ISPRS Journal of Photogrammetry and Remote Sensing 223 (2025) 173–187

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Semantically-Aware Contrastive Learning for multispectral remote sensing images

Leandro Stival [a], Ricardo da Silva Torres [b], Helio Pedrini [a,*]

[a] Institute of Computing, University of Campinas, Campinas, SP, Brazil
[b] Wageningen University and Research, Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Remote sensing images
Self-supervised learning
Multi-spectral image
Semantic information
Contrastive learning

ABSTRACT

Satellites continuously capture vast amounts of data daily, including multispectral remote sensing images (MSRSI), which facilitate the analysis of planetary processes and changes. New machine-learning techniques are employed to develop models to identify regions with significant changes, predict land-use conditions, and segment areas of interest. However, these methods often require large volumes of labeled data for effective training, limiting the utilization of captured data in practice. According to current literature, self-supervised learning (SSL) can be effectively applied to learn how to represent MSRSI. This work introduces Semantically-Aware Contrastive Learning (SACo+), a novel method for training a model using SSL for MSRSI. Relevant known band combinations are utilized to extract semantic information from the MSRSI and texture-based representations, serving as anchors for constructing a feature space. This approach is resilient against changes and yields semantically informative results using contrastive techniques based on sample visual properties, their categories, and their changes over time. This enables training the model using classic SSL contrastive frameworks, such as MoCo and its remote sensing version, SeCo, while also leveraging intrinsic semantic information. SACo+ generates features for each semantic group (band combination), highlighting regions in the images (such as vegetation, urban areas, and water bodies), and explores texture properties encoded based on Local Binary Pattern (LBP). To demonstrate the efficacy of our approach, we trained ResNet models with MSRSI using the semantic band combinations in SSL frameworks. Subsequently, we compared these models on three distinct tasks: land cover classification task using the EuroSAT dataset, change detection using the OSCD dataset, and semantic segmentation using the PASTIS and GID datasets. Our results demonstrate that leveraging semantic and texture features enhances the quality of the feature space, leading to improved performance in all benchmark tasks. The model implementation and weights are available at https://github.com/lstival/SACo — As of Jan. 2025.

1. Introduction

Today, technology integration into human tasks is nearly everywhere. This trend is especially notable in Earth Observation (EO), where satellites produce vast amounts of data that require proper storage, processing, and analysis to support various applications, including land cover state and change assessment (Shafique et al., 2022), forest monitoring (Fassnacht et al., 2024), food security/early warning systems (Krishnamurthy R et al., 2020), water management and soil protection (Arefin et al., 2020), as well as urban mapping (Yin et al., 2021). Therefore, the automation of these processes is a prominent and active research area in computer vision, particularly concerning Remote Sensing Images (RSI).

However, there is not just one type of RSI, as the sensors used to capture this information exhibit differences in how the data is encoded and made available. In this study, we examine multispectral images (MSI), which are defined mainly by the number of channels (or bands) captured by the sensors. In most cases, multispectral remote sensing images (MSRSI) capture information within the 0.433 μm - 0.2280 μm. In the case of the Satellite Sentinel 2, this range is divided into 13 bands (Phiri et al., 2020).

In addition to the numerous bands present in the images, this type of data is also generated in large quantities. This abundance creates extensive opportunities for constructing datasets that machine learning algorithms can utilize to address remote sensing tasks. However, labeling these images remains costly, as it requires remote sensing experts to identify land use types and their associated visual properties, and then annotate regions (Xu et al., 2022). This assertion is further corroborated

* Corresponding author.
E-mail addresses: leandro.stival@ic.unicamp.br (L. Stival), ricardo.dasilvatorres@wur.nl (R. da Silva Torres), helio@ic.unicamp.br (H. Pedrini).

https://doi.org/10.1016/j.isprsjprs.2025.02.024
Received 14 September 2024; Received in revised form 13 February 2025; Accepted 27 February 2025
Available online 18 March 2025

by comparing the sizes of datasets presented in the literature. Those with labels are significantly lower than those without (Mall et al., 2022).

Therefore, techniques for training machine learning models that do not demand labeling have gained attention (Wang et al., 2022; Tao et al., 2023; Jung et al., 2021). These approaches often produce results similar to those of models trained using supervised techniques with labeled data. This phenomenon is not limited to MSRSI processing but is also present in other areas of visual computing, such as image classification (Misra and Maaten, 2020), object detection (Huang et al., 2022), image segmentation (Ziegler and Asano, 2022), and activity recognition (Jaiswal et al., 2020).

However, the development of these techniques encounters inherent challenges stemming from *learning* in the absence of labeled data. The primary challenge lies in determining the most effective way to identify intrinsic visual properties that distinguish images from each other (Jaiswal et al., 2020; Newell and Deng, 2020). A promising avenue explored in the literature commonly employs contrastive techniques to guide the creation of image representations (Jaiswal et al., 2020). These techniques use vector representations of images, where the distance between them is minimized for 'similar' samples and maximized for 'different' ones.

Several proposals have emerged in recent years on generating and approximating image representations (Chen et al., 2020; Caron et al., 2020; Kalantidis et al., 2020; Chuang et al., 2022). These include methods such as applying random modifications to the image and approximating them with the original representation, as in the SimCLR method (Chen et al., 2020), as well as using an extensive list of negative samples simultaneously, as seen in MoCo (He et al., 2020) and its variations.

Using contrastive techniques in self-supervised learning has also been applied in MSRSI processing. The objective is to develop a model capable of representing these images, which could subsequently be employed in other remote sensing tasks, including detecting changes in regions and classifying land cover types. This approach was explored, for instance, in the studies involving SeCo (Manas et al., 2021), CACo (Mall et al., 2023), and MATTER (Akiva et al., 2022). Current methods yield satisfactory results, but they often rely on a single approach for semi-supervised learning, not fully benefiting from the potential of integrating multiple learning strategies seamlessly. Another issue relies on the fact that those methods overlook the intricate semantic information inherent in MSRSI band representations.

This work introduces Semantically-Aware Contrastive Learning (SACo+), a novel method for training a model using SSL for MSRSI that aims to address both limitations. SACo+ provides an integrated framework that supports SSL based on augmented instances and temporal variation. Moreover, SACo+ improves the representation of the MSRSI feature space by incorporating information from the relationships among MSRSI bands while also leveraging temporal and texture information about the region. To achieve this, we organize the bands of the MSRSI images into groups representing known categories documented by the European Space Agency (ESA) (GISGeography, 2024). SACo+ also explores textural properties as visual cues in the creation of feature spaces. This enables the creation of a feature space that remains invariant to seasonal and environmental changes. Each of these groups was individually processed to generate suitable representations for MSRSI. A visual example of *semantics*-informed band groups is shown in Fig. 1. The figure illustrates how different semantic groups encode visual properties of a *forest* and *residential* areas. Bands are combined according to distinct criteria, leading to the arrangement of five semantic groups: RGB, Infrared, Agriculture, Urban, and Atmospheric Penetration. It can be observed that the RGB group can provide information about the natural colors of the region, while the infrared is effective in emphasizing the boundaries of relevant objects within the image. Agricultural bands, on the other hand, are promising to encode

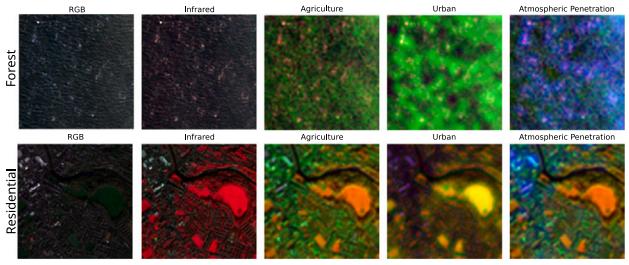


Fig. 1. The illustration depicts semantic groups derived from the Sentinel-2 bands, which encode the visual properties of *forest* and *r*esidential areas. The red, green, and blue (RGB) bands are capable of capturing natural colors, while the infrared (IR) band emphasizes object boundaries. These encoded characteristics are explored in the SACo†'s training process.

the different 'structures' connected with vegetation types found within the image.

However, creating a feature space rich in semantic information alone is not enough to ensure invariance to all types of changes. For example, semantic bands (e.g., near-infrared bands) can represent soil types well, but only texture representations can encode changes in patterns in regions where land use is similar (e.g., different types of crops, urban and industrial areas).

The issue of texture representation in MSRSI images is also due to the fact that band images typically have different resolutions. In some instances, this results in a low-resolution image with a corresponding reduction in the quality of the texture information. To address this challenge, some researchers have employed a texture extractor to enhance the fidelity of the texture data, such as the Gray Level Co-occurrence Matrix (GLCM) to perform feature fusion with machine learning models (Wang and Gu, 2024), and CNN trained to extract texture features (Xu et al., 2024).

In the proposed SACo+ approach, the semantic and texture features from the MSRSI are used to construct a feature space related to regions. Bands conveying information about vegetation, urban areas, and near-infrared exhibited similar features within the same image and texture details about the land cover. During the training process, the output of our encoder is compared to the mean vector of all semantic groups, making the method flexible to changes in the number of groups and how bands are combined. We integrate semantic and texture information into state-of-the-art training protocols of self-supervised learning methods for MSRSI, allowing us to address three research questions: RQ1: Would the use of semantic groups of bands lead to relevant feature spaces for MSRSI? RQ2: Would the combination of semantic features with texture from MSRSI increase the quality of features extracted? RQ3: Would the SACo+ encoder be effective when applied in downstream tasks, leading to superior results compared to other state-of-the-art self-supervised methods?

To address RQ1, we trained ResNet-18 and ResNet-50 models using the SACo+ methodology and compared results with the MoCo v2 (He et al., 2020), SeCo (Manas et al., 2021), CACo (Mall et al., 2023), SpectralGPT (Hong et al., 2024) and DINO-MC (Wanyan et al., 2024). To tackle RQ2, we explored different training procedures, where the encoder is trained just using the semantic information SACo and using a combination of semantic features and texture SACo+. Finally, to address RQ3, we tested our encoder in three different applications: land cover classification, change detection, and semantic segmentation. Using the EuroSAT dataset, we trained a single MLP layer on top of the encoder to perform land cover classification. For change detection using the OSCD dataset and semantic segmentation on the PASTIS and GID dataset, we used a decoder to predict changes in pixels and their classes. The obtained results demonstrate the effectiveness of incorporating semantic information in constructing the feature space.

The state-of-the-art self-supervised contrastive approaches present promising results for the extraction of features from the MSRSI, which can be applied in downstream tasks. However, these methods are designed to process images in general, which may result in the omission of certain intrinsic information unique to MSRSI images. Our primary contribution is to demonstrate that semantic and texture features in the RSI domain can be effectively utilized to generate a comprehensive feature space through the application of cutting-edge self-supervised contrastive learning techniques. Our approach, therefore, constitutes a substantial advancement in the extraction and representation of semantically-enriched features for Remote Sensing Imagery (RSI). We provide a versatile framework that can be seamlessly adapted to support diverse semantic representations of multi-spectral RSI (MSRSI).

