ProMM-RS: Exploring Probabilistic learning for Multi-Modal Remote Sensing Image Representations

Nicolas Houdré[1], Diego Marcos[2], Dino Ienco[2,3], Laurent Wendling[1], Camille Kurtz[1], Sylvain Lobry[1]
[1]LIPADE, Université Paris Cité, 75006 Paris, France
[2]INRIA, Univ. Montpellier, Montpellier, France
[3]INRAE, TETIS, Univ. Montpellier, Montpellier, France

Abstract

Remote sensing imagery offers diverse modalities, such as synthetic aperture radar and multispectral data, which can bring rich, complementary and valuable information about observed scenes. This information is of paramount importance for downstream applications (e.g. land cover mapping, natural resources monitoring, human settlement characterization) that may benefit from such complementarity. Remote sensing imagery often suffers from a lack of labeled data which can hamper the learning of good representations via state-of-the-art supervised methods. Self-supervised learning has thus emerged as a promising paradigm for remote sensing feature extraction, enabling the extraction of meaningful features without reliance on labeled data. While existing multi-modal contrastive models effectively capture shared information between modalities, they often struggle to account for the inherent heterogeneity of multi-modal remote sensing data. This limitation prevents them from fully leveraging the complementarity of multi-modal remote sensing data. Probabilistic representation learning has emerged as a powerful approach to capture the inherent uncertainty and diversity in multi-modal relationships. In this paper we present ProMM-RS, a novel multi-modal self-supervised training framework incorporating a joint probabilistic embedding space to explicitly model the uncertainty of representations between different inputs and modalities. We evaluate our learned representations with a scene classification downstream task from Sentinel optical and radar images, effectively showing the potential of probabilistic embeddings as a way to measure the relevancy of each modality representation, especially under an obstructed dataset.

1. Introduction

Given the vast amount and complexity of data collected by remote sensing (RS) sensors, there is an increasing demand for advanced methods to extract actionable knowledge from such vast amounts of data.

Traditional neuron-based supervised approaches, which are typically designed to optimize representations for a single downstream task come with significant limitations. They require large amounts of annotated data, which is costly and time-consuming to produce. Additionally, the resulting representations are often too task-specific, limiting their generalizability and transferability across different tasks [30].

Self-supervised learning (SSL) has emerged as a compelling solution in machine learning and computer vision for its ability to produce rich and meaningful representations of imagery without the need for labeled examples. Among prominent SSL methods, SimCLR [5] employs a contrastive framework that maximizes the agreement between differently augmented views of the same image, relying on a simple yet effective mechanism to create positive and negative pairs through data augmentations. In the field of remote sensing, the growing availability of multi-modal data has driven interest in adapting SSL techniques to exploit the complementary nature of different sensor modalities. Multi-modal SSL approaches aim to align representations across diverse data types, such as optical and Synthetic Aperture Radar (SAR) imagery. For instance, recent studies have successfully applied self-supervised methods to tasks like change detection [6, 7, 23] and land cover classification [13, 17], demonstrating the potential of SSL to learn from vast amounts of unlabeled RS data. However, existing multi-modal SSL approaches face limitations in fully leveraging the distinct features of each modality. Many methods emphasize shared information across modalities while overlooking modality-specific attributes that are critical for remote sensing applications, as investigated by [26, 27]. For instance, SAR imagery is valuable for its ability to capture structural information under all weather and lighting conditions, making it indispensable for consistent monitoring. In contrast, optical imagery provides detailed spectral information, offering insights into surface materials and vegetation but is significantly constrained by weather or lighting conditions. The complementary strengths of SAR and optical data highlight their potential when used together, as they provide multi-scale heterogeneity in terms of spatial resolution, spectral richness and consistency.

Recently, probabilistic embeddings have gained significant interest among computer vision researchers as a way to model the inherent uncertainty in multi-modal representations arising from many-to-many image-text matching [20]. Building on this idea, we aim to adapt a joint probabilistic embedding space to the specific challenges of RS data, where the diversity and variability of modalities can introduce substantial uncertainty in representations. This approach is particularly beneficial for remote sensing scenarios characterized by significant heterogeneity across modalities. By incorporating probabilistic embeddings into multi-modal SSL, our framework addresses key challenges in leveraging the complementary nature of different modalities and improves robustness, even when datasets are obstructed or corrupted (e.g. by cloud cover in optical imagery).

In this work, we make two main contributions:

- We introduce **Probabilistic Multi-Modal Learning for Remote Sensing (ProMM-RS)**, a method combining multi-modal contrastive learning with probabilistic embeddings to represent RS data as random variables. ProMM-RS is designed to process multi-source remote sensing data, specifically leveraging Sentinel-1 (SAR) and Sentinel-2 (optical) imagery, which are two widely available modalities in remote sensing data and exhibit complementary properties;

- We evaluate ProMM-RS on a scene classification downstream task, exploring various fusion mechanisms for probabilistic embeddings. By explicitly modeling uncertainty, we aim at improving reliability and robustness, particularly in scenarios involving incomplete or noisy data (e.g. cloud-obstructed optical imagery).

