International Journal of Remote Sensing, Volume 42, Issue 8, 2021, pp. 3017-3038
DOI: 10.1080/01431161.2020.1864059

RESEARCH ARTICLE

Replacing Human Interpretation of Agricultural Land in Afghanistan with a Deep Convolutional Neural Network

A. M. Hamer[a] and D. M. Simms[a] and T. W. Waine[a]

[a]Remote Sensing Group, School of Water, Energy and Environment, Cranfield University, Cranfield, Bedfordshire, MK43 0AL, UK

ARTICLE HISTORY
Compiled December 11, 2020

ABSTRACT
Afghanistans annual opium survey relies upon time-consuming human interpretation of satellite images to map the area of potential poppy cultivation for statistical sample design. Deep Convolutional Neural Networks (CNNs) have shown groundbreaking performance for image classification tasks by encoding local contextual information, in some cases outperforming trained analysts. In this study, we investigate the development of a CNN to automate the classification of agriculture from medium resolution satellite imagery as an alternative to manual interpretation. The residual network (ResNet50) CNN architecture was trained and validated for delineating the agricultural area using labelled multi-seasonal Disaster Monitoring Constellation (DMC) satellite imagery (32 m) of Helmand and Kandahar provinces. The effect of input image chip size, training sampling strategy, elevation data and multi-seasonal imagery were investigated. The best performing single year classification used an input chip size of 33×33 pixels, a targeted sampling strategy and transfer learning, resulting in high overall accuracy (94%). The inclusion of elevation data marginally lowered performance (93%). Multi-seasonal classification achieved an overall accuracy of 89% using the previous two years' data. Only 25% of the target year's training samples were necessary to update the model to achieve $> 94\%$ overall accuracy. A data-driven approach to automate agricultural mask production using CNNs is proposed to reduce the burden of human interpretation. The ability to continually update CNN models with new data has the potential to significantly improve automatic classification of vegetation across years.

1. Introduction

The United Nations Office on Drugs and Crime (UNODC) and Government of Afghanistan conduct an annual survey to estimate the production of opium in Afghanistan, a country responsible for 82% of global production (UNODC 2019). The opium trade fuels poverty, political instability and insurgency; hampering development efforts. The survey plays an important role in monitoring the extent and evolution of illicit opium production for the development of counter-narcotic policy. Within the survey, the accurate mapping of agricultural land, known as the agricultural mask, is essential for robust statistical sample design and production estimates. The mask is reviewed each year because of the large variation in the annual distribution of agricultural land (UNODC 2018). The current method uses unsupervised classification of medium resolution satellite imagery, such as Land-Remote Sensing Satellite System

CONTACT D. M. Simms. Email: d.m.simms@cranfield.ac.uk

(Landsat) 8 (30 m). This approach has difficulty separating natural vegetation from agriculture, so requires post-classification manual refinement. This is time-consuming and uses trained interpreters with knowledge of local agronomic practice in order to accurately map agricultural land.

Machine learning techniques have been shown to increase the accuracy of image classification (Belgiu and Drgu 2016; Pouliot et al. 2019; Lecun, Bengio, and Hinton 2015; Yamashita et al. 2018). Of particular importance are Convolutional Neural Networks (CNNs), which can outperform other machine learning classifiers such as Support Vector Machines (about 19%) at image labelling (Russakovsky et al. 2015) and Random Forests (7%) on mapping from medium resolution imagery (Pouliot et al. 2019), and can match human performance in certain image related tasks (Haenssle et al. 2018). They are inspired by the connections between neurons in the cerebral cortex (Ball, Anderson, and Chan 2017) and are made up of convolutional layers that encode the spatial and spectral elements of image features from large training datasets. The rapid improvements in accuracy have been achieved through the development of new CNN architectures for image classification (Zeiler and Fergus 2013; Simonyan and Zisserman 2015). CNNs are able to capture complex contextual information in a similar way to manual image-interpretation, where associated features and the spatial context of observations are used as interpretation keys (e.g. field patterns, irrigation canals and topography). This overcomes one of the limitations of pixel-based classification for mapping agricultural land in Afghanistan, where spectral separation alone is not able to discriminate between natural vegetation and agriculture.

The aim of this study is to determine whether CNNs can perform the role of a human interpreter in delineating agricultural land from medium-resolution imagery. Access to densely labelled agricultural masks from opium surveys in Afghanistan provide the necessary data to develop an optimal CNN training strategy for agricultural delineation and evaluate its performance across multiple years.

2. Convolutional Neural Networks

CNNs are widely used in image classification because of their high performance and ability to accept multi-dimensional pixel arrays (Lecun, Bengio, and Hinton 2015). These networks are designed to adaptively learn spatial features from a set of labelled examples through a backpropagation algorithm (Yamashita et al. 2018). Each convolutional layer in the neural network runs a fixed-sized filter matrix across the image at a defined spacing, or stride, to generate a feature map, which forms the input to the next layer. A rectified linear unit (ReLU) activation function is used to introduce non-linearity and avoid saturation during learning (Nogueira, Penatti, and dos Santos 2017). Pooling layers are used to reduce the dimensionality of feature maps, using a maximum or average filter matrix, by downsampling the spatial resolution of the input layers. Fully connected layers join each node from the previous layer, flattening them out into one-dimensional feature maps. The final layer is a fully connected layer that calculates class probabilities for each instance using a classification activation function, usually a Softmax (Goodfellow, Bengio, and Courville 2016).

The network is trained using a gradient-based optimisation algorithm, most commonly Adam (Kingma and Ba 2015), Stochastic Gradient Descent (SGD), Adaptive Gradient Algorithm (AdaGrad) (Duchi, Hazan, and Singer 2011) or Root Mean Square Propagation (RMSprop) (Tieleman and Hinton 2012), and a loss function that measures the agreement between the model predictions and the reference data labels.

Normalisation is used as a data pre-processing step, usually z-score normalisation (Chollet 2017), to scale input data to a common range. The optimisation algorithm minimises the loss by altering the layer weights for each batch of reference data fed into the CNN. Training stops once there is no longer any significant decrease in the loss, usually after many epochs, where an epoch represents one complete pass of the reference data through the network during training.

There are two approaches for CNN training, known as end-to-end and transfer learning. End-to-end learning uses the input data alone to identify the target object's features from randomly initialised filter weights. Transfer learning uses pre-trained filter weights from a previous application. ImageNet, a dataset of commercial photographs used for visual object recognition, is commonly used for image transfer learning, particularly where training data are limited (Shin et al. 2016). Transfer learning has been beneficial for classification as similar features often transcend individual image recognition tasks and reduce the requirement for large labelled datasets (Yosinski et al. 2014).

Common CNN architectures include Visual Geometry Group (VGG) 16 and VGG 19 that use small convolutional filters (3×3) across their 16 and 19 layer networks to achieve state-of-the-art classification accuracy on ImageNet (Simonyan and Zisserman 2015). The residual network (ResNet50) architecture, a 50 layer network, found deeper networks were beneficial to classification accuracy and has outperformed VGG CNNs (He et al. 2016).

Existing applications of CNNs for remote sensing data often use imagery benchmark datasets, including University of California (UC) Merced land use dataset (Yang and Newsam 2010), Aerial Image Dataset (AID) (Xia et al. 2017) and Brazilian coffee scenes (Penatti, Nogueira, and Santos 2015; Nogueira, Penatti, and dos Santos 2017; Deng et al. 2018; Zhang, Tang, and Zhao 2019). These use a similar approach to photography-based object recognition and classify whole images, or image subsets known as chips, with a single label. Across these benchmark datasets there are differences in training sample sizes, input image chip sizes and CNN model architectures. Image chip sizes vary greatly, with the UC Merced dataset (0.3 m) (Yang and Newsam 2010) at 256×256 pixels, AID dataset (0.5 to 0.8 m) (Xia et al. 2017) at 300×300 pixels and the Brazilian coffee scenes dataset (10 m) (Penatti, Nogueira, and Santos 2015) at 64×64 pixels. Sample sizes are constrained by the amount of labelled data available with samples ranging from 10s (Fu et al. 2017) to 100s (Deng et al. 2018) to 1000s per class (Cheng, Han, and Lu 2017). CNN image categorisation is used outside of benchmark datasets (Liu et al. 2018; Feng et al. 2019; Koga, Miyazaki, and Shibasaki 2018), but is again constrained by the requirement for large datasets. These applications use the normalised spectral bands exclusively for prediction, while for other machine learning algorithms, ancillary data, such as distance to water, elevation and economic indicators (Lucas et al. 2011; Gislason, Benediktsson, and Sveinsson 2006) are used to improve classification performance.

