# WHEN LESS IS MORE: SIMPLICITY BEATS COMPLEXITY FOR PHYSICS-CONSTRAINED INSAR PHASE UNWRAPPING

**Prabhjot Singh**
The University of Texas at Austin, USA
RediMinds Inc., USA
prabhjot.singh@utexas.edu

**Manmeet Singh**
The University of Texas at Austin, USA
Western Kentucky University, USA
manmeet.singh@utexas.edu

## ABSTRACT

Operational phase unwrapping is the primary computational bottleneck in InSAR-based volcanic and seismic monitoring. We challenge the industry trend of adopting high-complexity computer vision architectures, such as attention mechanisms, without validating their suitability for physics-constrained geophysical regression. We present the first large-scale architectural ablation study on a global LiCSAR benchmark (20 frames, 39,724 patches, 651M pixels). Our results reveal a significant "complexity penalty": a vanilla U-Net (7.76M parameters) achieves $R^2 = 0.834$ and $RMSE = 1.01$ cm, outperforming 11.37M-parameter attention-based models by 34% in $R^2$ and 51% in $RMSE$. Power Spectral Density (PSD) analysis provides the physical justification: while attention excels at capturing sharp semantic edges in natural images, it injects unphysical high-frequency artifacts ($> 0.3$ cycles/pixel) into geophysical fields, violating the fundamental smoothness constraints of elastic surface deformation. With a 2.92ms inference latency (a 2.5$\times$ speedup), the vanilla U-Net is the only candidate to comfortably meet the sub-100ms requirement for operational early-warning systems. This work bridges the "publication-to-practice" gap by proving that convolutional locality outperforms modern complexity for smooth-field regression, advocating for physics-informed simplicity in ML4RS. Code available at `github/prabhjotschugh`.

## 1 INTRODUCTION

Interferometric Synthetic Aperture Radar (InSAR) enables millimeter-precision surface deformation monitoring at continental scales, yet phase is measured modulo $2\pi$ and must be *unwrapped* to recover true displacement, the primary computational bottleneck in volcanic and seismic monitoring. While deep learning offers significant acceleration over traditional solvers like SNAPHU (Chen & Zebker, 2001), a concerning trend has emerged: the uncritical adoption of high-complexity architectures, such as attention mechanisms (Vaswani et al., 2023) and multi-scale aggregation (Chen et al., 2018), directly from computer vision benchmarks. However, a fundamental domain mismatch exists. Unlike natural images characterized by discrete semantic boundaries (Dosovitskiy et al., 2021), geophysical displacement is governed by elasticity and spatial autocorrelation, favoring continuous, smooth-field representations (Reichstein et al., 2019).

We investigate a critical question: *Do ImageNet-derived inductive biases transfer to InSAR, or do domain-specific constraints favor architectural simplicity?* Through a rigorous ablation study on a global LiCSAR benchmark, we reveal a "complexity penalty" where simpler models better align with geophysical priors. Our contributions are as follows:

- **Global Operational Benchmark:** We curate a benchmark of 39,724 patches (651M pixels) across six continents, employing frame-level splitting to strictly evaluate geographic generalization and prevent the spatial leakage common in existing literature.
- **Quantifying the Complexity Penalty:** We demonstrate empirically that a vanilla U-Net (7.76M params) achieves $R^2 = 0.834$, outperforming 47% larger attention-based models by 34% in $R^2$ and 51% in $RMSE$.

- **Physics-Grounded Diagnostics:** Using Power Spectral Density (PSD) analysis, we show that complex models inject unphysical high-frequency artifacts ($> 0.3$ cycles/pixel) that violate the elasticity-driven smoothness of surface deformation.
- **Operational Deployment:** We achieve a 2.92ms inference latency (a 2.5$\times$ speedup), meeting sub-100ms requirements for real-time volcanic and seismic early-warning systems.