The remainder of this article is organized as follows. Sec. 2 presents an overview of the state of the art in the context of learning RS image representations from a self-supervised and probabilistic perspective. The proposed ProMM-RS method is detailed in Sec. 3. An experimental study involving a scene classification task from optical and radar images is then described in Sec. 4 (with an ablation study provided in Sec. 5), followed by discussions and conclusions (Sec. 6).

2. Related work

2.1. Self-supervised learning

Recently, self-supervised learning has emerged as an interesting solution in both computer vision and remote sensing due to its ability to learn rich, generalizable data representations without the need of (costly) human annotations. This is particularly advantageous in the RS domain, where labeling large amounts of data can be prohibitively expensive and time-consuming.

Among the different families of SSL approaches (relying on pretext tasks, student-teacher strategies, etc.), a notable approach involves contrastive techniques. One of the pioneering approaches, SimCLR [5], relies on considering different augmented views representing the same data (seen as positive pairs) to train a model to bring them closer together in the representation space while negative pairs are moved away from each other. Methods such as MoCo [16] or DINO [2] were also proven efficient on natural and RS imagery to enhance feature extraction [28].

2.2. Multi-modal contrastive learning

Nowadays, remote sensing data are acquired daily from multiple modalities. Optical, multi-spectral, hyper-spectral, Light Detection and Ranging (Lidar) or SAR sensors produce different images and highlight specific characteristics. To leverage the complementary information of multiple modalities, recent publications investigate multi-modal self-supervised learning, mostly on optical and SAR images. A common strategy for multi-modal learning is to consider each modality as an augmented view of the same content to perform cross-modal contrastive learning. This technique was proven efficient for (cross-modal) image-text representations with vision-language models such as CLIP [22] and has been successfully used on various remote sensing tasks such as change detection [6, 7, 23] or classification [13, 17, 25].

However, these methods mainly focus on representing common features and tend to ignore modality-unique subtle features in the joint embedding space. To deal with this issue, *Wang et al.* [26] recently proposed a multi-modal self supervision structure based on BarlowTwins [29] that decouples common and unique embeddings to effectively learn multi-modal representations without losing intra-modal features. In [19], the authors introduced the FactorCL method aiming at capturing both shared and unique information by factorizing task-relevant information. Generative methods and especially masked autoencoders [15] are also commonly used to perform cross-modal pre-training. Both [14] and [3] explored different early/late fusion and masking techniques to combine SAR and optical images while the authors of [1] introduced OmniSat, a model employing both multi-modal contrastive learning and intra-modal reconstruction objectives.

Combining multiple sources of observation and aligning them on a common latent space can provide additional complementary characteristics compared to intra-modal learning. However cross-modal contrastive models typically learn to maximize mutual information shared between modalities and discard unique features [26, 27]. In this article, we aim to address this limitation; ProMM-RS exploits a probabilistic joint embedding space that explicitly models the inherent uncertainty of modality-specific representations. Probabilistic embeddings enable the representation of inputs as distributions rather than fixed points, capturing the variability and ambiguity within each modality. By incorporating uncertainty measurements, ProMM-RS highlights the distinct and valuable features contained in each modality, ensuring that unique characteristics are preserved and leveraged.

2.3. Probabilistic contrastive learning

Probabilistic embedding spaces emerged in the vision-language field as a potential solution to a fundamental challenge in multi-modal tasks such as Image-Text matching: the inherent ambiguity caused by many-to-many correspondences and sparse dataset annotations. Inspired by [20], the authors of [9] proposed the Probabilistic Cross-modal Embeddings (PCME) method to represent embeddings as Gaussian distributions and trained their model using a contrastive loss between Monte-Carlo sampled distributions. This work was then extended in PCME++ [8] by introducing a closed form probabilistic distance and multiple optimization techniques to reduce computational cost and improve efficiency for cross-modal retrieval. Recently, numerous works used such probabilistic embedding spaces for various multi-modal tasks such as soundscape mapping [18], action recognition and video retrieval [21] or face recognition [4, 24].

In ProMM-RS, we propose a novel probabilistic contrastive learning framework inspired by the principles introduced in PCME++ [8]. Our approach redefines the concepts of probabilistic contrastive learning and adapts them to suit the challenges of a multi-modal remote sensing setup.

3. Proposed method

This section presents our proposed probabilistic multi-modal contrastive learning (ProMM-RS) framework for remote sensing image representations. Our model aims at extracting rich and valuable features from modality-specific data while explicitly representing the uncertainty of each representation by modeling them as random variables. Figure 1 provides an overview of the ProMM-RS architecture, taking as input multi-modal RS data (SAR and optical imagery) and mapping them in a joint probabilistic latent space as random variables. The framework involves two main components performed consecutively: (1) a contrastive learning process, which operates both within and across modalities to align representations, (2) a probabilistic learning step, which quantifies and models the uncertainty inherent in the representations from each modality.

Compared to state-of-the-art multi-modal approaches for remote sensing representations such as CROMA [13] or OmniSAT [1], our framework distinguishes itself by explicitly integrating uncertainty into the learned representations. This approach allows us to leverage the complementary informations contained in both modalities and compare their importance for feature extraction. We demonstrate the potential of probabilistic learning by evaluating the joint embedding space on a classification task, experimenting with various fusion strategies to combine modality-specific features.

