TOWARDS LIFELONG CROP RECOGNITION USING FULLY CONVOLUTIONAL RECURRENT NETWORKS AND SAR IMAGE SEQUENCES

J. A. Chamorro[a,][*] R. Q. Feitosa[a], P. N. Happ[a], J. D. Bermudez[a]

[a] Dept. of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil

KEY WORDS: Fully convolutional recurrent networks, crop recognition, SAR, remote sensing

ABSTRACT:

Recent works have studied crop recognition in regions with highly complex spatio-temporal dynamics typical of a tropical climate. However, most proposals have only been evaluated in a single agricultural year, and their capabilities to generalize to dates outside the temporal sequence have not been properly addressed thus far. This work assesses the generalization capabilities of a recent convolutional recurrent architecture, testing it in a temporal sequence two years ahead of the sequence with which it was trained. Furthermore, a N-to-1 variant of such network is proposed, which is able to produce classification outcomes for every month in the agricultural year, and it is compared with two baselines designed in a more traditional approach, in which a separate specific network is trained for each month of the year. The approaches are evaluated on two public datasets from a tropical region. The first dataset comprehends the period from June 2017 to May 2018, while the second goes from October 2019 to September 2020. Results show a decrease of up to 24.6% in per-date average F1 score when training the network with data of an agricultural year different from the one it is tested on, which indicates a domain shift that demands further research. Additionally, the proposed approach presented only a slight decrease in performance compared to its baseline when trained on the same dataset, with a 2.7% drop in average F1 score. This performance drop is a small cost in exchange for its operational advantages, such as reduced training time and a more straightforward pipeline.

1. INTRODUCTION

Crop monitoring involves several factors that might vary from one region to another. In tropical areas, the favorable climate allows more flexibility regarding the seeding and harvest times, making the modeling of crops' dynamics more complex (Sanches et al., 2018b). Notably, tropical areas generally exhibit complex spatio-temporal dynamics of culture because of their favorable climate and crop rotation practices. In Brazil, for example, the favourable climate allows up to 3 harvests per year, and the plantation and harvesting dates for the same crop type may greatly vary. Thus, it is necessary to map crops in tropical regions several times throughout the year, unlike temperate areas, where a single classification result may be sufficient for the entire agricultural year.

Several works have addressed the crop recognition problem using classical machine learning techniques. In (Tardy et al., 2017), the authors used a Random Forest algorithm for multi-temporal crop mapping and evaluated fusion techniques to leverage information from multiple past agricultural years to classify unlabeled data from the current period. The authors in (Ajadi et al., 2021) proposed a large scale crop mapping scheme leveraging auxiliary data from the United States, using a XG-Boost model. Similarly, the authors in (Santos et al., 2021) used self-organizing maps (SOMs) to classify crop types using MODIS time series. However, their training features only considered the temporal dimension, leaving the spatial context unexplored.

In recent years, deep learning techniques have been successfully applied to crop recognition applications (Audebert et al.,

2017). Such methods can be roughly grouped into Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). A recent work compared a CNN and a more spatially efficient Fully Convolutional Network (FCN) for crop recognition in a tropical climate in the Municipality of Campo Verde, Brazil (La Rosa et al., 2018). Despite the reported high accuracy values, these approaches require specific training for each date in which a result is desired, which represents an operational disadvantage.

The second group of techniques involves recurrent neural networks (RNNs). A type of RNN called ConvLSTM was proposed by (Shi et al., 2015), which replaces the LSTM internal operations with convolutional layers, allowing it to consider the spatio-temporal context in the input sequence. ConvLSTM networks have recently been used for crop recognition applications (Rußwurm and Körner, 2018, Teimouri et al., 2019, Martinez et al., 2021). (Rußwurm and Körner, 2018) reports an evaluation of a ConvLSTM network for crop recognition in a temperate region. (Martinez et al., 2021) proposed a network that combines a ConvLSTM with an Unet-like encoder-decoder network (UNetConvLSTM). The authors tested the proposed network in public datasets from a tropical climate. This network was trained as a unique, end-to-end network and can produce classification results for every date represented in the temporal sequence.

