CHANGING MODALITIES: ADAPTING REMOTE SENSING MODELS TO NEW SATELLITES AND SENSORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models for remote sensing are trained and deployed on a static set of modalities. However, as we equip newer satellites with novel sensors and retire old ones, practitioners may wish to deploy an existing model on a substitution, superset, or subset of modalities with minimal retraining given data availability or practical computational constraints. We study the setting of updating existing models to changing modalities and identify three main scenarios: Modality Transfer (substitution), Addition (superset), and Peeking (subset). We propose DeluluNet, an architecture with modular components for all three changing modality scenarios. DeluluNet is trained end-to-end, learning a multi-modal model from a unimodal teacher and *un*labeled multimodal data via modality hallucination—predicting missing modality representations from those that are present. As a result, DeluluNet can keep predicting even when input modalities change, providing a practical alternative to re-labeling and re-training in a changing world.

1 INTRODUCTION: MODALITIES CAN CHANGE, SO MODELS MUST CHANGE

Earth observation (EO) satellites continually collect vast amounts of *diverse* data across modalities (optical (Goward et al., 2017), radar (ESA, 2013), lidar (Markus et al., 2017)). The quantity and diversity of data make machine learning for remote sensing promising but challenging. Satellite missions are regularly updated—sensors added or retired, new missions launched at higher resolutions, and formerly free data paywalled. It is difficult to make use of these new modalities with existing remote sensing models without more training and annotations. This poses a major challenge since models fail on modalities *not* seen during training, e.g., a model trained on Sentinel-2's 12-band imagery cannot make predictions on the newest hyperspectral sensor's hundreds of bands (NRC, 2019). Ideally, a model trained on labeled data from existing satellites could be *adapted* to make use of the new modality, without labeling new data or training from scratch. We identify three scenarios under which models must adapt to new modalities:

1. *Modality Transfer*: switching a model from an old modality to a new one while maintaining predictions. For example, when Sentinel-2 (Drusch et al., 2012) became available, many sought to convert LandSAT (Goward et al., 2017) models to take advantage of its higher spatial, spectral, and temporal resolution.
2. *Modality Addition*: incorporating a new modality into a model's predictions. When Sentinel-1 SAR data became available (ESA, 2013), it was combined with existing optical imagery from LandSAT—a process that required re-labeling, but could be avoided with label-free modality addition.
3. *Modality Peeking*: improving a model's predictions on an old modality through limited observation of a new one. Global hyperspectral imagery (NRC, 2019) is rarely freely available, but is rich enough that a model could peek at it to learn fine-grained spectral patterns and leverage the knowledge acquired to improve predictions on standard optical data.

These changing modality settings have been relatively unexplored in the literature. Existing machine learning techniques have limited applicability: a given method may work for one of the three settings but not the others. A key contribution of our work is recognizing that transfer, addition, and peeking share similarities that are better served by a unified approach. We thus propose **DeluluNet**, an architecture with modular uni-modal and multi-modal components that adapts an existing predictive

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.

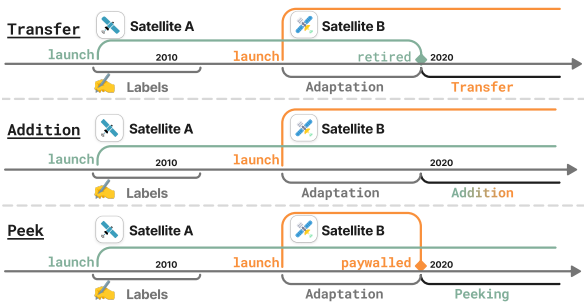

Figure 1: **Changing modalities:** When satellites are launched, paywalled, or retired, models must adapt to new sensor data without re-labeling. We consider three scenarios: *modality transfer* (predict using only new Satellite B data), *modality peeking* (predict using only old Satellite A data, informed by B), and *modality addition* (predict using both A and B).

model using unlabeled multi-modal data. DeluluNet leverages its multi-modal components even when test data is incomplete, imputing missing modality representations by predicting from the modalities that are present. We evaluate DeluluNet on two types of RS modality change: cross-band scenarios within Sentinel-2, and cross-satellite scenarios between Sentinel-1 and Sentinel-2. These experiments demonstrate its efficacy across transfer, addition, and peeking, covering changes in both channel bands and satellites, and all without requiring any labels on the new modality data.

2 Setting: Changing Modalities

Data splits and modalities. Consider a predictive task with labels $\mathcal{Y}$ and two possible input modalities: a *partially-labeled modality* $\mathcal{M}_A$ and an *unlabeled modality* $\mathcal{M}_B$. Data for learning is then divided into three splits:

1. **Labeled split** $\mathcal{D}_l \subset (\mathcal{M}_A \times \mathcal{Y})$ for the initial supervised training.
2. **Adaptation split** $\mathcal{D}_a \subset (\mathcal{M}_A \times \mathcal{M}_B)$ for unsupervised adaptation.
3. **Evaluation split** $\mathcal{D}_e$ contains either or both of $\mathcal{M}_A$ and $\mathcal{M}_B$, depending on the scenario.

Scenarios. Three main scenarios stem naturally from the changing modalities setting, each corresponding to a different evaluation split:

1. **Modality Transfer** $\mathcal{D}_e \subset \mathcal{M}_B$: The adapted model makes predictions using $\mathcal{M}_B$ alone, having transferred predictive ability from $\mathcal{M}_A$ to $\mathcal{M}_B$ without supervision.
2. **Modality Addition** $\mathcal{D}_e \subset (\mathcal{M}_A \cup \mathcal{M}_B)$: The adapted model makes predictions using $\mathcal{M}_A$ and $\mathcal{M}_B$ jointly, leveraging the two modalities jointly.
3. **Modality Peeking** $\mathcal{D}_e \subset \mathcal{M}_A$: The adapted model continues to make predictions using $\mathcal{M}_A$ alone, but with additional knowledge from observing unlabeled $\mathcal{M}_B$ during adaptation.

In all scenarios, we start with a uni-modal predictive model f_0 trained with supervision on the labeled split $\mathcal{D}_l$, adapt f_0 on $\mathcal{D}_a$ into a multi-modal DeluluNet, and evaluate on $\mathcal{D}_e$ (see Figure 1).

3 Background: Foundation Models, Scenarios, and Remote Sensing

Prior work adapts models on unlabeled data, but assumes a constant input modality. We review why existing models and methods are insufficient for the transfer, addition, and peeking scenarios.

Existing remote sensing foundation models cannot predict well on unseen modalities. To handle RS data from a variety of satellite modalities, RS Foundation Models (RSFMs) are pretrained on large multi-modal datasets and extract useful information across diverse downstream use cases (Lacoste et al., 2023). Most recently, the field has seen a wave of remote sensing foundation models (RSFMs) (Fuller et al., 2023; Xiong et al., 2024b; Tseng et al., 2025; Herzog et al., 2025; Labatie et al., 2025) following the success of language (Radford et al., 2021) and vision foundation models (Siméoni et al., 2025). These RSFMs are powerful and *multi-modal*, taking in various modalities of remote sensing, such as multi-spectral optical data (e.g. Sentinel-2, LandSAT 8&9) and Synthetic Aperture Radar data (e.g. Sentinel-1). Many such models utilize some modality-aware masked image

modeling objectives (Mizrahi et al., 2023) to enable usability when some modalities are absent (Cong et al., 2022; Herzog et al., 2025; Zhang et al., 2025), but they are constrained to a set of modalities seen during training. As a result, they require costly re-pretraining, relabeling, and re-finetuning when sensor modalities change (Tamazyan et al., 2025). While "sensor-agnostic" remote sensing models (Xiong et al., 2024a; Waldmann et al., 2025) are trained to handle arbitrary numbers of spectral channels at any wavelength, and thus are suitable for any satellite modality in theory, they become specialized to the labeled modality they are finetuned on. They cannot readily make use of an adaptation set to update a finetuned model given a new modality.

