

## FOREST COVER CLASSIFICATION USING GEOSPATIAL MULTIMODAL DATA

K. Suzuki \*, U. Rin, Y. Maeda, H. Takeda

Dept. of R&D, KOKUSAI KOGYO CO., LTD., 2-24-1 Harumi-cho, Fuchu-shi, Tokyo, 183-0057, JAPAN  
(kumiko\_suzuki, utei\_rin, yoshiko\_maeda, hiroshi\_takeda)@kk-grp.jp

Commission II, ICWG II/III

**KEY WORDS:** Forest Cover Classification, LiDAR, Airborne Imagery, Convolutional Neural Network, Multimodal Learning

### ABSTRACT:

To address climate change, accurate and automated forest cover monitoring is crucial. In this study, we propose a Convolutional Neural Network (CNN) which mimics professional interpreters' manual techniques. Using simultaneously acquired airborne images and LiDAR data, we attempt to reproduce the 3D knowledge of tree shape, which interpreters potentially make use of. Geospatial features which support interpretation are also used as inputs to the CNN. Inspired by the interpreters' techniques, we propose a unified approach that integrates these datasets in a shallow layer in the CNN network. With the proposed CNN, we show that the multi-modal CNN works robustly, which gets more than 80 % user's accuracy. We also show that the 3D multi-modal approach is especially suited for deciduous trees thanks to the ability of capturing 3D shapes.

### 1. INTRODUCTION

The Paris Agreement, adopted at the COP21 in 2015, set out a global action plan to reduce greenhouse-gas emissions, which not only puts the world on track to avoid dangerous climate change but also accelerates the Carbon Disclosure Project (CDP). CDP requests companies and cities to disclose the status of environmental actions against climate change. Under these circumstances, remote sensing, which enables us to observe the planetary surface, is expected to monitor the forest owners' effort such as sustainable forest management (e.g. organized logging, planting and thinning). To meet the purpose of carbon disclosure, not only monitoring but also frequent and low-cost monitoring is required. Since both these features would be difficult to achieve through manual work, it is urgent to establish an automated forest monitoring method.p

Currently, there already exists automated forest monitoring systems. Global Forest Watch (World Resources Institutes, 2014), a dynamic online forest monitoring and alert system, automatically produces annualized global tree cover change data based on Landsat satellite imagery. Global Forest/Non-forest Maps (Shimada et al., 2014) also show the forest cover with certain thresholds. Due to medium-resolution images and limited number of classes, neither of the two systems is suitable to monitor the forest management. Aiming for specific targets, a number of different methods have been developed for different types of forests using various remote sensing data, whereas forest cover classification using high-resolution data remains challenging.

To tackle the accurate forest cover classification, we propose a CNN (convolutional neural network) approach which is inspired by professional interpreters. Professional interpreters produce official forest maps by interpreting a forest from airborne images or satellite images. Interpreting requires the knowledge of forestry and in some cases geospatial features as well as RGB images. Inspired by that, we employ geospatial features in the proposed CNN. Following the interpreters'

techniques, where they not only consider the surface of forests but also recall the inside of forests, we propose to feed a 3D-voxel data originated from LiDAR (Light Detection and Ranging) to the CNN. To combine the geospatial input data, we propose a multi-modal CNN, where the input data is integrated in a shallow layer, which is a closer layer to the input than output in the CNN.

### 2. RELATED WORK

In the following, we review recent advances in remote sensing tasks with CNNs. Driven by powerful deep neural networks (Krizhevsky et al., 2012), remote sensing tasks, especially in land cover classification have started to make great progress. Using UC Merced Land Use Dataset (UCM) introduced by (Yang et al., 2010) which provides 21 land cover classes with 100 images each, (Penatti et al., 2015) shows that their CNN obtained 99.5% class accuracy. (Nogueira et al., 2017) points out that features of fine-tuned networks tend to perform well on UCM through comparing popular CNN algorithms. CNNs are thus reported to perform quite well in the remote sensing field.

