Farm Area Segmentation in Satellite Images Using DeepLabv3+ Neural Networks

Sara Sharifzadeh[1][0000-0003-4621-2917] and Jagati Tata[1] and Hilda Sharifzadeh[1], Bo Tan[2][0000-0002-6855-6270]

[1] Coventry University University, Coventry CV1 2JH, UK
[2] Tampere University, Tampere, Finland
ac8115@Coventry.ac.uk

Abstract. Farm detection using low resolution satellite images is an important part of digital agriculture applications such as crop yield monitoring. However, it has not received enough attention compared to high-resolution images. Although high resolution images are more efficient for detection of land cover components, the analysis of low-resolution images are yet important due to the low-resolution repositories of the past satellite images used for timeseries analysis, free availability and economic concerns. In this paper semantic segmentation of farm areas is addressed using low resolution satellite images. The segmentation is performed in two stages; First, detect local patches or Regions of Interest (ROI) that include farm areas are detected. Next, deep semantic segmentation strategies are employed to detect the farm pixels. For patch classification, two previously developed local patch classification strategies are employed; a two-step semi-supervised methodology using hand-crafted features and Support Vector Machine (SVM) modelling and transfer learning using the pretrained Convolutional Neural Networks (CNNs). For the latter, the high-level features learnt from the massive filter banks of deep Visual Geometry Group Network (VGG-16) are utilized. After classifying the image patches that contain farm areas, the DeepLabv3+ model is used for semantic segmentation of farm pixels. Four different pretrained networks, resnet18, resnet50, resnet101 and mobilenetv2, are used to transfer their learnt features for the new farm segmentation problem. The first step results show the superiority of the transfer learning compared to hand-crafted features for classification of patches. The second step results show that the model trained based on resnet50 achieved the highest semantic segmentation accuracy.

Keywords: Farm Detection, Semantic Segmentation, Satellite Image.

1 Introduction

Satellite image analysis is an important topic in land cover classification and remote sensing domain. In digital agriculture, farm detection is a key factor for different agricultural applications such as diagnosis of diseases and welfare-impairments, crop yield monitoring and surveillance and control of micro-environmental conditions [1–4].

While new high-resolution satellites are launched every day, it is still important to study and use Low-resolution satellite imagery that is being used since more than 30 years. That is due to the fact that the increased resolution offered by new sensors improve the accuracy and precision, yet the continuity of the existing low-resolution systems data is crucial for time series analysis. One important application of time series investigation is change detection, that requires comparison with low resolution images of the old databases [5, 6]. Another example of using low-resolution satellite images for crop monitoring and yield forecasting is [7], that uses Landsat imagery in order to expand the used operational systems. Furthermore, the processing time and cost of analyzing high resolution satellite images is more [8], while the variations in sensor angle and increase in shadows might influence the accuracy when using high resolution sensors [8]. Such factors challenge the precision of spatial rectification. Then, a compromise between accuracy and cost should be considered for the resolution of the satellite images depending on the application. Then, for land cover classification and semantic segmentation of large features such as farms, low resolution satellite images for instance, Landsat are appropriate [3].

Image segmentation methods address the problem of finding objects boundaries in images. This leads to assigning multiple sets of pixels in an image into different classes or objects [9].

There is a long history for land cover classification and semantic segmentation of meaningful objects from the scene. In early works when pixels were bigger than ground features due to very low resolution [10, 11], pixels , sub-pixel or object level analysis were carried out using unsupervised and supervised techniques such as Neural Networks (NN), decision trees and nearest neighbors and hybrid classification [12–17] were developed. Then, due to the significant increase in spatial resolution of images, objects include several pixels. Therefore, Object-Based Image Analysis (OBIA) was developed for the improved spatial resolution of images [11] to deal with complex classes [18]. OBIA assigns groups of pixels into shapes with a meaningful representation of objects [10]. For this aim, usually image segmentation is performed followed by feature extraction and classification. The segmentation step is more critical and influences the overall accuracy [19, 20].

In many cases software and computational tools such as ERDAS and Khoros 2.2 were used [17]. eCognition and ArcGIS softwares are recent examples in this case [8, 21]; Traditional hand-crafted feature extraction and discrimination techniques for object classification in remote sensing was reviewed in [22]. When using low resolution images such as Landsat 8, appropriate choice of training samples, segmentation parameters and modelling strategy is important. That is a challenge in using software-based strategies and limit their accuracy [21]. An example in this case is selection of a suitable segmentation scale to avoid over and under segmentation in Object Based Image Analysis OBIA. Although there are several reports of superior performance on different landscapes, due to the segmentation scale issue and lower resolution, OBIA is not very ideal for Landsat data [21].

Utilization of saliency maps for pixel level classification of high-resolution satellite images was performed based on spectral domain analysis such as Fourier and wavelet transforms for creation of local and global saliency maps [23, 24] . In another work

based on saliency analysis low level SIFT descriptors, middle-level features using locality-constrained linear coding (LLC) and high level features using deep Boltzmann machine (DBM) were combined [25].

In addition, the state of the art CNNs have been used recently for classification of satellite images [26–28]. Due to the limited effectiveness of manual low-level feature extraction methods in highly varying and complex images such as diverse range of land coverage in satellite images, deep feature learning strategies have been applied recently for ground coverage detection problems. One of the effective deep learning strategies is the deep CNNs due to its bank of convolutional filters that enables quantification of massive high-level spectral and spatial features. For semantic segmentation problems, the most recently developed methods are based on deep learning techniques [29]. Examples of such techniques are fully convolutional network (FCN) [30–33], encoder-decoder architectures such as Unet [34] and other similar architectures such as an subsample-upsample architecture in [35], LinkNet [36], ResNet [37] , AD-LinkNet [29]. Recently. deepLabv3 [38] and deepLabv3+ [39] methods based on atrous convolution have been developed for semantic segmentation.

In this paper, the problem of farm detection and segmentation using low resolution satellite images is addressed. In our previous contribution, a farm detection strategy was developed at patch level [40]. The analysis include two different strategies; the first one was a semi-supervised strategy based on hand-crafted features combined by classification modeling similar to [40–43]. The developed algorithm consists of an unsupervised pixel-based segmentation of vegetation area using Normalized Difference Moisture Index (NDMI), followed by a supervised step for texture area classification and farm detection; GLCM and 2-D DCT features are used in an SVM framework for texture classification and in then, object-based morphological features were extracted from the textured areas for farm detection. The second one was a CNN-based transfer learning strategy that uses the pre-trained VGGNet16 for local patch classification.

