ROOF TYPE SELECTION BASED ON PATCH-BASED CLASSIFICATION USING DEEP LEARNING FOR HIGH RESOLUTION SATELLITE IMAGERY

T. Partovi [a,*], F. Fraundorfer [a, b], S.Azimi [a], D. Marmanis [a], P. Reinartz [a]

[a] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234,Wessling, Germany-
{tahmineh.partovi, friedrich.fraundorfer, seyedmajid.azimi, dimitrios.marmanis, peter.reinartz}@dlr.de

[b] Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria

Commission II, WG II/6

KEY WORDS: Roof Reconstruction, High Resolution Satellite Imagery, Deep Learning Method, Convolutional Neural Networks

ABSTRACT:

3D building reconstruction from remote sensing image data from satellites is still an active research topic and very valuable for 3D city modelling. The roof model is the most important component to reconstruct the Level of Details 2 (LoD2) for a building in 3D modelling. While the general solution for roof modelling relies on the detailed cues (such as lines, corners and planes) extracted from a Digital Surface Model (DSM), the correct detection of the roof type and its modelling can fail due to low quality of the DSM generated by dense stereo matching. To reduce dependencies of roof modelling on DSMs, the pansharpened satellite images as a rich resource of information are used in addition. In this paper, two strategies are employed for roof type classification. In the first one, building roof types are classified in a state-of-the-art supervised pre-trained convolutional neural network (CNN) framework. In the second strategy, deep features from deep layers of different pre-trained CNN model are extracted and then an RBF kernel using SVM is employed to classify the building roof type. Based on roof complexity of the scene, a roof library including seven types of roofs is defined. A new semi-automatic method is proposed to generate training and test patches of each roof type in the library. Using the pre-trained CNN model does not only decrease the computation time for training significantly but also increases the classification accuracy.

1. INTRODUCTION

Today, due to the availability of satellite images with high spatial resolution, there is an increasing interest in developing algorithms for 3D point cloud generation by applying stereo image matching techniques. Although the accuracies of the Digital Surface Models (DSMs) obtained from stereo satellite images are generally lower than that obtained from LiDAR data or aerial images, it is assumed that they are still sufficient for building recognition and reconstruction. Furthermore, due to their rich semantic information content, satellite images can be used additionally for segmentation and classification. 3D building model reconstruction from satellite images is an active research topic which is still in its early stages. Therefore, there are still considerable challenges for improvement of the fully-automatic approaches. Buildings are elevated objects; therefore, the height information provided by DSM data can help to distinguish them from other objects with similar properties such as color or gray values (e.g., flat-roof buildings from asphalt roads). However, due to occlusions, clutter, and some deficiencies of the DSM data generation techniques (i.e. dense stereo matching), the resulting DSM data usually suffers from noise, artefacts, and imperfection (e.g. gaps), especially, at the building edges and roof break lines such as ridge-line and valleys.

Another problem with the use of DSM data for 3D building model reconstruction is that the existence noise in DSM makes it a difficult task to discover meaningful patterns (like roof planes) since neighboring roof points show high variability in height information. For example, in some regions, the geometric parameters (e.g., the slopes and normal vectors) of neighboring pixels in the same roof plane can be significantly different. For this reason, data-driven based methods cannot accurately segment the 3D roof planes and lead to several spurious small segments known as an over-segmentation (Khoshelham, 2005). This problem appears also in the model-driven based methods, where the selection of the correct roof type from the building library will most often fail. In the previous works, the model which exhibits the minimum distance to the 3D point cloud of each building roof in the library is selected as the best roof model. For instance, Lafarge et al. (2010) defined a complex building roof library and used Reversible jump Markov Chain Monte Carlo (RjMCMC) to fit a model to the DSM of the building. The roof type was detected based on its best fit to the point cloud. The correctness of the final results for the classified roof types was not evaluated. Huang et al. (2013) defined a simple and small library to reconstruct 3D building roof models. A generative RjMCMC approach was developed to fit the model into the LiDAR points. Then, a jump strategy started to go to the multi-dimensional space of roof parameters and select the one that has the best fit to the point cloud. As an improvement of previous work for DSM of satellite imagery, Partovi et al. (2015) used mean curvature and Gaussian curvature of DSM to classify building roof to pitched and flat roofs. Additionally, ridge lines were extracted to simplify model initialization and conduct a model fitting process to reduce the dependencies of the reconstruction techniques on DSM data. Partovi et al. (2014) applied clustering of the various local features of an optical image (e.g. pixel intensities, texture, geometrical structures) to classify the building roofs into two

