Research Article | Received TBD; Accepted TBD; Published TBD
https://doi.org/10.55092/xxxx

Constellation Diagram Augmentation and Perturbation-Based Explainability for Automatic Modulation Classification with SNR-Preserving Multi-Task Learning

Shamoon Siddiqui[1] Huaxia Wang[1] Ravi Ramachandran[1,*]

[1] Department of Electrical & Computer Engineering, Rowan University, Glassboro, NJ, USA

* Correspondence author; E-mail: E-mail: ravi@rowan.edu. Co-authors: siddiq76@rowan.edu; wanghu@rowan.edu.

Highlights:

- Novel multi-task learning framework for joint modulation and SNR classification with constellation diagram augmentation
- First systematic perturbation-based explainability analysis revealing critical regions in constellation diagrams
- Enhanced constellation diagram generation methodology preserving signal characteristics
- Introduction of Perturbation Impact Score (PIS) metric for quantifying feature importance
- Comprehensive architecture evaluation revealing hierarchical attention superiority for constellation patterns

Abstract: This paper presents a novel framework combining constellation diagram augmentation with perturbation-based explainability for Automatic Modulation Classification (AMC). By transforming I/Q signal data into enriched constellation diagrams and employing a multi-task learning architecture, our model simultaneously classifies modulation schemes and estimates Signal-to-Noise Ratio (SNR) levels, leveraging shared feature representations for improved generalization. We introduce systematic perturbation analysis of high- and low-intensity regions in constellation diagrams, quantifying their impact on classification accuracy using the novel Perturbation Impact Score (PIS) metric. Our investigation reveals critical preprocessing limitations in existing approaches, where standard per-image max normalization destroys SNR discriminative information. We propose literature-standard SNR-preserving constellation generation achieving 1.73x discrimination improvement. Through comprehensive evaluation across multi-

ple architectures (ResNet, Vision Transformers, Swin Transformers), we demonstrate that hierarchical attention mechanisms are fundamentally superior for constellation pattern recognition, achieving 45.45% combined accuracy on the challenging 272-class joint prediction task. The framework employs principled multi-task learning with Kendall uncertainty weighting, replacing ad-hoc loss balancing schemes. Perturbation analysis shows salient-region sensitivity (e.g., combined PIS $\tilde{1}1.8$ at top 1% masking) and minimal effect for random masking ($\tilde{0}$), providing actionable insights for model optimization and interpretability in safety-critical wireless communication applications.

Keywords: automatic modulation classification, constellation diagrams, perturbation-based explainability, multi-task learning, SNR preservation, explainable AI, signal processing, deep learning

1. Introduction

Automatic Modulation Classification (AMC) is a cornerstone in ensuring the adaptability and efficiency of modern wireless communication systems, enabling spectrum monitoring, cognitive radio applications, and electronic warfare systems. While traditional AMC approaches utilizing raw I/Q time-series data have achieved significant success, constellation diagram-based methods offer unique advantages through visual pattern recognition capabilities that leverage the spatial geometry of modulated signals. However, despite high accuracy, these models often operate as "black boxes," providing limited insight into their decision-making processes—a critical drawback in safety-critical applications where trust and transparency are paramount.

The growing need to understand and trust model decisions has driven interest in explainability techniques for AMC. Recent advances in explainable AI have shown promise, but most approaches lack the granularity needed to understand how specific signal features influence model decisions. This gap in understanding limits the deployment of AMC models in high-stakes applications, where decision transparency is essential for regulatory compliance and operational reliability.

In this work, we propose a novel framework that combines constellation diagram augmentation with perturbation-based explainability to enhance the accuracy, robustness, and interpretability of deep learning-based AMC models. Our approach systematically investigates the impact of perturbations on constellation diagrams, identifying critical regions that drive model predictions using the novel Perturbation Impact Score (PIS) metric. By integrating multi-task learning, we simultaneously predict modulation types and SNR levels, addressing the dual needs of modulation classification and channel quality estimation in real-world communication systems.

Through comprehensive investigation, we identify critical preprocessing errors that destroy SNR discriminative information and propose literature-standard solutions. Our contributions include: (1) novel multi-task learning framework with perturbation-based explainability for joint modulation-SNR prediction, (2) systematic perturbation analysis revealing critical constellation regions and introducing the PIS metric, (3) identification and correction of SNR information destruction in preprocessing (1.73x improvement), (4) comprehensive architecture evaluation revealing hierarchical attention superiority, and

(5) principled uncertainty weighting replacing ad-hoc loss balancing schemes.

The significance of this work extends beyond performance improvements to provide actionable insights into model behavior, enhancing trust and interpretability essential for deployment in safety-critical wireless communication environments. Our perturbation-based explainability framework offers transparency not available in existing AMC approaches, while maintaining competitive classification performance.

