WEB2EARTH: EXTENDING WEB VISION FOUNDATION MODELS ACROSS REMOTE SENSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models pretrained on large-scale web imagery learn rich visual representations, yet their RGB-centric training limits transfer to other sensing modalities. We ask whether this knowledge can be adapted to remote sensing imagery while extending beyond optical RGB to other sensing modalities. We introduce Web2Earth, a lightweight post-pretraining framework that extends web-pretrained vision models to remote sensing across modalities. It combines RGB self-distillation for optical domain adaptation with asymmetric cross-modal transfer from an RGB teacher to a non-RGB student. To support this transfer, we propose two complementary objectives: Cross-Modal Relational Distillation (CMRD) and Co-Registered Patch Distillation (CoPaD), which transfer RGB knowledge at the scene and region levels, respectively. Web2Earth uses 22K RGB images and approximately 5K co-registered RGB–SAR pairs, while updating only approximately 2% of the backbone parameters. Web2Earth achieves leading or competitive performance across diverse downstream tasks compared with existing remote sensing foundation models, while improving optical and SAR representations over its DINOv3 initialization. The gains are particularly pronounced in retrieval: cross-scale optical R@1 increases from 86.9% to 97.0%, and RGB-to-SAR R@1 from 9.5% to 34.4%. These results show that a web-pretrained vision model can gain cross-modal capabilities while improving its optical performance, using limited additional data and parameter updates.

1 INTRODUCTION

Remote sensing foundation models aim to support diverse applications, from land-cover mapping to flood monitoring and change detection, through representations that transfer across downstream tasks (Cong et al., 2022; Guo et al., 2024). Their utility also depends on learning from different observations of Earth's surface. Optical RGB imagery captures visible appearance and spatial structure, while synthetic aperture radar (SAR) provides complementary measurements through a different image-formation process. Even the same scene can appear substantially different across sensors and spatial resolutions. A broadly useful model should therefore support both task transfer within each modality and correspondence across these different views of the same scene.

Large-scale domain-specific pretraining has advanced this goal through optical, multispectral, and multimodal learning (Cong et al., 2022; Fuller et al., 2023; Guo et al., 2024). Alongside these efforts, web-pretrained vision models offer an opportunity to reuse visual knowledge acquired outside remote sensing. DINOv2 and DINOv3 learn image-level and dense representations that transfer across visual tasks (Oquab et al., 2024; Siméoni et al., 2026). This knowledge is already useful in the target domain. In our optical classification and segmentation evaluations, web-pretrained DINOv3 outperforms its satellite-pretrained counterpart (Table 1). This observation motivates adapting existing web representations rather than assuming that useful remote sensing features must be learned through domain-specific pretraining from scratch. Our central question is: *Can lightweight post-pretraining extend a web-pretrained model beyond RGB while also improving its optical remote sensing performance?*

Realizing this goal requires both adapting RGB representations to overhead imagery and extending their utility to another sensing modality. A straightforward approach is continual pretraining on mixed RGB and SAR images using the original self-supervised objective. However, simply mixing the modalities does not explicitly exploit their pairing to guide knowledge transfer between corre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

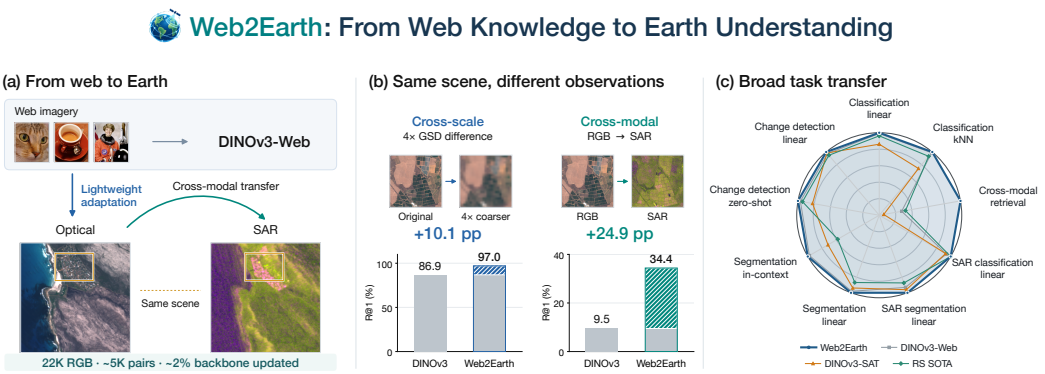


Figure 1: **Web2Earth: from web knowledge to Earth understanding.** (a) RGB self-distillation adapts DINOv3-Web to optical remote sensing, while asymmetric cross-modal transfer extends it to SAR with lightweight parameter updates. (b) Cross-scale optical ($4\times$ GSD) and RGB-to-SAR retrieval improve R@1 over DINOv3 by 10.1 and 24.9 percentage points, respectively. (c) Nine frozen-encoder evaluation settings, normalized to the best compared score on each axis. RS SOTA combines the best compared RS-specific baseline per axis.

sponding scenes, which may limit the model's ability to learn cross-modal correspondence. Exploiting these paired observations, however, requires deciding what should be transferred across sensors. Although co-registered RGB and SAR images depict the same ground regions, their different image-formation processes do not imply identical feature representations. This suggests transferring shared scene structure rather than enforcing equality between sensor embeddings. Scene-level relationships capture how different scenes relate to one another, while local predictions provide complementary supervision at corresponding regions.

We introduce **Web2Earth**, a lightweight post-pretraining framework that jointly adapts web-pretrained vision models to optical remote sensing and extends them across sensing modalities (Figure 1). RGB self-distillation adapts the model to the optical domain while two complementary objectives asymmetrically transfer knowledge from an RGB teacher to a non-RGB student. *Cross-Modal Relational Distillation (CMRD)* transfers similarities among scenes in a batch. *Co-Registered Patch Distillation (CoPaD)* complements this global supervision with local predictions at aligned patch positions. Both objectives accompany RGB self-distillation in joint training. Web2Earth uses 22K RGB images and approximately 5K co-registered RGB–SAR pairs, updating only approximately 2% of the backbone parameters.

We evaluate the resulting representations with frozen encoders on optical classification, segmentation, and change detection, as well as SAR classification and segmentation. Web2Earth achieves leading or competitive performance against existing remote sensing foundation models and improves optical and SAR performance over its DINOv3 initialization. We also evaluate whether the model recognizes the same scene when the observation changes. Cross-scale optical retrieval uses image pairs with a $4\times$ difference in ground sampling distance, while cross-modal retrieval matches RGB queries to SAR observations. Web2Earth improves R@1 from 86.9% to 97.0% and from 9.5% to 34.4%, respectively. Together, these results show that limited additional data and parameter updates can add cross-modal capabilities while improving optical performance.

Our contributions are threefold:

- We present Web2Earth, a lightweight post-pretraining framework that jointly adapts web-pretrained representations to optical remote sensing and extends their utility to SAR.
- We propose asymmetric cross-modal transfer through complementary CMRD and CoPaD objectives, using the RGB teacher's scene-level relations and local predictions to supervise non-RGB representations.
- We conduct extensive evaluations across diverse downstream tasks, including optical and SAR classification and segmentation, optical change detection, and cross-scale and cross-modal retrieval. These evaluations show that lightweight adaptation enables Web2Earth to achieve leading or competitive performance against existing remote sensing foundation

models across this broad task suite, while improving on DINOv3 under frozen-encoder evaluation.

2 Related Work

Vision and Remote Sensing Foundation Models Self-supervised vision foundation models have demonstrated strong transferability across diverse visual tasks. SimCLR introduced contrastive representation learning based on differently augmented versions of the same image (Chen et al., 2020), while MAE enabled scalable pretraining through masked image reconstruction (He et al., 2022). The DINO family further learned semantically structured patch-level representations via self-distillation, enabling dense prediction tasks such as semantic segmentation and depth estimation (Caron et al., 2021; Oquab et al., 2024). More recently, DINOv3-SAT extended this paradigm to large-scale satellite imagery, demonstrating the effectiveness of domain-specific pretraining for Earth observation (Siméoni et al., 2026).

