TFFNet: A Time–Frequency Fusion Network for Automatic Modulation Recognition

Qifan Zhang[1, 2], Lixiang Liu[1, 2], Xin Zhou[1, 2, *]
[1]The Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
Beijing, China
{zhangqifan2023, lixiang, zhouxin}@iscas.ac.cn

***Abstract*—The rapid proliferation of wireless communication devices has intensified spectrum congestion, highlighting the need for intelligent spectrum monitoring and management. Automatic modulation recognition (AMR) plays a critical role in this context by enabling modulation identification for demodulation, interference detection, and dynamic spectrum access. Traditional likelihood- and feature-based AMR methods suffer from high computational cost and limited adaptability, while recent deep learning approaches, though effective, often process different signal representations in isolation and struggle under low-SNR conditions. To address these limitations, we propose TFFNet, a time–frequency fusion network that jointly exploits temporal and spectral characteristics of modulation signals. TFFNet adopts a dual-branch architecture: the Time-Mamba branch extends the Mamba state space model with a Frequency-aware State Space Model (FSSM) for efficient temporal–spectral sequence modeling, while the frequency-domain CNN (Freq-CNN) branch leverages Stationary Wavelet Transform (SWT) to extract discriminative time–frequency features. A cross-attention fusion module adaptively aligns and integrates domain-specific features, enhancing consistency and robustness. Extensive experiments show that TFFNet achieves 63.6% accuracy on RML2016.10a and 56.1% on RML2018.01a, demonstrating competitive performance compared to representative methods.**

***Index Terms*—Automatic modulation recognition, deep learning, stationary wavelet transform, mamba**

I. Introduction

With the proliferation of wireless communication devices, modern spectrum resources face unprecedented congestion, intensifying the demand for efficient spectrum monitoring and management. Automatic modulation recognition (AMR) has emerged as a fundamental technology that enables the intelligent identification of modulation types under unknown channel conditions. By providing critical prior information for demodulation, interference detection, and dynamic spectrum access, AMR plays a pivotal role in both civilian and military applications, particularly in non-cooperative communication scenarios [1], [2].

Traditional AMR methods are typically divided into likelihood-based (LB-AMR) [3] and feature-based (FB-AMR) [4] approaches. While LB-AMR achieves strong theoretical classification performance, its high computational complexity renders it impractical for real-time deployment. FB-AMR relies heavily on handcrafted expert features, such as higher-order moments or instantaneous features, but suffers from limited scalability and adaptability to new modulation schemes. To overcome these limitations, recent research has shifted towards deep learning-based AMR (DL-AMR), which automatically learns discriminative representations from raw or transformed signals.

Among DL-AMR approaches, the most common representation is the in-phase and quadrature (I/Q) domain, where convolutional neural networks (CNNs) capture local spatial patterns [5]. However, CNNs lack temporal modeling capability, while recurrent neural networks (RNNs), such as LSTM and GRU, capture sequential dependencies but suffer from high computational cost and limited parallelization [6]. Hybrid CNN-RNN models integrate both representations [7], and transformer-based methods further capture long-range dependencies [8]. Despite these advances, balancing local feature extraction, temporal continuity, and global context modeling remains challenging.

To further exploit modulation characteristics, amplitude and phase (A/P) representations have been explored, providing insight into signal variations under different SNR conditions. Dual-branch networks process A/P and I/Q streams jointly, improving robustness through feature fusion [9]. However, their heterogeneous statistical properties often require more complex architectures for effective alignment.

Another line of research introduces time-frequency representations, with short-time Fourier transform (STFT) being widely used. By mapping signals into spectrogram-like matrices, STFT enables CNN-based models to exploit two-dimensional structures [10, 11]. Other representations, such as constellation and eye diagrams, further enhance discriminability but remain sensitive to channel impairments and low-SNR conditions [12].

More advanced transformations, such as the continuous wavelet transform (CWT), have been explored for AMR. CWT provides multi-resolution analysis that captures both local and global variations in the signal's frequency content, enabling robust feature learning under noisy conditions [13]. However, it introduces additional preprocessing complexity.

The discrete wavelet transform (DWT) offers a more compact multi-resolution decomposition through iterative filtering and down-sampling, enabling efficient extraction of time–frequency characteristics with reduced computational

cost [14]. Building upon DWT, the stationary wavelet transform (SWT) removes the down-sampling step to preserve alignment between transformed coefficients and the original signal, improving robustness to shifts and distortions. Although SWT introduces higher redundancy, it often achieves better performance under challenging conditions [15].