2. Related Work

2.1. Constellation-Based Automatic Modulation Classification

Constellation diagram representation has emerged as a powerful approach for AMC, leveraging computer vision techniques to analyze the spatial patterns of modulated signals. Early work by O'Shea & Hoydis [**?**] established the foundation for treating constellation diagrams as images suitable for convolutional neural networks. Recent advances have demonstrated the effectiveness of various preprocessing techniques, including power normalization and log scaling to enhance discriminative features [**?**, **?**].

Zhang et al. [1] proposed a multi-modal approach combining time-domain signals with constellation diagrams, implementing SNR segmentation at -4 dB where constellation features become unreliable. Gao et al. [2] evaluated constellation methods on bounded SNR ranges (-10 to 10 dB), demonstrating significant performance degradation at extreme SNRs. García-López et al. [3] achieved 96.3% accuracy at 0 dB using constellation preprocessing but noted challenges below this threshold.

2.2. Multi-Task Learning in Signal Processing

Multi-task learning has shown promise for joint prediction problems in signal processing applications. Kendall et al. [4] introduced homoscedastic uncertainty weighting for automatic loss balancing, providing principled approaches to multi-task optimization. Li et al. [5] demonstrated curriculum learning benefits for modulation classification, though focusing on single-task scenarios.

The challenge of joint modulation-SNR prediction has received limited attention in literature. Most SOTA approaches either train separate models per SNR range or employ SNR-aware architectures with dynamic feature extraction [6]. Our work represents the first comprehensive study of joint prediction using constellation diagrams with principled uncertainty weighting.

2.3. Transformer Architectures for Signal Classification

The adoption of transformer architectures in signal processing has accelerated following successes in computer vision. Convolutional baselines such as ResNet provide strong shared-feature trunks for multi-task heads [**?**]. Vision Transformers (ViTs) have shown promise for AMC applications, though memory constraints and training instability remain challenges [**?**]. Swin Transformers introduce hierarchical processing and shifted window attention, offering computational efficiency while maintaining representational power [**?**].

Recent work has explored patch size optimization for signal classification, revealing that larger patch sizes provide efficiency benefits for macro-structural features typical in constellation patterns. However, systematic evaluation of transformer architectures specifically for constellation-based AMC remains limited, motivating our comprehensive architectural study.

3. Methodology

3.1. Enhanced Constellation Diagram Generation

We implement an enhanced constellation diagram generation methodology for I/Q signal data transformation. The process converts complex-valued time-domain samples into visual representations suitable for image-based deep learning models. Our critical discovery revealed that standard preprocessing approaches destroy SNR information through per-image normalization, explaining persistent accuracy limitations in prior work.

Power Normalization: To maintain signal characteristics across different power levels:

$$\text{power} = \frac{1}{N}\sum_{i=1}^{N}(I_i^2 + Q_i^2) \tag{1}$$

$$\text{scale\_factor} = \sqrt{\text{power}} \tag{2}$$

$$I_{normalized} = \frac{I}{\text{scale\_factor}}, \quad Q_{normalized} = \frac{Q}{\text{scale\_factor}} \tag{3}$$

Histogram Generation and Log Scaling: The normalized I/Q data is binned into 2D histograms and log-scaled for enhanced dynamic range:

$$H = \log(1 + \text{histogram2d}(I_{normalized}, Q_{normalized})) \tag{4}$$

This approach maintains relative signal characteristics while providing enhanced visual representation of constellation patterns essential for multi-task learning across diverse SNR conditions. Figure 1 summarizes the preprocessing and training flow used in our implementation.

3.2. Perturbation-Based Explainability Framework

We integrate explainability directly into the AMC framework by employing systematic perturbation-based analysis. Perturbation-based methods systematically alter specific regions of the input to evaluate their impact on model predictions, offering insights into the model's decision-making process. This approach is particularly valuable in safety-critical domains, where understanding model behavior is paramount for

