Farm Parcel Delineation Using Spatio-temporal Convolutional Networks

Han Lin Aung
Department of Computer Science
Stanford University
hanlaung@cs.stanford.edu

Burak Uzkent
Department of Computer Science
Stanford University
buzkent@cs.stanford.edu

Marshall Burke
Department of Earth Systems Science
Stanford University
mburke@stanford.edu

David Lobell
Department of Earth Systems Science
Stanford University
dlobell@stanford.edu

Stefano Ermon
Department of Computer Science
Stanford University
ermon@cs.stanford.edu

Abstract

Farm parcel delineation (delineation of boundaries of farmland parcels/segmentation of farmland areas) provides cadastral data that is important in developing and managing climate change policies. Specifically, farm parcel delineation informs applications in downstream governmental policies of land allocation, irrigation, fertilization, greenhouse gases (GHG's), etc. This data can also be useful for the agricultural insurance sector for assessing compensations following damages associated with extreme weather events - a growing trend related to climate change [4]. Using satellite imaging can be a scalable and cost-effective manner to perform the task of farm parcel delineation to collect this valuable data. In this paper, we break down this task using satellite imaging into two approaches: 1) Segmentation of parcel boundaries, and 2) Segmentation of parcel areas. We implemented variations of U-Nets, one of which takes into account temporal information, which achieved the best results on our dataset on farm parcels in France in 2017.

1. Introduction

Farm parcel delineation has been a highly manual task before the use of computer vision and machine learning, incurring high costs and time for those who are labeling the data through 'theodolites, total stations, and GPS' [37]. Beyond costs, the cadastral information retrieved from de-lineation of farm parcels is particularly important in forming climate change policies for mitigation and adaptation. More specifically, such information is important in developing and managing incentive plans on environmental and climate change, allocation of water and irrigation, and policies around agricultural insurance for catastrophe. Overall, according to World Bank, land policies, specifically for farm parcels, are considered to be highly intertwined with climate change [4]. Hence, we believe that automating farm parcel delineation in a scalable manner can cut down costs and time by replacing the otherwise manual process of collecting parcel-level information. Doing so can help inform relevant policy makers and stakeholders on real-time cadastral data very rapidly.

On the technical front, deep convolutional neural networks (CNNs) have been successfully applied to different computer vision tasks including image recognition [18, 13, 31], object detection [36, 19, 25, 24], and object tracking [17, 2, 33, 34]. Following this trend, we investigate its application in delineating farm parcel boundaries/areas using satellite imagery. Combination of CNNs with satellite images is a scalable method that automates the otherwise manual process [28, 35]. We further take into account temporal data to improve results over the vanilla approach using a single image for our farm parcel delineation task.

2. Related Work

In recent years, deep learning has become very popular in computer vision tasks due to its incredible success.
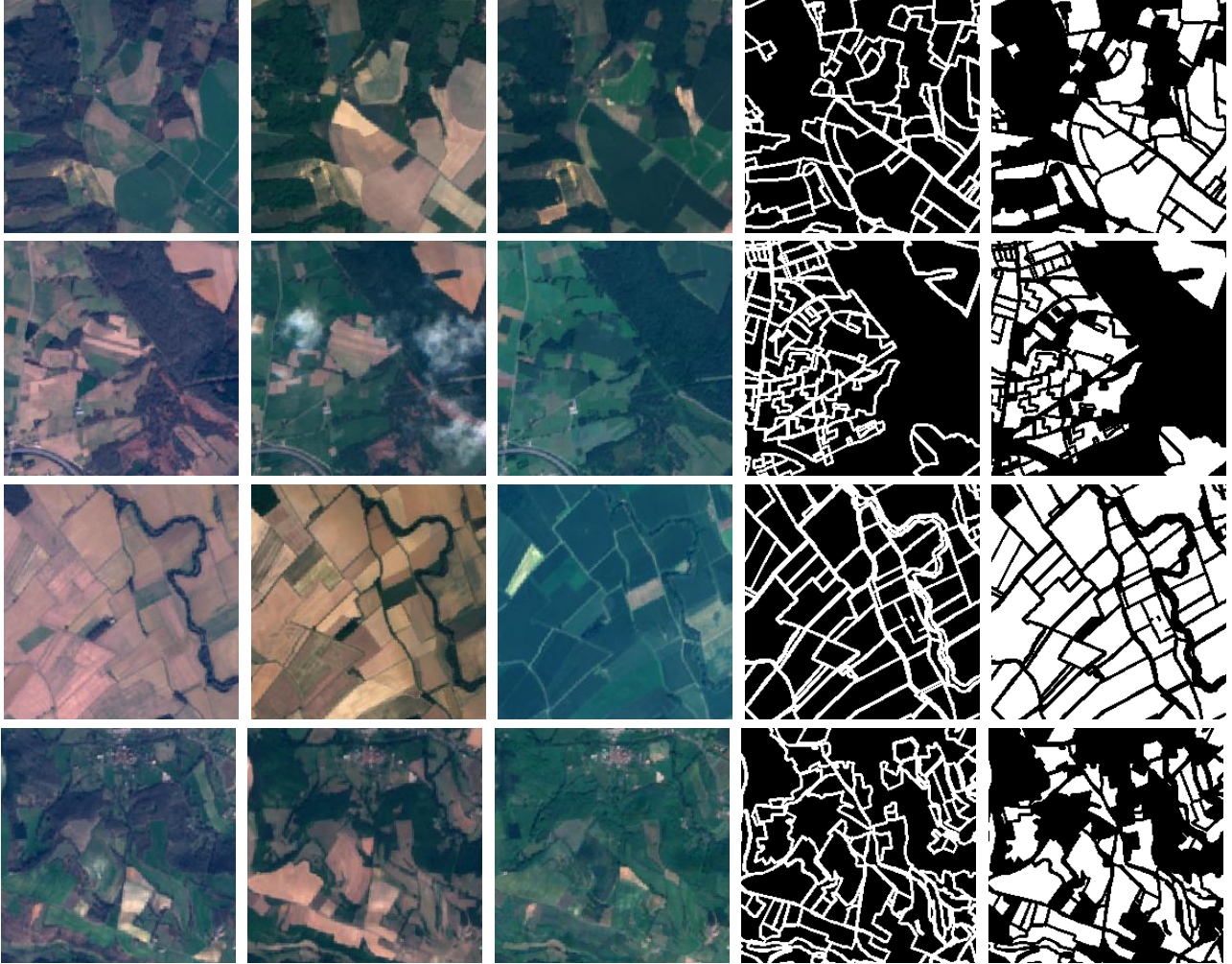
Figure 1: Each row is an instance from the dataset. From left to right, the first 3 images are satellite images from January-March, April-June, July-September respectively. The final 2 images are the binary image masks of boundaries and areas respectively.

