Lancaster Environment Centre | Lancaster University

MSc Data Science

Semantic Segmentation of Multispectral Satellite Images Using Deep Learning

LEC402 — Geoinformatics

35762970

Lancaster University

13th December 2023

1 Introduction

The increased availability of high resolution satellite imagery brings with it a data challenge to extract information. Multispectral sensors are able to capture image information across several wide bands, detailed in Table 2, across the electromagnetic spectrum, the collection of image information beyond that which the human eye can see is where significant exploits can be found within Infrared and hyperspectral imagery. Object detection in aerial imagery has seen significant growth in recent years from academics and industry alike, largely due to the decreased cost of aerial imagery and ability to process information with the motivation the unlocking of very detailed data with huge commercial and research implications. Aerial imagery is many scenarios, one being to document natural disasters and the progression of hurricanes, or the position of flooding. The possibility of accurately semantically segmenting these images into different types of objects could have huge implications for, not only the response efforts to these disasters, but also environment monitoring, urban planning and general mapping. The practical need for this is clear, however the environmental challenges this problem provides can be difficult to solve for experts, let alone image interpretation systems, which additionally offers some large computational hurdles.

This paper presents a solution to satellite feature detection from sparse data, provided by DSTL via kaggle *Dstl Satellite Imagery Feature Detection* 2021. First by working through the challenges of data sparsity, and computational problems then evaluating both supervised and unsupervised classification techniques and evaluating what these mean in practice. A summary of the work attempted is as follows:

1. Evaluating data merging techniques to inject more data on which to train the model (Pansharpening)
2. Evaluating semantic segmentation of Water and Vegetation
3. Adapting a Convolutional Neural Network for multispectral images and evaluating the output against other methodologies

2 Literature Review

Ding and Zhao 2018 attempted building extraction from satellite images using a mask based R-CNN leveraging building boundary regularisation which produced regularised polygons, which are beneficial in various applications, with high accuracy comparable to other state of the art methods with an F1 score of 0.717. It was tested in various locations such as Las Vegas, Paris and Shanghai to evaluate its performance in differing densities.

Perhaps the origin of U-Net architecture in satellite imagery applied it to the Massachusetts Roads and Buildings Dataset Hinton and Mnih 2013 to perform the binary segmentation of roads, while Iglovikov, Mushinskiy and Osin 2017 proposed the use of the U-Net architecture applied to satellite imagery feature detection, they briefly mention the use of semantic segmentation, which Xu Jiang 2021 does use, but does not discuss the strength of the algorithm, further more they exploit the use of reflectance indicies to capture further

information regarding specific classes and they go on to attempt a generate adversarial network to improve segmentation accuracy however they do not discuss the results of this.

Another approach uses a variation of U-Net called the Tiramisu Architecture to classify and segment objects using the SpaceNet multi-spectral dataset Howe, Casterline and Brown 2021. With others also attempting the Single Shot MultiBox Detector (SSD) method to segment and extract building footprints Chawda, Aghav and Udar 2018.

3 Methodology

3.1 Dataset

The dataset for this project is based upon the provided dataset by Kaggle for the DSTL Satellite Imagery Feature Detection (SIFD) competition *Dstl Satellite Imagery Feature Detection* 2021. The dataset is split into train and test, the train phase consists of 25 labelled images with the test phase consisting of 32, these are captured across 20 bands with varying channels and resolutions. The RGB-Band dataset is synthetically generated from the P-Band and M-Band using Pansharpening. The ground truth masks were generated using labelled polygonal shape data for 10 classes for each of the 25 images in the dataset, which includes buildings structures, roads, tracks, trees, crops, water-ways, standing waterbodies, trucks and cars. An example of the labelled data is shown in Figure 1, with a higher resolution example in Figure 2. An important note to add is the concealed true location of each satellite image, hence using further data such as LiDAR, or historic classification is unfortunately not possible.

Dataset Problems

The size of training set could be an issue, especially if there is any imbalance present, with only 25 labelled images there could be an information sparsity - exacerbated by the pixel depth giving too much information to reliably classify data. However obtaining these kinds of datasets requires massive amounts of manual work to label huge areas and assign the 10 class multi-polygons. The full test set relies on the Kaggle competition remaining open and hence obtaining the performance scores which additionally means using a validation set is not possible. An important note is the uncompressed size of the dataset is 90GB, the practicality of loading this amount of information into memory is low - further more by the fact that no access to a computer capable of storing that kind of dataset in RAM.

