Spatio-Temporal Deep Learning Approach to Map Deforestation in Amazon Rainforest

Raian V. Maretto, Leila M. G. Fonseca, Nathan Jacobs, Thales S. Körting, Hugo N. Bendini, and Leandro L. Parente

***Abstract*— We address the task of mapping deforested areas in the Brazilian Amazon. Accurate maps are an important tool for informing effective deforestation containment policies. The main existing approaches to this task are largely manual, requiring significant effort by trained experts. To reduce this effort, we propose a fully automatic approach based on spatio-temporal deep convolutional neural networks. We introduce several domain-specific components, including approaches for: image preprocessing; handling image noise, such as clouds and shadow; and constructing the training data set. We show that our preprocessing protocol reduces the impact of noise in the training data set. Furthermore, we propose two spatio-temporal variations of the U-Net architecture, which make it possible to incorporate both spatial and temporal contexts. Using a large, real-world data set, we show that our method outperforms a traditional U-Net architecture, thus achieving approximately 95% accuracy.**

***Index Terms*— Convolutional neural networks (CNNs), deep learning (DL), deforestation, spatio-temporal analysis, U-Net.**

I. Introduction

DESPITE the significant reduction in the deforestation rates in the Brazilian Amazon in the early 2000s, mainly due to the Brazilian policies and enforcement actions, thousands of square kilometers of forest are still being cleared every year. Producing accurate deforestation maps is critical for informing and enabling public policies aimed at combating deforestation. Since 1988, the PRODES program, developed by INPE,[1] has been estimating deforestation rates on an annual basis. Since 2000, digital maps have been produced, thus resulting in the most consistent and dense temporal series of maps of anthropic disturbance in primary forests in the Brazilian Amazon [1]. Together with the Near Real-time Deforestation Detection System (DETER), PRODES played an important role in the reduction of deforestation rates in the early 2000s [2]. PRODES and DETER are considered the main references on large-scale accurate mapping of deforestation in tropical forests [3]. Data generated by PRODES and DETER are used by the Brazilian government to support environmental surveillance actions, environmental protection, and public policies in the Brazilian Amazon. However, both systems still rely on remote sensing experts to perform the visual analysis of the satellite imagery [1]. This makes the task of producing deforestation maps being expensive, time-consuming, and strongly dependent on the expertise of the analysts. Many initiatives have been made to automate this process, such as the Global Land Analysis an Discovery (GLAD), developed by Global Forest Watch [4] and the Deforestation Alert System (SAD), developed by Imazon. However, none of them achieved the classification accuracy greater than 90% similar to PRODES and DETER. Therefore, there is still necessity to develop an automated method of deforestation detection in the Brazilian Amazon that can be operational, processing a large amount of data in an efficient way, and also having high classification accuracy.

Automating Land Use and Land Cover (LULC) mapping and change detection are difficult tasks. As a change detection problem, the deforestation mapping involves some challenges. The main challenges include: the presence of clouds and cloud shadows; integrating imagery from different sensors; spectral difference due to phenological changes; and various other imaging artifacts. Recently, deep learning (DL) methods have shown promise for LULC mapping and change detection tasks, with high accuracies, robustness to various sources of noise, and the ability to scale to large-scale mapping [5], [6].

This letter investigates the effectiveness of DL techniques for mapping deforestation in the Brazilian Amazon. We propose a fully automatic approach, using PRODES maps as ground truth, to train three variations of the U-Net [7] on Landsat 8 Operational Land Imager (OLI) images. Furthermore, we propose two spatio-temporal variations of the U-Net. We found that including temporal context is important for reducing false positives, as well as increasing the focus on changes. The approach was tested for a region comprising nearly 111 000 km^2 in southeastern Pará state, a well-known agricultural expansion frontier. The resultant deforestation maps, depicted in Fig. 1, achieved an accuracy of approximately 95% on a held-out testing set.