To our knowledge, all previous works evaluate the network performance in the same temporal sequence used for training. Such an analysis is insufficient to assess how well the networks generalize on dates outside the period covered by the training's temporal sequence.

The first contribution of this work is an evaluation of a Unet-ConvLSTM network at dates not used for training. A second contribution is a novel N-to-1 variant of the UnetConvLSTM

---

* Corresponding author

network from (Martinez et al., 2021), which is trained in a sliding window manner to produce outcomes for any date along the temporal sequence using a single, end-to-end network called VUnetConvLSTM.

This work also evaluates the performance of the proposed network at dates outside the temporal sequence used for training. We used two open-access datasets for this purpose. Both refer to the municipality of Luis Eduardo Magalhães, Bahia state, Brazil. The dataset refers to two temporal sequences separated in about one year from each other. The first dataset, called hereafter LEM 17/18, comprises SAR images acquired between 2017 and 2018. We used this dataset for training. The second dataset, called henceforth LEM 19/20, comprehends a temporal sequence between 2019 and 2020, and it was used for testing.

The remainder of this work is organized in the following sections. Section 2 briefly explains some fundamental concepts required to understand the evaluated networks, such as LSTM, ConvLSTM and UnetConvLSTM. Section 3 presents the proposed VUnetConvLSTM architecture. Section 4 details the datasets used in this work, the baseline architectures, and the experimental protocol. In Section 5, the results are presented and discussed. Finally, Section 7 describes the final conclusions and discusses possible future directions.

2. FUNDAMENTALS

2.1 LSTM and ConvLSTM

Recurrent Neural Networks are a class of neural networks designed to handle data that exhibits a time/sequence dependency. The RNN's output at a specific time step t, depends on the data from the current data observation x_t and the information stored in the network's hidden state h_t, which is updated at each time step. A modern type of RNN called Long Short-Term Memory (LSTM) introduces trainable gates in its architecture to control the information flow of the network's memory c_t (Hochreiter and Schmidhuber, 1997). In particular, gates are shallow neural networks with *sigmoid* activation function as output and parameterized by W_f, W_i, and W_o, respectively, which regulate the information accessed, cleared, and added to c_t (see Figure 1). This architecture allows LSTMs to be capable of modeling both long and short-term time dependencies.
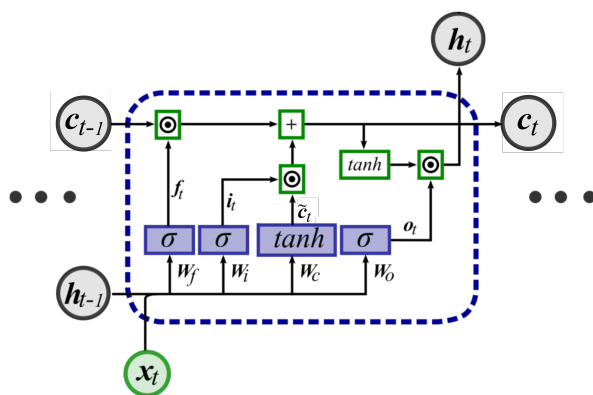

Figure 1. LSTM structure diagram.

The major drawback of LSTMs for image sequence processing is that their internal operations are fully connected layers, which means that their input, hidden states, and output correspond to

sequences of vectors. To overcome this issue, (Shi et al., 2015) proposed the ConvLSTM, which replaces all the LSTM internal fully connected layers with convolutional layers. As a result, its input, hidden states, and output correspond to sequences of tensors with spatial dimensions.