Figure 3 shows the percentage area for each class for each training image, where the total accumulation can be over 100% as classes are not mutually exclusive.

3.2 Multispectral Photography

Multispectral images can be leveraged to extract features beyond human vision, for example, the near infrared wavelength can be used to separate types of

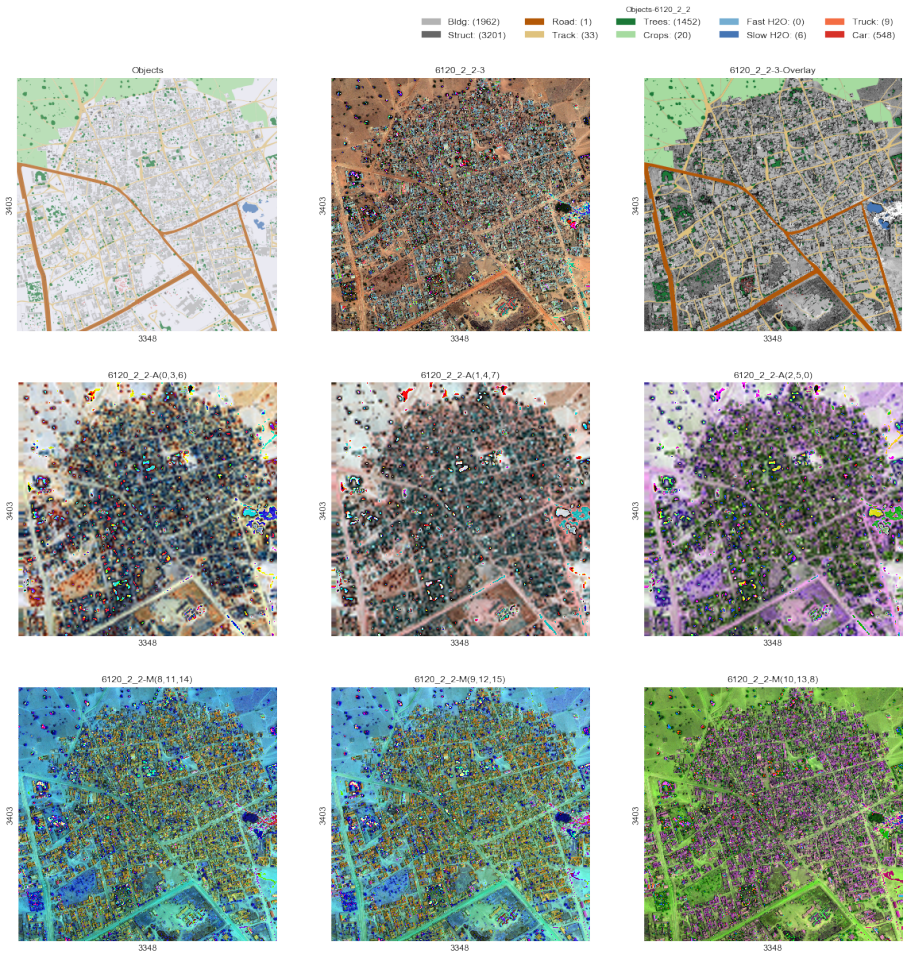

Figure 1: Example (6120_2_2) of training features with labels. Each image is different bands

vegetation due to the high reflection coefficients. Additionally, the colour depth on these WorldView-3 images is 11 and 14-bit, instead of the traditional 8-bit, this benefits a neural network as each pixel has the ability to store more information - conversely this can work against us too. The multispectral bands can be used to quickly detect specific classes of objects:

3.3 Vegetation Indices

Vegetation classification is commonly considered a well-solved problem, however due to the unpredictable location, shape and size of all vegetation - especially when compared to the detection of vehicles (which require a road, and are "vehicle" shaped, with the ability to distinguish between small and large) which have strict rules - and when a minor part of a larger solution, this motivates the leveraging of additional information contained within the multispectral bands. With the use of high resolution spectral instrumentation the number of

Figure 2: Example (6120_2_2) of training features. RGB bands

Name	Size	Description
Coastal Blue	400-450	Imaging shallow water and tracking fine particles
Blue	450-510	Penetration of water and differentiating rock with vegetation
Green	510-580	Separates ground from vegetation
Yellow	585-625	Separates ground from vegetation, also healthy vegetation
Red	630-690	
Red Edge	705-745	
NIR[1-2]	770-895	Seperates healthy vegetation, and for calculating NDVI
SWIR[1-8]	1195-2365	Differentiating wet and dry ground and types of rocks

