Scientific Research Publishing

Advances in Remote Sensing, 2023, 12, 71-87
https://www.scirp.org/journal/ars
ISSN Online: 2169-2688
ISSN Print: 2169-267X

Building Detection and Counting in Convoluted Areas Using Multiclass Datasets with Unmanned Aerial Vehicles (UAVs) Imagery

Shital Adhikari, Vaghawan Prasad Ojha

IKebana Solutions LLC, Tokyo, Japan
Email: shital.adhikari@ekbana.info, vaghawan.ojha@ekbana.net

How to cite this paper: Adhikari, S. and Ojha, V.P. (2023) Building Detection and Counting in Convoluted Areas Using Multiclass Datasets with Unmanned Aerial Vehicles (UAVs) Imagery. *Advances in Remote Sensing*, **12**, 71-87.
https://doi.org/10.4236/ars.2023.123004

Received: June 7, 2023
Accepted: August 18, 2023
Published: August 21, 2023

Abstract

This paper studies the effect of breaking single-class building data into multi-class building data for semantic segmentation under end-to-end architecture such as UNet, UNet++, DeepLabV3, and DeepLabv3+. Although, the already existing semantic segmentation methods for building detection work on the imagery of developed world, where the buildings are highly structured and there is a clearly distinguishable space present between the building instances, the same methods do not work as effectively on the developing world where there is often no clear differentiable spaces between instances of building thus reducing the number of detected instances. Hence as a noble approach, we have added building contours as new class along with building segmentation data, and detected the building contours and the inner building regions, hence giving the precise number of buildings existing in the input imagery especially in the convoluted areas where the boundary between the buildings are often hard to determine even for human eyes. Breaking down the building data into multi-class data increased the building detection precision and recall. This is useful in building detection where building instances are convoluted and are difficult for bare instance segmentation to detect all the instances.

Keywords

Multi-Class Segmentation, Building Segmentation, Remote Sensing, Semantic Segmentation, UNet

1. Introduction

Remote sensing images have a wide range of applications, including monitoring

and counting wild-life [1], detecting and classifying vegetation from grasslands and heavily forested areas [2], mining land cover classification [3], flood extent mapping [4], and multi-feature building extraction in urban area [5]. Most of the applications are based on high-resolution images which can be used for the pixel-level classification known as semantic classification [6] [7] [8]. The higher resolution Unmanned Aerial Vehicle (UAV) images can create better semantically segmented images [9] which help to understand the contents in the given images. Despite the higher resolution, it lacks spectral information of ground objects which is inherently available in satellite imagery, and thus possesses difficulty in segmentation [10].

The deep learning methodologies have been studied extensively in remote sensing [11]. With the rise of Deep Convolution Neural Networks (DCNN), they have been widely used in object detection and segmentation [10]. Architecture like UNet, based on DCNN, can be used to extract spatial features of required regions. The building region detection can also be treated as a feature detection problem where we can extract the features of the buildings using a Convolution Neural Network (CNN) in encoder-decoder network architecture [10].

In order to tackle this problem, many deep learning architectures [12] [13] extract the building's mask from the given satellite imagery. The building extraction can also be treated as an instance segmentation [14] problem where the building localization follows the segmentation. However, this method is computationally intensive because it requires running each instance of the building detection process every time. Furthermore, we can treat this problem as a semantic segmentation problem [15], but this gives us a very vague representation of building count. Especially in convoluted buildings, the identification of individual buildings is cumbersome. On top of that, the detected buildings lack a clear boundary between the buildings. While the existing methodologies [16] [17] for building segmentation work well in developed countries, where buildings are typically well-defined and distinct, they may not be suitable for developing countries where buildings are often disorganized and complex, making it difficult even for humans to differentiate. This possesses challenges when automating building segmentation using semantic segmentation for humanitarian aid, population estimation, or urban density calculations.

Existing methodologies have treated building detection as a binary classification problem between building inner segments and the background [18]. Since buildings have a clear boundary between the inner and the outer regions, we can treat it as a multi-class classification problem where the boundary of a building, inner regions, and background can be treated as different classes on their own. This converts the existing binary segmentation problem into multi-class segmentation. It can be applied to all existing datasets with minor modifications. The use of edge detection can help to separate a merged building into separate buildings.

Treating the building segmentation as multi-class segmentation, we increased

the building count while preserving the inner building segmentation regions. The erosion process in OpenCV [19] is applied to the building mask, resulting in the creation of inner regions. The new class for the multiclass dataset is defined by taking the difference between the original building mask and these inner regions. Further, we can use the existing model architecture, which does not significantly increase computation complexities. Moreover, the features regarding the building exterior are also already learned during the binary segmentation so, it does not pose any problem with the network convergence instead it converges faster to similar precision and recall metrics.

In summary, we experimented with multiple architectures with joint contour and structure learning. Our experiment showed a clear efficiency in building detection on often complex and convoluted areas, increasing the precision and recall on datasets such as Opencities [20] and Inria building dataset [21] along with our own dataset. The rest of the paper is organized as follows: Section 2 summarizes the previous work on building detection. Section 3 describes the proposed method, while Section 4 provides implementation details. In Section 5, experimental results are presented, followed by Section 6, 7, and 8, which include discussions, challenges, and conclusions, respectively.

2. Literature Review

Building segmentation from satellite imagery has been rigorously studied over the decade [9] [10] [18]. The abundance of data has led to the use of numerous machine-learning methods for building detection [12] [13]. The majority of the building detection techniques can be classified into either classical or deep learning approaches. While visually identifying various simple and complex building patterns may be easy, classifying buildings in remote sensing based on their diverse patterns and styles proves to be challenging for classical machine learning algorithms. Traditional remote sensing image processing methods like Support Vector Machine (SVM) do not perform well on UAV images, as they require a training sample from the multiple variations of building datasets [22]. Also, the building is extracted from high-resolution images using Normalized Difference Vegetation Index (NDVI) indices [23] or even further custom indices like Morphological Building indices (MBI) have been developed [24]. Such methods are prone to errors due to variations in the building's characters and properties. Buildings extractions from the top-view have also been evaluated to be prone to building complexities [25]. The building segmentation from high-resolution images is also done based on the binary mathematical morphology (MM) operator [26]. Furthermore, direct building extraction by ensembling models trained on multi-spectral images, OpenStreetMap (OSM) dataset, and the RGB images have also been experimented with [27]. The building can be extracted by combining results from two models trained on high-resolution satellite imagery and Digital Surface Model (DSM) data from the LIDAR dataset [28].