2.2 UnetConvLSTM

This architecture was introduced in (Martinez et al., 2021). It combines the ConvLSTM with the Unet network (Ronneberger et al., 2015), which considers the spatial context at multiple spatial scales. The block diagram for this network is presented in Figure 2. The input to this network is a sequence of images. First, an encoder stage is applied to each of the images in the sequence. This encoder consists of successive convolutional layers and downsampling operations, which extract more coarse feature representations. The resulting sequence of feature maps produced by the encoder is then passed to a unidirectional ConvLSTM network, which effectively considers the spatio-temporal context. The ConvLSTM is configured as N-to-1, and thus only the last element in the sequence produced at the RNN output is considered. This feature map is then passed to a decoder stage. In the decoder stage, successive convolutional layers and upsampling operations are applied to recover the original spatial resolution. Skip connections are used to preserve fine details. Afterward, a convolutional layer with a 1x1 filter and a softmax activation function produces the final class probabilities. At inference, the patch outcomes in the test area are stitched together in a mosaic to form the final result.
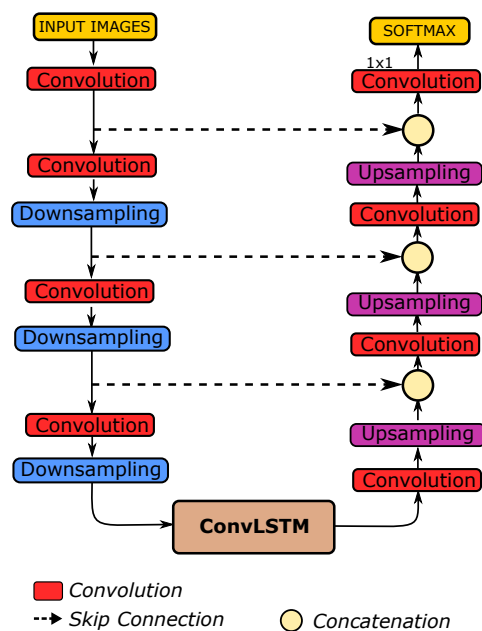

Figure 2. UnetConvLSTM architecture. Input is a sequence of images. The output corresponds to the class probabilities predicted for each pixel location in the last image.

3. PROPOSED VARIABLE SEQUENCE NETWORK

The majority of works on multi-date crop recognition for tropical regions are designed to produce classification outcomes for a single date in the dataset (Castro et al., 2017, La Rosa et al., 2018, Martinez et al., 2019). In those works, separate network

architectures need to be trained for each date to be predicted. However, this results in operational disadvantages such as larger training times and a more complex pipeline.

The proposed variable sequence network, called VUnetConvLSTM, attempts to solve this problem by training a unique, end-to-end architecture capable of producing classification outcomes for every month of the year. The UnetConvLSTM network is configured with an input sequence of length 12 with one image per month, corresponding to a year before the desired output month. Therefore, the network's output corresponds to the class probabilities for the last image in the sequence. During training, the network is trained with a different target month from the available labeled dates in each mini-batch. Considering our works focus on obtaining the classification outcomes for the last date in the sequence, we did not use the bidirectional variant because our model does not employ future information.

At test time, the network is capable of producing outcomes for every month of the year. Thus, this architecture addresses the issues presented in the aforementioned previous works, associated with the need to train separate networks for each month of the year. Nonetheless, this approach may result in a performance decrease, given that the network needs to learn the specific classification patterns for the entire year.

For each image in the input sequence, the day of the year was added as metadata for the network to better understand the sequence's temporal structure. First, the sine and cosine operations were computed on the day of the year to preserve its cyclical nature. The resulting values are of shape $T \times 2$, where T is the sequence length. Spatial dimensions were added to this representation, resulting in a sequence of shape $T \times 1 \times 1 \times 2$. The result was spatially upsampled, concatenated with the network encoder's outcomes, and then used as an input to the ConvLSTM block. The modified UnetConvLSTM is presented in Figure 3.

4. EXPERIMENTS

4.1 Study Area

The proposed approaches were evaluated in two publicly available multi-date crop recognition datasets with complex spatio-temporal dynamics due to their tropical climate. Both datasets are located in Luis Eduardo Magalhães municipality, Bahia state, Brazil (See Figure 4). The first dataset, called LEM 17/18 (Sanches et al., 2018a), comprises a monthly sequence of 12 pre-processed Sentinel-1 images and their ground truth labels, corresponding to the period between June 2017 and May 2018. In this work, the dataset was extended with additional monthly images, for a total of 21 monthly images from September 2016 to May 2018 (See Table 1). Reference information is not available for these additional images.