Table 1: WorldView-3 Multispectral Band Wavelength and Descriptions (*Radiometric Use of WorldView-3 Imagery* 2021; *WorldView-3* 2021)

Band	Range (nm)	Resolution (m)	Dynamic Range (b/pix)
Panchromatic	450-800	0.31[0.34]	11
MS 8-Bands	400-1040	1.24[1.38]	11
SWIR 8-Bands	1195-2365	3.70[4.10]	14

Table 2: WorldView-3 Band Specifications *WorldView-3* 2021

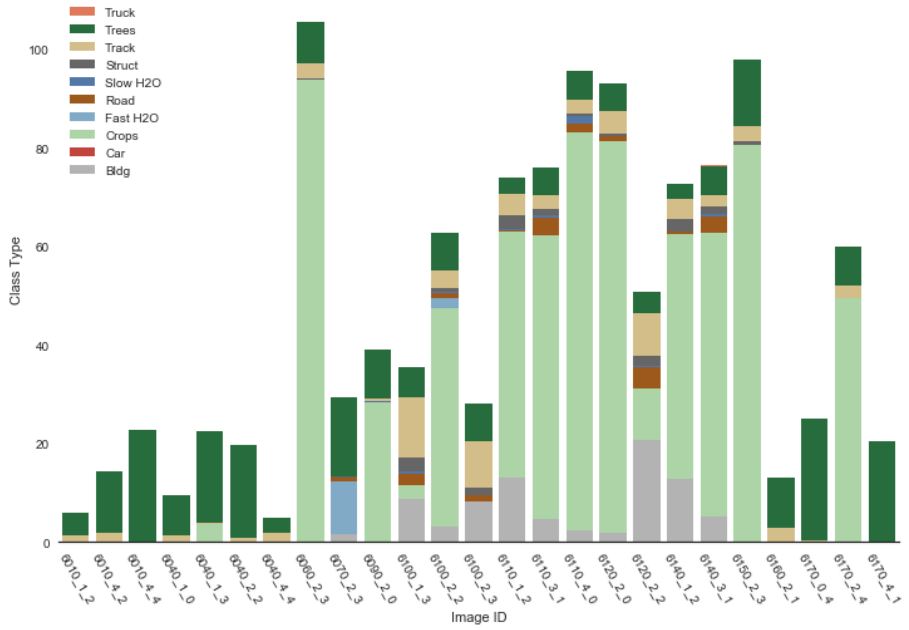

Figure 3: Breakdown of all classes for all training data (Total can be >100% due to non mutual exclusivity)

bands obtained by remote sensing is increasing and the bandwidth is becoming narrower Honkavaara et al. 2013; Xue and Su 2017. The most commonly known, used and implemented indicies calculated from multispectral information is the Normalised Difference Vegeration Index (NDVI) which is the normalised ratio between red and near infrared bands Karnieli et al. 2010. This can directly be applied to characterise canopy growth or vigor. A further index is the Normalised Difference Red Edge (NDRE) as it is a better marker of plant conditions for middle and late season crops that are post-maxima chlorophyll. These are combined to produce the Canopy Chlorophyll Content Index (CCCI), shown in Eq.1, which is proven to be a robust canopy control measure and so will work effectively in this scenario to attempt to quickly categorise vegetation.

$$\text{CCCI} = \frac{b_{\text{nir}} - b_{\text{edge}}}{b_{\text{nir}} + b_{\text{edge}}} \cdot \frac{b_{\text{nir}} + b_{\text{red}}}{b_{\text{nir}} - b_{\text{red}}} \tag{1}$$

As stated in the previous section, pixels store more information than solely colour due to the inclusion of infrared and other non-visible channels, in some circumstances this allows classification without context, which has big implications on computation time.

