A Multimodal and Multitemporal Deep Learning Semantic Segmentation Method based on Variational Autoencoder for Multimodal Remote Sensing Image Time Series

Luca Bergamasco[1], Marius Guichard[2], Mauro Dalla Mura[2], Francesca Bovolo[1]

[1]Fondazione Bruno Kessler, Trento, Italy - (lbergamasco, bovolo)@fbk.eu
[2]GIPSA-lab, Institut polytechnique de Grenoble, Grenoble, France -
marius.guichard@grenoble-inp.org, mauro.dalla-mura@gipsa-lab.grenoble-inp.fr

Keywords: Variational Autoencoder, Semantic Segmentation, Multimodal Analysis, Image time series, Remote Sensing.

Abstract

Multimodal Remote Sensing (RS) methodologies have been increasingly studied in recent years due to their capacity to analyze multimodal RS data acquired from different sensors, thereby providing improved temporal resolution and extracting richer information than single-modal RS data. Deep Learning (DL) methodologies have accelerated the study of multimodal RS methods, thanks to their ability to learn features during training automatically. Many multimodal DL methods exploit this capability to learn a shared domain across modalities. However, most of them struggle to align heterogeneous modalities in a common representation. For this reason, we propose a supervised multimodal DL method that analyzes image time series acquired by different sensors to perform semantic segmentation. The proposed DL method is based on a Variational Autoencoder (VAE) that models the spatio-temporal information of the multimodal input image time series, with encoders and decoders composed of 3D convolutional layers, and learns the probability distributions for each modality. The probability distributions are combined to derive a joint distribution used for semantic segmentation. Learning the joint probabilistic distribution is achieved by combining the probabilistic parameters across modalities using a Product of Experts (PoE) approach. The feature maps derived from the obtained latent space are processed through three decoders. Two decoders aim to reconstruct the input multimodal image time series. The third decoder performs a semantic segmentation based on the inputs. Experiments conducted on the MultiSenGE and Austria datasets, which comprise Sentinel-1 and Sentinel-2 image time series acquired in France and Austria and representing heterogeneous classes, yielded promising results.

1. Introduction

The increasing number and types of satellite sensors over the last decade have led to the development of a growing number of multimodal Remote Sensing (RS) methods. This methodology enhances the performance of RS tasks, such as classification, semantic segmentation, and change detection, by leveraging its capacity to extract and utilize information from RS data acquired by heterogeneous sensors. The multimodal RS methods also enable increasing the temporal resolution of the input data, since the use of different sensors allows for more frequent monitoring of a given geographical area. Thus, it is possible to model an object or a phenomenon with finer temporal information. Deep Learning (DL) methods have improved multimodal RS analysis by enabling automatic learning and feature extraction across diverse multimodal domains.

Multimodal DL RS methods can aggregate the various modalities using three approaches: early, intermediate, and late fusion (Zhao et al., 2024). Early-fusion approaches combine the different modalities at the input level. However, they are sensitive to differences in spatial resolution and cannot mitigate the domain gap between the modalities. Late-fusion approaches combine outputs from multiple modalities at the decision level. These models analyze unimodal data and may therefore yield suboptimal results due to the limitations of single modalities. Intermediate fusion approaches are the most widely used since they combine multimodal information at the feature level. However, the combination of multimodal features remains an open issue. Most multimodal DL methods exploit two-branch architectures to extract features for each modality and then combine them throughout the model (Hu et al., 2017, Benedetti et al., 2018, Zhao et al., 2025). In (Li et al., 2023), the authors use a sparse constraint to reduce the feature redundancy and improve the model generalization. Generative Adversarial Networks (GANs) are used to generate optical-like images from Synthetic Aperture Radar (SAR) images, or vice versa, and to analyze the images within the same data domain (Saha et al., 2021, Hong et al., 2021, Zhu et al., 2025). Cross-attention mechanisms allow DL models to fuse multimodal features by learning the multimodal interdependencies and global spatial context (Wang et al., 2022). In (Luppino et al., 2024), the authors propose a code-aligned autoencoder that aims to learn a common domain between the multimodal data. In recent years, Transformer-based methods have been developed to exploit the capacity of Transformer layers to analyze global context and improve data fusion across heterogeneous modalities (Roy et al., 2023, Cai et al., 2023, Chang et al., 2024). In (Ma et al., 2024), the authors combine the capacity of convolutional layers to extract local spatial information and that of transformer layers to extract global information to fuse multimodal features at different scales. State Space Models (SSMs), such as Mamba (Gu and Dao, 2024), have also been used to process multimodal RS images and reduce the computational load of the model compared to Transformer-based ones (He et al., 2025, Cui and Zhang, 2026). Recently, many multimodal RS methods fine-tune foundation models to improve the analysis of multimodal data, thanks to their generalization capacity, which enables them to fuse data across multiple modalities and learn models that can be applied to multiple downstream tasks (Yan et al., 2023b, Zhou et al., 2025). However, these methodologies are computationally intensive and challenging to fine-tune since they require a large amount of data. The previous methodologies do not handle the temporal relationship between bitemporal images, which is critical for change detection tasks, nor do they model temporal information in image time series, which is relevant

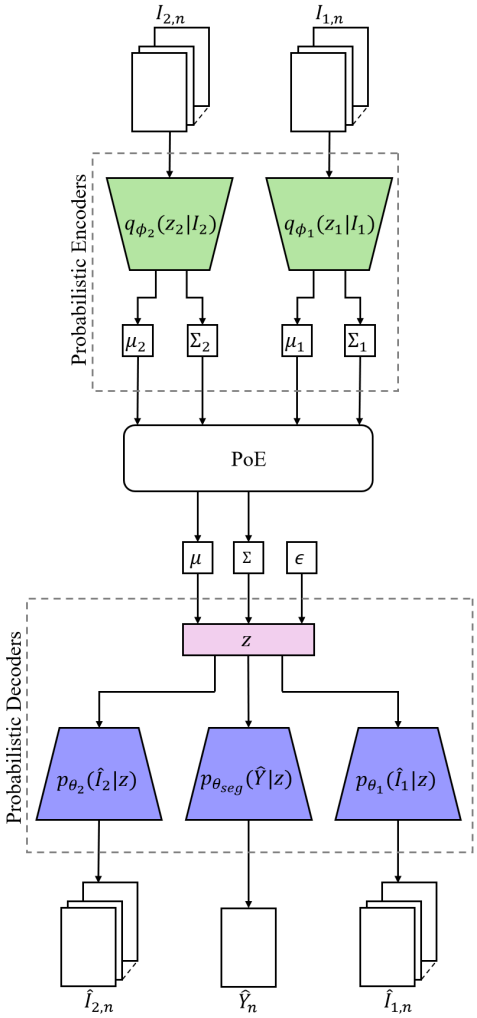


Figure 1. The flowchart of the proposed method. $q_{\phi_1}(z_1|I_1)$ and $q_{\phi_2}(z_2|I_2)$ are the two encoders learning the probability distribution parameters of the two input image time series belonging to the two modalities to be combined with a Product of Experts (PoE) strategy for learning a shared latent space z. $p_{\theta_1}(\hat{I}_1|z)$ and $p_{\theta_2}(\hat{I}_2|z)$ are the decoders that reconstruct the multimodal inputs and enable the learning of the multimodal probability distributions. $p_{\theta_{seg}}(\hat{Y}|z)$ is the decoder for semantic segmentation.

