



Multi-modal co-learning for Earth observation: enhancing single-modality models via modality collaboration

Francisco Mena, Dino Ienco, Cassio F Dantas, Roberto Interdonato, Andreas Dengel

► To cite this version:

Francisco Mena, Dino Ienco, Cassio F Dantas, Roberto Interdonato, Andreas Dengel. Multi-modal co-learning for Earth observation: enhancing single-modality models via modality collaboration. Machine Learning, 2025, 114 (12), pp.279. <10.1007/s10994-025-06903-0>. <hal-05326486>

HAL Id: hal-05326486

<https://hal.science/hal-05326486v1>

Submitted on 22 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

MULTI-MODAL CO-LEARNING FOR EARTH OBSERVATION: ENHANCING SINGLE-MODALITY MODELS VIA MODALITY COLLABORATION

ACCEPTED AT MACHINE LEARNING JOURNAL

Francisco Mena^{1,2} , Dino Ienco^{3,5} , Cassio F. Dantas^{3,5}  Roberto Interdonato^{4,5}  Andreas Dengel^{1,2} 

¹Department of Computer Science, University of Kaiserslautern-Landau (RPTU), Kaiserslautern, Germany

²SDS, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany

³INRAE, UMR TETIS, University of Montpellier, Montpellier, France

⁴CIRAD, UMR TETIS, University of Montpellier, Montpellier, France

⁵INRIA, EVERGREEN, University of Montpellier, Montpellier, France

f.menat@rptu.de

ABSTRACT

Multi-modal co-learning is emerging as an effective paradigm in machine learning, enabling models to collaboratively learn from different modalities to enhance single-modality predictions. Earth Observation (EO) represents a quintessential domain for multi-modal data analysis, wherein diverse remote sensors collect data to sense our planet. This unprecedented volume of data introduces novel challenges. Specifically, the access to the same sensor modalities at both training and inference stages becomes increasingly complex based on real-world constraints affecting remote sensing platforms. In this context, multi-modal co-learning presents a promising strategy to leverage the vast amount of sensor-derived data available at the training stage to improve single-modality models for inference-time deployment. Most current research efforts focus on designing customized solutions for either particular downstream tasks or specific modalities available at the inference stage. To address this, we propose a novel multi-modal co-learning framework capable of generalizing across various tasks without targeting a specific modality for inference. Our approach combines contrastive and modality discriminative learning together to guide single-modality models to structure the internal model manifold into modality-shared and modality-specific information. We evaluate our framework on four EO benchmarks spanning classification and regression tasks across different sensor modalities, where only one of the modalities available during training is accessible at inference time. Our results demonstrate consistent predictive improvements over state-of-the-art approaches from the recent machine learning and computer vision literature, as well as EO-specific methods. The obtained findings validate our framework in the single-modality inference scenarios across a diverse range of EO applications.

Keywords Multi-modal data · Co-learning · Earth observation · Multi-loss · Representation learning · Sensor data.

1 Introduction

The collaborative learning paradigm, named co-learning, has been largely studied in the machine learning field [1, 2]. The objective is to have multiple models (or layers) that share knowledge and cooperate with each other to improve their performance. This learning paradigm has been adopted for domain adaptation [3, 4], federated learning [5], learning with noisy labels [2], and knowledge distillation [6]. Recently, with the increasing availability of multi-modal data acquired through a plethora of different platforms and sensors, co-learning can play an even more crucial role [7].

Multi-modal data describing the same phenomena of interest have been used for different purposes in the literature. Multi-modal learning targets the effective exploitation of multiple data modalities for improving model performance and supporting informed decisions [8]. Recent strategies rely on data fusion mechanisms to improve accuracy or to

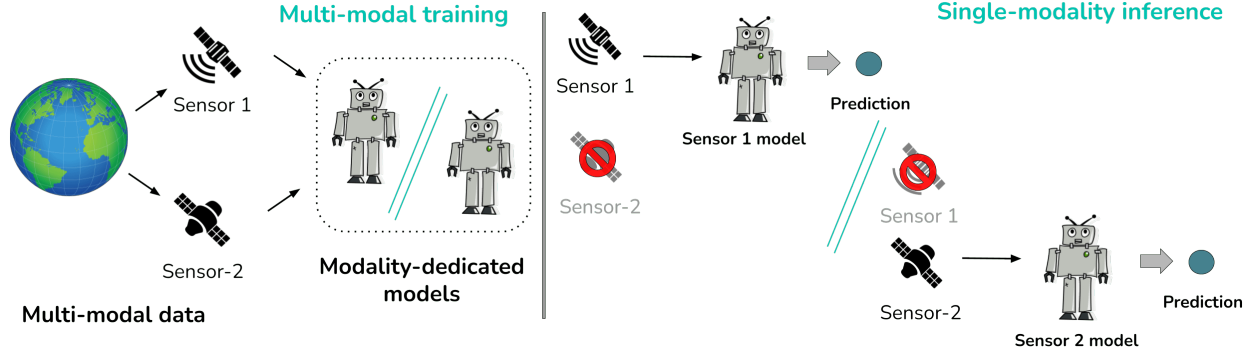


Figure 1: Illustration of multi-modal co-learning. The multi-modal data is available for training, but single-modality data is accessible for inference as an all-but-one missing modality scenario.

enhance synergy between modalities via self-supervision. However, we consider multi-modal data under a co-learning paradigm in this work, focusing on the scenario where only one of the modalities available at training is accessible at inference time [9], referred to as *all-but-one missing modality* [7]. More precisely, we do not make any assumption on which modality is accessible at inference time, as shown in Fig. 1. To address this point, Andrew et al. [10] learn a shared feature space between modalities by maximizing the cross-covariance, while Zhang et al. [11] share the last linear layer among single-modality models to force sharing information between modalities. The benefit of learning from multi-modal data has been demonstrated empirically for single-modality inference, as well as studied theoretically by Zadah et al. [12]. Thus, multi-modal co-learning has been applied to different domains where only a subset of the modalities is accessible at the inference stage, such as in Earth Observation (EO).

Nowadays, the EO domain is characterized by a vast amount of heterogeneous multi-modal data. Thanks to the advances in instruments and technology, numerous satellites are constantly sensing the Earth’s surface [13]. Such multi-modal data requires advanced methods to take advantage of the carried complementary information [14]. This is because the collected sensor data is diverse and heterogeneous due to: different acquisition modes (e.g. optical and radar), spectral characteristics (e.g. RGB and multi-spectral), and resolutions (e.g. spatial and temporal). Thus, EO sensor data go beyond standard benchmarks that mainly cover natural images in the standard computer vision domain [15], limiting the applicability of mainstream machine learning and vision models to the EO domain. Moreover, accessing the same set of sensor data during both the training and inference stages could be infeasible in scenarios characterized by operational constraints, resulting in missing modality scenarios [9].

The lack of systematically available sensor data covering the same region over the same period is an inherent problem in the EO domain [16]. This is because sensor data collection occurs under operational constraints in real-world environments, where geographical extent, weather acquisition conditions, deployment costs, and sensor failures may affect its consistent and systematic availability. Thus, sensor modalities in multi-modal scenarios can be partially or entirely missing during inference. For instance, the Landsat 7 ETM+ SLC-off problem after 2003 [17], the Sentinel-1b satellite that stopped operating at the end of 2021 [18], and the NAIP satellite that operates only in the United States. Thus, enhanced collaboration between modalities available in the training stage is essential if modalities are missing at inference.

Recently, multi-modal co-learning has been used in the EO domain to face the problem of missing modality during inference [19], as shown in Fig. 1. However, most of the co-learning exploration has been limited to frameworks tailored for a predefined missing modality during inference, named *dedicated training* [9]. In this way, training aims to use additional modalities only as a support for the main target modality. For instance, Kampffmeyer et al. [20] use a hallucination branch from an optical modality to simulate the depth modality that is expected to be missing, while MMEarth model [21] reconstructs various other sensor modalities from a specific one. Since these approaches define in advance the modalities that will be missing during inference, they are not general enough to cases when other modalities might be absent. Conversely, *non-dedicated training* avoids making in advance the firm choice on the modality available during inference. For instance, Zheng et al. [22] use a common model generator to instantiate single-modality models, while Mena et al. [23] share weights of the last layers between modality-dedicated models. Similarly to this latter family of approaches, we address the missing modality problem with a non-dedicated training strategy.

In this manuscript, we propose a multi-modal co-learning framework to handle the problem of all-but-one missing modality occurring at inference time with EO data. Our framework focuses on boosting the cross-modal knowledge

transfer between heterogeneous sensor modalities at the feature-level. To this end, we use modality-dedicated encoders to extract common and unique information. More precisely, we enforce a disentanglement of three feature spaces (shared, specific, and unused) per modality that are guided by several loss functions, named Multi-modal Disentanglement for Co-learning (MDiCo). Unlike prior co-learning approaches in the EO domain, which often focus on a single task, e.g. land-cover classification, or design frameworks tightly coupled to specific data modalities, we propose a framework that is task-agnostic and adaptable to the modality available at inference time. To achieve this flexibility, we incorporate dedicated prediction heads trained explicitly to predict the downstream task from the shared and specific feature representations per-modality, enabling inference from any of the modality available at the training stage.

To comprehensively assess the behavior of our framework, we compare several recent state-of-the-art approaches from the general domain of machine learning and computer vision, as well as four recent methods especially tailored for the EO domain. Our experimental assessment covers four multi-modal EO benchmarks, featured by different combinations of sensor modalities, spanning several downstream tasks such as binary, multi-class, and multi-label classification, as well as regression. The results prove the quality of MDiCo over the competing approaches across the different downstream tasks and single-modality scenarios at the inference stage. The systematic improvement exhibited by MDiCo, in all validation scenarios, positions our framework as an effective solution for multi-modal co-learning where all-but-one missing modality scenarios arise at the inference stage. Our code and related datasets are available at <https://github.com/fmenat/MDiCo>.

This manuscript is organized as follows: The related literature on co-learning, multi-modal data, and EO is described in Sec. 2. Sec. 3 introduces our multi-modal co-learning framework. The experimental evaluation and the related findings are reported and discussed in Sec. 4, while Sec. 5 draws the conclusions of our work.

