PFG (2023) 91:413–423
https://doi.org/10.1007/s41064-023-00256-w

ORIGINAL ARTICLE

Generating Virtual Training Labels for Crop Classification from Fused Sentinel-1 and Sentinel-2 Time Series

Maryam Teimouri[1,3] · Mehdi Mokhtarzade[1] · Nicolas Baghdadi[2] · Christian Heipke[3]

Received: 17 July 2023 / Accepted: 29 August 2023 / Published online: 26 September 2023

Abstract

Convolutional neural networks (CNNs) have shown results superior to most traditional image understanding approaches in many fields, incl. crop classification from satellite time series images. However, CNNs require a large number of training samples to properly train the network. The process of collecting and labeling such samples using traditional methods can be both, time-consuming and costly. To address this issue and improve classification accuracy, generating virtual training labels (VTL) from existing ones is a promising solution. To this end, this study proposes a novel method for generating VTL based on sub-dividing the training samples of each crop using self-organizing maps (SOM), and then assigning labels to a set of unlabeled pixels based on the distance to these sub-classes. We apply the new method to crop classification from Sentinel images. A three-dimensional (3D) CNN is utilized for extracting features from the fusion of optical and radar time series. The results of the evaluation show that the proposed method is effective in generating VTL, as demonstrated by the achieved overall accuracy (OA) of 95.3% and kappa coefficient (KC) of 94.5%, compared to 91.3% and 89.9% for a solution without VTL. The results suggest that the proposed method has the potential to enhance the classification accuracy of crops using VTL.

Keywords Virtual training labels · Fusion · Optical and radar image time series · 3D-CNN · Crop classification

1 Introduction

Agriculture, as a major source of food production, plays a crucial role in meeting the nutritional needs of the growing human population. In the face of limited agricultural land and an increasing population, enhancing the efficiency of agricultural production becomes imperative to meet the rising food demand. An essential requirement for effective agricultural management is up-to-date information on crop types and their spatial distribution. Knowledge about the specific crop types serves as a fundamental input for analysis in crop management, including crop growth monitoring (Mascolo et al. 2015), estimation of crop area (Ali et al. 2022; Hudait and Patel 2022), and assessment of water requirements (Foster et al. 2019).

The advent of satellite sensors with high spatial resolution has significantly improved the ability to rapidly create accurate crop maps. Consequently, extensive research has been dedicated to automating crop classification using various data sources, such as optical (Niazmardi et al. 2018; Vuolo et al. 2018; Hao et al. 2020; Sakamoto 2021; Xia et al. 2022; Teimouri and Mokhtarzade 2023) and radar images (Bargiel 2017; Hariharan et al. 2018). It was found that identifying and differentiating crops from images is challenging due to factors like diverse environmental conditions, spectral heterogeneity within a particular class as well as similarity among different classes, and small-scale management practices, such as varying planting and harvesting times, leading to complex and spatially diverse signatures across multiple seasons.

On the algorithmic side, deep learning (DL) approaches and in particular convolutional neural networks (CNNs) are currently considered the best methods in image classification (e.g., Heipke and Rottensteiner 2020). Also, methods based on attention mechanisms (Vaswani et al. 2017; Dosovitskiy

✉ Maryam Teimouri
mteimouri@mail.kntu.ac.ir; teimouri@ipi.uni-hannover.de

1 Department of Photogrammetry and Remote Sensing, K. N. Toosi University of Technology, Tehran, Iran

2 INRAE, UMR TETIS, University of Montpellier, 500 Rue François Breton, 34093 Montpellier CEDEX 5, France

3 Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Hannover, Germany

et al. 2021; Voelsen et al. 2023) have recently made a major impact in the field. However, DL methods require a vast amount of training data in the learning phase to yield good results, and these training data are not always available.

In this paper, we address this problem and suggest a method, which automatically generates labels for unlabeled samples, so called VTL, from a given amount of real training labels (RTL). We show that adding the VTL to the RTL improves crop classification using Sentinel 1 and 2 (S1 and S2) time series, i.e. fusing optical and radar imagery. The architecture proposed by Teimouri et al. (2022) is applied in this study to assess the impact of the VTL on the training of 3D-CNNs.

The remainder of this article is structured as follows. Section 2 gives an overview of the state-of-the-art in crop classification, Sect. 3 discusses the approach of VTL and the structure of the 3D-CNN for crop classification using a fusion of optical and radar time series. Section 4 presents the study area, input data, experiments, and the analysis of the results. Finally, Sect. 5 provides the conclusions of the study.

2 State-of-the-Art in Deep Learning for Crop Classification

2.1 CNNs of Various Dimensions

CNNs are capable of learning complex functions, making them a powerful tool for developing accurate classification. Depending on the dimension of the convolution operator (one-, two- or three-dimensional (1D, 2D, 3D), CNNs can extract various types of features, including spatial, spectral, temporal, spatial-spectral, and spatial–temporal features. For example, 1D convolutions have been used to extract spectral features in hyperspectral images (Li et al. 2016) and temporal features in image sequences (Pelletier et al. 2019), while the standard 2D convolutions are commonly used for extracting spatial features in single images. 3D convolution operators are typically applied to extract spatial–temporal or spectral-spatial features, as demonstrated in studies by Li et al. (2017), Ji et al. (2018), Han et al. (2020), Sellami et al. (2020), and Fernandez-Beltran et al. (2021). Additionally, some studies have used a combination of 1D and 2D convolutions (Kussul et al. 2017; Zhang et al. 2017) or of 2D and 3D convolutions (Ge et al. 2020; Voelsen et al. 2022).

2.2 Crop Classification Using Neural Networks

Research on crop classification using networks of different dimensions is briefly reported in this section. Most approaches rely on time series and, besides CNN operations, employ architectures developed for temporal data such as recurrent neural networks (RNN), long-short-term memory networks (LSTM), and transformers based on attention mechanisms.

1D-CNN: Rußwurm and Körner (2020) investigated the effectiveness of 1D-CNN, LSTM, and self-attention neural networks for crop classification from S2 time series. Their research demonstrated that both, the transformer and LSTM models outperformed the 1D-CNN. The study by Zhao et al. (2021) aimed at evaluating the performance of five different neural network models, namely 1D-CNN, LSTM, gated recurrent unit (GRU), LSTM-CNN, and GRU-CNN, in classifying crops using S2 time series images. The results showed that GRU-CNN and LSTM-CNN, as well as the 1D-CNN, performed significantly better than the other investigated models.