Known as a universal learning approach, the same deep neural network architecture can be applied with great success over multiple domains [1]. One particular domain in computer vision that deep learning has shown state-of-the-art results is on image segmentation tasks.

There are two general Convolutional Neural Network (CNN) architectures often used in image segmentation: patch-based networks and Fully Convolutional Networks (FCNs). The patch-based models [32] receive fixed-size patches based on a central pixel as inputs to make a prediction. Hence, to make predictions on a per-pixel level, these models receive patches for each of the pixels of each input image. However, using patch-based models to provide dense predictions on a per-pixel level is highly computationally expensive [23]. On the other hand, FCNs are built from locally connected layers and do not contain fully

connected layers. Hence, FCNs can perform training and inference on a per-pixel level efficiently [20]. A particular FCN known as U-Net shows competitive performance in image segmentation, especially in medical imaging but also in other domains such as robotics through the use of video data [30]. U-Net consists of a series of convolutional layers that downsamples the image (encoder phase), 2 convolutional layers (bottleneck), and a series of convolutional layers that upsamples the image (decoder phase) [26]. The downsampling path captures context while the symmetric upsampling path enables precise localization [26].

There are generally two approaches in training a CNN, from scratch and transfer learning [9]. Using the transfer learning approach, pretrained FCNs on ImageNet [6] generally provide improved performance over training models from scratch. This approach is especially useful when

limited data or computational resources is present [7]. In particular, TernausNet, a U-Net with a VG11 encoder pre-trained on ImageNet speeds up convergence and improves segmentation accuracy over the trained from scratch counterpart on boundary segmentation of cars [14].

Apart from the traditional method of using a single image as input and the ground truth segmented mask, approaches to incorporate additional relevant data sources show improved performance on image segmentation tasks. For instance, a recent approach using U-Nets to incorporate input images from multiple time frames as a means to incorporate temporal information provide more promising results over simply using a single image in the task of urban land use classification [21].

For our particular task, supervised machine learning and deep learning approaches using FCNs to delineate farm parcels perform better than classical computer vision methods and have produced state-of-the-art results [10, 11, 22]. The traditional methods such as edge-based detectors perform reasonably well on delineating boundaries of regularly-shaped farmland areas but fail in cases with more complex farmland shapes. Furthermore, many of these traditional methods do not delineate boundaries contextually, a relationship that can be utilized by deep learning methods [22].

3. Problem Statement

Given an input of satellite images represented with $x_i^t \in \mathcal{X}^t$ from an area with geo-coordinates $c_i = (lat_i, lon_i)$, we output a binary mask represented with $\hat{y}_i \in \hat{\mathcal{Y}}$. Here, the index t represents the images over time from the area centered at c_i. $t = 0, t = 1$, and $t = 2$ represent the time ranges Jan-March, April-June, and July-Sept respectively. The image captured in April-June is represented with $x_i^{t=1}$ and it aligns with ground truth mask y_i temporally. In our study, we model the output mask to contain pixel wise binary labels. Specifically, our models output two different masks: (1) delineated boundaries between farm parcels, (2) segmentation masks of farm parcels. In the first case, the goal is to assign the pixel with image coordinates, m and n, the label 1 if it lies on the boundary of a farm parcel as $f_b : x_i^t(m, n) \mapsto \hat{y}_i(m, n) = 1$ where f_b represents the boundary segmentation network. In the second task, our goal is to map a pixel in the image space with coordinates m and n to label 1 if it lies inside the farm parcel as $f_a : x_i^t(m, n) \mapsto \hat{y}(m, n) = 1$ where f_a represents the area segmentation network.

