RNN-based Multi-Source Land Cover mapping: An application to West African landscape *

Yawogan Jean Eudes Gbodjo[1], Louise Leroux[2], Raffaele Gaetano[4], and Babacar Ndao[5]

[1] IRSTEA, UMR TETIS, Univ. of Montpellier, France
jean-eudes.gbodjo@irstea.fr
[2] CIRAD, UPR AÏDA, Dakar, Sénégal
louise.leroux@cirad.fr
[3] CIRAD, UMR TETIS, Montpellier, France raffaele.gaetano@cirad.fr
[4] CSE, University Cheikh Anta Diop, Dakar, Sénégal
babacar.ndao@cse.sn

Abstract. In the southern countries, timely and accurate land cover mapping is crucial for food security monitoring. Nowadays, Earth Observation missions like Sentinel-1 (S1) and Sentinel-2 (S2) provide radar and optical imagery respectively, which can be organized in dense time series and leveraged for a wide range of applications such as land cover mapping. In this paper, a deep learning (DL) architecture is designed to combine S1 and S2 time series at object level with the aim to deal with heterogeneous agricultural landscape land cover mapping located in the southern part of the Senegalese groundnut basin. Both quantitative and qualitative results obtained demonstrate the significance of the proposal. In addition, we explore how the parameters learnt by the DL model can supply insights towards the explanation of the classifier decision.

1 Introduction

Nowadays, huge amount of heterogeneous Earth Observation (EO) data are made publicly available, and they represent a valuable source of information that can be easily leveraged for a wide range of land monitoring applications: agricultural management [14], ecology [13] or urban planning [11]. Among EO initiatives, the Copernicus programme developed by the European Space Agency provides radar and optical imagery through its Sentinel-1 (S1) and Sentinel-2 (S2) missions, respectively, with fine spatial resolution (up to 10-m) and high revisit time (around 5 days). These data, usually organized in Satellite Image Time Series (SITS), represent a practical tool to monitor human and physical environment through the production of precise and timely Land Use/Land Cover (LULC) maps. As regards LULC mapping, both S1 [19] (radar) and S2 [13] (optical) SITS have been employed. While their combination have shown to perform

* Supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 and the Programme National de Télédétection Spatiale grant no PNTS-2018-5.

better than using the single sensors in different contexts (e.g., change detection [6] and urban mapping [9]), how to profitably exploit multi-source data for LULC mapping remains a challenging task. In LULC-related contexts, most existing approaches rely on data fusion techniques [18] or on leveraging standard machine learning techniques (i.e. Random Forest, Support Vector Machine) on a simple concatenation of radar and optical input data [5]. In both cases, proposed methodologies threat the different data sources as completely independent from each other, also ignoring spatial and temporal dependencies that may be present in the data. Recently, deep learning (DL) approaches have become pervasive in several domains, including remote sensing [10]. A main attractive of DL models is that they are able to *learn* features optimized for a specific task (i.e. image classification), by simultaneously training the associated classifier. Moreover, they can be exploited to leverage temporal dependencies available in SITS data. Deep learning techniques tailored for multi-source (i.e. radar and optical) satellite data have been proposed to solve tasks such as optical image simulation [7] or change detection [15]. However, only marginal advances have been made in multi-source LULC mapping tasks [14]. Our hypothesis is that the complementarity carried out by radar and optical SITS can be effectively leveraged by DL based models compared to standard remote sensing techniques. In this context, we propose a deep learning architecture, named $OB2SRNN$ (Object-Based two-Stream RNN), to manage multi-temporal (SITS) and multi-source (radar and optical) data at object-level to deal with Land Cover mapping/classification with an application to a West African agricultural landscape. The proposed DL architecture involves the use of an attention-based RNN technique to effectively take into account time dependencies.

2 Data and Preprocessing

The analysis was carried out on the southern part of the Senegalese groundnut basin, one of the main agricultural regions of the country dominated by a smallholder agriculture in an heterogeneous landscape including isolated trees within plots. It covers a total area of 441 km^2 (21km×21km).

