

Response to RC1

This preprint carries out controlled comparative experiments between pixel-wise MLP and Swin U-Net to disentangle how polarimetric observables, spatial context and short-term temporal radar evolution jointly shape non-precipitation echo discrimination. The research addresses a long-standing operational demand for radar data quality control and historical single-polarization archive reprocessing, with well-designed ablation groups, multi-type precipitation case demonstrations and dual test sets featuring clean and flawed labels to support quantitative evaluation via CSI, POD and FAR. The core finding that spatial feature learning serves as the primary driver of classification accuracy is physically consistent, and the discovery that reflectivity-only Swin U-Net surpasses dual-polarization MLP delivers valuable practical implications.

Even so, several substantive scientific shortcomings limit the manuscript's depth and generalizability. The entire dataset originates from a single Guangzhou S-band radar without supporting cross-regional or cross-terrain validation; temporal modeling is confined to two consecutive radar volumes, leaving the potential of longer scan sequences unexplored; spatial attention mechanisms of Swin Transformer are not visualized or interpreted, making the model's multi-scale feature extraction a black box; computational overhead critical to real-time operational workflows is not quantified; the adverse effects of label noise on evaluation metrics are only qualitatively illustrated rather than quantitatively decomposed; performance metrics lump all non-precipitation echo subtypes together without separate statistics for clutter, anomalous propagation and biological echoes; no classic U-Net segmentation baseline is included to confirm performance increments brought by windowed self-attention. Major revisions are required to resolve these deficiencies prior to formal consideration.

General response: Thank you for your careful assessment of our manuscript and for recognizing its relevance to radar quality control and the reprocessing of historical single-polarization observations. We appreciate your positive comments on the comparative experimental design and the practical value of spatial-context learning. We also recognize the need for additional evidence and clearer discussion to better establish the scope and interpretation of our findings. We provide detailed point-by-point responses below, explaining the additional analyses, the corresponding manuscript revisions, and the limitations that remain.

Major Comments:

1) All experimental data are sourced merely from a single S-band radar in Guangzhou, such a narrow observational basis risks rendering the study's core conclusions case biased and one-sided. It is recommended to partition the research into different seasons to calculate stratified verification metrics, or thoroughly discuss how seasonal disparities may alter the accuracy gap between the two model architectures if supplementary radar data cannot be accessed.

Response: Thank you for this important suggestion. To assess whether the model comparison is sensitive to seasonal variability, we further stratified the Guangzhou Test Dataset A into spring, summer, autumn, and winter subsets and recalculated the critical success index (CSI) for each model configuration. The seasonal results for the MLP, Swin U-Net, and standard U-Net models are summarized in Table 1 and Fig. 1.

The seasonal evaluation shows that the absolute CSI varies among seasons, indicating that seasonal changes in precipitation characteristics and non-precipitation echo conditions affect the

classification difficulty. Nevertheless, the main conclusions remain consistent across all four seasons. Both encoder–decoder models, namely Swin U-Net and U-Net, consistently outperform their MLP counterparts, with the improvement being particularly pronounced for the single-polarization configurations. In addition, the standard U-Net achieves comparable or slightly higher CSI values than the Swin U-Net in most seasonal subsets. For example, the dual-polarization U-Net configurations reach CSI values of 0.954–0.993 across seasons, while the corresponding Swin U-Net configurations obtain 0.935–0.989. For the single-polarization configurations, U-Net also remains slightly higher overall than Swin U-Net.

In general, most configurations obtain their highest CSI in autumn, while the lower scores in summer indicate increased classification difficulty under more complex precipitation and NPE conditions. The MLP models show larger seasonal variability, especially under the single-polarization setting, whereas the U-Net-based models remain more stable across seasons. These additional results demonstrate that the advantage of spatial-context modeling is not attributable to a particular season. These results also support the use of the U-Net as the primary encoder–decoder model in the revised manuscript. We have added this discussion to Section 3.3 of the revised manuscript.

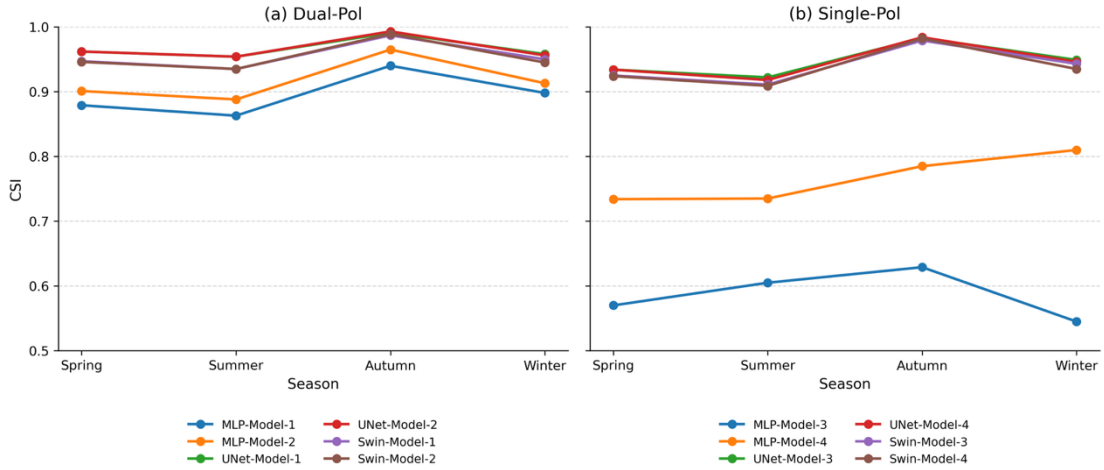


Figure 1: Seasonal CSI scores of the MLP, Swin U-Net and U-Net models evaluated using the Guangzhou S-band radar Test Dataset A for (a) dual-polarization and (b) single-polarization configurations.

Table 1: Seasonal critical success index (CSI) scores of the MLP, Swin U-Net and U-Net models evaluated using the Guangzhou S-band radar Test Dataset A.

Model	Spring	Summer	Autumn	Winter
MLP-Model-1	0.879	0.863	0.940	0.898
MLP-Model-2	0.901	0.888	0.965	0.913
MLP-Model-3	0.570	0.605	0.629	0.545
MLP-Model-4	0.734	0.735	0.785	0.810
Swin-Model-1	0.947	0.935	0.987	0.950

Swin-Model-2	0.946	0.935	0.989	0.945
Swin-Model-3	0.925	0.911	0.979	0.943
Swin-Model-4	0.924	0.909	0.982	0.935
UNet-Model-1	0.962	0.954	0.991	0.958
UNet-Model-2	0.962	0.954	0.993	0.956
UNet-Model-3	0.934	0.922	0.983	0.949
UNet-Model-4	0.934	0.918	0.984	0.946

We have added independent case studies from three geographically distinct locations, namely Nanjing, Huai'an, and Chengdu, to further examine the robustness of the models beyond the Guangzhou radar observations. The comparisons between the Swin U-Net and MLP models are shown in Figs. 2–4, and the corresponding results from the U-Net are provided in Figs. 15–17.

Across these independent cases, the encoder–decoder models generally show better separation of precipitation and non-precipitation echoes than the corresponding pointwise MLP models. In the Nanjing case, both architectures retain the main precipitation system, whereas the MLP predictions contain more residual non-precipitation echoes. In the Huai'an case, where precipitation coexists with localized non-precipitation contamination and linear interference-like echoes, the Swin U-Net and U-Net models suppress the contaminating echoes more effectively than the MLP while preserving the spatial structure of precipitation. The additional U-Net results in Figs. 15–17 further show that the standard U-Net produces comparable or cleaner precipitation-only fields than the Swin U-Net, particularly for localized and near-range NPE contamination. We have added this discussion to Section 3.3 of the revised manuscript.

The Chengdu case provides a more challenging scene dominated by non-precipitation echoes. In this case, the dual-polarization encoder–decoder models provide the most effective NPE removal, whereas the single-polarization configurations retain more residual false echoes because only Z_H is available. The standard U-Net still shows relatively stable suppression of these NPE-dominated structures compared with the MLP, further confirming the importance of spatial-context modeling. These cross-site experiments, together with the seasonal evaluation over Guangzhou, indicate that the advantage of spatial-context models is not restricted to a particular season or an individual Guangzhou case. The additional comparison also shows that the standard U-Net performs comparably to, and in several cases more cleanly than, the Swin U-Net. Nevertheless, we acknowledge that these additional experiments are based on representative cases and do not constitute a comprehensive multi-site climatological evaluation; this limitation has been explicitly noted in the revised manuscript.