* Corresponding author

categories: pitched and flat roofs. Zhang et al. (2015) demonstrated that synthetic data could be used to improve the performance of the classifier. Roof type classification was performed based on recognizing detected edges of the roofs. A library including six different roof types (e.g. flat, gable, gambrel, half-hip, hip and pyramid) were determined. Primarily, ridge lines or valley lines were extracted for each roof style. Then, synthetic data were generated by means of the Multichannel Autoencoder (MCAE). The extracted features were used to train a Lenet-5 CNN model and a SVM for roof type classification. Training and test patches were aligned using their footprint principal directions and were cropped inside of the building footprint. For simplicity, background and other objects were neglected inside of the aforementioned patches. The total numbers of training and test patches were 3860 and 5684 respectively. F1-score values of roof style classification reached 65% for the CNN method and 77% for the SVM method.

Alidoost and Arefi (2016) proposed three strategies to combine aerial image and DSM of LiDAR data to evaluate the efficiency of this fusion on the roof type classification using fine-tuning of a pre-trained CNN model. As a first strategy, a pre-trained CNN model was fine-tuned on RGB and DSM separately. Next, the label of the roof type with the highest probability of one of the determined models was selected as optimal label for the building roof region. In the second strategy, a fined-tuned CNN model trained on RGB data was used for training of the CNN on DSM data. This strategy improved the final accuracy and reduced the training time. In the third strategy, Principle Component Analysis (PCA) was used on RGB-D data to perform whitening of data as input to the pre-trained CNN model. They also defined a library including seven roof types such as flat, gable, hip, pyramid, mansard, complex and non-building roof. Each training and test patch was assigned to one roof style. Each patch was aligned based on the principle direction of the footprint. A segmentation method was used to remove the other objects inside of the patch and focus on the building roof.

Concerning roof shapes (such as flat, gable, hip, half hip, and etc.) can be recognized from high resolution satellite imagery and with inspiration of the previous methods, we propose a new strategy for roof type classification based on convolutional neural networks (CNNs). A CNN is one of the state-of-the-art classification methods, very successful in object recognition, classification, and object detection. As training of CNNs, from scratch needs large datasets with labels which are hard to obtain in the remote sensing community, using a pre-trained CNN models is suggested. Marmanis et al., (2016) used a pre-trained CNN model which was pre-trained on ImageNet dataset (Krizhevsky, et al., 2012) and successfully achieved good results in the classification of remote sensing datasets. To select the roof types, two methodologies are utilized for classification of roof patches. The first method uses fine-tuning of a pre-trained VGGNet (Simonyan and Zisserman, 2015) model on ImageNet in the Caffe framework (Jia, 2013).

The second method uses deep features obtained from output of the last fully-connected layers of three large pre-trained CNNs and concatenates the information into a new feature vector. The extracted feature vectors for training and test roof patches are employed by Support Vector Machines (SVM) classifier (Chang and Lin, 2011) in order to classify the roof types.

The first step of using a pre-trained CNN model in classification starts with preparing training and test datasets. Training and test patches including only one type of roof are prepared and labelled. We propose a new semi-automatic method for generating roof patches based on building skeleton and building outlines. To generate training and test patches, three channels (RGB) of a pansharpened image of WorldView-2 data from Munich city are used. Each patch is manually labelled related to the roof type. The main difference between this way of generating patches and previous methods (Zhang et al., 2015) and (Alidoost and Arefi, 2016) is that, the main direction of each roof is also considered inside the patch selection. In this way, the quality of roof patches cannot be degraded by rotation and resizing.

The rest of the paper is organized as follows: Section 2 introduces the developed CNN and the pre-trained model. In section 3, methodologies for roof type classification are explained. Section 4 provides experimental results and discussion. Finally, section 5 concludes the paper.