for distinguishing classes with spectral behavior that varies over time.

The temporal information derived from image time series can be exploited by a multimodal DL RS method to guide the translation from one domain to another using an architecture composed of two input branches, one for each modality, in which the spatiotemporal information extracted by one branch is influenced by the one extracted from the other one (Liu et al., 2022). Many State-of-the-Art (SoA) multimodal DL methods analyze the spatiotemporal and spatial context information separately for each modality using Convolutional Long-Short-Term-Memory (ConvLSTM) and convolutional-based models, respectively, and combine them with fully connected layers (Ienco et al., 2019) or U-Net-based models to perform the downstream task (Wenger et al., 2023). These methods give the same relevance to all the multimodal features. However, some are more relevant to a given task than others. Attention mechanisms improve the combination of multimodal spatiotemporal features by emphasizing the most relevant ones for a given RS task (Garnot and Landrieu, 2021, Garnot et al., 2022). Transformer-based methods have been used to extract long-range, time-dependent, and high-level semantic information from multimodal data (Yan et al., 2023a, Follath et al., 2024). Foundation models, such as SkySense (Zhang et al., 2025), can process multimodal RS image time series to extract spatiotemporal features for my downstream tasks. However, they comprise billions of parameters and require billions of images for fine-tuning, making them challenging to adapt to additional modalities. The approaches presented so far do not learn the probabilistic distributions of input modalities, but rather shared domains that are limited by the number of training samples. Thus, when input multimodal images do not fall within the learned discrete domain, the performance of the multimodal DL model drops.

For these reasons, we present a supervised multimodal DL method that analyzes the spatiotemporal information of image time series acquired from multispectral and SAR sensors to perform a semantic segmentation. The proposed method is based on a Variational Autoencoder (VAE) that models spatiotemporal information using an encoder/decoder structure composed of 3D convolutional layers. VAEs can learn the probabilistic distribution of the input data to generate a brand new output belonging to the same domain. We exploit this characteristic to learn the probabilistic distributions of the different modalities and to derive a probabilistic distribution over a common domain shared by them. The combination of multimodal probabilistic distributions into a joint one is performed using a Product of Experts (PoE) approach. The experiments conducted on the MultiSenGE dataset, a labeled semantic segmentation dataset composed of Sentinel-1 and Sentinel-2 image time series acquired in France, yielded promising results.

The paper has the following outline. Section 2 describes the proposed methodology. In Section 3, we present the dataset, experimental settings, and results. Finally, we draw our conclusion in Section 4.

2. Multimodal and Multitemporal Semantic Segmentation Using a Variational Autoencoder

The paper proposes a multimodal DL semantic segmentation method that uses a VAE to learn a joint domain across the input multimodal image time series (Figure 1). We use the VAE since it does not encode a single deterministic latent vector, like other autoencoders; instead, it learns a continuous latent space defined by the probabilistic distribution parameters. In this way, the model learns the probabilistic distributions of each modality and enables us to learn a new distribution common to all modalities. Let us assume having two image time series $I_1 = \{i_{t_1,1}, t_1 = 1, ..., T_1\}$ and $I_2 = \{i_{t_2,2}, t_2 = 1, ..., T_2\}$ composed of T_1 and T_2 images $i_{t_1,1}$ and $i_{t_2,2}$ having different modalities. $i_{t_1,1}$ has size of $h \times w \times s_1$ and $i_{t_2,2}$ has size of $h \times w \times s_2$, where h is the height, w is the width, and s_1 and s_2 are the number of spectral bands. The images for I_1 and I_2 have identical spatial resolutions, are spatially coregistered, and do not need to be temporally aligned. We also assume that no significant changes occur among the acquisitions of the images in the time series. We process I_1 and I_2 with a VAE to obtain a semantic segmentation map $\hat{Y}$ with size $h \times w \times C$, where C is the number of classes. To do so, we train the VAE using a labeled dataset $X = \{(I_{1,n}, I_{2,n}), Y_n\}$, where $n = 1, ..., N$, composed of N couples of multimodal image time series $I_{1,n}$ and $I_{2,n}$ and associated to a reference semantic segmentation

map Y_n with size $h \times w \times C$ presenting C classes. The VAE processes $I_{1,n}$ and $I_{2,n}$ using two encoders, one for each modality, and extracts statistical parameters for both modalities. Assuming that the modalities are independent, the multimodal statistical parameters are combined using a PoE approach. We sample the latent space derived from the PoE to generate a feature map in the joint domain, which is then processed by three decoders. Two decoders aim to reconstruct $I_{1,n}$ and $I_{2,n}$, whereas the third one processes $I_{1,n}$ and $I_{2,n}$ to obtain a semantic segmentation map $\hat{Y}_n$. Encoders and decoders model the spatiotemporal information of $I_{1,n}$ and $I_{2,n}$ using 3D convolutional layers.

