SAMODEJNO ZAJEMANJE ODTISOV STAVB IZ UAV PODOB Z UPORABO NEVRONSKIH MREŽ

AUTOMATIC BUILDING FOOTPRINT EXTRACTION FROM UAV IMAGES USING NEURAL NETWORKS

Zoran Kokeza, Miroslav Vujasinović, Miro Govedarica, Brankica Milojević, Gordana Jakovljević

UDK: UDK: 004.032.26:528.7/.8
Klasifikacija prispevka po COBISS.SI: 1.01
Prispelo: 18. 2. 2020
Sprejeto: 26. 10. 2020

DOI: 10.15292/geodetski-vestnik.2020.04.545-561
SCIENTIFIC ARTICLE
Received: 18. 2. 2020
Accepted: 26. 10. 2020

IZVLEČEK

Ažurni katastrski načrti so ključnega pomena za urbanistično načrtovanje, njihovo posodabljanje pa je drago in zahteva veliko časa. Razvoj daljinsko vodenih letalnikov (angl. unmanned aerial vehicles – UAV) je omogočil hiter zajem podatkov z veliko višjo stopnjo podrobnosti od klasične geodetske izmere. V raziskavi se ukvarjamo s samodejnim zajemom odtisov stavb iz ortofotov visoke ločljivosti. Cilja naše študije sta bila: (1) preskusiti možnosti uporabe različnih javno dostopnih podatkovnih nizov (Tanzania, AIRS in Inria) za učenje nevronskih mrež ter nato preskusiti zmožnosti generalizacije modela območja obravnave; (2) oceniti vpliv normaliziranega digitalnega modela površja na rezultate učenja in uporabo nevronskih mrež. Rezultati so pokazali, da modeli, ki smo jih učili na omenjenih podatkovnih nizih, niso zadovoljivi, saj je prekrivanje identificiranih odtisov stavb z referenčnimi podatki znašalo za podatkovni niz Tanzania 36,4%, za AIRS je bila vrednost 64,4% za Inria pa le 7,4%. Boljše rezultate smo dosegli v drugem delu raziskave, kjer je bilo učenje nevronskih mrež izvedeno na delih (256 x 256 pikslov) ortofota, pridobljenega na podlagi podatkov, zajetih z UAV. Pri kombiniranju ortofota z normaliziranim digitalnim modelom površja se je še povečal delež prostorskega ujemanja z referenčnimi podatki (IoU) in znašalo 90 %.

ABSTRACT

Up-to-date cadastral maps are crucial for urban planning. Creating those maps with the classical geodetic methods is expensive and time-consuming. Emerge of Unmanned Aerial Vehicles (UAV) made a possibility for quick acquisition of data with much more details than it was possible before. The topic of the research refers to the challenges of automatic extraction of building footprints on high-resolution orthophotos. The objectives of this study were as follows: (1) to test the possibility of using different publicly available datasets (Tanzania, AIRS and Inria) for neural network training and then test the generalisation capability of the model on the Area Of Interest (AOI); (2) to evaluate the effect of the normalised digital surface model (nDSM) on the results of neural network training and implementation. Evaluation of the results shown that the models trained on the Tanzania (IoU 36.4%), AIRS (IoU 64.4%) and Inria (IoU 7.4%) datasets doesn't satisfy the requested accuracy to update cadastral maps in study area. Much better results are achieved in the second part of the study, where the training of the neural network was done on tiles (256x256) of the orthophoto of AOI created from data acquired using UAV. A combination of RGB orthophoto with nDSM resulted in a 2% increase of IoU, achieving the final IoU of over 90%.

KLJUČNE BESEDE

nevronske mreže, globoko učenje, klasifikacija, grajenje strukture iz gibanja, daljinsko vodeni letalniki, zajemanje odtisa stavbe

KEY WORDS

neural network, deep learning, classification, Structure from Motion, unmanned aerial vehicles, building footprint extraction

1 INTRODUCTION

Automatic extraction of building footprints, the outer surface of building rooftop, on high-resolution orthophoto is one of the most challenging and essential tasks. This information finds its application in urban planning, especially in detailed regulatory planning and 3D city modelling. Although it is possible to extract building footprints manually it is very time consuming and expensive, especially in the case of large urban areas. With the development of computer technology, there was a significant number of attempts to complete this process automatically. Most of those attempts can be grouped into two categories, image classification and image segmentation.

Extraction of building footprints as image classification is most simply done with only two classes assigned, building and non-building. In the basis of all image classification algorithms lies the idea of assigning each image pixel to a certain class. Image classification algorithms are commonly classified into two categories: pixel-based and object-based methods. Pixel-based classification analyses each pixel individually and assigns it to a class based on its spectral similarities with the class. Thank you for your comment. Updated accordingly

The object-based classification was first introduced and demonstrated in the 1970s but was not widely accepted until the mid-1990s due to low computing power (de Kok, Schneider, and Ammer, 1999). Object-based classification, in comparison to pixel-based, does not operate on single pixels, but segments consisting of many pixels that have been grouped using some of the image-segmentation algorithms (Guo et al., 2018). Image segmentation algorithms separate buildings from their surrounding on orthophoto. Over the past decades, a significant amount of image segmentation algorithms has been proposed, that are commonly divided into the following groups: pixel, edge, region and artificial neural networks based techniques. Segments created by an image-segmentation algorithm can then be classified in many ways, for example, by comparing mean spectral characteristics of segments with class training samples, or by using a pixel-based classification of every segment, and assigning the whole segment to the class that most pixels belong to.

Development of deep learning architectures, especially Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), and state-of-the-art methods, such as:

- for image classification: AlexNet (Krizhevsky, Sutskever, and Hinton, 2012), GoogleNet (Szegedy et al., 2015), Residual Network (ResNet) (He et al., 2017b);
- for image semantic segmentation: Fully Convolutional Network (FCN) (Long, Shelhamer, and Darrell, 2015), U-Net (Ronneberger, Fischer, and Brox, 2015);
- for object detection: Region Convolutional Neural Network (R-CNN) (Girshick et al., 2014), You Only Look Once (YOLO) (Redmon et al., 2016);
- for instance image segmentation: Mask R-CNN (He et al., 2017a), Fully Convolutional Instance Segmentation (FCIS) (Li et al., 2017); etc.,

inspired researchers to apply those methods for building footprint extraction.

There is a lot of research conducted to determine the best neural network model for the extraction of building footprints. Wu et al. (2018) perform extraction of building footprints using the multi-constraint fully convolutional network (MC-FCN) and compare its performance with the U-Net model. Although the proposed model achieves better results than with U-Net, this difference is not significant in comparison to the longer time period needed to train the network.

Maltezos et al. (2018) attempted to detect changes between two different epochs of data acquisition by using CNN, Linear Supported Vector Machine (SVM) and Radial Base Function kernel SVM. Results indicated that CNN (average quality: 65% (new buildings) and 77% (unchanged buildings)) provides higher classification accuracy compared to the results provided by SVM (average quality: 57% (new buildings) and 74% (unchanged buildings)).

Ševo and Avramović (2016) attempted to automatically detect buildings on UAV images using CNN based approach on GoogleNet architecture, initialised with weights trained to convergence on the ImageNet (Deng et al., 2009) database. ImageNet database is probably the largest dataset for neural network training containing over 1.2 million high-resolution images, classified into the 1000 different classes (Krizhevsky, Sutskever, and Hinton, 2012). They achieved an accuracy of 98.61%, with training on UCMerced dataset (UAV images 256 x 256 pixels) and USGS national map (high-resolution image 5000 x 5000 pixels).