Land cover classifications are validated using hold-out samples to calculate accuracy metrics. Overall accuracy (OA) is the number of correctly classified pixels in comparison to the reference data and widely adopted within remote sensing (Foody 2002),

$$\mathrm{OA} = \frac{\sum_i n_{i,i}}{\sum_i t_i} \tag{1}$$

where $n_{i,i}$ is number of pixels predicted as class i belonging to class i and t_i is total number of pixels belonging to class i in the reference data. The Kappa coefficient (K) is used to quantify the statistical significance in comparison to random performance (Cohen 1960),

$$K = \frac{p_{\mathrm{o}} - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}} \tag{2}$$

where p_{o} is the empirical probability of correctly labelled samples and p_{e} is the expected probability of correctly labelled samples by random chance.

The Intersection over Union (IoU), also known as Jaccard index, gives the similarity between the predicted region and reference region by identifying overlapping regions (Long, Shelhamer, and Darrell 2015). Two variations of IoU commonly used for deep learning applications are the mean IoU (mIoU) (Russakovsky et al. 2015),

$$\mathrm{mIoU} = \frac{1}{k} \sum_i \frac{n_{i,i}}{t_i + \sum_j n_{j,i} - n_{i,i}} \tag{3}$$

and frequency-weighted IoU (fwIoU),

$$\mathrm{fwIoU} = \frac{1}{\sum_i t_i} \sum_i \frac{t_i n_{i,i}}{t_i + \sum_j n_{j,i} - n_{i,i}} \tag{4}$$

where $n_{j,i}$ is the number of pixels predicted as class j belonging to class i, and k is the number of classes in the reference data.

3. Materials and Methods

3.1. *Study site*

The study area is the provinces of Helmand and Kandahar in the south of Afghanistan, covering an area of 81,383 km^2 (Figure 1). These are the largest opium producing provinces in Afghanistan with an estimated 160,208 ha grown in 2018, accounting for 61% of national opium cultivation (UNODC 2018). They also contain the highest proportions of irrigated agricultural land in Afghanistan, 342,172 ha and 312,465 ha respectively (FAO 2016). The main area of cultivation is in the Helmand valley with large areas of natural vegetation in northern Kandahar (Figure 1). Helmand and Kandahar contain a wide range of agricultural landscapes, including rain-fed agriculture in lowland and highland areas, fruit trees, vineyards, marginal agriculture and natural needle leaved forests (FAO 2016). Agriculture in Afghanistan is predominately reliant on snowpack melt to supply sufficient groundwater for irrigation. Water availability is vital for agricultural production and is the main driver for changes in agricultural area (Shahriar Pervez, Budde, and Rowland 2014; UNODC 2019).

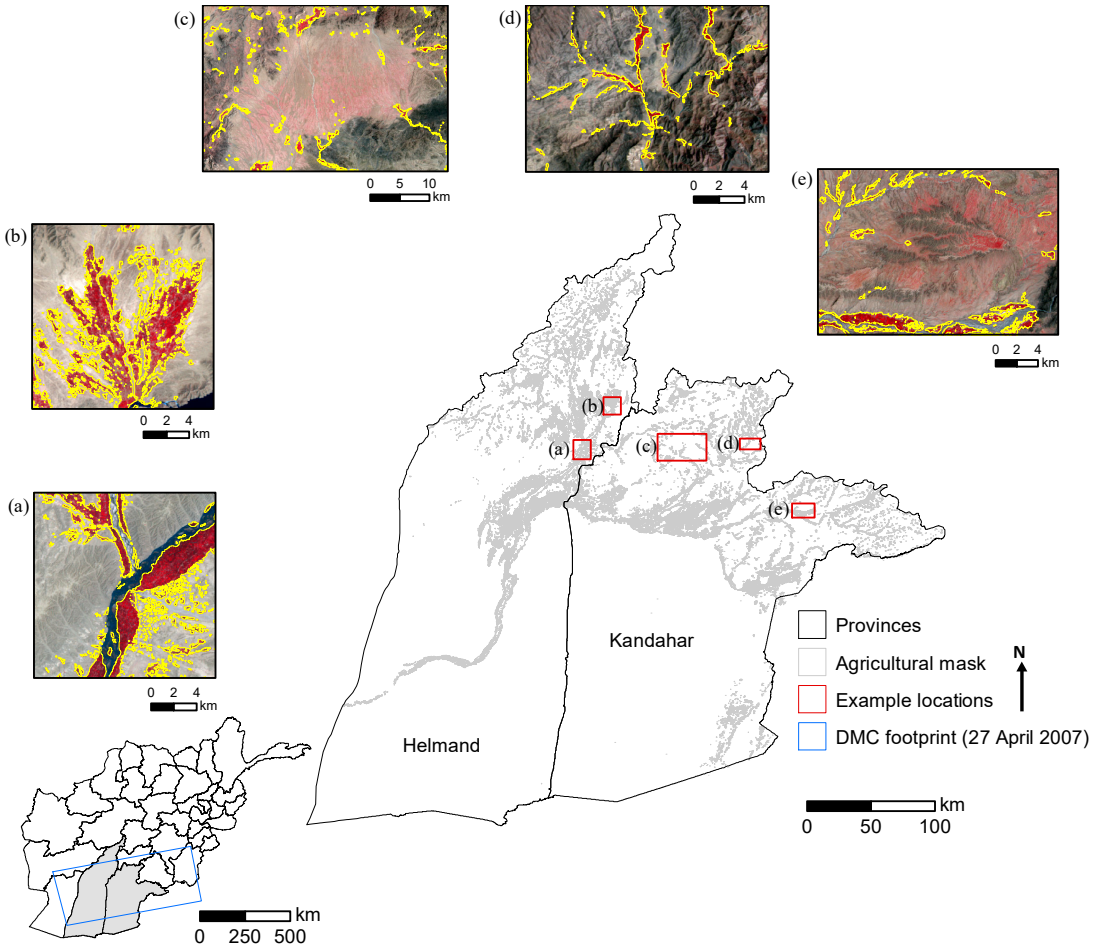


Figure 1. Helmand (centre 31.3636°N, 63.9586°E) and Kandahar (centre 31.6289°N, 65.7372°E) provinces, Afghanistan showing the agricultural area in 2007. Insets show locations used for detailed evaluation with 2007 agriculture delineated in yellow; (a) and (b) show areas of intensive agriculture, (c) and (e) show areas with natural vegetation and (d) shows agriculture in the highlands of Kandahar. Inset background is a false colour DMC image (NIR, R, G at 32 m) from 27 April 2007.

3.2. *Image data and agricultural masks*

The agricultural masks and associated Disaster Monitoring Constellation (DMC) imagery from the 2007 to 2009 opium cultivation surveys were used as labelled reference datasets of agricultural land. These densely labelled data were originally created from an unsupervised classification of orthorectified multispectral DMC imagery, with near-infrared (NIR) (0.76 to 0.90 µm), red (R) (0.63 to 0.69 µm), and green (G) (0.52 to 0.62 µm) bands at 32 m spatial resolution. DMC imagery was used because its high temporal frequency (up to daily) and wide area coverage were well suited to target the peak of opium poppy biomass. Images were collected on: 27 April 2007, 24 March, 7 April and 24 April 2008 and 25 March, 3 April and 8 April 2009. The same area was used for analysis between 2007 and 2009 based on the provincial boundaries of Helmand and Kandahar and the DMC footprint from 2007 (Figure 1), which resulted in multiple images for 2008 and 2009. The unsupervised classification was performed

using the Iterative Self-Organising Data Analysis Technique (ISODATA) with each output cluster manually labelled as agriculture or non-agriculture. These classifications were then manually edited as some clusters contained pixels of both agriculture and natural vegetation. Editing was done by trained interpreters with access to ancillary information from high resolution commercial IKONOS imagery (Taylor et al. 2010). Finally the masks were quality checked and compared with data from other years to ensure consistency of interpretation.

3.3. *Model selection*

The best performing CNN model for chip classification was selected from three state-of-the-art models: ResNet50, VGG16 and VGG19. Firstly, the input image from 27 April 2007 was split into 33 $\times$ 33 pixel chips, the smallest input image size based on these model architectures, using a non-overlapping grid. The class of the centre pixel was used as the label for each chip as the goal was to classify whole images pixel-by-pixel through reconstructing overlapping chips (Kampffmeyer, Salberg, and Jenssen 2016). All chips were z-score normalised and a 75% random sub-sample was selected for each class for training and the remaining 25% were used for independent validation. The agriculture samples in the training datasets were augmented using horizontal and vertical flipping to increase the number of samples by a factor of 2. The training and validation datasets were balanced by undersampling the majority class (non-agriculture) to match the number of samples in the agriculture class. This resulted in a total of 11,664 training samples and 1,944 validation samples across the two classes. Each model was then trained end-to-end and using an ImageNet transfer learning model with an Adam optimizer and a learning rate of 0.0001. Model performance was assessed on the validation samples using overall accuracy and the Kappa coefficient. All experiments were undertaken using Keras (Chollet 2015) with a TensorFlow (Abadi et al. 2015) backend on a workstation with a Intel Xeon E5-2687W v3 CPU, NVIDIA Quadro K2200 GPU and 64 GB of RAM. As a benchmark the CNNs were compared to a Random Forest, a pixel-based machine learning classifier, to provide comparison between a pixel-based and chip-based classifier. The number of trees (100), tree-depth (2) and maximum features used to split each internal node (10) were determined as the optimal hyper-parameters by grid search using a stratified 3 fold cross validation on the training data.