The main contribution of this paper is segmentation of farm areas semantically at pixel level. The analysis strategy consists of two main stages; first similar to our previous work [40], local image patches or ROIs that include farm areas are detected. Then, having found the local ROIs consisting the farm areas, in the next step, semantic segmentation of farm regions in the ROIs is performed using deepLabv3+ modelling strategy [39]. Based on transfer learning concept, labelled ROIs including farms are used together with four different pretrained networks, resnet18, resnet50, resnet101 and mobilenet and the transferred models results are compared.

The rest of paper is organized as follows; section 2 is about data description. Section 3 describes the both classification strategies. The experimental results are presented in section 4 and we finally conclude in section 5.

2 Dara Description

Landsat 8 is the latest earth imaging satellite of the Landsat Program operated by the EROS Data Centre of United States Geological Survey (USGS), in collaboration with NASA. The spatial resolution of the images is 30m. Landsat 8 captures more than 700

scenes per day. The instruments Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) in Landsat 8 have improved Signal to Noise Ratio (SNR). The products downloaded are 16-bit images (55,000 grey levels) [3, 44]. There are 11 bands out of which, the visible and infrared (IR) bands are used in this paper. The data set consist Landsat 8 image of an area near Tendales, Ecuador (See Fig. 1). In this work, different combinations of band are used for calculating vegetation and moisture indices used in estimation of vegetation green areas as well as visible RGB bands for classification analysis.

Fig. 1. Landsat 8 RGB image of Tendales, Ecuador.

3 Methodology

In this section the procedures used for classification of farm patches and segmentation of farm areas in the detected patch are described. Fig. 2 shows the overall analysis strategy in this work.

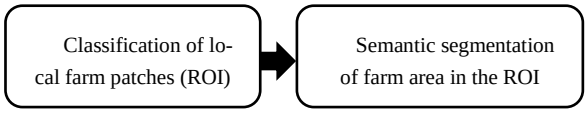


Fig. 2. Overall analysis strategy of this paper

3.1 Classification of Patches (ROIs)

Two strategies are used and compared in this paper for classification of local patches of satellite image into farm and non-farm. They are described in the following.

Hand-Crafted Features for Classification of Farm Patches.

First, the vegetation area is segmented using the NDMI image. Next, local patches are generated automatically, from the segmented green area. Then, textured areas including farms or any other pattern are classified by applying SVM on the extracted features using GLCM and 2-D DCT. Finally, the farm areas are detected by morphological analysis of the textured patches and SVM modelling. MATLAB 2018 was used for all implementations. Fig. 3 shows the block diagram of the analysis strategy.

Vegetation Segmentation.

There are two standard indexes for segmentation of vegetation green vegetation area. They are Normalized Difference Vegetation Index (NDVI) [45] and NDMI [46]. The pixels are segmented using spectral bands; the Near Infra-Red (NIR) in 851-879 nm range and Shortwave NIR (SWIR) in 1566-1651 nm range. However, NDMI [46] is a more suitable technique because it considers the moisture content of the soil and plants instead of the leaf chlorophyll content or leaf area. There are also similar works like [47], which have used NDMI and tasseled cap transformations on 30m resolution Landsat images for estimating soil moisture. Hence, the farm areas that went undetected by NDVI are well detected by thresholded NDMI strategy. NDMI uses two near-infrared bands

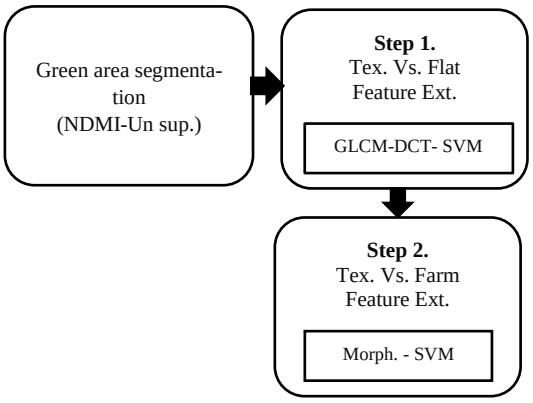


Fig. 2. Block diagram showing the overall classification process based on hand-crafted features.

(one channel of 1.24-μm that was never used previously for vegetation indices) to identify the soil moisture content. It is employed in forestry and agriculture applications [48]. This index has been used in this paper for the estimation of total vegetation including the agricultural lands and farms. For Lands imagery, NDMI is calculated as:

$$\mathrm{NDMI} = \frac{\mathrm{NIR}-\mathrm{SWIR}}{\mathrm{NIR}+\mathrm{SWIR}} \qquad (1)$$

NDMI is always in the range [-1, +1]. It is reported that NDMI values more than 0.10-0.20 indicate very wet or moist soil surfaces [46]. Then, based on this study, cultivable land is extracted for further classification.

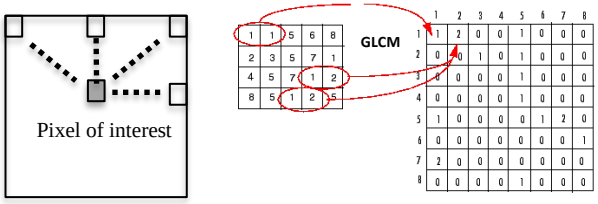
Texture Area Detection.

The detected green areas from the previous step are mapped on the RGB band images. Farm areas are part of the green areas of the image; therefore, the detected green areas are divided into small patches of 200×200 pixels. Then, feature extraction is performed for each patch of image to detect the textured patches. Patches with flat patterns do not include a farm area.

GLCM - One of the feature extraction techniques employed for texture areas is the GLCM that is widely used for texture analysis [49]. The GLCM studies the spatial correlation of the pixel grayscale and spatial relationship between the pixels separated by some distance in the image. It looks for regional consistency considering the extent and direction of grey level variation. Considering the characteristics of the flat regions and the textured regions (non-farm or farm) as shown in Fig. 4. GLCM is used for discrimination. Mathematically, the spatial relation of pixels in image matrix is quantified by computing how often different combinations of grey levels co-occur in the image or a section of the image. For example, how often a pixel with intensity or tone value i occurs either horizontally, vertically, or diagonally to another pixel at distance d with the value j (see Fig. 5-a). Depending on the range of intensities in an image, a number of scales are defined and a GLCM square matrix of the same dimensional size is formed. Then, image pixels are quantized based on the discrete scales and the GLCM matrix is filled for each direction. Fig. 5-b shows the formation process of a GLCM matrix based on horizontal occurrences at $d = 1$. The grayscales are between 1 to maximums 8 in this case.