Regarding forest cover classification, there is no public benchmark dataset yet. Researchers, therefore, explore the algorithms on their own datasets. (Lu et al., 2017) proposes a spatial-temporal-spectral data fusion framework over publicly available low-middle resolution images, leading to around 80% classification accuracy on seven-class classification task using support vector machine (SVM). (Kussul et al., 2017) reports around 85% class accuracy over 11 land cover and crop type classification using Landsat-8 and Sentinel-1A images. Thus, forest cover classification is a challenging task compared to land cover classification due to the similarity among classes.

Deep learning architectures have been developed for LiDAR datasets as well given it avoids feature engineering phase where discriminating features are designed as is common in traditional classification algorithms. Point clouds classification algorithms are especially discussed for daily scene such as Voxnet

\* Corresponding author

(Maturana et al., 2015) and PointNet (Qi et al., 2017). Although point clouds driven by airborne LiDAR is different from daily scene point clouds in a sense that it has tons of objects inside the dataset and not fine, several algorithms improves the result. (Yang et al., 2017) utilizes CNNs to transform from points to images and improves the urban objects classification on ISPRS 3D semantic labeling contest (Niemeyer et al., 2014). (Boulch et al., 2017) utilizes 2D CNNs to classify urban city point clouds: semantic3d.net (Hackel et al., 2016) and shows the efficient labelling algorithm. On the other hand, most of algorithms are aimed for urban objects classification and not for tree species classification.

As such, most of previous studies focus on images themselves or on standard indices such as NDVI (Normalized Difference Vegetation Index) to feed the classifier. However, LiDAR data can picture the characteristics inside the forest while remote sensing images only depict the surface of forests. As (Görgens et al., 2016) shows, some studies utilize airborne LiDAR data to classify forest cover yet images and LiDAR are not fed to the classifier simultaneously to extract information of both the surface and inside of forests.

### 3. METHODOLOGY

The basic principle of our methodology follows the traditional interpreters' techniques. As mentioned by (Ng, 2012), CNN itself is biologically inspired from human beings' brain system, meaning that the brain network is composed of neurons which, to interpret what we see from the sight, extract edges from pixels, primitive shapes from the edges, and object models from the several shapes. Accordingly, as (Russakovsky et al., 2015) shows, state-of-the-art CNN algorithms perform well for a general image such as a photograph taken at a close range. For forest classification from remote sensing images, on the other hand, not only picturing the shape of objects from a bird-eye view but also collecting geospatial information and capturing 3D shapes are required to interpret. CNN with just remote sensing images can thus be not sufficient to reproduce the quality of professionals' forest cover interpretation. Based on the assumption, in the following, we analyse how interpreters identify the tree species, and describe how the proposed CNN incorporates interpreters' techniques.

#### 3.1 Knowledge to Interpret Forestry

As we daily recognize objects in our sight, forest interpreters also utilize characteristics of images such as shape and colour as a key to classify forests. They, at the same time, recall how forest should appear from above and identify the tree species based on their knowledge of forestry. For instance, the difference between Hinoki (*Chamaecyparis obtusa*) and Sugi (*Cryptomeria japonica*) appeared in remote sensing images is not always obvious depending on the season or location. However, interpreters can differentiate them relatively easily based on their knowledge. They infer that Hinoki and Sugi are likely to have a different shape of crowns given that their tree shape is different as shown in Figure 1. On top of that, although interpreters have no information about the inside of the forest, they know that Sugi avoids touching other trees while Hinoki grows mixed with others, leading the tree crown shape of Sugi to tend to be clear and the one of Hinoki to be vague. Thus, interpreters utilize their knowledge of forestry and compensate the lack of information (i.e. the information about the inside of the forest) to classify the forest. We propose, in an

attempt to mimic the professional interpreters' strategy, an approach to feed LiDAR data as 3D information as well as remote sensing images to our CNN.