In summary, existing AMR research has evolved from raw I/Q-based learning to approaches that exploit amplitude-phase features, time-frequency distributions, and advanced signal transformations. While these representations provide complementary strengths, most methods process them independently and rely on late-stage fusion, limiting cross-domain interaction, particularly in low-SNR scenarios.

To overcome these challenges, we propose TFFNet, a time–frequency fusion network that jointly exploits temporal and spectral characteristics of modulation signals. The name TFFNet (Time–Frequency Fusion Network) is used in this work to denote the proposed architecture and is independent of other methods with similar naming that may exist in different domains. TFFNet adopts a dual-branch architecture, where the Time-Mamba branch extends the state space model (SSM) [16] with a Frequency-aware State Space Model (FSSM) for efficient sequence modeling from I/Q data, and the frequency-domain CNN (Freq-CNN) branch leverages SWT to extract discriminative time–frequency features using a CNN backbone. A cross-attention fusion module is further introduced to adaptively align and integrate features from both domains, thereby enhancing robustness and generalization.

The main contributions of this work are summarized as follows:

- We propose a time–frequency fusion network (TFFNet), a unified framework that integrates time-domain and frequency-domain features through a dual-branch design.
- We design a Frequency-aware State Space Model to enhance Mamba for joint temporal–spectral sequence modeling.
- We introduce a cross-attention fusion mechanism to adaptively combine domain-specific features, improving feature consistency in complex environments.
- We conduct extensive experiments on benchmark AMR datasets, demonstrating that TFFNet achieves strong performance while maintaining a compact architecture.

Open Science Statement: In line with the open science and reproducibility guidelines of IJCNN, the data and implementation code used in this study are publicly available at https://github.com/brand-zqf/TFFNet-for-AMR.

II. Preliminaries

A. Signal Model

We consider a wireless communication system where the transmitted modulated signal $s(t)$ passes through a fading channel. The received signal is given by

$$r(t) = H(t) * s(t) + n(t), \tag{1}$$

where $*$ denotes convolution, $H(t)$ is the channel impulse response, and $n(t)$ represents additive white Gaussian noise (AWGN).

The channel can be modeled as

$$H(t) = \alpha(t)e^{j(\omega t+\phi)}, \tag{2}$$

where $\alpha(t)$ denotes the time-varying channel gain, ω is the Doppler-induced frequency shift, and ϕ is the phase offset. Consequently, the received signal can be expressed as

$$r(t) = \alpha(t)e^{j(\omega t+\phi)}s(t) + n(t). \tag{3}$$

This model captures key channel impairments such as fading, Doppler effect, and additive noise, all of which complicate the AMR task.

B. Modulation Recognition

The AMR problem is formulated as a K-class classification task, where the objective is to map the received signal $r(t)$ into one of the modulation categories:

$$G : r(t) \mapsto C_k, \quad k \in \{1, 2, \ldots, K\}, \tag{4}$$

with G being the classifier and C_k the k-th modulation type.

In practice, $r(t)$ is often represented in I/Q components, which preserve amplitude, phase, and frequency characteristics of the baseband signal, thus providing a comprehensive representation for subsequent feature extraction and classification.

C. Stationary Wavelet Transform

Let $x[k] \in \mathbb{C}$ denote the discrete-time baseband signal with sample index k. The SWT decomposes $x[k]$ into approximation and detail coefficients at multiple scales without downsampling. At decomposition level $j = 1, \ldots, J$, the coefficients are obtained by

$$a_j[k] = \sum_m g^{(j)}[m]\ a_{j-1}[k-m], \tag{5}$$

$$d_j[k] = \sum_m h^{(j)}[m]\ a_{j-1}[k-m], \tag{6}$$

with $a_0[k] = x[k]$.

where m is the filter tap index, $g^{(j)}[\cdot]$ and $h^{(j)}[\cdot]$ are the level-j low-pass and high-pass analysis filters, obtained by inserting $2^{j-1} - 1$ zeros between the taps of the prototype filters. Since no decimation is applied, the lengths of $a_j[k]$ and $d_j[k]$ remain identical to that of $x[k]$.

III. Proposed Methods

In this section, we present the proposed framework, which is composed of four key components: a Frequency-aware State Space Model, a Time-Mamba branch for capturing long-range temporal dependencies, a frequency-domain CNN branch for extracting frequency features, and a cross-attention module designed to fuse temporal and spectral representations. The overall architecture is illustrated in Fig. 1.