2D-CNN: Moreno-Revelo et al. (2021) proposed a 2D-CNN for classifying ten agricultural crops in a tropical region from S1 and Landsat 8 images. A major limitation of this method is that the proposed architectures are shallow, which limits their ability to extract more complex features. Mazzia et al. (2019) classified fifteen different crops using a combination of RNNs and CNNs applied to S2 time series images. The proposed network architecture involved feeding the time series images into a LSTM module, followed by concatenating the extracted features and passing them through a 2D-CNN. The reported results were better than those for support vector machines and random forests. Seydi et al. (2022) applied a dual-stream network to classify agricultural and non-agricultural crops. The network consisted of convolutional blocks and attention models.

Garnot et al. (2020) proposed a method for crop classification from the S2 time series. The method involved extracting temporal features using an architecture that relied on self-attention mechanisms, while spatial features were obtained using a pixel-set encoder. Garnot and Landrieu (2021) introduced the Unet-Temporal Attention Encoder (U-TAE) model, which combines multi-scale spatial convolutions and temporal attention, enabling the extraction of spatial–temporal features at various resolutions. Ofori-Ampofo et al. (2021) integrated S1 and S2 time series for crop classification using an attention-based encoder. Garnot et al. (2022) explored various strategies for the fusion of optical and radar time series images, i.e. parcel-based classification, semantic, and panoptic segmentation, for crop classification. Finally, they proposed a mid-fusion scheme that utilizes separate spatial encoders and a shared temporal encoder. Finally, Tarasiou et al. (2023) introduced the spatial–temporal vision transformer, a model based on visual transformers (Dosovitskiy et al. 2021). Additionally, they proposed tokenization schemes to adapt the approach for modeling satellite image time series.

3D-CNN: In the study conducted by Ji et al. (2018) agricultural crops were classified using optical time series

images and a 3D-CNN. The network was designed by separately considering time series patches of spectral bands, and then combining the obtained features. Similarly, Teimouri et al. (2022) proposed a 3D-CNN architecture for crop classification using a fusion of S1 and S2 time series images. This architecture was able to extract temporal-spatial-radar-spectral features, and the results showed its high potential of this network for crop classification; it forms the basis of the work reported in this paper.

2.3 Data Augmentation for CNN Training

Studies have shown that the performance of CNNs improves with an increase in the amount of training data (Chen et al. 2016; Li et al. 2016), while traditional methods do not show significant improvements with the same increase in data (Sarker 2021). Thus, having a large number of training samples is essential for improving the accuracy of deep networks. However, labeling high-quality samples manually is expensive and time-consuming. To solve this problem, the related research can be divided into two main categories (Hao et al. 2023): data-driven methods and network-based methods.

Data-driven methods involve the generation of new samples by employing various techniques applied to real training data. These techniques include: (1) Geometric transformations such as rotation, scaling, flipping, and cropping (e.g., Zhang et al. 2017; Acción et al. 2020); (2) Noise disturbance (Ding et al. 2016); (3) Sharpness transformation (Ledig et al. 2017), albeit with limited success. (4) Generative adversarial networks (GANs) (Goodfellow et al. 2014); since then, various extensions have been suggested. However, these networks still require a large amount of training samples and have a high computational cost. (5) Virtual labels (Chen et al. 2016; Li et al. 2016). Labeling these samples is done before network training, thus reducing the high computing time required by GAN methods.

On the other hand, network-based data augmentation methods focus on modifying the architecture or learning process of CNNs. These methods include: (1) Transfer learning (Wurm et al. 2019; Cui et al. 2020), taking advantage of models pre-trained on large datasets and fine-tuning them for specific tasks. (2) Regularization (Yun et al. 2019), using techniques like dropout, weight decay, and batch normalization. (3) Meta-learning (Li et al. 2021) to train the model on multiple tasks to enhance its ability to better adapt to new tasks.

Despite the potential benefits of using virtual labels, limited research has focused on their application in remote sensing. Chen et al. (2016) proposed two novel approaches for generating VTL to improve the classification of hyperspectral images using a 3D-CNN. The first method involved multiplying a training sample with a real label by a random factor and adding random noise to create a VTL. The second method considered a combination of two RTL of the same class and added random Gaussian noise. Li et al. (2016) proposed a pixel-pair-based method for generating VTL from hyperspectral images. They utilized VTL along with real training labels for the classification of images. These techniques have the potential to increase the accuracy of classification models without requiring the costly and time-consuming process of manual labeling.

2.4 Summary of State-of-the-Art

In summary, promising results have been obtained in crop classification using image time series and CNNs for the spatial domain as well as CNNs or transformers for the temporal domain. In addition, fusing various data sources, such as optical and radar time series images led to improved classification results. Notably, the decision-level fusion yielded significantly better performance compared to the feature-level fusion approach. The lack of appropriate training data limits the success of these methods, however. To at least partly overcome this problem the generation of VTL (thus, a data-driven method) is suggested in this paper. We fuse optical and radar images and use a 3D-CNN architecture without transformers in our work.

3 Methodology

3.1 Overview

This research proposes a method to overcome the challenge of collecting sufficient training samples through field methods or other manual processes, which can be both, time-consuming and expensive. The aim is to generate VTL that can be used in the training of deep neural networks, enabling the network to accurately classify crops, ultimately improving its overall performance.

VTL are generated by first sub-dividing RTL of each class separately into different sub-classes in an unsupervised manner using SOM (Kohonen 1995). Unlabeled pixels are then associated with the sub-class they are most similar to using a set of similarity criteria, yielding the VTL.

Subsequently, the networks are trained using VTL and RTL together. Note that at this stage only the original classes, and not the sub-classes are considered, as otherwise the number of training samples would be too low. 3D convolution operators are used to extract feature vectors, which are then fed as input to the actual classifier. The employed architecture is the one proposed in Teimouri et al. (2022).

The study compares the results obtained by training the network using a combination of RTL and VTL with those achieved using RTL and with VTL only. The four evaluation

metrics are OA, KC, the F1-score per class, and the user accuracy (UA).

3.2 VTL Generation Using SOM

An agricultural crop is affected by many factors, such as sunlight, soil properties, irrigation, and other environmental factors. These effects can lead to differences in growth pattern. Xu et al. (2018) demonstrated how the reflection of a crop in different areas of a study region, as captured in an image, may vary. It can thus be beneficial to divide the training data of the individual classes into several sub-classes based on the highest degree of similarity in the growth cycle.

The method suggested in this paper is based on this observation. The training samples for each class are sub-divided into different sub-classes using an unsupervised classification of the time series images. This clustering step is guided by two constraints: the cluster centers should be as far away from each other as possible, and the clusters should be as compact as possible.

Once these clusters are found, unlabeled pixels are tested to belong to one of those sub-classes based on some metric, and are labeled according to the sub-class (and thus the class) with the minimum distance. These newly labeled pixels are the desired VTL, see Fig. 1 for an overview.

The procedure consists of three main steps: (1) Sub-division of the training samples of each class, (2) Similarity computation, (3) Majority voting and labeling.