Remote sensing foundation models further incorporate characteristics specific to Earth observation imagery. SatMAE (Cong et al., 2022) and Prithvi (Jakubik et al., 2023) model multispectral and multitemporal observations, whereas Scale-MAE explicitly accounts for spatial scale and ground coverage (Reed et al., 2023). GFM instead adapts a natural-image-pretrained model through continual pretraining on geospatial data (Mendieta et al., 2023). Recent multimodal approaches broaden the supported sensor space: CROMA jointly learns SAR and optical representations (Fuller et al., 2023), SkySense models multimodal spatiotemporal observations (Guo et al., 2024), and DOFA uses wavelength-conditioned dynamic embeddings to accommodate varying spectral configurations (Xiong et al., 2024).

These approaches incorporate remote sensing characteristics through domain-specific pretraining and multimodal learning. Our focus is complementary: we investigate how a web-pretrained RGB encoder can acquire useful SAR representations while also improving its optical remote sensing performance through lightweight post-pretraining. Rather than introducing a sensor-flexible input architecture, Web2Earth exploits paired RGB–SAR observations to transfer scene-level relations and local predictions.

Continual Pretraining for Domain Adaptation Continued pretraining adapts an existing model to a target distribution by further optimizing its pretraining objective on unlabeled in-domain data before downstream task training. This approach was explored early in language-model transfer: ULMFiT adapted a generally pretrained language model to the target-task corpus (Howard & Ruder, 2018), while Don't Stop Pretraining systematically established domain-adaptive and task-adaptive pretraining as effective mechanisms for domain specialization (Gururangan et al., 2020). In vision, Hierarchical PreTraining (HPT) showed that initializing target-domain self-supervised learning from an existing pretrained model can substantially accelerate convergence and improve downstream transfer across diverse visual domains (Reed et al., 2022).

This paradigm has subsequently been adopted in Earth observation. GFM transfers knowledge from natural-image pretraining to geospatial imagery through continued pretraining on GeoPile (Mendieta et al., 2023). ExPLoRA further introduces parameter-efficient self-supervised adaptation of natural-image-pretrained vision transformers to satellite imagery, unfreezing one or two pretrained transformer blocks while adapting the remaining layers with LoRA (Khanna et al., 2025). These works demonstrate that pretrained representations provide an effective basis for domain-specific learning, avoiding the cost of constructing a new foundation model from scratch.

Existing studies, however, largely address target-domain specialization or target-modality adaptation in isolation. It remains underexplored whether a web-trained RGB foundation model can be simultaneously specialized to Earth observation and endowed with multiple non-RGB sensing capabilities using only lightweight updates and small amounts of data. Web2Earth investigates this setting by combining RGB self-distillation with CMRD and CoPaD supervision from co-registered cross-sensor observations. This formulation transfers the semantic and spatial structure of the RGB teacher to non-RGB inputs while jointly improving representations in the original optical domain.

3 Method

Web2Earth adapts a web-pretrained encoder to optical remote sensing while extending its knowledge to non-RGB modalities (Figure 2). The central question is what to transfer: co-registered observa-

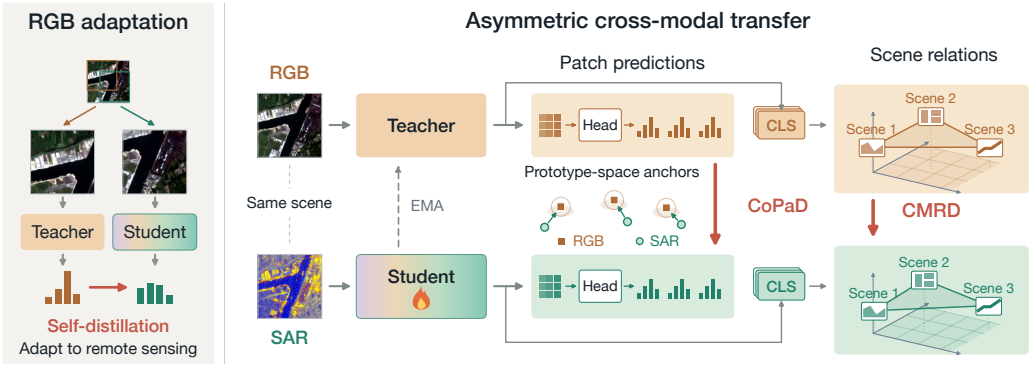


Figure 2: **Overview of Web2Earth.** RGB adaptation uses different crops of one image. Both pathways share the teacher (EMA updates only) and student (gradient updates). CMRD transfers relations among scene-level CLS embeddings within a batch, with one node per scene rather than per patch. CoPaD complements this scene-level supervision with patch predictions in prototype space.

tions depict the same scene, but their different sensing mechanisms need not yield identical embeddings. We therefore transfer the RGB teacher's scene relationships and local predictions, rather than directly equating feature vectors. This principle leads to scene-level relational supervision (CMRD), complemented by spatially aligned prototype supervision (CoPaD).

3.1 Asymmetric teacher–student transfer

Let $\{(\mathbf{x}_i^{\mathrm{rgb}}, \mathbf{x}_i^m)\}_{i=1}^B$ be B co-registered RGB–m pairs, where m denotes a non-RGB modality such as SAR. Each pair has N aligned patch positions. The teacher f_{t} and student f_{s} share the same web-pretrained initialization. Since this encoder was pretrained on RGB, the teacher's RGB outputs provide the reference for cross-modal transfer: $f_{\mathrm{t}}(\mathbf{x}_i^{\mathrm{rgb}})$ supplies stop-gradient targets for $f_{\mathrm{s}}(\mathbf{x}_i^m)$. We do not apply a reverse non-RGB-to-RGB objective. This asymmetry keeps RGB as the source of supervision instead of asking both modalities to reshape each other's targets. The teacher receives no gradients but evolves through an exponential moving average of the student.

RGB self-distillation. The direction of cross-modal transfer does not exclude RGB from student training. An independently sampled RGB batch adapts both encoders to optical remote sensing using augmented views of each image. DINO (Caron et al., 2021) supplies image-level self-distillation. For patch-level supervision, we use iBOT++ (Cao et al., 2026), an extension of iBOT (Zhou et al., 2022) covering masked and unmasked tokens. Thus, the RGB reference itself adapts to the target domain while providing supervision for the non-RGB student.

3.2 From scene relations to dense predictions

Cross-Modal Relational Distillation (CMRD). We first transfer how the RGB teacher relates different scenes. Rather than regressing each non-RGB embedding onto its paired RGB embedding, CMRD matches the similarities among scenes within a batch. Each RGB observation supplies a relational context for its paired non-RGB observation: its similarity to the other RGB scenes becomes a target for the corresponding non-RGB similarities. The supervision therefore comes from the teacher's organization of the batch, not a prescribed feature vector for each individual scene. Let $\mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^d$ be the unit-normalized CLS embeddings of $f_{\mathrm{s}}(\mathbf{x}_i^m)$ and $f_{\mathrm{t}}(\mathbf{x}_i^{\mathrm{rgb}})$, respectively, without a projection head. Their cosine similarities form student and teacher matrices $\mathbf{S}, \mathbf{T} \in \mathbb{R}^{B\times B}$:

$$S_{ij} = \mathbf{u}_i^\top \mathbf{u}_j, \qquad T_{ij} = \mathbf{v}_i^\top \mathbf{v}_j. \tag{1}$$

The relational objective is

$$\mathcal{L}_{\mathrm{CMRD}} = \frac{1}{B(B-1)} \sum_{i\neq j} \rho(S_{ij} - \mathrm{sg}(T_{ij})), \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator and ρ is the unit-threshold Huber penalty, reducing sensitivity to large pairwise discrepancies relative to squared error. The sum covers ordered pairs of distinct scenes; diagonal entries are excluded because self-similarity is always one. Crucially, a common orthogonal transformation of the student embeddings leaves $\mathbf{S}$ unchanged. CMRD thus constrains scene geometry without fixing the student's embedding coordinates to those of the teacher.

Co-Registered Patch Distillation (CoPaD). Although CMRD aligns scene-level relations, this alone does not guarantee spatially consistent representations at corresponding regions. CoPaD complements this global supervision with region-level targets at aligned patch positions. Co-registration and identical spatial preprocessing provide this correspondence without a learned patch-matching step. To avoid imposing the RGB feature coordinates on each non-RGB patch, we project patch embeddings into a common prototype space and transfer the resulting predictions instead.