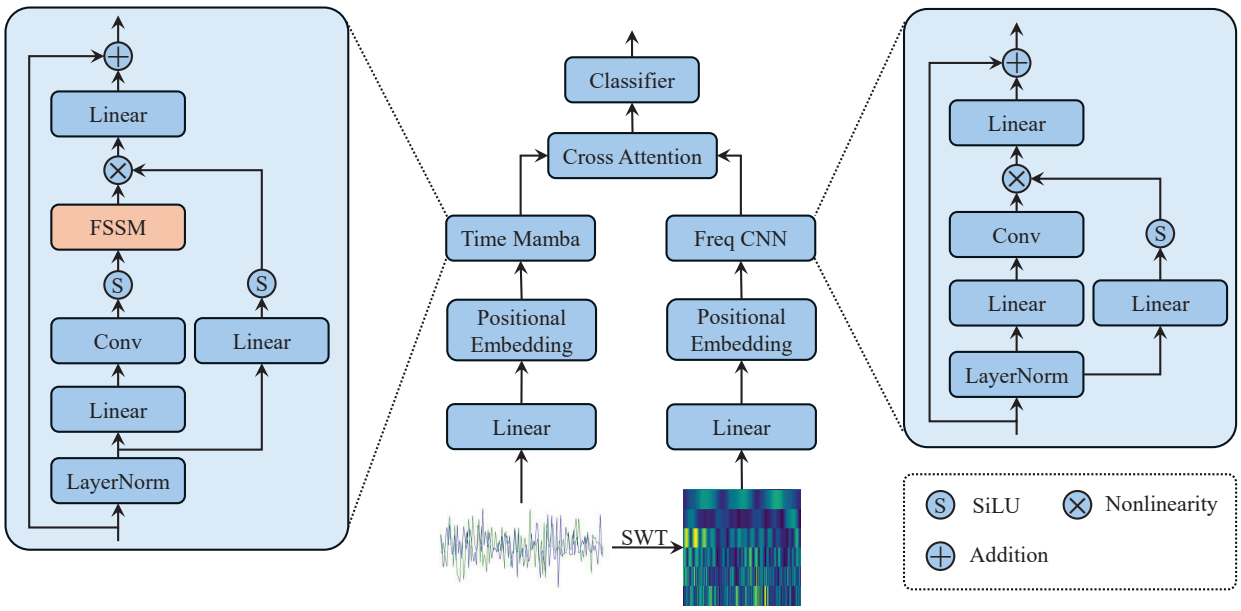

Fig. 1: The structure of the proposed TFFNet.

A. *Frequency-aware State Space Model*

We extend the standard SSM by incorporating frequency-domain operations into its recurrent formulation, resulting in a unified FSSM. A standard SSM is defined as

$$x_{k+1} = \bar{A}x_k + \bar{B}(u_k)u_k, \quad y_k = \bar{C}(u_k)x_k + \bar{D}u_k, \tag{7}$$

where $u_k \in \mathbb{R}^{d_{\text{in}}}$ denotes the input at time step k, $x_k \in \mathbb{R}^{d_{\text{state}}}$ is the hidden state, and $y_k \in \mathbb{R}^{d_{\text{out}}}$ is the output. In SSM, $\bar{B}(\cdot)$ and $\bar{C}(\cdot)$ are input-dependent, enabling adaptive dynamics. However, the recurrence in Eq. (7) is local in time and may struggle to capture long-range spectral patterns.

To address this, FSSM augments the recurrence with a spectral transformation. For an input segment $\{u_0, \ldots, u_{L-1}\}$, we first compute its discrete Fourier transform (DFT):

$$U(\omega) = \sum_{k=0}^{L-1} u_k \, e^{-j\omega k}, \quad \omega = \frac{2\pi n}{L}, \; n = 0, \ldots, L-1. \tag{8}$$

A learnable frequency response $H(\omega)$ is then applied:

$$\tilde{U}(\omega) = H(\omega) \cdot U(\omega), \tag{9}$$

followed by an inverse DFT to obtain the frequency-enhanced input:

$$\tilde{u}_k = \frac{1}{L} \sum_{\omega} \tilde{U}(\omega) e^{j\omega k}. \tag{10}$$

Here, $\tilde{u}_k$ denotes the spectrally transformed input derived from the original sequence u_k, which encodes global spectral context.