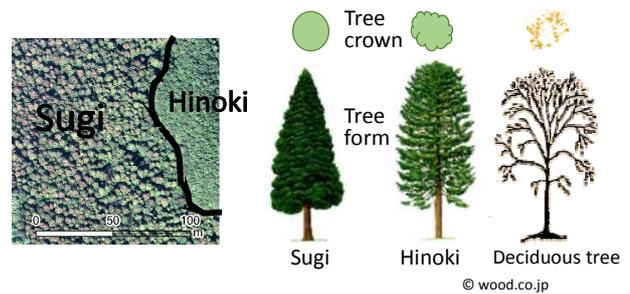


Figure 1. Knowledge of forestry for interpretation

LiDAR transmits a light pulse and records the time that the pulse returns, which creates a 3D point cloud of targets. Since typical convolutional neural networks require a regular shape for input data such as a collection of images, deep learning architectures with 3D voxels have been explored (Maturana et al., 2015). Although (Qi et al., 2017) shows that raw point clouds perform better than voxelization architecture on classification and segmentation by selecting informative points through the network, they assume point clouds taken from a close range such as CAD model, Kinect data, and structure from motion in proximity photographing as the input dataset. The point density of airborne LiDAR taken over a forest area, on the other hand, is generally around 4 points/m<sup>2</sup> in Japan which is different in the sense that the shape derived from point clouds is not necessarily clear. We thus use a basic voxel format as input.

#### 3.2 Geospatial Features

Since it is not practical for interpreters to check raw LiDAR data while examining images, geospatial features extracted from LiDAR have been developed to facilitate LiDAR data. As interpreters identify the tree species by attaching the observation on the geospatial features to their knowledge, we feed the geospatial features listed below to our CNN.

**Topographic openness:** the topographic openness (Yokoyama et al., 1999) is normally calculated from *digital terrain model* (DTM), whose value indicates the dominance or enclosure of a certain place, and the underground openness similarly indicates how much underground space is spreading. We compute them using *digital surface model* (DSM), which emphasizes the shape of tree crowns and supports to differentiate Sugi and Hinoki as shown in Figure 2.

**Slope:** slope of the ground is beneficial to classify trees which prefer a steep slope. For instance, Sugi tends to grow in the trough and Hinoki mainly grows in the ridge.

**Aspect:** direction of a downhill slope can imply how forests grow.

**Tree height:** tree height can imply the age of trees, and the appearance of trees varies according to the age.

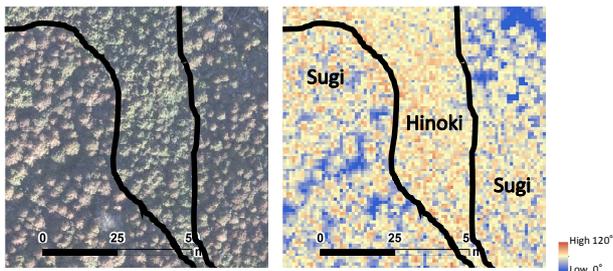


Figure 2. Topographic openness extracted from DSM

### 3.3 Multimodal Learning for Geospatial Datasets

With above analysis, we showed that interpreters empirically utilize features extracted from images, LiDAR which we consider is equivalent to our 3D reconstruction, and geospatial features properly. Accordingly, our classification algorithm, which mimics interpreters' strategy, needs to be capable of handling different types of data (i.e. airborne image and LiDAR). CNN can handle different types of input data through what is called multi-modal learning. As human beings process information from five senses and unify them to understand circumstances, multi-modal learning handles different modals for a given task such as video scene understanding using visual and audio signal. The main concept of multi-modal learning is to extract abstract and common information from each modal, which is in the same representation domain, and unify them to process information (Li et al., 2016). While handling geospatial datasets, on the other hand, interpreters associate data by overlaying over GIS since each data share its location. As such, we employ data fusion within a shallow layer in the CNN architecture where the layer is closer to the input and the spatial information has not been lost yet.