Let $\mathbf{p}_{i,n}, \mathbf{q}_{i,n} \in \mathbb{R}^K$ be the student and teacher distributions over K prototypes for patch n of scene pair i. They are obtained from the corresponding patch tokens through projection heads and temperature-scaled softmax, with centering of the teacher logits. We minimize cross-entropy with soft teacher targets:

$$\mathcal{L}_{\mathrm{CoPaD}} = -\frac{1}{BN} \sum_{i=1}^{B} \sum_{n=1}^{N} \mathbf{q}_{i,n}^{\top} \log \mathbf{p}_{i,n}. \tag{3}$$

All aligned patches contribute without masking in this cross-modal stream. Soft teacher targets retain relative probabilities over prototypes rather than assigning a single prototype to each region. Together, CMRD and CoPaD transfer complementary structure: relations between scenes and local predictions at corresponding regions. Section 4.5 compares this objective with direct cosine alignment of raw patch embeddings.

Total objective. We jointly optimize RGB self-distillation and the two cross-modal terms. These are concurrent training signals, not sequential stages: scene-level and patch-level transfer both accompany optical domain adaptation.

$$\mathcal{L} = \mathcal{L}_{\mathrm{DINO}} + \lambda_{\mathrm{ibot}} \mathcal{L}_{\mathrm{iBOT++}} + \lambda_{\mathrm{CMRD}} \mathcal{L}_{\mathrm{CMRD}} + \lambda_{\mathrm{CoPaD}} \mathcal{L}_{\mathrm{CoPaD}}. \tag{4}$$

Each auxiliary weight is calibrated against the DINO loss at its first training step and then fixed (Appendix B). At evaluation, the same adapted student processes RGB and non-RGB inputs, without requiring a paired RGB observation for a non-RGB image.

4 Experiments

We evaluate Web2Earth on optical tasks (Sec. 4.2), SAR transfer across datasets and unseen sensors (Sec. 4.3), and cross-scale and cross-modal retrieval (Sec. 4.4). We then analyze the effectiveness of our adaptation strategy through controlled ablations (Sec. 4.5).

4.1 Experimental setup

Baselines and training setup. Web2Earth uses web-pretrained DINOv3 ViT-L/16 as both its initialization and primary baseline. DINOv2 (Oquab et al., 2024) provides an additional web-pretrained reference, while DINOv3-SAT (Siméoni et al., 2026) provides a satellite-pretrained counterpart. The comparison also covers optical remote sensing models (Mendieta et al., 2023; Reed et al., 2023; Cong et al., 2022; Noman et al., 2024) and SAR or multimodal models (Liu et al., 2026; Fuller et al., 2023; Sumbul et al., 2025; Astruc et al., 2025; Xiong et al., 2024).

We train Web2Earth using 22K RGB images from xView2 (Gupta et al., 2019) for self-distillation and 5K paired Sentinel-1/2 observations from DFC2020 (Robinson et al., 2021) to strengthen cross-modal transfer. The RGB self-distillation data are separate from all evaluation datasets (Appendix C.1). With rank-16 LoRA (Hu et al., 2022), we update 6.29M backbone parameters (2.03%) and train 18.64M projection-head parameters that are discarded for evaluation. We train for 8,000 steps using AdamW. Training takes approximately one hour on a single NVIDIA B200 GPU. See Appendix B for training details.

Evaluation protocols. We freeze the encoder in all downstream evaluations and train only task-specific heads for linear evaluation. To probe the transferability of the frozen representations without task-specific learning, we also evaluate optical kNN classification, 1-shot in-context segmentation, and zero-shot change detection. In-context segmentation predicts a query mask given one annotated support image. We replace only the encoder in FROST (Park, 2026) and INSID3 (Cuttano et al., 2026), retaining their inference procedures, and report the mean of their scores. Zero-shot change detection applies change vector analysis (CVA) (Malila, 1980) to bi-temporal features.

Datasets. We evaluate on the following datasets, grouped by modality and task. Input configurations and evaluation details are provided in Appendix C.

Modality	Task	Datasets
Optical	Classification	DLRSD (Shao et al., 2018), RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), AID (Xia et al., 2017), fMoW-RGB (Christie et al., 2018)
	Segmentation	iSAID (Zamir et al., 2019), WHU (Ji et al., 2019), LoveDA (Wang et al., 2021), Vaihingen and Potsdam (Rottensteiner et al., 2012), LandCover.ai (Boguszewski et al., 2021), DeepGlobe (Demir et al., 2018)
	Change Detection	LEVIR-CD (Chen & Shi, 2020), WHU-CD (Ji et al., 2019), DSIFN (Zhang et al., 2020), EGY-BCD (Holail et al., 2023), SYSU-CD (Shi et al., 2022)
SAR	Classification	BigEarthNet-S1 (Sumbul et al., 2021), EuroSAT-S1 (Helber et al., 2019), So2Sat-S1 (Zhu et al., 2020)
	Segmentation	KuroSiwo (Bountos et al., 2024), SpaceNet6 (Shermeyer et al., 2020), Sen1Floods11 (Bonafilia et al., 2020), DFC2020-S1 (Robinson et al., 2021), OpenEarthMap-SAR (OEM-SAR) (Xia et al., 2025)

4.2 Optical evaluation

We first test whether joint optical and SAR adaptation improves optical performance relative to the original DINOv3 model. Table 1 reports classification, segmentation, and change detection results.

Table 1: **Performance on optical remote sensing tasks.** RS Pretrain Samples is the number of remote sensing images used for pretraining or adaptation; – denotes none and N/A an unavailable comparable count. Entries are dataset averages within each task. Web2Earth reports mean ± sd over five adaptation seeds; baselines use one checkpoint. Best and second-best results are bold and underlined.

Model	RS Pretrain Samples	Classification (Acc.)		Segmentation (mIoU)		Change Detection (F1)	
		kNN	Linear	In-context	Linear	Zero-shot	Linear
Remote sensing foundation models							
SatMAE	713K	69.5	80.1	25.0	56.6	28.9	69.1
GFM	600K	80.9	85.9	22.0	38.2	24.7	41.2
Scale-MAE	363.6K	70.3	81.5	27.4	55.9	30.2	69.3
DOFA	8.08M	78.6	85.4	26.8	48.3	29.3	66.0
SatMAE++	713K	68.7	81.0	31.0	59.0	32.2	70.8
AnySat	N/A	50.0	56.3	21.5	28.4	27.1	57.2
SMARTIES	496K	61.4	74.7	26.9	53.2	29.5	65.9
DINOv3-SAT	493M	64.2	76.9	38.3	63.8	28.0	<u>73.5</u>
Web-pretrained vision foundation models							
DINOv2	–	83.6	87.6	41.7	62.4	29.5	71.6
DINOv3	–	<u>85.8</u>	<u>88.5</u>	<u>52.2</u>	<u>66.7</u>	<u>33.1</u>	72.7
Web2Earth (Ours)	**32K**	**86.8 ± 0.3**	**89.3 ± 0.3**	**53.4 ± 0.7**	**68.0 ± 0.1**	**34.1 ± 0.2**	**74.6 ± 0.4**

Performance across optical tasks. Table 1 supports our central motivation: remote-sensing-specific pretraining does not necessarily yield consistently strong transfer across common downstream tasks. DINOv3-SAT, despite pretraining on 493M remote sensing images, remains below the

unadapted DINOv3 in linear classification (76.9 versus 88.5) and linear segmentation (63.8 versus 66.7). The broader set of remote sensing baselines exhibits the same pattern: individual models can be competitive on a particular task, but none matches the web-pretrained encoders across classification, segmentation, and change detection. In contrast, DINOv2 and DINOv3 provide consistently strong performance across all three task families.

Effect of lightweight adaptation. Web2Earth updates approximately 2% of the backbone parameters using only 32K images (22K RGB + 5K RGB–SAR pairs $\times 2$), improving on DINOv3 in all six summaries. Linear classification increases from 88.5 to 89.3, kNN classification from 85.8 to 86.8, linear segmentation from 66.7 to 68.0, and linear change detection from 72.7 to 74.6. Web2Earth consequently ranks first in all six evaluation settings in Table 1. Its gains under kNN classification, in-context segmentation, and zero-shot change detection further indicate that adaptation improves the representation itself rather than relying on a trained task head.

4.3 Cross-modal transfer to SAR

We next evaluate how effectively Web2Earth transfers to SAR across tasks and sensors. Table 2 evaluates three classification and five segmentation tasks with frozen encoders and linear heads. Web2Earth improves on DINOv3 on all eight tasks, raising the classification mean from 62.6 to 66.6 and the segmentation mean from 56.4 to 60.6.