The use of convolution neural networks has created state-of-the-art results on object identification, classification, and segmentation [29]. Deep learning architecture like Mask RCNN [30] is used for object detection and instance segmentation which is based on region proposal while architecture like UNet [31] is used in semantic segmentation originally used in the small datasets of biomedical images. On the other hand, UNet can extract feature masks from an input image that is equal to the original image in spatial resolutions. Further, the features extracted from the initial feature layers are also appended to the later layer preserving the information of predictions. Despite having fewer features, the UNet architecture consistently achieves state-of-the-art results in medical imagery segmentation [32]. For segmentation, medical images and satellite images, both have an issue of data deficiency [33] [34]. UNet has also been widely used in satellite imagery by many precursors. UNet architecture has been found to be beneficial in many competitions like Spacenet building detection where it achieved the leading score in the competition [35].

To enhance building detection and segmentation, the authors employed multiple UNet architectures and incorporated an attention layer to generate the final mask [36]. Here, the instance segmentation of the building is done using a multiple UNet where each model learns building contours, building regions, and background which is mixed later on. MAP-Net has been used for building segmentation [37] where the author tried to learn both features like building edges and the inner building regions from the same image.

3. Proposed Method

3.1. Model Architecture

We divided the ground truth into single-class and multi-class building datasets, effectively increasing the number of classes. In order to create a multi-class dataset from the single-class dataset, we applied an erosion operation to the building mask using a 15 × 15 pixel kernel as specified by Equation (1). This process generated eroded regions, which were assigned as a new building class, while the disparity between the inner building and the original mask was designated as another new class. Subsequently, we divided the region into three distinct classes: background, building edges, and inner building regions. This allows the segmentation architecture to learn edge information separately, thereby increasing the count of identified buildings. The following equation shows the erosion operation [19]

$$dst(x,y) = \min_{(x',y'):element(x',y')\neq 0} src(x+x', y+y') \tag{1}$$

To validate our methods, we experimented with UNet [31], UNet++ [38], DeepLabV3 [39], and DeepLabv3+ [40] architectures using efficientnet-b0 [41] as the encoder trained on imagenet datasets [42].

We used the UNet architecture as the baseline architecture for experimentation. The network architecture consists of an encoding network on the left and a

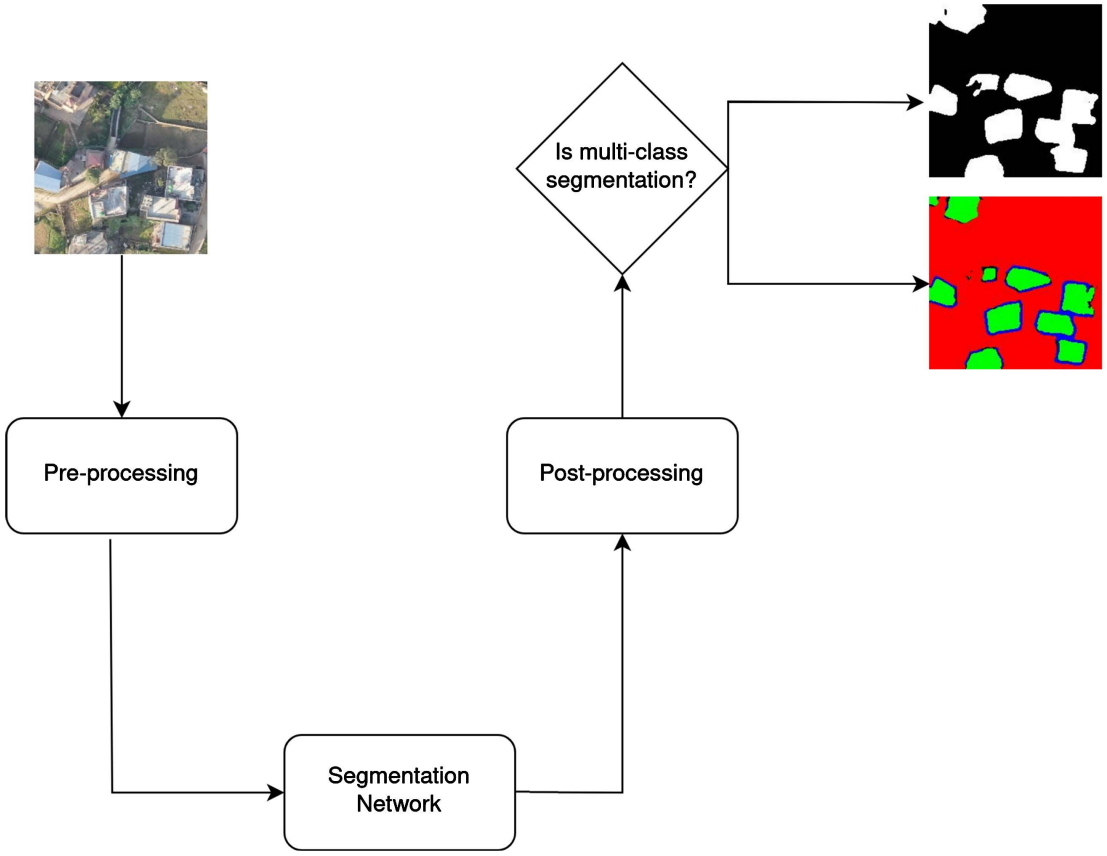

Figure 1. Overall system architecture: Multi-class segmentation.

decoder network on the right side. The input image is convoluted using a 3 × 3 convolution layer followed by ReLU and 2 × 2 max-pooling with stride 2 for downsampling. During upsampling, 2 × 2 up-convolution is used followed by 3 × 3 convolution and ReLU. During concatenation, zero padding is used to match the size of the feature. To match the number of features, a 1 × 1 convolution layer is used.