The objectives of this study are:

1. to determine the suitability of presented Res-U-Net-4 approach for automatic extraction of building footprints on high-resolution orthophotos,
2. to determine the possibility of using transfer learning and open datasets for automatic extraction of building footprints in different urban areas,
3. to evaluate the influence of heights (nDSM and Digital Terrain Model (DTM)) on the accuracy of building footprints extraction.

2 STUDY AREA

The AOI for building footprint extraction for urban planning is part of the city Banja Luka, Republic of Srpska, cadastral municipality Česma 2, which overview is illustrated in Figure 1. AOI is a suburban area on the right side of river Vrbas with small detached houses that consists of slope and flat, mainly red colour, roof surfaces. A detailed regulatory plan aimed to redesign the present settlement with single-family housing, planning new residential buildings, plots, and supporting facilities, new roads, and public spaces by using a comprehensive, integrated planning approach (Milojević, 2018). In the initial stage of the plan, it was necessary to perform an update of cadastral maps and mapping of buildings, among which some were not mapped.

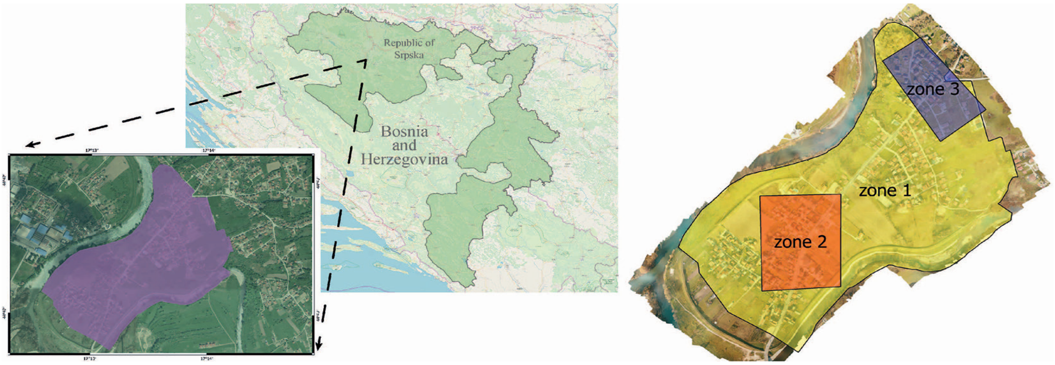


Figure 1: Overview of AOI. The right image shows the orthophoto of the AOI, *zone 1* is a testing area, *zone 2* is a training area, and *zone 3* is a validation area.

3 METHODOLOGY

In this paper, a pixel classification method to extract building footprints from UAV images based on a deep learning algorithm is proposed. We propose the workflow showed in Figure 2, which summarises the approach followed in this paper and consists of three main steps: pre-processing, image segmentation, and accuracy assessment.

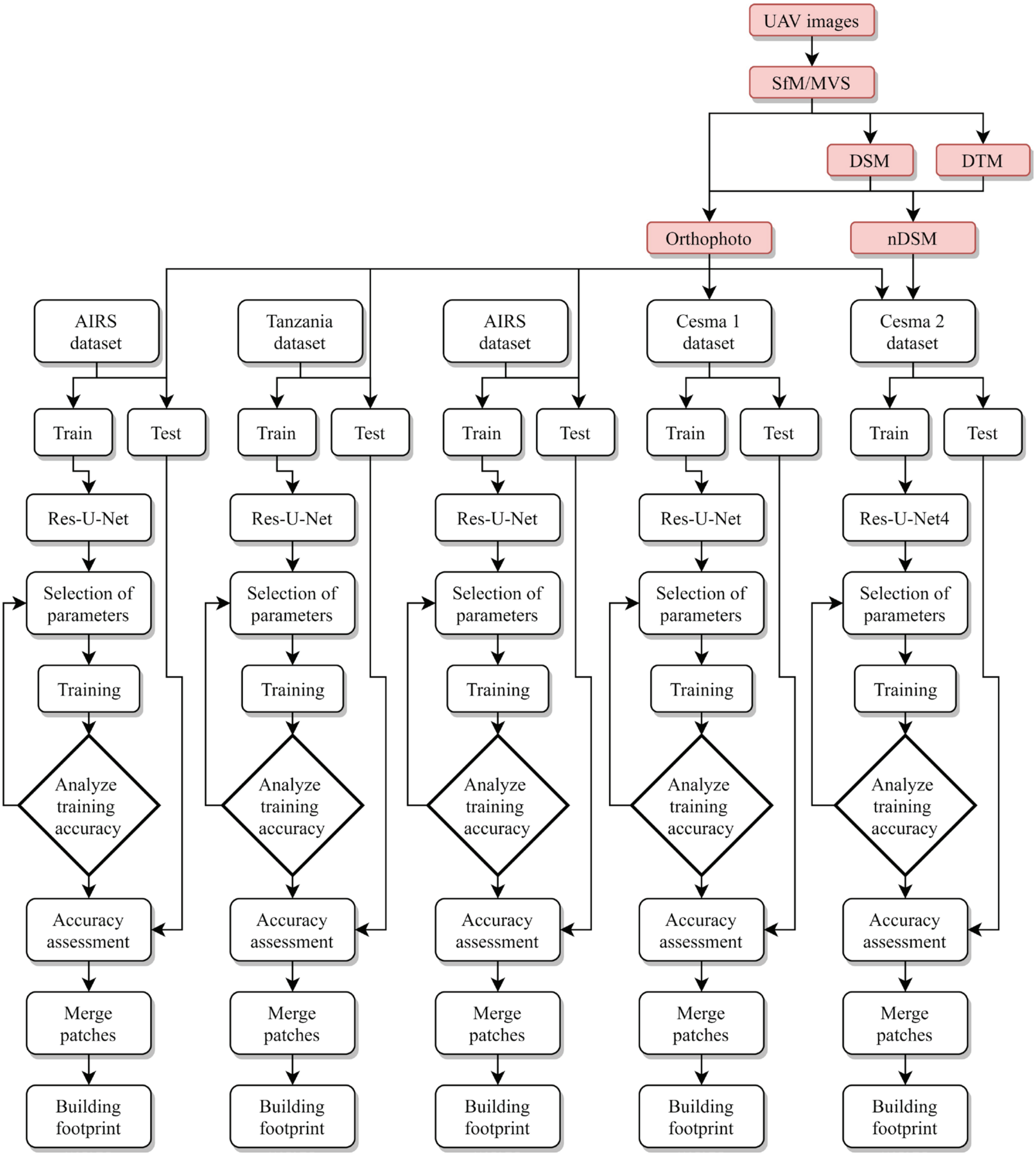


Figure 2: Workflow used in this research.

3.1 Pre-processing

UAV images of AOI are acquired using the small quadcopter DJI Mavic PRO (DJI Mavic Pro User Manual, 2017), in November 2018. The period of the year for image acquisition was determined by project requirements and natural conditions, to complete the process while the trees are without leaves but before the snow. The flight altitude was set to 87 m above the ground level, which provided images with a spatial resolution of 2.84 cm/pixel. We used the double grid flight pattern, which is suitable for 3D modelling (Martinez et al., 2019). The forward overlap and side overlap between images was 80%. A total of 4276 images was collected over the AOI. Acquired data is then processed applying the Structure from Motion (SfM) and multi-view stereo (MVS) algorithms (Figure 2), as described by Westoby et al. (2012) and Bianco, Ciocca and Marelli (2018). The result of processing is a dense point cloud, orthophoto, and Digital Surface Model (DSM) of AOI. In this study, the nDSM was chosen instead of DSM because it removes the terrain topography and leaves only object above the ground (Koc San and Turker, 2006). In order to create nDSM, which represents the difference between DSM and DTM, the ground points should be extracted from the created point cloud. Different approaches can be used for classification of ground points such as filtering based on geometrical features (Axelsson, 2000) or automatic raw point cloud classification by using Artificial Neural Networks (ANN) (Jakovljevic et al., 2019). In this paper, the Agisoft Metashape was used for processing of UAV images and point cloud classification (Agisoft Metashape User Manual, 2019).

