Deep Learning based Improved Automatic Building Extraction from Open-Source High Resolution Unmanned Aerial Vehicle (UAV) Imagery

Chintan Maniyar[1] and Minakshi Kumar[1]

[1]Indian Institute of Remote Sensing

November 21, 2022

Abstract

Automatically extracting buildings from remotely sensed imagery has always been a challenging task, given the spectral homogeneity of buildings with the non-building features as well as the complex structural diversity within the image. Traditional machine learning (ML) based methods deeply rely on a huge number of samples and are best suited for medium resolution images. Unmanned aerial vehicle (UAV) imagery offers the distinct advantage of very high spatial resolution, which is helpful in improving building extraction by characterizing patterns and structures. However, with increased finer details, the number of images also increase many fold in a UAV dataset, which require robust processing algorithms. Deep learning algorithms, specifically Fully Convolutional Networks (FCNs) have greatly improved the results of building extraction from such high resolution remotely sensed imagery, as compared to traditional methods. This study proposes a deep learning based segmentation approach to extract buildings by transferring the learning of a deep Residual Network (ResNet) to the segmentation based FCN U-Net. This combined dense architecture of ResNet and U-Net (Res-U-Net) is trained and tested for building extraction on the open source Inria Aerial Image Labelling (IAIL) dataset. This dataset contains 360 orthorectified images with a tile size of 1500m2 each, at 30cm spatial resolution with red, green and blue bands; while covering total area of 805km2 in select US and Austrian cities. Quantitative assessments show that the proposed methodology outperforms the current deep learning based building extraction methods. When compared with a singular U-Net model for building extraction for the IAIL dataset, the proposed Res-U-Net model improves the overall accuracy from 92.85% to 96.5%, the mean F1-score from 0.83 to 0.88 and the mean IoU metric from 0.71 to 0.80. Results show that such a combination of two deep learning architectures greatly improves the building extraction accuracy as compared to a singular architecture.

Deep Learning based Improved Automatic Building Extraction from Open-Source High Resolution Unmanned Aerial Vehicle (UAV) Imagery

Chintan B. Maniyar * and Minakshi Kumar

Indian Institute of Remote Sensing (IIRS), Indian Space Research Organisation (ISRO), India
Correspondence: chintanmaniyar@gmail.com

Abstract. Automatically extracting buildings from remotely sensed imagery has always been a challenging task, given the spectral homogeneity of buildings with the non-building features as well as the complex structural diversity within the image. Traditional machine learning (ML) based methods deeply rely on a huge number of samples and are best suited for medium resolution images. Unmanned aerial vehicle (UAV) imagery offers the distinct advantage of very high spatial resolution, which is helpful in improving building extraction by characterizing patterns and structures. However, with increased finer details, the number of images also increase many fold in a UAV dataset, which require robust processing algorithms. Deep learning algorithms, specifically Fully Convolutional Networks (FCNs) have greatly improved the results of building extraction from such high resolution remotely sensed imagery, as compared to traditional methods. This study proposes a deep learning based segmentation approach to extract buildings by transferring the learning of a deep Residual Network (ResNet) to the segmentation based FCN U-Net. This combined dense architecture of ResNet and U-Net (Res-U-Net) is trained and tested for building extraction on the open source Inria Aerial Image Labelling (IAIL) dataset. This dataset contains 360 orthorectified images with a tile size of 1500m^2 each, at 30cm spatial resolution with red, green and blue bands; while covering total area of 805km^2 in select US and Austrian cities. Quantitative assessments show that the proposed methodology outperforms the current deep learning based building extraction methods. When compared with a singular U-Net model for building extraction for the IAIL dataset, the proposed Res-U-Net model improves the overall accuracy from 92.85% to 96.5%, the mean F1-score from 0.83 to 0.88 and the mean IoU metric from 0.71 to 0.80. Results show that such a combination of two deep learning architectures greatly improves the building extraction accuracy as compared to a singular architecture.

Keywords: transfer learning, fully convolutional networks, image segmentation, building extraction.

1 Introduction

1.1Background

Remote sensing imagery, both satellite and aerial, contains a lot of terrain-feature specific information such as land-cover spread, building footprints, waterbody extent, vegetation and forest boundaries etc. Extracting this feature information without losing relative context within the image is a very important remote sensing image processing milieu [1], [2]. Feature extraction is usually done by identifying a common pattern among pixels and grouping them together, that group of pixels then being a feature [3]. One of the most crucial aspects for accurate image feature extraction is finer spatial details such as edges and corners. Primitive feature extraction methods were time consuming and required a lot of expensive human intervention [4]. This was mostly because of the unavailability of higher spatial resolution data in conjunction with the technical infrastructure at the time. However, with advancements in digital systems for image processing and also the increased availability and accessibility of high spatial resolution data from both satellites and Unmanned Aerial Vehicles (UAVs), image feature extraction has consistently been one of the hottest research topics in remote sensing image processing [5].

1.2 Previous Works

In remote sensing feature extraction, building extraction is one of the most vital aspects of research. With its applications spread in various pipelines of urban mapping and management, disaster

management, change detection, maintaining and updating geodatabases etc., building extraction has caught the attention of researchers worldwide for developing robust and accurate algorithms to automate the process.[6]. Primitive methods of building extraction were based on applying statistical and morphological operations on individual pixels to group them together [7], hence automating the task up to some extent. One of the most prevalent issues in building extraction that has propagated from early methods to the recent methods is the differentiation of foreground and background as well as building and non-building objects [8]. To be able to differentiate between these, spectral and geometrical cues such as colour, shape and line have been used to extract buildings from very high resolution imagery [9]. Another study combined distinctive corners while estimating building outlines to extract buildings [10], but was unable to extract irregular shaped buildings. In the beginning of the decade, a generic index called Morphological Building Index (MBI) was introduced to extract buildings from high resolution satellite imagery, based on spectral information [11]. While this method was able to successfully extract buildings with irregular shape, it failed in shadowy regions and also could not extract buildings located close-by (instance extraction). A consequent study to MBI proposed a Morphological Building/Shadow Index which defined a building index as well as a shadow index, and was specifically aimed at bridging the shortcomings of the MBI method [12].