A further measure will be used to quickly classify water using deep learning, the Normalised Difference Water Index (NDWI), shown in Eq.2, is able to quickly segement water and so it does not need to be classified later.

$$\text{NDWI} = \frac{b_{\text{grn}} - b_{\text{nir}}}{b_{\text{grn}} + b_{\text{nir}}} \tag{2}$$

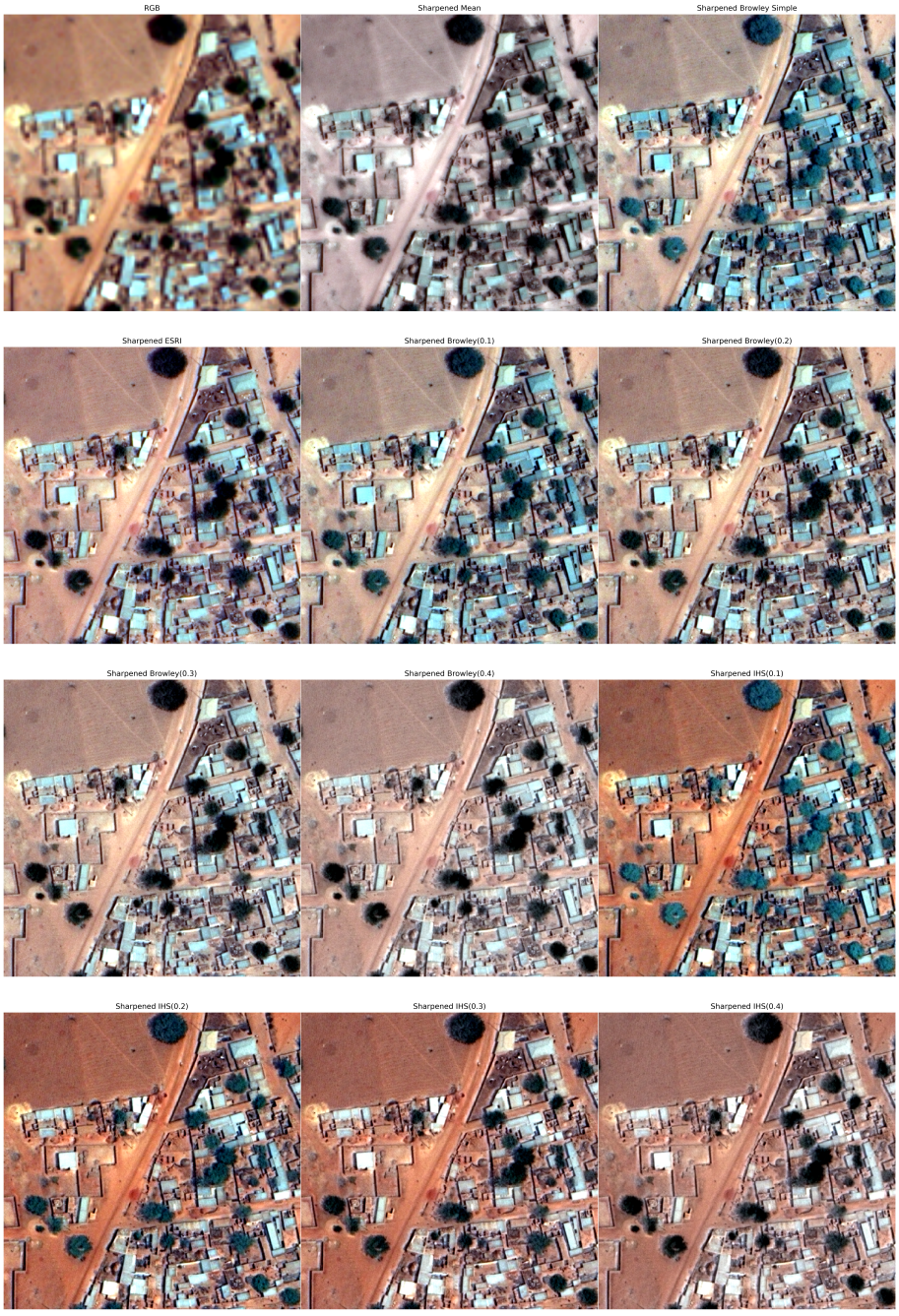

Figure 4: Pansharpening Examples. RGB shown Row 1 Left. Methods include Esri, Brownley, IHS

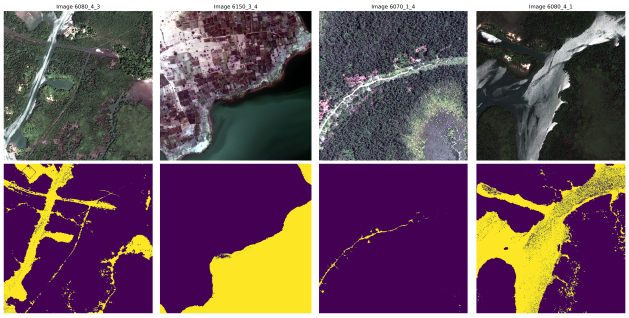

Figure 5: Watermask examples, demonstrating capabilities of fractured streams (left), clear coastlines (left middle), standard streams (right middle) and wide river systems (right)

Two somewhat difficult problems are the differentiation between Trees and Crops, and between Waterway and Standing Water. The Trees and Crops can be differentiated through further reflectance indicies, whereas due to Waterways being under-represented in the training data some modification of detection thresholds is required. Generally, bodies of water were detected very easily, with the boundaries well defined.

