

1 Appendix A: Training and Implementation Details

2 **Compute environment.** Training was conducted on two nodes equipped with 8 NVIDIA Quadro RTX
3 8000 GPUs (each with 48 GB memory). Mixed-precision training (AMP) was enabled throughout.
4 Distributed training was implemented via `torchrun` with NCCL backend, using socket-based
5 transport for stability. Inference evaluations were performed on a single RTX 8000 GPU and also a
6 single NVIDIA RTX 2000 Ada GPU (8 GB VRAM).

7 **Training configuration.** We summarize the key hyperparameters in Table 1.

Table 1: Training configurations for progressive pretraining.

| Parameter | Stage 1: FMoW-S2 | Stage 2: BigEarthNet-S2 |
|--------------------|-----------------------------|-----------------------------|
| Image resolution | 96×96 | 128×128 |
| Epochs | 200 | 100 |
| Batch size per GPU | 256 | 256 |
| Total batch size | 2048 | 2048 |
| Learning rate | 1×10^{-4} | 5×10^{-5} |
| Weight decay | 5×10^{-5} | 5×10^{-5} |
| Warmup epochs | 10 | 5 |
| Scheduler | Cosine decay | Cosine decay |
| Gradient clipping | <code>max_norm = 1.0</code> | <code>max_norm = 1.0</code> |
| Steps per epoch | 695 | 537 |
| Warmup steps | 6950 | 2685 |

8 Appendix B: Dataset Processing and Band Grouping

9 We apply a unified preprocessing pipeline across all Sentinel-2 datasets used in this work. Key
10 preprocessing steps include band selection, normalization, and spatial resizing, detailed in Table 2.
11 Bands B01, B09, and B10 are excluded due to their coarse 60 m resolution. All retained bands are
grouped into physically coherent spectral categories for structured encoding.

Table 2: The summary of common preprocessing and dataset-specific properties.

| Common Preprocessing Across All Datasets | | | |
|------------------------------------------|---------------------------------------------|-------------|------------------------------------------|
| Normalization | Pixel values divided by 10,000 | | |
| Valid reflectance range | Clipped to [0, 1.2] | | |
| Resizing method | Bicubic interpolation to fixed spatial size | | |
| Excluded bands | B01, B09, B10 (60 m resolution) | | |
| Retained bands | B02–B08A, B11, B12 (10 bands) | | |
| Spectral groupings | Visible: B02, B03, B04 | | |
| | Red-Edge/NIR: B05–B08, B8A | | |
| | SWIR: B11, B12 | | |
| Dataset-Specific Details | | | |
| Dataset | Tile Size | #Tiles | Task |
| FMoW-S2 | 96 × 96 | 712,874 | Pretraining & scene-level classification |
| BigEarthNet-S2 | 128 × 128 | 549,488 | Pretraining & multi-label classification |
| OSCD | 96 × 96 | 336 pairs | Change detection |
| SegMunich | 128 × 128 | 8,430 | Semantic segmentation |
| DynaS2 | 256 × 256 | 5,472 pairs | Multi-temporal change detection |
| EuroSAT | 64 × 64 | 27,000 | Land cover classification |

12

13 **Dataset licenses and sources.**

- 14 • **BigEarthNet-S2:** <https://bigearth.net>. Licensed under the **Community Data Li-**
15 **cense Agreement – Permissive, Version 1.0.**

- **FMoW-S2**: <https://github.com/fMoW/dataset>. Released under the **Functional Map of the World Challenge Public License**. Openly available for non-commercial research use.
- **OSCD (Onera Satellite Change Detection)**: <https://rcdaudt.github.io/oscd/>. Publicly released for academic benchmarking; **no explicit license** provided.
- **DynaS2 (DynamicEarthNet Sentinel-2)**: <https://mediatum.ub.tum.de/1650201>. Licensed under **Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)**.
- **EuroSAT**: <https://github.com/phelber/eurosat>. Licensed under the **MIT License**.
- **SegMunich**: <https://huggingface.co/datasets/earthflow/SegMunich>. Distributed under **Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)**.

Appendix C: Expanded Ablation Studies

All ablations in this section are conducted using the PhySwin-T model pretrained on the BigEarthNet-S2 dataset for 100 epochs. So, the reported results may slightly differ from those presented in Section 4. Evaluation is performed on three downstream tasks: SegMunich (semantic segmentation), OSCD (change detection), and EuroSAT (land cover classification).