3.1. Probabilistic visual encoders

To produce probabilistic embedding representing a multivariate normal distribution $Z(\mu, \sigma^2)$, we propose a visual encoder architecture (inspired from [8]) which contains two output heads returning two D-dimensional vectors representing mean and variance, respectively μ and $\log(\sigma^2)$. The overview of our Probabilistic Vision Transformer (PVT) architecture is summarized in Figure 2.

The employed visual backbone is a common Vision Transformer architecture [11] with the last transformer layer duplicated for μ and $\log(\sigma^2)$ heads. On the 12 layers wide vision transformers used in this project, 11 transformer layers are shared for feature extraction while the two output heads are 1-layer transformer blocks. As investigated by [8], increasing the number of unique layers for the $\log(\sigma^2)$ head only marginally improves the efficiency of the model. In practice we only considered 1-layer wide output heads for computational efficiency.

Feature aggregation is done by applying average pooling followed by L2-normalization for the μ head.

Having introduced the PVT for its ability to generate probabilistic embeddings, we integrate in the next section these encoders into our multi-modal contrastive learning framework.

3.2. Multi-modal contrastive learning

We encode each modality separately using dedicated encoders (Probabilistic Vision Transformers (PVT) defined in Sec. 3.1), which are first optimized through a contrastive learning objective to align their features in a shared representation space. Our overall multi-modal contrastive learning framework is illustrated in Figure 1, see box (1). The contrastive learning framework proposed in this section is adapted from SimCLR [5] and conceptually similar to [12]. A SimCLR-like strategy is applied to both intra-modal situation, where augmented views of the same modality are compared, and cross-modal situation, where views of the same geo-located image from different modalities serve as augmented views. This approach ensures that each encoder is specialized for its respective modality while learning representations that can be compared and aligned across modalities. We apply a projection head $g(\cdot)$ before computing the contrastive loss. As evidenced by SimCLR [5], this projection head facilitates alignment and improves the quality of learned representations.

Considering a positive pair of projected representations h_i, h_j from an input batch of size N (hence $2N$ augmented

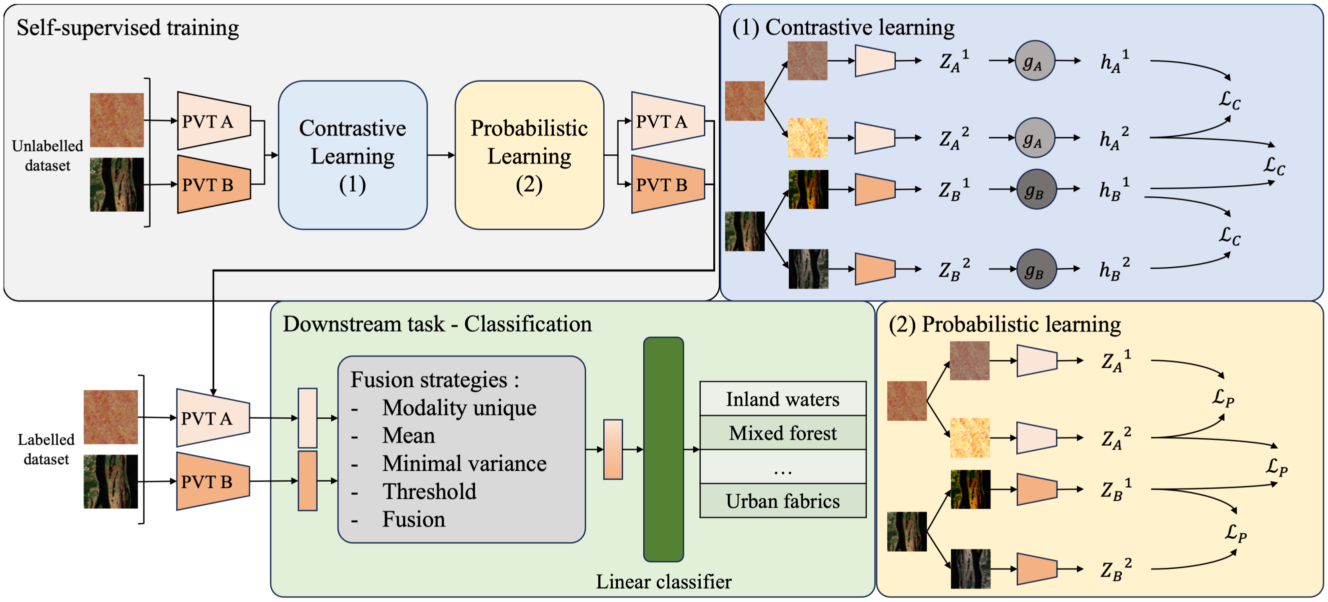

Figure 1. Visualization of the proposed ProMM-RS framework. The framework integrates a self-supervised training phase where unlabeled data are passed through modality-specific Probabilistic Vision Transformer (PVT) to obtain representations (Figure 2). Contrastive learning (1) aligns representations in both intra and inter-modal setup by minimizing the contrastive loss $\mathcal{L}_C$ while probabilistic learning (2) incorporates probabilistic modelling through a probabilistic loss $\mathcal{L}_P$. Representations are evaluated on a scene classification task where labeled data is processed through previously trained frozen encoders. Different combinations of obtained probabilistic embeddings are considered.

