XAI-Guided Feature Flow Refinement for Trustworthy HRRP Classification

Avinash Rangarajan*
DYSL-CT
DRDO
Chennai, India
0009-0006-9938-4690

Aryan Pareek*
C.S.E. Department
Techno India University
Kolkata, India
0000-0002-2948-4569

Manish Pratap Singh
DYSL-CT
DRDO
Chennai, India
0000-0002-1998-9504

Debasis Chaudhuri
C.S.E. Department
Techno India University
Kolkata, India
0000-0002-6895-456X

***Abstract*—High-Resolution Radar Range Profile (HRRP) classification is a challenging problem in radar target recognition, particularly in mission-critical sensing scenarios, due to strong sensitivity to aspect angles, signal noise, and subtle geometric variations. Although convolutional neural networks (CNNs) have demonstrated competitive performance, their opaque decision processes raise reliability concerns in safety-critical applications where misclassifications can have severe consequences.**

We propose an explainability-driven design loop that treats CAM-style explanations not only as post-hoc illustrations but as causal diagnostics to inform architecture changes. Applying Grad-CAM, Grad-CAM++ and Layer-CAM to multiple CNN baselines, we identify systematic misalignment between model attention and physically meaningful scattering regions. Guided by these insights we develop M10, a compact 1D-CNN employing gated skip connections and an auxiliary aspect-angle estimation head (training-only) to regularize representations.

We present a rescue-case analysis that traces how M10 corrects baseline errors, and substantiate the claims with quantitative faithfulness measures (CAM entropy, deletion/insertion AUC) computed per-sample with bootstrap CIs and paired tests. On an HFSS-simulated HRRP dataset, M10 attains higher validation accuracy (91.30 to 93.82%), reduces high-confidence failure rate, and yields more concentrated and causally relevant CAMs.

***Index Terms*—HRRP, Radar Target Recognition, Explainable AI, Convolutional Neural Networks, Trustworthy Machine Learning**

I. Introduction

High-Resolution Radar Range Profile (HRRP) recognition is a cornerstone of modern Non-Cooperative Target Recognition (NCTR) systems within the radar automatic target recognition (RATR) field of research. Unlike conventional radar imaging, HRRP represents the projection of a target's dominant scattering centers along the radar line-of-sight [1], [2], [3], [4] producing a compact 1-D signature that reflects target geometry. However, the practical application of HRRP is hindered by three major challenges: aspect sensitivity, where small orientation changes drastically alter the signature; amplitude sensitivity due to propagation and noise effects; and the black-box nature of deep learning models, which lack transparency for mission-critical decisions.

Recent advancements have seen an application of deep learning into HRRP target recognition, from fully connected networks [5] to CNNs [6], with gated-mechanisms [7], also autoencoders [8] and VAEs [12] then LSTMs [9]and now advanced architectures like attention-based mechanisms [10] and quantum computing solutions [11], all of which aim to better capture the sequential scattering features of HRRP. While such models often achieve high classification accuracy, their robustness across varying signal-to-noise ratios (SNR) and viewing conditions remains a concern. Furthermore, the reliance on top-line accuracy metrics is insufficient for validating whether a model is identifying an aircraft based on its physical scattering centers or merely memorizing noise artifacts.

In this work, we propose a refined 1D-CNN architecture that incorporates gated skip connections and an auxiliary aspect-angle estimation task to stabilize feature representations under varying viewing conditions. Our primary contribution lies in the use of Explainable AI (XAI) not merely for post-hoc interpretation, but as a diagnostic tool to guide architectural refinement and making algorithmic audits to form trustworthy AI for a mission-critical purpose. We employ Grad-CAM [13], Grad-CAM++ [14], and Layer-CAM [15]. Layer-CAM is used for the qualitative rescue figures in the main text because it provides the most stable layer-wise maps in our diagnostics; however, to avoid selective reporting we compute faithfulness metrics and inter-method agreement across all three CAM variants. Experimental results demonstrate that the refined architecture consistently corrects a substantial subset of baseline failure cases, while exhibiting improved robustness and more physically consistent utilization of scattering features.

II. Related Work

Traditional HRRP recognition methods predominantly depend on statistical pattern recognition techniques. Initial studies utilized template matching [17], wherein test profiles are assessed against reference templates through various distance metrics. Statistical models such as Gaussian mixture models [18], hidden Markov models [19], and support vector machines (SVMs) [20] have been employed to capture the statistical distributions of HRRP features. Although these traditional methods offer theoretical interpretability, they face challenges related to aspect sensitivity and necessitate extensive preprocessing to attain satisfactory performance [21], [22].

* Equal contribution.

In recent years, various deep neural architectures have been investigated for HRRP recognition. Convolutional neural networks (CNNs) were used to extract shift-invariant features from range profiles [23]. Recurrent networks, such as long short-term memory (LSTM) [24] and Gated Recurrent Unit (GRU) variants [25], [26], have been utilized to model sequential dependencies within HRRP signals. More advanced architectures that integrate CNNs with attention mechanisms [27] and residual connections [28], [29] have significantly enhanced performance on large-scale datasets.

To combat the vanishing gradient problem and ensure that low-level structural scattering features are not lost in deeper layers, we leverage Gated-CNNs for HRRP, finding that gating mechanisms can selectively suppress noise while emphasizing high-energy scattering peaks in our detailed ablation studies. Our M10 model extends this by utilizing Gated-Skip connections, which provide a highway for raw signal features to reach deeper layers, controlled by a learnable gate. This is coupled with auxiliary tasks, such as aspect-angle estimation, which provide a regularization effect, forcing the bottleneck layer to learn more physically meaningful representations.