The second dataset, called LEM 19/20, consists of 12 pre-processed Sentinel-1 images from October 2019 to September 2020, together with their corresponding ground truth (Oldoni et al., 2020). As in the previous case, this dataset was extended by using additional monthly images, without references, from November 2018 until the beginning of the dataset (See Table 2). Although it covers a larger area compared to the LEM 17/18 dataset, we cropped its input and ground truth images to match the same extension of LEM 17/18. The class distribution for each month in both cases is presented in Figure 5.
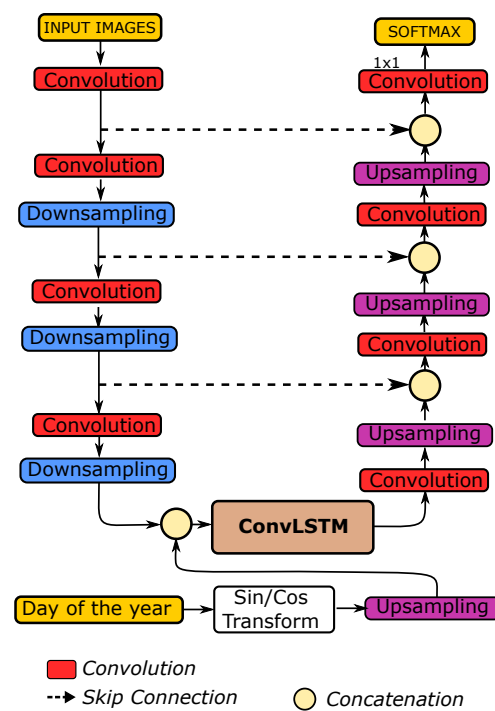

Figure 3. Modified UnetConvLSTM architecture, with the images' day of the year as additional input. The output corresponds to the class probabilities predicted for each pixel location in the last image.
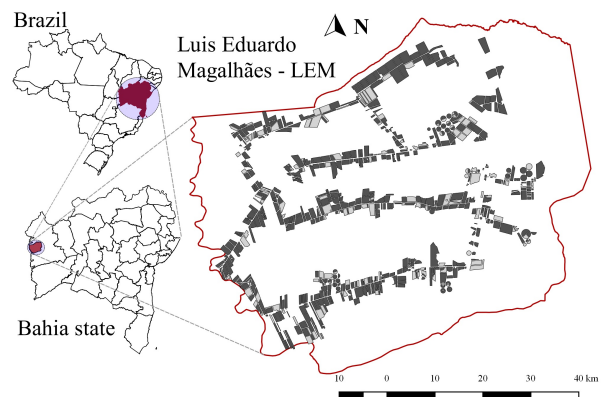

Figure 4. Luis Eduardo Magalhães (LEM) municipality. Polygons correspond to the ground truth parcels in LEM 17/18 dataset.

Experiments were carried out considering the most representative classes. We grouped classes with an overall percentage of samples lower than 5.3% into a single class called *Other classes*. In all cases, the Sentinel-1 images were resampled to a 10m resolution, and the VV and VH bands were used. No specific filtering was considered towards reducing speckle noise.

4.2 Baseline

We compared the proposed approach with two baseline methodologies. In the first one, called Baseline 1, we trained separate UnetConvLSTM networks specifically for each month of the year. The input sequence length is the same in each of these networks, corresponding to an input sequence of one year in the past with respect to the desired output month. At inference, outcomes for each month are produced using each of the separ-
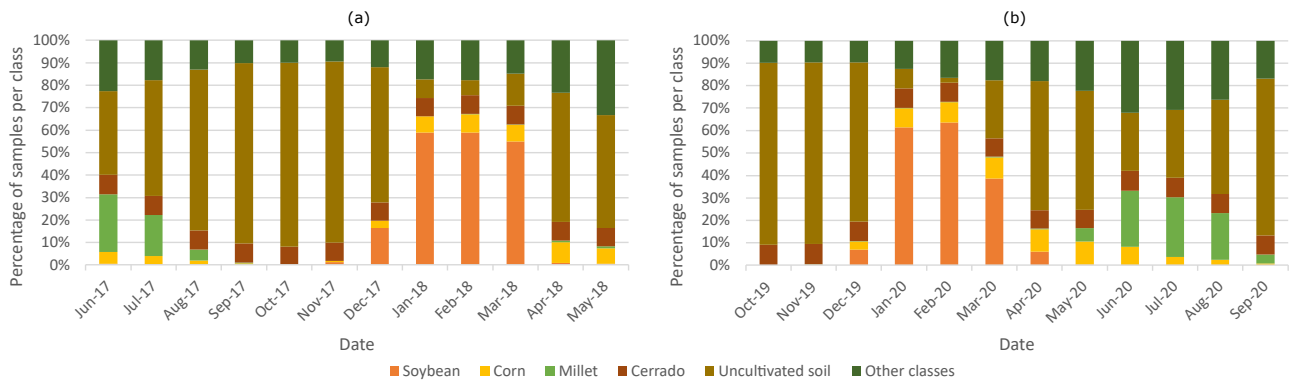
Figure 5. Class percentages per date for (a) LEM 17/18 and (b) LEM 19/20 datasets

