Supplementary Material

A Additional Ablations

Table 1: Complexity Loss Weight β (Sen1Floods11 - Frozen Backbone, 20 Epochs)

β	mIoU (%)	C. Std Dev	Time
0.0	86.94 %	0.08	2.3h
0.01	87.42 %	0.18	2.4h
0.05	87.58 %	**0.22**	**2.4h**
0.1	87.31 %	0.26	2.5h

Finding: $\beta = 0.05$ (used in main experiments) provided the best trade-off during these shorter ablation runs.

B Hyperparameter Sensitivity

Table 2: Learning Rate Sweep (GeoBench Mean mIoU - Full Fine-Tuning)

Initial LR	Warmup	Mean mIoU	Conv.
5×10^{-5}	5	85.72 %	68
1×10^{-4}	**8**	86.66 %	**52**
2×10^{-4}	10	86.19 %	47
5×10^{-4}	15	85.31 %	42

Finding: Initial LR=1×10^{-4} (used in main experiments) performed best for full fine-tuning.

C Computational Details

Hardware: All experiments were conducted on NVIDIA H100 GPUs with 80GB VRAM.

Efficiency Measurement Protocol (Section 4.5): Inference throughput and latency were measured using standard PyTorch CUDA event timing. We performed 200 warmup iterations followed by 500 measurement iterations for each model configuration, averaging the results. Measurements were taken with mixed precision (FP16) enabled, but without 'torch.compile'. Batch size was 1 for latency measurements.

Parameter Counts: Reported parameter counts represent the number of trainable parameters as reported by standard PyTorch model summary utilities (e.g., `sum(p.numel() for p in model.parameters() if p.requires_grad)`). For DARN (Decoder Only) and UPerNet (Decoder Only), this counts parameters in the decoder head. For TerraMind-L (Full Model), this counts parameters in the entire model (encoder + standard decoder).

D Robustness Evaluation Details

Scope: Robustness evaluations were performed in the **frozen backbone** setting, comparing the DARN decoder against the TerraMind-L baseline decoder. Evaluations were conducted on the combined validation sets of the 6 GeoBench tasks, with results averaged across tasks.

Seed: All reported robustness results are based on a **single training run** (the same run used for main results) for both DARN and the baseline reimplementation, due to computational constraints. While averaging over multiple seeds would be ideal, these single-run results provide a strong indication of relative robustness.

FGSM Attack: We used the standard Fast Gradient Sign Method (FGSM) adversarial attack. It was performed with a **single step** and an epsilon (ϵ) value of 8/255, applied to the input image before feature extraction by the frozen backbone. The model's prediction on the perturbed features was then evaluated.

Common Corruptions (mCE): We used the standard benchmark corruptions defined in Hendrycks & Dietterich. The Mean Corruption Error (mCE) was calculated by averaging the segmentation mIoU degradation across the 15 corruption types (grouped into Noise, Blur, Digital, Weather categories) and across **severities 1 through 5** for each corruption type, following the standard protocol. The implementation follows common benchmarks available online (e.g., robustness benchmarks associated with).