Table 2: **SAR downstream performance under linear evaluation.** BEN-S1 and Sen1F11 abbreviate BigEarthNet-S1 and Sen1Floods11, respectively. BEN-S1 reports mAP, EuroSAT-S1 and So2Sat-S1 report accuracy, and segmentation tasks report mIoU, all in percent. Best and second-best results are shown in bold and underlined, respectively.

	Classification				Segmentation					
	Multi-label	Single-label		Overall						
Model	BEN-S1	EuroSAT-S1	So2Sat-S1	Avg.	KuroSiwo	Sen1F11	DFC2020	SpaceNet6	OEM-SAR	Avg.
Remote sensing foundation models										
SARMAE	64.6	70.4	28.5	54.5	73.2	44.2	32.9	45.4	11.5	41.4
SMARTIES	65.8	74.1	32.3	57.4	**83.5**	<u>73.4</u>	47.4	45.9	14.2	52.9
AnySat	63.7	74.6	32.8	57.0	63.8	63.4	40.6	45.4	16.5	45.9
CROMA	**73.7**	**86.9**	<u>39.1</u>	**66.6**	80.1	68.0	49.8	45.8	14.1	51.6
DINOv3-SAT	66.2	81.1	38.3	61.9	80.2	68.4	<u>56.6</u>	56.5	**29.4**	<u>58.2</u>
Web-pretrained vision foundation models										
DINOv2	67.3	83.0	37.7	<u>62.7</u>	74.6	61.7	51.0	<u>58.1</u>	26.5	54.4
DINOv3	67.4	83.3	37.1	62.6	79.0	66.7	54.7	55.3	26.2	56.4
Web2Earth (Ours)	<u>72.5</u>	<u>86.5</u>	**40.9**	**66.6**	<u>82.0</u>	**73.6**	**60.0**	**58.9**	<u>28.7</u>	**60.6**

SAR representation quality. Web2Earth consistently improves on DINOv3 across Sentinel-1 classification and segmentation, with the largest gain on Sen1Floods11 (66.7 to 73.6 mIoU). The comparison includes SAR and multimodal foundation models pretrained with SAR imagery (Liu et al., 2026; Fuller et al., 2023; Sumbul et al., 2025; Astruc et al., 2025). Web2Earth matches CROMA for the highest classification mean (66.6). While CROMA leads on BEN-S1 and EuroSAT-S1, Web2Earth ranks first on So2Sat-S1. SMARTIES leads on KuroSiwo and remains competitive on Sen1Floods11. Both CROMA and SMARTIES were pretrained with Sentinel-1 imagery. Alongside its joint-best classification mean, Web2Earth achieves the highest overall segmentation mean (60.6), demonstrating strong performance across both scene-level and dense prediction tasks.

Evaluation on unseen SAR sensors. Beyond Sentinel-1, we evaluate sensors absent during adaptation using SpaceNet6's Capella X-band SAR (Shermeyer et al., 2020) and OEM-SAR's Umbra Spotlight imagery (Xia et al., 2025). The strong Sentinel-1 performance of CROMA and SMARTIES does not carry over to these benchmarks: they achieve 45.8–45.9 mIoU on SpaceNet6 and 14.1–14.2 on OEM-SAR, both below the unadapted DINOv3 (Table 2). In contrast, Web2Earth improves on DINOv3 from 55.3 to 58.9 and from 26.2 to 28.7, respectively, ranking first and second among the compared models. These results highlight a contrast between performance on familiar and unseen sensors.

4.4 SCENE CORRESPONDENCE ACROSS SCALES AND MODALITIES

Connecting observations across satellites requires recognizing the same scene despite changes in resolution or modality. We test this through *cross-scale retrieval* of optical pairs with a $4\times$ difference in ground sampling distance (GSD), and *cross-modal retrieval* from RGB to SAR. Unlike class prediction, these evaluations require identifying the corresponding scene across different observations.

Table 3: **Cross-scale and cross-modal retrieval.** Recall@K (%). Cross-scale averages iSAID, LandCover.ai, and Vaihingen at $4\times$ GSD; cross-modal matches RGB to SAR. Best/second-best results are bold/underlined.

	Cross-scale ($4\times$)			RGB→SAR		
Model	R@1	R@5	R@10	R@1	R@5	R@10
GFM	87.8	97.4	<u>99.4</u>	<u>11.1</u>	<u>21.5</u>	<u>26.7</u>
DOFA	<u>91.8</u>	<u>98.1</u>	98.8	5.2	12.8	17.5
SatMAE++	78.0	91.2	95.3	0.8	2.0	3.2
DINOv3	86.9	94.8	98.1	9.5	20.1	25.2
DINOv3-SAT	58.8	78.6	85.2	1.9	6.9	11.0
Web2Earth	**97.0**	**99.6**	**99.6**	**34.4**	**52.7**	**61.3**

Web2Earth leads both settings (Table 3). For cross-scale retrieval, average R@1 rises from DINOv3's 86.9% to 97.0%, also exceeding the strongest compared baseline, DOFA (91.8%), by 5.2 points. This improvement indicates stronger scene matching despite the loss of spatial detail at coarser resolution.

The margin is especially pronounced for RGB-to-SAR retrieval: Web2Earth reaches 34.4% R@1, compared with 9.5% for DINOv3 and 11.1% for the strongest baseline, GFM. The advantage persists at R@5 and R@10, where Web2Earth reaches 52.7% and 61.3%, versus GFM's 21.5% and 26.7%. Together, these results demonstrate improved scene correspondence across scales and modalities, beyond the gains in classification and segmentation.

4.5 WHAT DRIVES THE IMPROVEMENT?

We evaluate two aspects of Web2Earth: whether its gains go beyond simply continuing pretraining on RGB and SAR images, and how CMRD and CoPaD contribute individually and together (Table 4). The continual-pretraining baseline uses the same RGB+SAR data and the original DINOv3 objective without exploiting RGB–SAR pairing, isolating the benefit of additional data from that of cross-modal supervision. Optical classification is averaged over DLRSD, RESISC45, EuroSAT, and AID.

As shown in Table 4, continual pretraining alone does not improve the two modalities jointly. SAR classification rises from 62.6 to 65.0 and segmentation from 56.4 to 58.4, but optical classification remains at 96.0 and segmentation falls from 66.7 to 63.9. Additional SAR data therefore benefits SAR without a corresponding optical gain.

Table 4: **Comparison of adaptation strategies.** Continual pretraining: same RGB+SAR data and DINOv3 objective, no pairing or cross-modal supervision. Parentheses: changes from DINOv3.

Adaptation strategy	Optical		SAR	
	Cls.	Seg.	Cls.	Seg.
DINOv3	96.0	66.7	62.6	56.4
Continual pretraining	96.0 (0.0)	63.9 (-2.8)	65.0 (+2.4)	58.4 (+2.0)
+ CMRD	96.8 (+0.8)	67.9 (+1.2)	65.8 (+3.2)	59.4 (+3.0)
+ CoPaD	96.8 (+0.8)	67.6 (+0.9)	65.2 (+2.6)	60.2 (+3.8)
Web2Earth	**97.0** (+1.0)	**68.0** (+1.3)	**66.6** (+4.0)	**60.6** (+4.2)

In contrast, both cross-modal objectives individually improve all four evaluation groups over DINOv3 and continual pretraining. The individual objectives show complementary strengths. CMRD yields higher SAR classification than CoPaD (65.8 versus 65.2), whereas CoPaD yields higher SAR segmentation (60.2 versus 59.4). This pattern is consistent with CMRD transferring scene-level relations and CoPaD supplying spatially corresponding patch supervision. Their benefits are not task-exclusive, as both improve classification and segmentation in both modalities.

Combining the objectives gives the highest scores in all four groups: 97.0 and 68.0 for optical classification and segmentation, and 66.6 and 60.6 for SAR. Relative to continual pretraining, the gains are 1.0, 4.1, 1.6, and 2.2 points, respectively. These results support complementary cross-modal supervision beyond simply adding SAR images to the pretraining data.

Choice of patch-level supervision. We test whether directly encouraging similar RGB and SAR patch embeddings is sufficient for effective cross-modal transfer. Table 5 compares raw embedding alignment, which maximizes cosine similarity between corresponding patch tokens without projection heads, with CoPaD, which matches the RGB teacher's local prototype predictions through projection heads.

Table 5: **Choice of patch-level supervision.** Raw embedding alignment matches raw patch tokens using cosine similarity; CoPaD matches prototype predictions through projection heads. Parentheses show changes from DINOv3.

