BUILDING DETECTION FROM AERIAL IMAGERY USING INCEPTION RESNET UNET AND UNET ARCHITECTURES

S. Aghayari [1*], A. Hadavand [1], S. Mohamadnezhad Niazi [1], M. Omidalizarandi [2]

[1] Photogrammetry and computer vision department, Ideh Pardazan Tosseah co, Tehran, Iran - {Saleh.aghayari, hadavand, mohamadnezhad}@idehpardazan.ir
[2] Geodetic Institute, Leibniz University Hannover, Germany - zarandi@gih.uni-hannover.de

Commission IV, WG IV/3

KEY WORDS: Large-scale monitoring, Building detection, Image segmentation, Residual blocks, Skip connection.

ABSTRACT:

Buildings are one of the key components in change detection, urban planning, and monitoring. The automatic extraction of the building from high-resolution aerial imagery is still challenging due to the variations in their shapes, structures, textures, and colours. Recently, the convolutional neural networks (CNN) show a significant improvement in object detection and extraction that surpasses other methods. To extract building, in this paper two segmentation architectures, the UNet and the Inception ResNet UNet are implemented and then tested on the Inria aerial image datasets. The Inception ResNet UNet utilizes the Inception architecture and residual blocks. This makes the model wide and deep, though there are a few differences between numbers of UNet and Inception ResNet UNet parameters. The analyses show that UNet has a high rate of metrics in the training progress. However, on the unseen dataset, Inception ResNet UNet extracts buildings more accurately (97.95% accuracy and 0.96 in the dice metric) in comparison with UNet (94.30% accuracy and 0.55 in the dice metric).

1. INTRODUCTION

The development of remote sensing earth observation systems led to the availability of aerial images at almost all times and locations. It opened numerous applications in computer vision and photogrammetry, e.g., change detection (Gomroki et al., 2022, Isaienkov et al., 2021; Zhang et al., 2020), long-term large-scale monitoring (Immerzeel et al., 2009; Lehmann et al., 2015), and urban management (Mignard and Nicolle, 2014). One of the vital elements that can be extracted from the aforementioned aerial images are buildings. For this task, some datasets and benchmarks have been developed, such as the Inria aerial image dataset (Maggiori et al., 2017) and the Massachusetts buildings dataset (Mnih, 2013). The aim of these processes is to detect the features of buildings or other urban elements (binary or multiple) in aerial images by semantic segmentation (Huang et al., 2018; Ji et al., 2018; Li et al., 2021; Pan et al., 2019).

Semantic segmentation is a crucial task in computer vision and remote sensing community, which deals with assigning a label to each pixel in an image (Yuan et al., 2021). Different machine learning algorithms including artificial neural networks (ANN) have been used to perform this task in the recent years (Mas and Flores, 2008). Over the previous years, researchers have proposed many methods to deal with spatial dependency algorithms (Tarabalka et al., 2009), geographical object-based image analysis (Blaschke, 2010), feature extraction algorithms (Yang et al., 2010), and super-pixel algorithms (Hadavand et al., 2019). These methods could be considered as pre-processing steps for the task of building extraction.

CNN revolutionized a new way to deal with this problem by involving a mathematical convolution with the traditional ANN algorithm. Mathematical convolution in image processing is a matrix operation that works by applying a kernel to each pixel and its neighbours to produce a new value for the centre pixel (Gonzalez, 2009). Nowadays, researchers paid more attention by introducing the AlexNet (Krizhevsky et al., 2012) and showing the good performance on the ImageNet dataset (Deng et al., 2009).

The reasons for super-passing CNN algorithms are that they provide an end-to-end solution and object-based classification (Diakogiannis et al., 2020). In the CNN architecture, any convolution layer generates a new feature from the original image data and uses it as extra information to get a better result. Due to the use of plenty of convolutional layers in the CNN algorithms, they are usually known as "deep CNN," "deep networks," or "deep learning algorithms". Deep learning models are successfully applied in different computer vision and remote sensing tasks such as object detection (Wu et al., 2020; Zhao et al., 2019), image segmentation (Ghosh et al., 2019; Wang et al., 2019), human activity monitoring (Toshev and Szegedy, 2014; Zheng et al., 2019), object tracking (Ciaparrone et al., 2020; Zhai et al., 2018) and also the semantic segmentation.

Semantic segmentation is the essential input for plenty of applications in computer vision and remote sensing, including scene understanding for autonomous driving (Siam et al., 2018), augmented reality (Ko and Lee, 2020), and different environmental monitoring applications such as precision agriculture (Anand et al., 2021), change detection (Venugopal, 2020), and urban mapping and monitoring (Du et al., 2021). In urban remote sensing, discriminating different elements of a city, including different kinds of buildings, paved areas, water bodies, trees and grasslands, cars and clutter are challenging due to variations in shapes, structures, textures, and colours differences (Diakogiannis et al., 2020). In object-based image analysis, this problem is solved by defining several subclasses

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W1-2022-9-2023 | © Author(s) 2023. CC BY 4.0 License.

for a specific class such as building, and therefore, in the post-processing step, they will merge to get a map of buildings (Benz et al., 2004). However, having an algorithm able to detect a class of objects with different characteristics is still a difficult task in remote sensing image analysis.

Among the various existing architectures, UNet (Ronneberger et al., 2015) is a well-known and powerful architecture that shows prominent results in labelling remote sensing imagery in different applications (Feng et al., 2018; Freudenberg et al., 2019; Yang et al., 2019). The UNet structure was originally developed by Ronneberger et al. (2015) to segment biomedical images consisting of an encoder-decoder block to label the pixels of the input image. This model aims to distinguish between the disease location and the corresponding total area in biometrical images to obtain the size and location of the disease in the body.