All CNN models were able to classify agricultural chips better than the Random Forest classifier with up to a 9% improvement. The ResNet50 architecture achieved the highest overall accuracy and Kappa coefficient using transfer learning, 99.02% and 0.98 respectively (Table 1) and was used for all further experiments.

Table 1. Summary of ResNet50, VGG16 and VGG19 CNN model performance for end-to-end (EE) and transfer learning (TL) training parameters for 10 epochs using 33 × 33 pixel image chips from DMC (NIR, R, G) imagery for 27 April 2007 across Helmand and Kandahar. The best performing model and metrics are highlighted in **bold**.

CNN model	Training approach	Model performance (10 epochs) (%)	
		Overall accuracy	K
ResNet50	EE	96.81	0.94
	TL	**99.02**	**0.98**
VGG16	EE	98.82	**0.98**
	TL	97.48	0.95
VGG19	EE	98.30	0.97
	TL	96.40	0.93
Random Forest	N/A	89.22	0.88

3.4. *Experiment 1: Image chip size and CNN training strategy*

Image chipping is an important pre-processing step for CNNs with fully connected layers. Three sets of fixed non-overlapping grids (33 × 33 pixels, 65 × 65 pixels and 129 × 129 pixels) were created to provide model input data at different spatial scales to investigate the effect of chip size. The chips were z-score normalised and the reference data label for each chip was assigned based on the centre pixel.

The agriculture class is heavily under-represented in the data and accounts for only 5% of samples for all image chip sizes (Table 2). Data augmentation was used for all experiments to increase the number of agriculture samples by a factor of 2. Datasets for all experiments were balanced in number between non-agriculture and augmented agriculture samples using a stratified random sample from the non-agriculture group. Sub-stratification was carried out within the non-agriculture group to select chips at the boundary of agricultural land and areas of natural vegetation, which are known confusion areas for agricultural mapping (Simms et al. 2016). Non-agriculture chips with natural vegetation were selected using an Otsu threshold (Otsu 1979) on a Normalised Difference Vegetation Index (NDVI) image of the study extent with the reference agricultural mask applied. Masking agricultural land forces the NDVI threshold to focus on selecting samples located in areas of natural vegetation.

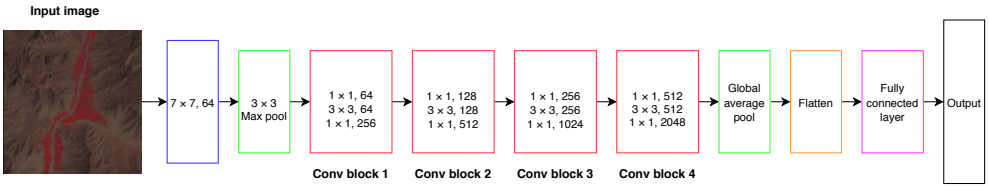


Figure 2. The ResNet50 model architecture used for agricultural mask prediction.

Three separate strategies were used to identify how best to train the ResNet50 CNN (Figure 2). These were: (1) random sampling of agriculture and non-agriculture; (2) sampling from boundary cases; and (3) targeted sampling from non-agriculture chips containing natural vegetation and boundary cases. Boundary cases are defined as chips with a non-agriculture label that contain agriculture within the chip. These samples have been introduced to provide more difficult interpretation cases to train

and validate the model.

The number of selected training and validation samples (from the total number in Table 2) for each chip size remained consistent, regardless of training strategy. Chip size 33 × 33 used 11,664 training and 1,944 validation samples, 2,868 training and 478 validation samples were used for chip size 65 × 65 and 724 training and 120 validation samples were used for chip size 129 × 129.

Table 2. Total number of chips (n) in the study area for agriculture, non-agriculture and boundary samples (non-agriculture chip label, but with agriculture present) for each size of image chip for 2007 data with the percentage of total area.

Input image chip size (pixels)	Agriculture		Boundary		Non-agriculture	
	n	Percentage area (%)	n	Percentage area (%)	n	Percentage area (%)
33 × 33	3899	5.4	10699	14.8	57556	79.8
65 × 65	1012	5.5	4414	24.0	12961	70.5
129 × 129	250	5.5	1599	35.1	2711	59.5

A separate ResNet50 CNN model was also trained to include Shuttle Radar Topography Mission (SRTM) elevation data (resampled to 32 m and min-max normalised) to investigate the effect of ancillary data. Most CNN architectures are developed for photographs, with input channels restricted to three. The green band was substituted for elevation as the NIR and R spectral bands were considered to be of higher importance for monitoring of vegetation (Panda, Ames, and Panigrahi 2010).

The CNN outputs a single prediction for each image chip so reconstruction is required to classify a whole image pixel-by-pixel. The reconstruction process used in this study applies the trained CNN model to each pixel in the image using an overlapping sliding window the same dimensions as the image chip used during training. This achieves a pixelwise classification by labelling the centre pixel of each sliding window with the model prediction for the chip (Figure 3).

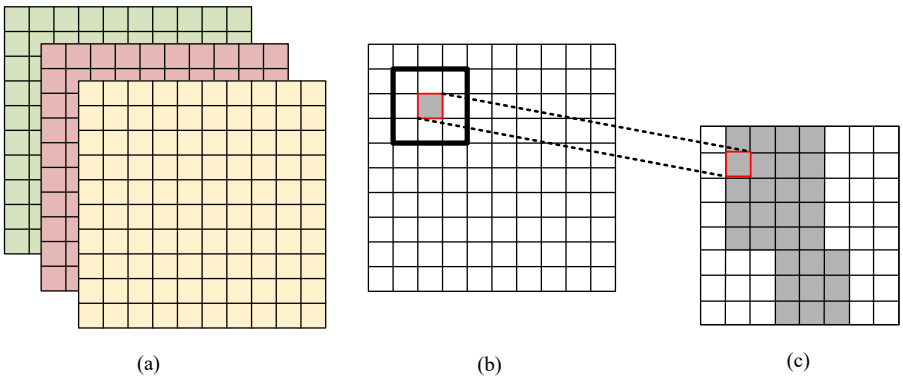


Figure 3. The process for pixelwise agricultural mask prediction using a sliding window with a trained CNN model. (a) 3 band satellite image chip, (b) sliding window applying CNN model e.g. 3 × 3 and (c) pixelwise agricultural mask production.

3.5. *Experiment 2: Transfer learning across multiple seasons*

The ability to retrain CNN models with new data is a desirable attribute for image classification. The transferability of agricultural features for continual refinement of a multi-seasonal classifier was explored using 2007, 2008 and 2009 data. Transfer learning was used to update the model previously trained on 2007 data with data from 2008, the combined model was then updated with 2009 data. The best performing training strategy from experiment 1 was used to create the 2008 and 2009 input data, which were balanced using the same image augmentation as the 2007 dataset. The total number of training samples used for 2008 and 2009 were 11,960 and 12,032 with 1,994 and 2,006 validation samples, respectively. The proportion of training data was varied (25%, 50%, 75% and 100%) to identify the number of samples required to update each year's model to a similar level of accuracy.

4. Results

4.1. *Training strategy selection*

The CNN model outputs for agricultural delineation were found to consistently achieve higher accuracy with ImageNet transfer learning across all chip sizes than end-to-end learning (Figure 4). The major difference between transfer and end-to-end learning is shown during the initial 5 epochs with higher initial training and overall accuracy, where the training accuracy is the overall accuracy of the training data. Training accuracies for all image chip sizes were found to achieve a similar accuracy after 50 epochs, unlike validation accuracies. The 129 $\times$ 129 chips were found to plateau faster than the other chip sizes across all training strategies and 65 $\times$ 65 chips took the longest to train in both transfer and end-to-end learning. End-to-end CNN models across all training strategies were also found to require more epochs of training than transfer learning.