The three scenarios are scattered in different sub-fields of deep learning. Variants of the three changing modalities scenarios have been investigated in different sub-fields of deep learning. Below we discuss methods that, although not designed for changing modalities in general, can nevertheless be applied to one of the scenarios.

- *Knowledge Distillation* (Bucila et al., 2006; Hinton et al., 2015) studies transferring knowledge from one deep network to another, typically assuming a shared input modality between teacher and student. The transfer scenario we propose is an under-explored special case of knowledge transfer, where we transfer predictive power on the old modality to the new, unlabeled modality. The closest sub-branch of KD to our *modality transfer* setting is the cross-modality distillation (Gupta et al., 2016; Zhao et al., 2018; Zhou et al., 2023), which assumes labeled multimodal data, enabling a supervised hard loss term in addition to soft distillation loss.
- *Semi-Supervised Learning* (Semi-SL) (Blum and Mitchell, 1998) combines labeled data and unlabeled data, often leveraging one or several of entropy minimization (Grandvalet and Bengio, 2004), pseudo-labeling (Lee et al., 2013; Cascante-Bonilla et al., 2021) and consistency regularization (Sohn et al., 2020; Wang et al., 2023b; Berthelot et al., 2019). Semi-SL, similar to *modality peeking*, learns from labeled and unlabeled data splits, but assumes that labeled and unlabeled data come from the same modality.
- *Multimodal Knowledge Expansion* (MKE) (Xue et al., 2021) sits on the intersection of knowledge distillation and Semi-SL, and corresponds to our modality addition setting, learning a multimodal student from a unimodal teacher. Our setting is more general by additionally considering transfer and peeking scenarios that flexibly predict with any subsets of the considered modalities.

No existing method addresses all three scenarios jointly. Each sub-field has developed solutions in isolation, and none are designed for the label-free setting of changing modalities that arises naturally in remote sensing. DeluluNet bridges this gap with a single unified framework.

Adjacent fields assume fixed modalities or multimodal labels. *Continual Learning* (McCloskey and Cohen, 1989; Van de Ven et al., 2022; Lu et al., 2025) and *Domain Adaptation* (Ganin et al., 2016; Sun and Saenko, 2016; Jaritz et al., 2020; Munro and Damen, 2020; Hafner et al., 2022) handle new tasks or shifting data distributions, but assume the input modality stays fixed. *Learning with Privileged Information* (Vapnik and Vashist, 2009; Hoffman et al., 2016; Ma et al., 2021; Wang et al., 2023a) leverages modalities available only at training time, but requires *labeled* multimodal data and typically trains from scratch rather than building on pretrained encoders. We discuss how the changing modality setting differs from these fields in greater details in Appendix A.

Modality change is broadly relevant in applied machine learning—especially in remote sensing. This work focuses on changing modalities for remote sensing, where it is relevant due to the abundance of unlabeled multi-modal data from orbiting satellites that continually capture data. However, the changing modalities setting is relevant in many applied machine learning fields. Modality transfer could be useful for adapting existing models to new medical devices (Phillips et al., 2020), modality addition is useful for adding new sensors for autonomous driving (Hegde et al., 2025), and modality peeking could leverage high-cost imaging such as CT scans during training and still deploy with only more accessible alternatives like chest X-rays in resource-constrained settings (Cao et al., 2024).

4 Method: Modality Hallucination (DeLulu is the Solulu)

DeluluNet operates under the changing-modality setting by augmenting an existing uni-modal model with modular components for inter- and intra-modality learning, handling transfer, addition, and peeking within a single end-to-end framework. Starting from a ViT (Dosovitskiy et al., 2021) on

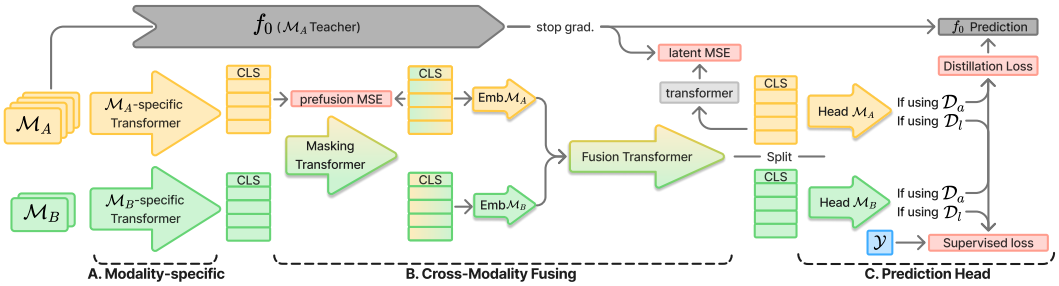

Figure 2: **DeluluNet Architecture and Training Strategy.** *A.* Shallow modality-specific transformers extract low-level information from available modalities; *B.* Masking and fusion transformers facilitate cross-modality fusion through masked image modeling, and enables prediction with incomplete input modalities at test time; *C.* Predictor heads learning switches between distilling from uni-modal f_0 teacher on $\mathcal{D}_a$ and learning from real labels from $\mathcal{D}_l$.

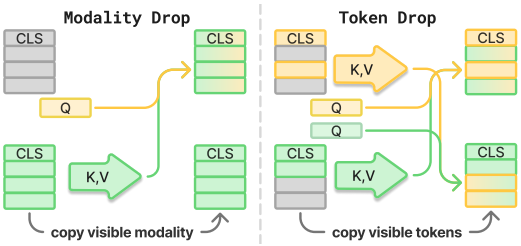

Figure 3: **Masking transformer cross-attention under modality-drop (left) and token-drop (right).** In both cases, each modality's learned query Q attends to visible tokens from the other modality as K, V to predict masked tokens. Transfer and peeking at test time naturally follow the modality-drop path, while addition requires no masking transformer.

$\mathcal{M}_A$-input as the predictive model f_0 trained on labeled $\mathcal{D}_l$, DeluluNet introduces three modular components to support feature extraction and prediction with the new modality $\mathcal{M}_B$ (Figure 2):

1. *Modality-Specific Extraction:* a randomly-initialized $\mathcal{M}_B$-specific tokenizer and f_0-initialized transformer layers for modality-specific representation of $\mathcal{M}_B$.
2. *Cross-Modality Fusion:* randomly-initialized modality embeddings Emb$_{\mathcal{M}_B}$ and masking transformers that predict cross-modal features between $\mathcal{M}_A$ and $\mathcal{M}_B$.
3. *Predictor Head:* a task-specific Head$_{\mathcal{M}_B}$ that predicts from $\mathcal{M}_B$ representations.

At test time, when both modalities are available (addition), each passes through its respective extractor before cross-modality fusion. When only one modality is present (transfer or peeking), the masking transformers hallucinate intermediate representations for the missing modality, enabling the model to treat it as multi-modal input. We next describe how these components are trained from unlabeled $\mathcal{D}_a$.

4.1 Learning Features and Predictions From Unlabeled Modalities.

DeluluNet is trained with the same end-to-end strategy for all three scenarios, with losses at different network depths supervising each modular component.

Modality-specific feature learning. To extract low-level features tailored to each modality's characteristics, DeluluNet employs shallow modality-specific transformers: patch embedders that project each modality's inputs into a shared dimension d, followed by n_{ms} transformer blocks. This design is common among multimodal RSFMs, which vary in fusion depth from early-fusion ($n_{\text{ms}} = 0$) (Tseng et al., 2025; Herzog et al., 2025) to mid-fusion (Astruc et al., 2024) and dynamic fusion (Labatie et al., 2025); we fix $n_{\text{ms}} = 3$ and ablate fusion depth in Appendix B. The $\mathcal{M}_A$-specific components are initialized from f_0, while $\mathcal{M}_B$-specific components are randomly initialized.