Method	Optical		SAR	
	Cls.	Seg.	Cls.	Seg.
DINOv3	96.0	66.7	62.6	56.4
Raw embedding alignment	96.4 (+0.4)	**68.0** (+1.3)	60.3 (-2.3)	54.7 (-1.7)
CoPaD	**96.8** (+0.8)	67.6 (+0.9)	**65.2** (+2.6)	**60.2** (+3.8)

Raw embedding alignment improves optical classification and segmentation over DINOv3, but reduces SAR classification from 62.6 to 60.3 and segmentation from 56.4 to 54.7. Compared with CoPaD, its optical scores remain close (96.4 versus 96.8 for classification and 68.0 versus 67.6 for segmentation), while SAR classification and segmentation are lower by 4.9 and 5.5 points, respectively. Directly enforcing cross-sensor embedding similarity therefore does not necessarily yield better transferable features, even when optical performance improves. This contrast supports CoPaD's design: it transfers the teacher's local predictions rather than explicitly pulling raw RGB and SAR patch vectors together.

4.6 Qualitative analysis

Figure 3 examines optical–SAR consistency through PCA maps of a coastal scene and a river crossed by a bridge, comparing DINOv3, RGB+SAR continual pretraining, and Web2Earth. See Appendix E for PCA visualization details and additional examples.

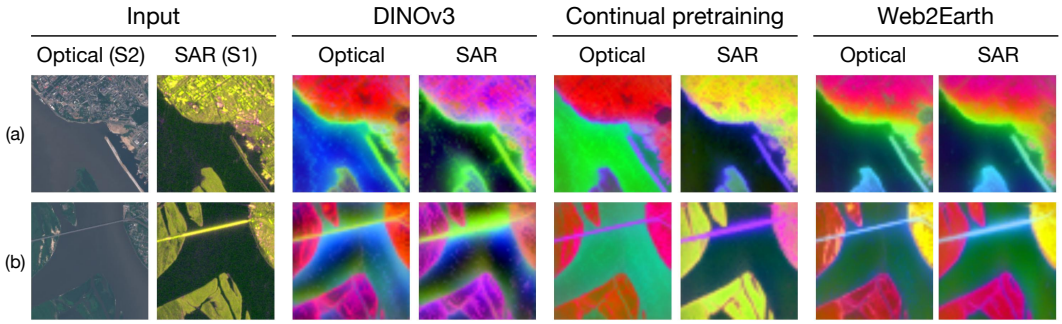


Figure 3: **PCA visualization of optical and SAR features.** (Best viewed in color.)

Web2Earth shows more consistent optical–SAR PCA patterns in these examples. In row (a), water and coastal regions have similar color assignments across sensors, unlike the pronounced shifts under continual pretraining. Row (b) shows a similar contrast around the bridge and river islands. These observations provide qualitative support for the intended effect of our cross-modal objectives. By transferring scene-level relations and local predictions, the proposed supervision appears to encourage consistent representations of corresponding scene content across sensing modalities.

5 Conclusion

In this work, we presented Web2Earth, a lightweight post-pretraining framework for extending web-pretrained vision models to remote sensing across modalities. It combines RGB self-distillation with asymmetric cross-modal transfer through CMRD and CoPaD, transferring scene-level relations and local predictions rather than directly equating sensor embeddings. With limited additional data and updates to approximately 2% of the backbone parameters, Web2Earth improves optical and SAR representations over DINOv3 across diverse downstream tasks. The pronounced gains in cross-scale optical and RGB-to-SAR retrieval further demonstrate improved correspondence between different observations of the same scene. Together, these findings show that selective knowledge transfer can add cross-modal capabilities while strengthening the model's optical performance.

AI USE STATEMENT

The authors used ChatGPT solely for language editing, manuscript drafting, improving readability, and literature retrieval and discovery (e.g., finding related work). These tools were not used to generate data, formulate mathematical claims, design the methodology or experiments, implement the proposed method, or interpret the experimental results. All AI-assisted content was independently reviewed and revised by the authors, who take full responsibility for the final manuscript, experimental findings, and conclusions.

REPRODUCIBILITY STATEMENT

Section 3 defines the training objectives and Section 4.1 summarizes the experimental setup. Appendix B provides the backbone configuration, trainable parameters, optimization settings, input preprocessing, augmentation procedures, and loss balancing. Appendix C describes the data sources, downstream metrics, and evaluation protocols. Appendix D reports dataset-level optical results to complement the aggregate comparisons in the main text.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loic Landrieu. AnySat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19530–19540, 2025.

Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. LandCover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1102–1110, 2021.

Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 210–211, 2020.

Nikolaos Ioannis Bountos, Maria Sdraka, Angelos Zavras, Ilektra Karasante, Andreas Karavias, Themistocles Herekakis, Angeliki Thanasou, Dimitrios Michail, and Ioannis Papoutsis. Kuro Siwo: 33 billion m^2 under the water. a global multi-temporal satellite dataset for rapid flood mapping. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and André Araujo. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29325–29335, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020. doi: 10.3390/rs12101662.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018. doi: 10.1109/CVPR.2018.00646.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, volume 35, pp. 197–211, 2022. doi: 10.52202/068431-0015.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. *arXiv preprint arXiv:2603.28480*, 2026.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 172–181, 2018. doi: 10.1109/CVPRW.2018.00031.

Anthony Fuller, Koreen Millard, and James Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5506–5538, 2023. doi: 10.52202/075280-0241.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27672–27683, 2024.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Shimaa Holail, Tamer Saleh, Xiongwu Xiao, and Deren Li. AFDE-Net: Building change detection using attention-based feature differential enhancement for satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. doi: 10.1109/LGRS.2023.3283505.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1031.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore,

Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023. doi: 10.48550/arXiv.2310.18660.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. doi: 10.1109/TGRS.2018.2858817.

Samar Khanna, Medhanie Irgau, David B. Lobell, and Stefano Ermon. ExPLoRA: Parameter-efficient extended pre-training to adapt vision transformers under domain shifts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29799–29818. PMLR, 2025.

Danxu Liu, Di Wang, Hebaixu Wang, Haoyang Chen, Wentao Jiang, Yilin Cheng, Haonan Guo, Wei Cui, and Jing Zhang. SARMAE: Masked autoencoder for SAR representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

William A. Malila. Change vector analysis: An approach for detecting forest changes with Landsat. In *LARS Symposia*, 1980.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16806–16816, 2023.

Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27811–27819, 2024. doi: 10.1109/CVPR52733.2024.02627.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Junghwan Park. FROST: Training-free few-shot segmentation with frozen features and nonparametric statistics. *arXiv preprint arXiv:2606.31136*, 2026.

Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, Kurt Keutzer, and Trevor Darrell. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2584–2594, 2022.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, Changlin Xiao, Michael Schmitt, Pedram Ghamisi, Ronny Hänsch, and Naoto Yokoya. Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3185–3199, 2021. doi: 10.1109/JSTARS.2021.3063849.

Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3:293–298, 2012. doi: 10.5194/isprsannals-I-3-293-2012.

Zhenfeng Shao, Ke Yang, and Weixun Zhou. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sensing*, 10(6):964, 2018. doi: 10.3390/rs10060964.

Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hänsch, Alexei Bastidas, Scott Soenen, Todd Bacastow, and Ryan Lewis. SpaceNet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 196–197, 2020.

Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. doi: 10.1109/TGRS.2021.3085870.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mário Caetano, Begüm Demir, and Volker Markl. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021. doi: 10.1109/MGRS.2021.3089174.

Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. SMARTIES: Spectrum-aware multi-sensor auto-encoder for remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5569–5578, 2025.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. doi: 10.1109/TGRS.2017.2685945.

Junshi Xia, Hongruixuan Chen, Clifford Broni-Bediako, Yimin Wei, Jian Song, and Naoto Yokoya. OpenEarthMap-SAR: A benchmark synthetic aperture radar dataset for global high-resolution land cover mapping [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 13(4):476–487, 2025. doi: 10.1109/MGRS.2025.3599512.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for Earth observation. *arXiv preprint arXiv:2403.15356*, 2024. doi: 10.48550/arXiv.2403.15356.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–37, 2019.

Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020. doi: 10.1016/j.isprsjprs.2020.06.003.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias H"aberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020. doi: 10.1109/MGRS.2020.2964708.