An example of the water masking is presented in Figure 5, it is able to quickly differentiate water from vegetation even in complex siutations, the far left hand figures show this, the middle left demonstrates simple borders with an area of confusion in the middle - perhaps a section of high ground beneath the water, the middle right demonstrates some problems with slim waterways where data is sparse and the far right shows some potential problems with thresholds or shallow water however this model is very adequate for segmentation of water.

3.4 Semantic Segmentation

Semantic Segmentation is an important concept, where the aim is to label each pixel with a corresponding class of what it is representing as opposed to image classification, in which an entire image is classified according to a label *Matlab Semantic Segmentation Guide* 2021. Additionally, segmentation differs from object detection in that it works at the pixel level to determine the contours of objects within an image, in the case of satellite imagery, these objects may be buildings, roads, cars, or trees. Applications of this type of aerial imagery labelling are widespread, from analysing traffic to monitoring environmental changes taking place due to global warming. An extension, not covered in this but could prove to be a great extension, is instance segmentation whereby each individual class object is separated into individual labels.

3.5 Convolutional Neural Network (CNN)

A convolutional neural network is a type of deep learning model for processing data, a thorough explanation of these is provided in Yamashita et al. 2018. These networks outperform other state of the art methods for visual recognition tasks,

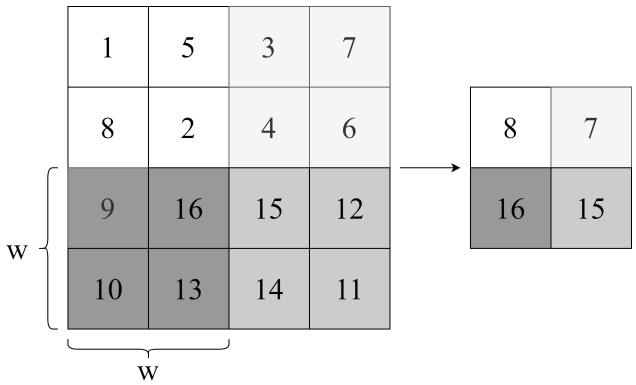

Figure 6: An example of pooling operation with a kernel size of 2x2, no padding using a maximisation function

e.g Girshick et al. 2013; Krizhevsky, Sutskever and Hinton 2012 by adaptively learning spatial hierarchies of features from low to high level patterns. They are mathematical constructs made up of layers; convolution, pooling and fully connected layers. Convolution and pooling layers are a type of feature extraction, where as the fully connected layer maps extracted features onto a final output, such as a classification task. A demonstration of a pooling layer is shown in Figure 6.

While these method have existed for a long time Ioffe and Szegedy 2015, their success usually directly correlates to the available size of training sets and the size of the network (in some applications the size of the problem scope).

3.6 U-Net Neural Network

There is consensus that the size of training data directly correlates to the performance of deep networks, the heavy reliance on thousands of annotated samples is often a heavy burden. This led to the inception of U-Net Ronneberger, Fischer and Brox 2015, where the main concept is to supplement a traditional contracting network by successive layers, see Figure 7, where pooling operations are replacing by upsampling operators hence increasing output resolution. Localisation is achieved by combining high resolution features from the contracting path with the upsampled output. Successive convolution layers can then learn to assemble an output on this information that is more precise.