## 2 Related Work & Task Formulation

**InSAR Phase Unwrapping.** Traditional solvers like SNAPHU (Chen & Zebker, 2001) incur $O(N^2)$ complexity and error propagation in low-coherence regions. Deep learning (DL) has mitigated these bottlenecks, evolving from the vanilla U-Net of PhaseNet (Spoorthi et al., 2019) toward high-complexity architectures like ResDANet (dual-attention) and Unwrap-Net (ASPP) (Zhou et al., 2021). However, while attention-based designs excel at capturing discontinuous semantic boundaries in natural images, geophysical displacement is governed by elasticity and spatial autocorrelation (Tobler's First Law (Tobler, 1970)). We hypothesize that high-frequency computer vision (CV) priors are mismatched for smooth-field regression and may introduce unphysical artifacts.

**Operational Task Formulation.** We define unwrapping as a physics-constrained regression. The input is a 6-channel tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times 6}$ containing wrapped phase components $(\sin\phi, \cos\phi)$, interferometric coherence $\gamma$, and unit look vectors $[\mathbf{e}_E, \mathbf{e}_N, \mathbf{e}_U]$. The model predicts a continuous line-of-sight (LOS) displacement map $\hat{\mathbf{y}}$, where the physical displacement $d_{\text{LOS}}$ relates to the absolute phase via $d_{\text{LOS}} = \frac{\lambda\phi}{4\pi}$ (for Sentinel-1, $\lambda = 5.6$cm).

**Physics-Aligned Objective.** To penalize unphysical discontinuities while remaining robust to decorrelation noise, we optimize a composite loss:

$$\mathcal{L} = \text{Huber}_{\delta=1}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_{\text{grad}} \sum_{i\in\{x,y\}} \|\nabla_i \hat{\mathbf{y}} - \nabla_i \mathbf{y}\|_1 \tag{1}$$

where $\lambda_{\text{grad}} = 0.1$. This combination is selected over standard $L_2$ or Laplacian regularization to better handle the heavy-tailed noise distributions typical of real-world LiCSAR products while explicitly enforcing the first-order smoothness priors required for geophysical validity.

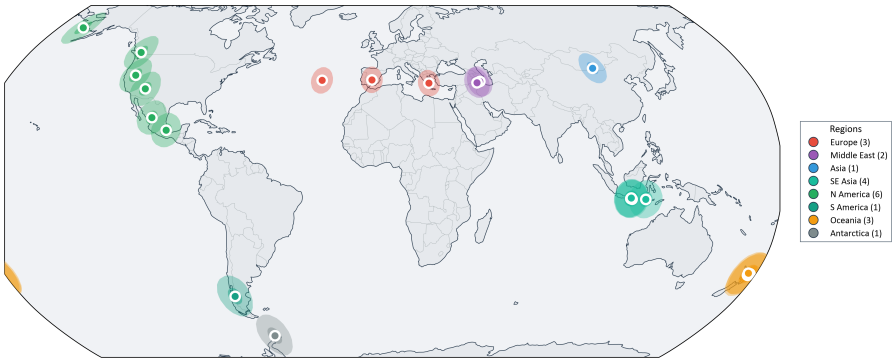



Figure 1: Geographic distribution of 20 LiCSAR frames across 6 continents.

## 3 Experimental Framework

**Operational Benchmark Construction.** We curate a global InSAR dataset from 350 operational LiCSAR interferograms (2020–2025) (Lazecký et al., 2020) spanning 20 frames across six continents (Fig. 1). The dataset encompasses diverse volcanic (e.g., White Island, Pico de Orizaba), tectonic (Middle Gobi, Sahand), and glacio-tectonic (Deception Island) regimes. Each sample integrates wrapped phase, SNAPHU-unwrapped ground truth, coherence ($\gamma \in [0, 1]$), and East-North-Up look vectors. From full-frame products, we extract $128 \times 128$ patches (stride = 64) and apply strict quality filters ($\bar{\gamma} > 0.5$, max displacement $>$ 1mm), yielding 39,724 high-quality patches (651M pixels).

