UNIVERSITY OF CALIFORNIA

Santa Barbara

Convolutional Neural Network Based Approaches for Instance Segmentation of Irrigated Agriculture in Satellite Imagery

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

by

Ryan Barry Avery

Committee in charge:

Professor Kelly Caylor, Chair

Professor Dar Roberts

Professor Joseph McFadden

June 2020

The thesis of Ryan Barry Avery is approved.

_____

Dar Roberts

_____

Joseph McFadden

_____

Kelly Caylor, Committee Chair

June 2020

Convolutional Neural Network Based Approaches for Instance Segmentation of Irrigated Agriculture in Satellite Imagery

ACKNOWLEDGMENTS

Thank you to my parents, Steve and Erin Avery, for their support, guidance, and trust.

ABSTRACT

Convolutional Neural Network Based Approaches for Instance Segmentation of Irrigated Agriculture in Satellite Imagery

by

Ryan Barry Avery

Irrigated agriculture makes up the large majority of consumptive water use, and demand for water has greatly increased over the 21st century. To date, fine scale information about where irrigated agriculture is expanding is difficult to acquire due to: 1) a lack of manually delineated field boundaries due to: 1) a lack of centralized planning and management of agricultural development and 2) the limited ability of shallow machine learning models based on these limited data to generalize beyond small geographies or accurately map instances using remotely sensed imagery. Convolutional neural networks (CNNs) have been demonstrated to outperform shallower models with less learned non-linear feature transformations, such as decision tree based ensembles or SVMs, in image classification and segmentation of true color photography. To date there has been little research on the performance of CNN-based models for segmentation in the field of remote sensing. This thesis examines the performance of Mask R-CNN and Fully Convolutional Instance Aware Segmentation, two CNN-based methods for segmenting objects in images. There is a need to evaluate how these new methods perform in the remote sensing domain, given that images in these datasets tend to have a higher variance of objects and the geospatial dataset labels are

more limited in number compared to large image corpora, like ImageNet and COCO, which are used to test the performance of deep learning image recognition algorithms. Results show that true color Landsat 5 scenes can be used to produce sufficiently accurate instance detections of center pivot fields, even with a coarser resolution relative to newer sensors such as Sentinel-2. This opens the possibility of using Landsat's long historical record for longitudinal studies of irrigation and cropping dynamics of center pivot agriculture.

TABLE OF CONTENTS

1. Introduction

1.1. Problem Statement & Research Objectives

Irrigated agriculture exerts strong controls on global food production yields and the water cycle while accounting for 85-90 percent of human consumption (Shiklomanov 2000; Döll 2009). However, little is known about the spatial distribution of agricultural fields, their crop types, or their methods of irrigation. Spatially explicit knowledge of these field attributes is necessary in order to implement more water-efficient agricultural practices and plan for more sustainable economies (Foster et al. 2019). However, most remote sensing based mapping efforts cover limited political boundaries, on the order of U.S. states or smaller, and usually only cover a snapshot in time. Agriculture maps in developing countries are even more lacking in semantic detail, coverage, and resolution, which is a particularly acute problem given that agricultural expansion in these regions tends to be decentralized and without a guiding management plan for water sustainability. While a fine scale and up-to-date census of global agriculture does not yet exist, it is feasible that we can map a subset of fields that are spectrally and visually distinct from surrounding landcover and numerous enough to train a machine learning model to map a substantial subset of global agriculture.

Center pivot irrigation fits these criteria; they are relatively uniform in shape, have a narrow range of sizes within particular geographies, and in drylands, can be strongly contrasted with the surrounding lack of vegetation (Figure 1). Like all remote sensing applications for agriculture detection, image obstruction by clouds, fallow periods and the growth of non-agricultural natural vegetation in the landscape all pose challenges for models that detect center pivots. In more humid environments, center pivots have distinctive

growing season patterns and larger amplitude changes in vegetation “greenness” compared to other vegetation types, making them readily identifiable with time series of multispectral images. However, it’s difficult to assemble a comprehensive time series of imagery over many parts of the world, primarily due to cloud occlusion and scene availability. Scene availability is particularly low for dates prior to the launch of the Sentinel-2, so a method to map center pivots using single date imagery in many parts of the world is desirable. Such a method is particularly valuable given that center pivots are one of the most ubiquitous irrigation sprinkler systems employed in large scale commercial agriculture and make up a considerable fraction of the unplanned agricultural expansion in developing countries.

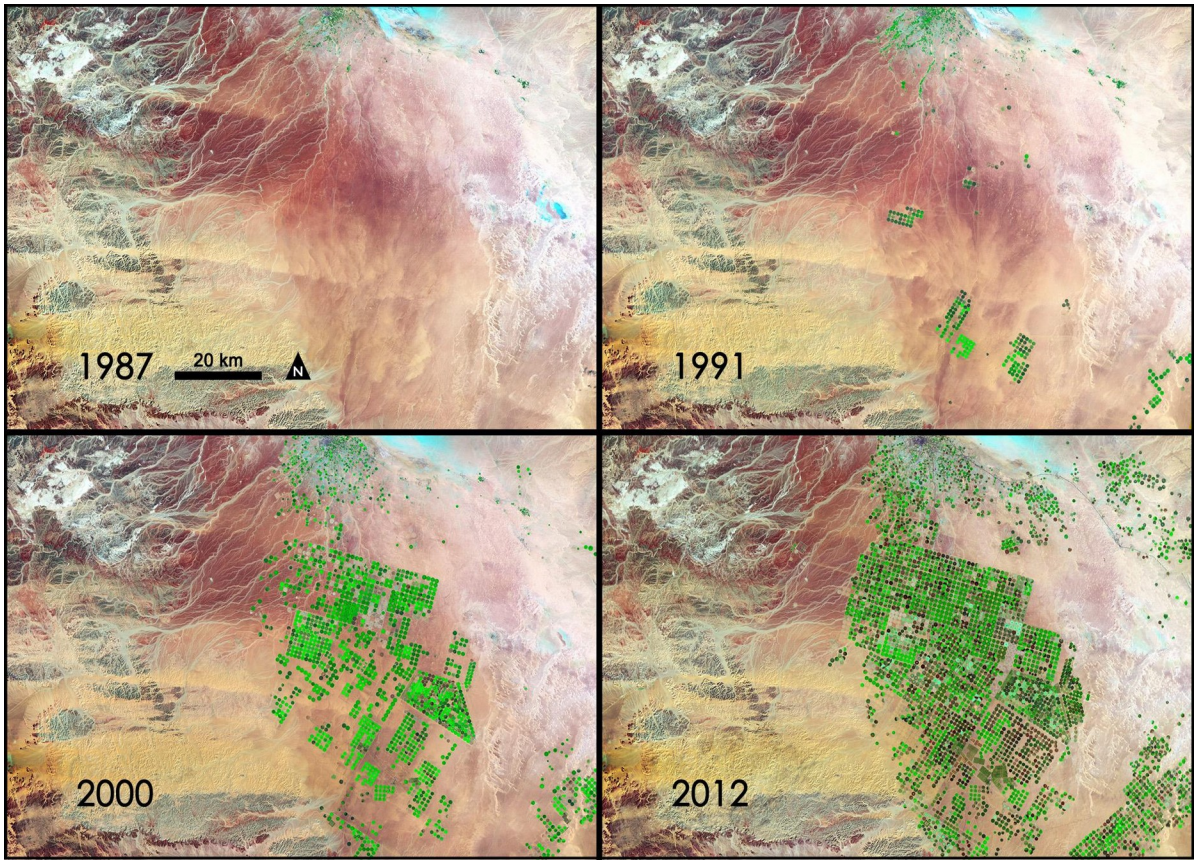

Fig. 1: This Landsat time series of a SWIR, NIR, Green Band composite shows the massive development of center pivot fields by Saudi Arabia in the Syrian Desert. Each field measures approximately 1 kilometer in diameter ("NASA Sees Fields of Green Spring up in Saudi Arabia" 2012).

There are a variety of methods for mapping objects using remotely sensed imagery, but object based classification has been shown to outperform per-pixel classifiers in cases where the object of interest contains enough pixels to distinguish objects by textural or shape properties (Myint et al. 2011; Blaschke and Strobl 2001). Because center pivots are large relative to the resolution of public sensors like Landsat, they are amenable to being mapped

using object based classifiers. The traditional approach to object based classification in remote sensing has been to use manually tuned algorithms to delineate edges or engineered features that capture texture and shape properties combined with per-pixel machine learning algorithms (Rydberg and Borgefors 2001; Mathieu, Freeman, and Aryal 2007). However, traditional object based classifiers in remote sensing have a tendency to overfit and must be manually tuned or supplemented with region specific post processing to arrive at a suitable result. Another class of methods which make use of convolutional neural networks have achieved great success on complex image recognition problems in true color photography. These have not been thoroughly evaluated for mapping agricultural fields as instances across a large climate gradient using Landsat imagery.

The goals of this research are twofold. First, I evaluate the performance of convolutional neural network (CNN) based instance segmentation models on Landsat imagery. This experiment determines if CNN based models can make use of Landsat's 30 meter resolution to provide reliable predictions of the locations and extents of center pivot agriculture in various states of development, including cleared, growing, and fallow. I test this approach by using the current most popular and near state of the art Mask R-CNN model, an approach based on a lineage of regional CNNs which jointly minimize prediction loss on region proposals, object class, refined object bounding boxes, and an object's instance mask. This model is tested on the 2005 CALMIT Nebraska Center Pivot Dataset, which is divided into geographically independent samples that were partitioned into a training, validation and testing set. Multiple model runs with varying hyperparameters and preprocessing steps were conducted to arrive at the most accurate result on the validation set, and the final most accurate model was applied to the test set to produce the final reported accuracy. The model

is also evaluated on the full training data set and 50% of this dataset in order to examine the effect of reduced training data on model accuracy over a large region.

Second, I compare these results to the Fully Convolutional Instance Aware Segmentation model, which was previously the state of the art in instance segmentation in true color photography prior to Mask R-CNN. My goal is to determine whether Mask R-CNN and FCIS produce substantially different results when applied to the task of segmenting center pivots, and if these differences are as pronounced as their performance on segmentation tasks in true color photography. This can inform the field of remote sensing on the value of adopting new CNN based architectures and to what extent performance on datasets like COCO translates to the remote sensing domain. The FCIS model is tested on the same Nebraska dataset with minimally altered hyperparameters based on Rieke (2017), which applied FCIS to the task of segmenting large agricultural fields in Denmark. It is evaluated on a validation set, as the validation set was not used to make decisions to change hyperparameters. Both the FCIS model and the Mask R-CNN model performances on the Nebraska dataset are compared relative to their results on Common Objects in Context, a large image dataset similar to Imagenet but with added annotations for segments of objects. Imagenet is a dataset with 3.2 million images labeled with categories appearing in the images, while COCO or Common Objects in Context, contains 328,000 images with 2.5 million labeled instance segments (Deng et al. 2009; Lin et al. 2014). In summary, this thesis addresses four research questions:

1. Does Mask R-CNN outperform the FCIS model on the Nebraska dataset?
2. Does Mask R-CNN outperform FCIS to a similar extent that it outperforms FCIS on the COCO dataset?

3. Does the decision of Rieke (2017) to incorporate more training data relative to this study compensate for having a more difficult detection target (fields of more variable shape instead of center pivots)?
4. To what extent is performance reduced if training data is reduced by 50%, given that CNN's typically require large training datasets to be effective?

Answering these questions can help inform users of these models of their compared to one another and whether differences in performance are consistent across image domains. The ultimate goal of this research is to improve the mapping of field boundaries that are either actively cultivated, recently cultivated, or soon to be cultivated using the public Landsat record. The dataset used does not allow for classifying particular crop types, though the Mask R-CNN method allows for this. Nevertheless, accurate field boundaries obtained from CNNs could make it much easier to classify crop type by allowing crop type classification procedures to focus on known field locations.

1.2. Characteristics of center pivots and their environmental and political relevance

The popularity of center pivots stems from their low energy consumption relative to the amount of water applied, low maintenance costs, and uniformity (Waller and Yitayew 2016). These advantages make center pivot irrigation productive in both temperate regions and drylands with access to groundwater. However, while center pivots have been highly productive, their unchecked expansion in water-scarce regions has contributed to water shortages, over-tapped aquifers, and even political conflict. These risks are exemplified by Saudi Arabia in the 1970s and 80s, when center pivot development expanded and into the present day, where domestic production has slowed and must eventually decline ("FAO Country Profiles:Saudi Arabia" 2012, Luckey et al. 1981). Like many of the world's most

agriculturally productive regions that depend on fossil water from aquifers, Saudi Arabia's overall rate of irrigation has long since eclipsed groundwater recharge rates back to the Arabia Aquifer. As early as 1951, it was noted that the first government agricultural project at Al Kharj was dependent on ancient aquifer waters, and that parts of the project were already "... operating close to the margin of safety in regard to the current (1948) water supply." (Crary 1951). Since then, the government of Saudia Arabia has transformed large tracts of it's deserts into fields of center pivots (Figure 1, ("NASA Sees Fields of Green Spring up in Saudi Arabia" 2012), resulting in a tripling of total water use between 1980 and 2006, 936% of total renewable water (Frenken and Others 2009).

This boom in irrigated area has been followed by a curtailing of domestic production, with the government pledging to eliminate water intensive wheat production by 2016 (DeNicola et al. 2015), though this commitment has since been postponed. Though researchers have called for more investment in demand side controls on water resource sustainability, the government has invested heavily in direct foreign investments in water intensive agriculture abroad. Some of these investments are ongoing in areas experiencing extreme water scarcity, including southern Arizona, which depends on declining flows from the Colorado river, and Gambela, Ethiopia, where a government mandate to relocate smallholder farmers for a Saudi agriculture project led to armed conflict (Pearce 2012; Halverson 2015). There is a lack of public information on the distribution and ownership of large scale commercial agriculture, particularly in developing countries. Rectifying this knowledge gap will allow us to make more informed decisions of how to allocate water resources to best serve ecosystems, industry, and local populations. In regions where this information is available, it can enable improved monitoring of water rights compliance and

experimentation with potentially more sustainable water management practices. A potentially successful strategy for mapping agriculture fields in developing countries where data is scarce is to train machine learning models using geospatial labels of fields located in regions with similar climates.

The High Plains Aquifer (HPA) region encompasses Nebraska, Kansas, Colorado, and Texas, spanning a climatological gradient that contains semi-arid and humid regions and supplies nearly one third of all groundwater irrigation applied in the United States. Since the invention of center pivot irrigation in Nebraska in the 1950s, the HPA has experienced significant water level declines as a result of groundwater pumping for irrigation (Dennehy, Litke, and McMahon 2002). These were noted as early as the 1980s (Luckey et al. 1981), yet substantial expansions in irrigated area continued to occur into the 2000's, with Nebraska adding 1.3 million hectares of irrigated area, a 16.3% increase in total irrigated area for the state (Brown and Pervez 2014). Projections indicate that, assuming pumping at the same linearly extrapolated historic rate, overuse of groundwater will leave 50% of the southern and central HPA dry by 2025 and 2065 respectively (Haacker, Kendall, and Hyndman 2016).

Since the 1980s, management strategies have been evaluated on the basis of whether or not they increase the sustainability of groundwater resources in the HPA while preserving local economies. These strategies fall into two broad categories: water conservation strategies for increasing water use efficiency or limiting irrigation, and water productivity strategies that focus on increasing yields. Specifically, these strategies have ranged from switching irrigation methods from high pressure applicators to low energy precision agriculture or subsurface drip irrigation (Colaizzi et al. 2009), converting land use from

water intensive wheat to cotton (Colaizzi et al. 2009), converting irrigated farmland to dryland agriculture (Deines et al. 2020), and variable rate irrigation. However, (Pfeiffer and Lin 2014)) found that the dominant crop for drop strategy in Kansas between 1995 and 2005, switching from highly pressurized applicators to low energy precision agriculture (LEPA) systems, still coincided with an unsustainable increase in the total volume of irrigation. The increased profits and water savings generated by the technology switch enabled growers to expand irrigated acreage as well as switch to water intensive corn, alfalfa and soybeans, contributing to a consistent declining trend in the southern and central HPA. Colaizzi et al. 2009 also recounts that for the southern HPA in Texas between 1958 and 2000, despite many center pivot systems switching to more efficient LEPA irrigation methods, total irrigation volumes increased with total irrigated acreage. Deines et al. (2020) underscores the consequences of a continued present trend of unsustainable groundwater management; they estimate that of the current irrigated acreage across the HPA, possibly 24% would undergo a forced transition to dryland agriculture by 2100, 13% of which is likely unsuitable for crop production, resulting in substantial negative impact to local economies. This body of research suggests that strategies to increase water use efficiency or yields are not by themselves sufficient to sustainably manage finite ground water resources.