With the recent availability of strong computing systems as well as finer resolution data, artificial intelligence based deep learning algorithms such as Convolutional Neural Networks (CNNs) are being aggressively used for building extraction given their advantage of hierarchical feature extraction without losing any contextual information [13]–[15]. In general, a deep learning architecture consists of a network structure with many hidden layers leading to hierarchical feature extraction thus, eliminating the problem of inadequate representation of learning features [16]. Building-A-Nets is an adversarial network to for robust extraction of building rooftops. Multiple Feature Reuse Network (MERN) is a resource efficient rich CNN to detect building edges from high spatial resolution satellite imagery [17]. A special type of pre-trained CNN, called a Fully Convolutional Network (FCN) is also being widely used for transfer learning based building extraction. A few such popular FCNs are VGG-16 [18], ResNet [19], Deeplab [20], DenseNet [21], SegNet [22] and U-Net [23]. Studies specifically on building extraction from UAV images have also increased of late. SegNet and U-Net have been used in an ensemble manner to improve building footprint extraction from high resolution UAV imagery [24]. Techniques such as dilated spatial pyramid pooling [25], multi-stage multi-task learning [26], channel attention mechanisms [27] have been used to improve the building segmentation accuracy from UAV data. Variants of U-Net architecture have also been tested for building extraction and a studies indicate that the U-Net is the most suitable for dense image building extraction [15], [28], [29].

1.3 Objective and Summary

Sometimes, the FCN based segmentation is visually degraded in case of blurred building boundaries [30]. Moreover, high spatial resolution data is generally restricted to three or four spectral channels, which makes it difficult to differentiate buildings and other spatially similar features [24]. To address these issues, this study proposes a deep learning based segmentation approach that combines a pre-trained FCN with a U-Net being trained for building extraction, to extract buildings from high resolution RGB UAV imagery. The learning of a deep Residual Network (ResNet) trained on the ImageNet dataset is transferred to the segmentation based FCN U-Net, hence forming a combined Res-U-Net architecture. In this Res-U-Net, the pre-trained ResNet helps capture more context in case of features spatially similar to buildings while the U-Net learns building segmentation based on a unique loss function (discussed in Section 2.3) that simultaneously accounts for crispness as well as the region of a segmented building, hence preventing prediction leakage outside of feature in case of blurred boundaries. Consequent

sections of the paper discuss the dataset details, data preparation and training methodology, results and their inferences, and conclude the study.

2 Dataset Details

This study uses the Inria Aerial Image Labelling (IAIL) dataset. This dataset contains a total of 360 orthorectified images (180 for training and 180 for testing) with a tile size of 1500m^2 each, at 30cm spatial resolution with red, green and blue bands. Each image is of size 5000 x 5000 pixels. While covering an area of 81km^2/city in select US cities of Austin, Chicago, Kitsap County and select Austrian cities of Vienna and West Tyrol, this dataset contains 36 images from each city having high variance in terms of urban density and building spacing. Moreover, numerous instances of shadowy features and shadowy background are present, especially in the images from Chicago, US. The ground truth of the training set is provided as a binary feature image with only two classes namely building and non-building. Since ground truth is provided only for the training set of 180 images, we use only those 180 images to train and validate our model. Figure 1 shows the UAV image and its corresponding ground truth as available from the IAIL training set, for each of the five cities.

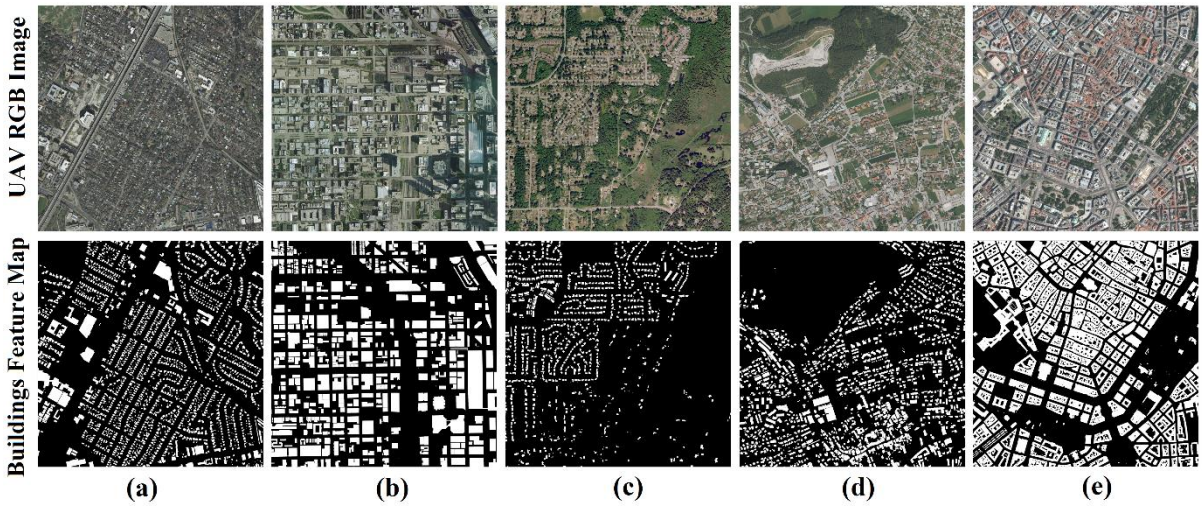


Fig. 1. Data samples from the IAIL dataset, one from each city (a) Austin, USA (b) Chicago, USA (c) Kitsap County, USA (d) West Tyrol, Austria (e) Vienna, Austria