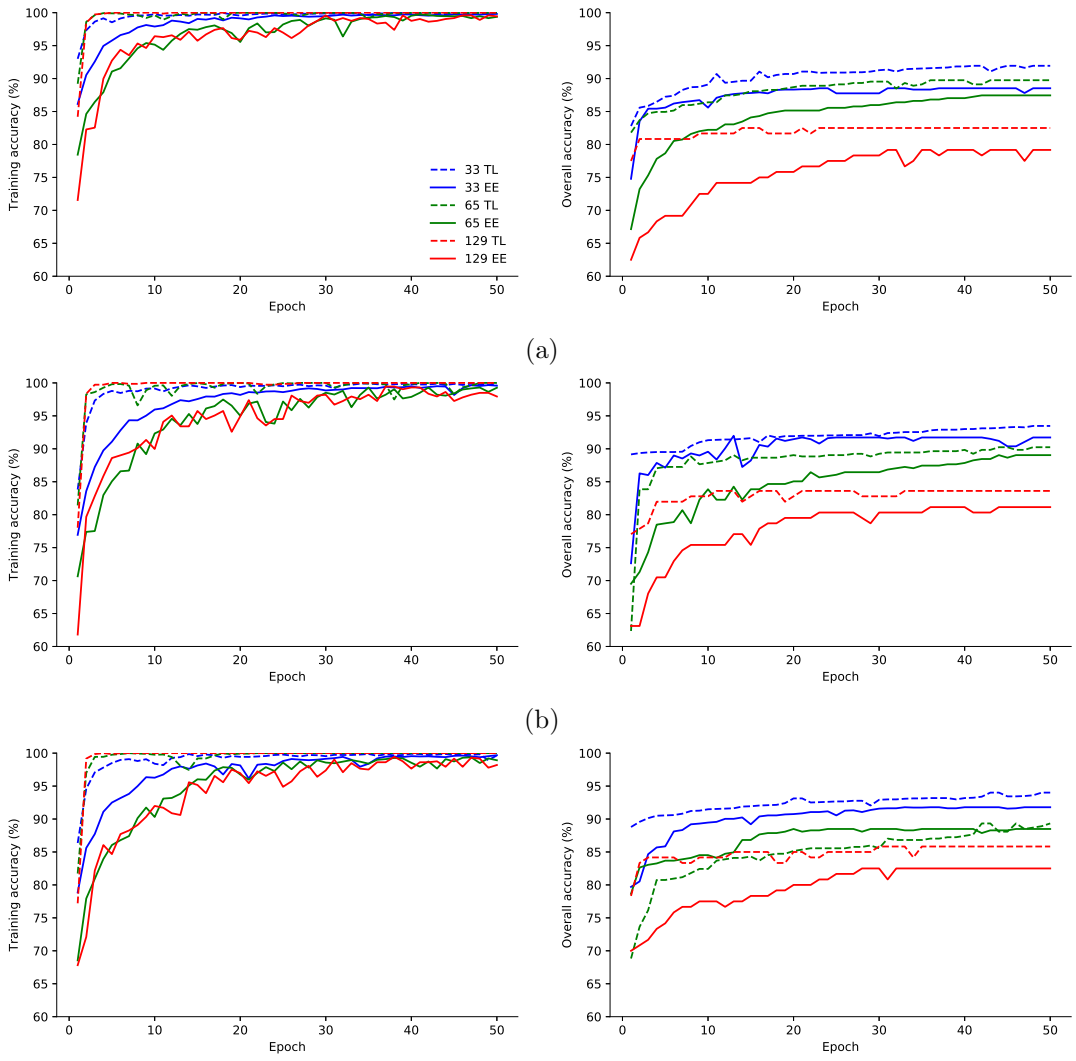


Figure 4. Evaluation of training and validation overall accuracy for three different training strategies for different image chips sizes (33 × 33, 65 × 65 and 129 × 129 pixels) using transfer (TL) and end-to-end learning (EE): (a) strategy 1: random agriculture and non-agriculture classes; (b) strategy 2: random agriculture and boundary classes; and (c) strategy 3: random agriculture class, boundary cases and NDVI targeted non-agriculture class.

Smaller chips were found to be less generalised than larger chips across various agricultural landscapes after image reconstruction (Figure 5). The 129 × 129 chip classification delineates the overall agricultural boundary extent, but performs poorly on smaller non-agricultural areas and edge cases. This can be seen as a buffering effect along agricultural boundaries. Boundaries between the agriculture and non-agriculture class were found to be well-defined for the 33 × 33 chip size. High visual agreement with the reference agricultural mask was achieved, particularly for strong edge cases such as the river valley.

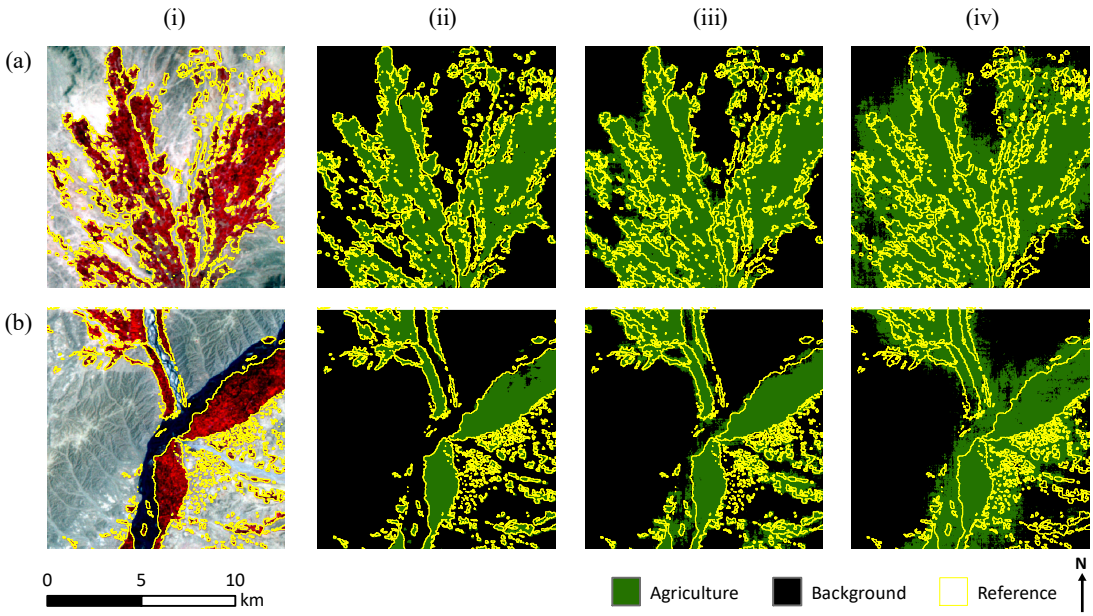


Figure 5. False colour DMC imagery (NIR, R, G at 32 m) from 27 April 2007 (i) for an (a) agriculture dominated area and (b) non-agriculture dominated area with corresponding agriculture delineation for three image chip sizes ((ii) 33 × 33, (iii) 65 × 65 and (iv) 129 × 129) using the best-performing ResNet50 CNN (strategy 3, random agriculture, boundary cases and NDVI targeted non-agriculture class with transfer learning).

The ResNet50 CNN model performance summarised in Table 3 shows the overall accuracy, Kappa coefficient, mean IoU and frequency weighted IoU, for the three different chip sizes, two model training methods (end-to-end and transfer learning) and the three different training strategies. Overall accuracy is highest for 33 × 33 chip size, with transfer learning outperforming all end-to-end training strategies (Figure 4). The best performing model was transfer-trained using data from strategy 3 and 33 × 33 chips, with an overall accuracy of 94.01%, Kappa coefficient of 88.02% and mean and weighted IoU of 50.33 and 67.61 respectively (Table 3). Strategy 3 produced the best performing models with the exception of the 65 × 65 size chips, which achieved marginally higher overall accuracy (+ 0.5% improvement using transfer learning), Kappa coefficient, mIoU and fwIoU for both end-to-end and transfer learning with strategy 2.

Table 3. Evaluation of input chip sizes (129 × 129, 65 × 65 and 33 × 33 pixels) and strategies for CNN models trained using both end-to-end (EE) and transfer learning (TL) across Helmand and Kandahar provinces on DMC (NIR, R, G) imagery in April 2007. Strategy 1: random sampling of agriculture and non-agriculture classes; strategy 2: random sampling of agriculture and boundary classes; strategy 3: random agriculture, boundary cases and NDVI targeted non-agriculture class. The best performing validation metrics for each training strategy are highlighted in **bold**.

Training strategy	Validation metric (50 epochs) (%)	Training approach					
		EE (pixels)			TL (pixels)		
		129	65	33	129	65	33
Strategy 1							
	OA	79.17	87.45	88.53	82.50	89.75	**91.94**
	K	58.33	74.90	77.07	65.00	79.50	**83.88**
	mIoU	30.31	34.80	47.75	30.04	40.32	**49.03**
	fwIoU	34.56	41.73	63.70	33.86	52.04	**65.35**
Strategy 2							
	OA	81.15	89.04	91.72	83.61	90.24	**93.47**
	K	62.30	78.09	83.44	67.21	80.48	**86.93**
	mIoU	32.05	37.36	48.56	28.31	41.66	**49.18**
	fwIoU	36.96	45.40	63.79	31.89	52.95	**65.42**
Strategy 3							
	OA	82.50	88.49	91.79	85.83	89.33	**94.01**
	K	65.00	76.10	83.57	71.67	78.66	**88.02**
	mIoU	28.13	36.33	48.57	29.63	38.17	**50.33**
	fwIoU	31.78	44.43	64.63	33.29	47.58	**67.61**

Table 4. Best performing CNN training strategies based on overall accuracy for each image chip size with prediction times using DMC (NIR, R, G at 32 m) imagery samples across Helmand and Kandahar provinces in April 2007. Strategy 2: random sampling agriculture and boundary classes and strategy 3: random agriculture, boundary cases and NDVI targeted non-agriculture class.