Chhor et al. (2017) used slightly modified version of UNet by considering the following modifications:

- replacing the stochastic gradient decent with Adam optimizer, which converge faster
- using 'same' padding
- adding batch normalization after every ReLU
- using Dice coefficient of cross entropy as loss
- not utilizing drop out though no overfitting and in training process
- removing down-sampling layer for ease of use in optimization and tackle vanishing gradient.

The loss is set to negative value of Dice. This leads to 0.75 in Dice coefficient and IOU 0.60.

Emek and Demir (2020) used the Sentinel SAR images of Sentinel-1 SAR and Sentinel multi spectral images that cover 120 km^2. Their model is a CNN-based on the UNet architecture. They achieve an implementation accuracy of 81%. The output mask of the model detects some other elements such as buildings, e.g., in wooded areas, some wood is classified as buildings because of its high reflectance value. In addition, it is powerful enough to deal with building extraction problems in complex urban landscapes (Pan et al., 2020).

Wang and Miao (2022) developed RS-UNet. This architecture is based on incorporating the Residual Learning in UNet and combination of Focal Loss (FL) and the Atrous Spatial Pyramid Pooling (ASPP). Focal Loss was used for connection between encoder and decoder, and ASPP as a loss function. For the extraction of more features in the images, a larger size of images has been used in training which was implemented at the size of 512×512 px. However, this increases the training time. In the architecture, the encoder and decoder parts five layers have been used. FL was used for balancing the encoder and decoder parts. The results of various sizes of images with a larger size (512×512) have 97.66% precision in 200 epochs, which is reduced to 97.41% in 128×128 px. The selected best size for training is 256x256 px with consideration of time and precision.

In this paper, the UNet and Inception ResNet UNet architectures are trained and analysed on the Inria aerial image dataset. All buildings are categorized in one class. Our analyses show that by using UNet with the same kernel size in convolution, leads to inability to detect the very large and very small building in the image. In addition, the detection is limited into number of building size. Furthermore, it is not deep enough to detect all kind of building. In some cases, this architecture detects shadows as part of building. The Inception ResNet UNet is a deep and wide. Due to using of various kinds of kernel size, the architectures could detect all kind of building with various shapes, structures, textures, and colours. More details of architectures are presented in the following sections.

This paper is organized as follows: Section 2 presents the methodology. In section 3, experimental results are discussed and interpreted. The summary and conclusions are represented and discussed in section 4.

2. METHODOLOGY

Our proposed deep learning structure is based on the UNet and Inception ResNet UNet architectures. Inception ResNet UNet is an improvement on UNet to solve the convergence problem for deeper encoder-decoder layers. The bing deeper and wider of Inception ResNet UNet allows for precise detection of the object in the image, which is why the Inception ResNet UNet is selected. In the following, the UNet and Inception ResNet UNet architectures are explained.

2.1 UNet architecture

UNet is a convolutional network were proposed in 2015 for medical image segmentation to obtain the precise location and area of objects in a class (Ronneberger et al., 2015). The architecture has a U-shaped structure consisting of two main paths called contracting and expansive paths by the authors and is known as encoder and decoder.

The contracting or encoder path uses repeated convolutions with 3×3 kernel size, same padding, and stride one with Rectified Linear Unit (ReLU), followed by batch normalization and max pooling, which increases the number of feature layers and decreases the size of the image simultaneously. There are no fully connected layers in this model. This part of the architecture is a typical CNN that can be replaced with any pre-trained model. In every step of down sampling, the number of features is doubled. Contrarily, in an expansive or decoder path, up-convolution is used to decrease the number of features and take the image size back to the original input image. Every up-convolution step halved the number of features. To prevent losing the details, concatenating the features from the contracting path is considered in up sampling. In this step, typical convolution layers are applied to the concatenated features. This procedure continues until mask image creation and getting the result. Figure 1 depicts the UNet network in the proposed paper. The encoder is considered on the left and the decoder on the right side. 512×512 px is the input size of the image in the model. Dataset images are cropped to this size, e.g. the 5000×5000 px size, which is the size of any image in the Inria aerial image dataset. In the preprocessing step, they are cropped to a size of 512×512. This operation yields 100 images with a 12-px overlap in images and their side images. The output image size is 512×512 px.

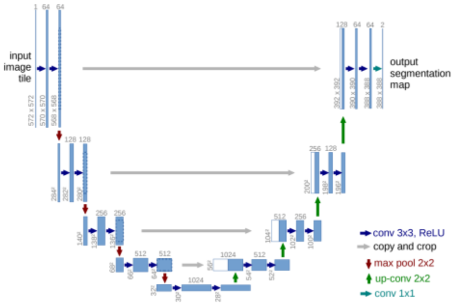

Figure 1. UNet architecture (Ronneberger et al., 2015)

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.

2.2 Inception ResNet UNet architecture

The main structures of network architecture are represented in Figure 3. The Inception ResNet UNet architecture is a modification of the UNet and the Inception ResNet v2 (Szegedy et al., 2016b). Inception ResNet is a combination of the Inception architecture (Szegedy et al., 2015) and residual blocks (He et al., 2016).