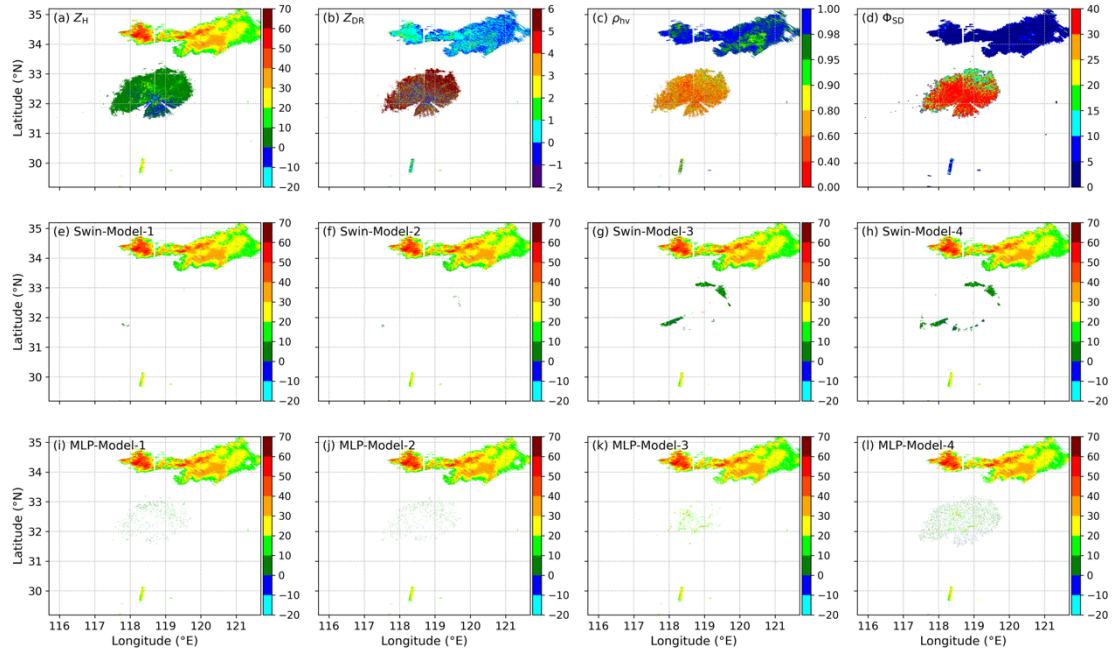


Figure 2: Comparison of precipitation-echo identification results from the Swin U-Net and MLP models for a dual-polarization radar case over Nanjing at 23:40 UTC on 15 July 2025. (a–d) Observed Z_H (dBZ), Z_{DR} (dB), ρ_{hv} , and $\sigma_{\phi_{DP}}$ (°), respectively. (e–h) Precipitation-only Z_H fields predicted by Swin-Model-1 to Swin-Model-4, respectively. (i–l) Same as (e–h), but for MLP-Model-1 to MLP-Model-4.

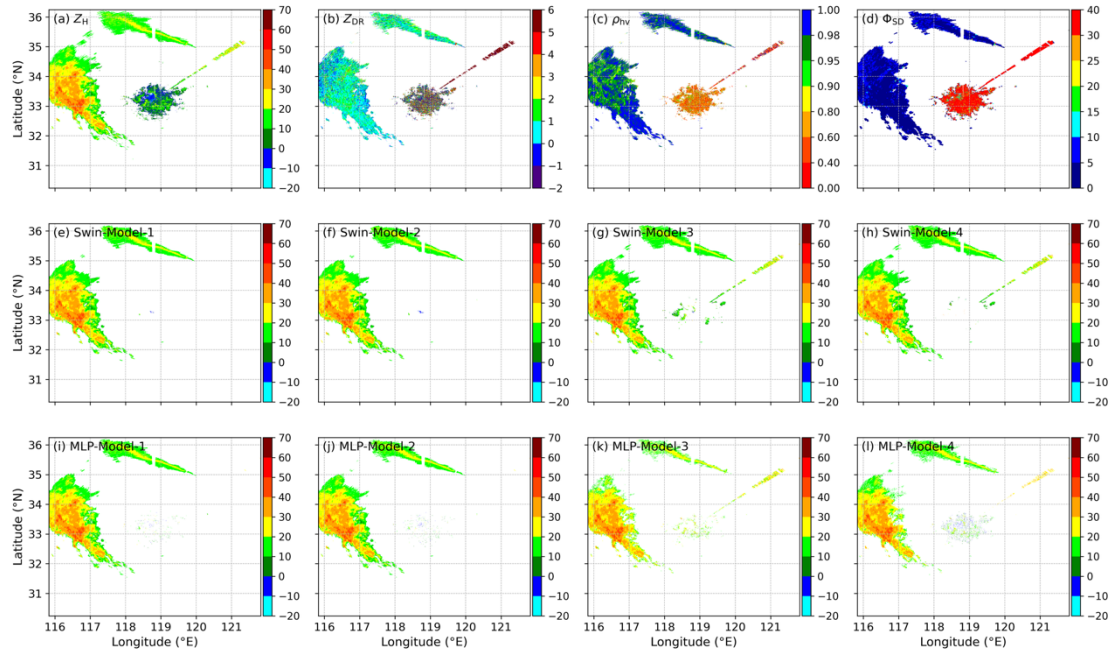


Figure 3: Same as Figure 2, but for a radar case over Huai'an at 11:07 UTC on 25 June 2025.

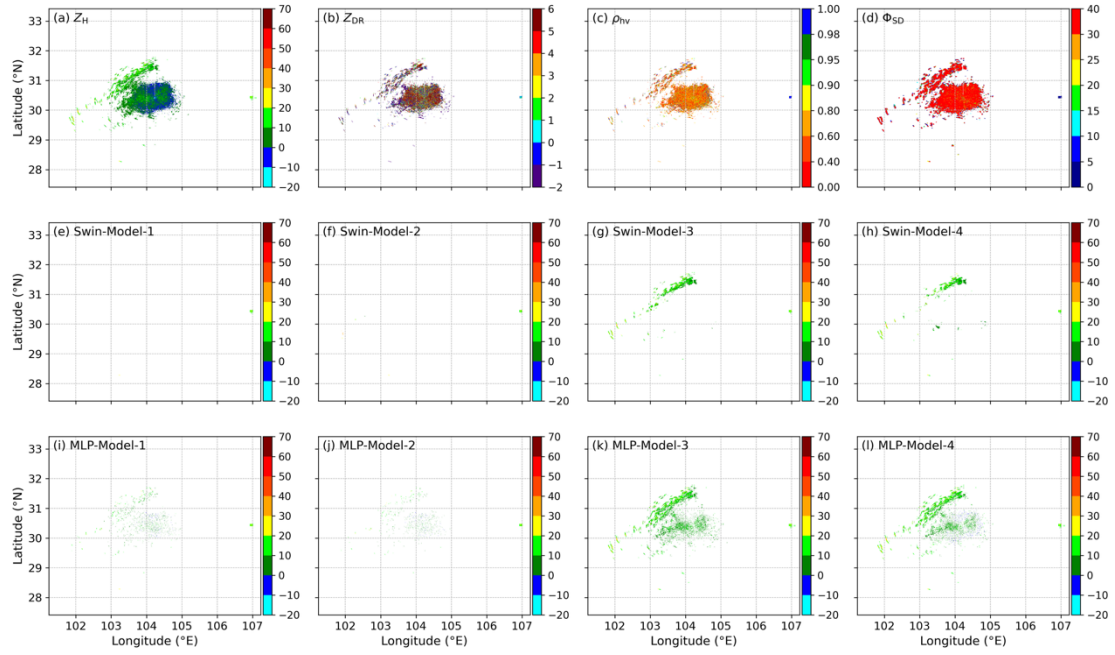


Figure 4: Same as Figure 2, but for a radar case over Chengdu at 02:52 UTC on 28 June 2025.

2) Only two sequential radar scans are fed into the model to capture temporal variation. It is suggested to extending input sequences to three or four consecutive volumes for additional ablation tests which would help quantify marginal accuracy gains from prolonged temporal context, while splitting metrics by clutter echo categories can further reveal the unique value of temporal information for different artifact types.

Response: Thank you for this valuable suggestion. Following the reviewer’s recommendation, we extended the input sequence from the original one and two consecutive radar volumes to three and four volumes and performed additional ablation experiments for both dual- and single-polarization configurations. The newly added experimental results, including the CSI, POD, and FAR scores, are presented in Table 2, while a summary of the experiments using different numbers of input radar volumes is shown in Figure 5.

The results show that the benefit of increasing the temporal context is architecture-dependent. For the MLP models, extending the sequence from two to three and four volumes increases the CSI from 0.913 to 0.922 and 0.928 for the dual-polarization configuration, and from 0.749 to 0.795 and 0.819 for the single-polarization configuration, respectively. The improvement is therefore more pronounced when only single-polarization observations are available.