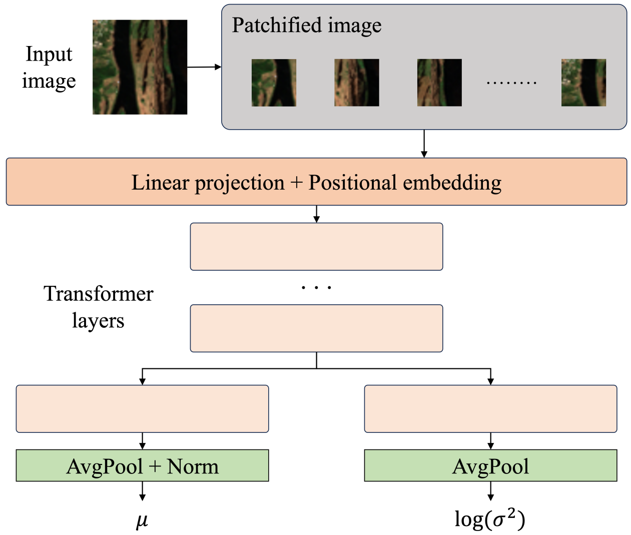

Figure 2. Probabilistic Vision Transformer (PVT) architecture overview. The last layer of a classic vision transformer is duplicated to map the input to a normal distribution parameterized by the μ and $\log(\sigma^2)$ probabilistic heads.

representations), we define the contrastive loss used for pretraining as:

$$l_{ij} = -\log \frac{\exp(\text{sim}(h_i, h_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(h_i, h_k)/\tau)} \tag{1}$$

with τ a temperature coefficient, $sim(\cdot)$ the cosine similarity function defined as $sim(u, v) = \frac{u^T v}{\|u\|\|v\|}$ and $\mathbb{1}_{[k\neq i]}$ the indicator function. The loss for the entire mini-batch can be defined as:

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [l_{2k-1,2k} + l_{2k,2k-1}] \tag{2}$$

We apply this contrastive loss both for intra-modal learning (between augmented views from the same modality) and inter-modal learning (between views of two different modalities).

In a second time, we extend the learned encoders to generate probabilistic embeddings that explicitly model uncertainty, enhancing their capability to distinguish which modality contains the most valuable informations.

3.3. Probabilistic distances

Given two input images x_A and x_B, we encode two probabilistic embeddings $Z_A(\mu_A, \sigma_A^2)$ and $Z_B(\mu_B, \sigma_B^2)$ defined as multivariate normal distributions with output μ and $\log(\sigma^2)$ being D-dimensional vectors.

The purpose of our model is to learn probabilistic embeddings satisfying the following conditions:

- If the embedding contains specific and valuable informations, its variance should be low;

- If the embedding does not contain any characteristic information (e.g. coming from an obstructed image), its variance should be high;
- If an embedding Z_A contains more valuable informations and a better representation than Z_B, its variance σ_A^2 should be lower than σ_B^2.

To adapt our multi-modal framework to probabilistic learning, a probabilistic measure of similarity between two embeddings is required. Closed-form Sampled Distance (CSD) is a probabilistic distance introduced in [8] to perform cross-modal retrieval on probabilistic embeddings. Given two probabilistic embeddings $Z_A(\mu_A, \sigma_A^2)$ and $Z_B(\mu_B, \sigma_B^2)$, the CSD distance is defined as:

$$D_{CSD}(Z_A, Z_B) = \|\mu_A - \mu_B\|_2^2 + \|\sigma_A^2 + \sigma_B^2\|_1, \quad (3)$$

Even though the CSD distance was initially designed for cross-modal retrieval, its properties transfer well to our more general self-supervised learning paradigm:

- Considering a positive pair of inputs x_A and x_B mapped to two probabilistic embeddings $Z_A(\mu_A, \sigma_A^2)$ and $Z_B(\mu_B, \sigma_B^2)$ with fixed means. If the pair is largely resembling, it should be mapped close in the joint embedding space, i.e. $\|\mu_A - \mu_B\|_2^2 \approx 0$ and the variances σ_A^2 and σ_B^2 should collapse towards 0;
- Now considering a negative pair of embeddings with fixed means, if the pair is resembling (i.e. $\|\mu_A - \mu_B\|_2^2 \approx 0$), the contribution of $\|\sigma_A^2 + \sigma_B^2\|_1$ in the loss will be greater and σ_A^2 and σ_B^2 will be increased. If the pair is largely dissimilar (i.e. $\|\mu_A - \mu_B\|_2^2 \gg 0$), the variances will still be increased but their overall contribution in the loss will be lower.

Having defined a probabilistic framework with encoders mapping to normal distributions output and proper probabilistic distances to assess the similarity of our learned embeddings, we propose to apply a probabilistic contrastive learning strategy to model the variance embeddings.