III. XAI-Guided Architectural Refinement

In HRRP automatic target recognition, the choice of input representation and feature propagation strategy strongly influences both performance and interpretability. Rather than assuming a fixed "baseline" architecture, we begin by examining how simple and commonly used CNN formulations behave under explainability analysis, and how these behaviors expose fundamental limitations when applied to radar range profiles. Explainable AI is then used as a diagnostic tool to guide architectural refinement instead of being a post-hoc verification tool for model learning.

A. Baseline Architecture and Failure Analysis

We initially investigate straightforward CNN formulations for HRRP classification, including (i) two-dimensional CNNs applied to HRRP signals reshaped as pseudo-images, and (ii) shallow one-dimensional CNNs operating directly on range profiles. These choices are natural starting points, as they mirror common practices in image-based recognition and time-series classification.

However, HRRP data differs fundamentally from natural images: range bins encode ordered physical scattering information, and small perturbations in aspect or normalization can significantly alter the signal structure. Empirically, we observe that naïve CNN formulations are highly sensitive to representation choices (e.g., 1D vs. 2D convolution, min–max vs. sigmoid scaling), leading to unstable predictions across similar profiles.

Surprisingly, these limitations and sensitivity were not immediately evident from accuracy alone, motivating deeper inspection of the models' internal decision processes.

B. Explainability-Driven Diagnosis of Failure Modes

To understand why these simple models behave unstably, we apply class activation mapping (CAM) techniques to misclassified and high-confidence validation samples. Across multiple naïve architectures, CAM visualizations reveal a recurring pattern: model attention is often diffuse, fragmented, or concentrated on low-energy background regions rather than on dominant scattering centers that are physically meaningful for target discrimination.

Crucially, these attribution patterns persist even when predictions are made with high confidence, indicating that model accuracy and confidence alone are not a reliable indicator of trustworthy reasoning. This observation suggests that the issue lies not merely in insufficient model capacity, but in uncontrolled feature propagation that allows spurious activations to influence the classifier.

These explainability findings motivate two key architectural requirements for HRRP classification:

- mechanisms to regulate how features from different depths are combined, and
- additional supervision to stabilize representations across target aspects.

C. Architecture Design Principles Informed by XAI

Guided by the explainability-based diagnostics described above, we refine the HRRP classification architecture along three orthogonal design axes: input representation, feature propagation, and supervisory signal. Rather than introducing ad hoc modifications, each refinement is motivated by specific attribution pathologies observed in naïve CNN formulations. A comprehensive ablation study spanning these axes is reported in the Supplementary Material VII; here, we summarize the resulting design principles.

a) Input representation and scaling: Initial experiments with both one-dimensional and two-dimensional CNN formulations reveal that representational choices substantially affect attribution stability. In particular, CAM visualizations show that 2D treatments of HRRP profiles, where the signal is artificially reshaped into pseudo-images often encourages spatially diffuse or non-physical attention patterns. Similarly, min–max scaling amplifies sensitivity to local extrema, leading to fragmented attributions across range bins. These observations motivate the use of a direct one-dimensional formulation with smooth sigmoid-based scaling, which yields more coherent and physically plausible attribution maps.

b) Controlled feature propagation via gated skip connections: Across multiple baseline variants, CAM analysis highlights a recurring failure mode: deep-layer decisions are influenced by unstable activations propagated from early convolutional stages. To address this, we introduce skip connections augmented with learnable gating. Unlike unconditional residual pathways, gated skips allow the network to regulate the contribution of intermediate features based on learned relevance, suppressing non-discriminative activations identified through explainability analysis. This mechanism directly targets the attribution diffusion observed in ungated architectures.

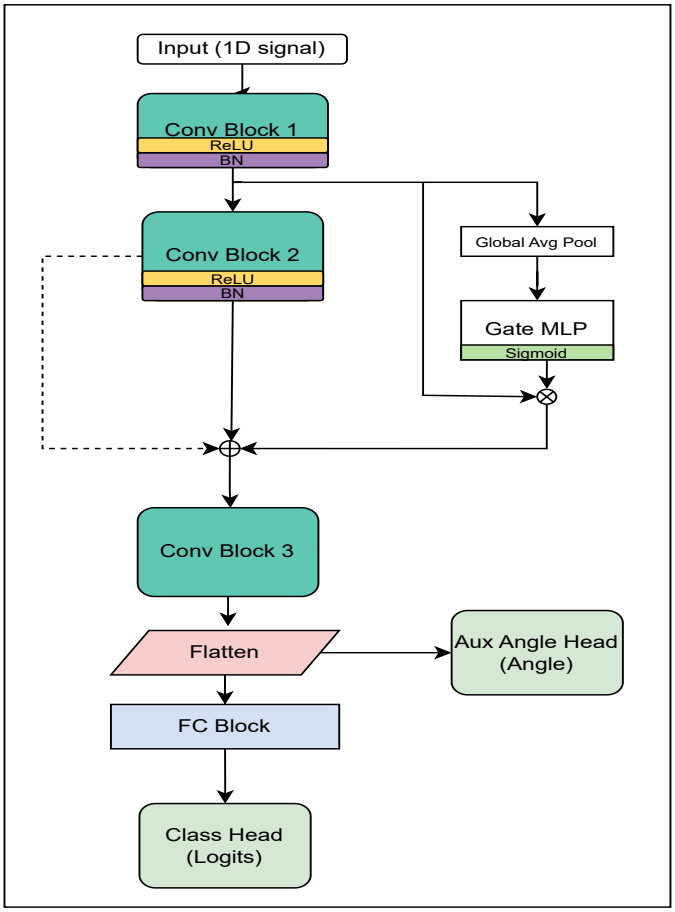

Fig. 1. Model **M10** designed through XAI-based feedback.