A LIMITATIONS AND FUTURE WORK

Our experiments evaluate Web2Earth using DINOv3 ViT-L/16 as the web-pretrained initialization. Although the evaluation spans diverse downstream tasks and SAR sensors, whether the same gains hold across backbone families and model sizes remains an open question. Studying additional initializations would help clarify how the quality and structure of the RGB teacher's representations affect cross-modal transfer, and how much adaptation different backbones require.

Another direction is to extend the framework beyond optical–SAR transfer to multispectral and hyperspectral observations. The current three-channel input configuration does not establish transfer to full-band inputs. Such extensions should investigate input mappings that retain the additional spectral information while remaining compatible with a web-pretrained encoder, together with suitable scene-level and local supervision. Evaluation should jointly assess performance in the new modalities, cross-modal correspondence, and retention of optical capabilities, rather than considering target-modality gains in isolation.

B TRAINING DETAILS

B.1 BACKBONE AND OPTIMIZATION

We initialize Web2Earth from the web-pretrained (LVD) DINOv3 ViT-L/16 checkpoint, with 16×16 patches and embedding dimension 1024. Rank-16 LoRA (Hu et al., 2022) adapters are applied to the joint query–key–value projection, the attention output projection, and both feed-forward linear layers of every transformer block. The final LayerNorm and mask token are also trainable; all other backbone parameters, including the patch embedding and register tokens, remain frozen. LoRA uses $\alpha = 32$, giving a scale of $\alpha/r = 2$; its factors are initialized as $A \sim \mathcal{N}(0, 0.02^2)$ and $B = 0$.

The trainable backbone contains 6.29M parameters, or 2.03% of ViT-L. Projection heads add 18.64M trainable parameters, bringing the total optimized parameter count to approximately 24.93M. These heads are used only during self-supervised training and discarded for all downstream evaluations. Thus, the approximately 2% figure in the main text refers specifically to the backbone, not to all parameters optimized during training.

Setting	Value
Optimizer	AdamW, $\beta = (0.9, 0.95)$
Learning rate: LoRA / final LayerNorm	4×10^{-4}
Learning rate: projection heads	10^{-3}
Weight decay	0.04
Schedule	200-step linear warm-up, followed by cosine decay
Training duration	8,000 steps
Gradient clipping threshold	3.0
RGB self-distillation batch	12 images per step
Cross-modal batch	48 RGB–SAR pairs per step
Teacher parameter EMA momentum	0.996
Teacher-logit center momentum	0.9
Hardware and runtime	One NVIDIA B200 GPU; approximately one hour
Observed GPU memory usage	Approximately 44GB

B.2 INPUT STREAMS AND PREPROCESSING

RGB self-distillation and cross-modal transfer use independently sampled batches, so their images need not depict the same scenes at a given step. The two streams also use different input resolutions and augmentation procedures.

RGB self-distillation. Source images are resized to 512 pixels before cropping. Each image yields two 256×256 global views with crop-area ratios in $[0.4, 0.9]$, and four 112×112 local views with ratios in $[0.10, 0.35]$. The second global crop box is jittered by up to 25% relative to the first.

Brightness, contrast, and channel-gain augmentation use photometric strength 0.3. For ViT-L/16, the global and local views yield 16×16 and 7×7 patch grids, respectively.

Cross-modal transfer. The DFC2020 cache provides pixel-aligned 256×256 RGB–SAR tiles at 10m ground sampling distance. We load both modalities at the same tile index and resize each to 224×224 using bilinear interpolation. No random cropping, flipping, or rotation is applied, and the optical self-distillation crop boxes are not reused for this stream. Both modalities therefore retain the same spatial grid: CoPaD matches the 196 corresponding patch positions in their 14×14 grids one to one. During cross-modal training only, SAR inputs undergo channel-wise mean and standard-deviation matching to the paired RGB image. This changes image statistics without changing spatial alignment. No mixup is used in the cross-modal stream.

Evaluation preprocessing. The paired-RGB statistics matching is disabled for SAR downstream evaluation and RGB-to-SAR retrieval; SAR preprocessing does not use statistics from the corresponding RGB image in these evaluations. RGB-to-SAR retrieval uses 224×224 inputs, yielding 14×14 patch grids. Downstream input configurations are described in Appendix C.

B.3 Distillation targets and loss balancing

The teacher parameters are updated from the student by $\theta_{\mathrm{t}} \leftarrow 0.996\,\theta_{\mathrm{t}} + 0.004\,\theta_{\mathrm{s}}$, without gradient updates to the teacher. This parameter EMA is distinct from logit centering, which updates $c \leftarrow 0.9\,c + 0.1\,\bar{z}_{\mathrm{t}}$, where $\bar{z}_{\mathrm{t}}$ is the batch mean of teacher prototype logits. The center is subtracted from the teacher logits before the softmax.

CMRD and global relational alignment. CMRD matches similarities among scene-level CLS embeddings across the batch. Its unit-threshold Huber penalty is defined as $\rho(a) = a^2/2$ for $|a| < 1$ and $\rho(a) = |a| - 1/2$ otherwise. The teacher similarity matrix is treated as a stop-gradient target at each training step.

CoPaD targets and local alignment. CoPaD complements scene-level relations with predictions at corresponding patch positions. For cross-modal losses, the teacher processes RGB and the student processes the paired non-RGB observation; teacher predictions are stop-gradient targets. Patch heads use an MLP, ℓ_2 normalization, and a weight-normalized prototype layer. The student temperature is fixed at 0.1, while the teacher temperature increases from 0.03 to 0.05 with a cosine warm-up schedule. Teacher-logit centering uses the distinct momentum 0.9 specified above.

Balancing the objectives. The auxiliary losses have different raw scales, so we calibrate their initial contributions before joint optimization. For each auxiliary term $k \in \{\mathrm{ibot}, \mathrm{CMRD}, \mathrm{CoPaD}\}$, we set $\lambda_k = 0.5\,|\mathcal{L}_{\mathrm{DINO}}^{(0)}|/|\mathcal{L}_k^{(0)}|$, where (0) denotes evaluation at the term's first training step and $\mathcal{L}_{\mathrm{ibot}}$ denotes $\mathcal{L}_{\mathrm{iBOT++}}$. The resulting weights remain fixed for the rest of adaptation, preserving a stable trade-off between optical retention and cross-modal transfer.

C Evaluation protocols

Our evaluation is designed around two questions: whether adaptation preserves useful optical representations, and whether the same encoder transfers to SAR. To make both questions about representation quality rather than downstream fine-tuning capacity, the encoder is frozen in every experiment. Linear evaluation optimizes only a task-specific head, while kNN classification, in-context segmentation, and zero-shot change detection introduce no task-specific training.

C.1 Adaptation setup

Web2Earth is initialized from DINOv3-Web ViT-L. RGB self-distillation uses 22,068 images drawn exclusively from xView2 (xBD; Gupta et al., 2019). We choose this source to adapt the web-pretrained encoder to fine spatial detail in high-resolution overhead imagery, while keeping the RGB self-distillation set separate from all downstream evaluation datasets. No images from those datasets are included in this RGB set, and xView2 annotations are not used.

Cross-modal transfer uses 5,128 paired Sentinel-1/2 observations from DFC2020. We use the original DFC2020 test split (5,128 pairs) for self-supervised adaptation and reserve the original validation split (986 observations) for downstream evaluation. These names refer to the original contest splits; the 986 evaluation observations are excluded from adaptation. Adaptation images are therefore disjoint from downstream evaluation images. Counting both images in each RGB–SAR pair gives $22{,}068 + 5{,}128 \times 2 = 32{,}324$ images in total, rounded to 32K in the main text. The backbone and optimizer settings are provided in Appendix B.1; input sampling and preprocessing are detailed in Appendix B.2.

C.2 Dataset coverage and metrics

The evaluation progresses from image-level semantics to spatial and temporal reasoning. We begin with single-label optical classification on DLRSD, EuroSAT, AID, RESISC45, and fMoW, reporting accuracy for each dataset and their aggregate.

We next test whether the representation retains spatial detail. The optical segmentation suite comprises iSAID, WHU, LoveDA, Vaihingen, Potsdam, LandCover.ai, and DeepGlobe, with mIoU computed after tiled predictions are reassembled. Vaihingen uses its IRRG configuration. Change detection adds a temporal dimension through LEVIR-CD, WHU-CD, DSIFN, EGY-BCD, and SYSU-CD. We report F1 for the changed region as the positive class on a 0–100 scale, separately for zero-shot and linear evaluation.