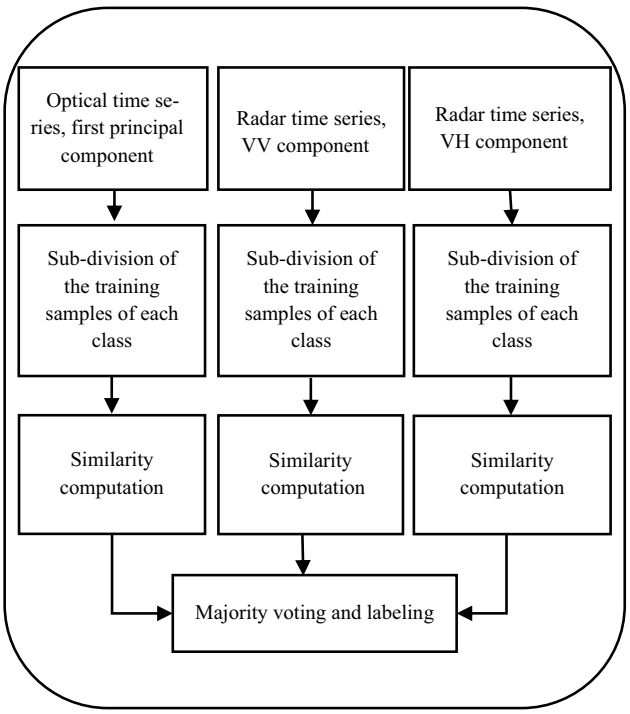

Fig. 1 The flowchart of VTL generation, for details see text

Sub-Division of the Training Samples of Each Class: For this clustering problem we use the traditional SOM, as they are a highly notable unsupervised neural network classifier. To reduce the computational load, we only use the first principal component of each optical image of each epoch, as well as the VV and VH polarizations of radar time series images as input for clustering, as shown in Fig. 1. The SOM output layer consists of m neurons, meaning that we subdivide a given class into m different sub-classes (or clusters). Different values for m (i.e., 2, 3, and 4) have been explored for each class separately, and the best number of clusters was selected, again for each class separately.

To determine this best number m of sub-classes for each class, a scoring criterion is defined, which considers two constraints: the first constraint, referred to as BC, is based on the different sub-class centers. Equation (1) is used to calculate the distance between the centers of sub-classes and the class center.

$$BC = \sum_{i=1}^{m} ||\mu^i - \mu||^2 \tag{1}$$

In this context, μ and μ^i represent the center of all samples belonging to a class, computed by averaging the positions of all samples within that class, and the center of sub-class i, respectively.

The second constraint, termed WC, is related to the cluster compactness of each sub-class, see Eq. (2):

$$WC = \sum_{i=1}^{m} \sum_{j=1}^{L} ||x_j^{(i)} - \mu^i||^2 \tag{2}$$

μ^i denotes the center of the ith sub-class, L represents the number of samples belonging to sub-class i, and $x_j^{(i)}$ refers to jth sample in sub-class i.

Using the scoring criterion given in Eq. (3), based on BC and WC defined above, the best number of sub-classes m is selected for each crop. The number of sub-classes is considered best, if the distance between the cluster centers, and thus BC, is largest and the compactness is highest, resulting in a small value for WC. Thus, we minimize the $Score$ for each training class as a function of m:

$$Score(m) = MIN(\frac{WC(m)}{BC(m)}) \tag{3}$$

Similarity Computation: Next, n unlabeled pixels are randomly chosen in the study area, and virtual labels are assigned to these pixels. To do so, the distance to the center of all generated sub-classes is computed according to the following four similarity criteria (Thenkabail et al. 2007): spectral angle mapper, spectral correlation similarity, Euclidean distance, and spectral similarity value. A threshold is then determined for each similarity criterion for each sub-class.

These thresholds (Eq. (4)) are calculated as the average value of each similarity criterion among RTL.

$$T^{(i)} = \frac{1}{L}\sum_{j=1}^{L} V_j^{(i)} \tag{4}$$

$V_j^{(i)}$ denotes the value derived from the similarity criterion, which measures the similarity between sample j and the center of sub-class i. L represents the number of samples belonging to that subclass, and T denotes the threshold value for the similarity criterion of that subclass.

Only pixels with a value below the established thresholds for all four similarity criteria are chosen for further processing, and for each criterion the class with the smallest score is retained.

Majority Voting and Labeling: These computations are carried out separately for the principal components of the optical images, and the VV component and the VH component of the radar images, resulting in one, two or three sets (for the optical, the VV, and the VH bands, respectively) of classes with four entries each (one for each similarity criterion), i.e., up to 12 possible labels for each pixel. The final label is chosen according to the majority voting method. In case of ambiguity, i.e., two or more classes have the same number of votes, and no class has a higher number, the pixel is rejected.

3.3 3D-CNN Architecture for Combined Optical and Radar Time Series Image Classification

In this study, 3D-CNNs were employed to extract spatial–temporal, spectral, and intensity features from the optical and radar data; the architecture used in our previous research (Teimouri et al. 2022) was employed. In this previous study, 3D-CNNs were trained using RTL only for crop classification. Here, we extend our approach by incorporating both, RTL and VTL together to train the network and investigate the impact of VTL on crop classification.

As shown in Fig. 2, to fuse the optical and radar time series images, a 3D-CNN with two input branches is used. Each branch takes the optical and radar time series images, linearly normalized to [0, 1], as input, respectively. Each consists of twelve 3D convolutional operators. Finally, the features extracted from each data source are concatenated and fed into the fully connected layer.

More specifically, the input channels for each dataset (optical, radar, and fused) are processed separately, where the 3D convolution operators are applied to a sequence of three images with stride one in the temporal direction for each channel. Next, the features generated from each time series of a specific channel are concatenated. The network architectures for radar and optical data consist of three convolutional blocks with 32, 32, and 64 kernels, respectively. Each block is followed by a ReLu activation function and maximum pooling with dimensions of $2\times2\times1$. The final layers of this architecture consist of two fully connected layers with 128 and 64 neurons, respectively.

This architecture is used for pixel-wise classification, the central pixel of each 7×7 patch is classified. Patches for training are randomly extracted from the scene, taking care to avoid any overlap between patches in order to decrease possible correlations. While for the generation of the semantic segmentation map, each pixel of the scene was classified independently, the test sample patches were again selected randomly in the scene and in a way that they did not have any overlap either.

The learning rates, number of epochs, and mini-batch sizes used in this study are 0.001, 1000, and 500, respectively. Dropout layers with a rate of 0.4 are used after each fully connected layer to reduce the effect of overfitting. The network in this research was trained using the cross-entropy loss function with adaptive Moment Estimation (Adam) optimizer (Kingma and Ba 2014) and early stopping. The early stopping criterion was considered to be satisfied when the validation accuracy had consistently decreased for ten consecutive iterations.