Input image chip size (pixels)	Best performing training strategy	Prediction time (s) (n = 250,000)
33 × 33	Strategy 3: transfer learning	384
65 × 65	Strategy 2: transfer learning	677
129 × 129	Strategy 3: transfer learning	1677

The fastest prediction times were found using smaller image chips for the same area extent (800 ha). The shortest time was achieved using 33 × 33 chips (384 seconds for 250,000 samples), as opposed to 1677 seconds for 129 × 129 chips (Table 4).

Visual evaluation of the Helmand and Kandahar agricultural mask for April 2007 found most complex agricultural areas and highland vegetation were well delineated (Figure 6). The CNN identified the distinct difference between features of the background class and agriculture. Large extents of natural vegetation were found to be correctly classified, with small commission errors of agriculture (Figure 6 (a)). There are also regions of high commission in low lying areas surrounding highland regions in Kandahar (Figure 6 (c)).

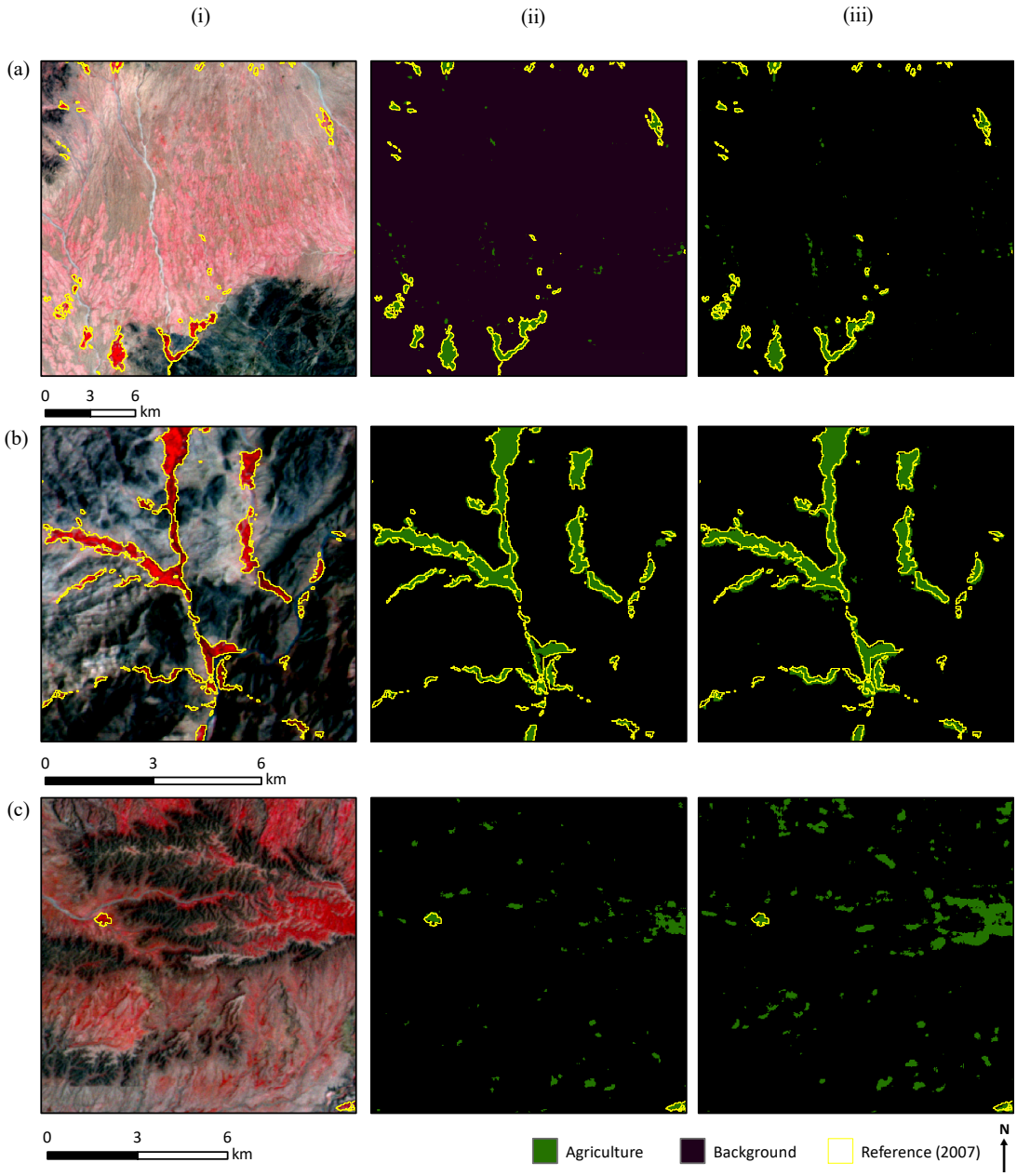


Figure 6. Visual interpretation of (i) DMC imagery for the best performing CNN classification model (training strategy 3: random agriculture, boundary cases and NDVI targeted non-agriculture class with transfer learning) with input image chip size 33 × 33 pixels for (ii) spectral and (iii) SRTM elevation data (resampled to 32 m). Image extents have been selected based on prior knowledge of confusion areas for interpretation. (a) Large extent of natural vegetation, (b) well-delineated agriculture in highland areas and (c) commission of agriculture surrounding highland areas in Kandahar, Afghanistan. False colour DMC imagery (NIR, R, G at 32 m) for 27 April 2007.

Substituting elevation data (SRTM) with the green spectral band (Table 5) and training with the best performing strategy resulted in marginally lower performance than using only spectral data (- 0.63% overall accuracy). The visual comparison of spectral and elevation CNNs for 2007 (Figure 6) show little difference between natural

vegetation in highland areas and an increase in the commission error of agriculture in the mountains of Kandahar province (Figure 6 (c)).

Table 5. Evaluation of using elevation data (SRTM) and DMC imagery (NIR, R, G at 32 m) across Helmand and Kandahar provinces in April 2007 for agricultural delineation using transfer learning, targeted background sampling (training strategy 3) and image chip size 33 × 33 pixels.

Training data	Model performance (50 epochs) (%)			
	Overall accuracy	K	mIoU	fwIoU
NIR, R, G	94.01	88.02	50.33	67.61
NIR, R, SRTM	93.38	86.78	49.43	65.47

4.2. *Multi-seasonal CNN application*

A multi-season model was trained starting with the best performing 2007 CNN (Table 5). The black dotted line in Figure 7 (a) shows the overall accuracy of the 2007 model on classifying the 2008 validation data (80.16%). The 2007 model trained faster over the first 10 epochs when updated with 75% (977 ha) of the available 2008 data than the other proportions (25%, 50%, and 100%) and achieved a similar overall accuracy to a single-season 2008 model (92.83% and 91.78% green line and red dotted line respectively in Figure 7 (a)). Adding data from 2008 increased the overall accuracy of the 2007 model by + 12.67% to 92.83% and the Kappa coefficient, mean and weighted IoU also increases by 24.04%, 7.29 and 11.46, respectively (Table 6).

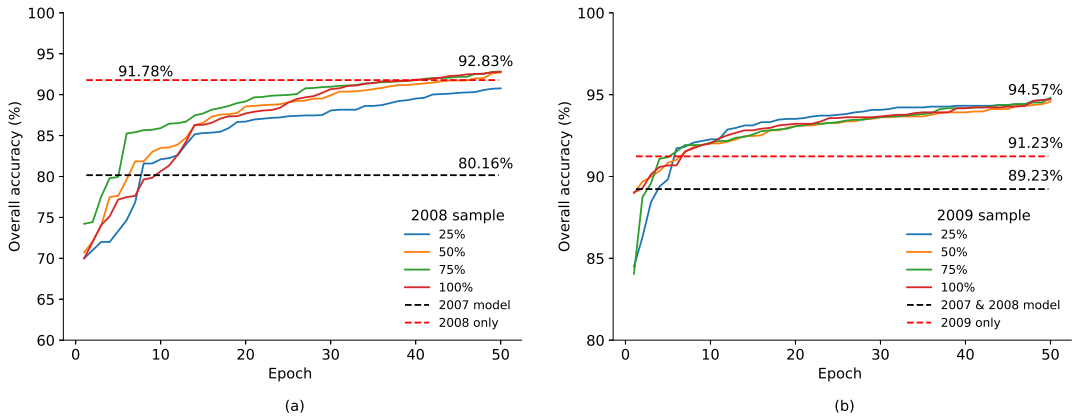


Figure 7. Evaluation of updating the best performing 2007 model using multiple training sample proportions (25%, 50%, 75% and 100% of total training data available) from 2008 and 2009 training data. (a) Transfer learning of 2008 training data using the 2007 model and (b) transfer learning of 2009 training data using the 2007 model updated with 75% of 2008 training data. Previous year's model with no additional training is shown by the black dashed line. Target year's model trained from the ImageNet dataset using 100% of available training data is shown by the red dashed line.