3.4. Probabilistic contrastive learning

To keep dimension-wise correspondence between our mean and variance representations, we apply contrastive learning directly on the embeddings $Z_A(\mu_A, \sigma_A^2)$ and $Z_B(\mu_B, \sigma_B^2)$ without passing them through the projection head $g(\cdot)$. Our overall probabilistic multi-modal contrastive learning framework is illustrated in Figure 1, see box (2).

Following [8] implementation of the previously defined Closed form sampled distance loss (Equation 3), we consider a soft probabilistic matching loss as follows:

$$\begin{aligned}\mathcal{L}_{\text{match}} &= m_{AB} \log \text{sigmoid}(-aD(Z_A, Z_B) + b) \\ &\quad - (1 - m_{AB}) \log \text{sigmoid}(aD(Z_A, Z_B) - b)\end{aligned} \quad (4)$$

where $m_{AB} \in \{0, 1\}$ is a matching indicator and a, b learnable scalar values. This $\mathcal{L}_{\text{match}}$ function is computed for all pairs in the mini batch as a contrastive learning objective.

We also consider a pseudo-positive strategy as formulated by [8] to account for the numerous false negatives. For a positive match (x_A, x_B), we consider $x_B^{'}$ to be a pseudo-positive match with x_A if $D_{CSD}(Z_A, Z_B^{'}) < D_{CSD}(Z_A, Z_B)$. $\mathcal{L}_{\text{pseudo-match}}$ is then computed for all pseudo-positive pairs following Equation 4.

To prevent the collapse of σ, we add a Variational Information Bottleneck loss (VIB) as evidenced by [20] and [9]. In practice, this is done by minimizing the KL divergence between the learned distribution and $\mathcal{N}(0, 1)$.

Our objective function becomes:

$$\mathcal{L}_P = \mathcal{L}_{\text{match}} + \alpha \mathcal{L}_{\text{pseudo-match}} + \beta \mathcal{L}_{\text{VIB}} \quad (5)$$

with α the pseudo-match coefficient and β the VIB coefficient.

To incorporate both inter-modal uncertainty coming from different modality representations and intra-modal uncertainty potentially coming from noisy image capture, we train our soft contrastive loss on both intra-modal setup (between augmented views from the same modality) defined as $\mathcal{L}_{P-intra}$ and inter-modal setup (between views of two different modalities) defined as $\mathcal{L}_{P-inter}$:

$$\mathcal{L} = \mathcal{L}_{P-inter} + \delta(\mathcal{L}_{P-intra}^A + \mathcal{L}_{P-intra}^B) \quad (6)$$

with δ the intra-modal coefficient.

3.5. Fusion strategies for probabilistic embeddings

To leverage the uncertainty information of our probabilistic embeddings, we propose multiple data fusion strategies.

Given two probabilistic embeddings from two modalities A and B defined by multivariate normal distribution $Z_A(\mu_A, \sigma_A^2)$ and $Z_B(\mu_B, \sigma_B^2)$, we highlight in Tab. 1 the different combination strategies used to define the resulting Y vector representing a geo-located pair of images.

These strategies provide a foundational approach to fusing probabilistic embeddings, focusing on leveraging the uncertainty of learned representations.

4. Experimental study

As a preliminary experimental study, we evaluate our learned representations with a scene classification downstream task from Sentinel optical and radar images, experimenting with various fusion strategies to combine modality-specific features. We aim to showcase the potential of probabilistic embeddings as a way to measure the relevancy of each modality representation, especially when one of the modalities suffers from occlusions.

Table 1. Fusion strategies for probabilistic embeddings. Given two probabilistic embeddings $Z_A(\mu_A, \sigma_A^2)$ and $Z_B(\mu_B, \sigma_B^2)$ from modalities A and B: (1) *Modality Unique*: uses only the mean embedding of one modality. (2) *Mean*: averages the mean embeddings from both modalities. (3) *Fusion*: concatenates the embeddings from both modalities. (4) *Minimal Variance*: selects the mean value from the modality with the smallest variance (highest confidence) for each dimension. (5) *Threshold*: uses the mean from modality A if its variance is below a given threshold; otherwise, it uses the mean from modality B.

Strategy	**Formula**
Modality Unique	$Y_{unique} = \mu$
Mean	$Y_{mean} = \frac{1}{2}(\mu_A + \mu_B)$
Fusion	$Y_{fusion} = (\mu_A, \mu_B)$
Minimal Variance	$Y_{min-var} = (y^1, \dots, y^i, \dots, y^d)$ where $y^i = \begin{cases} \mu_A^i & \text{if } \sigma_A^i < \sigma_B^i, \\ \mu_B^i & \text{otherwise.} \end{cases}$
Threshold	$Y_{threshold} = (y^1, \dots, y^i, \dots, y^d)$ where $y^i = \begin{cases} \mu_A^i & \text{if } \sigma_A^i < \tau, \\ \mu_B^i & \text{otherwise.} \end{cases}$

4.1. Data

All experiments were conducted on the newly introduced refined BigEarthNet dataset [10] (also referred to as reBEN), which contains 549488 multi-modal (Sentinel-1 and Sentinel-2) image pairs, with additional 69450 cloudy/snowy pairs. We considered for the experiments: i) the VV and VH bands for the Sentinel-1 images, ii) the spectral bands associated with 10m and 20m spatial resolution for Sentinel-2. All images are of dimension 120×120 pixels. Input data is transformed to produce a corresponding pair of images following classic data augmentation techniques [5, 28]: RandomResizedCrop, horizontal flip and color jittering. We used for all experiments the train/validation/test splits provided by [10].