II. Deep Learning-Based Land Use/Cover Mapping

The feature representation learned by deep neural networks (DNNd), especially the convolutional neural networks (CNNs) and the end-to-end fully convolutional networks (FCNs) [8], have shown to be greatly effective in scene classification and semantic segmentation tasks. Through a recurrent attention structure, the ARCNet was able to focus selectively on key regions, thus demonstrating the importance of high-level features and achieving over 99% accuracy on target

Manuscript received November 5, 2019; revised February 3, 2020; accepted February 13, 2020. Date of publication April 28, 2020; date of current version April 22, 2021. This work was developed as part of the project "Development of systems to prevent forest fires and monitor vegetation cover in the Brazilian Cerrado", with financial support of the Forest Investment Program (World Bank Project #P143185) and in part by the National Science Foundation under Grant IIS-1553116. *(Corresponding author: Raian V. Maretto.)*

Raian V. Maretto, Leila M. G. Fonseca, Thales S. Körting, and Hugo N. Bendini are with the National Institute for Space Research (INPE), São José dos Campos 12227-010, Brazil (e-mail: raian.maretto@inpe.br).

Nathan Jacobs is with the Department of Computer Science, University of Kentucky, Lexington, KY 40506 USA.

Leandro L. Parente is with the LAPIG, Department of Computer Science, Federal University of Goiás, Goiânia 74690-900, Brazil.

Color versions of one or more of the figures in this letter are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.2986407

[1]National Institute for Space Research, Brazil.

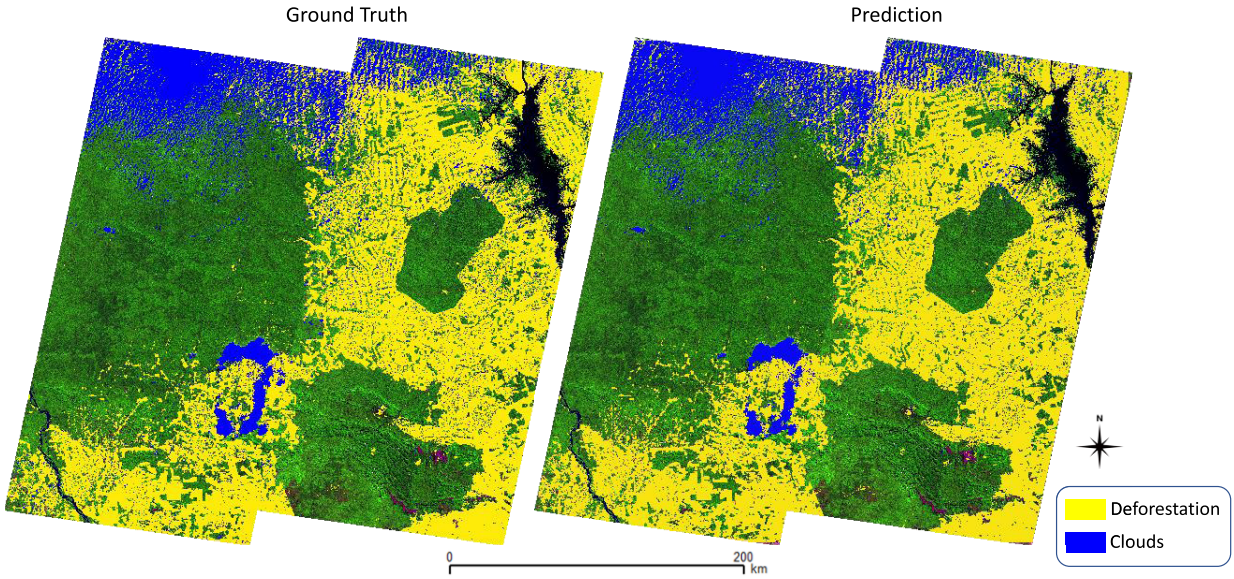

Fig. 1. Overview results for the Late Fusion U-Net, which achieved the best performance among the tested models. The *deforestation* and *cloud* classes are here overlaid to the images with a color composition on Landsat-OLI bands R(6)G(5)B(4).

detection tasks [9]. Wang *et al.* [10] developed a weekly supervised adversarial approach able to learn domain-invariant features, thus improving the semantic segmentation accuracy with synthetically produced training data. Several works have demonstrated the effectiveness of DNN for LULC mapping and LULC change detection [6]. Syrris *et al.* [11] evaluated different variations of four DNN models to map eight different land cover classes from the Infrastructure for Spatial Information in Europe (INSPIRE) TOP10NL data set over Sentinel-2 images, thus reaching an overall accuracy of approximately 87% compared to 81% obtained by the Random Forest.