2 Related work

2.1 Multi-modal learning and missing modalities

Multi-modal learning has proven to be effective in enhancing model performance and generalization. This paradigm involves training deep learning models on multiple data modalities simultaneously [8], such as images, texts, and sensor signals, enabling models to leverage complementary information across heterogeneous data. Beyond performance improvements, research works have explored how multi-modal approaches influence the learning of distinct feature representations per modality. To extract common (or shared) feature representations across modalities, various loss functions have been employed, including correlation maximization [10], contrastive learning [24], and weight-sharing mechanisms across modality-specific models [25]. For instance, Poklukar et al. [26] demonstrate that shared features derived via contrastive loss remain robust to missing all-but-one modality. Other approaches aim to disentangle shared and specific features jointly by combining multiple loss functions. The MISA model [27] learns shared features through both weight-sharing and minimizing inter-modality distances, while specific features are extracted using separate encoders and enforced by minimizing similarity across modalities. In contrast, the ShaSpec model [28] employs a domain discriminator to learn modality-specific features. However, multi-modal learning can still suffer from the missing modality problem.

The challenge of missing modalities at inference time has been addressed through various strategies in the literature [9]. One common approach involves simulating missing modalities during training, as demonstrated in works such as [28] and [29]. A different strategy, explored by Choi et al. [30], involves randomly selecting a single modality during training, at each feature dimension, to encourage robustness. Furthermore, knowledge distillation and self-distillation frameworks have been employed, where a full-modality teacher guides student models trained with incomplete modalities, as introduced in the approaches by McKinzie et al. [31] and Lin & Hu [32].

2.2 Multi-modal co-learning

Co-learning has been used in diverse fields and applications by having multiple models that cooperate among them. In domain adaptation, Ganin et al. [3] introduce per-domain models with shared encoders and a domain classifier for image classification, while Obrenovic et al. [4] introduce a similar approach by sharing only the middle layers among per-domain models in a heterogeneous domain adaptation setting. In learning with noisy labels, co-teaching [2] uses two models trained simultaneously that supervise each other, selecting samples with potential clean labels. Similarly, MentorNet [33] uses a teacher network to guide the training via selecting reliable instances to automate curriculum learning. Under the lens of mutual distillation, Zhang et al. [34] introduce a framework with two identical models that are cross-guided based on predicted probabilities matching for image classification tasks.

Co-learning has emerged in the multi-modal setting to handle noisy modalities and unreliable label scenarios [7]. The co-learning process between single-modality models can be done at the feature-, decision-, or model-level. For instance, learning a shared feature space between modalities has been implemented via deep learning models as a feature-based co-learning strategy [10] and for cross-modal distillation [6]. Recently, it has been used in contrastive learning for self-supervision [24]. Moreover, decision-based co-learning, i.e. knowledge transfer between single-modality model predictions, has been used in a (multi-modal) semi-supervised setting [35] and for mutual distillation [36]. Furthermore, the model-based co-learning involves sharing layers among the single-modality models. For instance, Zhang et al. [11] use a shared prediction head among single-modality models with orthogonal directions in the gradient trajectory. On the other hand, Zadah et al. [12] theoretically show that learning from multi-modal information is better even if testing scenarios have a single-modality setting, i.e. the model can benefit from additional information available only during training.

2.3 Multi-modal co-learning in EO

Multi-modal data has been crucial in the EO domain to analyze complex phenomena on Earth [13]. This is because sensors capture different information about the Earth’s surface, complementing individual observations, such as optical and radar data. Thus, most of the research leveraging multi-modal data focuses on data fusion to enhance model accuracy in different applications [37]. One example is the data fusion contest hosted each year by the Geoscience and Remote Sensing Society [38, 39]. Moreover, Mena et al. [14] discuss different fusion strategies used in the EO domain and underline the tendency to design models (and their combinations) with increasing levels of complexity. However, efforts have focused on designing simple models employing cross-distillation strategies involving a multi-modal teacher and single-modal students [40]. For instance, Pande et al. [41] use an adversarial approach with a hallucination network on multi-spectral and panchromatic images.

Multi-modal co-learning has been effectively used to handle missing sensor modalities in the EO domain [19]. The most common scenario is feature-based co-learning, where features are forced to be similar, such as with hallucination networks [20], cross-modal retrieval [42], and contrastive learning [43]. Recently, contrastive learning has been leveraged for multi-modal model pretraining, such as in Heidler et al. [44] and OmniSat [45]. On the model-based co-learning aspect, Hong et al. [46] introduce the sharing of prediction heads between hyper- and multi-spectral optical image models. Then, Zheng et al. [22] propose a meta-model that generates the parameters of the single-modality models in an optical-radar sensor setting. Moreover, Xie et al. [47] use a co-training strategy to obtain pseudo-labels from one modality to another in a semi-supervised setting with point cloud data and optical images. Another family of strategies available in the literature relies on cross-reconstruction. For instance, MMEarth [21] reconstructs various sensor modalities from a specific one. On the same track, Xiong et al. [48] share layers between modalities to set up a reconstruction task.

Previous co-learning works in the EO domain commonly focus on single tasks and design specific methods for the used modalities [22, 47] or assume a modality-dedicated training [20, 41, 40, 21]. In contrast to these works, we introduce a general multi-modal co-learning approach that is flexible enough to be used for any available modality across a diverse set of downstream tasks.

3 Method

3.1 Notation

Let us consider the multi-modal co-learning under a supervised setting as follows. There are N training samples available with corresponding multi-modal data and ground truth information. Without loss of generality, we consider a two-modality setup as $\mathbb{D} = \{\mathbf{X}_1^{(i)}, \mathbf{X}_2^{(i)}, y^{(i)}\}_{i=1}^N$. During inference, only one of the modalities is available, either $\mathbf{X}_1$ or $\mathbf{X}_2$, as illustrated in Fig. 1. Each modality $\mathbf{X}_m$, with $m \in \{1, 2\}$ provides per-modality specific information. Here, the objective is to derive single-modality models $\mathcal{G}_m(\cdot)$, where each model is composed of an encoder and prediction head, that supply the prediction for test data, i.e. $\hat{y}_m^{(i)} = \mathcal{G}_m(\mathbf{X}_m^{(i)})$. For simplicity, we avoid the superscript of the sample index i in the following.

3.2 Framework description

To address the problem of all-but-one missing modality at the inference stage (see Fig. 1), we introduce a framework that facilitates the collaboration between heterogeneous sensor modalities and improves the cross-modal learning at the feature-level.

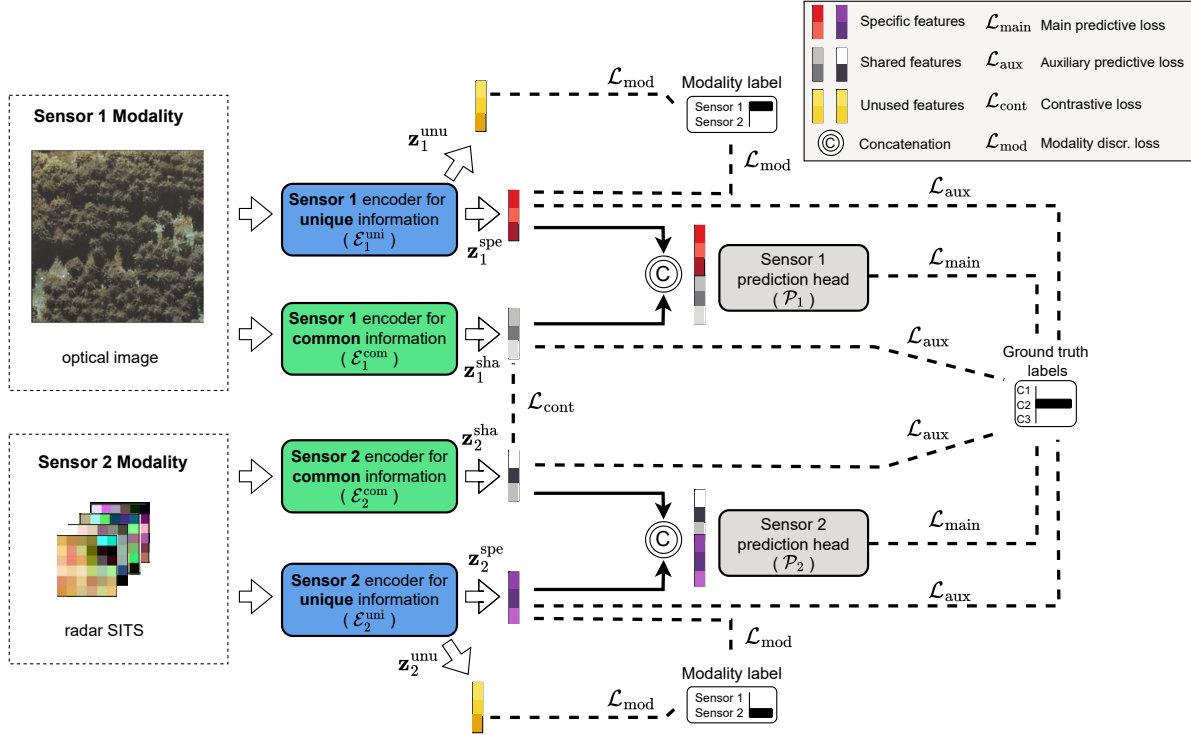


Figure 2: Illustration of our MDiCo framework with shared, specific, and unused features. The main prediction per modality is illustrated in regular arrows, while the losses are shown with dashed lines. Two sensor modalities are shown: an optical image and a radar Satellite Image Time Series (SITS).

3.2.1 Overview

Our framework, shown in Fig. 2, considers modality-dedicated encoders that extract three representations per modality. Two of these representations are used for the downstream task, considering the modality-shared and modality-specific feature space, while the third one, referred to as unused, is successively discarded. To disentangle these feature spaces and guide the co-learning process, we use multiple loss functions. Two of these losses are at feature-level while the remaining two are at prediction-level. Thus, our **Multi-modal Disentanglement for Co-learning (MDiCo)** framework is task-agnostic, adopting a predictive loss function driven by each downstream task.

During inference, the target is predicted from each single-modality model based on the shared and specific features coming from the per-modality encoders. Thus, our multi-modal formulation operates with arbitrary single-modality available at the inference stage. This corresponds to a more general case than dedicated training in the literature [20, 40], which considers a predefined modality accessible at inference time.