3 Methodology

3.1 Data Preparation Methodology

A single image is of size 5000 x 5000 pixels. We further split it into small data chips of size 224 x 224 pixels in accordance to the proposed network architecture. This results into 484 such tiles from a single image. However, certain number of chips contain no buildings or hardly any buildings at all, creating a bias in the type of data which could result in model misfit. To ensure uniformity of 224 x 224 chips in terms of buildings, we further filter the 484 chips using a High Label Filter (Equation 1). This is basically a ratio of the number of labelled pixels to the total number of pixels in a 224 x 224 chip. We use a threshold of 0.3 in the High Label Filter to further filter these 484 chips. This excludes the chips having label density less than 30% and hence the earlier bias in the data is now removed. Figure 2 shows the data preparation methodology for a single image. This process is performed for all 180 images as

well as labels. Passing the 87,120 224x224 chips obtained from 5000x5000 180 images (180*484) through the High Label Filter, we get 27,164 224x224 chips. The proposed model is trained and validated on these 27,164 chips and entire images of size 5000x5000 are used for testing.

$$HLF = \frac{\sum_{i=0}^{224*224} building\_pixel_i}{\sum_{i=0}^{224*224} image\_pixel_i} \quad (1)$$

Fig 2. Data preparation methodology for a single image

3.2 Network Architecture

In this study, the U-Net architecture is implemented with a dynamic decoder to learn building extraction as a fully convolutional network (FCN). The whole architecture essentially consists of two major operations – image contraction performed by the encoder and image expansion performed by the decoder (Fig 3). The encoder is responsible for pooling out the necessary information from within the convolution kernel which is done by max pooling operations. The decoder helps preserve precise local information such as building edges in case of blurred images which is done by upsampling and convoluting over transposed kernels. Each step of encoder is connected with the corresponding inverse step of the decoder using successive skip connections. The advantage of using a dynamic network is the automatic creation of the decoder based on how the encoder is initialized [31] as well as working with almost any patch-size [32].

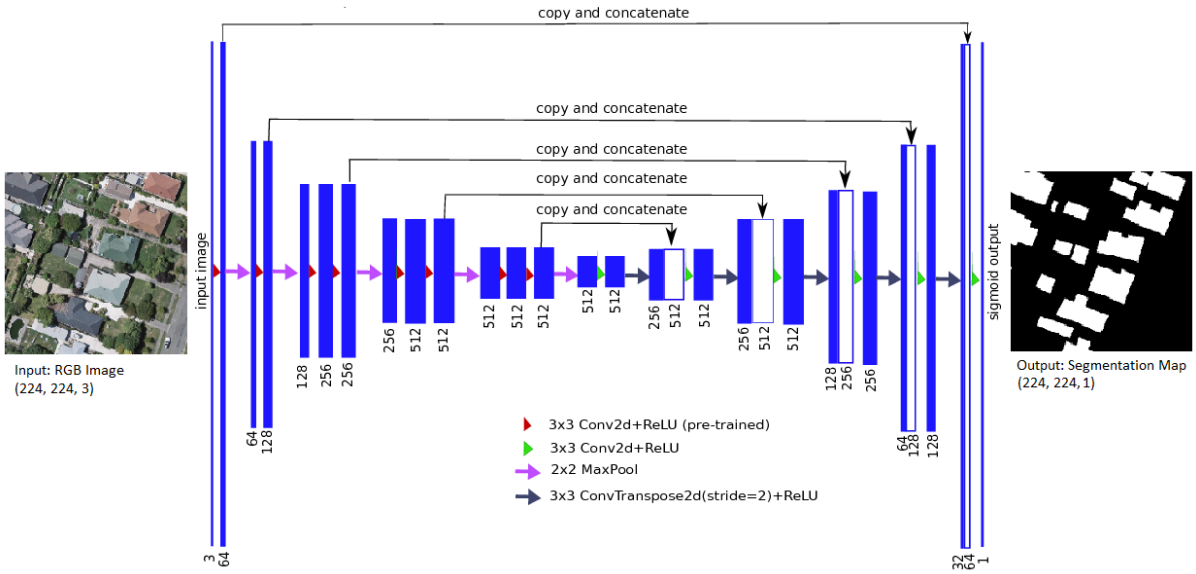


Fig 3. Proposed Res-U-Net architecture described in terms of U-Net encoders and decoders, along with the pre-trained ResNet34 layers

U-Net being an end-to-end FCN can easily initialized with the weights of a deeper CNN. We further initialize the proposed dynamic U-Net architecture with the weights of ResNet34 trained on ImageNet,

forming a Res-U-Net. The proposed Res-U-Net comprises of multiple sequential blocks as well as dynamic U-Net blocks initialized with ResNet34. Each encoder-decoder block of the architecture consists of a series of 2D batch normalization and ReLU activations which extract the trainable features from the data. Table 1 shows the specific network architecture of the proposed Res-U-Net architecture. The input to the network is an RGB image of shape (224, 224, 3) to which the network segments buildings and outputs segmented maps of shape (224, 224, 2). Here, the prediction contains two channels, one of which is a boolean array having discrete prediction for every pixel being a building or not and the other is a float32 array which contains the logit probability score for every pixel being a building. This is helpful in refining the results by further pooling the probability scores with bounded functions such as sigmoid.