*Critical Innovation:* To prevent spatial leakage, we implement frame-level stratified splitting, assigning entire geographic regions exclusively to train (14 frames), validation (3 frames), or test (3 frames) sets, ensuring generalization to unseen geographic provinces.

**Systematic Architectural Ablation.** To isolate the impact of recent computer vision (CV) advancements on geophysical regression, all models utilize an identical 4-level U-Net backbone (Ronneberger et al., 2015) (base channels $C = 32$). We evaluate four levels of increasing complexity:

- **V-UNet (Vanilla, 7.76M params):** Standard $2 \times$ (Conv3 $\times$ 3 $\rightarrow$ BN $\rightarrow$ ReLU) blocks with skip connections; our primary baseline for local inductive bias.
- **E-UNet (Enhanced, 8.29M params):** Incorporates Squeeze-Excitation blocks (Hu et al., 2019) after each encoder stage for channel-wise recalibration.
- **A-UNet (Attention, 11.37M params):** Integrates 4-head self-attention at the bottleneck and spatial attention gates at skip connections (Schlemper et al., 2019) for global context.
- **H-UNet (Hybrid, 17.21M params):** Combines SE blocks, MHSA, and an Atrous Spatial Pyramid Pooling (ASPP) (Chen et al., 2018) bottleneck to capture multi-scale features (see Appendix A).

**Training Protocol.** Models are optimized using AdamW with a OneCycleLR scheduler. To ensure a fair comparison, we perform a validation grid search for each model to determine optimal dropout (0.0–0.2) and weight decay ($5 \times 10^{-5}$–$10^{-4}$), accounting for the increased capacity of larger variants. Attention and Hybrid models use mixed-precision (FP16); Vanilla and Enhanced use full FP32. All models use batch size 32 and early stopping (patience = 100). (see Appendix B)

## 4 Results & Analysis

**Quantitative Performance.** Table 1 summarizes performance across 5,961 geographically held-out patches. The **Vanilla U-Net consistently achieves the best performance** despite having 32–122% fewer parameters, revealing a systematic "complexity penalty": attention mechanisms lead to a 25% $R^2$ drop ($0.834 \rightarrow 0.622$) and 51% RMSE increase. Vanilla U-Net reaches the operational threshold ($<$1cm error) in 88% of predictions versus only 67.5% for the Hybrid, confirming convolutional locality better aligns with geophysical regression.

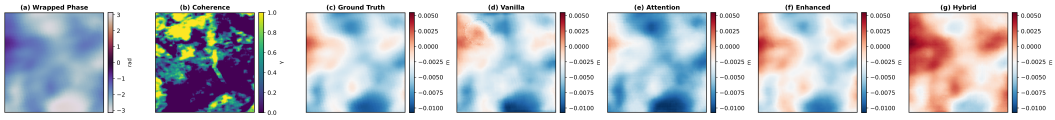



Figure 2: Representative predictions across test regimes.

| Model | RMSE (cm) | MAE (cm) | $R^2$ | P@1.0 (%) | 95th %ile (cm) | Params (M) |
|---|---|---|---|---|---|---|
| **Vanilla** | **1.009** | **0.492** | **0.834** | **88.0** | **2.03** | 7.76 |
| Enhanced | 1.149 | 0.647 | 0.786 | 80.7 | 2.64 | 8.29 |
| Attention | 1.528 | 0.844 | 0.622 | 75.6 | 3.40 | 11.37 |
| Hybrid | 1.595 | 1.001 | 0.588 | 67.5 | 3.49 | 17.21 |

Table 1: Test set performance on 5,961 held-out patches. Bold indicates best performance.

