CHANGING MODALITIES BY CROSS-BAND TRANSFER, ADDITION, AND PEEKING

Tim G. Zhou[1,2] **Anthony Fuller**[3,2] **Geoff Pleiss**[1,2] **Evan Shelhamer**[1,2]
[1]University of British Columbia [2]Vector Institute [3]Carleton University

ABSTRACT

Machine learning models for remote sensing typically assume a static set of modalities. However, as we equip newer satellites with novel sensors and retire old ones, practitioners may wish to deploy a model on a substitution, superset, or subset of modalities given data availability or practical constraints. We formulate the setting of changing modalities and identify three main scenarios: Modality Transfer, Addition, and Peeking. We propose Delulu-Net, an architecture with modular components for all three changing modality scenarios. Delulu-Net learns a multi-modal model from a unimodal teacher and *un*-labeled multimodal data, providing a practical alternative to re-labeling and re-training.

1 INTRODUCTION: MODALITIES CAN CHANGE, SO MODELS MUST CHANGE

Earth observation (EO) satellites continually collect vast amounts of *diverse* data across modalities (optical, radar, lidar). The quantity and diversity of data makes machine learning for remote sensing (ML4RS) promising but challenging. Satellite missions are regularly updated by adding new sensors while retiring old ones, new missions are launched with higher resolutions, formerly free data can become paywalled, etc. It is difficult to make use of these new modalities with existing remote sensing models. One solution is to create new annotated training data that includes the new modality, which is costly and sometimes impossible. We instead propose changing the model.

We formalize the **changing modalities** setting and identify key scenarios: (1) *Transfer*: switching from an old modality to a new modality; (2) *Addition*: pairing old and new modalities; and (3) *Peeking*: improving predictions on an old modality by training with a new modality.

We propose **Delulu-Net**, an architecture with modular uni-modal and multi-modal components that adapts an existing predictve model from unlabeled multi-modal data. We test Delulu-Net on grouped bands from Sentinel-2, showing its efficacy in transfer, addition, and peeking settings *without* a single label on the new modality. We further show Delulu-Net's compatibility with foundation models, equipping the state-of-the-art DINOv3 (Siméoni et al., 2025) model with multi-spectral capabilities.

2 SETTING: CHANGING MODALITIES

Data splits and modalities. Consider a predictive task with labels $\mathcal{Y}$. We consider two modalities: a *partially-labeled modality* $\mathcal{M}_A$ and an *unlabeled modality* $\mathcal{M}_B$. Our data is 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. **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. **Addition** $\mathcal{D}_e \subset (\mathcal{M}_A \times \mathcal{M}_B)$: The adapted model makes predictions using $\mathcal{M}_A$ and $\mathcal{M}_B$ jointly, leveraging the two modalities in synergy.
3. **Peeking** $\mathcal{D}_e \subset \mathcal{M}_A$: The adapted model continues to make predictions using $\mathcal{M}_A$ alone, but with the additional knowledge from observing unlabeled $\mathcal{M}_B$ during adaptation.

