

Multi-Modal Temporal Attention Models for Crop Mapping from Satellite Time Series

Vivien Sainte Fare Garnot¹, Loic Landrieu¹, and Nesrine Chehata²

¹LASTIG, ENSG, IGN, Univ Gustave Eiffel, Saint-Mande, France

²EA G&E Bordeaux INP, Univ Bordeaux Montaigne, Bordeaux

December 15, 2021

Abstract

Optical and radar satellite time series are synergetic: optical images contain rich spectral information, while C-band radar captures useful geometrical information and is immune to cloud cover. Motivated by the recent success of temporal attention-based methods across multiple crop mapping tasks, we propose to investigate how these models can be adapted to operate on several modalities. We implement and evaluate multiple fusion schemes, including a novel approach and simple adjustments to the training procedure, significantly improving performance and efficiency with little added complexity. We show that most fusion schemes have advantages and drawbacks, making them relevant for specific settings. We then evaluate the benefit of multimodality across several tasks: parcel classification, pixel-based segmentation, and panoptic parcel segmentation. We show that by leveraging both optical and radar time series, multimodal temporal attention-based models can outmatch single-modality models in terms of performance and resilience to cloud cover. To conduct these experiments, we augment the PASTIS dataset (Garnot and Landrieu, 2021) with spatially aligned radar image time series. The resulting dataset, PASTIS-R, constitutes the first large-scale, multimodal, and open-access satellite time series dataset with semantic and instance annotations.

1 Introduction

The multiplication of Earth Observation satellites with various sensors represents an opportunity to improve the automated analysis of remote sensing data for tasks such as crop mapping. Indeed, different modalities capture information of different natures and distinct spatial and temporal resolutions and have varying resilience to atmospheric conditions. Machine learning models can leverage these complementary characteristics to learn richer and more robust representations. In particular, C-band radar and optical images possess well-known synergies for automated crop mapping (Van Tricht et al., 2018; Steinhagen et al., 2018; Campos-Taberner et al., 2019), the driving application of this paper. More specifically, multispectral time series contain highly relevant information for monitoring the evolution of plant phenology (Vrieling et al., 2018; Segarra et al., 2020). For example, the study of red and infrared reflectances helps monitoring photosynthetic activity (Tucker, 1979). However, passive optical sensors are highly susceptible to cloud cover and atmospheric distortion (Sudmanns et al., 2020). Conversely, due to the influence of extrinsic factors such as humidity and terrain, it is harder to extract discriminative information from radar images for crop mapping. On the other hand, the high revisit frequency and imperviousness to cloud cover make them uniquely well-suited for monitoring the rapid-changing biological processes of agricultural parcels (McNairn et al., 2014).

Automated crop mapping is necessary for various applications carrying crucial economic and ecological stakes, such as environmental monitoring, subsidy allocation, and food price prediction. For example, the Common Agricultural Policy is responsible for the allocation of over 57 billion euros each year of agricultural subsidies in the European Union (Commission, 2016). This application is at the center of an effort towards algorithmic solutions for machine learning-based crop monitoring

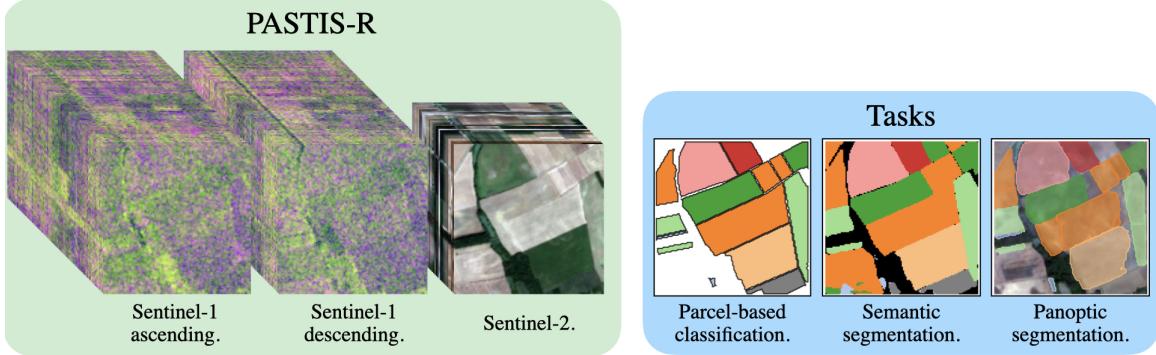


Figure 1: We introduce the PASTIS-R dataset containing 2433 multimodal image time series of Sentinel-2 and Sentinel-1 data. On PASTIS-R, we evaluate different fusion strategies and enhancements on parcel-based classification, semantic segmentation, and panoptic segmentation.

(Koetz et al., 2019). This endeavor is aided by the accessibility of high-quality satellite data worldwide (Drusch et al., 2012) and individual parcel annotations for some European countries such as France (EtaLab, 2017), which is particularly conducive to training large-scale deep networks.

In the context of crop type mapping, the fusion of optical and radar time series has been extensively explored with traditional machine learning methods (Van Tricht et al., 2018; Steinhausen et al., 2018; He and Yokoya, 2018; Campos-Taberner et al., 2019; Orynbaikyzy et al., 2020; Giordano et al., 2020), and more recently recurrent neural networks (Ienco et al., 2019). However, despite the significant performance gain offered by methods based on temporal attention (Rußwurm and Körner, 2020; Garnot et al., 2020; Kondmann et al., 2021; Garnot and Landrieu, 2020), these approaches are mostly restricted to the analysis of optical Satellite Image Time Series (SITS). Recently, Ofori-Ampofo et al. (2021) proposed a first exploration of the benefit of fusion strategies for parcel-based crop type classification from Sentinel-1 and Sentinel-2 time series with attention-based methods. The present paper extends their analysis to a broader set of crop mapping tasks, including semantic and panoptic segmentation (Kirillov et al., 2019; Garnot and Landrieu, 2021) of agricultural parcels. We also study the performance benefit of standard enhancements such as auxiliary supervision and temporal dropout.

To train and evaluate our models, we augment the open-access PASTIS optical time series dataset (Garnot and Landrieu, 2021) with corresponding Sentinel-1 radar acquisitions for each 2433 time series for a total of 339 174 radar images. We demonstrate that the right choice of fusion scheme can lead to improvement across the board for all tasks and increased robustness to varying cloud cover. The main contributions of this paper are as follows:

- We present a complete reformulation of fusion strategies in the context of temporal attention-based SITS encoders, as well as standard model enhancements.
- We present PASTIS-R, the first large-scale, multimodal, open-access SITS dataset with panoptic annotations.
- We evaluate our fusion schemes and enhancements on parcel classification, semantic and panoptic segmentation, defining a new state-of-the-art for all tasks.
- We show that combining optical and radar imagery grants significant improvement in terms of robustness to varying cloud cover.

2 Related Work

In the following paragraphs, we review the recent literature on fusion approaches for the multi-temporal fusion of SITS. In particular, we detail commonly implemented fusion strategies.

Traditional Approaches for Multimodal SITS Multiple traditional machine learning approaches such as random forest or support vector machines have been adapted to handle information from optical and radar images. As highlighted by the review of Joshi et al. (2016), the joint

processing of both modalities can mitigate the sensitivity of optical images to cloud cover. Most methods use an early fusion scheme in which the radar and optical features are stacked before being processed by the model (Van Tricht et al., 2018; Mercier et al., 2019). This approach can be further improved by selecting the most relevant acquisitions (Steinhausen et al., 2018) or features (Campos-Taberner et al., 2019; Giordano et al., 2020). Orynbaiyzy et al. (2020) compare this feature concatenation approach with a decision fusion approach in which two separate random forest classifiers predict posterior probabilities over classes, and the most confident prediction is retained as the final classification. Their results show that decision fusion performs slightly worse than early feature concatenation.

Deep learning for MultiModal SITS The first multimodal deep learning models advocated for an *early fusion* scheme: the channels of all acquisitions from optical and radar time series are concatenated to form a single image with both multimodal and multitemporal pixel features. The resulting images are then processed pixelwise (Tarpanelli et al., 2018) or with convolutional networks (Kussul et al., 2017). In contrast, Ienco et al. (2019) propose to encode each radar and an optical time series separately using a combination of dedicated convolutional and recurrent-convolutional networks. In a *late-fusion* fashion, all resulting embeddings are concatenated channelwise and classified pixelwise by a Multi-Layer Perceptron (MLP). They observe that, as long as each branch is also supervised separately with auxiliary loss terms, this fusion scheme outperforms early fusion. More recently, Ofori-Ampofo et al. (2021) studied four fusion strategies for parcel-based classification with a PSE-TAE architecture (Garnot et al., 2020). Early fusion yields the best improvement on their dataset of Sentinel-2 time series and Sentinel-1 observations in descending orbit. We extend their analysis by evaluating the impact of multimodality for different tasks, evaluate the effects of typical enhancements such as auxiliary classifiers, and use both Sentinel-1 orbits in our analysis.

Other Fusion settings In a different setting, Benedetti et al. (2018) use a late fusion approach to combine mono-temporal high spatial resolution images with low spatial resolution time series, and Tom et al. (2021) exploit three different mono-temporal modalities for lake ice monitoring by training three encoders to map the different acquisitions to a common feature space. Liu et al. (2016) explore multimodal change detection on mono-temporal pairs. They propose to train two encoders in an unsupervised fashion to map simultaneously-acquired images of different modalities to a common feature space. More broadly, the synergy between radar and optical SITS has motivated other exciting applications such as the regression of optical signals from radar images (Garioud et al., 2020; Meraner et al., 2020; He and Yokoya, 2018).