Groundwater management areas that have the power to tax, educate, and regulate jurisdictions were evaluated on their effectiveness throughout the HPA in Nebraska, Kansas, and Texas. This analysis used well hydrographs, which were missing or absent in some jurisdictions, making a complete analysis of management area effects difficult (Haacker et al. 2019). This data gap can be supplemented by satellite and machine learning derived estimates of field locations and water use, where irrigation supplied is sourced from

groundwater. In regions of the world without good, publicly available records of well hydrographs, remote sensing and machine learning will need to play a role in monitoring and evaluating groundwater use and groundwater management policies. Understanding the spatial location and extent of center pivot agriculture is critical to monitoring, enforcement, and evaluating the effectiveness of groundwater management. Machine learning models trained on remotely sensed imagery can provide real time predictions on location and extent.

1.3. Traditional segmentation approaches in remote sensing

In 1986, Strahler et al. identified two classes of models remote sensing - low resolution models where the pixel size is larger than the object of interest and high resolution models where an object is resolved by many pixels (Strahler, Woodcock, and Smith 1986). For machine learning problems in remote sensing, the scale of the object of interest determines whether a low resolution or a high resolution model can be employed to locate and classify the object of interest. In many cases, low resolution approaches have been used to classify pixels using only spectral information, even in cases where the high resolution method is available to map objects of interest in finer spatial and semantic detail.

Random forest has been the most popular pixel based classifier of choice in remote sensing and has been used to produce global and regional, coarse resolution LULC maps from MODIS imagery (H. K. Zhang and Roy 2017; Thenkabail, Schull, and Turral 2005). The spatial resolution of the imagery used for global land cover mapping (500 meters-1 km) obfuscates any spatial features that can be used to distinguish land cover classes. For global, coarse, land cover mapping, random forest is a more appropriate method than a CNN model pretrained to interpret hierarchies of spatial features. Random Forest has also been applied for regional mapping of particular land cover categories, including irrigated area (Xu et al.

2019; Deines et al. 2019). While pixel-wise, irrigated area maps are useful, the method used to generate them discards valuable information embedded in the labels, which distinguish between unique fields.

Instead of using pixel-wise methods, which only consider spectral information at a point as a feature to classify a pixel, object oriented classification methods that make use of spatial features provided by higher spatial resolution can be used to map objects in finer spatial and semantic detail. These techniques require information about the size, distribution, temporal pattern, and ranges of reflectance in order to be effective, and this can lead to overfitting to particular regions. For circular feature detection, Hough transform methods have been employed, which require prior knowledge of the radii of the objects. However, the Hough transform must be tuned to changes in illumination that happen over a scene or multiple scenes, and also has problems handling changing topographic relief. While it has been successfully used to extract very visually distinct, perfectly circular oil tanks, as well as circular geologic features, each case required visual inspection, manual tuning (Cross 1988; Argialas and Mavrantza 2004). In the case of (Cross 1988; Argialas and Mavrantza 2004), the scenes tested were very small (two 314 by 260 pixel SPOT-5 images), allowing for easier manual tuning and less scene diversity for the method to handle. In the case of Cross 1988, there were substantial commission errors. Generally, center pivots are not amenable to this family of techniques since they are not always perfect circles, come in a varying range of sizes which makes the Hough transform more computationally burdensome, can have variable brightness within the field, and are present across the world in varying topographic relief and illumination.

All of the aforementioned techniques have a limited capacity to learn complex patterns for object recognition relative to the current state-of-the-art, CNNs. This was demonstrated clearly in 2012 on the task of classifying photographic scenes, with CNNs outperforming the previous state of the art (Krizhevsky, Sutskever, and Hinton 2012). There is still a need to evaluate how CNN based instance segmentation models perform on coarser resolution sensors like Landsat OLI, given the more limited size of labeled datasets compared to large image datasets like Imagenet and COCO.

1.4. Neural Networks and CNN Model Implementations

The simplest model of a neural network can be described as a pipeline which takes as input training data and performs a sequence of linear regressions and nonlinear functions, with each layer in each sequence operating on the output of the previous layer (Figure 2). When used in a supervised machine learning task, the goal in training a neural network is to learn the set of all parameters *w* and *b* across all layers of the neural network in order to achieve minimal loss relative to a reference dataset.

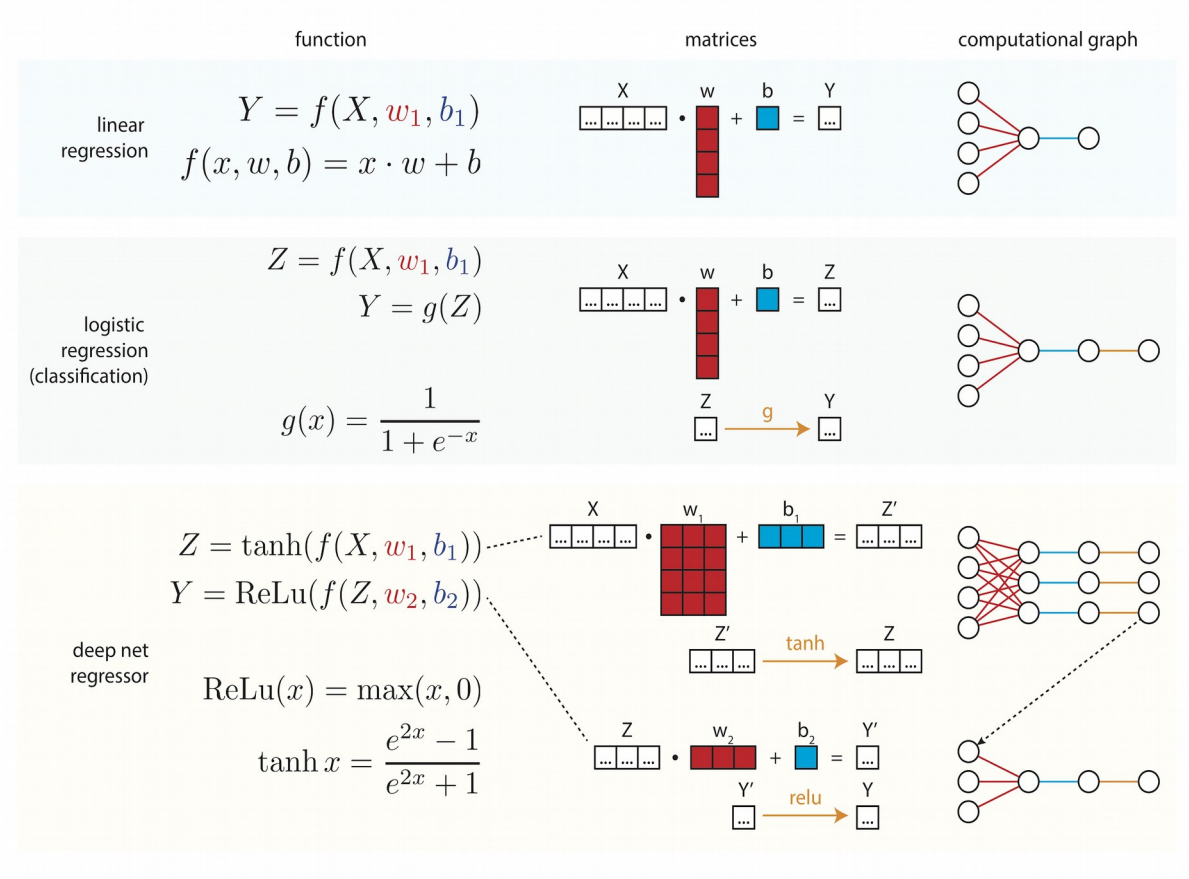

Fig. 2: This depicts the components of a neural network and a full neural network in 3 different visual formats: as mathematical functions, matrices, and computational graphs. In each case, *f* is the linear regression function. Red lines represent weight parameters, blue lines represent bias parameters, and orange lines represent the application of a nonlinear function (Ma 2020).

The learning process is conducted through gradient descent, where a model's parameters are first initialized and then the model is tested on data that has corresponding reference labels. The difference between the results and the reference labels are then used to adjust model parameters by a learning rate through a method called backpropagation, a

method for efficient auto differentiation. Model weights are either randomly assigned or taken from a previously trained model, which is referred to as “pretraining” a neural network model. The difference between the prediction and the reference labels is measured by a loss function, which varies depending on the task the neural network is applied toward. For example, a common loss function for regression tasks is mean squared error. If the task is complex, the sum of multiple loss functions can be used to allow the model to learn to achieve multiple performance targets.

Neural networks can have an infinite number of parameters, and at this limit, a neural network model can fit any function. However, more parameters equates to more complex models, longer training times, and a higher tendency to overfit the data. For visual processing tasks, the simple neural network described above is quite inefficient, as each the first neural network layer, or the set of neurons that take as an input the pixels of the flattened multichannel image, would need to have a number of weights equal to the product of the height, width, and number of channels in the input. Because of memory limitations in modern computing hardware, a fully connected neural network with multiple layers could easily max out available memory.

CNNs solve this problem by replacing linear layers with convolutional layers (Figure 3). A trained CNN model learns a hierarchy of features rather than a set of weights and biases, and the assumption of locality, that nearby pixels will be highly correlated, and that the features learned are translation invariant, improves both model accuracy and computational efficiency. Compared to fully connected neural networks, CNNs are more efficient representations of the spatial patterns commonly found in imagery and are a suitable option for complex image recognition tasks (Lecun and Bengio 1995).

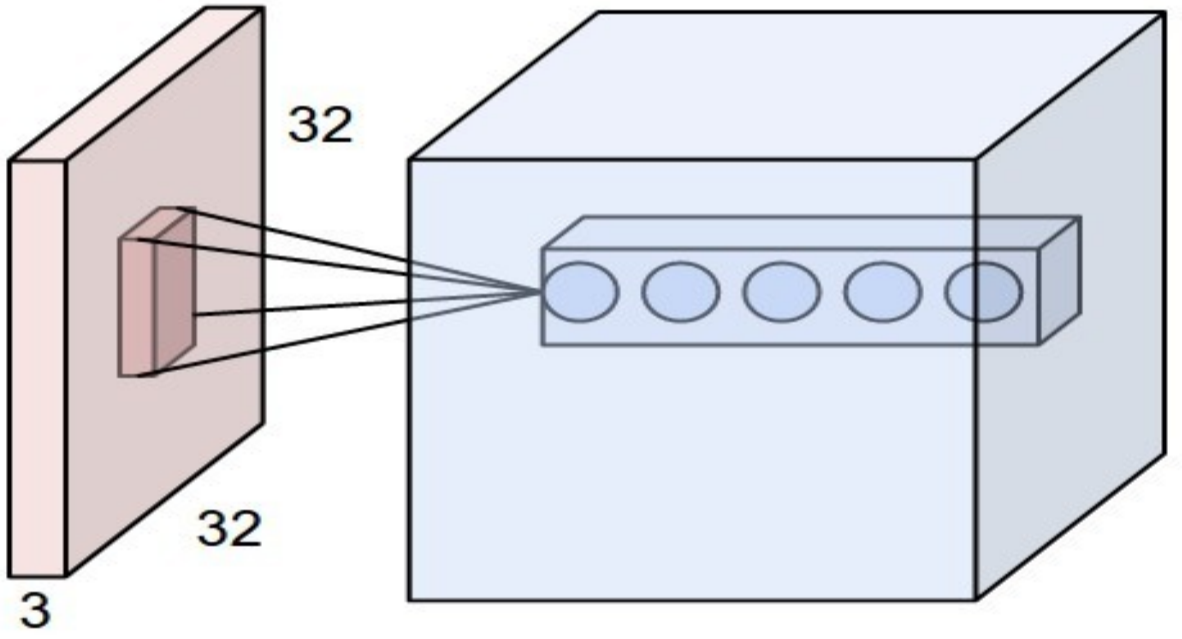

Fig. 3: Depicts a first and second layer of a CNN. The red layer represents a 32x32x3 image and the blue block is a convolutional layer. The height and width of the convolution layer are the same dimensions as the image and the depth is the number of features learned. In this case, each of the five circles represents five features that correspond to a particular location in the original image. Each of these five circles is like a neuron in a fully connected neural network, except each represents a set of neurons which are locally connected to the input pixels in a local spatial region rather than all pixels in the flattened image. This local region is called the receptive field and is depicted by the black lines describing the receptive field of the feature map (“CS231n Convolutional Neural Networks for Visual Recognition” n.d.).

The shape of a convolutional layer depends on three model hyperparameters: the stride of the moving window filter that computes the feature maps, the depth, or number of feature maps to be computed in a given layer, and the amount of zero padding to add to the

output feature map. These hyperparameters control the size of the output of each layer in the network and are thus crucial for determining the shape of the ultimate result. Setting these network hyperparameters such that the shape of the output of the network matches the input of the network is the first requirement for a CNN to learn to predict the category of each input pixel. Alternatively, if the shape of the output is the length of the potential categories, each element of the output array will correspond to the likelihood scores that a given input image corresponds to each category. Besides convolutional layers, there are two other layers used in basic CNNs. These include a rectified linear unit (RELU) layer, a nonlinear function also used in fully connected networks to give neural networks the capacity to learn non linear features (Glorot, Bordes, and Bengio 2011), and a max pooling layer, a type of downsampling which takes the maximum value within each feature map of a convolutional layer to reduce the dimensionality of deeper convolutional layers to provide more computationally tractable learning by resulting less parameters to learn (Goodfellow, Bengio, and Courville 2016). These two layers and the convolutional layer depicted above are the primary building blocks used in model architectures for image classification, bounding box detection, semantic segmentation, and instance segmentation. For a complete review of the historical development of different CNN architectures for semantic and instance segmentation, refer to Rieke 2017.

Semantic segmentation and instance segmentation are more complex detection tasks than image classification and thus require further changes to hyperparameters, the basic convolutional layer, and additional layers to preserve alignment between input pixels and output pixels. The historical progression of Regional CNNs (R-CNNs) illustrates how specific model architectures are imperative to accurate instance segmentation. Girshick et al.

(2013) showed that by training CNNS to extract features from precomputed proposals, they could achieve state of the art performance on bounding box detection and segmentation (Girshick et al. 2013). However, in a follow on paper, (Girshick 2015), the authors note several drawbacks to the R-CNN method: it is computationally slow because features are computed separately for each precomputed region, the training process requires that the bounding box location and category of the region are trained sequentially (rather than jointly), and applying the model to an image takes a long time (47 seconds using a GPU).

Fast R-CNN rectified these setbacks by training a set of feature maps up front, then extracting feature vectors from feature maps in a step called ROIPooling, using precomputed region proposals, which ensures that feature vectors correspond to regions in an image. These feature vectors are then used to jointly train a network to minimize two loss functions, one which relates to the likelihood scores for each category of interest and another which relates to the 4 coordinates of the bounding box of each of the possible categories. This effectively reduces N sets of feature maps for N objects to a single feature map (Girshick 2015). However, with Fast R-CNN there is still a computational bottleneck in generating the precomputed region proposals, which are trained separately from the rest of the Fast R-CNN network. Faster R-CNN rectifies this by using a region proposal network, or RPN (Figure 4). The RPN circumvents a limitation of neural networks in that they typically return an output of fixed size, whereas there can be a variable number of objects within a given image. The RPN operates by first distributing a predefined number of uniformly distributed potential bounding boxes across an image (which are referred to as anchors in the literature), and then training a convolutional network to 1) recognize if a given bounding box contains an object of any class of interest and 2) to adjust the coordinates of the bounding box to better fit the

object. Once the regions have been trained, the next step of the network is effectively ROIPooling and Fast R-CNN (Ren et al. 2017). This improvement made it possible to apply an object detection model at near frame rate speeds.

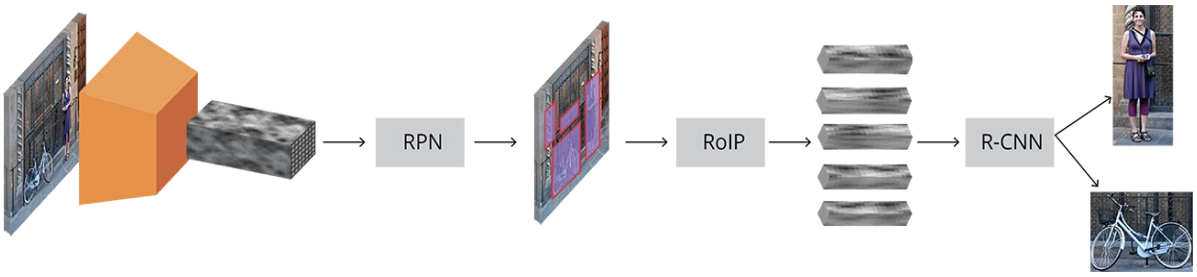

Fig. 4. This illustrates the steps in Faster R-CNN. Starting from the left side, the orange box represents the set of convolutional layers used to produce a set of output features (The grey and black box). To generate initial uniform bounding boxes, this set of features is sampled such that the input image is sampled uniformly, since the shape of the features is proportional to the input image. Then, the RPN network predicts the change in the center of the bounding box and the change in it's width and height to produce regions, which are shown as blue boxes in the middle. ROIPooling extracts the feature vectors for each region, and then ultimately R-CNN is used to learn the object location and class (Tryolabs 2018).