Cross-modality hallucination via masking transformers. To facilitate cross-modal feature learning and prepare DeluluNet for missing modalities at test time, we introduce **masking transformers** inspired by masked image modeling (He et al., 2022; Cong et al., 2022; Tseng et al., 2025; Herzog et al., 2025) that hallucinate masked representations from visible ones, and cross-modality representation hallucination (Hoffman et al., 2016). For each mini-batch, we adopt masking at two granularities (Figure 3): *token-level masking* randomly masks tokens across all modalities, providing

cross-modality learning signals when both modalities are partially available; *modality-level masking* treats an entire modality as absent, preparing the model for the missing-modality inference. The masking transformer imputes masked tokens and modalities from visible ones via cross-attention: a learned per-modality query attends to visible tokens from other modalities, with one CLS and one patch token, each of dimension d. In forward passes, the patch token is broadcast to (n, d) and augmented with f_0's position encodings, keeping the design parameter-efficient. A PREFUSION MSE loss directly supervises this hallucination by encouraging predicted CLS and patch tokens to match the true tokens of the masked modality.

Modality fusion learning. To combine features across modalities and teach DeluluNet to leverage hallucinated representations, a fusion transformer operates on the concatenation of all modality tokens via standard transformer blocks, allowing tokens from different modalities to attend to each other. Since transformers are applicable to any sequence of tokens sharing a hidden dimension, this naturally accommodates both real and hallucinated representations, so the fusion transformer always receives a full multi-modal input regardless of which modalities are present at test time. To anchor fused representations in a task-relevant feature space, a LATENT MSE loss is applied between the fusion transformer output and f_0's frozen penultimate representations via a 1-layer transformer decoder.

Prediction head learning. To ensure accurate predictions that incorporate features from both modalities, we alternate mini-batches between the multi-modal unlabeled $\mathcal{D}_a$ and the uni-modal labeled dataset $\mathcal{D}_l$. As a result, DeluluNet receives two sources of supervision for predictor heads, a technique we refer to as **batch mixing**, made possible by the masking transformer:

- On batches from the *unlabeled* $\mathcal{D}_a$, DeluluNet applies token- and modality-level masking and distills from the predictions made by the uni-modal teacher f_0 using a DISTILLATION LOSS. The unlabeled batches allow learning from pseudo-labels using real multi-modal inputs.
- On batches from the *labeled* $\mathcal{D}_l$, DeluluNet treats these as inputs where $\mathcal{M}_B$ is masked, and a supervised CROSS-ENTROPY LOSS is applied to predictions. The labeled batches allow learning from real labels using pseudo multi-modal inputs from the masking transformer.

We ablate masking transformers, MSE losses, and batch mixing in Section 5.1, showing each is crucial for DeluluNet's performance across scenarios. Further architecture details are in Appendix B.

5 EXPERIMENTS ACROSS CHANGING BANDS AND SATELLITES

Datasets and bands. We evaluate DeluluNet across single-label classification, multi-label classification, and semantic segmentation with remote sensing benchmarks. We study both cross-band modality changes using channel groups within Sentinel 2 (e.g., RGB to Vegetation Red-Edge bands), and cross-satellite modality changes between Sentinel-1 and Sentinel-2 (e.g., Synthetic Aperture Radar data to Multi-Spectral Images). **EuroSAT** (Helber et al., 2019) is the canonical 10-way *classification* dataset with 13 spectral bands from the Sentinel-2 satellites. We select two of the major groups of bands from EuroSAT, namely RGB (B-02,B-03,B-04) and VRE (B-05, B-06, B-07) to evaluate DeluluNet's performance for cross-band modality change, and present more results on cross-band transfer, peeking, and addition in Appendix E. **reBEN** (Clasen et al., 2025) is the latest update of the classic BigEarthNet dataset (Sumbul et al., 2019), a large European multi-modal (2-band Sentinel-1 and 12-band Sentinel-2) dataset for *multi-label Land Cover Classification*. We use reBEN from Geo-Bench-2 (Simumba et al., 2025) to study DeluluNet's ability to adapt to an entirely new satellite without labels. **DFC2020** (Robinson et al., 2021) is a 8-class land-cover *semantic segmentation* benchmark, multi-modal with Sentinel-1 (2 bands) and Sentinel-2 (13 bands) data. We use it to study DeluluNet's ability to adapt dense prediction tasks. For all benchmarks, we randomly split the official training data in half: *labeled uni-modal* $\mathcal{D}_l$ and *unlabeled multi-modal* $\mathcal{D}_a$.

Architecture. We train supervised **ViT-Base** (Dosovitskiy et al., 2021) on the labeled $\mathcal{D}_l$, initialized with pre-trained *DINO v3* (Siméoni et al., 2025) weights for the modality-fusion blocks and uni-modal transformers. These supervised models serve as the initial uni-modal f_0 for DeluluNet. We contextualize the efficacy of DeluluNet by fully fine-tuning the recent and high-accuracy multi-modal RSFMs OlmoEarth (Base) (Herzog et al., 2025) and Panopticon (Base) (Waldmann et al., 2025) on labeled uni- and multi-modal data as oracles. These RSFMs represent the most competitive performance achievable when models are pre-trained with massive multi-modal data and fine-tuned with labeled multi-modal inputs.

Table 1: **Modality Transfer** results show that DeluluNet transfers more effectively than knowledge distillation (KD) and transformed teacher matching (TTM), and at times even surpasses supervised fine-tuned DINOv3 on M_B.

Dataset (metric)		reBEN (mAP)				DFC2020 (mIoU)				EuroSAT (Acc)
Start (M_A)		$S2_{rgb}$		S1	S2	$S2_{rgb}$		S1	S2	RGB
Transfer (M_B)		S1	$S2_{\neg rgb}$	S2	S1	S1	$S2_{\neg rgb}$	S2	S1	VRE
$f_0(M_A)$	f_0	63.4	63.4	49.2	62.4	59.8	59.8	48.0	67.0	98.6
Baselines	KD	49.3±0.2	59.3±0.6	46.5±0.8	51.9±0.6	47.1±0.7	58.7±0.2	48.2±0.6	**47.1±0.3**	88.5±0.2
	TTM	48.8±0.4	59.3±0.9	48.4±0.6	51.0±1.1	46.8±0.5	57.9±0.3	48.1±0.3	46.8±0.3	88.9±0.7
Ours	Delulu	**52.4±0.2**	**62.0±0.4**	**50.6±0.9**	**52.9±0.8**	**48.8±0.5**	**59.6±0.1**	**49.2±0.0**	46.7±0.2	**90.6±0.4**
Oracle (M_B)	DINOv3	49.2	59.2	62.4	49.2	48.0	66.4	67.0	48.0	90.7
	Panopticon	55.1	61.4	62.1	55.1	53.0	58.4	58.4	53.0	97.8
	OlmoEarth	54.1	69.6	65.1	54.1	52.7	63.2	61.7	52.7	96.4

Table 2: **Modality Addition** results show that DeluluNet is able to extract and fuse inputs from both $\mathcal{M}_A$ and $\mathcal{M}_B$ through modality addition, not only generally surpassing multimodal knowledge expansion (MKE) and DINO-init M_A f_0, but sometimes even supervised RSFMs.