The analysis was repeated for the 2009 agriculture mask classification (Figure 7 (b)). The CNN was trained using only 2009 data to provide a single-year model with and overall accuracy of 91.23% (the red dotted line in Figure 7 (b)). The previous years' combined model (trained on 2007 and 75% of 2008 data) with no training from

2009 achieved an overall accuracy of 89.23% (the black dotted line in Figure 7 (b)). This was an increase of + 9.07% compared with using the 2007 model on 2008 imagery, showing a year-on-year improvement with additional data across seasons. Updating the previous years' model with 25% (317 ha) of available data trained faster than other sample proportions, improving the overall accuracy by 5.34% with the Kappa coefficient, mean and weighted IoU increasing by 10.68, 9.86 and 16.18, respectively (Figure 7 (b) and Table 6).

Table 6. Evaluation of transfer learning (TL) using a CNN trained on DMC imagery (NIR, R, G at 32 m) from 2007, 2008 and 2009 for Helmand and Kandahar provinces. *Retrain* uses the optimal percentage of available training samples with CNN TL from the previous year. The retrained 2008 model uses the 2007 model as a starting point and the retrained 2009 model uses the retrained 2007 model with 2008 data (75%).

Training year & model	Model performance (50 epochs) (%)			
	Overall accuracy	K	mIoU	fwIoU
2008				
2007 TL only	80.16	61.62	39.39	48.63
Retrain (75%)	92.83	85.66	46.68	60.09
2009				
2008 TL only	89.23	78.44	41.05	52.23
Retrain (25%)	94.57	89.12	50.91	68.41

Visual inspection of the multi-seasonal classifications in areas of known confusion between agriculture and natural vegetation show differences between years (see Figure 6 to Figure 9). In 2008 there are fewer agriculture commission errors than in 2009, despite the higher overall accuracy of the 2009 model. There are noticeable seasonal differences in natural vegetation between all three years. Larger areas of natural vegetation are found in 2007 (Figure 6 (a) and (c)) and 2009 (Figure 9 (a) and (c)), but little natural vegetation is found across the same area for 2008 (Figure 8 (a) and (c)).

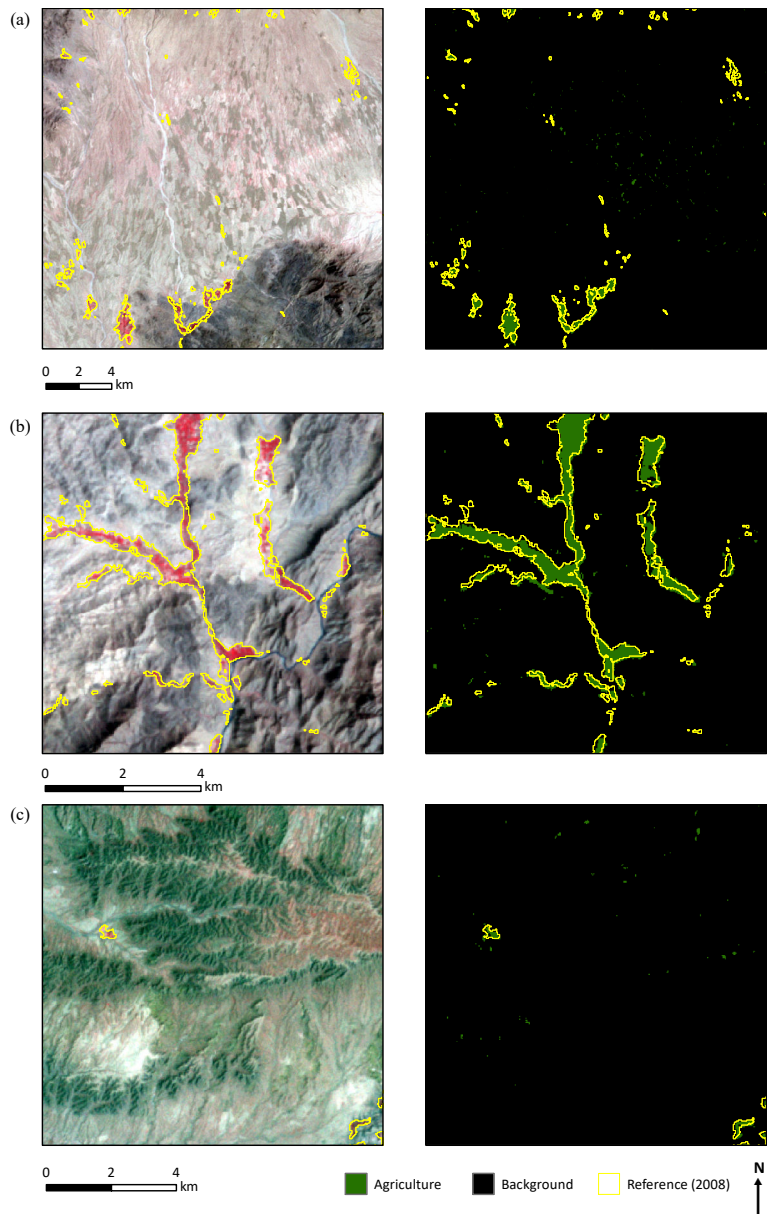


Figure 8. Visual interpretation of the 2008 CNN model with 75% of available training data using training strategy 3 (random agriculture, boundary cases and NDVI targeted non-agriculture classes with transfer learning) with input image chip size 33 × 33 pixels. Image extents have been selected based on prior knowledge of confusion areas for interpretation. (a) Large extent of natural vegetation, (b) well-delineated agriculture in highland areas and (c) commission of agriculture in highland areas in Kandahar, Afghanistan. False colour DMC imagery (32 m) for April 2008.

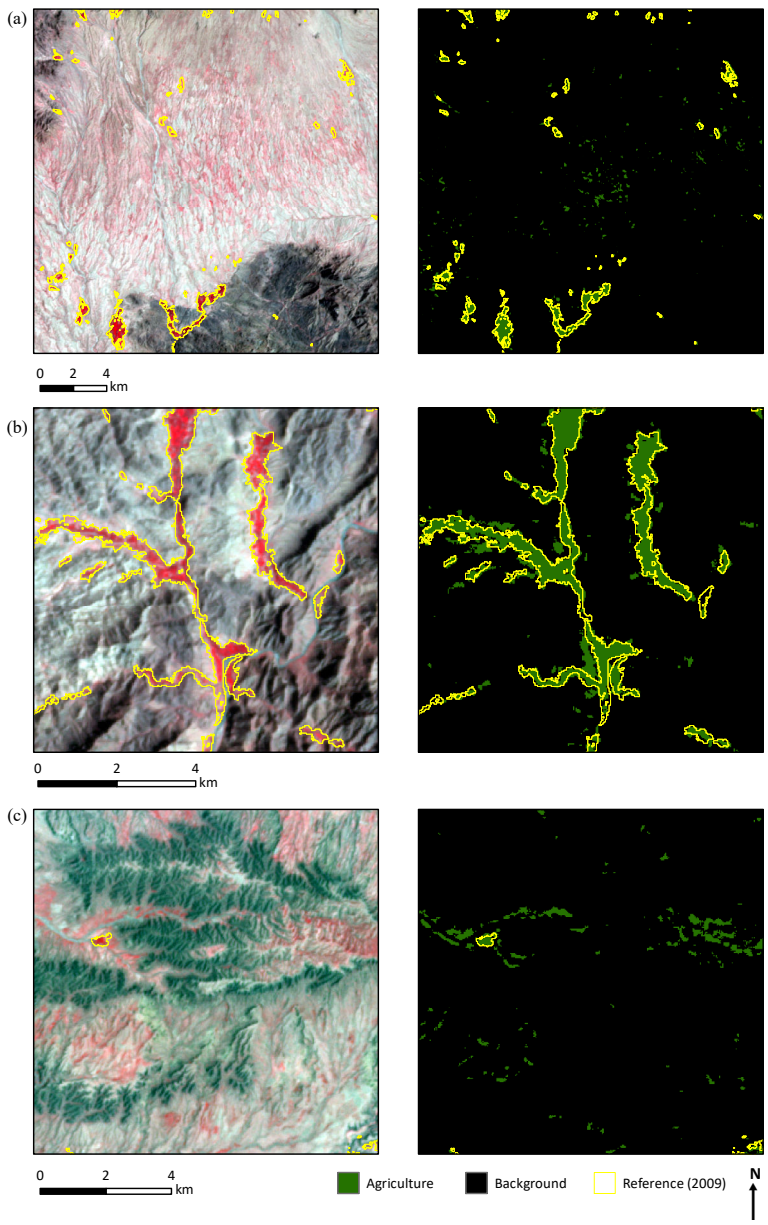


Figure 9. Visual interpretation of the 2009 CNN model with 25% of available training data using training strategy 3 (random agriculture, boundary cases and NDVI targeted non-agriculture classes with transfer learning) with input image chip size 33 × 33 pixels. Image extents have been selected based on prior knowledge of confusion areas for interpretation. (a) Large extent of natural vegetation, (b) well-delineated agriculture in highland areas and (c) commission of agriculture in highland areas in Kandahar, Afghanistan. False colour DMC imagery (32 m) for April 2009.