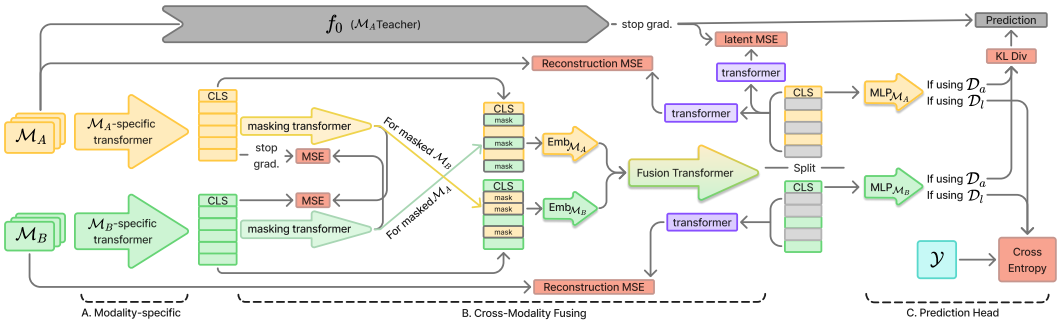

Figure 1: **Delulu-Net Architecture and Training Strategy.**

Example use cases. Each scenario corresponds to a realistic use case. First, we may want to update an existing model to process data from a new satellite for improved predictions (Tamazyan et al., 2025). For example, when Sentinel-1 data became available (ESA, 2013), it was combined with existing optical imagery from LandSAT (Goward et al., 2017)—a process that typically required re-labeling paired data, but could be avoided with label-free modality *addition*. Second, we may want to replace one source with another while maintaining or improving predictions. When Sentinel-2 data became available, many wished to convert their LandSAT models to Sentinel-2 (which provides higher spatial, spectral, and temporal resolutions) via *transfer*. Finally, we may want to exploit temporarily available or expensive data to improve predictions on our own data. For example, global hyperspectral imagery (NRC, 2019) is rarely available freely, but is information-rich enough that models could *peek* at it to learn fine-grained spectral patterns (e.g., of agriculture) and transfer that knowledge back to improve predictions on standard data. More details about the setting and discussions on its connection to other machine learning paradigms can be found in Appendix A.

3 Method: Modality Hallucination (Delulu is the Solulu)

We introduce Delulu-Net, designed to operate under the changing-modality setting. Delulu-Net starts with any uni-modal $\mathcal{M}_A$-input ViT f_0 (Dosovitskiy et al., 2021) trained on labeled data from $\mathcal{D}_l$, and converts f_0 into a multi-modal model to also process unlabeled $\mathcal{M}_A$ and $\mathcal{M}_B$ data from $\mathcal{D}_a$.

3.1 Augmenting the Model Architecture from Uni-Modal to Multi-Modal

Delulu-Net augments the uni-modal f_0 with $\mathcal{M}_B$ capabilities using modular multi-modal components for feature extraction and prediction compatible with modality transfer, addition, and peeking:

1. *Modality-Specific Extraction:* randomly-initialized $\mathcal{M}_B$-specific tokenizer and f_0-initialized $\mathcal{M}_B$-specific transformer layers that attend to tokens extracted from $\mathcal{M}_B$ for modality-specific representation. We denote this $\mathcal{M}_B$-specific transformer.
2. *Cross-Modality Fusion:* randomly-initialized modality embeddings Emb$_{\mathcal{M}_B}$, and masking transformers that learn to predict cross-modal features between $\mathcal{M}_A$ and $\mathcal{M}_B$.
3. *Class Predictor:* classifier head MLP$_{\mathcal{M}_B}$ that predicts from CLS token from $\mathcal{M}_B$.

3.2 Learning Features and Predictions From Unlabeled Modalities.

Delulu-Net is trained with the same end-to-end strategy for all three scenarios. Losses at different depths for different components encourage learning features from $\mathcal{M}_B$ that are suitable for fusion. At the heart of it are the **masking transformers** that learn to predict cross-modal features from present modalities to missing ones. The predicted features serve as educated hallucinations that replace missing or masked input modality during training and testing, and are essential in improving performance on top of the starting supervised uni-modal model f_0.

Modality-specific feature learning. Raw multi-modal inputs are passed independently through modality-specific transformers initialized for each modality's number of bands. We further exploit

the asymmetry in the initial model's training of different modalities, and supervise the $\mathcal{M}_B$-specific components using features extracted from $\mathcal{M}_A$.

Cross-modal feature learning. We use a feature-distillation loss for $\mathcal{M}_A$ using f_0 as teacher, and use input-reconstruction loss for both modalities. In addition to token masking, we randomly mask entire modalities during training. Unique to Delulu-Net, we use the masking transformers to predict and replace the masked tokens. This allows the model to handle inputs with missing modalities and encourages cross-modality alignment beneficial to the modality addition setting.

Prediction head learning. To ensure accurate predictions that incorporate features from both modalities, the training minibatches alternate between the multi-modal unlabeled $\mathcal{D}_a$ and the uni-modal labeled dataset $\mathcal{D}_l$. On the unlabeled $\mathcal{D}_a$, we use knowledge distillation (Hinton et al., 2015) to learn from the predictions made by the uni-modal supervised f_0 as teacher model. On the labeled $\mathcal{D}_l$, the Delulu-Net treats these as multi-modal input where $\mathcal{M}_B$ is masked, and the cross-modal masking transformers enable Delulu-Net to in turn directly learn from the labeled uni-modal data. The ratio of training batches between $\mathcal{D}_a$ and $\mathcal{D}_l$ is a hyper-parameter denoted batch-mixing ratio.