Dataset (metric)		reBEN (mAP)				DFC2020 (mIoU)				EuroSAT (Acc)
Start (M_A)		$S2_{rgb}$		S1	S2	$S2_{rgb}$		S1	S2	RGB
New (M_B)		S1	$S2_{\neg rgb}$	S2	S1	S1	$S2_{\neg rgb}$	S2	S1	VRE
$f_0(M_A)$	f_0	63.4	63.4	**49.2**	62.4	59.8	59.8	48.0	**67.0**	98.6
Baselines	MKE	55.4±0.4	58.0±0.2	47.9±0.6	60.3±0.2	59.2±0.0	59.0±0.3	48.2±0.9	66.0±0.1	93.6±0.2
Ours	Delulu	**64.7±0.8**	**65.1±0.5**	47.2±0.3	**64.8±0.5**	**60.0±0.1**	**60.1±0.1**	**49.3±0.3**	65.1±0.1	**99.0±0.1**
Oracle (M_A+M_B)	DINOv3	65.3	65.1	61.5	61.5	63.4	66.6	68.4	68.4	98.6
	Panopticon	59.4	62.5	63.4	63.4	56.9	58.5	58.4	58.4	98.7
	OlmoEarth	63.2	69.5	64.6	64.6	62.3	63.1	62.8	62.8	98.3

Table 3: **Modality Peeking** results show that DeluluNet is able to learn from observations of paired unlabeled $\mathcal{M}_A$ and $\mathcal{M}_B$ data $\mathcal{D}_a$ to make better predictions on $\mathcal{M}_A$ alone, and generally outperform Semi-SL baselines.

Dataset (metric)		reBEN (mAP)				DFC2020 (mIoU)				EuroSAT (Acc)
Start (M_A)		$S2_{rgb}$		S1	S2	$S2_{rgb}$		S1	S2	RGB
New (M_B)		S1	$S2_{\neg rgb}$	S2	S1	S1	$S2_{\neg rgb}$	S2	S1	VRE
$f_0(M_A)$	f_0	63.4	63.4	**49.2**	62.4	59.8	59.8	48.0	**67.0**	98.6
Baselines	MixMatch	56.5±0.6	56.5±0.6	48.3±0.6	60.1±0.4	59.2±0.1	59.2±0.1	48.6±0.3	60.9±0.7	98.5±0.2
	FreeMatch	63.3±0.7	63.3±0.7	47.9±0.4	57.6±1.0	59.4±0.1	59.4±0.1	47.5±1.4	59.1±0.3	97.0±1.2
Ours	Delulu	**64.4±0.2**	**64.5±0.7**	48.1±0.5	**63.9±0.2**	**60.7±0.2**	**61.3±0.2**	**50.5±0.4**	64.1±0.1	**99.0±0.1**
Oracle (M_A)	Panopticon	60.0	60.0	55.1	62.1	56.1	56.1	53.0	58.4	98.4
	OlmoEarth	59.7	59.7	54.1	65.1	59.2	59.2	52.7	61.7	97.7

Optimization and hyper-parameter tuning. The main hyper-parameters to tune for DeluluNet include learning rate, loss-weighting, masking ratio (token and modality), and batch-mixing ratio. Since DeluluNet can handle all three scenarios with one end-to-end strategy, we separately tune masking ratio and batch-mixing ratio for the three scenarios on DFC2020 dataset with Sentinel-1 as $\mathcal{M}_A$ and $S2_{\neg rgb}$ as $\mathcal{M}_B$. Validation scoring is also label-free for multi-modal evaluation via agreement-based metric detailed in Appendix C. All DeluluNet experiments are trained with these hyper-parameters fixed, and only learning rate and loss-weighting are tuned per dataset and modality to avoid full per-dataset-architecture-M_A-M_B tuning. Even though better performance is achievable through independent hyper-parameter tuning for each dataset and $\mathcal{M}_A$-$\mathcal{M}_B$ pairing, this simple and cheap approach is already effective across datasets, modalities, and architectures.

Modality transfer results. We compare DeluluNet in modality transfer scenario to canonical knowledge distillation (KD) (Hinton et al., 2015) and the recent Transformed Teacher Matching (TTM) (Zheng and Yang, 2024) as label-free baselines, and additionally fine-tune fully supervised DINOv3, Panopticon, and OlmoEarth as label-oracles. Table 1 shows that DeluluNet not only generally transfers more effectively than KD and TTM in the changing modality setting, but often surpasses the DINOv3-init supervised label-oracle models. This is particularly true when the starting modality is RGB or the RGB bands from Sentinel-2—these are consistent with the DINOv3's RGB

pretraining, providing a stronger starting point for transfer. For example, when transferring from $S2_{rgb}$ to S1, DeluluNet is able to reach 52.4 mAP on reBEN and 48.8 mIoU on DFC2020, a respective $+3.2$ and $+0.8$ improvement compared to a label-oracle fine-tuned model, and in both cases reaching performance close to those of fine-tuned RSFMs.

Modality addition results. We compare DeluluNet in modality addition scenario to Multimodal Knowledge Expansion (MKE) (Xue et al., 2021), the natural baseline for modality addition which learns a multi-modal model from a uni-modal teacher, aiming to make better predictions. We found that while MKE can sometimes outperform the uni-modal f_0, DeluluNet provides more consistent and sizable improvements. Table 2 shows that across all modality settings we tested, DeluluNet almost always improves performance on f_0, suggesting that it is effective at improving predictive performance when new complementary modalities become available. For example, when adding Sentinel-1 and Sentinel-2's multi-spectral bands to a $S2_{rgb}$ f_0, DeluluNet is able to improve mAP on reBEN by 1.3% and 1.7% respectively, matching supervised DINOv3 oracle and even achieving performance comparable to supervised fine-tuned RSFM oracles.

Modality peeking results. We compare DeluluNet in modality peeking scenario to the popular Semi-SL methods MixMatch (Berthelot et al., 2019) and FreeMatch (Wang et al., 2023b). MixMatch trains a model on both labeled and unlabeled data using consistency regularization and pseudo labeling with various augmentations including Mixup (Zhang et al., 2018); and FreeMatch use a dynamic confidence threshold to simulate hard pseudo-labels. Table 3 shows that DeluluNet effectively learns from paired $\mathcal{M}_A, \mathcal{M}_B$ from $\mathcal{D}_a$ and improves on its uni-modal f_0 initialization. For example, DeluluNet initialized with an $S2_{rgb}$ f_0 after peeking at $S2_{\neg rgb}$ achieves a 1.5% improvement in mIoU on DFC2020, achieving performance close to RSFM oracles supervised with actual labeled multi-modal data.

5.1 Under The Hood of DeluluNet: Masking Transformer Analysis and Component Ablations

We analyze three key design choices of DeluluNet—*masking transformers*, *batch mixing*, and *auxiliary MSE loss terms*—to validate that each contributes to performance as intended.

Masking transformer analysis. Masking transformers facilitate cross-modality learning through modality hallucination during training, and enable DeluluNet to make accurate predictions by generating pseudo multi-modal representations even when only a subset of modalities is available during testing. Figure 4 shows that masking transformers impute representations of missing modalities effectively: hallucinated tokens occupy the same representational space as real ones (left) and reproduce their spatial structure (right). Masking transformers can hallucinate Sentinel-2 representations from Sentinel-1 input and vice versa, resulting in grounded pseudo multi-modal representations suitable for fusion. In Appendix D, a per-feature pearson correlation analysis confirms this quantitatively: hallucinated features correlate more strongly with those of the paired sample than with those of mis-paired samples, showing that the imputation is specific to the observed input. Table 4 shows that replacing masking transformers with per-modality mask tokens hurts DeluluNet on modality transfer and peeking tasks—the settings where masking transformers impute the missing modality. This verifies that accurate hallucinations of multi-modal intermediate feature maps are beneficial for downstream performance. Beyond hallucination quality, our ablation in Appendix B shows that an earlier fusion strategy is preferable to a late one, consistent with recent works in RSFMs (Tseng et al., 2025; Labatie et al., 2025), and further verifies that DeluluNet performs meaningful fusion using hallucinated multi-modal representations.

Batch mixing: learning from real label with pseudo data and from pseudo label with real data. Standard distillation methods like KD and TTM learn from teacher predictions (pseudo labels) with real input data when no labels are available. DeluluNet additionally learns from the unimodal labeled data when the input modality is incomplete. This is achievable for DeluluNet because of masking transformers' ability to impute missing modality with predicted intermediate feature maps, i.e. learn from real labels with pseudo inputs. The two directions of batch mixing serve distinct purposes: unlabeled batches from $\mathcal{D}_a$ ground the model in real multi-modal inputs, preventing it from overfitting to hallucinated representations; labeled batches from $\mathcal{D}_l$ provide clean ground truth supervision, preventing the model from overfitting to the teacher's prediction mistakes. Table 4 shows that batch mixing improves DeluluNet's performance across all three scenarios.