In contrast, the Swin U-Net models show only marginal changes with increasing sequence length. The dual-polarization CSI reaches 0.953 (0.953 for three volumes and 0.952 for four volumes), while the single-polarization CSI varies only slightly between 0.927 and 0.930. The POD and FAR values likewise show no systematic improvement with further extension of the input sequence. These results indicate that additional temporal information substantially benefits the MLP architecture, whereas the Swin U-Net obtains only limited additional gains from longer sequences. Longer input sequences also increase the input data volume and introduce additional requirements for data availability, temporal alignment, and quality consistency. We have added these results and the corresponding discussion to the revised manuscript.

The corresponding experiments with the conventional U-Net, adopted as the primary spatial-context

model in the revised manuscript, are presented in Table 8 of this response. For U-Net, extending the input sequence beyond two volumes also produces only small changes in CSI under the tested configurations.

Table 2: CSI, POD, and FAR of single- and dual-polarization MLP and Swin U-Net models using three or four consecutive radar volumes as input.

Model	Polarization	Volume	CSI	POD	FAR
MLP-Model-5	Dual	3	0.922	0.967	0.056
MLP-Model-6	Dual	4	0.928	0.971	0.054
MLP-Model-7	Single	3	0.795	0.902	0.119
MLP-Model-8	Single	4	0.819	0.935	0.126
Swin-Model-5	Dual	3	0.953	0.976	0.033
Swin-Model-6	Dual	4	0.952	0.966	0.025
Swin-Model-7	Single	3	0.927	0.956	0.048
Swin-Model-8	Single	4	0.930	0.951	0.039

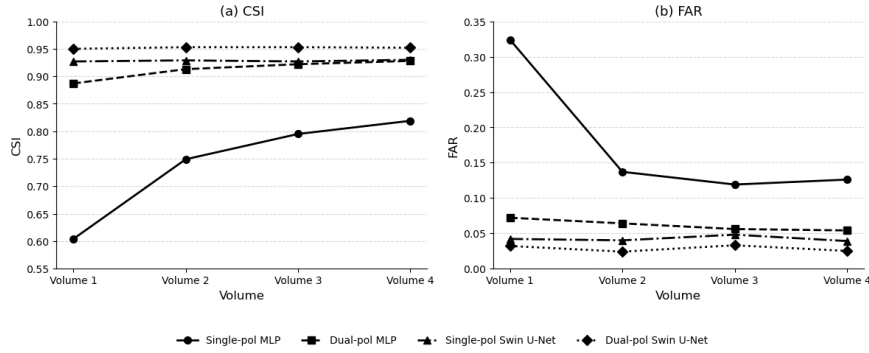


Figure 5: Ablation study of the number of consecutive radar volumes used as model input. Variations in (a) critical success index (CSI) and (b) false alarm ratio (FAR) are shown for the single-polarization MLP, dual-polarization MLP, single-polarization Swin U-Net, and dual-polarization Swin U-Net using one to four consecutive radar volumes.

We agree that category-specific evaluation could provide additional insight into the role of temporal information for different types of non-precipitation echoes. However, a reliable quantitative separation of all non-precipitation echo categories is difficult for the present dataset.

Strong and persistent ground-clutter echoes, as well as some obvious anomalous-propagation contamination, had already been largely suppressed during the radar signal-processing and quality-control procedures. The remaining non-precipitation echoes therefore mainly consist of relatively weak residual clutter and biological echoes. In particular, residual ground clutter and anomalous-propagation echoes can exhibit highly overlapping radar signatures and spatial structures, making them difficult to distinguish objectively using the available radar variables. More importantly,

independent ground truth is unavailable for these categories; therefore, manually assigning category-specific labels would introduce substantial uncertainty, and the correctness of such labels could not be independently verified. For this reason, we did not calculate separate quantitative verification scores for residual ground clutter and anomalous propagation, as doing so could lead to potentially unreliable conclusions based on uncertain category labels.

To partially address the reviewer’s suggestion, we instead constructed a conservative biological-echo subset using characteristic dual-polarization signatures. Candidate biological echoes were identified using the following criteria:

$$-5 < Z_H < 30 \text{ dBZ}, Z_{DR} > 5 \text{ dB}, \rho_{hv} < 0.85, SNR > 5 \text{ dB}.$$

These thresholds were used to isolate echoes exhibiting relatively weak reflectivity, low copolar correlation, and enhanced differential reflectivity, which are characteristic of biological scatterers. A representative example has been added as Figure 6, showing the original dual-polarization observations, the candidate biological-echo subset, and the precipitation-only fields predicted by the MLP- and Swin U-Net models. This additional analysis provides a more physically interpretable assessment of model behavior for a specific non-precipitation echo type while avoiding potentially unreliable subdivision of poorly distinguishable clutter categories.

As illustrated in Fig. 6, the candidate biological echoes are characterized by relatively weak Z_H , enhanced Z_{DR} , and reduced ρ_{hv} . The comparison further shows that the Swin U-Net models suppress these spatially coherent biological echoes more effectively than the corresponding MLP models while retaining the major precipitation structures.

Compared with Fig. 11, the single-polarization MLP models using three and four consecutive radar volumes show fewer biological echoes misclassified as precipitation than the two-volume model, indicating that longer temporal context provides useful information for distinguishing biological echoes from precipitation. In contrast, the single-volume MLP model relies solely on Z_H and lacks temporal and polarimetric information. Although it removes most non-precipitation echoes, it also erroneously suppresses a substantial fraction of precipitation echoes, indicating a poorer balance between non-precipitation echo removal and precipitation preservation.

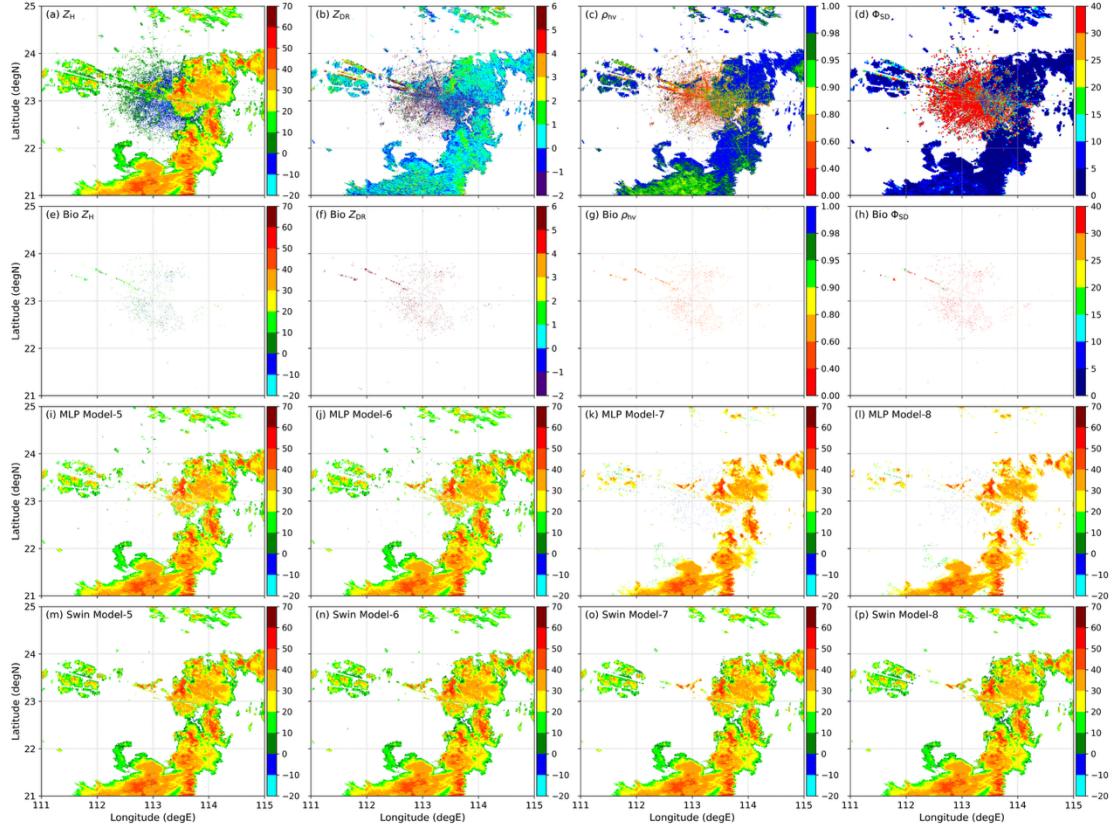


Figure 6: Dual-polarization radar observations, candidate biological echoes, and model-predicted precipitation echoes for a representative case over Guangzhou at 00:36 UTC on 28 May 2019. Panels (a–d) show the observed horizontal reflectivity Z_H (dBZ), differential reflectivity Z_{DR} (dB), copolar correlation coefficient ρ_{hv} , and differential-phase texture σ_{FDP} (°), respectively. Panels (e–h) show the corresponding biological-echo fields for Z_H , Z_{DR} , ρ_{hv} , and σ_{FDP} , respectively. Panels (i–l) show the precipitation-only predicted by MLP-Model-5 to MLP-Model-8, respectively, and panels (m–p) show the corresponding results predicted by Swin-Model-5 to Swin-Model-8.