Radar processing Data analysis from Synthetic-Aperture Radar (SAR) relies on either extracting backscattering coefficients, interferometric, or polarimetric features from a measured radar signal (Richards et al., 2009). Backscattering coefficients are most commonly used for crop type mapping applications (Orynbaiyzy et al., 2019). These approaches derive information on the observed surface's geometric properties and dielectric constant from the amplitude of the complex SAR signal, and discard the phase information. In contrast, interferometric SAR measure phase shift to detect potentially small deformations between two acquisitions. Interferometric features are traditionally used in geodesy (Simons and Rosen, 2007) and surface (Monserrat et al., 2014; Tarchi et al., 2003) or structural (Tomás et al., 2012; Tarchi et al., 1997; Tison et al., 2007) monitoring, but also proved discriminative for crop type mapping. Indeed, coherence estimation in interferometry can help detecting mowing, harvesting, and seeding events (Tamm et al., 2016; Mestre-Quereda et al., 2020; Shang et al., 2020), as well as providing information on crop height and density (Srivastava et al., 2006). Lastly, polarimetric SAR data analysis relies on target decomposition of polarimetric information (Cloude and Pottier, 1996; Yamaguchi et al., 2005) to provide additional terrain information, and can be used for canopy structure estimation (Srikanth et al., 2016), topography (Schuler et al., 1996), or land cover estimation (Tupin et al., 1998; Kourgli et al., 2010). However, such approaches require full polaristion radar images, *i.e.*, acquired with a sensor emitting radar waves along both polarisation directions. In this paper, we focus on crop type mapping from data of the open acces Sentinel-1 sensor which does not allow such full polarimetric analyses. Furthermore, to limit the complexity of our experiments and avoid downloading very large Single Look Complex datasets, we focus on SAR backscattering coefficients and leave the extension to interferometric features to further work.

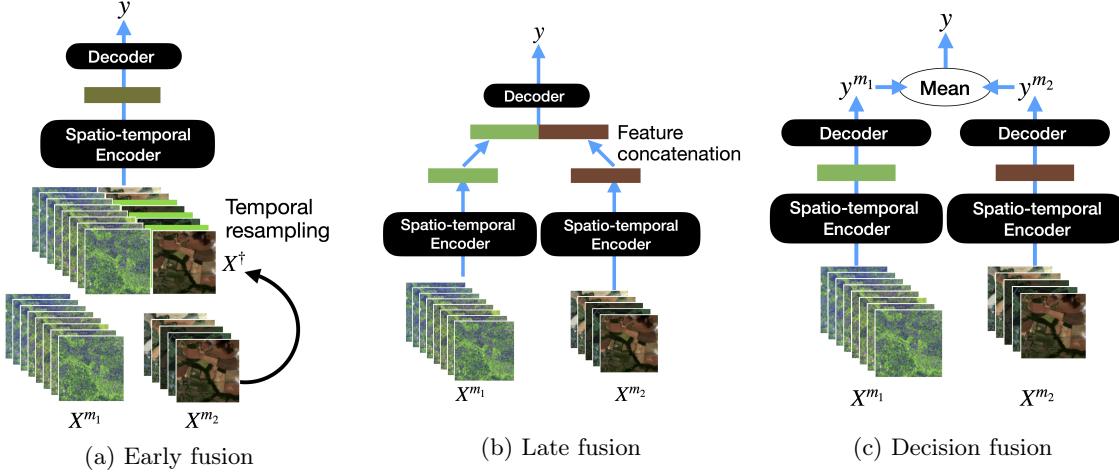


Figure 2: **Fusion Schemes.** We represent the three fusion strategies commonly found in the recent literature. (a) the raw features are interpolated and concatenated into a single sequence. (b) the learned spatio-temporal features of each modality are concatenated prior to classification. (c) each modality is processed independently and the resulting decision averaged.

3 Methods

We consider a set of M image time series $\{X^m\}_{m=1}^M$ corresponding to M distinct modalities for a single geo-referenced patch containing one or several agricultural parcels. We assume that all modalities are resampled to the same resolution for simplicity’s sake. Each time sequence X^m can be expressed as a tensor of size $T^m \times C^m \times H \times W$ with T^m the number of available temporal acquisitions for modality m , C^m the feature size for each pixel for the modality m , and $H \times W$ the spatial extent of the patch.

3.1 Fusion Strategies

The methods reviewed previously can be categorised into three main strategies: *early*, *late*, and *decision* fusion, all represented in Figure 2. We also present *mid*-fusion, a novel fusion scheme specifically adapted for multimodal time sequences. Certain terms—such as “features”—have seen their accepted meaning evolve with the gradual adoption of the deep learning paradigm, leading to ambiguity in terms such as “early” or “late” feature fusion. We propose redefining the terminology of fusion schemes unambiguously for analyzing temporal sequences of images in the following.

Early Fusion This approach combines the different modalities at the raw feature level. In our context, this amounts to concatenating the modalities *channel-wise* at each observation date. If the different acquisitions are simultaneous, and since the resolutions are identical, this is a straightforward step. However, when the modalities are captured at different times, a preprocessing step is required to interpolate all modalities to a standard temporal sampling. We denote by T^\dagger the number of time steps in the chosen temporal sampling and by X^\dagger the resulting aggregated tensor of size $T^\dagger \times C^\dagger \times H \times W$ with $C^\dagger = \sum_m C^m$ as defined in Equation 1.

This interpolation step can be costly in terms of computation and memory. Furthermore, the relevance of temporal interpolation for fast-changing processes such as plant growth and harvesting is questionable, and this is only made worse by clouds obstructing the optical modalities. However, an advantage of this approach is the simplicity of encoding X^\dagger : a single spatio-temporal encoder $\mathcal{E}_{\text{spatio-temporal}}$ can be used to learn a truly cross-modal representation, and a unique decoder \mathcal{D} produces the final prediction:

$$X^\dagger = \text{merge}^{(C)} \left(\{\text{interpolate}(X^m) \text{ to } T^\dagger\}_{m=1}^M \right) \quad (1)$$

$$y^{\text{early}} = \mathcal{D} \circ \mathcal{E}_{\text{spatio-temporal}}(X^\dagger). \quad (2)$$

Late Feature Fusion This fusion scheme starts by encoding each modality m separately with dedicated spatio-temporal encoders $\mathcal{E}_{\text{spatio-temporal}}^m$ into embeddings of size F^m . These vectors are then concatenated for all modalities along the channel dimension into a vector of size $\sum_m F^m$, which is ultimately mapped to a prediction y^{late} by a unique decoder \mathcal{D} :

$$y^{\text{late}} = \mathcal{D} \circ \text{merge}^{(C)} \left(\left\{ \mathcal{E}_{\text{spatio-temporal}}^m (X^m) \right\}_{m=1}^M \right), \quad (3)$$

with $\text{merge}^{(C)}$ the channelwise concatenation operator. While each latent features are derived from a single modality, this method allows the decoder to make decisions taking all modalities into account simultaneously.

Decision Fusion This approach ignores the interplay between modalities and makes a prediction for each modality independently. A set of M spatio-temporal encoder $\mathcal{E}_{\text{spatio-temporal}}^m$ maps each sequence of size $T^m \times C^m \times H \times W$ to a latent space of size F^m . Then, a set of M decoders \mathcal{D}^m maps each spatio-temporal feature into a prediction. Finally, an aggregation rule is applied to combine all M predictions into a final prediction y^{decision} . Typically, predictions are averaged across all available modalities :

$$y^{\text{decision}} = \frac{1}{M} \sum_{m=1}^M \mathcal{D}^m \circ \mathcal{E}_{\text{spatio-temporal}}^m (X^m). \quad (4)$$

Mid-Fusion Specific network architectures used to process temporal sequences such as SITS can be broken down into a spatial and a temporal encoder. In such cases, the spatial features can be interwoven, *i.e.* temporally stacked, into a single multimodal sequence, see Figure 3. This approach can be seen as a compromise between early and late fusion and combines three of their advantages: (i) the temporal encoder can leverage all modalities simultaneously, (ii) only one temporal encoder is needed, (iii) no heavy preprocessing is necessary to merge the feature sequences as they are stacked.

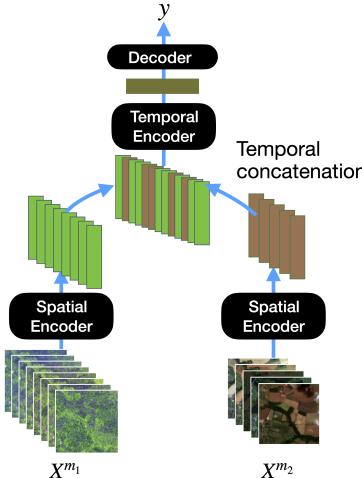


Figure 3: **Mid-Fusion.** A dedicated spatial encoder processes each modality, and the resulting features are stacked into a single sequence of features.

Each modality m has a dedicated spatial encoder $\mathcal{E}_{\text{spatial}}^m$ mapping images to a feature vector of size F^m . These vectors are then concatenated chronologically along the temporal dimension into a unique sequence of length $\sum_m T^m$. A unique temporal encoder $\mathcal{E}_{\text{temporal}}$ maps this sequence of features into a unique vector, which is in turn classified by a unique decoder \mathcal{D} :

$$y^{\text{mid}} = \mathcal{D} \circ \mathcal{E}_{\text{temporal}} \circ \text{merge}^{(T)} \left(\left\{ \mathcal{E}_{\text{spatial}}^m (X^m) \right\}_{m=0}^M \right), \quad (5)$$

with $\text{merge}^{(T)}$ the operator concatenating a set of tensors along the temporal dimension.

3.2 Auxiliary Supervision

We denote by $\text{criterion}(\cdot, \cdot)$ the function used to compare the prediction y with the target signal \hat{y} . This is typically the cross-entropy for parcel or pixel classification and can be more complex for panoptic or instance segmentation (Garnot and Landrieu, 2021). The resulting function \mathcal{L}_{obj} is called the objective loss and supervizes the prediction y of the network to realize the sought task:

$$\mathcal{L}_{\text{obj}} = \text{criterion}(y, \hat{y}) \quad (6)$$

A common problem in deep feature fusion is encountered when most (but not all) discriminative information is concentrated among a reduced number of modalities. In this case, the other modalities yield predictions and less relevant features for the considered task. Consequently, the final decision taken by the multimodal network focuses on the *better* modalities, and the parts of the network operating on the *lesser* modalities receive a weaker supervisory signal. This results in a network that may not fully leverage the inter-modal patterns that would otherwise allow the multimodal prediction to outperform the *best* modality. This is typically the case for Sentinel SITS, as multispectral optical acquisitions are often more conducive to capture phenological patterns than SAR information. Sentinel-1 signal is indeed affected by local terrain angle (Kaplan et al., 2021), humidity (Garkusha et al., 2017), and is subject to speckle (Abramov et al., 2017).