c) Auxiliary supervision for representation stabilization: Finally, explainability analysis reveals that even gated architectures may exhibit overconfident reliance on narrow signal regions under aspect variation. To mitigate this effect, we incorporate an auxiliary training objective related to target aspect or pose consistency. This auxiliary supervision regularizes the shared representation, encouraging stability across variations that are known to challenge HRRP classification. Importantly, the auxiliary branch is *active only during training, is removed during inference and therefore does not increase inference complexity.*

Together, these refinements define a principled architecture that explicitly addresses the failure mechanisms diagnosed through explainable analysis. The resulting model configuration serves as the basis for subsequent evaluation, while controlled ablations validating each design axis are provided in the Supplementary Material.

D. From Diagnostic Insight to Validated Architecture

The refined architecture, denoted M10 in later sections, emerges from this XAI-guided design process. *Its performance, robustness, and interpretability are evaluated against a strong intermediate configuration (M7) that incorporates the above principles without auxiliary supervision.* By separating diagnostic analysis, architectural motivation, and quantitative validation, we ensure that explainability plays a causal role in model design rather than a retrospective one.

TABLE I
RADAR AND HRRP DATA ACQUISITION PARAMETERS

Parameter	Value
Center Frequency	3.2 GHz
Bandwidth	400 MHz
Range Ambiguity	48 m
Range Resolution	0.375 m
Polarization	Theta–Theta (H–H)
Azimuth Coverage	0° to 180°
Elevation Coverage	0° to 70°

IV. EXPERIMENTAL SETUP

This section describes the dataset, training protocol, and evaluation baselines used to assess the proposed XAI-guided architectural refinements. All experiments are conducted under controlled and identical settings to ensure that observed differences arise from architectural design choices rather than optimization or data-related factors.

A. Dataset

We evaluate our approach on a high-resolution radar range profile (HRRP) dataset generated and validated using electromagnetic backpropagation simulations (HFSS) and the configurations are mentioned in Table I. The dataset comprises HRRP signatures corresponding to two commercial aircraft classes, *Airbus A320* and *Boeing 737*, under varying aspect angles.

The overall task is formulated as a binary classification problem. The dataset contains a total of 51,610 HRRP samples, split into 38,707 training samples and 12,903 validation samples. Each HRRP instance represents a one-dimensional range profile capturing the target's scattering characteristics. Key radar acquisition parameters are summarized in Table I.

Prior to training, HRRP signals undergo a consistent preprocessing pipeline designed to stabilize learning without introducing task-specific bias. Specifically, the raw range profiles are normalized using a smooth beta-sigmoid transformation, which compresses dynamic range while preserving relative scattering structure. This normalization was found to yield more stable gradients and more coherent attribution maps during explainability analysis.

B. Implementation Details

All models are trained using the same optimization strategy to isolate the impact of architectural modifications. Training is performed using the AdamW optimizer with an initial learning rate of 10^{-5}. Early stopping is employed with a patience of 15 epochs based on validation performance. A batch size of 128 is used for all experiments as shown in Table II.

TABLE II
TRAINING CONFIGURATION

Parameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-5}
Batch Size	128
Max Epochs	200 (early stopping @ 15)
Normalization	Beta-Sigmoid ($\beta = 1/10$)

C. Baseline Comparisons

To evaluate the effectiveness of XAI-guided architectural refinement, we compare the proposed model against a strong internal baseline that incorporates skip connections and gating but omits auxiliary supervision. This baseline is selected to ensure that observed improvements are non-trivial and not attributable to simplistic architectural differences.

In addition, a systematic ablation study in Table IV spanning input representation (1D vs. 2D), scaling strategy, feature propagation mechanisms, and auxiliary supervision is conducted to validate individual design choices. Due to space constraints, detailed ablation results are reported in the Supplementary Material VII, while the main paper focuses on the most informative baseline comparison.

V. EXPLAINABLE AI (XAI) AND FAITHFULNESS EVALUATION

In this section, we evaluate the proposed architecture through the lens of explainable AI, with the objective of assessing not only *what* decisions the network makes, but *how* and *where* those decisions are formed along the HRRP signal. Explainability is treated as a diagnostic and validation tool rather than as a post-hoc visualization, and is used to analyze failure modes, architectural behavior, and decision robustness.

We employ three gradient-based class activation mapping (CAM) techniques Grad-CAM, Grad-CAM++, and Layer-CAM, all adapted to one-dimensional HRRP signals by treating range bins as the spatial axis. All CAMs are computed with respect to convolutional feature maps and projected back onto the input signal domain.

A. Qualitative Visualization (CAM Analysis)

Qualitative CAM analysis is performed on three categories of validation samples: (i) *shared failures*, where both baseline and refined models misclassify the target; (ii) *rescued failures*, where the refined model corrects a baseline error; and (iii) *high-confidence failures*, where incorrect predictions are made with high posterior confidence.