3) Though the manuscript attributes superior classification performance of Swin U-Net to multi-scale spatial self-attention, no attention map visualization is provided to verify what spatial textures the model prioritizes during decision-making. Generating attention heatmaps for representative precipitation and non-precipitation cases and contrasting learned spatial features with conventional handcrafted texture indicators will enhance the interpretability of the Transformer-based framework.

Response: Thank you for this valuable suggestion. Following the reviewer’s recommendation, we performed an interpretability analysis to examine the spatial regions contributing to NPE identification and compared the model attribution with conventional radar texture indicators.

We used class-specific Grad-CAM as a gradient-based spatial attribution method. We note that Grad-CAM does not directly visualize the internal self-attention weights of the Swin Transformer; rather, it identifies the spatial regions that contribute strongly to the predicted NPE class. We adopted this method as a consistent spatial diagnostic because the revised study additionally includes a conventional U-Net, allowing the spatial attribution of the two encoder–decoder architectures to be examined within the same framework.

We first applied the analysis to Swin-Model-2, which uses dual-polarization variables and two consecutive radar volumes. The Grad-CAM responses were derived from multi-scale encoder

features and compared with two conventional texture indicators: differential-phase texture, $\sigma_{\Phi_{DP}}$, and the local standard deviation of Z_H . For the two representative cases, the Grad-CAM responses were enhanced mainly over NPE-dominated regions and precipitation–NPE transition zones. The Spearman correlations between Grad-CAM and $\sigma_{\Phi_{DP}}$ were 0.719 and 0.509, respectively, whereas the corresponding correlations with the local standard deviation of Z_H were -0.261 and -0.150 . Because $\sigma_{\Phi_{DP}}$ is explicitly included as an input variable, this correspondence is interpreted as consistency between the model attribution and a physically motivated texture feature, rather than as evidence that the model independently learns this specific physical mechanism.

The Grad-CAM maps do not completely overlap with either handcrafted texture indicator, suggesting that the spatial attribution cannot be fully represented by a single conventional texture measure. However, Grad-CAM alone cannot uniquely separate the contributions from individual input variables or learned spatial features. We therefore interpret these results as a spatial attribution diagnostic rather than a decomposition of the contributions from individual input variables.

In response to another reviewer suggestion, we additionally introduced a conventional U-Net as a segmentation baseline. The comparison showed that the standard U-Net achieves comparable or slightly better performance than the Swin U-Net in most experiments. Therefore, the revised manuscript no longer attributes the performance improvement over the MLP specifically to Swin Transformer self-attention. Instead, the results indicate that explicit encoder–decoder spatial-context modeling is the primary contributor. Accordingly, the standard U-Net is adopted as the principal spatial-context model in the revised manuscript, while the Swin U-Net is retained as an additional comparison.

We also performed the same Grad-CAM analysis for the adopted U-Net. Its spatial attribution shows a similar tendency to emphasize NPE-dominated regions and precipitation–NPE transition zones. Because the standard U-Net is now adopted as the principal spatial-context model in the revised manuscript, the corresponding U-Net Grad-CAM analysis and its limitations are presented in Sect. 3.5. The Swin U-Net analysis described above was retained as a direct response to the reviewer’s original request concerning the Transformer-based model.

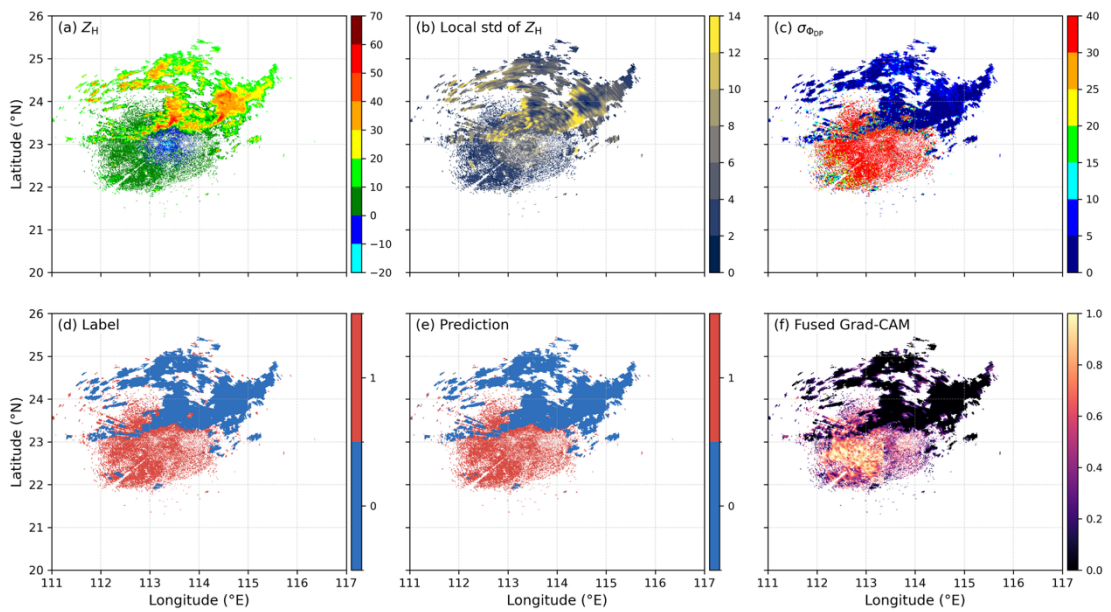


Figure 7: Comparison of Swin U-Net spatial responses with conventional radar texture indicators for a representative case over Guangzhou at 15:00 UTC on 5 March 2019. Panels show (a) Z_H (dBZ), (b) local standard deviation of Z_H , (c) $\sigma_{\Phi_{DP}}$ texture, (d) ground-truth label classification, (e) Swin-Model-2 prediction, and (f) fused Grad-CAM response.

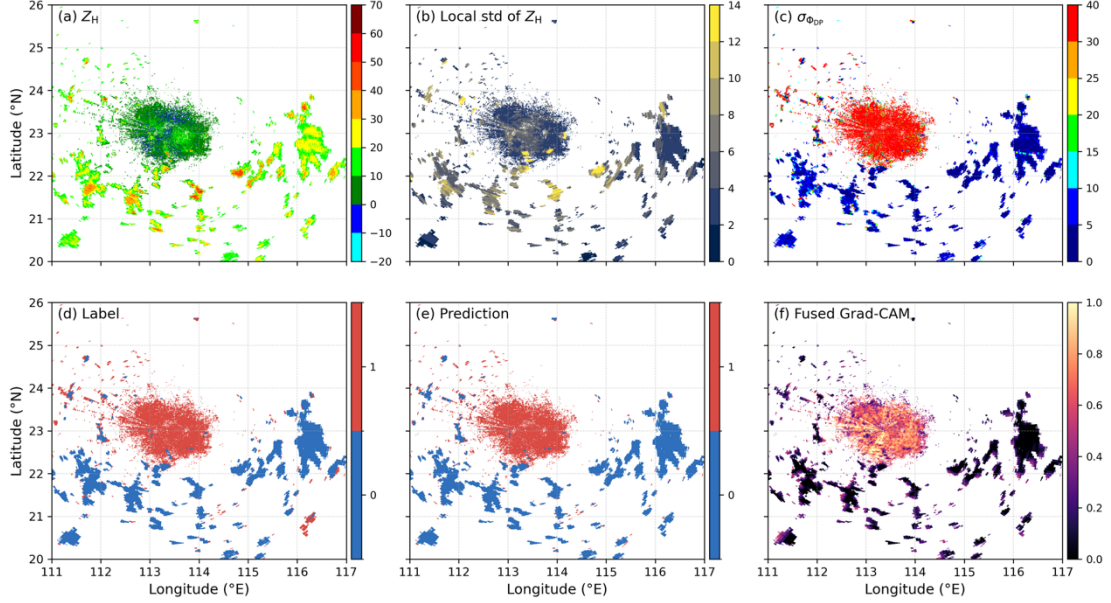


Figure 8: Same as Figure 7, but at 03:00 UTC on 29 July 2019.