The Inception architecture has convolution with multiple kernel sizes at the same level (Szegedy et al., 2015). In other words, instead of transforming a single convolution, the architecture considers multiple convolutions with different kernel sizes in parallel at every block (Figure 5), and at the end of every block, they are concatenated to form a single layer of features (Szegedy et al., 2016a). Due to utilizing multiple kernel sizes, objects of different sizes will be detected in the image. In other words, the model gets wider and deeper (Szegedy et al., 2016a). Sub-blocks depicted in Figure 5 contain parallel convolutions. For improvement in accuracy and reducing computational complexity in Inception ResNet v2, the modifications in architecture are summarised as follows:

Factorizing the 5×5 kernel size into the two 3×3 kernel sizes (as well as 7×7 into the three 3×3 kernels). Instead of using the 5×5 kernel size, we utilize the two 3×3 kernel sizes. The analysis shows they yield the same results, though in this case the number of parameters is reduced. This procedure is repeated in the 7×7 kernel, which is replaced with three 3×3 kernels (Szegedy et al., 2016b).

Factorizing the n×n into n×1 and 1×n convolutions. Every n×n convolution consists of two linear kernels in the horizontal and vertical directions. If we combine these two kernels, we get a squared kernel. In this model, every square kernel is divided into two linear parts in the horizontal and vertical directions. This again leads to reducing the parameters without losing the accuracy (Szegedy et al., 2016b). Typically, the accuracy is increased by making the network deeper by adding more layers. This may aid the network in learning the basic and complex details of an image. By adding more layers, due to the overfitting, the accuracy starts to degrade. The number of layers and designing deeper layers is a challenge to obtain optimum results, especially in building detection and segmentation. An aerial image consists of various building types. The model should be able to detect all kinds of buildings (varying in shapes, structures, textures, and colours) that are all considered in one class. We can make the network deeper by considering the residual block. In a residual block, each layer feeds into the next layer and directly into the next layer.

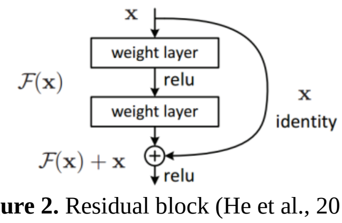
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual block (He et al., 2016).

Figure 2 depicts the residual black, which helps to design a deeper model without overfitting. The model is learned the simple and complex elements in the image by designing the model with this concept. The residuals block in deep architecture helps to avoid gradient vanishing in backpropagation, especially for the deeper architectures with plenty of layers (He et al., 2016).

As previously stated, the Inception ResNet is a combination of the Inception architecture and residual blocks. It consists of 164 layers of very deep and wide CNN. In Inception ResNet the performance is optimised by balancing the filter at every stage. There are 37 blocks in the encoder and six blocks in the decoder part of the Inception ResNet UNet network (Figure 3). The result of convolutions in encoder is concatenated into three different parts in decoder, as in the original UNet algorithm (Figure 1), to form the final result. The number of inputs and outputs of each block is included in Figure 3. The output size of features of every block is mentioned in Figure 3 in every block unit. Block 3 is then 10 times repeated, and finally, the output size is mentioned in the related section. The output feature size is 61×61×320, which is the input of the next block. We have the same conditions in block 5. The outcome is presented after 20 iterations of this block. The output feature size is 30×30×1088 which feeds into the next section. The details of sub-blocks of Figure 3 are displayed in Figure 5. In Figure 5, block 1 simply shows the typical convolution, batch normalization, and an activation function. Blocks 3 and 5 have the skip connection and block 1 and others have block 1 in their structure.

In Inception ResNet UNet architecture, we have one to four parallel convolutions in every block (Diakogiannis et al., 2020). The common property of all blocks is to concatenate the results of all internal operations of the block, similar to residual connections, to produce the output, which is usually the input of another block. Block 4 shares two outputs, one for concatenation and the second for zero-padding, which is reserved for use in the decoder part. Block 6 is the core part of the output of the network, which uses the concatenation of the input image and the output of the UNet encoder-decoder block to produce the result. The proposed structure has 36 million parameters which need to be trained. In the 2014 ILSVRC classification challenge (Russakovsky et al., 2015), VGGNet (Simonyan and Zisserman, 2015) and GoogLeNet (Szegedy et al., 2015) produced comparable high performance. VGGNet needs more resources for computations; in other words, this architecture has 138 million parameters. The computation cost of VGGNet is higher than GoogLeNet, with 5 million parameters. In the 2014 ILSVRC classification challenge (Russakovsky et al., 2015), VGGNet (Simonyan and Zisserman, 2015) and GoogLeNet (Szegedy et al., 2015) produced comparable high performance.

VGGNet needs more resources for computations; in other words, this architecture has 138 million parameters. The computation cost of VGGNet is higher than GoogLeNet, with 5 million parameters. The observations (Szegedy et al., 2015) show the quality of Inception ResNet v2 is higher than GoogLeNet and needs very low resources as we need in VGGNet and the computation cost of Inception is lower than the VGGNet (Szegedy et al., 2015). These are the main reasons for selecting the Inception Resnet v2 over other networks.

3. IMPLEMENTATION AND EXPERIMENTS

3.1 Dataset

The Inria aerial image labelling dataset (Maggiori et al., 2017) was used in our experiments. The dataset covers 405 squares kilometres with a 0.30 meter ground sampling distance (GSD), consists of 180 images and a mask image with 5000×5000 px dimensions. Existing masks divide the image area into two semantic classes: building and non-building.