3.1. Dataset

3.1.1 Data Description

The dataset represented with $(\mathcal{X}^t, \mathcal{Y})$ consists of Sentinel-2 satellite imagery ($224px \times 224px$ RGB image correspond-

Figure 2: Spatial coverage of the shapefile used for the dataset to sample polygons.

ing to $2.24km \times 2.24km$ of land area) along with corresponding binary masks of boundaries and areas of farm parcels [8]. Sample instances from the dataset are shown in Fig. 1. Sentinel-2 is used over other satellite imaging datasets such as DigitalGlobe as Sentinel-2 covers a much larger area of coverage per image, allowing more farm parcels to be delineated in a single image. Furthermore, Sentinel-2 is freely available and has a relatively short revisit time globally, making the dataset a suitable choice in collecting the most recent cadastral information through satellite imaging.
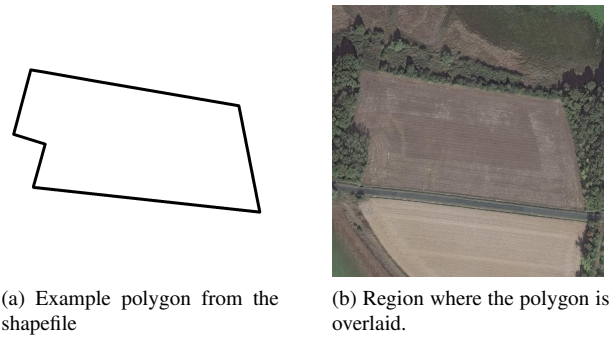
(a) Example polygon from the shapefile

(b) Region where the polygon is overlaid.

Figure 3: Sample polygon from the shapefile and the corresponding region to be overlaid. The satellite image is zoomed-in to better visualize where the polygon would be overlaid.
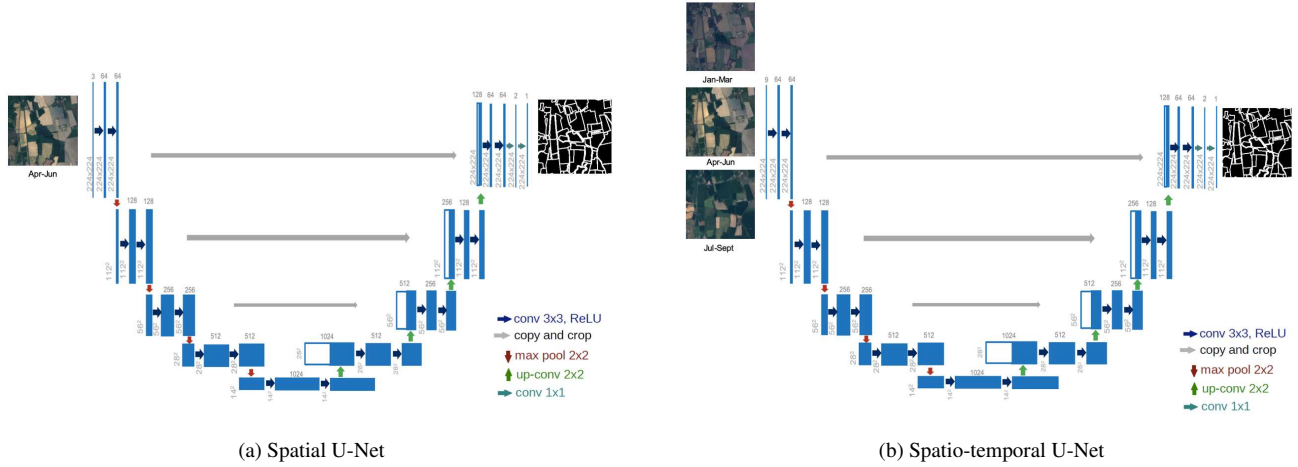
(a) Spatial U-Net

(b) Spatio-temporal U-Net

Figure 4: U-Net models (a type of Fully Convolutional Network) used for our boundary and area segmentation task. The Spatial U-Net takes in a single satellite image as input and outputs the binary mask whereas the Spatio-temporal U-Net takes into account 3 images from different timestamps to output the binary mask.

3.1.2 Data Generation

First, we started with downloading the shapefile consisting of the polygons (in Lambert-93 coordinate system) of farm parcel boundaries of France (2017) with the spatial coverage shown in Fig. 2 from a publicly available resource [1]. Fig. 3 presents a sample polygon from the shapefile. Afterwards, we projected the existing coordinate system of all the polygons to the geographic coordinate system of latitude/longitude. We sampled 2000 regions by randomly sampling 2000 $c_i(lat_i, lon_i)$ that have at least one farm parcel in its vicinity (within 2.24km radius) in France. Our image geolocation parameter $c_i(lat_i, lon_i)$ represents the center of the square region i that spans $2.24km \times 2.24km$. To ensure that there is at least one farm parcel in each region i, we sampled random polygons in the downloaded shapefile and calculated the centroid (in latitude and longitude) of each polygon until we collected 2000 centroids that do not overlap. We then selected polygons in the original shapefile that are within the bounds of these 2000 square regions to create a filtered shapefile. To generate the binary boundary mask for each square region i, we parsed the filtered shapefile to combine the set of polygons of farmland boundaries that are only relevant for that particular square region i. Afterwards, we projected the geographic coordinate system (lat_i, lon_i) onto the $224px \times 224px$ image based on linear-scaling. To draw the boundaries, we used OpenCV's polylines functionality [3]. We set the thickness of the boundary to be 2 pixels. For the corresponding binary mask of farmland areas, we used OpenCV's fillPoly functionality [3].

