# Adapting Semi-Supervised Segmentation methods to Multimodal Remote Sensing Data

Itza Hernandez-Sequeira[1], Damian Ibanez[1], Ruben Fernandez-Beltran[2], Filiberto Pla[1]

[1] Institute of New Imaging Technologies, University Jaume I, 12071 Castellón de la Plana, Spain - (isequeir, ibanezd, pla)@uji.es
[2] Dept. of Computer Science and Systems, University of Murcia, 30100 Murcia, Spain - rufernan@um.es

**Keywords:** Pseudo-labeling, Consistency Regularization, Contrastive Learning, Multimodal Fusion, Aerial Imagery.

**Abstract**

Remote sensing (RS) imagery is important for applications ranging from land cover and land use (LCLU) mapping to agriculture and forest monitoring. However, there is a limited availability of high-quality labeled data to use as a reference to train supervised learning (SL) models. Semi-supervised learning (SSL) frameworks, such as UniMatch (Yang et al., 2023), use pseudo-labeling and consistency regularization methods to address this limitation. Similar works have been adapted to RS: LSST (Lu et al., 2022) refines pseudo-labels with adaptive class-specific thresholds, while RS-DWL (Huang et al., 2024) mitigates noise and class imbalance through decoupled learning and confidence-based weighting. Despite these advances, SSL applications to multimodal RS imagery remain underexplored. We address this gap by adapting the SSL framework UniMatch to incorporate diverse encoders and multimodal remote sensing data for LCLU segmentation. We experimented on FLAIR-2 (Garioud et al., 2023), a dataset that combines very high-resolution aerial imagery (RGB) with near-infrared (NIR) data and elevation measurements (above-ground height). Key findings reveal that we achieved the best segmentation results using a transformer encoder for SL and SSL scenarios. When comparing RGB-only data and multimodal data, we observed that some classes, like "buildings", "water", and "coniferous", benefited from the inclusion of NIR and elevation information. In the semi-supervised experiments, where only half of the data was labeled, and the remaining half was used as unlabeled (simulating a real-world scenario), the multimodal SSL approach outperformed the fully supervised learning (FSL) approach using only the labeled subset (1/2). These results highlight the strong potential of data fusion in RS applications with limited labeled data.

## 1. Introduction

Land cover refers to the physical materials on the Earth's surface, like vegetation, bare soil, and water, while land use is used to describe human activities like agriculture and urban areas. Timely and consistent land cover and land use classification (LCLU) maps are essential to monitor rapid urbanization, deforestation, and agricultural expansion. Remote Sensing (RS) allows us to obtain images of the Earth -typically through satellites, aircraft, or drones- making it easier to perform large-scale and cost-effective LCLU (Read and Torrado, 2009). Long-term satellite missions like NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) have allowed the annual production of global land cover maps (e.g., MCD12Q1) since 2001, although at a coarse spatial resolution of 500 meters (Friedl and Sulla-Menashe, 2019). More recently, the European Space Agency (ESA) delivered WorldCover with a spatial resolution of 10 m based on Sentinel-1 and Sentinel-2 data (Zanaga et al., 2021). In parallel, countries strive to create country-scale submeter-level LCLU maps at 0.25–0.5 m spatial resolution derived from national aerial survey programs (Yokoya et al., 2024).

Deep Learning (DL) for image processing and data analysis has provided innovative solutions to detect and classify objects on Earth. Its ability to learn directly from raw data reduces reliance on manual feature engineering. Although raw data is generally abundant, insufficient high-quality human-labeled information makes classifying large reference datasets a persistent challenge. Some promising techniques that aim to bypass this barrier include transfer learning, self-supervised learning, semi-supervised learning, few-shot learning, active learning, weakly supervised learning, and more (Safonova et al., 2023).

In order to reduce the costly and labor-intensive process of manual pixel annotations on high-resolution imagery, some studies have developed workflows that use OpenStreetMap (OSM)(Wan et al., 2017), which provides information concerning roads and building footprints worldwide. Also, RS workflows often incorporate Vegetation indices (VI), such as the widely used Normalized Difference Vegetation Index (NDVI), to improve classification accuracy. VIs are combinations of red and near-infrared reflectance and help monitor green vegetation (Rojas et al., 2011). Furthermore, the fusion of RS images and digital surface model (DSM) data has the potential to detect changes in urban areas (Tian et al., 2022). Although multimodal data fusion has been widely explored to improve LCLU classification, its integration with SSL remains significantly underexplored, thus motivating our research.

This study addresses these gaps by integrating SSL with multimodal data and assessing its efficacy in remote sensing segmentation tasks. We adapt UniMatch (Yang et al., 2023), a state-of-the-art SSL framework, to process multimodal remote sensing data, enabling more effective segmentation by leveraging both labeled and unlabeled samples. To evaluate its effectiveness, we benchmark five encoder architectures, comparing their performance in supervised and semi-supervised settings. Our experiments demonstrate that SSL significantly improves segmentation accuracy, especially in scenarios with limited labeled data, underscoring its potential for large-scale remote sensing applications.

Our contributions are as follows:

- **Multimodal SSL Adaptation** – We extend UniMatch to incorporate multimodal inputs, integrating additional spectral bands (NIR) and elevation data to improve very

high-resolution land cover classification based on submeter aerial imagery. This integration enables richer feature learning for land cover segmentation and represents one of the first applications of UniMatch to multimodal RS data.