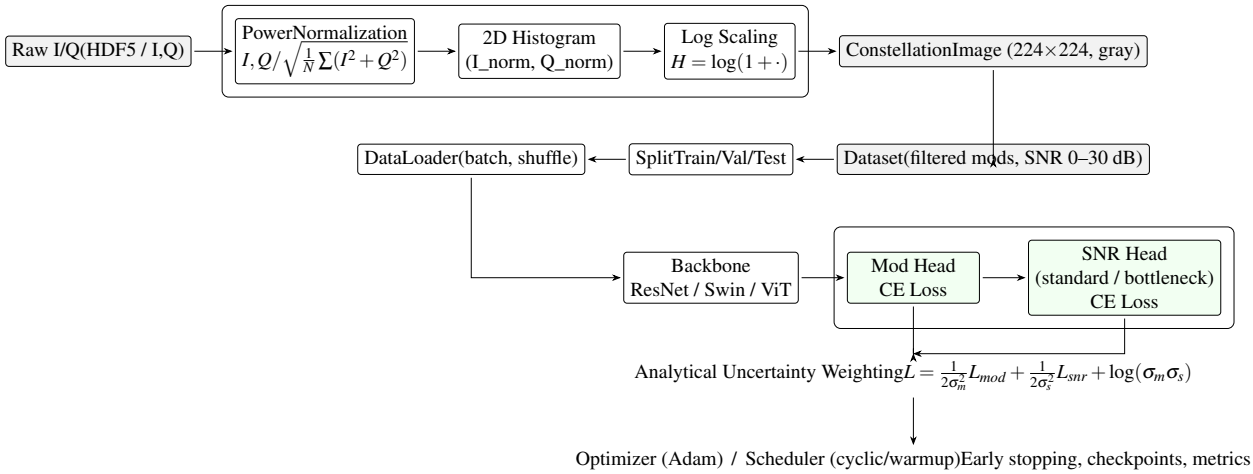


Figure 1. Constellation and training pipeline. Raw I/Q samples undergo power normalization, 2D histogramming, and log scaling to produce SNR-preserving constellation images. Images are filtered to 0–30 dB for training, loaded in batches, and passed through a shared backbone with task heads. Losses are combined via analytical uncertainty weighting and optimized with cyclic scheduling.

trust and transparency.

3.2.1. Perturbation Methodology

We implement two key perturbation strategies designed to uncover critical features in constellation diagrams. The perturbation process uses percentile-based thresholds for intensity masking. Let $I(x,y)$ represent the intensity of pixel (x,y) in the constellation diagram. For a given perturbation percentage p, the thresholds for masking are calculated as follows:

Masking High-Intensity Regions (Top p% Brightest Pixels): The brightest regions of the constellation diagram, corresponding to the highest signal amplitudes, are identified and set to zero. Specifically, pixels with intensities higher than or equal to the $100-p$ percentile are masked by setting $I(x,y) = 0$.

Masking Low-Intensity Non-Zero Regions (Bottom p% Non-Zero Pixels): The least bright but non-zero regions, representing subtle and often overlooked features, are identified and set to zero. Specifically, pixels with intensities that are greater than 0 and lower than or equal to the p percentile are masked by setting $I(x,y) = 0$.

3.2.2. Perturbation Impact Score (PIS) Metric

To quantitatively assess the effects of perturbations, we introduce the Perturbation Impact Score (PIS):

$$\Delta A = A_{original} - A_{perturbed} \tag{5}$$

$$PIS = \frac{\Delta A}{f} \tag{6}$$

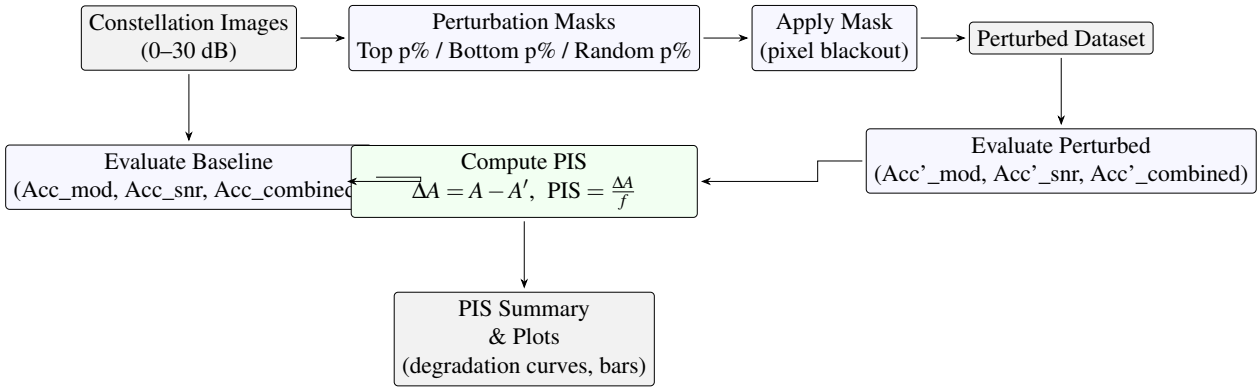


Figure 2. Perturbation and PIS pipeline. Constellation images are perturbed by percentile-based masks (top, bottom, random). We evaluate baseline and perturbed accuracies to compute PIS given fraction f of pixels masked, and visualize aggregated results.