(a) (b) (c)

Fig. 4. Examples of (a) Flat (b) Textured-Farm (c) Textured Non-Farm patches [40].



(a) (b)

Fig. 5. (a) Illustration of forming GLCM matrices in four directions i.e., $0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$. (b) Computation of GLCM matrix based on horizontal occurrences at $d = 1$ for an image [50].

Two order statistical parameters: Contrast, Correlation, Energy and Homogeneity samples are used to define texture features in the vegetation. Considering a grey co-occurrence matrix p , they are defined as:

$$\text{Contrast} = \sum_{i,j} |i - j|^2 p(i,j) \quad (2)$$

$$\text{Correlation} = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \quad (3)$$

$$\text{Energy} = \sum_{i,j} p(i,j)^2 \quad (4)$$

$$\text{Homogenity} = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \quad (5)$$

where, i, j denote row and column number, μ_i, μ_j , σ_i, σ_j are the means and standard deviations of p_x and p_y, so that, $p_x(i) = \sum_{j=0}^{G-1} p(i,j)$ and $p_y(j) = \sum_{i=0}^{G-1} p(i,j)$. G is the number of intensity scales, used for GLCM matrix formation.
Further detailed information can be found in [51]. The GLCM features are calculated in directions 0°, 45°, 90°, and 135° as shown in Figure 2-a. The calculated GLCM features in the four directions are averaged for each parameter and used as input to the classification model $\text{GLCM} = [\text{Cont}_{\text{av}}, \text{Corr}_{\text{av}}, \text{Eng}_{\text{av}}, \text{Hom}_{\text{av}}]$.

2D DCT - DCT sorts the spatial frequency of an image in ascending order and in the form of cosine coefficients. Most significant coefficients lie in the lower order, corresponding to the main components of the image, while the higher order coefficients correspond to high variation in images. Since the variation in a textured patch is higher than a flat one, the DCT map can help to distinguish them. For this aim, the original image patch I_{org} is transformed into DCT domain I_{DCT} and a hard threshold is applied to the DCT coefficients to remove the high order coefficients $I_{DCT(th)}$. Then, the inverse 2D-DCT of the thresholded image I_{iDCT} is computed. In both original and DCT domain, the reduction in the entropy of the textured patches is more significant than the flat areas representing smooth variations. Therefore, the ratio of coefficients' entropy before and after thresholding $[\frac{ent(I_{DCT})}{ent(I_{DCT(th)})}, \frac{ent(I_{org})}{ent(I_{iDCT})}]$ are calculated in both domains. For textured patches the entropy ratios are greater compared to flat patches due to the significant drop in entropy after thresholding the large amount of high frequency information.

Morphological Features

To recognize if a detected textured patch contains farm areas, first the patch image is converted to grayscale image. Then, the Sobel edge detection followed by morphological opening and closing by reconstruction are performed. This highlights the farm areas, keeping the boundaries and shapes in the image undisturbed. Next, the regional maxima were found to extract only the areas of maximum intensity (or the highlighted foreground regions). Further, the small stray blobs, disconnected or isolated pixels, and pixels having low contrast with the background in their neighborhood are discarded. This is because there is a contrast between the farm regions (marked as foreground) and

their surrounding boundary pixels. The same procedure is performed for a non-farm sample. The area of the foreground as well as the entropy for a patch including farm is higher compared to a non-farm due to the higher number of connected foreground pixels.

SVM Modelling

SVM classifiers are trained using the four GLCM and the two DCT features at step 1 and morphological features at step 2. The first model is capable to detect textured versus the flat patches and the second one detects the patches including farms from the textured patches with no farm areas. The LibSVM [52] is used. In this paper, the 5-fold cross-validation [53], is used to find the optimum kernel and the corresponding parameters. It helps to avoid over-fitting or under-fitting. The choice of kernel based on cross validation allows classifying data sets with both linear and non-linear behaviour. SVM was used for remote-sensing and hyperspectral image data analysis previously [54].

Transfer Learning Strategy for VGGNet16.

CNN is a popular classification method based on deep learning different levels of both spectral and special features using the stack of filter banks at several convolutional layers. However, training a CNN requires large data sets and heavy time-consuming computations and is prone to over-fitting using small data sets. A versatile approach in this case is transfer learning; The high-level deep features learnt over several layers of convolution, pooling and RELU using million images of massive ranges of scenes and objects are kept. That is based on the fact that the weighted combination of these activation maps of high-level features are the underlying building blocks of different objects of the scenes. While, the end layers called fully connected layers (FC) should be re-trained using hundreds of new training images. These layers are used to evaluate the strong correlation of the previous layers high-level features to particular classes of the task (in training images) and calculate the appropriate weights giving high probabilities for correct classifications. Fig. 6 shows the transfer learning concept.

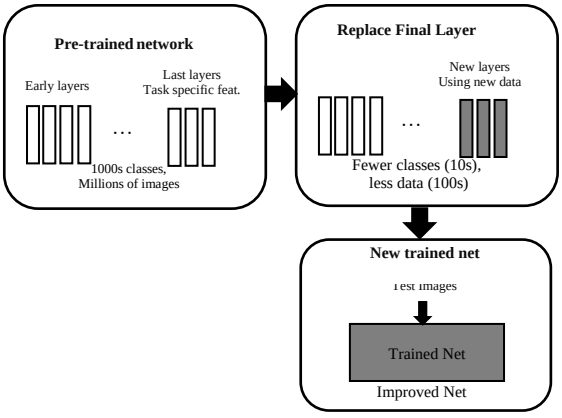


Fig. 6. Block diagram showing the transfer learning strategy [40].

The recent works on utilization of this technique [55, 56] shows suitability of transference of the learnt activation vectors for a new image classification task. Therefore, new patches of satellite images are used to retrain the final FC layers of VGG-16 CNN.