Labelling of the buildings was done manually in QGIS. The Cesma A and Cesma B training dataset consist of only 82 buildings footprint (Figure 1, Zone 1), while the validation part (Figure 1, Zona 3) consists of 40 buildings. Based on the data collected in the AOI, two datasets were created. Cesma A dataset (consists of orthophoto) and Cesma B dataset (consists of orthophoto and nDSM) (Figure 2).

In order to test the generalisation possibilities of already trained neural network in a different part of the world, some publicly available datasets are used. Due to the rise of open source and interest in the automatic mapping of buildings, more and more high-quality UAV imagery datasets are available. Some of them are presented inTable 1. The comparison (Table 1) indicates that there are significant differences between datasets in terms of spatial resolution and coverage area. When using transfer learning the basic idea is to reduce the need for large datasets (Soekhoe, van der Putten, and Plaat, 2016), which indicates that the area covered by the dataset is not as crucial as other parameters, although all datasets feature relatively broad coverage. The more significant difference is a spatial resolution which should not be a limitation when it comes to object-level recognition, but when it comes to the automatic high-precision mapping of buildings, it will significantly affect the final result, especially for smaller buildings (Chena et al., 2019).

Table 1: Comparison of publicly available datasets that can be used for building footprint extraction.

	Data type	Target classes	Area covered (km^2)	Spatial resolution (cm)	Location
AIRS (Chena et al., 2019)	RGB	building	457	7.5	Christchurch, New Zeland
Inria (Maggiori et al., 2017)	RGB	building	810	30	Ten regions in the USA and Austria
Tanzania	RGB	building	100	7.7	Zanzibar

Besides the datasets presented in Table 1, there are many more of which important to mention are

SpaceNet Challenge datasets (Etten, Lindenbaum, and Bacastow, 2018) and Massachusets Buildings Dataset (Mnih, 2013). SpaceNet Challenge datasets are the largest dataset available for satellite image segmentation, featuring high-resolution images for four cities (Etten, Lindenbaum, and Bacastow, 2018). Massachusets Buildings Dataset is one of the first aerial image datasets, used for training of CNNs, that is made publicly available. SpaceNet was not used because it provides off-nadir while the labels are provided using nadir images, which makes precise mapping practically impossible. At the same time, Massachusets dataset adopts the OpenStreetMap data for ground truths (Mnih, 2013), which may bring significant noise into the data, such as missing or incorrect labelling, caused by crowdsourcing.

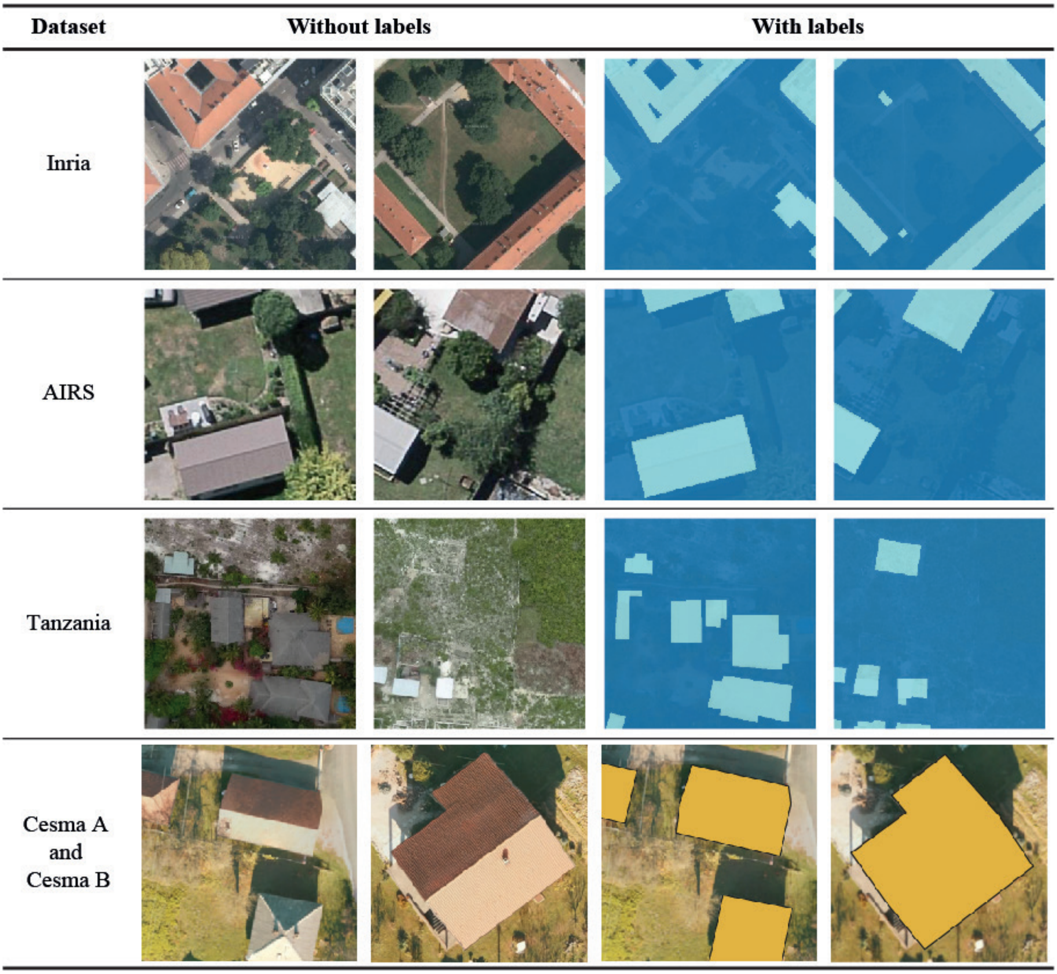


Figure 3: Examples of images in datasets used for training of each model.

On the other hand, the analysis of datasets mentioned in Table 1 detected several problems, such as:

- Ground truths in Inria dataset are not aligned with building footprints, but with ground part of the building, which brings considerable noise in data when there are buildings that are not projected perfectly orthogonal.
- Tanzania dataset features the buildings that are not finished or that are demolished, which

brings a significant noise in data. Those buildings are featured because the dataset is created mainly for the classification of buildings in different categories such as complete, incomplete and foundation.

Examples of input images for Inria, AIRS, Tanzania and Cesma dataset are given in Figure 3.

For the training of models, five datasets were used:

1. Inria dataset (RGB images – 3 bands), Inria model,
2. AIRS dataset (RGB images – 3 bands), AIRS model,
3. Tanzania dataset (RGB images – 3 bands), Tanzania model,
4. Cesma dataset A (RGB images – 3 bands), Cesma model A, and
5. Cesma dataset B (RGB + nDSM images – 4 bands), Cesma model B.