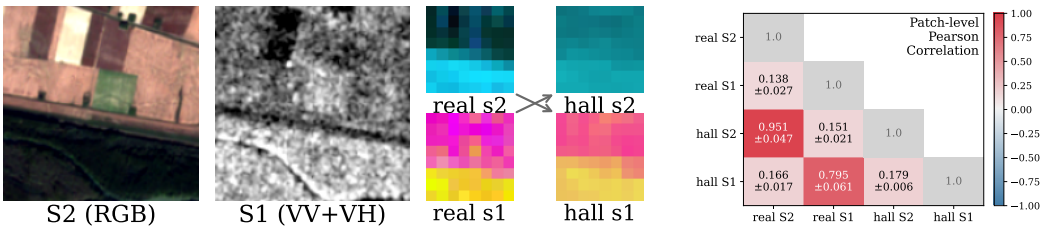

Figure 4: **(left) DeluluNet modality hallucinations (qualitative).** Top-3 PCA components visualized as RGB show high similarity between real and hallucinated feature maps for both S1→S2 and S2→S1. **(right) Patch-level Pearson Correlation between real and hallucinated patch representations (quantitative).** Despite low correlation between real S1 and S2 tokens, masking transformers achieve high linear correlation between real and hallucinated patch tokens across both modalities.

Configuration	Transfer	Peek	Addition
DeluluNet (ours)	51.8	49.2	50.9
+ Masking Token	50.7 (-1.1)	48.2 (-1.0)	50.8 (-0.1)
− Prefusion MSE	49.2 (-2.5)	48.3 (-0.9)	50.6 (-0.3)
− Latent MSE	48.8 (-3.0)	48.6 (-0.6)	49.7 (-1.2)
− Batch Mixing	50.9 (-0.8)	48.5 (-0.7)	48.8 (-2.1)

Table 4: **DeluluNet Components Ablation on reBEN** ($\mathcal{M}_A$:S1, $\mathcal{M}_B$:S2), deltas relative to DeluluNet. *+Masking token config* uses per-modality learned mask tokens to replace masking transformers; *−Prefusion MSE config* removes direct supervision for masking transformers to hallucinate cross-modality representations; *−Latent MSE config* removes latent reconstruction of the teacher; *−Batch Mixing config* removes labeled batch learning. Each component improves its intended objective; all configurations are hyperparameter tuned with 32 runs over a shared search space.

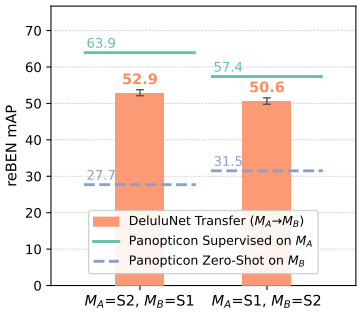

Figure 5: **Panopticon Zero-Shot vs. Delulu modality transfer.** DeluluNet takes advantage of widely available unlabeled paired multi-modal data, achieving significantly better transfer performance.

Model	Modality	Params (M)	GMACs	Max Batch Size	Throughput
DINOv3	s1 (2)	85.5	6.0	8192	1759
	s2 (12)	87.4	6.1	8192	1719
DeluluNet	s1 (2)	113.9	11.3	4096	959
	s2 (12)	115.8	11.4	4096	952
	s1+s2 (14)	109.1	12.2	4096	877
Panopticon	s1 (2)	99.0	9.0	4096	1261
	s2 (12)	99.0	18.0	512	660
	s1+s2 (14)	99.0	19.8	512	607

Table 5: **Computation cost comparison among uni-modal DINO, DeluluNet, and Panopticon on an NVIDIA L40S GPU.** DINOv3 is low in GMACs and high in throughput, but is uni-modal. DeluluNet is comparable or cheaper in GMACs and achieves higher throughput than multi-modal RSFMs.

Loss ablations: Pre-fusion MSE and Latent MSE. In addition to the two loss terms for prediction heads that enable batch mixing (distillation loss and supervised loss), DeluluNet receives pre-fusion MSE and latent MSE as supervision. Prefusion MSE directly encourages masking transformers to impute the masked tokens/modality with predictions similar to the actual tokens. Latent MSE takes advantage of having a starting uni-modal f_0 that can extract useful latents from uni-modal $\mathcal{M}_A$ inputs, and is similar in construction to feature-based distillation methods (Heo et al., 2019). Table 4 shows that each term is crucial as intended: Pre-fusion MSE improves performance for transfer and peeking, when masking transformers are used to hallucinate the modality unavailable in inputs, and Latent MSE largely improves performance across the three scenarios.

Comparison to Panopticon zero-shot transfer. Unlike regular remote sensing models, a small group of sensor-agnostic models can extract features from unseen modalities. Notably, Panopticon (Waldmann et al., 2025) can perform modality transfer zero-shot. However, paired unlabeled data is abundant in remote sensing due to the continuous collection of orbiting satellites, and Figure 5 shows that leveraging it allows DeluluNet to significantly outperform zero-shot transfer.

Computational cost of DeluluNet. The amount of computation required by DeluluNet depends on the changing modality scenario. For modality peeking and transfer, where DeluluNet hallucinates multi-modal representation to perform pseudo-fusion, DeluluNet is generally more expensive than uni-modal foundation models like DINO due to this hallucination and token concatenation. However, even computationally light multi-modal remote sensing models such as Panopticon (Waldmann et al., 2025) tend to incur higher computational cost than uni-modal RGB models. This is due to design choices common in RSFMs such as per-band embedding (Waldmann et al., 2025), channel-group concatenation (Cong et al., 2022; Tseng et al., 2025), and small patch size (Clay Foundation, 2025; Feng et al., 2026). In Table 5, we use Panopticon as a representative RSFM and show that DeluluNet's computation is comparable for both uni-modal and multi-modal inputs, while achieving higher throughput. Note that the number of active parameters in DeluluNet varies across modes due to its modality-specific patch embedders and conditionally-activated masking transformers. All measurements are taken on a single NVIDIA L40S GPU, with GMACs reported per sample.

6 Discussion: Limitations and Extensions

DeluluNet demonstrates that a single unified framework can adapt existing uni-modal models to changing satellite modalities—without re-labeling—across transfer, addition, and peeking scenarios. The current work establishes this in the two-modality setting, with experiments restricted to static land-use tasks and paired unlabeled adaptation data. Below we discuss these limitations, and the broader opportunities DeluluNet opens up for the remote sensing community in future orbits.

Adapting foundation models to new sensor modalities. While we initialize DeluluNet from fine-tuned RGB DINOv3 to avoid remote sensing modality leakage in our experiments, in principle DeluluNet applies to adapting existing RSFMs to entirely new and future sensor modalities. As satellite constellations grow and sensors diversify, even the most capable RSFMs today will face modalities unseen during pretraining—and retraining from scratch each time is unsustainable. DeluluNet offers a practical path forward: augmenting an existing fine-tuned RSFM with modular components and unlabeled paired data, rather than rebuilding it. For practitioners wishing to work with sensor modalities not supported by existing RSFMs, DeluluNet offers an immediately deployable solution: rather than waiting for a new foundation model to be pretrained, one can augment an existing model with modular components using unlabeled paired data.