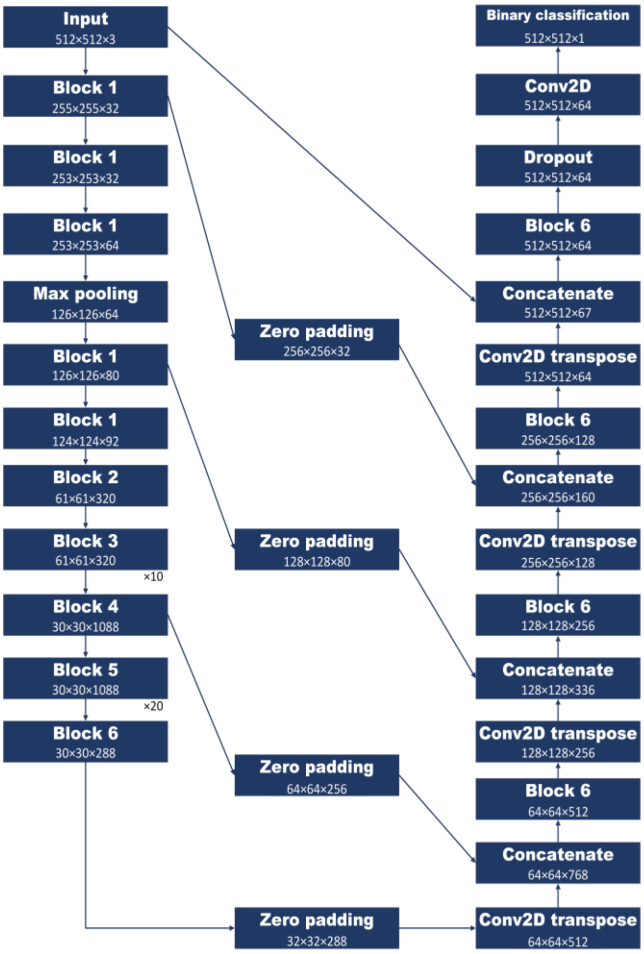

Figure 3: An Overview of the Inception ResNet UNet architecture. The left part is a regular convolution neural network, called an encoder. The right side is called the decoder, which consists of convolution blocks, convolution transpose, and concatenation. Zero padding is utilized to get the same size for concatenating features. Details of blocks 1 to 6 have been shown in Figure 5. The output of every block is represented at the end of every section.

The images are captured across various urban landscapes and illumination. The dataset is gathered from the US and Austrian areas, including Bellingham, Innsburck, San Francisco, Tyrol, and Chicago (Figure 4). These cities contain both high and low densities urban features. There is higher density in Chicago, San Francisco, Vienna, and Innsbruck, and lower density in Kistap, Bloomington, and West and East Tyrol. Every image is divided into 512×512-px sub-images for training the algorithm, leading to a total of 30,000 training and validation images and masks.

3.2 Implementation details

Inception ResNet UNet and UNet have 36 and 34 million parameters, respectively. The UNet is a standard CNN, but the Inception ResNet UNet is made up of Inception architecture and residual blocks. By considering the number of layers in Inception ResNet UNet, there are a few differences between their parameters. The reason, as mentioned in the methodology section, is updates in the GoogLeNet (Szegedy et al., 2015) that led to a deeper and wider model with a few variations in the number of parameters.

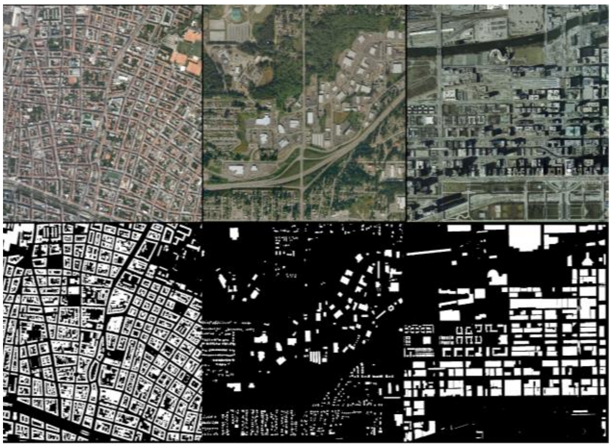

Figure 4. The images and their related masks from the Inria aerial image dataset belong to the Vienna, Kitsap, and Chicago regions. The images have a dimension of 5000×5000 px.