C.1 Physics Loss Weights

Table 3: Sensitivity to physics-informed loss weights (λ, β) . Mixing ratio is fixed at 50%, and spectral grouping follows the default setting in Section 4. The best results are highlighted in **bold**.

| (λ, β) | SegMunich mIoU (%) | OSCD F1 (%) | EuroSAT OA (%) |
|---------------------|--------------------|--------------|----------------|
| (0.1, 0.05) | 47.97 | 56.33 | 96.64 |
| (0.25, 0.1)* | 48.46 | 56.98 | 96.88 |
| (0.5, 0.2) | 48.21 | 57.03 | 96.27 |

C.2 Mixing Ratio Effects

Table 4: Ablation on MixMAE spatial mixing ratio. Physics-informed loss weights are fixed at $(\lambda = 0.25, \beta = 0.1)$, and spectral grouping follows the default setting in Section 4. The best results are highlighted in **bold**.

| Mixing Ratio | SegMunich mIoU (%) | OSCD F1 (%) | EuroSAT OA (%) |
|--------------|--------------------|--------------|----------------|
| 50%* | 48.46 | 56.98 | 96.88 |
| 67% | 45.29 | 55.63 | 96.01 |
| 75% | 42.97 | 53.61 | 94.79 |

C.3 Spectral Grouping Variants

Table 5: Performance of different spectral grouping strategies. Physics loss weights are fixed at $(\lambda = 0.25, \beta = 0.1)$, and the MixMAE spatial mixing ratio is fixed at 50%. Default grouping is: Visible (B02–B04), RedEdge+NIR (B05–B08A), SWIR (B11–B12). The best two results are highlighted in **bold**.

| Grouping Scheme | #Groups | SegMunich mIoU (%) | OSCD F1 (%) | EuroSAT OA (%) |
|--------------------------------------|---------|--------------------|--------------|----------------|
| Visible RedEdge+NIR SWIR* | 3 | 48.46 | 56.98 | 96.88 |
| Visible+RedEdge NIR SWIR | 3 | 48.37 | 57.41 | 96.04 |
| Visible+NIR RedEdge SWIR | 3 | 47.95 | 56.74 | 97.11 |
| Visible+SWIR RedEdge NIR | 3 | 47.48 | 55.23 | 96.49 |
| Visible+SWIR RedEdge+NIR | 2 | 46.77 | 54.45 | 96.73 |
| Visible RedEdge NIR SWIR | 4 | 48.24 | 57.20 | 96.07 |

36 Appendix D: Model Configurations

37 D.1 MixMAE Decoder Configuration

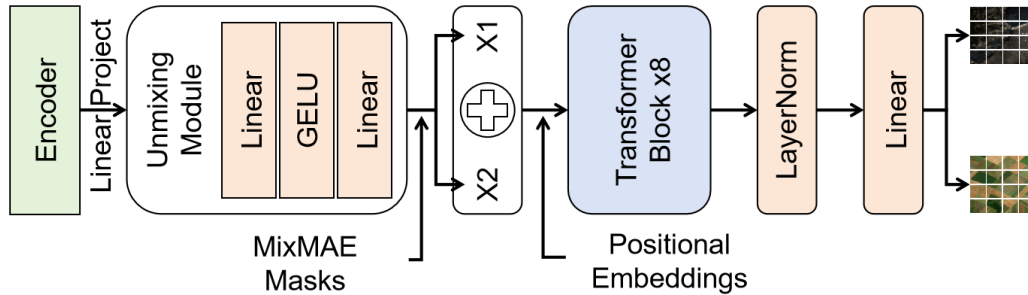


Figure 1: PhySwin Decoder Structure.

PhySwin’s decoder maps the encoder’s hidden states (dimension equal to the encoder’s hidden size) into per-patch reconstructions via a lightweight Transformer stack, as shown in Fig. 1. Concretely, the decoder first projects hidden features into a $D_{\text{dec}} = 512$ -dimensional space, then applies an *unmixing module* (two linear layers with GELU nonlinearity) to disentangle mixed tokens. Two streams are built by masking/unmasking this output according to the MixMAE mask, then concatenated and summed with bicubically-interpolated 2D sine-cosine positional embeddings. This fused sequence is processed by 8 Transformer blocks, followed by LayerNorm and a final linear prediction head that emits $\text{stride}^2 \times C$ values per token (where $\text{stride} = 4$ and $C = 10$ bands). All linear layers are Xavier-initialized and biases zeroed.