Figure 3 illustrates the base architecture of our CNN, where we feed a patch of images around a targeted LiDAR point and get class predicted by the CNN. The main architecture is inspired by AlexNet (Krizhevsky et al., 2012). AlexNet, which is a basic structure of deep learning with a small number of layers, suits this case since our input data size is small so that it is difficult to apply the deeper network such as ResNet (He et al., 2015). In case of voxels, 3D-CNN is additionally executed (i.e. the 3 dimension convolution over height, width and band) to extract features along z-axis, leading to dimension reduction along z-axis as well.

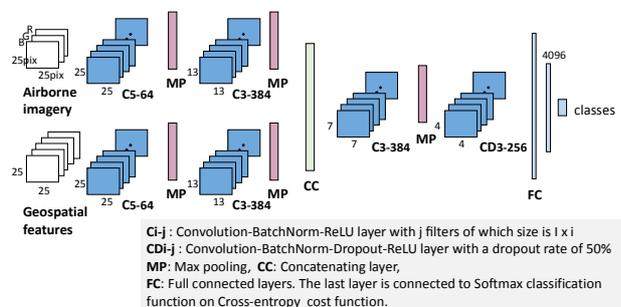


Figure.3 Base architecture of multi-modal CNN

## 4. EXPERIMENTS

We evaluate our proposed method in this section on the dataset of Japanese planted forest from two perspectives: (1) the contribution from each modality (i.e. airborne images, LiDAR voxels, and geospatial features) to the forest cover classification, and (2) the effects of 3D information (i.e. 3D-CNN over voxel grids) driven from LiDAR.

### 4.1 Dataset

Since there are no publicly available datasets for forest cover classification with high resolution images, we at first create the labelled dataset.

#### 4.1.1 Data Acquisition

The dataset is acquired over the forest in Tenryu area, a traditionally famous planted forest located in Shizuoka Prefecture, middle of Japan as shown in black lines of Figure 4. As a typical Japanese planted forest, there is a cycle of planting, growing, thinning, and logging, so that monitoring is crucial to evaluate the forest management. Given an area size of around 2.5km<sup>2</sup>, we adopt airborne measurements to acquire RGB images and LiDAR data as listed in Table 1.

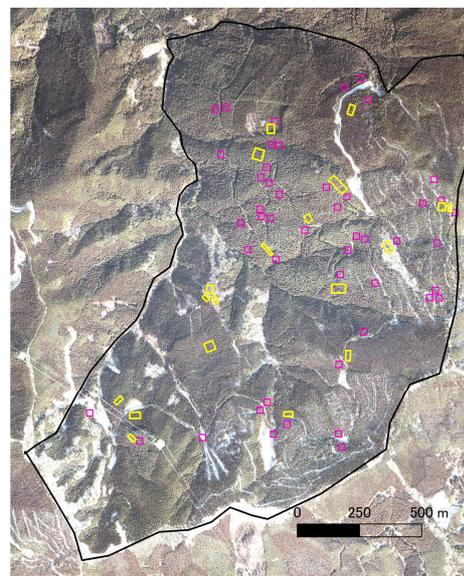


Figure. 4 Acquired dataset for the experiments

platform	airborne
date	13th March 2015
flying height	4500 ft
<b>Aerial Imageries</b>	
band	red, green, blue
data type	unsigned integer 8
ground resolution	20cm
<b>LiDAR</b>	
FOV	20 degrees
density	3.37 points/m <sup>2</sup>

Table. 1 Data acquisition

The label is allocated by a field survey. The yellow areas in Figure 4 are used for training and pink rectangles for testing, where we successfully conducted a field survey. The classes are determined by the basic inventory as shown in Figure 5: Sugi,

Hinoki, Deciduous trees, and Others. The number of samples is Sugi: 2500, Hinoki: 2475, and deciduous tree: 2499 respectively. Note that class Others is not trained in the experiments since we can filter them just by DTM.