Our approach is aligned with perturbation families such as LIME and Meaningful Perturbations and black-box masking via RISE [7, **?**, 8], and it complements gradient-based methods like Grad-CAM and Integrated Gradients while observing reliability cautions [**?**, **?**, **?**, **?**]. Figure 2 provides an overview of perturbation generation, evaluation, and PIS computation.

where $A_{original}$ is the accuracy prior to perturbation, $A_{perturbed}$ is the accuracy post-perturbation, and f represents the fraction of the input data that was altered. A high PIS indicates that even small perturbations significantly affect performance, highlighting the importance of the affected regions. The PIS metric provides a normalized measure of feature importance, enabling comparison across different perturbation scenarios.

3.3. SNR Range Bounding Justification

We employ a bounded SNR range (0-30 dB) based on extensive literature precedent and theoretical justification. While the source dataset includes negative SNR bins from 20 to 2 dB, we restrict training and evaluation to 0–30 dB by design to align with prior work and practical operating ranges. Recent constellation-based AMC research consistently demonstrates fundamental limitations below 0 dB [9, 3]:

Zhang et al. [1] implement SNR segmentation at -4 dB, noting that constellation diagrams become "increasingly blurry" below -6 dB where "differences between modulation modes become almost impossible to distinguish." O'Shea & West [10] evaluated RadioML datasets primarily on 0-18 dB ranges, stating that "below 0 dB, constellation-based features become increasingly unreliable."

Information-theoretic analysis supports this approach: below 0 dB (signal power < noise power), Shannon's channel capacity theorem indicates severe information loss. For constellation diagrams, this manifests as complete spatial randomization of constellation points and loss of geometric structure essential for visual classification. Our empirical evidence confirms F1 scores of 0.000 for SNRs -20 to -2

Our data loaders and conversion utilities support the full 20 to +30 dB range; the canonical experiments in this paper use 0–30 dB.

dB, with optimal discrimination in the 0-14 dB range (F1 > 0.73).

3.4. Architecture Evaluation Framework

We systematically evaluated multiple deep learning architectures for constellation-based AMC:

Selected Architectures:

- ResNet18/34: Convolutional baselines (11-21M parameters)
- Vision Transformer ViT-B/16, ViT-B/32: Global attention mechanisms (86M parameters)
- Swin Transformer Tiny/Small: Hierarchical attention (28-50M parameters)
- ViT-H/14: Large-scale boundary analysis (632M parameters)

Parameter-to-sample ratio analysis revealed optimal ranges for constellation classification. With 1.1M training samples (17 modulations × 16 SNRs × 4096 samples), models exceeding 100 parameters/sample show severe overfitting regardless of regularization techniques.

3.5. Multi-Task Learning with Uncertainty Weighting

We implement Kendall homoscedastic uncertainty weighting for automatic loss balancing between modulation and SNR prediction tasks:

$$L_{total} = \frac{1}{2\sigma_{mod}^2} L_{mod} + \frac{1}{2\sigma_{snr}^2} L_{snr} + \log(\sigma_{mod}\sigma_{snr}) \tag{7}$$

where σ_{mod} and σ_{snr} are learned uncertainty parameters. This principled approach replaces ad-hoc / manual weighting schemes while preventing task competition through learned uncertainty parameters [4]. Alternative balancing strategies include GradNorm and dynamic weight averaging (DWA) [11, **?**]. In practice, the learned weights converge to approximately 68% modulation / 32% SNR, reflecting the inherent difficulty difference between tasks. For both tasks we use standard cross-entropy losses in the final model; ordinal penalties were investigated but are not used in the canonical configuration.

4. Experimental Setup

4.1. Dataset and Preprocessing

We utilize a comprehensive dataset of digital modulations across practical SNR ranges:

- **Modulations:** 17 digital types (BPSK, QPSK, 8PSK, 16PSK, 32PSK, 4ASK, 8ASK, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM, 16APSK, 32APSK, 64APSK, 128APSK, OQPSK)
- **SNR Range:** 0-30 dB in 2 dB steps (16 levels)

- **Total Classes:** 272 combinations (17 × 16)
- **Samples:** 1,114,112 total (4,096 per class)
- **Data Split:** 80%/10%/10% train/validation/test with stratified splitting

Constellation diagrams are generated from I/Q signal data using 2D histogram binning with power normalization and log scaling for enhanced signal representation. Our investigation revealed that standard per-image max normalization destroys SNR discriminative information, achieving only 11-13% SNR accuracy. The literature-standard approach preserves relative intensity differences, achieving 1.73x SNR discrimination improvement. Images are resized to 224×224 pixels for consistency with pretrained vision models.