Subsequently, (He et al. 2017) released Mask R-CNN, which adds on top of Faster R-CNN the ability to compute an instance segmentation mask to describe not just the bounding boxes around objects but also the locations and classes of all pixels associated with individual objects. This was achieved by improving ROIPool using bilinear interpolation to more accurately extract feature maps corresponding to region proposals, and then using these feature maps to compute a binary mask for each class, as well as a class score vector, and bounding box coordinates.

Fully Convolutional Instance Aware Segmentation is an alternative instance segmentation algorithm similar to Mask R-CNN, released a year prior, and the state of the art prior to the release of Mask R-CNN. Instead of using a mask branch that is jointly trained alongside a region proposal network, bounding box regressor, and region classifier, FCIS produces a semantic segmentation within each region proposal, which is used to calculate portions of the mask that are part of an object and portions of a mask that are not part of an object (Li et al. 2016).

For this study, I used the FCIS model distributed by researchers affiliated with Microsoft Research Asia and written in an older 2017 version of the mxnet deep learning library (Li et al. 2016). This implementation was also used by (Rieke 2017), and is difficult to reproduce and configure for new detection tasks, while also being two years older than the current, near state of the art Mask R-CNN model. The results from Reiki 2017 were reproduced and results from the FCIS model were compared against a Mask R-CNN model implementation distributed by Facebook AI Research (Wu et al 2019, (He et al. 2017)). The Mask R-CNN model is implemented in Pytorch with Facebook AI Research's Detectron2 framework for object detection, and is packaged as a python library and in a Docker image, making it easily configurable and reproducible.

1.5. Deep learning approaches applied in remote sensing

Previous work has been done to map agricultural fields, including center pivots. However, in many cases these efforts have been directed toward simpler detection targets, such as pixel-wise classification or bounding boxes, or they have not adequately evaluated the ability of machine learning models to generalize to other, similar geographies. In one case, different CNN backbone architectures were tested for mapping center pivots based on

the assumption that fields are perfect circles (C. Zhang et al. 2018). However, these models were not properly validated or tested on geographically independent scenes. Instead the models were tested on the same locations at different growing seasons, which likely allowed the models to be substantially overfit to the study location. In another study, the popular U-net architecture was applied for pixel-wise mapping of center pivot agriculture (Saraiva et al. 2020). An independent validation set was used in this case. The U-net model architecture was used to generate pixel-wise classifications, which is a simpler detection target than instance segmentation and the underlying imagery used was Planetscope, a commercial dataset from Planet Labs, San Francisco, CA that has 3 meters resolution (https://assets.planet.com/docs/combined-imagery-product-spec-april-2019.pdf). There has yet to be a thorough evaluation of Landsat as a viable image source for instance segmentation tasks on large agricultural fields.

The FCIS model has been evaluated on Sentinel-2A image chips for segmenting field boundaries in Denmark. The author found that the FCIS model produced encouraging results, with high recall and precision for large fields (any field with greater than .1 km^2 area) and medium size fields (between .025km^2 and .1 km^2 area), but performed poorly for smaller fields with less than .025 km^2 area (Rieke 2017).

Rieke 2017 found that transfer learning and finetuning can speed up the training process enormously and enable learning on small datasets. However, the model was tested on a single image mosaic composed of 2 Sentinel-2A scenes over a smaller area with less geographic diversity than Nebraska, which is about 7 times the area of Denmark. Furthermore, Sentinel-2A's historical record is much more limited than Landsat, and it

remains to be tested to what extent the Landsat record can be used for instance segmentation of commercial agriculture, given its coarser spatial resolution of 30 meters.

1.6. Challenges when predicting field boundaries

A supervised machine learning model for predicting field boundaries is only as good as the data that is used to train it. For geospatial datasets describing the locations and timings of presence of agricultural fields, there are unique sources of bias compared to computer vision datasets, such as Imagenet.

Geospatial labels describe regions undergoing change, and are often not distributed with the corresponding image scene or scene metadata taken at the time that the labels were created. This can necessitate a "best-guess" approach for pairing labels with imagery based on the metadata that is available. It is worth noting that a new geospatial data standard, Spatio-Temporal Asset Catalog, corrects this shortcoming in geospatial data formats, but it's adoption is in the early stages at the time of this writing ("SpatioTemporal Asset Catalog" n.d.). Besides issues with metadata not always describing scene correspondence, geospatial labels are often only valid for a single time period. Since these datasets are rarely audited and updated, it may be difficult or impossible to link a satellite sensor's operating lifespan with both the availability of the geospatial labels and the date when predictions need to be made. Virtual constellations that combine the high resolution and high temporal revisit rate of Sentinel-2 and Landsat make field level mapping more achievable, but only since Sentinel-2A has been in operation (Wulder et al. 2015). Therefore, the limited temporal availability of a sensor record can make it difficult or impossible to apply a single machine learning model over a particular region, for a particular time period of interest.

Spatial resolution is another limiting factor to consider when creating accurate training data of field boundaries to develop machine learning models. Depending on the fuzziness of the field's boundary and the spatial resolution used to resolve it, a sensor may be unable to resolve the boundary of a field in all environments, leading to more within class variability for the class of interest and more similarity between the class of interest and the background class. While there has been a proliferation of new privately owned satellite data providing resolutions of 3 meters or less which can resolve smaller objects, there has not been a corresponding growth in public, high quality geospatial labels which can be used to train machine learning models.

Along with all of these sources of error or bias which are unique to earth observations, there are also more common challenges, including the need for undefined but large amounts of training data, image interpretation error during training data creation, and unknown/unreported error in reference data (Elmes et al. 2020). Because the cost of acquiring, processing, and labeling earth observations is much higher compared to photographic images, generating large training data sets for machine learning algorithms is much more difficult. Difficulties in image interpretation during the labeling stage are often much higher when working with remotely sensed images, due to spatial resolutions that are at the limit of being able to resolve objects, or because agriculture field classes appear similar to other field classes; crop types in particular are difficult to visually distinguish. In an ongoing study referenced by Elmes et al. (2020), the accuracy of trained field annotators was evaluated against an expert generated reference set. A skill statistic that accounts for true positive and false positive rate was generated for each annotator, and each image was mapped by at least five annotators. The study found that there was a large range in

agreement with the reference data among annotators, indicating that the costs of generating high quality training data from earth observations may also need to cover multiple annotators per unit of annotation as well as expert derived reference data to assess the accuracy of difficult annotations.

Because of the high cost and lack of efficient tools for developing high quality training geospatial labels, most research using machine learning and earth observations develop or make use of small, regional datasets that are constrained to arbitrary political boundaries. These constraints bias models toward geographies which contain political entities with the means to produce labeled training data. It's an open question to what extent these data can be leveraged to make predictions outside of the limited geography and time they were created.

2. Methods

2.1. Study Area

Nebraska is a high producer of cereal crops such as corn, millet, and soybean. These crops are used for biofuels, animal feed, and human consumption. Center pivots are a dominant feature across the state, particularly along the Platte river. However, there are multiple confounding features that make monitoring of field level landscape change difficult, including non-irrigated row crops, forested areas, circular hills, and the variability of center pivots themselves, which can be multi cropped, semicircular, and in various states of development ranging from cleared, to irrigated, to fallowing, to fallow. The majority of these fields are irrigated using groundwater from the Northern High Plains Aquifer (NHPA). While the Northern High Plains has not experienced as much groundwater depletion as the

Southern and Central HPA, portions of it, including the Upper Republic Natural Resource District, have experienced water table declines up to 40 meters, mostly to irrigate corn (Haacker, Kendall, and Hyndman 2016; Foster et al. 2019). Compared with the Southern HPA (SHPA) and Central HPA (CHPA), the NHPA is less depleted due to historically later stage irrigation development, higher recharge, and higher precipitation. The NHP also contains larger volumes of water relative to the SHPA and CHPA, 2940 km3 versus 636 km3 and 171 km3, respectively (Haacker, Kendall, and Hyndman 2016).

2.2. Training and Reference Dataset and Imagery

I used the Nebraska 2005 Center Pivot Inventory dataset, developed by the Center for Advanced Land Management Technologies and Nebraska Department of Natural Resources ("[No Title]" n.d.). This dataset represents a full census of center pivot field vectors for the growing season of 2005, which spans June-September (Figure 5). A limitation of this dataset's metadata is that the exact Path/Row and Date combinations used for basemaps is unknown, so the full combination of Landsat 5 dates and scene path/row IDs used to generate the Nebraska 2005 Center Pivot Inventory were used to train, validate, and test the model. The least cloudy Landsat 5 scenes out of this set of scenes were selected such that all scenes with greater than 5% cloud cover were discarded. Because NAIP images served as basemaps to digitize the center pivot field vectors and because it is ambiguous exactly which individual or multiple Landsat 5 scenes were used to label center pivots, some inaccuracies in correspondence between the imagery and the labels inevitably exist. In total, 32 mostly cloud free Landsat 5 Analysis Ready Data (ARD) scenes were used to generate samples to train,validate, and test the model.

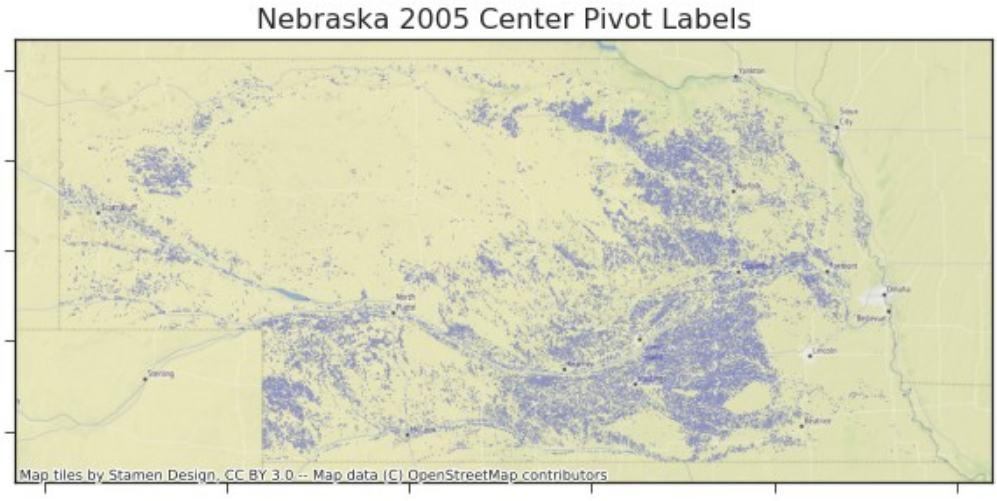

Fig. 5. This map shows the state of Nebraska with all recorded center pivots for the 2005 growing season. Center pivot fields, colored blue, are distributed throughout the state, from the semi-arid western half of the state to the more temperate and humid eastern half.

In total, 52,127 center pivot fields are labeled across the state of Nebraska. The majority of these pivots are between .6 and .8 kilometers squared (Figure 6). The large majority of these are complete circles, though a substantial portion of these labels represent semicircles.

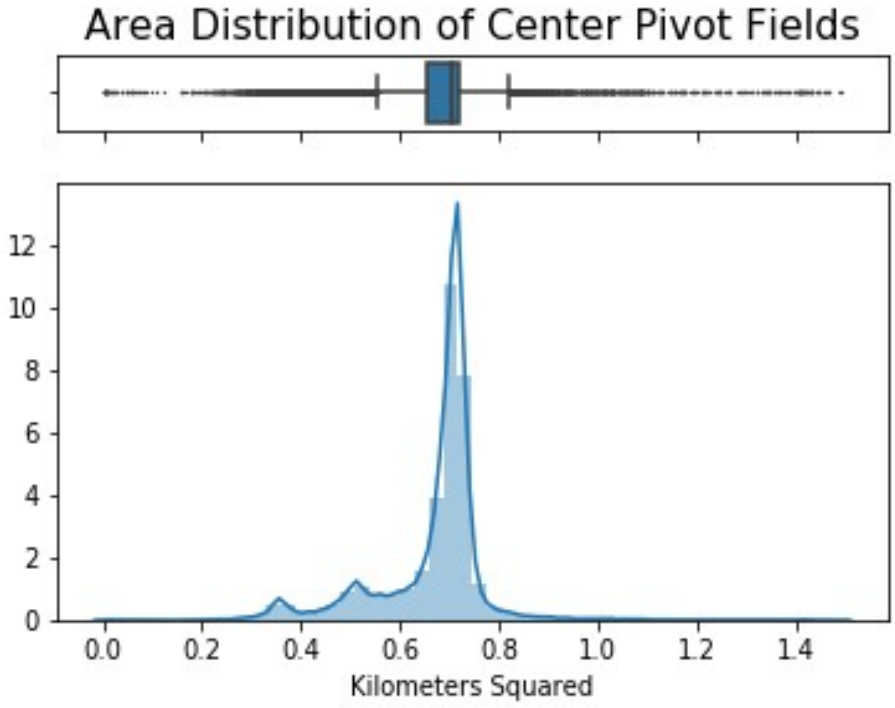

Fig. 6. This histogram and box plot of the area distribution of center pivot fields shows a strongly unimodal pattern, with much smaller peaks between .3 and .6 kilometers squared. There is a long tail representing very large center pivots.

I downloaded Landsat 5 ARD scenes for all path/row combinations intersecting Nebraska and dates within June 2005 and September 2005 in geotiff format. XML metadata for each scene was used to filter and select scenes below a 5% cloudy pixel percentage. No cloud masking was performed, as the Landsat 5 cloud mask was visually inspected and found to be substantially confused by other landscape features, including agricultural fields.

2.3. Preprocessing

Before training the model on the dataset and Landsat 5 images, inputs went through preprocessing steps to be formatted and sized small enough for the Mask R-CNN and FCIS model to be able to train (Figure 7).

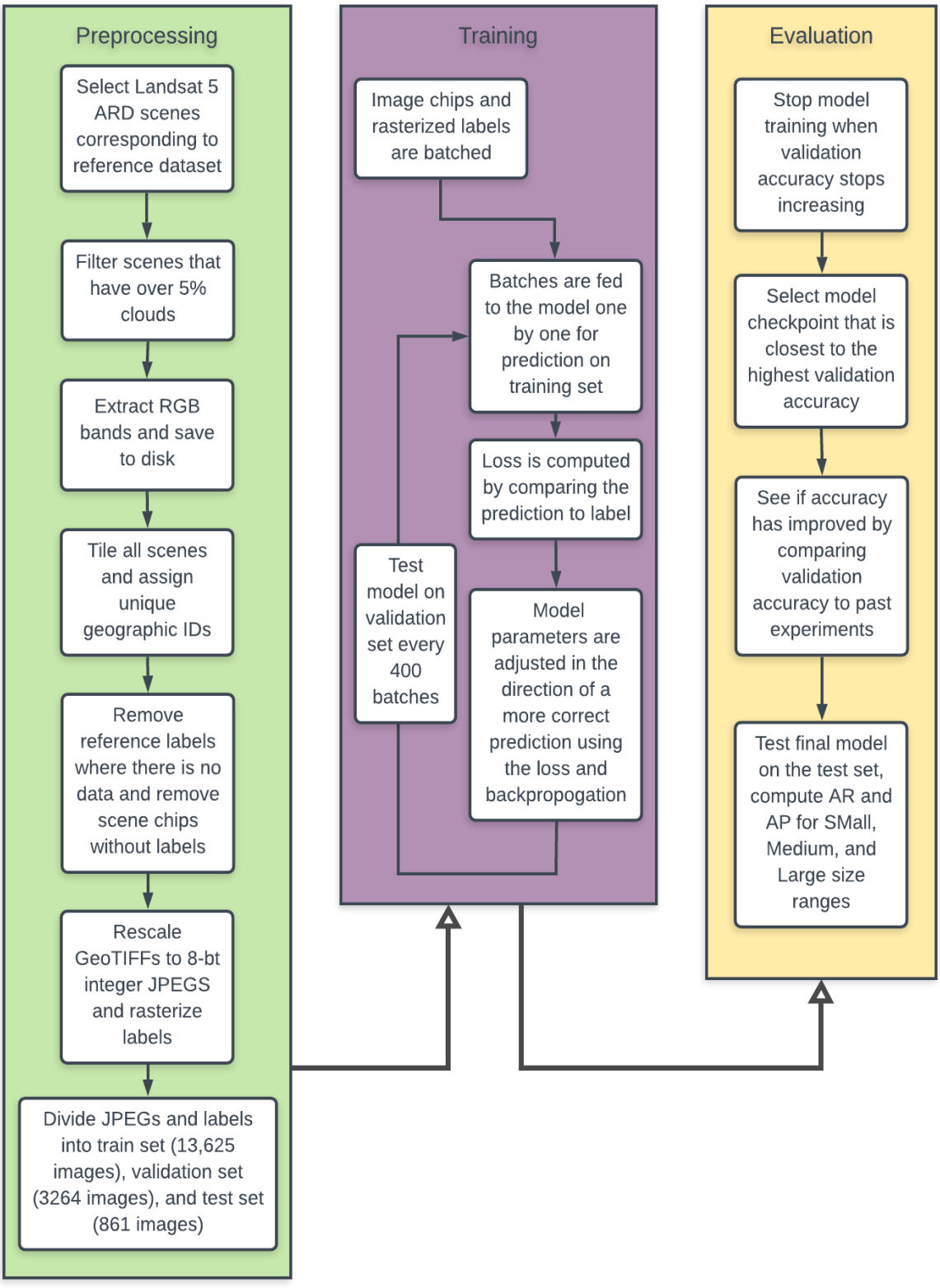
Preprocessing
Select Landsat 5 ARD scenes corresponding to reference dataset
Filter scenes that have over 5% clouds
Extract RGB bands and save to disk
Tile all scenes and assign unique geographic IDs
Remove reference labels where there is no data and remove scene chips without labels
Rescale GeoTIFFs to 8-bt integer JPEGS and rasterize labels
Divide JPEGs and labels into train set (13,625 images), validation set (3264 images), and test set (861 images)
Training
Image chips and rasterized labels are batched
Batches are fed to the model one by one for prediction on training set
Loss is computed by comparing the prediction to label
Model parameters are adjusted in the direction of a more correct prediction using the loss and backpropogation
Test model on validation set every 400 batches
Evaluation
Stop model training when validation accuracy stops increasing
Select model checkpoint that is closest to the highest validation accuracy
See if accuracy has improved by comparing validation accuracy to past experiments
Test final model on the test set, compute AR and AP for SMall, Medium, and Large size ranges

Fig. 7. This details the preprocessing, training, and evaluation workflow for the Mask R-CNN model runs. The preprocessing steps for the Mask R-CNN model runs and the FCIS model were identical. The FCIS model implementation is older and has limited functionality for profiling validation or training set accuracy during model training. Therefore, the FCIS model was only tested on the validation set after training with the configuration used by Rieke 2017.