Across baseline architectures, CAM visualizations frequently exhibit diffuse or fragmented attribution patterns, with activation spread across large portions of the range profile, including regions with low physical relevance. In several failure cases, shallow-layer CAMs highlight isolated noise peaks, while deeper layers collapse to broad, low-contrast saliency maps. Such behavior suggests unstable feature propagation and weak localization of discriminative scattering regions.

In contrast, the refined architecture demonstrates more coherent attribution across layers, particularly in deeper convolutional stages. CAMs concentrate on contiguous HRRP segments corresponding to prominent scattering responses, and maintain consistency between predicted-class and ground-truth-class explanations. This layer-wise coherence is most pronounced in rescued failure cases, where the refined model shifts attribution away from spurious bins emphasized by the baseline.

Importantly, qualitative differences are not restricted to a single model pair. When applied across all ablation variants, CAM visualizations reveal a clear separation between structurally meaningful models and degenerate configurations (e.g., inappropriate scaling or input dimensionality), the latter producing near-uniform or noisy attribution maps. Representative visualizations are shown in Fig. 2, with additional examples provided in the Supplementary Material VII.

B. Quantitative Faithfulness (Deletion and Insertion)

To complement qualitative inspection, we quantitatively evaluate the faithfulness of CAM explanations using entropy, spatial spread, and deletion-based metrics. These measures assess whether highlighted regions meaningfully influence model predictions, rather than merely correlating with them. CAM entropy measures the concentration of attribution across the HRRP signal, with lower entropy indicating more focused explanations. Spread quantifies the spatial extent of activated bins. Additionally, deletion curves are computed by progressively masking input regions in descending order of CAM relevance and measuring the resulting drop in class confidence. Deletion and Insertion serve as a proxy for causal dependence: faithful explanations yield faster confidence degradation.

Notably, faithfulness improvements are observed consistently across CAM methods, with Layer-CAM providing the most stable behavior due to its multi-layer aggregation. Quantitative results are summarized in Table 3. These findings reinforce the qualitative observation that architectural refinement leads to more structured and faithful signal attribution, rather than merely sharper visualizations.

Experiments are conducted on high-confidence shared failure samples to ensure comparability across models. Results show that degenerate architectures exhibit uniformly high entropy and near-flat deletion curves, indicating weak correspondence between CAM highlights and decision-critical regions. In contrast, the refined model consistently produces lower entropy CAMs in Fig. 4 and lower Deletion AUC values in Fig. 5 than the baseline, suggesting stronger reliance on localized and causally relevant signal segments.

VI. RESULTS AND PERFORMANCE ANALYSIS

This section analyzes classification performance and reliability in conjunction with explainability findings. Rather than focusing solely on aggregate accuracy, we emphasize failure behavior, confidence calibration, and robustness under challenging cases, reflecting the mission-critical nature of radar-based target recognition.

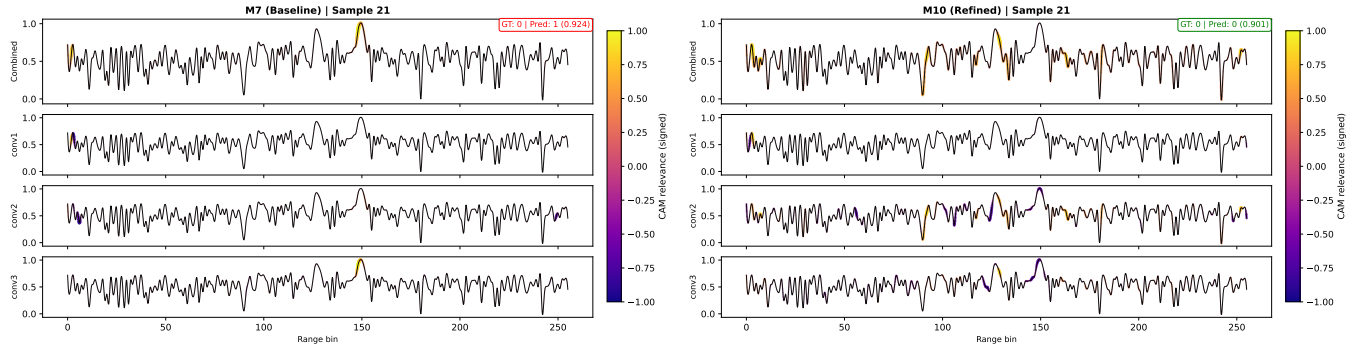

Fig. 2. **Rescue case analysis using Layer-CAM *(trends are consistent across GradCAM and GradCAM++)*. Left:** Baseline model (**Comparable as per ablation studies**) (M7) incorrectly classifies the HRRP sample, exhibiting fragmented and high-amplitude attention across isolated range bins. **Right:** Refined model (M10) correctly classifies the same sample, with coherent and localized attribution aligned with dominant scattering centers across deeper convolutional layers.