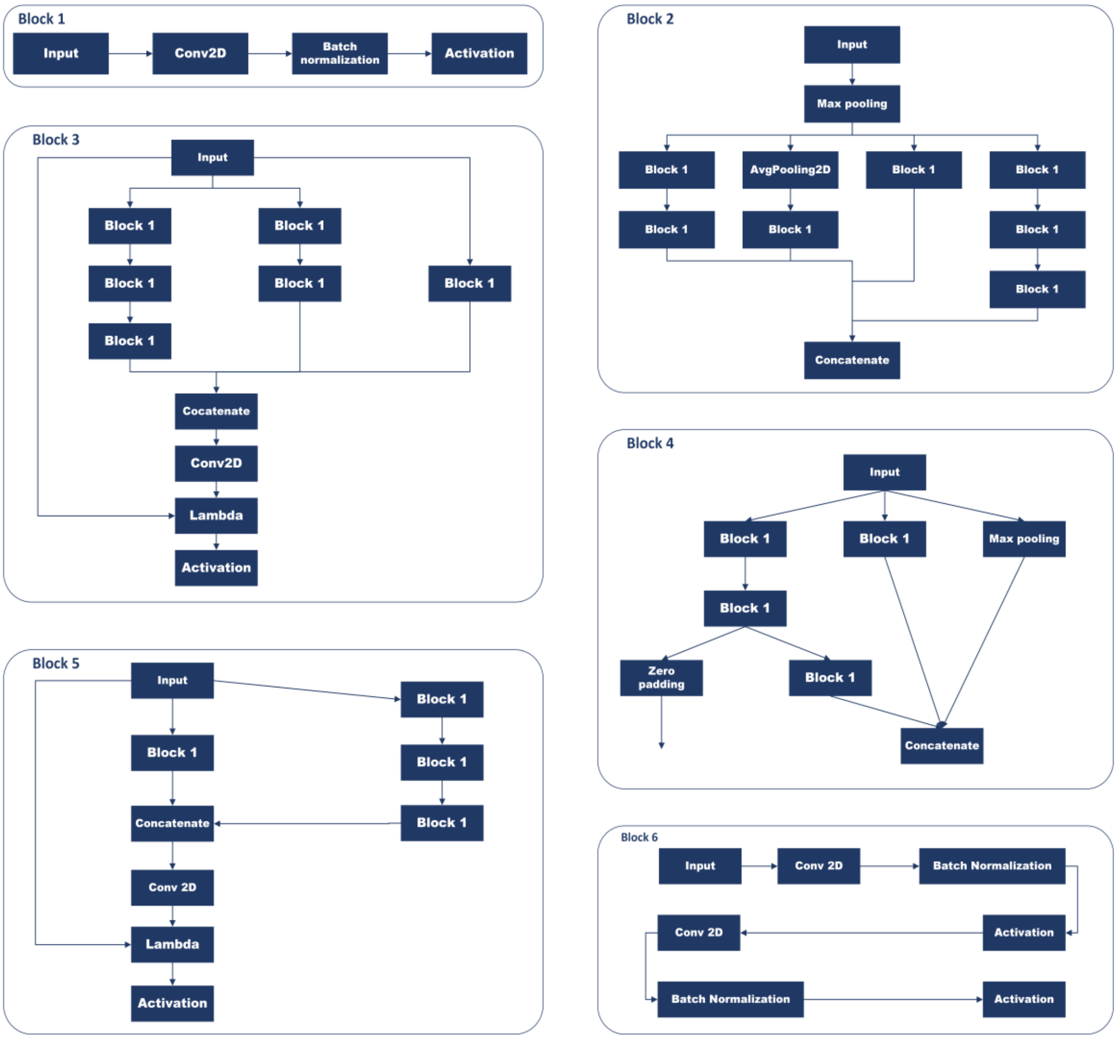

Figure 5. Every block in detail has been used in the ResNet UNet network in Figure 3. Blocks3 and 5 have the residual network and Block 1 in their structure. In architecture, we have one to four parallel convolutions in every block. We use binary classification in this paper, but it can also be used for multi-labeled classes.

The networks are analysed using training and validation data during the training process in every epoch. The results demonstrate that overfitting or underfitting doesn’t occur in the training of the models. After completion of the training, the unseen dataset (form the same distribution as models initial input) is segmented by trained models and compared the metric (accuracy and dice) results. The main aim of trained models is to perform well in unseen datasets. Therefore, our main focus is to analyse the performance of models on the unseen datasets. Figure 6-9 show the results of applying models to these kinds of datasets. The loss function and metrics during training are cross-entropy, dice, and accuracy, respectively. These metrics will be explained in the following. The cross-entropy, which measures the difference between two probability distributions, is used as the loss function of the models (De Boer et al., 2005). Its mathematical equation is as follows:

$$crossEntropy = -\sum_{classess} y_{true}\ log(y_{pred})$$

where: y_{true}: the mask image
y_{pred}: the predicted image

The dice and accuracy metrics are computed and used to evaluate the results of experiments. The Dice coefficient measures the overlap between the model's prediction results and the corresponding mask (Milletari et al., 2016). The dice metric returns a value between 0 and 1, and its maximum values coincide with the ideal prediction result. The dice metric is computed using the following equation:

$$Dice = {}^{\left(2|y_{true} \cap y_{pred}|\right)}\!\big/_{\left(|y_{true}| + |y_{pred}|\right)}$$

Accuracy is based on the confusion matrix. The confusion matrix represents counts from predicted and actual values. This is a 2-dimensional matrix that includes true positives (TP), the number of positive examples classified accurately; true negatives (TN), the number of negative examples classified accurately; false positives (FP), the number of actual negative examples classified as positive; and false negatives (FN), the number of actual positive examples classified as negative to describe model performance. A related mathematical equation is presented as follows (Pan et al., 2019):