2. CNN AND PRE-TRAINED MODELS

A convolutional neural network (CNN) is one of the deep learning architectures which have been applied in different fields of computer vision and machine learning. A CNN is a kind of feedforward neural network which is composed of several convolutional, subsampling layers, fully connected layers, activation functions and classifier layers. The convolution layer is the main part of the CNN architecture. Convolution layers consist of a set of learnable filters which are convolved across the width and height of input values and produce a 2-D feature map of that filter. The convolution layer is formed by stacking the feature maps of all filters along the depth of the input volume. Another important part of CNN is pooling which is a kind of non-linear down sampling. There are several functions such as average pooling, L2-norm pooling and max-pooling for implementing pooling among which max-pooling is the most common. The pooling layer reduces the spatial size of the representation. Subsequently it reduces the number of parameters in the network to control overfitting and also reducing the computation time. After several convolutional and max-pooling layers, the last fully-connected layer holds the output, such as the class scores. Neurons in a fully connected layer have full connections to all feature maps in the previous layer. The loss layer is the last layer of the CNN which is employed to penalize the deviation between the predicted and true labels during network training phase. Various loss functions appropriate for different tasks may be used. Cross Entropy loss is the most widely used loss function in object recognition tasks. As classifier layer, softmax is the most common one to produce on output probability for each object class to predict a single class of several potential classes. To train large numbers of parameters in deep CNN networks, starting from random values of weights and bias vectors, a high capacity of memory and time as well as a huge dataset in various classes are needed. Therefore, without having large real datasets for training, the training process from scratch will fail due to the overfitting problem. To solve this problem, a pre-trained CNN model, already trained on a big dataset, is used as a starting point of the training process. Pre-trained CNN models can be adapted to small datasets by re-adjusting their learned weights and bias using a back-propagation algorithm. This method is called “transfer learning”. This is motivated by the observation that the earlier layers of CNN model will have learned almost the same basic features such as edge, color or gradient features for visual recognition tasks, while the upper layer has more specific features related to the classes and properties of the original dataset. Therefore, upper layers can be fine-tuned to match to a different but correlated problem. Pre-trained model for CNNs like Alexnet (Krizhevsky and Sutskever, 2010), VGGNet or GoogleNet (Szegedy and Liu,

2015) that have been trained on large dataset such as ImageNet can be used for other visual recognition tasks without any needing to train the first few layers. Such a property is very useful for classification tasks in remote sensing, where the acquisition of large sets of training data need a lot of effort and cost (Marmanis et al., 2016). In addition to fine-tuning a pre-trained model for a new classification task, a pre-trained CNN can be treated as fixed feature extractor. In this structure, the classification layer of the CNN is removed and the rest of the CNN is treated as a feature extractor for the new dataset. When these features are extracted from all images, a classifier like SVM or softmax is used in the end to classify images based on extracted features. These features are known as DeCAF features (Donahue, et al., 2014).

3. METHODOLOGY

Building roof type classification is one of the important steps for model-driven based 3D building reconstruction. In model-driven methods, first a library of building roof types is created. Depending on the roof complexity of a city area, we designed a library consisting of seven types of different roofs such as flat, gable, half hip, hip, pyramid, mansard and complex roofs. The automatic selection of the roof type from the library is a classification problem. To select the correct roof type from the library, we evaluate two common strategies based on using a pre-trained CNN model instead of training the CNN models from scratch.

3.1 Roof type classification based on pre-trained CNN model

The first strategy is fine-tuning of the pre-trained VGGNet (16-layer version) model on the ImageNet dataset (The actual size of the training data set consists of about 1.2 million images with 1000 distinct classes). Among many pre-trained model, VGGNet adopts the simplest kernel and pooling windows. Only 3×3 convolutions and 2×2 pooling are used throughout the whole network. VGG also shows that the depth of the network plays an important role and gives better results. In this strategy, we fine tune the higher level portion of the network on our training patches.

3.2 Roof type classification based on deep features and SVM classifier

The idea of deep features is to use the first stage of a CNN only. The weights of internal layers are used as feature vector to be classified with traditional methods, e.g. SVM. Thus for the second strategy, we remove the last fully connected layer of the VGG (16 and 19-layer version) and also GoogleNet pre-trained model which act as classifier layer together with softmax layer, then treat the rest of CNNs as fixed feature extractors for the training and test patches. The extracted features from fc7 layer of the VGGNet (16 and 19-layer versions) and pool5/7x7_s1 from GoogleNet pre-trained model are concatenated to the single vector and employed SVM classifier using RBF kernel in order to classify roof types.