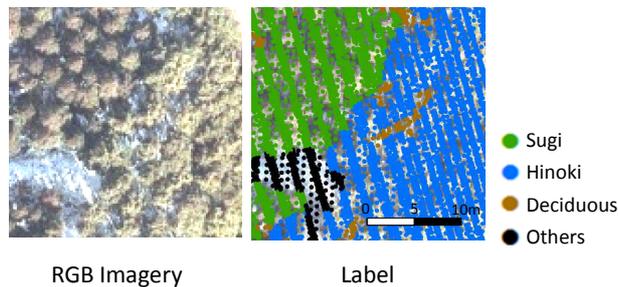


Figure 5 Labels (ground truth)

#### 4.1.2 Data Pre-processing

As Figure 6 describes, we extract geospatial features and voxels from LiDAR data. Geospatial features are composed of five elements: tree height, aspect, slope and underground/ground openness. Tree height is the subtraction of DSM and DTM, whose range is between 0m to 50m. Aspect and slope are computed by Esri ArcGIS Tools. Aspect is the orientation which shows the largest slope angle within 3 x 3 pixels, ranges from 0 to 360 degrees where 0 degree means the north and 180 means the south. Slope is the maximum gradient of the elevation within 3 x 3 pixels, which ranges from 0 to 90 degrees. Topographic openness is calculated as the average of maximum angle of elevation for each orientation within the search range. An angle of elevation is replaced by an angle of depression in underground openness. The range is between 0 degree and 180 degrees. As for voxels, we create 50cm grids given the LiDAR point density, and aggregate the number of points inside each grid, which is in the end transformed to 25 x 25 x 50 voxels around each LiDAR point through resampling. Note that the input data fed in the CNN including images is cropped around a target LiDAR point since the proposed network is for classification. The cropping size is set to 5m which is likely to be equivalent to a crown size.

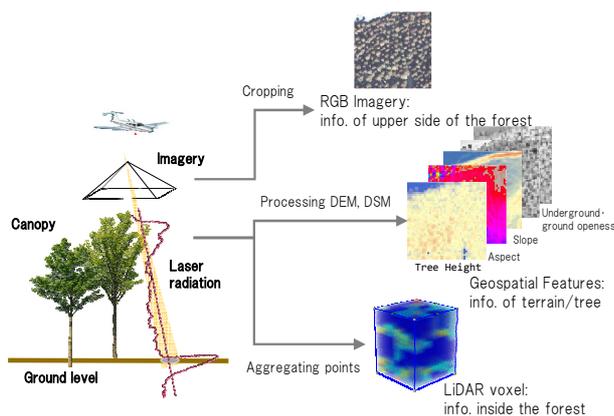


Figure 6 Data pre-processing flow

(a) Single-modal: images

		Reference			User's Accuracy
		Sugi	Hinoki	Dec.	
Prediction	Sugi	37081	5133	810	86.2%
	Hinoki	17588	75081	1171	80.0%
	Dec.	439	297	2856	79.5%
Producer's Accuracy		67.3%	93.3%	59.0%	81.9%

(b) Single-modal: geospatial features

		Reference			User's Accuracy
		Sugi	Hinoki	Dec.	
Prediction	Sugi	32601	9338	799	76.3%
	Hinoki	16290	52628	939	75.3%
	Dec.	6217	18545	3099	11.1%
Producer's Accuracy		59.2%	65.4%	64.1%	62.9%

(c) Single-modal: voxels

		Reference			User's Accuracy
		Sugi	Hinoki	Dec.	
Prediction	Sugi	31223	22580	1014	57.0%
	Hinoki	20437	53359	952	71.4%
	Dec.	3387	2627	2592	30.1%
Producer's Accuracy		56.7%	67.9%	56.9%	63.1%