In short, the main contributions of the paper are:

- we demonstrate that the use of semantic information encoded in band groups of multi-spectral RSI (MSRSI) and texture information contribute to the creation of effective representations;
- we introduce a new contrastive learning self-supervision approach that explores band combination, texture, augmentation, and temporal information at the same time;
- we demonstrate that the proposed self-supervision approach produces representations that can lead to effective models when fine-tuned to three downstream tasks (e.g., change detection, land cover classification, and semantic segmentation).

2. Related work

This section overviews concepts and relevant studies related to the topic investigated in this work.

2.1. Self-supervised learning

Given the substantial volume of data generated daily and the requirement for high levels of expertise and experience to label this data, self-supervised learning (SSL) approaches have gained prominence. Typically, these approaches are trained based on contrastive learning methods, where the model aims to keep the representations (feature vectors) of similar samples (*positive*) close while simultaneously maximizing differences for samples known as dissimilar (*negative*) (Tao et al., 2023; Chen et al., 2022).

Studies have focused on the proposal of SSL methods that are expected to be effective, therefore leading to results that are close to those observed for supervised methods (Tao et al., 2023). Self-supervised methods, more specifically the contrastive methods, can compare the samples at different levels of abstraction to create a latent space able to represent them. SSL methods can be categorized into three main categories (Tao et al., 2023): instance-level, category-level, and time-series-level.

- **Instance:** This is the classical method to realize contrastive learning, where the samples are manipulated individually and the positive pairs are usually defined based on random augmented versions of the original sample.
- **Categorical:** This formulation relies on creating groups or categories able to represent samples and pull these groups close or apart. For example, we can use a cluster of samples for the same region and use this cluster to guide how representations could be closer in the feature space.
- **Temporal:** This approach explores different temporal versions of samples. The goal is to refine image representations by contrastive learning in such a way that samples related to the same regions at different time stamps would be paired as similar.

Fig. 2 illustrates these three SSL methods.

A classical contrastive learning approach in deep learning applications is the SimCLR method (Chen et al., 2020), which aims to bring the features generated from an image and its augmented version closer together while treating negative samples as two different randomly selected images from the same dataset.

One of the fundamental techniques employed in MSRSI is the Momentum Contrast (MoCo v2) (He et al., 2020) methodology. This approach represents a robust framework for contrastive learning, wherein an input sample, denoted by x, is augmented to generate two distinct versions: x^q and x^{k+}. The contrastive method then aims to align the representations of these two augmented versions, while a *queue* or memory bank comprising a thousand negative samples, typically denoted as x^{k-} is utilized to achieve this alignment. The learning process is based on two instances of the same encoder. After training the first encoder, its weights are slightly modified and used in the second encoder for inference.

The Seasonal Contrast (SeCo) (Manas et al., 2021) methodology uses the same idea to create a memory bank with the negative samples x^{k-} and a momentum parameter to update the instances of the encoder. However, the main difference is the presence of temporal samples x^t, where these images are from the same local of x^q but at a different time stamp (weeks or months ahead). The created feature space is expected to encode both seasonal and augmentation variations.

As a direct evolution of SeCo, we can highlight Change-Aware Sampling (CACo) (Mall et al., 2023), where the trained model generates the feature space to detect changes within images. Different from SeCo, CACo uses a long time range (e.g., years) for the definition of samples used in the contrastive learning process. The goal is to encode *permanent* changes in a region instead of only the seasonal ones.

Representing features for SSL methods is a primary and extensively explored topic in the literature for MSRSI. The presence of numerous bands with varying resolutions makes tasks with this type of data challenging. However, some authors have explored SSL in a different direction, focusing on encoding texture information (Akiva et al., 2022). The proposed formulation explores pixel-wise similarities defined in terms of textural properties. Different from our formulation, their study does not account for temporally-defined samples in the contrastive learning process.

Another trend in the SSL area for RSI applications relies on the use of learning procedures based on masked autoencoders (MAE) (He et al., 2022), different from the contrastive methods, these methods do not use the feature vectors to create positive or negative samples, the learning process occurs predicting missing parts (the masked regions) of the input RSI. One recent example is the MultiSSL (Xue et al., 2025) method, in which a multimodal training approach is applied using hyperspectral images combined with very high-resolution (VHR) images in an asymmetric encoder–decoder training process. In this process, the encoder is trained to produce features from each kind of image using a cross-attention method among the feature extracted and the positional embeddings. Later, the decoder is trained to reconstruct the masked parts of these inputs. This asymmetric training process ensures that the encoder is capable of producing features from distinct data modalities. Another study that employs the masked approach is the S2MAE (Li et al., 2024a), which applies a 3D transformer autoencoder with a 90% mask ratio to training samples. A lightweight decoder was subsequently employed to reconstruct the information and predict the masked parts. The employment of the 3D transformer network has enabled the use of a high mask ratio in a training process comprising over one million images, resulting in a robust feature representation of spectral remote sensing images.

Another extension to the MAE (He et al., 2022) methodology was proposed by Li et al. (2024b). Their study incorporated a scaling center crop operation into the training process. This resulted in the generation of a new cropped and rotated version of the sample while maintaining the integrity of the primary scene. This process demonstrated that

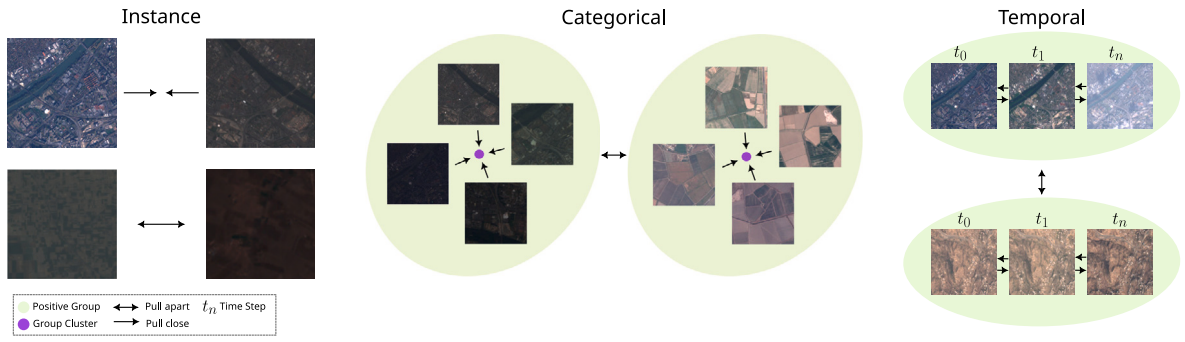


Fig. 2. Illustration of SACo+'s training procedures: instance learning (left) aligns augmented region representations, categorical learning (middle) clusters samples by region or land use, and temporal learning (right) aligns representations across time, integrating instance and categorical learning.

features associated with different rotation angles can enhance the quality of the final representation, provided that the encoder training enables it to reconstruct the masked-rotated patch using the optimal transport (OT) loss introduced by the authors. SpectralGPT (Hong et al., 2024) is also another MAE-based method. Its formulation considers a 3D spectral cube made of the MSRSI bands as input. The encoder is then trained using a 3D vision transformer. The training process relies on a lightweight decoder that is used to reconstruct the visible tokens and predict the masked ones. Unlike Masked Auto Encoder (MAE)-based methods, which infer semantic information indirectly, SACo+ prioritizes the construction of a feature space based on the integration of spectral and textural information.

Multiple modalities have also been explored in SSL in recent studies, once the combination of different types of information from the same RSI can increase the quality of the features. For example, Wang et al. (2024) proposed a method, Model Via Multitask Pretraining (MTP), with one encoder and multiple decoders. The encoder receives data with multiple modalities that are processed in a multiscale way. While vision-transformer-based decoders are trained in distinct downstream tasks, each one for a different modality. The downstream tasks considered are semantic and instance segmentation, as well as rotated object detection. To some extent, SACo+ is similar to MTP, as it also explores different "views" of the same data (e.g., semantic band groups and texture) in its training process. However, our method is distinguished by the inclusion of temporal information, in addition to the incorporation of texture information into the encoding process.

Another relevant family of self-supervision methods relies on the use of distillation knowledge. One example is the method based on the DINO (Caron et al., 2021) and its extension for multi-spectral remote sensing images (DINO-MC (Wanyan et al., 2024)). In the DINO-MC formulation, variations in crop size are applied to the images and their corresponding timestamps. This results in the processing of both global and local views of the same region in a contrastive manner, thereby enhancing the encoder training and improving the temporal representation of these regions. The authors claim that the integration of global and local views contributes to the encoding of semantic information. Different from DINO-MC, SACo+ encodes semantic information by exploring well-established knowledge regarding the interpretation of MSRSI band combination, which leads to more effective discriminating features.

The state-of-the-art methods discussed above effectively generate representations from MSRSI data. However, these methods either utilize information solely at the local level (Wanyan et al., 2024) or exclusively at the global level (Li et al., 2024a,b), resulting in a feature space that lacks a training approach capable of simultaneously integrating both types of information. Furthermore, none had explored temporal information in the training process, as we explore in our SACo+ approach.

We follow the principles of MoCo and SeCo training methods in our framework. However, we have also integrated semantic information into our data processing approach, thereby empowering the model to construct a more effective feature space. Another difference refers to the integration of textural properties in the representation learning process.

2.2. Semantic and texture in remote sensing images

While current machine learning models with millions or even billions of parameters have shown remarkable capability to learn various representations, increasing the size and complexity of models is not the only path to improving results in deep learning applications. Developing effective methods to guide the model in learning features has significant potential to enhance downstream tasks. This highlights the importance of innovative approaches and techniques in feature representation learning.

Various authors have explored methods to enhance the quantity or quality of information that can be extracted from MSRSI. For instance, low contrast in some MSRSI makes it difficult to extract information from it. Akiva et al. (2022), for example, used textons to preprocess these images. Diffusion methods applied (Florindo and Abreu, 2023) and cross-modal texture (Yang et al., 2023) were also explored.

The absence of labels poses a challenge in utilizing MSRSI data. Some authors have attempted to mitigate this issue by leveraging semantic information present in the images or related contextual information, such as *where* or *when* the image was captured.

One of the most recent examples of utilizing external information for SSF in RSI is the work by Ayush et al. (2021). In their study, the authors employed contrastive learning to train a method using geolocalization as a pre-text task. The experiments were conducted on the fMoW and GeoImageNet datasets, and the training framework was based on the classical MoCo v2 architecture. However, it included two loss functions: the first compared the embeddings of the original image of the region at different temporal moments, while the second compared the features extracted from the location information (latitude, longitude, and country code) with the original image space.

Xu et al. (2022) tested an approach employing multiple models to generate embeddings. Their study used a contrastive SSL method with only a single input and no augmentation. They employed different networks, such as ResNet and VGG, to compare feature spaces and train the model to learn how maintain proximity between features extracted from optical and SAR images across embedding of various pre-trained deep learning models.