3.2.2 Shared, Specific and Unused Feature Spaces

Consider per-modality encoders $\mathcal{E}_m^{\text{com}}(\cdot)$, with $m \in \{1, 2\}$, that extract common information between modalities. This common information is obtained as follows:

$$\mathbf{z}_m^{\text{sha}} = \mathcal{E}_m^{\text{com}}(\mathbf{X}_m) \quad \forall m \in \{1, 2\}, \quad (1)$$

with $\mathbf{z}_m^{\text{sha}} \in \mathbb{R}^d$ the **shared features**. To achieve the learning of shared features between modalities, we use a contrastive learning strategy (cf. Sec. 3.3.3). Furthermore, consider modality-dedicated encoders $\mathcal{E}_m^{\text{uni}}(\cdot)$ that extracts information that is unique to the modality m . This unique information can be separated into task-discriminative and non-discriminative, similar to [6], modeled as follows:

$$\mathbf{z}_m^{\text{spe}}, \mathbf{z}_m^{\text{unu}} = \mathcal{E}_m^{\text{uni}}(\mathbf{X}_m) \quad \forall m \in \{1, 2\}, \quad (2)$$

with $\mathbf{z}_m^{\text{spe}} \in \mathbb{R}^d$ the **specific features**, and $\mathbf{z}_m^{\text{unu}} \in \mathbb{R}^d$ the **unused features**. The guidance for learning this unique per-modality information is achieved via a modality discriminator strategy discussed in Sec. 3.3.4. The difference between

these two feature spaces is that we adopt an explicit predictive loss over the specific features. We opt to discriminate the unused features in this unique information, based on the hypothesis that each modality could have noise that is unrelated to the downstream task. Moreover, we did not assume any fixed shape for the per-modality encoders, $\mathcal{E}_m^{\text{com}}$ and $\mathcal{E}_m^{\text{uni}}$, as they tightly depend on the input data (e.g. image or time series).

3.2.3 Per-Modality Prediction

We use a per-modality prediction head $\mathcal{P}_m(\cdot)$ that takes the concatenation of the task-discriminative feature spaces, i.e. $\mathbf{z}_m^{\text{sha}}$ and $\mathbf{z}_m^{\text{spe}}$, and estimates the target value, expressed by

$$\hat{y}_m = \mathcal{P}_m([\mathbf{z}_m^{\text{sha}} \parallel \mathbf{z}_m^{\text{spe}}]) \quad \forall m \in \{1, 2\}. \quad (3)$$

In this way, the framework splits into single-modality models used for the single-modality prediction at inference time, i.e. $\mathcal{G}_m = \{\mathcal{P}_m, \mathcal{E}_m^{\text{com}}, \mathcal{E}_m^{\text{uni}}\}$, given by

$$\hat{y}_m = \mathcal{G}_m(\mathbf{X}_m) \quad (4)$$

$$= \mathcal{P}_m([\mathcal{E}_m^{\text{com}}(\mathbf{X}_m) \parallel \mathcal{E}_m^{\text{uni}}(\mathbf{X}_m)_{\text{spe}}]) \quad (5)$$

where $\mathcal{E}_m^{\text{uni}}(\mathbf{X}_m)_{\text{spe}}$ is the specific features in the unique information, i.e. $\mathbf{z}_m^{\text{spe}}$.

3.3 Training losses

To disentangle the learning of the three per-modality feature spaces, we employ several loss functions. The resulting final loss function optimized during training is designed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{mod}}. \quad (6)$$

We use an unweighted sum of the loss functions in our MDiCo framework (Eq. (6)). In the following, each of the loss functions is described.

3.3.1 Main Predictive loss

As the main loss in our framework, we use a task-discriminative loss to guide the learning process of the per-modality shared and specific features. Considering the per-modality prediction $\hat{y}_m$ in Eq. (3), with $m \in \{1, 2\}$, and the associated ground truth y , the task-discriminative loss is computed as follows:

$$\mathcal{L}_{\text{main}} = \frac{1}{2} \sum_{m \in \{1, 2\}} \mathcal{L}_{\text{pred}}(y, \hat{y}_m), \quad (7)$$

with $\mathcal{L}_{\text{pred}}(\cdot, \cdot)$ the cross-entropy loss function commonly used in multi-class classification tasks, i.e. $\mathcal{L}_{\text{CE}}(p, q) = -\sum_k \mathbb{1}(p = k) \log q_k$ with $\mathbb{1}(\cdot)$ the indicator function, or the squared error in regression tasks, i.e. $\mathcal{L}_{\text{SE}}(p, q) = (p - q)^2$.

3.3.2 Auxiliary Predictive loss

Additionally, we consider an auxiliary predictive loss to enhance the learning of task-discriminative information for both shared and specific per-modality features. Here, we consider a single auxiliary prediction head $\mathcal{P}^{\text{aux}}(\cdot)$ that predicts the target $\hat{y}^{\text{aux}}$ across modalities and that is used on both feature spaces, $\mathbf{z}_m^{\text{sha}}$ and $\mathbf{z}_m^{\text{spe}}$ with $m \in \{1, 2\}$. Thus, the auxiliary predictive loss is computed as:

$$\mathcal{L}_{\text{aux}} = \frac{1}{2} \sum_{m \in \{1, 2\}} \sum_{s \in \{\text{sha}, \text{spe}\}} \mathcal{L}_{\text{pred}}(y, \mathcal{P}^{\text{aux}}(\mathbf{z}_m^s)), \quad (8)$$

with $\mathcal{L}_{\text{pred}}(\cdot, \cdot)$ the same loss function used in Eq. (7).

3.3.3 Contrastive loss

For learning the features shared among modalities, i.e. $\mathbf{z}_m^{\text{sha}}$ with $m \in \{1, 2\}$, we use the contrastive learning [49] by forcing the paired features between modalities to be similar in the learned latent space. In this way, the shared features are guided to group themselves and be modality invariant. We use the InfoNCE [49] criteria given by:

$$\mathcal{L}_{\text{info}}(\mathbf{z}_1^{\text{sha}}, \mathbf{z}_2^{\text{sha}}; \tau) = -\log \frac{\exp(\mathcal{S}(\mathbf{z}_1^{\text{sha}}, \mathbf{z}_2^{\text{sha}})/\tau)}{\sum_{\tilde{\mathbf{z}}_2^{\text{sha}} \in \mathbb{Z}_2} \exp(\mathcal{S}(\mathbf{z}_1^{\text{sha}}, \tilde{\mathbf{z}}_2^{\text{sha}})/\tau)}, \quad (9)$$

with the cosine similarity as $\mathcal{S}(a, b) = \langle a, b \rangle / (\|a\|_2 \cdot \|b\|_2)$ and $\langle \cdot, \cdot \rangle$ the inner product between two vectors, $\tau > 0$ a temperature hyper-parameter that we set to 0.07 following [43], and $\mathbb{Z}_*$ the set of all other cross-modality pairs in the batch. Here, positive pairs (numerator in Eq. (9)) are the two modalities from the same sample, while the negative ones (denominator in Eq. (9)) are the complementary modality features that come from all other samples in the batch. Thus, we consider each modality $m \in \{1, 2\}$ to be the anchor in the contrastive loss, which is given by:

$$\mathcal{L}_{\text{cont}} = \mathcal{L}_{\text{info}}(\mathbf{z}_1^{\text{sha}}, \mathbf{z}_2^{\text{sha}}; \tau) + \mathcal{L}_{\text{info}}(\mathbf{z}_2^{\text{sha}}, \mathbf{z}_1^{\text{sha}}; \tau). \quad (10)$$

3.3.4 Modality Discriminant loss

For learning the features proper to each modality, i.e. $\mathbf{z}_m^{\text{spe}}$ and $\mathbf{z}_m^{\text{unu}}$ with $m \in \{1, 2\}$, we use an auxiliary classifier that discriminates from which modality the features are coming from. To this end, we use two classifiers, one for the specific $\mathcal{P}^{\text{spe}}(\cdot)$ and another for the unused $\mathcal{P}^{\text{unu}}(\cdot)$ feature spaces, where the loss function is designed as:

$$\mathcal{L}_{\text{mod}} = \frac{1}{2} \sum_{m \in \{1, 2\}} \sum_{s \in \{\text{spe}, \text{unu}\}} \mathcal{L}_{\text{CE}}(m, \mathcal{P}^s(\mathbf{z}_m^s)), \quad (11)$$

with $\mathcal{L}_{\text{CE}}(\cdot, \cdot)$ the standard cross-entropy loss function.

4 Experiments

We evaluate our framework against several approaches from recent machine learning, computer vision, and EO literature. The analysis encompasses four multi-modal EO benchmarks covering classification (binary, multi-class, and multi-label) and regression tasks. Concretely, we first introduce the benchmarks (Sec. 4.1) and competing approaches (Sec. 4.2). Then, we describe the evaluation protocol and setup (Sec. 4.3). Subsequently, we present and discuss the obtained results (Sec. 4.4). To provide deeper insights, we examine individual components in our framework through ablation studies (Sec. 4.5 and 4.6). Furthermore, we assess our framework using various encoder backbones (Sec. 4.7). Finally, we analyze the internal representations learned by MDiCo (Sec. 4.8), and the evolution of the individual losses during training (Sec. 4.9).

4.1 Datasets

For the binary and multi-class classification task, we consider a crop recognition problem by using the CropHarvest benchmark [50]. This benchmark contains samples around the globe between 2016 and 2021. The benchmark contains two multi-temporal sensor modalities, corresponding to Satellite Image Time Series (SITS), at a spatial resolution of $10[m]$: optical Sentinel-2 SITS (considering spectral bands and vegetation indices), and radar Sentinel-1 SITS (with polarization bands). Thus, we consider the crop/non-crop estimation at a specific region as a cropland (binary) classification task. This dataset, named **CropHarvest binary (CropH-b)** has 69 800 samples. For multi-class classification, we use a subset of the CropHarvest benchmark including 29 642 samples with crop-type labels associated. More precisely, it covers ten different classes for a downstream crop-type classification task. This variant is named **CropHarvest multi (CropH-m)**.

For the multi-label classification task, we consider a tree-species identification problem introduced in [45], referred to as **TreeSatAI-Time-Series (TSAITS)**. It involves a multi-label classification task covering 15 tree species in Germany. This benchmark includes a multi-temporal and a mono-temporal modality: Sentinel-2 SITS at a $10[m]$ spatial resolution (with multi-spectral bands), and an aerial (mono-temporal) image at a high spatial resolution of $0.2[m]$ (with RGB and infrared bands). This benchmark includes 38 520 samples for training, 6 810 for validation, and 5 044 for testing, collected between 2017 and 2020.

For the regression task, we consider a moisture estimation task introduced in [51], referred to as **Live Fuel Moisture Content (LFMC)**. It involves the prediction of vegetation water (moisture) content per dry biomass in the western US. For this task, we have access to two multi-temporal modalities at a spatial resolution of $250[m]$: Landsat 8 SITS (with spectral bands and vegetation indices), and Sentinel-1 SITS (with polarization bands and indices). This dataset contains 2 578 samples collected between 2015 and 2019.

A summary of the benchmark information with the corresponding modality characteristics and the data format is reported in Table 1.