This research contributed new functionality to the python library solaris, which was used to tile large Landsat 5 scenes into 128 by 128 pixel image chips for training models while also excluding portions of scenes that fell outside of the Nebraska state boundary, where no labels were digitized. A 128 pixel x 128 pixel sample was selected in order to utilize the maximum amount of gpu memory to speed up the training process and to provide the model with more geographic diversity in each batch during the training step. Since many features are computed for each image that is used to train the model, GPU memory use for each image is much larger than the actual image memory footprint, and therefore all images must be served iteratively in the training process. Landsat 5 scene chips were saved as both geotiffs to preserve geospatial metadata, and rescaled, 8-bit JPEG files. The JPEG files were used for model training, since detectron2's default training logic expects JPEG formatted images at the time of this writing. Landsat 5 ARD scenes contain large amounts of NoData at scene edges, so vector labels that overlapped with these NoData areas were masked out. As tiff and JPEG files were tiled, vectors were also tiled and saved as geojson files. Fields at scene edges that had an area smaller than 4000 m^2 (approximately 4 Landsat pixels) were discarded in order to remove processed labels that do not represent half or full center pivot

fields. Solaris python functions were also used to rasterize geospatial vectors into multi channel tiffs, where each channel represents a binary mask of a unique field instance for training. These rasterized masks were paired with and tiled to the same size as the corresponding imagery.

Two different training sets were used to evaluate the effect of training data size on model accuracy. The first set of processed scene chips were taken from the least cloudy scenes across the state of Nebraska during the 2005 growing season, such that no samples overlapped geographically. These image chips were divided into a set of training, validation, and test chips, with 13,625, 3264, and 861 samples assigned, respectively. The second training data set used 50% of the training data set used by the first set. These splits were determined by collecting geospatial IDs for all image chips and randomly assigning chips with particular geospatial IDs to each set. This means that all sets are geographically independent from each other, though there is inevitably scene similarity since most of Nebraska has a humid climate and agricultural landscape. A given set will have some geospatial IDs represented multiple times, in cases where Landsat 5 captured multiple scenes during the study period. This provides more representation of center pivot fields at different stages of development, from cleared for development, to cultivating, to harvested.

Blue, Green, Red, and NIR bands were used to train two sets of models in RGB and GRNIR 3-band combinations. In each case, the channel-wise means for the whole training set were calculated in order to normalize the models inputs by subtracting the channel mean from each individual scene's respective channels. A Resnet-50 CNN backbone loaded with pretrained weights from Imagenet was downloaded and used as a starting point for training and fine-tuning the FCIS and Mask R-CNN models (He et al. 2015). Fine tuning from

pretrained weight is a common practice in deep learning that has been shown to lead to faster model convergence with similar ultimate model accuracies and less hyperparameter tuning, though recent work has shown that pretraining does not always lead to higher overall accuracies (He, Girshick, and Dollár 2018).

2.4. Training process

All of the following Mask R-CNN and FCIS results were obtained from a Standard NC6 Microsoft Azure virtual machine, with 6 virtual cpus, 56 GB memory and 4 Titan V GPUs (Wikipedia contributors 2020). Mask R-CNN models took roughly two hours each to train, while the FCIS model took 8 hours to train, given the implementation's limitation of having a batch size of one image and only being able to use one GPU at a time.

For the Mask R-CNN model, a standard detectron2 training loop was used. Table 1 contains the hyperparameters used for the model, including the optimizer, learning rate, batch size, etc. Particular configurations were changed from the defaults in order to adapt the training process for the Nebraska dataset. Relative to Imagenet or the COCO dataset, the Nebraska dataset has a larger average and a wider range in the number of instances per image. Therefore, the maximum number of detections was set to 100. The default number of warmup iterations was decreased in order to stop the learning rate from increasing to an amount that would make convergence difficult, and the initial base learning rate was increased by an order of magnitude for a faster learning process, given that we started from pretrained weights on a training dataset size that is more limited than Imagenet or COCO. All decisions to adjust these hyperparameters were made after reviewing average precision (AP) and average recall (AR) metrics and loss on the validation set relative to the training loss during the training process (Table 1).

Referring to Table 1, the non max suppression configurations were increased to improve detections on scenes with multiple center pivot images. Non max suppression is a step that removes low confidence region proposals prior to later steps in the model that refine the bounding box around an object and generate a final object mask. However, the result from this setting did not substantially differ from the default. Likewise, detections per image was increased so that scenes with many instances did not have missed detections because of an arbitrary limit. The Model Freeze setting determines which parameters are allowed to be learned at different stages of the pretrained model. By changing the setting from 0 to 2, parameters that represent higher level features were learned. This did not impact model accuracy compared to using the default but it did speed the training of the model. The max iteration for warmup was adjusted so that the learning rate increased during training faster, given that the model was run for comparatively less iterations than the default model, which is set up to run on the larger COCO dataset.

Table 1: Hyperparameters that were changed from the defaults for the Mask R-CNN model. The full model configuration file is available at https://github.com/ecohydro/CropMask_RCNN/blob/master/config.yaml. And the pretrained weights used can be found at https://dl.fbaipublicfiles.com/detectron2/COCO-InstanceSegmentation/mask_rcnn_R_50_FPN_3x/137849600/model_final_f10217.pkl

Number of region proposals to generate during training prior to Non Max Suppression	Number of region proposals to generate during testing prior to Non	Model weights	Detections per image	Max iteration for warmup learning (determines rate of	Pretrained model frozen at iteration number

	Max Suppression			learning rate increase)	
12000	6000	See caption	200	4000	2

The AP performance metric and loss curves informed when to stop the model from training, the condition being when AP and validation loss began to show a consistently increasing trend. The final model configuration exhibited the lowest validation loss and highest validation performance metrics, and this model was used to generate the following figures and metrics on the test set.

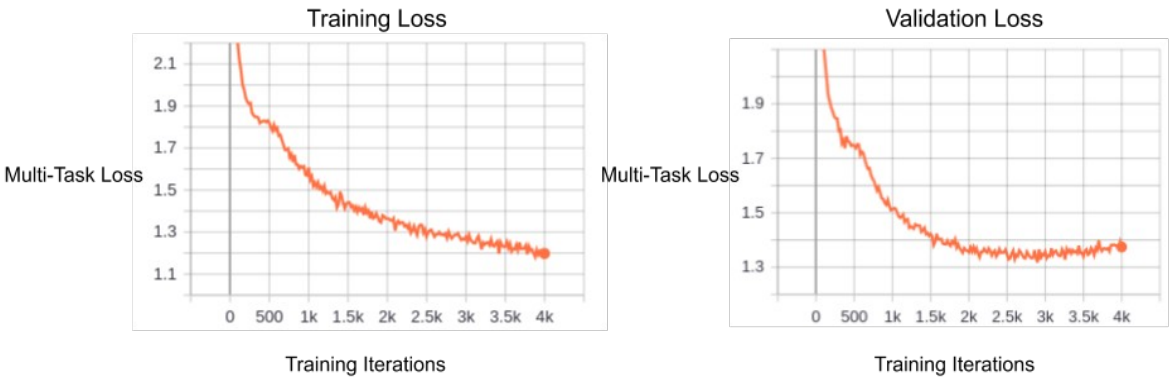

Fig. 8. Training loss continues to decrease to the point of overfitting, since the model is trained to improve loss against reference labels in the training set. Validation loss starts to increase around iteration 2500, which indicates overfitting. The model checkpoint used for testing is selected that is at the minimum validation loss.

For the FCIS model, no hyperparameters were altered from those used by Rieke 2017, except for the maximum number of detection instances, which was set to 50 per 128x128 pixel image. This upper bound was increased based on the larger number of instances present

in each sample. The validation set was not used to tune hyperparameters for the FCIS model since only one model was run to reproduce the results of Rieke 2017 and one model was run to test the FCIS architecture on the Nebraska Dataset. No test set was evaluated for the FCIS model since no validation set was used to tune hyperparameters.

2.5. Evaluation

Precision is calculated as the count of correctly detected positives over all detected positives (Equation 1). Recall is calculated as the count of correctly detected positives over all true positives (Equation 2). A correct identification is defined by the intersection over union (IoU) ratio between the detection and reference label, in order to account for the model's ability to not only correctly ascribe a class to a group of homogenous pixels but also to accurately localize the field.

$$precision = \frac{TP}{TP + FP}$$

Equation 1. Where TP stands for the count of true positives. A detection is considered a true positive if it has an IoU score greater than 0.5, and has the highest likelihood score among all instances that have an IoU greater than 0.5 with the reference label. FN stands for False negative, a lack of a detection where there is a reference instance.

$$recall = \frac{TP}{TP + FN}$$

Equation 2. For recall, the denominator is the sum of true positives, TP, and actual positives erroneously classified as negatives, FN.

Average Precision (AP) and Average Recall (AR) are metrics that account for the variation in confidence scores associated with model detections as well as the trade-off between precision and recall (Equation 3). Each detection has a confidence score ascribed to it by the model, or a sequence of confidence scores in the case of multi class classification. A confidence score threshold determines whether a prediction should be considered or not considered when calculating precision or recall. Lower confidence scores tend to increase recall by potentially keeping more predictions while higher confidence scores decrease recall as lower confidence predictions that match the reference labels are not considered. To calculate Average precision, predictions are generated for all images and ranked by confidence score. Precision and recall are then calculated for a range of confidence score thresholds (Equation 4, Equation 5). The range of precision and recall values are then averaged to provide an Average Precision and Average Recall for an entire validation or test dataset. This process is repeated for IoU values ranging from .5 to .95, and the results are then averaged to report an average of averages, which is commonly referred to as AP:.05-.95 and AR:.05-.95, or just AP and AR. Henceforth, we will refer to AP:.05-.95 and AR:.05-.95 as AP and AR. All metrics are computed using the Microsoft COCO API ((Lin et al. 2014), https://github.com/cocodataset/cocoapi).

$$\mathrm{AveP} = \int_0^1 p(r)dr$$

Equation 3. Average precision is defined on the left as the integral of the precision-recall curve, which describes the relationship between precision and recall at varying confidence score thresholds.

$$\text{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

Equation 4. Average precision is calculated as a finite sum of the product of the change in recall and the precision value at rank k, where rank k relates to the confidence score threshold used to calculate precision and recall.

$$\text{AveP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1.0\}} p_{\text{interp}}(r)$$

$$where\ p_{\text{interp}}(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r})$$

Equation 5. In practice, the Microsoft COCO API interpolates the precision-recall curve at 11 points corresponding to 11 confidence score thresholds such that the maximum precision score within a range of recalls is used to calculate average precision. This is done in order to reduce the effect that ranking order has on the final result. AR is calculated in the same manner.

AP and AR are calculated for specific object size ranges to evaluate the model's ability to generate high confidence predictions across different field sizes. Field ranges used for both the FCIS and Mask R-CNN models are divided as follows: the Small size range is 0 - 0.43 km^2, Medium is 0.43 - 0.52 km^2 (the interquartile range), and >0.52 km^2 is the Large size range. These metrics aren't sufficient by themselves to resolve whether the Mask R-CNN or FCIS model performed better on the Nebraska dataset. Therefore, to understand this aspect of performance, I inspected multiple examples visually, stratifying by qualitative landscape complexity. I plotted and compared high confidence detection results (above a

90% confidence score) from Mask R-CNN and the FCIS model in order to evaluate the most accurate predictions from each model that would be most likely to be used in subsequent analysis.

3. Results

After training the models and stopping based on the condition of maximum validation accuracy (or in the case of the FCIS model, 8 model epochs), comparisons were made between 1) Mask R-CNN on the Large Training Set vs FCIS on the Large Training Set, 2) Mask R-CNN on the Large Training Set vs Mask R-CNN on the Small Training Set, and 3) Different Object Size Ranges for Mask R-CNN on the Large Training Set. For each comparison, I evaluated differences in average precision and average recall (Tables 2, 3, and 4) to compare overall model performance across different field size ranges.

3.1. Metrics

Table 2: Mask R-CNN average precision and average recall metrics for the Large Training Set run. Metrics are computed on the Test set for each size range.

Test Set Performance Metrics for Mask R-CNN at IoU .50 : .95 for Large Training Set		
Field Size Range	Average Precision	Average Recall
Small (<.43 km^2, 477 Landsat 5 pixels, 25% of fields)	0.420	0.563
Medium (.43 - .52 km^2, 477 - 578 Landsat 5 pixels, 50% of fields)	0.732	0.814
Large (.52< to a max of 2.83 km^2, 578< to a max of 3144 Landsat 5 pixels, 25% of fields)	0.734	0.763

Table 3: FCIS model's average precision and average recall metrics for the Large Training Set run. Metrics are computed on the Test set for each size range.

Test Set Performance Metrics for FCIS at IoU .50 : .95 for Large Training Set		
Field Size Range	Average Precision	Average Recall
Small (<.43 km^2, 477 Landsat 5 pixels, 25% of fields)	0.429	0.512
Medium (.43 - .52 km^2, 477 - 578 Landsat 5 pixels, 50% of fields)	0.764	0.817
Large (.52< to a max of 2.83 km^2, 578< to a max of 3144 Landsat 5 pixels,25% of fields)	0.693	0.771

Metrics from the FCIS and Mask R-CNN models were also compared to Rieke 2017's results on detecting fields as a single class using Sentinel-2 10 meter RGB imagery. Since I used Landsat in this study, the size ranges were converted to units of pixel area for comparison.

Table 4: Results from Rieke (2017) for detecting fields as a single category.

Validation Set Performance Metrics for FCIS at IoU .50 : .95 for Reike 2017, Single Field Category		
Field Size Range	Average Precision	Average Recall
Small (<.025 km^2, <250 Sentinel-2A pixels, 55.9% of fields)	0.281	.395
Medium (.025 - .1 km^2, 250 - 1000 Sentinel-2A pixels, 32.1% of fields)	0.473	0.589
Large (> .1 km^2, >1000 Sentinel-2A pixels, 11.9% of fields)	0.51	0.686

The final comparison made was between Mask R-CNN results on the full training set vs 50% training data in order to evaluate the ability of Mask R-CNN to handle smaller training datasets that are more common in remote sensing.

Table 5: Mask R-CNN model's average precision and average recall metrics for the Small Training Set (50% training data) run. Metrics are computed on the Test set for each size range.