47 D.2 Downstream Plug-and-Play Heads

To ensure fair comparisons across backbones, we adopt a “plug-and-play” evaluation strategy: all models, including PhySwins and baselines, share the same task-specific head types, and only the encoder varies across models. This design isolates the effect of pretraining and encoder quality.

Table 6: Downstream task heads and objectives. UPerNet = Unified Perceptual Parsing Network; FPN = Feature Pyramid Network; MLP = multi-layer perception. All models use a frozen encoder with plug-and-play heads.

| Dataset | Task | Head Type | Objectives |
|---------------|----------------------------|---------------------------------------------------|----------------------|
| SegMunich | Semantic segmentation | UPerNet decoder (FPN + classifier) | Cross-entropy |
| Dyna.-S2 | Semantic segmentation | UPerNet decoder (FPN + classifier) | Cross-entropy |
| OSCD | Binary change detection | FPN + 3-layer Conv | Binary cross-entropy |
| Dyna.-S2 (CD) | Semantic change detection | FPN + 3-layer Conv | Cross-entropy |
| FMoW-S2 | Scene classification | GlobalAvgPool + LayerNorm + 3-layer MLP | Cross-entropy |
| EuroSAT | Scene classification | GlobalAvgPool + LayerNorm + 3-layer MLP | Cross-entropy |
| BigEarthNet | Multi-label classification | GlobalAvgPool + LayerNorm + 3-layer MLP + Sigmoid | Binary cross-entropy |

50

51 Appendix D: Qualitative Examples

In this section, we provide additional qualitative examples for both semantic segmentation and change detection tasks. These remain highly challenging for RSFMs, and only a few of the predictions achieve acceptable accuracy. Nevertheless, PhySwin consistently delivers superior overall quality than competing methods, as seen in Figures 2, 4 and 5, and in many cases finer detail.

That said, there is still considerable room for improvement, especially on the most demanding benchmarks such as the Dyna.-S2 change detection challenge. Although PhySwin leads all SOTA

Table 7: Downstream task training configurations.

| Dataset | Input Size | Batch Size | Optimizer / LR | Epochs |
|---------------|------------|------------|----------------|--------|
| SegMunich | 128×128 | 96 | AdamW / 5e-4 | 70 |
| Dyna.-S2 | 256×256 | 36 | AdamW / 1e-4 | 120 |
| OSCD | 96×96 | 128 | AdamW / 1e-3 | 60 |
| Dyna.-S2 (CD) | 96×96 | 64 | AdamW / 5e-4 | 70 |
| FMoW-S2 | 96×96 | 128 | AdamW / 1e-5 | 100 |
| EuroSAT | 96×96 | 128 | AdamW / 1e-3 | 100 |
| BigEarthNet | 128×128 | 96 | AdamW / 5e-5 | 100 |

58 baselines in the quantitative metrics reported in Table 2, its combined mask visualizations in Figure
59 5 remain visually inconsistent. Bridging this gap between numerical performance and perceptual
60 fidelity will be a key focus for future RSFM development.