5. Discussion

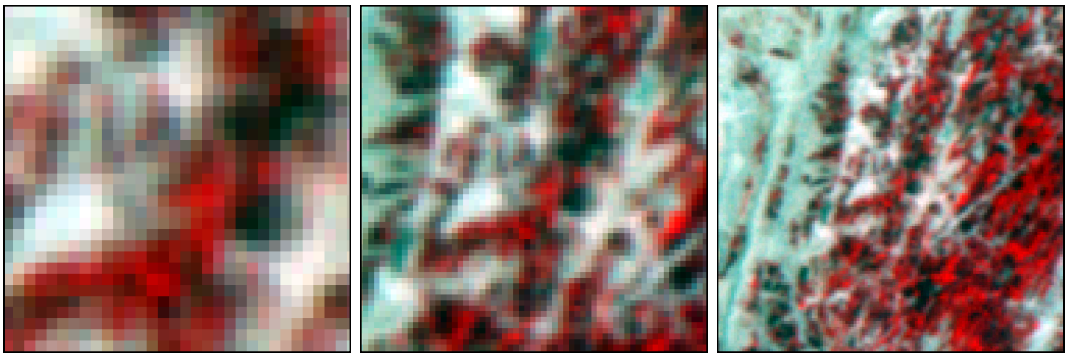
5.1. *Importance of contextual information*

In the operational opium survey, image interpreters manually refine the agricultural area defined by an unsupervised classification to adjust boundaries and remove areas

of confusion, using contextual information to support their decisions (UNODC 2018). Contextual information includes field boundaries contrasting with desert and rock, buildings and river channels. The CNN overcomes the limitations of pixel-based unsupervised classification by encoding the surrounding landscape features within each chip, with the scale of features dependent on the chip size. For example, in Figure 10 there are local field parcels and texture visible in the 33 × 33 chip (Figure 10 (a)), many more fields and the surrounding desert in the 65 × 65 chip (Figure 10 (b)) and a much greater proportion of desert and lower proportion of local information (relative to the centre pixel) within the 129 × 129 chip (Figure 10 (c)). Smaller chip sizes result in a set of more localised features in the CNN, which is analogous to how a human interpreter will use local context (using a larger mapping scale) to refine a boundary.

In the CNN output each prediction is based on a single chip, with pixel-by-pixel classification achieved by sliding a chip-sized window across the input image, assigning the prediction to the centre pixel. A one pixel shift of a small chip results in a more substantial change to the surrounding contextual information than larger chips, explaining why smaller chips are more sensitive to localised change and result in increased classification accuracy (Kroupi et al. 2019). Also classification for the whole image is much more efficient with smaller chips as prediction for each sliding window is faster, despite the longer training times. A limitation of the fixed sampling grid is that a different number of samples are created for each chip size. An equal number of samples for each chip scale could be produced by introducing an overlapping sampling grid with the same centre pixel, but would result in non-independent samples.

Making the image chip smaller might further improve the accuracy of the agricultural mask but there is likely to be a trade-off as the amount of contextual information decreases. The ResNet50 architecture limits the smallest chip size to 32 × 32 (33 × 33 was used here to accommodate a centre pixel), but to investigate smaller chips would require changing the network architecture. Alternatively, Fully Convolutional Networks (FCNs) could be used to extract agricultural features from arbitrary chip sizes and provide pixel-by-pixel predictions, overcoming this limitation and improving the speed of image classification (Paisitkriangkrai et al. 2015).



(a) (b) (c)

Figure 10. Examples of the sliding window sizes (a) 33 × 33, (b) 65 × 65 and (c) 129 × 129 pixels using the same centre pixel for agriculture. False colour DMC imagery (NIR, R, G at 32 m) for 27 April 2007.

The influence of the centre pixel is an important consideration for pixel-by-pixel classification (Zhang and Lu 2019) as it is used to label training and validation samples into agriculture and non-agriculture classes. As an experiment into pixel bias, the

centre pixels for agriculture validation samples for each chip size were altered to the mean average of non-agriculture chips for each input channel (NIR, R, G). There was a negligible difference in overall pixel accuracy (- 0.36%) for the smallest chip (33 × 33), and no differences were found with the other chip sizes. This shows the centre pixel plays no individual role in prediction and why pixel-level classification using image chips generalises as chip size increases.

Human interpreters have other sources of information at their disposal to help in the delineation of agricultural land. In our study area, using elevation data with the CNN model made little difference to classification accuracy (- 0.63%). Highland and lowland areas are visually different and it is likely that the spatial and spectral information in the chip is not improved further by adding explicit height information. In future work, other image sensors (e.g. hyperspectral (Dell'Acqua et al. 2006) and Synthetic Aperture Radar (SAR) (Liu, He, and Li 2017)), more ancillary data, and new deep learning architectures with an increased number of input channels could be investigated to improve agricultural land classification.

5.2. *Year-on-year transfer learning for agricultural mask production*

Transfer learning has similarities in the way a trained interpreter gains experience as they are both able to build upon existing knowledge. Whereas end-to-end learning is similar to an inexperienced interpreter with no prior knowledge of image classification. Transfer learning was consistently faster than end-to-end learning demonstrating some similarity between the underlying features across years. The accuracy for CNN models trained by transfer learning were also generally higher than their end-to-end counterparts, which is consistent with previous studies of transfer learning for remote sensing data (UC Merced land use, RS19 and Brazilian coffee scenes) (Nogueira, Penatti, and dos Santos 2017).

Transfer learning increases the total number of samples used to train the CNN. However, sampling of the inter-annual changes between 2007 and 2009 was still required to refine the model. Even with very little (25%), or no training data from the target year, the model's performance increases. Fewer samples are required each year for training as multiple years worth of different landscapes and examples of agricultural features adequately extract and predict common features. This alleviates the burden for complete labelled datasets for CNN classification, which remains a challenge in remote sensing. Transfer learning from remote sensing data provides the opportunity to provide timely initial predictions without the need for additional labelled datasets. Updating the model for subsequent years may only require 25% or less of the total sample fraction.

Targeted sampling using boundary samples and an NDVI threshold (strategy 3) to identify samples containing natural vegetation was the optimal training strategy. By directing the training to areas of known confusion in the background class the CNN was better able to separate edge cases, which supports other studies reporting a decrease in CNN performance with random sampling compared to a selective strategy (Van Grinsven et al. 2016). A human interpreter would also improve their delineations with more experience of difficult interpretation cases. Future CNN applications could include additional samples using post-classification refinement to further improve classification performance.

Using historical information to inform predictions of agricultural land has the potential to substantially decrease the manual effort associated with current agricultural

mask production by the UNODC. Change detection techniques could be used by interpreters to focus manual editing on those areas identified as having changed. Misclassified areas could then be used as training samples to improve the model for the next year's agricultural mask. The outlined rationale could be used to develop a data-driven classification based solely on historical knowledge of agricultural land within Afghanistan. Utilising existing knowledge to derive upcoming agricultural masks without the need for additional data is an exciting prospect for timely image classification.

6. Conclusions

The overall accuracy for the ResNet50 CNN was $> 94\%$ for agricultural land classification in all years (2007 to 2009). The best results were achieved using a chip size of 33×33 pixels and a NDVI-based sampling strategy, which targeted the main source of confusion between natural and agricultural vegetation. Transfer learning using a pre-trained model (ImageNet) was found to achieve higher overall accuracy than end-to-end learning (+ 2.2%). Substitution of the green spectral band for elevation data achieved marginally lower performance (- 0.6%).

When considering transfer learning of the CNN model year-on-year, the classification of 2008 imagery using the 2007 ResNet50 model, with no additional training, resulted in an accuracy of 80.2%, improving to 89.3% for 2009 imagery using a combined 2007 and 2008 model. Using only 25% of the 2009 data to update the combined model further improved classification accuracy to 94.6%. High classification performance coupled with continual model refinement from additional data shows the potential for CNNs to replace human interpreters for the UNODC's agricultural mask production. Reducing the manual effort in the production of the mask to a small proportion of the total area would improve the speed and efficiency of the survey, reducing the overall cost. Deep transfer learning across multiple years presents an exciting opportunity for timely and efficient land cover classification.