Table 1. Specific proposed network architecture with each layer parameters

Layer		**Kernel Size**	**Output Shape**	**Stride**
Conv2d		7 x 7	64 x 112 x 112	2
Sequential Block 1	Conv2d	3 x 3	64 x 56 x 56	1
	Conv2d	3 x 3	64 x 56 x 56	1
	Conv2d	3 x 3	64 x 56 x 56	1
Sequential Block 2	Down Block 1	1 x 1	128 x 28 x 28	2
	Conv2d	3 x 3	128 x 28 x 28	1
	Conv2d	3 x 3	128 x 28 x 28	1
	Conv2d	3 x 3	128 x 28 x 28	1
	Conv2d	3 x 3	128 x 28 x 28	1
Sequential Block 3	Down Block 2	1 x 1	256 x 14 x 14	2
	Conv2d	3 x 3	256 x 14 x 14	1
	Conv2d	3 x 3	256 x 14 x 14	1
	Conv2d	3 x 3	256 x 14 x 14	1
	Conv2d	3 x 3	256 x 14 x 14	1
	Conv2d	3 x 3	256 x 14 x 14	1
Sequential Block 4	Down Block 3	1 x 1	512 x 7 x 7	2
	Conv2d	3 x 3	512 x 7 x 7	1
	Conv2d	3 x 3	512 x 7 x 7	1
Conv2d		3 x 3	1024 x 7 x 7	1
Conv2d		3 x 3	512 x 7 x 7	1
U-Net Block 1	Conv2d	3 x 3	1024 x 7 x 7	
	Conv2d	3 x 3	512 x 14 x 14	1
	Conv2d	3 x 3	512 x 14 x 14	1
U-Net Block 2	Conv2d	3 x 3	1024 x 14 x 14	
	Conv2d	3 x 3	384 x 28 x 28	1
	Conv2d	3 x 3	384 x 28 x 28	1
U-Net Block 3	Conv2d	3 x 3	768 x 28 x 28	
	Conv2d	3 x 3	256 x 56 x 56	1
	Conv2d	3 x 3	256 x 56 x 56	1
	Conv2d	3 x 3	512 x 56 x 56	

U-Net Block 4	Conv2d	3 x 3	128 x 112 x 112	1
	Conv2d	3 x 3	96 x 112 x 112	1
	Conv2d	3 x 3	96 x 112 x 112	1
Sequential Extension	Conv2d	3 x 3	99 x 224 x 224	1
	Conv2d	3 x 3	99 x 224 x 224	1
Conv2d		1 x 1	2 x 224 x 224	1

3.3 Training the Network

After weight initialization of the proposed Res-U-Net, transfer learning methodology was used to train for building extraction. Figure 4 shows the step-by-step training methodology. Out of 27,164 image-label pairs, the network was trained on 23,089 pairs (85%) and was validated on the remaining 4075 (15%) pairs of images and their corresponding labels. The network was trained with a batch size of 6 and a patch size of 224x224 for 30 epochs, with roughly 1200 batches being processed per epoch. The training was cut-off based on loss convergence (Figure 5(a)). The learning was carried out on nearly 20 million parameters extracted at different layers of the network. The network was optimized with ADAM optimizer at a learning rate of 0.0001 and a decay rate of 0.9.

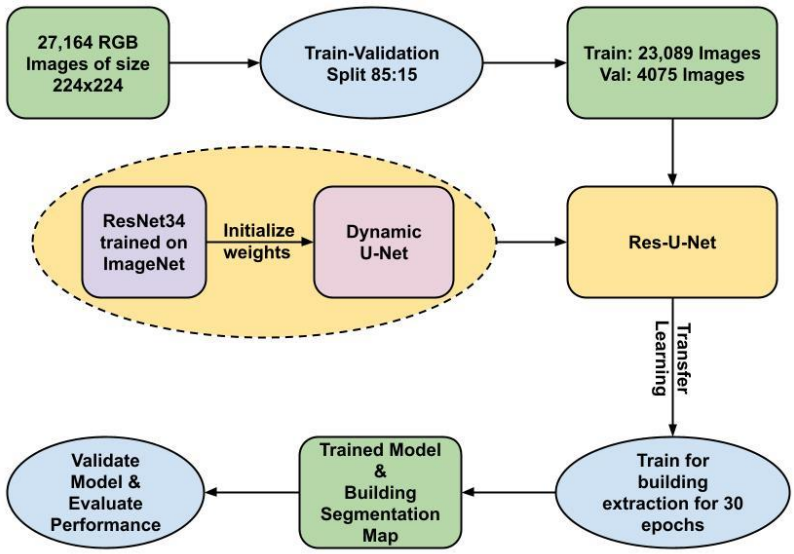


Fig. 4. Network training methodology for building extraction using transfer learning

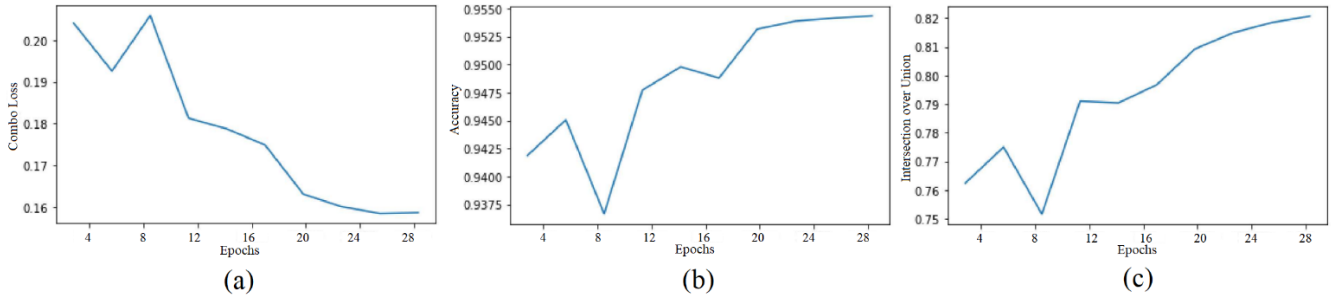
A unique combination of Binary Cross Entropy (BCE) loss (Equation 2) and dice loss (Equation 3) was used to train the network. BCE is a probability distribution based loss [33] and hence was used to minimize the entropy between the prediction and the ground truth in terms of buildings as features. It was also helpful in preserving the crispness near the boundary regions. Dice loss is a region based Intersection-over-Union like metric [34] and it was used to maximize the overlap and similarity between the predicted region and the ground truth of the feature region. Hence, a combo loss was defined (Equation 4) which focused on both boundary and region preservation. Fig. 5(a) shows the loss-based convergence of the model after 30 epochs of training. After training for 30 epochs and processing 36,000 batches the model began to converge and was saved at the end of 30 epochs with an overall accuracy of 95.7% and mean Intersection over Union (IoU) of 0.83.