3.2. Loss Function and Evaluation Metrics

We used dice loss [43] instead of Intersection over Union (IOU) [44] as the loss function, as it results in higher consistency between the predicted segmentation mask and the labels, without favoring common regions. The following is the formula for dice loss [43],

$$diceloss = 1 - \frac{2\sum_{i=1}^{N} y_i * \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \tag{2}$$

We employed cross-entropy loss [31] in addition to dice loss to detect differences between the predicted and original masks. The binary cross-entropy loss was utilized for binary segmentation, while the cross-entropy loss was used for multi-class segmentation. To assess the predicted results, we used the dice coefficient [43] to measure the similarities among the building's datasets. Addition-

ally, the building mask was assessed using precision and recall [37] [45], which were computed based on the number of building counts in each image slice.

$$CrossEntropy = -\sum_{k}^{K} y^{(k)} \log \hat{y}^{(k)} \tag{3}$$

$$TotalLoss = diceloss + CrossEntropy \tag{4}$$

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

4. Implementation Details

4.1. Training Pipeline

We used an image size of 224 × 224 which we resized from the original image of 1024 × 1024. A batch size of 8 was used for the 20 epochs with an initial learning rate of 0.001 which decreases by a factor of 10 at each 30% of epochs with SGD optimization. We used the threshold pixel of 100 pixel area in the image for the building size.

Training Configurations, all the models use *efficient-net-b0* as an encoder which is trained on Imagenet dataset [42]. We implemented the model using Python and PyTorch [46] for deep learning model architecture using two GPUs (NVIDIA RTX 3090), CPU (i9 10th gen), and RAM 32GB.

We also implemented the model parallelism using the PyTorch data parallelism [46] pipeline which enabled us to experiment with larger models and large batch sizes. Although it did not improve the training time, this allowed us to train larger models. During the training phase, the datasets were split into multiple GPUs and merged on each epoch.

4.3. Datasets

The experimentation of segmenting the building was carried out in the UAV imagery of Kathmandu, Nepal. The area covered by the dataset is sparsely populated with medium size houses. Along with it, there are tunnels that are used for farming in the regions. We also used the regional datasets from the Nima regions of Accra, Ghana [20] which has densely populated buildings. This region is represented by *tier_1_source_acc_d41d81* in the open cities tier1 sample datasets. The building structures were arduous for humans to identify building instances due to limited space between building roof structures. Furthermore, to validate our methodology, we also utilized the Inria Aerial Image Labeling Dataset [21].

While datasets from Kathmandu regions have been sliced in the 850 × 850 px RGB images, the open cities datasets are sliced on 1024 × 1024 px RGB images, and inria aerial images are randomly cropped at 224 × 224 px since they were less magnified. This variation in image sizes and cropping methods ensured that the datasets were appropriately prepared for the building segmentation task in

different regions. The corresponding mask is created from each of these slices and the segmentation error produced while creating a raster image has been reduced to a minimum. Thus created samples are separated on the training and testing samples with 80% and 20% ratio and the validation size is taken as 20% of the training sample with all the images resized during training.

4.4. Data Augmentation

We used the Albumentation [47] library for dataset augmentation, which involved rotation, brightness and contrast manipulation, and RGB shift operations. Augmentation increases the diversity of datasets thus improving the generalization capability of the model [48], leading to a better performance in building detection. The image was subjected to rotation with a 50% probability, while variations in brightness and contrast ranged from 0 to 20%. Additionally, horizontal and vertical flips, as well as color shifts within the range of 0 to 20%, were performed with a 50% probability.

4.5. Post Processing

Once the image are predicted using the above pipeline, all the images are post-processed to remove any noisy building regions. We use binary thresholding and created the contour from the given segmentation mask using the marching square algorithm [49] using skimage [50] library. Then created contour is processed using Ramers Douglas Peuckar algorithm [51] which minimizes the number of points for the contour. The predicted contour is considered building if its pixelated area exceeds 100 pixels. This threshold is calculated from the smallest building contour area in the given dataset. Once the building cases are identified, we calculate the precision and recall only if they cover an area greater than the minimum building threshold and have an IOU score greater than 0.5. In all the following cases, building recall and building precision are calculated based on the number of buildings following the previous assumptions.

5. Experiments & Results

We trained UNet, UNet++, DeepLabV3, and DeepLabV3+ on datasets from regions of Kathmandu, Opencities datasets [32], and Inria Aerial Image Labeling Dataset [46] individually with image resized to 224 using *efficientnet-b0* as an encoder pre-trained on ImageNet for 20 epochs. All the models had plateaued building precision and building recall scores at 20 epochs. Data was grouped into binary segmentation and multiclass classification which consists of background, building contour, and inner building segments.

The segmentation results on the regions of Kathmandu, the Open Cities datasets and Inria building datasets are displayed in **Figures 2-4** using four columns. There has been an increase in the number of buildings, and it is apparent that there is a distinct separation between building boundaries in the multi-class scenario compared to the single-class scenario.