(d) Multi-modal: images + voxels

		Reference			User's Accuracy
		Sugi	Hinoki	Dec.	
Prediction	Sugi	39688	6652	1048	83.8%
	Hinoki	15014	71771	871	81.9%
	Dec.	345	143	2639	84.4%
Producer's Accuracy		72.1%	91.4%	57.9%	82.6%

(e) Multi-modal: images + geospatial features + voxels

		Reference			User's Accuracy
		Sugi	Hinoki	Dec.	
Prediction	Sugi	38277	7682	1091	81.4%
	Hinoki	16369	70779	913	80.4%
	Dec.	401	105	2554	83.5%
Producer's Accuracy		69.5%	90.1%	56.0%	80.8%

(f) 3D Multi-modal: images + voxels

		Reference			User's Accuracy
		Sugi	Hinoki	Dec.	
Prediction	Sugi	34800	15810	1079	67.3%
	Hinoki	17612	61394	301	77.4%
	Dec.	2635	1362	3178	44.3%
Producer's Accuracy		63.2%	78.1%	69.7%	71.9%

'Single-modal' denotes a CNN with 2D convolution on single modality.  
 'Multi-modal' denotes a CNN with 2D convolution on multiple modalities.  
 '3D Multi-modal' denotes a CNN with 3D convolution on multiple modalities.

Table 2. Results

#### 4.1.3 Data Transformation

As CNNs require highly regulated input data formats (e.g. images with 3 bands whose pixels are corresponding over the bands), the dataset acquired above need to be transformed while keeping geospatial information. The geospatial features, therefore, are transformed to 20cm ground resolution which is the same size as the imagery, the highest ground resolution. Voxel grids are also transformed to 20cm × 20cm grids. The value range is normalized from 0 to 1 for all features.

#### 4.2 Implementation Detail

Our implementation is based on the public platform Chainer (Tokui et al., 2015). We use the Adam optimization with base learning rate of 0.001 as a basic method, which adaptively arrange the learning rate and known to be converged relatively fast. The epoch number is set to 200 at a maximum, where we check the convergence of test data every experiment. The mini-batch size is 126.

#### 4.3 Results

To evaluate our methodology, we conduct experiments with several settings as listed in Table 2, including single-modal or multi-modal, and 2D or 3D convolution on multi-modal method. For evaluation, class accuracy is used.

As for the contribution from each modality, the results with the single-modal CNN show that images contribute a lot to Sugi and Hinoki classification. Geospatial features work well for deciduous trees, which indicates images are not capable of identifying bared deciduous trees. As for the multi-modal CNN, the performance of classifying Sugi is improved when compared to each single modality, and the result of classifying Hinoki keeps as good accuracy as obtained in the single-modal CNN with images. The reason that the multi-modal CNN works robustly is that multi-modal datasets potentially compensate each modality which has a different kind of information. As for the 3D multi-modal CNN, the result shows a slightly decreased accuracy over Sugi and Hinoki classification, which implies that the 3D convolution over potentially different resolutions might weaken the performance. On the other hand, the performance of deciduous trees classification is improved, which indicates that the 3D multi-modal CNN can learn the shape of branches captured from LiDAR data.

### 5. CONCLUSION AND FUTURE WORK

In this study, we proposed a CNN which mimics professional interpreters' manual techniques. Using simultaneously acquired airborne images and LiDAR data, we fed the 3D knowledge of tree shape (i.e. voxel) and geospatial features as well as RGB images to the proposed CNN. Inspired by the interpreters' techniques, our network provides a unified approach that integrates these datasets in a shallow layer. The proposed CNN shows that the multi-modal CNN works robustly, and the 3D multi-modal approach is especially suited for deciduous trees. The results of this study suggest that the 3D multi-modal learning over voxels is a promising approach for forest cover classification tasks, especially those involving a forest with a complex 3D structure.