- **Comprehensive Encoder Benchmarking** – We perform an extensive comparison of five encoder architectures (ResNet, EfficientNet, and MiT variants) across supervised and semi-supervised settings, identifying MiT-B2 as the most effective model for multimodal segmentation.
- **SSL Performance in Label-Scarce Scenarios** – We assess SSL performance under different labeled-to-unlabeled data ratios (50%, 25%, and 12.5%) to evaluate how SSL scales with reduced supervision. Our results demonstrate that SSL significantly boosts segmentation accuracy, maintaining high segmentation accuracy even when labeled data is minimal.

These findings highlight the potential of SSL and multimodal data fusion for remote sensing applications, providing a scalable approach to LCLU in label-constrained scenarios.

## 2. Related work

Firstly, we investigate recent progress in image segmentation, followed by a discussion on semi-supervised semantic segmentation, particularly emphasizing remote sensing images.

### 2.1 Deep Learning for Image Segmentation

Deep learning has significantly advanced image classification, with convolutional neural networks (CNNs) forming the backbone of many state-of-the-art models. **ResNet** (He et al., 2016) introduced deep residual learning to address vanishing gradients, enabling the training of very deep networks, while **EfficientNet** (Tan and Le, 2019) proposed a compound scaling approach that optimally balances network depth, width, and resolution, improving both accuracy and efficiency.

Semantic segmentation extends image classification from image-level to pixel-wise predictions. Fully convolutional networks (FCNs) (Long et al., 2015) pioneered this task by eliminating fully connected layers, enabling end-to-end dense classification. DeepLab (Chen et al., 2018a) improved upon FCN by introducing atrous (dilated) convolutions, which expand the receptive field without increasing computational cost, allowing multi-scale context capture. To further enhance segmentation, DeepLab introduced atrous spatial pyramid pooling (ASPP), which applies multiple parallel atrous convolutions with different rates, effectively capturing objects and contextual information at multiple scales. Another major approach to semantic segmentation involves encoder-decoder architectures such as **U-Net** (Ronneberger et al., 2015), which gradually recover fine-grained spatial details. Initially designed for biomedical image segmentation, U-Net introduced skip connections that preserve high-resolution features, making it highly effective for remote sensing and other segmentation tasks. **DeepLabV3+** (Chen et al., 2018b) incorporates elements from both spatial pyramid pooling and encoder-decoder structures, using a DeepLabV3 encoder with a lightweight decoder module to recover spatial details lost in deep feature extraction, improving boundary delineation.

Recent methods have proved the effectiveness of transformer-based architectures. **Vision transformers** (ViTs) (Kolesnikov et al., 2021) achieved state-of-the-art performance in image classification by utilizing self-attention for global feature extraction. However, traditional ViTs lack the inductive biases of CNNs, making them data-hungry and computationally expensive. Hybrid approaches such as SegFormer (Xie et al., 2021) addressed this by integrating a hierarchical transformer encoder with a lightweight multi-layer perceptron (MLP) decoder, achieving efficient and accurate segmentation. It employs mix transformers (MiT), a series of encoders designed to generate multi-level hierarchical features crucial for dense prediction. These advancements highlight the ongoing shift from purely convolutional architectures to hybrid and transformer-based models, significantly improving semantic segmentation across various domains, including remote sensing.

### 2.2 Semi-supervised semantic segmentation

Semantic segmentation requires labeling each pixel in an image, which makes the task very time-consuming and labor-intensive. Semi-supervised learning (SSL) helps solve this problem by using a small set of labeled examples along with a large amount of unlabeled data to improve segmentation performance. Recent progress in SSL for segmentation is primarily classified into two methodologies: consistency regularization and pseudo-labeling (Huang et al., 2024).

**Consistency regularization** ensures that model predictions remain stable under image, feature, and network perturbations by applying a regularization term to the final loss (e.g., cross-entropy). CutMix (Yun et al., 2019) is an image perturbation technique that enhances generalization by cutting and pasting patches between images while mixing their labels proportionally. Whereas cross-consistency training (CCT) (Ouali et al., 2020) enforces consistency by training auxiliary decoders on perturbed inputs and aligning their predictions with the main decoder.

**Pseudo-labeling and self-training** iteratively refines high-confidence predictions on unlabeled data to enhance learning. ST++ (Yang et al., 2022) improves traditional self-training by using strong data augmentations (SDA) to reduce overfitting and selectively retraining on reliable unlabeled images. U2PL (Wang et al., 2022) ensures that all pixels contribute to training by classifying low-confidence predictions as negative samples rather than discarding them. It dynamically adjusts reliability thresholds, adapting to model evolution.

Hybrid methodologies enhance model robustness by integrating multiple techniques. **FixMatch** (Sohn et al., 2020) employs weak-to-strong consistency regularization, generating pseudo-labels from weakly augmented images and enforcing consistency on strongly perturbed versions when predictions exceed a 0.95 confidence threshold. **CoMatch** (Li et al., 2021) extends FixMatch by using weakly augmented images to generate pseudo-labels, which then guide class predictions on two strongly augmented versions. It further integrates contrastive and graph-based learning, leveraging InfoNCE loss (Oord et al., 2019) to enforce consistency across augmented views while improving feature separation across categories.