Year	Month	Date	Labeled
2016	September	27	
	October	15	
	November	20	
	December	14	
2017	January	19	
	February	12	
	March	08	
	April	13	
	May	19	
	June	12	✓
	July	06	✓
	August	11	✓
	September	16	✓
	October	10	✓
	November	15	✓
	December	09	✓
2018	January	14	✓
	February	19	✓
	March	15	✓
	April	20	✓
	May	14	✓

Table 1. Acquisition dates for LEM 17/18 dataset, comprising 21 images, whereby the ground truth is available for the last 12 dates.

Year	Month	Date	Labeled
2018	November	10	
	December	16	
2019	January	21	
	February	14	
	March	22	
	April	15	
	May	21	
	June	14	
	July	20	
	August	13	
	September	18	
	October	12	✓
	November	17	✓
	December	23	✓
2020	January	16	✓
	February	21	✓
	March	16	✓
	April	21	✓
	May	15	✓
	June	20	✓
	July	14	✓
	August	19	✓
	September	12	✓

Table 2. Acquisition dates for LEM 19/20 dataset, comprising 23 images, whereby the ground truth is available for the last 12 dates.

ately trained networks. The networks were trained and tested in the same LEM 19/20 dataset. Because this approach does not need to generalize to unseen dates at test time, it is expected to outperform the remaining approaches, and we used it as an upper bound for comparison purposes.

As in the previous approach, Baseline 2 consists of multiple identical UnetConvLSTM networks, each one trained for a specific month of the year. Input continues to be a sequence of images corresponding to one year in the past with respect to the corresponding output month. In this case, the networks are trained on the LEM 17/18 dataset, while at inference the networks are tested on the LEM 19/20 dataset. Thus, this model evaluates the baseline's generalization capabilities in the temporal domain. One should expect Baseline 2 to perform better than our proposed network because it trains a specialized classifier for each specific date. Notice that the advantage of the proposed model compared to this baseline is operational because it requires a unique network to produce classification outcomes for all dates.

4.3 Experimental protocol

For the UnetConvLSTM network, we used average pooling as downsampling operator, and transposed convolution as up-

sampling operator. Table 3 presents the filter number we used at every stage.

In each dataset, agricultural crop parcels were represented as polygons. We used 100% of the LEM 17/18 dataset polygons for training the Baseline 1 and the proposed approach. In LEM 19/20, the polygons intersecting with the train area of LEM 17/18 were used for training, and the remaining polygons were used for testing. The train and test areas for both datasets are presented in Figure 6 and Figure 7.

Patches of spatial dimensions 32-by-32 were used, with an input sequence length of 12. We used zero padding for the input images that were not available and were required during the training of the LEM 17/18 early dates. Such unavailable images correspond to the months of July and August 2016.