To mitigate this issue, we can add auxiliary losses to supervise each modality independently on top of the objective loss \mathcal{L}_{obj} . Ienco et al. (2019) has shown this strategy to help to combine optical and radar imagery. To this end, we associate a prediction y^m to each modality, which is supervised by the auxiliary loss \mathcal{L}_{aux} :

$$\mathcal{L}_{\text{aux}} = \sum_{m=1}^M \lambda^m \text{criterion}(y^m, \hat{y}), \quad (7)$$

with λ_m the strength associated with each modality. Note that, depending on the chosen fusion scheme, computing the single-modality prediction y^m may imply adding new modules to the backbone network. This requires M decoders \mathcal{D}^m , in the case of late fusion. For mid-fusion, we must add M temporal encoders $\mathcal{E}_{\text{temporal}}^m$ as well. No additional modules are necessary for decision fusion as single-modality predictions y^m are already necessary to produce the final prediction y . In contrast, auxiliary supervision in the case of early fusion would amount to duplicating the entire network, making it both fruitless and costly.

Since the added modules do not participate in the multimodal decision, only the gradients they propagate are beneficial to the training and not their predictions. Hence, auxiliary supervision only benefits the modules that receive gradients from both the main and auxiliary losses. Consequently, auxiliary loss with early supervision would not affect the performance. By the same reasoning, we can expect auxiliary supervision to be beneficial for late and decision fusion than mid-fusion for which only the spatial decoders are affected by the auxiliary losses.

3.3 Temporal Dropout

We propose a simple data augmentation strategy called temporal dropout to promote a multimodal model that leverages all available modalities. Inspired by the classic dropout strategy (Srivastava et al., 2014), we randomly drop observations from the input sequences. The idea is to prevent the network from over-relying on a single modality since its presence is never assured. Formally, we associate a dropout probability $p^m \in [0, 1]$ for each modality $m \in [1, M]$. During training, each observation of the sequence is dropped with probability p^m . At inference time, the network can use all available observations. Note that this technique can also be used on models operating on a single modality by randomly dropping some acquisitions.

3.4 Implementation.

As we set out to evaluate the benefit of multimodality for several tasks, we detail how our fusion schemes can be integrated into temporal attention-based, state-of-the-art networks.

Parcel-based classification We first implement the different fusion strategies for parcel-based crop type classification. In this setting, the contour of parcels is known, and the task is to classify the cultivated crop from a corresponding yearly SITS. The spatial and temporal encoding modules we selected are the Pixel-Set Encoder (PSE) and Lightweight Temporal Attention Encoder (L-TAE), whose accuracy and computational efficiency have been solidified in recent studies (Schneider and Körner, 2020; Kondmann et al., 2021; Garnot and Landrieu, 2020; Garnot et al., 2020), and whose implementations are available.¹ All spatio-temporal encoders $\mathcal{E}_{\text{spatio-temporal}}$ are a combination of a PSE encoding all images of the time series simultaneously and an L-TAE processing the resulting sequence of embeddings, in the manner of Garnot et al. (2020). All decoders \mathcal{D} are simple Multi-Layer Perceptrons (MLP). All models are trained with the cross-entropy loss. Since we are using these networks in a straightforward manner and without parameter alteration, we refer the reader to the (Garnot et al., 2020, Sec. 3) for more details on their configuration. As explained in Section 3.2, auxiliary supervision does not affect the early supervision beyond an increased memory load, and we do not evaluate it.

Semantic segmentation In this setting, the contours of the parcels are unknown, and the model predicts a crop type for each pixel of a given patch from the corresponding yearly SITS. For this task, we use the state-of-the-art U-TAE architecture (Garnot and Landrieu, 2021) as spatio-temporal encoder². The fact that this network’s spatial and temporal encoders are intertwined prevents us from applying the mid-fusion scheme. We use a 2-layer convolutional neural net as decoder \mathcal{D} . The models are trained with cross-entropy loss.

Panoptic segmentation Panoptic segmentation amounts to retrieving both the boundary and crop type of each agricultural parcel. In practice, we predict a set of non-overlapping instance masks and their semantic labels (Kirillov et al., 2019). In our setting, pixels that do not belong to a predicted parcel are classified as background. For this task, we also use U-TAE for spatio-temporal encoding, combined with the instance segmentation module Parcel-as-Points (PaPs) and its associated loss function for supervision (Garnot and Landrieu, 2021).² Averaging instance masks predicted by different modality-specific modules is not straightforward and costly in terms of memory, hence we do not evaluate the decision-fusion scheme for this task.

4 Experiments

We present in this section our numerical experiments to assess the benefit of multimodality for crop mapping with temporal attention-based networks. We evaluate several modality-fusion schemes and several mapping tasks. We also introduce a new large-scale open-access and multimodal dataset with annotations fit for all tasks.

4.1 Pastis-R

To evaluate the benefit of multimodality, we extend the open-access PASTIS dataset (Garnot and Landrieu, 2021) with corresponding Sentinel-1 observations. PASTIS is composed of 2433 time series of multi-spectral patches sampled in four different regions of France. Each patch has a spatial extent of $1.28\text{km} \times 1.28\text{km}$ and contains all available Sentinel-2 observations for the 2019 season for a total of 115k images. Note that PASTIS does not filter out observations with high cloud cover, hence, certain patches can be partially or entirely obstructed by clouds.

We use Sentinel-1 in Ground Range Detected format processed into σ_0 backscatter coefficient in decibels, orthorectified at a 10m spatial resolution with Orfeo Toolbox (Christophe et al., 2008). We do not apply any spatial or temporal speckle filtering, nor radiometric terrain correction: following the deep learning paradigm, we limit data preprocessing to the minimum. We assemble each Sentinel-1 observation into a 3-channel image: vertical polarization (VV), horizontal polarisation (VH), and the ratio of vertical over horizontal polarization (VV/VH). We separate observations made in ascending and descending orbit into two distinct time series. Indeed, the incidence angle of space-borne radar can significantly influence the return signal (Singhroy and Saint-Jean, 1999). As

¹[VSainteuf/lightweight-temporal-attention-pytorch](https://github.com/VSainteuf/lightweight-temporal-attention-pytorch)

²github.com/VSainteuf/utae-paps

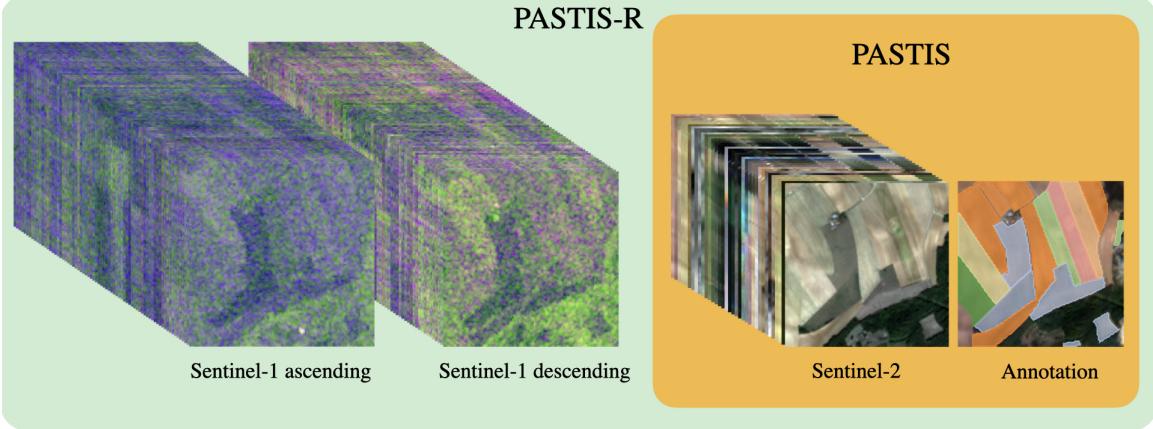


Figure 4: **Pastis-R.** We extend the PASTIS dataset with radar time series corresponding to ascending and descending orbits of Sentinel-1. For each square patch of $1.28\text{km} \times 1.28\text{km}$, PASTIS-R thus provides the image time series of 3 different modalities, along with semantic and instance annotation for each pixel.

represented in Figure 4, each time series comprises around 70 radar acquisitions for each of the 2433 patches. This amounts to a total of 339k added radar images. We use the annotations of PASTIS: semantic class and instance identifier for each pixel, allowing us to evaluate models for parcel-based classification, semantic segmentation, and panoptic segmentation. We make the PASTIS-R dataset (Garnot and Landrieu, 2021) publicly available at: github.com/VSainteuf/pastis-benchmark.

4.2 Implementation details

As detailed in Section 3.4, we use the official PyTorch implementations of PSE+LTAE and U-TAE with default hyperparameters. We use the official 5 cross-validation folds of PASTIS (Garnot and Landrieu, 2021) to evaluate the performance of the different models. We train our models using the Adam optimizer (Kingma and Ba, 2015) with default parameters $\text{lr} = 0.001$, $\beta = (0.9, 0.999)$ unless specified otherwise, and train all networks on a single TESLA V100 GPU with 32Gb of VRAM.

Multimodality Configuration We consider the two orbits of Sentinel-1 as separate modalities to account for their difference in incident angle, which corresponds to $M = 3$. When using auxiliary loss terms, we set $\lambda^m = 0.5$ for all modalities. When using temporal dropout, we set $p_0 = 0.4$ for the optical modality and $p_1 = p_2 = 0.2$ for the radar time series. For early fusion, we interpolate the Sentinel-1 observations to the dates of the Sentinel-2 time series. Indeed, the opposite interpolation strategy would imply tripling the temporal length of the Sentinel-2 time series, which would significantly increase memory usage. Interpolation is computed on the fly when loading dataset samples.

Parcel Classification For this problem, we train the models for 100 epochs in batches of 128 parcels. We use the $K = 18$ class nomenclature of PASTIS and report the classification Intersection over Union macro-averaged over the class set (mIoU) to evaluate the parcel-level predictions.