Although FixMatch and CoMatch's primary goal is image classification, recent SSL frameworks incorporate these into image segmentation. The Unified Dual-Stream Perturbations approach (**UniMatch**) (Yang et al., 2023) builds upon FixMatch

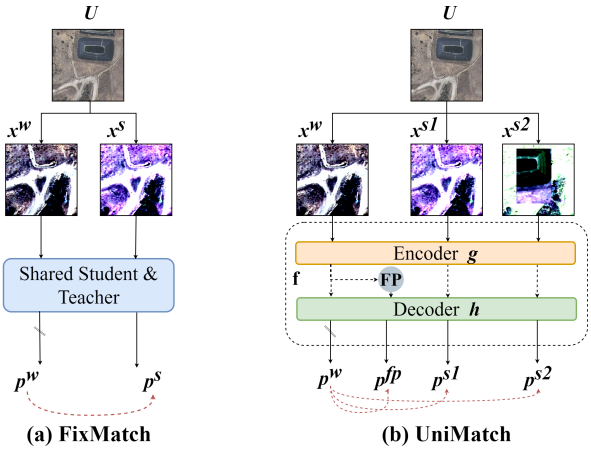


Figure 1. Semi-supervised Semantic Segmentation Framewoks: (a) FixMatch and (b) UniMatch.

by introducing dual-stream perturbations at both the image level and the feature level. It enforces two strong augmented data to a single weakly augmented data and uses unlabeled data for representation learning. In figure 1, we can observe a comparison between FixMatch and UniMatch. Given an unlabeled image $U$, for UniMatch, there are four potential outcomes. First, they apply a **weak augmentation** $x^w$ to resize, randomly crop, and flip the image. Then, they apply a **feature perturbation** $x^{fp}$ to the weakly augmented sample by introducing a channel dropout with a 50% probability between the encoder and decoder. Finally, they apply two **strong augmentations** $x^{s1}$ and $x^{s2}$: one using the ST++ augmentation strategy and the other using CutMix. Finally, they use consistency regularization between the prediction $p^w$ of $x^w$ and the predictions $p^{fp}$, $p^{s1}$, and $p^{s2}$ from its augmented versions $x^{fp}$, $x^{s1}$ and $x^{s2}$, respectively. In the case of FixMatch, there's only one weak augmentation $x^w$ and one strong augmentation $x^{s1}$, and the consistency regularization is performed between $p^w$ and $x^s$.

### 2.3 Remote sensing Semantic Image Segmentation

Some works apply SSL in the remote sensing domain and coin the term remote sensing semi-supervised semantic segmentation (RS-SSS) (Huang et al., 2024).

**Linear sampling selftraining** (LSST) (Lu et al., 2022), whose work is based on (Yang et al., 2022), investigated that the application of strong data augmentations (SDA) to RS images tends to corrupt data distribution and impair model performance. Therefore, they constructed SDA applicable to RS images from the following three aspects: color transformation (CT), geometric transformation (GT), and CutOut.When comparing a baseline model without augmentations and models using different SDAs on unlabeled images, they found that using only CT led to a 1.15 improvement in mIoU. Similarly, using GT alone boosted the mIoU by 1.37. However, the best results were achieved when combining all three augmentations —CT, GT, and Cutout—resulting in the largest gain of 1.91 mIoU.

Other studies investigate pseudo-label confidence in remote sensing data because performance can degrade due to the inevitable memorization of wrong pseudo-labeling of unlabeled data. **Decoupled weighting learning** (DWL) (Huang et al., 2024) proposes two modules: decoupled learning and ranking weighting. During training, the decoupled learning module separates the predictions of the labeled and unlabeled data to decrease the negative impact of the self-training of the wrongly pseudo-labeled unlabeled data on the supervised training of the labeled data.

Because most of the studies on SSL are based on aerial datasets for image classification, recent works have advanced RS-SSS by creating datasets with dedicated partitions for labeled and unlabeled data. **MiniFrance** (Castillo-Navarro et al., 2022) was the first aerial semi-supervised dataset for semantic segmentation, but it only contemplated RGB data. The Data Fusion Contest (DFC2022) (Hänsch et al., 2022) extended the dataset by incorporating elevation bands. Nevertheless, the absence of test set labels limits its utility for benchmarking. More recently, the **FLAIR-2** dataset (Garioud et al., 2023) integrates very high-resolution (0.2m) mono-temporal aerial imagery with high-resolution Sentinel-2 satellite time series for multimodal remote sensing image segmentation.

## 3. Methodology

This section describes the remote sensing dataset used in our study, focusing on the train/validation/test splits and the ratio of labeled to unlabeled data. We also outline the semi-supervised learning (SSL) training framework, including details of the encoder-decoder strategy and the configurations required for our experiments.

### 3.1 Training splits for supervised and semi-supervised

For semi-supervised learning (SSL) experiments, we created labeled and unlabeled partitions of the training data, following SSL protocols such as those in Unimatch (Yang et al., 2023) and LSST (Lu et al., 2022). We experimented with three **labeled-to-unlabeled ratios**: 50% labeled (1/2 split), where half of the training dataset is labeled while the other half remains unlabeled; 25% labeled (1/4 split); and 12.5% labeled (1/8 split), maximizing the use of unlabeled data.

