Fusing Time-Domain and Constellation Views: A Multimodal MAE for Wireless Signals (Student Abstract)

Agniva Banerjee and Arijit Sen

Department of Electrical Engineering & Computer Science
Indian Institute of Science Education and Research Bhopal
{agniva24, ajsen}@iiserb.ac.in

Abstract

This paper introduces a multi-modal masked autoencoder (MMAE) that jointly denoises and classifies signals by fusing time-domain IQ sequences and constellation diagrams within a cross-attentive transformer. This approach treats noise as a learnable modality to enhance robustness, a dynamic masking curriculum combined with domain regularization training and a hybrid loss function to promote domain-invariant features. Experimentation on the RadioML 2018.01A and RadioML22 datasets demonstrates superior accuracy across different SNR levels while using substantially less labeled data than state-of-the-art approaches.

Introduction

Automatic modulation classification (AMC) is a critical task in cognitive radio and spectrum monitoring (Dong et al. 2024). Many learning-based AMC models have shown promising results but are hampered by real-world channel impairments, especially in low SNR conditions (Zargari et al. 2023). Earlier convolutional neural networks (CNNs) (Gama et al. 2018) and later transformer-based architectures (Zeng et al. 2024) are used to capture features from signals. Although they are effective in controlled settings and struggle with generalization, they require large labeled datasets (Rehman et al. 2024). Self-supervised learning methods, particularly masked autoencoders (MAE), have shown promising performance for unlabeled data (Zhang et al. 2023). However, existing methods in AMC are typically limited to single modalities and treat noise as an unstructured disturbance rather than a learnable component (Huynh-The et al. 2021). These limitations highlight the need for approaches that can integrate richer signal representations beyond a single view (Dosovitskiy et al. 2020; He et al. 2022). In this work, we propose an MMAE framework that integrates time-domain signals, constellation diagrams, and explicit noise as distinct modalities. Combined with advanced training strategies, it learns robust, generalizable representations for improved performance in challenging wireless environments. To validate the generalization and robustness of our framework, we utilize the RadioML 2018.01A (RML18) (O'Shea, Roy, and Clancy 2018) and RadioML22 (RML22) (Sathyanarayanan, Gerstoft, and El Gamal 2023) datasets, which are public benchmarks for modulation classification. These datasets are crucial, as they provide complex, over-the-air signal captures that represent realistic channel impairments and serve as a standard for comparing model performance.

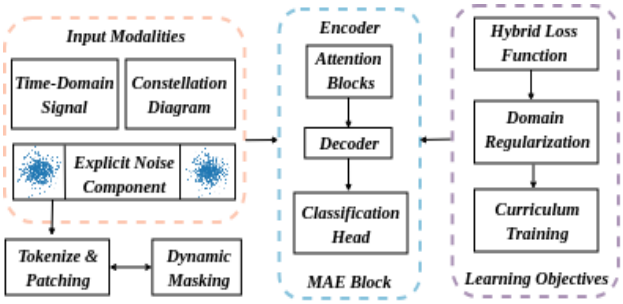


Figure 1: Pipeline of the proposed MMAE for robust wireless signal denoising and classification.

Methodology

The proposed framework operates on five correlated modalities, such as noisy and clean time-domain signals, noisy and clean constellation diagrams, and the explicit noise component (see Figure 1). Each input is first divided into non-overlapping patches, linearly embedded into tokens, and enriched with positional encoding to preserve structural information. A dynamic masking strategy is applied during pre-training, where the ratio of masked tokens follows a cosine annealing schedule. The visible tokens are then processed by a shared cross-modal Transformer encoder that employs both self-attention and cross-attention to capture intra- and inter-modality dependencies. A lightweight transformer decoder reconstructs the complete set of tokens, and modality-specific projection heads generate the reconstructed outputs.