We used focal loss cost function and Adam optimizer with a learning rate of 0.001. Mini-batches of 16 samples were used during training. The networks were trained on an NVIDIA GTX 1080Ti GPU.[1]

---

[1] Code available at `https://github.com/DiMorten/FCN_ConvLSTM_Crop_Recognition_Generalized`

Layer	Processing
Input	$T \times H \times W \times D$
Conv	$T \times H \times W \times 16$
Downsampl.	$T \times H/2 \times W/2 \times 16$
Downsampl.	$T \times H/4 \times W/4 \times 32$
Downsampl.	$T \times H/8 \times W/8 \times 64$
ConvLSTM	$T \times H/8 \times W/8 \times 256$
Upsampl.	$H/4 \times W/4 \times 64$
Upsampl.	$H/2 \times W/2 \times 32$
Upsampl.	$H \times H \times 16$
Conv	$H \times H \times 16$
Conv	$H \times H \times C$

Table 3. Parameter configuration used in all experiments. T is the temporal sequence length, H and W indicate height and width, D is the input dimensionality, and C is the class number (H=32, W=32, D=2, C=6). In VUnetConvLSTM, the day of the year was used as an additional input, which was concatenated to the input of the ConvLSTM layer.
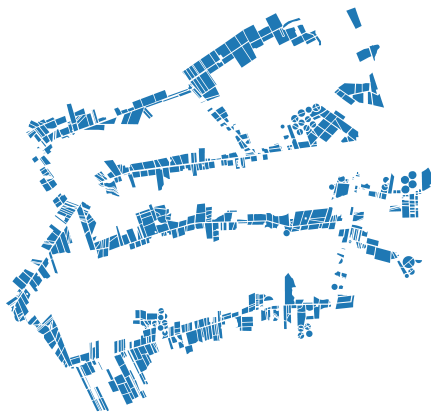

Figure 6. Reference crop parcels in LEM 17/18 dataset. All the polygons were used for training the Baseline 2 and VUnetConvLSTM networks
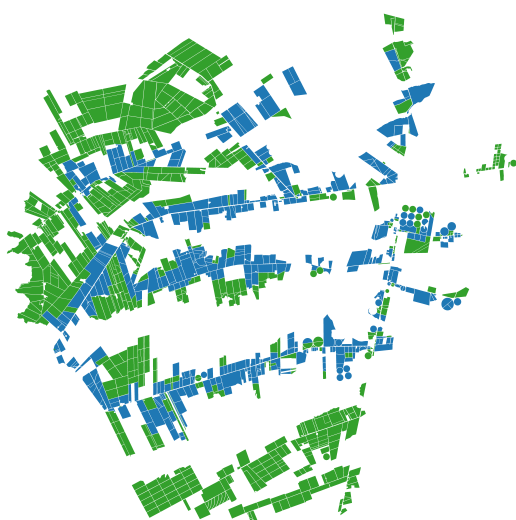

Figure 7. Reference crop parcels in LEM 19/20 dataset. Blue polygons were used for training the Baseline 1. Green polygons were used for testing the proposed networks.

5. RESULTS

Figure 8 and Figure 9 summarize the performance of the assessed methods in terms of per-date Overall Accuracy (OA) and F1-score, respectively. The corresponding rates were obtained by evaluating the models on the LEM 19/20 dataset on the sequence from October 2019 to September 2020. As expected, Baseline 1 produced the highest OA metrics with values up to 93.3%. Even so, it produced relatively low OA values for some months such as March, with an OA value of 58.2%.

The other two methods, Baseline 2 and VUnetConvLSTM, performed similarly for most dates, but produced lower OA values than Baseline 1. However, the drop in performance regarding Baseline 1 is not constant over the dates for both methods. The differences are more remarkable in January, February, and between May and July, where the decrease can be up to 39.9%. These results are understandable considering the time-shift of two years between the training and test sets, which is also accentuated by the high dynamics that characterize the tropical regions. In this sense, domain adaptation techniques would help in reducing the gap between both datasets.

Recall that Baseline 2 involves specific classifiers for each date, while VUnetConvLSTM consists of a single classifier for all dates. Note that the performance gap between these two approaches was relatively small, with an average reduction of 2.7% in the F1 score.