Sentinel-1 Data The radar dataset consists of 16 Sentinel-1 (S1) SITS acquired between May and October 2018 in C-band Interferometric Wide Swath (IW) mode with dual polarization (VV+VH). All images, as retrieved at level-1C Ground Range Detected (GRD) from the PEPS platform [5], are radiometrically calibrated in backscatter values (decibels, dB) using parameters included in metadata files and then coregistered with the Sentinel-2 (see below) grid and orthorectified at the same 10-m spatial resolution. Finally, a multitemporal filtering was applied removing artefacts resulting from speckle effect.

Sentinel-2 Data The optical data consists of a 19 Sentinel-2 (S2) SITS acquired between May and October 2018. All images are retrieved from the THEIA pole platform [6] and calibrated from digital number to level-2A top of canopy (TOC)

[5] https://peps.cnes.fr/
[6] http://theia.cnes.fr

reflectance. Only 10-m spatial resolution bands were considered (i.e. Blue, Green, Red and Near Infrared spectrum). Since the main issue with optical data, especially in tropical areas, is cloudiness, a preprocessing was performed over each band to replace cloudy observations as detected by the supplied cloud masks through a multi-temporal gapfilling [11]. Cloudy pixel values were linearly interpolated using the previous and following cloud-free dates. Then, the Normalized Difference Vegetation Index (NDVI) was calculated for each date. NDVI was chosen as supplementary optical descriptor since it describes the photosynthetic activity and the metabolism intensity of the vegetation.

Field data The field database was built from GPS records collected during the 2018 agricultural campaign and the visual interpretation of a very high spatial resolution (3-m) PlanetScope image acquired in October 4, 2018. The ground truth database includes 734 polygons distributed over 9 land cover classes (Table 1). The SITS analysis was conducted at object-level. Working with objects instead of pixels has two main advantages : i) objects represent more representative spectral information and potentially feature-rich pieces of information and ii) object based approaches facilitate data analysis scale-up since, for the same area, the number of objects is usually smaller than the number of pixels by several order of magnitude. To analyse SITS data at object-level, a segmentation was performed on the PlanetScope image which has been coregistered with the S2 time series. The PlanetScope image was segmented via the Large Scale Generic Region Merging (LSGRM) Orfeo Toolbox remote module obtaining 116 937 segments.

The obtained segments were spatially intersected with the ground truth data to provide radiometrically homogeneous class samples and finally it resulted in new comparable size 3 084 labeled objects. Finally, the mean value of the pixels corresponding to each segment was calculated over all the timestamps in the time series, resulting in 127 variables per segment (19×5 for S2 $+ 16 \times 2$ for S1).

Table 1: Field database characteristics

Class	Label	Polygons	Segments
0	*Bushes*	50	100
1	*Fallows*	69	322
2	*Ponds*	33	59
3	*Banks and bare soils*	35	132
4	*Villages*	21	767
5	*Wet areas*	22	156
6	*Valley*	22	56
7	*Cereals*	260	816
8	*Legumes*	222	676
Total		734	3084

3 Object based Multi-Source RNN Land Cover Classification

Figure 1 depicts the proposed $OB2SRNN$ deep learning architecture for the multi-source SITS classification process. The architecture involves two twin streams: one for the radar and one for the optical time series. Each stream of the $OB2SRNN$ architecture can be roughly decomposed in three parts: i) data preprocessing and enrichment ii) time series analysis and iii) multi-temporal combination to