**Operational Efficiency.** The Vanilla U-Net achieves 2.92ms latency, a 2.5$\times$ speedup over the Hybrid model (Table 2). The 2.2$\times$ lower memory footprint (29.62MB) is critical for deployment on resource-constrained observatory edge-nodes. While all variants meet the sub-100ms requirement for early warning, Vanilla enables continental-scale monitoring at a fraction of computational cost.

**Physics-Grounded Failure Analysis.** Power Spectral Density (PSD) analysis (Figure 3b) reveals that Vanilla and Enhanced models accurately preserve the ground-truth spectrum. In contrast, Attention and Hybrid models inject spurious high-frequency power at $> 0.3$ cycles/pixel. Given that crustal deformation is governed by elasticity, true signals rarely exhibit sub-wavelength variations at

| Model | Memory (MB) | FLOPs (G) | Latency (ms) | Params (M) |
|---|---|---|---|---|
| **Vanilla** | **29.62** | **1017.63** | **2.92 ± 0.06** | 7.76 |
| Enhanced | 31.61 | 1086.21 | 6.35 ± 0.07 | 8.29 |
| Attention | 43.38 | 1490.66 | 7.08 ± 0.07 | 11.37 |
| Hybrid | 65.64 | 2255.24 | 7.13 ± 0.17 | 17.21 |

Table 2: Operational efficiency profiling (NVIDIA GH200).

the 14m Sentinel-1 scale. Consequently, the high-frequency content produced by complex models represents hallucinated unphysical artifacts rather than legitimate geophysical signal.

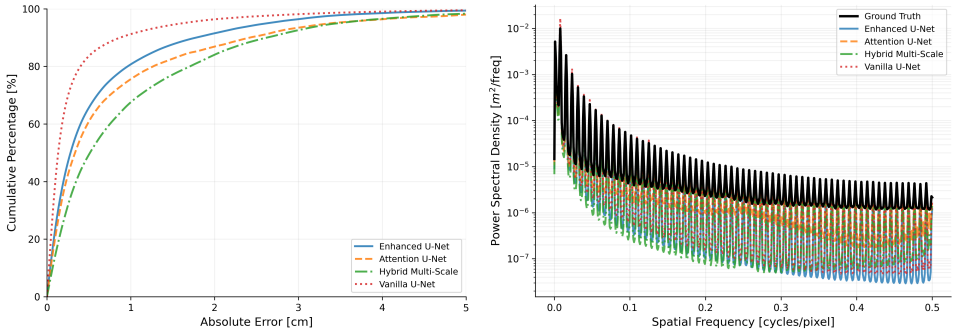



Figure 3: **(a)** Cumulative error distribution. **(b)** Power spectral density analysis

**Root Causes of Failure.** We identify three mechanisms driving this divergence: (1) *Inductive bias mismatch*: Attention mechanisms excel at detecting discrete boundaries in natural images (Dosovitskiy et al., 2021; Vaswani et al., 2023); however, InSAR displacement is characterized by high spatial autocorrelation, making the global flexibility of attention a liability for continuous fields, disrupting local autocorrelation structure and introducing spurious long-range dependencies. (2) *Capacity-data mismatch*: The 17M-parameter Hybrid models appear to overfit frame-specific atmospheric noise rather than underlying physics, evidenced by degraded generalization to held-out test frames despite strong training performance. (3) *Multi-scale misapplication*: ASPP-driven aggregation introduces aliasing artifacts when regressing the smooth spectral decay characteristic of geophysical deformation. These failure modes are visually confirmed in Figures 2 and 5: across volcanic, tectonic, and vegetated regimes, Vanilla predictions closely match the smooth gradients of SNAPHU ground truth, whereas attention-based models exhibit unphysical discontinuities and localized artifacts, particularly near patch boundaries and in low-coherence regions.

## 5 Discussion & Conclusion

**Design Principles for ML4RS:** (1) Domain ablations are mandatory: ImageNet winners fail when geophysical physics dominates. (2) Match inductive bias to physics: Convolutional locality beats global attention for autocorrelated fields. (3) Validate with physics diagnostics: Spectral analysis reveals violations invisible to RMSE. (4) Simplicity generalizes better: Vanilla models learn physics rather than scene-specific noise. Complexity may suit temporal or multi-modal tasks, but for smooth-field regression, domain physics must guide design.