Finally, we evaluate whether the optical-to-SAR transfer extends beyond a single task. BigEarthNet-S1 measures multi-label classification, EuroSAT-S1 measures land-use and land-cover classification, and So2Sat-S1 measures local climate-zone classification. KuroSiwo, SpaceNet6, Sen1Floods11, DFC2020-S1, and OEM-SAR measure segmentation. DFC2020-S1 and OEM-SAR denote the SAR-only subsets of their corresponding multimodal benchmarks.

C.3 Preprocessing and evaluation heads

Three-channel SAR inputs. For DINOv3 and Web2Earth, SAR observations are converted to three-channel images during dataset construction to match the pretrained encoder's input interface. This conversion is fixed, requires no learned input adapter, and is shared by the two models. For the dual-polarization Sentinel-1 datasets DFC2020, Sen1Floods11-S1, BigEarthNet-S1, and KuroSiwo, the channels are $[\mathrm{VV}, \mathrm{VH}, \mathrm{VV} - \mathrm{VH}]$. The third channel is computed by subtracting the raw VH measurement from the raw VV measurement in dB, before clipping or rescaling either channel. The three channels are independently clipped to $[-25, 0]$, $[-32, -5]$, and $[0, 25]$, respectively, and each is linearly rescaled to unsigned 8-bit values in $[0, 255]$. SpaceNet6 uses $[\mathrm{HH}, \mathrm{HV}, \mathrm{VV}]$ channels from Capella X-band SAR, with dB values clipped to $[-5, 20]$. For single-channel amplitude imagery such as OEM-SAR, the same channel is replicated three times. These input channels encode SAR measurements, not visible RGB colors. This fixed channel construction is distinct from the paired-RGB statistics matching used during cross-modal training (Appendix B.2). No paired RGB image is needed to construct SAR inputs at evaluation.

Spatial preprocessing and heads. Preprocessing follows the spatial structure of each task instead of forcing all datasets through a single input format. For classification, images are resized so that the shorter side is 256 pixels and then center-cropped to 224×224. This convention is shared by all five classification datasets.

Dense prediction preserves substantially more spatial information. Datasets provided at 256 or 512 pixels are evaluated without additional resizing. LoveDA, Vaihingen, and Potsdam use 512-pixel random training crops and 512-pixel sliding-window test tiles from 1024-pixel inputs; DeepGlobe is processed analogously from 2448-pixel inputs. The tile predictions are reassembled before mIoU is computed. SAR segmentation follows the same dense-prediction principle, with inputs resized to 448 pixels.

Change detection must preserve correspondence across time as well as space. LEVIR-CD is tiled from 1024 pixels to 512 pixels, WHU-CD and DSIFN use 512-pixel inputs, and EGY-BCD and SYSU-CD use 256-pixel inputs. The two temporal images and their target mask always share crop

coordinates. The linear head receives the two feature maps and their absolute difference, whereas zero-shot evaluation applies a per-image Otsu threshold to the last-layer cosine distance.

Although the input construction varies with the task, the representation being evaluated does not: the encoder is always frozen. Classification uses the global image representation, segmentation uses the final spatial feature map, and change detection uses the paired temporal descriptor described above. Table 6 summarizes the number of epochs used to train these lightweight heads; these budgets do not include additional encoder adaptation.

Table 6: **Downstream linear-evaluation budgets.** Epoch counts refer to training the task-specific head while keeping the encoder frozen.

Dataset group	Epochs
Optical classification	30
Optical segmentation	70
Change detection	100
SAR classification	30
SAR segmentation	70

C.4 Retrieval protocols

Retrieval evaluates whether frozen representations identify the same scene across changes in modality or spatial resolution. We use only the final-layer CLS token, apply ℓ_2 normalization, and rank gallery embeddings by cosine similarity. No projection head or retrieval-specific training is used. Recall@K is the percentage of queries whose corresponding tile appears among the top K retrieved candidates, with $K \in \{1, 5, 10\}$.

Cross-modal retrieval. We use all 986 co-registered tiles in the original DFC2020 validation split. This split is held out from adaptation, which draws paired observations exclusively from the original DFC2020 test split. For the RGB-to-SAR results in Table 3, each RGB tile is a query and all 986 SAR tiles form the gallery. The SAR observation with the same tile identity is the positive match. We compute the full 986×986 similarity matrix, so retrieval is evaluated against the entire held-out gallery rather than a smaller sampled minibatch. Inputs are resized to 224×224, and SAR preprocessing does not use statistics from the paired RGB image. The evaluation routine also supports the reverse direction and RGB–multispectral retrieval using false-color images, and computes median rank and cosine-similarity diagnostics. The paper reports RGB-to-SAR Recall@1, Recall@5, and Recall@10.

Cross-scale retrieval. We evaluate the test splits of iSAID, Vaihingen, and LandCover.ai. For each optical tile, we create a coarser observation by downsampling both spatial dimensions by a factor of four using area interpolation, adding Gaussian noise with standard deviation $\sigma = 0.02$, and upsampling to the original image dimensions using bilinear interpolation. This procedure simulates a $4\times$ increase in GSD while retaining the same field of view and tile identity. The resampling kernels differ from those used in the training augmentations, aiming to probe robustness beyond a degradation pattern encountered during adaptation. We match original and degraded tiles using their CLS embeddings, with the corresponding version of the same tile as the positive match. Retrieval is evaluated separately within each dataset, and Table 3 reports the unweighted mean across the three datasets for each of Recall@1, Recall@5, and Recall@10.

D Detailed downstream results

The aggregate scores in the main text establish the overall trend, but they do not show whether an improvement is broad or driven by a small number of datasets. We therefore expand the optical classification, segmentation, and change-detection results below. These views separately examine image-level semantics, spatial detail, and temporal comparison. All experimental results in this appendix are from single-seed runs. Web2Earth is evaluated using a single checkpoint, whereas Table 1

reports means over five adaptation seeds. The appendix averages therefore need not match the multi-seed means in the main table.

D.1 Optical classification

Table 7 reports single-seed classification results on DLRSD, RESISC45, EuroSAT, AID, and fMoW, the five datasets summarized in Table 1.

Table 7: **Per-dataset optical classification performance (%).** Each entry reports kNN / linear-probe accuracy. Best and second-best scores for each dataset and evaluation regime are bold and underlined, respectively.

Model	DLRSD kNN / Linear	RESISC45 kNN / Linear	EuroSAT kNN / Linear	AID kNN / Linear	fMoW kNN / Linear
SatMAE	60.7 / 81.9	73.2 / 85.0	92.2 / 96.6	80.4 / 89.5	41.1 / 47.5
GFM	92.6 / 95.5	86.7 / 91.7	92.9 / 96.8	87.8 / 95.2	44.5 / 50.6
Scale-MAE	58.8 / 82.9	77.6 / 86.7	79.9 / 87.7	88.0 / 94.8	46.9 / 55.6
DOFA	88.6 / 94.3	84.1 / 91.9	93.8 / 97.2	86.7 / 93.9	39.7 / 49.5
SatMAE++	58.1 / 84.0	77.4 / 87.7	92.0 / 96.6	80.8 / 91.3	35.0 / 45.1
AnySat	40.7 / 51.4	53.3 / 59.2	83.1 / 84.6	51.8 / 62.2	21.2 / 24.0
SMARTIES	51.0 / 72.6	66.7 / 81.3	89.1 / 93.9	68.5 / 84.6	31.5 / 41.1
DINOv3-SAT	46.4 / 64.5	77.1 / 87.5	93.0 / 97.3	62.2 / 85.0	42.3 / 50.3
DINOv2	91.4 / 96.7	90.5 / 93.7	94.3 / 97.2	92.5 / 95.0	49.3 / 55.5
DINOv3	95.0 / 95.5	92.6 / 94.8	95.5 / 97.2	93.4 / 96.5	**52.6 / 58.5**
Web2Earth	**95.2 / 96.9**	**94.4 / 95.4**	**96.8 / 98.2**	**94.5 / 97.3**	50.8 / 55.7

Web2Earth leads on DLRSD, RESISC45, EuroSAT, and AID under both evaluation regimes, while DINOv3 remains strongest on fMoW in both regimes. The aggregate improvement in Table 1 therefore reflects broad gains across the aerial scene benchmarks rather than uniform dominance on every dataset.

D.2 Optical segmentation

Table 8 reports single-seed optical segmentation results on the seven constituent datasets. This view allows us to distinguish a consistent shift from an isolated gain on a single benchmark.