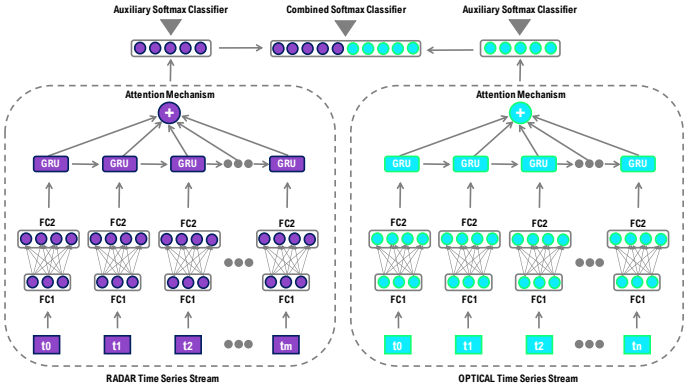

Fig. 1: *OB2SRNN* takes as input two time series (radar and optical SITS) and provides as output the land cover classification. It is composed by two twin streams. Each stream firstly processes the SITS by means of fully connected layers and successively an RNN with attention is employed. Finally, the set of features extracted is combined to provide the final decision.

generate per-source features. Finally, the radar (left stream) and optical (right stream) information are combined together. The features extracted from each stream (rectangles in top) are concatenated and directly leveraged to produce the final land cover classification. Such learned features named $feat_{rad}$ (resp. $feat_{opt}$) indicate the output of the radar (resp. optical) stream. The first part of each stream is represented by two fully connected layers (FC1 and FC2) that take as input one time stamp of the object time series (radar or optical) and combine the input data. Such stage allows the architecture to extract an useful input combination for the classification task enriching the original data representation. A ReLU non-linearity activation function [17] is associated to each fully connected layer. The second part is constituted by Gated Recurrent Units (GRUs) [3], a kind of Recurrent Neural Networks (RNNs) which have demonstrated its effectiveness in the field of remote sensing [1,16] among others. Unlike standard feed forward networks (i.e. CNNs), RNNs explicitly manage temporal data dependencies since the output of the neuron at time $t-1$ is used, together with the next input, to feed the neuron itself at time t. Furthermore, this approach explicitly models the temporal correlation presents in the object time series and is able to focus its analysis on the useful portion of the time series (i.e., discarding less useful information). The third and last stage consists of a neural attention model [2] on top of the outputs produced by the GRUs. Attention mechanisms are widely used in automatic signal processing [2] (1D signal or language) as they allow to join together the information extracted by the RNN model at different timestamps via a convex combination of the input sources. The attention formulation used is the same as in [1]. The purpose of this procedure is to learn a set of weights that allows the contribution of each timestamp

to be weighted through a linear combination. The $SoftMax$ function is used to normalize weights λ so that their sum is equal to 1. In our context, two different sets of attention weights are obtained: λ_{rad} and λ_{opt}. The former refers to the attention over the radar SITS while the latter represents the attention over the optical ones. The output of each stream is a feature vector encoding the temporal information related to input sources. Once each stream has processed the corresponding time series information, the concatenation of the extracted radar and optical features is used to perform the classification. To strengthen the complementarity as well as the discriminative power of the learned features for each stream, we use an adaptation of the technique proposed in [8], by introducing an auxiliary classifier for each set of learned features ($feat_{rad}$ and $feat_{opt}$). The goal of this extra classifiers is to stress the fact that the learned features need to be discriminative alone (i.e., independently from each other). Then, the learning process involves the optimization of three classifiers at the same time: one specific to $feat_{rad}$, one related to $feat_{opt}$ and one that considers the concatenation $[feat_{rad}, feat_{opt}]$. The associated cost function is :

$$L_{total} = 0.5 * L_1(feat_{rad}) + 0.5 * L_2(feat_{opt}) + L_{fus}([feat_{rad}, feat_{opt}]) \quad (1)$$

where $L_i(feat)$ is the loss function (in our case the categorical Cross-Entropy) associated to the classifier fed with the features $feat$. The contribution of each auxiliary classifier was empirically weighted by a weight of 0.5 to enforce the discriminative power of the per-source learned features. The final land cover class is derived combining the three $SoftMax$ classifiers with the same weight schema employed in the learning process. In addition, dropout with a rate equal to 0.4 was employed for the GRU unit and between the two Fully Connected layers. The model is learned end-to-end from scratch.