To ensure fidelity to reconstruction and semantically meaningful representations, MMAE is optimized with a hybrid objective. The loss function combines: (i) reconstruction loss of the mask patch (Z_r) based on the MSE, (ii) perceptual loss (Z_p) derived from the VGG-16 feature maps, (iii) L_1 and structural similarity losses (Z_{L_1},Z_{SSIM}) to preserve the sharpness and global structure, and (iv) Contrastive

Method	Samples		SNR (dB)																				
	Pretrain	Finetune	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
VIT	10000	1000	65	66.2	66.5	68	69	71	72.5	76	77.5	78.5	80	80.2	80.4	80.5	80.6	80.8	81	81.5	82.5	83	83.2
MAE	10000	1000	65.2	67.5	68	71	72.5	73.5	75	77.5	79	79.5	80.2	80.5	80.7	81	81.2	81.5	81.7	82	82.5	83	83.2
DenoMAE 2.0	10000	1000	**66**	71	71	77	**77**	80	**82**	**82**	82	83	83	82.2	83.3	83.5	83.8	84	84.2	84.3	84.5	84.6	84.7
Ours	10000	1000	63.6	**71.5**	**72.9**	**77.1**	75.6	**80.06**	81	81.5	**82.5**	**83.1**	**83.3**	**83.6**	**83.8**	**84.3**	**84.5**	**84.8**	**84.4**	**84.5**	**84.8**	**84.7**	**85**

Table 1: Classification accuracy on the custom multimodal dataset constructed from five correlated modalities.

loss (Z_c) to align latent representations across paired modalities. The overall multi-modal loss is defined as,

$$Z = \lambda_r Z_r + \lambda_p Z_p + \lambda_{L_1} Z_{L_1} + \lambda_S Z_{SSIM} + \lambda_c Z_c \qquad (1)$$

Moreover, domain regularization is introduced to improve robustness under diverse channel conditions (Duan, Xu, and Tsang 2012). A gradient reversal layer and domain classifier are integrated to encourage the encoder to learn domain-invariant features across different SNR levels. Thus, the loss function (1) is reformulated as, $Z_{total} = Z + \lambda_d Z_d$, where Z_d represents the domain regularization loss. Finally, a curriculum learning-based SNR scheduling strategy is employed (Bengio et al. 2009). Training begins with high-SNR samples, which facilitates optimization, and gradually progresses to more challenging low-SNR conditions, ensuring stable convergence and improved generalization.

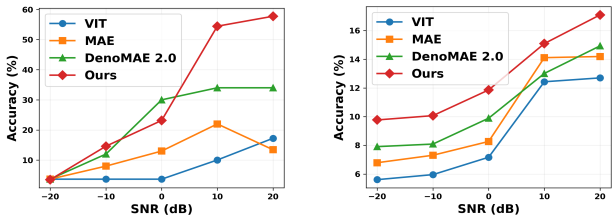
Experiment and Discussion

The experimental process begins with a foundational pretraining phase, where all models are trained on the proposed multimodal synthetic dataset, containing 10,000 samples. Following this, the models are finetuned using 1000 samples across ten modulation classes, with performance evaluated across SNR levels from -10 dB to $+10$ dB (see Table 1). To validate generalization on real benchmark datasets,

Loss Coefficient	λ_r	λ_p	λ_{L_1}	λ_S	λ_c	λ_d
Value	1.0	0.25	0.01	0.01	0.1	0.02

Table 2: Coefficients for the hybrid loss function.

the framework is subsequently finetuned on RML18 and RML22, using their respective train/test splits of 3276/820 and 1600/400. The model architecture processes five input modalities using 224×224 patches and features a 12-layer Transformer encoder and an 8-layer decoder, with 12 attention heads. The training regimen employs a batch size of 8 for 100 epochs, incorporates dynamic masking from 25–75%, and uses the Adam optimizer with a learning rate of 10^{-4}. Early stopping is applied for stability, and all loss coefficients are detailed in Table 2. The evaluation of the models across different SNR levels reveals a clear hierarchy in performance and robustness. The ViT (Dosovitskiy et al. 2020) and MAE (He et al. 2022) baselines show steady improvement as the SNR increases, but their performance degrades sharply in low-SNR conditions, confirming they have limited robustness against severe channel impairments. In contrast, DenoMAE 2.0 (Faysal et al. 2025) demonstrates stronger resilience, particularly in -9 dB to $+7$ dB



(a) RML18 (b) RML22

Figure 2: Classification accuracy of different models on RML18 and RML22 across SNR levels after fine-tuning.

range, which is attributed to its use of denoising objectives. However, the MMAE consistently outperforms all baselines across the entire SNR levels, achieving significantly better accuracy at both low-SNR levels (such as, 72.9% at -8 dB) and high-SNR levels (up to 85% at $+10$ dB). These results validate the benefits of the proposed approach, which jointly models multiple correlated modalities, including explicit noise, within a unified cross-attentive architecture.