Optimization and hyper-parameter tuning. Figure 1 summarizes our backbone's training scheme. All architectural design details and hyper-parameter sweeping range can be found in Appendix B.

4 Experiment

Dataset and Bands. We evaluate performance of Delulu-Net on EuroSAT (Helber et al., 2019), a 10-way classification dataset with 13 spectral bands from the Sentinel-2 satellites. We select the four major groups of bands from EuroSAT: RGB (B-02,B-03,B-04), VRE (B-05, B-06, B-07) , NIR (B-08,B-8A), and SWIR (B-10, B-11, B-12). We sample half of the official training split as labeled uni-modal $\mathcal{D}_l$ and the other half as unlabeled multi-modal $\mathcal{D}_a$. Appendix C includes additional results on cross-satellite experiments on reBEN multi-label classification task (Clasen et al., 2025).

f_0 Architecture. We train supervised **ViT-Small** (Dosovitskiy et al., 2021) on the labeled $\mathcal{D}_l$, these supervised models serve as either initiating uni-modal f_0 for Delulu-Net, or as oracle baselines depending on the evaluation scenario.

Modality Transfer. Modality Transfer aims to transfer f_0's performance from $\mathcal{M}_A$ onto $\mathcal{M}_B$. We compare Delulu-Net to knowledge distillation (Hinton et al., 2015) as a label-free baseline, and additionally train fully supervised uni-modal ViT-S as oracles. Figure 2 shows that Delulu-Net not only consistently transfers better than regular knowledge distillation, but even sometimes surpass the supervised input-oracle or label-oracle models.

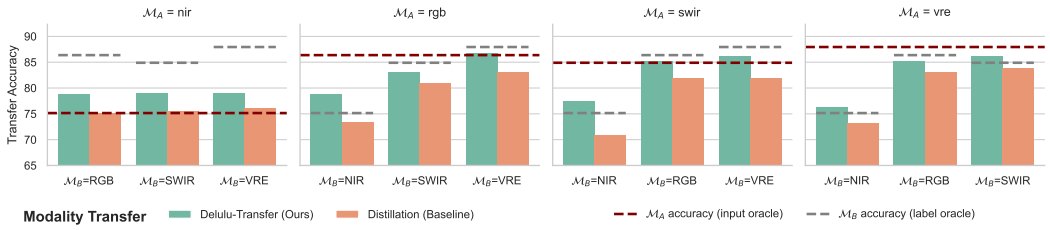

Figure 2: **Transfer: Delulu-Net can switch modalities without labels.** When transferring from RGB, VRE, or SWIR to NIR, the transferred models even surpass label-oracle fully-finetuned uni-modal models.

Modality Addition. Modality Addition aims to improve predictions with both $\mathcal{M}_A$ and $\mathcal{M}_B$. We found that the Delulu-Net is best at addition tasks when ensembling two Delulu-Nets with hyperparameters tuned for peeking and transfer respectively. We hence compare Delulu-Net to the ensemble of f_0 and KD-transferred $\mathcal{M}_B$ model as baseline. Figure 3 shows that Delulu-Net outperforms both initializing f_0 and Ens$(f_0, \text{distilled})$ baseline.

Modality Peeking. Modality Peeking aims to improve prediction accuracy from observing an unlabeled modality. Figure 4 shows Delulu-Net effectively learns from paired $\mathcal{M}_A, \mathcal{M}_B$ from $\mathcal{D}_a$ and achieved up to 5% improvement in accuracy from the f_0 initialization without any supervision.