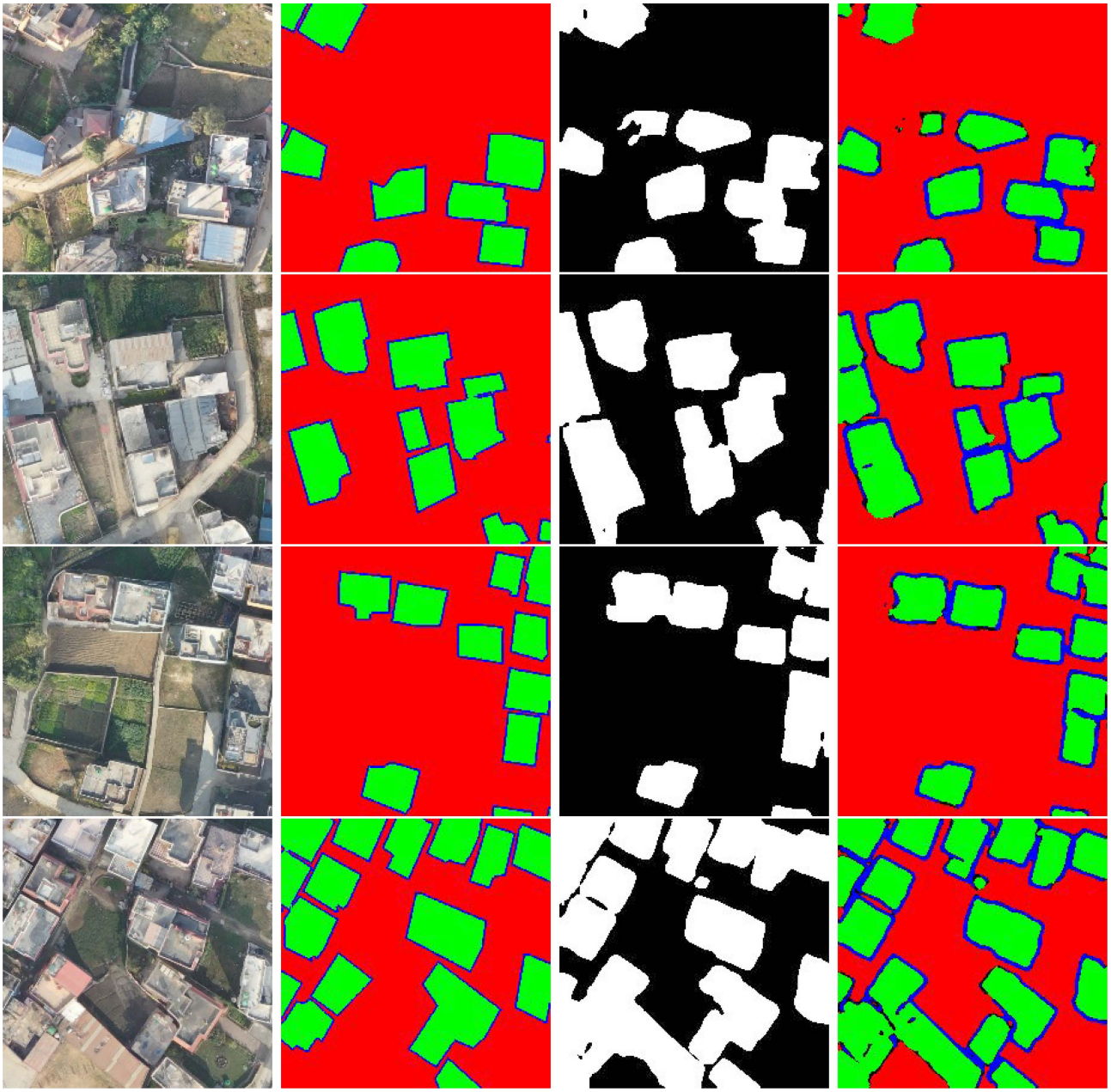

Figure 2. Using UNet, Ground Image, Mask, Predicted mask in a single channel, Predicted mask with three channel on Kathmandu regions.

Similarly, the evaluation metrics for the regions of Kathmandu, the Opencities datasets, and the Inria Aerial Image Labeling Dataset on different architectures are shown in **Tables 1-3**. The dice loss is computed by comparing the actual building pixels with the predicted building pixels. Building precision and Building Recall are determined by comparing the building count in the actual image with that in the predicted image. IOU and Accuracy are calculated based on pixel counts for comparison with existing results.

In terms of building recall for multi-class segmentation, Unet++ demonstrated the best performance, achieving a recall rate of 72% on the Opencities dataset

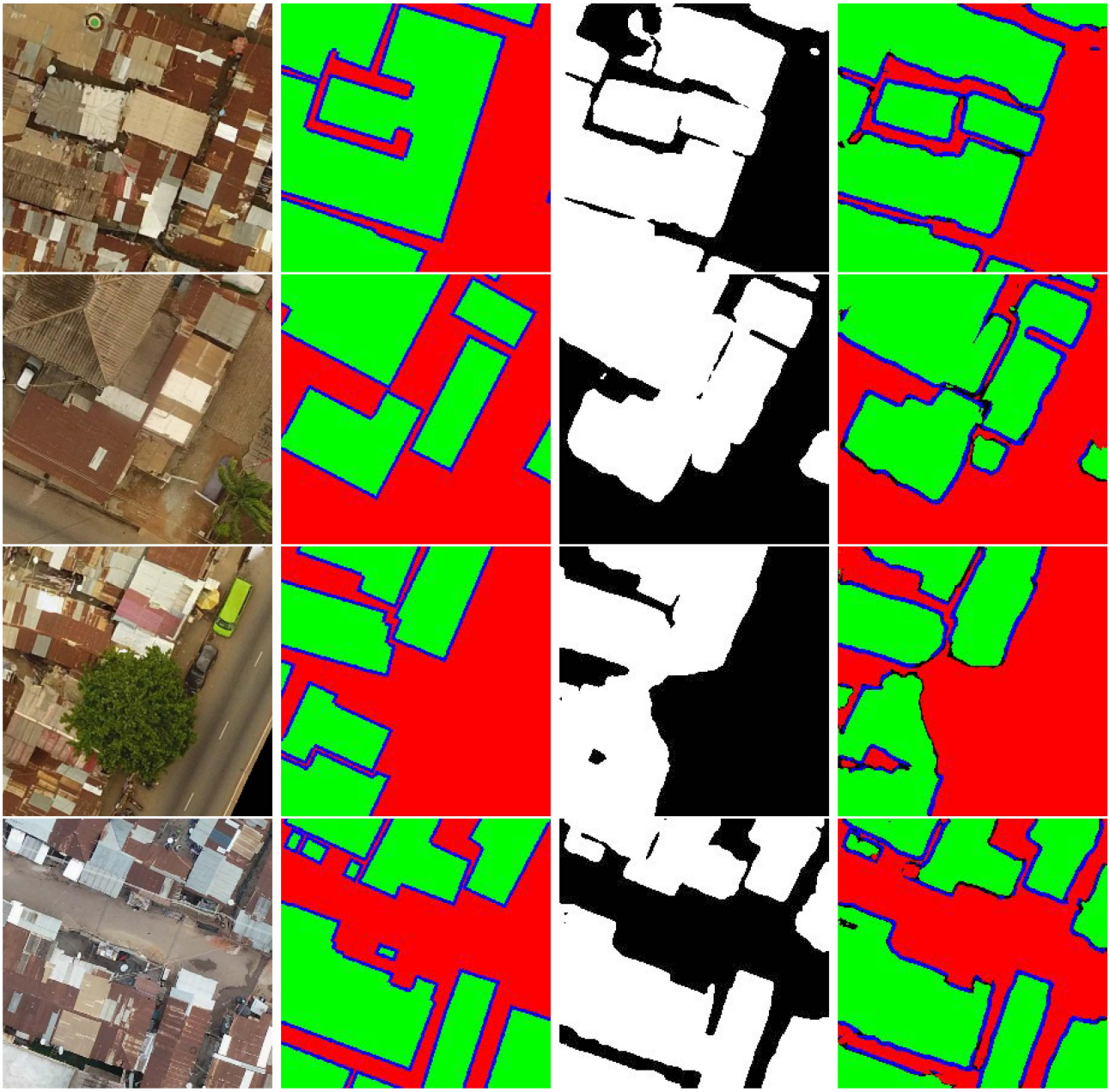

Figure 3. Using UNet: Ground Image (first column), Ground Truth Mask (second column), Predicted mask in a single channel (third column), Predicted mask with three channel on Open Cities Dataset (last column).