Note that baseline tests in [10] were conducted on the reBEN dataset without cloudy and snowy patches. To assess the potential of probabilistic embeddings on noisy/corrupted datasets, we evaluate our models on both the same dataset as [10] and the full reBEN with cloudy/snowy patches.

4.2. Experimental setup

We considered as our base encoders two modality-specific Vision Transformers. As shown in Figure 1, we trained these encoders in two steps: i) A multi-modal contrastive learning strategy for 50 epochs for intra-modal learning and 50 epochs for inter-modal learning. We set the batch size to 1024 under an AdamW optimization strategy, with an initial learning rate of 10^{-4} and weight decay of 0.05. A cosine scheduler with 10 warm-up epochs is applied; ii) Probabilistic contrastive learning for 20 epochs with a batch size of 512 under an AdamW optimization strategy, with an initial learning rate of 10^{-4} and weight decay of 0.05. A cosine scheduler with 5 warm-up epochs is applied. As discussed in Sec. 5.2, mean embeddings and backbone are frozen for the probabilistic training step, only the layers specific to the $\log(\sigma^2)$ head are trained. For probabilistic training, we set the pseudo-match coefficient α to 0.1, the VIB coefficient β to 0.001 and the intra-modal coefficient δ to 0.2. We perform all training experiments on 8 NVIDIA V100-32Gb GPU.

4.3. Numerical results and discussions

We report hereinafter the results and performances of our different training strategies compared to linear probing on classical self-supervised learning framework. To highlight the value of probabilistic embeddings under a dataset where one of the modalities may suffer from occlusions, we evaluate in Tab. 2 our models on the reBEN dataset including cloudy and snowy patches. We use as metrics Average Precision (AP^M/AP^μ), F1 score ($F1^M/F1^\mu$), and Precision ($Precision^M/Precision^\mu$), evaluated both as macro-averaged (M) and micro-averaged (μ) values. Macro-averaging computes the metric independently for each class and then averages the results, while micro-averaging aggregates contributions from all classes to calculate a single overall metric. We observe several key aspects:

- **Our cross-modal model is biased on Sentinel-2 representations.** The results indicate that the learned embeddings of Sentinel-2 contain richer and more relevant features, as the use of averaged Sentinel-1 and Sentinel-2 embeddings does not perform as well as using Sentinel-2 alone;

- **Probabilistic learning effectively learns dimension-wise embedding relevance.** However, we observe that our model is performing consistently better than *S1+S2(mean)* for the probabilistic strategies considered. This highlights the fact that our learned variance is measuring dimension-wise relevance of each modality representation. It enables the model to focus on Sentinel-1 inputs only when Sentinel-2 does not carry useful information. This selective capability allows our model to leverage the complementary information of both modalities only when needed, rather than combining them indiscriminately.

To compare our model with existing fully supervised baselines of [10], we provide in Tab. 2 the results of the different training strategies considered in this project on a Vision Transformer Base backbone, trained on the reBEN dataset **without cloudy/snowy patches**.

We observe that our learned multi-modal representations perform better under a linear probing evaluation than fully supervised models, highlighting the benefits of self-supervised

Table 2. Performances for scene classification on reBEN dataset using a multi-modal self-supervised Vision Transformer backbone trained under contrastive learning and probabilistic learning with frozen mean embeddings. We evaluate the different fusion strategies highlighted in Tab. 1. We report the following macro and micro metrics : Average Precision (AP), F1 score and Precision. *Prob* refers to the use of probabilistic embeddings for fusing both modalities. For fair comparison, we do not consider the results of *S1+S2* for best performing strategy as the linear classifier receives doubled information from this modality.

		Results					
Modality	Prob?	AP$^{\mathrm{M}}$	AP$^{\mu}$	F_1^{M}	F_1^{μ}	Precision$^{\mathrm{M}}$	Precision$^{\mu}$
Evaluated on reBEN dataset **with cloudy/snowy images**							
S1	✗	61.28	78.01	50.31	67.06	63.32	74.82
S2	✗	67.97	81.99	**56.71**	71.74	67.16	**77.33**
S1+S2(mean)	✗	65.00	80.64	54.72	70.26	**67.46**	77.08
S1+S2(fusion)	✗	69.49	82.93	57.55	72.55	68.08	77.94
Minimal variance	✓	**69.04**	**82.44**	56.24	71.74	65.67	77.17
Threshold 0.7	✓	67.92	81.95	56.32	**71.84**	67.27	**77.33**
Evaluated on reBEN dataset **without cloudy/snowy images**							
S1	✗	64.51	79.65	53.01	68.55	63.35	75.38
S2	✗	**71.34**	**83.67**	**59.94**	**73.55**	68.49	**78.26**
S1+S2(mean)	✗	69.26	82.66	58.70	72.55	**70.61**	77.92
Minimal variance	✓	71.30	83.60	59.50	73.37	68.91	78.15
Fully supervised model trained on reBEN dataset **without cloudy/snowy images** [10]							
S1	✗	50.94	70.23	39.78	58.98	58.45	71.09
S2	✗	63.42	81.41	57.46	71.92	68.29	76.50
S1+S2	✗	66.09	83.37	59.81	73.54	70.87	77.74

learning for reliable feature extraction. However, not evaluating the snow or cloud covered patches decreases the efficiency of our probabilistic minimal variance strategy, which is under performing compared to Sentinel-2 embeddings.