VGG-16 Network

The VGG-16 network is a pretrained network using more than a million images from the ImageNet database [57]. There are 16 deep layers and 1000 different classes of objects, e.g. keyboard, mouse, pencil, and many animals. This network has learned rich high-level feature representing wide ranges of objects. The size of input image is $224 \times 224 \times 3$ where the three color layers are RGB bands. The last three FC layers are trained for classification of 1000 classes. As explained, these three layers are retrained using our satellite image patches of the same size for farm classification while all other layers are kept.

3.2 Semantic Segmentation of Farm area using DeepLabv3+

As described in Introduction Section, after classifying the local patches, the pixels that include farm area are segmented. For this aim DeepLabv3+ model is used that utilizes an Encoder-Decoder architecture with atrous Convolution [39]. They are used in both DeepLabv3 and DeepLabv3+. They address two main challenges of semantic segmentation with deep CNN models, (1) the reduced feature resolution caused by consecutive pooling operations or convolution striding and (2) existence of objects at multiple scales [38].

The first challenge causes to learn increasingly abstract feature representations and invariance to local image transformation that makes issues in prediction tasks [38]. That is due to the loss of detailed spatial features that influences the prediction performance. To overcome this problem, atrous convolution also known as dilated convolution is used in both DeepLabv3 and DeepLabv3+ architecture. The resolution of extracted deep features can be controlled explicitly using atrous convolution (see Fig. 7). Given a two-dimensional image, for each location i on the output feature map y and a convolution filter w, atrous convolution is applied over the input feature map x according to the following equation:

$$y[i] = \sum_k x[i + rk]w[k]$$

where the atrous rate r determines the stride used to sample the input signal. If $r = 1$, it is the standard convolution.

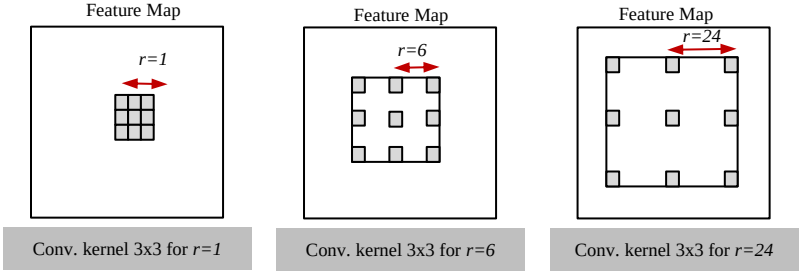


Fig. 7. Illustration of atrous convolution concept, with kernel size 3 × 3 and different rates. Standard convolution corresponds to atrous convolution with $r = 1$, while with higher atrous rates, the model's field-of-view enlarges and allows multi-scale feature extraction.

Using atrous also allows, adjusting the filter's field-of-view in order to capture multi-scale information which addresses the second challenge. Reviewing recent literatures shows that several methods have been proposed to address the issue with objects at multiple scales [58–61]. In DeepLabv3+, the spatial pyramid pooling is embedded into an encoder-decoder architecture as shown in Fig. 8. While the early layers include convolution and down-sampling operations (similar to Deep CNN), the down sampling operations are removed from the last few layers and instead, up-sampling of the corresponding filter kernels is performed and multiple parallel atrous convolutions are applied in different rates. This results in denser feature maps and capturing context at several ranges compared to Deep CNN (see Fig. 8).

As stated above, DeepLabv3+ utilizes an encoder-decoder structure. The encoder-decoder networks have been successfully used for different computer vision problems including semantic segmentation for example in [62, 63]. There are two main modules in encoder-decoder networks structure (1) an encoder module that gradually extracts semantic features and reduces the feature maps, and (2) a decoder module that gradually recovers the spatial information [39]. The encoder module that includes the spatial pyramid pooling has been described above. The last feature map in the top left side of Fig. 8 is the encoder output. The encoder features from DeepLabv3 [38] are usually computed with output stride = 16 and the features are then bilinearly up-sampled by a factor of 16. That is described as a naive decoding module and may not successfully recover object segmentation details [39]. Therefore, in DeepLabv3+, a simple yet effective decoder module is proposed as shown in Fig. 8 right side modules. Instead of up-sampling directly by a factor of 16, the encoder features are first bilinearly up-sampled by a factor of 4 and then concatenated with the corresponding low-level features from the left side encoder module that have the same spatial resolution. Then, few 3 × 3 convolutions followed by another simple bilinear up-sampling by a factor of 4 is performed. For further details, we refer to [39].

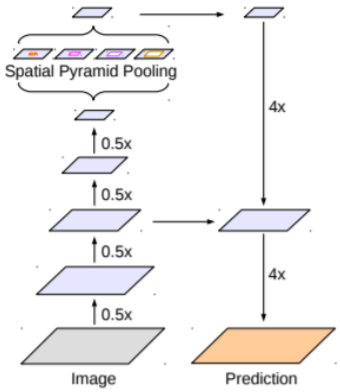


Fig. 8. The DeepLabv3+ encoder-decoder structure. The encoder module encodes multi-scale contextual information by applying atrous convolution at multiple scales, while the simple yet effective decoder module refines the segmentation results along object boundaries [39].

In this paper, four different pretrained networks, resnet18, resnet50, resnet101 and mobilenetv2, are used to transfer their learnt features into a DeepLabv3+ structure and train a new network for farm segmentation problem.

4 Experimental Results

In this section first the ROI classification results obtained from the both applied techniques, hand-crafted features and transfer learning using VGGNet16 will be presented. Then, the results of semantic segmentation of farm areas will be shown.

4.1 Hand-crafted Features and Classification Modelling Results

Fig. 9 shows the result of vegetation green area detection using NDMI. This image was further utilized for making patches (from green areas) that are used for the two-step classification framework.

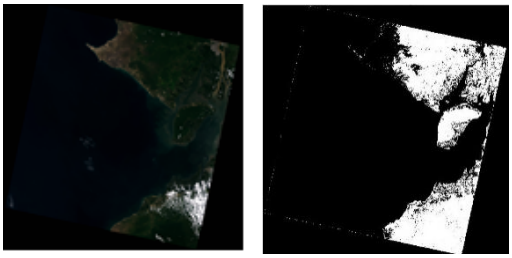


Fig. 9. (a) Landsat 8 image of Tendales, Ecuador (b) Result of thresholding using NDMI [40].