To map four different classes in urban spaces on the high-resolution airborne images from the ISPRS data set Vaihingen, Häufel *et al.* [12] evaluated a traditional CNN on a superpixel segmentation approach against the pixel-wise DeepLabV3+, proposed in [13], achieving promising results, with 82% and 88% overall accuracy, respectively. As originally proposed in [7] to perform semantic segmentation on medical images, the U-Net and its variations are the most successful DNN architectures for remote sensing applications. Zhang *et al.* [14] proposed a Residual U-Net to perform road extraction on high-resolution aerial images from the Massachusetts roads data set.

However, despite the successful results, most applications work on ready-made training data sets with preprocessed data. Our approach encompasses not only the classification task but also the data preprocessing and data set generation, thus providing a fully automatic approach from preprocessing to classification.

III. Methodology

The main purpose of this letter is to apply DL-based semantic segmentation techniques to automate the deforestation detection in the Brazilian Amazon. We propose a fully automatic approach to preprocess input data, deal with ground-truth labels under clouds, train the proposed DNN classifiers, and perform the prediction, using PRODES maps as ground truth. Fig. 2 represents a simplified flowchart

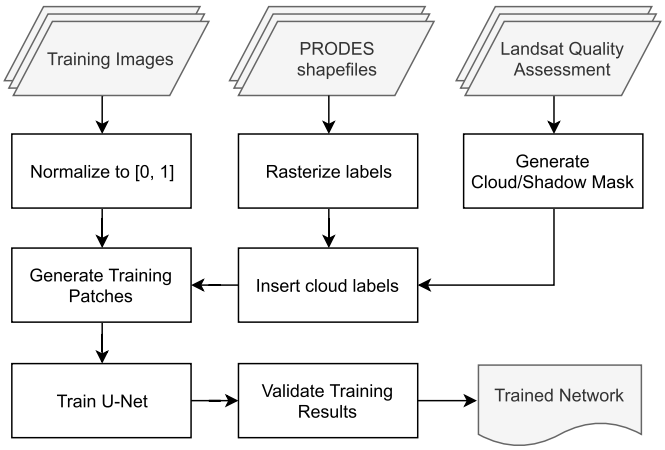

Fig. 2. Overview of the DNN training methodology.

of data preprocessing and training processes, as described in Section III-A.

Since deforestation is a land cover change phenomenon, it is necessary to take into account not only the spatial context but also temporal dynamics. To accomplish that, we propose two variations of U-Net that take into account short-term temporal dynamics. These two variations were tested with two different loss functions, the average soft dice (SD) (ASD) score, and the weighted cross-entropy, which are described in Section III-C. These spatio-temporal variations were then compared with our implementation of the traditional U-Net.

A. Preprocessing and Training

To take better advantage of the neuron's activation, which is done through the rectified linear unit (ReLU) function, the input images were normalized to the interval [0, 1]. PRODES maps are produced from Landsat-8 OLI images and distributed as shape files through the TerraBrasilis Platform.[2]