4 Experimental Evaluation

In this section, we present and discuss the results obtained on the study site. To assess the quality of $OB2SRNN$, we compare its performance with those of Random Forest (RF) classifier. The competing method, namely $RF(S1, S2)$, consists in a Random Forest classifier trained on the concatenation of S1 and S2 SITS. In addition, we provide an inspection of the attention parameters learnt by $OB2SRNN$, to investigate how such side information can be exploited to get insights from the data itself.

Experimental Settings To learn $OB2SRNN$ parameters the Adam optimizer [12] was used with a learning rate equal to 1×10^{-4}. The training process was conducted over 1000 epochs with a batch size equal to 32. The number of hidden units for the RNN module was fixed to 512 (resp. 256) for the optical (resp. radar) branch while, 16 and 32 (resp. 32 and 64) neurons were employed for the first and the second Fully Connected layers for S1 (resp. S2) stream. The $RF(S1, S2)$ model was optimized using a grid search procedure on the maximum depth of each tree (in the range {20,40,60,80,100}) and the number of trees in

the forest (in the set {100, 200, 300,400,500}). The dataset was split into training, validation and test set with an object proportion of 50%, 20% and 30% respectively. The values were normalized, per band (resp. indices) considering the time series, in the interval $[0, 1]$. The assessment of the classification performances was done considering global precision (*Accuracy*), *F-Measure* and *Kappa* metrics. Training set was used to learn the $OB2SRNN$ and $RF(S1, S2)$ models. For each method, the model achieving the best accuracy on the validation set was subsequently employed to classify the test set. Since the performances of the models may vary depending on the split of the data due to simpler or more complex samples involved in the training/test set, all metrics were averaged over ten random splits following the strategy previously reported.

Comparative analysis Table 2 reports on the average results of the evaluation metrics for the two competing methods. Considering the average behavior, $OB2SRNN$ clearly outperforms $RF(S1, S2)$, with a gain of more than 5 points on each metric. Table 3 reports the per-class *F-Measure* obtained by each method. The $OB2SRNN$ approach achieves the best performance on 7 classes over 8. This behavior is particularly accentuated for 0–*Bushes*, 1–*Fallows and Uncultivated areas*, 2–*Ponds*, 5–*Wet areas*, 7–*Cereals* and 8–*Legumes* classes. The biggest gap between the two methods concerns the 2–*Ponds* class, corresponding to a gain of more than 30 points. Given that the *Ponds* class is one of the less represented in terms of examples in the ground truth, the fact that Random Forest is known to be sensible to class imbalance [10] (e.g. giving more chance to more representative classes like 5–*Wet areas*) can be a possible explanation behind its poor behavior on such land cover. On the other hand, this result demonstrates the ability of $OB2SRNN$ to effectively leverage on the different information sources, thus being able to deal with the land cover classification task even in a class imbalance scenario. It should be noted that the imbalance in the data also affects the performance on the over represented classes, i.e., $OB2SRNN$ largely outperforms $RF(S1, S2)$ also on 5–*Wet areas*, i.e., a class closely related to 2–*Ponds* but more represented in the input data. Concerning agricultural classes (i.e., 7–*Cereals* and 8–*Legumes*) the RNN based approach achieves better scores due to the fact that it is especially tailored to leverage temporal dependencies. This behavior particularly suits our context, where crop classes follow specific temporal patterns due to cropping practices, and depending on the the timing of the different phenological stages in which the plants are sowing, grow, reach maturation and are successively harvested. Regarding other classes, the gap between the performances of two methods are less significant, however, the $RF(S1, S2)$ method achieves a slightly better but still comparable score on class 3–*Banks and bare soils*.

Table 2: F-Measure, Kappa and Accuracy considering the competing methods.