The number of training patches of both classes (textured verses flat and farm verses non- farm) were almost balanced at both feature extraction step and classification with SVM step. That is to avoid discriminative hyperplanes found by SVM that favores the more populated class. Totally from total patches, around 75% was used for training and the rest were kept as unseen data for test. In the first classification, 111 samples were used for training and 15 samples for test. In the second classification, there were 83 training samples and 22 test samples.
First, the four GLCM features and two DCT features were extracted from patches and combined. Fig. 10 visualizes the 2D DCT maps of a flat and textured patch before thresholding the higher frequencies coefficients and after thresholding. As can be seen, the textured patch has high energies in both low frequencies as well as high frequencies,

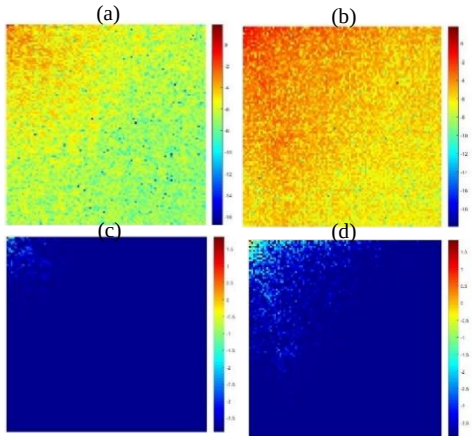


Fig. 10. DCT ap before thresholding (a) flat patch, (b) textured patch. After thresholding (c) flat patch (d) textured patch [40].

Table 1. GLCM and one of the DCT features used for classification of Flat and Textured Areas. (values shown are averaged over 20 samples) [40].

Class	Cont.	Eng.	Hom.	Ent.	DCT Ent. Ratio
Flat	0.0041	0.991	0.9979	3.014	0.1202
Tex.	0.067	0.847	0.9671	4.761	0.3337

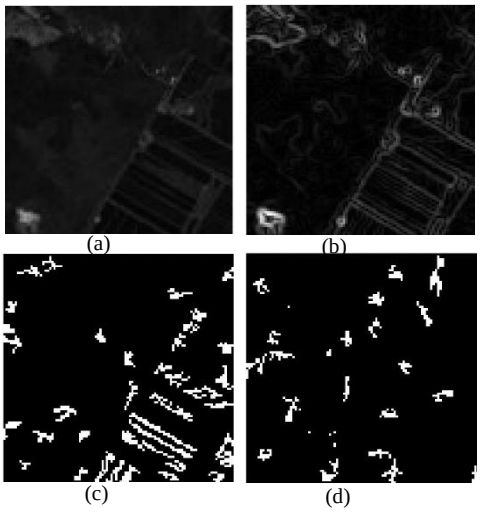


Fig. 11. (a) Grayscale image of a farm patch (b) Result of Sobel edge detection (c) Detected farm area by morphological foreground detection (d) Detected area of a textured non-farm patch shown in Fig. 4-c [40].

while in the flat patch DCT map, only low coefficients show high energy values. Therefore, the thresholded DCT map of the textured patch shows significant drop of energies in high frequencies. This influences the entropy ratios. Table 1 presents the average of the GLCM and DCT features over 20 patches for textured and flat classes. All the classified textured patches from this step were used to extract the morphology features at the second step, as shown in Fig.11.

The performance of classifiers is evaluated based on the number of correctly classified samples. Results are presented in Table 2. As can be seen, the first texture classification step is very robust. However, the performance is reduced for the second farm classifier based on morphology features.

Table 2. Accuracy results of the two-step hand-crafted features and classification modelling strategy for farm detection [40].

Classification Step	Train Accuracy (%)	Test Accuracy (%)
1	96.39 (107/111)	93.33 (14/15)
2	83.1325 (69/83)	81.8182 (18/22)

4.2 Transfer Learning Strategy Results

In order to retrain the three FC layers of VGG-16 net, hundreds of images are required. Then, further number of patches were used compared to the hand-crafted features and modelling strategy to fulfil the requirements of the second patch classification strategy. Transfer learning was performed using three different sets of more than 300 patches.

- The first set includes image patches from any general area of the satellite image, including ocean patches, mountains, residential areas, green flat and textured areas and farms. The last three FC layers of VGG-16 were retrained for the two-class farm detection problem.
- In the second set, the same number of patches were used excluding the non-green areas based on NDMI. This means the patches can include one of the flat green area, green textured non-farm area or a farm area.
- Finally, in the third set of the same size, only green textured non-farm patches as well as farm ones were used.

In all three cases, 75% of patches were used for training and the remaining was used as the test unseen data. There were 72 farm patches and the rest were non-farm in all three sets. Due to random selection, the number of patches of each class are different in the generated sets. The average and standard deviation of the results over 5 randomly generated train and test sets are reported in Table 3. As expected, no significant difference can be seen between the results of the three studies. That is, the high-level features acquired from the stack of filter banks include all those spectral, special, structural and color features extracted using the manual feature extraction strategy. Due to inclusive level of features extracted using the deep convolutional layers, the CNN results outperform the two-step feature extraction strategy.

Table 3. Average and standard deviation of the training and test accuracy of the CNN using transfer learning on the three different sets of patches [40].

Classification type	Train Accuracy (%)	Test Accuracy (%)
Farm vs. general areas	99.55 ± 0.64	96.76 ± 2.26
Farm vs. green areas	99.37 ± 0.76	95.95 ± 2.87
Farm vs. green tex. area	98.91 ± 0.52	96.76 ± 2.80

Fig. 12 shows the confusion matrix of one of the five test sets results using the transferred CNN models. The first experiment data set, that classifies farm patches from any general patch was used. As shown, only one general non-farm patch was misclassified as a farm patch.

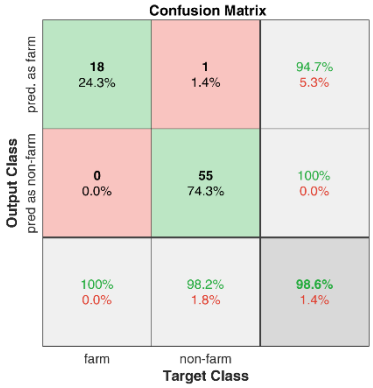


Fig. 12. The confusion matrix of one of the five test sets results from the first data set (classification of farm patches from any general patch) [40].