Even though the overall performance of the minimal variance strategy is weaker, we assess that our probabilistic strategy is still learning valuable information from Sentinel-1 images by studying, in Tab. 3, the Average Precision scores of different labels.

Typically, Sentinel-1 images are known to work well in urban areas, as the various mix of buildings, roads and infrastructures creates strong and distinct returns as the waves are reflected back to the SAR sensor. Similarly with water bodies, the strong contrast difference between the dark appearance of water surfaces and the surrounding land/man-made features makes it easy to identify water bodies in SAR images. On the opposite, vegetated environments tend to absorb the radar signal and their homogeneous structure make SAR images less distinct. Given these considerations, we selected five labels showcasing valuable differences between both modalities: Urban fabrics, Industrial units, Inland waters, Coniferous forest and Mixed forest. As expected, we observe that the average precision score of our minimal variance strategy is outperforming Sentinel-2 modality on urban and water environments, since it effectively captures the relevant information in Sentinel-1 to perform classification, while Sentinel-2 is still the strongest option for discriminating between vegetation environments like forests.

5. Ablation studies

The experimental study is completed below by a selected set of ablations about pseudo-positive matches and the effect of freezing the mean embeddings.

5.1. Effects of pseudo-positive matches

Tab. 4 reports the results of the probabilistic strategy with and without incorporation of the pseudo-positive matches. When training without pseudo-positive match loss term (Equation 5), we observe a 0.55% point drop in performance on macro-average precision and 1.17% point drop on micro-average precision. Theses results illustrate the importance of considering more nuanced positive and negative matches for uncertainty estimations, to adress efficiently the multiplicity of labels and false negatives in contrastive techniques.

5.2. Effects of freezing mean embeddings

In contrast to the probabilistic contrastive learning approach presented in [8], which does not use frozen mean embeddings, our results in Tab. 4 demonstrate that allowing the probabilistic training to modify mean embeddings significantly reduces classification performance. This effect suggests that unfreezing mean embeddings during training could result in incorrect and less stable representations. While letting mean embeddings be adjusted during probabilistic training would theoretically help the model further distinguish positive and negative matches, our probabilis-

Table 3. Average Precision performances on reBEN dataset **without cloudy/snowy images** on specific labels using a multi-modal self-supervised Vision Transformer backbone trained under contrastive and probabilistic learning strategies with frozen mean embeddings. Experimental setup and model used is the same as Tab. 2.

		Average Precision				
Modality	Prob?	Urban fabric	Industrial units	Inland waters	Coniferous forest	Mixed forest
S1	✗	80.07	46.18	87.42	83.24	79.52
S2	✗	81.43	49.20	89.02	**89.25**	**84.70**
S1+S2 (mean)	✗	81.61	**49.66**	88.95	87.92	83.91
Minimal variance	✓	**82.05**	49.39	**89.07**	87.92	84.33

Table 4. Ablation study : Impact of the pseudo-positive (PP) matching loss and freezing mean embeddings on performances. Experimental setup and model used is the same as Tab. 2. We report the following macro and micro metrics : Average Precision, F1 score and Precision.

			Results					
Modality	PP?	Frozen mean?	AP^{M}	AP^{μ}	F_1^{M}	F_1^{μ}	Precision^{M}	Precision^{μ}
Minimal variance	✗	✓	67.87	81.89	**56.29**	**71.74**	66.09	77.18
	✓	✗	65.58	79.81	53.75	69.29	**68.84**	**77.24**
	✓	✓	**69.04**	**82.44**	56.24	**71.74**	65.67	77.17

tic model may struggle to keep the pre-trained embeddings aligned and relevant, ultimately leading to worsened performances. Freezing the mean embeddings on the other hand keep the consistent and reliable joint embedding space obtained through the multi-modal pre-training, while focusing on determining the variations between inputs rather than altering the input centroid. Further research and hyperparameter sweeps would be beneficial to analyze the effect of CSD-based contrastive loss on the mean embeddings.

6. Conclusion and perspectives

In this work, we introduced ProMM-RS (Probabilistic Multi-Modal Learning for Remote Sensing), a novel method that combines multi-modal contrastive learning with probabilistic embeddings to represent remote sensing data as random variables. ProMM-RS is specifically designed to process multi-source remote sensing data, leveraging the complementary properties of Sentinel-1 (SAR) and Sentinel-2 (optical) imagery. By modeling data as probabilistic embeddings, we aim to enhance the robustness and reliability of representations, particularly in challenging scenarios such as cloud-obstructed optical imagery or incomplete data.