As future work, we plan to improve the performance and robustness of the 3D multi-modal CNN by means of (1) optimization of the weight to integrate modalities and (2)

ensemble learning to combine the effective models appropriately. Aside from that, we intend to use different cropping sizes and resolutions as input data to investigate the effect of information amount. To verify the ability of 3D feature extraction of the proposed method, it is also considerable to incorporate the complexity of 3D shapes such as TIN surface area which derived from point clouds. Finally, the experience on a different area is essential to show the robustness of the proposed approach.

#### ACKNOWLEDGEMENTS

We would like to thank Nakayama Forester for their help in a field survey. This work was supported by JST ACT-I Grant Number JPMJPR16UE, Japan.

#### REFERENCES

- Boulch, A., Saux, B. L., and Audebert, N. 2017. Unstructured Point Cloud Semantic Labeling Using Deep Segmentation Networks. In *Proceedings of Eurographics Workshop on 3D Object Retrieval*.
- Görgens, E.B., Soares, C.P., Nunes, M.H., and Rodriguez, L.C. 2016. Characterization of Brazilian forest types utilizing canopy height profiles derived from airborne laser scanning. *Applied Vegetation Science*.
- Guan, H., Yu, Y., Ji, Z., Li, J. and Zhang, Q. 2015. Deep Learning-Based Tree Classification using Mobile LiDAR Data, *Remote Sensing Letters*, Vol 6, pp.864-873
- Hackel, T., Savinov, N., Ladicky, L., Wegner, J., Schindler, K., Pollefeys, M., 2017. Semantic3d.net: A new large-scale point cloud classification benchmark. arXiv preprint arXiv:1704.03847.
- He, K., Zhang, X., Ren, S. and Sun, J. 2015. Deep Residual Learning for Image Recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105.
- Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A. 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), pp. 778-782.
- Li, Y., Yang, M., Zhang, Z. 2016. Multi-View Representation Learning: A Survey from Shallow Methods to Deep Methods, *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lu, M., Chen, B., Liao, X., Yue, T., Yue, H., Ren, S., Li, X., Nie, Z., and Xu, B. 2017. Forest Types Classification Based on Multi-Source Data Fusion. *Remote Sensing*. 9(11), 1153.
- Maturana, D., and Scherer, S. 2015. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.922-928.

Niemeyer, J., Rottensteiner, F., Sorgel, U. 2014. Contextual classification of lidar data and building object detection in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 87, pp 152-165

Ng, A. 2012. Deep Learning, Self-Taught Learning and Unsupervised Feature Learning, <https://www.youtube.com/watch?v=n1ViNeWhC24> (9 Jan. 2018).

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

Penatti, O.A.B., Nogueira K., and Santos, J.A.D. 2015. Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp.44-51.

Qi, C.R., Su, H., Mo, K., and Guibas, L.J. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*. 115(3), pp.211-252.

Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Rajesh, T., and Lucas, R. 2014. New Global Forest/Non-forest Maps from ALOS PALSAR Data (2007-2010). *Remote Sensing of Environment*, 155, pp. 13-31.

Tokui, S., Oono, K., Hido, S. and Clayton, J. 2015. Chainer: a Next-Generation Open Source Framework for Deep Learning, In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

World Resources Institute, 2014. Global Forest Watch. <http://www.globalforestwatch.org> (9 Jan. 2018).

Yang, Y., and Newsam, S. 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*, pp. 270-279.

Yang, Z., Jiang, W., Xu, B., Zhu, Q., Jiang, S. and Huang, W. 2017. A Convolutional Neural Network-Based 3D Semantic Labelling Method for ALS Point Clouds. *Remote Sensing*, Vol 9, 936

Yokoyama, R., Shirasawa, M. and Kikuchi, Y., 1999, Representation of topographical features by openness. *Journal of the Japan Society of Photogrammetry and Remote Sensing*, 38(4), pp. 26-34. (In Japanese)