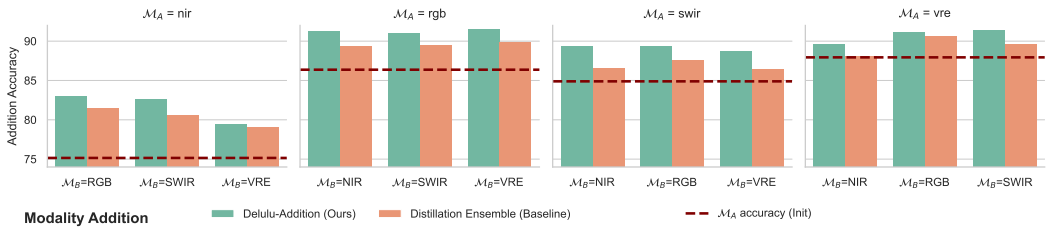

Figure 3: **Addition: Delulu-Net learns and predicts with both labeled $\mathcal{M}_A$ and unlabeled $\mathcal{M}_B$,** consistently outperforms the f_0-distillation ensemble, and improves accuracy over uni-modal f_0 by up to 8%.

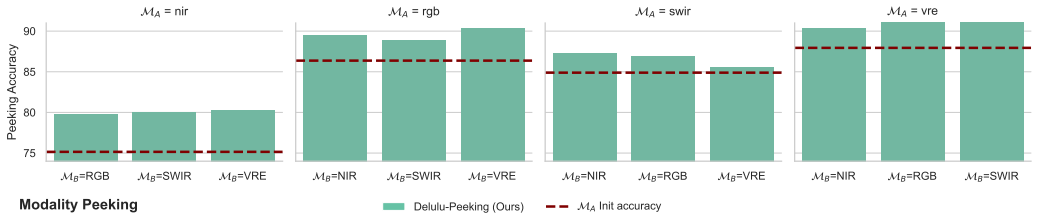

Figure 4: **Peeking: Delulu-Net learns to make uni-modal predictions with multi-modal data** by only observing unlabeled paired $\mathcal{M}_A$ and $\mathcal{M}_B$ inputs during training.

Compatibility with Pre-training on External Data (DINOv3). While remote sensing data contains distinct modalities from RGB imagery (Rolf et al., 2024), there is an opportunity to make use of the massive efforts on image foundation models. We therefore apply Delulu-Net to the setting where f_0 is a fine-tuned DINOv3 RGB model, the leading vision foundation model. Figure 5 shows that Delulu-Net can augment DINOv3 with multi-spectral capabilities for better transfer, addition, and peeking than direct fine-tuning and knowledge distillation.

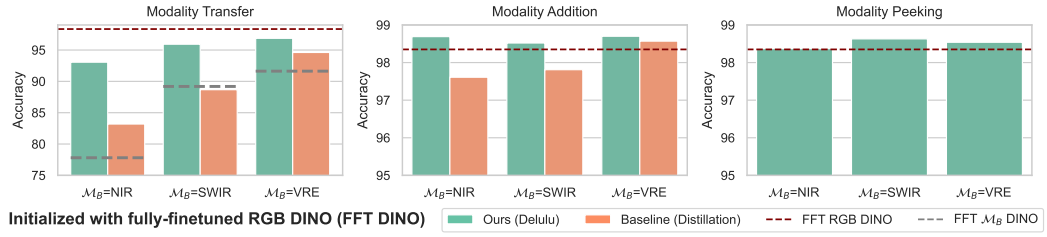

Figure 5: **Delulu-Net can use RGB pre-training from DINOv3,** improving over direct fine-tuning and distillation without new modality labels. Pre-training on external data via DINOv3 improves accuracy over training from scratch on remote sensing data (Figures 2–4).

5 Discussion: Limitations and Extensions

Beyond classification. Our experiments focus on classification, but the Delulu-Net architecture is applicable to dense prediction tasks such as semantic segmentation. Incorporating temporal encoding into the model would expand its applicability to time-series analysis and change detection tasks.

Toward ever-changing modalities. Delulu-Net takes a promising first step toward sustainable deployment of remote sensing models subject to the reality of evolving satellite constellations. While we restrict our experiments to the two-modality setting to first understand the minimal case, our method naturally extends to more modalities. Even sequences of changing modalities could be handled by applying transfer, addition, and peeking in sequence as sensors are introduced or retired. Investigating such a setting of ever-changing modalities is an important direction for future work.

REFERENCES

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*, pp. 1264–1268. IEEE, 2025.

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.

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

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.

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.

Stephen Grossberg. Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks*, 37:1–47, 2013. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2012.09.017. Twenty-fifth Anniversay Commemorative Issue.