and 67% on the Inria dataset. On the other hand, UNet achieved a higher recall rate of 76% on the region of Kathmandu dataset. Notably, when transitioning from a single class to a multiclass setup, Deeplabv3+ exhibited the most significant improvement in building recall, with an impressive increase of 20%. These findings highlight the effectiveness of different architectures in accurately identifying and segmenting buildings in various datasets.

All model architectures achieved an average accuracy of 95% on both single and multiclass datasets, which is comparable to the existing leaderboards on the Inria Building dataset. Similarly, the IOU values obtained were also comparable.

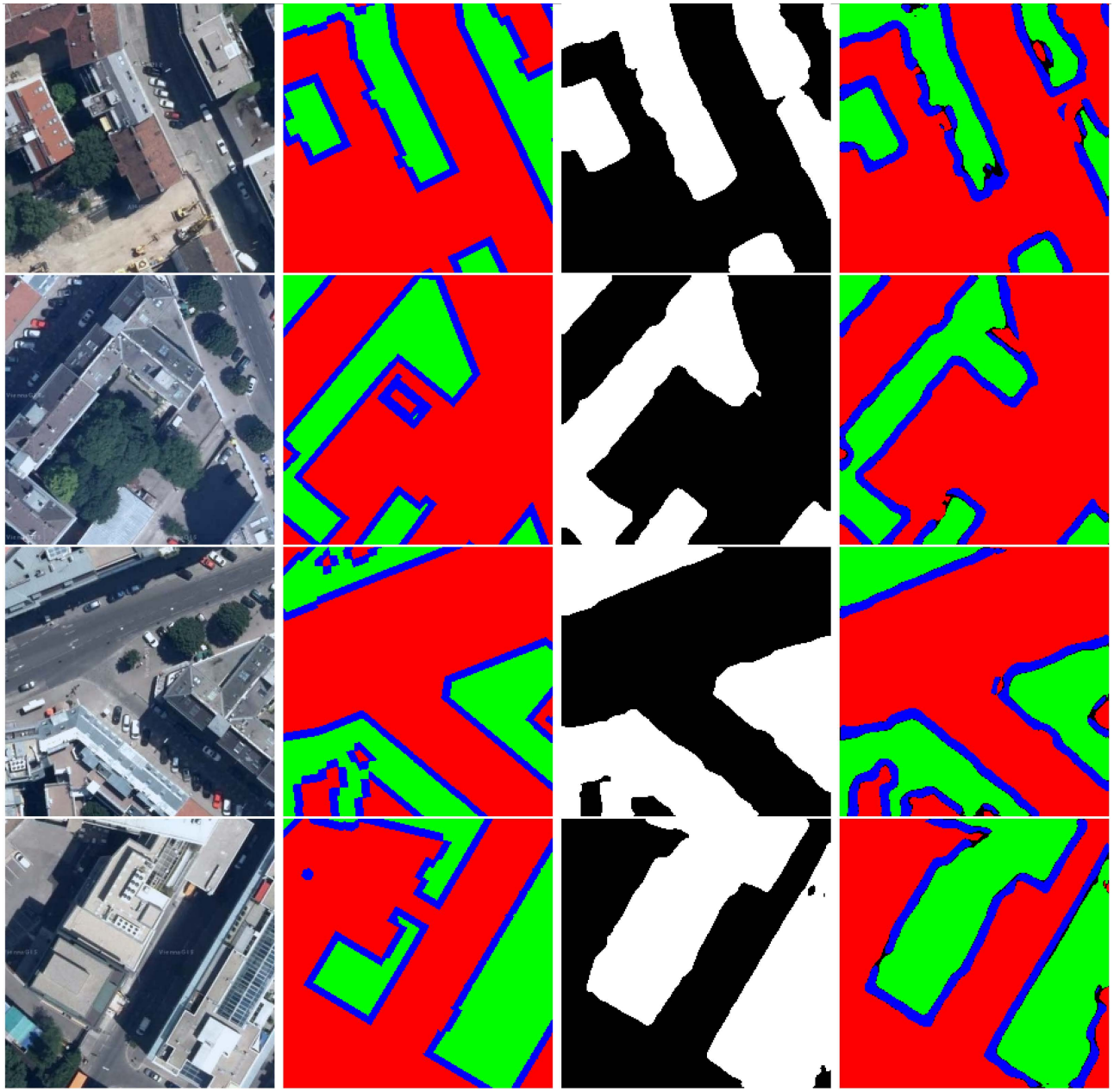

Figure 4. Using UNet: Ground Image (first column), Ground Truth Mask (second column), Predicted mask in a single channel (third column), Predicted mask with three channels on Inria Building Dataset (last column)

These results indicate that our models performed at a similar level of accuracy as the established benchmarks in the field.

We conducted a comparison of F1 scores among various model architectures on the Inria Building Datasets in **Table 5**. Specifically, we evaluated the performance using an IOU (Intersection over Union) threshold of 0.5, which is the same threshold that [10] uses. The comparison revealed that the utilization of multi-class datasets resulted in improved F1 scores compared to single-class datasets. DeepLabV3+ model exhibited the highest improvement of 12% in F1 score, while the Unet model showed a comparatively lower improvement of 3% in F1 score.

Table 1. Result on Kathmandu region.