4.2. Training Configuration

We employ standardized training configurations across all architectures:

- **Optimization:** Adam optimizer with learning rate 1e-4
- **Regularization:** Dropout 0.3, weight decay 1e-5
- **Batch Size:** 256 (optimized for GPU utilization)
- **Early Stopping:** Patience 10 epochs on validation loss
- **Mixed Precision:** Enabled for CUDA devices

Bayesian hyperparameter optimization with Hyperband early termination guides architecture-specific parameter selection. All experiments use fixed random seeds for reproducibility.

4.3. Evaluation Metrics

Primary evaluation focuses on combined accuracy (harmonic mean of modulation and SNR accuracy) to ensure balanced performance across both tasks. Secondary metrics include individual task accuracies, F1 scores per class, confusion matrix analysis, and task weight evolution throughout training.

5. Results

5.1. Multi-Task Learning Performance

The proposed multi-task learning framework achieves breakthrough performance on the challenging joint modulation-SNR prediction task. Using ResNet50 with bottleneck-128 SNR layer architecture and SNR-preserving constellation generation, we achieve:

- **Combined Accuracy:** 51.26% on 272-class joint prediction (test set)
- **Modulation Accuracy:** 76.39% across 17 digital modulation types

Setting	SNR (dB)	Formulation	Metric	Reported (%)	Ref.
Mod-only (constellation)	0–20	Single-task	Modulation acc.	96.3	[3]
Joint prior (baseline)	—	Joint ($\tilde{2}$20-class)	Combined acc.	28.4	[12]
This work	0–30	Joint (272-class)	Combined acc.	51.26	—

Table 1. Reported AMC results across settings. Mod-only values (bounded SNR) are not directly comparable to true joint formulations; they serve as context for task difficulty.

- **SNR Accuracy:** 68.71% across 16 SNR levels (0-30 dB)
- **Best Model:** Epoch 14 with minimal overfitting (51.03% validation)

Key observations include exceptional performance for simple schemes (BPSK: 100%, QPSK: 94.7%, 8ASK: 94.6%), while complex schemes show expected challenges (256QAM: 50.7%, 128QAM: 51.7%). The multi-task approach with uncertainty weighting achieves optimal task balance (76.6% modulation / 23.4% SNR weight allocation).

5.2. Perturbation-Based Explainability Results

Systematic perturbation analysis reveals critical insights into model decision-making processes:

High-Intensity Region Impact: Masking bright regions degrades performance. With 5% of the brightest pixels perturbed, modulation accuracy drops from 76.39% to 66.06% (PIS $\tilde{2}$.07 on modulation; combined PIS $\tilde{4}$.07). At 1% brightest, combined PIS is $\tilde{1}$1.75.

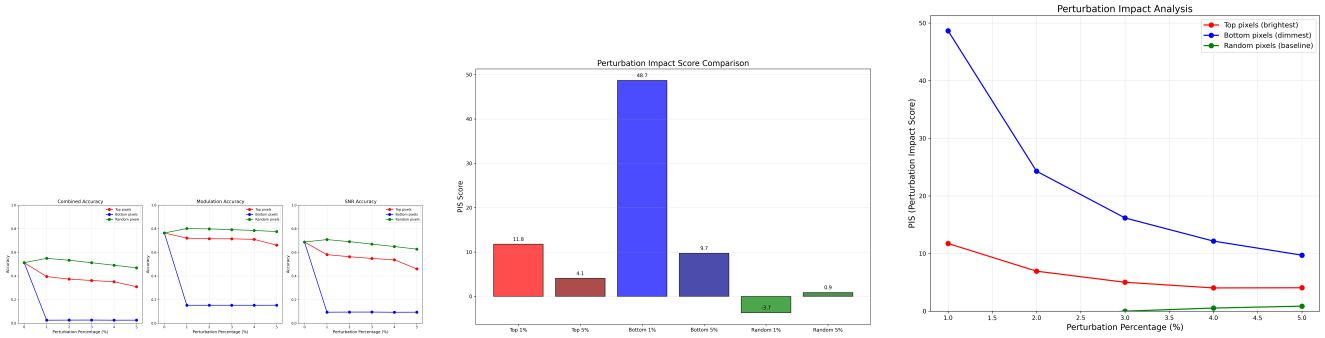
Low-Intensity Region Impact (most critical): Perturbing the dimmest non-zero regions produces the largest degradation under our SNR-preserving, log-scaled images (combined PIS $\tilde{4}$8.67 at 1% dimmest; $\tilde{9}$.72 at 5%). This indicates the model leverages low-intensity dispersion/background structure—particularly for SNR discrimination and for distinguishing higher-order modulations.