Test Set Performance Metrics for Mask R-CNN at IOU .50 : .95 for Small Training Set		
Field Size Range	Average Precision	Average Recall
Small (<.43 km^2, 477 Landsat 5 pixels)	0.306	0.416
Medium (.43 - .52 km^2, 477 - 578 Landsat 5 pixels)	0.676	0.785
Large (>.52, >578 Landsat 5 pixels)	0.690	0.761

Tables 1 and 2 show that FCIS slightly outperformed Mask R-CNN in AP and AR for the medium category and AR for the large category, while Mask R-CNN performed slightly better in terms of AP on the large category. Mask R-CNN achieved 0.732 AP and 0.814 AR while the FCIS model achieved 0.764 AP and 0.817 AR on the medium size category, which contains 50% of all groundtruth instances. For the large size category (containing 25% of fields), Mask R-CNN achieved 0.734 AP and 0.763 AR while the FCIS model achieved 0.693 AP and 0.771 AR. These results show that FCIS and Mask R-CNN's ability to approximately locate center pivot fields to within 50% IoU is nearly equivalent, given their similar metrics for 75% of fields. Since the AP for each model's large and medium size range is equal to or greater than 69.3%, both models can be expected to correctly detect reference fields to within an IoU of 50% at least 73.2% of the time for the

medium size category. The AR value for each model's medium size range indicates that the proportion of true positives to all positives is at least 81.3%. Performance for small fields is low for both models, substantially underperforming detections in the medium and large category in terms of both AP and AR (Table 1, Table 2). The FCIS model achieved an AP and AR of 0.429 and 0.512 respectively, while Mask R-CNN achieved an AP of 0.420 and an AR of 0.563. FCIS had a lower AR value for the small category by 5.1% while Mask R-CNN had a lower AP by 0.9%. This means that Mask R-CNN is the more successful model in terms of being able to minimize false positives when detecting small center pivots while still maintaining a high rate of true positives relative to all reference labels. As expected, the large area category was mapped with greater accuracy than the small area category, but slightly less accurately than the medium category, in terms of overall AR and AP metrics.

From inspecting visual results and given the reduced occurrence of small false positives in the Mask R-CNN model, the Mask R-CNN model was selected for further analysis of the effect that training has on model accuracy. Performance on 50% of the training dataset resulted in an AP of 0.676 and an AR of 0.785 on the medium category, an AP of 0.306 and an AR of 0.416 on the small category, and an AP of 0.690 and an AR of 0.761 on the large category. This amounts to a decrease in performance by 5.6% AP and 2.9% AR for the medium category, -4.2% AP and -0.02% AR for the large category and -11.4% AP and -14.7% AR for the small category, relative to the results that used the full training set. Decreasing the amount of training data degraded performance the most for the small size category, which saw a large increase in false positives and false negatives. The medium category experienced a slightly larger decrease in performance than the large size category in terms of both AP and AR percentage.

3.2. Mask R-CNN and FCIS Visual Results

While these metrics illustrate the model's ability to roughly map an object's boundary, given the minimum criteria for an object detection is an IoU of 50%, it is also informative to inspect visual differences in how well object boundaries match the reference labels. For comparisons 1 and 3, examples were chosen across scenes of varying landscape complexity in order to provide a more comprehensive survey of the CNN models' ability to map fields using relatively coarse resolution satellite imagery. This analysis primarily focuses on examples where the model fails to correctly detect fields with high confidence. I also evaluated the distribution of likelihood scores associated with predictions for each example image to determine if there were differences in how confident model predictions were across difficult to classify scenes (Figures 10-19). The red line in the boxplot represents the median of all confidence scores for detections in the small, medium, and large categories, while the red triangle represents the mean across all detections. In each boxplot of detection likelihood scores, the distribution shown is for all detections made for a particular scene, including center pivots that did not meet the 90% confidence threshold and were not displayed.

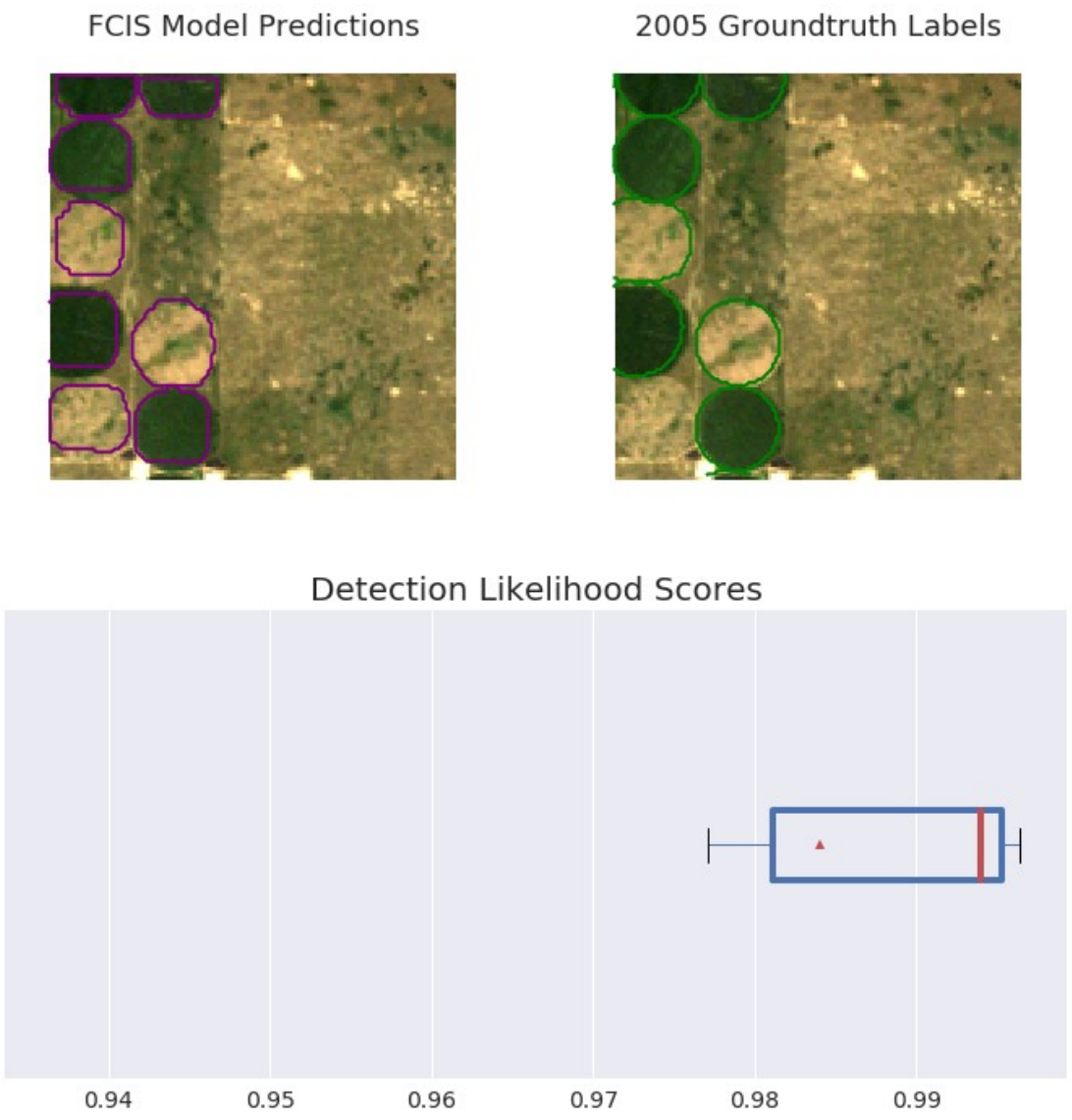

Fig. 10. This sample represents a slightly difficult scene, given the contrasting cleared or fallow pivots and cultivated pivots. Even though the detection confidence scores are high (a consistent tendency in all FCIS results), it is clear that a reference label was missed and that some boundaries are not accurately resolved relative to the reference labels.

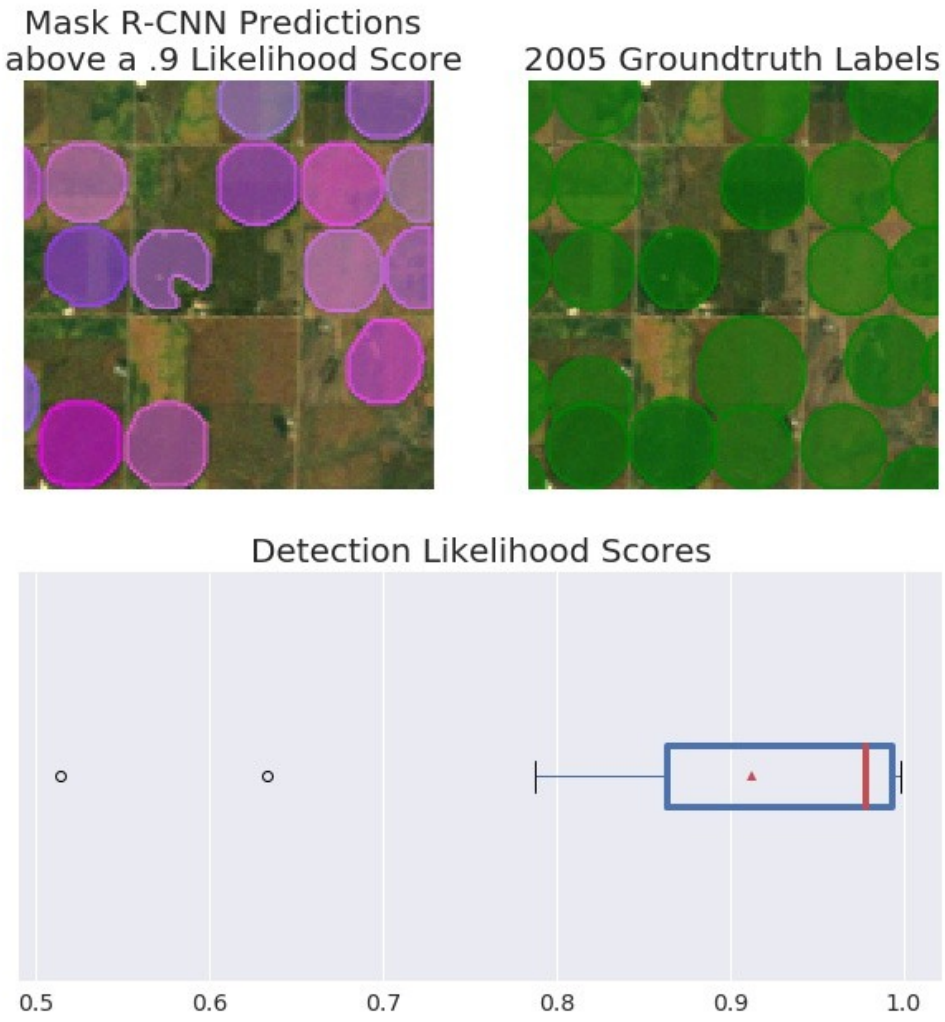

Fig. 11. This scene contains many reference labels of center pivots in close proximity to each other. Some fields are missed but most fully circular fields are classified with high confidence. Some reference labels appear to not match the imagery in that center pivots are absent in the top left and bottom right of the scene. Multiple center pivot detections are

mapped with confidence lower than 90%, these are primarily fields that are brown in color and are not actively cultivated, as well as truncated fields at the scene boundary.

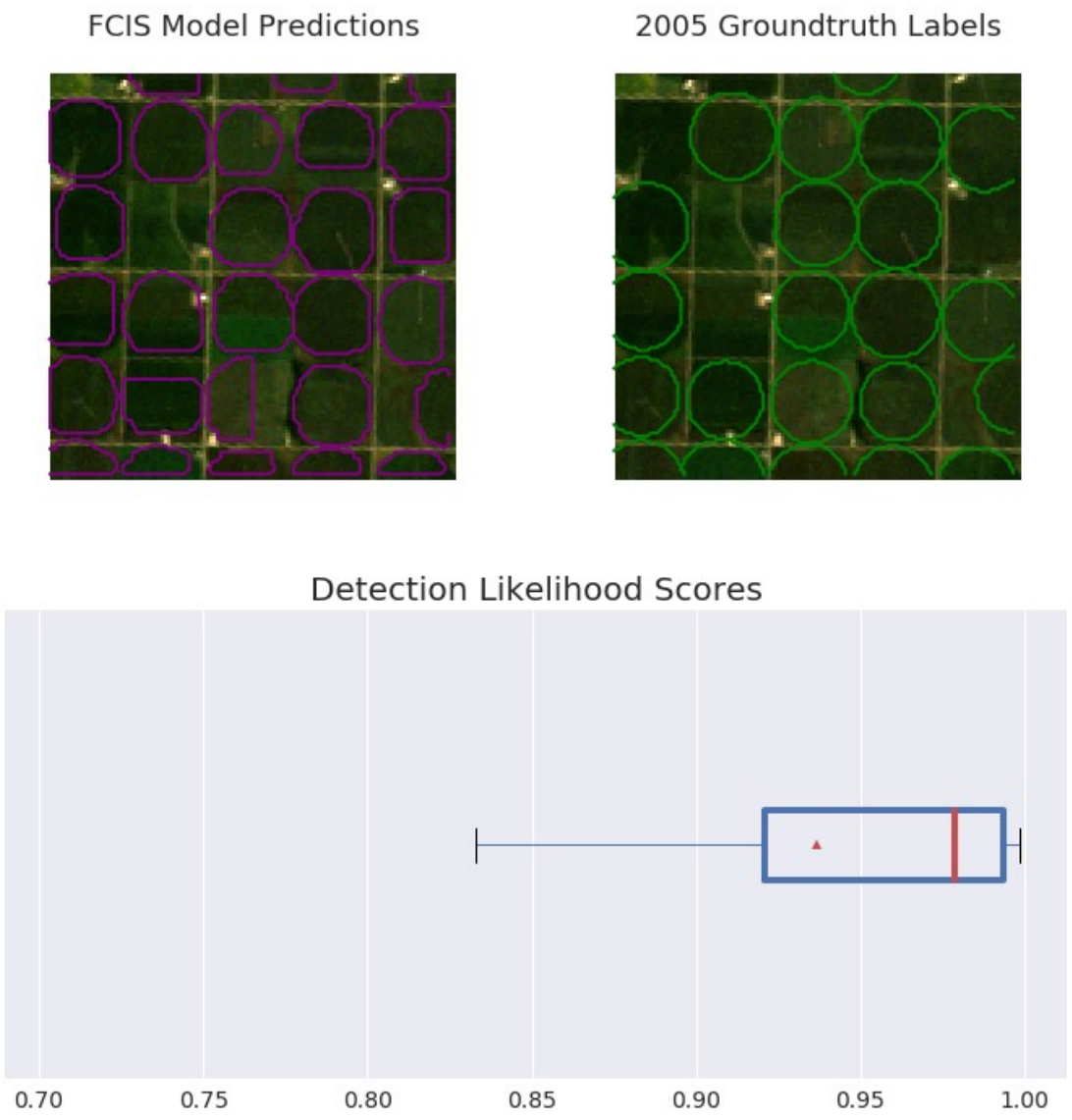

Fig. 12. This sample represents a scene with many center pivots in close proximity to each other. Most are correctly classified, with a few exceptions and some false positives, which

looks to be a result of indistinct boundaries. Even though most are correctly classified, almost all have eccentric boundaries. In contrast to Figure 11, there are no low outlier detection scores below 70%. This could be a result of the easier landscape qualities of this scene relative to Figure 10 or because of the tendency of the FCIS model to not generate low confidence detections.

The following two scenes are most difficult to classify due to the surrounding non center pivot fields.

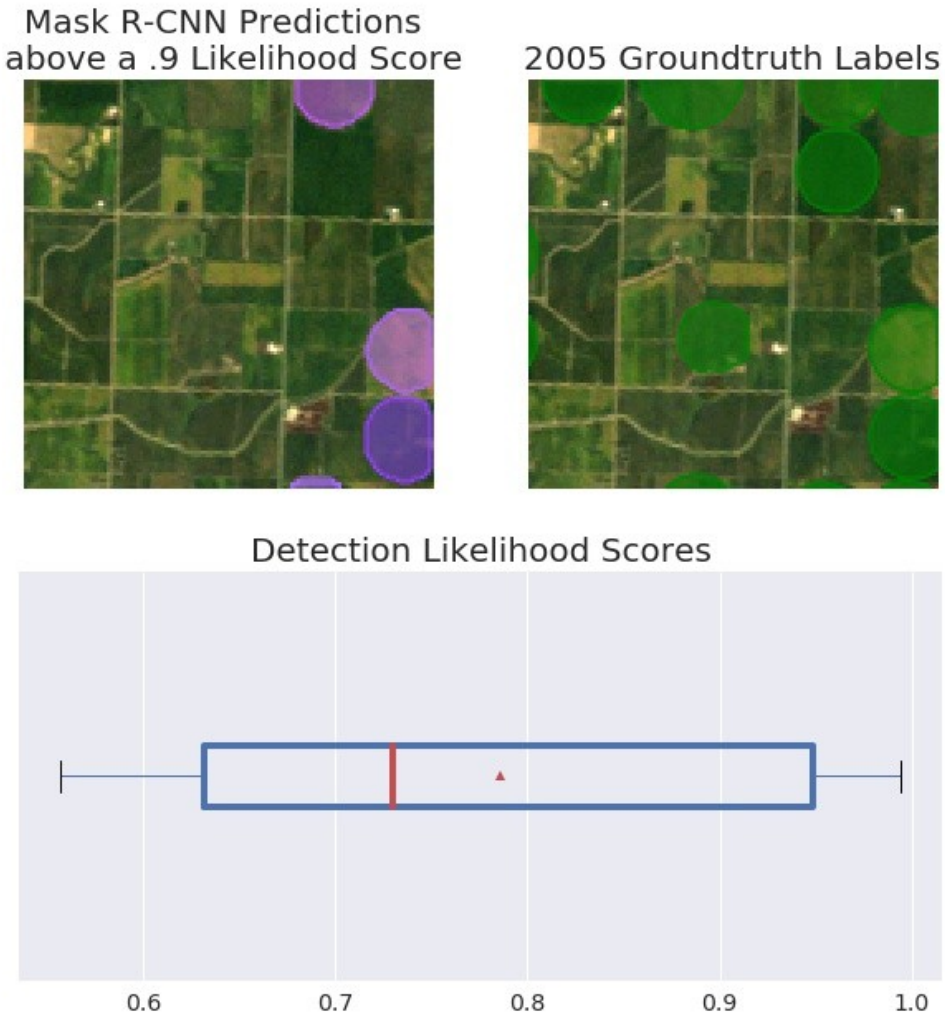

Fig. 13. This sample contains a variety of field shapes with heterogeneous reflectances. Many truncated fields are present in the reference dataset and are not detected with a confidence score over 90%. At least 50% of all confidence detections for this scene are mapped with confidence lower than ~73%. Two full center pivots are also missed, one in the

center of the scene that is visually distinct but small and irregularly shaped, and another in the top right that does not have a visually distinct boundary.