Model Architecture	Dice Loss		Building Precision		Building Recall		IOU		Accuracy	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
UNet	0.107	0.109	0.756	0.734	0.698	0.767	0.846	0.835	0.944	0.946
UNet++	0.104	0.09	0.771	0.746	0.683	0.759	0.832	0.813	0.937	0.942
DeepLabV3	0.098	0.095	0.795	0.822	0.620	0.678	0.814	0.810	0.932	0.939
DeepLabV3+	0.086	0.106	0.776	0.758	0.626	0.694	0.825	0.808	0.940	0.942

Table 2. Result on Opencities dataset.

Model Architecture	Dice Loss		Building Precision		Building Recall		IOU		Accuracy	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
UNet	0.165	0.167	0.664	0.681	0.621	0.702	0.629	0.650	0.851	0.852
UNet++	0.163	0.167	0.651	0.695	0.617	0.721	0.625	0.638	0.834	0.856
DeepLabV3	0.165	0.175	0.597	0.673	0.524	0.699	0.636	0.579	0.874	0.851
DeepLabv3+	0.172	0.174	0.601	0.678	0.553	0.679	0.644	0.658	0.820	0.850

Table 3. Result on Inria building dataset.

Model Architecture	Dice Loss		Building Precision		Building Recall		IOU		Accuracy	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
UNet	0.184	0.287	0.721	0.704	0.552	0.629	0.697	0.575	0.950	0.953
UNet++	0.173	0.222	0.712	0.721	0.569	0.675	0.712	0.649	0.948	0.961
DeepLabV3	0.190	0.230	0.757	0.735	0.528	0.654	0.677	0.637	0.947	0.960
DeepLabv3+	0.178	0.239	0.671	0.674	0.469	0.671	0.694	0.613	0.944	0.954

Table 4. Building count metrics using UNet architecture on different datasets.

Datasets	True Positive		False Positive		False Negative	
	Single	Multi	Single	Multi	Single	Multi
Inria Building dataset	1251	1080	558	488	978	610
Kathmandu Dataset	3358	3980	1111	1442	1506	1217
Opencities dataset	450	569	238	268	293	184

Table 5. Building instance F1 score on Inria building dataset.

Model Architecture	F1 Score		Improvement
	Single	Multi	
Unet	0.625	0.664	0.039
UNet++	0.633	0.697	0.064
DeepLabV3	0.622	0.692	0.07

Continued

DeepLabV3+	0.552	0.672	0.12
MapNet		0.755	NA
Joint Learning [9]		0.776	NA

6. Discussion

We conducted inference on a randomly selected subset of the test dataset to calculate the metrics presented in Building Counts using Unet. The table reveals a significant reduction in false negative cases when utilizing a multi-class dataset as opposed to a single-class dataset, resulting in improved building recall results. Conversely, false positive cases have increased, which can be attributed to the absence of a comprehensive ground truth dataset or nearby similar structures for reference. In order to validate the functionality of our model, we trained it on the Inria building dataset and achieved comparable IOU and accuracy metrics compared to the Inria leaderboard [52]. This further confirms that our methodology can be employed without compromising IOU and accuracy results on the leaderboard, while simultaneously enhancing overall building recall and precision. The results indicate that the UNet model performed well in building segmentation tasks across different datasets, and the use of multi-class datasets consistently improved the building count. However, there was a trade-off observed between false negatives and false positives, with some datasets experiencing an increase in false positives when using the multi-class dataset. Nonetheless, the improved building count with the multi-class dataset indicates its effectiveness in capturing the complexity and diversity of buildings in different urban environments. Overall, the results highlight the importance of dataset selection and the potential benefits of utilizing multi-class datasets in building segmentation tasks with UNet.

The comparison of F1 scores across different model architectures, **Table 5** highlights the improvement achieved when utilizing multi-class datasets. To specifically validate the enhancement between single and multi-class datasets, we conducted the training for all models on 20 epochs, as opposed to the MAP-Net and Joint Learning approaches. This approach allowed us to assess the impact of dataset composition on the F1 score and validate the benefits of utilizing multi-class datasets in the context of building segmentation. So that opens up a possibility of experimentation on MAP-Net and Joint learning using multi-class datasets.

When evaluating our model on the Inria building dataset, we achieved comparable performance in terms of metrics like IOU and accuracy to the leaderboard [52]. This indicates that our methodology can effectively generalize to the dataset used for benchmarking. Additionally, in the experimental results with other datasets such as the Kathmandu Dataset and Opencities dataset, we observed similar trends and performance patterns, further highlighting the generalization potential of our approach across different datasets. These findings demonstrate the robustness and adaptability of our methodology, enabling accurate building segmentation across multiple datasets while maintaining comparable

performance to the established leaderboard.

However, due to lack of similar studies done using multi-class dataset which not just utilize the background and building instances, but also make use of building contours as a class to improve the efficacy of the model, we could not directly compare with the existing studies on the same basis.

7. Challenges and Limitations

Dataset bias in our study arises from two main factors: the similarity of regions surrounding the buildings and the limited availability of comprehensive ground truth data. The presence of similar structures or elements in the surrounding regions poses a challenge for our model, leading to an increase in false positive cases. This occurs when neighboring structures exhibit visual characteristics or patterns resembling buildings. Additionally, the lack of accurate annotations for certain buildings or structures in the vicinity contributes to dataset bias, affecting the precision of our model's predictions. Addressing these biases requires careful consideration and further research to overcome the challenges posed by regional similarities and the need for improved ground truth data in building segmentation tasks. One another particular limitation of this study is that we could not experiment with multi-class dataset on architecture like MAP-Net and Joint-learning presented in [9], in which case, our study would have beend directly comparable to [9].

8. Conclusion

In this paper, we experimented the building detection and counting problem with multiple architectures (UNet, UNet++, DeepLabV3, and DeepLabV3+) on single and multi-class datasets and calculated the evaluation metrics. Using the multi-class method instead of ensembling the network reduces the computational complexities and inference time by making the model learn both features, building contour and inner building segments, on a single model. Building recall has increased substantially on the convoluted regions which helps us to decrease the human effort during the labeling process. And not only that, our work can be used to detect, segment and count the building instances in the convoluted areas more effectively than the other existing methods to our knowledge.