Temporal data and time-sensitive tasks. DeluluNet's paired adaptation split $\mathcal{D}_a$ assumes that co-located observations from different satellites capture the same targets—a reasonable assumption for static land-use tasks, but one that breaks down for time-sensitive applications such as disaster monitoring (Revankar et al., 2025) or unlicensed fishing detection (Paolo et al., 2022). In these settings, fast-shifting targets may no longer be present when a second satellite passes over the same region, introducing temporal misalignment that undermines reliable modality imputation. Explicitly modeling the temporal offset between satellite captures—and its effect on cross-modal correspondence—is therefore an important direction for extending DeluluNet to dynamic remote sensing tasks.

Continual self-supervised learning. While DeluluNet adapts an existing task model to new modalities on demand, a complementary direction is to continually update the foundation model itself as new sensors come online—eliminating the need to retrain from scratch each time a new modality is added. Remote sensing is uniquely suited for this: the continuous data collection nature of satellites means unlabeled paired observations accumulate naturally as new missions launch, providing a steady stream of data to update on without any additional annotation effort.

Toward *ever*-changing modalities. The three scenarios of changing modalities in this work—transfer, addition, and peeking—provide the elements to describe how many modality changes can unfold over time. While we restrict our experiments to the two-modality setting to first understand the minimal case, as new sensors are launched and old ones are retired, transfer, addition, and peeking can be applied in sequence, each building on the adapted model from the previous step. DeluluNet shows a promising first step toward sustainable, relabeling-free deployment of remote sensing models subject to the reality of evolving satellite constellations. We hope that our work motivates the community to develop long-term benchmarks covering diverse modalities and prediction tasks, and to design methods for adapting remote sensing models to change along with the sensors they rely on.

ICLR Reproducibility Statement All datasets we reported in this research are free and publicly available. In Appendix B we provide sufficient details on the proposed model architecture for reproducibility. We have included an anonymous code base to reproduce our main method as supplementary material for this submission.

AI Use Statement In this work, we used generative AI tools for implementation of methods and proofreading manuscript. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis or interpret results.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570. doi: 10.1145/279943.279962.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. volume 2006, pages 535–541, 08 2006. doi: 10.1145/1150402.1150464.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247, 2024.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6912–6920, 2021.

Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 865–872, 2019.

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pages 1264–1268. IEEE, 2025.

Clay Foundation. Clay foundation model, 2025. URL `https://clay-foundation.github.io/model/index.html`. Accessed: 2025-07-21.

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. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

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. In *International Conference on Learning Representations*, 2021.

Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*, 2018.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2011.11.026. URL https://www.sciencedirect.com/science/article/pii/S0034425712000636. The Sentinel Missions - New Opportunities for Science.

ESA, 2013. URL https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Free_access_to_Copernicus_Sentinel_satellite_data. Accessed: 2026-02-04.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 34818–34831, 2026.

Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Samuel N. Goward, Laura E. P. Rocchio, Darrel L. Williams, Terry Arvidson, James R. Irons, Carol A. Russell, and Shaida S. Johnston. Landsat's enduring legacy: Pioneering global land observations from space. *Photogrammetric Engineering and Remote Sensing*, 84:9–10, 2017. URL https://api.semanticscholar.org/CorpusID:134711498.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

Sebastian Hafner, Yifang Ban, and Andrea Nascetti. Unsupervised domain adaptation for global urban extraction using sentinel-1 sar and sentinel-2 msi data. *Remote Sensing of Environment*, 280: 113192, 2022. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2022.113192. URL https://www.sciencedirect.com/science/article/pii/S0034425722003029.

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*, pages 16000–16009, 2022.

Niharika Hegde, Shishir Muralidhara, René Schuster, and Didier Stricker. Modality-incremental learning with disjoint relevance mapping networks for image-based semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5540–5549. IEEE, 2025.

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.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1921–1930, 2019.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, Joshua Hansen, Andrew Howe, Patrick Alan Johnson, Mark Otterlee, Ted Schmitt, Hunter Pitelka, Stephen Daspit, Rachel Ratner, Christopher Wilhelm, Sebastian Wood, Mike Jacobi, Hannah Kerner, Evan Shelhamer, Ali Farhadi, Ranjay Krishna, and Patrick Beukema. Olmoearth: Stable latent image modeling for multimodal earth observation, 2025. URL https://arxiv.org/abs/2511.13655.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12605–12614, 2020.

Antoine Labatie, Michael Vaccaro, Nina Lardiere, Anatol Garioud, and Nicolas Gonthier. Maestro: Masked autoencoders for multimodal, multitemporal, and multispectral earth observation data, 2025. URL https://arxiv.org/abs/2508.10894.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

Aojun Lu, Hangjie Yuan, Tao Feng, and Yanan Sun. Rethinking the stability-plasticity trade-off in continual learning from an architectural perspective. In *Forty-second International Conference on Machine Learning*, 2025.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2302–2310, 2021.

Thorsten Markus, Tom Neumann, Anthony Martino, Waleed Abdalati, Kelly Brunt, Beata Csatho, Sinead Farrell, Helen Fricker, Alex Gardner, David Harding, Michael Jasinski, Ron Kwok, Lori Magruder, Dan Lubin, Scott Luthcke, James Morison, Ross Nelson, Amy Neuenschwander, Stephen Palm, Sorin Popescu, CK Shum, Bob E. Schutz, Benjamin Smith, Yuekui Yang, and Jay Zwally. The ice, cloud, and land elevation satellite-2 (icesat-2): Science requirements, concept, and implementation. *Remote Sensing of Environment*, 190:260–273, 2017. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2016.12.029. URL https://www.sciencedirect.com/science/article/pii/S0034425716305089.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 119–129. IEEE, 2020.

NRC. Hyperspectral imaging technologies, 2019. URL https://nrc.canada.ca/en/research-development/products-services/technical-advisory-services/hyperspectral-imaging-technologies. Accessed: 2026-02-04.

Fernando Paolo, Tsu-ting Tim Lin, Ritwik Gupta, Bryce Goodman, Nirav Patel, Daniel Kuster, David Kroodsma, and Jared Dunnmon. xview3-sar: Detecting dark fishing activity using synthetic aperture radar imagery. *Advances in Neural Information Processing Systems*, 35:37604–37616, 2022.

Lawrence G Phillips, David B Grimes, and Yihan Jessie Li. Teacher-student domain adaptation for biosensor models. *ICLR*, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Shreelekha Revankar, Utkarsh Mall, Cheng Perng Phoo, Kavita Bala, and Bharath Hariharan. Monitrs: Multimodal observations of natural incidents through remote sensing. *arXiv preprint arXiv:2507.16228*, 2025.

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.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun 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, 2025. URL https://arxiv.org/abs/2508.10104.

Naomi Simumba, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, Geeth De Mel, Salman Khan, Manil Maskey, Nicolas Longepe, Xiao Xiang Zhu, Hannah Kerner, et al. Geo-bench-2: From performance to capability, rethinking evaluation in geospatial ai. *arXiv preprint arXiv:2511.15658*, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *arXiv preprint arXiv:1902.06148*, 2019.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

Hakob Tamazyan, Ani Vanyan, Alvard Barseghyan, Anna Khosrovyan, Evan Shelhamer, and Hrant Khachatrian. Geocrossbench: Cross-band generalization for remote sensing. *arXiv preprint arXiv:2511.02831*, 2025.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *International Conference on Machine Learning*, pages 60280–60300. PMLR, 2025.

Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2204–2214, 2025.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15878–15887. IEEE, 2023a.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. In *Eleventh International Conference on Learning Representations*. OpenReview. net, 2023b.

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*, 2024a.

Zhitong Xiong, Yi Wang, Fahong Zhang, and Xiao Xiang Zhu. One for all: Toward unified foundation models for earth vision. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 2734–2738, 2024b. doi: 10.1109/IGARSS53475.2024.10641637.

Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 854–863, 2021.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. Skysense v2: A unified foundation model for multi-modal remote sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9136–9146, 2025.

Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. *arXiv preprint arXiv:2402.11148*, 2024.

Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5116–5125, 2023.