**Limitations & Future Work.** Parameter count differences (7.76M–17.21M) and single split are acknowledged. Future research should explore capacity-matched variants, multi-sensor generalization (ALOS-2/NISAR), and physics-hybrid layers embedding elasticity constraints.

**Conclusion.** We presented the first systematic architectural ablation for operational InSAR across 20 frames and 651M pixels. Vanilla U-Net outperforms complex variants by 34% in $R^2$ with 2.5$\times$ faster inference. PSD analysis confirms that architectural complexity injects high-frequency artifacts via inductive bias mismatch. For physics-constrained regression, domain physics, not architectural sophistication, should guide ML4RS design. Less is more.

## REFERENCES

Curtis W. Chen and Howard A. Zebker. Two-dimensional phase unwrapping with use of statistical models for cost functions in nonlinear optimization. *J. Opt. Soc. Am. A*, 18(2):338–351, Feb 2001. doi: 10.1364/JOSAA.18.000338. URL https://opg.optica.org/josaa/abstract.cfm?URI=josaa-18-2-338.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 833–851, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. URL https://arxiv.org/abs/1709.01507.

Milan Lazecký, Karsten Spaans, Pablo J. González, Yasser Maghsoudi, Yu Morishita, Fabien Albino, John Elliott, Nicholas Greenall, Emma Hatton, Andrew Hooper, Daniel Juncu, Alistair McDougall, Richard J. Walters, C. Scott Watson, Jonathan R. Weiss, and Tim J. Wright. Licsar: An automatic insar tool for measuring and monitoring tectonic and volcanic activity. *Remote Sensing*, 12(15), 2020. ISSN 2072-4292. doi: 10.3390/rs12152430. URL https://www.mdpi.com/2072-4292/12/15/2430.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, feb 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0912-1. URL https://doi.org/10.1038/s41586-019-0912-1.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S1361841518306133.

G. E. Spoorthi, Subrahmanyam Gorthi, and Rama Krishna Sai Subrahmanyam Gorthi. Phasenet: A deep convolutional neural network for two-dimensional phase unwrapping. *IEEE Signal Processing Letters*, 26(1):54–58, 2019. doi: 10.1109/LSP.2018.2879184.

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(sup1):234–240, 1970. doi: 10.2307/143141. URL https://www.tandfonline.com/doi/abs/10.2307/143141.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Lifan Zhou, Hanwen Yu, Yang Lan, and Mengdao xing. Artificial intelligence in interferometric synthetic aperture radar phase unwrapping: A review. *IEEE Geoscience and Remote Sensing Magazine*, 9(2):10–28, 2021. doi: 10.1109/MGRS.2021.3065811.

# A Architectural Specifications and Design Rationales

To ensure full reproducibility and provide a technical basis for the "Complexity Penalty" observed in our experiments, we detail the implementation of all four architectural variants (Fig. 4).

## A.1 Variant Specifications

**Vanilla U-Net (7.76M parameters):** Our baseline serves as the minimalist control group, adhering strictly to the original U-Net topology but with modern normalization.

- **Encoder:** 4-level hierarchy. Each level consists of two $3 \times 3$ convolutions (padding=1).
- **Block Structure:** Conv3 $\times$ 3 $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ Conv3 $\times$ 3 $\rightarrow$ BatchNorm $\rightarrow$ ReLU.
- **Downsampling:** $2 \times 2$ Max Pooling with stride 2.
- **Channel Progression:** $[32, 64, 128, 256, 512]$.
- **Decoder:** Up-convolutions via ConvTranspose2d followed by concatenation-based skip connections from the corresponding encoder stage.
- **Output Head:** $1 \times 1$ Convolution mapping to a single-channel displacement map.