The name U-Net comes from the shape of the diagram, which is a representation of an important modification allowing the network to propagate contextual information to higher resolution layers - the expansive path is mostly symmetric to the contracting path hence the U shape. Another motivation for the use of this model is the seamless segementation of arbitrarily large images by an overlap-tile strategy. Prediction of pixels in the border region leverages extrapolation by mirroring the input image this allows the method to not be limited by GPU memory. The input is a tensor of multispectral bands, RGB bands and reflection coefficient. Further more, using an exponential linear unit as the activation function, which decreases computational burden by reducing

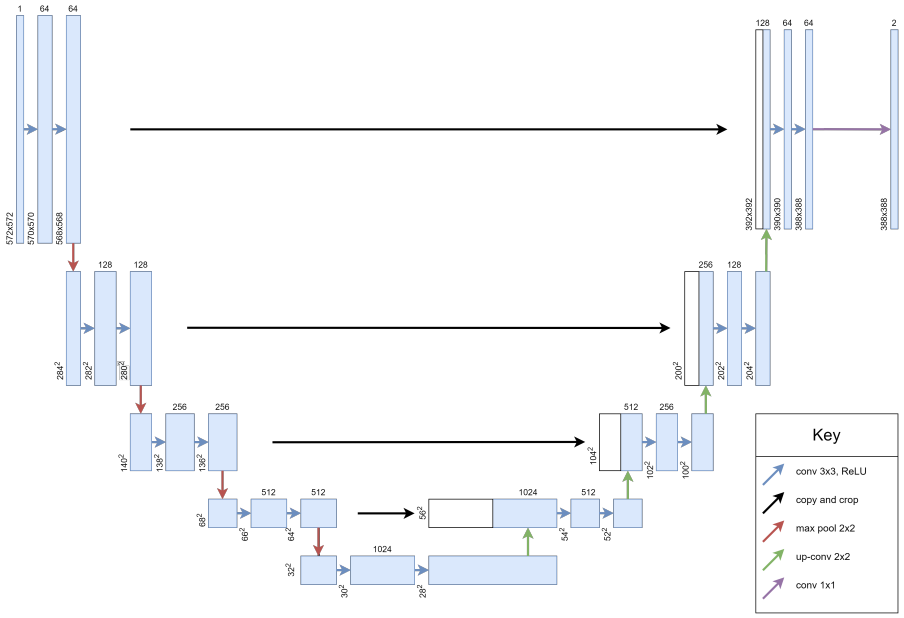

Figure 7: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box represents a multi-channel feature map. The number of channels is denoted on top of the box. The x-y size is the lower left edge. White boxes represent copied feature maps. The arrows are operations described in the key

noise sensitivity, and as mentioned previously, batch normalisation is used to accelerate convergence.

3.7 Evaluation Metrics

The proposed evaluation metric is the Jaccard Index which is effectively a similarity measure over a finite number of sets, this is defined as,

$$J(A,B) = \frac{|A \cap B|}{|\mathbb{A} \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|},$$

this can have problems with differentiability so this is approximated with the following,

$$J_m(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \hat{y}_i y_i + \hat{y}_i - y_i \cdot \hat{y}_i.$$

The loss function that is used is binary cross entropy as the classes are not necessarily mutually exclusive - whilst a road cannot be a vehicle semantically, in this case a group of pixels can be road and vehicle; this is defined as ,

$$H = -\frac{1}{n}\sum_{i=1}^{n} y \log \hat{y} + (1-y)\log(1-\hat{y}).$$

It logically follows that in order to get improved results, the objective function - that is a function of which to measure the subject of importance; and evaluation

metric must be as close to equivalent as possible. Combining the Jaccard approximation with the loss function gives the joint loss function:

$$L = H - \log J_m.$$

3.8 Optimisation Method

The optimisation procedure used is Adam with Nesterov momentum, Adam is used for its implementation simplicity and Nesterov momemtum gives convergence performance benefits. The network is trained for 50 epochs with a learning rate of $1e-4$, each epoch is trained on 400 batches, where a batch is 128 random images - created by randomly cropping 112 x 112 patches and applying a random transformation.

These models were developed and trained on AWS using a p2.xlarge instance, with K80 Tesla GPU and 61GB RAM - a requirement of this solution due to keeping large datasets in RAM for quick access.

4 Results

The testing phase brought some fruitious results, attaining an evaluation score of 0.402921 which would place it 15th on the Leaderboard of submissions *Dstl Satellite Imagery Feature Detection* 2021. An example of the predicted vs truth labels for the buildings class is presented in Figure 8, this model performs lower during validation and similar with the validation set by looking at the intersection over union, which is to be expected. The segmentation results for different classes are presented in Table 3, the notable results are the high detection of waterways which is performed with ease using the NDWI however this may be affecting the relatively poor Standing Water class as the distinction between them is nuanced from a 2D still image. The building class performs strongly, as shown in Figure 8, with most buildings detected and in some cases the model is able to correct some imprecise labels from training data where in some cases the training data shows shadows as part of buildings. The distinction of crops and trees is strong, perhaps the model gets confused with which class it should choose and some kind of probabilistic method could improve correct detection rates. Vehicle detection is very poor in both the train and testing phase which could be due to the small training sample combined with lower resolution images leading to training data sparsity.