[2]www.terrabrasilis.dpi.inpe.br/downloads

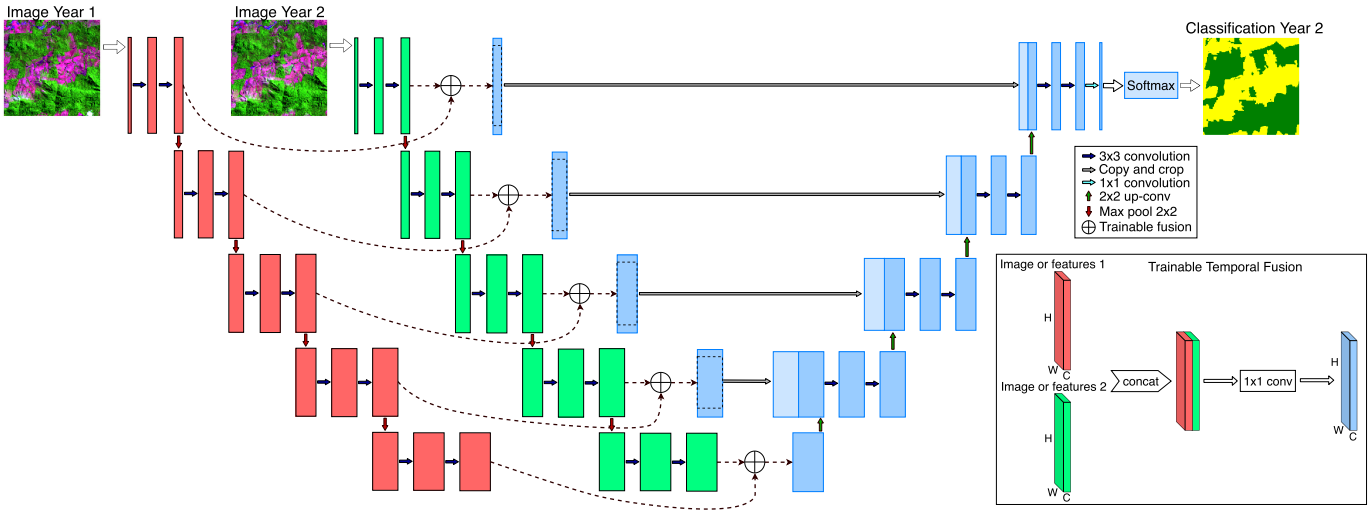

Fig. 3. U-Net with late spatio-temporal fusion. (Bottom-right frame) Trainable temporal fusion.

To keep the consistency with PRODES methodology, our method was developed over images from the same sensor. To develop our method, first, the shape files were rasterized with the same 30 m of spatial resolution of the input Landsat-OLI Images to keep the correspondence between the image pixel and ground-truth labels. PRODES maps the increment in the deforestation on primary forests, using a mask to ensure that older deforested areas will not be mapped again, thus keeping the consistency of the temporal series. PRODES uses images from several dates to produce the maps for each year, due to the high occurrence of clouds in the Amazon region. This brings a problem to train the DNN because merging all images in a cloudless image could produce undesired artifacts, potentially confusing the classifier. To overcome this problem, we used the Landsat quality assessment channel to create a new class related to the clouds and cloud shadow in the ground-truth data, thus generating a mask and then replacing the labels for the corresponding pixels. With this approach, we reduced the cloudy noise on the ground-truth data.

After that, the images and labels were sequentially sliced into 286 × 286 patches to reduce the computational cost. In addition, all patches containing pixels with no data values were removed from the training data set. The generated patches were divided into three data sets: 60% for training, 20% for evaluation, and 20% for validation.

B. Spatio-Temporal U-Net With Trainable Temporal Fusion

As aforementioned, three U-Net variations were compared in this letter, a baseline model that follows the same architecture as described in [7] and two U-Net extensions with a trainable temporal fusion approach, to consider not only the spatial context but also the temporal dynamics between N timestamps, with N being considered the temporal depth. The baseline method consists of the same structure of the original U-Net, changing the input size to 286 × 286 pixels. As in the original U-Net, the output image is smaller than the input, due to the unpadded convolutions, with a size of 100 × 100 pixels.

1) Trainable Temporal Fusion Component: Used in both U-Net variations proposed in this letter, the temporal fusion is depicted in the bottom-right frame of Fig. 3. It consists of concatenating the N time-stamps and then performing a 1 × 1 convolution with the same number of filters as the number of channels of each individual input image. This process aims to reduce the amount of data being processed by the network, as well as work as a trainable change detector. The main difference between the two spatio-temporal approaches is the way this temporal fusion is applied.

2) Early Fusion (EF) Spatio-Temporal U-Net: In the EF version, the temporal fusion of the N timestamps is performed as an extra layer before starting the network encoder, with the resulting feature maps following the traditional U-Net flow.

3) Late Fusion (LF) Spatio-Temporal U-Net: In the LF version, depicted in Fig. 3, the U-Net encoder is duplicated and each image is processed by its corresponding encoder. After each convolutional block, the temporal fusion is applied, fusing the feature maps generated by the N encoders, and then, the fused feature maps are cropped and copied to be concatenated on the corresponding block on the decoder.