**Enhanced U-Net (8.29M parameters):** This variant investigates whether channel-wise recalibration can improve phase ambiguity resolution in low-coherence regions.

- **Base:** Identical to Vanilla U-Net.
- **Addition:** Squeeze-and-Excitation (SE) blocks integrated after each encoder stage.
- **SE Formulation:** Let $\mathbf{h}$ be the input feature map. The gated scale $\mathbf{s}$ is:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \text{GlobalAvgPool}(\mathbf{h}))) \tag{2}$$

  where the final output is $\tilde{\mathbf{h}} = \mathbf{s} \odot \mathbf{h}$.
- **Reduction Ratios:** $r = \{4, 8, 8, 16\}$ for levels 1 through 4, respectively.

**Attention U-Net (11.37M parameters):** Designed to capture global dependencies and focus on salient deformation regions through spatial gating.

- **Bottleneck:** Multi-head self-attention (4 heads, $d_k = 128$).
- **Self-Attention:**

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

- **Skip Connections:** Replaced standard concatenation with Gated Spatial Attention. The attention coefficient $\alpha$ is derived from the gating signal $\mathbf{g}$ (lower level) and the skip feature $\mathbf{x}$ (encoder):

$$\alpha = \sigma(\psi^T \cdot \text{ReLU}(\mathbf{W}_g\mathbf{g} + \mathbf{W}_x\mathbf{x})) \tag{4}$$

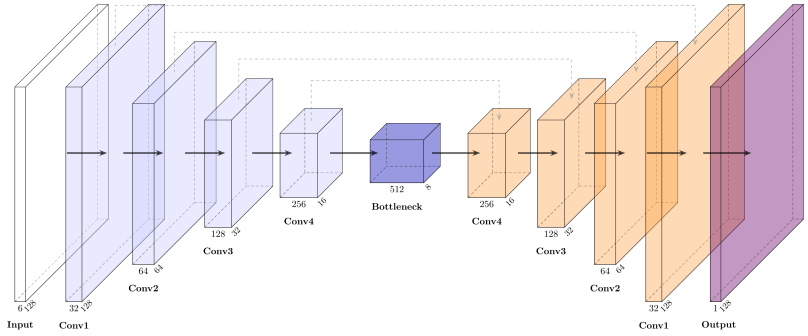
  The gated output is $\tilde{\mathbf{x}} = \alpha \odot \mathbf{x}$.

**Hybrid Multi-Scale U-Net (17.21M parameters):** Our most complex variant, combining the SE encoder, ASPP bottleneck, and Gated Attention skips to maximize receptive field multi-modality.

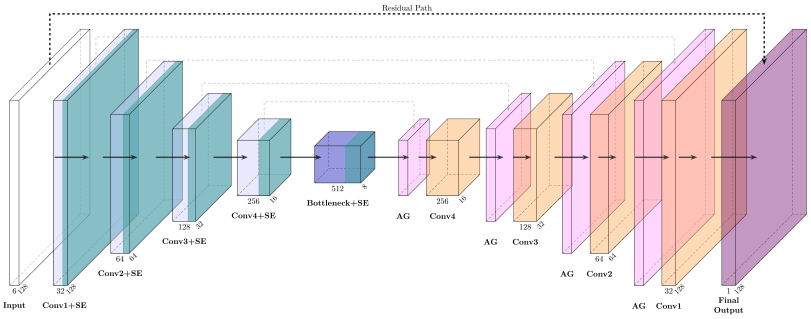
- **Bottleneck:** Atrous Spatial Pyramid Pooling (ASPP) utilizing parallel dilated convolutions.
- **ASPP Structure:**

$$\mathbf{f}_{\text{ASPP}} = \text{Concat}(\mathbf{f}_1^{1\times1}, \mathbf{f}_6^{3\times3}, \mathbf{f}_{12}^{3\times3}, \mathbf{f}_{18}^{3\times3}, \mathbf{f}_{\text{global}}) \tag{5}$$