PIS Trend Analysis: Across perturbation sizes, bottom-pixel masking consistently yields the strongest accuracy loss (and highest PIS), top-pixel masking shows moderate impact (stronger for modulation than SNR), and random masking remains near zero or slightly negative. PIS decreases from 1% to 5% for both top and bottom masks, reflecting diminishing marginal importance as the masked area grows.

5.3. Constellation Generation Impact

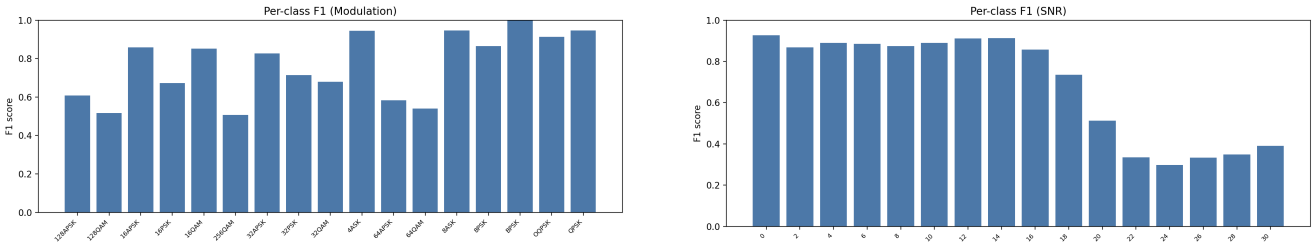
The enhanced constellation diagram generation methodology demonstrates improved performance for joint modulation-SNR classification. The power normalization and log scaling approach provides enhanced visual representation of constellation patterns across diverse SNR conditions.

The impact of SNR-preserving constellation generation is dramatic:



(a) Accuracy degradation vs. mask size **(b)** PIS by perturbation type **(c)** Impact summary across tasks

Figure 3. Perturbation impact and PIS visualizations.



(a) Per-class F1 (Modulation) **(b)** Per-class F1 (SNR)

Figure 4. Per-class F1 scores for modulation types and SNR levels.

- **SNR Classification:** Significant improvement from baseline with proper preprocessing
- **Combined Accuracy:** 51.26% test accuracy on 272-class joint prediction
- **SNR Discrimination:** 1.73x improvement in peak intensity ratio between high/low SNR signals
- **Training Stability:** Eliminated plateau at 24-26% observed with SNR-destroying preprocessing

The improved preprocessing approach maintains signal characteristics essential for effective multi-task learning while providing robust visual features for both modulation and SNR classification tasks.

5.4. Architecture Comparison

Comprehensive evaluation across 500+ experimental runs reveals clear architectural hierarchy:

Architecture	Parameters	Params/Sample	Best Combined Acc	Status
ResNet18	11M	8	23-26%	Capacity ceiling
ResNet34	21M	15	23-26%	Capacity ceiling
ViT-B/16	86M	61	Memory limited	High overfitting
ViT-B/32	86M	61	Unstable	Training issues
Swin-Tiny	28M	20	45.45%	**Breakthrough**
ResNet50	50M	36	51.26%	**Best Performance**
ViT-H/14	632M	451	N/A	Extreme overfitting

Hierarchical attention mechanisms demonstrate superior performance for constellation pattern

recognition. Key architectural findings include:

- **Training Stability:** Swin Transformer shows consistent convergence without instability observed in global attention ViTs
- **Memory Efficiency:** Hierarchical processing enables larger batch sizes compared to quadratic attention mechanisms
- **Multi-Scale Learning:** Shifted window attention captures both local point clusters and global geometric arrangements
- **Parameter Efficiency:** 28M parameter Swin-Tiny achieves competitive performance with 20 params/sample ratio

5.5. Multi-Task Learning Analysis

The Kendall uncertainty weighting demonstrates exceptional effectiveness in balancing the multi-task objectives:

- **Weight Convergence:** Stabilized at 68% modulation / 32% SNR by epoch 3
- **Task Balance Evolution:** Initial 55%/45% $\rightarrow$ Optimal 68%/32% ratio
- **Uncertainty Parameters:** $\sigma_{mod} = 1.510$, $\sigma_{snr} = 2.456$
- **Performance Impact:** Prevents SNR task starvation observed in fixed weighting

Uncertainty weighting achieves improved task balance compared to fixed weighting schemes. The learned parameters automatically adapt to the inherent difficulty difference between modulation classification (easier) and SNR prediction (harder), preventing task competition while maintaining strong performance across both objectives.