$$BCE_{Loss} = -\frac{1}{patchsize}\sum_{i=1}^{patchsize} g_i \times \log p_i + (1 - g_i) \times \log(1 - p_i) \quad (2)$$

$$DiceLoss = \frac{2\times\sum_{i=0}^{patchsize} p_i g_i}{\sum_{i=0}^{patchsize} p_i^2 + \sum_{i=0}^{patchsize} g_i^2} \quad (3)$$

$$ComboLoss = BCE_{Loss} + DiceLoss \quad (4)$$

Where g = ground truth image and p = predicted image



(a) (b) (c)

Fig 5. (a) Combo loss variation, (b) Accuracy variation and (c) IoU variation in 30 epochs of training

4 Results and Discussion

Figure 6 shows the results for building extraction for select RGB images from each city of the IAIL dataset. The first column is the input to the model, the second column is the ground truth, the third column is the segmented building map as predicted by the model and the fourth column shows the evaluation of the prediction with True Positives (TP) in white, True Negatives (TN) in black, False Positives (FP) in red and False Negatives (FN) in yellow. These are original images of size 5000x5000 from the IAIL dataset. The predictions are obtained by clipping to chips of 224x224, segmenting buildings and then again merging to the original size of 5000x5000. In Figure 6 we try to show all different conditions for building extraction such as the surrounding land-cover classes, urban density, shadows etc. from each city. Figure 6(a,c,f) show successful building extraction in case of high urban density with closely spaced buildings, with rare instance segmentation challenges. Figure 6(b) shows effective building extraction even in shadowy regions. It can be noted that the shadows are not falsely classified as buildings, which has been a very popular challenge in building extraction [12]. Figure 6(a,b,f) show successful building extraction in presence of spectrally similar features such as cemented roads and parking lots as well as spatially similar features such as roads, open grounds and vegetation patches having shape similar to buildings. The model is also able to segment buildings even when the dominant land cover in the image is not urban - Figure 6(d,e) contain a large cover of vegetation, Figure 6(b,e) contain a large area of water.

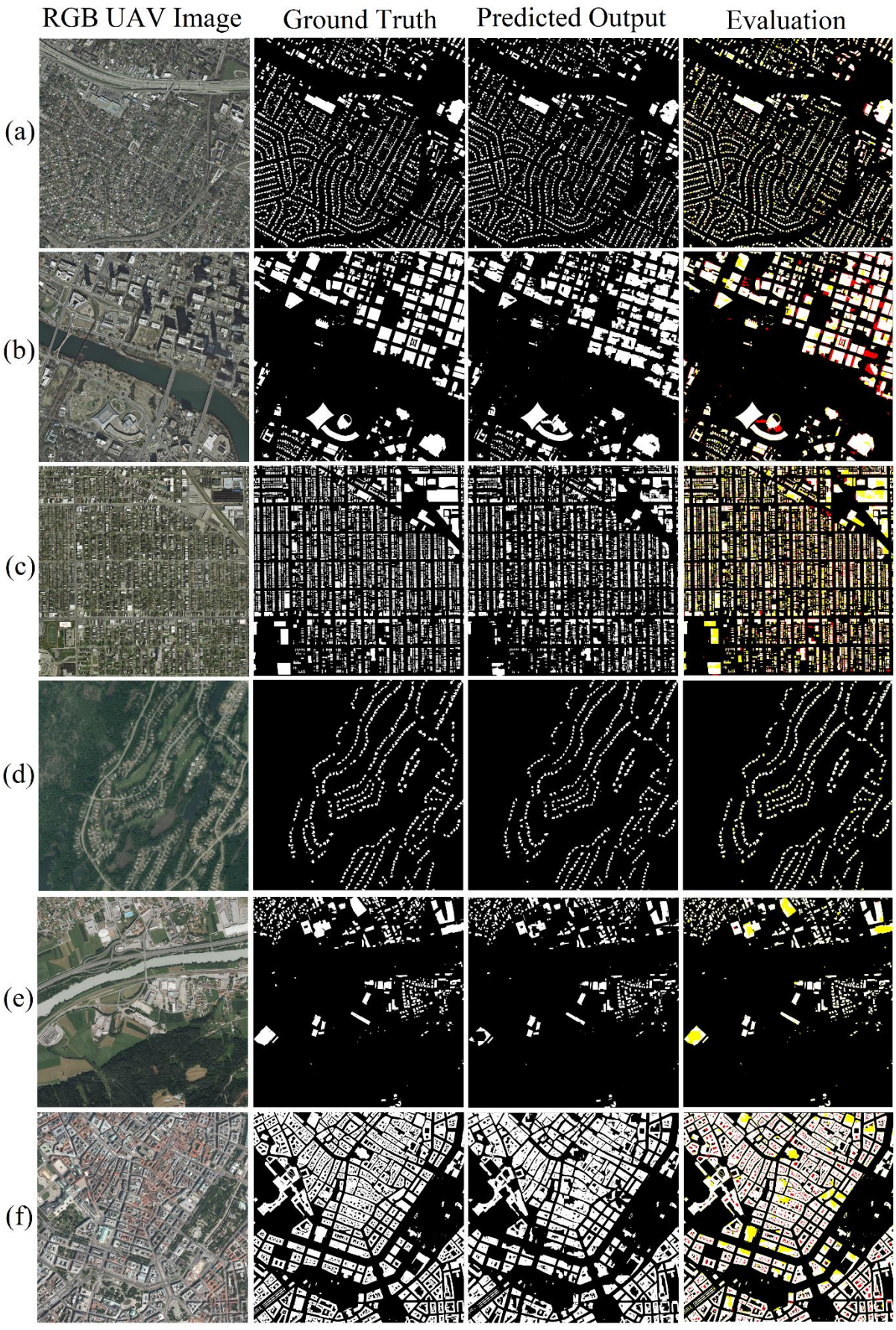

RGB UAV Image
Ground Truth
Predicted Output
Evaluation
(a)
(b)
(c)
(d)
(e)
(f)