Table 1: Description of data modalities used in each dataset. The input format corresponds to (time-steps, features, height, width).

Dataset	Modality	Sensor type	Spatial resolution	Temporal resolution	Input shape
CropH-b & CropH-m	Sentinel-1	radar SITS	10 [m]	1 per month	(12, 11, 1, 1)
	Sentinel-2	optical SITS	10 [m]	1 per month	(12, 2, 1, 1)
TSAITS	Aerial	optical image	0.2 [m]	None	(1, 4, 320, 320)
	Sentinel-2	optical SITS	10 [m]	~10 per month	(150, 12, 6, 6)
LFMC	Sentinel-1	radar SITS	250 [m]	1 per month	(4, 3, 1, 1)
	Landsat 8	optical SITS	250 [m]	1 per month	(4, 8, 1, 1)

4.2 Competing methods

For the assessment of our framework, we consider several families of competitors. Using a dedicated training, we include models individually trained on each modality as single-modality baselines, named *Individual*. In addition, we consider models fusing the multi-modal sensor data.

Incorporating a non-dedicated training, we select methods from the recent literature that are especially tailored for handling the different downstream tasks we validate on. From the field of multi-modal fusion with missing modalities, we include two methods: EmbraceNet [30], a feature-level fusion model employing feature sampling as a merge function, and ShaSpec [28], a feature-level fusion model using shared and specific features to handle missing modalities. We include two methods from the multi-modal co-learning field: DeCuR [52], a self-supervised approach that learns common and unique features in single-modality models, DML [34], an ensemble using mutual distillation to match the predictions of single-modality models. From the cross-modal distillation field, we include DisCoM-KD [6], a recent method using invariant, informative, and irrelevant features for the cross-modal learning of paired modalities. Finally, we include four recent methods from the EO literature, AnySat [53], a geospatial foundational model pre-trained on several remote sensing modalities, including the ones covered by some benchmarks considered in this study¹, FMod-Drop [54] a multi-modal model using modality dropout over transformer layers, FCoM-av [29], a multi-modal model simulating all combinations of missing modalities at training, and ESensI [23], an ensemble with shared prediction heads in the single-modality models.

4.3 Experimental setting

We train all competitors having access to all multi-modal data during training, while the inference evaluation is performed considering a single modality, considering the all-but-one missing modality scenario; see Fig. 1 for an illustration. To measure performance, we use the Weighted F1 (F1) score for classification tasks and the Coefficient of Determination (R^2) for the regression task. For the TSAITS dataset, we use the validation set to select models and the test set for final evaluation. As CropH-b, CropH-m, and LFMC datasets do not have a predefined test partition available, we use a standard 10-fold cross-validation for evaluation, following common practices in the literature [51, 55]. For each method, we report results averaged over 5 runs.

All the input data is rescaled via z-score normalization. For the modality-specific encoders, we adopt standard architectures commonly used in the EO literature. Specifically, for all multi-temporal modalities, we use TempCNN [56], a widely adopted backbone based on 1D convolutions over the temporal dimension. For the mono-temporal modality, i.e., the aerial image in TSAITS, we use ResNet-50 [57], a 2D convolutional neural network with residual connections. In each encoder, an additional projection layer is used—a linear layer of 128 units with 20% dropout. For parameters optimization, we use the Adam optimizer with a learning rate of 10^{-3} , batch size of 128, and early stopping with a patience of 5, across 100 training epochs. For all competitors, we retain the default hyperparameter settings as reported in their original works.

In our framework, we intentionally opt for a simple and consistent design across modalities to ensure reproducibility and reduce model complexity. Thus, all prediction heads—i.e., $\mathcal{P}_m(\cdot)$, $\mathcal{P}_m^{\text{aux}}(\cdot)$, $\mathcal{P}_m^{\text{spe}}(\cdot)$, and $\mathcal{P}_m^{\text{un}}(\cdot)$ for both modalities $m \in \{1, 2\}$ —are implemented as single linear layers. We present results using alternative encoder backbones in Sec. 4.7, showing that the proposed framework performs consistently across architectures. We did not perform an extensive hyperparameter search, as our architectural choices are grounded in standard practices from the literature.

¹The AnySat model is fine-tuned in each single-modality inference case.

Table 2: F1 scores in the CropH-b dataset, cropland (binary) classification. The **best** and second-best values are highlighted. *Based on pre-training.

Method	Field	Sentinel-1	Sentinel-2
MMGF	Multi-modal fusion	82.3	
Individual	Unimodal learning	71.5	81.9
EmbraceNet	Fusion with missing modalities	68.8	81.4
ShaSpec	Fusion with missing modalities	68.9	76.1
DeCuR*	Multi-modal co-learning	71.5	81.6
DML	Multi-modal co-learning	<u>71.7</u>	81.5
DisCoM-KD	Cross-modal distillation	70.1	78.6
AnySat*	(EO) Fusion with missing modalities	70.2	<u>82.4</u>
FModDrop	(EO) Fusion with missing modalities	70.6	79.8
FCoM-av	(EO) Fusion with missing modalities	71.4	82.1
ESensI	(EO) Multi-modal co-learning	<u>71.7</u>	81.7
MDiCo	Multi-modal co-learning	73.5	83.3

Table 3: F1 scores in the CropH-m dataset, crop-type (multi-class) classification. The **best** and second-best values are highlighted. *Based on pre-training.

Method	Field	Sentinel-1	Sentinel-2
MMGF	Multi-modal fusion	73.3	
Individual	Unimodal learning	55.4	72.2
EmbraceNet	Fusion with missing modalities	40.6	69.4
ShaSpec	Fusion with missing modalities	44.3	63.9
DeCuR*	Multi-modal co-learning	55.1	71.9
DML	Multi-modal co-learning	55.1	71.7
DisCoM-KD	Cross-modal distillation	<u>55.6</u>	70.2
AnySat*	(EO) Fusion with missing modalities	52.2	<u>73.8</u>
FModDrop	(EO) Fusion with missing modalities	52.5	69.7
FCoM-av	(EO) Fusion with missing modalities	54.9	72.3
ESensI	(EO) Multi-modal co-learning	<u>55.6</u>	71.8
MDiCo	Multi-modal co-learning	58.3	74.2

We use two criteria for balancing the predictive losses ($\mathcal{L}_{\text{main}}$ and $\mathcal{L}_{\text{aux}}$) in our framework. In the classification tasks, we have used a weighted cross-entropy with per-class weights that are inversely proportional to the number of samples. This is used to handle class imbalanced scenarios. In the regression task, we have applied a z-score normalization to the target. This is used to rescale the squared error loss and prevent it from dominating over others, like $\mathcal{L}_{\text{mod}}$. Moreover, to avoid overfitting to any specific configuration, we adopt a uniform sum for aggregating the loss terms (Eq. 6). In Sec. 4.5, we report a comparative analysis with an adaptive weighting scheme, which further supports the effectiveness of our design.

4.4 Results

We report the predictive performance results in Table 2, 3, 4, and 5, for datasets CropH-b, CropH-m, TSAITS, and LFMC respectively.

For the classification datasets (CropH-b, CropH-m, TSAITS), we observe that methods proposed in the EO domain, like AnySat, FCoM-av, and ESensI, tend to work better than approaches proposed in the more generic computer vision and machine learning domains. This is reasonable, as data in the EO domain have a heterogeneous nature and require approaches that explicitly consider the peculiar characteristics of this kind of information [15]. Concerning the rest of the competitors, we can note that ShaSpec, a state-of-the-art method for missing image modalities, clearly exhibits poor results on TSAITS and LFMC, as well as EmbraceNet in the LFMC dataset. Moreover, existing literature has shown that methods introduced for data fusion in the context of missing modalities are not sufficiently robust and generic for single-modality predictions [29]. This fact underscores the need for specialized frameworks for multi-

Table 4: F1 scores in the TSAITS dataset, tree (multi-label) classification. The **best** and second-best values are highlighted. *Based on pre-training. †It only predicts a no-label pattern.

Method	Field	Aerial	Sentinel-2
OmniSAT	Multi-modal fusion		73.3
MMGF	Multi-modal fusion		68.6
Individual	Unimodal learning	64.7	64.9
EmbraceNet	Fusion with missing modalities	63.2	47.3
ShaSpec	Fusion with missing modalities	†	†
DeCuR*	Multi-modal co-learning	62.4	64.2
DML	Multi-modal co-learning	65.2	60.5
DisCoM-KD	Cross-modal distillation	60.4	47.9
AnySat*	(EO) Fusion with missing modalities	60.0	74.8
FModDrop	(EO) Fusion with missing modalities	59.1	54.0
FCoM-av	(EO) Fusion with missing modalities	<u>65.9</u>	60.2
ESensI	(EO) Multi-modal co-learning	<u>64.3</u>	57.9
MDiCo	Multi-modal co-learning	66.4	<u>66.3</u>

Table 5: R^2 scores in the LFMC dataset, regression. The **best** and second-best values are highlighted. *Based on pre-training.

Method	Field	Sentinel-1	Landsat-8
InputFu	Multi-modal fusion		0.520
Individual	Unimodal learning	0.242	0.432
EmbraceNet	Fusion with missing modalities	0.110	0.079
ShaSpec	Fusion with missing modalities	0.053	0.189
DeCuR*	Multi-modal co-learning	0.248	<u>0.417</u>
DisCoM-KD	Cross-modal distillation	0.226	<u>0.416</u>
AnySat*	(EO) Fusion with missing modalities	0.185	0.353
FModDrop	(EO) Fusion with missing modalities	0.080	0.280
FCoM-av	(EO) Fusion with missing modalities	0.214	0.374
ESensI	(EO) Multi-modal co-learning	<u>0.233</u>	0.393
MDiCo	Multi-modal co-learning	0.248	0.467

modal co-learning that can effectively handle both the complexity and heterogeneity of multi-modal data, as well as the missing modality scenario at inference time.

Throughout the benchmarks, our MDiCo framework consistently outperforms both Individual baselines and all competing methods in nearly all considered cases. The improvement over the single-modality baseline (Individual) is not straightforward, as several competitors fail to ameliorate these baselines. For example, this is the case for DisCoM-KD on CropH-b and EmbraceNet on CropH-m. Our approach achieves comparable performances to DeCuR in the Sentinel-1 evaluation of the LFMC dataset, and is only outperformed by AnySat in the Sentinel-2 evaluation of the TSAITS dataset, which can be attributed to AnySat’s pre-training exposure to this specific benchmark data. Despite these isolated cases, our framework achieves superior results in the other inference modality (Aerial). Furthermore, our performance improvement is substantial in several cases, particularly with both modalities in CropH-m and with Landsat-8 modality in the LFMC dataset. Regarding the CropHarvest benchmarks (CropH-b and CropH-m), our approach even improves the results of the multi-modal reference, MMGF, that leverages both data modalities when the Sentinel-2 modality is considered for inference.