Another example of utilizing semantic information is the work by Xu et al. (2023). Their study used the MoCo architecture to train ColorSelf, an architecture capable of achieving good results in remote sensing tasks, such as classification and segmentation. The main contribution of the model lies in considering that color contains important semantic information about regions and can be used to instruct the model in understanding the patterns present

In addition to semantic information, texture features have also been explored in MRSI-related tasks. Once semantic information can effectively represent and highlight the presence of materials or regions in images, texture information can be employed to enhance representations of region visual properties defined in terms of edges, contours, and local patterns. Some authors have indicated that the incorporation of texture into their methods has led to improvements in quality, as evidenced by studies in algae area estimation (Rahul et al., 2024), pixel-level image classification (Xu et al., 2024), and chlorophyll estimation (Wang et al., 2023).

2.3. Multispectral bands

Semantic information can be derived from the image by examining the bands present in MSRSI. Each band represents a specific wavelength, and by grouping them appropriately, we can highlight relevant properties regarding image content, such as vegetation, infrastructure, and water bodies. In the literature, authors have applied different strategies to group bands that possess semantic coherence.

A classical approach that focuses on spectral bands is highlighted by Bigdeli et al. (2013), in which an ensemble of Support Vector Machine (SVM) is employed for clustering hyperspectral bands. This approach aims to mitigate the redundancy of information during the training process. These clusters are formed considering the correlation of information and were applied to perform land cover classification in the Indian Pines dataset (Baumgardner et al., 2015), achieving state-of-the-art results at the time of publication.

Lu et al. (2022), in turn, grouped the bands of hyperspectral images to train an ensemble of models for hyperspectral image classification. The main contribution of their paper lies in demonstrating that training each model in the ensemble to specialize in certain bands of the image can yield better results than a single model learning all bands in the data. In another study focusing on hyperspectral images, Liu et al. (2021) proposed an approach where attention methods were applied in combination with classical Convolutional Neural Networks (CNNs) to perform super-resolution in the images. The main contribution of their paper was demonstrating the effectiveness of incorporating prior information about the bands of hyperspectral images. By training a network to extract spatial–spectral features from similar spectral signatures, groups are created considering spectral correlation. Furthermore, clustering techniques have been applied to hyperspectral images in the work of Li et al. (2021). In their study, a feature extraction method was trained to cluster sub-spaces of the feature space to represent the images using pseudo-labels produced by a Multi-Layer Perceptron (MLP) at the top of the encoder.

The division of the learning process into groups of bands, rather than processing all channels simultaneously, is a commonly explored approach for hyperspectral data but is absent in MSRSI. Therefore, applications could utilize band information to create clusters capable of training models for better generalization with a more informative feature space. One of the few works that address this topic in MSRSI is the paper by Segarra et al. (2020). Their study investigated band combinations to describe regions. The goal was to tackle region classification tasks based on calculated vegetation indices, leaf area coverage, and water absorption.

These aforementioned works highlighted the potential of extracting significant semantic information by grouping and processing image bands based on their wavelengths. Building on this approach, we grouped the Sentinel 2 bands, considering how combining wavelengths can produce semantically coherent and valuable information for the learned feature space.

3. Semantically-aware contrastive learning

This section introduces Semantically-Aware Contrastive Learning (SACo+) and outlines its main components. It also details its associated training procedure.

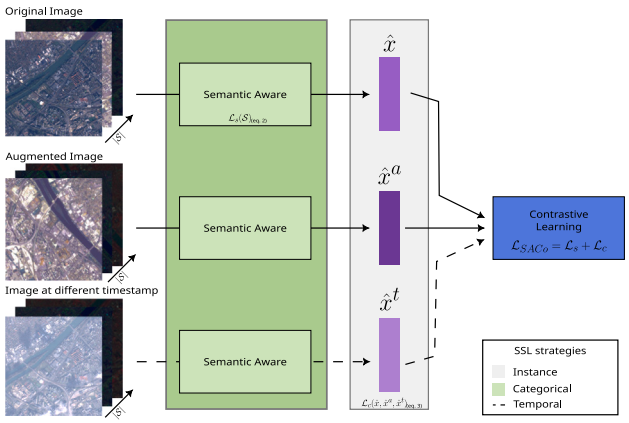


Fig. 3. SACo+ training pipeline. As input, SACo+ receives the original image of a region, an augmented version of the same region, and an image of the same region but at a different time stamp. These images are processed by the Semantic Aware model (Fig. 4) to produce a representation in the feature space for each image. x represents the original image, $\hat{x}^a$ relates to the augmented version, and $\hat{x}^t$ refers to the image at a different time stamp. Later, we employ contrastive learning to bring these representations closer together, ensuring they represent the same region. This process is repeated for multiple images of an input dataset. The SSL strategy involving temporal instances (dashed line) relies on Instance and Categorical learning procedures.

3.1. Overview

Fig. 3 provides an overview of the pipeline. Our approach utilizes self-supervised learning methods to create a feature space capable of representing MSRSI. SACo+ comprises two main components: firstly, a novel methodology that leverages semantic and texture processing of MSRSI bands, and secondly, a training method that integrates all three forms of self-supervised learning training: instance, categorical, and temporal. The categorical approach is explored in the Semantic Aware module where the representation of semantic groups is combined. In contrast, the instance and temporal aspects are presented in the contrastive method, which is applied to the features obtained after the semantic groups have been processed by the Semantic Aware module.

SACo+ aims to process groups of semantically related bands from the Sentinel 2 satellite and project them onto a feature space, forming clusters using the mean of all groups. At the same time, we process the texture representation of these groups. In the first training phase, our goal is to approximate all features related to groups, i.e., by approximating the mean representation of all semantic groups of the MSRSI. In the end, the created feature space encodes both semantic and texture information about the image. In the second training step, our objective is to ensure that the representation generated by processing groups of bands remains as close as possible to the representation generated from the augmented version of the image and the same image at another time point. This combination enables the feature space to remain invariant to seasonal changes (time samples), augmentation changes (random changes in the region), and maintain semantic texture correlation (keeping each semantic group close to others processed in the same image). The goal is to ensure that the semantic features are positioned close to the mean region representation (the mean of all feature representation for that region), minimizing the distance between each semantic group and its corresponding mean. Simultaneously, we aim to maximize the similarity among the clusters of the image its augmented version, and its counterpart at a different time.

3.2. Semantic aware architecture

The Semantic Aware module is pivotal in our architecture as it generates the feature representations for each multispectral image (MSRSI) band group. This process utilizes a ResNet model as an encoder. Each

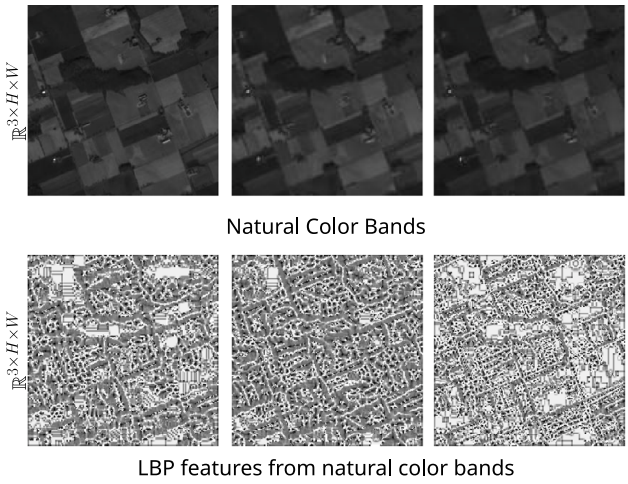


Fig. 4. The Semantic Aware module trains encoders with shared weights based on seven semantic groups (s_g) composed of three-band images and seven three-band texture-based images (t_g). Each input is processed individually by the Semantic Aware model and produces semantic features. After processing all input data, a mean representation is computed. The semantic features and the mean are utilized to compute the semantic loss of the encoder.

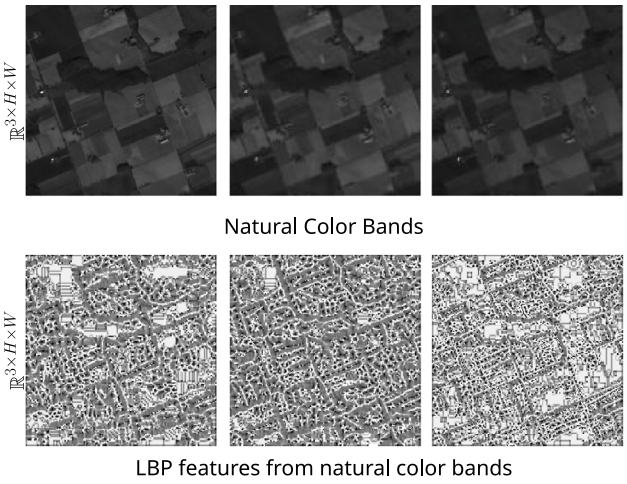


Fig. 5. The top row depicts the original bands used to estimate the LBP features, while the bottom displays matrices of LBP values with the same dimensions as the original image. Highlighted patterns such as edges and contours can be observed in the images.

Table 1
Semantic grouping of Sentinel 2 bands. Each group comprises three bands that can highlight and represent characteristics of the ground, such as vegetation, urban areas, and visible colors.

Groups	Bands		
	R	G	B
Natural Colors	B04	B03	B02
Near-Infrared	B08	B04	B03
Urban	B12	B11	B04
Agriculture	B11	B8A	B02
Atmospheric Pen.	B12	B11	B8A
Complementary 1	B01	B05	B06
Complementary 2	B07	B08	B10

semantic group s_g, such as natural colors, urban, and agricultural areas, and texture group t_g is processed by this encoder. Following the encoding of individual groups, we compute the mean of these encoded outputs to establish the overall output of the Semantic Aware model. This semantic encoding process is depicted in Fig. 4.

3.2.1. Semantic band grouping

Multi-spectral images represent various wavelengths per band, offering valuable insights into different characteristics of regions based on the absorption and reflection properties of these spectra. The image bands were grouped according to established categorizations found in the field literature (Drusch et al., 2012).

In our implementation, we used the 12 bands available on the Sentinel 2 satellite; however, our approach can be easily adapted for any number of bands and groups. We used five groups based on semantic information, as described in Drusch et al. (2012). Additionally, we included two groups to complement the bands that were not represented in the semantic categories. The groups were named as follows: natural colors, near-infrared, urban, agriculture, atmospheric penetration, complementary 1, and complementary 2. Table 1 describes how the Sentinel 2 bands are distributed across each group.

3.2.2. Texture

Texture information has been successfully explored for the effective representation of RSIs. In the SACo+ formulation, while the semantic groups s_g are responsible for providing semantic information to the feature space, we explore texture features to refine their representations.