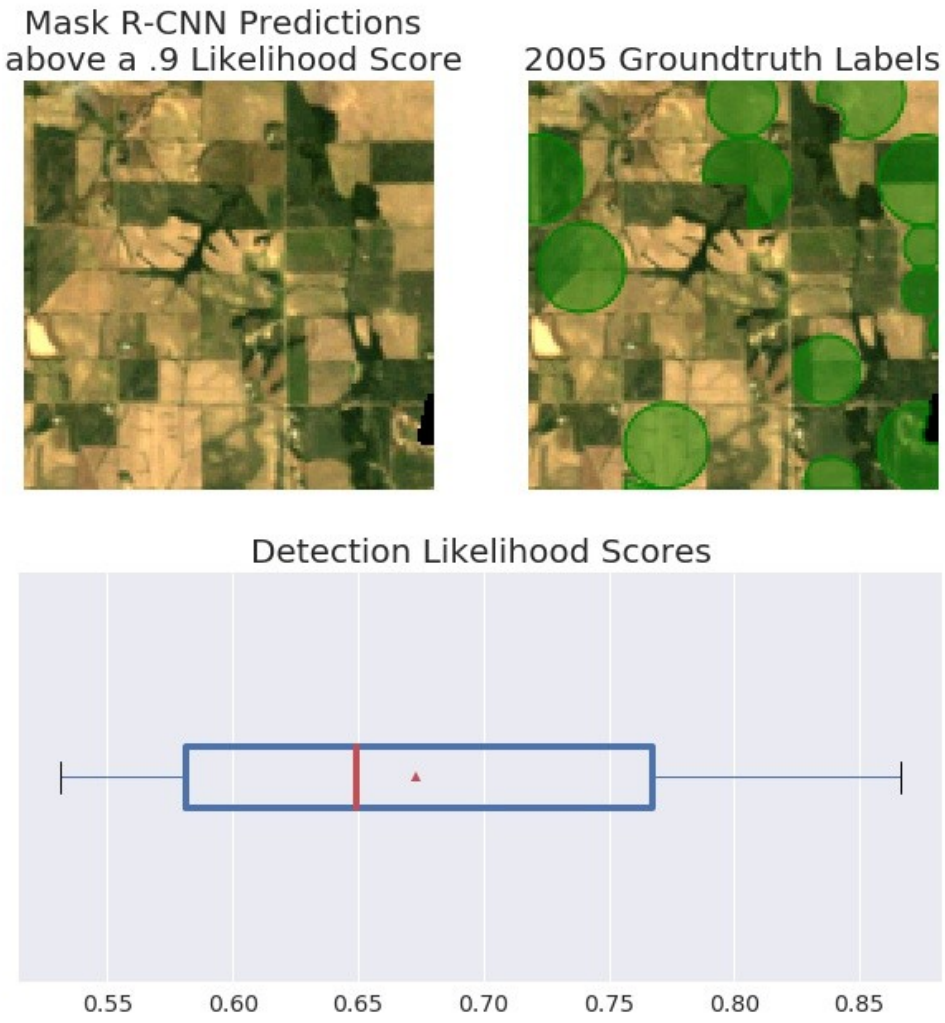

Fig. 14. In this image no correct high confidence detections were made. However, 25% of detections had a greater than 75% confidence score. Because the threshold chosen for plotting is 90% confidence, these detections are not displayed. Mixed pivots are present.

There is also a partial pivot in the top middle of the image. Finally, in the lower right is a reference label that is truncated because of a No Data region in the Landsat ARD image. Each of these complexities is less represented in the training dataset and thus detection confidences are lower for these more difficult instances.

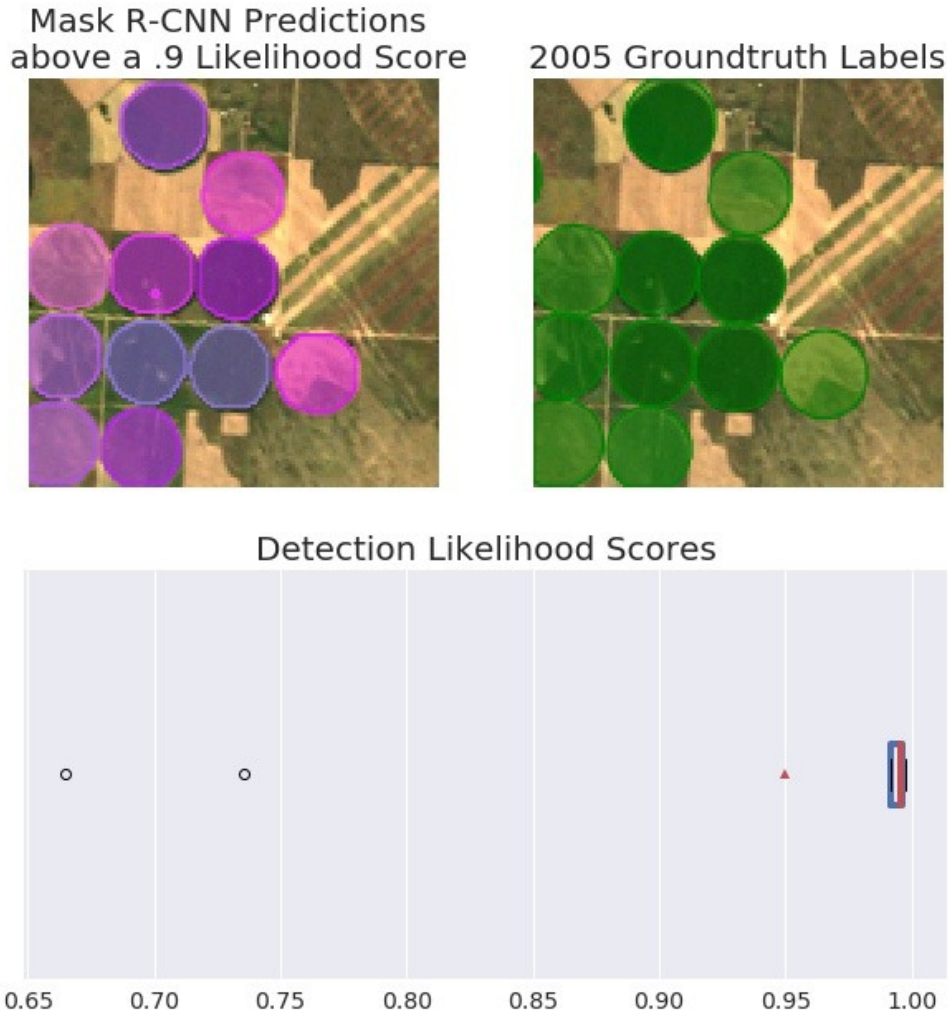

Fig. 15. This sample represents one of the easiest scenes to segment and classify. Medium, equal sized center pivots in the growing stage contrast strongly with the surrounding dry, grassland background. The scene is segmented perfectly relative to the reference data. A few extra detections have very low confidence scores below 0.75 (denoted by the circles on the boxplot) and are excluded by the 90% confidence threshold.

Figures 16 through 18 were selected to evaluate the impact that the size of center pivot fields have on detection accuracy. Selected samples are shown which specifically include fields large and small size categories. The figures are roughly ordered by increasing landscape complexity.

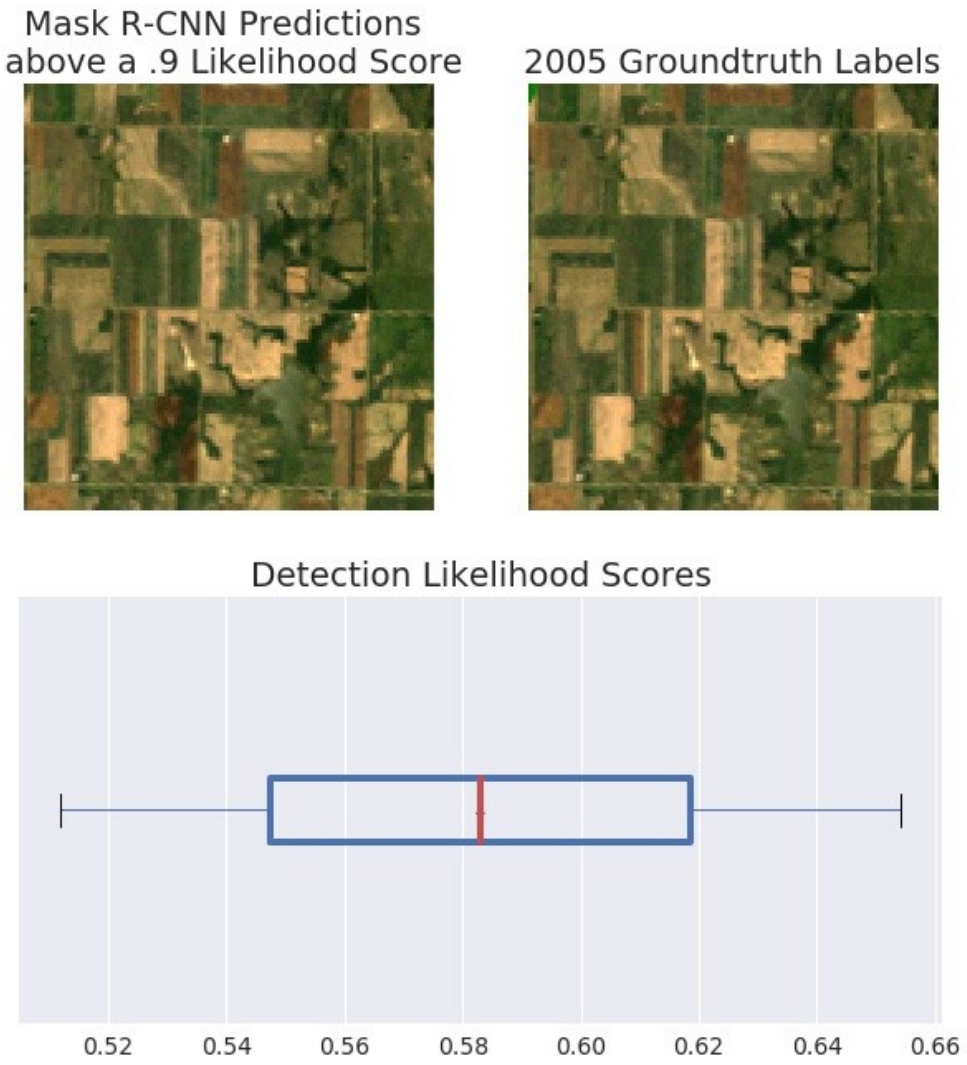

Fig. 16. This scene hasd no center pivots, therefore the lack of detections is encouraging. The only low large center pivot detection was well below the 90% score threshold, and so it was not deemed to be an accurate detection. All detections had confidence scores below 66%. The scene had no reference center pivots, so this is an accurately mapped scene.

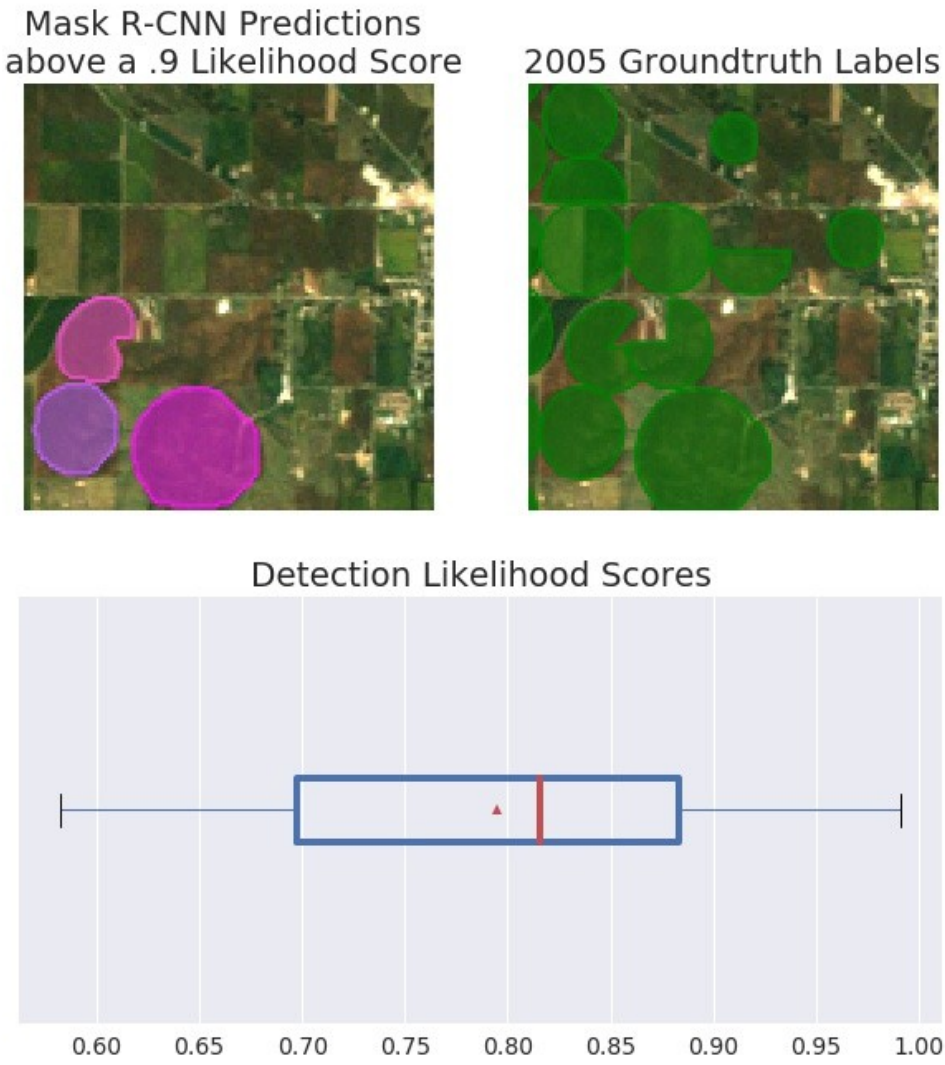

Fig. 17. This is a difficult scene to classify due to many non-pivot fields, mixed pivots, and partial pivots that are adjacent to roads (in the upper left). What's more, many of these fields are in close proximity. The large pivot is mapped correctly but many of the other smaller pivots are detected with low confidence or are undetected. At least 75% of detections are below the 90% confidence threshold.