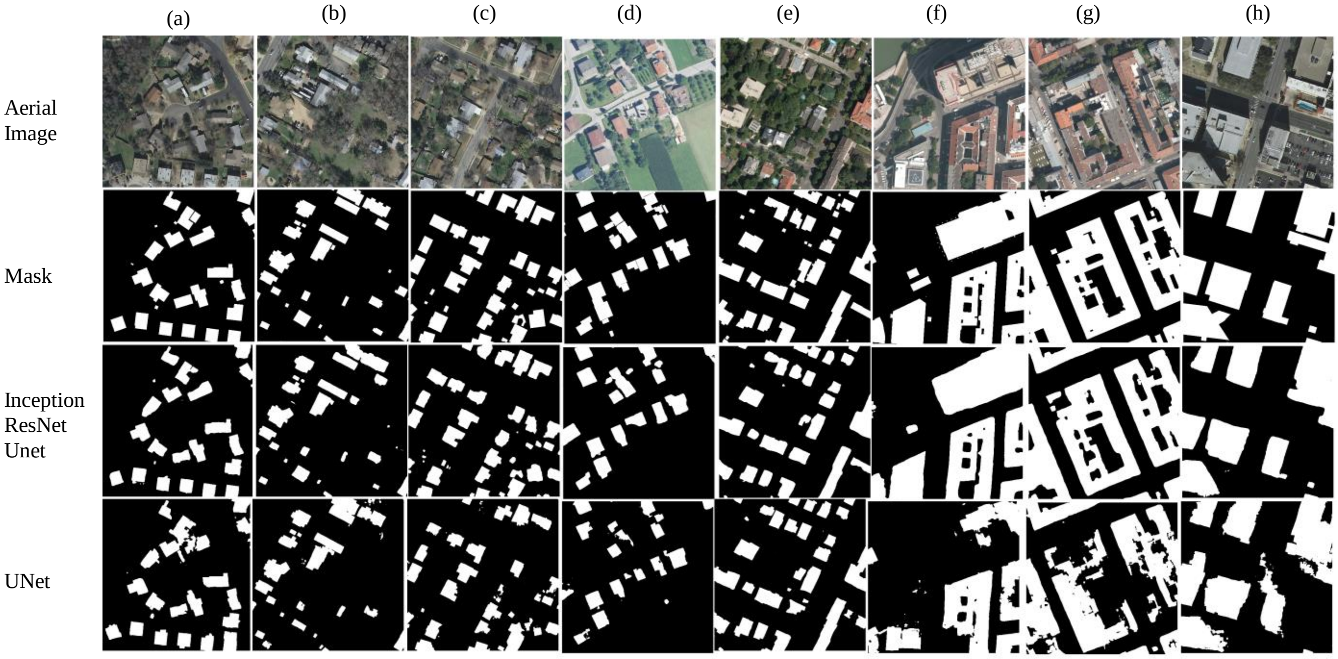

Figure 6. From (a) to (d), some buildings of small size are detected in Inception ResNet UNet, but UNet couldn't detect them precisely. In (f), large buildings are not detected in UNet properly, though the Inception ResNet UNet detects them more accurately. Some shadows were detected as part of building in (h) in UNet, whereas Inception Resnet UNet could handle relief displacement.

$$Accuracy = {(TP+TN)}\Big/{(TP+TN+FP+FN)}$$

The results of computed metrics for the training phase and testing on the unseen datasets for two networks are presented in Table 1. In accordance with it, the results show that UNet reaches higher accuracy on the training dataset; the accuracy and Dice are 99.77% and 0.98, respectively. But its performance is reduced when it is utilized on the unseen dataset; the accuracy and Dice are reduced to 94.30% and 0.55.

Methods	Type	Accuracy (%)	Loss	Dice
Inception ResNet UNet	Training	98.17	0.0449	0.973
	Unseen dataset	97.95	-	0.965
UNet	Training	99.77	0.0035	0.984
	Unseen dataset	94.30	-	0.558

Table 1. Result of training for both network architectures

Contrarily, Inception ResNet UNet results in training are lower than the UNet (98.17% accuracy with 0.97 in Dice), while it performs better than the UNet when dealing with the unseen dataset. The final results of accuracy and Dice are as follows: 97.95% and 0.96 on the unseen dataset. The results of the building detection using Inception ResNet UNet and the UNet on the unseen dataset are depicted in Figure 6. From top to bottom, the images are the aerial images (first row), their corresponding masks (second row), and predicted masks in Inception ResNet UNet (third row) and UNet (last row). The edges and building footprints in the third (Inception ResNet UNet) and last row (UNet) show that Inception ResNet UNet has high accuracy in building detection and edge extraction. The visual comparison and inspection of the results (Figure 7) show that the UNet has issues in detecting the building borders accurately.

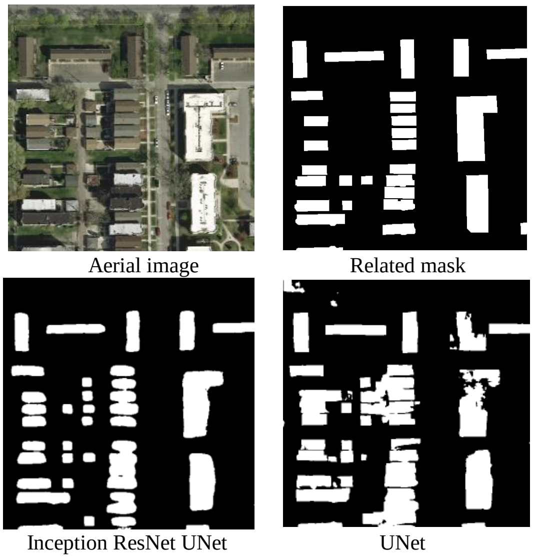

Figure 7. The edges in Inception ResNet UNet are detected more precisely in comparison with UNet.

Inception ResNet UNet performs significantly better and is successful in detecting details of building borders. The Inception ResNet UNet architecture consists of various kernel sizes and residual blocks (Figure 5). By utilizing them, Inception ResNet UNet detects objects with varying shapes, structures, textures, and colour. As illustrated in Figure 8, the Inception ResNet UNet can detect very small and large-scale buildings, though the UNet couldn't detect those buildings accurately. Especially in very large-scale buildings in the Vienna region, the UNet (Figure 6, 6th and 7th columns and Figure 8) couldn't detect footprints. In medium-sized buildings in the Austin region, the models detect almost the same level.