3.3 Dataset generation

The quality of training and test patches are an important issue to obtain higher accuracies. In roof type classification, each patch should be related to only one roof type. To reduce the computation time of generating training and test patches, a new semi-automatic method is proposed. First the skeleton of the building mask which is extracted from cadastral building footprint is computed by morphological operator. After that, the junction points of the skeleton are projected on pansharpened satellite images. Three channels of pansharpened images are used to generate the patches. Around each junction point of the skeleton, a square box with fixed size crops the image. The size of the square box is selected so that other building parts cannot involve into the patch of the selected building since only one roof type should be inside of each patch.

4. EXPRIMENTS AND RESULTS

4.1 Satellite roof classification dataset

The main dataset of the experiments is provided from an orthorectified panchromatic and multispectral WorldView-2 satellite image from Munich city. In order to use a RGB image with high spatial resolution, a pan-sharpened image is generated by fusion of the panchromatic image with 50 cm spatial resolution and 8 channels of the orthorectified multispectral image with 2 m spatial resolution. Then 3 channels (channel 5 is red, channel 3 is green and channel 2 is blue) from 8 channels are selected to generate the RGB image. To separate buildings from other objects, a building mask is extracted from a building footprint mask. Based on the prominence of roof types in Munich city, we define a building roof library consisting of seven roof types (flat, gable, half hip, hip, pyramid, mansard and complex roof shape). Satellite data exhibit great challenges even for visual analysis tasks. First, the quality of some satellite images is degraded because of blurring which occurs during the capturing process. Second roofs are covered by various kinds of facilities such as air conditioning, chimneys and other roof build-ups.
In addition, some roofs in satellite imagery are occluded by shadow and trees. These imperfections are significant obstacles to reliable visual and computerized analysis tasks. Furthermore, the number of instances for some roof classes is extremely low in comparison to other classes (such as mansard and pyramid roofs). Although augmentation methods such as converting color space to HSV and flipping can modulate the shortage of a dataset partly, there is still need for more data to have a better balance among classes.
These unbalanced distributions of data before and after data augmentation is shown in Table 1.

Roof type	Data Augmentation	Training #	Test #
Flat	yes/ no	2491/624	116
Gable	yes/no	2603/651	193
Half hip	yes/no	2599/650	176
Hip	yes/no	1751/438	46
Pyramid	yes/no	303/76	0
Mansard	yes/no	99/25	3
Complex	yes/no	44/11	0

Table 1. The distribution of the training and test sets used in the experiment

Training patches are selected from different areas of Munich and test patches are from a completely new area and are therefore totally independent from the training data. Figure 1 shows a library of roof patches generated by the proposed method in section 3. The pre-trained VGGNet (16-19) layer model architectures require inputs of a fixed 224×224×3 patch size. All of the training and test patches are generated with a