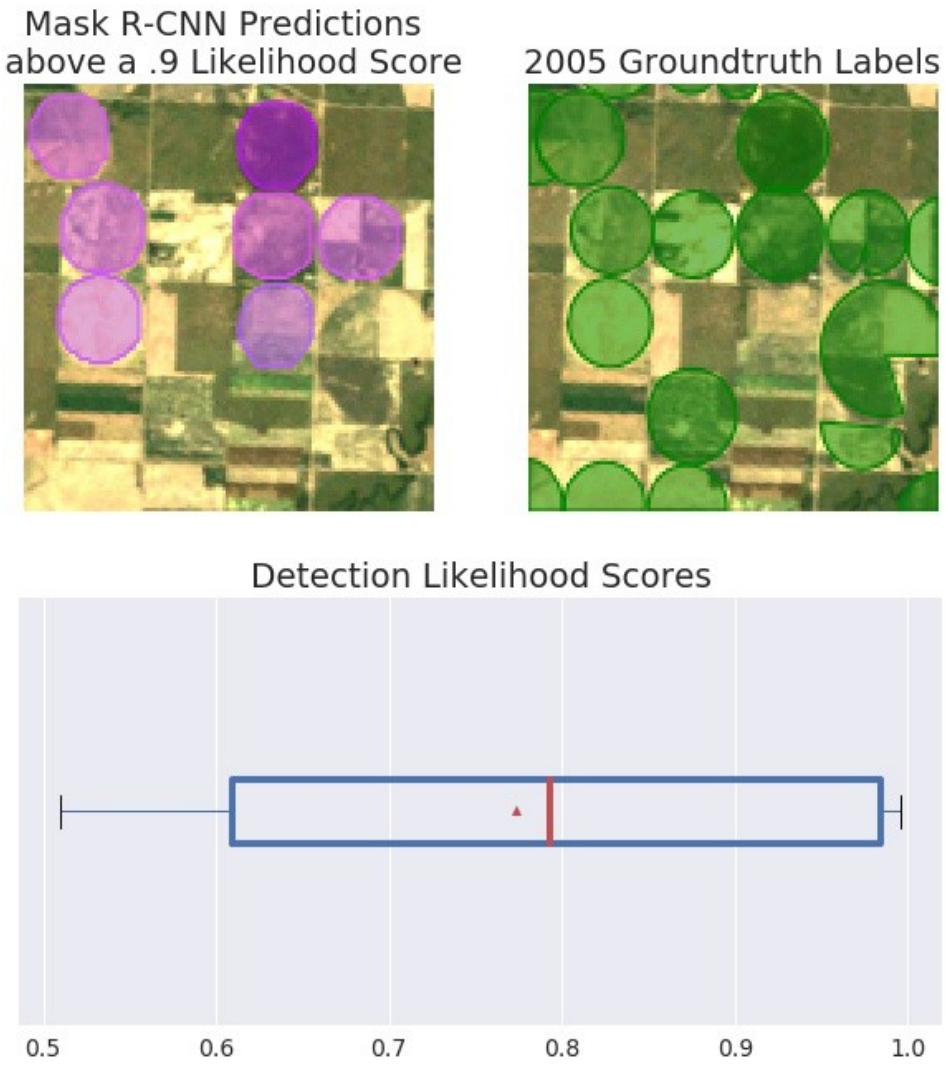

Fig. 18. The large pivot in this scene was not detected with high confidence, possibly because of its irregular shape as well as having semicircular portions in different stages of development. Some medium sized center pivots are also not detected with high confidence because of either indistinct boundaries, development stage being cleared instead of cultivated, or because of mismatch between the imagery and the reference label.

Figure 10 and 12 highlight that the FCIS model generates boundaries that are eccentric relative to the reference label boundaries and actual field boundaries in the image. The plotted detections in Figure 10 and Figure 12 are counted in the overall AP and AR metrics as successful detections because the IoU of the detection and reference label is greater than 50%. Yet, many pixels that belong to individual center pivot fields are not included in the detection. Across similar scenes, Mask R-CNN does not appear to have the consistent boundary eccentricity bias that FCIS has (Figure 15). Because Mask R-CNN showed better boundary accuracy along with comparably high performance metrics relative to the FCIS model, further comparisons and visual examples comparing accuracy across different field size ranges and with different training dataset sizes used Mask R-CNN instead of FCIS.

Many scenes are more complex than the arid landscape with fully cultivated center pivots shaped like full circles in Figure 10. Figure 11 is a representative example of a scene with more complex fields. These include non center pivot fields, center pivots containing two halves, quarters, or other fractional portions in different stages of development (mixed pivots), and partial pivots, which is semicircular not because the rest of the circle is in a different stage of development or cultivation, but because of another landscape feature that restricts the field's boundary (such as a river or road). In this scene, at least 25% of the detections in this scene are below the 90% confidence threshold, and many atypical pivots are missed based on this threshold (Figure 11).

Figure 12 is a simpler scene that like Figure 12, has a high density of center pivot fields. In this case the detections more closely match the reference labels and detection confidence scores are higher, either because of the FCIS model's tendency to produce higher

confidence scores or because the scene has less variation in center pivot field type. Figure 13 highlights another common issue when testing both models, that reference labels are truncated by the tiling grid used to make 128 by 128 sized samples from Landsat 5 scenes. These tend to be missed detections based on the 90% confidence threshold since they are not mapped with a high confidence score. There are cases where no high confidence detections above 90% are produced, such as in Figure 14. In this scene, No Data values in the Landsat 5 imagery, partial pivots near non-center pivot fields, and mixed pivots with indistinct boundaries all result in a scene that is not mapped with high confidence. However, in cases where there is high contrast between center pivot fields and their surrounding environment, they are mapped nearly perfectly by the Mask R-CNN model with high confidence scores (Figure 15).

Figures 16 through 18 were selected to illustrate the impact that size range has on detection accuracy, since center pivots can come in various semicircular shapes and sizes. Figure 16 shows that in a scene with no reference labels, no high confidence detections were produced for any size category. The highest confidence score associated with an erroneous detection in this case was at most ~0.66, which is relatively low for both models. Figure 17 shows a case where a large center pivot is mapped accurately, whereas smaller and medium center pivots in the scene are not. This example shows the scale invariance of the Mask R-CNN model in that it can accurately map the large center pivot because it looks similar to a medium sized center pivot, only larger. On the other hand, smaller center pivots, partial pivots, and mixed pivots are detected with lower confidences (half of the detections are less than ~82%) or not detected. Figure 18 highlights a case where a large center pivot is not mapped with a high confidence score above 90%. Unlike Figure 17 where the large center

pivot is uniform in appearance, the large center pivot in Figure 18 has a mixture of three different land cover types. This indicates that the inaccuracies from large center pivot detection may come from large center pivots that were partially cultivated or divided into multiple portions with different crop types or cultivation stages. Many false negatives in this scene and others are the result of partial pivots, mixed pivots or pivots that had not been annotated yet in the 2005 dataset.

4. Discussion

4.1. Metrics Discussion

Small fields are more difficult to detect than large ones and so as expected, removing 50% of the training data available to train the Mask R-CNN model caused large drops in performance. Having more training data available to improve features that are attuned to detect small fields is particularly important with regard to overall model performance. The metric results for small fields is likely biased toward a worse result because many full pivots overlapped a sample image boundary, leading to small, partial areas of pivot irrigation at scene edges being overrepresented after Landsat scenes were tiled into 128x128 image chips. Since these fields have a less distinctive shape, some full pivots at scene edges were missed.

However, since small fields make up a minority of the total population of fields and the medium and large category were more accurate by 20 or more percentage points for both AR and AP, this shows that both the FCIS and Mask R-CNN models can map a substantial majority of pivots with greater than 50% intersection over union. Zhang et al. (2020) tested their model on the same geographic locations at a different year and used samples produced from two Landsat scenes to train their model over a 21,000 km^2 area versus the Nebraska

dataset which spans 200,520 km^2. These results extend upon the work by Zhang et al. (2020), as the test set is geographically independent from both the training and validation set and 32 Landsat 5 scenes across a large geographic area were used to train and test the model. Furthermore, Zhang et al. (2020)'s approach produces bounding boxes for each field, while Mask R-CNN produces instance segmentations that can be used to count fields and identify pixels belonging to individual fields.

4.2. FCIS vs Mask R-CNN Comparison

While comparing metrics is useful, they don't indicate how performance varies across different landscapes or how well the quality of the boundary matches reference given that a detection is determined to be correct by having an IoU over 50%. While the FCIS model slightly outperformed the Mask R-CNN model in terms of the medium size category average precision, it also exhibited poorer boundary fidelity. Figure 10 demonstrates arbitrarily boxy boundaries that appear to be truncated by the position sensitive score map voting process that is the final stage of the FCIS model. The Mask R-CNN model's high confidence detections that had a confidence score of 0.9 or higher matched the reference boundaries much more closely, showing that this model can be usefully applied to delineate center pivot agriculture in a complex, humid landscape. While the FCIS model could also be employed with post processing to assign a perfect circle to the center of a FCIS detection (essentially discarding the segmentation produced), this would lead to further errors that would overestimate the size and wrongly estimate the shape of partial center pivots.

The results from Mask R-CNN on medium sized fields are encouraging because it indicates that the model could potentially generalize well to semi-arid and arid regions outside of Nebraska (Figure 15). In addition, the results from Figures 10, 12, and 15 indicate

that where many uniform center pivots are densely colocated and not many non center pivots are present, a higher quantity will be mapped correctly. This is encouraging, since in many parts of the world, center pivots are densely colocated, or are cultivated in semiarid or arid environments, where contrast is high. Therefore, the model can be expected to generalize well outside of Nebraska, though it remains future work to test the model in other regions. False negatives are present for many scenes that are heavily cultivated, and in many cases it is ambiguous whether the absence of a high confidence detection is due to the absence of a center pivot or because of Landsat's inability to resolve fuzzier boundaries between a field and its surrounding environment (Figure 11).

A time series based approach similar to Deines et al. (2019) could improve detections so that only pivots that exhibited a pattern of increased greenness would be detected in a given year. However, this requires multiple images within a growing season from a Landsat sensor, which are not always available due to clouds, and also precludes the use of traditional CNN methods and pretrained networks which ease the computational burden of training and detection. Furthermore, this approach is difficult to incorporate into a CNN based method for segmentation, as it precludes the use of pretraining. Another alternative is to develop higher quality annotated datasets which make meaningful semantic distinctions between agriculture in different stages of cultivation. For example, in Figure 11, brown center pivots that are not detected could instead be labeled as "fallow" or "uncultivated", and this information could be used to refine the training samples used to train a model to segment pivots within specific cultivation stages.

4.3. Comparison to COCO Detection Baselines

With 4 hours of training on 8 GPUs, the original implementation of Mask R-CNN achieved 37.1 AP using a ResNet-101-FPN backbone on the COCO dataset, a large improvement over the best FCIS model tested, which achieved 33.6 AP (He et al. 2017, Li et al. 2016). This amounts to a difference of 3.5 AP percentage points, with Mask R-CNN performing better on the COCO dataset. On the Nebraska dataset, for the medium size category, the difference in AP was 3.2 AP percentage points, with the FCIS model outperforming Mask R-CNN. However Mask R-CNN outperformed FCIS in terms of AR by 5.1%. These results indicate that COCO detection baselines are not necessarily reflective of overall metric performance, given that FCIS outperformed Mask R-CNN in the more numerous size category. The improvements on the COCO baseline do reflect the improved boundary accuracy of MAsk R-CNN relative to the FCIS model.

4.4. Comparison to Rieke 2017

The AR and AP results on the Nebraska center pivot dataset are higher relative to Reike (2017), which is to be expected since center pivots are a simpler detection target than fields in the Denmark dataset, which come in more various shapes and sizes. What's especially notable is that even though Rieke (2017) trained the FCIS model on approximately 11 times the training data compared to the training data used in this study, the AP results and AR results on the Nebraska dataset were about 10 to 20 points higher for each of the size categories. The difference for the AP for the small category was 0.42-0.28 = 0.14 , the difference for the medium category was 0.732 - 0.473 = 0.259, and the difference for the large category was 0.734 - 0.51 = 0.224 (Table 2 and Table 4). Rieke (2017) used 159042 samples compared to 13625 samples used in this study. These samples were equivalently sized to this study, at 128x128 pixels.

Even though the size categories used in this study and Rieke (2017) are not exactly the same, the fact that each of the categories saw substantially better performance for the FCIS model on the Nebraska dataset indicates that the relative simplicity of the center pivot detection target played a substantial role in the jump in performance. Given that Rieke (2017) used 11 more training samples, this indicates that the simplicity of the detection target played an even larger role in the performance difference. This is an important lesson for remote sensing researchers looking to use CNN techniques to map fields or other land cover objects, the feasibility of mapping the detection target can be even more important than using an order of magnitude more training data to improve a model's ability to generalize.

4.5. Comparison to other remote sensing detection results

These results are comparable to other results achieved for other detection targets. Wen et al. (2019) applied a slightly modified version of Mask R-CNN which can produce rotated bounding boxes to segment building footprints in Fujian Province, China from Google Earth imagery. The model was trained on manually labeled annotations across a range of scenes containing buildings with different shapes, materials, and arrangements. Though an independent test set was not used that was separate from the validation set (which is used to inform how to change hyperparameters), the model was tested on half of the imagery collected, while the other half was used to train the model, providing a large amount of samples to test the model. The total dataset used to split between training and testing/validation amounted to 2000 500x500 pixel images containing 84,366 buildings (Wen et al. 2019). The Mask R-CNN model produced an AP of 0.8996. Results were not stratified by size category. This result is 8 AP percentage points higher than the Mask R-CNN results on the Nebraska dataset for the medium sized category. This could be due to a number of

factors. First, the imagery used was 0.26 meter, which is much higher resolution than Landsat 5's 30 meter resolution. Therefore, each building instance could have a higher amount of pixels with which to compute informative features. A larger model with more parameters was also used, Resnet-101, which increases learning capacity as well as the tendency to overfit. Finally, some figures referenced in Wen et al. (2019) indicate that bounding boxes encompass multiple individual buildings. If these groups of buildings were used to represent a single instance to calculate mAP, the score would be higher than if each building was considered as a separate instance.

Non-CNN approaches have also been used to map irrigated and rainfed agriculture. Deines et al (2019) used a random forest model to classify irrigated pixels across the HPA using spectral indices, GRIDMET precipitation, SSURGO soil water content, a topographic DEM, and other features. The random forest model was trained on 40% of the data, validated on 30%, and tested on the remaining 30%. Since this method classified pixels and not objects, metrics are not directly comparable, however, the result shows that the model was quite successful in accurately mapping irrigated area across a wide climatological gradient. Pixelwise, overall accuracy for classified irrigated area was 91.4%, and results were visually assessed to correspond well with GCVI computed from Landsat scenes. Given that this model performed so well and that we are able to inspect feature importance for random forest models more easily than with CNN models, it is useful to determine if the features that were successful in the random forest model were or were not shared by the Mask R-CNN model so future models may take advantage of both approaches.

The top six most important features in the random forest model that led to this accurate classification were, in order of importance: latitude, slope, longitude, growing

degree days (days between 10 and 30 degrees Celsius), minimum NDWI, and the day-of-year at peak greenness. The authors note that the random forest model uses latitude and longitude to separate scenes by climatological gradients, which can improve detection as different climatological areas contain different agricultural patterns. Of these features, only minimum Normalized Difference Water Index can be directly learned by Mask R-CNN, though the model likely recognizes scene qualities such as bare soil that are correlated with a drier climatology. This indicates that pixel-wise features besides reflectances can contribute, and be even more important, than reflectance alone. Future approaches based on CNN's could make use of not just reflectance information, but also the features named above.

However, this would currently require training CNN models from scratch. This could be prohibitively expensive, increase training times, or lead to lower accuracies since weights are initialized randomly and training from scratch may not lead to as informative features as those that are arrived at after pretraining. The results from Deines et al. (2019) do not separate individual fields from each other; many clearly distinct fields are mapped as a single irrigated area unit. The method used here is useful for tracking pixel level changes in irrigation, yet it cannot be used for tracking field level statistics from year to year.

4.6. Other Mask R-CNN Results

The amount of training data was the second largest factor in determining the resulting performance metrics. Decreasing the amount of training data by 50% decreased performance metrics by a substantial amount. The large performance drop for the small size category, which had an 11.4% difference in terms of AP percentage points, indicates that more training data is especially important for more uncommon center pivot size categories, which highlights the importance of using as much training data as possible when training CNN-

based models. Using a NIR-R-G vs RGB composite did not affect the validation AP, which is expected since center pivots are visually defined by shape rather than spectra. Correct preprocessing choices were also important in order to achieve good results, and the most important of these was to convert image values to 8-bit integers before normalizing by the mean. This step clips the very large values present in Landsat 5 OLI scene chips from high reflectance from snow and clouds and adjusts the range of values of the training set to more closely match that of the pretrained model's original dataset, which in this case was Imagenet.

Models trained on the original TIFF imagery performed very poorly even when these images were normalized by the mean (results not shown). Adjusting other hyperparameters did not affect the model performance as much as converting the data type of each image sample to 8 bit integer and using the largest training data size. Examples of hyperparameters tested include doubling the number of region proposals generated during the training stage and increasing the amount of non-max suppression to prune low-confidence region proposals. I expected that more region proposals and more aggressive pruning of low confidence proposals would lead to a better performing model, however making these adjustments did not impact model performance, which is likely because the default number of regions generated was sufficient to intersect with most potential instances of center pivots in each sample.

4.7. Limitations

These results clearly show that segmenting small center pivots will be a challenge for Landsat RGB imagery, given it's coarse resolution and the less resolving power for smaller boundaries. However, Figure 13 shows that in a scene where there are many fields in full

cultivation, they are for the most part all accurately mapped. This implies that inaccuracies from misdetected small center pivots in a stage of pre-cultivation could be remedied by applying the model to multiple dates and merging the output results in order to capture center pivots in a stage of cultivation during the growing season.

The results from the medium category are overall much more accurate, though even in this size category, Figure 11 shows that there are many false negatives due to many pivots being in a stage of pre-cultivation or showing half in cultivation and half pre-cultivation. Figure 11 also highlights another source of error in segmenting center pivots, where corner areas surrounding center pivots are also cultivated. While these regions are not annotated in the Nebraska dataset as belonging to part of the center pivot, they could either be irrigated along with the center pivot as a single field using an extension to the irrigation apparatus or separately. This has relevance for accurately estimating individual field water use, and highlights the difficulty in resolving individual units using satellite imagery.