Semantic Segmentation We train the semantic segmentation models for 100 epochs in batches of 4 multi-temporal patches. In this setting, the models also predict *background* pixels, resulting in a $K = 19$ class nomenclature. We report the mIoU of the pixel-level predictions:

$$\text{mIoU} = \frac{1}{K} \sum_{k=1}^K \frac{TP_k}{TP_k + FP_k + FN_k}, \quad (8)$$

with TP_k , FP_k , and FN_k the count of true positives, false positives, and false negatives for the binary class predictions defined by a class k .

Panoptic Segmentation We follow the training procedure recommended by Garnot and Landrieu (2021) to train the PaPs network: the learning rate starts at 0.01 for the first 50 epochs, and decreases to 0.001 for the last 50 epochs. We report the class-averaged panoptic metrics introduced in Kirillov et al. (2019): Segmentation Quality (SQ), Recognition Quality (RQ), and Panoptic Quality (PQ). The RQ corresponds to the F_1 score for the problem of combined detection and classification: to be counted as a true positive, a parcel must be both detected (the intersect over union of the predicted and true instance masks is above 0.5) and have its crop type correctly classified. The SQ corresponds to the intersect over union between the true and predicted masks for correctly detected and classified parcels. Finally, the PQ is the product of both values, thus simultaneously combining information on the quality of detection, classification, and delineation. We report the unweighted classwise average of the three quality measurements. We refer the reader to Kirillov et al. (2019) for more details on these metrics.

4.3 Parcel Classification Experiment

We first implement and evaluate the different fusion schemes and their training enhancements for parcel classification. In this setting, the contour of parcels is known in advance, and the model predicts the type of crop cultivated during the period covered by the SITS.

Table 1: **Parcel Classification.** We evaluate the performance of models operating on a single modality (top) and for different fusion strategies for parcel-based classification (bottom). We evaluate each model’s baseline performance and the impact of the temporal dropout and auxiliary classifiers enhancements, when applicable. We report the 5-fold cross-validated classification scores in terms of mean classwise Intersect over Union, the base model’s parameter count, and, when relevant, of the model with auxiliary classifiers.

	Base		Temp. dropout	Auxiliary supervision	Auxiliary & Temp. dropout	Parameter Count
	OA	mIoU				
S2	91.7	73.9	74.5	-	-	114k
S1D	87.0	64.5	64.7	-	-	114k
S1A	86.4	63.3	62.9	-	-	114k
Early Fusion	91.8	74.9	76.5	-	-	117k
Mid Fusion	92.0	75.1	75.9	75.0	76.5	152k/185k
Late Fusion	91.1	73.0	73.6	76.1	77.2	254k/287k
Decision Fusion	91.0	72.5	72.8	75.2	75.8	259k

Analysis In Table 1, we report the performance of all fusion schemes with and without enhancements. We first observe that the optical satellite S2 significantly outperforms the two radar time series by a margin of almost 10 points of mIoU, confirming the relevance of Sentinel-2 for crop type mapping. We remark that, without enhancement (first column), multimodal models trained with early or mid-fusion schemes improve the performance compared to single optical modality networks, while decision and late fusion perform slightly worse consistently with the results of Ofori-Ampofo et al. (2021). This highlights the benefit of learning to mix modality features early on. In contrast, auxiliary supervision and temporal dropout improve the later models. This shows that these enhancements can encourage attention-based models to combine features and decisions efficiently from different modalities, as observed in Ienco et al. (2019) for recurrent networks. All things considered, late fusion with both enhancements performs best with +3.3 mIoU compared to a network operating purely on the optical modality, see Figure 5 for a classwise comparison. Mid-fusion without enhancement provides good performances with a lower parameter count and none of the preprocessing necessary for early fusion. In practice, the mid-fusion scheme is 20% faster at inference time than late fusion, making it a valid choice when operating with limited computational resources.

Auxiliary Supervision and Gradient Flow Motivated by the impact of auxiliary supervision on the performance of the late fusion approach, we propose to study its effect on the learning process further. Specifically, we wish to evaluate the different spatio-temporal encoders’ contribution to the

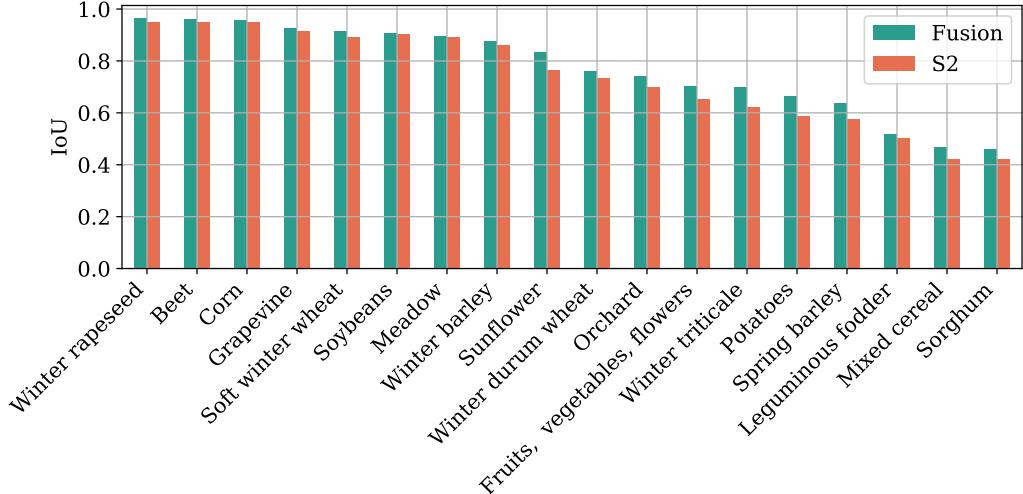


Figure 5: **Classwise Performance for Parcel Classification.** We report the IoU of the late fusion model with auxiliary supervision and temporal dropout and of the model trained purely on the optical modality. Multimodality brings a consistent benefit across all classes, which is more notable for some of the most challenging classes such as *Potatoes*, or *Winter triticale*.

reduction of the objective loss \mathcal{L}_{obj} , with and without auxiliary supervision, and for the parcel classification task. Note that, as auxiliary decisions are not computed at inference time, we only consider the decrease of \mathcal{L}_{obj} : a decrease in the auxiliary losses does not directly affect the model’s performance.

Following the insights of Wang et al. (2020), we consider the following first-order approximation of the decrease of \mathcal{L}_{obj} incurred by taking a gradient step:

$$\Delta \mathcal{L}_{\text{obj}} = \eta \langle \nabla \mathcal{L}, \nabla \mathcal{L}_{\text{obj}} \rangle , \quad (9)$$

with η the current learning rate. The term $\nabla \mathcal{L}$ of the scalar product in (9) corresponds to the step size in the gradient descent and the term $\nabla \mathcal{L}_{\text{obj}}$ to the slope of the objective loss. Their scalar product approximates the decrease in objective loss when taking a single gradient step. Note that this approximation, called gradient flow, is only valid when using stochastic gradient descent (SGD) and does not hold for momentum or adaptive optimization schemes such as ADAM (Kingma and Ba, 2015). We thus retrain the late fusion model with SGD for parcel classification. By considering each term in the scalar product in Equation 9, we can estimate the contribution of each parameter of the network to the decrease of the objective loss \mathcal{L}_{obj} .

In Figure 6, we represent the evolution of the gradient flow for different modules of our architecture by summing the contribution of their corresponding parameters. We observe that, as expected, the gradient flow is concentrated in the modules dedicated to the optical modality. Interestingly, the spatial encoders contribute as much or even more than the temporal encoders despite having four times fewer parameters.

We remark that auxiliary losses lead the model to a different training regime. While auxiliary supervision results in an increase of the proportion of gradient flow in some radar modules such as PSE-S1A, the flow also increases in proportion in some optical modules as well. We conclude that auxiliary supervision affects all modalities, not only the weaker modalities.

4.4 Semantic Segmentation

In this section, we evaluate the performance of the late fusion scheme compared to single modality baselines for semantic segmentation. While the mid-fusion scheme yields promising results on parcel-based experiments, its implementation into a semantic segmentation architecture is not trivial. Indeed, the state-of-the-art network for this task (Garnot and Landrieu, 2021) relies on a U-Net architecture with temporal encoding. In this architecture, spatial and temporal encoding are

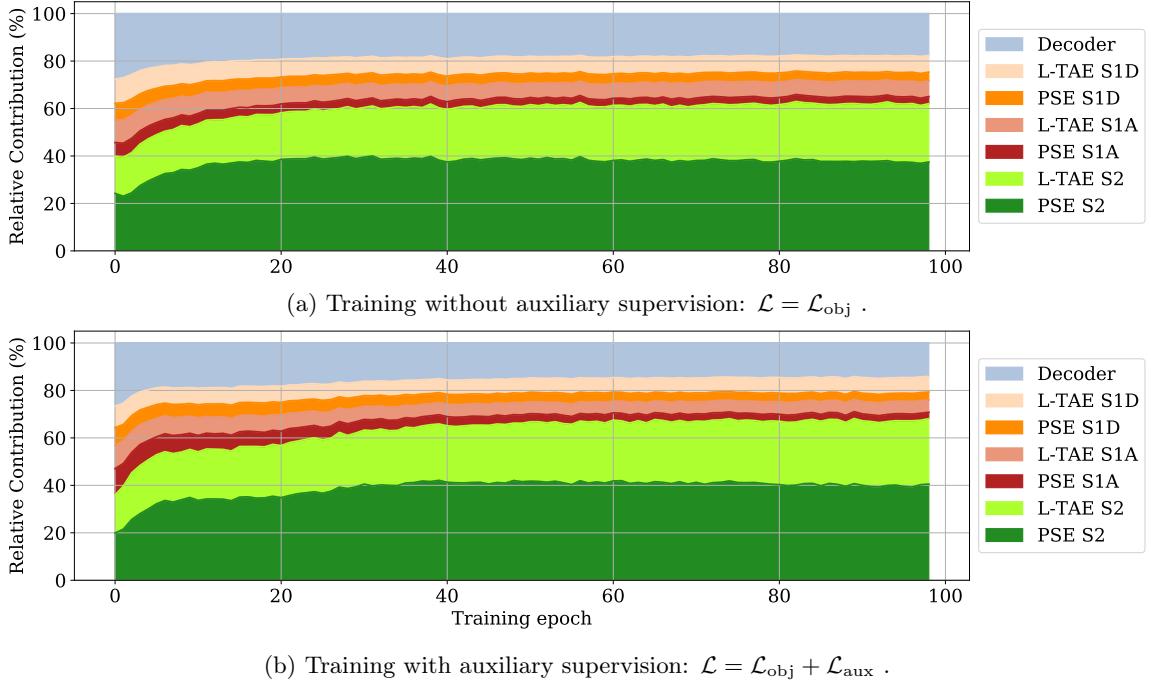


Figure 6: **Gradient Flow.** Evolution of the gradient flow for different modules of the late fusion model. The contribution of each modality is plotted as a fraction of the total flow, without auxiliary loss terms (top) and with the additional \mathcal{L}_{aux} term (bottom). We report the flow for the spatial encoders (PSE), temporal encoders (LTAE), and the MLP-based decoders.

performed conjointly. After several unsuccessful attempts, we limit our study to the other fusion schemes for this task.