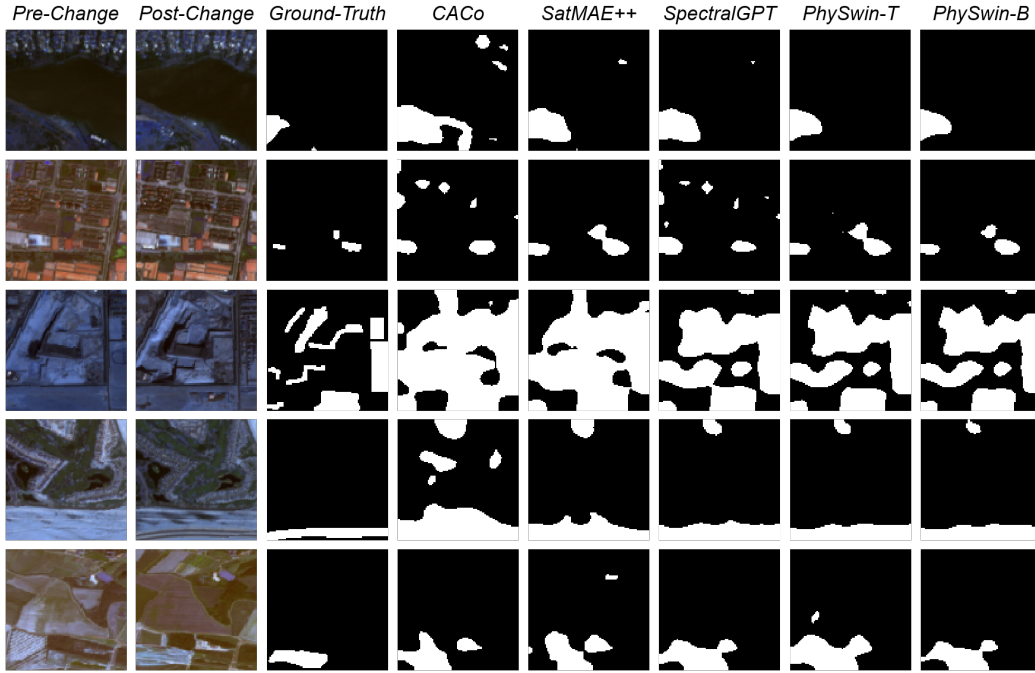


Figure 2: Additional visualizations on the OSCD dataset. Although pixel-level change detection remains challenging, PhySwin-B achieves the best performance among all baselines, producing fewer false positives.

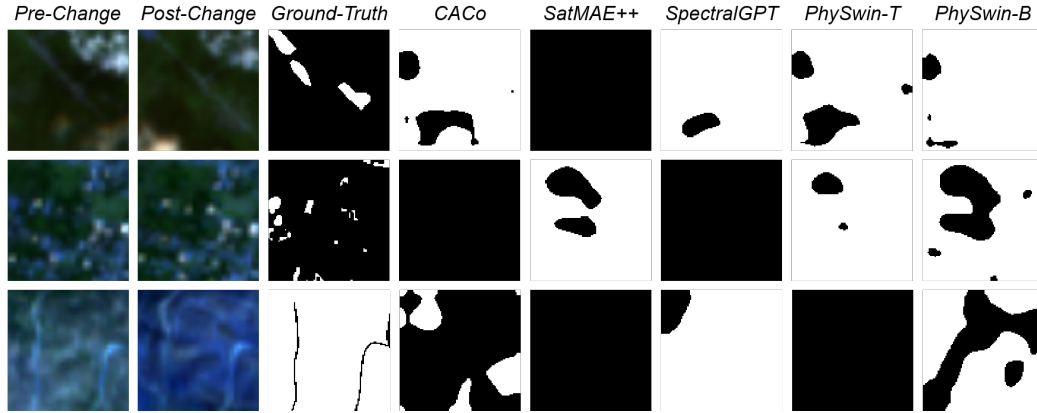


Figure 3: Additional visualizations on Dyna-S2 change detection. This benchmark is even more challenging, and visually all RSFMs fail to produce coherent change masks. Although PhySwin-B attains the best quantitative results, these examples highlight the substantial room for improvement in RSFMs.