The final step of the pre-processing stage was splitting all datasets into training, validation and testing datasets. Inria, AIRS, and Tanzania dataset consists of training (initial training set) and test datasets (initial test dataset). The initial training dataset was split into 80% of the data for training and 20% for the validation, while the test dataset was used during the testing phase. For Cesma A and Cesma B dataset, Zone 2 (Figure 1) was used for training, while Zone 3 (Figure 1) was used for validation. In order to provide insight into the generalisation ability of algorithms trained on the initial dataset (Inria, AIRS, and Tanzania), Cesma A Zona 1 was used as a final test dataset. Due to the processing power available on Colab, it was necessary to convert all input images to 256 x 256 pixel patches. Also, during this stage, each dataset was normalised so that all values are in the range [0, 1], as described by Sane and Agrawal (2017).

3.2 Image segmentation

Image segmentation is the main stage in the extraction of the buildings' footprints from UAV images. In this stage, the architecture of the proposed model is defined. After defining the model architecture, data augmentation is performed on all images to avoid overfitting by increasing the number of training images. The appropriate metrics and loss function are chosen as described in the following sections. After the definition of all parameters needed for model training, the implementation of the network is performed.

3.2.1 Res-U-Net and Res-U-Net-4

In the first part of the study, a combination of original ResNet34 (He et al., 2017b) and U-Net (Figure 4) (Ronneberger, Fischer, and Brox, 2015) architectures, called Res-U-Net (Figure 5) is used. The U-Net network consists of two parts: downsampling (left) and upsampling part (right). In the Res-U-Net the downsampling part is ResNet34 which is used for extraction of the features from the input data. In this part of the study, the network architecture is used without any modifications.

In the second part of the study, the combination of modified and extended architectures, ResNet34 and U-Net, called Res-U-Net-4 is used (Figure 5 (a)). The downsampling part of Res-U-Net-4 is ResNet34 architecture, which is modified to accept 4-channels of the input data.

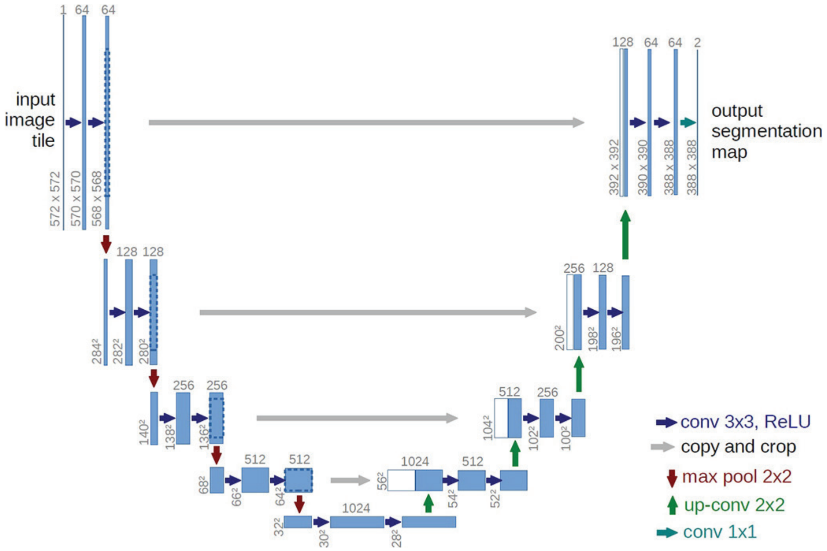
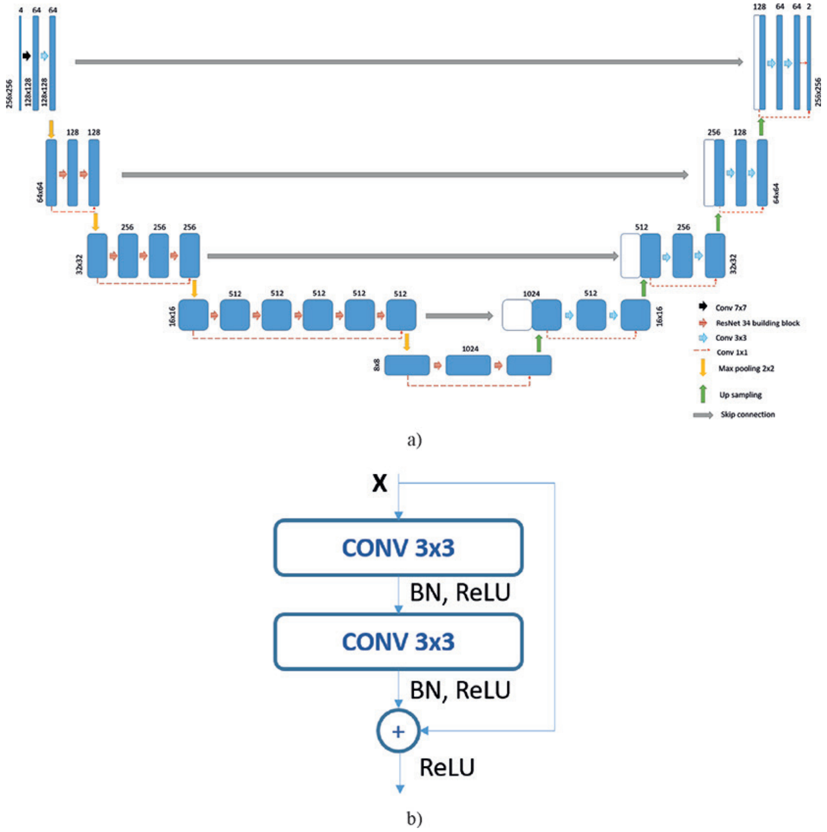


Figure 4: Original U-Net architecture (Ronneberger, Fischer, and Brox, 2015).



a)

b)

Figure 5: (a) The architecture of the Res-U-Net-4 used in this work, (b) the building block of ResNet 34.

The input layer is followed by a convolution block that performs convolution with a 7 × 7 kernel and stride 2 (Figure 5(a)). This block is followed by normalisation, activation, and max-pooling layer. The activation layer consists of a ReLU. The following part of the downsampling consists of four encoder blocks, and every encoder block includes several repetitive residual blocks (Figure 5 (b)). In every residual block, repetitive convolution and normalisation are applied to provide downsampling. The upsampling part aims to extract the buildings using the feature maps and consists of several decoder blocks which are connected with the corresponding encoder block using skip connections. Each decoder blocks consists of convolution layer, batch normalisation and transposed convolution. The number of encoder and decoder blocks are the same.

3.2.2 Data augmentation

The performance of the deep neural network is highly limited by the low number of training data. To avoid overfitting, it was necessary to perform data augmentations. Data augmentation is a technique that is often used to artificially increase the size of the training set by creating modified versions of images by performing different types of transformations (Mikołajczyk and Grochowski, 2018). This technique is applied randomly during the training stage and consisted of the following transformations horizontal- and vertical-flip, rotations and change of the brightness level.

3.2.3 Metrics

The metrics used to evaluate the results of the detection of building footprints was the *Dice* score and Intersection over Union (*IoU*). *Dice* score often referred to as F_1 score, is used during the training stage and is calculated using the equation (1):

$$Dice = F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{1}$$

where *TP*, *FP*, *FN* denotes the true positive, false positive and false negative respectively (Powers, 2011).

The *IoU* score is standard metrics used for semantic segmentation problems. *IoU* is used due to its high correlation to geometry representation of the object, it measures the similarity between the predicted region and the ground-truth region for objects presented in the image and is defined by the equation (2):

$$IoU = \frac{|T \cap P|}{|T \cup P|} = \frac{TP}{TP + FP + FN} \tag{2}$$

In the equation (2), *T* and *P* denote ground-truth and prediction regions, respectively, while *TP*, *FP* and *FN* are the same as for *Dice* (Powers, 2011). The precision and recall score are explained in details by Powers (2011).