2.1 Probabilistic Encoders

The probabilistic encoders aim to model the spatiotemporal information in the input multimodal image time series $I_{1,n}$ and $I_{2,n}$ and to extract the probabilistic distribution parameters for each modality. The encoders model the spatiotemporal information of the multimodal image time series using a series of L_q 3D convolutional layers with a fixed kernel size of $k \times k \times k$ followed by 3D batch normalization and LeakyRelu layers. We use 3D convolutional layers since they can analyze 3-dimensional data. In the proposed method, we use them to jointly model the spatial context and temporal information of image time series, and we exploit the kernel size value for both modalities since they have similar spatial and temporal resolutions. The encoders can be defined as $q_{\phi_1}(z_1|I_1)$ and $q_{\phi_2}(z_2|I_2)$, where q_{ϕ_1} and q_{ϕ_2} are the approximate posterior probabilities learned by the encoders and are used to indicate the encoders, ϕ_1 and ϕ_2 are the parameters of the two encoders. $z_1 \in \mathbb{R}^{h_{L_q} \times w_{L_q} \times T_{1,L_q} \times f_{L_q}}$ and $z_2 \in \mathbb{R}^{h_{L_q} \times w_{L_q} \times T_{2,L_q} \times f_{L_q}}$ are the sampled latent variables corresponding to modality domains, h_{L_q}, w_{L_q}, and f_{L_q} are the height, width, and the number of filters of the two matrices at the L_q-th layer, and T_{1,L_q} and T_{2,L_q} are the temporal size of the two modalities at the L_q-th layer. An hidden encoder layer $H_{l_q,m}$, $l_q = 1, ..., L_q$, belonging to the encoder analyzing the modality $m = 1, 2$ can be defined by $H_{l_q,m} = \alpha(W_{l_q,m} * H_{l_q-1,m} + b_{l_q,m})$, where $W_{l_q,m}$ and $b_{l_q,m}$ are the weights and bias of the l_q-th layer in the encoder processing the modality m, α is the LeakyReLu activation function, and $*$ is the convolutional operation. Each layer l_q compresses spatiotemporal information to increase the receptive field of the model.

From the spatiotemporal features $H_{L_q,m}$ obtained from the encoder outputs, it is possible to derive the sampled latent variables corresponding to modality domains z_m through a sampling operation:

$$z_m \sim q_{\phi_m}(z_m|I_m). \tag{1}$$

However, this sampling operation is non-differentiable. To make the model differentiable and, as a simplification, assuming that the modality probabilistic distributions are Gaussian, the probabilistic distributions of the two modalities z_m are defined:

$$z_m = \mu_m + \Sigma_m \epsilon, \epsilon \sim N(0, 1), \tag{2}$$

where $\mu_m \in \mathbb{R}^{f_{L_q} \times T_{m,L_q}}$ and $\Sigma_m \in \mathbb{R}^{f_{L_q} \times f_{L_q} \times T_{m,L_q}}$ are the mean and coviariance matrices of the two modalities, and ϵ is a random number generator with a standard normal distribution. μ_m and Σ_m are learned by the encoders using parametrization layers. Each parametrization layer is a 3D convolutional layer with a kernel size of $1 \times 1 \times 1$. Thus, $\mu_m, \Sigma_m = \alpha(W_{p,m} * H_{L_q,m} + b_{p,m})$, where $W_{p,m}$ and $b_{p,m}$ are the weights and biases of the parametrization layers processing the spatiotemporal features $H_{L_q,m}$ of the modality m. We use 3D convolutional layers to represent μ_m and Σ_m, rather than fully connected or monodimensional layers, because they provide a better representation of the probabilistic distribution parameters of image time series across more dimensions than monodimensional layers.

The sample latent variable z representing the joint domain between the modalities can be defined by considering Eq. 1 as:

$$z \sim q_\phi(z|I_1, I_2). \tag{3}$$

Since we assume that the various modalities are statistically independent, we can represent $q_\phi(z|I_1, I_2)$ as the product of the approximate posterior probabilities learned for each modality:

$$q_\phi(z|I_1, I_2)\alpha \prod_{m=1}^{2} q_{\phi_m}(z_m|I_m). \tag{4}$$

This formulation corresponds to the PoE ones. Since Eq. 2 proves that we can represent a sample latent variable z_m using its mean and variance in the case it is assumed to have a Gaussian distribution, the sampled latent variable z represents the joint domain, which can be expressed as:

$$z = \mu + \Sigma\epsilon, \epsilon \sim N(0, 1), \tag{5}$$

where μ and Σ are the mean and covariance matrices derived from the PoE using the μ_m and Σ_m of each modality and computed with the following closed forms:

$$\mu = \left(\sum_{m=1}^{2} \mu_m \Sigma_m^{-1} \right) \left(\sum_{m=1}^{2} \Sigma_m \right), \Sigma = \left(\sum_{m=1}^{2} \Sigma_m \right). \tag{6}$$

2.2 Probabilistic Decoders

The output of the sampled latent variable z is passed to the three decoders. Two decoders $p_{\theta_m}(\hat{I}_m|z)$, where θ_m are the decoder parameters for the modality m and $\hat{I}_m$ is the reconstruction of the input image time series belonging to the modality m, aim to the reconstruct the input image time series I_m. The reconstruction process helps the VAE learning spatiotemporal features and the parameters of the probabilistic distributions (i.e., μ_m, Σ_m, μ, and Σ). The third decoder $p_{\theta_{seg}}(\hat{Y}|z)$, where θ_{seg} are the decoder parameters and $\hat{Y}$ are output segmentation map, uses the features deriving from z to process the multimodal image time series and obtain a segmentation map. The three decoders process the features from z using three 3D convolutional layers, one for each decoder, with a kernel of $1 \times 1 \times 1$ to adjust the feature size to match the decoder ones. Afterwards, the decoders upsample and process their spatiotemporal information using 3D convolutional layers with a fixed kernel size of $k \times k \times k$ followed by 3D batch normalization layers and LeakyReLU activation functions. The final layers of the reconstruction decoders $p_{\theta_m}(\hat{I}_m|z)$ are 3D convolutional layers with as many filters as s_1 and s_2, followed by sigmoid activation functions and obtain the reconstrunction of the multimodal image time series $\hat{I}_{m,n}$. Whereas, the segmentation decoder $p_{\theta_{seg}}(\hat{Y}|z)$ has a final layer composed of a 3D convolutional layer with a filter number equal to the number of classes C and uses a softmax activation function to derive the segmentation map $\hat{Y}_n$.