Second, we collected the Sentinel-2 RGB satellite imagery x_i^t ($224px \times 224px$) corresponding to each square region i over 3 time ranges $t = 0$ (January - March), $t = 1$ (April - June), and $t = 2$ (July-September) in 2017. Each x_i^t is a composite satellite image over the 3-month time range.

4. Methods

Given our dataset represented with $(\mathcal{X}^t, \mathcal{Y})$ we propose two methods to segment farmland boundaries and areas. Our first method uses only one image that is representative of time range (April-June), $x_i^{t=1}$. The second method, on the other hand, utilizes 3 images, $x_i^{t=0,1,2}$, that is representative of 3 time ranges mentioned previously.

4.1. Spatial U-Net

We decided to use U-Net as it has shown competitive performance over multiple domains in image segmentation [26]. We coin our U-Net model [26] for this task as the *Spatial U-Net* which includes 3 channels for the input. The model takes in a single RGB satellite image x_i^1 (April-June) as input and outputs a binary mask of boundaries/areas. We experiment with both a Spatial U-Net trained from scratch and also a pretrained Spatial U-Net whose encoder is initialized with weights learned on the ImageNet classification task [7].

4.2. Spatio-temporal U-Net

In addition to the Spatial U-Net, we propose a *Spatio-temporal U-Net* model that utilizes 9 channels representing 3 RGB images over time as $(t = 0, t = 1, t = 2)$ from the area i centered at c_i. In some cases we found that an image
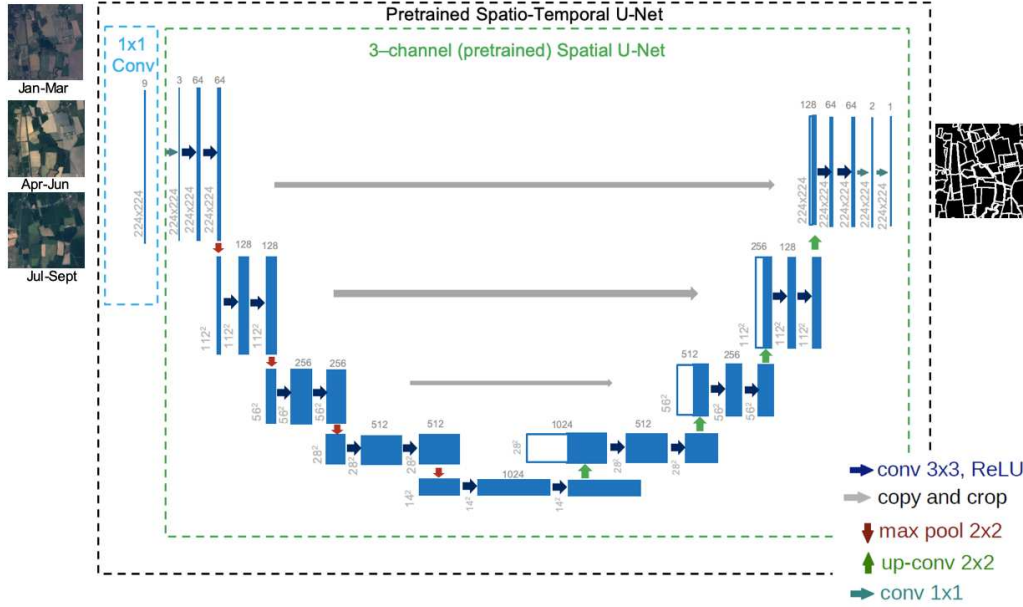
Figure 5: Pretrained Spatio-temporal U-Net. The pretrained Spatio-temporal U-Net differs from the (pretrained) Spatial U-Net (and non-pretrained Spatio-temporal U-Net) in that this model adds an additional convolution layer to map the 9 channels from 3 RGB satellite images into the 3-channel pretrained Spatial U-Net.

out of 3 images is not available. In such cases, we replace the missing image with x_i^1, which always exists.

We also experiment with a pretrained Spatio-temporal U-Net. To use the pretrained weights, the architecture of the pretrained Spatio-temporal U-Net instead includes a convolution layer that simply maps the 9 channels into the 3-channel pretrained Spatial U-Net as shown in Fig. 5.

5. Experiments

5.1. Metrics

We evaluated all the models on the Dice score and per-pixel accuracy. The Dice score is also known as the F1 score and formulated as

$$DICE = \frac{2TP}{2TP + FP + FN} \quad (1)$$

where TP = True Positive, FP = False Positive and FN = False Negative.