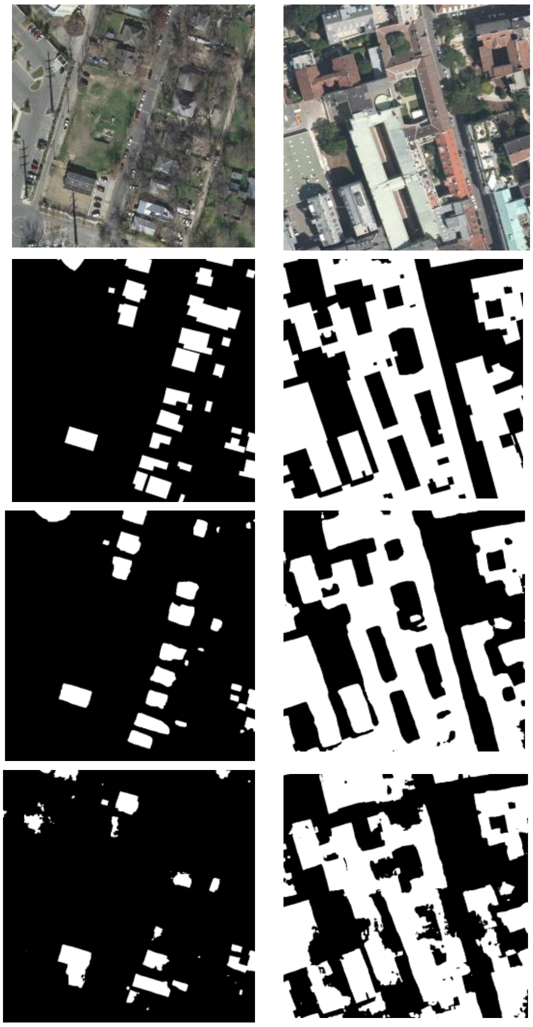

Figure 8. From top to bottom: the aerial images (first row), related masks (second row), and predicted masks in Inception ResNet UNet (third row) and UNet (last row). Inception Resnet UNet detects the large and small building more precisely.

Another issue with working with buildings is the relief displacement that occurs for elevated objects such as tall buildings in aerial and satellite imagery. To remove this effect, the image should be processed to generate a true orthophoto, which is a hard process due to the need for precise 3D models of buildings. Therefore, the relief displacement usually appears in orthorectified images, especially in areas with tall buildings. The results of Figure 9 show that the Inception ResNet UNet enables us to extract tall building footprints more precisely. In the UNet, the shadows are detected as part of the building.

4. DISSCUSSION AND CONCLUSIONS

In this research, two deep network architectures, UNet and Inception ResNet UNet, are implemented and evaluated in automatic building detection from aerial imagery. The Inception Resnet UNet could detect buildings of different shapes, structures, textures, and colours in images in almost all regions, though UNet couldn't detect very large buildings, e.g., in the Vienna region. That is because the Inception ResNet UNet model is wide and deep with few variations in the number of parameters in comparison to UNet. Our experiment results demonstrate that Inception ResNet UNet with 97.95% is preferable in comparison with UNet with 94.30% accuracy in unseen data. Our future work includes in improvement in architectures for multiclass classification of aerial images and precise boundary detection of objects for vectorization of objects.

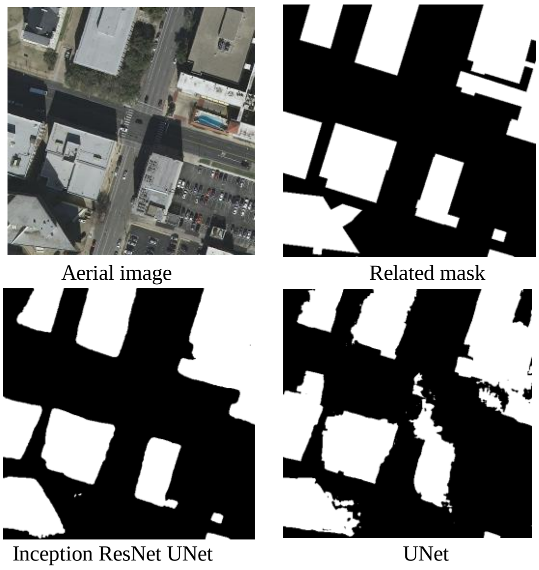

Figure 9. Tall buildings in Inception ResNet UNet are detected more precisely in comparison with UNet. This model detects some shadows as part of the buildings.

ACKNOWLEDGEMENTS

This research was supported by Ideh Pardazan Tosseah Consulting Engineering Company. We are grateful to all of those with whom we have had the pleasure to work during this and other related projects.

REFERENCES

Anand, T., Sinha, S., Mandal, M., Chamola, V., Yu, F.R., 2021. AgriSegNet: Deep aerial semantic segmentation framework for IoT-assisted precision agriculture. IEEE Sensors Journal, 21(16), pp.17581-17590.

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS Journal of photogrammetry and remote sensing, 58(3-4), pp.239-258.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS journal of photogrammetry and remote sensing, 65(1), pp.2-16.

Chhor, G., Aramburu, C.B., Bougdal-Lambert, I., 2017. Satellite image segmentation for building detection using UNet. Web: http://cs229. stanford. edu/proj2017/final-reports/5243715. pdf.

Ciaparrone, G., Sánchez, F.L., Tabik, S., Troiano, L., Tagliaferri, R., Herrera, F., 2020. Deep learning in video multi-object tracking: A survey. Neurocomputing, 381, pp.61-88.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162, pp.94-114.

De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A Tutorial on the Cross-Entropy Method. Annals of operations research, 134(1), pp.19-67.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Du, S., Du, S., Liu, B., Zhang, X., 2021. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. Remote Sensing of Environment 261, p.112480.

Emek, R.A., Demir, N., 2020. building detection from sar images using unet deep learning method. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.

Feng, W., Sui, H., Huang, W., Xu, C., An, K., 2018. Water body extraction from very high-resolution remote sensing imagery using deep UNet and a superpixel-based conditional random field model. IEEE Geoscience and Remote Sensing Letters, 16(4), pp.618-622.