Fig 6. Select instances of building extraction results from each city of the IAIL dataset. First column is RGB input to the model, second column is model prediction for building segmentation, third column is ground truth and fourth column is evaluation image showing TP (white), TN (black), FP (red) and FN (yellow). (a), (b) From Austin, USA (c) From Chicago, USA (d) From Kitsap County, USA (e) From Tyrol West, Austria (f) From Vienna Austria

To quantify the prediction made by the model in terms of binary segmentation, the metrics of accuracy (4), precision (5), recall (6) and F1-score (7) were used. To further perform a feature-based evaluation, object-based metrics such as branching factor (8), miss factor (9), detection percentage (10) and IoU or quality percentage (11) (otherwise also popularly known as jaccard index) were used. Table 2 shows the metrics of the individual images in Figure 6.

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (5)$$

$$precision = \frac{tp}{tp+fp} \quad (6)$$

$$recall = \frac{tp}{tp+fn} \quad (7)$$

$$f1 = 2 \times \frac{precision*recall}{precision+recall} \quad (8)$$

$$branchingFactor = \frac{fp}{tp} \quad (9)$$

$$missFactor = \frac{fn}{tp} \quad (10)$$

$$detectionPercentage = 100 \times \frac{tp}{tp+fn} \quad (11)$$

$$qualityPercentage = IoU = 100 \times \frac{tp}{tp+fn+fp} \quad (12)$$

Where tp = True Positive, fp = False Positive, tn = True Negative and fn = False Negative

Table 2. Metrics for individual images of Figure 6

	Accuracy	Precision	Recall	F1-Score	Branching Factor	Miss Factor	Detection Percentage	Quality Percentage/ IoU
Fig 6(a)	0.961	0.943	0.802	0.867	0.060	0.246	0.802	0.765
Fig 6(b)	0.942	0.850	0.764	0.859	0.177	0.152	0.868	0.752
Fig 6(c)	0.870	0.833	0.666	0.745	0.200	0.500	0.666	0.589
Fig 6(d)	0.991	0.997	0.851	0.920	0.001	0.176	0.851	0.845
Fig 6(e)	0.982	0.994	0.796	0.889	0.006	0.257	0.796	0.801
Fig 6(f)	0.927	0.938	0.893	0.915	0.066	0.120	0.893	0.843

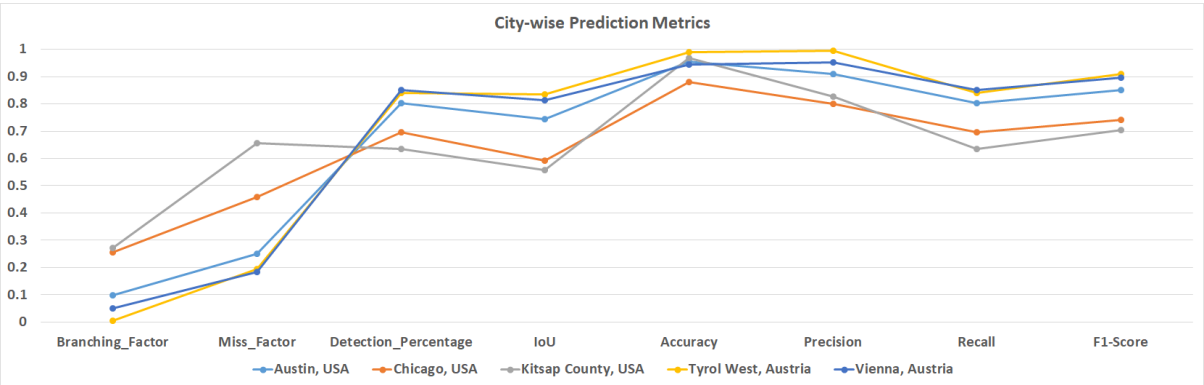


Fig 7. City-wise prediction metrics from the IAIL dataset validation part

Figure 7 shows the city-wise metrics of model validation. Tyrol West and Vienna from the IAIL dataset exhibit highly favourable conditions for building extraction. Extracting buildings from Chicago and Kitsap has been the most challenging. This is due to shadowy regions, typically the shadows being cast on other buildings. Though the proposed model successfully discriminates between shadowy regions and buildings and avoids shadows as false positives, it faces significant challenges in extracting the buildings which are under shadows. This drastically increases the rate of false negatives, as the model excludes the buildings under shadows as only shadowy regions (Figure 8(a,b)). A potential reason for this could be loss of spectral variance as well as spatial distinction of a building that is under shadow. Moreover, another isolated issue encountered in a Kitsap image is a patch of waterbody being falsely segmented as building, resulting into a high number of false positives (Figure 8(c)). This could be due to multiple reasons such as spectral similarity of the waterbody area due to turbidity, or saturation of DN values in those areas due to direct glint on sensor. Such instances of shadowed buildings and typical water areas are prominent in the images from Chicago and Kitsap and hence the extraction results are lowest for these two cities from the IAIL dataset. Figure 8 shows select instances buildings under shadows which result in a high number of false negatives.