While accuracy is an intuitive metric, the Dice score is considered as a better metric in the cases with class imbalance between boundary/non-boundary pixels. It has been widely used in image segmentation tasks across different domains [41, 15, 29].

5.2. Implementation Details

We implemented all the U-Net models in Keras [5]. The U-Net models that are not pretrained incorporates dilated convolutions for the encoder instead of regular convolutions [39] as dilated convolutions can better aggregate the context of the whole image. We initialized the weights of the encoder of pretrained models from the ImageNet classification task with the ResNet-34 backbone by extracting the same number of layers from the ResNet-34 as the number of encoder layers in our models [38, 7]. We trained all the models using the Adam optimizer [16], a learning rate of 1e-4, batch size of 6 over 200 epochs with NVIDIA GeForce GTX 1080 Ti. Finally, we experimented with both the binary cross entropy and the Dice loss functions to train the models.

For experiments with the Spatial U-Net, each data instance is represented by (x_i^1, y_i), which corresponds to the satellite image and ground truth mask collected in April-June. As for experiments with the Spatio-temporal U-Net, each data instance is represented by $(\{x_i^0, x_i^1, x_i^2\}, y_i)$. The dataset is split using a random distribution and the split percentages for train/validation/test are 80/10/10, respectively. The implementation of the whole pipeline can be found in our github repository [2].

---

[2]Code repo: https://github.com/sustainlab-group/ParcelDelineation

	Spatial U-Net	U-Net (*Pretrained on ImageNet*)	Spatio-temporal U-Net	Spatio-temporal U-Net (*Pretrained on ImageNet*)
Dice Score - *Boundary*	0.56	0.54	0.60	**0.61**
Dice Score - *Area*	0.72	0.75	0.80	**0.81**
Accuracy - *Boundary*	0.76	0.78	0.81	**0.82**
Accuracy - *Area*	0.71	0.77	0.82	**0.83**

Table 1: Boundary and area segmentation results in terms of the Dice score and accuracy.

5.3. Results

We provide the Dice score and per-pixel accuracy for both boundary and area segmentation tasks for all the models trained using the binary cross entropy loss in Table 1. We also provide qualitative results (i.e. samples of predictions of output binary masks) for the pretrained Spatio-temporal U-Net in Fig. 6 and a qualitative comparison between the pretrained Spatio-temporal U-Net and pretrained Spatial U-Net in Fig. 7.

As shown in Table 1, the Spatial U-Net performs the worst on the area segmentation task with a Dice score of **0.72** and accuracy of **0.71**. The pretrained Spatial U-Net performs the worst on the boundary segmentation task with a Dice score of **0.54** though the pretrained Spatial U-Net does have a higher accuracy of **0.75** compared to the Spatial U-Net's accuracy of **0.72**.

The pretrained Spatio-temporal U-Net performs best for both boundary and area segmentation tasks on all metrics, achieving **0.61** Dice score and **0.82** accuracy on the boundary segmentation task and **0.81** Dice score and **0.83** accuracy on the area segmentation task.

6. Analysis

6.1. Quantitative Analysis

Area vs. Boundary Segmentation The area segmentation results seem to indicate a very powerful performance from the models. However, a large region of each satellite image is farmland area. Therefore, the quantitative border segmentation results provide a more indicative measure of the performance of each model. Hence, we put most of the emphasis on quantitative analysis on the boundary segmentation task. However, the area segmentation results are still useful for qualitative comparisons.

Pretrained vs. Trained From Scratch Spatial U-Net Despite limited amount of training data, we observe from Table 1 that even the baseline Spatial U-Net performs reasonably well with a Dice score of **0.56** and **0.72** on the boundary and area segmentation tasks respectively. However, we are surprised that the pretrained Spatial U-Net has a lower Dice score for the boundary segmentation task

(though the accuracy is higher for the pretrained Spatial U-Net). One main difference between the pretrained and trained from scratch Spatial U-Net is the use of dilated convolutions in the latter model. The dilated convolutions are found to be effective in improving the learned contextual information in satellite images [12, 40].

Spatio-temporal vs. Spatial U-Net When we compare the Spatio-temporal U-Net with the Spatial U-Net, we observe an increase in performance on the boundary segmentation task with the Spatio-temporal U-Net. This improvement indicates the utility of temporal information. The Spatio-temporal U-Net has an increase of 7% in Dice score and an increase of 11% in accuracy over the Spatial U-Net. Across the pretrained counterparts, the pretrained Spatio-temporal U-Net also has an increase of 13% in Dice score and an increase of 5% in accuracy over the pretrained Spatial U-Net.

On the other hand, pretraining seems to only slightly boost the performance of the model though it does greatly increase the convergence rate during training. The effects of pretraining can not be completely isolated, however, as the architecture of the pretrained and the non-pretrained counterparts slightly differ for the Spatio-temporal U-Net models.