We employed the Local Binary Pattern (LBP) (Ojala et al., 2002) in each of the seven semantic groups to produce a texture group representation, t_g. This representation also serves as an anchor to estimate the representative of the region. LBP has been successfully employed in several applications, including tasks related to remote sensing image analysis and understanding. Recall that our formulation is generic and can consider other more effective texture descriptors.

Eq. (1) defines the calculation of LBP, where I represents a single-channel image with dimensions $H \times W$. The LBP value at a pixel location (x, y) is determined by the following expression:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(v_p - v_c) \cdot 2^p \tag{1}$$

where P denotes the number of sampling points, R represents the radius of the circle on which the sampling points are positioned, v_c stands for the intensity of the center pixel, v_p denotes the intensity of the pth neighbor of the center pixel, and $s(v_c - v_p)$ is a sign function. This function takes a numerical input and returns 1 if it is greater than or equal to 0, and 0 otherwise. In our study, we set $P = 16$ and $R = 2$.

Generating a texture representation assists the model in learning to construct a feature space where texture serves as one of the guiding principles for estimating the optimal group representation. By processing the extracted LBP for each channel in the same order as the semantic groups, we maintain the semantic integrity of the band information. Fig. 5 illustrates some examples of how the texture features are represented in our model.

3.2.3. Feature space construction

Fig. 6 illustrates the contrastive learning process that is conducted in two stages. In the initial stage, the categorical level is processed (e.g., a single sample is sought, and categorical information is created based on its semantics). "Positive pairs" are created based on s_g and t_g associated with a same region. The goal of the training process is to approximate their representations to the mean representation, designated as $\hat{x}$. At the end of this first stage, three means ($\hat{x}$, $\hat{x}^a$, and $\hat{x}^t$) have been computed, considering three inputs: original images, and their augmented and temporal versions. In the second stage, another contrastive learning procedure is employed to align the representations. For that, we use a memory bank comprising thousands of negative samples (middle points from additional MSRSI images within the dataset).

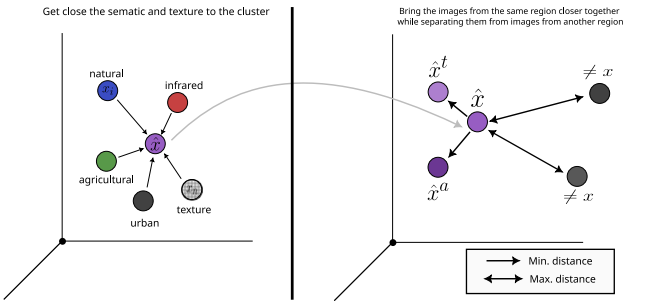


Fig. 6. On the left side, the semantic and texture groups are positioned close to the mean representation of the region (mean representation of all band groups). On the right side, the model is trained to decrease the distance between the region representation $\hat{x}$ and its counterpart at a different time $\hat{x}^t$, as well as the augmented version of the region $\hat{x}^a$, while increasing the distance from all other regions present in the queue of negative samples.

3.3. Loss functions

3.3.1. Semantic loss $\mathcal{L}_s$

The goal of $\mathcal{L}_s$ is to measure how close representations computed from different semantic groups are, i.e., $\mathcal{L}_s$ encodes how compact the cluster formed by the representations extracted from the semantic groups is.

Let $S = \{s_1, s_2, \ldots, s_{|S|}\}$ be a set of semantic groups.

$$\mathcal{L}_s(S) = \frac{\sum_{i=1}^{|S|}(1 - \cos(x_i, \hat{x}))}{|S|} \tag{2}$$

where $\hat{x}$ is the mean of the representations obtained from all semantic groups and x_i is a representation computed from a semantic group s_i using a ResNet-based feature extractor.

3.3.2. Contrastive loss $\mathcal{L}_c$

The goal of $\mathcal{L}_c$ is to measure how close representations computed from samples and their augmented and temporal versions are. $\mathcal{L}_c$ is defined as follows:

$$\mathcal{L}_c = \text{InfoNCE}(\hat{x}, \hat{x}^a) + \text{InfoNCE}(\hat{x}, \hat{x}^t) \tag{3}$$

$\hat{x}^a$ is the mean representation of the augmented version of x and $\hat{x}^t$ is the mean representation of x at a different time stamp.

InfoNCE loss function is a noise contrastive estimator (Oord et al., 2018). This function approximates representations of the same image at different times, as well as augmented versions of the same image. InfoNCE is defined as follows:

$$\text{InfoNCE} = -\log \frac{\exp(x_q \cdot x_{k+}/\tau)}{\sum_{i=0}^{K} \exp(x_q \cdot x_{k-_i}/\tau)} \tag{4}$$

where τ is a temperature hyperparameter (set to 0.05 in our experiments). The sum is over one positive and K negative samples. Intuitively, this loss is the log loss of a $(K+1)$-way softmax-based classifier that tries to classify x_q as x_{k+}.

3.3.3. SACo loss $\mathcal{L}_{SACo}$

The final loss is computed by summing up the Semantic Loss and Contrastive Loss:

$$\mathcal{L}_{SACo} = \mathcal{L}_s + \mathcal{L}_c \tag{5}$$

3.4. Conceptual comparison with other proposals

Fig. 7 illustrates the key differences and innovations of SACo+ compared to the MoCo and SeCo models.

MoCo introduced an SSL training procedure that employs augmented versions of an input MSRSI (He et al., 2020). Examples of augmentation operations are random crop, zoom, rotation, and flips. SeCo, in turn, introduced the use of a training procedure involving instances of a region at different time stamps (e.g., the same region on a different day, week, or month).

In Fig. 7(a), $\mathcal{W}_0$ are the encoder weights while $\mathcal{W}_{\xi_a}$ are those weights after applying moment (a technique that makes a copy of the encoder while applies a small change in the original weights). The second encoder will produce features with a small difference when compared with the $\mathcal{W}_0$. These features are provided to the projection head. Each encoder has its projection head ($\mathcal{P}_0$ and $\mathcal{P}_\xi$). The output of the projection heads ($\mathcal{Z}_a$ and $\mathcal{Z}_t$) are compared using a contrastive method. The SeCo formulation (Manas et al., 2021) introduced the use of temporal data in the learning process, illustrated in Fig. 7(b) by the incorporation of the encoder $\mathcal{W}_{\xi_t}$ and the projection head $\mathcal{P}_{\xi_t}$. Again, contrastive learning is used to align the different representations. As illustrated in Fig. 7(c), SACo+ also uses augmented and temporal instances. However, SACo+ employs an encoder training that explores band combination and texture information (encoders highlighted in green).

Table 2 summarizes the main differences among the MoCo (He et al., 2020), SeCo (Manas et al., 2021), and SACo+ SSL methods regarding the kind of information explored in the contrastive learning procedures employed for each method. In the table, SACo is a variation of SACo+ that does not consider texture information (used in ablation analysis).

4. Experiments

After training our encoders to generate robust representations from the semantic and texture groups, we utilize these representations in three downstream tasks within the remote sensing domain. Additionally, we compare our results with those of current state-of-the-art methods and techniques in self-supervised learning.

4.1. Training datasets for the SSL methods

In the employed training process, we utilized the dataset proposed by CACo (Mall et al., 2023), which comprises 100k multispectral images from the Sentinel 2 satellite, each containing 12 bands of information. Additionally, we fine-tuned the model using the 1M dataset proposed by Manas et al. (2021). Classical random augmentations were applied to the images, including horizontal flipping, cropping, resizing, and jitters. We refrained from using color jitter, as small changes in color can alter the semantic meaning of the bands (Xu et al., 2023).

In our training process, we considered three versions of the same region x to use as positive pairs: the original region x, its augmented version (random crop, flip, etc.) x^a, and another version at a different temporal moment x^t. As negative samples, we follow the methodology employed in CACo and SeCo, utilizing 65,536 samples (represented as $\neq x$ in Fig. 6).

We define our dataset as X to represent all the images. Thus, the set of images can be represented by $\{x_1, x_2, \ldots, x_n\}$, where n is the number of images in X. Given that the dataset comprised images of the same region but at k different times, each sample x_i had k-time variations denoted as $\{x_i^1, x_i^2, \ldots, x_i^k\}$.

4.2. Land cover classification

4.2.1. Experimental protocol

Land cover classification is an essential tool for managing critical tasks such as urbanization and environmental conservation (Avtar et al., 2019). To perform land cover classification, we utilized the EuroSAT dataset (Helber et al., 2019), commonly employed in downstream tasks to validate the quality of feature extractors in multispectral remote sensing images. This dataset comprises 27,000 images from the Sentinel 2 satellite, categorized into 10 distinct classes (e.g., Annual

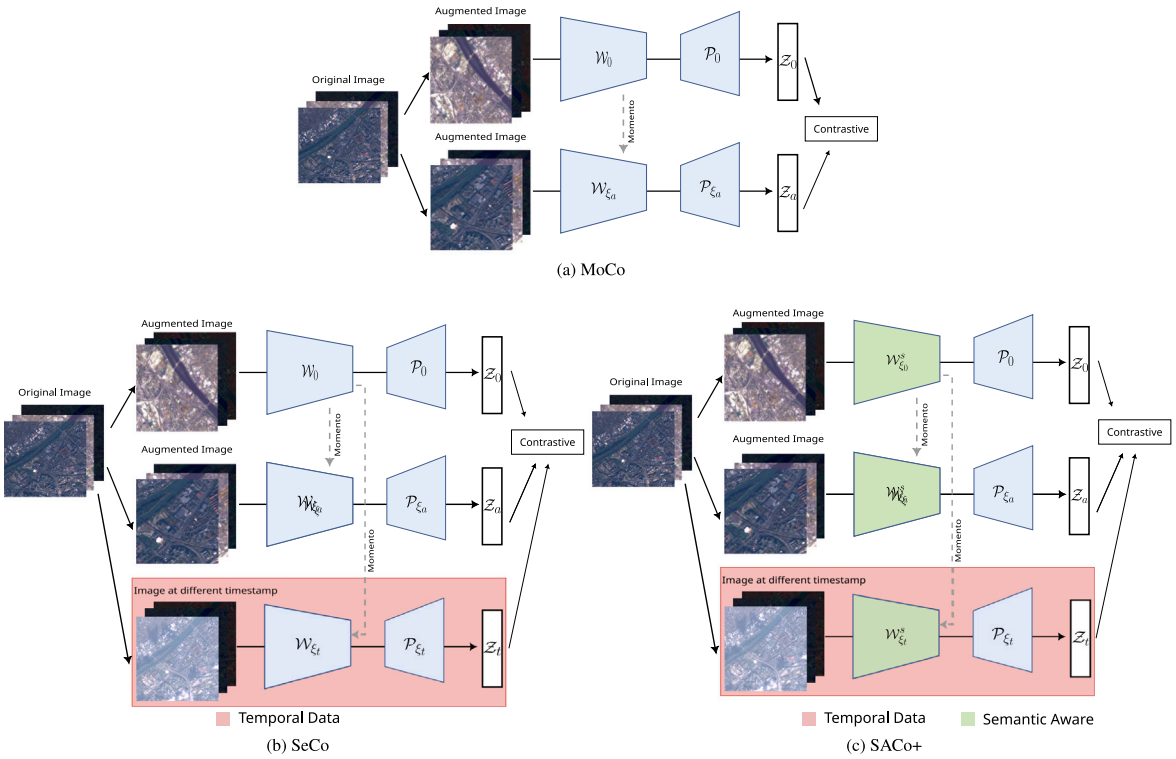


Fig. 7. (a) MoCo, (b) SeCo, and (c) SACo+ SSL methods. All methods rely on a contrastive learning process. SeCo introduces the use of temporal information in the learning process, while SACo+ employs Semantic Aware training involving band combination and texture information.