Acknowledgements

The authors would like to thank IKebana Solutions LLC for letting them use time and resources required for this research project and the constant support throughtout the project. The authors would like to thank our colleagues at IKebana for their fruitful discussion, support and feedbacks. The authors would also like to thank the reviewers for their time and valuable comments on the work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Barbedo, J.G.A., Koenigkan, L.V., Santos, P.M. and Ribeiro, A.R.B. (2020) Counting Cattle in UAV Images—Dealing with Clustered Animals and Animal/Background Contrast Changes. *Sensors*, **20**, Article No. 2126. https://doi.org/10.3390/s20072126

[2] Hamdi, Z.M., Brandmeier, M. and Straub, C. (2019) Forest Damage Assessment Using Deep Learning on High Resolution Remote Sensing Data. *Remote Sensing*, **11**, Article No. 1976. https://doi.org/10.3390/rs11171976

[3] Giang, T.L., Dang, K.B., Le, Q.T., Nguyen, V.G., Tong, S.S. and Pham, V.-M. (2020) U-Net Convolutional Networks for Mining Land Cover Classification Based on High-Resolution UAV Imagery. *IEEE Access*, **8**, 186257-186273. https://doi.org/10.1109/ACCESS.2020.3030112

[4] Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P. and Langan, T. (2019) Deep Convolutional Neural Network for Flood Extent Mapping Using Unmanned Aerial Vehicles Data. *Sensors*, **19**, Article No. 1486. https://doi.org/10.3390/s19071486

[5] Boonpook, W., Tan, Y. and Xu, B. (2021) Deep Learning-Based Multi-Feature Semantic Segmentation in Building Extraction from Images of UAV Photogrammetry. *International Journal of Remote Sensing*, **42**, 1-19. https://doi.org/10.1080/01431161.2020.1788742

[6] Orfanidis, G., Ioannidis, K., Avgerinakis, K., Vrochidis, S. and Kompatsiaris, I. (2018) A Deep Neural Network for Oil Spill Semantic Segmentation in Sar Images. 2018 25*th IEEE International Conference on Image Processing* (*ICIP*), Athens, Greece, Athens, 7-10 October 2018, 3773-3777. https://doi.org/10.1109/ICIP.2018.8451113

[7] Kumar, L., Sinha, P. and Taylor, S. (2014) Improving Image Classification in a Complex Wetland Ecosystem through Image Fusion Techniques. *Journal of Applied Remote Sensing*, **8**, Article ID: 083616. https://doi.org/10.1117/1.JRS.8.083616

[8] Wu, M., Zhang, C., Liu, J., Zhou, L. and Li, X. (2019) Towards Accurate High Resolution Satellite Image Semantic Segmentation. *IEEE Access*, **7**, 55609-55619. https://doi.org/10.1109/ACCESS.2019.2913442

[9] Mesner, N. and Ostir, K. (2014) Investigating the Impact of Spatial and Spectral Resolution of Satellite Images on Segmentation Quality. *Journal of Applied Remote Sensing*, **8**, Article ID: 083696. https://doi.org/10.1117/1.JRS.8.083696

[10] Liu, J., Li, P. and Wang, X. (2015) A New Segmentation Method for Very High Resolution Imagery Using Spectral and Morphological Information. *ISPRS Journal of Photogrammetry and Remote Sensing*, **101**, 145-162. https://doi.org/10.1016/j.isprsjprs.2014.11.009

[11] Osco, L.P., Junior, J.M., Marques Ramos, A.P., De Castro Jorge, L.A., Fatholahi, S.N., De Andrade Silva, J., Matsubara, E.T., Pistori, H., Gonçalves, W.N. and Li, J. (2021) A Review on Deep Learning in UAV Remote Sensing. *International Journal of Applied Earth Observation and Geoinformation*, **102**, Article ID: 102456. https://doi.org/10.1016/j.jag.2021.102456

[12] Diakogiannis, F., Waldner, F., Caccetta, P. and Wu, C. (2020) ResUNET-A: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **162**, 94-114. https://doi.org/10.1016/j.isprsjprs.2020.01.013

[13] Lateef, F. and Ruichek, Y. (2019) Survey on Semantic Segmentation Using Deep Learning Techniques. *Neurocomputing*, **338**, 321-348. https://doi.org/10.1016/j.neucom.2019.02.003

[14] Zhao, K., Kang, J., Jung, J. and Sohn, G. (2018) Building Extraction from Satellite Images Using Mask R-CNN with Building Boundary Regularization. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Salt Lake City, 18-22 June 2018, 242-2424. https://doi.org/10.1109/CVPRW.2018.00045

[15] Li, W., He, C., Fang, J., Zheng, J., Fu, H. and Yu, L. (2019) Semantic Segmentation-Based Building Footprint Extraction Using Very High-Resolution Satellite Images and Multi-Source GIS Data. *Remote Sensing*, **11**, Article No. 403. https://doi.org/10.3390/rs11040403

[16] Xu, Y., Wu, L., Xie, Z. and Chen, Z. (2018) Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sensing*, **10**, Article No. 144. https://doi.org/10.3390/rs10010144

[17] Audebert, N., Le Saux, B. and Lefèvre, S. (2017) Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing*, **9**, Article No. 368. https://doi.org/10.3390/rs9040368

[18] Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y. and Shibasaki, R. (2018) Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. *Remote Sensing*, **10**, Article No. 407. https://doi.org/10.3390/rs10030407

[19] OpenCV (2023) Eroding and Dilating. https://docs.opencv.org/3.4/db/df6/tutorial_erosion_dilatation.html

[20] Open Cities AI Challenge Dataset. Version 1.0, Radiant MLHub. https://mlhub.earth/10.34911/rdnt.f94cxb

[21] Maggiori, E., Tarabalka, Y., Charpiat, G. and 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*), Fort Worth, 23-28 July 2017, 3226-3229. https://doi.org/10.1109/IGARSS.2017.8127684

[22] Dixon, B. and Candade, N. (2008) Multispectral Landuse Classification Using Neural Networks and Support Vector Machines: One or the Other, or Both? *International Journal of Remote Sensing*, **29**, 1185-1206. https://doi.org/10.1080/01431160701294661

[23] Singh, D., Maurya, R., Shukla, A., Sharma, M. and Gupta, P. (2012) Building Extraction from Very High Resolution Multispectral Images Using NDVI Based Segmentation and Morphological Operators. 2012 *Students Conference on Engineering and Systems*, Allahabad, 16-18 March 2012, 1-5. https://doi.org/10.1109/SCES.2012.6199034