On some dates, the proposed VUnetConvLSTM outperformed Baseline 2 in terms of F1 score and OA, e.g., in May. On this date, there were few training samples of Class Millet for LEM 17/18 dataset, which might have led to a low F1 score at that date for Baseline 2. In contrast, the proposed network leveraged the labeled data from other dates to learn a proper representation for Millet, leading to a better F1 score in May. The same behaviour occurred in August, where the Millet percentage was 4.9% in the LEM 17/18 dataset and 14.7% in the LEM 19/20 dataset.
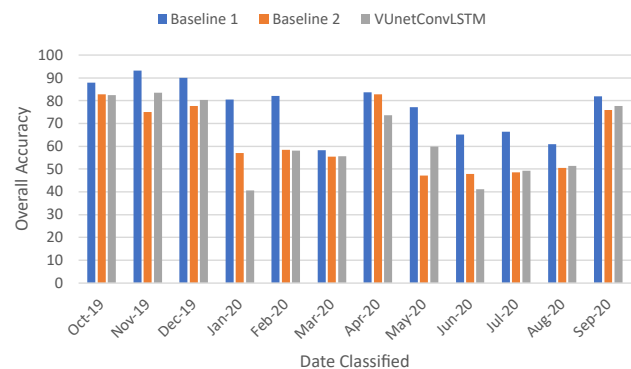

Figure 8. Per-date Overall Accuracy for the proposed approaches.

Results regarding the F1 score metric are consistent with the ones reported in terms of OA. For most of the dates, the highest F1 scores were obtained with Baseline 1, followed by Baseline 2 and VUnetConvLSTM approaches, which performed similarly on most dates. However, it can be noticed that the decrease in performance for both methods, regarding Baseline 1, is significant for the entire sequence. An exception occurred in April, in which Baseline 2 outperformed Baseline 1 by a margin of 7.5%. For Baseline 2, May was the most affected month, with a

drop of 31% in F1 score, while for VUnetConvLSTM, the major reduction occurred in July and August, with a decrease of 24.1% and 24.6% respectively.
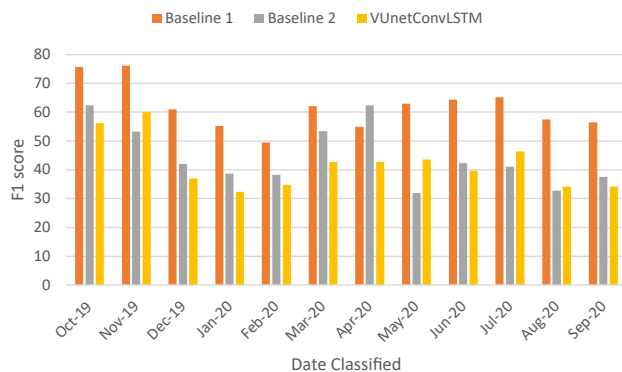

Figure 9. Per-date average F1 score for the proposed approaches.

Table 4 presents the result in terms of F1 score and OA averaged over the entire temporal sequence. Notice that metrics are consistent with results reported in Figure 8 and Figure 9, where the highest values of F1 and OA were obtained by Baseline 1. Baseline 2 resulted in a decrease of 17% and 14% in terms of F1 score and OA, respectively. Interestingly, the proposed VUnetConvLSTM only presented a slight reduction in both metrics regarding Baseline 2, with a drop of 2.7% and 0.5% for F1 score and OA, respectively. The aforementioned table also presents training times. Both Baseline 1 and Baseline 2 presented the highest times because they required to train a separate network for each date in the sequence. The proposed VUnetConvLSTM reduced training times by approximately 74.3% compared to the baseline approaches.

Method	F1	OA	Training [h]
Baseline 1	61.7	77.3	14.6
Baseline 2	44.7	63.3	14.6
VUnetConvLSTM	42	62.8	3.8

Table 4. Performance metrics averaged over the temporal sequence, and training time.

6. CONCLUSIONS

This work evaluated the capabilities of a convolutional recurrent crop mapping architecture adapted to tropical regions, to generalize to future agricultural years in dates unseen during training. Results suggest that merely training a network in a specific agricultural year might not be enough for the architecture to correctly generalize to future unseen dates. The Baseline 2 and the proposed VUnetConvLSTM model, which were trained in LEM 17/18 and tested two years ahead in time using the LEM 19/20 dataset, presented a significant performance drop compared to Baseline 1, which was trained and tested on the same temporal sequence. This indicated a domain gap between the datasets, which was not addressed in this work.