Table 2: **Semantic Segmentation Experiment.** We evaluate the semantic segmentation performance of models operating on a single modality and multimodal models trained with early, late, and decision fusion strategies. We evaluate each model’s baseline performance and the impact of temporal dropout and auxiliary classifiers, when applicable. We report the 5-fold cross-validated classification scores in mean classwise Intersect over Union (- not applicable). Note that temporal dropout is necessary for the late and decision fusion models to fit in memory.

	Base	Temporal dropout	Auxiliary & Temporal dropout	Parameter Count
S2	63.1	63.6	-	1 087k
S1D	54.9	54.7	-	1 083k
S1A	53.8	53.3	-	1 083k
Early Fusion	64.9	65.8	-	1 602k
Late Fusion	-	65.8	66.3	1 709k
Decision Fusion	-	64.7	64.3	1 742k

Analysis We report the performance of the different models in Table 2. In our experimental setup, the late fusion model with over ~ 200 total multimodal observations did not fit in the 32Gb of memory of our GPU with a batch size of 4 image time series. By reducing the size of the input sequences, temporal dropout allowed us to train this memory-intensive model. The late fusion model improves the performance of the unimodal models by 2.7 mIoU points. The performance is further improved by another 0.5 point with the addition of auxiliary supervision. The early fusion model performs slightly below late fusion, even with temporal dropout. As represented in Figure 7,

the radar modality allows for prediction with crisper contours, in particular between adjacent or nearly adjacent parcels. This suggests that the image rugosity of the radar acquisitions is can be valuable to detect inter-parcel zones. These areas, often of sub-pixel extent, may display optical reflectances similar to their neighboring parcels but often present surfaces such as fences or groves with a volumetric scatter and thus a distinct radar response .

Note that the performance of our models on semantic segmentation is around 10pts below that for parcel classification. This result was expected as the semantic segmentation task prevents us from exploiting knowledge about the contour of parcels and has the supplementary class *background*, corresponding to non-agricultural land.

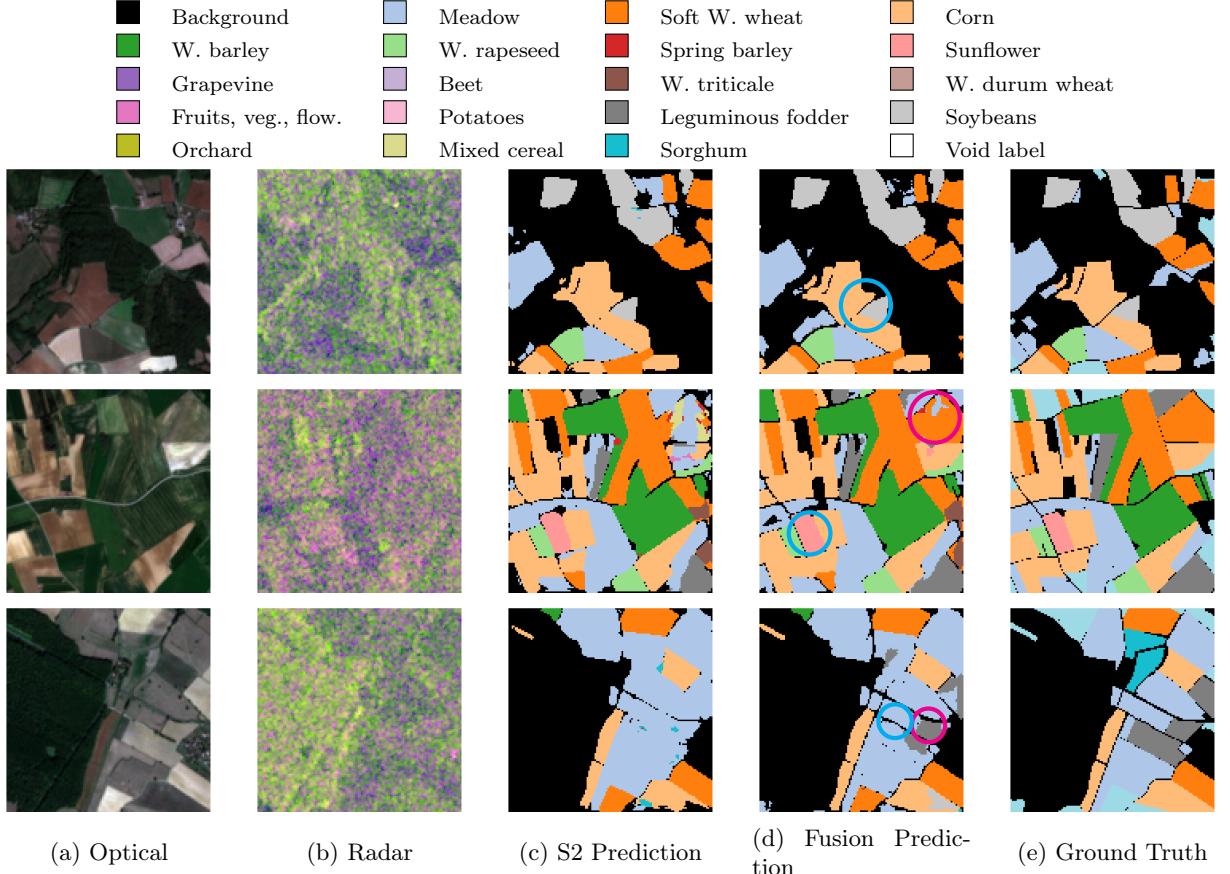


Figure 7: Qualitative Results for Semantic Segmentation. We show one observation from the optical time series in (a) and one from the radar time series in (b). The prediction for the unimodal optical model is represented in (c), aour late fusion multimodal model in (d), and finally the ground truth in (e). We observe that the multimodal model produces results with clearer and more distinct borders between close parcels (cyan circle \circlearrowleft). The multimodal model also displays fewer errors for hard and ambiguous parcels, showing the benefit of learning intermodal features (magenta circle \circlearrowright). Crop types are represented according to the color code above (W. stands for Winter). The same legend is used in all subsequent figures representing crop labels.

Varying Cloud Cover Experiment One of the motivations for using both optical and radar images in the context of crop type mapping is to exploit the imperviousness of radar signals to cloud cover. This potentially allows our model to rely on the radar signal when optical observations are obstructed by clouds, which is particularly crucial in countries with pervasive cloud cover, such as subtropical regions (Orynbaikyzy et al., 2020). The parcel-based and semantic experiments allow for a first exploration of this capacity, but remain bound to the specific cloud conditions of the French metropolitan territory and the year of acquisition (2019). We propose to further investigate this benefit of multimodality by artificially simulating increased cloud obstruction on the test set. To do

so, we evaluate the performance of models when removing random optical acquisitions while leaving the radar time series unchanged. We report the performance of the models in Figure 8, for different ratios of remaining optical observations, corresponding to different levels of cloud obstruction.

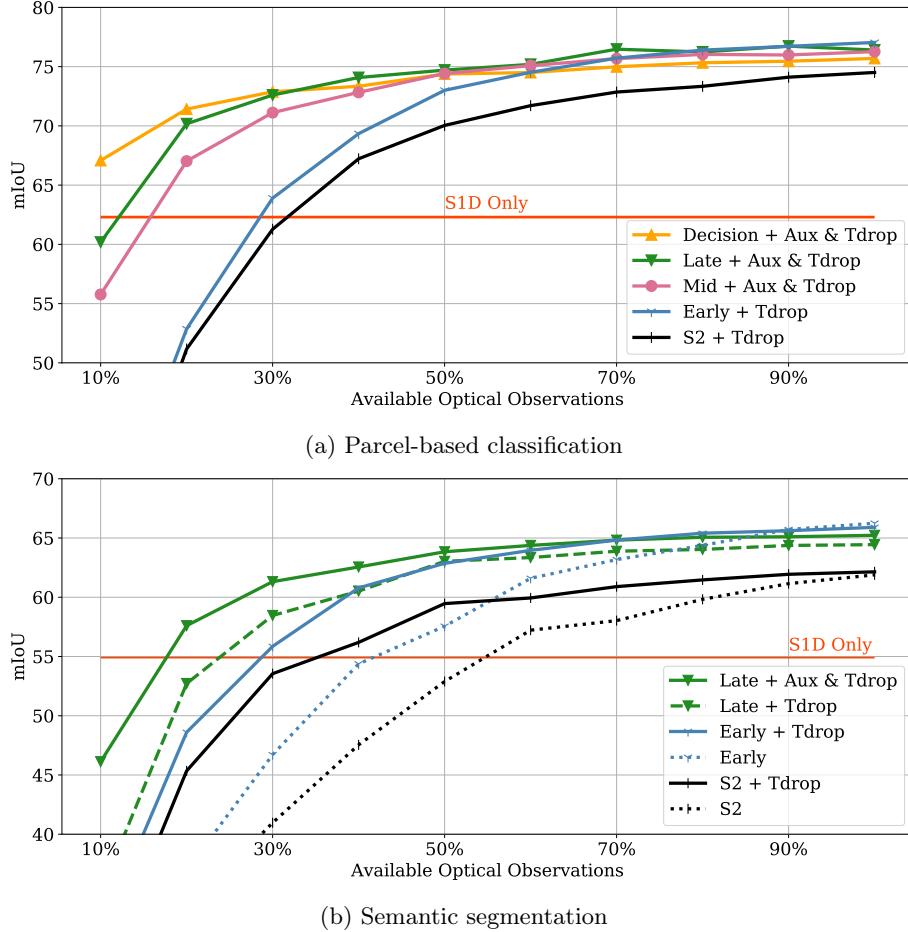


Figure 8: Varying Cloud Cover Experiment. We evaluate the different models with varying ratios optical observations remaining. In both parcel-based classification (a) and semantic segmentation (b), the fusion models prove more robust to a reduced number of optical observations.