4) No quantitative comparisons of inference speed, GPU memory consumption and total parameter volume across all eight model configurations are presented, yet computational efficiency is a key concern for operational radar platforms. A supplementary summary table documenting such hardware and latency metrics is suggested, alongside preliminary trials of lightweight distilled Swin variants to balance classification precision and real-time processing capability.

Response: Thank you for this important suggestion. We agree that computational efficiency is an essential consideration for the operational application of radar echo classification models. We have therefore conducted tests using an NVIDIA L40S and added a quantitative comparison of all twelve model configurations in Table 3, Table 4 and Table 9, including the total number of parameters, model size, peak GPU memory usage, inference latency, and throughput. We further clarified in Sect. 2.3 that the reported latency represents full-PPI inference over all six azimuthal sectors, using an inference batch size of 1, FP32 precision, 10 warm-up runs, and 30 repeated measurements with CUDA synchronization.

The MLP models contain only approximately 0.176–0.179 million parameters, with inference latencies of 29.55–33.92 ms PPI^{-1} and throughputs of 29.48–33.84 PPI s^{-1} . The standard U-Net models contain approximately 3.350–3.352 million parameters, with model sizes of 12.79–12.80 MB and peak GPU memory usage of 112.82–131.58 MB. Despite having more trainable parameters than the MLP models, the U-Net models achieve substantially lower inference latencies of 7.48–7.62 ms PPI^{-1} , corresponding to throughputs of 131.19–133.74 PPI s^{-1} . In comparison, the Swin U-Net models contain approximately 41.65–41.68 million parameters and require 46.86–48.18 ms PPI^{-1} , corresponding to throughputs of approximately 20.76–21.33 PPI s^{-1} . These results indicate that the standard U-Net provides a better balance between classification performance and

computational efficiency than the Swin U-Net under the hardware configuration used in this study. The full-PPI inference latency of the U-Net remains below approximately 8 ms PPI^{-1} , indicating its potential compatibility with near-real-time radar echo processing under the tested implementation and hardware configuration.

The revised manuscript reports the computational efficiency of the eight primary MLP and U-Net configurations in Tables 4 and 5. For completeness and to directly address the reviewer’s original comment concerning the Swin U-Net models, the corresponding Swin U-Net efficiency results are additionally provided in Table 4 of this response. The computational-efficiency analysis of the primary MLP and U-Net models has been added to Sect. 3.4 of the revised manuscript.

We also appreciate the reviewer’s suggestion regarding lightweight or distilled Swin variants. Given that the conventional U-Net already achieves comparable or better classification performance while requiring substantially fewer parameters and lower inference latency than the Swin U-Net, we did not further introduce a distilled Swin variant in the present revision. Developing and optimizing a dedicated teacher–student framework would constitute a separate model-compression study and is therefore left for future work.

Table 3: Comparison of parameter count, model size, peak GPU memory, inference latency and throughput for MLP models using single- or dual-polarization data and one or two consecutive radar volumes.

MLP-Model Efficiency	MLP-Model-1	MLP-Model-2	MLP-Model-3	MLP-Model-4
Total parameters (M)	0.177	0.179	0.176	0.176
Model size (MB)	0.683	0.691	0.677	0.679
Peak GPU memory (MB)	501.753	512.261	492.560	495.374
Latency (ms)	29.737 ± 0.226	29.550 ± 0.252	33.923 ± 0.099	30.495 ± 0.288
Throughput (PPI/s)	33.628	33.841	29.479	32.792

Table 4: As in Table 3, but for Swin U-Net model.

Swin-Model Efficiency	Swin-Model-1	Swin-Model-2	Swin-Model-3	Swin-Model-4
Total parameters (M)	41.666	41.684	41.652	41.657
Model size (MB)	160.467	160.537	160.414	160.432
Peak GPU memory (MB)	280.119	290.814	271.878	273.833
Latency (ms)	47.369 ± 0.579	48.176 ± 0.580	47.190 ± 1.139	46.863 ± 1.039
Throughput (PPI/s)	21.111	20.757	21.191	21.325

5) Test Dataset B contains mislabeled samples to assess model robustness, yet the current analysis cannot distinguish performance degradation induced by intrinsic model defects versus label inaccuracies. Injecting artificial label noise at varying proportions into the high-confidence Test Dataset A enables systematic quantification of each model’s tolerance to labeling errors.

Response: Thank you for this valuable suggestion. Following the reviewer’s suggestion, we conducted a controlled reference-label perturbation experiment using a randomly selected subset of 100 cases from Test Dataset A. Artificial perturbations were introduced only into the reference labels, while the radar observations and model predictions were kept unchanged. The experiment therefore evaluates how verification scores respond to uncertainty in the reference labels under fixed model predictions.

As the perturbation fraction increases, CSI generally decreases, demonstrating that verification scores are sensitive to reference-label uncertainty. However, Test Datasets A and B consist of different cases and may differ not only in label quality but also in echo characteristics, classification difficulty, and class distribution. Therefore, this experiment does not estimate the label-error rate in Test Dataset B and does not quantitatively separate the effects of label uncertainty from case difficulty or intrinsic model performance. We have revised Sect. 4.3 accordingly and interpret this experiment as a sensitivity analysis of the verification metrics.

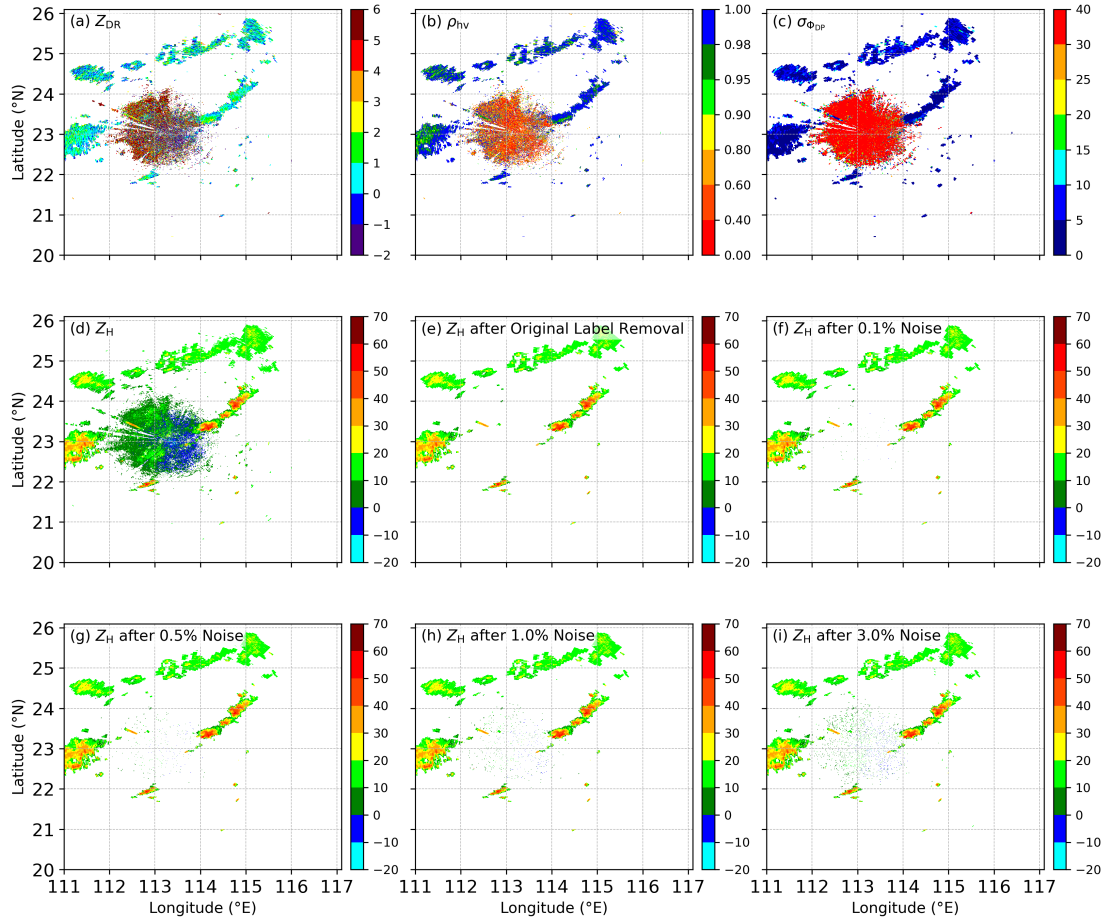


Figure 9: Illustration of the artificial label-noise experiment based on a high-confidence sample from Test Dataset A. Panels (a–d) show the original radar observations of (a) differential reflectivity Z_{DR} (dB), (b) copolar correlation coefficient ρ_{hv} , (c) differential-phase texture $\sigma_{\Phi_{DP}}$ (°), and (d) horizontal reflectivity Z_H (dBZ). Panel (e) presents the precipitation field derived from the original high-confidence reference labels, while panels (f–i) show the corresponding fields after introducing artificial label noise at proportions of 0.1%, 0.5%, 1.0%, and 3.0%, respectively. These controlled perturbations are used to evaluate the sensitivity of verification scores to reference-label uncertainty while holding the model predictions fixed.