Overall, we notice that the methods obtaining the second-best results, across the benchmarks, vary depending on which modality is available for inference, while MDiCo obtains consistent improvements no matter the considered modality. This indicates that our framework takes advantage of modality cooperation in the multi-modal training stage, enhancing the single-modality model’s performance at inference time.

Table 6: F1 scores in the binary and multi-class classification tasks for different variations in our framework. The **best** and second-best values are highlighted.

Variant	CropH-b		CropH-m		Avg.
	Sentinel-1	Sentinel-2	Sentinel-1	Sentinel-2	
Individual	71.5	81.9	55.4	72.2	70.3
MDiCo	73.5	83.3	58.3	74.2	72.3
→ w/o modality loss	<u>73.3</u>	<u>83.0</u>	<u>58.1</u>	<u>74.0</u>	<u>72.1</u>
→ w/o auxiliary loss	73.1	82.9	58.0	74.2	72.0
→ w/o contrastive loss	70.4	80.2	56.5	71.6	69.7
→ w/o unused features	72.8	82.8	56.5	72.7	71.2
→ unpaired data	71.8	82.1	57.4	<u>74.0</u>	71.3
→ shared encoders	72.8	82.8	56.0	72.4	71.0
→ weighted loss	72.4	82.7	57.8	74.2	71.2

Table 7: Predictive performance when using only the shared or specific features in the MDiCo framework. The **best** and second-best values between our variants are highlighted.

Variant	CropH-b		CropH-m		TSAITS		LFMC	
	S1	S2	S1	S2	Aerial	S2	S1	L8
MDiCo	73.5	83.3	58.3	74.2	66.4	66.3	0.248	0.467
→ shared features	<u>71.7</u>	81.6	58.3	70.5	35.3	28.4	0.226	<u>0.445</u>
→ specific features	71.5	<u>82.3</u>	<u>55.3</u>	<u>72.8</u>	<u>65.3</u>	<u>63.5</u>	<u>0.230</u>	0.421

4.5 Ablation

We conduct an ablation study to isolate the key factors characterizing the behavior of MDiCo. To this end, we individually remove components in our framework, considering the CropH-b and CropH-m datasets. These results are reported in Table 6.

We note that the contrastive loss has the greatest impact on model performance when removed, while the modality-discriminant and auxiliary predictive losses demonstrate less individual impact. One possible explanation is that the contrastive loss is greatly enhancing the cross-modal interaction by aligning the representation of heterogeneous sensor modalities. In comparison, the other loss terms may regulate modality-specific representations in a more localized manner. We also note that the contrastive loss typically holds higher values during training (Sec. 4.9), suggesting its dominant role in driving the co-learning process. Nevertheless, the combination of all the loss terms enables our approach to fully exploit the available multi-modal data during training and clearly improves single-modality inference overall.

Furthermore, we consider an additional ablation analysis of our framework where we discard the unused feature component (Sec. 3.2.2). This experiment leads to a drop of around 1 point in the F1 score. The same relative drop is observed when experimenting using unpaired modalities. Moreover, we consider an ablation case where the same encoder is used to extract the per-modality unique and common information (i.e. $\mathcal{E}_m^{\text{com}} = \mathcal{E}_m^{\text{uni}}$ in Sec. 3.2.2). This variation produces a considerable decrease in performance in the CropH-m dataset of around 2 points in the F1 score.

Additionally, we include a comparison to a version of our framework including an adaptive weighting strategy to combine all loss terms (Eq. 6). More precisely, we follow the uncertainty-based criteria by Kendall et al. [58]. Notably, this strategy yields inferior results compared to the simple summation we have adopted. This reflects that our framework benefits from an unbiased aggregation of the loss terms to improve the co-learning process and thereby, single-modality inference.

The ablation analysis indicates that all the components on which our MDiCo framework is built, and their interplay, contribute to its superior behavior in the (all-but-one) missing modality scenario. Moreover, this occurs regardless of the considered benchmark and available modality at the inference stage.

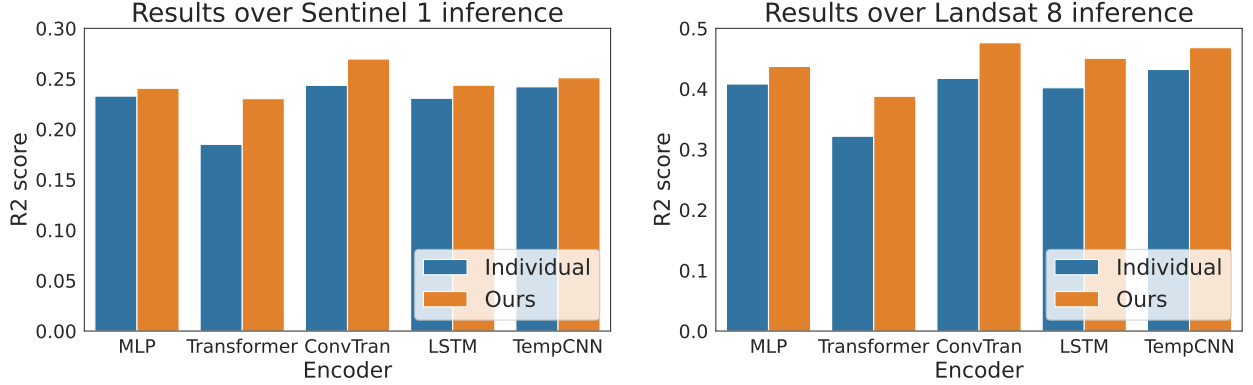


Figure 3: Predictive performance by using different encoder architectures in the MDiCo framework. Results are from the LFMC dataset (regression).

4.6 Usefulness of shared and specific features

Table 7 shows the performance when using only the shared or specific features learned by MDiCo. Overall, the method achieves the best results when both feature sets are considered, highlighting the benefit of leveraging complementary information. In most cases, specific features contribute more positively to performance than shared ones. An exception is observed in the Sentinel-1 modality of CropH-m, where shared features slightly outperform specific ones. Interestingly, results using only shared features are notably lower for the TSAITS dataset. This may be due to the challenge of aligning multi-modal data captured by sensors with substantial differences in spatial and temporal resolutions, such as Sentinel-2 and Aerial. In contrast, sensor pairs with more similar characteristics (e.g., Sentinel-1 and Sentinel-2) tend to encode more common information relevant to the task. These observations indicate that the relationship between modality-shared and modality-specific features is both sensor- and task-dependent, with no absolute rule about the balance between these two spaces.

4.7 Effect of different encoders

To analyze the sensitivity of our MDiCo framework to the modality encoders, we evaluate its performance using various backbones in the multi-temporal modalities. The goal is to verify whether the learning process remains consistent across architectures. To this end, we consider several temporal encoders with diverse characteristics in addition to the TempCNN backbone. These are: i) a two layer Multi-Layer Perceptron (MLP) that ignores the temporal dimension in the input modalities, ii) a standard Transformer [59] that explicitly models temporal dependencies, iii) ConvTran [60], a recent backbone for multi-variate time series classification that combines both 1D convolutions with transformer blocks, and iv) an LSTM [61] (Long Short Term Memory) recurrent neural network. The results on the LFMC dataset are shown in Fig. 3, while others are in the appendix (Sec. A.1).

We observe that the performance of our framework can be influenced by the encoder architecture, with the best results coming from the ConvTran, TempCNN, and LSTM encoders, respectively. Surprisingly, the worst results are obtained with the Transformer architecture. Although its layer structure is comparable to other models (see Sec. A.1), this may be attributed to the over-parametrization of this model—over one million parameters (Table A1)—which can hamper the learning process from the relatively small LFMC dataset (fewer than 3 thousand samples). Notably, the Transformer and ConvTran architectures achieve the top performances on the other datasets (see Sec. A.1 in the appendix). Moreover, we notice that the improvement of MDiCo, compared to the individual trained models, is consistent across different encoder architectures. This relative gain is higher when recent backbone encoders (i.e., ConvTran) are employed, while it is moderate with TempCNN and MLP. Overall, this analysis demonstrates the general applicability of our framework regardless of the considered encoder architecture.

4.8 Qualitative analysis of learned features

We visually inspect the learned features using our MDiCo framework in a 2D projection via the t-SNE method [62], illustrated in Fig. 4. We compare the features extracted by our framework (concatenation of shared and specific ones) with the ones derived by the Individual methods in the multi-label dataset TSAITS. Thus, we observe that the features

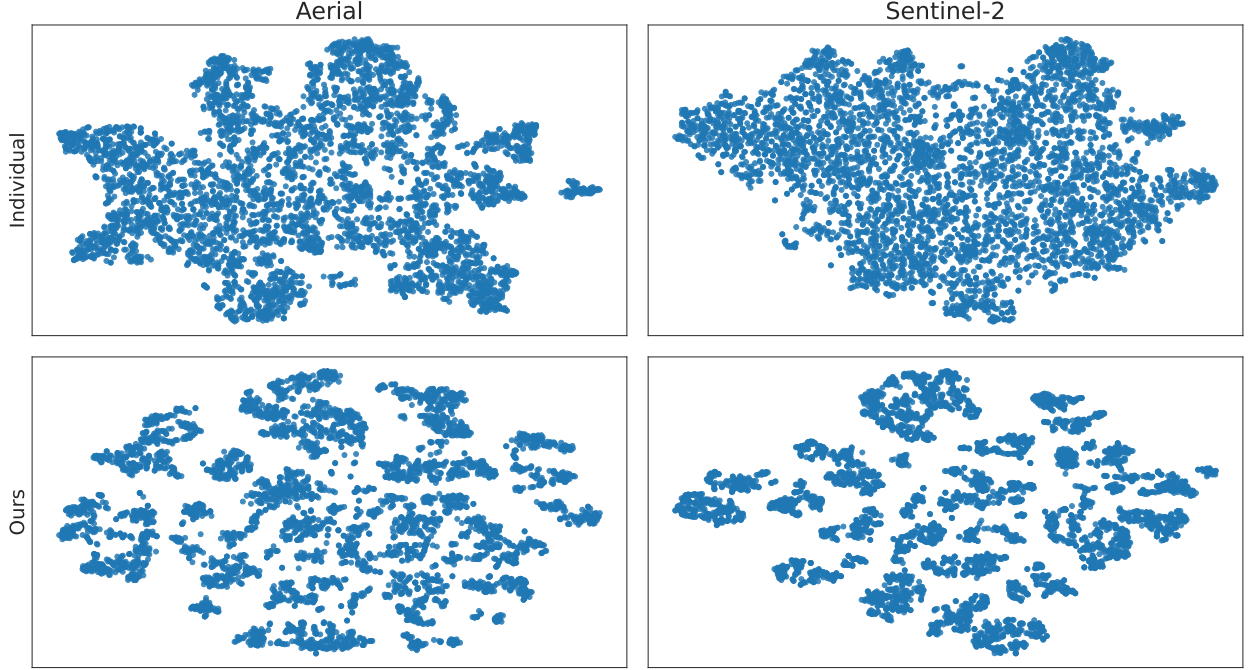


Figure 4: t-SNE projection of the learned features by the Individual method and our MDiCo (concatenation of shared and specific features) on the TSAITS dataset.