As expected, the performance of the S2-only model drops drastically as the number of available optical observations decreases for both parcel classification and semantic segmentation, performing worse than unimodal radar models for a ratio of 70% of artificial occlusion. Multimodal fusion models can maintain an almost constant level of performance for up to 50% missing optical acquisitions. For more extreme ratios, the performances of the multimodal models eventually drop. The magnitude of the drop seems to be related to the amount of interplay between modalities in the network. Early fusion proves the least robust to missing optical observations. Mid-fusion, and to a lesser extent, the late fusion are also affected by obstruction. These models rely on multimodal encoders and decoders, which are likely to be affected by a severe decrease in the quality of the optical sequence. In contrast, the decision fusion scheme comprises independent classifiers and proves to be the most resilient: even with 90% of optical images removed, it still outperforms the radar modality by several mIoU points. We conclude that such models should be favored in regions with pervasive or inconsistent cloud cover.

We also observe that auxiliary supervision and temporal dropout make both unimodal and multimodal models more resilient to missing optical acquisitions for semantic segmentation. The same phenomenon can be observed for parcel classification but was not represented for clarity.

Table 3: **Panoptic Segmentation Experiment.** We evaluate the panoptic segmentation performance of models operating on a single modality and multimodal models trained with the early and late fusion strategy with temporal dropout.

	SQ	RQ	PQ	Parameter count
S2	81.3	49.2	40.4	1 318k
S1D	77.0	39.3	30.9	1 318k
S1A	77.4	38.8	30.6	1 318k
Early Fusion + Tdrop	82.2	50.6	42.0	1 791k
Late Fusion + Tdrop	81.6	50.5	41.6	2 390k

4.5 Panoptic segmentation experiment

In this section, we evaluate the performance of the early and late fusion schemes compared to single modality baselines for panoptic segmentation. We do not assess auxiliary losses on the late fusion model as the use of auxiliary decoders in this setting comes at a prohibitive computational cost. Indeed, the auxiliary decoders would be PaPs instance segmentation modules which already significantly impact training times on single modality architectures. Decision fusion is not evaluated here for the same reason. Like in the semantic segmentation experiment, temporal dropout proved necessary to train the late fusion model.

Analysis We report the results of this experiment on Table 3. Overall, the early and late fusion schemes increase the panoptic quality by 1.6pt and 1.2pt, respectively, compared to the optical baseline. This improvement is mostly driven by an increase in recognition quality, while the segmentation quality remains almost unchanged. This suggests that the radar modality helps correctly detect additional agricultural parcels rather than refining the delineation of their boundaries. Although modest, this improvement is valuable for this notoriously complex task.

We show on Figure 9 the qualitative evaluation of the panoptic fusion model compared to the optical baseline. In practice, the fusion model seems to retrieve more agricultural parcels successfully and manages to retrieve small parcels missed by the optical model. We also display the predictions made by the unimodal models and the predictions of the fusion model in Figure 11. These qualitative results show how the radar modality helps detect more parcels than the optical baseline or improve the fusion model’s semantic predictions. Additionally, given the relative noisiness of radar observations, the radar-only models retrieve surprisingly well the parcel boundaries. As mentioned previously, this could be attributed to the distinct volumetric radar response on parcel boundaries. We report the per-class performances on Figure 10.

Regarding robustness to clouds, when performing inference on only 30% of the optical observations, the S2 baseline model drops to 33.0 PQ. In contrast, our late fusion model maintains a score of 37.6 PQ. Consistently with the previous experiments, the addition of the radar modality helps improve the panoptic predictions with reduced availability of optical observations.

5 Discussion

In this section, we discuss the relevance of the different modality fusion strategies, with a focus on Sentinel-1 & 2 data for crop mapping. Our experiments showed that combining optical and radar imagery allowed for an increase in performance for all tasks considered (Table 1, Table 2, Table 3) as well as robustness to cloud cover (Figure 8).

5.1 Recommendations.

Our experiments showed that each fusion scheme has advantages and limitations influencing when its use is most relevant:

- **Early Fusion.** It is the most compact of the fusion models and shows competitive performance on all three tasks. The main drawback of this approach is the necessity of an expensive interpolation. As reported in Table 4, this preprocessing makes the early fusion scheme slower

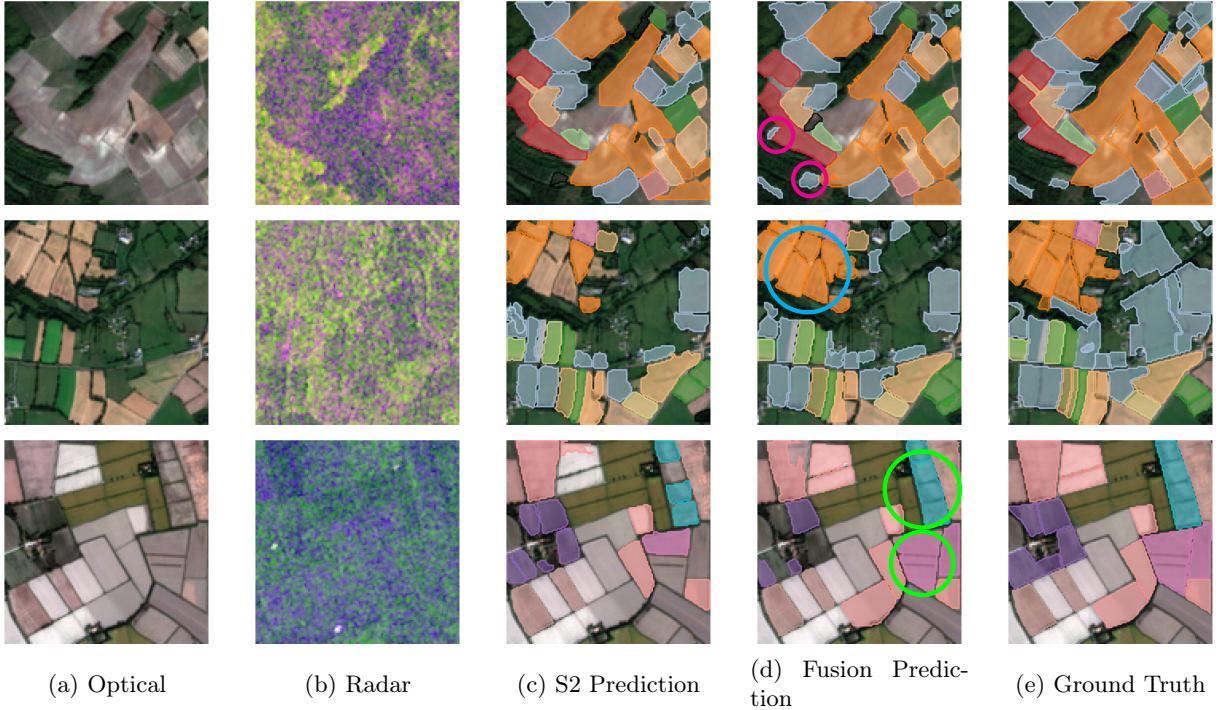


Figure 9: **Qualitative Results for Panoptic Segmentation.** We show one observation from the optical time series in (a) and one from the radar time series in (b). The prediction for the unimodal optical model is represented in (c) and the multimodal model in (d), and finally the ground truth in (e), with the same colormap as in Figure 7. The fusion model retrieves more parcels (cyan circle $\textcolor{cyan}{\circ}$), and even small parcels that were missed by the purely optical model (magenta circle $\textcolor{magenta}{\circ}$). We also note that the fusion model seems to handle parcels with internal subdivisions (green circle $\textcolor{green}{\circ}$) better than the optical model.

than late fusion despite relying on a smaller network for parcel classification and semantic segmentation. Early fusion is the least robust fusion scheme to cloud cover.

- **Mid Fusion.** Of all methods without preprocessing, this strategy leads to the fastest run time and the lowest memory requirement. It yields the second-best performance for parcel-based classification but suffers more than late and decision fusion when the cloud cover is extensive. Its dependence on separate spatial and temporal encoders prevents its straightforward adaptation to pixel-based tasks. We recommend using this scheme for parcel classification in areas without extensive cloud cover and when inference speed is critical.
- **Late Fusion.** This fusion method, when combined with enhancement schemes, leads to the best performance and the highest adaptability, as well as excellent resilience to even extreme cloud cover. This method is our default recommendation when using temporal attention methods with multimodal time series.
- **Decision Fusion.** Despite having the highest parameter count, this method lags in terms of performance and is prohibitively costly for panoptic segmentation. However, it is the most resilient to cloud cover. We recommend using decision fusion when it is expected that only a few optical observations may be available for inference.

We also have evaluated the influence of two enhancement schemes:

- **Auxiliary Supervision.** This method consists in adding alongside the main prediction auxiliary predictions based on one modality alone. The rationale is to help each specialized module to learn meaningful features regardless of the interplay with other modalities. We observe a strong effect in precision for late and decision fusion, which have dedicated encoding modules for each modality.