In the fully supervised learning (FSL) scenario, we use all available labeled samples (100%). However, we also conduct experiments using 50%, 25%, and 12.5% of the dataset in a supervised manner, without an unlabeled portion. When discussing baselines, we differentiate between **Baseline_SL (supervised learning)**, which refers to training with 100%, 50%, 25%, or 12.5% of the data in a supervised manner without any unlabeled samples, and **Baseline_SSL (semi-supervised learning)**, which follows a semi-supervised approach.

### 3.2 SSL framework and encoder-decoder strategy

**Framework.** We employed the Unified Dual-Stream Perturbations approach (UniMatch) (Yang et al., 2023), an SSL framework designed for tasks like image segmentation.

The framework algorithm utilizes one weakly augmented sample ($x_w$), and two strongly augmented samples ($x_{s1}$ and $x_{s2}$). Following the recommended configuration from (Yang et al., 2023), the weakly augmented sample is generated by resizing the raw unlabeled image $x$ within a scaling range of 0.5 to 2.0 for both height and width. For example, when resizing images with an original size of $512 \times 512$, this modification results in new dimensions between $256 \times 256$ and $1024 \times 1024$. The resized images are then randomly cropped to $128 \times 128$ and flipped to produce the weakly augmented sample $x_w$. This smaller crop reduces GPU memory usage and accelerates training.

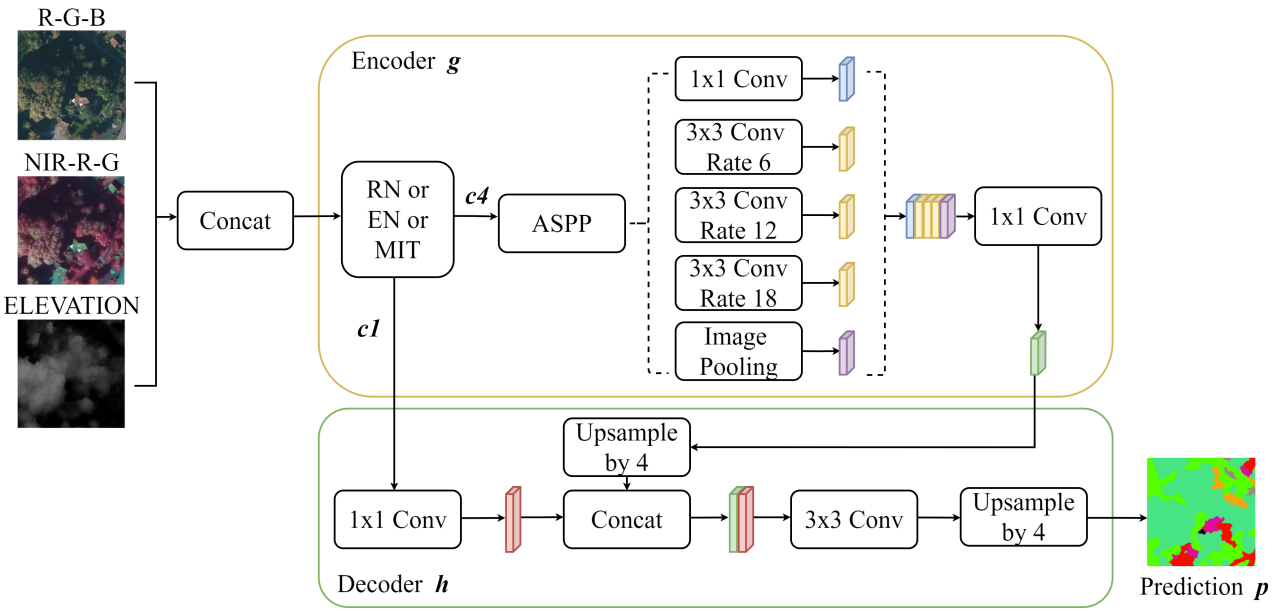


Figure 2. Multimodal encoder (ResNet (RN), EfficientNet (EN) or Mix Transformer (MIT) with DeepLabV3+ decoder

For strong augmentations, UniMatch adopts the strategy proposed in ST++ (Yang et al., 2022), incorporating color transformations along with CutMix (Yun et al., 2019) to enhance model robustness. The augmentation pipeline includes color jittering to introduce variations in brightness, contrast, saturation, and hue, as well as Gaussian blur, applied with a 50% probability. Additionally, CutMix is employed by randomly selecting a rectangular patch within an image and replacing it with a corresponding patch from another image. These transformations introduce variations in color, texture, and spatial structure, this diversity helps in improving the model's generalization across diverse data distributions.

The toolbox includes a segmentation model $f$, which can be decomposed into an encoder $g$ and a decoder $h$, both of which have been modified as described in the following sections.

**Encoder ($g$).** We integrated the Segmentation Models PyTorch (SMP) (Iakubovskii, 2019) into the UniMatch toolbox. This modification enabled the use of advanced encoders like EfficientNet and Mix Vision Transformers. We conducted experiments on the FLAIR-2 dataset (13 classes) to compare RGB-only and multimodal configurations across five encoders: ResNet50 (RN50), EfficientNet-B3 (ENB3), EfficientNet-B4 (ENB4), MIT-B2, and MIT-B3, with DeepLabV3+ as the decoder.