2.3 Loss Function

The proposed multimodal VAE learns spatiotemporal features to perform semantic segmentation and latent spaces that represent the probabilistic distributions of the two modalities during training. To do so, we train the proposed model using a combination of a Mean Squared Error (MSE) loss function for reconstructing the multimodal image time series, a categorical cross-entropy loss $\mathcal{L}_{seg}$ for the semantic segmentation, and a Kullback Leibler (KL) divergence to encourage the learning of a well-structured latent space. We used a weighted categorical cross-entropy loss function to alleviate the problem of the imbalance between the dataset classes, and it is defined as follows:

$$\mathcal{L}_{seg} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c Y_{c,n} log(\hat{Y}_{c,n}), \tag{7}$$

where w_c is the weight for the class $c = 1, ..., C$, and $Y_{c,n}$ and $\hat{Y}_{c,n}$ are the true label and predicted label for the class c. The MSE loss function for the modality m can be defined as:

$$MSE_m = \frac{1}{N}\sum_{n=1}^{N}(I_{m,n} - \hat{I}_{m,n})^2. \tag{8}$$

Whereas, the KL divergence controls the learning of the latent space, and therefore, the probabilistic distribution represented by the statistical parameters μ and σ, and is defined as:

$$KL = \frac{1}{2}\sum_{j=1}^{J}(\mu_j^2 + \sigma_j^2 - log\sigma_j^2 - 1), \tag{9}$$

where μ_j and σ_j are the mean and the standard deviation of the j-th latent dimension with $j = 1, ..., J$ and J is the total number of latent dimensions. The loss function used for the training of the multimodal VAE is:

$$\mathcal{L} = \sum_{m=1}^{2} \lambda_m MSE_m + \gamma_{seg}\mathcal{L}_{seg} + \beta KL, \tag{10}$$

where γ_{seg} is the constraint variable used to weight the categorical cross-entropy loss, λ_m is the weight for the reconstruction loss of each modality m, and β controls the contribution of the KL divergence.

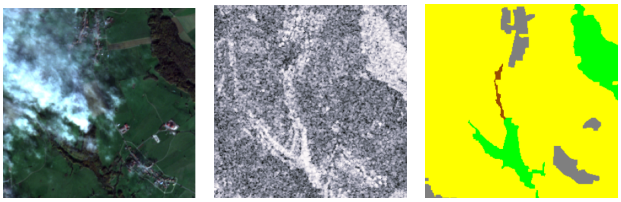
3. Experimental settings and results

3.1 Dataset

We evaluated the proposed method using two datasets. The first dataset is the MultiSenGE Dataset (Wenger et al., 2022) that is exploited for LandUse and LandCover (LULC) tasks and is composed of Sentinel-1 (Figure 2(b)) and Sentinel-2 (Figure 2(a)) image time series, covering the Grand-Est region in France (57433 km^2). The dataset originally had 14 classes, but we aggregated them into five more conservative classes (Figure 2(c)). The aggregated classes are not equally distributed (Table 1). Agricultural areas and Forests represent the majority, followed by Urban Areas, while Wetlands and Water Surfaces have very few samples. In particular, Wetlands tend to be small areas similar to their surroundings, which makes them challenging to segment accurately.
This aggregation was chosen because it provides a more readable, interpretable set of categories while remaining sufficiently diverse to evaluate the model. It includes classes with both small and large areas, classes that vary over time, and classes that exhibit more stable temporal behavior.

Category	Distribution
	%
Urban	7.5
Agriculture	58.7
Forests	32.5
Wetlands	0.31
Water Surfaces	0.89

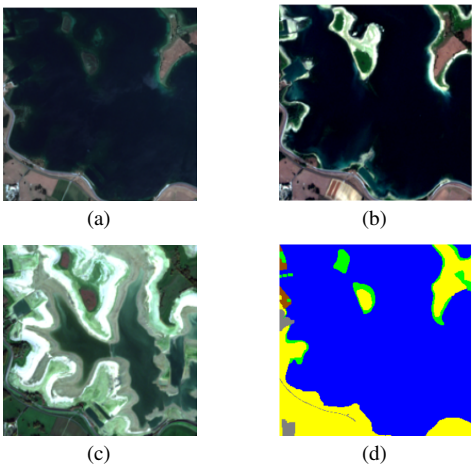
Table 1. Aggregated categories of the MultiSenGE dataset with their distribution.



(a) (b) (c)

Figure 2. Example of MultiSenGE (a) Sentinel-2 and (b) Sentinel-1 patches, and (c) the corresponding reference map (Grey is Urban, yellow is Agriculture, green is Forests, brown is Wetlands, and blue is Water Surfaces).

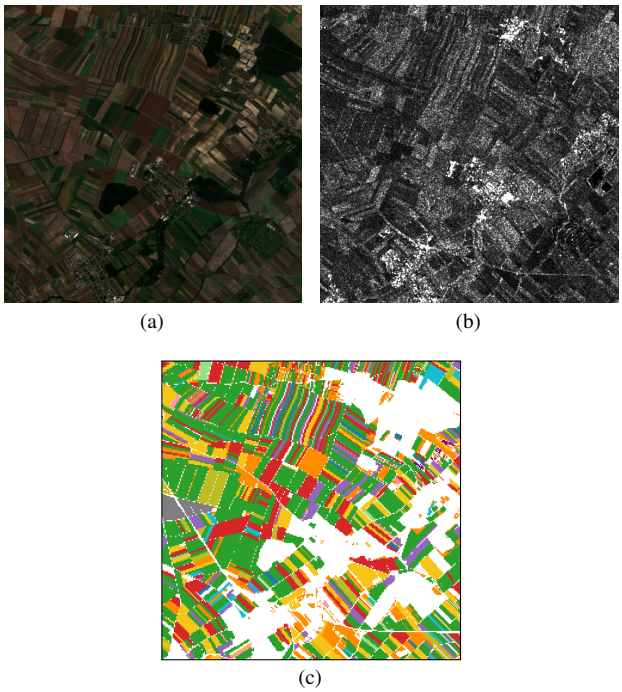
Some classes have more temporal variation in the spectral values than others (e.g., Agriculture or Wetlands). As shown in Figure 3, the RGB images reveal crop color changes: brown/green in July at harvest time, brown in September after harvest, and green in November during regrowth. This temporal variation underscores the crucial need for image time series in distinguishing classes with high temporal variation in spectral values, which is otherwise impossible in single-date semantic segmentation. Figure 3 also shows that only a segmentation map is provided for the whole image time series, even though some classes, such as water surfaces, vary in size across the image time series, making the semantic segmentation challenging.



(a) (b) (c) (d)

Figure 3. Three examples of Sentinel-2 images belonging to the same time series acquired on (a) July, (b) September, and (c) November, and (d) the corresponding reference map (Grey is Urban, yellow is Agriculture, green is Forests, brown is Wetlands, and blue is Water Surfaces).