Table 2
Information explored by different MoCo, SeCo, and SACo SSL methods in their training process. SACo is a variation of SACo+ that does not consider texture information (used in the ablation analysis).

Model	Augmented Instances	Temporal Instances	Band Combination Information	Texture Information
MoCo	✓			
SeCo	✓	✓		
SACo	✓	✓	✓	
SACo+	✓	✓	✓	✓

Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, Sea Lake). The dataset was divided into training, validation, and test sets, following a ratio of 60%, 20%, and 20%, respectively, as proposed in the experimental protocol of SeCo (Manas et al., 2021).

The training process was executed using the Adam optimizer, setting the learning rate at $1e^{-4}$ and applying a weight decay of $1e^{-5}$. We employed a batch size of 64 samples. The training spanned over 100 epochs, with scheduled learning rate reductions by a factor of 0.1 at epochs 60 and 80, following the approach described in SeCo (Manas et al., 2021). This configuration was used to train an MLP to predict the classes.

We considered recently proposed baselines including MoCo V2 (He et al., 2020), SeCo (Manas et al., 2021), and CACo (Mall et al., 2023). Another baseline refers to the use of a retrained ResNet-18 model. Also, we considered two additional baselines whose encoder uses ResNet-50: DINO-MC (Wanyan et al., 2024) and SpectralGPT (Hong et al., 2024). To show the impact of using the Semantic Aware module (Fig. 4), we also consider a formulation of MoCo that includes this module. We refer to this model as SACo (see Table 3). Finally, we computed the results of SACo+, which also accounts for training with temporal instances.

4.2.2. Results

Table 3 presents the achieved results. The use of SACo+ led to the highest results, with an accuracy performance reaching 94.72% with ResNet-18 and 95.77% for the ResNet-50. Also, as can be observed, the use of the Semantic Aware module (SACo) improved the classification accuracy performance of MoCo in our experiment with the ResNet-18 encoder.

Table 3
Classification accuracy of different models in the land cover classification task, considering the EuroSAT dataset.

Land Cover Classification	
Model	Accuracy (%)
ResNet-18 Encoder	
MoCo V2 (He et al., 2020)	83.72
SACo (Ours)	85.79
SeCo (Manas et al., 2021)	90.05
CACo (Mall et al., 2023)	93.08
SACo+ (Ours)	**94.72**
ResNet-50 Encoder	
SeCo (Manas et al., 2021)	93.12
CACo (Mall et al., 2023)	94.48
DINO-MC (Wanyan et al., 2024)	95.70
SpectralGPT (Hong et al., 2024)	99.15
SACo+ (Ours)	**95.77**

The SpectralGPT model shows remarkable effectiveness. As a vision transformer-based model, however, it contains at least 10× more

parameters than our largest ResNet model (ResNet-50), making direct comparisons challenging. Nonetheless, the SACo+ approach has achieved results that are closely comparable to this significantly larger model. This suggests that the semantic-aware training method effectively enhances feature quality, even when implemented using less robust architectures. Future work may consider the investigation of the use of semantic information in combination with the SpectralGPT model.

In addition to the quantitative analysis of accuracy values, we also performed a qualitative analysis of cases of success and failure. Fig. 8, below, presents these results. As an example, the sample of class Highway shown in Fig. 8(a) contains regions with crop fields, being visually similar to samples of the class Permanent Crops (Perma. Crop) (Figs. 8(d) and (e)). Another challenging classification scenario refers to distinguishing samples of Perma. Crop from those belonging to the Pasture class, as illustrated in Figs. 8(d), (e), and (f). The presence of extensive green areas in residential regions also poses challenges (see Figs. 8(g), (h), and (i)). The success in classifying these examples demonstrates the ability of SACo+ feature space to well represent the MSRSI for land cover classification tasks.

We also analyzed some of the cases of failure. We could observe, for example, that misclassifications occur mainly for samples where the real class had huge visual similarity with other classes in the dataset, the examples of misclassification are present in the fourth row, with Figs. 8(j), 8(k), 8(l). Note that Figs. 8(j) and 8(k) share similar soil visual properties (e.g., small patches of vegetation), which makes predictions harder to all methods, including SACo+. Furthermore, Fig. 8(l) contains a Highway sample. In this example, the highway is surrounded by dense vegetation and some crop fields, which leads to a complex prediction scenario for the definition of the correct class.

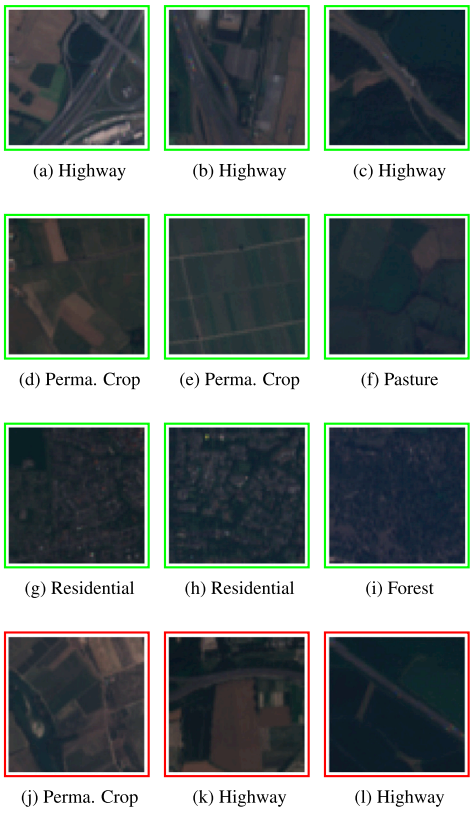


Fig. 8. Examples of challenging classification scenarios for SACo+. Green and red boxes refer to correct and wrong prediction results, respectively. The first row shows images of Highways (a–c). The second row displays images of Permanent Crop (Perma. Crop) (d, e), and Pasture (f). The third row contains images of Residential areas (g, h), and Forest (i). For all those challenging scenarios, SACo+ made correct predictions. The fourth row shows a mix of classification failure cases from the classes Permanent Crop (j) and Highway (k, l). For those samples, all methods failed in assigning the correct label.

4.3. Change detection

4.3.1. Experimental protocol

Additionally, our encoder was assessed in the context of changing detection on the Onera Satellite Change Detection (OSCD) dataset, as referenced by Caye Daudt et al. (2019). This dataset comprises 24 pairs of images from Sentinel 2 between 2015 and 2018, with the data divided into 14 samples for training and 10 for testing. To ensure comparability with other methods, we have adopted the same data partitioning strategy proposed by the authors of this dataset.

To predict the change of a region, we employed the SeCo methodology proposed by Manas et al. (2021), wherein the model input comprises two images of the same region at disparate times, each processed independently by the encoder with the frozen weights. In our case, the ResNet-18 and ResNet-50 models, trained using SACo+, also utilized the intermediate outputs of the convolutional layers stored for each image to facilitate the identification of skip connections.

Subsequently, the absolute difference between the intermediate features and the SACo+ output was calculated, with the dimensions of the vectors maintained. The result of the subtraction was employed as input to the decoder network. The decoder was composed of five convolutional layers, followed by a rectified linear unit (ReLU) activation and an upscale factor to resize the features to the original image dimensions. As the output of the decoder, a segmentation mask was generated, wherein a threshold of 0.5 was applied to classify the pixels as either with or without change.

The training process was conducted over 100 epochs using the Adam optimizer with learning rate $1e^{-3}$ with weight decay of $1e^{-4}$, batch size of 32 and dropout between the CNN layers of decoder with 0.5, where each input image was split in patches with size of 96×96 pixels with non-overlapping, as described in SeCo (Manas et al., 2021).

4.3.2. Results

To assess the quality of the change prediction at the pixel level, we utilized precision, recall, and F1 scores. Results are compared with state-of-the-art methods. Table 4 shows the results obtained on the test set, wherein SACo+ exhibited the highest F1 score (52.78%), surpassing the other SSL methods when the ResNet-18 architecture was used. When the ResNet-50 was used, the F1 score reached 53.11%, achieving a result comparable to models based on transformers. Additionally, training MoCo with the semantic awareness approach (SACo) led to enhanced performance when compared to the original MoCo. This result suggests that the use of semantic information was beneficial for this method as well.

Table 4
Change detection results on the OSCD dataset.

Change Detection			
ResNet-18 Encoder	Precision ↑	Recall ↑	F1 ↑
ImageNet	56.48	13.70	22.05
MoCo V2 (He et al., 2020)	62.21	27.57	38.21
SACo (Ours)	37.63	48.60	42.42
SeCo (Manas et al., 2021)	**64.15**	38.89	46.84
CACo (Mall et al., 2023)	60.68	42.94	50.29
SACo+ (Ours)	53.51	**48.78**	**52.78**
ResNet-50 Encoder	Precision ↑	Recall ↑	F1 ↑
MoCo V2 (He et al., 2020)	44.35	34.70	38.94
SeCo (Manas et al., 2021)	63.21	38.26	47.67
CACo (Mall et al., 2023)	62.87	44.49	52.11
DINO-MC (Wanyan et al., 2024)	51.94	54.04	52.46
SpectralGPT (Hong et al., 2024)	51.65	**56.15**	**53.51**
SACo+ (Ours)	**66.98**	44.00	53.11

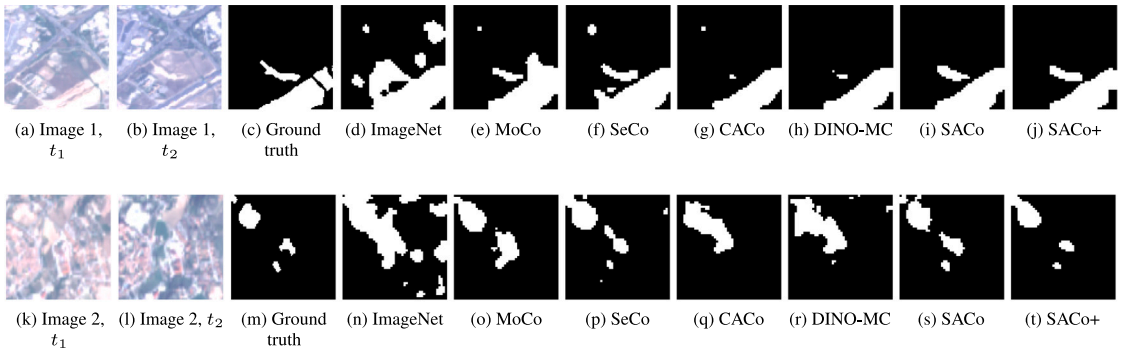


Fig. 9. Change detection results on the OSCD dataset. Changes associated with two images (Images 1 and 2) are considered, across two timestamps (t_1 and t_2). The ground truth for changes in Images 1 and 2 can be found in (c) and (m), respectively. Results in (d–h) refers to those obtained by baselines on Image 1. Results (i, j) refer to the use of our formulations (SACo and SACo+). Similarly, results in (n–r) refer to the use of baselines for Image 2. The results related to the SaCo-based formulations can be found in (s, t).