4.3 Semantic segmentation of farm regions results

In order to apply the semantic segmentation based on DeepLabv3+ the patch images pixels need to be labelled. That is due to the fact that it is supervised strategy and requires a label for every pixel of the image patch. For this aim, 72 local image patches that include farm areas in some parts were manually labelled. Totally seven different objects could be seen in the patch images and labelled accordingly. We refer to this data set as *Tendales_farm*. As we are only interested on farm area segmentation in this paper, all labels apart from farm are merged in this work and only two labels namely *farm* and *non-farm* are considered. Fig. 13 shows sample patches and the corresponding labels.

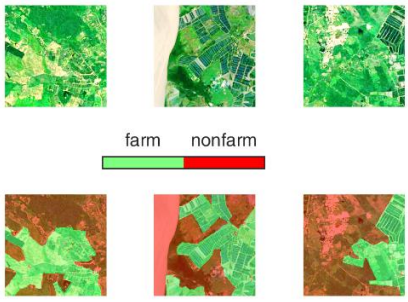


Fig. 13. Sample patches in Tendales_farm (top), the corresponding farm, non-farm labeled areas (down).

In order to do semantic segmentation, the data set is divided into training (70%), validation (15%) and test sets (15%). Then DeepLabv+ network is considered using the four different pretrained networks, resnet18, resnet50, resnet101 and mobilenetv2. The pixel classification layer was replaced based on farm classification problem classes and retrained using the training and validation images and their corresponding label sets. To compensate for class imbalance, the farm and non-farm classes weights are calculated. First the number of pixels in each class is calculated and divided by the total number of pixels, yielding 0.4029 and 0.5971 for the farm and non-farm classes. Then, the median of these frequencies is divided by the individual frequencies yielding the weights 1.2411, 0.8373 corresponding to the farm and non-farm classes. These weights are used in the cross entropy loss function that is used in the pixel classification layer. The Stochastic Gradient Descend with Momentum (SGDM) with piecewise learning rate was used for training. The number of epochs before the varied between 30 to 50 for the four models, and the training stopped afterwards due to no further improvements and to avoid overfitting.

In order to evaluate the performance of the models xx different metric factors are calculated. The first factor is accuracy. It is calculated for each class, based on the ratio of correctly classified pixels to the total number of pixels in that class, according to the ground truth. Given the number of True Positives (TP), False Positives (FP) and False Negatives (FN) as shown in Fig. 14, accuracy is calculated as follows:

$$Accuracy\ score\ = TP/(TP\ +\ FN) \qquad (6)$$

It indicates how well each class correctly identifies pixels. Besides that, the global accuracy is calculated which is the ratio of correctly classified pixels, to the total number of pixels regardless of their class. This metric is computationally less expensive compared to each class accuracies.

Another metric is Intersection Over Union (IoU), that is also called *Jaccard similarity coefficient*. It is a statistical accuracy measurement that penalizes false positives and is commonly used. For each class, IoU is the ratio of correctly classified pixels to the total number of ground truth and predicted pixels in that class. Then, an IoU equal to one shows a perfect segmentation while an IoU smaller than one shows an increase in FP or FN.

$$IoU\ score\ = TP/(TP\ +\ FP + FN) \quad (7)$$

The training and test results obtained for the four different models are presented in Table 4 and 5. The most successful models in terms of global and class accuracy as well as the IoU factor are resnet18 and resnet50.

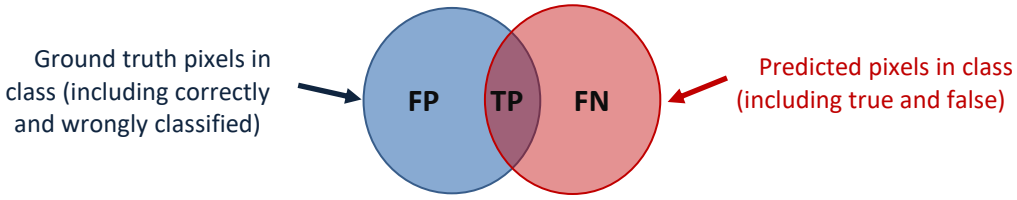


Fig. 14. Illustration of the relationship between TP, FP and FN.

Table 4. comparison of the four semantic segmentation models results on train set

Train set	Class Accuracy		Global Accuracy	IoU	
	farm	Non-farm		farm	Non-farm
resnet18	0.9411	0.8461	0.8854	0.7726	0.8123
resnet50	0.9413	0.8774	0.9038	0.8019	0.8425
resnet101	0.8942	0.8067	0.8429	0.7019	0.7507
mobilenetv2	0.8998	0.7972	0.8396	0.6989	0.7445

Table 5. comparison of the four semantic segmentation models results on test set

Test set	Class Accuracy		Global Accuracy	IoU	
	farm	Non-farm		farm	Non-farm
resnet18	0.8631	0.7905	0.82594	0.7077	0.6991
resnet50	0.8430	0.8145	0.8284	0.7059	0.7084
resnet101	0.8591	0.7040	0.7797	0.6557	0.6205
mobilenetv2	0.8520	0.6968	0.7726	0.6466	0.6106

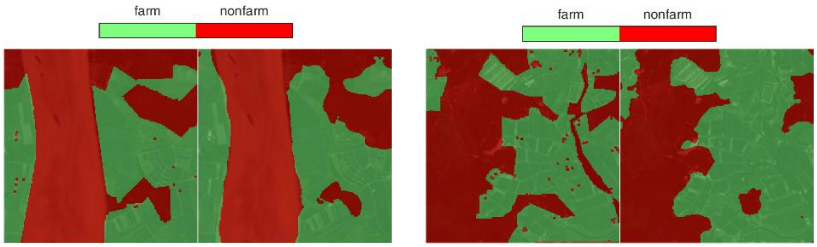


Fig. 15. From left to right, a ground truth labeled image for a training sample and the predicted labels, then a ground truth labeled image for a test sample and the predicted labels.