5 Conclusions

The pansharpening technique used to merge high-resolution panchromatic images with lower resolution multispectral images to create a single high-resolution colour image was extremely successful in creating higher detailed images on which to train, a thorough comparison of the pansharpening methods would be required to evaluate which method is best for feature detection however the difference is likely negligible, or perhaps circumstantial. The use of vegetation indicies to quickly segment the training data, hence reducing the area of the images on which the CNN needs to be applied, gives positive results however they bring a significant number of false positives such as metal building

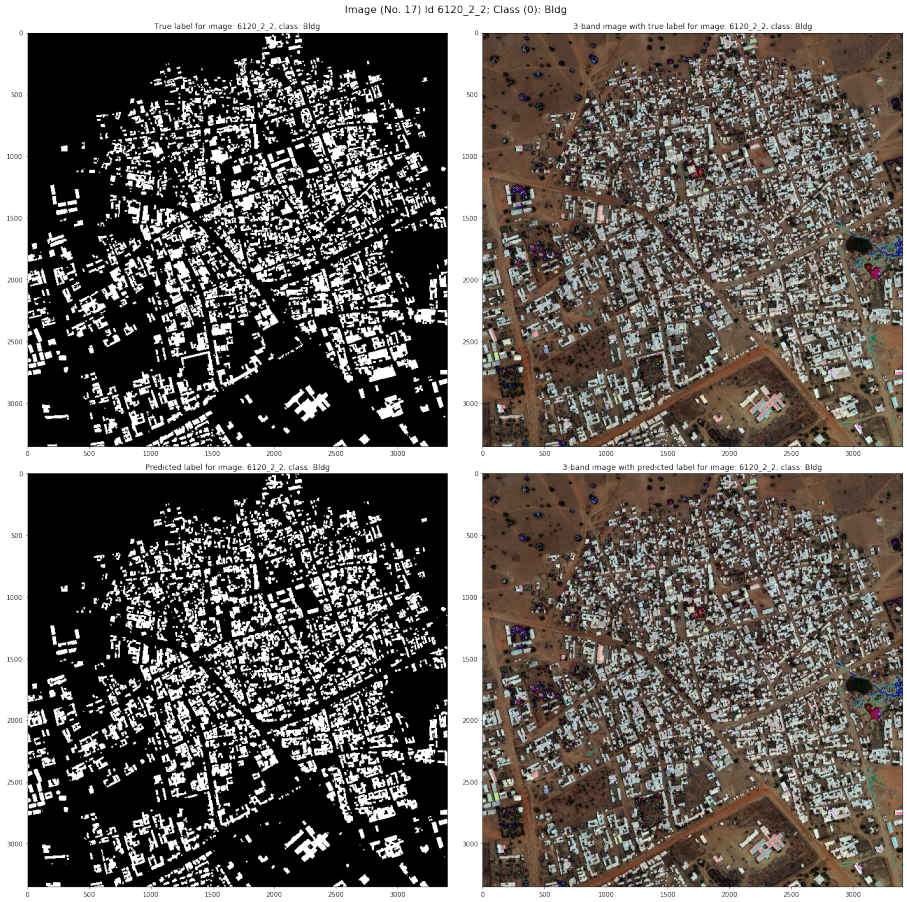

Figure 8: Comparison of true label and predicted label for Image 6120_2_2, True label top right, predicted label bottom right, 3-band image with true label top left and 3-band image predicted bottom left

Class	Train	Test
Buildings	0.7353	0.6290
Structures	0.1905	0.2015
Roads	0.8005	0.5605
Tracks	0.3281	0.3965
Trees	0.5018	0.6984
Crops	0.8251	0.8280
Waterway	0.9697	0.9131
Standing Water	0.6081	0.5272
Vehicle (Large)	0.2964	0.0331
Vehicle (Small)	0.0186	0.0000

Table 3: Jaccard Score for Train and Test set for each individual segmentation category

roofs (highly reflective surfaces) however with the very little computation required for these images it would require little effort to calculate several more and the potential use of multi-objective vegetation indicies could alleviate the false-positive issues. U-Net architecture, although originally designed for medical applications with data sparsity, has proven that it can perform strongly in any situation with training data sparsity which is important in satellite feature detection due to the huge effort required to adequately label images. For traditional CNN models, the final output is only as good as the passed training data, but not in the case of U-Net. The final results are strong compared to others that have attempted this problem, the high scores must be attributed to the pre-processing techniques applied where information is enhanced with the use of multispectral techniques. This solution is easily translatable to any WorldView3 images, and also any other satellite imagery semantic-segmentation task.