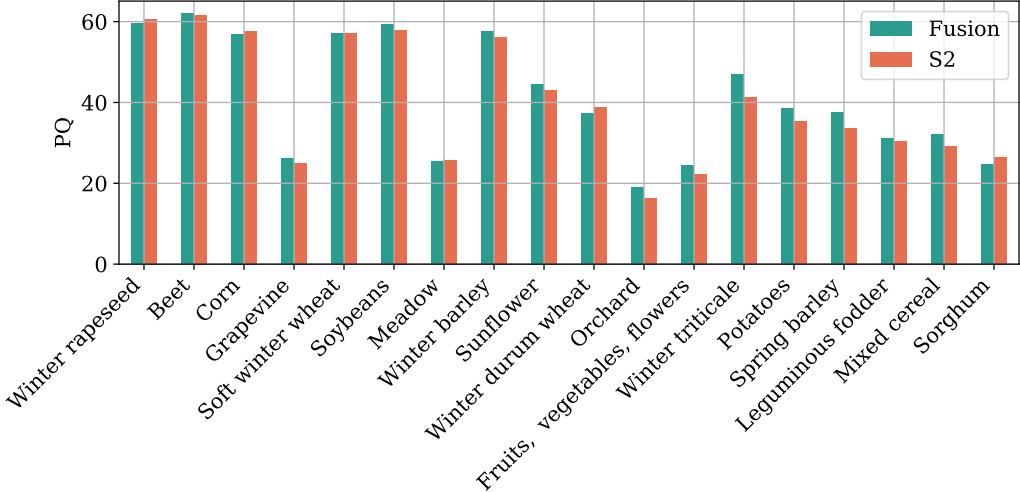


Figure 10: **Classwise Performance for Parcel Classification.** We report the Panoptic Quality of the late fusion model with temporal dropout and the model trained purely on the optical modality. The classes are ordered as in Figure 5. In the panoptic setting, the radar modality is also specifically beneficial for hard classes such as *Winter triticale*.

Table 4: **Inference times.** We report the inference times in seconds of Early and Late fusion for one fold of PASTIS (500 patches, 820km²). We measure the combined data loading and prediction time to account for the interpolation step in early fusion

	Parcel classification	Semantic segmentation	Panoptic segmentation
Early	192	280	414
Late	149*	259*	819

* with auxiliary loss.

- **Temporal Dropout.** This simple method consists in randomly dropping acquisitions of the time series considered. Its effect was beneficial to all fusion schemes and the optical baseline across our experiments. Another benefit of this scheme is that it reduces the memory footprint of networks during training.

5.2 Limitations.

Our study hinged on the PASTIS dataset, which contains annotated agricultural parcels from four different regions of the French metropolitan territory. In this regard, our results are most relevant for crop mapping applications with the same meteorological context, terrain conditions, and crop types as this region. Certain crop types not observed in PASTIS could benefit even more from the radar modality than our results show. For instance, rice fields are often filled with water and thus have a distinctive SAR response but are not represented in PASTIS.

Furthermore, our evaluation of cloud robustness focused on assessing the effect of a reduced number of optical observations *at inference time*. This corresponds to artificially increasing the cloud cover in the test set without affecting crop growth. A more rigorous approach would constitute a dataset comprising truly observed cloud coverage by varying the regions and years of acquisition. This is complicated by the lack of harmonization between LPIS across different countries in nomenclature and open-access policy. Lastly, we only used backscattering coefficients from the SAR data in our experiments, as is commonly done in the crop type mapping literature (Orynbaiyzy et al., 2019). Mestre-Quereda et al. (2020) found that the addition of interferometric radar features is beneficial to crop classification when using only radar inputs. Further work is needed to assess the benefit of interferometric radar features in a fusion setting with optical imagery. Moreover, we chose to prepare the SAR inputs with limited preprocessing. We do not apply speckle filtering or

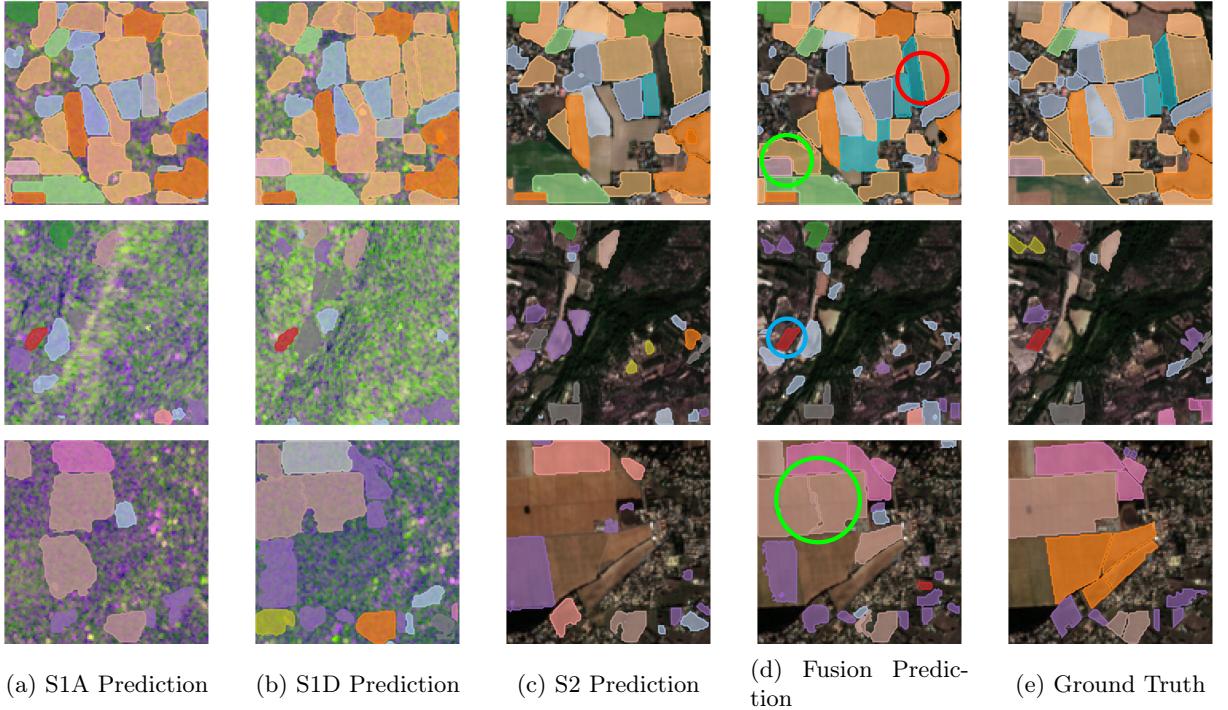


Figure 11: **Qualitative Results for Panoptic Segmentation.** We compare the predictions made by unimodal models operating on S1A (a), S1D (b), S2 (c), and the predictions made by the late fusion model (d). We also show the ground truth annotations (e). We observe cases where the optical model does not detect parcels but successfully predicted by the radar-only models and by the fusion model as well (green circle $\textcolor{green}{\circ}$). We also note that the optical model detects some parcels, but the crop type is corrected by the addition of the radar modality (red circle $\textcolor{red}{\circ}$). Conversely, some parcels are detected by the radar-only model with an incorrect crop type and not detected by the optical model. Combining both modalities in the fusion model leads to a correct prediction. (cyan circle $\textcolor{cyan}{\circ}$)

radiometric terrain correction to compensate for the effect of the local incident angle. Interestingly, our experiments showed that this does not prevent the radar modality from benefiting crop mapping models. However, further studies could evaluate the benefit of adding speckle filtering, elevation information, or meteorological context to networks using radar images for crop mapping.

6 Conclusion

This article formulated and explored different schemes to design fusion architectures using temporal attention for predicting agricultural crop type maps from radar and optical satellite time series.

Across the three tasks of parcel-based classification, semantic segmentation, and panoptic segmentation, we experimentally confirmed that the multispectral information of Sentinel-2 proves more discriminative than the SAR signal of Sentinel-1. Yet, for all three tasks, leveraging both modalities led to improvements in the overall performance and the robustness to cloud obstruction. The late fusion scheme, where the learned representations of each modality are concatenated before decoding, outperformed the other approaches on parcel-based classification. Our subsequent exploration of this approach on semantic and panoptic segmentation confirmed its validity to leverage optical and radar time series for crop type mapping. Our experiments also showed that models with less interplay in the encoding of the modalities are most robust to changes in cloud obstruction. In this regard, decision fusion may be favored in contexts with highly unpredictable cloud conditions. Yet, both late and decision fusion approaches proved computationally costly as they incur distinct spatio-temporal encoders for each modality. We introduced a mid-fusion scheme that circumvents this problem by using separate spatial encoders and a shared temporal encoder. This approach

performed marginally worse than late fusion on parcel-based classification while having close to half the trainable parameters. Mid-fusion can thus be a valid choice for applications with limited computational resources. Furthermore, the extension of this approach to semantic and panoptic segmentation should be explored in future works. We release PASTIS-R, the augmented version of PASTIS with radar time series, to encourage further endeavors in multi-temporal fusion for Earth Observation.

Acknowledgements We thank Hugo Lecomte and Nicolas David for their help in the preparation of the PASTIS-R dataset. This work was partly supported by ASP, the French Payment Agency.

References

- V. S. F. Garnot, L. Landrieu, Pastis - panoptic segmentation of satellite image time series, 2021.
URL: <https://zenodo.org/record/5012942>. doi:10.5281/ZENODO.5012942.
- K. Van Tricht, A. Gobin, S. Gilliams, I. Piccard, Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: a case study for Belgium, *Remote Sensing* (2018).
- M. J. Steinhausen, P. D. Wagner, B. Narasimhan, B. Waske, Combining Sentinel-1 and Sentinel-2 data for improved land use and land cover mapping of monsoon regions, *International journal of applied earth observation and geoinformation* (2018).
- M. Campos-Taberner, F. J. García-Haro, B. Martínez, S. Sánchez-Ruiz, M. A. Gilabert, A Copernicus Sentinel-1 and Sentinel-2 classification framework for the 2020+ European common agricultural policy: A case study in València (Spain), *Agronomy* (2019).
- A. Vrieling, M. Meroni, R. Darvishzadeh, A. K. Skidmore, T. Wang, R. Zurita-Milla, K. Oosterbeek, B. O'Connor, M. Paganini, Vegetation phenology from sentinel-2 and field cameras for a dutch barrier island, *Remote sensing of environment* (2018).
- J. Segarra, M. L. Buchaillot, J. L. Araus, S. C. Kefauver, Remote sensing for precision agriculture: Sentinel-2 improved features and applications, *Agronomy* (2020).
- C. J. Tucker, Red and photographic infrared linear combinations for monitoring vegetation, *Remote Sensing of Environment* (1979).
- M. Sudmanns, D. Tiede, H. Augustin, S. Lang, Assessing global sentinel-2 coverage dynamics and data availability for operational earth observation (eo) applications using the eo-compass, *International Journal of Digital Earth* (2020).
- H. McNairn, A. Kross, D. Lapen, R. Caves, J. Shang, Early season monitoring of corn and soybeans with terrasar-x and radarsat-2, *International Journal of Applied Earth Observation and Geoinformation* 28 (2014).
- E. Commission, The common agricultural policy at a glance, ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/cap-glance_en, 2016. Accessed: 2021-09-24.
- B. Koetz, P. Defourny, S. Bontemps, K. Bajec, C. Cara, L. de Vendictis, L. Kucera, P. Malcorps, G. Milcinski, L. Nicola, et al., Sen4cap sentinels for cap monitoring approach, in: JRC IACS Workshop, 2019.
- M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al., Sentinel-2: Esa's optical high-resolution mission for gmes operational services, *Remote sensing of Environment* (2012).
- EtaLab, Registre parcellaire graphique (rpg) : contours des parcelles et îlots culturaux et leur groupe de cultures majoritaire, data.gouv.fr/en/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-ilots-culturaux-et-leur-groupe-de-cultures-majoritaire/, 2017.