3.2.4 Loss function

For training the network, a dual loss function which combines binary cross-entropy and *Dice* loss is used. This approach showed that it slightly outperforms *IoU*-loss, and it is the main reason why two similar metric functions are used (Zhou, Zhang, and Wu, 2018). Minimisation of the loss was done using the Adam (Kingma and Ba, 2015) optimiser.

3.2.5 Implementation details and training

Due to lack of processing power available at authors end, training of network was done using publicly available cloud platform Colaboratory, hosted on Google Cloud. Google Colaboratory, often referred to as Google Colab, is based on Jupyter Notebooks and works like Google Docs object. Colab provides either Python2 and Python3 runtimes, which comes with pre-configured essential machine learning and artificial intelligence libraries, such as Pythorch, Tensorflow, Keras and Matplotlib (Carneiro et al., 2018). Colab provides access to fully configured Graphics Processing Unit (GPU)-accelerated runtime, which makes neural network training practical and comfortable. The only problem is that the specific runtime, called a virtual machine, is deactivated after a specific period of time, often said that it is 12 hours but during the testing stage, that time varied a lot, from few hours to over 20 hours. After runtime is deactivated, all user data and configuration are lost, but it is possible to connect Colab with users Google Drive account or Google Cloud Storage and resume tasks after activating the new virtual machine.

For the training stage, a separate neural network for each dataset is trained, in total, five different models. Each model gets a name according to the dataset used for its training. The training process consisted of epoch training. Iteratively feeding mini-batches to the network, computing the gradients and updating the weights, until each patch in the training data has been processed once by the network, is considered as one epoch. The number of iterations in one epoch varies depending on the batch size. Limitation of the graphic card memory available on Colab, limited the batch size, which was chosen as big as possible for each network. To avoid overfitting, the early stopping is used. Early stopping is a commonly used form of regularisation which interrupts the training process when there is not the improvement of validation loss for a predefined number of epochs. Initial training hyperparameters are obtained empirically and are presented in Table 2. The learning rate was multiplied by 0.1 every few epochs to ensure the quality of the results.

Table 2: Initial values of hyperparameters used for training. All parameters are obtained empirically, during the testing of the training process on the validation set.

Initial learning rate	Weight decay	Momentum
0.001	0.004	0.9

The number of epochs and batch size for the training of each neural network is given in Table 3.

Table 3: Batch size and number of epochs in the training stage by the model.

Model name	Batch size	Res-U-Net	Res-U-Net-4	Time per epochs [min]
Inria model (Inria dataset)	8	45	-	04:50
AIRS model (AIRS dataset)	8	42	-	08:50
Tanzania model (Tanzania dataset)	8	47	-	07:50
Cesma model A (Cesma dataset A)	4	40	-	01:20
Cesma model B (Cesma dataset B)	4	-	40	03:20

All weights are initialised with the weights from ResNet34 pre-trained on ImageNet dataset. In the ResUNet, all weights are set to be the same as the ones in ResNet34. The same way, in the Res-U-Net-4 for the first three channels weights, are the same as in the ResNet34, while the weights for the fourth channel are set to be same as the weights from the first channel of ResNet34.

Workflow for the training of all five models is presented in Figure 2. For the first four models, Cesma A, AIRS, Inria and Tanzania, in the first step of the training stage input dataset consisting of training and validation imagery data is fed to the Res-U-Net network, which is initialised with hyperparameters as shown in Table 2. After that, the training process is started, during which fine-tuning of hyperparameters is performed, according to the analysis of training and validation loss function.

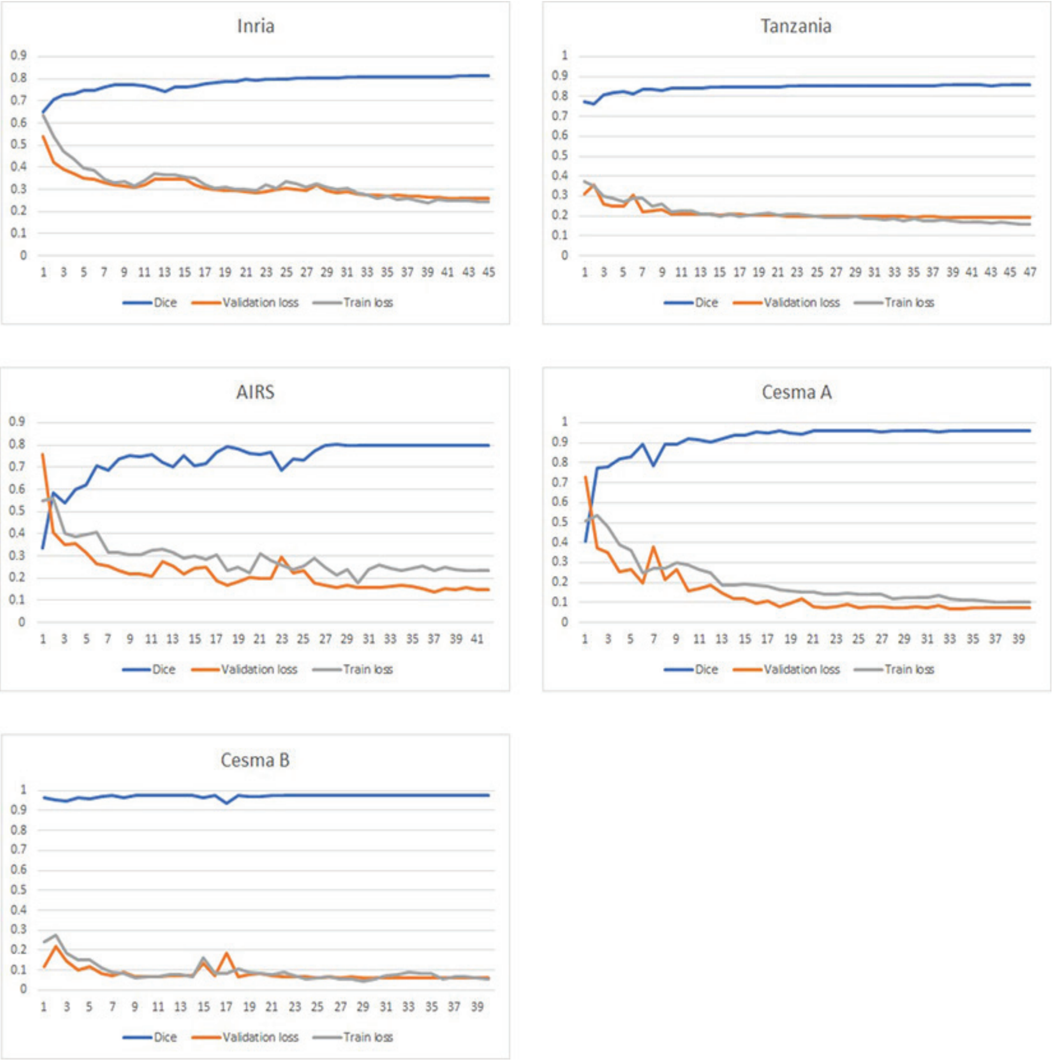


Figure 6: Training and validation loss and accuracy for the five models and dataset.

After the training stage is finished, testing and evaluation of the results are performed. The trained model is applied to predict building footprints firstly on the original test set, and then on the orthophoto (Cesma A) test set. Predictions are then compared to the ground truth labels, and values of the statistics are calculated. More details on statistics calculation are given in section 4. All models are evaluated on the same area of the orthophoto to provide consistency of results.