4 Experimental Results

4.1 Test Site and Preprocessing

We use images from the region of Catalonia, located in the northeastern part of Spain. The majority of the area is covered by agricultural lands, as shown in Fig. 3a. The selected area is dominated by seven different crops (alfalfa, oat, corn, beans, triticale, wheat, and rapeseed). As is usual in crop monitoring, one image per month was used in this work, resulting in seven images between February and August 2018 (Table 1). The optical image of March is covered by a few clouds and therefore was ignored in this study. Each S1 image consists of two polarizations (VV, VH), which were acquired in Ground Range Detection (GRD) mode. The preprocessing applied to these images included accurate geolocation, removing thermal noise to enhance image quality, performing radiometric calibration to normalize the intensity values, applying speckle filtering to reduce the granular noise, and conducting range doppler terrain correction to correct geometric distortions caused by topography. All preprocessing steps were executed using Sentinel's Application Platform (SNAP) software, the necessary parameters were taken from the available orbit files. The radar images were resampled to a spatial resolution of 10×10 m^2. Four spectral bands (red, green, blue, and near-infrared) of each S2 level 2A image were chosen, as these hold significant potential for crop classification (Defourny et al. 2019; Dhau et al. 2021;

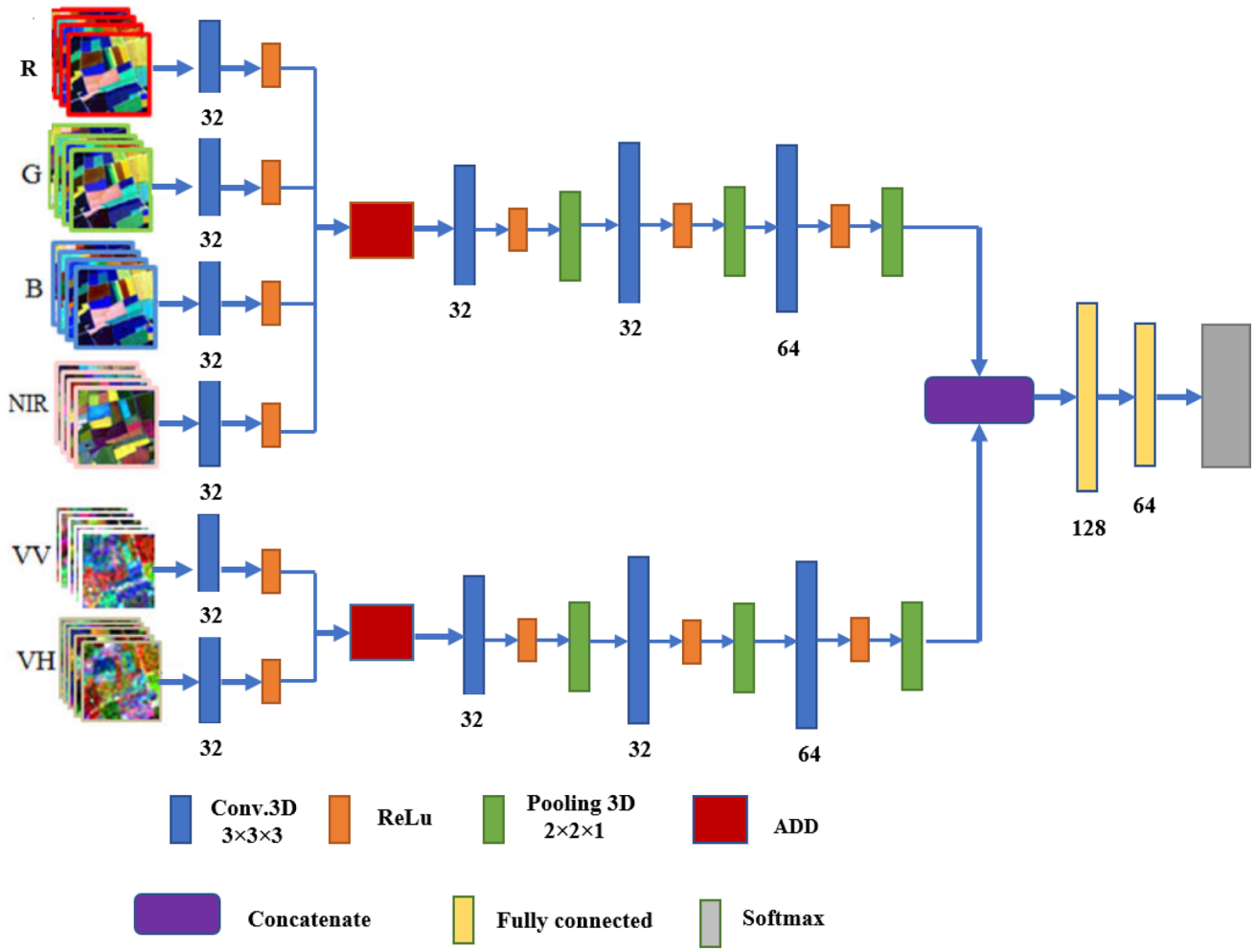

Fig. 2 The network structure for fusion of S1 and S2 (adopted from Teimouri et al. 2022)

You et al. 2021). All images consisted of 1593 × 2516 pixels. A ground truth map was produced by the Department of Agriculture, Livestock, Fishing and Food of the Generalitat of Catalonia. This map was resampled to 10 × 10 m^2 (Fig. 3b) as well. For network training, 1050 training samples and 490 validation samples were used, and 3500 test samples were employed to evaluate the algorithms (where each sample is an individual pixel). Training as well as validation and test samples were randomly distributed across the study area. An equal distribution of the number of samples was ensured across all classes.

4.2 Generating VTL

As described in chapter 3, there are two different inputs for VTL generation: the six first principal components derived from the optical time series images (one for each epoch), and seven VV polarization as well as seven VH polarization channels for the radar images. The 1050 pixels, all coming from the training data, were then employed as RTL to generate VTL.

We generated three different sets of VTL, one each for classifying only optical and radar data separately, and one for the classification of both image types together. For the VTL for classifying the optical data, only RTL from the optical channels were used, and analogously for the radar data and the fused image set.

The number of selected unlabeled pixels was chosen to be approximately three times as large as the number of RTL. The reason was that in the end we wanted to have approximately the same number of RTL and VTL, however, some VTL were rejected due to ambiguous results, as mentioned before. The factor of three turned out to be a good choice. Finally, a total of 2100 samples (i.e., RTL + VTL) was used for training in each run.