The recurrence is then updated with the frequency-enhanced input:

$$x_{k+1} = \bar{A}x_k + \bar{B}(u_k)\tilde{u}_k, \quad y_k = \bar{C}(u_k)x_k + \bar{D}\tilde{u}_k. \tag{11}$$

Compared with the original SSM in Eq. (7), FSSM replaces the raw input u_k with $\tilde{u}_k$. This design enables the model to capture both long-range dependencies and local temporal dynamics within a unified recurrent framework. While the recurrent structure is retained, the spectral transformation introduces additional computational overhead, and the current implementation is primarily designed for offline processing.

The additional frequency-domain operations incur a computational complexity of $O(L \log L)$ due to the DFT. In practice, this overhead is moderate and does not dominate the overall computation, as the model remains largely governed by the recurrent state updates. Moreover, since the DFT is applied to finite-length signal segments, the additional cost is manageable in practical scenarios.

B. *Time-Mamba*

The Time-Mamba branch is built upon the Mamba architecture, which is designed for efficient long-range sequence modeling by leveraging SSM. Unlike attention mechanisms that suffer from quadratic complexity, Mamba achieves linear-time complexity with respect to sequence length, making it suitable for wireless signals that often exhibit long temporal dependencies. However, the original SSM formulation in Mamba primarily focuses on time-domain correlations and may underutilize frequency-domain priors that are crucial for communication signals.

To address this limitation, we replace the standard SSM with our proposed FSSM. The FSSM augments the recurrence formulation of Mamba by integrating frequency-domain operations, enabling the model to simultaneously capture temporal dynamics and spectral patterns within a unified formulation. This design improves the robustness of long-sequence model-

ing in the presence of noise and fading, while maintaining the efficiency and scalability advantages of Mamba. In this way, the Time-Mamba branch serves as the backbone for extracting global temporal–spectral dependencies across sequences.

C. Freq-CNN

Recent work MambaOut [17] questioned the necessity of Mamba in vision tasks by removing its core SSM modules. Their study showed that while Mamba offers clear benefits for tasks requiring long-term dependency modeling, its superiority over convolutional architectures diminishes on short-sequence tasks such as image classification. This finding suggests that not all modalities or feature domains require heavy long-range modeling, and simpler local operators may be more effective.

Inspired by this insight, we introduce a complementary frequency-domain branch, termed Freq-CNN, to capture localized spectral structures. Specifically, Freq-CNN employs convolutional kernels directly on frequency-domain feature maps, effectively modeling neighborhood spectral correlations such as harmonics, sub-carrier dependencies, and narrowband interference. This branch provides a lightweight yet effective complement to the Time-Mamba branch, focusing on local spectral consistency and enhancing discriminative power in challenging modulation scenarios. By combining Freq-CNN with Time-Mamba, our framework achieves a balance between global long-range modeling and local frequency-aware representation learning.

D. Cross Attention

To effectively integrate the time-domain and frequency-domain representations in our model, we adopt a *cross-attention* mechanism. The key idea is to allow each branch to selectively focus on informative patterns from the other domain, enhancing the fused feature representation while preserving the structural integrity of each branch.

Specifically, let $X \in \mathbb{R}^{B\times C\times L_x}$ denote the time-domain feature map from the Time-Mamba branch, and $Y \in \mathbb{R}^{B\times C\times L_y}$ denote the frequency-domain feature map from the Freq-CNN branch. Both feature maps are first projected to queries, keys, and values via 1D convolution or linear layers. Cross-attention is then computed in both directions: X attends to Y and Y attends to X. Residual connections are applied to retain the original features, and the resulting feature maps are concatenated along the channel dimension to produce the fused representation $\hat{Z} \in \mathbb{R}^{B\times(C_X+C_Y)\times L}$.

The detailed procedure is summarized in Algorithm 1.