Overall, the recall rate is high, which is a result of the semantic-aware module. By visualizing the various spectral wavelengths during SSL training, the encoder was able to more accurately represent the images.

Fig. 9 illustrates the qualitative predictions of change in the test images, demonstrating that approaches utilizing semantic awareness in SSL training (SACo and SACo+) lead to prediction results that are very similar to the ground truth masks. The improved performance of these representations stems from incorporating texture information through a semantic-aware training method. As a result, the projected feature space guides the decoder by preserving relevant spatial information.

As shown in Fig. 9, the encoder trained with semantic-aware SSL achieved a better fit with the actual shapes of the changed regions. For example, in the first row, SACo+ (Fig. 9(i)) and SACo (Fig. 9(j)) images do not show the random false positive change detected in the top-left corner, which is present in the other predictions for this region (ImageNet — Fig. 9(d), MoCo — Fig. 9(e), SeCo — Fig. 9(f), CACo — Fig. 9(g)). Also, SACo and SACo+ are the only ones to keep the similar shape of the change area in the middle of the image. On the other hand, in the second row, the methods struggled more to follow the exact shape of the changes. However, once again, the semantic methods (SACo and SACo+) produced more accurate predictions, as seen in SACo (Fig. 9(s)) and SACo+ (Fig. 9(t)), with a clear separation of the actual changes. In contrast, the other methods without semantic training predicted changes in regions without actual alterations (SeCo — Fig. 9(o)) or connected distinct regions of change into a single region (ImageNet — Fig. 9(l), MoCo — Fig. 9(m), CACo — Fig. 9(q) and DINO-MC — Fig. 9(r)).

4.4. Semantic segmentation

Another common method for assessing the quality and utility of SSL encoders is the semantic segmentation task. To correctly predict the class of pixels, it is essential to have a robust feature representation that can effectively distinguish between different land cover types, such as soil, vegetation, and other potential land cover classes at the pixel level. To assess the efficacy of the SACo+ encoder in this context, we used the Panoptic Agricultural Satellite Time Series (PASTIS) (Sainte Fare Garnot and Landrieu, 2021) dataset and the Gaofen Image Dataset (GID) (Tong et al., 2020). The PASTIS dataset comprises 2,433 time series of MSRSI from four distinct regions across France, these regions are divided into 5 sets, where 1, 2, and 3 are used for training, 4 for validation, and 5 for testing. Each of these regions encompasses between 38 and 61 different time samples and 20 potential classes for the pixels. Among these classes, 18 represent crop types, while the remaining 2 represent background and uncertainty areas. The GID dataset comprises 150 high-resolution RSI, split into 100, 10, and 40 images for training, validating, and testing, respectively. These images were captured by Gaofen satellites with 15 different pixel classes.

Table 5
Our method's performance on the semantic segmentation task on the PASTIS dataset is reported in terms of overall accuracy (OA) and mean intersection over union (mIoU) for the ResNet-18 and ResNet-50 backbones.

Semantic Segmentation		
ResNet-18 Encoder	OA ↑	mIoU ↑
ImageNet	43.23	23.42
MoCo V2 (He et al., 2020)	45.23	24.88
SACo (Ours)	45.70	25.01
SeCo (Manas et al., 2021)	49.23	25.30
CACo (Mall et al., 2023)	49.20	26.47
SACo (Ours)	**54.67**	**29.15**
ResNet-50 Encoder	OA ↑	mIoU ↑
MoCo V2 (He et al., 2020)	50.60	24.45
SeCo (Manas et al., 2021)	52.60	25.65
CACo (Mall et al., 2023)	54.04	26.21
DIN0-MC (Wanyan et al., 2024)	54.88	27.47
SACo+ (Ours)	**56.40**	**31.10**

4.4.1. Experimental protocol

The same network implementation used for the change detection task was employed here. The encoder was trained with two different architectures (ResNet-18 and ResNet-50), both trained using the SACo+ approach, while the decoder consists of five CNN layers. The only differences are the addition of skip connections between the decoder layers (as in the original dataset implementation (Sainte Fare Garnot and Landrieu, 2021)) and the number of output channels. Rather than a binary mask, the output consists of 20 channels, each representing a possible class for PASTIS, and 16 channels for GID.

The training was carried out over 100 epochs with a batch size of 32 images, using the Adam optimizer with a learning rate of $1e^{-3}$ with cross entropy as the loss function, the training only updates the weights of the decoder once we have frozen the weights of the encoder. We also used the weights provided by the other methods (MoCo, SeCo, and CACo) to train the different decoders, making it possible to compare the results of SACo+ with these state-of-the-art approaches. In our assessment, we considered the ResNet-18 for all methods but for DINO-MC (Wanyan et al., 2024), which used a ResNet-50 encoder.

4.4.2. Results on PASTIS

The evaluation was conducted using the fifth set, as defined in Sainte Fare Garnot and Landrieu (2021). The overall accuracy (OA) and intersection over the union (mIoU) were estimated for each MSRSI. Table 5 shows the results. As we can observe, SACo+ exhibits the highest values for OA and mIoU in comparison to the other methods. Additionally, the integration of semantic awareness in MoCo (SACo) enhances the quality of features in SSL training, as evidenced by the superior performance of SACo relative to the original MoCo.

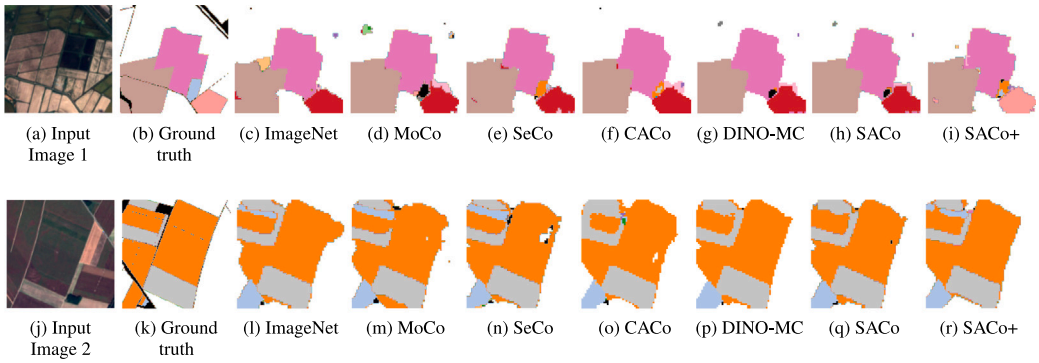


Fig. 10. Visual comparisons of semantic segmentation predictions for two input images from the PASTIS dataset, showcasing results from SACo+ and baseline methods.

When considering the qualitative results, as illustrated in Fig. 10, it is evident that the trained models with semantic aware methods demonstrate superior performance in terms of semantic segmentation, particularly in the separation of regions and predicting the correct class, i.e., the semantic awareness training strategy leads to improved accuracy in aligning with the ground truth of pixel classes.

We can observe that all models perform well in reproducing the areas covered by crops in both examples of Figs. 10(a) and (j). These visual representations also highlight the positive impact of the semantic-aware method within the encoder's SSL training, as misclassifications and artifacts appear in smaller quantities. For instance, in Figs. 10(h) and (i), the latter shows better separation of crop areas (with fewer black regions within classes and fewer "random" points classified in the background). Similarly, between Fig. 10(h) and (i), SACo+ is the only one that correctly predicts the class in the lower-right region while maintaining minimal misclassification in other areas. We observe a similar effect in the second row of Fig. 10, where the semantic models show better classification performance compared to the non-semantic versions, as illustrated in Figs. 10(q) and 10(r), where the upper-left part exhibits improved class separation.

The experiments conducted on the PASTIS dataset demonstrated that combining semantic information with texture features led to more realistic results in pixel classification. However, there were instances where the model struggled to differentiate between regions that were only narrowly separated from another class. Employing more robust techniques, such as convolutional neural networks (CNNs), to extract texture features could potentially address these gaps in feature space construction.

4.4.3. Results on GID

This experiment aims to demonstrate whether the weights learned during the encoder training can yield effective results on a different dataset for the semantic segmentation task. We followed the same evaluation protocol used for the PASTIS dataset, but instead applied it to the ResNet-50 model trained on the Gaofen Image Dataset (GID) (Tong et al., 2020), which features high spatial-resolution remote sensing images. For our evaluation metrics, we considered Overall Accuracy (OA) and mean Intersection over Union (mIoU) on the test set.

Table 6 presents the results for the encoder trained using the ResNet-50 architecture. We compared the performance of SACo and SACo+ with state-of-the-art methods. As observed, the networks trained with the semantic-aware method achieved superior results.

Moreover, we performed a qualitative analysis of our results, using two samples of the GID test set, where the output of the networks are compared side by side with the ground truth. Fig. 11 presents these results. An analysis of the results shows that most models performed satisfactorily in segmentation tasks using the GID dataset. The SACo+ model produced outcomes that were most similar to the ground truth mask in both input instances. Similarly, the model that incorporates semantic features, SACo, also demonstrated a high degree of similarity between the predicted mask and the ground truth. In contrast, the MoCo (Figs. 11(c) and (k)) and DINO-MC (Figs. 11(f) and (n)) encoders demonstrated less favorable results, especially in cases where certain classes were not represented in the outcomes, and the shapes of the regions differed from the ground truth. It is reasonable to hypothesize that these observed results may be attributed to the lack of temporal characteristics in the training processes of these methodologies.

Table 6
Performance on the GID dataset using ResNet-50 as the backbone.

Pre-training	Encoder	OA ↑	mIoU ↑
MoCo V2	ResNet-50	25.39	67.98
SeCo	ResNet-50	25.53	70.20
CACo	ResNet-50	25.50	69.71
DINO-MC	ResNet-50	24.04	54.85
SACo	ResNet-50	25.26	70.06
SACo+ (Ours)	ResNet-50	**25.56**	**70.22**

4.5. Representation quality analysis

4.5.1. Experimental protocol

Downstream tasks are the primary method to assess the quality of an encoder, but they may not fully reveal the separation of features; rather, they demonstrate how the features can be applied in different tasks. In this section, we assess the quality of produced representations using visualization and clustering methods.