	F-Measure	*Kappa*	*Accuracy*
$RF(S1, S2)$	81.61 ± 1.22	0.77 ± 0.02	81.80 ± 1.24
$OB2SRNN$	**87.04** ± 0.93	**0.84** ± 0.01	**87.02** ± 0.92

Furthermore, Figures 2 and 3 report the confusion matrices of the different approaches for the classification task at hand. Overall, the heat map (Figure 2) related to $OB2SRNN$ has a more visible diagonal structure than the one depicting the results produced by $RF(S1, S2)$ (Figure 3) (the darker red blocks concentrated on the diagonal). On the other hand, RF exhibits major confusions on class 2–*Ponds*, which is often wrongly classified as 5–*Wet areas*, mainly due to similar spectral characteristics. This fact can also explain the poor performance of $RF(S1, S2)$ on the *Ponds* class (see Table 3). We can also note that RF tends to overestimate 7–*Cereals*, so decreasing its effectiveness on 1–*Fallows and Uncultivated areas*, 2–*Ponds* and 8–*Legumes*.This behavior can be explained by the fact that RF does not leverage the temporal dependencies available in the time series, i.e., it may take some of its decision by considering particular timestamps in which such classes exhibit similar characteristics. To sum up, we can state that, in most cases, $OB2SRNN$ shows a more effective behavior with respect to RF, considering any of the classes involved in the land cover mapping task.

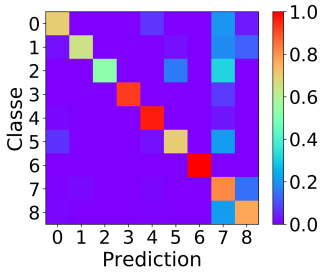

Fig. 2: Obtained confusion matrix for $RF(S1, S2)$

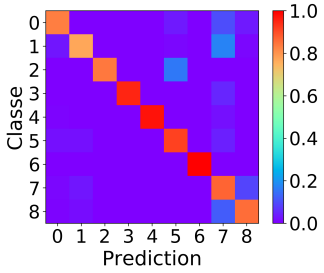

Fig. 3: Obtained confusion matrix for $OB2SRNN$

Table 3: Per-Class *F-Measure* of the competing methods (average over ten different random splits). Class definitions are reported in Table 1.

Method	0	1	2	3	4	5	6	7	8
$RF(S1, S2)$	65.90	75.29	56.69	**91.57**	96.58	73.67	90.00	76.49	77.51
$OB2SRNN$	**77.16**	**81.00**	**84.16**	91.37	**98.52**	**86.77**	**92.13**	**81.50**	**83.99**

Qualitative inspection of LULC maps Figure 4 reports two representative details of the land cover maps produced by $RF(S1, S2)$ and by $OB2SRNN$. We remind that land cover maps are produced by labeling each of the 116 937 segments obtained by the segmentation process. The first detail (Figures 4a to 4c) depicts a wet area close (in the North) to the Toukar village, a medium-size urban area. We have observed that $RF(S1, S2)$ clearly underestimates the presence of 5–*Wet areas* class (purple) and tends to confuse it with 7–*Cereals* class. Conversely, $OB2SRNN$ exhibits a more effective behavior on such complex land cover class detection. The second detail (Figures 4d to 4f) focuses on fallows located near (in the West) the Diohine village. Also in this case, $OB2SRNN$ allows a better detection of 1–*Fallows and Uncultivated areas* class than $RF(S1, S2)$

which confuse it with 8–*Legumes* class. In addition, $RF(S1, S2)$ also overestimates the real extent of the villages in the expense of 7–*Cereals* class. Overall, the qualitative inspection is in concordance with the quantitative results in terms of evaluation metrics obtained above.