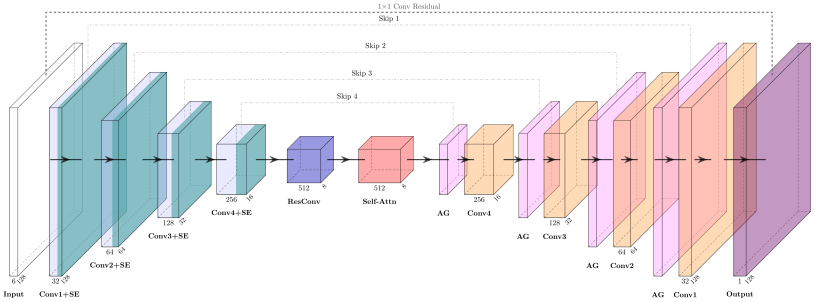
- **Effective Receptive Fields:** Rates of $\{3, 13, 25, 37\}$ pixels plus a global average pooling branch.



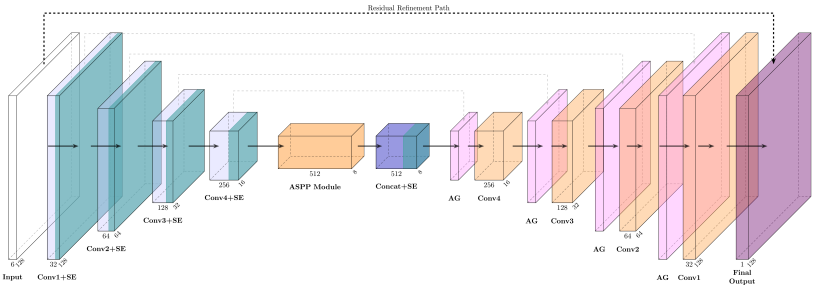

(a) Vanilla U-Net (7.76M params)




(b) Enhanced U-Net (8.29M params)




(c) Attention U-Net (11.37M params)




(d) Hybrid Multi-Scale (17.21M params)

Figure 4: Detailed architectural variants evaluated in this study

## B Training Regimes and Hyperparameters

All models were trained using a standardized protocol to ensure the performance differences are attributable solely to architectural capacity.

### B.1 Hyperparameter Configuration

| Model | Max LR | Weight Decay | Dropout | Params (M) |
|---|---|---|---|---|
| Vanilla U-Net | $8 \times 10^{-5}$ | $5 \times 10^{-5}$ | 0.0 | 7.76 |
| Enhanced U-Net | $8 \times 10^{-5}$ | $5 \times 10^{-5}$ | 0.10 | 8.29 |
| Attention U-Net | $8 \times 10^{-5}$ | $1 \times 10^{-4}$ | 0.20 | 11.37 |
| Hybrid Multi-Scale | $5 \times 10^{-5}$ | $1 \times 10^{-4}$ | 0.20 | 17.21 |

Table 3: Hyperparameter configurations for all architectural variants.

**Learning Rate Schedule:** We employed the OneCycleLR scheduler with a 10% linear warmup phase, followed by a cosine annealing decay to $\eta_{\text{max}}/25$.

**Optimization:** AdamW with $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We applied gradient clipping with a maximum norm of 1.0 to ensure stability in the Attention and Hybrid models.

**Mixed-Precision Details:** Attention and Hybrid models were trained with PyTorch's automatic mixed-precision (AMP) with FP16 forward passes and FP32 master weights, with loss scaling handled automatically via `GradScaler`. BatchNorm layers are kept at FP32 by PyTorch's autocast by default. The Vanilla and Enhanced models were trained in full FP32 precision given their smaller memory footprint.

**Convergence:** Maximum epochs set to 1000 with an early stopping patience of 100 validation epochs. Typical convergence occurred between 300–500 epochs.