Algorithm 1 Cross-Attention

Require: Time-domain feature map $X \in \mathbb{R}^{B\times C\times L_x}$;
Frequency-domain feature map $Y \in \mathbb{R}^{B\times C\times L_y}$.
Ensure: Fused feature map $\hat{Z} \in \mathbb{R}^{B\times(C_X+C_Y)\times L}$.
Project X to $Q_X, K_X, V_X \in \mathbb{R}^{B\times L_x\times d}$.
Project Y to $Q_Y, K_Y, V_Y \in \mathbb{R}^{B\times L_y\times d}$.
Cross-Attention: X **attends to** Y
$A_{X\leftarrow Y} = \text{Softmax}\left(\frac{Q_X K_Y^\top}{\sqrt{d}}\right)$
where $A_{X\leftarrow Y} \in \mathbb{R}^{B\times L_x\times L_y}$.
$\hat{X} = A_{X\leftarrow Y} V_Y$, where $\hat{X} \in \mathbb{R}^{B\times L_x\times d}$.
Cross-Attention: Y **attends to** X
$A_{Y\leftarrow X} = \text{Softmax}\left(\frac{Q_Y K_X^\top}{\sqrt{d}}\right)$.
$\hat{Y} = A_{Y\leftarrow X} V_X$, where $\hat{Y} \in \mathbb{R}^{B\times L_y\times d}$.
Apply residual connections: $\hat{X} \leftarrow \hat{X} + X$, $\hat{Y} \leftarrow \hat{Y} + Y$.
Align sequence lengths (e.g., via interpolation or pooling) to obtain a unified length L.
$\hat{Z} = \text{Concat}(\hat{X}, \hat{Y})$ (channel-wise)

IV. Experiments

A. Evaluation Datasets

In this work, we evaluate our proposed method using two benchmark datasets: RML2016.10a [18] and RML2018.01a [19], which are commonly used in the field of AMR. The RML2016.10a dataset contains 11 modulation schemes, including BPSK, QPSK, 8PSK, 16QAM, 64QAM, GFSK, CPFSK, WBFM, AM-SSB, AM-DSB, and PAM4, with a signal-to-noise ratio (SNR) range from -20 dB to 18 dB in 2 dB increments. Each modulation type contains 1,000 samples per SNR. Each sample is represented as a 2×128 matrix, where the first dimension corresponds to the in-phase (I) component, and the second dimension corresponds to the quadrature (Q) component.

The RML2018.01a dataset extends the modulation diversity significantly, introducing 24 modulation schemes. This diverse set of modulations allows the dataset to evaluate models in a broader context, particularly in terms of scalability and performance in complex signal environments. The dataset contains a total of two million samples, each with a signal length of 1024, and is represented in a 2×1024 matrix format.

For consistency across datasets, we adopt a unified SNR range of -20 dB to 18 dB for both RML2016.10a and RML2018.01a. This choice aligns with the standard setting of RML2016.10a and ensures a consistent evaluation protocol. In addition, performance on RML2018.01a is known to saturate at high SNR levels (e.g., above 20 dB), where most methods achieve near-perfect accuracy and performance differences become marginal. Therefore, focusing on the -20 dB to 18 dB range provides a more challenging and practically relevant evaluation setting, particularly for assessing robustness under low-SNR conditions.

For both datasets, the data is split into training, validation, and test sets with a 6:2:2 ratio, ensuring proper evaluation across all modulation types and SNR levels.

B. Implementation Details

The experiments are conducted in an offline setting, where full-sequence information is available during inference. All experiments are conducted on a server equipped with a GeForce GTX 4090 GPU, providing sufficient computational resources for model training and evaluation. The initial learning rate is set to 0.001, and the model parameters are optimized using the AdamW optimizer with a batch size of 128. A dynamic learning rate adjustment strategy is adopted: if the

validation loss does not decrease within 10 consecutive epochs, the learning rate is reduced by a factor of 0.1 to facilitate convergence. Furthermore, an early stopping mechanism is employed with a patience threshold of 20 epochs to prevent overfitting and reduce unnecessary computations.

C. Comparison with representative AMR methods

To evaluate the effectiveness of the proposed TFFNet, we compare it with several representative and competitive methods that explicitly exploit time–frequency information, including CLDNN [20], TFA [21], MFF [22], and MFCA [23], on the RML2016.10a and RML2018.01a datasets.

As shown in Table I, TFFNet consistently achieves superior recognition performance on both datasets while maintaining a more compact architecture. Compared with both classical deep learning models and recent time–frequency–aware approaches, TFFNet demonstrates a clear advantage in terms of accuracy, model size, and inference efficiency. In particular, it achieves better performance with fewer parameters and lower inference latency, indicating a more effective utilization of time–frequency representations.

Overall, these results confirm that explicit time–frequency feature fusion is crucial for improving modulation recognition. By jointly modeling temporal dynamics and spectral characteristics within a unified framework, TFFNet achieves a favorable balance between recognition accuracy and computational efficiency, highlighting its potential for practical and real-time AMR applications.