C. Loss Functions

The input data are unbalanced, in terms of the number of pixels, between the classes of interest. For that reason, we tested two different losses that have been successfully applied in the literature for unbalanced data, the *weighted cross-entropy* and the *ASD*.

Weighted Cross-Entropy (WCE): The cross-entropy evaluates the class prediction for each pixel vector individually, asserting equal importance for every pixel in the learning process. This may bring problems if the classes have an unbalanced distribution across the image. To overcome that, Ronneberger *et al.* [7] and Long *et al.* [8] successfully applied different weighting strategies to the cross-entropy. The first loss function evaluated was the WCE, described as follows:

$$\mathrm{WCE} = -\sum_{c=1}^{C} w_c \sum_{i=1}^{N} y_{\mathrm{true}} \log(y_{\mathrm{pred}}). \tag{1}$$

where the weight w_c of each class, described in the following equation, is defined as the mean across the proportions of all classes (μ_P) divided by the proportion p_c of that class. y_{true} represents the ground truth and y_{pred} is the prediction

generated by the network:

$$w_c = \frac{\mu_P}{p_c}. \tag{2}$$

1) Average Soft Dice: Based on the Dice Score (DS), it is also commonly used as the loss function for semantic segmentation tasks. As originally developed for binary data, the DS is essentially a measure of overlap between two sample sets, which ranges in $[0, 1]$. The adaptation of the DS to be used as a loss function is called *SD* and is described in the following equation. The numerator represents the measure of the common activations between the predicted map and the ground truth, while the denominator represents the measure of the number of activations in each one. This has an effect of normalizing the loss according to the size of the target mask, making it less affected by imbalanced data.

$$SD_c = 1 - \frac{2\sum_{i=1}^{N} y_{\text{true}} y_{\text{pred}}}{\sum_{i=1}^{N} y_{\text{true}}^2 + \sum_{i=1}^{N} y_{\text{pred}}^2 + \epsilon}. \tag{3}$$

The SD score is computed for each class separately and then averaged across all classes. An essential difference between the WCE and the ASD is that the first is computed over the network logits, while the second is computed over the predicted probabilities.

D. Implementation and Optimization Details

DL methods are highly prone to overfitting, thus bringing the need for strategies to avoid it. We used three strategies for that: *batch normalization*, *L2 regularization*, and *data augmentation*. The batch normalization was applied after each convolution operation, before the ReLU activation. The L2 regularization was applied to all convolution layers, with a factor of 5×10^{-4}. The data augmentation is used to artificially increase the number of training samples. To accomplish that, four operations were applied to each patch on the training data set, randomly chosen between three rotations (90°, 180°, and 270°) and three flips (left–right, up–down, and transpose). Taking advantage of the parallelism of the TensorFlow input data pipeline [15], the data augmentation operations were applied on the fly during the training process. For the EF U-Net and the LF U-Net, we used a total of 4723 patches for training, increasing to 23 615 with the data augmentation procedure. For each of the evaluation and validation, 1574 patches were used. For the baseline U-Net, once each year was processed separately instead of fusing pairs of years, we had a total of 5904 patches for training (29 520 after the data augmentation) and 1968 for each of the evaluation and validation.

The network weights were initialized through the Xavier initializer. For all experiments, it was run for 100 epochs, with batches of 80 patches. Once the networks were trained from scratch, the learning rate was adjusted through exponential decay, starting from 0.1, and decaying with a rate of 0.95. Due to the high influence of atmospheric effects on lower wavelength bands, we used the five bands corresponding to the green, red, near-infrared, and the two short-wave infrared bands. The methods were developed into the DeepGeo Toolbox [16].

IV. Study Area and Experiments

Our study area corresponds to four Landsat 8 OLI scenes, comprising an area of nearly 111 000 km^2, in southeastern Pará state, a well-known agricultural expansion frontier.