The MultiSenGE dataset uses the ten spectral bands at 10 m/pixel and 20 m/pixel for the Sentinel-2 image time series, interpolated at 10 m/pixel, and exploits both VV and VH polarizations for the Sentinel-1 image time series. For each Sentinel-2 image time series, we considered six images acquired in January, June, July, August, September, and November. We chose these months because they have the most images in the dataset. For the Sentinel-1 image time series, we selected the images closest in time to the Sentinel-2 ones. From the MultiSenGE dataset, we considered 8157 patches containing pairs of Sentinel-1/Sentinel-2 image time series with a size of 256×256. To increase the dataset size, we split the patches into 16 sub-patches, yielding 130512 patches of size 64×64. We divided the dataset into training (80%), validation (10%), and test (10%) sets, resulting in a training set of 104410 samples and validation and test sets of 13051 samples each.



(a) (b)

(c)

Figure 4. Example of the Austria dataset: (a) Sentinel-2 and (b) Sentinel-1 patches, and (c) the corresponding reference map (white areas are unclassified and not included in the quantitative results, orange is Grassland, light blue is Small-seeded grass, yellow is Maize, red is Soy, green is Wheat, Lilac is Sunflower, brown is Potato, purple is Wine, gray is Colza, light green is beet, cyan is Cucurbitacea, pink is Other cereals, and leek green is Legumes).

The second dataset consists of Sentinel-1 (Figure 4(b)) and Sentinel-2 (Figure 4(a)) image time series acquired from September 2020 to August 2021 over an agricultural area in Austria. Each image time series consists of 12 images, one per month, and the images from the two modalities are not temporally correlated. Thus, unlike the previous dataset, the Sentinel-1 images are not acquired as close in time as possible to the Sentinel-2 images. The reference map (Figure 4(c)) is derived and adapted from the Invekos dataset, which comprises various agricultural classes and fields reported by farmers (Austria, 2021, Weikmann et al., 2023). The dataset comprises 13 unbalanced agricultural classes (Table 2), whose spectral behavior varies widely throughout the year. We used 10 spectral bands for the Sentinel-2 data, excluding the atmospheric bands at 60 m/pixel of spatial resolution, and the VV and VH polarizations of Sentinel-1 images. All the Sentinel-2 spectral bands are interpolated at 10 m/pixel of spatial resolution, whereas Sentinel-1 images have a spatial resolution of 10 m/pixel. The Sentinel-2 and Sentinel-1 image time series are spatially coregistered. We sampled the image time series to obtain patches of size $64 \times 64 \times 12 \times 10$ for the Sentinel-2 data and $64 \times 64 \times 12 \times 2$ for the Sentinel-1 data, and composed a training set of 24159 patches and a validation set of 3611. As a test area, we consider a 750×750 pixel region, disjoint from the training and validation sets.

Category	Distribution
	%
Grassland	9.27
Small-seeded Grass	1.15
Maize	13.65
Soy	14.34
Wheat	44.02
Sunflower	4.70
Potato	1.72
Wine	1.47
Colza	0.81
Beet	5.08
Cucurbitacea	2.37
Other cereals	0.50
Legumes	0.92

Table 2. Categories of the Austria dataset with their distribution.

3.2 Experimental settings

We trained the model using the Adam optimizer with a learning rate of 5×10^{-4}, a batch size of 128, and 100 training epochs. The number of epochs is sufficient, allowing the model to reach a performance plateau before the end of training. The hyperparameters of the loss function presented Section 2.3 are $\beta = 0.1$, $\gamma_{seg} = 5$, and $\lambda_m = (0.5, 0.5)$. $w_c = (1, 1, 1, 5, 1)$ in the MultiSenGE dataset, whereas all values of w_c are equal to 1 for the Austria dataset. In the MultiSenGe dataset, a higher weight w_c was assigned to the Wetlands class to address class imbalance in the dataset, as it is the least represented class. During the experiments, these hyperparameter values provided the best results. We implemented the model using the PyTorch library. All experiments were executed on NVIDIA Tesla V100 GPUs. To ensure reproducibility, we run all experiments with a fixed random seed.

We demonstrated the effectiveness of incorporating multimodal and multitemporal information into the proposed model through an ablation study. This study compares the proposed method with two models processing single-modal image time series (i.e., one for Sentinel-1 and one for Sentinel-2) and one model analyzing multimodal single-date images. All the models exploit the same VAE architecture, but the model processing single-modal image time series had a probabilistic encoder and two decoders: one for the reconstruction and one for the semantic segmentation; whereas, the model processing multimodal single-date images used 2D convolutional layer instead of the 3D ones since it did not process the temporal information. The ablation study was done using the MultiSenGe dataset. We then compared the proposed method with ConvLSTM+Inception-S1S2 (CIS) (Wenger et al., 2023) and U-TAE (Garnot et al., 2022). CIS is a SoA multimodal architecture based on a U-Net backbone and enhanced with two feature extractors. First, it uses a ConvLSTM module to capture spatiotemporal features from satellite image sequences. Second, a Naive Inception module extracts spatio-spectral features using convolution filters of different sizes. The

outputs of both extractors are concatenated and fed into the U-Net decoder to produce the semantic segmentation map. U-TAE analyzes spatiotemporal information using a U-Net-based architecture and employs a temporal attention encoder to handle temporal information.

To evaluate the performance of our model on the segmentation maps, we used Overall Accuracy (OA), F1-score, and Intersection over Union (IoU).

3.3 Effectiveness of the multimodal and multitemporal information

This ablation study examined the contribution of multimodal and multitemporal information to the proposed method. As illustrated in Table 3, the model processing only the Sentinel-1 image time series achieved the lowest performance (e.g., OA of 67.38%). The model, which processed only the Sentinel-2 image time series, improved the OA by 9.47% compared to the Sentinel-1 case. This is because Sentinel-2 images provide more information and are less noisy than Sentinel-1. The combination of Sentinel-1 and Sentinel-2 image time series allowed the proposed method to achieve the best performance (e.g., OA of 77.19%), slightly improving the results compared to the Sentinel-2 case. This improvement demonstrated that the proposed method can effectively combine Sentinel-1 and Sentinel-2 image time series and extract more informative features, thereby improving performance. We can also observe that including temporal information in the proposed method improved the F1 score by 3.58% compared to the multimodal single-date case. The temporal information enabled the model to distinguish better classes characterized by spectral variation over time, which are challenging to categorize with the multimodal single-date setup. However, the proposed model was the most computationally demanding since it used multiple encoders and decoders composed of 3D convolutional layers to analyze the multimodal and spatiotemporal information. The model processing multimodal single-date images had the fewest parameters because it used 2D convolutional layers rather than 3D ones.