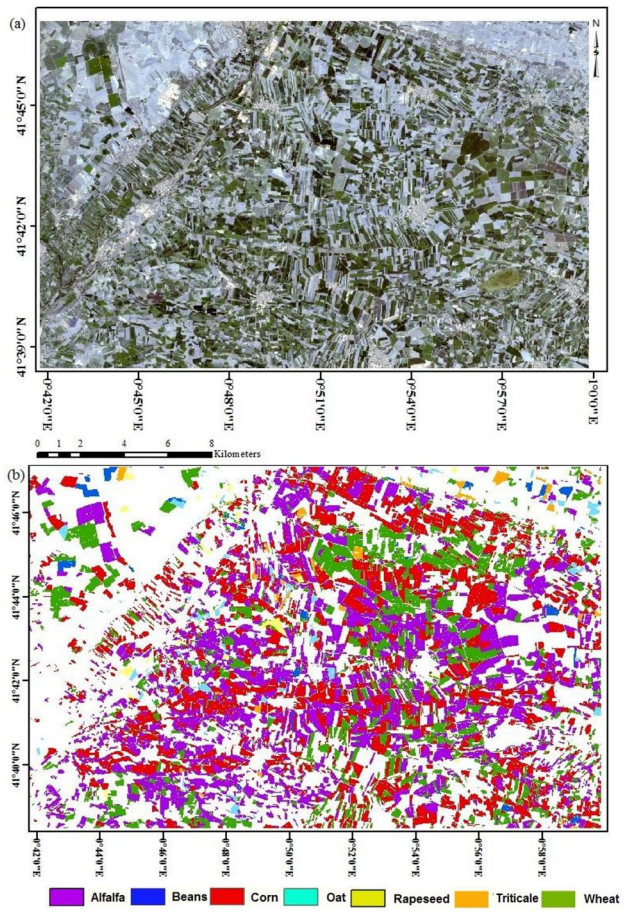

Fig. 3 **a** True color S2 image of the study area in July (Red: B4 band, Green: B3, Blue: B2), **b** ground truth map, both with a size of 1593×2516 pixels at a resolution of 10×10 m^2 (from Teimouri et al. 2022)

Table 1 SAR and optical data collection

Sensor	Acquisition date	Sensor	Acquisition date
S1	February 12, 2018	S2	February 10, 2018
	March 20, 2018		–
	April 25, 2018		April 21, 2018
	May 07, 2018		May 16, 2018
	June 24, 2018		June 20, 2018
	July 18, 2018		July 20, 2018
	August 23, 2018		August 19, 2018

4.3 Results

A comparison of the results for the 3500 test samples achieved with only RTL and with a combination of RTL and VTL is presented in Tables 2 and 3. It shows that for most classes the combination leads to an increase in classification accuracy for crops in optical and radar data sources, as well as their fusion. Achieving an OA of 92.6% and a KC of 91.4% for the S2 images demonstrates the performance of combining RTL and VTL. The inclusion of VTL improved the OA by 4.0% and the KC by 4.7%. In particular, the VTL generated for corn, oat, wheat, and triticale were highly effective, with corn showing the largest improvement of 15.6% in UA. These improvements are further supported by the F1-score analysis (Table 3), which confirms the enhanced performance of these classes when using RTL and VTL together. However, it should be noted that the addition of VTL resulted in a decrease in UA for beans, although the F1-score indicates an improvement of 2.1%.

Similarly, for radar images the integration of VTL and RTL led to an increase in OA (3.3%) and KC (3.9%). Interestingly, for alfalfa the VTL generated using radar data showed more significant improvements compared to those generated using optical data, with an improvement of approximately 7.2% in UA and 5.0% in F1-score, while the impact of VTL generated from optical data was only a 1% in UA and showed a decrease in the F1-score of 0.2%. The integration of VTL also led to improvements in UA of beans, corn, oat, triticale, and wheat; the largest improvement in F1-score was observed in corn, triticale, and wheat, with approximately 11.5%, 2.3%, and 7.5% improvements, respectively. However, for rapeseed, oat, and beans the F1-scores decreased somewhat.

The combination of VTL and RTL also yielded significant improvements in the OA and KC when fusing optical and radar data. With OA and KC scores of 93.0% and 91.9%, respectively. Notably, UA of wheat, corn, beans, and oat improved by 4.4%, 3.2%, 2.6%, and 2.4%, respectively, demonstrating the effectiveness of VTL in accurately identifying and distinguishing between different crops. Furthermore, the F1-score displayed improvements in all crops. These results highlight the potential of using VTL, which can significantly improve the accuracy of crop classification, especially for certain crops.

To test the quality of the generated VTL, we also trained the networks with only VTL. While the results were still acceptable, in general, a decrease of about 10% in OA and KC was observed, as was to be expected.

Figures 4 and 5 depict two large subsets of the study area, which were chosen for visual interpretation. The maps produced using a combination of VTL and RTL exhibit a significant level of map uniformity with reduced noise. Additionally, the yellow ellipses illustrate the impact of VTL on crop classification, leading to improved results in most regions.

Finally, although care was taken to only generate correct VTL, there is obviously a probability that some virtual samples have incorrect labels, potentially introducing erroneous information during training. To tackle this challenge, a strategy was designed, where the virtual sample set was randomly divided into multiple subsets. The subsets were

Table 2 OA, UA and KC, all in %, for classification with only RTL, combination of RTL + VTL and differences (based on 3500 test samples, RTL only from Teimouri et al. 2022)

Class	RTL only			RTL + VTL			Difference		
	Optical	Radar	Fusion	Optical	Radar	Fusion	Optical	Radar	Fusion
	UA	UA	UA	UA	UA	UA	UA	UA	UA
Alfalfa	76.8	66.6	82.0	77.8	73.8	80.8	1.0	7.2	-1.2
Beans	95.6	81.8	93.0	94.6	84.0	95.6	-1.0	2.2	2.6
Corn	71.8	85.6	87.0	87.4	87.2	90.2	15.6	1.6	3.2
Oat	90.8	87.4	91.8	96.6	91.2	94.2	5.8	3.8	2.4
Rapeseed	97.2	97.2	97.4	98.2	96.8	98.6	1.0	-0.4	1.2
Triticale	97.6	95.4	99.2	100.0	97.8	98.6	2.4	2.4	-0.6
Wheat	90.6	83.0	88.8	93.6	89.6	93.2	3.0	6.6	4.4
OA	88.6	85.3	91.3	92.6	88.6	93.0	4.0	3.3	1.7
KC	86.7	82.8	89.9	91.4	86.7	91.9	4.7	3.9	2.0

The colors represent the impact of the method on the results, which can be either positive or negative

Table 3 F1-scores in % for classification with only RTL, combination of RTL + VTL and differences (based on 3500 test samples, RTL only from Teimouri et al. 2022)