In Fig. 15, two training and test labelled images together with the prediction result is illustrated using the resnet50 transferred network. Some of the prominent misclassified nan-farm pixels as farms (FN) are the rivers and residential areas. The misclassified river can be seen in the test image. The rivers are very similar to the sharp edges connecting several farms and the tiny connected components as residential areas were segmented as farm areas. In addition, in the local patches, there are also parts of green areas that show some sharp corners similar to farms but different in color compared to the majority adjacent farms. They might be farms left uncultivated for some time and caused uncertainties in labeling stage and can also influence the accuracy of the models. That can be considered as one of the limitations low resolution images for appropriate labeling.

5 Conclusions

This paper is focused on farm detection using low resolution satellite images. The overall frame work consists of local patch or Region of Interest (ROI) classification followed by semantic segmentation of detected farm patches in order to find farm pixels. Two main patch classification strategies were employed in the first stage of the frame work; first a traditional hand-crafted feature extraction and modelling strategy was developed. In this method, unsupervised thresholding using Normalized Difference Moisture Index (NDMI) was used for green area detection. Then, a two-step algorithm was developed using Grey Level Co-occurrence Matrix (GLCM), 2D Discrete Cosine Transform (DCT) and morphological features as well as Support Vector Machine (SVM) modelling to discriminate the farms patches from other patches (non-textured or textured) that do not include any farm. The second patch classification strategy is based on deep high-level features learnt from the pre-trained Visual Geometry Group Network (VGG-16) networks. In order to use these features for farm classification, transfer learning strategies were employed. Then in the second stage of the framework, farm pixels were semantically segmented from the local patches. For this aim, the Tendales_farm data set was created by manual labelling of the images. The deepLabv+ semantic segmentation modelling strategy based on transfer learning was employed.

Four different pretrained networks, resnet18, resnet50, resnet101 and mobilenet together with labelled patches were used to retrain the networks. Experimental results showed that for the first stage of the framework, Convolutional Neural Networks (CNN) models are superior in terms of patch classification accuracy (99.55% and 96.76% for train and test respectively). For the second stage of the frame work, the resnet50 achieved the highest global accuracy for semantic segmentation (90.38% and 82.84% for train and test respectively).

References

1. Stephanie Van Weyenberg, Iver Thysen, Carina Madsen, J.V.: ICT-AGRI Country Report. (2010).
2. Schmedtmann, J., Campagnolo, M.L.: Reliable crop identification with satellite imagery in the context of Common Agriculture Policy subsidy control. Remote Sensing. (2015).
3. Leslie, C.R., Serbina, L.O., Miller, H.M.: Landsat and Agriculture — Case Studies on the Uses and Benefits of Landsat Imagery in Agricultural Monitoring and Production:U.S. Geological Survey Open-File Report. 27 (2017).
4. Vorobiova, N.S.: Crops identification by using satellite images and algorithm for calculating estimates. In: CEUR Workshop Proceedings. pp. 419–427 (2016).
5. Canty, M.J., Nielsen, A.A.: Visualization and unsupervised classification of changes in multispectral satellite imagery. International Journal of Remote Sensing. 27, 3961–3975 (2006).
6. Tian, J., Cui, S., Reinartz, P.: Building change detection based on satellite stereo imagery and digital surface models. IEEE Transactions on Geoscience and Remote Sensing. 52, 406–417 (2014).
7. Rembold, F., Atzberger, C., Savin, I., Rojas, O.: Using low resolution satellite imagery for yield prediction and yield anomaly detection. Remote Sensing. 5, 1704–1733 (2013). https://doi.org/10.3390/rs5041704.
8. Fisher, J.R.B., Acosta, E.A., Dennedy-Frank, P.J., Kroeger, T., Boucher, T.M.: Impact of satellite imagery spatial resolution on land use classification accuracy and modeled water quality. Remote Sensing in Ecology and Conservation. 4, 137–149 (2018).
9. Lee, L.W., Francisco, S.: Perceptual information processing system, U.S. Patent 10 618 543, (2004).
10. Hossain, M.D., Chen, D.: Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. ISPRS Journal of Photogrammetry and Remote Sensing. 150, 115–134 (2019).
11. Blaschke, T.: Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing. 65, 2–16 (2010).
12. Paola, J.D., Schowengerdt, R. a: The Effect of Neural-Network Structure on a Classification. Americal Society for Photogrammetry and Remote Sensing. 63, 535–544 (1997).
13. Hansen, M., Dubayah, R., Defries, R.: Classification trees: An alternative to traditional land cover classifiers. International Journal of Remote Sensing. (1996).
14. HARDIN. P.J: Parametric and nearest-neighbor methods for hybrid classification. a

comparison of pixel assignment accuracy. Photogramtnetric Engineering and Remote Sensing. 60. 60, 1439–1448 (1994).
15. Foody, G.M., Cox, D.P.: Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions. International Journal of Remote Sensing. (1994).
16. Ryherd, S., Woodcock, C.: Combining Spectral and Texture Data in the Segmentation of Remotely Sensed Images. Photogrammetric engineering and remote sensing. (1996).
17. Stuckens, J., Coppin, P.R., Bauer, M.E.: Integrating contextual information with per-pixel classification for improved land cover classification. Remote Sensing of Environment. (2000).
18. Lang, S.: Object-based image analysis for remote sensing applications: modeling reality - dealing with complexity. Springer (2008).
19. G. Mountrakis, J.I. and C.O.: Support vector machines in remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 66, 247–259 (2011).
20. Zhang, T.S. and S.: Local and global evaluation for remote sensing image segmentation. ISPRS J. Photogramm. Remote Sens. 130, 256–276 (2017).
21. Juniati, E., Arrofiqoh, E.N.: Comparison of pixel-based and object-based classification using parameters and non-parameters approach for the pattern consistency of multi scale landcover. In: ISPRS Archives. pp. 765–771. International Society for Photogrammetry and Remote Sensing (2017).
22. Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance, (2007).
23. Zhang, L., Yang, K.: Region-of-interest extraction based on frequency domain analysis and salient region detection for remote sensing image. IEEE Geoscience and Remote Sensing Letters. 11, 916–920 (2014).
24. Zhang, L., Li, A., Zhang, Z., Yang, K.: Global and local saliency analysis for the extraction of residential areas in high-spatial-resolution remote sensing image. IEEE Transactions on Geoscience and Remote Sensing. 54, 3750–3763 (2016).
25. Junwei Han, Dingwen Zhang, Gong Cheng, Lei Guo, and J.R.: Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning. IEEE Transactions on Geoscience and Remote Sensing. 53, 3325–3337 (2015).
26. Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q.: Classification for high resolution remote sensing imagery using a fully convolutional network. Remote Sensing. 9, 1–21 (2017). https://doi.org/10.3390/rs9050498.
27. Muhammad, U., Wang, W., Chattha, S.P., Ali, S.: Pre-trained VGGNet Architecture for Remote-Sensing Image Scene Classification. Proceedings - International Conference on Pattern Recognition. 2018–Augus, 1622–1627 (2018).
28. Albert, A., Kaur, J., Gonzalez, M.: Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. (2017).
29. Wu, M., Zhang, C., Liu, J., Zhou, L., Li, X.: Towards Accurate High Resolution Satellite Image Semantic Segmentation. IEEE Access. 7, 55609–55619 (2019). https://doi.org/10.1109/ACCESS.2019.2913442.
30. J. Long, E. Shelhamer, and T.D.: Fully convolutional networks for semantic segmentation. In: IEEE conference on Computer Vision and Pattern Recognition