To adapt UniMatch for multimodal SSL, we modified the first convolutional layer to accept five input channels: RGB, near-infrared (NIR), and elevation. We adjusted augmentation pipelines to ensure spatial transformations were uniformly applied across all bands while color transformations were restricted to RGB. Preprocessing included min-max normalization for NIR and elevation bands and standard ImageNet normalization for RGB.

**Decoder ($h$).** Unimatch uses DeepLabV3+ (Chen et al., 2018b) as its segmentation model. The high-level features (c4) extracted from the encoder move directly to the ASPP Module while the low-level features (c1) go directly to the decoder, as can be appreciated in 2.

We recreated the FLAIR-2 baseline for both supervised (_SL) and semi-supervised (_SSL) learning with our training splits. The Base_SL was reproduced using ResNet34 (RN34) with a U-Net decoder (Ronneberger et al., 2015). For the Base_SSL, we incorporated modifications to the U-Net, including the feature dropout between the encoder and the decoder as done by (Yang et al., 2023). Feature perturbation is applied at the encoder-decoder intersection for the weakly augmented images. It is implemented as channel dropout with a 50% probability in Pytorch.

### 3.3 Implementation details

**Dataset.** In this study, we used the FLAIR-2 dataset (Garioud et al., 2023), focusing exclusively on the georeferenced multimodal aerial imagery, which consists of five spectral bands: blue, green, red, near-infrared, and elevation. The dataset consists of 77,762 image patches (512×512 pixels) annotated into 12 classes, plus an "other" category that we excluded from metric computation due to its low representation (<1% of the dataset). The training set contains 61,712 aerial imagery patches, while the official test set includes 16,050 samples used exclusively for inference. We followed the FLAIR-2 suggested train/test split to maintain benchmark consistency. We partitioned the training data into 80% (51,097 samples) for training and 20% (10,615 samples) for validation while keeping the official test dataset separate, as shown in table 1

| Split | Labeled | Unlabeled | Validation |
|---|---|---|---|
| FSL (100%) | 51,097 | 0 | 10,615 |
| SSL 1/2 (50%) | 25,450 | 25,647 | 10,615 |
| SSL 1/4 (25%) | 12,775 | 38,322 | 10,615 |
| SSL 1/8 (12.5%) | 7,000 | 44,097 | 10,615 |

Table 1. Fully supervised learning (FSL) samples and the labeled-unlabeled splits used for semi-supervised learning (SSL) experiments.

**Configuration** To ensure a fair comparison, we doubled the batch size to 20 for supervised experiments while keeping it at 10 for semi-supervised training. Since UniMatch applies two

strong augmentations to each image, the larger batch size is justified in supervised settings. We adopted training parameters from the Cityscapes setup, using an initial learning rate of 0.005 with the SGD optimizer and online hard example mining (OHEM) loss (Shrivastava et al., 2016). We trained each model for 100 epochs with a crop size of 128.

We evaluated both supervised and semi-supervised settings using mean Intersection over Union (mIoU) as the primary performance metric, which quantifies the average overlap between predicted and ground truth segmentation masks across all classes.

## 4. Results

We evaluated the encoder's performance when using RGB versus multimodal inputs across two frameworks: supervised and semi-supervised, considering different labeled-to-unlabeled data ratios.

### 4.1 Encoder evaluation on RGB and Multimodal data

Based on the evaluation of five different encoders (table 2), we can observe that models using multimodal imagery consistently outperform their RGB-only counterparts, particularly in semi-supervised settings with limited labeled data. Among the tested encoder architectures, MiT-B2 demonstrated the highest overall mean Intersection over Union (mIoU), achieving 56.28% in fully supervised learning (100% of labels) and 55.65% in the semi-supervised 1/2 scenario (50% labels). When comparing the FSL scenarios, MiT-B2 outperforms the baseline by +3.65 percentage points. In the most label-scarce setting (1/8 - 12.5%), ENB4 ranked second with 45.68 mIoU and was the top RGB-only performer. Additionally, while MiT-B3 demonstrated strong performance, MiT-B2 was ultimately preferred due to its balance between accuracy and computational efficiency.

| Type | Encoder | FSL | 1/2 | 1/4 | 1/8 |
|---|---|---|---|---|---|
| RGB | RN50 | 52.70 | 50.82 | 50.29 | 42.89 |
| | MITB2 | **54.37** | 52.76 | 48.33 | 41.59 |
| | MITB3 | 52.43 | 53.20 | 48.61 | 40.32 |
| | ENB3 | 49.44 | 48.55 | 46.28 | 38.38 |
| | ENB4 | 51.01 | 48.89 | 47.94 | 44.30 |
| Multi | Base_SL | 52.63 | 50.61 | 46.20 | 40.57 |
| | Base_SSL | - | 50.13 | 47.37 | 30.00 |
| | RN50 | 54.68 | 54.14 | 49.10 | 42.05 |
| | MITB2 | **56.28** | 55.65 | 52.69 | 45.98 |
| | MITB3 | 55.60 | 54.36 | 52.62 | 45.13 |
| | ENB3 | 53.94 | 49.77 | 47.37 | 42.94 |
| | ENB4 | 52.24 | 51.37 | 45.31 | 45.68 |

Table 2. Performance comparison of mIoU on the test dataset for RGB and Multimodal under different supervision levels.