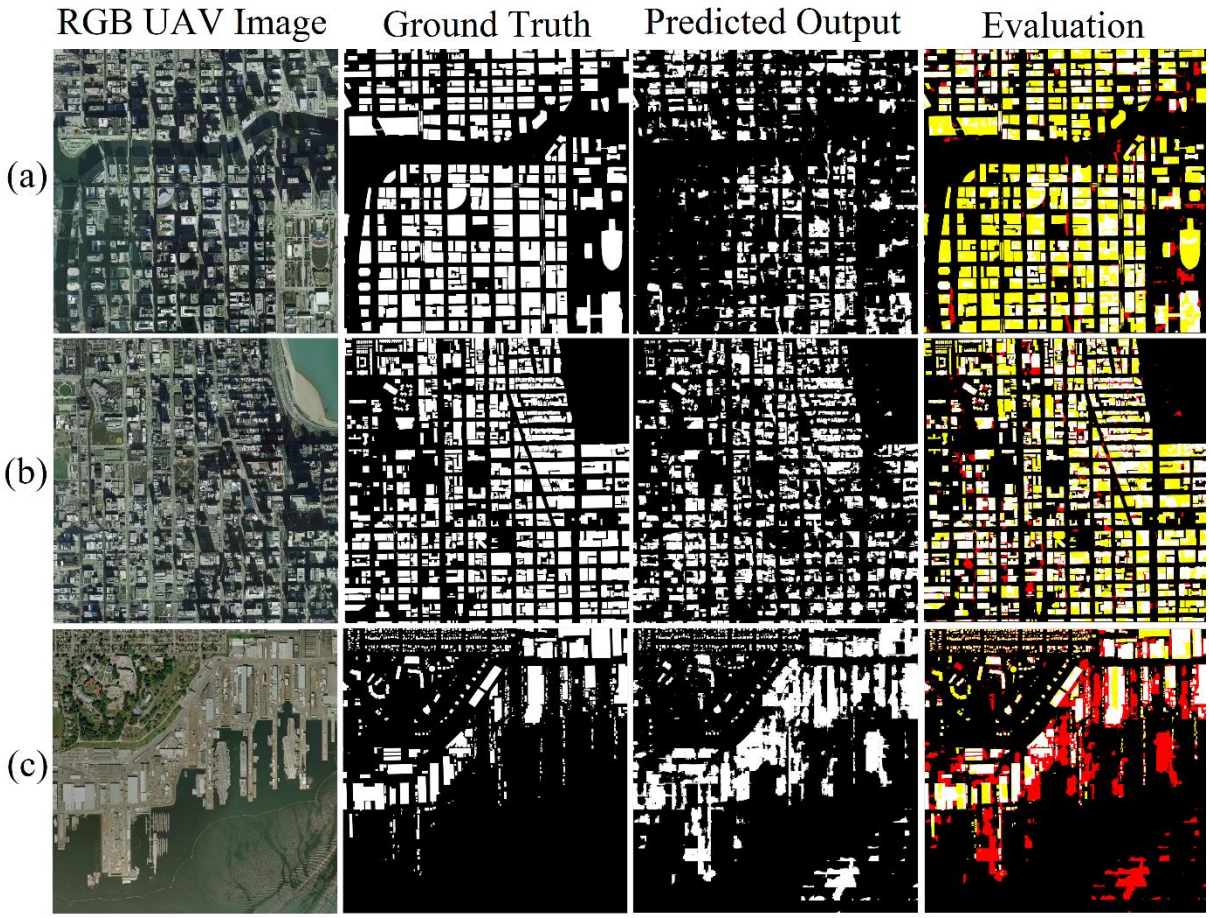


Fig 8. Select instances where buildings are covered under shadows, leading to high false negative rate. First column is RGB input to the model, second column is model prediction for building segmentation, third column is ground truth and fourth column is evaluation image showing TP (white), TN (black), FP (red) and FN (yellow). (a), (b) From Chicago, USA (c) From Kitsap county

Despite these specific challenges and rare instance segmentation issues, the overall performance of the model when evaluated on the validation set of 4075 images is highly favourable. The high values of the evaluation metrics, especially IoU, also indicate that the proposed model is able to segment buildings well within the feature edges and there is no region loss except for when the building itself is under a shadow. When compared with other deep learning based approaches, the proposed model increases the average IoU to 0.80 and average F1 score to 0.86. Table 3 shows the overall evaluation metrics of the model for the validation set as well as a comparison of those metrics with other studies on the same IAIL dataset.

Table 3. Overall metrics of the proposed approach and their comparison with existing approaches

Method	Proposed Res-U-Net	Dilated Spatial Pyramid Pooling [25]	GAN-SCA [27]	MSMT-Stage-1 [26]	AMLL [28]	Dilated CNN [29]
Overall Accuracy	0.965	0.894	0.966	0.961	0.959	0.928
Precision	0.883	-	-	-	-	-
Recall	0.861	-	-	-	-	-
Mean F1-Score	0.88	-	-	-	-	0.83
Branching Factor	0.193	-	-	-	-	-
Miss Factor	0.230	-	-	-	-	-
Detection Percentage	93.42	-	-	-	-	-
Mean IoU	0.80	-	0.777	0.733	0.725	0.710

5 Conclusion

In this research work, building extraction from UAV imagery was explored using deep learning and transfer learning methodology. A Res-U-Net architecture consisting of U-Net blocks initialized with pre-trained ResNet34 weights and was used to learn building extraction from the IAIL dataset. The combination of ResNet and U-Net was used in an attempt to overcome the problems of blurred building boundaries and limited spectral resolution in building extraction. Moreover, a combined loss function that accounts both for the building region as well as building boundaries was used to train the proposed Res-U-Net. The model was trained and validated on 180 images from across five different cities of US and Austria. These images depicted high variance in terms of urban density and dominant land cover of the image. The proposed model was successfully able to segment buildings in all cases with rare instance segmentation issues. Model performance was measured using quantitative metrics of confusion matrix as well as object based metrics such as branching factor, miss factor and IoU. When comparing these metrics with those of existing deep learning based methods, highly favourable results were noted. Specific challenges such as extracting buildings lying under shadow and excluding turbid/active waterbody as a building were also identified and are open for research.

References

[1] H. Momm and G. Easso, "Feature Extraction from High-Resolution Remotely Sensed Imagery using Evolutionary Computation," *Evol. Algorithms*, no. October 2014, 2011, doi: 10.5772/15915.

[2] L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, and J. Calpe, "Semi-supervised image classification with huberized Laplacian Support Vector Machines," *IEEE Geosci. Remote Sens.*

Lett., vol. 5, no. 3, pp. 336–340, 2008, doi: 10.1109/ICET.2013.6743545.