We used UMAP (McInnes et al., 2018) to visualize the features generated by our encoder, aiming to observe the clustering of features for two SSL methods that use the proposed SA formulation (SACo and SACo+) and two state-of-the-art methods (CACo and DINO-MC). In this visualization, points with the same colors refer to representations of regions with the same land cover. In our analysis, we selected the EuroSAT test set. This allows us to evaluate how well the model can represent the different types of land cover in an unseen dataset.

In addition, we applied several clustering metrics to assess how well encoders can create data representations. We evaluated the quality of the cluster of points generated for the different classes of the EuroSAT dataset. We employed four different cluster quality metrics: Davies–Bouldin (D.B), Cohesion index, Separation index, and Silhouette score. The Silhouette score (Rousseeuw, 1987) is in the range of −1 to 1, representing the similarity level between samples and their neighbors within a cluster. Higher values indicate that points inside the cluster are closer to each other. The Davis–Bouldin Davies and Bouldin (1979) index, in turn, encodes how well the clusters are separated and with no overlapping. The cohesion index assesses how similar points belonging to the same cluster are, while the separation index measures how well-separated clusters are (Lee et al., 2012).

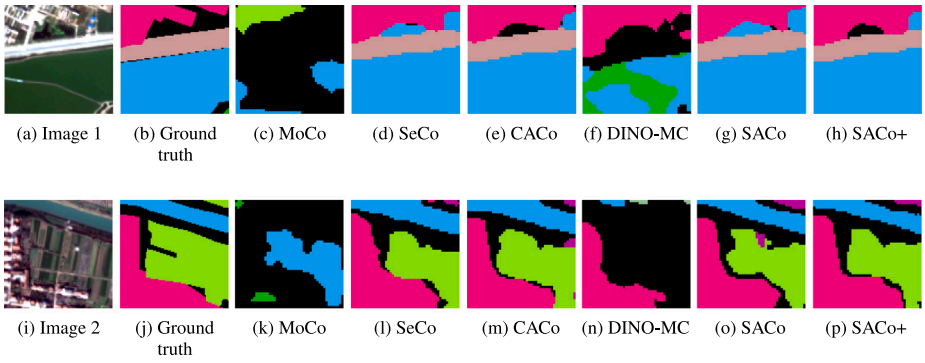


Fig. 11. Visual comparisons of semantic segmentation predictions for two input images from the GID dataset, showcasing results from SACo+ and baseline methods.

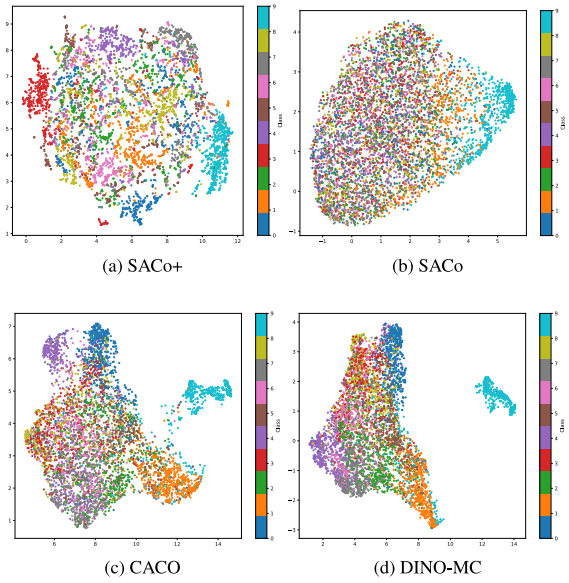


Fig. 12. Projection of the features from different encoders using the ResNet-50 architecture for (a) SACo+, (b) SaCo, (c) CACo, and (d) DINO-MC.

Table 7
Values for the Davies–Bouldin (D.B) index, Cohesion, Separation, and Silhouette for the ResNet-18 trained following the MoCo and SeCo, SA + MoCo, and SACo.

	D.B. ↓	Cohesion ↑	Separation ↑	Silhouette ↑
MoCo	0.96	0.21	0.08	−0.010
SeCo	0.80	0.46	0.19	0.008
SACo	0.84	0.48	0.17	0.011
SACo+	**0.73**	**0.57**	**0.22**	**0.234**

4.5.2. Results

Fig. 12 displays the UMAP visualization of the features obtained from training using the semantic bands processing based on SACo, SACo+, CACo, and DINO-MC approaches. We can observe that the use of the Semantic Aware module led to compact and separated clusters, especially for the SACo+ formulation. The results of SACo demonstrate the relevance of training using texture, once the separation of the cluster presents the worst results when compared with SACo+.

Table 7 presents the cluster metric results for different SSL methods. For all metrics, the use of SACo+ led to compact and separated clusters when compared to SACo, MoCo, and SeCo. It is also worth noting the performance of SACo when compared to MoCo. Those results suggest that the SA module, i.e., the use of semantic information (band combination) is effective in improving the performance of MoCo.

Table 8
Performance comparison of different SACo encoders across the downstream tasks.

Encoder	EuroSAT	PASTIS		OSCD
	Acc ↑	OA ↑	mIoU ↑	F1 ↑
SACo (Semantic)	94.09	51.22	25.88	43.36
SACo (Texture)	94.59	51.04	24.53	46.71
SACo+ (Augmented Images)	95.01	52.25	27.21	49.58
SACo+ (Temporal Images)	94.61	53.85	27.59	47.96
SACo+	95.77	56.40	31.10	53.11

4.6. Ablation

4.6.1. Relevance of semantic aware components

In this section, we investigate the impact of using semantic information and texture in the encoder. Table 8 shows the results related to the assessment of the impact of the use of semantic information, texture, temporal, and augmented data in the downstream tasks. SACo (Semantic) stands for the SACo formulation that accounts for only the band combination. SACo (Texture) — only SACo in Table 2, in turn, employs only texture information. SACo+ (Augmented Images) employs semantic and texture information but is trained using only augmented images. SACo+ (Temporal Images) utilizes semantic and texture information but is trained using only temporal instances. Finally, SACo+ stands for the complete method. We can observe that the use of semantic information and texture in isolation can produce good results when the encoder is trained using the semantic-aware method. However, the table shows that the combination of semantic and texture using not only augmented or temporal images can produce a better encoder to be applied in downstream tasks, once the best results are presented with the SACo+.

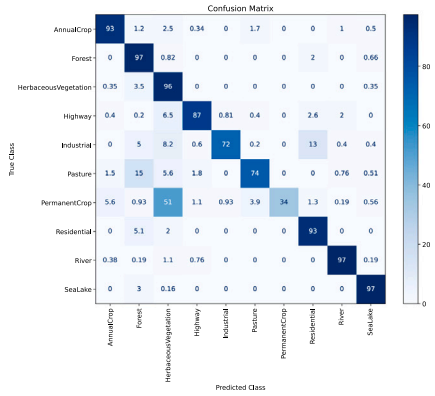
4.6.2. Relevance of texture features

In this section, we conducted an evaluation of the design choices in the SACo+ framework. For instance, we compared the impact of the use of textural information in the land cover classification task (EuroSAT dataset) and examined the most efficient method for evaluating the quality of semantic groups.

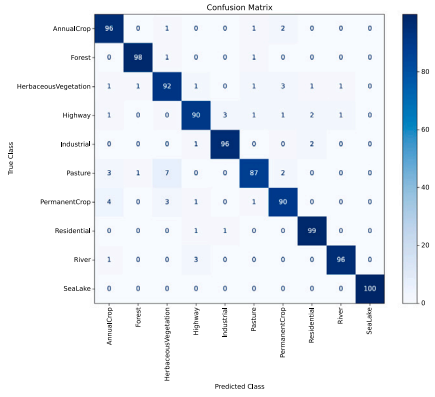
Fig. 13(a) shows the confusion matrix for the scenario where only band combination is employed, while Fig. 13(b) also considers textural information. As we can observe, the inclusion of texture features led to significant gains. It is worth mentioning how relevant those features were to distinguish, for instance, samples belonging to the class Permanent Crop from those of Herbaceous Vegetation.

4.7. Computational efficiency

In this section, we compare the time for training each epoch using our Semantic Aware module. Time is estimated for training using only the semantic groups from MSRSI, texture, and both combined.



(a) No Texture



(b) With Texture

Fig. 13. (a) The confusion matrix related to the land cover classification task, when only semantic band combinations are used (SACo). (b) The confusion matrix when not only semantic band combinations but also texture properties are explored by SACo+.

A comparison of the SACo+ model with other state-of-the-art networks reveals no significant difference in the number of parameters within the network. This is due to the fact that all networks utilize a ResNet-18 as their encoder, and the inference time and complexity are also similar, as the inputs have the same shape for all tasks.

However, the incorporation of semantic and texture features in successive training rounds proved indispensable for guiding the feature space. This resulted in a significant increase in the number of iterations required at each training epoch to process each ground individually, thereby rendering this process somewhat more time-consuming. Nevertheless, the training time is essentially comparable to that of the other methods. To illustrate this phenomenon, we trained all the models on 10% of the CACo 100k dataset for 100 epochs and estimated the training time. All experiments were conducted on the same hardware, a Windows 11 computer with the following specifications: CPU (AMD Ryzen-5600g(12) @ 4.00 GHz, 16 GB RAM) and GPU (NVIDIA GeForce GTX 1080 Ti, and 11000 MB GDDR5).

The results in Table 9 demonstrates that increasing the number of semantic groups does not lead to a significant impact on training time. Recall that the use of semantic information also does not impact the number of parameters of the encoder.

5. Discussion

Examining the results for land cover classification (Table 3), change detection (Table 4), and semantic segmentation (Tables 5 and 6), it becomes evident that leveraging semantic awareness to construct the feature space through the training protocols of current state-of-the-art self-supervised learning methods for remote sensing images yields effective representations. By creating distributions of the samples in the feature space based on anchors of each semantic group of bands, our approach teaches the model how to represent these groups in the same space. This modeling of the space enhances its robustness to variations and augments its generalization capabilities. These results support responding positively RQ1: *Would the use of semantic groups of bands lead to relevant feature spaces for MSRSI?*.

Another important point was to investigate how the semantic features can be complemented by additional information available within MSRSI to increase the quality of encoded features (Table 8). Thus, the use of texture from the MSRSI when applied together with the semantic features increased the performance of the encoder in clustering the samples and in the downstream tasks. In order to respond to RQ2: *Would the combination of semantic features with texture from MSRSI increase the quality of features extracted?*

Furthermore, the SACo+ had demonstrated superior or comparable results in all downstream tasks, whereas the quantitative and qualitative results presented superior results when compared with the other contrastive self-supervised learning methods for multi-spectral remote sensing, responding positively RQ3: *Would the SACo+ encoder be effective when applied in downstream tasks, leading to superior results compared to other state-of-the-art self-supervised methods? Furthermore, the ablation shows the positive impact of the semantic and texture features combined to train the encoder while the computational time does not show a large increase in the time to process.*

Table 9
A comparative analysis of the training time for the model using different SSL approaches reveals that SACo+ had the highest time per epoch. However, when compared to the other methods, the difference is only a few minutes, which can be considered a favorable trade-off given the improvement in feature quality for the tasks.