Note: Experiments trained with a crop of 128. The baseline is an RN34 encoder with a U-Net decoder, while all others use DeepLabV3+.

We observed that for both baseline models that use RN34 with a U-Net, the semisupervised (Base_SSL), which included a feature dropout at the intersection between encoder and decoder, underperformed compared to the supervised Base_SL under diverse labeled-to-unlabeled data ratios. We theorize that this is due to U-Net's reliance on skip connections for detailed feature propagation, where dropout disrupts spatial consistency. Meanwhile, DeepLabV3+ benefits from such regularization due to the atrous spatial pyramid pooling (ASPP) module's ability to retain multi-scale context.

Overall, multimodal data proves highly beneficial, especially in low-label settings, helping models retain higher performance compared to RGB-only models. Semi-supervised training configurations outperformed the supervised baseline (Base_SL) when labeled data was limited (1/2, 1/4, 1/8). This highlights the value of semi-supervised learning, particularly in real-world applications where labeled data is limited.

### 4.2 Quantitative and qualitative analysis of the best model

**Quantitative.** The table 3 compares the Intersection over Union (IoU) scores per class of the best model MiT-B2 in a fully supervised training (FSL) for RGB and multimodal. Then, we also include its performance for semi-supervised learning (SSL) with diverse labeled-to-unlabeled data splits but only using multimodal data. Additionally, we define **Δ** RGB-MM as the difference between class IoU from RGB and Multimodal and **Δ** SSL as the change between the lowest SSL setting (1/8) and the highest (1/2). It's important to mention that "other" class is excluded from the final mIoU calculation, following FLAIR-2 guidelines.

| Class | RGB | Multimodal | | | | Δ | Δ |
|---|---|---|---|---|---|---|---|
| | FSL | FSL | 1/2 | 1/4 | 1/8 | RGB-MM | SSL |
| Building | 73.52 | 81.64 | 80.46 | 74.24 | 76.84 | **+8.12** | -3.62 |
| Pervious | 50.08 | 48.54 | 47.89 | 44.49 | 39.94 | -1.54 | -7.95 |
| Impervious | 68.3 | 70.84 | 70.47 | 66.68 | 66.18 | +2.54 | -4.29 |
| Bare soil | 49.81 | 52.77 | 51.86 | 38.33 | 25.11 | +2.96 | -26.75 |
| Water | 78.09 | 83.93 | 86.44 | 79.51 | 80.70 | **+5.84** | -5.74 |
| Coniferous | 44.98 | 52.98 | 52.58 | 57.68 | 34.93 | **+8.00** | -17.65 |
| Deciduous | 64.86 | 68.90 | 69.57 | 68.41 | 59.30 | +4.04 | -10.27 |
| Brushwood | 19.68 | 19.45 | 21.60 | 21.53 | 19.45 | -0.23 | -2.15 |
| Vineyard | 63.6 | 60.85 | 52.09 | 48.57 | 39.66 | -2.75 | -12.43 |
| Herbaceous | 49.03 | 47.06 | 44.83 | 44.57 | 34.65 | -1.97 | -10.18 |
| Agricultural | 57.67 | 54.70 | 52.94 | 53.25 | 46.35 | -2.97 | -6.59 |
| Plowed land | 32.81 | 33.67 | 37.03 | 35.02 | 28.66 | +0.86 | -8.37 |
| Other | 10.79 | 12.65 | 9.68 | 10.19 | 1.00 | +1.86 | -8.68 |
| **mIoU** | 54.37 | **56.28** | 55.65 | 52.69 | 45.98 | +1.91 | -9.67 |

Table 3. IoU per class for MiT-B2+DLV3+ considering RGB and multimodal data and under fully supervised and semi-supervised learning (1/2 to 1/8).

The three classes that had the most gains between RGB and Multimodal were "Building" (+8.12), "Coniferous" (+8.00), and "Water" (+5.84). This gain is expected as the elevation channel provides information on the above-ground height, which is highly useful for distinguishing vertical structures like buildings and trees (e.g., coniferous), which have characteristic height profiles that RGB alone cannot capture (Wang et al., 2021, Tian et al., 2022). On the other hand, water surfaces are flat and typically return near-zero elevation values. While traditional water indices such as NDWI rely on SWIR bands, these are often absent in high-resolution imagery, limiting their effectiveness. Instead, bands B1 (R), B2 (G), and B4 (NIR) allow for better spectral distinction, as water typically appears dark and saturated due to its low reflectivity (Chen et al., 2018c).