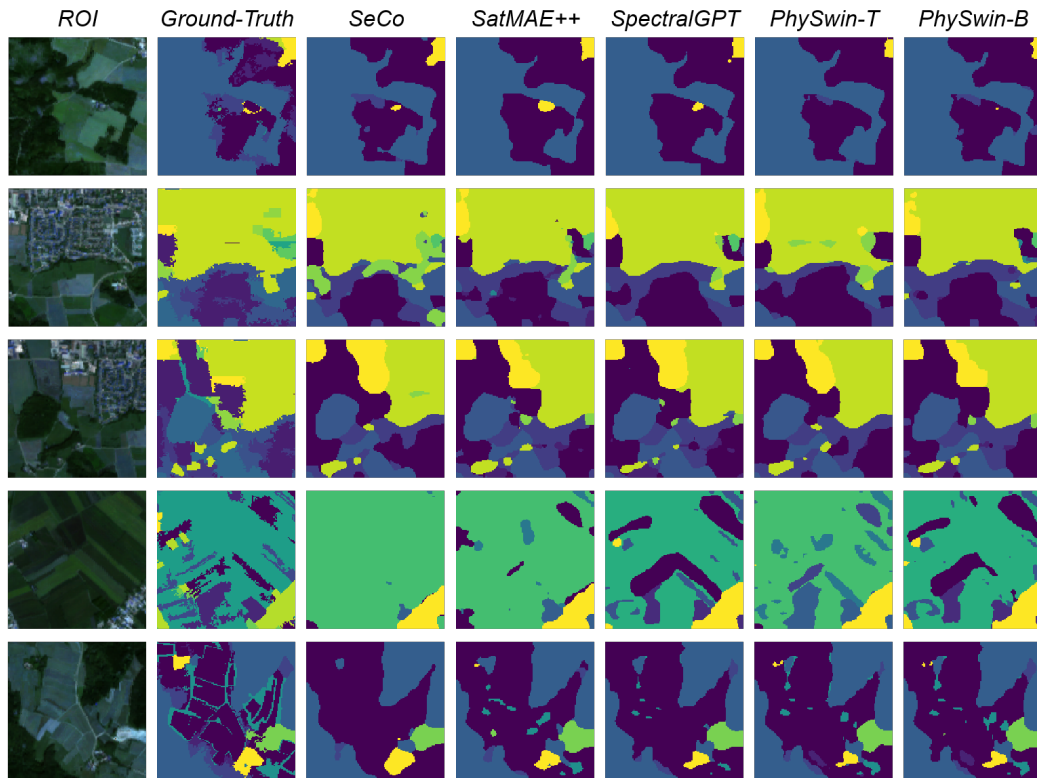


Figure 4: Additional visualizations on SegMunich semantic segmentation. While most models capture the overall scene layout, PhySwin-B more accurately delineates object boundaries and small regions, demonstrating finer detail.

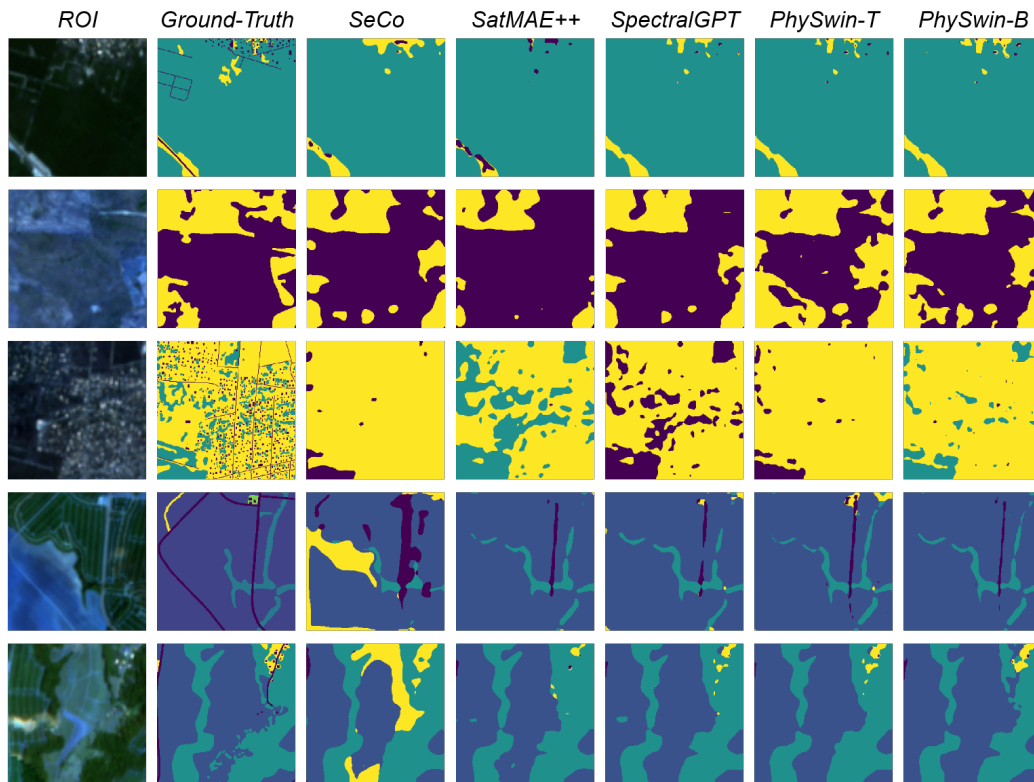


Figure 5: Additional visualizations on Dyna.-S2 semantic segmentation. Despite the overall challenge, most models capture the coarse layout. PhySwin-B presents finer details.