Tim G Zhou, Evan Shelhamer, and Geoff Pleiss. Asymmetric duos: Sidekicks improve uncertainty. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A The Changing Modality Scenarios And The Adjacent Fields

The changing modality setting shares ingredients with several established research areas, but departs from each in at least one key assumption. Continual Learning and Domain Adaptation assume that the input modality remains fixed; cross-modal Domain Adaptation allows the modality to change but assumes unpaired data; and Learning with Privileged Information assumes labeled multimodal data. In contrast, our setting involves a change in sensing modality, requires labels only for the old modality, exploits paired but unlabeled multi-modal observations, and builds on existing pretrained encoders. We discuss each field in more detail below.

Continual Learning. Continual Learning focuses on continuously acquiring new knowledge while preserving previously learned knowledge, typically organized into task-, class-, and domain-incremental learning (McCloskey and Cohen, 1989; Van de Ven et al., 2022; Lu et al., 2025). In all of these variants, new knowledge arrives as labeled data, and the input modality remains the same throughout. In the changing modality setting, by contrast, it is the sensing modality itself that changes, and the new modality comes without labels.

Domain Adaptation. Domain Adaptation (Ganin et al., 2016; Sun and Saenko, 2016), including multi-modal Domain Adaptation (Jaritz et al., 2020; Munro and Damen, 2020; Hafner et al., 2022), transfers knowledge from a labeled source domain to a different, unlabeled target domain. However, it assumes a fixed set of input modalities shared across domains, with only the underlying data distribution shifting. This is orthogonal to our setting, where the sensing modality itself changes. *Cross-modal Domain Adaptation* (Dou et al., 2018; Chen et al., 2019) allows the train and test modalities to differ, and is therefore closest to our *modality transfer* scenario. However, it assumes *unpaired* data and adapts via adversarial feature alignment (Ganin and Lempitsky, 2015). In contrast, we exploit spatiotemporally co-located, *paired* multi-modal observations (e.g., co-registered Sentinel-1/Sentinel-2 patches), which are cheap to acquire even when labels are not.

Learning with Privileged Information. Learning with Privileged Information (Vapnik and Vashist, 2009; Hoffman et al., 2016; Ma et al., 2021; Wang et al., 2023a) leverages modalities that are available only at training time, and is closest in spirit to our *modality peeking* scenario. Among these methods, modality hallucination (Hoffman et al., 2016) is the closest match, but it requires *labeled* multimodal data for training, which implies relabeling whenever a new modality becomes available. It also trains its convolutional hallucination network from scratch, without leveraging pretrained backbones. In contrast, modality peeking only requires labels for the old modality and builds on existing encoders.

B DeluluNet Architecture Details

Modality Specific Feature Extractor. In Vision Transformers (ViT) (Dosovitskiy et al., 2021), image inputs are divided into patches and projected into token embeddings via a 2D convolution layer, where `in_chan` is set to 3 for RGB inputs. Since remote sensing modalities have varying numbers of input channels, it is common to use modality-specific patch embedders (Tseng et al., 2025; Herzog et al., 2025). DeluluNet follows this design: each modality receives a dedicated 2D convolution layer for patch embedding, followed by 3 transformer blocks where tokens from the same modality attend to each other.

Masking Transformers. Masking transformers are lightweight 2-layer transformers consisting of a standard self-attention layer followed by a custom cross-attention layer. The cross-attention layer contains per-modality queries for $\mathcal{M}_A$ and $\mathcal{M}_B$, each of dimension 768 and length 2—one for the CLS token and one for the patch token. At forward time, the patch query is broadcast into n copies, where $n = (\texttt{input-size}/\texttt{patch-size})^2$ is the number of tokens for that modality. Each broadcast patch query is then augmented with ROPE positional encoding shared with the rest of DeluluNet, ensuring a principled and parameter-efficient design.

Modality Embeddings and Fusion Blocks. Modality embeddings encode modality information, since naive transformer layers treat all tokens equally and are blind to which modality a given token belongs to. These are learned $d = 768$ vectors added directly to intermediate representations from modality-specific feature extractors (or hallucinated ones from masking transformers). The Fusion Transformer consists of 9 self-attention blocks; together with the 3-layer modality-specific transformers, this matches the ViT-Base architecture, which we initialize with DINOv3-Base weights.

Predictor Heads. We use simple linear heads across all evaluated tasks. For DeluluNets, separate heads predict from the CLS token (for classification) or patch tokens (for segmentation) from both $\mathcal{M}_A$ and $\mathcal{M}_B$. This resembles Shallow Ensemble (Lee et al., 2015), where multiple heads predict from the same backbone—with the key difference that DeluluNet's heads operate on features from different modalities. Although this potentially introduces an implicit asymmetry in predictive power when some modality features are hallucinated by masking transformers, joint optimization of the predictor heads during training (Zhou et al., 2025) mitigates this concern.

Train-Time-Only Component. DeluluNet includes a lightweight one-layer transformer that predicts the f_0 representation from fused $\mathcal{M}_A$ representations. This provides flexibility in what DeluluNet learns, rather than forcing it to directly match the teacher's representation.

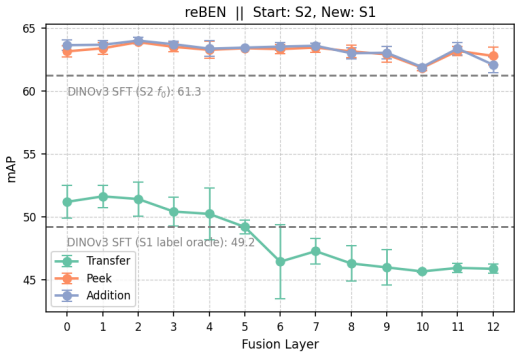

Figure 6: **DeluluNet benefits from more allocation of layers for modality fusion**: earlier fusion tends to lead to better results across transfer, peeking, and addition. For modality transfer, earlier fusion even leads to transfer results that surpass DINOv3 supervised fine-tuned on Sentinel-1 data.

Fusion Time The depth at which hallucination and modality fusion is done is a hyperparameter in DeluluNet. Figure 6 shows that earlier fusion tends to lead to better results across scenarios. This is especially clear for modality transfer, where earlier modality hallucination and fusion lead to DeluluNet outperforming the label-oracle supervised fine-tuned baseline. This shows that DeluluNets benefit from fusion of real and hallucinated tokens.

C Label-Free Early Stopping for DeluluNet Transfer and Addition

Similar to training splits $\mathcal{D}_l$ and $\mathcal{D}_a$, we assume a labeled uni-modal $\mathcal{M}_A, \mathcal{Y}$ validation set and an unlabeled multi-modal $\mathcal{M}_A, \mathcal{M}_B$ validation set. Since peeking mode only takes in $\mathcal{M}_A$, checkpoint selection follows standard validation evaluation. For *transfer* and *addition*, which operate on the unlabeled multimodal split, we instead use each mode's agreement with the supervised uni-modal f_0 as a proxy metric.

There are many ways to define "agreement", both hard (using peek prediction as pseudo labels) or "soft" (with divergence or similarity metrics). Figure 7 shows that this proxy correlates well with the true test metric when validation score is defined as top-1 agreement. It is thus useful for both hyper-parameter tuning and early stopping without access to labels on the multimodal split.

D Measuring Pearson Correlation Between Real and Hallucinated Features

The token-wise correlations of Fig.4 establish that hallucinated tokens occupy the same representational space as real ones, and the real-vs-hallucination feature map visualization further shows that the cross-modality hallucinated features meaningfully imitate real features. Here we present the per-feature pearson correlation analysis to further emphasize that the hallucination is sample-dependent by measuring correlation across the sample axis, per (patch position, feature) coordinate, which subtracts the dataset mean token and assigns zero to any projector that ignores its input. Table6 reports this alongside a shuffled control that hallucinates each sample's missing modality from a different sample's observed modality.