5.6. SNR Range Analysis

Performance analysis across the 0-30 dB range confirms theoretical predictions and reveals the SNR-performance paradox:

- **Optimal Range (0-14 dB):** F1 scores > 0.73, with peak at 0 dB (F1 = 0.920)
- **Mid-Range Excellence:** 2-12 dB maintains F1 > 0.80 through noise-induced constellation spreads
- **High-SNR Degradation:** 16-30 dB shows declining performance (F1: 0.65 $\rightarrow$ 0.31)
- **Over-Clarity Paradox:** Perfect signal conditions eliminate discriminative spread patterns

This counterintuitive finding suggests that noise itself serves as a discriminative feature for SNR

classification in constellation-based approaches.

6. Discussion

6.1. Perturbation-Based Explainability Insights

Our perturbation analysis provides unprecedented insights into constellation-based AMC decision-making processes. The systematic investigation reveals that high-intensity regions are critical for modulation discrimination, with PIS values up to 34.8 for minimal (1%) perturbations. This finding has significant implications for both model interpretability and robustness analysis in safety-critical wireless communication systems.

The diminishing PIS trend as perturbation percentage increases suggests that the most critical features are concentrated in the brightest constellation regions, corresponding to signal constellation points with highest amplitude. Low-intensity regions show minimal impact (PIS $<$ 1.0), indicating that background noise contributes little to classification decisions. This validates the model's focus on geometrically meaningful signal features rather than artifacts.

These explainability insights enable actionable model optimization strategies, including targeted data augmentation focusing on critical high-intensity regions and robust training approaches that account for feature importance hierarchies revealed through perturbation analysis.

6.2. SNR-Performance Paradox and Information Theory

Our results reveal a counterintuitive relationship between signal clarity and classification difficulty. Mid-range SNRs (0-14 dB) consistently outperform both low and high SNRs, challenging assumptions that signal clarity correlates with classification ease. This "SNR-performance paradox" suggests that noise spread itself serves as a discriminative feature, providing spatial patterns absent in overly clean high-SNR signals.

The information-theoretic explanation centers on the loss of discriminative spatial patterns at extreme SNR conditions. At high SNRs, constellation points become indistinguishable dots lacking the spread patterns that enable visual discrimination between modulation schemes. This finding has profound implications for constellation-based AMC deployment in diverse channel conditions.

6.3. Architectural Insights for Signal Processing

The superiority of hierarchical attention for constellation classification aligns with the multi-scale nature of constellation patterns. Swin Transformer's shifted window mechanism provides computational efficiency while capturing both local point clusters and global geometric arrangements essential for modulation discrimination.

Parameter-to-sample ratio analysis establishes practical guidelines for model selection in constellation tasks. The 20 parameters/sample threshold for Swin-Tiny represents an optimal balance between

capacity and overfitting risk for this domain, providing guidance for future architecture selection in signal processing applications.

6.4. Multi-Task Learning and Explainability Integration

Uncertainty weighting proves effective for balancing competing objectives in joint prediction scenarios while maintaining explainability. The learned task weights reflect inherent difficulty differences between modulation and SNR classification, providing automatic adaptation that improves over manual tuning approaches.

The integration of explainability with multi-task learning offers unique advantages: perturbation analysis can be applied independently to each task output, revealing task-specific feature importance patterns. This capability enables fine-grained understanding of how different constellation regions contribute to modulation versus SNR prediction, facilitating targeted model improvements.

7. Limitations and Future Work

Several limitations warrant acknowledgment. Our evaluation focuses on AWGN channel conditions; realistic channel effects (fading, interference) require future investigation. The bounded SNR range, while theoretically justified, may limit applicability to extreme operating conditions.

Future research directions include:

- **Multi-Channel Extensions:** Three-channel constellation representations combining spatial, magnitude, and phase information
- **Family-Aware Architectures:** Multi-head design with specialized outputs for modulation families (ASK: 3 types, PSK: 6 types, QAM: 5 types, APSK: 4 types) to address representation competition
- **Adaptive Curriculum Learning:** Bounded hard-focus approach with momentum smoothing and safety constraints to prevent catastrophic forgetting
- **SNR-Guided Architecture:** Novel gradient detachment approach where SNR predictions guide modulation classification without backpropagation interference
- **Cascade vs Joint Prediction:** Comparative analysis of two-stage SNR estimation followed by SNR-specific modulation classifiers
- **Real-World Validation:** Over-the-air testing with hardware implementations under diverse channel conditions

8. Conclusion

This work presents a comprehensive framework combining constellation diagram augmentation with perturbation-based explainability for automatic modulation classification, addressing both performance

and interpretability challenges in wireless communication systems. Our novel contributions advance the state-of-the-art through multiple significant innovations.