The training stage for the Cesma model B only differs in the segment of the input data fed to the network. For the Cesma model B training, validation and test sets are created by combining orthophoto and nDSM data as presented in Figure 2. The rest of the details regarding the workflow are the same as for the training of the first four models, as already described. In the testing stage, the trained model is applied to predict building footprints on the orthophoto test set combined with nDSM. In Figure 6, the loss and accuracy results are shown for the training and validation of the five different models.

As it can be seen, both training and validation loss decrease to the point of stable performance with the minimal gap between them indicating the good fit of the models (Figure 6). From the training and validation loss curves, it can be concluded that Inria, Cesma A and Cesma B model shows optimal fit, providing the best accuracy (Figure 6). The Tanzania model's losses highlight that overfitting is still observed, but it has a minimal impact on accuracy. For the AIRS model, slight underfitting is noticed; however, both curves show stable performance over the last ten epochs. As we explained earlier, the number of epochs was determined by early stopping to avoid overfitting and increase generalisation ability. A low number of epochs were expected, taking into account the size of the network and the size of the training dataset. However, the graphs and high accuracy indicated that the number of epochs was enough for fine-tuning of the pre-trained network for this specific task. The shape of curves doesn't suggest that the increase in the number of epochs would increase accuracy. Additionally, the accuracy of building footprint detection can be increased by increasing the training dataset's size and network depth. Tanzania, Inria, and AIRS models are tested using two different test sets (initial test set and Cesma A test set) to evaluate the approach used in this paper, while Cesma A and Cesma B models are tested using only one test set (Cesma A and Cesma B test set respectively), as described in the previous section. For each model, several metrics are used to evaluate models on the Cesma test data (Figure 1). As the primary evaluation metric, the *IoU* was used, due to its high correlation to geometry representation of the object, and it is the only metric presented for evaluation on the original test set. On the other hand, for evaluation on the orthophoto measures for Overall Accuracy (*OA*), Precision, Recall, F_1 and *IoU* metric are presents. The results of the accuracy assessment are presented in Table 4. The visual inspection of merged classified patches is presented on Figure 7.

As for the comparison among the freely available datasets, Inria model provided the highest accuracy in the test phase (test phase based on original test dataset). The work done by Khalel and El-Saban (2018) produced the *IoU* of 74.6% by using 2-levels U-Net and Inria dataset while Pan et al. (2019) used the generative adversarial network with spatial and channel attention mechanisms on the same dataset achieving the *IoU* of 74.92%. The reported results of *IoU* prove that the proposed approach achieved the state-of-the-art performance on the Inria dataset.

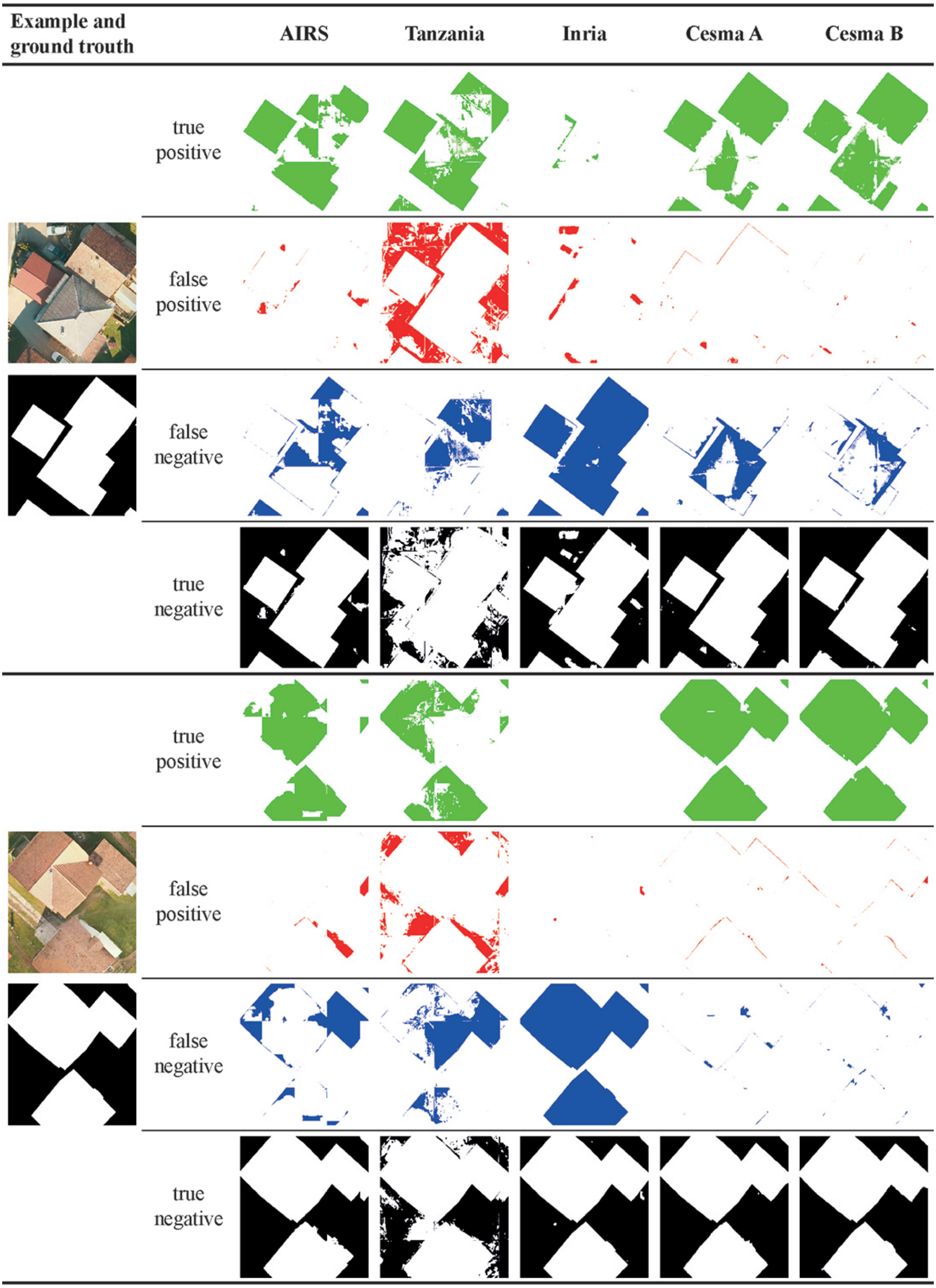


Figure 7: Examples of prediction results.

Table 4: Evaluation of the segmentation results. Measures of the Intersection over Union (*IoU*), overall accuracy (*OA*), precision and F_1 score for different models using original test data and test data created from orthophoto of the study area.

Model	**Original test set**	**Orthophoto**				
	IoU	*OA*	**Precision**	**Recall**	F_1	*IoU*
Tanzania model	70%	0.7986	0.6565	0.4495	0.5336	0.3639
Inria model	81%	0.7883	0.0966	0.2420	0.1381	0.0742
AIRS model	80%	0.9302	0.7204	0.8589	0.7836	0.6442
Cesma model A	Same as orthophoto	0.9793	0.9151	0.9652	0.9394	0.8858
Cesma model B	Same as orthophoto	0.9795	0.9233	0.9793	0.9505	0.9056

In the test phase based on Cesma A test dataset, the Inria model almost wholly omits the buildings producing low accuracy (Table 4). The low accuracy of Inria model is expected due to lower spatial resolution (30 cm vs 2.5 cm) and differences in ground truth data. Bad quality of the ground truth can make testing difficult, influence the results and cause uncertainty on incorrectly labelled features (Chena et al., 2019; Schuegraf and Bittner, 2019) and low precision (Table 4) indicating the significant underestimation of the area covered by buildings (7).