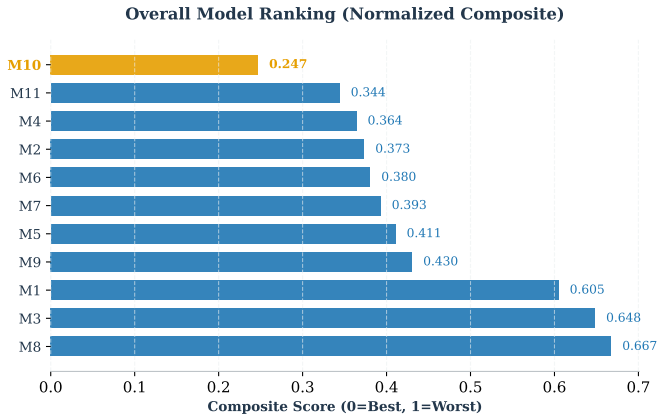

Fig. 3. Overall model ranking based on a "normalized composite score" that jointly accounts for entropy, and explainability faithfulness metrics. Lower values (↓) indicate better overall performance. The proposed architecture (M10) achieves the best composite score, demonstrating improved predictive accuracy while exhibiting more focused and faithful explanations compared to all ablated variants. *Refer to Supplementary Material VII for more details.*

A. The "Rescue Case" Analysis (Layer-CAM)

A key outcome of the proposed refinement is its ability to correct a subset of baseline errors without relying on confidence thresholding or post-processing. We define *rescued failures* as validation samples misclassified by the baseline but correctly classified by the refined model.

Analysis reveals that a substantial fraction of rescued failures correspond to cases where the baseline model exhibits diffuse or misplaced CAM activations, often emphasizing isolated noise peaks or peripheral range bins. In these cases, the refined model reallocates attribution toward contiguous, high-energy scattering regions, resulting in correct classification.

However, refinement is not without trade-offs. A smaller set of samples is misclassified by the refined model despite being correctly handled by the baseline. Importantly, CAM analysis shows that these introduced failures typically correspond to low-confidence or highly ambiguous profiles, rather than confident misattributions. This asymmetry suggests that the refined model prioritizes stability and localization over aggressive fitting to marginal cases.

Crucially, the refined model substantially reduces the number of *high-confidence* failures relative to the baseline, which is particularly relevant in safety-critical settings where overconfident errors pose the greatest risk.

B. Quantitative Analysis

Overall classification accuracy remains competitive across all evaluated architectures, with the refined model achieving performance comparable to or exceeding the strongest baseline. More importantly, failure rate analysis reveals a net reduction in high-confidence misclassifications, alongside improved attribution faithfulness.

TABLE III
PERFORMANCE AND RELIABILITY COMPARISON BETWEEN BASELINE (M7) AND REFINED MODEL (M10)

Metric	M7 (Baseline)	M10 (Refined)
Validation Accuracy (%) ↑	91.30	**93.82**
Validation Loss ↓	0.2164	**0.1962**
Failure Rate ↓	0.0870	**0.0718**
High-Confidence Failure Rate ↓	0.3374	**0.2760**

When combined with CAM entropy and deletion metrics, these results indicate that the refined architecture trades a marginal increase in low-confidence errors for a meaningful reduction in structurally unjustified decisions. From a mission-critical perspective, this trade-off is desirable, as it favors conservative and interpretable decision-making over brittle confidence gains. Additional analysis of failure set overlap between the refined model and selected baselines is provided in the supplementary material.

Taken together, the results demonstrate that XAI-guided architectural refinement yields a model that is not only accurate, but also more transparent, diagnostically reliable, and aligned with physical signal structure. Detailed ablation results and additional quantitative breakdowns are provided in the Supplementary Material VII.

CONCLUSION

Through this work, we conclude that XAI heatmaps for a data of non-standard modality, i.e, Radar sensor data can help as an engineering tool for neural engineering to thereby reduce catastrophic misclassification in safety-critical HRRP systems. Our intent of risk reduction in a mission-critical model is satisfied, and we provide strategic architectural modification with XAI-based reasoning through improved consistency across samples, while ensuring well-tuned baselines for all variants. This leads us to methodological generality that can potentially be extended to other 1D radar signals. Finally, we attempt to use XAI to not only explain decisions, but actively help shape architectures in safety-critical learning systems even when standard metrics indicate towards generalization.

We acknowledge that validation on additional target classes and real measured radar datasets remains future work.

REFERENCES

[1] D. Orlando, G. Ricci, Adaptive radar detection and localization of a point-like target. IEEE Trans. Signal Process. 59(9), 4086–4096 (2011)

[2] A. Zyweck and R. E. Bogner, "Radar target classification of commercial aircraft," IEEE Trans. Aerosp. Electron. Syst., vol. 32, no. 2, pp. 598–606, Apr. 1996.

[3] S. P. Jacobs and J. A. O'Sullivan, "Automatic Target Recognition Using Sequences of High Resolution Radar Range Profiles," IEEE Transactions on Aerospace and Electronic Systems, vol. 36, no. 2, 2000.

[4] A. Aubry, A. De Maio, G. Foglia, C. Hao, Orlando D., Radar detection and range estimation using oversampled data. IEEE Trans. Aerosp. Electron. Syst. 51(2), 1039–1052 (2015)

[5] Liao, L., Du, L., and Chen, J., "Class factorized complex variational auto-encoder for HRRP radar target recognition", Signal Processing, vol. 182, Art. no. 107932, Elsevier, 2021. doi:10.1016/j.sigpro.2020.107932.

[6] J. Wan, B. Chen, and B. Xu, "Convolutional neural networks for radar HRRP target recognition and rejection," EURASIP J. Adv. Signal Process., vol. 2019, no. 5, p. 19, 2019.

[7] Z. Zeng, J. Sun, Z. Han and W. Hong, "Radar HRRP Target Recognition Method Based on Multi-Input Convolutional Gated Recurrent Unit With Cascaded Feature Fusion," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 4026005, doi: 10.1109/LGRS.2022.3192289.