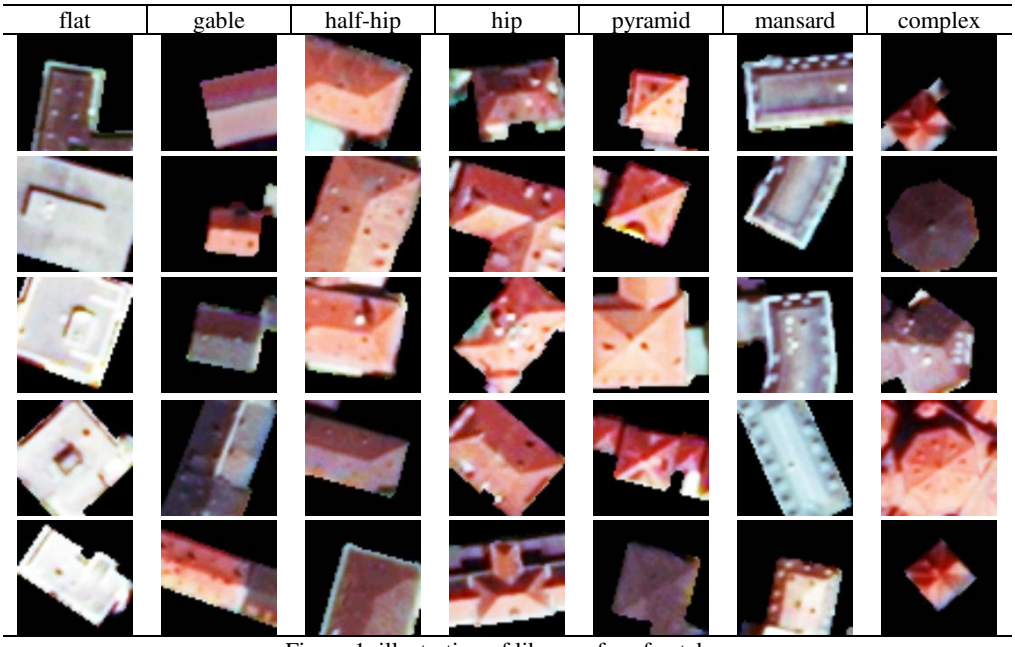

Figure 1. illustration of library of roof patches

fixed size of 50×50×3 and resized to the required input size of network without any rotation. Since the CNN pre-trained on ImageNet ILSVRC benchmark dataset, all of the training data and test data are subtracted from the mean image which is computed from 1.2 million images with 1000 different categories of ILSVRC challenge dataset which is ImageNet dataset.

4.1.1 Training step: For the training phase of the first strategy based on fine-tuning the pre-trained model, 20% of each class of training data are considered as validation data and separated from training data. Next, new weights of the pre-trained CNN model are fine-tuned based on the rest of the training data. Figure 2 shows how the accuracy of validation data is improved through the fine-tuning process. It also shows that the pre-trained model on ImageNet can transfer learning to the new dataset by fine-tuning without many number of iterations. In the second strategy after extracting deep features of training and test patches, a 5-fold cross validation is used on the training features to train a SVM in order to find the best parameters (gamma=7.0711 and c=0.0001953) of a RBF kernel (Chang and Lin, 2011). Since the goal of the second strategy is to test the use of deep features, training patches are used without any augmentation step. This is to speed up the classification process.

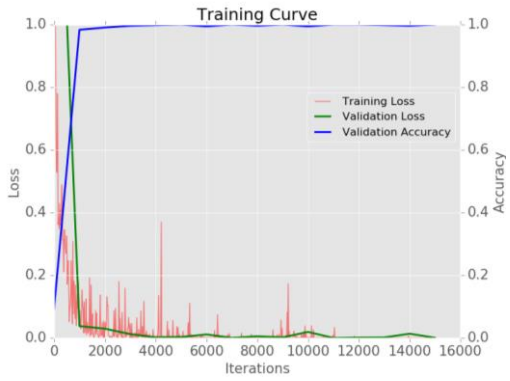

Figure 2: Learning curve (red curve: training Loss, green curve: Validation Loss, blue curve: Validation accuracy)

4.1.2 Testing step: To predict the class of each roof type in the test data which is selected from different areas, the pre-trained model and the SVM using determined RBF parameters obtained from the previous step are employed. The validation of the proposed methods for roof type classification is performed through quantitative comparison of the predicted classes and ground truth (Vakalopoulou, et al., 2015). The standard measures of completeness, correctness and quality of the predicted classes for test data are calculated as given in Equations (1).

$$Coml. = \frac{TP}{TP + FN};$$
$$Corr. = \frac{TP}{TP + FP}; \quad (1)$$
$$Qual. = \frac{TP}{TP+FN+FP}$$

where *TP* is True Positive, *FP* is False Positive, and *FN* is False Negative. Table 2 and 3 show the results of these measures for each class of test area for the two strategies respectively.

Images	TP	FN	FP	Compl.	Corr.	Qual.
flat	98	18	29	84%	77%	68%
gable	149	45	35	77%	80%	65%
half-hip	122	55	35	69%	77%	58%
hip	36	10	30	78%	54%	47%
pyramid	0	0	0	---	---	---
mansard	0	3	0	0	---	0
Complex	0	0	2	---	---	---
All classes	405	131	131	76%	76%	60 %

Table 2. Quantitative evaluation results for test data using the first strategy

In the test data, no patch belongs to pyramid and complex roof, therefore their accuracies cannot be evaluated. We only found three patches for mansard roof in the region which are classified as flat and half-hip roof. The reason for this misclassification is due to the low number of training patches and the pre-trained

model that cannot distinguish mansard roofs from hip roofs which have a similar structure. As Table 1 and Table 2 show, a high number of training patches results in high quantitative evaluation measures so that gable and flat roofs have higher completeness, correctness and quality compared with other roof types in the test area.