D. Explanation of SNR-dependent Performance

Fig. 2 and Fig. 3 present the recognition accuracy of different methods under varying SNR conditions on RML2016.10a and RML2018.01a, respectively. These results correspond to the overall performance comparison reported in Table I.

Across the full SNR range, TFFNet consistently achieves superior performance compared with the baseline models on both datasets. At low SNR levels, where signal distortion and noise interference are severe, TFFNet maintains a clear advantage, indicating stronger robustness to adverse channel conditions. As the SNR increases, the performance of all methods improves; however, TFFNet exhibits a more stable and sustained lead, demonstrating its ability to effectively exploit informative signal structures.

Compared with CLDNN and other time–frequency–aware baselines, TFFNet shows smoother accuracy transitions across different SNR regimes, suggesting better generalization and reduced sensitivity to noise fluctuations. This behavior is consistent across both datasets, despite their different modulation configurations and channel characteristics.

E. Confusion Matrix Analysis

To further evaluate the effectiveness of the proposed model, we analyze the confusion matrices on two datasets at 0 dB SNR, as shown in Fig. 4 and Fig. 5.

On RML2016.10a, most modulation types are recognized with high accuracy. For example, BPSK, QPSK, and 8PSK

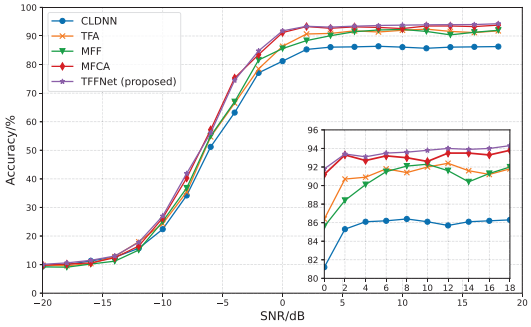

Fig. 2: Accuracy of different methods on RML2016.10a.

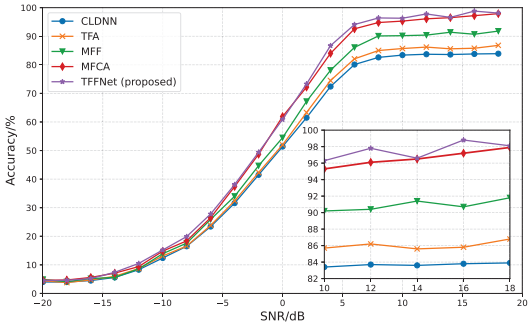

Fig. 3: Accuracy of different methods on RML2018.01a.

achieve recognition rates of over 95%, demonstrating that the joint representation effectively separates low-order phase modulations. However, residual confusion is observed between higher-order QAM signals, where part of the 16QAM samples are predicted as 64QAM due to less distinguishable time–frequency energy distributions in noisy conditions. Moreover, PAM4 occasionally overlaps with QPSK, indicating that amplitude-varying signals may still interfere with phase-based ones at low SNR.

On RML2018.01a, the classification difficulty increases due to the larger set of 24 modulation classes. The proposed model still achieves near-perfect recognition for low-order PSK modulations, benefiting from the complementary I/Q and SWT features. Nevertheless, significant confusion occurs among higher-order APSK and QAM classes (e.g., 32APSK, 64APSK, 128APSK, and 32QAM, 64QAM), where the model tends to misclassify adjacent orders, as their fine-grained spectral-temporal differences are strongly affected by noise.

Overall, the confusion matrix analysis verifies that the dual-modal input strategy effectively improves the robustness of low- and mid-order modulations at low SNR. However, the classification of higher-order QAM/APSK remains challenging, suggesting that more discriminative time–frequency representations or modulation-specific feature regularization may be required.

TABLE I: Comparison of Model Performance on RML2016.10a and RML2018.01a.