Method	OA	F1	IoU	Params.
	%	%	%	
Single-date	75.13	71.70	62.88	**2.30M**
S1 time series	67.38	67.87	57.98	3.74M
S2 time series	76.85	74.58	65.36	3.75M
Proposed	**77.08**	**76.76**	**67.13**	6.24M

Table 3. Ablation study on the multitemporal and multimodal information effectiveness.

Class	S1	S2	Single-date	Proposed
	%	%	%	%
Urban	72.40	**82.37**	80.72	79.39
Agriculture	93.39	95.25	94.54	**95.40**
Forests	91.55	**94.04**	93.03	93.67
Wetlands	14.41	27.98	15.69	**34.92**
Water Surfaces	67.61	73.28	74.50	**80.39**

Table 4. F1-score per class for Proposed, S1, S2 and Static models.

Table 4 shows the F1 score of the analyzed model for the five considered classes. The proposed method outperformed all other models, except for Wetlands, where the model processing Sentinel-2 image time series achieved a slightly higher score (e.g., 0.34%). The slightly lower performance of the proposed method in Wetland identification compared to the Sentinel-2 case may be due to the geometrical distortion introduced by the Sentinel-1 image time series. The two other models generally struggle in the smaller classes, such as Water and Wetlands. The Sentinel-2 case achieved a comparable F1-score to the proposed method for Wetlands. However, the former detected more false alarms (Figure 5(c)) than the proposed method (Figure 5(e)), where we can observe that the small Water surface is incorrectly predicted as Wetlands. Thus, including temporal information is beneficial for distinguishing classes with a spectral variation over time, such as Wetlands and Agriculture.

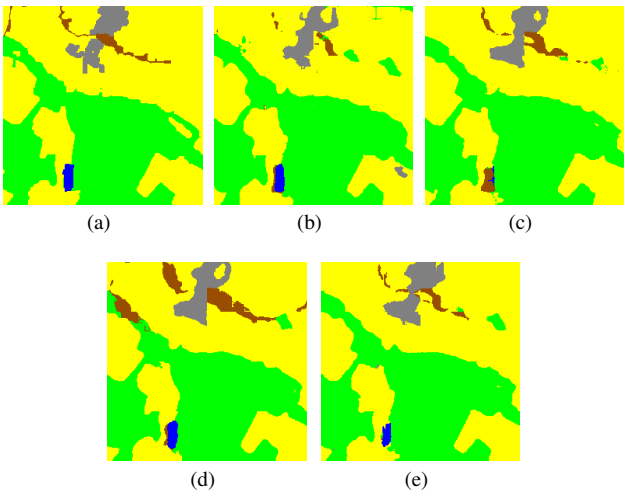


Figure 5. Comparison between (a) the reference map and the segmentation maps obtained using a VAE processing (b) Sentinel-1 image time series, (c) Sentinel-2 image time series, (d) multimodal single-date images, and (e) the proposed method (Grey is Urban, yellow is Agriculture, green is Forests, brown is Wetlands, and blue is Water Surfaces).

3.4 State-of-the-Art comparison: MultiSenGE dataset

The comparison between the proposed method and the SoA multimodal method CIS showed that the proposed method consistently outperformed CIS (Table 5). It improved the OA of 7.88%, the F1 score of 14.98%, and the IoU of 15.50%. The effectiveness of the proposed method compared to CIS in the semantic segmentation can also be observed in Table 6, where the proposed method sharply improved the identification of both majority (e.g., improvement in Agriculture by 8.31%) and minority classes (e.g., improvement in Wetlands by 30.31%). Thus, the quantitative results demonstrated that the proposed method more effectively modeled the multimodal and spatiotemporal information than CIS. This improvement is due to the proposed method capability to learn a joint latent domain across multiple modalities, whereas CIS learned a shared domain, having a domain gap, which reduced the CIS performance. This is achieved thanks to the VAE capacity to learn the probabilistic distributions of the modalities and to PoE, which combines them. The proposed method achieved results comparable to U-TAE using a temporal attention encoder (Table 5). U-TAE achieved slightly better results in identifying water bodies than the proposed method (Table 6), likely due to better modelling of temporal information, since the class varies significantly over time, as seen in Figure 3. However, the proposed method achieved slightly better performance in identifying a minor and mixed class, such as Wetlands, thanks to its combination of multiple modalities.

The qualitative results confirm that the proposed model significantly outperforms CIS. Even though the proposed method under-

Method	OA	F1	IoU	Params.
	%	%	%	
CIS	69.20	61.78	51.63	3.74M
U-TAE	76.71	**77.97**	**69.07**	**2.16M**
Proposed	**77.08**	76.76	67.13	6.24M

Table 5. Quantitative results of CIS, U-TAE, and the proposed method (MultiSenGe dataset).

Category	CIS	U-TAE	Proposed
	%	%	%
Urban	65.65	**81.85**	79.39
Agriculture	87.09	**95.92**	95.40
Forests	89.04	**94.31**	93.67
Wetlands	4.61	34.39	**34.92**
Water Surfaces	62.50	**82.84**	80.39

Table 6. F1-score per class for CIS, U-TAE, and the proposed method (MultiSenGe dataset).

estimated some small areas of Urban and Forests due to spatial compression in the 3D convolutional layers, it accurately detected most classes (Figure 6(d)). In contrast, CIS over-predicted the Water Surfaces, Wetlands, and Forests classes (Figure 6(b)). The lower performance of CIS appears to be related to its loss calibration and class aggregation strategy. The superior performance of the proposed model can be attributed to its VAE-based architecture, which jointly encodes multimodal and multitemporal information in a shared latent space, capturing correlations across modalities and over time before decoding. In contrast, the CIS model processed each modality separately and merged features only at a later stage, thereby limiting its ability to exploit cross-modal and temporal dependencies. As observed in the quantitative results, the U-TAE (Figure 6(c)) and the proposed method (Figure 6(d)) achieved similar qualitative results. However, it is worth noting that the proposed method identified the wetland areas, whereas U-TAE misclassified them as Forest and Water surfaces. This is due to the better modelling of the multiple modalities and their combination performed by the proposed method.