Furthermore, this work compared the generalization capabilities of a network trained to produce an outcome on a specific date (Baseline 2) and the proposed VUnetConvLSTM, trained to produce outcomes for all the months of the agricultural year in an end-to-end fashion. Results showed that the proposed VUnetConvLSTM presented only slight performance decreases compared to Baseline 1, which might be a small cost in exchange for the operational advantages brought by the end-to-end approach.

Future works will focus on evaluating domain adaptation techniques such as Colormapgan (Tasar et al., 2020) to address the domain gap between the assessed datasets. Data fusion between optical and SAR images will also be considered for improving the classification metrics. Likewise, the inclusion of other deep learning networks such as Unet3D and Transformers will be evaluated.

REFERENCES

Ajadi, O. A., Barr, J., Liang, S.-Z., Ferreira, R., Kumpatla, S. P., Patel, R., Swatantran, A., 2021. Large-scale crop type and crop area mapping across Brazil using synthetic aperture radar and optical imagery. *International Journal of Applied Earth Observation and Geoinformation*, 97, 102294.

Audebert, N., Boulch, A., Randrianarivo, H., Le Saux, B., Ferecatu, M., Lefèvre, S., Marlet, R., 2017. Deep learning for urban remote sensing. *2017 Joint Urban Remote Sensing Event (JURSE)*, IEEE, 1–4.

Castro, J. D. B., Feitosa, R. Q., La Rosa, L. C., Diaz, P. M. A., Sanches, I. D. A., 2017. A comparative analysis of deep learning techniques for sub-tropical crop types recognition from multitemporal optical/sar image sequences. *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, 382–389.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, 9(8), 1735–1780.

La Rosa, L. E. C., Happ, P. N., Feitosa, R. Q., 2018. Dense fully convolutional networks for crop recognition from multi-temporal sar image sequences. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 7460–7463.

Martinez, J. A. C., Happ, P. N., Feitosa, R. Q., 2019. A comparison of fully convolutional and recurrent networks for multi-temporal crop recognition using sar images. *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*.

Martinez, J. A. C., La Rosa, L. E. C., Feitosa, R. Q., Sanches, I. D., Happ, P. N., 2021. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171, 188–201.

Oldoni, L. V., Sanches, I. D., Picoli, M. C. A., Covre, R. M., Fronza, J. G., 2020. LEM+ dataset: For agricultural remote sensing applications. *Data in Brief*, 33, 106553.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Rußwurm, M., Körner, M., 2018. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4), 129.

Sanches, I., Feitosa, R., Achanccaray, P., Montibeller, B., Luiz, A., Soares, M., Prudente, V., Vieira, D., Maurano, L., 2018a. Lem Benchmark Database for Tropical Agricultural Remote Sensing Application. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 621, 387–392.

Sanches, I., Feitosa, R., Achanccaray, P., Soares, M., Luiz, A., Schultz, B., Maurano, L., 2018b. Campo Verde Database: Seeking to Improve Agricultural Remote Sensing of Tropical Areas. *IEEE Geoscience and Remote Sensing Letters*, 15(3), 369–373.

Santos, L. A., Ferreira, K., Picoli, M., Camara, G., Zurita-Milla, R., Augustijn, E.-W., 2021. Identifying Spatiotemporal Patterns in Land Use and Cover Samples from Satellite Image Time Series. *Remote Sensing*, 13(5), 974.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214*.

Tardy, B., Inglada, J., Michel, J., 2017. Fusion approaches for land cover map production using high resolution image time series without reference data of the corresponding period. *Remote Sensing*, 9(11), 1151.

Tasar, O., Happy, S., Tarabalka, Y., Alliez, P., 2020. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10), 7178–7193.

Teimouri, N., Dyrmann, M., Jørgensen, R. N., 2019. A Novel Spatio-Temporal FCN-LSTM Network for Recognizing Various Crop Types Using Multi-Temporal Radar Images. *Remote Sensing*, 11(8), 990.