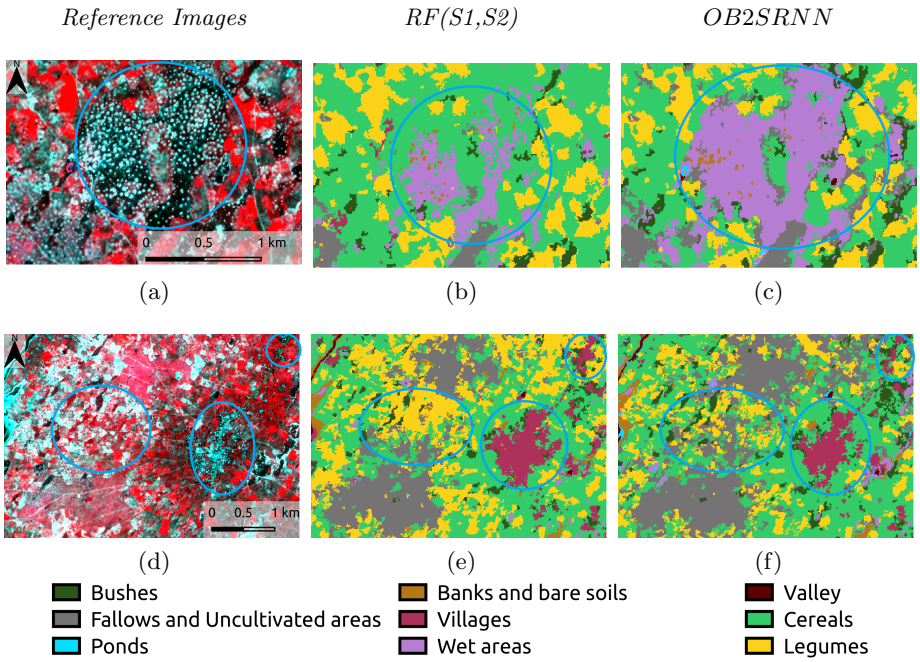

Fig. 4: Qualitative analysis of land cover maps produced by $RF(S1, S2)$ and $OB2SRNN$ in 2 different zones (top: focus on a wet area, bottom: focus on fallow and uncutivated areas). The references supplied come from a PlanetScope image acquired on October 4, 2018.

Inspection of the Attention parameters In this experiment we inspect the side information provided by our model via the analysis of the attention parameters. As discussed in Section 3, $OB2SRNN$ learns two set of parameters, λ_{rad} and λ_{opt}. The attention parameters are usually employed, in the field of NLP [2,4], to explain which are the contributions to the final decision of the different parts of the signal. With the aim to set up a similar investigation in our context, we depict in Figure 5 the different λ values considering the two sources of information via a box plots visualization. For each SITS, we consider its length (16 for the radar SITS and 19 for the optical SITS) and, for each timestamps, we collect the λ values learned by the model, leveraging box plots to draw values distribution. At first look, we can observe that the attention distributions, considering both data sources, are skewed towards the last portion of the time series. Such behavior seems to indicate that the majority of the information is contained within the second part of the considered period. This is inline with the

agronomic knowledge about the study area since, considering the year spanned by the time series (2018), in the southern part of the Senegalese groundnut basin, most of the activity related to agriculture happens between the period August - October with the maximum vegetation (chlorophyll) activity peak occurring at the end of August. Such period is characterized by an heavy rain activity. This characteristic induces sharp contrasts among agricultural and non agricultural land cover classes and differences in the canopy structures. Furthermore, regarding only agricultural classes: *Cereals* and *Legumes*, the former is harvested mid-September while the latter is harvested at the end of the considered Satellite Image Time Series. Both meteorological conditions and agricultural practices can explain why, according to the attention parameters analysis, the second portion of the time series contains the major part of the useful information.

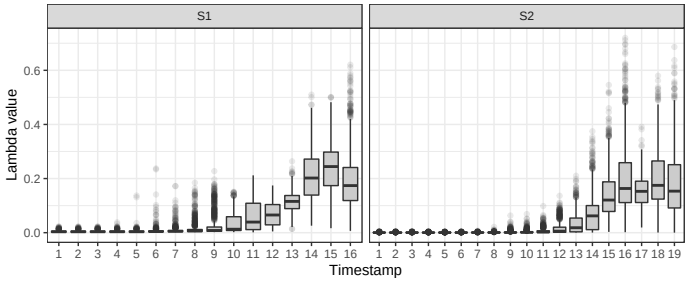

Fig. 5: Box Plots of the Attention parameters for the radar (S1) and the optical (S2) SITS considering different timestamps.