Moving on, we compare the performance of MiT-B2 in the semi-supervised training with varying labeled-to-unlabeled data ratios (1/2, 1/4, 1/8). When comparing the reduction from 1/8 against the 1/2 scenario (**Δ** SSL): "bare soil" (-26.75), "coniferous" (-17.65) and "vineyard" (-12.43) experienced the

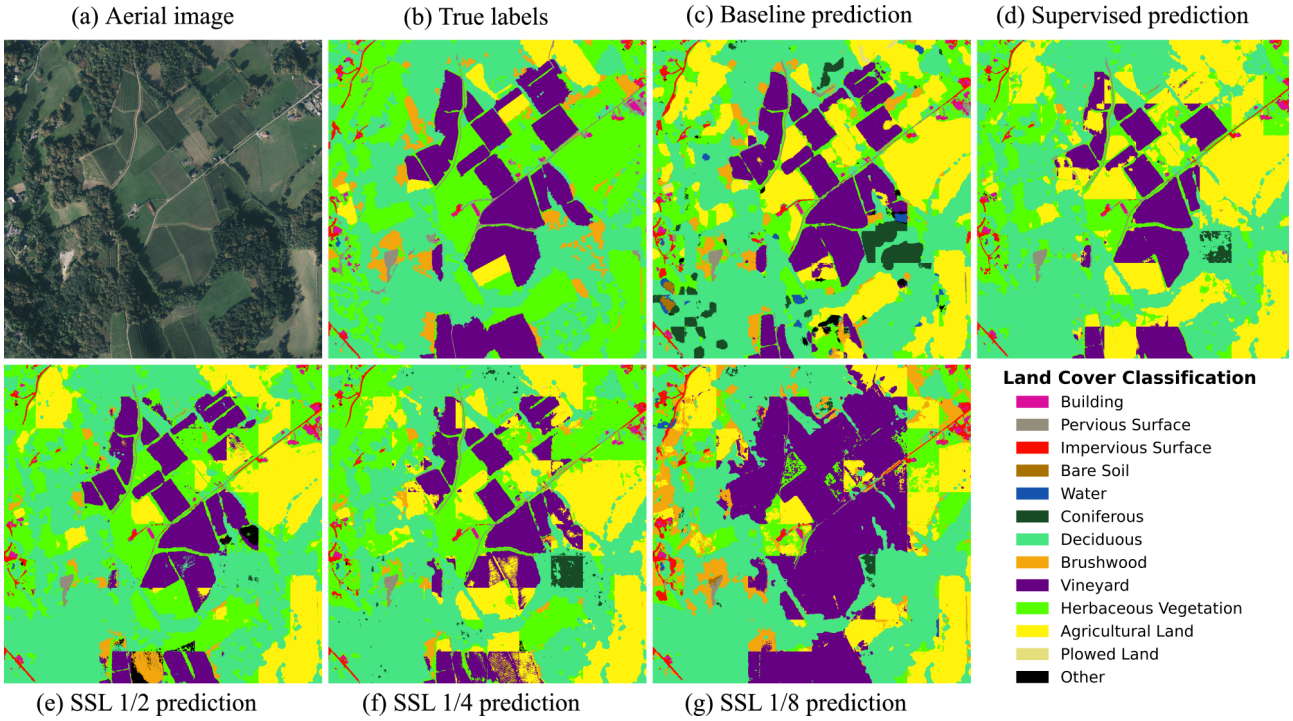


Figure 3. Predictions over test set region D064-Z1-AA (10x10 patches of 512x512) using Supervised baseline RN34+UNet(c) and MITB2+DLV3+ on Supervised (d) and semi-supervised scenarios (e-g).

largest declines. This demonstrates that some of the classes are particularly sensitive to data scarcity. Part of this sensitivity can be attributed to the variability introduced by sample selection in SSL, where the quality of the examples received from the labeled set can condition the training. Additionally, the FLAIR-2 dataset itself reflects this imbalance: coniferous represents only 2.74% of the labeled training data, followed by vineyard (3.13%), bare soil (3.47%), and plowed land (3.88%). Therefore, reducing the labeled data further disproportionately affects these already underrepresented classes, amplifying the challenge for the model to learn robust representations for them.

**Qualitative.** Based on Figure 3, the fully supervised MiT-B2 with DLV3+ model (d) provides a more accurate segmentation than the reproduced baseline RN34+UNet (c). Nevertheless, both models show confusion between "herbaceous vegetation" and "agricultural land". This overlap can be partly attributed to the definition of the classes, where "herbaceous vegetation" includes non-cultivated grass in agricultural areas, and "agricultural land" also includes permanent and temporary grasslands with agricultural use (Garioud et al., 2023).

In contrast, semi-supervised models (e-g, from figure 3) progressively lose detail as labeled data decreases. SSL 1/2 (e) retains much of the supervised model's accuracy, while SSL 1/8 (g) introduces noticeable noise and misclassification. "Buildings" (vivid magenta) remain well-preserved but slightly degrade in SSL 1/8. As supervision diminishes, class boundaries blur, partly due to class definitions and overlaps. For example, "vineyards", while it is an agricultural use, are treated as a distinct class due to their unique land cover structure. However, they represent only a tiny portion of the dataset (3.13%), making them more prone to misclassification, especially in low-label regimes.

## 5. Conclusions

This study demonstrates the effectiveness of semi-supervised learning (SSL) in multimodal land cover classification, particularly in scenarios with limited labeled data. We enhance segmentation accuracy through multimodal fusion by adapting UniMatch to process aerial imagery, near-infrared (NIR), and elevation data. Among the five encoder architectures evaluated, MiT-B2 consistently achieved the highest segmentation accuracy, striking a balance between computational efficiency and performance. The semi-supervised models outperformed their supervised counterparts when labeled data was limited, confirming the value of SSL in remote sensing applications.