Referring to Figure 2, the proposed MMAE framework consistently outperforms all baselines on both the RML18 and RML22 datasets across all tested SNR levels. This performance advantage is particularly notable at specific points. At 0 dB, the MMAE framework achieves a classification accuracy of 23.16 on RML18 and 11.86 on RML22, which is nearly double the accuracy of the DenoMAE 2.0 baseline. The gap is even more pronounced at $+10$ dB, where the proposed MMAE reaches 54.44 (on RML18) and 15.09 (on RML22), far exceeding 34 and 13.01 achieved by DenoMAE 2.0, respectively. These results strongly confirm that the combined strategy of multimodal fusion and explicit noise modeling is highly effective for robust and data-efficient modulation classification, especially under realistic channel conditions.

Conclusions

This paper introduced a novel MMAE framework designed for robust wireless signal classification. The proposed architecture successfully achieves state-of-the-art performance on public AMC benchmarks by integrating a hybrid loss function with advanced training strategies. This approach is shown to be effective in challenging low-SNR conditions. Given its high data efficiency, the framework represents a scalable and practical solution poised for implementation in next-generation wireless communication systems.

Acknowledgments

The authors acknowledge partial support for A. Banerjee through the Visvesvaraya PhD Fellowship, Government of India.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Dong, P.; He, C.; Gao, S.; Zhou, F.; and Wu, Q. 2024. Edge learning based collaborative automatic modulation classification for hierarchical cognitive radio networks. *IEEE Internet of Things Journal*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duan, L.; Xu, D.; and Tsang, I. W.-H. 2012. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on neural networks and learning systems*, 23(3): 504–518.

Faysal, A.; Rostami, M.; Boushine, T.; Roshan, R. G.; Wang, H.; and Muralidhar, N. 2025. DenoMAE2. 0: Improving Denoising Masked Autoencoders by Classifying Local Patches. *arXiv preprint arXiv:2502.18202*.

Gama, F.; Marques, A. G.; Leus, G.; and Ribeiro, A. 2018. Convolutional neural network architectures for signals supported on graphs. *IEEE Transactions on Signal Processing*, 67(4): 1034–1049.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Huynh-The, T.; Pham, Q.-V.; Nguyen, T.-V.; Nguyen, T. T.; Ruby, R.; Zeng, M.; and Kim, D.-S. 2021. Automatic modulation classification: A deep architecture survey. *IEEE Access*, 9: 142950–142971.

O'Shea, T. J.; Roy, T.; and Clancy, T. C. 2018. Over-the-air deep learning based radio signal classification. *IEEE Journal of Selected Topics in Signal Processing*, 12(1): 168–179.

Rehman, A.; Zhovmer, A.; Sato, R.; Mukouyama, Y.-s.; Chen, J.; Rissone, A.; Puertollano, R.; Liu, J.; Vishwasrao, H. D.; Shroff, H.; et al. 2024. Convolutional neural network transformer (CNNT) for fluorescence microscopy image denoising with improved generalization and fast adaptation. *Scientific Reports*, 14(1): 18184.

Sathyanarayanan, V.; Gerstoft, P.; and El Gamal, A. 2023. RML22: Realistic dataset generation for wireless modulation classification. *IEEE Transactions on Wireless Communications*, 22(11): 7663–7675.

Zargari, S.; Hakimi, A.; Tellambura, C.; and Maaref, A. 2023. Enhancing AmBC systems with deep learning for joint channel estimation and signal detection. *IEEE Transactions on Communications*, 72(3): 1716–1731.

Zeng, D.; Xiao, Y.; Liu, W.; Du, H.; Zhang, E.; Zhang, D.; Wang, Y.; Zhang, M.; and Chen, W. 2024. Efficient automatic modulation classification in non-terrestrial networks with snn-based transformer. *IEEE Internet of Things Journal*.

Zhang, C.; Zhang, C.; Song, J.; Yi, J. S. K.; and Kweon, I. S. 2023. A Survey on Masked Autoencoder for Visual Self-supervised Learning. In *IJCAI*, 6805–6813.