6.2. Qualitative Analysis

6.2.1 Qualitative performance of the Spatio-temporal U-Net

Fig. 6 presents samples of predictions using our pretrained Spatio-temporal model. We observe that the pretrained Spatio-temporal model provides predictions reasonably well on the regularly shaped farmland areas as shown in Fig. 6 (a)-(c) for both the border and area segmentation tasks. However, the model misses more positive predictions of farmland pixels for irregularly shaped and densely packed farmland areas as shown in Fig. 6 (d)- (f) and (j)-(l), it still provides reasonable predictions. In particular, for the area segmentation task, we observe that the model is often able to predict areas that may seem like farmland areas at a cursory glance to the human eye but are instead buildings/barren areas/bushes correctly.

(a) Input (Apr-Jun, Jul-Sept) (b) Border prediction (c) Area prediction

(d) Input (Jan-Mar, Apr-Jun, Jul-Sept) (e) Border prediction (f) Area prediction

(g) Input (Apr-Jun, Jul-Sept) (h) Border prediction (i) Area prediction

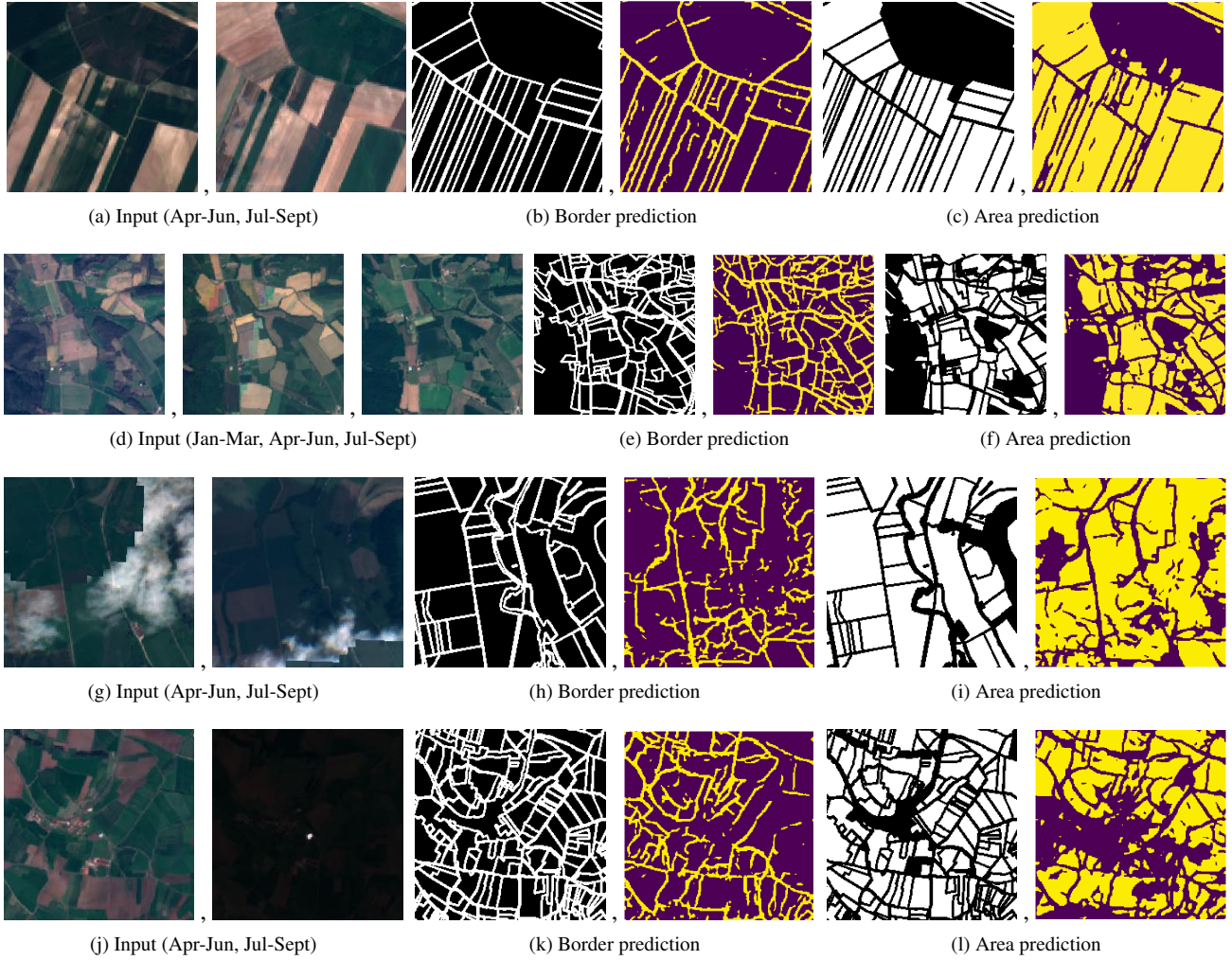(j) Input (Apr-Jun, Jul-Sept) (k) Border prediction (l) Area prediction

Figure 6: Example predictions by the Spatio-temporal U-Net. From left to right, the first 2 or 3 images (depending on availability of Jan-Mar image) are the input satellite images. The next set of two images are the ground truth binary mask and the model's prediction respectively for the border segmentation task. The final set of two images are the ground truth binary mask and the model's prediction respectively for the area segmentation task.