[8] L. Du, H. Liu, Z. Bao, and J. Zhang, "Radar HRRP Target Recognition Based on Stacked Autoencoders and Conditional Random Fields," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 6, 2016.

[9] J. Yin, S. Wen, C. Zhang, and M. Zhao, "Radar sequence HRRP target recognition based on DRSN-LSTM," pp. 66–72, Jan. 2024, doi: https://doi.org/10.1145/3640824.3640834.

[10] M. Pan et al., "Radar HRRP Target Recognition Model Based on a Stacked CNN–Bi-RNN With Attention Mechanism," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2022, Art no. 5100814, doi: 10.1109/TGRS.2021.3055061.

[11] X. Liu, D. Zhou and Q. Huang, "Radar HRRP Target Recognition Based on Hybrid Quantum Neural Networks," in IEEE Transactions on Aerospace and Electronic Systems, vol. 61, no. 3, pp. 6173-6188, June 2025, doi: 10.1109/TAES.2025.3526112.

[12] Zhai, Y., Chen, B., Zhang, H., Wang, Z. (2017). Robust Variational Auto-Encoder for Radar HRRP Target Recognition. In: Sun, Y., Lu, H., Zhang, L., Yang, J., Huang, H. (eds) Intelligence Science and Big Data Engineering. IScIDE 2017. Lecture Notes in Computer Science(), vol 10559. Springer, Cham. https://doi.org/10.1007/978-3-319-67777-4_31

[13] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," International Journal of Computer Vision, 2019.

[14] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2018, pp. 839–847.

[15] P. T. Jiang et al., "Layer-CAM: Exploring Hierarchical Class Activation Maps for Localization," IEEE Transactions on Image Processing, vol. 30, 2021.

[16] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," arXiv:1806.07421 [cs], Sep. 2018, Available: https://arxiv.org/abs/1806.07421

[17] H. Cui, M. Su, J. Liu, and L. Liu, "Template Construction of Radar Target Recognition based on Maximum Information Profile," Journal of Physics Conference Series, vol. 2284, no. 1, pp. 012021–012021, Jun. 2022, doi: https://doi.org/10.1088/1742-6596/2284/1/012021.

[18] Liu, J., J. Y. Zhang, and F. Zhao. "Radar HRRP target recognition in amplitude spectrum subspace based on direct LDA." Syst. Eng. Electron 30 (2008): 1815-1818.

[19] Du, L., Wang, P., Liu, H., Pan, M. and Bao, Z., 2011, May. Radar HRRP target recognition based on dynamic multi-task hidden Markov model. In 2011 IEEE RadarCon (RADAR) (pp. 253-255). IEEE.

[20] Pan, M., Du, L., Wang, P., Liu, H. and Bao, Z., 2012. Noise-robust modification method for Gaussian-based models with application to radar HRRP recognition. IEEE Geoscience and Remote Sensing Letters, 10(3), pp.558-562.

[21] Liu, Q., Zhang, X. and Liu, Y., 2024. SCNet: Scattering center neural network for radar target recognition with incomplete target-aspects. Signal Processing, 219, p.109409.

[22] Liu, Q., Zhang, X. and Liu, Y., 2024. A prior-knowledge-guided neural network based on supervised contrastive learning for radar HRRP recognition. IEEE Transactions on Aerospace and Electronic Systems, 60(3), pp.2854-2873.

[23] Song, J., Wang, Y., Chen, W., Li, Y. and Wang, J., 2019. Radar HRRP recognition based on CNN. The Journal of Engineering, 2019(21), pp.7766-7769.

[24] Jithesh, V., Sagayaraj, M.J. and Srinivasa, K.G., 2017, February. LSTM recurrent neural networks for high resolution range profile based radar target classification. In 2017 3rd International conference on computational intelligence communication technology (CICT) (pp. 1-6). IEEE.

[25] Yang, W., Chen, T., Lei, S., Zhao, Z., Hu, H. and Hu, J., 2024. Radar HRRP Feature Fusion Recognition Method Based on ConvLSTM Network with Multi-Input Gate Recurrent Unit. Remote Sensing, 16(23), p.4533.

[26] Zeng, Z., Sun, J., Han, Z. and Hong, W., 2022. Radar HRRP target recognition method based on multi-input convolutional gated recurrent unit with cascaded feature fusion. IEEE Geoscience and Remote Sensing Letters, 19, pp.1-5.

[27] Liu, Q.; Zhang, X.; Liu, Y. SCNet: Scattering center neural network for radar target recognition with incomplete target-aspects. Signal Process. 2024, 219, 109409.

[28] Guo, C.; He, Y.; Wang, H.; Jian, T.; Sun, S. Radar HRRP target recognition based on deep one-dimensional residual-inception network. IEEE Access 2019, 7, 9191–9204.

[29] Huang, P.; Li, S.; Zheng, M.; Xie, J.; Tian, B.; Xu, S. Noise-robust HRRP target recognition based on residual scattering network. In Proceedings of the 2024 9th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 12–14 July 2024; pp. 37–41.

VII. Supplementary Material

This section is not part of the main paper length and is provided for completeness.