TABLE I
AVERAGE METRICS ACROSS ALL CLASSES

Model	Loss Function	F1-score	Avg. accuracy	AuC
Baseline U-Net	WCE	0.9448	0.9447	0.9883
EF U-Net		0.9427	0.9425	0.9873
LF U-Net		**0.9471**	**0.9470**	**0.9887**
Baseline U-Net	ASD	0.9417	0.9417	0.9857
EF U-Net		0.9388	0.9388	0.9858
LF U-Net		0.9460	0.9461	0.9876

The training data set is composed of a five-year temporal series, from 2013 to 2017, and the spatio-temporal networks were configured with a temporal depth of two years. Thus, to achieve the data samples for each year, the images from that year and the previous were used. For example, for the year of 2017, images from 2016 and 2017 were used. PRODES classify four main classes, *non forest*, *forest*, *hidrography*, and *deforestation*. For simplification purposes, once *hidrography* and *non forest* are invariant in time, we grouped them with the class *forest* in one single class, named *not deforestation*.

Due to the availability of PRODES temporal series for the entire Amazon, it is more important to ensure the network generalization in time than in space. To produce the results in this letter, we used the images from 2018 as a test data set. As shown in Table I, the LF U-Net presented better scores for all metrics with both losses, when compared with the baseline and the EF U-Net, with the baseline network performing slightly better than the EF U-Net. Furthermore, we can also observe that all network configurations presented better scores with the WCE than with the ASD, thus demonstrating a better ability to deal with the data imbalance. Although the LF U-Net worked well in improving the performance of the network, the EF U-Net did not have the same success. This may be explained by the fact that applying the temporal fusion only in the beginning, the reduced amount of filters was not enough to properly deal with the temporal dynamics.

When performing a classwise comparison, shown in Table II, we observe that although the WCE loss function performed better for most classes and metrics, the ASD presented a better precision for the deforestation and cloud classes. This might be explained by the fact that the ASD loss function is derived from an overlap measure reducing the number of false positive detections, once the precision measures the purity of the positive detections. However, the higher recall for the WCE shows that the number of true positives was also reduced when using the ASD, which explains the better overall accuracy for the versions using the WCE.

The main goal of PRODES is to map anthropic disturbance on primary forests, thus creating a consistent temporal series of the deforested areas. For this reason, the deforested areas that are abandoned and regenerated, becoming secondary forest, still being mapped as deforestation. Nevertheless, the areas abandoned for longer time present a similar spectral behavior as the primary forest, acting then as a kind of noise in the ground truth for the training process. Taking a closer look at the classification produced by our model, we observed that several false negatives happened due to these regenerated areas. This behavior can be observed in the red circles in the left column on Fig. 4. Therefore, we believe that the real error, especially in more recently deforested areas, was overrated. In the right column on Fig. 4, it is possible to observe that the ground truth is speckled with some nonexistent clouds, which

TABLE II
CLASSWISE METRICS

		Baseline U-Net	Early Fusion U-Net	Late Fusion U-Net	Baseline U-Net	Early Fusion U-Net	Late Fusion U-Net
Metric	Class	Weighted cross-entropy			Soft Dice Score		
F1-score	Not Deforestation	0.9536	0.9520	**0.9558**	0.9522	0.9507	0.9557
	Deforestation	0.9241	0.9215	**0.9274**	0.9189	0.9182	0.9248
	Clouds	**0.9662**	0.9617	0.9650	0.9599	0.9395	0.9636
Precision	Not Deforestation	0.9636	0.9630	**0.9657**	0.9484	0.9449	0.9531
	Deforestation	0.9101	0.9074	0.9147	0.9204	0.9223	**0.9286**
	Clouds	0.9593	0.9498	0.9537	**0.9795**	0.9617	0.9660
Recall	Not Deforestation	0.9438	0.9413	0.9461	0.9561	0.9564	**0.9584**
	Deforestation	0.9386	0.9362	**0.9405**	0.9173	0.9142	0.9209
	Clouds	0.9732	0.9739	**0.9766**	0.9411	0.9183	0.9612
AuC	Not Deforestation	0.9881	0.9868	**0.9891**	0.9843	0.9847	0.9870
	Deforestation	0.9871	0.9855	**0.9879**	0.9829	0.9838	0.9859
	Clouds	**0.9994**	0.9989	0.9990	0.9981	0.9971	0.9886