Model	Batch size	Time per epoch
MoCo-V2	32	2.71 s
SeCo	32	4.40 s
SACo (Texture)	32	3.68 s
SACo (Semantic)	32	4.85 s
SACo+ (Texture)	32	5.13 s
SACo+ (Semantic)	32	5.58 s
SACo+	32	6.20 s

6. Conclusions

This paper introduced Semantically-Aware Contrastive Learning (SACo+), a novel self-supervised learning (SSL) methodology for multispectral remote sensing imagery (MSRSI). SACo+ combines semantic information encoded in the MSRSI bands with texture features, utilizing training strategies based on augmented, categorical, and temporal instances. Experiments conducted with multiple datasets and three downstream tasks demonstrate that incorporating semantic and texture information is essential for creating effective encoders. Additionally, SACo+ has been shown to produce results that are either better than or comparable to several state-of-the-art methods.

As directions for future work, we aim to enhance the semantic information explored in our training procedure by including MSRSI indices, such as NDVI, EVI, and SAVI, as band groups to improve the anchor points for guiding the encoder training, which may lead to a better cluster as output. That may contribute to tackling the limitations identified (Sections 4.2, 4.3, and 4.4). Another possible venue is to apply the concept of semantic information to other domains in which SSL has been explored, such as medical image analysis, to test the generalization of the SSL methodology proposed in this study.

CRediT authorship contribution statement

Leandro Stival: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Ricardo da Silva Torres:** Writing – review & editing, Validation, Supervision, Conceptualization. **Helio Pedrini:** Writing – review & editing, Validation, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank CAPES, CNPq (grant #304836/2022-2), and FAPESP (grants #2022/12294-8 and #2023/11556-1) for their financial support. This work was also partially funded by the NorDark project, supported by NordForsk (grant #105116).

References

Akiva, P., Purri, M., Leotta, M., 2022. Self-supervised material and texture representation learning for remote sensing tasks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8203–8215.

Arefin, R., Mohir, M.M.I., Alam, J., 2020. Watershed prioritization for soil and water conservation aspect using GIS and remote sensing: PCA-based approach at northern elevated tract Bangladesh. Appl. Water Sci. 10, 1–19.

Avtar, R., Tripathi, S., Aggarwal, A.K., Kumar, P., 2019. Population–urbanization–energy nexus: A review. Resources 8 (3), 1–21.

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S., 2021. Geography-aware self-supervised learning. In: IEEE/CVF International Conference on Computer Vision. pp. 10181–10190.

Baumgardner, M.F., Biehl, L.L., Landgrebe, D.A., 2015. 220 band aviris hyperspectral image data set: June 12, 1992 Indian pine test site 3. Purdue Univ. Res. Repos. 10 (7), 991.

Bigdeli, B., Samadzadegan, F., Reinartz, P., 2013. Band grouping versus band clustering in SVM ensemble classification of hyperspectral imagery. Photogramm. Eng. Remote Sens. 79, 523–533.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. 33, 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers. In: IEEE/CVF International Conference on Computer Vision. pp. 9650–9660.

Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y., 2019. OSCD - Onera satellite change detection. ArXiv.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Chen, H., Li, W., Chen, S., Shi, Z., 2022. Semantic-aware dense representation learning for remote sensing image change detection. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Chuang, C.-Y., Hjelm, R.D., Wang, X., Vineet, V., Joshi, N., Torralba, A., Jegelka, S., Song, Y., 2022. Robust contrastive learning against noisy views. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16670–16681.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 2, 224–227.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36.

Fassnacht, F.E., White, J.C., Wulder, M.A., Næsset, E., 2024. Remote sensing in forestry: Current challenges. Considerations Dir. Forestry: An Int. J. For. Res. 97 (1), 11–37.

Florindo, J.B., Abreu, E., 2023. A pseudo-parabolic diffusion model to enhance deep neural texture features. Multimedia Tools Appl. 83 (4), 1–22.

GISGeography, 2024. Sentinel 2 bands and combinations. (Accessed 13 June 2024).

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

Helber, P., Bischke, B., Dengel, A., Borth, D., 2019. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 12 (7), 2217–2226.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., Plaza, A., Gamba, P., Benediktsson, J.A., Chanussot, J., 2024. SpectralGPT: Spectral remote sensing foundation model. IEEE Trans. Pattern Anal. Mach. Intell. 46 (8), 5227–5244.

Huang, G., Laradji, I., Vazquez, D., Lacoste-Julien, S., Rodriguez, P., 2022. A survey of self-supervised and few-shot object detection. IEEE Trans. Pattern Anal. Mach. Intell. 45 (4), 4071–4089.

Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F., 2020. A survey on contrastive self-supervised learning. Technologies 9 (1), 1–22.

Jung, H., Oh, Y., Jeong, S., Lee, C., Jeon, T., 2021. Contrastive self-supervised learning with smoothed representation for remote sensing. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D., 2020. Hard negative mixing for contrastive learning. Adv. Neural Inf. Process. Syst. 33, 21798–21809.

Krishnamurthy R, P.K., Fisher, J.B., Schimel, D.S., Kareiva, P.M., 2020. Applying tipping point theory to remote sensing science to improve early warning drought signals for food security. Earth' s Futur. 8 (3), e2019EF001456.

Lee, K.M., Lee, K.M., Lee, C.H., 2012. Statistical cluster validity indexes to consider cohesion and separation. In: International Conference on Fuzzy Theory and its Applications. pp. 228–232.

Li, X., Hong, D., Chanussot, J., 2024a. S2MAE: A spatial-spectral pretraining foundation model for spectral remote sensing data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24088–24097.

Li, Z., Hou, B., Ma, S., Wu, Z., Guo, X., Ren, B., Jiao, L., 2024b. Masked angle-aware autoencoder for remote sensing images. In: 18th European Conference on Computer Vision. Springer-Verlag, Milan, Italy, pp. 260–278.

Li, K., Qin, Y., Ling, Q., Wang, Y., Lin, Z., An, W., 2021. Self-supervised deep subspace clustering for hyperspectral images with adaptive self-expressive coefficient matrix initialization. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 3215–3227.

Liu, D., Li, J., Yuan, Q., 2021. A spectral grouping and attention-driven residual dense network for hyperspectral image super-resolution. IEEE Trans. Geosci. Remote Sens. 59 (9), 7711–7725.

Lu, H., Su, H., Hu, J., Du, Q., 2022. Dynamic ensemble learning with multi-view kernel collaborative subspace clustering for hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 15, 2681–2695.

Mall, U., Hariharan, B., Bala, K., 2022. Change event dataset for discovery from spatio-temporal remote sensing imagery. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), In: Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., pp. 27484–27496.

Mall, U., Hariharan, B., Bala, K., 2023. Change-aware sampling and contrastive learning for satellite images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5261–5270.

Manas, O., Lacoste, A., Nieto, X.Giró-i., Vazquez, D., Rodriguez, P., 2021. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: IEEE/CVF International Conference on Computer Vision. pp. 9414–9423.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. pp. 1–63, arXiv preprint arXiv:1802.03426.

Misra, I., Maaten, L., 2020. Self-supervised learning of pretext-invariant representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717.

Newell, A., Deng, J., 2020. How useful is self-supervised pretraining for visual tasks? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7345–7354.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 971–987.

Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V., Murayama, Y., Ranagalage, M., 2020. Sentinel-2 data for land cover/use mapping: A review. Remote. Sens. 12, 1–10.

Rahul, A., Lokesh, G., Goswami, S., Ponnalagu, R., Sudha, R., 2024. Automatic area estimation of algal blooms in water bodies from UAV images using texture analysis. Water Sci. Eng. 17 (1), 62–71.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Sainte Fare Garnot, V., Landrieu, L., 2021. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: Internation Conference on Computer Vision. pp. 4872–4881.

Segarra, J., Buchaillot, M.L., Araus, J.L., Kefauver, S.C., 2020. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. Agronomy 10 (5), 641.

Shafique, A., Cao, G., Khan, Z., Asad, M., Aslam, M., 2022. Deep learning-based change detection in remote sensing images: A review. Remote. Sens. 14 (4), 871.

Tao, C., Qi, J., Guo, M., Zhu, Q., Li, H., 2023. Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. IEEE Trans. Geosci. Remote Sens. 61, 1–26.

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322.

Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., Zhu, X.X., 2022. Self-supervised learning in remote sensing: A review. IEEE Geosci. Remote. Sens. Mag. 10 (4), 213–247.

Wang, Y., Gu, M., 2024. Classification methods for hyperspectral remote sensing images with weak texture features. J. Radiat. Res. Appl. Sci. 17 (3), 1–12.
Wang, Y., Tan, S., Jia, X., Qi, L., Liu, S., Lu, H., Wang, C., Liu, W., Zhao, X., He, L., 2023. Estimating relative chlorophyll content in rice leaves using unmanned aerial vehicle multi-spectral images and spectral–textural analysis. Agronomy 13 (6), 1541.
Wang, D., Zhang, J., Xu, M., Liu, L., Wang, D., Gao, E., Han, C., Guo, H., Du, B., Tao, D., Zhang, L., 2024. MTP: Advancing remote sensing foundation model via multitask pretraining. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 11632–11654.
Wanyan, X., Seneviratne, S., Shen, S., Kirley, M., 2024. Extending global-local view alignment for self-supervised learning with remote sensing imagery. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2443–2453.
Xu, Y., Guo, W., Zhang, Z., Yu, W., 2022. Multiple embeddings contrastive pretraining for remote sensing image classification. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.
Xu, Z., Jiang, W., Geng, J., 2024. Texture-aware causal feature extraction network for multimodal remote sensing data classification. IEEE Trans. Geosci. Remote Sens. 62, 1–12.
Xu, G., Jiang, X., Liu, X., 2023. Color-aware self-supervised learning for scene classification and segmentation of remote sensing images. In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 5049–5052.
Xue, Z., Yang, G., Yu, X., Yu, A., Guo, Y., Liu, B., Zhou, J., 2025. Multimodal self-supervised learning for remote sensing data land cover classification. Pattern Recognit. 157, 1–12.
Yang, R., Zhang, D., Guo, Y., Wang, S., 2023. A texture and saliency enhanced image learning method for cross-modal remote sensing image-text retrieval. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 4895–4898.
Yin, J., Dong, J., Hamm, N.A., Li, Z., Wang, J., Xing, H., Fu, P., 2021. Integrating remote sensing and geospatial big data for urban land use mapping: A review. Int. J. Appl. Earth Obs. Geoinf. 103, 102514.
Ziegler, A., Asano, Y.M., 2022. Self-supervised learning of object parts for semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14502–14511.