Table 5: Critical success index (CSI) scores of different models evaluated on Test Dataset A with varying proportions of artificially injected label noise and on Test Dataset B using the Guangzhou S-band radar dataset.

Model	0.1%	0.5%	1%	3%	Test Dataset B
MLP-Model-1	0.870	0.866	0.861	0.840	0.866
MLP-Model-2	0.912	0.907	0.901	0.876	0.906
MLP-Model-3	0.553	0.551	0.549	0.540	0.629
MLP-Model-4	0.687	0.685	0.682	0.671	0.720
Swin-Model-1	0.967	0.961	0.955	0.927	0.952
Swin-Model-2	0.971	0.965	0.958	0.929	0.951
Swin-Model-3	0.948	0.943	0.937	0.912	0.939
Swin-Model-4	0.949	0.944	0.937	0.911	0.935

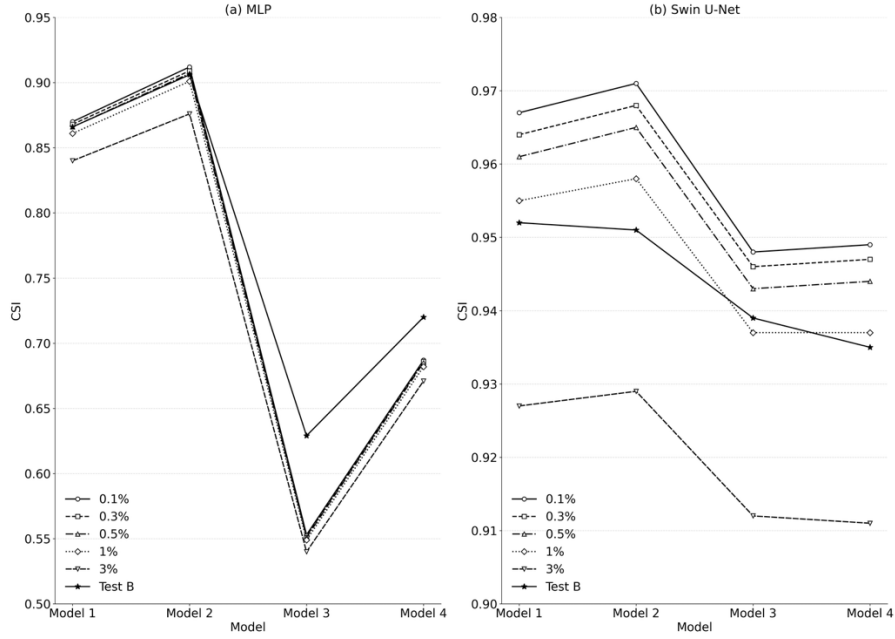


Figure 10: Sensitivity of CSI to reference-label perturbations for (a) MLP and (b) Swin U-Net, with model predictions held fixed. Different proportions of label noise (0.1%, 0.3%, 0.5 %, 1.0%, and 3.0%) were introduced into Test Dataset A, and the resulting performance was compared with that on Test Dataset B for Models 1–4.

6) All non-precipitation echoes are treated as a unified binary target without differentiated assessment of distinct artifact categories, which weakens the targeted guidance for operational quality control workflows. Stratifying CSI, POD and FAR values by ground clutter, anomalous propagation streaks and biological echoes will clarify which model-input combination optimally mitigates each interference source.

Response: As discussed above (Response 2), reliable category-specific labels for ground clutter and anomalous propagation are difficult to obtain because their residual signatures overlap and independent reference truth is unavailable.

To show the capability of our algorithm, we added a representative candidate biological-echo case selected using characteristic dual-polarization signatures. For the specific identification method, please refer to the description provided in Response 2. Figure 11 presents the one- and two-volume model results, complementing the three- and four-volume experiments discussed in Figure 6. The Swin Transformer models generally remove biological echoes more effectively while preserving precipitation structures. For the single-polarization MLP models, additional temporal context reduces biological-echo misclassification, whereas the one-volume model, which relies only on Z_H tends to remove both non-precipitation echoes and part of the precipitation signal.

This case provides qualitative evidence of the distinct effects of temporal context and model architecture on biological-echo suppression.

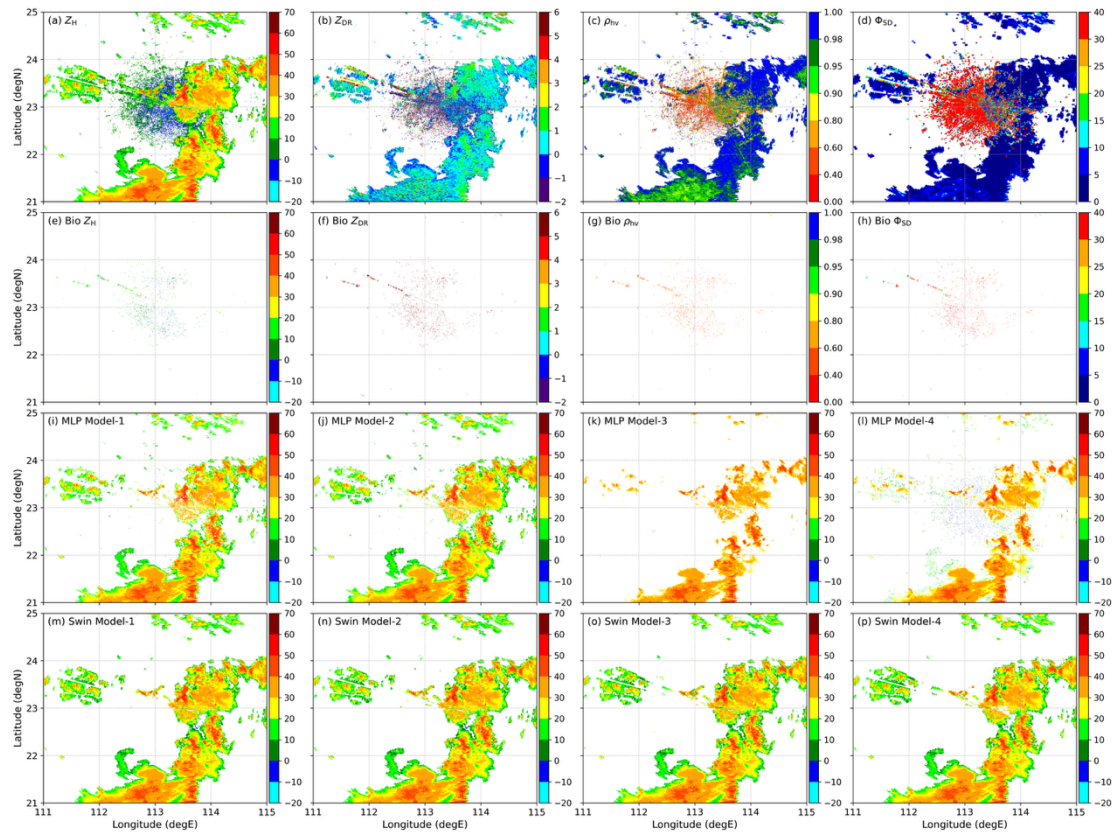


Figure 11: Similar to Figure 6, but for MLP-Model-1 to MLP-Model-4 (i–l) and Swin-Model-1 to Swin-Model-4 (m–p).

7) The paper only compares MLP and self-built Swin U-Net without benchmarking against widely adopted radar segmentation backbones. Incorporating a standard U-Net into ablation experiments can isolate performance improvements stemming from Swin Transformer windowed self-attention, rather than generic encoder-decoder segmentation structures.

Response: Thank you for this important suggestion. Following the reviewer’s recommendation, we introduced a standard U-Net as an additional benchmark and conducted the same experiments under identical input configurations and evaluation settings. The corresponding quantitative results and

representative cases have been added as Tables 6–10 and Figs. 12–17.

The comparison shows that the standard U-Net generally performs slightly better than the Swin U-Net. For example, U-Net achieves higher CSI values across the validation and test datasets, and this advantage remains consistent in the seasonal evaluation and in the experiments using three and four consecutive radar volumes. For the three- and four-volume experiments, for example, the U-Net achieves CSI values of 0.966–0.967 for dual-polarization input and 0.938–0.939 for single-polarization input, compared with 0.952–0.953 and 0.927–0.930, respectively, for the Swin U-Net. The independent case studies from different locations further suggest that the U-Net provides qualitative robustness, particularly for the suppression of near-range and localized non-precipitation echoes.