Class	RTL only			RTL + VTL			Difference		
	Optical	Radar	Fusion	Optical	Radar	Fusion	Optical	Radar	Fusion
	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score
Alfalfa	82.4	73.8	81.5	82.2	78.8	87.3	-0.2	5.0	5.8
Beans	93.9	88.0	94.6	96.0	86.9	96.3	2.1	-1.1	1.7
Corn	78.3	74.8	87.8	89.1	86.3	88.9	10.8	11.5	1.1
Oat	92.7	91.3	94.7	97.2	89.9	95.0	4.5	-1.4	0.3
Rapeseed	95.4	95.9	97.5	96.1	94.0	97.7	0.7	-1.9	0.2
Triticale	93.3	92.1	95.7	94.6	94.4	96.6	1.3	2.3	0.9
Wheat	83.3	81.7	87.5	92.2	89.2	89.3	8.9	7.5	1.8

The colors represent the impact of the method on the results, which can be either positive or negative

then iteratively injected into RTL. If the OA of the validation improved, the corresponding VTL subset was combined with the RTL. Otherwise, the subset was rejected, this happened in about 30% of the cases. While there is a possibility that, in this way, some virtual samples with correct labels fell into rejected subsets, the primary objective was to identify VTL subsets with high accuracy. The best number of subsets was experimentally found to be 10, with an equal number of samphles in each subset.

The selected virtual samples are referred to as VTL*. Table 4 presents the results obtained by training the 3D-CNN using RTL + VTL and RTL + VTL*, respectively, for the classification of crops from the fusion of S1 and S2 time series images.

According to Table 4, the proposed method demonstrates another significant improvement in classification accuracy. Additionally, the comparison between RTL + VTL and RTL + VTL* highlights that although the generated VTL enhances classification accuracy, there is a possibility of some samples having incorrect labels. By excluding these samples, higher accuracy samples were utilized for training the 3D-CNN, resulting in improvements in OA and KC by approximately 2.3% and 2.6%, respectively. The results obtained from RTL + VTL* exhibit stronger performance in comparison to RTL + VTL, with improved UA and the F1-scores compared to the results presented in Tables 2 and 3 across nearly all classes.

5 Conclusion

This paper presents a novel method for generating VTL to enhance the training of 3D-CNNs for crop classification using fused Sentinel S1 and S2 time series data. The study revealed that incorporating both, VTL and RTL during training leads to higher classification accuracy and a better F1-score for nearly all classes. By training the network using VTL + RTL for fusing S1 and S2 time series images, the results obtained demonstrate an improvement in OA and KC by 1.7% and 2.0%, respectively, compared to training the network solely with RTL. Furthermore, it was observed that among the generated virtual labels, some had incorrect labels. However, by iteratively adding only VTL, which increased OA, and training the network

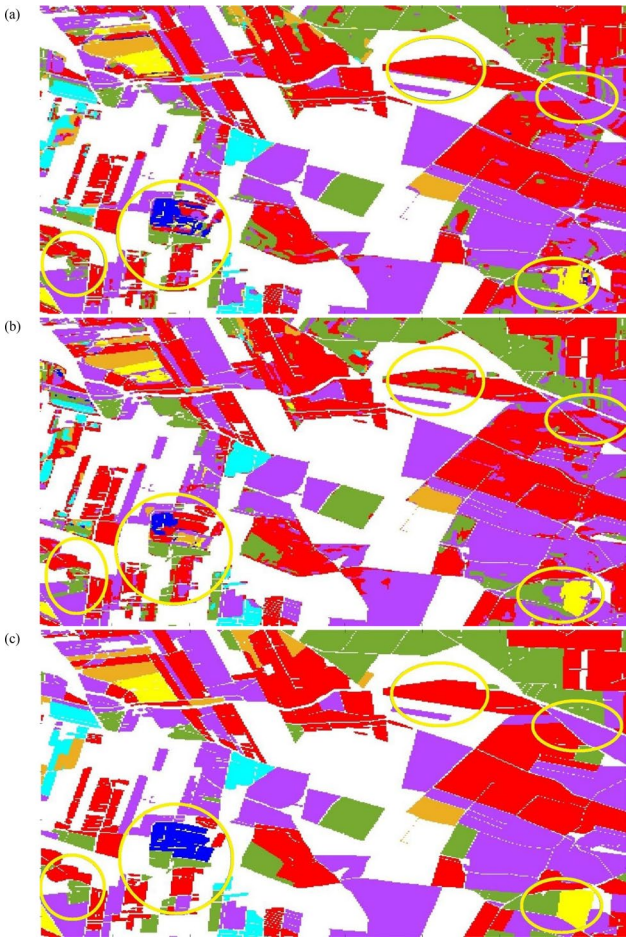

Fig. 4 Results of the fusion of optical and radar time series images from first subset of the study area. **a** VTL + RTL, **b** RTL only, **c** ground truth. Yellow ellipses show areas with particular differences

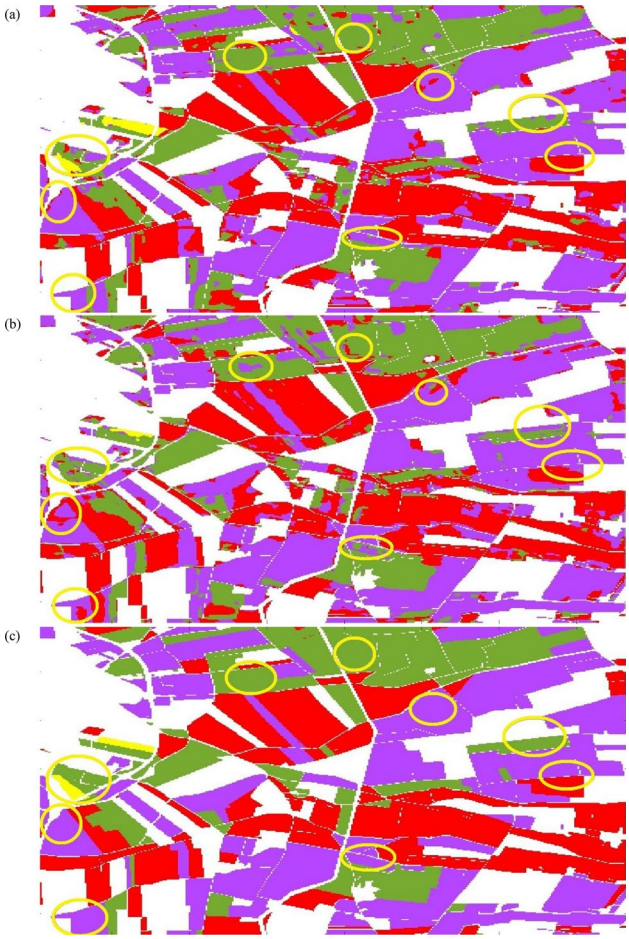

Fig. 5 Results of the fusion of optical and radar time series images from the second subset. **a** VTL + RTL, **b** RTL only, **c** ground truth. Yellow ellipses show areas with particular differences

with a reduced number of VTL, OA and KC improve by 2.3% and 2.6%, respectively. Consequently, the proposed method can be said to significantly contributing to the improvement of crop classification.