Center pivots that are composed of multiple sections will be hard to segment, given the limited amount of training data that contains similar representations. As with small center pivots in stages of pre cultivation, one approach to segmenting these could be to apply a model to multiple dates in a growing season and then merge the highest confidence detections in order to reduce the amount of false negatives.

A major limitation to this study is that the reference dataset was not entirely accurate. Because the dataset contained 52,127 instances of center pivots, it was infeasible to examine each with respect to 32 Landsat scenes as an image reference. In some cases existing center pivots went unlabeled (Figure 11). This impacted both the model training, leading to lower performance due to greater within class variability and similarity between the center pivots

category and background category, and less certain evaluation metrics, since the results were evaluated on an independent test subset of the reference dataset. Furthermore, center pivots as a detection target represent some of the most visually distinct agricultural features, and therefore it is expected that these models would perform more poorly on small agricultural fields that do not conform to such a consistent shape and range of sizes. Finally, the reference dataset for Nebraska used a broad interpretation of center pivots to include center pivots in multiple stages of cultivation (fallow, cleared, cultivated, senescing). This led to a considerable amount of within class variability during training which impacted model performance. This limitation could be handled better by assigning more specific semantic categories to the center pivot labels based on greenness indices, in order to distinguish between different developmental stages. Or, the model produced from the original dataset could be applied for multiple dates throughout the seasons and results could be merged based on detection confidence in order to map pivots when they are at their most discernible, i.e. cultivated.

5. Conclusions

The following experiments were conducted: 1) a comparison of the Mask R-CNN and FCIS models to evaluate how a newer CNN model compares to a previously state-of-the-art model, 2) a comparison between performance differences between these model on the COCO dataset vs the Nebraska dataset to evaluate how well benchmarks in other image domains translate to the remote sensing domain, 3) a comparison between FCIS model results on the Nebraska dataset and the Denmark dataset used in Rieke (2017) to evaluate the importance of more training data relative to a simpler detection target, and 4) the ability of

the best model to, MAsk R-CNN, to generate high confidence predictions using 50% less training data.

These experiments demonstrated 1) that without image augmentation or post processing, the Mask R-CNN model is comparable to the FCIS model in terms of overall AR and AP on the Nebraska dataset, 2) that benchmarks in other imaging domains are not necessarily indicative of performance in the remote sensing domain in terms of metrics, 3) that having a simpler detection target (center pivots vs fields of more various shape and size) was more impactful for overall AR and AP performance than having an order of magnitude more training image samples, and 4) that reducing training data by 50% had a substantial impact on model AR and AP. The results indicate that Mask R-CNN can be an important new tool for mapping distinct field objects, but that care must be taken to apply it during periods where landcover or land use patterns are most apparent relative to their surroundings. Furthermore, small center pivots and semi circular pivots are more difficult to detect and classify due to their less distinct boundaries and less represented, irregular shape, while the Mask R-CNN and FCIS models perform well for both the predominant medium and large size category. The limitations highlighted by visual inspection show that the model should be applied during periods when center pivots are being actively cultivated, as detection confidence was low for fields that were not actively cultivated since they had low contrast with their surroundings.

Despite these limitations, Mask R-CNN model can produce field boundaries with human level detail for a subset of center pivot fields of varying size, providing that they have sufficient contrast with their surroundings and have uniform crop cover. This model has been demonstrated to be useful for mapping the uniform center pivots with high confidence across

semi arid and humid climates, while avoiding generating high confidence false positives. Given these results, I expect that this model can generalize well to regions outside of Nebraska where center pivot agriculture is expanding, particularly in semi-arid and arid regions.

6. Future Work

Data providers should ensure that geospatial labels describing boundaries are labeled as distinct instances to make it possible to derive field level statistics. Labelling pixels as opposed to distinct objects effectively discards useful information present in satellite imagery, and data providers should consider tools and workflows that enable them to label instances. Another benefit of object oriented approaches that output instances (vector polygons representing fields) is that they can be readily incorporated into different machine learning models, and provide more detailed, field level information, compared with pixel-wise outputs. Whereas pixel-wise inputs necessarily can only be used to characterize a snapshot in time or a single period of time, instance outputs can be used to train future machine learning models, or can be incorporated into rule based classifiers to determine higher fidelity field characteristics such as growth state, crop type, or rate of water use. Vector datasets derived from instances can be more readily complemented with such attributes that are related to objects, and can be more easily extended when coupled with other open source geospatial data. By training models that support instance segmentation when mapping agriculture, the results can be more easily and accurately converted to geospatial vectors and then incorporated into services that make geospatial data more discoverable. OpenStreetMap is a primary example of an open geo data portal that can be used to publish geodata that is searchable, joinable, and complemented by annotation tools to

refine vectors or add additional information about objects (such as ownership). Vectors can also be easily encoded in non-spatial formats, such as tabular formats, which increases their usability in domains outside of the geographic sciences.

Pretraining from large image datasets like Imagenet is both an asset and a limitation when training models that are applied to remotely sensed images. It is an asset because of the reduced computational costs, speed, and potential accuracy gains in training models, but is a limitation as it requires the model input to have the same dimensions as the original dataset that produced the model. In a step in the right direction, pretrained models derived from one of the largest remote sensing image datasets, BigEarthNet v2, have recently been made available in both Pytorch and Tensorflow compatible model formats (Sumbul et al. 2019). TA comparison should be made between models derived from pretrained models from remote sensing image datasets and computer vision datasets, to assess if any gains in performance can be made.

The reproducible code for this work is available at https://github.com/ecohydro/CropMask_RCNN and in the future, the models used to produce these results will be made available as a Python and Flask API service for generating center pivot detections from Landsat scenes. This application will convert model results to geospatial shapefiles to allow for analysis in GIS software.

References

1. Argialas, D. P., and O. D. Mavrantza. 2004. “Comparison of Edge Detection and Hough Transform Techniques for the Extraction of Geologic Features.” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 34 (Part XXX). http://www.cartesia.org/geodoc/isprs2004/comm3/papers/376.pdf.
2. Blaschke, Thomas J., and Josef Strobl. 2001. “What’s Wrong with Pixels? Some Recent Developments Interfacing Remote Sensing and GIS.” https://www.semanticscholar.org/paper/51b65077ce150e717f21ee8ce1d96e4790d5779c.
3. Brown, Jesslyn F., and Md Shahriar Pervez. 2014. “Merging Remote Sensing Data and National Agricultural Statistics to Model Change in Irrigated Agriculture.” *Agricultural Systems* 127 (May): 28–40.
4. Colaizzi, P. D., P. H. Gowda, T. H. Marek, and D. O. Porter. 2009. “Irrigation in the Texas High Plains: A Brief History and Potential Reductions in Demand.” *Irrigation and Drainage* 58 (3): 257–74.
5. Crary, Douglas D. 1951. “Recent Agricultural Developments in Saudi Arabia.” *Geographical Review* 41 (3): 366–83.
6. Cross, A. M. 1988. “Detection of Circular Geological Features Using the Hough Transform.” *International Journal of Remote Sensing* 9 (9): 1519–28.
7. “‘CS231n Convolutional Neural Networks for Visual Recognition.” n.d. Accessed June 6, 2020. https://cs231n.github.io/convolutional-networks/.
8. Deines, Jillian M., Anthony D. Kendall, Morgan A. Crowley, Jeremy Rapp, Jeffrey A. Cardille, and David W. Hyndman. 2019. “Mapping Three Decades of Annual Irrigation across the US High Plains Aquifer Using Landsat and Google Earth Engine.” *Remote Sensing of Environment* 233 (November): 111400.
9. Deines, Jillian M., Meagan E. Schipanski, Bill Golden, Samuel C. Zipper, Soheil Nozari, Caitlin Rottler, Bridget Guerrero, and Vaishali Sharda. 2020. “Transitions from Irrigated to Dryland Agriculture in the Ogallala Aquifer: Land Use Suitability and Regional Economic Impacts.” *Agricultural Water Management* 233 (April): 106061.
10. Deng, J., W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55.

11. DeNicola, Erica, Omar S. Aburizaiza, Azhar Siddique, Haider Khwaja, and David O. Carpenter. 2015. “Climate Change and Water Scarcity: The Case of Saudi Arabia.” *Annals of Global Health* 81 (3): 342–53.
12. Dennehy, K. F., D. W. Litke, and P. B. McMahon. 2002. “The High Plains Aquifer, USA: Groundwater Development and Sustainability.” *Geological Society, London, Special Publications* 193 (1): 99–119.
13. Döll, Petra. 2009. “Vulnerability to the Impact of Climate Change on Renewable Groundwater Resources: A Global-Scale Assessment.” *Environmental Research Letters: ERL [Web Site]* 4 (3): 035006.
14. Elmes, Arthur, Hamed Alemohammad, Ryan Avery, Kelly Caylor, J. Ronald Eastman, Lewis Fishgold, Mark A. Friedl, et al. 2020. “Accounting for Training Data Error in Machine Learning Applied to Earth Observations.” *Remote Sensing* 12 (6): 1034.
15. “FAO Country Profiles:Saudi Arabia.” n.d. Accessed April 20, 2020. http://www.fao.org/countryprofiles/index/en/?lang=ar&iso3=SAU&paia=4.
16. Foster, T., I. Z. Gonçalves, I. Campos, C. M. U. Neale, and N. Brozović. 2019. “Assessing Landscape Scale Heterogeneity in Irrigation Water Use with Remote Sensing and in Situ Monitoring.” *Environmental Research Letters: ERL [Web Site]* 14 (2): 024004.
17. Frenken, Karen, and Others. 2009. “Irrigation in the Middle East Region in Figures AQUASTAT Survey-2008.” *Water Reports*, no. 34. https://www.cabdirect.org/cabdirect/abstract/20093247971.
18. Girshick, Ross. 2015. “Fast R-CNN.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1504.08083.
19. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1311.2524.
20. Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. “Deep Sparse Rectifier Neural Networks.” Edited by Geoffrey Gordon, David Dunson, and Miroslav Dudík, Proceedings of Machine Learning Research, 15: 315–23.
21. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
22. Haacker, Erin M. K., Kayla A. Cotterman, Samuel J. Smidt, Anthony D. Kendall, and David W. Hyndman. 2019. “Effects of Management Areas, Drought, and Commodity Prices on Groundwater Decline Patterns across the High Plains Aquifer.” *Agricultural Water Management* 218 (June): 259–73.
23. Haacker, Erin M. K., Anthony D. Kendall, and David W. Hyndman. 2016. “Water Level Declines in the High Plains Aquifer: Predevelopment to Resource Senescence: Ground Water Xx, No. X: Xx-Xx.” *Groundwater* 54 (2): 231–42.
24. Halverson, N. 2015. “What California Can Learn from Saudi Arabia’s Water Mystery.” *Reveal News. April* 22.
25. He, Kaiming, Ross Girshick, and Piotr Dollár. 2018. “Rethinking ImageNet Pre-Training.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1811.08883.

26. He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. “Mask R-CNN.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1703.06870.
27. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. “Deep Residual Learning for Image Recognition.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1512.03385.
28. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105. Curran Associates, Inc.
29. Lecun, Yann, and Yoshua Bengio. 1995. “Convolutional Networks for Images, Speech, and Time-Series.” In *The Handbook of Brain Theory and Neural Networks*. MIT Press.
30. Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. “Microsoft COCO: Common Objects in Context.” In *Computer Vision – ECCV 2014*, 740–55. Springer International Publishing.
31. Li, Yi, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. 2016. “Fully Convolutional Instance-Aware Semantic Segmentation.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1611.07709.
32. Ma, Eric. 2020. “Deep Learning Workshop.” https://doi.org/10.6084/m9.figshare.12424571.v1.
33. Mathieu, Renaud, Claire Freeman, and Jagannath Aryal. 2007. “Mapping Private Gardens in Urban Areas Using Object-Oriented Techniques and Very High-Resolution Satellite Imagery.” *Landscape and Urban Planning* 81 (3): 179–92.
34. Myint, Soe W., Patricia Gober, Anthony Brazel, Susanne Grossman-Clarke, and Qihao Weng. 2011. “Per-Pixel vs. Object-Based Classification of Urban Land Cover Extraction Using High Spatial Resolution Imagery.” *Remote Sensing of Environment* 115 (5): 1145–61.
35. “NASA Sees Fields of Green Spring up in Saudi Arabia.” n.d. Flickr. Accessed April 28, 2020. https://www.flickr.com/photos/gsfc/7027523783/in/set-72157623424324229.
36. “[No Title].” n.d. Accessed May 21, 2020. https://www.arcgis.com/home/item.html?id=8e7b99d90da84c82889e00ba8f90ef41.
37. Pearce, Fred. 2012. “Saudi Arabia Stakes a Claim on the Nile.” *National Geographic*, December 19, 2012. https://www.nationalgeographic.com/news/2012/12/121217-saudi-arabia-water-grabs-ethiopia/.
38. Pfeiffer, Lisa, and C-Y Cynthia Lin. 2014. “Does Efficient Irrigation Technology Lead to Reduced Groundwater Extraction? Empirical Evidence.” *Journal of Environmental Economics and Management* 67 (2): 189–208.
39. Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2017. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (6): 1137–49.
40. Rieke, Christoph. 2017. “Deep Learning for Instance Segmentation of Agricultural Fields,” September. http://dx.doi.org/.

41. Rydberg, A., and G. Borgefors. 2001. “Integrated Method for Boundary Delineation of Agricultural Fields in Multispectral Satellite Images.” *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 39 (11): 2514–20.
42. Saraiva, Marciano, Églen Protas, Moisés Salgado, and Carlos Souza. 2020. “Automatic Mapping of Center Pivot Irrigation Systems from Satellite Images Using Deep Learning.” *Remote Sensing* 12 (3): 558.
43. Shiklomanov, Igor A. 2000. “Appraisal and Assessment of World Water Resources.” *Water International* 25 (1): 11–32.
44. “SpatioTemporal Asset Catalog.” n.d. Accessed June 12, 2020. https://stacspec.org/.
45. Strahler, Alan H., Curtis E. Woodcock, and James A. Smith. 1986. “On the Nature of Models in Remote Sensing.” *Remote Sensing of Environment* 20 (2): 121–39.
46. Sumbul, G., M. Charfuelan, B. Demir, and V. Markl. 2019. “Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding.” In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 5901–4.
47. Thenkabail, Prasad S., Mitchell Schull, and Hugh Turral. 2005. “Ganges and Indus River Basin Land Use/land Cover (LULC) and Irrigated Area Mapping Using Continuous Streams of MODIS Data.” *Remote Sensing of Environment* 95 (3): 317–41.
48. Tryolabs. 2018. “Faster R-CNN: Down the Rabbit Hole of Modern Object Detection | Tryolabs Blog.” Tryolabs. January 18, 2018. https://tryolabs.com/blog/2018/01/18/faster-r-cnn-down-the-rabbit-hole-of-modern-object-detection/.
49. Waller, Peter, and Muluneh Yitayew. 2016. “Center Pivot Irrigation Systems.” In *Irrigation and Drainage Engineering*, edited by Peter Waller and Muluneh Yitayew, 209–28. Cham: Springer International Publishing.
50. Luckey, Richard, Gutentag, Edwin, Weeks, John 1981. “Water-Level and Saturated-Thickness Changes, Predevelopment to 1980, in the High Plains Aquifer in Parts of Colorado, Kansas, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wyoming.” https://doi.org/10.3133/ha652.
51. Wen, Qi, Kaiyu Jiang, Wei Wang, Qingjie Liu, Qing Guo, Lingling Li, and Ping Wang. 2019. “Automatic Building Extraction from Google Earth Images under Complex Backgrounds Based on Deep Instance Segmentation Network.” *Sensors* 19 (2). https://doi.org/10.3390/s19020333.
52. Wikipedia contributors. 2020. “Microsoft Azure.” Wikipedia, The Free Encyclopedia. June 12, 2020. https://en.wikipedia.org/w/index.php?title=Microsoft_Azure&oldid=962146509.
53. Wulder, Michael A., Thomas Hilker, Joanne C. White, Nicholas C. Coops, Jeffrey G. Masek, Dirk Pflugmacher, and Yves Crevier. 2015. “Virtual Constellations for Global Terrestrial Monitoring.” *Remote Sensing of Environment* 170 (December): 62–76.
54. Xu, Tianfang, Jillian M. Deines, Anthony D. Kendall, Bruno Basso, and David W. Hyndman. 2019. “Addressing Challenges for Mapping Irrigated Fields in Subhumid Temperate Regions by Integrating Remote Sensing and Hydroclimatic Data.” *Remote Sensing* 11 (3): 370.

55. Zhang, Chenxiao, Peng Yue, Liping Di, and Zhaoyan Wu. 2018. “Automatic Identification of Center Pivot Irrigation Systems from Landsat Images Using Convolutional Neural Networks.” *Collection FAO: Agriculture* 8 (10): 147.
56. Zhang, Hankui K., and David P. Roy. 2017. “Using the 500m MODIS Land Cover Product to Derive a Consistent Continental Scale 30m Landsat Land Cover Classification.” *Remote Sensing of Environment* 197 (August): 15–34.