## C Data Preprocessing and Quality Control

**Patch Extraction:** Input interferograms (2000–3000 pixels squared) were decomposed into $128 \times 128$ patches with a 64-pixel stride (50% overlap) to preserve spatial continuity.

**Quality Filtering:** To prevent model bias from decorrelated noise, we applied the following strict inclusion criteria:

- **Mean Coherence:** $\bar{\gamma} > 0.5$ (eliminates water bodies and dense vegetation).
- **Signal Threshold:** Maximum displacement $> 1$ mm.
- **Data Integrity:** $> 95\%$ valid pixels per patch.

**Normalization:** Channel-wise training statistics were computed once: $\mathbf{x}_{\text{norm}} = (\mathbf{x} - \mu_{\text{train}})/(\sigma_{\text{train}} + 10^{-8})$.

## D Computational Resources and Efficiency

**Hardware:** All experiments were conducted on an NVIDIA GH200 GPU (120GB VRAM). **Training Time:**

- **Vanilla:** $\sim$8 hours.
- **Enhanced:** $\sim$10 hours.
- **Attention:** $\sim$11 hours.
- **Hybrid:** $\sim$16 hours.

Total computational budget: $\sim$100 GPU-hours including search and validation runs.

## E Extended Visual Results

Figure 5 provides a systematic visual comparison across four distinct geographic and tectonic settings, highlighting the robustness of the Vanilla baseline compared to the artifact-prone Hybrid model.

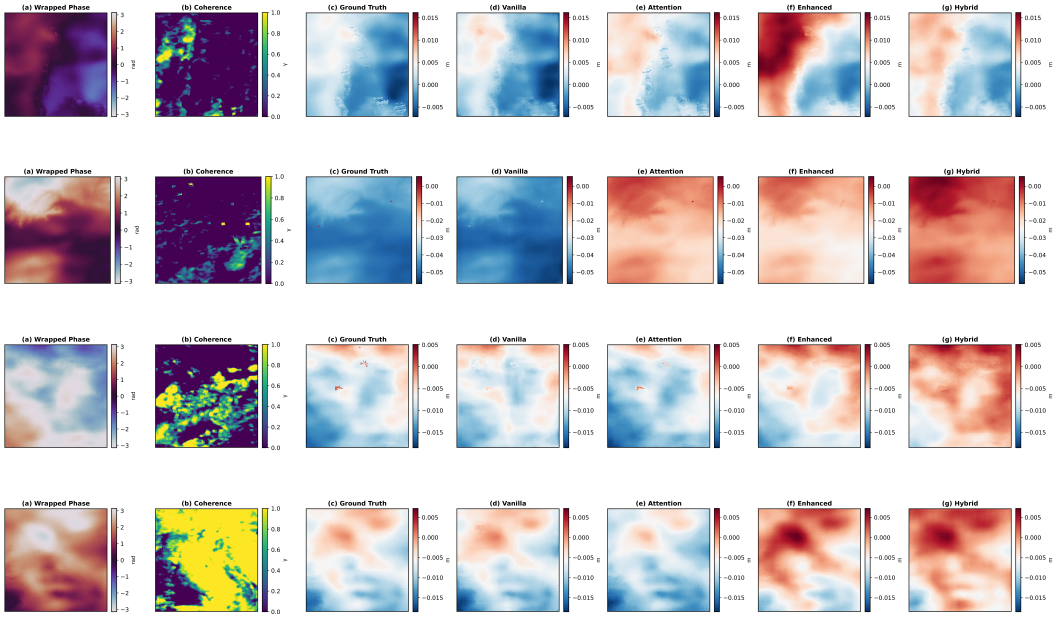



Figure 5: Visual comparison of phase unwrapping results. Each grid presents: (a) Wrapped Phase, (b) Coherence, (c) Ground Truth, (d) Vanilla, (e) Attention, (f) Enhanced, and (g) Hybrid predictions.