Model	Parameters (k)		Accuracy (%)		Inference Time (ms/sample)	
	RML2016.10a	RML2018.01a	RML2016.10a	RML2018.01a	RML2016.10a	RML2018.01a
CLDNN	164	367	58.1	48.4	1.48	1.86
TFA	104	231	61.7	50.6	0.76	0.92
MFF	286	529	61.2	52.3	0.82	0.98
MFCA	350	636	62.7	55.3	0.85	1.13
TFFNet	84	153	**63.6**	**56.1**	0.51	0.74

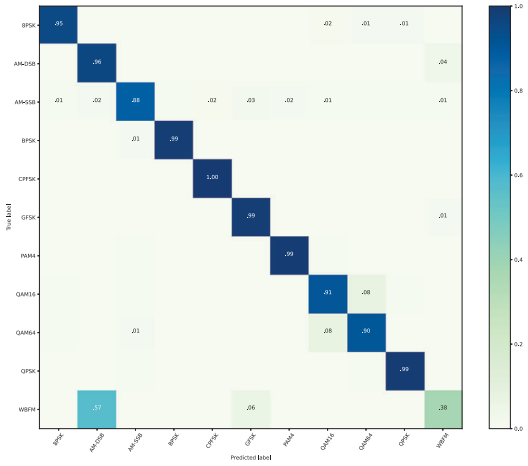

Fig. 4: The Confusion Matrix of TFFNet on RML2016.10a at 0 dB SNR.

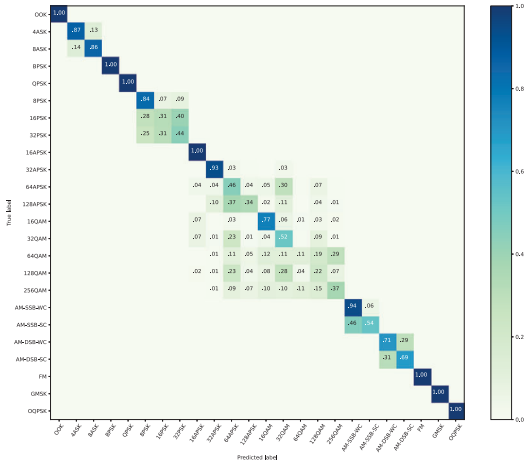

Fig. 5: The Confusion Matrix of TFFNet on RML2018.01a at 0 dB SNR.

F. Ablation Study

An ablation study is performed to analyze the contribution of each component in TFFNet, as summarized in Table II. Removing any module consistently degrades performance on both datasets, indicating that all components contribute positively to the final model.

Among them, removing the Time-Mamba branch leads to the most notable degradation, highlighting the importance of long-range temporal modeling over raw I/Q signals. Eliminating the frequency-domain branch also results in a clear decline, confirming that wavelet-based time–frequency representations provide complementary information to temporal features.

Replacing cross-attention with simple concatenation causes a moderate but consistent performance drop, suggesting the benefit of explicit cross-modal interaction. In addition, substituting SWT with DWT degrades performance, demonstrating the advantage of shift-invariant time–frequency decomposition under noisy conditions.

TABLE II: The Ablation Results of TFFNet.

Model	Accuracy (%)	
	RML2016.10a	RML2018.01a
w/o Time-Mamba	60.3	52.1
w/o Freq-CNN	62.1	54.4
w/o Cross-Attention (Concat only)	63.0	54.8
w/o FSSM (standard SSM)	62.7	55.3
SWT $\rightarrow$ DWT (db4, $J = 6$)	62.5	54.1
SWT wavelet: Haar, $J = 6$	62.8	55.2
SWT wavelet: sym4, $J = 6$	63.4	55.8
SWT wavelet: coif1, $J = 6$	63.3	55.7
SWT wavelet: db4, $J = 2$	63.0	55.3
SWT wavelet: db4, $J = 4$	63.3	55.8
SWT wavelet: db4, $J = 6$ (default)	63.6	56.1
SWT wavelet: db4, $J = 8$	63.5	56.0
TFFNet	**63.6**	**56.1**

V. Conclusion

In this paper, we propose TFFNet, a unified time-frequency fusion network for automatic modulation recognition. TFFNet integrates temporal and spectral representations through a dual-branch design, where the Time-Mamba branch incorporates a Frequency-aware State Space Model (FSSM) to enhance temporal–spectral sequence modeling, and the Freq-CNN branch exploits Stationary Wavelet Transform (SWT) for discriminative feature extraction. A cross-attention fusion mechanism was further introduced to adaptively align and integrate domain-specific features, improving robustness in complex environments. Extensive experiments on benchmark datasets demonstrate that TFFNet achieves competitive and consistent performance, with 63.6% accuracy on RML2016.10a and 56.1% accuracy on RML2018.01a under a unified SNR range (-20 dB to 18 dB).

