing its ability to distinguish subtle differences among modulation patterns.

Classifier: The final feature map is input into a fully connected layer for classification. This layer outputs the probability distribution corresponding to various modulation signals, with the predicted modulation type being the one with the highest probability.

3.3. Model Training and Optimization

We use the cross-entropy loss function as the optimization objective of the model. The model training process uses the Adam optimizer with the following hyperparameters:

Learning Rate: The initial learning rate is set to 0.001, and a cosine annealing strategy is used to dynamically adjust the learning rate to ensure stability and fast convergence during training.

Batch Size: The batch size is set to 4, striking a good balance between training speed and model performance.

Epochs: The model was trained for 250 epochs. After each epoch, the model's performance was evaluated on the validation set, and the model with the highest validation accuracy was saved as the final model.

Weight Decay: To prevent overfitting, a weight decay strategy with a decay rate of 0.0001 was employed.

Gradient Clipping: To avoid the gradient explosion problem, a gradient clipping strategy was used with a threshold set at 1.0.

Additionally, we applied an early stopping strategy during training. If the validation loss did not decrease for 10 consecutive epochs, the training process was terminated early[12].

3.4. Experimental Setup

To validate the effectiveness of the proposed improved RT-DETR model, a series of experiments were conducted on multiple public wideband signal datasets. These datasets cover a variety of modulation types and signal-to-noise ratio (SNR) conditions to ensure the robustness and generalization of the model.

We compared the performance of models equipped with different backbone networks, including ResNet-18 and the proposed EfficientFormerV2. Additionally, we evaluated the impact of replacing the original AIFI-DAttention module with the Dynamic-range Histogram Self-Attention (DHS-Attention) from HistoFormer.

Model performance was evaluated using standard metrics such as accuracy, precision, and recall. These experiments aim to demonstrate the benefits of the proposed architectural improvements in both recognition accuracy and efficiency under various SNR conditions.

3.5. Dataset Generation

The dataset for this study was generated through MATLAB simulations, with the signal-to-noise ratio (SNR) for all signals set at 30 dB. This SNR level represents a low-noise environment, suitable for testing the model's recognition ability under ideal conditions. The dataset generation process is as follows:

Modulation Types: The dataset includes multiple common digital modulation types, including BPSK, QPSK, 8PSK, 16QAM, and 64QAM.

Signal Length and Sampling Rate: Each signal length is set to 1024 sample points, with a sampling rate of 1 MHz. The generated signals are stored in complex I/Q data format for subsequent processing and feature extraction.

Data Augmentation: Although all signals have an SNR set at 30 dB, other simulation conditions such as frequency offset and phase shift were introduced through data augmentation, increasing the diversity and complexity of the data.

Dataset Splitting: The generated signal data is split into training, validation, and test sets in a 70%, 15%, and 15% ratio, ensuring an even distribution of modulation types across different datasets.

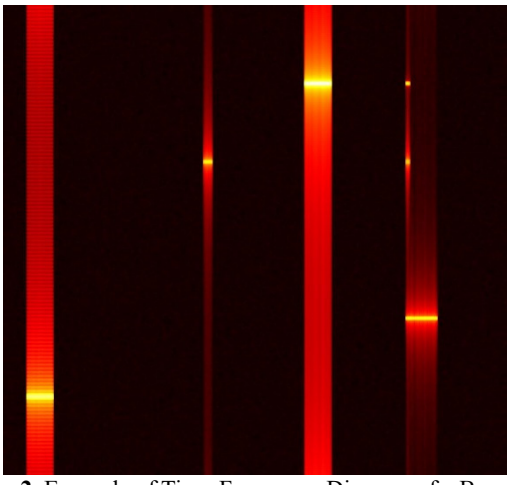

Figure 2. Example of Time-Frequency Diagram of a Broadband Signal

4. Experimental Process and Results Analysis

4.1. Training Process

In the training process, the data features of communication signals are first extracted using convolution windows in the convolutional neural network (CNN). The categorical cross-entropy loss function is employed, and the loss function $L_i$is defined as follows

$$L_i = -\sum_j t_{i,j} \log(p_{i,j}) \quad (1)$$

where t represents the true labels, i represents the input data, j represents the categories, and p represents the predicted results. This loss function is commonly used in multi-class classification tasks, such as when using the softmax function as the final output. The optimizer continuously reduces the loss function to update the parameters of the hidden layers. In this case, we selected the Adam optimizer, which is currently one of the most widely used optimizers[7]. According to the literature[16], this optimizer offers advantages such as low memory requirements, simple implementation, and high computational efficiency. The learning rate for the optimizer was set to a fixed value of 0.001, and a dropout rate of 0.5 was used to randomly deactivate neurons, preventing overfitting due to the large number of parameters in the fully connected layer.

The confusion matrix in Figure 3 shows the classification performance at a signal-to-noise ratio (SNR) of 30dB. The horizontal axis represents the nine predicted modulation categories, while the vertical axis represents the actual modulation categories. This confusion matrix provides a visual representation of how well the model correctly classifies each modulation type under the given conditions[8].

This process, combining the loss function, Adam optimizer, and dropout techniques, helped achieve better generalization and avoided overfitting during training.