Future work should explore extending SSL techniques to additional remote sensing datasets and refining augmentation strategies to enhance model robustness. Additionally, integrating advanced self-training methods and contrastive learning could improve performance, particularly in challenging environments with highly imbalanced class distributions. By continuing to develop and optimize SSL frameworks for remote sensing, we can make large-scale land cover classification more efficient, accurate, and accessible in real-world applications.

## Acknowledgments

This work was partially supported by the Ministerio de Ciencia e Innovación (project PID2021-128794OB-I00), the Generalitat Valenciana (project CIAICO-2023-032), and the University Jaume I (PREDOC/2020/50).

## References

Castillo-Navarro, J., Le Saux, B., Boulch, A., Audebert, N., Lefèvre, S., 2022. Semi-supervised semantic segmentation in

Earth Observation: the MiniFrance suite, dataset analysis and multi-task network study. *Machine Learning*, 111(9), 3125–3160. https://doi.org/10.1007/s10994-020-05943-y.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018a. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(04), 834-848. https://doi.ieeecomputersociety.org/10.1109/TPAMI.2017.2699184.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, Springer-Verlag, Berlin, Heidelberg, 833–851.

Chen, Y., Fan, R., Yang, X., Wang, J., Latif, A., 2018c. Extraction of Urban Water Bodies from High-Resolution Remote-Sensing Imagery Using Deep Learning. *Water*, 10(5). https://www.mdpi.com/2073-4441/10/5/585.

Friedl, M., Sulla-Menashe, D., 2019. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. `https://doi.org/10.5067/MODIS/MCD12Q1.006`. Distributed by NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed 2025-04-07.

Garioud, A., Gonthier, N., Landrieu, L., Wit, A. D., Valette, M., Poupée, M., Giordano, S., Wattrelos, B., 2023. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems (NeurIPS) 2023*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, W., Shi, Y., Xiong, Z., Zhu, X. X., 2024. Decouple and weight semi-supervised semantic segmentation of remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212, 13-26.

Hänsch, R., Persello, C., Vivone, G., Castillo Navarro, J., Boulch, A., Lefèvre, S., Le Saux, B., 2022. 2022 IEEE GRSS Data Fusion Contest: Semi-Supervised Learning [Technical Committees]. *IEEE Geoscience and Remote Sensing Magazine*.

Iakubovskii, P., 2019. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`.

Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., Zhai, X., 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Li, J., Xiong, C., Hoi, S. C. H., 2021. Comatch: Semi-supervised learning with contrastive graph regularization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9455–9464.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Lu, X., Jiao, L., Liu, F., Yang, S., Liu, X., Feng, Z., Li, L., Chen, P., 2022. Simple and Efficient: A Semisupervised Learning Framework for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–16.

Oord, A. v. d., Li, Y., Vinyals, O., 2019. Representation Learning with Contrastive Predictive Coding.

Ouali, Y., Hudelot, C., Tami, M., 2020. Semi-supervised semantic segmentation with cross-consistency training. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12671–12681.

Read, J., Torrado, M., 2009. Remote sensing. R. Kitchin, N. Thrift (eds), *International Encyclopedia of Human Geography*, Elsevier, Oxford, 335–346.

Rojas, O., Vrieling, A., Rembold, F., 2011. Assessing drought probability for agricultural areas in Africa with coarse resolution remote sensing imagery. *Remote Sensing of Environment*, 115(2), 343-352.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 234–241.

Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., Ryo, M., 2023. Ten deep learning techniques to address small data problems with remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 125, 103569.

Shrivastava, A., Gupta, A., Girshick, R., 2016. Training Region-Based Object Detectors with Online Hard Example Mining . *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos, CA, USA, 761–769.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., Raffel, C., 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv preprint arXiv:2001.07685*.

Tan, M., Le, Q. V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, ICML, 97, 6105–6114.

Tian, S., Zhong, Y., Ma, A., Zhang, L., 2022. Three-Dimensional Change Detection in Urban Areas Based on Complementary Evidence Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-13.

Wan, T., Lu, H., Lu, Q., Luo, N., 2017. Classification of High-Resolution Remote-Sensing Image Using OpenStreetMap Information. *IEEE Geoscience and Remote Sensing Letters*, 14(12), 2305-2309.

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X., 2022. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4238-4247. https://api.semanticscholar.org/CorpusID:247315180.

Wang, Y., Zhang, X., Guo, Z., 2021. Estimation of tree height and aboveground biomass of coniferous forests in North China using stereo ZY-3, multispectral Sentinel-2, and DEM data. *Ecological Indicators*, 126, 107645.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. Segformer: simple and efficient design for semantic segmentation with transformers. *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Curran Associates Inc., Red Hook, NY, USA.

Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y., 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. *CVPR*.

Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y., 2022. St++: Make self-training work better for semi-supervised semantic segmentation. *CVPR*.

Yokoya, N., Xia, J., Broni-Bediako, C., 2024. Submeter-level land cover mapping of Japan. *International Journal of Applied Earth Observation and Geoinformation*, 127, 103660.

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., Choe, J., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6022–6031.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, L., Tsendbazar, N.-E., Ramoino, F., Arino, O., 2021. ESA WorldCover 10 m 2020 v100.