- W. He, N. Yokoya, Multi-temporal Sentinel-1 and-2 data fusion for optical image simulation, *ISPRS Journal* (2018).
- A. Orynbaikyzy, U. Gessner, B. Mack, C. Conrad, Crop type classification using fusion of Sentinel-1 and Sentinel-2 data: Assessing the impact of feature selection, optical data availability, and parcel sizes on the accuracies, *Remote Sensing* (2020).
- S. Giordano, S. Bailly, L. Landrieu, N. Chehata, Improved crop classification with rotation knowledge using sentinel-1 and-2 time series, *Photogrammetric Engineering & Remote Sensing* (2020).
- D. Ienco, R. Interdonato, R. Gaetano, D. H. T. Minh, Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture, *ISPRS Journal* (2019).
- M. Rußwurm, M. Körner, Self-attention for raw optical satellite time series classification, *ISPRS Journal* (2020).
- V. S. F. Garnot, L. Landrieu, S. Giordano, N. Chehata, Satellite image time series classification with pixel-set encoders and temporal self-attention, in: *CPVR*, 2020.
- L. Kondmann, A. Toker, M. Rußwurm, A. C. Unzueta, D. Peressuti, G. Milcinski, P.-P. Mathieu, N. Longépé, T. Davis, G. Marchisio, et al., Denethor: The dynamicearthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space (2021). URL: openreview.net/forum?id=uUa4jNMLjrL.
- V. S. F. Garnot, L. Landrieu, Lightweight temporal self-attention for classifying satellite images time series, in: *AALTD*, 2020.
- S. Ofori-Ampofo, C. Pelletier, S. Lang, Crop type mapping from optical and radar time series using attention-based deep learning, *Remote Sensing* (2021).
- A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic segmentation, in: *CVPR*, 2019.
- V. S. F. Garnot, L. Landrieu, Panoptic segmentation of satellite image time series with convolutional temporal attention networks, in: *ICCV*, 2021.
- N. Joshi, M. Baumann, A. Ehamer, R. Fensholt, K. Grogan, P. Hostert, M. R. Jepsen, T. Kuemmerle, P. Meyfroidt, E. T. Mitchard, et al., A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring, *Remote Sensing* (2016).
- A. Mercier, J. Betbeder, F. Rumiano, J. Baudry, V. Gond, L. Blanc, C. Bourgoin, G. Cornu, M. Marchamalo, R. Poccard-Chapuis, et al., Evaluation of Sentinel-1 and 2 time series for land cover classification of forest-agriculture mosaics in temperate and tropical landscapes, *Remote Sensing* (2019).
- A. Tarpanelli, E. Santi, M. J. Tourian, P. Filippucci, G. Amarnath, L. Brocca, Daily river discharge estimates by merging satellite optical sensors and radar altimetry through artificial neural network, *IEEE Transactions on Geoscience and Remote Sensing* (2018).
- N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov, Deep learning classification of land cover and crop types using remote sensing data, *Geoscience and Remote Sensing Letters* (2017).
- P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, S. Dupuy, M3Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion, *JSTARS* (2018).
- M. Tom, Y. Jiang, E. Baltsavias, K. Schindler, Learning a sensor-invariant embedding of satellite data: A case study for lake ice monitoring, *arXiv preprint arXiv:2107.09092* (2021).
- J. Liu, M. Gong, K. Qin, P. Zhang, A deep convolutional coupling network for change detection based on heterogeneous optical and radar images, *Transactions on neural networks and learning systems* (2016).

- A. Garioud, S. Valero, S. Giordano, C. Mallet, On the joint exploitation of optical and SAR satellite imagery for grassland monitoring, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (2020).
- A. Meraner, P. Ebel, X. X. Zhu, M. Schmitt, Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion, *ISPRS Journal* (2020).
- J. A. Richards, et al., *Remote sensing with imaging radar*, volume 1, Springer, 2009.
- A. Orynbaikyzy, U. Gessner, C. Conrad, Crop type classification using a combination of optical and radar remote sensing data: a review, *International Journal of Remote Sensing* (2019).
- M. Simons, P. Rosen, Interferometric synthetic aperture radar geodesy, *Treatise on Geophysics - Geodesy* (2007).
- O. Monserrat, M. Crosetto, G. Luzi, A review of ground-based sar interferometry for deformation measurement, *ISPRS Journal of Photogrammetry and Remote Sensing* (2014).
- D. Tarchi, N. Casagli, R. Fanti, D. D. Leva, G. Luzi, A. Pasuto, M. Pieraccini, S. Silvano, Landslide monitoring by using ground-based sar interferometry: an example of application to the tessina landslide in italy, *Engineering geology* 68 (2003) 15–30.
- R. Tomás, J. García-Barba, M. Cano, M. P. Sanabria, S. Ivorra, J. Duro, G. Herrera, Subsidence damage assessment of a gothic church using differential interferometry and field data, *Structural Health Monitoring* (2012).
- D. Tarchi, E. Ohlmer, A. Sieber, Monitoring of structural changes by radar interferometry, *Journal of Research in Nondestructive Evaluation* (1997).
- C. Tison, F. Tupin, H. Maître, A fusion scheme for joint retrieval of urban height map and classification from high-resolution interferometric sar images, *IEEE Transactions on Geoscience and remote Sensing* 45 (2007) 496–505.
- T. Tamm, K. Zalite, K. Voormansik, L. Talgre, Relating sentinel-1 interferometric coherence to mowing events on grasslands, *Remote Sensing* (2016).
- A. Mestre-Quereda, J. M. Lopez-Sanchez, F. Vicente-Guijalba, A. W. Jacob, M. E. Engdahl, Time-series of sentinel-1 interferometric coherence and backscatter for crop-type mapping, *Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2020).
- J. Shang, J. Liu, V. Poncos, X. Geng, B. Qian, Q. Chen, T. Dong, D. Macdonald, T. Martin, J. Kovacs, et al., Detection of crop seeding and harvest through analysis of time-series sentinel-1 interferometric sar data, *Remote Sensing* (2020).
- H. S. Srivastava, P. Patel, R. R. Navalgund, Application potentials of synthetic aperture radar interferometry for land-cover mapping and crop-height estimation, *Current Science* (2006).
- S. R. Cloude, E. Pottier, A review of target decomposition theorems in radar polarimetry, *Transactions on geoscience and remote sensing* (1996).
- Y. Yamaguchi, T. Moriyama, M. Ishido, H. Yamada, Four-component scattering model for polarimetric sar image decomposition, *Transactions on Geoscience and Remote Sensing* (2005).
- P. Srikanth, K. Ramana, U. Deepika, P. K. Chakravarthi, M. S. Sai, Comparison of various polarimetric decomposition techniques for crop classification, *Journal of the Indian Society of Remote Sensing* (2016).
- D. L. Schuler, J.-S. Lee, G. De Grandi, Measurement of topography using polarimetric sar images, *IEEE Transactions on Geoscience and Remote Sensing* (1996).
- F. Tupin, H. Maitre, J.-F. Mangin, J.-M. Nicolas, E. Pechersky, Detection of linear features in sar images: Application to road network extraction, *Transactions on geoscience and remote sensing* (1998).

- A. Kourgli, M. Ouarzeddine, Y. Oukil, A. Belhadj-Aissa, Land cover identification using polarimetric SAR images, na, 2010.
- V. S. F. Garnot, L. Landrieu, Panoptic segmentation of satellite image time series with convolutional temporal attention networks, ICCV (2021).
- G. Kaplan, L. Fine, V. Lukyanov, V. Manivasagam, J. Tanny, O. Rozenstein, Normalizing the local incidence angle in Sentinel-1 imagery to improve leaf area index, vegetation height, and crop coefficient estimations, Land (2021).
- I. N. Garkusha, V. Hnatushenko, V. V, Research of influence of atmosphere and humidity on the data of radar imaging by sentinel-1 (2017).
- S. Abramov, O. Rubel, V. Lukin, R. Kozhemiakin, N. Kussul, A. Shelestov, M. Lavreniuk, Speckle reducing for sentinel-1 sar data, IGARSS (2017).
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research (2014).
- M. Schneider, M. Körner, [re] satellite image time series classification with pixel-set encoders and temporal self-attention, in: ML Reproducibility Challenge 2020, 2020.
- E. Christophe, J. Inglada, A. Giros, Orfeo toolbox: a complete solution for mapping from high resolution satellite images, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37 (2008) 1263–1268.
- V. Singhroy, R. Saint-Jean, Effects of relief on the selection of radarsat-1 incidence angle for geological applications, Canadian Journal of Remote Sensing (1999).
- V. S. F. Garnot, L. Landrieu, Pastis-r - panoptic segmentation of radar and optical satellite image time series, 2021. URL: <https://zenodo.org/record/5735646>. doi:10.5281/ZENODO.5735646.
- D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, ICLR (2015).
- C. Wang, G. Zhang, R. Grosse, Picking winning tickets before training by preserving gradient flow, ICLR (2020).