REFERENCES

[1] F. Meng, P. Chen, L. Wu, and X. Wang, "Automatic modulation classification: A deep learning enabled approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 760–10 772, 2018.

[2] Y. Wu, X. Li, and J. Fang, "A deep learning approach for modulation recognition via exploiting temporal correlations," in *2018 IEEE 19th international workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[3] F. Hameed, O. A. Dobre, and D. C. Popescu, "On the likelihood-based approach to modulation classification," *IEEE transactions on wireless communications*, vol. 8, no. 12, pp. 5884–5892, 2009.

[4] A. Hazza, M. Shoaib, S. A. Alshebeili, and A. Fahad, "An overview of feature-based methods for digital modulation classification," in *2013 1st international conference on communications, signal processing, and their applications (ICCSPA)*. IEEE, 2013, pp. 1–6.

[5] S. Peng, H. Jiang, H. Wang, H. Alwageed, and Y.-D. Yao, "Modulation classification using convolutional neural network based deep learning model," in *2017 26th Wireless and Optical Communication Conference (WOCC)*. IEEE, 2017, pp. 1–5.

[6] Z. Ke and H. Vikalo, "Real-time radio technology and modulation classification via an lstm auto-encoder," *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 370–382, 2021.

[7] Z. Zhang, H. Luo, C. Wang, C. Gan, and Y. Xiang, "Automatic modulation classification using cnn-lstm based dual-stream structure," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 521–13 531, 2020.

[8] J. Cai, F. Gan, X. Cao, and W. Liu, "Signal modulation classification based on the transformer network," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 3, pp. 1348–1357, 2022.

[9] S. Chang, S. Huang, R. Zhang, Z. Feng, and L. Liu, "Multitask-learning-based deep neural network for automatic modulation classification," *IEEE internet of things journal*, vol. 9, no. 3, pp. 2192–2206, 2021.

[10] X. Hao, Z. Feng, S. Yang, M. Wang, and L. Jiao, "Automatic modulation classification via meta-learning," *IEEE Internet of Things Journal*, vol. 10, no. 14, pp. 12 276–12 292, 2023.

[11] N. Daldal, Z. Cömert, and K. Polat, "Automatic determination of digital modulation types with different noises using convolutional neural network based on time–frequency information," *Applied Soft Computing*, vol. 86, p. 105834, 2020.

[12] Y. Mao, Y.-Y. Dong, T. Sun, X. Rao, and C.-X. Dong, "Attentive siamese networks for automatic modulation classification based on multitiming constellation diagrams," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 5988–6002, 2021.

[13] M. Walenczykowska and A. Kawalec, "Type of modulation identification using wavelet transform and neural network," *Bulletin of the Polish Academy of Sciences. Technical Sciences*, vol. 64, no. 1, pp. 257–261, 2016.

[14] D. H. Al-Nuaimi, I. A. Hashim, I. S. Zainal Abidin, L. B. Salman, and N. A. Mat Isa, "Performance of feature-based techniques for automatic digital modulation recognition and classification—a review," *Electronics*, vol. 8, no. 12, p. 1407, 2019.

[15] M. Zeng, Z. Liu, Z. Wang, H. Liu, Y. Li, and H. Yang, "An adaptive specific emitter identification system for dynamic noise domain," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25 117–25 135, 2022.

[16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[17] W. Yu and X. Wang, "Mambaout: Do we really need mamba for vision?" in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4484–4496.

[18] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *International conference on engineering applications of neural networks*. Springer, 2016, pp. 213–226.

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

[20] N. E. West and T. O'shea, "Deep architectures for modulation recognition," in *2017 IEEE international symposium on dynamic spectrum access networks (DySPAN)*. IEEE, 2017, pp. 1–6.

[21] S. Lin, Y. Zeng, and Y. Gong, "Learning of time-frequency attention mechanism for automatic modulation recognition," *IEEE Wireless Communications Letters*, vol. 11, no. 4, pp. 707–711, 2022.

[22] H. Zhang, Y. Kuang, R. Huang, S. Lin, Y. Dong, and M. Zhang, "Modulation recognition method based on multimodal features," *Frontiers in Communications and Networks*, vol. 6, p. 1453125, 2025.

[23] X. Hu, M. Chen, X. Zhang, J. Rao, S. Li, and X. Song, "Mfca-transformer: Modulation signal recognition based on multidimensional feature fusion," *Sensors*, vol. 25, no. 16, p. 5061, 2025.