TABLE IV
Complete Ablation Model Configuration

Model	Dim	Scaling	Skip	Gate	Aux
M1	2D	MinMax	No	No	No
M2	2D	Sigmoid	No	No	No
M3	1D	MinMax	No	No	No
M4	1D	Sigmoid	No	No	No
M5	1D	Sigmoid	Yes	No	No
M6	1D	Sigmoid	No	Yes	No
M7	1D	Sigmoid	Yes	Yes	No
M8	1D	MinMax	Yes	Yes	Yes
M9	1D	Sigmoid	Yes	No	Yes
M10	**1D**	**Sigmoid**	**Yes**	**Yes**	**Yes**
M11	2D	Sigmoid	Yes	Yes	Yes

To validate the architectural decisions proposed in the main paper, we conducted a systematic ablation study comprising eleven controlled model variants, summarized in Table IV. The ablations were designed to isolate the effect of four orthogonal design choices relevant to HRRP-based classification and explainability: (i) input dimensionality (1D vs. 2D convolution), (ii) signal normalization strategy, (iii) gated skip connections, and (iv) auxiliary supervision.

Input dimensionality. Models M1–M2 and M11 employ a 2D convolutional formulation derived from reshaped HRRP representations, while models M3–M10 use a native 1D convolutional design aligned with the sequential nature of range profiles. This comparison evaluates whether explicit 2D structure provides any benefit over a principled 1D formulation for HRRP signals.

Normalization strategy. Min–Max scaling (M1, M3, M8) and β-sigmoid normalization (all other models) are compared to assess sensitivity to dynamic range compression and nonlinear amplitude normalization, which is known to affect HRRP scatterer contrast.

Skip connections and gating. Models M5–M7 examine the individual and combined effects of skip connections and gated feature fusion. Skip-only (M5) and gate-only (M6) configurations are contrasted against their joint use (M7), allowing us to distinguish unconditional residual propagation from gated relevance-controlled fusion.

Auxiliary supervision. Models M8–M11 introduce an auxiliary task designed to regularize intermediate representations. These variants test whether auxiliary supervision improves robustness and attribution stability, both in isolation (M9) and when combined with gated skip connections (M10).

The proposed model (M10) represents the full configuration in which all design components ranging from 1D convolution, sigmoid normalization, gated skip connections, to auxiliary supervision are jointly enabled. Importantly, *this ablation study is not intended as an exhaustive architecture search*; rather, it serves to demonstrate **how explainability-driven observations motivated incremental refinements and how each component contributes to performance, failure reduction, and attribution faithfulness**. Quantitative performance metrics and explainability measures for all ablations are reported in the subsequent supplementary tables.

A. Extended Results

TABLE V
Validation Performance Across All Ablation Models

Model	Val Acc ↑	Val Loss ↓	Train Acc↑	Failures ↓
M10	**0.9382**	**0.1962**	0.9915	**855**
M11	0.9180	0.2182	**0.9931**	1058
M6	0.9165	0.2035	0.9691	1077
M4	0.9138	0.2143	0.9829	1112
M7	0.9130	0.2164	0.9868	1123
M5	0.9107	0.2159	0.9678	1152
M9	0.9101	0.2145	0.9783	1160
M2	0.9100	0.2161	0.9780	1161
M8	0.5017	1.0597	0.5036	6429
M1	0.5017	2.4466	0.5015	6429
M3	0.5017	4.1731	0.4996	6429

B. Quantitative Faithfulness Metrics

To further assess the faithfulness of the proposed explainability method, we evaluate three complementary metrics on shared failure cases across all ablation models:

- **CAM Entropy** — measuring spatial concentration of attribution maps (lower is better).
- **Deletion AUC** — quantifying how rapidly model confidence decreases when the most relevant HRRP bins are removed (lower is better).
- **Insertion AUC** — measuring how quickly model confidence is recovered when the most relevant bins are progressively inserted into a blank signal (higher is better).

All metrics are computed per-sample and reported as mean values across shared failure instances.

C. Composite Faithfulness Score

To summarize performance across multiple faithfulness criteria, we report a composite score that aggregates CAM entropy, deletion AUC, and insertion AUC. Each metric is standardized across models using z-score normalization.

We emphasize that composite rankings are sensitive to (i) the per-sample distribution of scores, (ii) inclusion/exclusion of extreme models (outliers), and (iii) the metric alignment (lower vs higher better). Therefore we compute per-sample composites and report bootstrap CIs and paired tests to communicate uncertainty.

$$C = \frac{1}{3}\left(\hat{E} + \hat{D} + \hat{I}\right) \tag{1}$$

where the normalized terms are defined as:

$$\hat{E} = \frac{E - E_{\min}}{E_{\max} - E_{\min}} \quad \text{(Entropy: lower is better)} \tag{2}$$

$$\hat{D} = \frac{D - D_{\min}}{D_{\max} - D_{\min}} \quad \text{(Deletion: lower is better)} \tag{3}$$

$$\hat{I} = 1 - \frac{I - I_{\min}}{I_{\max} - I_{\min}} \quad \text{(Insertion: higher is better)} \tag{4}$$

where, C is the final Composite Score (lower values are better for overall faithfulness), $\hat{E}, \hat{D}, \hat{I}$ are min-max normalized scores for Entropy, Deletion, and Insertion, respectively, E, D, I are the raw values for a specific model, E_{min}, E_{max} are minimum and maximum values observed across all compared models for that specific metric, Insertion Inversion (1 - ...): Since higher Insertion AUC is better, the term is inverted so that 0 becomes the best score, consistent with Entropy and Deletion.