We evaluated the proposed approach on a scene classification downstream task, exploring various fusion mechanisms for probabilistic embeddings. This evaluation demonstrated the potential of explicitly modeling uncertainty to effectively address the heterogeneity of multi-modal RS data.

Even though we achieve overall good results with probabilistic training and simple fusion methods such as *Minimal variance* and *Threshold*, these methods focus solely on which input is the most similar and has the least uncertainty of representation compared to the other inputs in the batch, without accounting for the specific strengths of each modality in extracting relevant features. This is particularly visible when testing our models on the dataset without occluded images (as presented in Tab. 2 and Tab. 3), where the inherent superiority of Sentinel-2 representations makes the minimal variance strategy weaker on labels with less distinctive SAR image features (e.g. vegetation environments). To address this issue, future work could explore more sophisticated fusion techniques that better leverage the complementary information in multi-modal probabilistic embeddings by integrating variances more effectively and incorporating each modality's intrinsic value, potentially enhancing the representation quality for downstream tasks.

Additionally, extending this framework to other modalities or time-series could enhance the obtained representations.

For example, integrating temporal dynamics could help the model distinguish temporary or persistent features, further improving uncertainty quantification. Finally, more thorough analysis of the obtained probabilistic embeddings would be necessary to guide new studies and to use the potential of these representations to optimize multi-modal learning frameworks, enhancing interpretability, and to improve the robustness of models in diverse Earth observation scenarios.

Acknowledgment. This work was supported in part by ANR project ANR-23-IAS1-0002 GEO ReSeT and was granted access to the HPC resources of IDRIS under the allocation 2024-AD010615611 made by GENCI.

References

[1] Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loïc Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *ECCV, Procs*, volume 15086 of *Lecture Notes in Computer Science*, pages 409–427, 2024. 2, 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV, Procs.*, pages 9630–9640, 2021. 2

[3] Hugo Chan-To-Hing and Bharadwaj Veeravalli. FUS-MAE: A cross-attention-based data fusion approach for masked autoencoders in remote sensing. In *IGARSS, Procs.*, pages 6953–6958, 2024. 2

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR, Procs.*, pages 5709–5718, 2020. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML, Procs.*, pages 1597–1607, 2020. 1, 2, 3, 6

[6] Yuxing Chen and Lorenzo Bruzzone. A self-supervised approach to pixel-level change detection in bi-temporal RS images. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–11, 2022. 1, 2

[7] Yuxing Chen and Lorenzo Bruzzone. Self-supervised change detection in multiview remote sensing images. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–12, 2022. 1, 2

[8] Sanghyuk Chun. Improved probabilistic image-text representations. In *ICLR, Procs.*, 2024. 3, 5, 7

[9] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR, Procs.*, pages 8415–8424, 2021. 3, 5

[10] Kai Norman Clasen, Leonard W. Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. *CoRR*, abs/2407.03653, 2024. 6, 7

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR, Procs.*, 2021. 3

[12] Zhixi Feng, Liangliang Song, Shuyuan Yang, Xinyu Zhang, and Licheng Jiao. Cross-modal contrastive learning for remote sensing image classification. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–13, 2023. 3

[13] Anthony Fuller, Koreen Millard, and James R. Green. CROMA: remote sensing representations with contrastive radar-optical masked autoencoders. In *NeurIPS, Procs.*, 2023. 1, 2, 3

[14] Jakob Hackstein, Gencer Sumbul, Kai Norman Clasen, and Begüm Demir. Exploring masked autoencoders for sensor-agnostic image retrieval in remote sensing. *CoRR*, abs/2401.07782, 2024. 2

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR, Procs.*, pages 15979–15988, 2022. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR, Procs.*, pages 9726–9735, 2020. 2

[17] Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks. *CoRR*, abs/2209.02329, 2022. 1, 2

[18] Subash Khanal, Eric Xing, Srikumar Sastry, Aayush Dhakal, Zhexiao Xiong, Adeel Ahmad, and Nathan Jacobs. PSM: learning probabilistic embeddings for multi-scale zero-shot soundscape mapping. In *ACM Multimedia, Procs.*, pages 1361–1369, 2024. 3

[19] Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS, Procs.*, 2023. 2

[20] Seong Joon Oh, Kevin P. Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C. Gallagher. Modeling uncertainty with hedged instance embeddings. In *ICLR, Procs.*, 2019. 2, 3, 5

[21] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *CVPR, Procs.*, pages 14691–14701, 2022. 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML, Procs.*, volume 139, pages 8748–8763, 2021. 2

[23] Sudipan Saha, Patrick Ebel, and Xiao Xiang Zhu. Self-supervised multisensor change detection. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–10, 2022. 1, 2

[24] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *ICCV, Procs.*, pages 6901–6910, 2019. 3

[25] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *CVPR, Procs.*, pages 1182–1191, 2021. 2

[26] Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decur: decoupling common & unique representations for multimodal self-supervision. *CoRR*, abs/2309.05300, 2023. 1, 2

[27] Yi Wang, Conrad M. Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS, Procs.*, pages 139–142, 2022. 1, 2

[28] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. *CoRR*, abs/2211.07044, 2022. 2, 6

[29] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML, Procs.*, volume 139, pages 12310–12320, 2021. 2

[30] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5:8–36, 2017. 1