6.2.2 Cloud-covered areas

In cases where the input image in a particular timestamp t is obscured by clouds (as in Fig. 6 input image (g)), the pretrained Spatio-temporal model performs worse compared to cases of non-obscured input images. However, the model still provides correct predictions in certain cloud-covered regions. This may be because even though the rest of the image(s) may have slightly different farmland shapes, they still help 'fill in the gap' for the cloud-obscured image. We notice that the model still segments some farmland areas and boundaries correctly despite image(s) being obscured by the clouds.

Fig. 7 provides a more concrete comparison between the pretrained Spatio-temporal U-Net and the pretrained Spa-

tial U-Net for the same cloud-covered input satellite images presented in Fig. 6 (g). As observed for both border and area segmentation predictions in Fig. 7, the pretrained Spatial U-Net misses positive predictions of boundaries and areas in regions where clouds are present (i.e. on the upper right corner of the satellite image in April-June in Fig. 7 (a)). However, the pretrained Spatio-temporal model is able to 'fill in the gap' using images from two different time ranges (April-June and July-Sept).
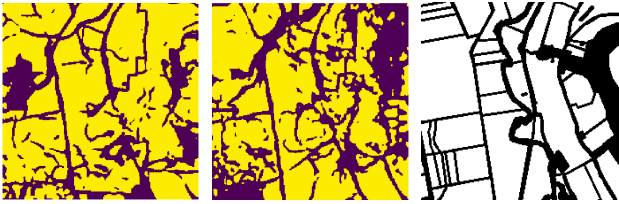
For boundary segmentation, the pretrained Spatio-temporal model outputs smoother and more connected boundaries in these cloud-covered areas compared to the pretrained Spatial U-Net. For area segmentation, the pretrained Spatio-temporal U-Net correctly predicts a large
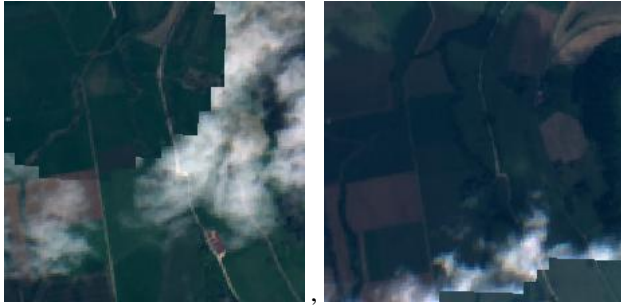
(a) Input satellite images (Apr-Jun on the left and Jul-Sept on the right)
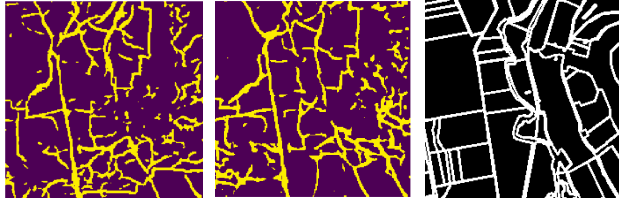
(b) Border predictions by the pretrained Spatio-temporal U-Net, pretrained Spatial U-Net and ground truth

(c) Area predictions by the pretrained Spatio-temporal U-Net, pretrained Spatial U-Net and ground truth

Figure 7: Border and area segmentation results comparison between the pretrained Spatial U-Net and pretrained Spatio-temporal U-Net. Row (a) presents satellite images with cloud coverage. Row (b) and (c) are border and area predictions from the two compared models respectively. For row (b) and row (c), from left to right, the first image is the prediction from the pretrained Spatial U-Net; the second image is the prediction from the pretrained Spatio-temporal U-Net; the third image is the ground truth.

part of the cloud-covered area as farmland pixels whereas the pretrained Spatial U-Net mostly predicts cloud-covered areas to be non-farmland area pixels. This further validates the hypothesis of Spatio-temporal models being able to 'fill in the gap'. To better handle clouds, we can use a cloud removal model as proposed in [27] that utilizes 3 Sentinel-2 images over time in an area to generate cloud-free image of the same area. Similarly to [27], we use 3 Sentinel-2 images over time and we can design a multi-task model that segments parcel boundary/areas and generates a cloud-free image jointly.

6.2.3   Error Analysis

While the pretrained Spatio-temporal performs reasonably well across different shapes of farmland areas, the model mislabels pixels for farm boundaries/areas as non-farm boundaries/areas in certain cases. Specifically, for the boundary segmentation task, the model often predicts 'broken' boundaries that should instead be connected. We notice a similar phenomenon for the area segmentation task in that the model incorrectly predicts smaller regions inside the farmland area as non-farmland pixels. We observe many of these errors in inputs where densely packed farmland regions and cloud-covered areas are present as shown in Fig. 6 (g)-(i) and Fig. 6 (j)-(l). One reason for these errors may have been due to the limitation of our dataset. An example of this could be occasional misalignment of polygons from the public dataset with the corresponding Sentinel-2 satellite image. This slight misalignment arises since we use 3-month composites for the Sentinel-2 images and a simple linear-based coordinate projection.

Furthermore, in Fig. 6 (j)-(k), where the different farmland regions have very similar colors, the model is not able to delineate boundaries and segment areas well, especially for smaller farmland areas. This error most likely arises from the limitation of using only RGB bands from the satellite images.