5 Conclusion

In this work, we introduced a deep learning architecture to perform land cover classification mapping, from multi-temporal and multi-source Satellite Image Time Series data. The proposed method, named $OB2SRNN$, outperforms standard remote sensing approaches and enables the possibility to explicitly leverage temporal as well as multi-source dependencies. Furthermore, we investigated how the attention parameters, learnt by our model, can be employed to perform temporal reasoning about the time series data, thus contributing to the interpretability of the decisions made by the neural network architecture.

References

1. Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R.G., Dupuy, S.: M3 fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion. IEEE JSTARS **11**(12), 4939–4949 (2018)
2. Britz, D., Guan, M.Y., Luong, M.: Efficient attention using a fixed-size memory representation. In: EMNLP. pp. 392–400 (2017)

3. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP. pp. 1724–1734 (2014)
4. Choi, H., Cho, K., Bengio, Y.: Fine-grained attention mechanism for neural machine translation. Neurocomputing **284**, 171–176 (2018)
5. Erinjery, J., Singh, M., Kent, R.: Mapping and assessment of vegetation types in the tropical rainforests of the western ghats using multispectral sentinel-2 and sar sentinel-1 satellite imagery. Remote Sensing of Environment **216**, 345–354 (2018)
6. Gao, Q., Zribi, M., Escorihuela, M.J., Baghdadi, N.: Synergetic use of sentinel-1 and sentinel-2 data for soil moisture mapping at 100 m resolution. Sensors **17**(9), 1966 (2017)
7. He, W., Yokoya, N.: Multi-temporal sentinel-1 and -2 data fusion for optical image simulation. ISPRS Int. J. Geo-Inf. **7**(10), 389 (2018)
8. Hou, S., Liu, X., Wang, Z.: Dualnet: Learn complementary features for image recognition. In: ICCV. pp. 502–510 (2017)
9. Iannelli, G.C., Gamba, P.: Jointly exploiting sentinel-1 and sentinel-2 for urban mapping. In: 2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018, Valencia, Spain, July 22-27, 2018. pp. 8209–8212 (2018)
10. Ienco, D., Gaetano, R., Dupaquier, C., Maurel, P.: Land cover classification via multitemporal spatial data by deep recurrent neural networks. IEEE Geosc. and Rem. Sens. Letters **14**(10), 1685–1689 (2017)
11. Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I.: Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sensing **9**(1), 95 (2017)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
13. Kolecka, N., Ginzler, C., Pazur, R., Price, B., Verburg, P.H.: Regional scale mapping of grassland mowing frequency with sentinel-2 time series. Remote Sensing **10**(8), 1221 (2018)
14. Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A.: Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci. Remote Sensing Lett. **14**(5), 778–782 (2017)
15. Liu, J., Gong, M., Qin, K., Zhang, P.: A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans. Neural Netw. Learning Syst. **29**(3), 545–559 (2018)
16. Mou, L., Ghamisi, P., Zhu, X.X.: Deep recurrent neural networks for hyperspectral image classification. IEEE Trans. on Geosc. and Rem. Sens. **55**(7), 3639–3655 (2017)
17. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML. pp. 807–814 (2010)
18. Sharma, R.C., Hara, K., Tateishi, R.: Developing forest cover composites through a combination of landsat-8 optical and sentinel-1 SAR data for the visualization and extraction of forested areas. J. Imaging **4**(9), 105 (2018)
19. Zhou, T., Zhao, M., Sun, C., Pan, J.: Exploring the impact of seasonality on urban land-cover mapping using multi-season sentinel-1a and GF-1 WFV images in a subtropical monsoon-climate region. ISPRS Int. J. Geo-Inf. **7**(1), 3 (2018)