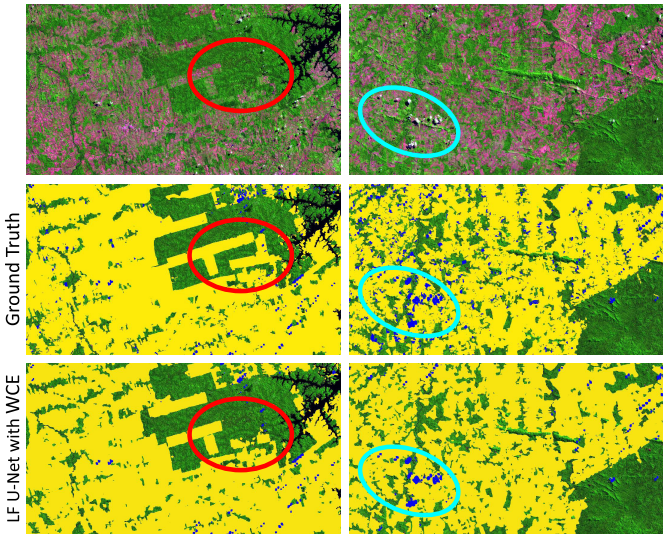

Fig. 4. Highlights on special cases on regenerated areas (red) and clouds (cyan), showing robustness to noises on training data.

also acts as noise in the training data. In this case, our network was also able to classify only the existent clouds, circled in cyan, also indicating an overrated error. This highlights the robustness of our model to noise in the ground truth.

V. CONCLUSION

We developed a fully automatic approach to mapping deforested areas in the Brazilian Amazon using Landsat 8 OLI imagery. Our approach uses modern learning-based classification techniques tailored to the particulars of the task and the available data sets. Through extensive evaluation, we demonstrated that our approach successfully generalizes from year-to-year, thus achieving an overall accuracy of approximately 95%. We also demonstrated that our approach is somewhat robust to noise in the ground truth, as depicted in Fig. 4.

In future work, we aim to expand our study area to the entire Brazilian Amazon in an operational application. Furthermore, we aim to evaluate its generalization in space, by testing the methodology for other biomes and regions. We also believe that by extending the methodology to include images from additional sensors, we may be able to produce highly accurate deforestation maps every few days. This would likely be an important tool for combating illegal deforestation.

REFERENCES

[1] INPE, *INPE, Amazon Program—Monitoring the Brazilian Amazon by satellite: The PRODES, DETER, DEGRAD and TerraClass Systems*, National Institute for Space Research, São José dos Campos, Brazil, 2019. [Online]. Available: http://www.obt.inpe.br/prodes

[2] D. Boucher, S. Roquemore, and E. Fitzhugh, "Brazil's success in reducing deforestation," *Tropical Conservation Sci.*, vol. 6, no. 3, pp. 426–445, Aug. 2013.

[3] D. Nepstad *et al.*, "Slowing amazon deforestation through public policy and interventions in beef and soy supply chains," *Science*, vol. 344, no. 6188, pp. 1118–1123, Jun. 2014.

[4] M. C. Hansen *et al.*, "High-resolution global maps of 21st-century forest cover change," *Science*, vol. 342, no. 6160, pp. 850–853, Nov. 2013.

[5] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[6] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE CVPR*, Jul. 2015, pp. 3431–3440.

[9] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[10] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.

[11] V. Syrris, P. Hasenohr, B. Delipetrev, A. Kotsev, P. Kempeneers, and P. Soille, "Evaluation of the potential of convolutional neural networks and random forests for multi-class segmentation of Sentinel-2 imagery," *Remote Sens.*, vol. 11, no. 8, p. 907, 2019.

[12] G. Häufel, L. Lucks, M. Pohl, D. Bulatov, and H. Schilling, "Evaluation of CNNs for land cover classification in high-resolution airborne images," *Proc. SPIE Remote Sensing*, vol. 10790, Oct. 2018, Art. no. 1079003.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.

[14] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[15] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[16] R. V. Maretto, T. S. Korting, and L. M. G. Fonseca, "An extensible and easy-to-use toolbox for deep learning based analysis of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 9815–9818.