The values of precision and recall (Table 4) show that the model trained on Tanzania dataset cannot extract the building footprints well. The Tanzania model produces a high false-positive rate and low recall (0.45) value due to the misclassification of low albedo surfaces and shadows. Khalel and El-Saban (2018) reported a similar problem. The roofs in Tanzania dataset has more structural complexities (Figure 3) which cause more classification errors (Boonpook et al., 2018). Besides that, the dataset contains the significant number of not finished or demolished buildings producing high noise level and confusion of algorithm.

In addition, AIRS model is capable of extracting building footprints with high accuracy (0.86) however it also tends to misclassify the ground point as buildings (0.72) producing the moderate F_1 value (0.78).

Visual inspection shows that Cesma dataset is much more affected by illumination variance compared to the open datasets (Figure 3). Part of the rooftop oriented opposite of the sun was in the shadow and had a low reflection while the opposite side is bright and shows a high reflection in all bands. Different reflection confused the algorithm and caused false-negative pixels at rooftops. Also, the shadows caused by rooftop infrastructure such as antennas, chimney were misclassified.

Visual inspection shows the AIRS, and Cesma has similar characteristics regarding the rooftop shape and materials. On the other hand, the Inria and Tanzania have flat concrete or more complex rooftops made mostly from metal which is significantly different compared to Cesma datasets (Figure 3). Since the AIRS provided significantly higher *IoU* (0.64) comparing with Inria and Tanzania dataset (0.07 and 0.36), it is essential to note that rooftop geometry and roughness has a significant influence to the algorithm performance.

In addition, the visible structural organisation in different urban morphologies also causes models performance loss (Demir et al., 2018). The Cesma A and Cesma B produce significantly larger accuracy compared to publicly available datasets.

Comparison of the Cesma A and Cesma B evaluation results shows that the *IoU* has been improved by 1.98% when nDSM was used. Xu et al. (2018) obtained the accuracy improvement of 1.64% by using DSM. Usage of nDSM improves building segmentation on pixels that belong to the rooftop but have different spectral values. Although the pixels that belong to the rooftop exposed to the sun and rooftop out of the sun have different spectral values, they have same nDSM values, so it will improve accuracy in those cases. Some pixels belonging to roads have similar spectral values to the building rooftops (Xu et al., 2018), but they have small nDSM values and will not be classified as buildings. As a result, usage of nDSM improves the capabilities and accuracy of the model in building segmentation. The results can be observed in 7.

4 CONCLUSION

Modern society requires accurately mapping of the buildings in the shortest time period possible, with the least amount of resources used. Although new technologies have made the quick acquisition of data possible, a considerable gap still exists in mapping this data. To overpass this gap and automate the process, the possibility of applying the transfer learning approach, using the U-Net model based on ResNet-34 architecture on different datasets and then applying this model to map buildings in Cesma was tested.

The usage of existing datasets increases the time efficiency of extraction of building footprints, but it is not proven that such datasets can efficiently be used to generalise in different parts of the world. The training and evaluation of the neural network proved that the Tanzania dataset contains a significant amount of noise with an *IoU* of 70%. Better results are achieved with AIRS and Inria datasets, with nearly identical *IoU*'s of 80% and 81% respectively.

The application of neural networks trained on Tanzania, Inria and AIRS datasets showed low generalisation capability for the city of Banja Luka. Among the tested datasets, the poorest generalisation capability was shown by Inria model (*IoU*, 7.42%) although it had the best *IoU* on the original validation set. On the other hand, the best generalisation was achieved with AIRS model, with an *IoU* value of 64.42%. Although this value looks promising in comparison to that of the Inria and Tanzania models it is still not suitable to be applied for extraction of building footprints for urban planning because to perform quality urban planning; it is necessary to have an accurate location for each existing object. The former shows that it is not possible to efficiently apply the existing publicly available datasets for the training of neural networks that will be used for detection of building footprints in Banja Luka. The analysis of detected footprints provides deeper insight into the level of cadastral maps update, which is vital information to determine the usability of cadastral maps for urban planning. Based on accuracy assessment and visual comparison of results, it can be concluded that the difference in material, colour, and structure of rooftop significantly limits the generalisation ability of the proposed model. Much better results were achieved with a neural network trained on datasets created from the AOI orthophoto *IoU* 88.58%. A Combination of RGB orthophoto with nDSM resulted in a 2% increase in *IoU* due to the resolution of images and errors that could not be avoided during the creation of the orthophoto, which then caused errors in extraction of building footprints. In the future, the possibility of integration of building 3D models, created from point clouds, and detected building footprints for updating of cadastral maps need to be examined.

Literature and references:

Agisoft Metashape User Manual. (2019). Retrieved from Agisoft: https://www.agisoft.com/pdf/metashape-pro_1_5_en.pdf, accessed 25. 3. 2019.

Axelsson, P. (2000). DEM generation from laser scanner data using adaptive TIN models. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 33, 110–117.

Bianco, S., Ciocca, G., Marelli, D. (2018). Evaluating the Performance of Structure from Motion Pipelines. Journal of Imaging, 4(8), 98. DOI: https://doi.org/10.3390/jimaging4080098

Boonpook, W., Tan, Y., Ye, Y., Torteeka, P., Torsri, K., Dong, S. (2018). A Deep Learning Approach on Building Detection from Unmanned Aerial Vehicle-Based Images in Riverbank Monitoring. Sensors, 18 (11), 3921. DOI: https://doi.org/10.3390/s18113921

Carneiro, T., Medeiros Da Nóbrega, R. V., Nepomuceno, T., Bian, G.-B., De Albuquerque, V. H. C., Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. IEEE Access, 6, 61677-61685. DOI: https://doi.org/10.1109/access.2018.2874767

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., Waslander, S. L. (2019). Aerial Imagery for Roof Segmentation: A Large-Scale Dataset towards Automatic Mapping of Buildings. ISPRS Journal of Photogrammetry and Remote Sensing, 147, 42–55. DOI: https://doi.org/10.1016/j.isprsjprs.2018.11.011

de Kok, R., Schneider, T., Ammer, U. (1999). Object-Based Classification And Applications In The Alpine Forest Environment. International Archives of Photogrammetry and Remote Sensing, 32.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., … Raska, R. (2018). DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). DOI: https://doi.org/10.1109/cvprw.2018.00031

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255), 20-25 June 2009, Miami, FL. DOI: https://doi.org/10.1109%2Fcvpr.2009.5206848

DJI Mavic Pro User Manual. (2017). https://dl.djicdn.com/downloads/mavic/Mavic%20Pro%20User%20Manual%20V2.0-.pdf, accessed 3. 11. 2018.

Etten, A. V., Lindenbaum, D., Bacastow, T. M. (2018). SpaceNet: A Remote Sensing Dataset and Challenge Series. arXiv. https://arxiv.org/abs/1807.01232, accessed 25. 3. 2019.

Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 23-28 June 2014, Columbus, OH, USA. DOI: https://doi.org/10.1109/cvpr.2014.81

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., Feris, R. S. (2018). SpotTune: Transfer Learning through Adaptive Fine-tuning. CoRR, abs/1811.08737. https://arxiv.org/abs/1811.08737, accessed 25. 3. 2019.