[3] J. A. Benediktsson, M. Pesaresi, and K. Arnason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9 PART I, pp. 1940–1949, 2003, doi: 10.1109/TGRS.2003.814625.

[4] A. Sowmya and J. Trinder, "Modelling and representation issues in automated feature extraction from aerial and satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 55, no. 1, pp. 34–47, 2000, doi: 10.1016/S0924-2716(99)00040-4.

[5] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep Learning Based Feature Selection for Remote Sensing Scene Classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015, doi: 10.1109/LGRS.2015.2475299.

[6] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "BUILDING EXTRACTION FROM MULTI-SOURCE REMOTE SENSING IMAGES VIA DEEP DECONVOLUTION NEURAL NETWORKS Zuming Huang , Guangliang Cheng , Hongzhen Wang , Haichang Li , Limin Shi , Chunhong Pan National Laboratory of Pattern Recognition (NLPR) Institute of Au," *2016 IEEE Int. Geosci. Remote Sens. Symp.*, pp. 1835–1838, 2016.

[7] V. Karathanassi, C. Iossifidis, and D. Rokos, "A texture-based classification method for classifying built areas according to their density," *Int. J. Remote Sens.*, vol. 21, no. 9, pp. 1807–1823, 2000, doi: 10.1080/014311600209751.

[8] R. Eskandarpour and A. Khodaei, "Leveraging accuracy-uncertainty tradeoff in SVM to achieve highly accurate outage predictions," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 1139–1141, Jan. 2018, doi: 10.1109/TPWRS.2017.2759061.

[9] W. Li, C. He, J. Fang, and H. Fu, "Semantic segmentation based building extraction method using multi-source GIS map datasets and satellite imagery," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Dec. 2018, vol. 2018-June, pp. 233–236, doi: 10.1109/CVPRW.2018.00043.

[10] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, 2013, doi: 10.1109/TGRS.2012.2200689.

[11] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectralgeoeye-1 imagery," *Photogramm. Eng. Remote Sensing*, vol. 77, no. 7, pp. 721–732, 2011, doi: 10.14358/PERS.77.7.721.

[12] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 5, no. 1, pp. 161–172, 2012, doi: 10.1109/JSTARS.2011.2168195.

[13] X. X. Zhu *et al.*, "Deep learning in remote sensing: a review," no. december, 2017, doi: 10.1109/MGRS.2017.2762307.

[14] L. Zhang, G. S. Xia, T. Wu, L. Lin, and X. C. Tai, "Deep Learning for Remote Sensing Image Understanding," *J. Sensors*, vol. 2016, no. june, 2016, doi: 10.1155/2016/7954154.

[15] F. Erdem and U. Avdan, "Comparison of Different U-Net Models for Building Extraction from High-Resolution Aerial Imagery," *Int. J. Environ. Geoinformatics*, vol. 7, no. 3, pp. 221–227, 2020, doi: 10.30897/ijegeo.684951.

[16] F. Chollet, *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG, 2018.

[17] L. Li, J. Liang, M. Weng, and H. Zhu, "A Multiple-Feature Reuse Network to Extract Buildings from Remote Sensing Imagery," *Remote Sens. 2018, Vol. 10, Page 1350*, vol. 10, no. 9, p. 1350, Aug. 2018, doi: 10.3390/RS10091350.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2015, Accessed: Jan. 11, 2021. [Online]. Available:

http://www.robots.ox.ac.uk/.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2016, vol. 2016-December, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[20] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Feb. 2018, vol. 11211 LNCS, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[21] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building Extraction in Very High Resolution Imagery by Dense-Attention Networks," *Remote Sens.*, vol. 10, no. 11, p. 1768, Nov. 2018, doi: 10.3390/rs10111768.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, May 2015, vol. 9351, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[24] A. Abdollahi, B. Pradhan, and A. M. Alamri, "An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images," *Geocarto Int.*, vol. 0, no. 0, p. 000, 2020, doi: 10.1080/10106049.2020.1856199.

[25] M. E. Morocho-cayamcela, "Increasing the Segmentation Accuracy of Aerial Images with Dilated Spatial Pyramid Pooling," *ELCVIA Electron. Lett. Comput. Vis. image Anal.*, vol. 19, no. 2, pp. 17–21, 2020.

[26] A. Marcu, D. Costea, E. Slusanschi, and M. Leordeanu, "A multi-stage multi-task neural network for aerial scene interpretation and geolocalization," *arXiv*, no. March 2019, 2018.

[27] X. Pan *et al.*, "Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms," *Remote Sens.*, vol. 11, no. 8, p. 917, 2019, doi: 10.3390/rs11080917.

[28] B. Huang *et al.*, "Large-scale semantic classification: Outcome of the first year of inria aerial image labeling benchmark," *Int. Geosci. Remote Sens. Symp.*, vol. 2018-July, pp. 6947–6950, 2018, doi: 10.1109/IGARSS.2018.8518525.

[29] M. Khoshboresh-Masouleh, F. Alidoost, and H. Arefi, "Multiscale building segmentation based on deep learning for remote sensing RGB images from different sensors," *J. Appl. Remote Sens.*, vol. 14, no. 03, p. 1, 2020, doi: 10.1117/1.jrs.14.034503.

[30] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-Resolution Aerial Image Labeling with Convolutional Neural Networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, 2017, doi: 10.1109/TGRS.2017.2740362.

[31] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation," *arXiv*, Jan. 2018, Accessed: Mar. 01, 2021. [Online]. Available: http://arxiv.org/abs/1801.05746.

[32] H. Lamba, "Understanding Semantic Segmentation with UNET," *Towards Data Science*, Feb. 17, 2019. https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47 (accessed Jan. 21, 2021).

[33] Z. Zhang and M. R. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," 2018.

[34] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," *Lect. Notes Comput. Sci.*

(including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10553 LNCS, pp. 240–248, Jul. 2017, doi: 10.1007/978-3-319-67558-9_28.