These results indicate that the performance improvement cannot be attributed specifically to the windowed self-attention mechanism of the Swin Transformer; the encoder–decoder segmentation framework itself plays a major role. More importantly, the standard U-Net provides comparable or better classification performance with a simpler and more widely adopted architecture. Therefore, in the revised manuscript, we replace the Swin U-Net with the standard U-Net as the primary encoder–decoder model and revise the corresponding analyses and conclusions accordingly.

Table 6: Quantitative performance of the UNet variants on the validation set, the high-confidence test subset (Test Dataset A) and Test Dataset B. We report the Critical Success Index (CSI), Probability of Detection (POD), and False Alarm Ratio (FAR).

Model	Validation Dataset			Test Dataset A			Test Dataset B		
	CSI	POD	FAR	CSI	POD	FAR	CSI	POD	FAR
UNet-Model-1	0.971	0.983	0.012	0.964	0.977	0.018	0.962	0.991	0.029
UNet-Model-2	0.974	0.986	0.012	0.967	0.980	0.020	0.960	0.988	0.029
UNet-Model-3	0.942	0.970	0.030	0.937	0.961	0.036	0.945	0.989	0.046
UNet-Model-4	0.949	0.970	0.022	0.936	0.955	0.033	0.941	0.986	0.046

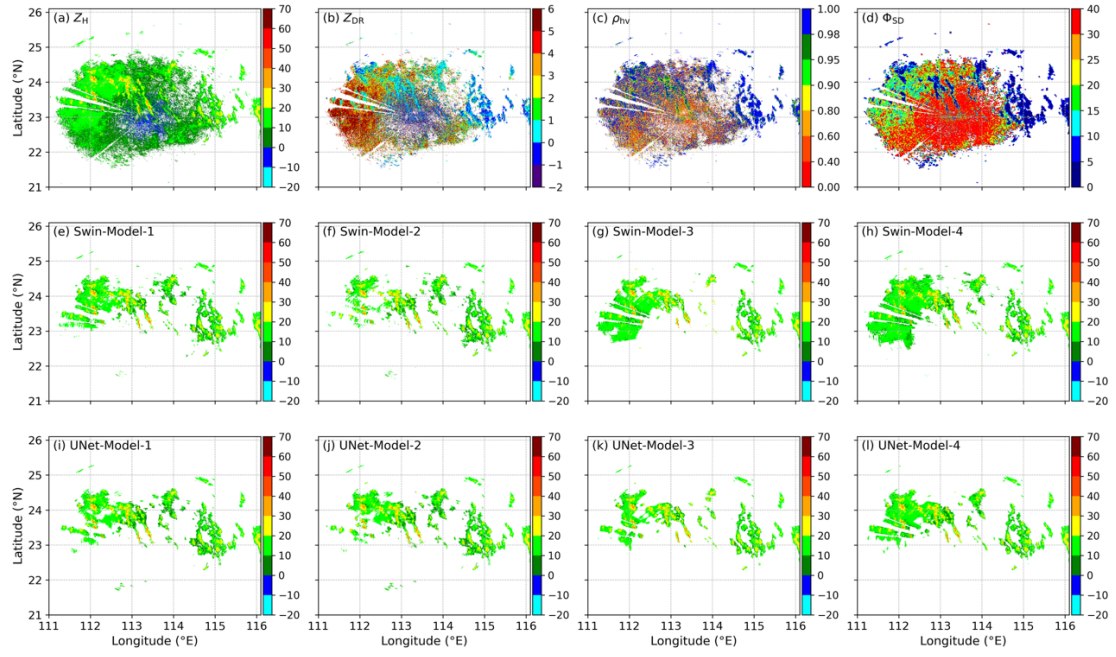


Figure 12: Dual-polarization radar observations and model-predicted precipitation echoes for a representative case over Guangzhou at 14:00 UTC on 5 May 2019. Panels (a–d) show the observed horizontal reflectivity (Z_H ; dBZ), differential reflectivity (Z_{DR} ; dB), copolar correlation coefficient (ρ_{hv}), and differential-phase texture (σ_{DP} ; °), respectively. Panels (e–h) show the precipitation-only predicted by Swin-Model-1 to Swin-Model-4, respectively, and panels (i–l) show the corresponding results predicted by UNet-Model-1 to UNet-Model-4.

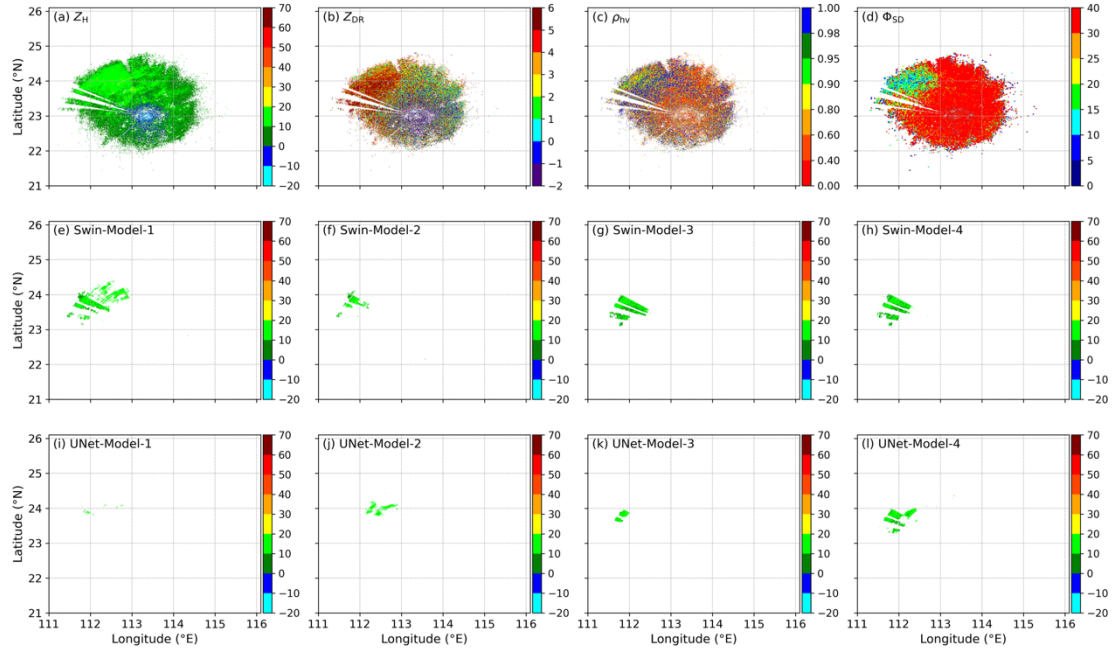


Figure 13: Similar to Figure 12, but at 21:00 UTC on 9 May 2019.

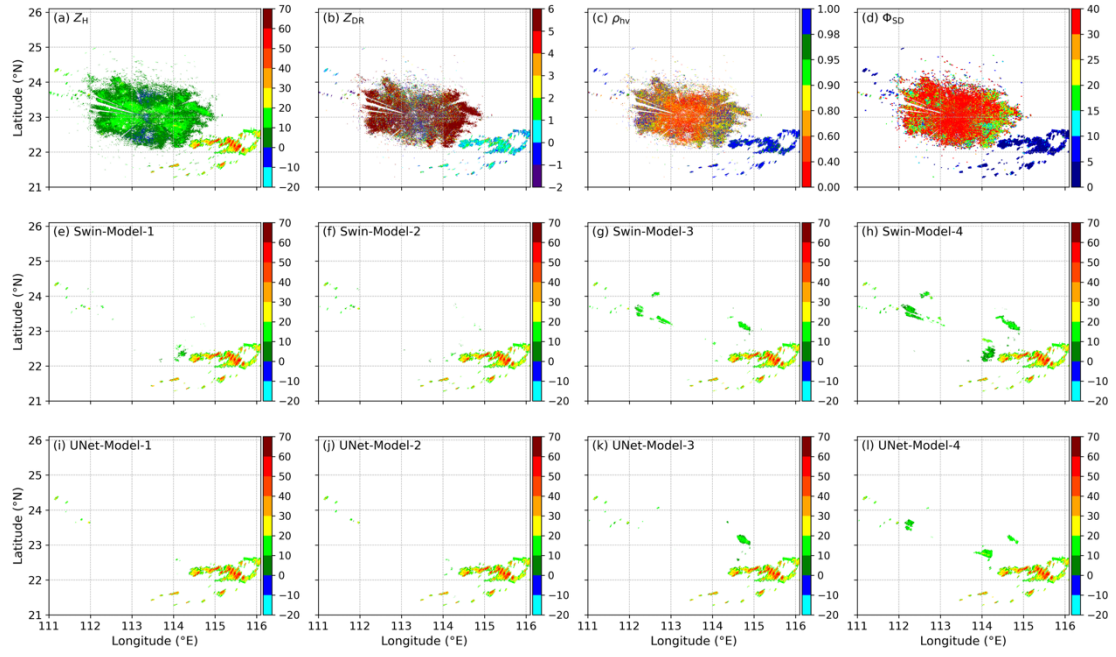


Figure 14: Similar to Figure 12, but at 23:00 UTC on 15 Aug 2019.