(CVPR). pp. 3431–3440 (2015).
31. S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y.B.: The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: IEEE Computer Vision and Pattern Recognition Workshops. pp. 11–19 (2017).
32. L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A.L.Y.: Attention to scale: Scale-aware semantic image segmentation. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 3640–3649 (2016).
33. Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S.Y.: Object Semantic, region mining with adversarial erasing: A simple classification to segmentation approach. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 1568–1576 (2017).
34. Olaf Ronneberger, authorPhilipp, F.B.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Springer Lecture Notes in Computer Science (2015).
35. Volpi, M., Tuia, D.: Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks. IEEE Transactions on Geoscience and Remote Sensing. 55, 881–893 (2017). https://doi.org/10.1109/TGRS.2016.2616585.
36. Culurciello, A.C.: LinkNet: Exploiting encoder representations for efficient semantic segmentation. In: IEEE Visual Communications and Image Processing (VCIP) (2017).
37. He, K., Sun, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.
38. Adam, L.-C.C.G.P.F.S.H.: Rethinking Atrous Convolution for Semantic Image Segmentation. Arxive. (2017).
39. Chen, L., Zhu, Y., Papandreou, G., Schroff, F.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.
40. Sharifzadeh, S., Tata, J., Tan, B., Weyenberg, S. Van, Thysen, I.: Farm Detection Based on Deep Convolutional Neural Nets and Semi- Supervised Green Texture Detection using VIS-NIR Satellite Image important topic in digital agriculture domain. In: Data2019. pp. 100–108 (2019).
41. Jake Bouvrie , Tony Ezzat, and T.P.: Proceedings of International Conference On Acoustics, Speech and Signal Processing. In: Localized Spectro-Temporal Cepstral Analysis of Speech. pp. 4733–4736 (2008).
42. Sharifzadeh, S., Skytte, J.L., Clemmensen, L.H., Ersboll, B.K.: DCT-based characterization of milk products using diffuse reflectance images. In: 2013 18th International Conference on Digital Signal Processing, DSP 2013 (2013).
43. Sharifzadeh, S., Serrano, J., Carrabina, J.: Spectro-temporal analysis of speech for Spanish phoneme recognition. In: 2012 19th International Conference on Systems, Signals and Image Processing, IWSSIP 2012 (2012).
44. Landsat.usgs.gov. Landsat 8 | Landsat Missions, https://landsat.usgs.gov, last accessed 2018/05/17.
45. Ali, A.: Comparison of Strengths and Weaknesses of NDVI and Landscape-Ecological Mapping Techniques for Developing an Integrated Land Use Mapping Approach A case study of the Mekong delta , Vietnam Comparison of Strengths and Weaknesses of NDVI and Landscape-Ecolo, (2009).
46. Ji, L., Zhang, L., Wylie, B.K., Rover, J.: On the terminology of the spectral vegetation

index (NIR - SWIR)/(NIR+SWIR). International Journal of Remote Sensing. 32, 6901–6909 (2011).
47. Li, B., Ti, C., Zhao, Y., Yan, X.: Estimating soil moisture with Landsat data and its application in extracting the spatial distribution of winter flooded paddies. Remote Sensing. 8, (2016).
48. Gao, B.: NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment. 266, 257–266 (1996).
49. Tuceryan, M.: Moment based texture segmentation. In: Proceedings - International Conference on Pattern Recognition. pp. 45–48. Institute of Electrical and Electronics Engineers Inc. (1992).
50. MATLAB: Graycomatrix.
51. Haralick, R.M., Dinstein, I., Shanmugam, K.: Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics. (1973).
52. Chang, C., Lin, C.: LIBSVM : A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2, 1–39 (2011).
53. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer, New York (2009).
54. Petropoulos, G.P., Kalaitzidis, C., Prasad Vadrevu, K.: Support vector machines and object-based classification for obtaining land-use/cover cartography from Hyperion hyperspectral imagery. Computers and Geosciences. 41, 99–107 (2012).
55. Li, E., Xia, J., Du, P., Lin, C., Samat, A.: Integrating Multilayer Features of Convolutional Neural Networks for Remote Sensing Scene Classification. IEEE Transactions on Geoscience and Remote Sensing. (2017).
56. Chaib, S., Liu, H., Gu, Y., Yao, H.: Deep Feature Fusion for VHR Remote Sensing Scene Classification. IEEE Transactions on Geoscience and Remote Sensing. 55, 4775–4784 (2017).
57. Image Net, http://www.image-net.org/, last accessed 2019/01/12.
58. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L., Yuille: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv:1606.00915. (2016).
59. O. Ronneberger, P. Fischer, and T.B.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015).
60. S. Zheng, S. Jayasumana, B. Romera-Paredes, V.V., Z. Su, D. Du, C. Huang, and P.T.: Conditional random fields as recurrent neural networks. In: ICCV (2015).
61. H. Zhao, J. Shi, X. Qi, X. Wang, and J.J.: Pyramid scene parsing network. arXiv:1612.01105. (2016).
62. Fu, J., Liu, J., Wang, Y., Lu, H.: Stacked deconvolutional network for semantic segmentation. arXiv:1708.04943. (2017).
63. Zhang, Z., Zhang, X., Peng, C., Cheng, D., Sun, J.: Enhancing feature fusion for semantic segmentation. arXiv:1804.03821. (2018).