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)*, pp. 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.

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.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in neural information processing systems*, 32, 2019.

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.

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.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pp. 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.

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.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: mission critical - satellite data is a distinct modality in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60280–60300. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tseng25a.html.

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

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*, pp. 2734–2738, 2024. doi: 10.1109/IGARSS53475.2024.10641637.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3014–3023, 2021.

A THE CHANGING MODALITIES SETTING

Under the minimal setup used in this paper when there are only two modalities in total, the three settings can be illustrated as such in Table 1, where Labeled Split $\mathcal{D}_l$ is used to train the initiating uni-modal f_0 under supervision, and Adaptation Split $\mathcal{D}_a$ is used to enable transfer/addition/peeking, before finally predicting on the evaluation split $\mathcal{D}_e$.

Table 1: Test Settings under the ever-changing modalities setup.

	Labeled Split	Adaptation Split	Evaluation Split
Transfer	$\mathcal{M}_A, \mathcal{Y}$	$\mathcal{M}_A$&$\mathcal{M}_B$	$\mathcal{M}_B$
Addition			$\mathcal{M}_A$& $\mathcal{M}_B$
Peeking			$\mathcal{M}_A$

As we've discussed in Section 2, each scenario corresponds to a common real-world use case within remote sensing applications.

A.1 CONNECTIONS TO CONTINUAL LEARNING.

Continual Learning aims to enable deep networks to continuously acquire new knowledge while preserving learned knowledge, broadly categorized into Task/Class/Domain incremental learning (Lu et al., 2025; Van de Ven et al., 2022). Recently, modality-incremental learning was proposed under the autonomous driving setting where new sensors similar to those used in remote sensing are constantly being introduced for robustness of self-driving vehicles (Hegde et al., 2025).

Importantly, continual learning works with streams of **labeled** distribution shifts. Our setting fundamentally differ as we focus on bypassing the need for labeling data collected with new input modalities. Nonetheless, the core ideas from Incremental Continual Learning are applicable here for drawing useful parallels between the changing modality (CM) setting and continual learning (CI).

One of the main focuses of CI is the plasticity-stability dilemma (Grossberg, 2013; Lu et al., 2025), where plasticity refers to a model's ability to learn from new task/distribution, and stability refers to a model's ability to preserve previously learned task/distribution (McCloskey & Cohen, 1989). There exists a natural tradeoff between stability and plasticity, for which many learning algorithms have been proposed over the years. These methods mainly revolve around replay (mixing data from previous tasks into data for new tasks), regularization (restraining new tasks from updating the model too much), and architecture-based methods that assign task-specific dynamic instantiation or allocation of new methods (Yan et al., 2021; Hung et al., 2019; Lu et al., 2025). Among these methods, Delulu-Net is the most similar to the architecture-based approach.

Under our CM setting, transfer and peeking closely correspond to plasticity and stability. As we've noticed during hyper-parameter tuning, the checkpoint selected for modality peeking is often different from the checkpoint selected from modality transfer, resembling the tradeoff between plasticity and stability.

B DELULU-NET: ARCHITECTURE AND OPTIMIZATION

B.1 BACKGROUND: REMOTE SENSING FOUNDATION MODELS

The field has seen a recent flux of remote sensing foundation models (RSFMs) following the success of language and vision foundation models (VFMs). These RSFMs are powerful generalists and uniquely *multi-modal*, taking in various modalities of remote sensing, with the most common ones being multi-spectral optical data (e.g. Sentinel-2, LandSAT 8&9) and Synthetic Aperture Radar data (e.g. Sentinel-1).

To handle raw inputs of different modalities, RSFMs use separate patch embedders (tokenizers) for different input modalities due to their intrinsic differences (Tseng et al., 2025; Herzog et al., 2025) before fusing different modalities together (often through concatenation). Although these are different

modalities with sometimes even different resolution, they can generally be considered as multi-band imagery. For this reason, the most recent wave of RSFMs have shown that these modalities can share a large portion of the encoder after the initial per-modality processing, and have converged to this weight-sharing strategy to different degrees as a desirable trade-off between capacity and efficiency (Fuller et al., 2023; Xiong et al., 2024; Tseng et al., 2025; Herzog et al., 2025; Labatie et al., 2025). These design principles heavily motivate the design of Delulu-Net.