Images	TP	FN	FP	Compl.	Corr.	Qual.
flat	102	14	24	87%	80%	72%
gable	72	122	19	37%	79%	34%
half-hip	130	47	116	73%	52%	44%
hip	31	15	42	67%	42%	35%
pyramid	0	0	0	---	---	---
mansard	0	3	0	0	---	0
Complex	0	0	0	---	---	---
All classes	335	198	201	63%	62%	45 %

Table 3. Quantitative evaluation results for test data using the second strategy

In the second method, the goal is to use deep features from different pre-trained CNN models and to compare the performance of SVM classifier with the pre-trained CNN classifier. As Table 3 shows, the overall performance of using deep features and SVM classifier is lower than the first method in average, and only flat roof obtained higher correctness and quality measures than the first strategy. Although standard procedures to prevent overfitting have been used, overfitting seems to be responsible for the unsatisfying performance. In the work of Zhang et al. (Zhang, et al., 2015), the SVM classifier performed better in roof type classification compared to a classifier based on a CNN only. This behaviour could not be confirmed with our experiments. However, since our dataset is highly different from the dataset used in (Zhang et al., 2015), strong conclusions cannot be drawn.

5. CONCLUSION

In this paper, we investigated the potential of pre-trained ImageNet models and their deep features using the Caffe framework for roof type classification from satellite imagery. We defined a library of the roof models based on the complexity of roofs in Munich city including flat, gable, half-hip, hip, pyramid, mansard and complex roofs. We also proposed a new semi-automatic method for training and test patch generation using building masks. In the first method, a pre-trained VGGNet model is used for classifying roof models. In the second method, deep features are extracted from three pre-trained models (such as VGGNets 16-19 layers and GoogleNet) to classify roof types by SVM classifier. Since the initial results of the two methods on a small dataset are promising, we will investigate their performance on a much larger dataset in future work.

REFERENCES

Alidoost, F., Arefi, H., 2016. Knowledge based 3D building model recognition using convolutional neural networks from LiDAR and aerial imageries: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Prague, Czech Republic, Vol. XLI-B3, part. XXIII, pp. 833-839.

Chang, C.-C., and Lin, C.-J., 2011. "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1-27:27, software available at htto://www.csie.ntu.edu.tw/cjlin/libsvm.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. DeCAF: A deep convolutional activation feature for generic visual recognition, *In Icml* Vol. 32, pp. 647-655.

Huang, H., Brenner, C. and Sester, M., 2013. A generative statistical approach to automatic 3D building roof reconstruction from laser scanning data. *ISPRS journal of photogrammetry and remote sensing*, pp. 29-43.

Jia, Y., 2013. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.Berkeleyvision.org/

Khoshelham, K., 2005. Region refinement and parametric reconstruction of building roofs by integration of image and height data. *In: CMRT05*, vol. XXXVI, part. 3/W24. pp. 3-8.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. *NIPS*, pp. 1106-1114.

Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M., 2010. Structural approach for building reconstruction from a single DSM. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (1), 135–147.

Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep Learning Earth Observation classification using ImageNet pre-trained Networks, IEEE Geoscience and Remote Sensing Letters, Vol. 13, No. 1, pp. 105-109.

Partovi, T., Huang, H., Krauß, T., Mayer, H., Reinartz, P., 2015. Statistical building roof reconstruction from WorldView-2 stereo imagery: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Munich, Germany, Vol. XL-3/W2, pp. 161-167.

Partovi, T., Bahmanyar, R., Krauß, T., Reinartz, P., 2014. Building roof component extraction from panchromatic satellite images using a clustering-based method: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Zurich, Switzerland, Vol. XL-3, pp. 247-252.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, Scott., Anguelov, D., Erhan, D., Vincent, V., Rabinovich, A., 2015. Going deeper with convolution, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9.

VakaLopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, pp. 1873-1876

Zhang, X., Fu, Y., Zang, A., Sigal, L., Agam, G., 2015. Learning classifiers from synthetic data using a multichannel autoencoder: *arXiv preprint arXiv:1503.03163*,