3.5 State-of-the-Art comparison: Austria dataset

The comparison between the proposed method and the SoA methods using the Austria dataset showed that the proposed method outperformed CIS by increasing OA, F1-score, and IoU by 10.22%, 12.29%, and 12.07%, respectively. This is mostly due to better modelling of multiple modalities, which allowed better identification of most classes compared to CIS (Table 8). The proposed method achieved comparable performance with U-TAE (Table 7). The proposed method increased the F1-score by 2.30% compared to U-TAE, particularly improving the detection of Cucurbitacea and Other Cereals (Table 8). This overall increase in class identification indicates that it better models information across multiple modalities. However, U-TAE better identified the shape of agricultural fields (e.g., an IoU increase of 2%) than the proposed method, thanks to the U-Net-based architecture and better modelling of temporal information.

Method	OA	F1	IoU	Params.
	%	%	%	
CIS	39.72	39.94	28.96	3.74M
U-TAE	**50.97**	49.93	**43.03**	**2.16M**
Proposed	49.94	**52.23**	41.03	6.24M

Table 7. Quantitative results of CIS, U-TAE, and the proposed method (Austria dataset).

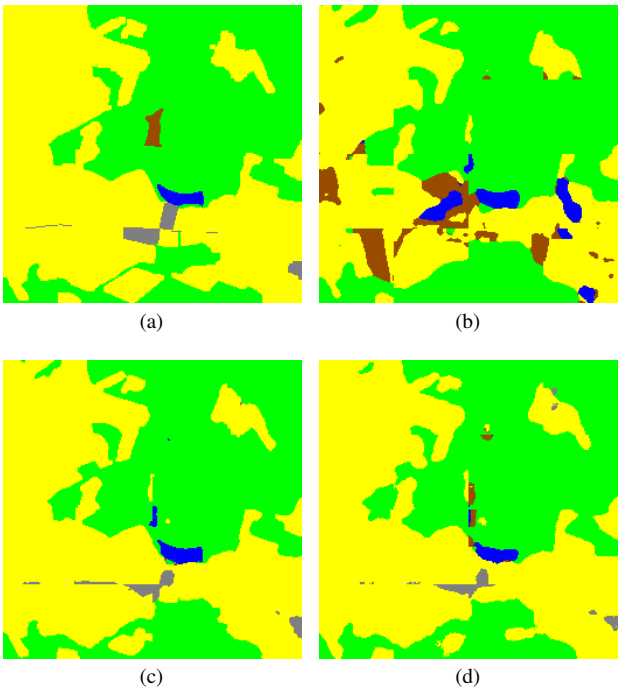


Figure 6. Comparison between (a) the reference map and the segmentation maps obtained using (b) CIS, (c) UTAE, and (d) the proposed method (Grey is Urban, yellow is Agriculture, green is Forests, brown is Wetlands, and blue is Water Surfaces).

From the qualitative results, we observed that the proposed method (Figure 7(d)) and U-TAE (Figure 7(c)) correctly classified most agricultural fields and performed similarly, with the proposed method classifying the agricultural fields, such as Cucurbitacea (see the upper right part of Figures 7(c) and 7(d)), slightly more accurately than U-TAE. Whereas, CIS misclassified some agricultural fields by overestimating some classes, such as Small-seeded grass (see upper left part of Figure 7(b)). This is due to better modelling of the multimodal and temporal information in U-TAE and the proposed method compared to CIS. We can also observe that the proposed method and U-TAE did not preserve the agricultural field borders, resulting in the loss of some thin-shaped fields.

4. Conclusions

We proposed a supervised DL multimodal method to analyze the spatiotemporal information of image time series acquired from various sensors and to perform semantic segmentation. The proposed method employed a VAE with 3D convolutional layers to model spatiotemporal information in the input and learn probabilistic distributions across various modalities. These distributions are combined using a PoE approach into a probabilistic distribution that characterizes a latent space domain common to the input modalities. The experimental results demonstrated the effectiveness of jointly analyzing multimodal and multitemporal information to improve the semantic segmentation performance of classes by extracting more informative features from the multimodal input and by better modeling classes characterized by spectral variation over time. We also showed that learning the probabilistic distributions of multiple modalities and their combination using the PoE approach to create a joint domain improved modeling of multiple modalities compared to SoA methodologies that process each modality separately and merge

Category	CIS	U-TAE	Proposed
	%	%	%
Grassland	44.05	**72.60**	71.49
Small-seeded Grass	9.72	**28.73**	16.47
Maize	59.64	**73.60**	71.07
Soy	71.11	**76.93**	74.46
Wheat	83.35	**90.44**	88.02
Sunflower	73.95	**81.19**	76.73
Potato	**3.84**	0.00	0.00
Wine	5.47	**24.75**	22.33
Colza	44.05	**93.72**	90.38
Beet	63.22	**77.19**	74.86
Cucurbitacea	17.37	36.33	**48.94**
Other cereals	3.97	4.33	**7.92**
Legumes	**39.45**	39.27	36.35

Table 8. F1-score per class for CIS, U-TAE, and the proposed method (Austria dataset).

features only at a later stage. This mechanism enabled the proposed method to achieve performance comparable to or better than that of SoA methods.

As future work, we plan to explore the learned latent spaces of the joint domain and various modalities in greater depth to understand which features are shared across modalities and which are specific to each. Since the assumption of the SAR data probability distribution as Gaussian-like is strong and can degrade model performance, we plan to improve the modelling of the SAR probability distribution to learn a more efficient shared domain, thereby increasing model performance. We also aim to improve the proposed method by adding ConvLSTM layers to better model temporal information, incorporating attention mechanisms to refine multimodal feature learning, and testing it on other datasets with different modalities.

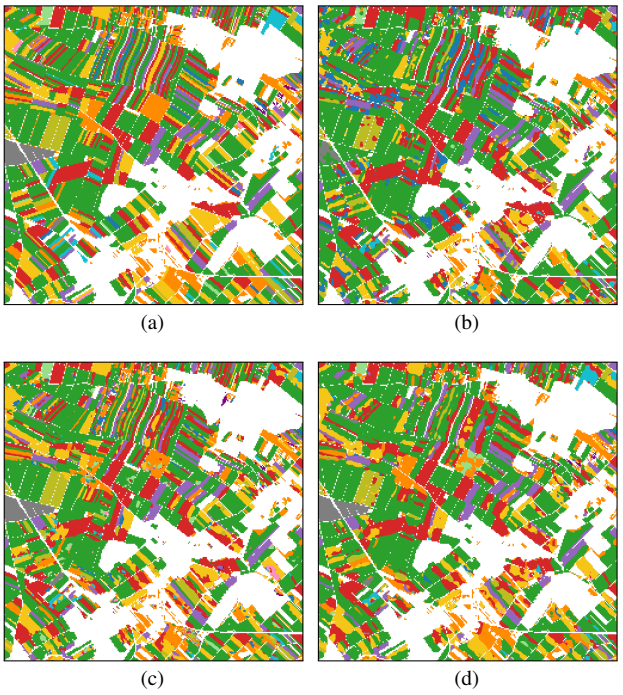


Figure 7. Comparison between (a) the reference map and the segmentation maps obtained using (b) CIS, (c) UTAE, and (d) the proposed method (white areas are unclassified, orange is Grassland, light blue is Small-seeded grass, yellow is Maize, red is Soy, green is Wheat, Lilac is Sunflower, brown is Potato, purple is Wine, gray is Colza, light green is beet, cyan is Cucurbitacea, pink is Other cereals, and leek green is Legumes)).