In future works, we will test the method on additional and larger datasets. We also plan to incorporate attention-based approaches for the temporal domain, as well as knowledge on plant phenology, the latter by conditioning the network on this prior information in a suitable way.

Acknowledgements The authors would like to express their gratitude to the European Space Agency (ESA) for supplying the Sentinel 1 and Sentinel 2 data, as well as to the Department of Agriculture, Livestock,

Table 4 Comparison of performance between RTL + VTL and RTL + VTL* obtained from the fusion of S1 and S2 time series

Class		RTL + VTL		RTL + VTL*	
		UA	F1-score	UA	F1-score
1	Alfalfa	80.8	87.3	85.6	89,5
2	Beans	95.6	96.3	95.0	97,0
3	Corn	90.2	88.9	95.4	92,4
4	Oat	94.2	95.0	96.6	97,6
5	Rapeseed	98.6	97.7	98.8	98,8
6	Triticale	98.6	96.6	100.0	97,5
7	Wheat	93.2	89.3	95.6	94.1
OA		93.0		95.3	
KC		91.9		94.5	

Fishing, and Food of the Generalitat of Catalonia for supplying the field data.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability All images used in this study were downloaded from "https://scihub.copernicus.eu/dhus/#/home," and ground truth data was generated by the Department of Agriculture, Livestock, Fishing, and Food of the Generalitat of Catalonia.

References

Acción Á, Argüello F, Heras DB (2020) Dual-window superpixel data augmentation for hyperspectral image classification. Appl Sci 10(24):8833. https://doi.org/10.3390/app10248833

Ali AM, Abouelghar M, Belal A, Saleh N, Yones M, Selim AI, Amin ME, Elwesemy A, Kucher DE, Maginan S (2022) Crop yield prediction using multi sensors remote sensing. Egypt J Remote Sens Space Sci 25(3):711–716. https://doi.org/10.1016/j.ejrs.2022.04.006

Bargiel D (2017) A new method for crop classification combining time series of radar images and crop phenology information. Remote Sens Environ 198:369–383. https://doi.org/10.1016/j.rse.2017.06.022

Chen Y, Jiang H, Li C, Jia X, Ghamisi P (2016) Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans Geosci Remote Sens 54(10):6232–6251. https://doi.org/10.1109/TGRS.2016.2584107

Cui B, Chen X, Lu Y (2020) Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection. IEEE Access 8:116744–116755. https://doi.org/10.1109/ACCESS.2020.3003914

Defourny P, Bontemps S, Bellemans N, Cara C, Dedieu G, Guzzonato E, Hagolle O, Inglada J, Nicola L, Rabaute T (2019) Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sens Environ 221:551–568. https://doi.org/10.1016/j.rse.2018.11.007

Dhau I, Dube T, Mushore TD (2021) Examining the prospects of sentinel-2 multispectral data in detecting and mapping maize streak virus severity in smallholder Ofcolaco farms South Africa. Geocarto Int 36(16):1873–1883. https://doi.org/10.1080/10106049.2019.1669724

Ding J, Chen B, Liu H, Huang M (2016) Convolutional neural network with data augmentation for SAR target recognition. IEEE Geosci Remote Sens Lett 13(3):364–368. https://doi.org/10.1109/LGRS.2015.2513754

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16 × 16 words: transformers for image recognition at scale. ICLR. https://doi.org/10.48550/arXiv.2010.11929

Fernandez-Beltran R, Baidar T, Kang J, Pla F (2021) Rice-yield prediction with multi-temporal sentinel-2 data and 3D CNN: a case study in Nepal. Remote Sens 13(7):1391. https://doi.org/10.3390/rs13071391

Foster T, Gonçalves IZ, Campos I, Neale C, Brozović N (2019) Assessing landscape scale heterogeneity in irrigation water use with remote sensing and in situ monitoring. Environ Res Lett 14(2):024004. https://doi.org/10.1088/1748-9326/aaf2be

Garnot VSF, Landrieu L (2021) Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. https://doi.org/10.48550/arXiv.2107.07933

Garnot VSF, Landrieu L, Giordano S, Chehata N (2020) Satellite image time series classification with pixel-set encoders and temporal self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1911.07757

Garnot VSF, Landrieu L, Chehata N (2022) Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS J Photogram Remote Sens 187:294–305. https://doi.org/10.48550/arXiv.2112.07558

Ge Z, Cao G, Li X, Fu P (2020) Hyperspectral image classification method based on 2D–3D CNN and multibranch feature fusion. IEEE J Select Top Appl Earth Observ Remote Sens 13:5776–5788

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, 27

Han Y, Wei C, Zhou R, Hong Z, Zhang Y, Yang S (2020) Combining 3D-CNN and squeeze-and-excitation networks for remote sensing sea ice image classification. Math Probl Eng 2020:1–15. https://doi.org/10.1155/2020/8065396

Hao P, Di L, Zhang C, Guo L (2020) Transfer learning for crop classification with cropland data layer data (CDL) as training samples. Sci Total Environ 733:138869. https://doi.org/10.1016/j.scitotenv.2020.138869

Hao X, Liu L, Yang R, Yin L, Zhang L, Li X (2023) A review of data augmentation methods of remote sensing image target recognition. Remote Sens 15(3):827. https://doi.org/10.3390/rs15030827

Hariharan S, Mandal D, Tirodkar S, Kumar V, Bhattacharya A, Lopez-Sanchez JM (2018) A novel phenology based feature subset selection technique using random forest for multitemporal PolSAR crop classification. IEEE J Select Top Appl Earth Observ Remote Sens 11(11):4244–4258. https://doi.org/10.1109/JSTARS.2018.2866407

Heipke C, Rottensteiner F (2020) Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. Geospatial Inf Sci 23(1):10–19. https://doi.org/10.1080/10095020.2020.1718003

Hudait M, Patel PP (2022) Crop-type mapping and acreage estimation in smallholding plots using Sentinel-2 images and machine learning algorithms: some comparisons. Egypt J Remote Sens Space Sci 25(1):147–156. https://doi.org/10.1016/j.ejrs.2022.01.004

Ji S, Zhang C, Xu A, Shi Y, Duan Y (2018) 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. Remote Sens 10(1):75. https://doi.org/10.3390/rs10010075