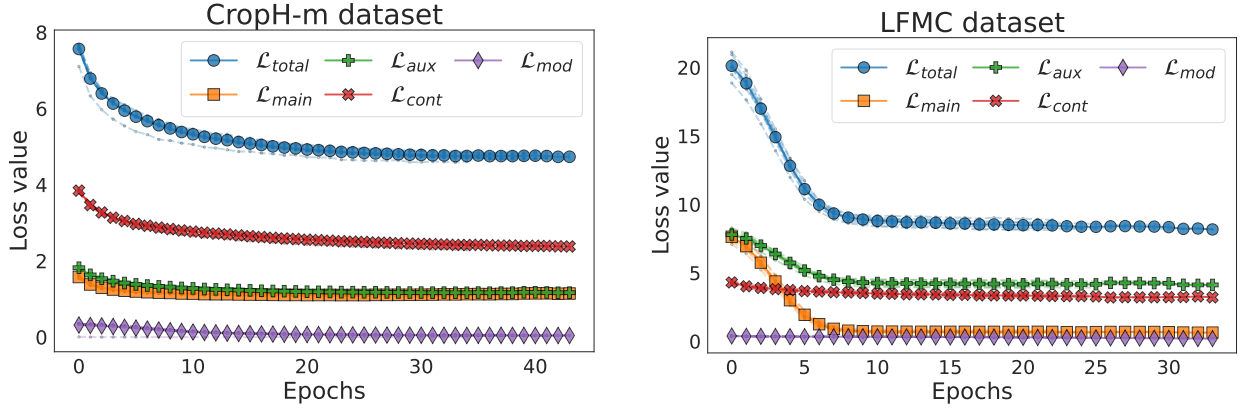


Figure 5: Individual loss functions of our framework across the training. The average across folds and multiple runs is shown for each loss function.

learned by MDiCo exhibit a clear cluster structure compared to the features of the Individual single-modality models. This pattern is observed in the features learned for both Aerial and Sentinel-2 modalities.

4.9 Loss functions during training

To analyze the relative magnitudes of the different loss functions employed in our framework, we plot their trends during training in the CropH-m and LFMC datasets in Fig. 5. The results of the CropH-b dataset are shown in the appendix (Sec. A.3). Under the classification task, all loss functions exhibit similar behaviors. For example, the modality discriminant minimization reaches the lowest magnitudes, whereas the contrastive loss maintains the highest ones. In the regression task, the main predictive loss exhibits rapid convergence during the initial training epochs, whereas the contrastive and auxiliary losses maintain relatively high values throughout the training process. This behavior demonstrates that our framework achieves balanced optimization across all loss components, effectively preventing any single loss function from dominating the learning process.

5 Conclusion

The possible lack of sensor modalities at inference time limits the applicability of multi-modal models in the EO domain. In this work, we demonstrate that multi-modal co-learning, by allowing interaction between available modalities at training time, can enhance single-modality model performance at inference. To achieve this, we introduce a novel task-agnostic framework that combines multiple loss functions to enable effective feature-based collaboration among heterogeneous sensor modalities. Our approach aims to disentangle modality-shared, modality-specific, and unused information, thereby improving the performance of single-modality models. We validate our framework on several EO multi-modal benchmarks spanning binary, multi-class, multi-label classification, and regression tasks. Our framework, unlike recent methods from the literature, systematically improves upon models trained on individual modalities and consistently outperforms state-of-the-art methods from the general fields of computer vision and machine learning, as well as EO-specific strategies, across various downstream tasks. Our research contribution to multi-modal co-learning advances the state of the art in the EO domain under missing modality scenarios, precisely when only a single training modality remains accessible during inference.

Limitations. The following points discuss the limitations of our work and suggest potential future research directions: i) The current framework has been validated for benchmarks with only two modalities available at the training stage. A natural extension would be to scenarios involving multiple modalities during training. ii) Our approach has been validated exclusively on EO data. Future work should extend the evaluation to general multi-modal computer vision and machine learning benchmarks.

Statements and Declarations

Funding

Francisco Mena acknowledges support through a scholarship of the University of Kaiserslautern-Landau.

Competing interests

The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.
- [2] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [3] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International Conference on Machine Learning*, pp. 1180–1189, 2015.
- [4] M. Obrenović, T. Lampert, M. Ivanović, and P. Gañarski, “Learning domain invariant representations of heterogeneous image data,” *Machine Learning*, vol. 112, no. 10, pp. 3659–3684, 2023.
- [5] C. Carrascosa, J. Rincón, and M. Rebollo, “Co-learning: Consensus-based learning for multi-agent systems,” in *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pp. 63–75, 2022.
- [6] D. Ienco and C. F. Dantas, “DisCoM-KD: Cross-modal knowledge distillation via disentanglement representation and adversarial learning,” in *British Machine Vision Conference*, 2024.
- [7] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, “Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions,” *Information Fusion*, vol. 81, pp. 203–239, 2022.
- [8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [9] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, “Deep multimodal learning with missing modality: A survey,” *arXiv preprint arXiv:2409.07825*, 2024.
- [10] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in *International Conference on Machine Learning*, pp. 1247–1255, 2013.

- [11] X. Zhang, J. Yoon, M. Bansal, and H. Yao, “Multimodal representation learning by alternating unimodal adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2024.
- [12] A. Zadeh, P. P. Liang, and L.-P. Morency, “Foundations of multimodal co-learning,” *Information Fusion*, vol. 64, pp. 188–193, 2020.
- [13] G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein, *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*. -: John Wiley & Sons, 2021.
- [14] F. Mena, D. Arenas, M. Nuske, and A. Dengel, “Common practices and taxonomy in deep multi-view fusion for remote sensing applications,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 4797 – 4818, 2024.
- [15] E. Rolf, K. Klemmer, C. Robinson, and H. Kerner, “Mission critical–satellite data is a distinct modality in machine learning,” *arXiv preprint arXiv:2402.01444*, 2024.
- [16] H. Shen, X. Li, Q. Cheng, C. Zeng, G. Yang, H. Li, and L. Zhang, “Missing information reconstruction of remote sensing data: A technical review,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 61–85, 2015.
- [17] B. L. Markham, J. C. Storey, D. L. Williams, and J. R. Irons, “Landsat sensor performance: History and current status,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 12, pp. 2691–2694, 2004.
- [18] P. Potin, O. Colin, M. Pinheiro, B. Rosich, A. O’Connell, T. Ormston, J.-B. Grataudour, and R. Torres, “Status and evolution of the Sentinel-1 mission,” in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 4707–4710, 2022.
- [19] N. Kieu, K. Nguyen, A. Nazib, T. Fernando, C. Fookes, and S. Sridharan, “Multimodal co-learning meets remote sensing: Taxonomy, state of the art, and future works,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [20] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, “Urban land cover classification with missing data modalities using deep convolutional neural networks,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1758–1768, 2018.
- [21] V. Nedungadi, A. Kaririyaa, S. Oehmcke, S. Belongie, C. Igel, and N. Lang, “MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning,” in *Proceedings of the European Conference on Computer Vision*, pp. 164–182, 2024.
- [22] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, “Deep multisensor learning for missing-modality all-weather mapping,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, pp. 254–264, 2021.
- [23] F. Mena, D. Arenas, and A. Dengel, “Increasing the robustness of model predictions to missing sensors in Earth observation,” *arXiv preprint arXiv:2407.15512*, 2024.
- [24] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, “Multimodal contrastive training for visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7004, 2021.
- [25] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, *et al.*, “Multimodal deep learning,” in *International Conference on Machine Learning*, vol. 11, pp. 689–696, 2011.
- [26] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic, “Geometric multimodal contrastive representation learning,” in *International Conference on Machine Learning*, pp. 17782–17800, 2022.
- [27] D. Hazarika, R. Zimmermann, and S. Poria, “MISA: Modality-invariant and-specific representations for multimodal sentiment analysis,” in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.
- [28] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, “Multi-modal learning with missing modality via shared-specific feature modelling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.
- [29] F. Mena, D. Arenas, and A. Dengel, “Missing data as augmentation in the Earth observation domain: A multi-view learning approach,” *Neurocomputing*, vol. 638, 2025.
- [30] J.-H. Choi and J.-S. Lee, “EmbraceNet: A robust deep learning architecture for multimodal classification,” *Information Fusion*, vol. 51, pp. 259–270, 2019.
- [31] B. McKinzie, V. Shankar, J. Y. Cheng, Y. Yang, J. Shlens, and A. T. Toshev, “Robustness in multimodal learning under train-test modality mismatch,” in *International Conference on Machine Learning*, pp. 24291–24303, 2023.