Acknowledgement(s)

The authors would like to thank the UK Natural Environment Research Council (NERC) for providing the funding for this project. The authors would also like to acknowledge the United Nations Office on Drugs and Crime (UNODC) for providing support during this research.

Disclosure statement

The authors declare no conflict of interest.

Funding

This research was supported by the UK Natural Environment Research Council (NERC) sponsored Data, Risk and Environmental Analytical Methods (DREAM) Centre for Doctoral Training [NERC Ref: NE/M009009/1].

Data availability statement

The data used in this study are subject to licensing restrictions or are confidential. For more information see doi:10.17862/cranfield.rd.13228634.

References

Abadi, M, A Agarwal, P Barham, E Brevdo, Z Chen, C Citro, G S Corrado, et al. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." .

Ball, J E, D T Anderson, and C S Chan. 2017. "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community." *Journal of Applied Remote Sensing* 11: 042609.

Belgiu, M, and L Drgu. 2016. "Random forest in remote sensing: A review of applications and future directions." *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31.

Cheng, G, J Han, and X Lu. 2017. "Remote sensing image scene classification: benchmark and state of the art." *Proceedings of the IEEE* 105: 1865–1883.

Chollet, F. 2015. "Keras." `https://github.com/keras-team/keras`.

Chollet, F. 2017. *Deep learning with Python*. New York: Manning Publications.

Cohen, J. 1960. "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement* 20 (1): 37–46.

Dell'Acqua, F., P. Gamba, V. Casella, F. Zucca, J. A. Benediktsson, G. Wilkinson, A. Galli, et al. 2006. "HySenS data exploitation for urban land cover analysis." *Annals of Geophysics* 49 (1): 311–318.

Deng, Z, H Sun, S Zhou, J Zhao, L Lei, and H Zou. 2018. "Multi-scale object detection in remote sensing imagery with convolutional neural networks." *IPRS Journal of Photogrammetry and Remote Sensing* 145: 3–22.

Duchi, J, E Hazan, and Y Singer. 2011. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization." *Journal of Machine Learning Research* 12: 2121–2159.

FAO. 2016. *The Islamic Republic of Afghanistan: Land cover atlas*. Technical Report. FAO.

Feng, Y, W Diao, X Sun, M Yan, and X Gao. 2019. "Towards automated ship detection and category recognition from high-resolution aerial images." *Remote Sensing* 11 (1901).

Foody, G M. 2002. "Status of land cover classification accuracy assessment." *Remote Sensing of Environment* 80 (1): 185–201.

Fu, G, C Liu, R Zhou, T Sun, and Q Zhang. 2017. "Classification for high resolution remote sensing imagery using a fully convolutional network." *Remote Sensing* 9 (5): 498.

Gislason, P O, J A Benediktsson, and J R Sveinsson. 2006. "Random Forests for land cover classification." *Pattern Recognition Letters* 27 (4): 294300.

Goodfellow, I, Y Bengio, and A Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Haenssle, H. A., C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, et al. 2018. "Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists." *Annals of Oncology* 29 (8): 1836–1842.

He, K, X Zhang, S Ren, and J Sun. 2016. "Deep Residual Learning for Image Recognition." In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Kampffmeyer, M, A B Salberg, and R Jenssen. 2016. "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks." In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Las Vegas, USA, 680–688.

Kingma, D P, and J L Ba. 2015. "Adam: A method for stochastic optimization." .

Koga, Y, H Miyazaki, and R Shibasaki. 2018. "A CNN-based method of vehicle detection from aerial images using hard example mining." *Remote Sensing* 10 (124).

Kroupi, E, M Kesa, V D Navarro-Snchez, S Saeed, C Pelloquin, B Alhaddad, L Moreno, A Soria-Frisch, and G Ruffini. 2019. "Deep convolutional neural networks for land-cover classification with Sentinel-2 images." *Journal of Applied Remote Sensing* 13: 024525.

Lecun, Y, Y Bengio, and G Hinton. 2015. "Deep learning." *Nature* 521: 436–444.

Liu, H, L He, and J Li. 2017. "Remote sensing image classification based on convolutional neural networks with two-fold sparse regularization." In *IEEE International Geoscience and Remote Sensing Symposium*, Texas, USA, 992–995.

Liu, X, Y Tian, C Yuan, F Zhang, and G Yang. 2018. "Opium Poppy Detection Using Deep Learning." *Remote Sensing* 10 (12): 1886.

Long, J, E Shelhamer, and T Darrell. 2015. "Fully convolutional networks for semantic segmentation." In *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, 3431–3440.

Lucas, R, K Medcalf, A Brown, P Bunting, J Breyer, D Clewley, S Keyworth, and P Blackmore. 2011. "Updating the Phase 1 habitat map of Wales, UK, using satellite sensor data." *ISPRS Journal of Photogrammetry and Remote Sensing* 66: 81–102.

Nogueira, K, O A B Penatti, and J A dos Santos. 2017. "Towards better exploiting convolutional neural networks for remote sensing scene classification." *Pattern Recognition* 61: 539–556.

Otsu, N. 1979. "A threshold selection method from gray-level histograms." *IEEE Transactions on Systems, Man, and Cybernetics* 9 (1): 62–66.

Paisitkriangkrai, S, J Sherrah, P Janney, and A Van-Den Hengel. 2015. "Effective semantic pixel labelling with convolutional networks and Conditional Random Fields." In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Boston, USA, 36–43.

Panda, S S, D P Ames, and S Panigrahi. 2010. "Application of vegetation indices for agricultural crop yield prediction using neural network techniques." *Remote Sensing* 2 (3): 673–696.

Penatti, A B, K Nogueira, and J A Santos. 2015. "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 44–51.

Pouliot, D, R Latifovic, J Pasher, and J Duffe. 2019. "Assessment of Convolution Neural Networks for Wetland Mapping with Landsat in the Central Canadian Boreal Forest Region." *Remote Sensing* 11 (7): 772.

Russakovsky, O, J Deng, H Su, J Krause, S Satheesh, S Ma, Z Huang, et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–252.

Shahriar Pervez, M, M Budde, and J Rowland. 2014. "Mapping irrigated areas in Afghanistan over the past decade using MODIS NDVI." *Remote Sensing of Environment* 149: 155–165.

Shin, H C, H R Roth, M Gao, L Lu, Z Xu, I Nogues, J Yao, D Mollura, and R M Summers. 2016. "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning." *IEEE Transactions on Medical Imaging* 35 (5): 1285–98.

Simms, D M, T W Waine, J C Taylor, and T R Brewer. 2016. "Image segmentation for improved consistency in image-interpretation of opium poppy." *International Journal of Remote Sensing* 37 (6): 1243–1256.

Simonyan, K, and A Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." .

Taylor, J C, T W Waine, G R Juniper, D M Simms, and T R Brewer. 2010. "Survey and monitoring of opium poppy and wheat in Afghanistan: 2003-2009." *Remote Sensing Letters* 1 (3): 179–185.

Tieleman, T, and G Hinton. 2012. *Lecture 6.5 - RMSProp*. Technical Report. COURSERA: Neural Networks for Machine Learning.

UNODC. 2018. *Afghanistan Opium Survey 2018*. Technical Report. UNODC.

UNODC. 2019. *World Drug Report 2019: Depressants*. Technical Report. UNODC.

Van Grinsven, M.J.J.P, B Van Ginneken, C.B Hoyng, T Theelen, and C.I Sanchez. 2016. "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images." *IEEE Transactions on Medical Imaging* 35 (5): 1273–1284.

Xia, G, J Hu, F Hu, B Shi, X Bai, and Y Zhong. 2017. "AID: a benchmark data set for performance evaluation of aerial scene classification." *IEEE Transactions on Geoscience and Remote Sensing* 55 (7): 3965–3981.

Yamashita, R, M Nishio, R Kinh, G Do, and K Togashi. 2018. "Convolutional neural networks: an overview and application in radiology." *Insights into Imaging* 9: 611–629.

Yang, Y, and S Newsam. 2010. "Bag-of-visual-words and spatial extensions for land-use classification." In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, California, USA, 270–279.

Yosinski, J, J Clune, Y Bengio, and H Lipson. 2014. "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems* 27: 3320–3328.

Zeiler, M D, and R Fergus. 2013. "Stochastic Pooling for Regularization of Deep Convolutional Neural Networks." .

Zhang, W, and X Lu. 2019. "The spectral-spatial joint learning for change detection in multispectral imagery." *Remote Sensing* 11 (240).

Zhang, W, P Tang, and L Zhao. 2019. "Remote Sensing Image Scene Classification Using CNN-CapsNet." *Remote Sensing* 11 (5): 494.