Kingma DP, Ba J (2014) Adam: a method for stochastic optmization. arXiv preprint https://arXiv.org/1412.6980. https://doi.org/10.48550/arXiv.1412.6980

Kohonen T (1995) Self-organising maps. Springer series in information, sciences, vol 30. Springer, Berlin, Heidelberg

Kussul N, Lavreniuk M, Skakun S, Shelestov A (2017) Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci Remote Sens Lett 14(5):778–782. https://doi.org/10.1109/LGRS.2017.2681128

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Li W, Wu G, Zhang F, Du Q (2016) Hyperspectral image classification using deep pixel-pair features. IEEE Trans Geosci Remote Sens 55(2):844–853. https://doi.org/10.1109/TGRS.2016.2616355

Li Y, Zhang H, Shen Q (2017) Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. Remote Sens 9(1):67. https://doi.org/10.3390/rs9010067

Li Y, Shao Z, Huang X, Cai B, Peng S (2021) Meta-FSEO: a meta-learning fast adaptation with self-supervised embedding optimization for few-shot remote sensing scene classification. Remote Sens 13(14):2776. https://doi.org/10.3390/rs13142776

Mascolo L, Lopez-Sanchez JM, Vicente-Guijalba F, Mazzarella G, Nunziata F, Migliaccio M (2015) Retrieval of phenological stages of onion fields during the first year of growth by means of C-band polarimetric SAR measurements. Int J Remote Sens 36(12):3077–3096. https://doi.org/10.1080/01431161.2015.1055608

Mazzia V, Khaliq A, Chiaberge M (2019) Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN). Appl Sci 10(1):238. https://doi.org/10.3390/app10010238

Moreno-Revelo MY, Guachi-Guachi L, Gómez-Mendoza JB, Revelo-Fuelagán J, Peluffo-Ordóñez DH (2021) Enhanced convolutional-neural-network architecture for crop classification. Appl Sci 11(9):4292. https://doi.org/10.3390/app11094292

Niazmardi S, Homayouni S, Safari A, McNairn H, Shang J, Beckett K (2018) Histogram-based spatio-temporal feature classification of vegetation indices time-series for crop mapping. Int J Appl Earth Obs Geoinf 72:34–41. https://doi.org/10.1016/j.jag.2018.05.014

Ofori-Ampofo S, Pelletier C, Lang S (2021) Crop type mapping from optical and radar time series using attention-based deep learning. Remote Sens 13(22):4668. https://doi.org/10.3390/rs13224668

Pelletier C, Webb GI, Petitjean F (2019) Temporal convolutional neural network for the classification of satellite image time series. Remote Sens 11(5):523. https://doi.org/10.3390/rs11050523

Rußwurm M, Körner M (2020) Self-attention for raw optical satellite time series classification. ISPRS J Photogramm Remote Sens 169:421–435. https://doi.org/10.1016/j.isprsjprs.2020.06.006

Sakamoto T (2021) Early classification method for US corn and soybean by incorporating MODIS-estimated phenological data and historical classification maps in random-forest regression algorithm. Photogramm Eng Remote Sens 87(10):747–758. https://doi.org/10.14358/PERS.21-00003R2

Sarker IH (2021) Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci 2(6):420. https://doi.org/10.1007/s42979-021-00815-1

Sellami A, Abbes AB, Barra V, Farah IR (2020) Fused 3-D spectral-spatial deep neural networks and spectral clustering for hyperspectral image classification. Pattern Recogn Lett 138:594–600. https://doi.org/10.1016/j.patrec.2020.08.020

Seydi ST, Amani M, Ghorbanian A (2022) A dual attention convolutional neural network for crop classification using time-series Sentinel-2 imagery. Remote Sens 14(3):498. https://doi.org/10.3390/rs14030498

Tarasiou M, Chavez E, Zafeiriou S (2023) ViTs for SITS: vision transformers for satellite image time series. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Teimouri M, Mokhtarzade M (2023) Investigating three-dimensional convolutional and recurrent neural networks for crop classification using time-series optical images. J Geomat Sci Technol 12(3):1–15 (in Persian)

Teimouri M, Mokhtarzade M, Baghdadi N, Heipke C (2022) Fusion of time-series optical and SAR images using 3D convolutional neural networks for crop classification. Geocarto Int. https://doi.org/10.1080/10106049.2022.2095446

Thenkabail P, GangadharaRao P, Biggs T, Krishna M, Turral H (2007) Spectral matching techniques to determine historical land-use/land-cover (LULC) and irrigated areas using time-series 0.1-degree AVHRR pathfinder datasets. Photogramm Eng Remote Sens 73(10):1029–1040

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Voelsen M, Teimouri M, Rottensteiner F, Heipke C (2022) Investigating 2D and 3D convolutions for multitemporal land cover classification using remote sensing images. ISPRS Ann Photogramm Remote Sens Spatial Inform Sci 2022:271–279. https://doi.org/10.5194/isprs-annals-V-3-2022-271-2022

Voelsen M, Lauble S, Rottensteiner F, Heipke C (2023)Transformer models for multitemporal land cover classification using remote sensing images. In: ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences VI-3

Vuolo F, Neuwirth M, Immitzer M, Atzberger C, Ng W-T (2018) How much does multi-temporal Sentinel-2 data improve crop type classification? Int J Appl Earth Obs Geoinf 72:122–130

Wurm M, Stark T, Zhu XX, Weigand M, Taubenböck H (2019) Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. ISPRS J Photogramm Remote Sens 150:59–69. https://doi.org/10.1016/j.isprsjprs.2019.02.006

Xia T, He Z, Cai Z, Wang C, Wang W, Wang J, Hu Q, Song Q (2022) Exploring the potential of Chinese GF-6 images for crop mapping in regions with complex agricultural landscapes. Int J Appl Earth Obs Geoinform 107:102702. https://doi.org/10.1016/j.jag.2022.102702

Xu L, Zhang H, Wang C, Zhang B, Liu M (2018) Crop classification based on temporal information using sentinel-1 SAR time-series data. Remote Sens 11(1):53. https://doi.org/10.3390/rs11010053

You N, Dong J, Huang J, Du G, Zhang G, He Y, Yang T, Di Y, Xiao X (2021) The 10-m crop type maps in Northeast China during 2017–2019. Sci Data 8(1):41. https://doi.org/10.6084/m9.figshare.13567526

Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision

Zhang H, Li Y, Zhang Y, Shen Q (2017) Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. Remote Sens Lett 8(5):438–447. https://doi.org/10.1080/2150704X.2017.1280200

Zhao H, Duan S, Liu J, Sun L, Reymondin L (2021) Evaluation of five deep learning models for crop type mapping using Sentinel-2 Time Series Images with Missing information. Remote Sens 13(14):2790. https://doi.org/10.3390/rs13142790