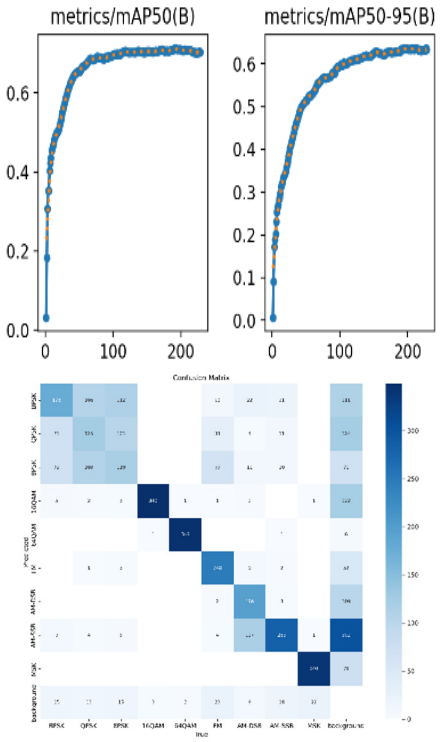

Figure 3. The result of a signal-to-noise ratio (SNR) of 30 dB.

4.2. Results Analysis

The neural network was trained for 250 epochs using the training dataset, while the test dataset was used to evaluate the model's performance after each epoch. As shown in Figure 4, the loss steadily decreases throughout the training process, reaching a minimum value of approximately 0.2. Figure 5 presents the training and test accuracy curves. The training accuracy peaks at 71.7%, while the test accuracy reaches a maximum of 72%. Both metrics increase in tandem during training and gradually stabilize, with no indication of severe divergence between the two curves.

These results suggest that the model does not suffer from overfitting or underfitting. The neural network effectively

learned representative features from the training data, and the consistent performance on the test set demonstrates strong generalization capabilities of the trained model.

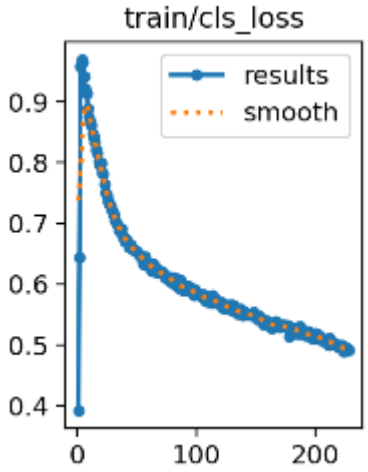

Figure 4. The variation curve of the classification loss

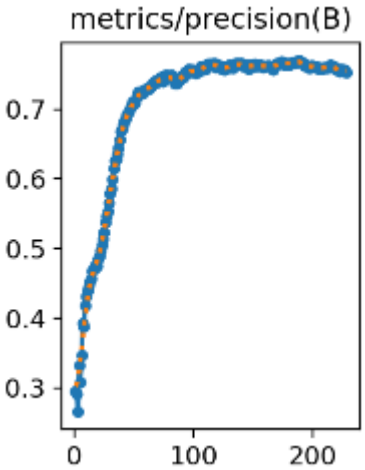

Figure 5. The accuracy of the model during training.

5. Conclusion

This This paper conducted extensive experiments to explore various hyperparameters and iteratively optimize the network architecture, ultimately resulting in the redesign of a new deep neural network model. Experimental results validate the effectiveness of the proposed approach, showing significant improvements in the recognition of various modulation types compared to traditional methods. Most notably, this method achieves end-to-end signal recognition, eliminating the need for manual feature extraction and simplifying the processing pipeline.

The model achieves high test accuracy and demonstrates strong generalization capability, indicating its potential for practical application. With ongoing research efforts and the continuous advancement of deep learning, it is expected that further breakthroughs will emerge in modulation recognition for signal and information processing, leading to even greater accuracy.

Nevertheless, there remain areas for improvement. The current experiments were limited by the scale of available data, and further optimization of the network architecture and hyperparameters is still possible. Future work will focus on expanding the dataset, enhancing model robustness, and refining key components of the network through continued experimentation.

References

[1] Zell O, Pålsson J, Hernandez-Diaz K, et al. Image-Based Fire Detection in Industrial Environments with YOLOv4[J]. arXiv preprint arXiv:2212.04786, 2022.

[2] Dobre, O. A., et al. (2007). Survey of automatic modulation classification techniques: Classical approaches and new trends. IEEE Communications Surveys & Tutorials.

[3] O'Shea, T. J., & Hoydis, J. (2017). An introduction to deep learning for the physical layer. IEEE Transactions on Cognitive Communications and Networking.

[4] Kulin, M., et al. (2018). End-to-end learning from spectrum data IEEE Access.

[5] Zhang, F., et al. (2022). Deep learning based automatic modulation recognition: Models, datasets, and challenges.

[6] Li, Y., et al. (2022). EfficientFormerV2: Efficient and Accurate Lightweight Vision Transformers.

[7] Zhang, Y., et al. (2024). HistoFormer: Histogram Transformer for Lightweight Image Recognition. ECCV 2024.

[8] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization.

[9] Amini, A., Shirani-Mehr, H., & Karbasi, A. Digital Modulation Classification: A Deep Learning Approach. IEEE Transactions on Communications, (2017) 65(11), 4658-4668.

[10] Hochreiter, S., & Schmidhuber, J. Long Short-Term Memory. Neural Computation, (1997) 9(8), 1735-1780.

[11] LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient BackProp. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade (pp. 9-48). Springer.

[12] Shorten, C., & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(60), (2019). 1-48.

[13] Prechelt, L. (1998). Early Stopping - But When? In G. B. Orr & K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade (pp. 55-69). Springer.

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Proceedings of the International Conference on Learning Representations (ICLR).