B.2 Components of Delulu-Net

Delulu-Net is initialized with a uni-modal ViT (Dosovitskiy et al., 2021) as f_0, and augment f_0 with the following components for a new modality $\mathcal{M}_B$, as listed out in Section 3.

1. *Modality-Specific Components:* $\mathcal{M}_B$-specific tokenizer, $\mathcal{M}_B$-specific transformer layers.
2. *Cross-Modality Fusion:* Modality encodings $\text{Enc}_{\mathcal{M}_B}$ and masking transformers that learn to predict cross-modal features between $\mathcal{M}_A$ and $\mathcal{M}_B$.
3. *Task Prediction:* Task-specific $\text{Head}_{\mathcal{M}_B}$ that predicts from CLStoken extracted from $\mathcal{M}_B$.

$\mathcal{M}_B$-specific tokenizer is a randomly-initialized 2D-Convolution layer with $\mathcal{M}_B$-dependent number of input channels; and the $\mathcal{M}_B$-specific transformer layers are initialized with the first m layers of the initial f_0.

$\mathcal{M}_B$ modality encodings is a randomly initialized, learnable vector matching the *hidden-dimension* of the initiating ViT, that is added to tokens from $\mathcal{M}_B$ before modality fusion. Masking transformers are 2-layer transformers going between modality-specific features from $\mathcal{M}_A$ and $\mathcal{M}_B$.

$\text{Head}_{\mathcal{M}_B}$ is a 2-layer MLP, same as the original predictor head from f_0 for $\mathcal{M}_A$.

Beside these components that are active during both training and testing, three additional transformers are used for feature-distillation and input-reconstruction as shown in Figure 1. These are all 2-layer transformers for seq-2-seq projection, and are randomly initialized.

B.3 Hyper-Parameter Tuning

As discussed in Appendix A, there is a natural tradeoff between different scenarios. Luckily, the three scenarios are parallel to each other in practice, and so we can optimize for each scenario separately. We do light hyper-parameter sweep over the following configurations:

1. *mask ratio* [0.1,0.9] controls the ratio of tokens that are randomly masked.
2. *modality dropout* [0.1,0.5] controls the ratio of each modality being entirely masked.
3. *batch-mixing frequency* [0,1] controls the proportion of batches coming from $\mathcal{D}_a$, where we treat it as $\mathcal{M}_B$-dropped input and hallucinate $\mathcal{M}_B$ intermediate features using masking transformers, and supervise with ground truth directly.
4. *batch-mixing scheduler* [0,1] controls when batch mixing starts.
5. *learning rate* [8e-5, 1e-3] sets the learning rate for AdamW optimizer.
6. *weight decay* [1e-5, 1e-1] sets the weight decay for AdamW optimizer.

We report the best checkpoint selected by validation set for each of transfer, addition, and peeking; where transfer and addition validation performance is evaluated by f_0-agreement on unlabeled multi-modal val set, and peeking validation performance is evaluated with ground truth from labeled uni-modal val set.

C Changing Satellites

To evaluate Delulu-Net under a more realistic deployment scenario, we conduct experiments on cross-satellite modality transfer using reBEN (Clasen et al., 2025), a popular benchmark for multi-label land cover classification. Specifically, we test adaptation between Sentinel-2 optical imagery and Sentinel-1 SAR modality, reflecting a common practical need when newer or complementary satellite sources become available (Tamazyan et al., 2025).

Figure 6: DELULU-NET achieves consistent improvements on the reBEN dataset across cross-satellite (Sentinel-1 and Sentinel-2) modality transfer, addition, and peeking settings.

As shown in Figure 6, DELULU-NET yields consistent gains across all three settings: modality transfer, addition, and peeking. These observations hold whether starting from Sentinel-1 or Sentinel-2 data and transferring, adding, or peeking at the other modality. This experiment serves as a first step to demonstrate that DELULU-NET generalizes beyond single-sensor scenarios, enabling the continual deployment of models trained on legacy satellite sources to new modalities, as well as the iterative refinement of existing models using unlabeled, paired, multi-modal satellite data.