- [32] R. Lin and H. Hu, “MissModal: Increasing robustness to missing modality in multimodal sentiment analysis,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1686–1702, 2023.
- [33] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, “MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels,” in *International Conference on Machine Learning*, pp. 2304–2313, 2018.
- [34] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep mutual learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018.
- [35] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, “Deep co-training for semi-supervised image recognition,” in *Proceedings of the European Conference on Computer Vision*, pp. 135–152, 2018.
- [36] S. Black and R. Souvenir, “Multi-view classification using hybrid fusion and mutual distillation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 270–280, 2024.
- [37] D. Tuia, R. Roscher, J. D. Wegner, N. Jacobs, X. Zhu, and G. Camps-Valls, “Toward a collective agenda on AI for Earth science data analysis,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 88–104, 2021.
- [38] C. Persello, R. Hänsch, G. Vivone, K. Chen, Z. Yan, D. Tang, H. Huang, M. Schmitt, and X. Sun, “2023 IEEE GRSS data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 1, pp. 94–97, 2023.
- [39] C. Persello, S. Prasad, G. Vivone, V. Lonjou, F. Bretar, R. Rodriguez-Suquet, P. Guntzburger, V. Poulain, J. L. Moigne, B. Smith, *et al.*, “2024 IEEE GRSS data fusion contest: Rapid flood mapping,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 2, pp. 109–112, 2024.
- [40] N. Bakalos, S. Sykiotis, A. Temenos, I. Rallis, A. Doulamis, and N. Doulamis, “Segmentation of remote sensing data with missing modalities through prototype knowledge distillation,” in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 10015–10018, 2024.
- [41] S. Pande, A. Banerjee, S. Kumar, B. Banerjee, and S. Chaudhuri, “An adversarial approach to discriminative modality distillation for remote sensing image classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- [42] Y. Chen, X. Lu, and S. Wang, “Deep cross-modal image–voice retrieval in remote sensing,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7049–7061, 2020.
- [43] C. F. Dantas, R. Gaetano, C. Paris, and D. Ienco, “Reuse out-of-year data to enhance land cover mapping via feature disentanglement and contrastive learning,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 1681–1694, 2024.
- [44] K. Heidler, L. Mou, D. Hu, P. Jin, G. Li, C. Gan, J.-R. Wen, and X. X. Zhu, “Self-supervised audiovisual representation learning for remote sensing data,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103130, 2023.
- [45] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu, “OmniSat: Self-supervised modality fusion for Earth observation,” in *Proceedings of the European Conference on Computer Vision*, pp. 409–427, 2025.
- [46] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “CoSpace: Common subspace learning from hyperspectral-multispectral correspondences,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4349–4359, 2019.
- [47] Y. Xie, J. Tian, and X. X. Zhu, “A co-learning method to utilize optical images and photogrammetric point clouds for building extraction,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, 2023.
- [48] Z. Xiong, Y. Wang, F. Zhang, and X. X. Zhu, “One for all: Toward unified foundation models for Earth vision,” in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 2734–2738, 2024.
- [49] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning*, pp. 1597–1607, 2020.
- [50] G. Tseng, I. Zvonkov, C. L. Nakalembe, and H. Kerner, “CropHarvest: A global dataset for crop-type classification,” *Proceedings of NIPS Datasets and Benchmarks Track*, 2021.
- [51] K. Rao, A. P. Williams, J. F. Flefil, and A. G. Konings, “SAR-enhanced mapping of live fuel moisture content,” *Remote Sensing of Environment*, vol. 245, p. 111797, 2020.
- [52] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, “DeCUR: Decoupling common & unique representations for multimodal self-supervision,” in *Proceedings of the European Conference on Computer Vision*, pp. 286–303, 2025.

- [53] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu, “AnySat: An Earth observation model for any resolutions, scales, and modalities,” in *Accepted at the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [54] Y. Chen, M. Zhao, and L. Bruzzone, “A novel approach to incomplete multimodal learning for remote sensing data fusion,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2024.
- [55] F. Mena, D. Pathak, H. Najjar, C. Sanchez, P. Helber, B. Bischke, P. Habelitz, M. Miranda, J. Siddamsetty, M. Nuske, *et al.*, “Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction,” *Remote Sensing of Environment*, vol. 318, 2025.
- [56] C. Pelletier, G. I. Webb, and F. Petitjean, “Temporal convolutional neural network for the classification of satellite image time series,” *Remote Sensing*, vol. 11, no. 5, 2019.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- [58] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.
- [59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [60] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, “Improving position encoding of transformers for multi-variate time series classification,” *Data Mining and Knowledge Discovery*, vol. 38, no. 1, pp. 22–48, 2024.
- [61] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [62] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

A Further Results

A.1 Encoders

We replicate the analysis of our MDiCo framework with different encoder architectures, made in Sec. 4.7, but for the other datasets with multi-temporal modalities. In addition, we display the number of learnable parameters in each of the encoders used along the datasets and modalities in Table A1. We use 2 layers with 128 units in all encoders, with a few variants in the specific parameters. In concrete, we use a kernel size of 5 in the TempCNN, batch normalization layers in the MLP, and 8 head attentions in the Transformer and ConvTran. Thus, Fig. A1 and Fig. A2 displays the results for the CropH-b and CropH-m datasets, respectively. We observe a more stable performance of our framework across encoder architectures, compared to the ones in the LFMC dataset (Fig. 3). In all cases, the MDiCo consistently outperforms the results of the individual trained models. Moreover, the best results of MDiCo are obtained between TempCNN and ConvTran architectures.

Table A1: Number of learnable parameters along different encoder architectures tested. The values are in thousands.

Encoder	CropH-b & CropH-m		LFMC		TSAITS	
	Sentinel 1	Sentinel 2	Sentinel 1	Landsat 8	Aerial	Sentinel 2
ResNet-50 [57]					23 773	
TempCNN [56]	1 052	1 059	496	496		10 077
MLP	36	52	35	37		
Transformer [59]	1 203	1 204	1 203	1 204		
ConvTran [60]	335	991	399	727		
LSTM [61]	216	221	217	219		

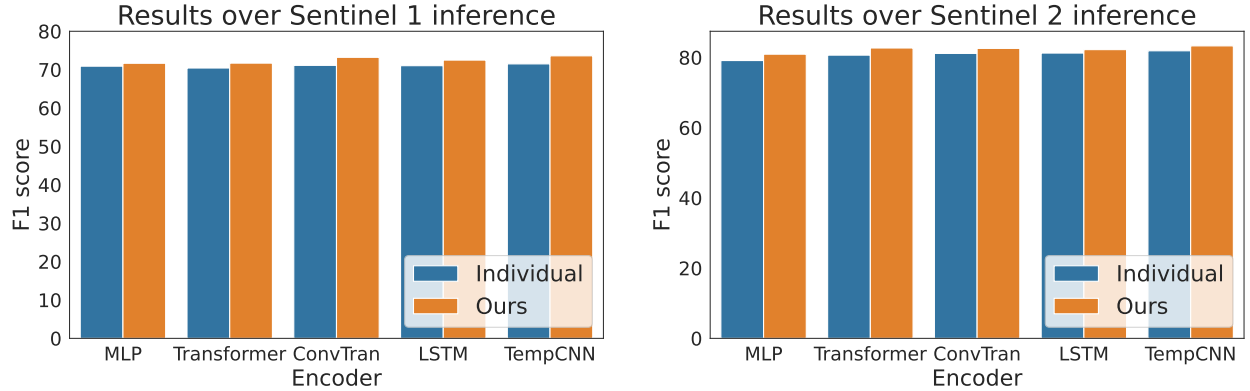


Figure A1: Predictive performance of the MDiCo framework by using different encoder architectures in the CropH-b dataset.

A.2 Visualization of learned features

We display the 2D projection of the learned features of our MDiCo framework for the CropH-b dataset in Fig. A3 and the CropH-m dataset in Fig. A4. The same findings are obtained from this analysis compared to the LFMC dataset (in Fig. 4). Moreover, we notice that the features learned by MDiCo are grouped into subclusters of smaller size.

A.3 Loss functions

We show the relative magnitudes of the different loss functions employed in our framework in the CropH-b dataset in Fig. A5. We notice the same behavior described in Sec. 4.9 for the classification case, i.e. all loss functions exhibit similar minimization trends. Besides, the contrastive loss maintains the higher magnitudes, while modality discriminant reaches the lowest ones.

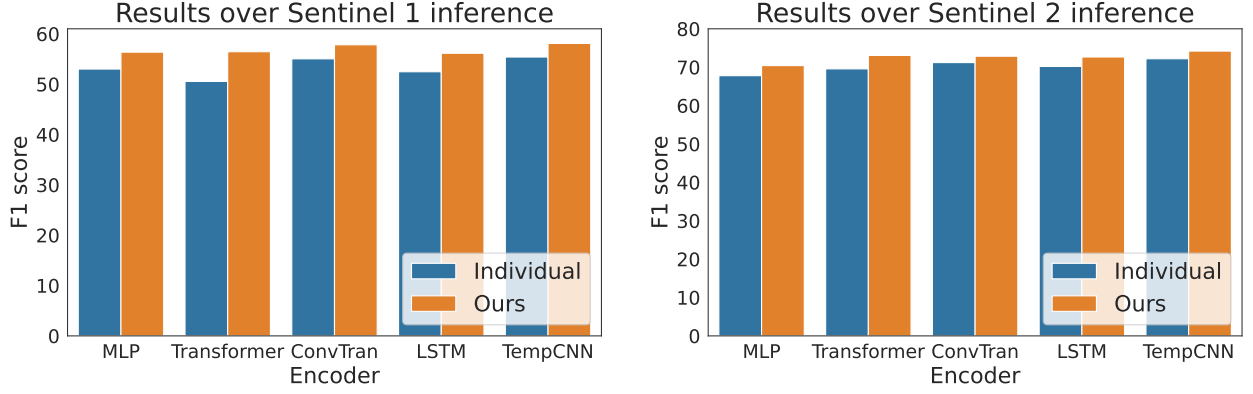


Figure A2: Predictive performance of the MDiCo framework by using different encoder architectures in the CropH-m dataset.

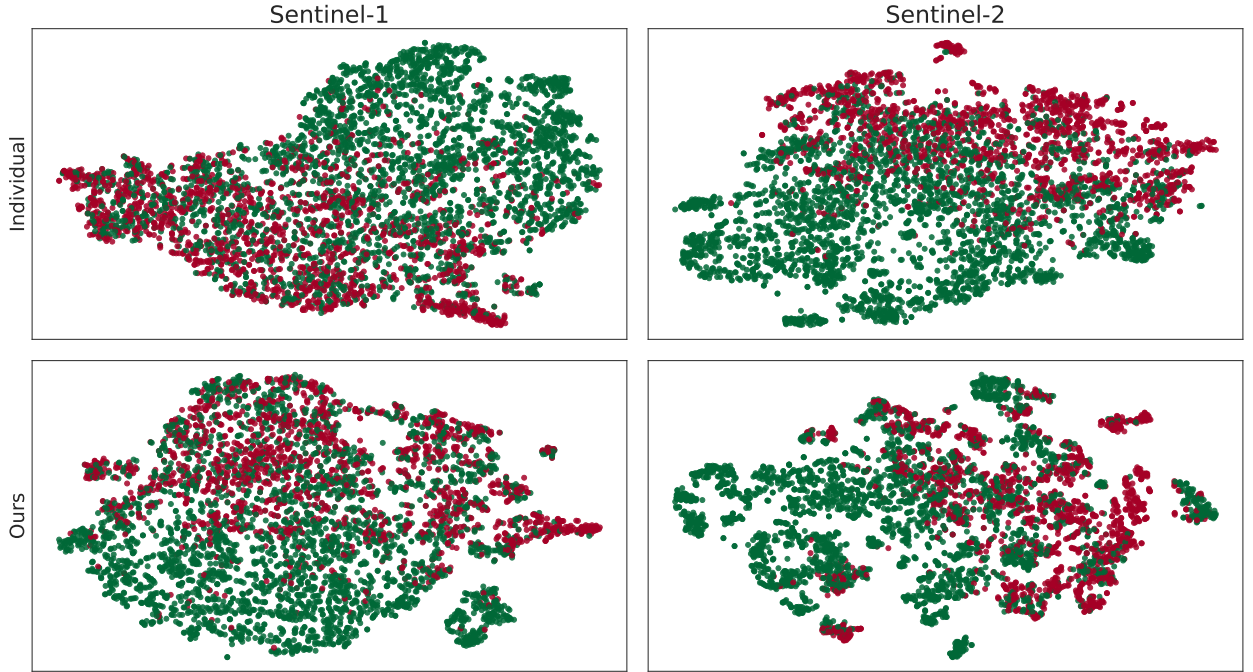


Figure A3: t-SNE projection of the learned features by the Individual method and our MDiCo (concatenation of shared and specific features) on the CropH-b dataset.