References

Austria, 2021. Invekos schl¨age ¨ osterreich 2021.

Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R. G., Dupuy, S., 2018. M^3Fusion: A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12), 4939-4949.

Cai, Y., Zhang, Z., Ghamisi, P., Rasti, B., Liu, X., Cai, Z., 2023. Transformer-based contrastive prototypical clustering for multimodal remote sensing data. *Information Sciences*, 649, 119655.

Chang, H., Bi, H., Li, F., Xu, C., Chanussot, J., Hong, D., 2024. Deep Symmetric Fusion Transformer for Multimodal Remote Sensing Data Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15.

Cui, X., Zhang, L., 2026. CADSM: crossing-attention dual-scanning Mamba for multimodal remote sensing image classification. *Journal of Applied Remote Sensing*, 20(2), 021406–021406.

Follath, T., Mickisch, D., Hemmerling, J., Erasmi, S., Schwieder, M., Demir, B., 2024. Multi-modal vision transformers for crop mapping from satellite image time series. *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 1937–1941.

Garnot, V. S. F., Landrieu, L., 2021. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4872–4881.

Garnot, V. S. F., Landrieu, L., Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187, 294–305.

Gu, A., Dao, T., 2024. Mamba: Linear-time sequence modeling with selective state spaces. *First conference on language modeling*.

He, X., Han, X., Chen, Y., Huang, L., 2025. A light-weighted fusion vision mamba for multimodal remote sensing data classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Hong, D., Yao, J., Meng, D., Xu, Z., Chanussot, J., 2021. Multimodal GANs: Toward Crossmodal Hyperspectral–Multispectral Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6), 5103-5113.

Hu, J., Mou, L., Schmitt, A., Zhu, X. X., 2017. Fusionet: A two-stream convolutional neural network for urban scene classification using polsar and hyperspectral data. *2017 Joint Urban Remote Sensing Event (JURSE)*, 1–4.

Ienco, D., Interdonato, R., Gaetano, R., Ho Tong Minh, D., 2019. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning

architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 11-22.

Li, W., Gao, Y., Zhang, M., Tao, R., Du, Q., 2023. Asymmetric Feature Fusion Network for Hyperspectral and SAR Image Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10), 8057-8070.

Liu, X., Hong, D., Chanussot, J., Zhao, B., Ghamisi, P., 2022. Modality Translation in Remote Sensing Time Series. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14.

Luppino, L. T., Hansen, M. A., Kampffmeyer, M., Bianchi, F. M., Moser, G., Jenssen, R., Anfinsen, S. N., 2024. Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1), 60-72.

Ma, X., Zhang, X., Pun, M.-O., Liu, M., 2024. A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15.

Roy, S. K., Deria, A., Hong, D., Rasti, B., Plaza, A., Chanussot, J., 2023. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-20.

Saha, S., Bovolo, F., Bruzzone, L., 2021. Building Change Detection in VHR SAR Images via Unsupervised Deep Transcoding. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3), 1917-1929.

Wang, J., Li, W., Gao, Y., Zhang, M., Tao, R., Du, Q., 2022. Hyperspectral and SAR image classification via multiscale interactive fusion network. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12), 10823–10837.

Weikmann, G., Marinelli, D., Paris, C., Migdall, S., Gleisberg, E., Appel, F., Bach, H., Dowling, J., Bruzzone, L., 2023. Multi-year Mapping of Water Demand at Crop Level: An End-to-End Workflow Based on High-Resolution Crop Type Maps and Meteorological Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 6758-6775.

Wenger, R., Puissant, A., Weber, J., Idoumghar, L., Forestier, G., 2022. Multisenge: A Multimodal and Multitemporal Benchmark Dataset for Land Use/Land Cover Remote Sensing Applications. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2022, 635–640. https://isprs-annals.copernicus.org/articles/V-3-2022/635/2022/.

Wenger, R., Puissant, A., Weber, J., Idoumghar, L., Forestier, G., 2023. Multimodal and Multitemporal Land Use/Land Cover Semantic Segmentation on Sentinel-1 and Sentinel-2 Imagery: An Application on a MultiSenGE Dataset. *Remote Sensing*, 15(1). https://www.mdpi.com/2072-4292/15/1/151.

Yan, J., Liu, J., Liang, D., Wang, Y., Li, J., Wang, L., 2023a. Semantic Segmentation of Land Cover in Urban Areas by Fusing Multisource Satellite Image Time Series. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-15.

Yan, Z., Li, J., Li, X., Zhou, R., Zhang, W., Feng, Y., Diao, W., Fu, K., Sun, X., 2023b. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-16.

Zhang, Y., Ru, L., Wu, K., Yu, L., Liang, L., Li, Y., Chen, J., 2025. Skysense v2: A unified foundation model for multi-modal remote sensing. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9136–9146.

Zhao, F., Zhang, C., Geng, B., 2024. Deep Multimodal Data Fusion. *ACM Comput. Surv.*, 56(9). https://doi.org/10.1145/3649447.

Zhao, W., Zhao, Z., Xu, M., Ding, Y., Gong, J., 2025. Differential multimodal fusion algorithm for remote sensing object detection through multi-branch feature extraction. *Expert Systems With Applications*, 265, 125826.

Zhou, G., Lihuang, Q., Gamba, P., 2025. Advances on multimodal remote sensing foundation models for Earth observation downstream tasks: A survey. *Remote Sensing*, 17(21), 3532.

Zhu, Z., Yuan, X., Gan, S., Li, R., Bi, R., Luo, W., Chen, C., 2025. PS-GAN: A Novel Pseudo-Siamese Generative Adversarial Network for Multimodal Remote Sensing Image Change Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 16609-16632.