References

Chawda, C., Aghav, J. and Udar, S. (Sept. 2018). 'Extracting Building Footprints from Satellite Images using Convolutional Neural Networks'. In: pp. 572–577.

Ding, S. and Zhao, K. (Mar. 2018). 'Research on Daily Objects Detection Based on Deep Neural Network'. In: *IOP Conference Series: Materials Science and Engineering* vol. 322, p. 062024.

Dstl Satellite Imagery Feature Detection (2021). https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection. Online; accessed 29 February 2021.

Girshick, R. B. et al. (2013). 'Rich feature hierarchies for accurate object detection and semantic segmentation'. In: *CoRR* vol. abs/1311.2524. arXiv: 1311.2524.

Hinton, G. E. and Mnih, V. (2013). 'Machine learning for aerial image labeling'. In:

Honkavaara, E. et al. (Oct. 2013). 'Processing and Assessment of Spectrometric, Stereoscopic Imagery Collected Using a Lightweight UAV Spectral Camera for Precision Agriculture'. In: *Remote Sensing* vol. 5, pp. 5006–5039.

Howe, J., Casterline, M. and Brown, A. (2021). *Solving SpaceNet Road Detection Challenge With Deep Learning | NVIDIA Developer Blog*.

Iglovikov, V., Mushinskiy, S. and Osin, V. (2017). 'Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition'. In: *CoRR* vol. abs/1706.06169. arXiv: 1706.06169.

Ioffe, S. and Szegedy, C. (2015). 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift'. In: *CoRR* vol. abs/1502.03167. arXiv: 1502.03167.

Karnieli, A. et al. (Feb. 2010). 'Use of NDVI and Land Surface Temperature for Drought Assessment: Merits and Limitations'. In: *Journal of Climate* vol. 23, no. 3, pp. 618–633.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Advances in Neural Information Processing Systems 25*. Ed. by Pereira, F. et al. Curran Associates, Inc., pp. 1097–1105.

Matlab Semantic Segmentation Guide (2021). https://uk.mathworks.com/help/vision/ug/getting-started-with-semantic-segmentation-using-deeplearning.html. Online; accessed 24 March 2021.

Radiometric Use of WorldView-3 Imagery (2021). https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/207/Radiometric_Use_of_WorldView-3_v2.pdf. Online; accessed 5 May 2021.

Ronneberger, O., Fischer, P. and Brox, T. (2015). 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. In: *CoRR* vol. abs/1505.04597. arXiv: 1505.04597.

WorldView-3 (2021). https://earth.esa.int/eogateway/missions/worldview-3. Online; accessed 29 February 2021.

Xu Jiang, R. (2021). *Neural network for satellite image segmentation.*

Xue, J. and Su, B. (2017). 'Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications'. In: *Journal of Sensors* vol. 2017, pp. 1–17.

Yamashita, R. et al. (June 2018). 'Convolutional neural networks: an overview and application in radiology'. In: *Insights into Imaging* vol. 9, no. 4, pp. 611–629.

Word Count : 2493

Description (a brief description of what the data represents)	Created by (who generated the data origininally)	Sourced from (name or link to data provider)	Scale\ re	Format (e.g. .tiff, .shp. .gdb, .ascii)	Date (that the data relates to)	use restrictions or copyright info	Other notes
WKT format of all the training labe	DSTL	kaggle	N/A	.wkt	Various	no distribution	data fully anonymised
3-band satellite images. The three	DSTL	kaggle	1km x 1k	.zip (.tiff)			
16-band satellite images. The 16 l	DSTL	kaggle	1km x 1k	.zip (.tiff)			
the sizes of grids for all the image	DSTL	kaggle	N/A	.csv			
geojson format of all the training la	DSTL	kaggle	N/A	.zip (.geojson)			

Figure 9: Metadata List