Freudenberg, M., Nölke, N., Agostini, A., Urban, K., Wörgötter, F., Kleinn, C., 2019. Large scale palm tree detection in high resolution satellite images using UNet. Remote Sensing, 11(3), p.312.

Ghosh, S., Das, N., Das, I., Maulik, U., 2019. Understanding deep learning techniques for image segmentation. ACM Computing Surveys (CSUR) 52, pp.1-35.

Gomroki, M., Hasanlou, M. and Reinartz, P., 2022. IUNet-UCD: Improved U-Net with weighted binary cross-entropy loss function for urban change detection of Sentinel-2 satellite images.

Gonzalez, R.C., 2009. Digital image processing. Pearson Education India.

Hadavand, A., Saadat Seresht, M., Homayouni, S., 2019. A novel density-based super-pixel aggregation for automatic segmentation of remote sensing images in urban areas. Earth Observation and Geomatics Engineering 3(1), pp.84-91.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

Huang, B., Lu, K., Audeberr, N., Khalel, A., Tarabalka, Y., Malof, J., Boulch, A., Le Saux, B., Collins, L., Bradbury, K., 2018. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark, in: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 6947–6950.

Immerzeel, W.W., Droogers, P., De Jong, S.M., Bierkens, M.F.P., 2009. Large-scale monitoring of snow cover and runoff simulation in Himalayan river basins using remote sensing. Remote Sens. Environ. 113, pp.40–49.

Isaienkov, K., Yushchuk, M., Khramtsov, V., Seliverstov, O., 2021. Deep Learning for Regular Change Detection in Ukrainian Forest Ecosystem With Sentinel-2. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, pp.364–376.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57, pp. 574–586.

Ko, T.-y., Lee, S.-h., 2020. Novel method of semantic segmentation applicable to augmented reality. Sensors 20, p. 1737.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25.

Lehmann, E.A., Caccetta, P., Lowell, K., Mitchell, A., Zhou, Z.-S., Held, A., Milne, T., Tapley, I., 2015. SAR and optical remote sensing: Assessment of complementarity and interoperability in the context of a large-scale operational forest monitoring system. Remote Sensing of Environment, 156, pp.335-348

Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2021. Building footprint generation through convolutional neural networks with attraction field representation. IEEE Trans. Geosci. Remote Sens. 60, pp. 1–17.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp. 3226-3229.

Mas, J.F., Flores, J.J., 2008. The application of artificial neural networks to the analysis of remotely sensed data. International Journal of Remote Sensing 29, pp. 617-663.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016 fourth international conference on 3D vision (3DV). IEEE, pp. 565-571.

Mignard, C., Nicolle, C., 2014. Merging BIM and GIS using ontologies application to urban facility management in ACTIVe3D. Comput. Ind. 65, pp. 1276–1290.

Mnih, V., 2013. Machine learning for aerial image labeling. University of Toronto (Canada).

Pan, Z., Xu, J., Guo, Y., Hu, Y., Wang, G., 2020. Deep learning segmentation and classification for urban village using a worldview satellite image based on UNet. Remote Sensing 12, p. 1574.

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 Sens. 11, p. 917.

Ronneberger, O., Fischer, P., Brox, T., 2015. UNet: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234-241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, pp. 211–252.

Shamsolmoali, P., Zareapoor, M., Wang, R., Zhou, H., Yang, J., 2019. A novel deep structure UNet for sea-land segmentation in remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, pp. 3219-3232.

Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M., Zhang, H., 2018. A comparative study of real-time semantic segmentation for autonomous driving, Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 587-597.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition (No. arXiv:1409.1556).arXiv.https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016a. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2016b. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning (No. arXiv:1602.07261) .arXiv. https://doi.org/10.48550/arXiv.1602.07261

Tarabalka, Y., Benediktsson, J.A., Chanussot, J., 2009. Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. Geoscience and Remote Sensing, IEEE Transactions on 47, pp. 2973-2987.

Toshev, A., Szegedy, C., 2014. Deeppose: Human pose estimation via deep neural networks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1653-1660.

Venugopal, N., 2020. Automatic semantic segmentation with DeepLab dilated learning network for change detection in remote sensing images. Neural Processing Letters 51, pp. 2355-2377.

Wang, H., Miao, F., 2022. Building extraction from remote sensing images using deep residual UNet. European Journal of Remote Sensing 55, pp. 71-85.

Wang, S., Yang, D.M., Rong, R., Zhan, X., Xiao, G., 2019. Pathology image analysis using segmentation deep learning algorithms. The American journal of pathology 189, pp. 1686-1698.

Wu, X., Sahoo, D., Hoi, S.C., 2020. Recent advances in deep learning for object detection. Neurocomputing 396, 39-64.
Yang, J.-M., Yu, P.-T., Kuo, B.-C., 2010. A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data. Geoscience and Remote Sensing, IEEE Transactions on 48, pp. 1279-1293.

Yang, X., Li, X., Ye, Y., Lau, R.Y., Zhang, X., Huang, X., 2019. Road detection and centerline extraction via deep recurrent convolutional neural network UNet. IEEE Transactions on Geoscience and Remote Sensing 57, pp. 7209-7220.

Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications 169, p. 114417.

Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J. Photogramm. Remote Sens. 166, pp. 183–200.

Zhai, M., Chen, L., Mori, G., Javan Roshtkhari, M., 2018. Deep learning of appearance models for online object tracking, Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0-0.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., Wu, X., 2019. Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems 30, pp. 3212-3232.

Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Shen, J., Kehtarnavaz, N., Shah, M., 2019. Deep Learning-Based Human Pose Estimation: A Survey. ArXiv abs/2012.13392.