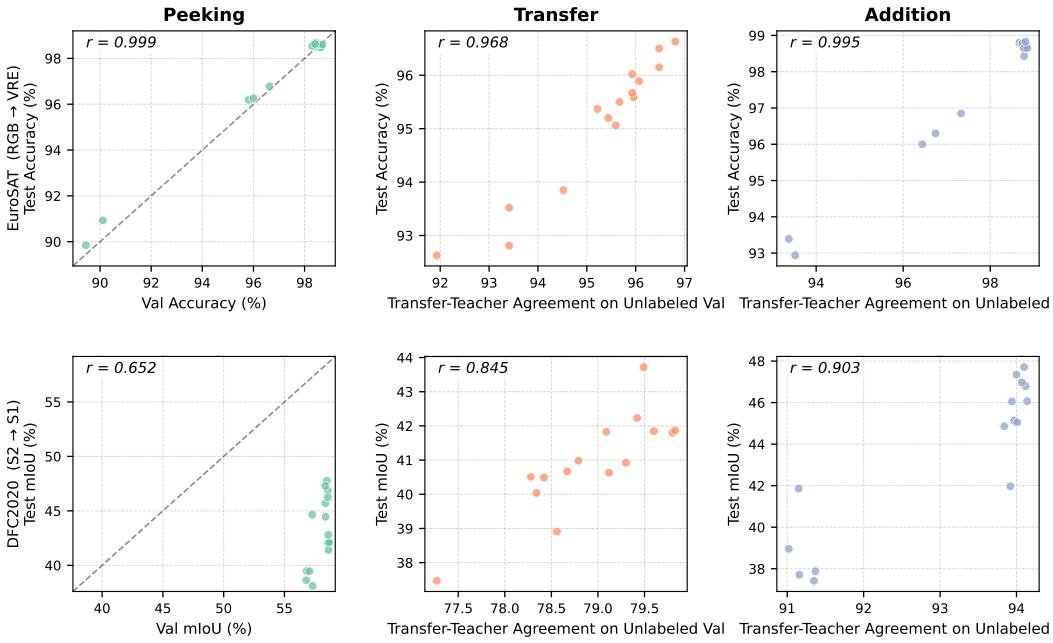

Figure 7: **Label-Free Validation Vs. Test metric** (left) peeking uses standard validation evaluation since we have labels for $\mathcal{M}_A$. (center&right) Transfer and Addition uses their agreement with teacher f_0 times peeking performance as val metric, providing a validation score that doesn't require labels for multimodal validation split, and corresponds to test split metric reasonably well.

Table 6: **Per-feature Pearson correlation between hallucinated and real patch tokens** on DFC2020, computed per (patch position, feature) over the test split. *Paired* hallucinates each sample's missing modality from its own observed modality; *Shuffled* hallucinates it from a different sample's to serve as a baseline. Centering over the sample axis removes the dataset mean token, so a projector that ignores its input scores 0. Paired correlations of 0.33–0.49 against a null indistinguishable from zero show the hallucination is driven by each sample's own content, not by shared representational geometry.

Modalities	Target	Paired r	Shuffled r	Margin
S1 ↔ S2	S1	0.330 ± 0.194	$+0.004 \pm 0.031$	10.5σ
	S2	0.395 ± 0.167	$+0.002 \pm 0.032$	12.5σ
S1 ↔ $S2_{noRGB}$	S1	0.488 ± 0.235	$+0.002 \pm 0.031$	15.9σ
	$S2_{noRGB}$	0.434 ± 0.191	$+0.004 \pm 0.031$	13.8σ

Paired correlations range from 0.33 to 0.49, while the shuffled control never exceeds 0.004 in magnitude. The margin is 10.5–15.9σ beyond the shuffled baseline. The masking transformer therefore recovers genuine per-sample content: it does not merely produce tokens that resemble the target modality in aggregate, but tokens that track the specific sample it was given. Using Pearson correlation is a strict criterion here, as the training objective does not optimize it directly.

The spread across coordinates is wide ($\pm$0.17–0.24) and is dominated by variation across feature channels rather than across patch positions (per-channel s.d. 0.19 versus per-position s.d. 0.004): some channels are hallucinated at $r > 0.8$ while others approach chance, this is expected since different sensors have different physical properties, data of different modality will carry unique information invisible to other modalities. This aligns well with the Modality Focusing Hypothesis (Xue et al., 2022), and should be insightful for future method design.

Table 7: **Cross-Band Modality Transfer** results show that DeluluNet transfers more effectively than knowledge distillation (KD) and transformed teacher matching (TTM) for cross-band setting, at times even surpasses supervised fine-tuned DINOv3 (DINO SFT) on M_B.

Dataset (metric)		EuroSAT (Acc)					
Start (M_A)		RGB		VRE		NIR	
Transfer (M_B)		VRE	NIR	RGB	NIR	RGB	VRE
$f_0(M_A)$	f_0	98.6	98.6	90.7	90.7	80.9	80.9
Baselines	KD	88.6±1.3	74.8±7.4	86.2±2.0	75.3±3.9	81.4±0.8	80.7±0.6
	TTM	89.5±1.0	77.4±5.7	88.2±0.7	75.0±3.7	81.4±0.8	80.8±1.0
Ours	Delulu	**96.4±0.2**	**88.4±0.6**	**89.0±0.5**	**78.9±1.7**	**85.0±0.2**	**82.3±0.4**
Oracle (M_B)	DINOv3	90.7	80.9	98.6	80.9	98.6	90.7

Table 8: **Cross-Band Modality Peeking** results show that DeluluNet is able to learn from observations of unlabeled and paired $\mathcal{M}_A$ and $\mathcal{M}_B$ data to make better predictions on $\mathcal{M}_A$, generally surpassing Semi-SL baseline and DINO-init M_A f_0.

Dataset (metric)		EuroSAT (Acc)					
Start (M_A)		RGB		VRE		NIR	
New (M_B)		VRE	NIR	RGB	NIR	RGB	VRE
$f_0(M_A)$	f_0	98.6	98.6	90.7	90.7	80.9	80.9
Baselines	MixMatch	**99.0±0.1**	**99.0±0.1**	87.9±4.1	87.9±4.1	66.7±5.7	66.7±5.7
Ours	Delulu	98.5±0.1	98.5±0.0	**91.6±0.3**	**91.6±0.2**	**82.8±0.3**	**82.3±0.4**

Table 9: **Cross-Band Modality Addition** results show that DeluluNet is able to extract and fuse inputs from both $\mathcal{M}_A$ and $\mathcal{M}_B$ through modality addition.

Dataset (metric)		EuroSAT (Acc)					
Start (M_A)		RGB		VRE		NIR	
New (M_B)		VRE	NIR	RGB	NIR	RGB	VRE
$f_0(M_A)$	f_0	98.6	98.6	90.7	90.7	80.9	80.9
Baselines	MKE	97.5±0.9	97.9±0.5	90.2±0.2	89.4±0.3	77.4±4.0	77.0±2.0
Ours	Delulu	**98.7±0.1**	**98.7±0.1**	**91.5±0.1**	**91.6±0.1**	**82.3±0.3**	**81.9±0.2**
Oracle (M_A+M_B)	DINOv3	98.6	98.8	98.6	88.7	98.8	88.7

E Full Cross-Band EuroSAT Results

The main experiments in Section 5 focus on cross-satellite modality changes, which are more practically relevant and challenging given the greater distinction between modalities. Here we complement these with full cross-band results, evaluating transfer, addition, and peeking across the three major Sentinel-2 band groups from EuroSAT: RGB (B-02, B-03, B-04), VRE (B-05, B-06, B-07), and NIR (B-08, B-8A). Tables 7, 8, 9 confirm that findings from the $\mathcal{M}_A$ =RGB, $\mathcal{M}_B$ =VRE setting in Section 5 generalize across Sentinel-2 sub-bands, and that DeluluNet serves as a reliable framework for both cross-satellite and cross-band transfer, peeking, and addition.