TABLE VI
QUANTITATIVE EVALUATION OF MODEL FAITHFULNESS AND EFFICIENCY. **ENTROPY** AND **DELETION** (LOWER IS BETTER) MEASURE FOCUS AND TRUTHFULNESS, WHILE **INSERTION** (HIGHER IS BETTER) MEASURES EFFICIENCY. THE **COMPOSITE SCORE** (LOWER IS BETTER) AGGREGATES THESE METRICS USING MIN-MAX NORMALIZATION.

Model	Entropy ↓	Deletion ↓	Insertion ↑	Composite ↓
M10	**4.137**	**0.373**	0.532	**0.247**
M11	4.170	0.419	0.422	0.344
M4	4.209	0.406	0.393	0.364
M2	4.224	0.626	0.602	0.373
M6	4.155	0.448	0.379	0.380
M7	4.244	0.453	0.406	0.393
M5	4.213	0.484	0.390	0.411
M9	4.202	0.552	0.417	0.430
M1	5.106	0.961	0.961	0.605
M3	5.261	0.968	**0.968**	0.648
M8	5.327	0.834	0.834	0.667

This composite score is provided for descriptive comparison only; the main analysis relies on individual metrics and qualitative visualizations. **Note:** We include the composite score only in supplementary material as a descriptive summary. The primary conclusions of the paper rely on qualitative CAM visualizations and individual faithfulness metrics, not on the composite ranking.

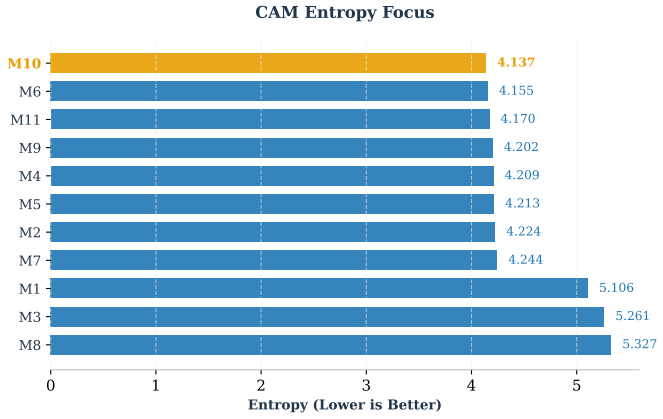

Fig. 4. Comparison of Layer-CAM entropy across all ablation models on shared high-confidence failure samples. Lower entropy indicates more concentrated and decisive attribution over HRRP bins. The proposed model (M10) exhibits consistently lower entropy than baseline and partially refined variants, while degenerate models show diffuse, near-uniform explanations.

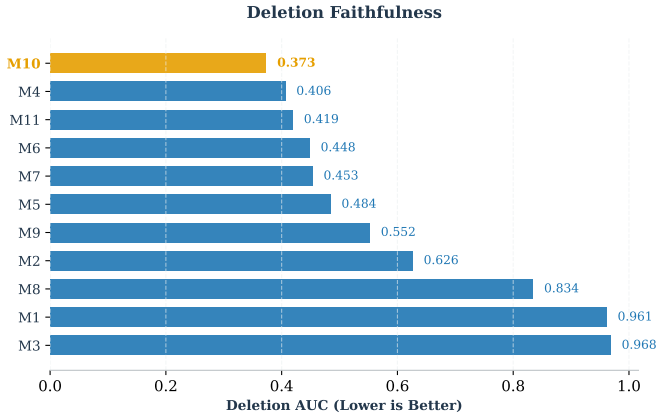

Fig. 5. Deletion-based faithfulness evaluation using Layer-CAM. The confidence of each model is measured as progressively more salient HRRP bins are removed. Lower area under the deletion curve indicates stronger causal reliance on highlighted regions. The proposed model (M10) demonstrates the most rapid confidence degradation, suggesting more faithful explanations.

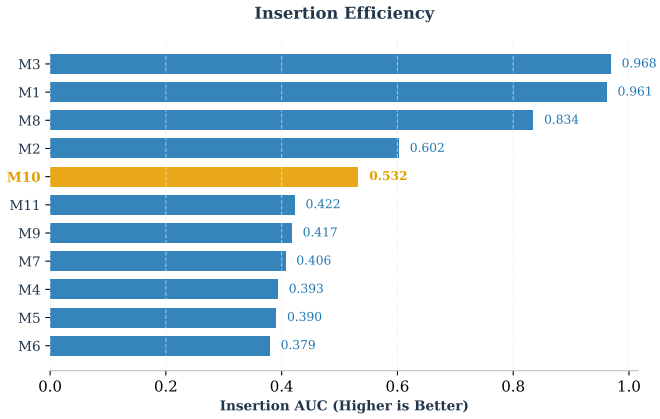

Fig. 6. Insertion-based evaluation measuring how rapidly model confidence is recovered as salient HRRP bins (identified by Layer-CAM) are progressively reintroduced. Higher area under the insertion curve indicates that the highlighted regions are sufficient to support the prediction. The refined architecture (M10) achieves a favorable trade-off between insertion efficiency and deletion faithfulness.

D. CAM inter-relation

Agreement is predominantly positive across samples, with no systematic anti-correlation observed. Variability reflects known sensitivity of gradient-weighting schemes to activation scaling in 1-D radar signals. Importantly, qualitative rescue patterns remain consistent across methods despite moderate correlation magnitude.