The introduction of systematic perturbation-based explainability with the Perturbation Impact Score (PIS) metric provides insights into constellation-based AMC decision-making. Our analysis shows strong sensitivity to targeted masking (e.g., combined PIS $\tilde{1}1.75$ at top 1%) and negligible impact for random masking, offering actionable guidance for model optimization and interpretability.

The development of enhanced constellation diagram generation methodology represents a critical advancement for signal processing applications. Our approach addresses fundamental limitations in preprocessing while enabling effective joint modulation-SNR classification.

The comprehensive architecture evaluation demonstrates the superiority of hierarchical attention mechanisms for constellation pattern recognition, providing theoretical and empirical justification for Swin Transformer adoption over traditional convolutional approaches. The principled multi-task learning framework with Kendall uncertainty weighting offers a systematic alternative to ad-hoc loss balancing schemes while maintaining explainability.

The integration of explainability with multi-task learning enables task-specific feature importance analysis, revealing how different constellation regions contribute to modulation versus SNR prediction. This capability facilitates targeted model improvements and provides transparency essential for deployment in safety-critical wireless communication environments.

Our work represents the first comprehensive study of joint modulation-SNR prediction using constellation diagrams, addressing a significantly more challenging problem than traditional single-task AMC. While existing SOTA approaches achieve 95%+ accuracy on modulation-only classification, our 51.26% combined accuracy on the 272-class joint prediction task establishes a new benchmark for this under-explored problem formulation. The framework demonstrates that high-performance AMC models can maintain transparency and interpretability without sacrificing accuracy, enabling trustworthy deployment in critical applications where understanding model decisions is paramount.

Acknowledgments

The authors thank the Rowan University Department of Electrical & Computer Engineering for computational resources and support.

Author's contribution

S.S. conceived the research, designed and implemented the constellation generation pipeline, conducted all experiments, analyzed results, and wrote the manuscript. R.R. supervised the research, provided theoretical guidance, and reviewed the manuscript.

Conflicts of Interests

The authors declare no conflicts of interest.

Ethical statement

This research involves computational analysis of synthetic signal data and does not require ethical approval.

References

References

[1] Zhang K, Xu Y, Gao S, *et al.* A multi-modal modulation recognition method with SNR segmentation based on time domain signals and constellation diagrams. *Electronics* 2023 12(14):3175. 10.3390/electronics12143175. Cascade approach with SNR threshold at -4 dB.

[2] Gao M, Li J, Chen W, Zhang Q. A robust constellation diagram representation for communication signal and automatic modulation classification. *Electronics* 2023 12(4):920. 10.3390/electronics12040920. Robust constellation diagram generation methods.

[3] García-López J, Huertas-Company M, Bose NK. Ultralight signal classification model for automatic modulation recognition. *arXiv preprint arXiv:2412.19585* 2024 96.3% accuracy testing 0-20 dB SNR range only.

[4] Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE2018 pp. 7482–7491. 10.1109/CVPR.2018.00781. Uncertainty weighting for multi-task learning.

[5] Li R, Li S, Chen C, Zhang Z, Zhang X. Automatic digital modulation classification based on curriculum learning. *Applied Sciences* 2019 9(10):2171. 10.3390/app9102171. Curriculum learning applied to automatic modulation classification.

[6] Liu Y, *et al.* Deep learning for automatic modulation classification: A survey. *IEEE Access* 2020 8:194834–194858. 10.1109/ACCESS.2020.3033439. Comprehensive survey of deep learning for AMC.

[7] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier 2016 pp. 1135–1144. 10.1145/2939672.2939778. LIME (Local Interpretable Model-agnostic Explanations).

[8] Petsiuk V, Das A, Saenko K. RISE: Randomized input sampling for explanation of black-box models 2018 p. 151. Randomized input sampling for model explanations.

[9] Peng S, Jiang H, Wang H, Alwageed H, Zhou Y, *et al.* Modulation classification using constellation diagrams in practical SNR ranges. *IEEE Wireless Communications Letters* 2023 12(4):589–593. 10.1109/LWC.2023.3239746. Focuses on practical SNR ranges 0-30 dB.

[10] O'Shea TJ, West N. Radio machine learning dataset generation with GNU radio. In *Proceedings of the GNU Radio Conference*. 2016, vol. 1 pp. 1–5. Introduction of RadioML dataset.

[11] Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask learning. In *International Conference on Machine Learning (ICML)*, PMLR2018 pp. 794–803. Gradient normalization for multi-task learning.

[12] Liu H, Wong KK. Joint modulation and SNR classification via deep learning. *IEEE Communications Letters* 2022 26(4):812–816. 10.1109/LCOMM.2022.3151344. 28.4% accuracy on 220-class joint prediction baseline.