A.4 Case study

We illustrate the prediction map of our approach and the individual baseline using the CropH-b dataset in Fig. A6-A8. We present three cases of the global (and sparse) binary prediction maps to qualitatively assess the model's spatial patterns and highlight different regions where there is higher difference. We observe that the maps generated by our MDiCo method are more similar to the ground truth in comparison to those produced by the Individual baseline. This highlights the better spatial patterns obtained through the collaboration of multi-modal data during training.

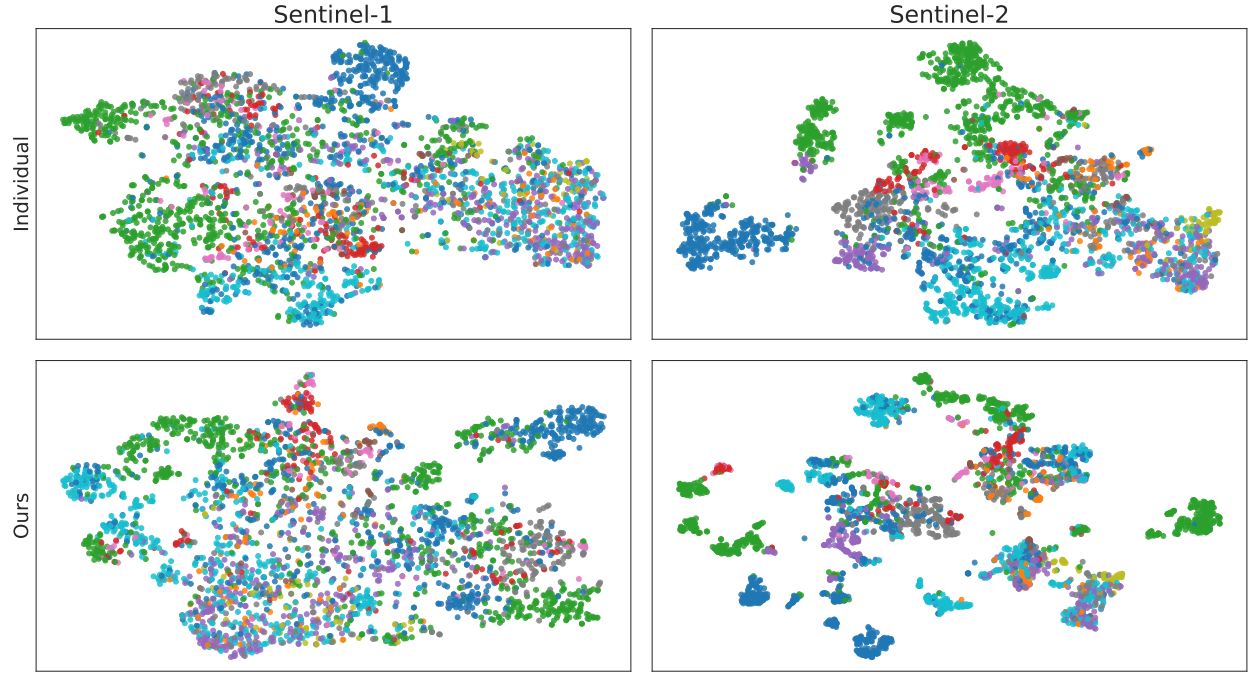


Figure A4: t-SNE projection of the learned features by the Individual method and our MDiCo (concatenation of shared and specific features) on the CropH-m dataset.

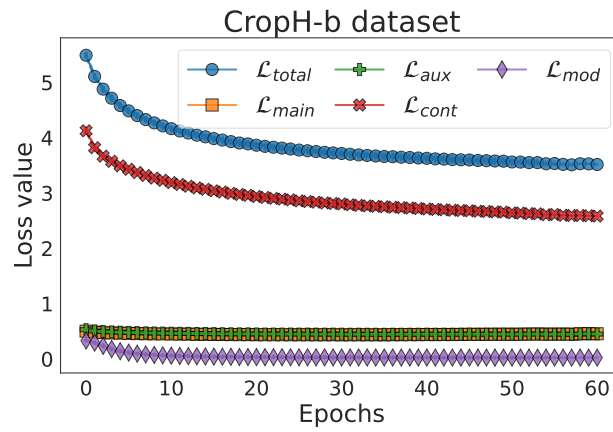


Figure A5: Individual loss functions of our framework across the training. The average across folds and multiple runs is shown for each loss function.

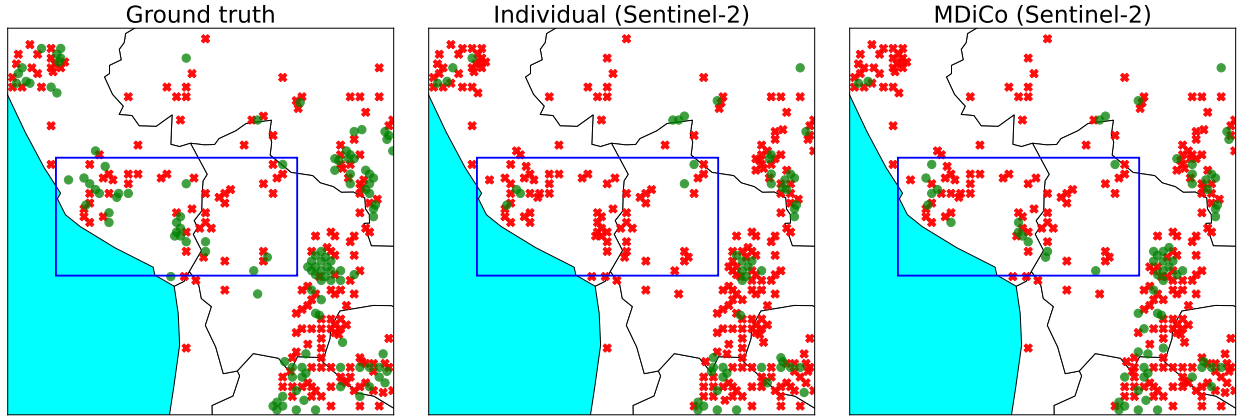


Figure A6: Prediction map of our MDiCo method in comparison to the ground truth labels and the Individual baseline using the Sentinel-2 modality. The region is in the Midwest of South America.

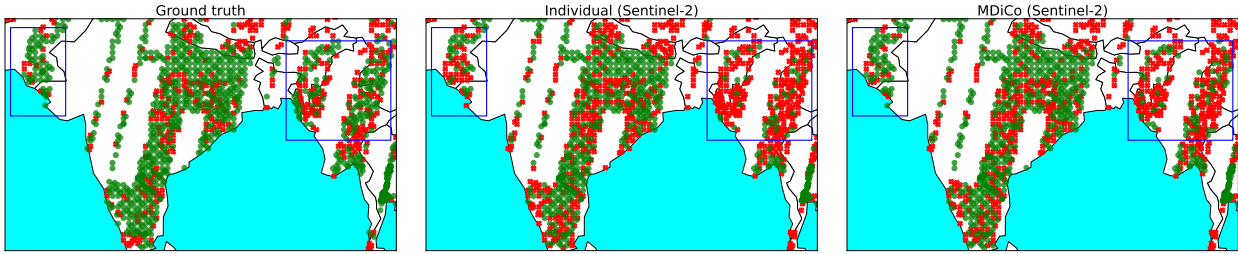


Figure A7: Prediction map of our MDiCo method in comparison to the ground truth labels and the Individual baseline using the Sentinel-2 modality. The shown region is in South Asia.

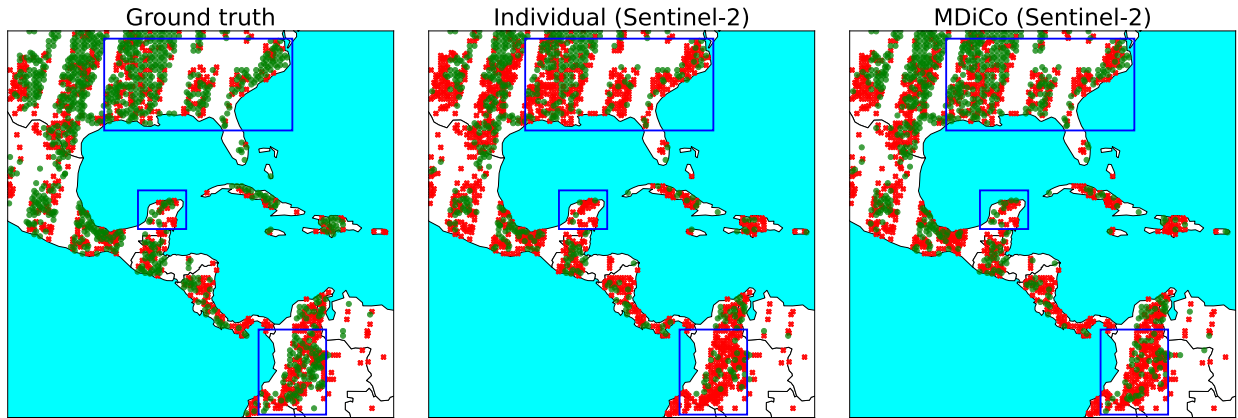


Figure A8: Prediction map of our MDiCo method in comparison to the ground truth labels and the Individual baseline using the Sentinel-2 modality. The shown region is in Central America.