Table 7: As in Table 1, but for UNet model.

Model	Spring	Summer	Autumn	Winter
UNet-Model-1	0.962	0.954	0.991	0.958
UNet-Model-2	0.962	0.954	0.993	0.956
UNet-Model-3	0.934	0.922	0.983	0.949
UNet-Model-4	0.934	0.918	0.984	0.946

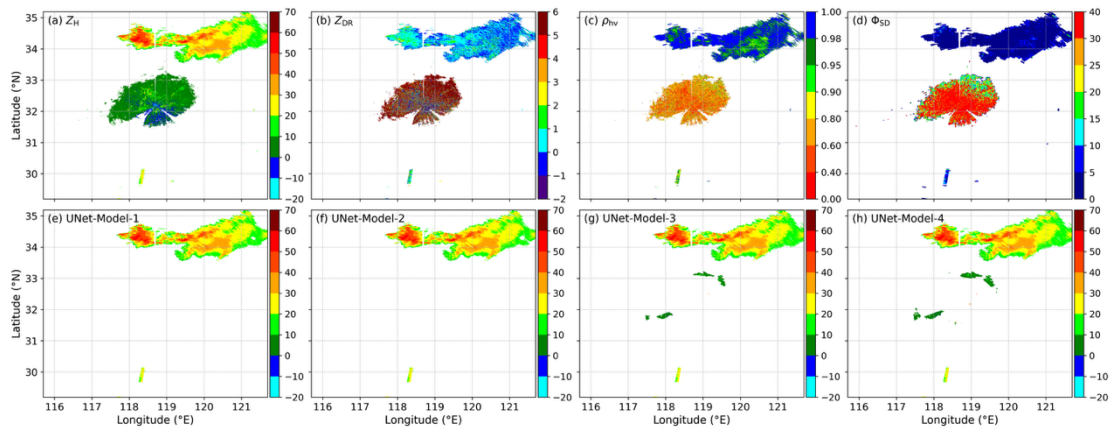


Figure 15: Same as Figure 2, but (e–h) predicted by U-Net-Model-1 to U-Net-Model-4, respectively.

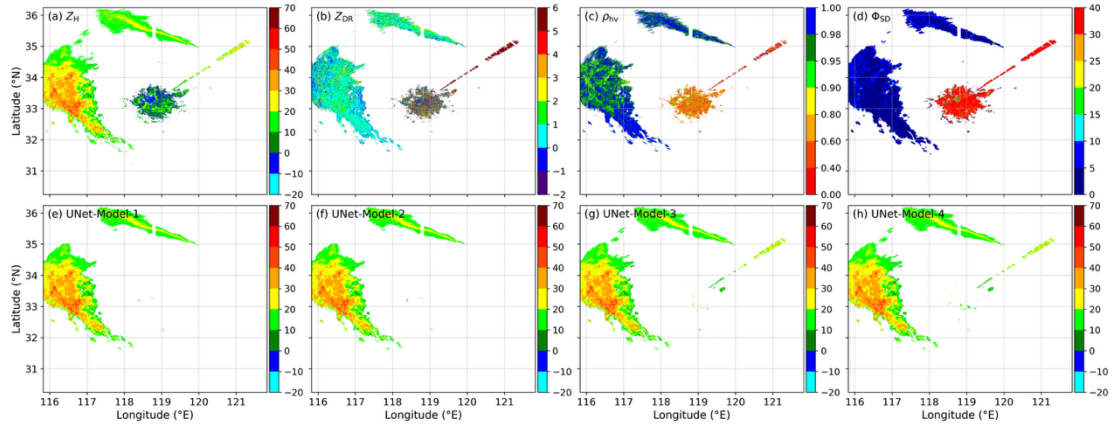


Figure 16: Same as Figure 3, but (e–h) predicted by U-Net-Model-1 to U-Net-Model-4, respectively.

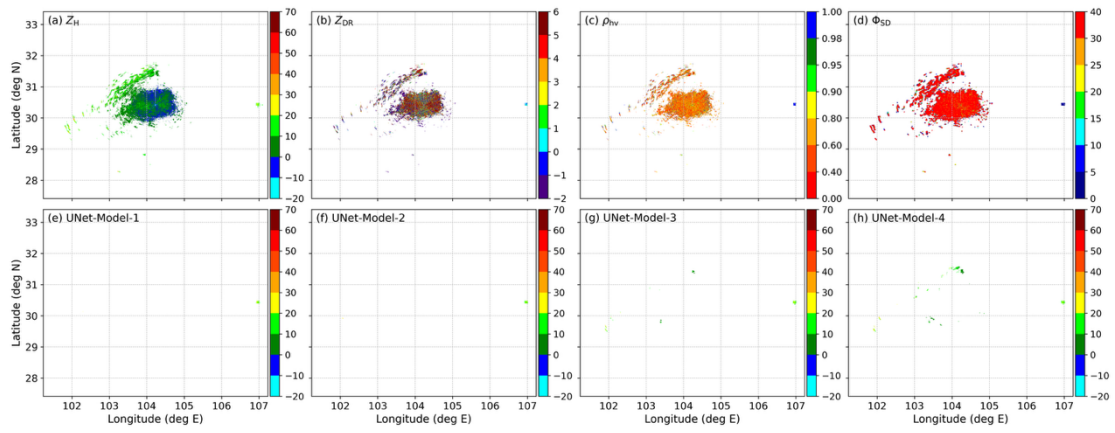


Figure 17: Same as Figure 4, but (e–h) predicted by U-Net-Model-1 to U-Net-Model-4, respectively.

Table 8: CSI, POD, and FAR of single- and dual-polarization U-Net models using three or four consecutive radar volumes as input.

Model	Polarization	Volume	CSI	POD	FAR
U-Net-Model-5	Dual	3	0.966	0.979	0.018
U-Net-Model-6	Dual	4	0.967	0.983	0.021
U-Net-Model-7	Single	3	0.938	0.960	0.037
U-Net-Model-8	Single	4	0.939	0.960	0.036

Table 9: As in Table 3, but for the U-Net models.

UNet-Model Efficiency	UNet-Model-1	UNet-Model-2	UNet-Model-3	UNet-Model-4
Trainable parameters (M)	3.351	3.352	3.350	3.350
Model size (MB)	12.797	12.801	12.793	12.795
Peak GPU memory (MB)	119.323	131.578	112.820	114.696

Latency (ms)	7.504 ± 0.033	7.622 ± 0.101	7.518 ± 0.210	7.477 ± 0.105
Throughput (PPI/s)	133.267	131.193	133.006	133.735

Table 10: As in Table 5, but for U-Net models.

Model	0.1%	0.5%	1%	3%
UNet-Model-1	0.978	0.972	0.965	0.937
UNet-Model-2	0.980	0.974	0.966	0.937
UNet-Model-3	0.956	0.951	0.944	0.919
UNet-Model-4	0.958	0.952	0.945	0.919

Minor comments:

1) Rectify distorted axis labels and misaligned longitude tick marks in all case figures, and supplement explicit feature channel dimension annotations for the unlabeled intermediate layers in Figure 1(a).

Response: Thank you for these careful suggestions. We have revised the manuscript accordingly. Specifically, (1) the axis labels and longitude tick marks in all case-study figures have been corrected, and the feature-channel dimensions of the intermediate layers in Fig. 1(a) have been explicitly annotated.

2) Fix textual typographical errors such as “filed” to “field”, trim repetitive boilerplate language across figure captions, and briefly define the drop-path regularization term for readers with limited deep learning background.

Response: Typographical errors have been corrected, repetitive wording in the figure captions has been reduced. In the revised manuscript, the original Swin U-Net has been replaced by a convolutional U-Net. As a result, the drop-path regularization used in the previous Swin Transformer architecture is no longer part of the current model and the corresponding description has been removed.

3) Round all CSI, POD and FAR values in Tables 2 and 3 to three decimal places for consistent numerical precision.

Response: All CSI, POD, and FAR values in Tables 2 and 3 have been rounded to three decimal places for consistency.