7. Conclusion

In this study, we proposed the use of deep learning and open source datasets to segment farm parcel areas and boundaries in satellite images. In particular, we trained variants of the U-Net model on the Sentinel-2 images given the corresponding area/boundary masks. We showed that the proposed Spatio-temporal U-Net achieves 83% Dice score outperforming all Spatial-only models by around $3 - 5\%$. This shows that the additional temporal data can better highlight the faint boundaries of farmland parcels. As a future work, to improve predictions, we will experiment with other variations of U-Nets and models along with different transfer learning approaches. To analyze more extensively the impact of temporal information, we hope to experiment with different levels of granularity of temporal information. Finally, to further validate our results, we will apply our models on other regions that may have vastly different agricultural regions using a cross-spatial split.

The proposed use of deep learning and open source datasets provides a further step towards efficient and cheap automated cadastral data collection to provide faster and more accurate land policies decisions on agriculture, and climate change mitigation and adaptation.

References

[1] Md. Zahangir Alom, Tarek Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Nasrin, Mahmudul Hasan, Brian Essen, Abdul Awwal, and Vijayan Asari. A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8:292, 03 2019. 2

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 1

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[4] Malcolm D Childress. Linking land policy with climate change: A multi-dimensional landscape approach to territorial development with a focus on the europe and central asia (eca) region. 2014. 1

[5] François Chollet et al. Keras. https://keras.io, 2015. 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, USA, June 2009. IEEE. 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3, 4, 5

[8] ESA. Observing the earth, copernicus, sentinel-2. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/Satellite_constellation, 2017. 3

[9] Dario Garcia-Gasulla, Ferran Parés, Armand Vilalta, Jonatan Moreno, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. On the behavior of convolutional nets for feature extraction. *Journal of Artificial Intelligence Research*, 61:563–592, 2018. 2

[10] A. García-Pedrero, C. Gonzalo-Martín, and M. Lillo-Saavedra. A machine learning approach for agricultural parcel delineation through agglomerative segmentation. *International Journal of Remote Sensing*, 38(7):1809–1819, 2017. 3

[11] A. García-Pedrero, M. Lillo-Saavedra, D. Rodríguez-Esparragón, and C. Gonzalo-Martín. Deep learning for automatic outlining agricultural parcels: Exploiting the land parcel identification system. *IEEE Access*, 7:158223–158236, 2019. 3

[12] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1442–1450. IEEE, 2018. 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. 1

[14] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. *CoRR*, abs/1801.05746, 2018. 3

[15] Konstantinos Kamnitsas, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. *CoRR*, abs/1603.05959, 2016. 5

[16] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 5

[17] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Gustav Hager, Alan Lukezic, Abdelrahman Eldesokey, et al. The visual object tracking vot2017 challenge results. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1972, 2017. 1

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[21] Lamiae El Mendili, Anne Puissant, Mehdi Chougrad, and Imane Sebari. Towards a multi-temporal deep learning approach for mapping urban fabric using sentinel 2 images. *Remote Sensing*, 12(3):423, 2020. 3

[22] Gideon Mutua Musyoka. Automatic delineation of small holder agricultural field boundaries using fully convolutional networks. 2018. 3

[23] Maria Papadomanolaki, Maria Vakalopoulou, and Konstantinos Karantzalos. A novel object-based deep learning framework for semantic segmentation of very high-resolution remote sensing data: Comparison with convolutional and fully convolutional networks. *Remote Sensing*, 11:684, 03 2019. 2

[24] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, Honolulu, HI, USA, July 2017. IEEE. 1

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 2, 4

[27] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *The IEEE Winter Confer-*

ence on Applications of Computer Vision, pages 1796–1805, 2020. 8

[28] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2698–2706. ACM, 2019. 1

[29] Evan Sheehan, Burak Uzkent, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. Learning to interpret satellite images using wikipedia. *arXiv preprint arXiv:1809.10236*, 2018. 5

[30] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[32] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2014. 2

[33] Burak Uzkent, Aneesh Rangnekar, and Matthew Hoffman. Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–48, 2017. 1

[34] Burak Uzkent and YoungWoo Seo. Enkcf: Ensemble of kernelized correlation filters for high-speed object tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141. IEEE, 2018. 1

[35] Burak Uzkent, Evan Sheehan, Chenlin Meng, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. Learning to interpret satellite images in global scale using wikipedia. *arXiv preprint arXiv:1905.02506*, 2019. 1

[36] Burak Uzkent, Christopher Yeh, and Stefano Ermon. Efficient object detection in large images using deep reinforcement learning. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1824–1833, 2020. 1

[37] Paul van Der Molen. The use of the cadastre among the members states: Property rights, land registration and cadastre in the european union. In *Seminar 1. The use of the cadastre. 1st congress on cadastre in the European Union*, pages 227–230, 2002. 1

[38] Pavel Yakubovskiy. Segmentation models. https://github.com/qubvel/segmentation_models, 2019. 5

[39] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 5

[40] Lichen Zhou, Chuang Zhang, and Ming Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *CVPR Workshops*, pages 182–186, 2018. 6

[41] Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells III, Ferenc A Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology*, 11(2):178–189, 2004. 5