Table 8: **Per-dataset optical segmentation performance (%).** Each entry reports linear-probe mIoU. The final column gives the one-decimal mean. Best and second-best scores in each column are bold and underlined, respectively.

Model	iSAID	WHU	LoveDA	Vaih.	Pots.	LC.ai	DeepG.	Mean
DINOv3	58.6	84.5	50.1	69.3	65.5	77.4	61.6	66.7
DINOv2	52.9	84.0	45.5	62.0	60.5	74.4	57.7	62.4
DINOv3-SAT	51.1	84.0	48.5	61.7	58.2	**79.5**	**63.8**	63.8
SatMAE	45.8	79.5	40.7	52.1	50.1	70.3	57.9	56.6
SatMAE++	49.4	83.3	41.3	56.5	52.6	72.6	57.9	59.0
Scale-MAE	43.7	**85.1**	42.4	51.5	48.1	62.3	58.6	55.9
DOFA	42.3	69.0	31.6	43.6	40.7	57.4	53.8	48.3
AnySat	11.0	45.5	15.8	21.8	24.3	38.6	41.9	28.4
GFM	30.4	52.0	28.0	30.9	31.5	47.4	47.8	38.2
SMARTIES	42.8	76.8	36.0	49.4	48.9	65.1	53.6	53.2
Web2Earth	**62.9**	84.9	**50.3**	**70.3**	**66.0**	78.6	63.1	**68.0**

Relative to the DINOv3 starting point, Web2Earth improves on every listed dataset, with the largest gain on iSAID (62.9 versus 58.6). The direction of improvement is therefore consistent across aerial scenes, building footprints, and land-cover benchmarks. Adaptation does not, however, make Web2Earth uniformly dominant over every specialized model: Scale-MAE remains strongest on

WHU, while DINOv3-SAT leads on LandCover.ai and DeepGlobe. The result is best read as broad improvement over the web-pretrained initialization rather than universal dominance.

D.3 CHANGE DETECTION

Segmentation evaluates spatial transfer within a single observation; change detection asks whether the same representation also supports comparison across time. Table 9 reports both zero-shot and linear regimes. The former relies directly on feature distance, whereas the latter measures how readily a lightweight head can separate changed from unchanged regions. Unlike the aggregate comparison in Table 1, this table exposes the contribution of each of the five change-detection datasets.

Table 9: **Per-dataset change detection performance (%).** Each entry reports zero-shot / linear-probe F1 for the changed region. Avg. is the unweighted mean across the five datasets. Best and second-best scores within each column and evaluation regime are bold and underlined, respectively.

Model	LEVIR-CD ZS / Linear	WHU-CD ZS / Linear	DSIFN-CD ZS / Linear	EGY-BCD ZS / Linear	SYSU-CD ZS / Linear	Avg. ZS / Linear
DINOv3	<u>23.5</u> / 70.6	17.2 / <u>93.3</u>	**49.3** / 58.8	<u>19.7</u> / 63.0	<u>55.7</u> / <u>78.0</u>	<u>33.1</u> / 72.7
DINOv2	17.4 / 70.2	23.4 / 92.4	39.6 / <u>59.6</u>	17.2 / 62.5	49.9 / 73.1	29.5 / 71.6
DINOv3-SAT	17.6 / <u>76.0</u>	17.7 / 92.4	39.1 / 58.7	18.6 / **66.4**	46.9 / 74.0	28.0 / <u>73.5</u>
SatMAE	16.6 / 65.6	**32.4** / 88.5	38.7 / **60.3**	14.4 / 59.3	42.5 / 72.0	28.9 / 69.1
SatMAE++	17.4 / 68.3	29.5 / 91.1	43.7 / 59.1	19.6 / 62.3	51.0 / 73.4	32.2 / 70.8
Scale-MAE	17.0 / **77.8**	<u>29.7</u> / 89.2	39.6 / 55.1	15.1 / 50.2	49.4 / 73.9	30.2 / 69.3
DOFA	15.3 / 55.9	26.8 / 88.2	40.2 / 57.7	16.3 / 55.4	47.8 / 73.0	29.3 / 66.0
AnySat	12.0 / 48.6	27.6 / 82.0	36.3 / 50.6	15.0 / 40.7	44.4 / 64.2	27.1 / 57.2
GFM	14.5 / 23.6	22.3 / 58.0	28.9 / 37.0	14.9 / 33.1	42.9 / 54.5	24.7 / 41.2
SMARTIES	14.9 / 56.8	28.5 / 86.2	37.8 / **60.3**	<u>19.7</u> / 53.4	46.6 / 72.7	29.5 / 65.9
Web2Earth	**24.8** / 73.8	16.6 / **93.6**	<u>47.2</u> / 55.6	**20.1** / <u>66.0</u>	**61.0 / 79.5**	**33.9 / 73.7**

Web2Earth achieves the strongest average in both regimes, but the per-dataset view shows that the gains are not uniform. In zero-shot evaluation, it leads on LEVIR-CD, EGY-BCD, and SYSU-CD, while SatMAE leads on WHU-CD and DINOv3 on DSIFN-CD. Under linear probing, Web2Earth leads on WHU-CD and SYSU-CD; Scale-MAE is strongest on LEVIR-CD, SatMAE and SMARTIES tie on DSIFN-CD, and DINOv3-SAT leads on EGY-BCD. The aggregate improvement therefore reflects complementary gains across datasets rather than dominance on every benchmark.

E ADDITIONAL PCA VISUALIZATIONS

Figure 4 provides four selected scenes spanning coastal, river, linear-network, and urban structures. The coastal example also appears in Figure 3; the other three provide additional scene contexts. All panels retain the original feature-map colors.

PCA visualization protocol. We visualize final-layer patch tokens on paired DFC2020 validation scenes. For each model and scene, the per-position mean token is subtracted and each modality is centered before fitting. The first three principal components are then fit jointly on optical and SAR tokens, and the two modalities share one color normalization. Similar colors therefore indicate similar coordinates in this shared three-dimensional projection, not necessarily equality of the full feature vectors. Color comparisons are made between modalities within each model and scene, rather than across separately fitted PCA spaces for different models.

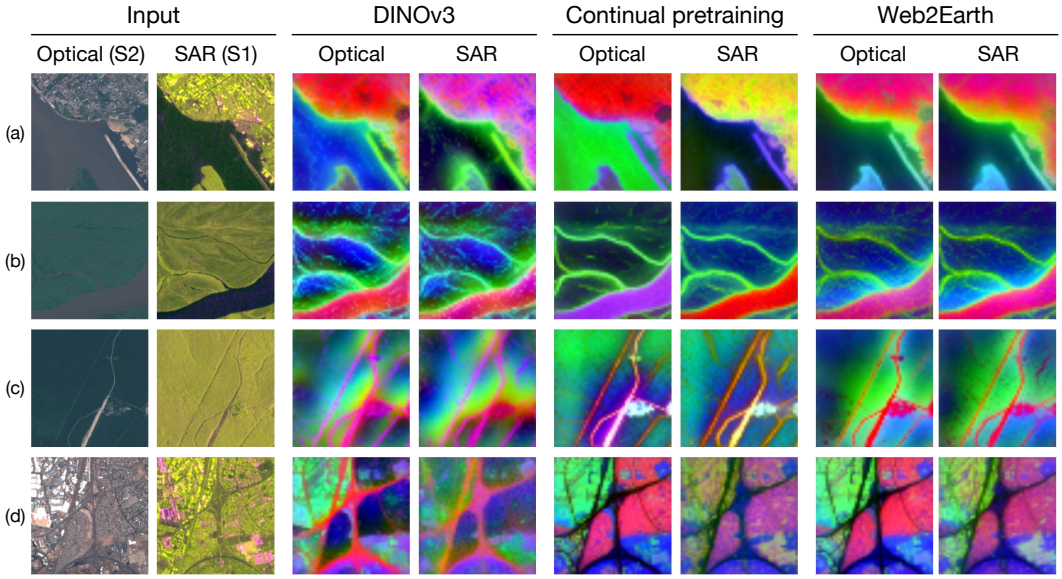


Figure 4: **Extended PCA visualization of optical and SAR features.** Paired DFC2020 validation scenes show (a) a coastal boundary, (b) river channels, (c) a branching linear network, and (d) urban regions. Row (a) also appears in Figure 3(a). PCA is fitted jointly on optical and SAR final-layer patch tokens for each model and scene, with shared color normalization after per-position mean subtraction and modality-wise centering. (Best viewed in color.)