He, K., Gkioxari, G., Dollár, P., Girshick, R. B. (2017). Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV), 2980-2988. DOI: https://doi.org/10.1109/iccv.2017.322

He, K., Zhang, X., Ren, S., Sun, J. (2017). Deep residual learning for image recognition. 2017 IEEE Conference on Computer Vision (ICCV). Venice, Italy. DOI: https://doi.org/10.1109/iccv.2017.322

Jakovljevic, G., Govedarica, M., Alvarez-Taboada , F., Pajic, V. (2019). Accuracy Assessment of Deep Learning Based Classification of LiDAR and UAV Points Clouds for DTM Creation and Flood Risk Mapping. Geosciences, 9 (7), 323. DOI: https://doi.org/10.3390/geosciences9070323

Khalel, A., El-Saban, M. (2018). Automatic Pixelwise Object Labeling for Aerial Imagery Using Stacked U-Nets. ArXiv, abs/1803.04953. https://arxiv.org/abs/1803.04953, accessed 25. 3. 2019.

Kingma, D. P., Ba, J. (2015). Adam: A Method for Stochastic Optimisation. CoRR, abs/1412.6980. https://arxiv.org/abs/1412.6980, accessed 25. 3. 2019.

Koc San, D., Turker, M. (2006). Automatic building detection and delineation from high resolution space images using model-based approach. ISPRS Workshop on Topographic Mapping from Space.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Neural Information Processing Systems. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf, accessed 25. 3. 2019.

Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y. (2017). Fully Convolutional Instance-Aware Semantic Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4438–4446), 21-26 July 2017, Honolulu, HI, USA. DOI: https://doi.org/10.1109/cvpr.2017.472

Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 7-12 June 2015, Boston, MA, USA. DOI: https://doi.org/10.1109/cvpr.2015.7298965

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P. (2017). Can semantic labeling methods generalise to any city? the inria aerial image labeling benchmark. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 23-28 July 2017, Fort Worth, TX, USA. DOI: https://doi.org/10.1109/igarss.2017.8127684

Maltezos, E., Ioannidis, C., Doulamis, A., Doulamis, N. (2018). Building Change Detection Using Semantic Segmentation on Analogue Aerial Photos. FIG Congress 2018, 6-11 May 2018, Istanbul, Turkey. https://www.fig.net/resources/proceedings/fig_proceedings/fig2018/papers/ts11c/TS11C_maltezos_ioannidis_et_al_9252.pdf

Martinez, L. I., Santos-Berbel, C. D., Pascual, V., Castro, M. (2019). Using Small Unmanned Aerial Vehicle in 3D Modeling of Highways with Tree-Covered Roadsides to Estimate Sight Distance. Remote Sensing. DOI: https://doi.org/10.3390/rs11222625

Mikołajczyk, A., Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop (IIPhDW), 9-12 May 2018, Swinoujście, Poland. DOI: https://doi.org/10.1109/iiphdw.2018.8388338

Milojević, B. (2018). Integrated urban planning in theory and practice. Contemporary theory and practice in construction XIII, (pp. 318–337), 24-25 May 2018, Banja Luka, B&H. DOI: http://dx.doi.org/10.7251/STP1813323M

Mnih, V. (2013). Machine Learning for Aerial Image Labeling. PhD Thesis, Toronto: University of Toronto. https://www.cs.toronto.edu/~vmnih/docs/Mnih_Volodymyr_PhD_Thesis.pdf, accessed 5. 5.2 019.

Pan, X., Yang, F., Gao, L., Chen, Z., Zhang, B., Fan, H., Ren, J. (2019). Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms. Remote Sensing, 11 (8), 917. DOI: https://doi.org/10.3390/rs11080917

Powers, D. M. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37–63. https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf

Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016, Las Vegas, NV, USA. DOI: https://doi.org/10.1109/cvpr.2016.91

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, (pp. 234–241), 5-9 October 2015, Munich, Germany. DOI: http://dx.doi.org/10.1007/978-3-319-24574-4_28

Sane, P., Agrawal, R. (2017). Pixel normalisation from numeric data as input to neural networks: For machine learning and image processing. 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 2221–2225), 22-24 March 2017, Chennai, India. DOI: https://doi.org/10.1109/wispnet.2017.8300154

Schuegraf, P., Bittner, K. (2019). Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN. ISPRS International Journal of Geo-Information, 8 (4), 191. DOI: https://doi.org/10.3390/ijgi8040191

Ševo, I., Avramović, A. (2016). Convolutional Neural Network Based Automatic Object Detection on Aerial Images. IEEE Geoscience and Remote Sensing Letters, 13 (5), 740–744. DOI: https://doi.org/10.1109/lgrs.2016.2542358

Soekhoe, D., van der Putten, P., Plaat, A. (2016). On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. Advances in Intelligent Data Analysis XV, 50–60. DOI: https://doi.org/10.1007/978-3-319-46349-0_5

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7-12 June 2015, Boston, MA, USA. DOI: https://doi.org/10.1109/cvpr.2015.7298594

Westoby, M. J., Brasington, J., Glasser, N. F., Hambrey, M. J., Reynolds, J. M. (2012). "Structure-from-Motion" photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology, 179, 300–314. DOI: https://doi.org/10.1016/j.geomorph.2012.08.021

Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., ... Shibasaki, R. (2018). Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. Remote Sensing, 10 (3), 407. DOI: https://doi.org/10.3390/rs10030407

Xu , Y., Wu, L., Xie, Z., Chen, Z. (2018). Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. Remote Sensing, 10 (1), 144. DOI: https://doi.org/10.3390/rs10010144

Zhou, L., Zhang, C., Wu, M. (2018). D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 18-22 June 2018, Salt Lake City, UT, USA. DOI: https://doi.org/10.1109/cvprw.2018.00034

Kokeza Z., Vujasinović M., Govedarica M., Milojević B., Jakovljević G. (2020). Automatic building footprint extraction from UAV images using neural networks. Geodetski vestnik, 64 (4), 545-561.
DOI: https://doi.org/10.15292/geodetski-vestnik.2020.04.545-561

Zoran Kokeza, MSc student
University of Banja Luka, Faculty of Architecture,
Civil Engineering and Geodesy
Bulevar vojvode Stepe Stepanovića 77/3
BiH-78000 Banja Luka, Bosnia and Herzegovina
e-mail: zorankokeza95@gmail.com

Miroslav Vujasinović, MSc student a
University of Banja Luka, Faculty of Architecture,
Civil Engineering and Geodesy
Bulevar vojvode Stepe Stepanovića 77/3
BiH-78000 Banja Luka, Bosnia and Herzegovina
e-mail: miroslav.vujasinovic@aggf.unibl.org

prof. Miro Govedarica, Ph.D.
University of Novi Sad, Faculty of Technical Sciences
Trg Dositeja Obradovića 6
SRB-21000 Novi Sad, Serbia
e-mail: miro@uns.ac.rs

assoc. prof. Brankica Milojević, Ph.D.
University of Banja Luka, Faculty of Architecture,
Civil Engineering and Geodesy
Bulevar vojvode Stepe Stepanovića 77/3
BiH-78000 Banja Luka, Bosnia and Herzegovina
e-mail: brankica.milojevic@aggf.unibl.org

assist. Gordana Jakovljević
University of Banja Luka, Faculty of Architecture,
Civil Engineering and Geodesy
Bulevar vojvode Stepe Stepanovića 77/3
BiH-78000 Banja Luka, Bosnia and Herzegovina
e-mail: gordana.jakovljevic@aggf.unibl.org