[24] Huang, X. and Zhang, L. (2012) Morphological Building/Shadow Index for Building Extraction from High-Resolution Imagery Over Urban Areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **5**, 161-172. https://doi.org/10.1109/JSTARS.2011.2168195

[25] You, Y., *et al.* (2018) Building Detection from VHR Remote Sensing Imagery Based on the Morphological Building Index. *Remote Sensing*, **10**, Article No. 1287. https://doi.org/10.3390/rs10081287

[26] Lefevre, S., Weber, J. and Sheeren, D. (2007) Automatic Building Extraction in VHR Images Using Advanced Morphological Operators. 2007 *Urban Remote Sensing Joint Event*, Paris, 11-13 April 2007, 1-5. https://doi.org/10.1109/URS.2007.371825

[27] Kulathunga, G.P. and Afanasyev, I. (2018) Deep Learning Approach for Building Detection in Satellite Multispectral Imagery.

https://www.researchgate.net/publication/328899673_Deep_Learning_Approach_for_Building_Detection_in_Satellite_Multispectral_Imagery

[28] Zhang, P., Du, P., Lin, C., Wang, X., Li, E., Xue, Z. and Bai, X. (2020) A Hybrid Attention-Aware Fusion Network (HAFNet) for Building Extraction from High-Resolution Imagery and LiDAR Data. *Remote Sensing*, **12**, Article No. 3764. https://doi.org/10.3390/rs12223764

[29] Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C.J., Bottou, L. and Weinberger, K.Q., Eds., *Advances in Neural Information Processing Systems* 25, Curran Associates, Inc., Red Hook, 1097-1105.

[30] He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017) Mask R-CNN. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.322

[31] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. and Frangi, A., Eds., *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015. *MICCAI* 2015. *Lecture Notes in Computer Science*, Vol. 9351, Springer, Cham, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[32] Liu, X., Song, L., Liu, S. and Zhang, Y. (2021) A Review of Deep-Learning-Based Medical Image Segmentation Methods. *Sustainability*, **13**, Article No. 1224. https://doi.org/10.3390/su13031224

[33] Yeung, H.W.F., Zhou, M., Chung, Y.Y., Moule, G., Thompson, W., Ouyang, W., Cai, W. and Bennamoun, M. (2022) Deep-Learning-Based Solution for Data Deficient Satellite Image Segmentation. *Expert Systems with Applications*, **191**, Article ID: 116210. https://doi.org/10.1016/j.eswa.2021.116210

[34] Kebaili, A., Lapuyade-Lahorgue, J. and Ruan, S. (2023) Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. *Journal of Imaging*, **9**, Article No. 81. https://doi.org/10.3390/jimaging9040081

[35] Weir, N., Lindenbaum, D., Bastidas, A., Etten, A., Kumar, V., Mcpherson, S., Shermeyer, J. and Tang, H. (2019) Spacenet MVOI: A Multi-View Overhead Imagery Dataset. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October 2019 - 02 November 2019, 992-1001.

[36] Wagner, F.H., Silva, R., Tarabalka, Y., Segantine, T., Thomé, R. and Hirye, M. (2020) U-Net-Id, an Instance Segmentation Model for Building Extraction from Satellite Images—Case Study in the Joanópolis City, Brazil. *Remote Sensing*, **12**, Article No. 1544. https://doi.org/10.3390/rs12101544

[37] Liao, C., Hu, H., Li, H., Ge, X., Chen, M., Li, C. and Zhu, Q. (2021) Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sensing*, **13**, Article No. 1049. https://www.mdpi.com/2072-4292/13/6/1049 https://doi.org/10.3390/rs13061049

[38] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. and Liang, J. (2018) UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: Stoyanov, D., *et al.*, Eds., *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS* 2018. *Lecture Notes in Computer Science*, Vol. 11045, Springer, Cham, 3-11. https://doi.org/10.1007/978-3-030-00889-5_1 https://link.springer.com/chapter/10.1007/978-3-030-00889-5_1

[39] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. (2017) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. http://arxiv.org/abs/1606.00915

[40] Chen, L.-C., Papandreou, G., Schroff, F. and Adam, H. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. ArXiv: 1706.05587.

[41] Tan, M. and Le, Q. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, Long Beach, 9-15 June 2019, 6105-6114. https://proceedings.mlr.press/v97/tan19a.html

[42] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Li, F.-F. (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[43] Sudre, C.H., Li,W., Vercauteren, T., Ourselin, S. and Jorge Cardoso, M. (2017) Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In: Cardoso, M., *et al.*, *Eds.*, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA ML-CDS* 2017. *Lecture Notes in Computer Science*, Vol. 10553, Springer, Cham, 240-248. https://doi.org/10.1007/978-3-319-67558-9_28

[44] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. and Savarese, S. (2019) Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 658-666. https://doi.org/10.1109/CVPR.2019.00075

[45] Google for Developers (2023) Classification: Precision and Recall. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

[46] Paszke, A., *et al.* (2019) Pytorch: An Imperative Style, High-Performance Deep Learning Library. *33rd Annual Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[47] Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M. and Kalinin, A.A. (2020) Albumentations: Fast and Flexible Image Augmentations. *Information*, **11**, Article No. 125. https://doi.org/10.3390/info11020125

[48] Perez, L. and Wang, J. (2017) The Effectiveness of Data Augmentation in Image Classification Using Deep Learning. ArXiv: 1712.04621.

[49] Lorensen, W.E. and Cline, H.E. (1987) Marching Cubes: A High Resolution 3d Surface Construction Algorithm. *ACM SIGGRAPH Computer Graphics*, **21**, 163-169. https://doi.org/10.1145/37402.37422

[50] van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E. and Yu, T. (2014) Scikit-Image: Image Processing in Python. *PeerJ*, **2**, e453. https://doi.org/10.7717/peerj.453

[51] Ramer, U. (1972) An Iterative Procedure for the Polygonal Approximation of Plane Curves. *Computer Graphics and Image Processing*, **1**, 244-256. https://doi.org/10.1016/S0146-664X(72)80017-0

[52] Leaderboard—Inria Aerial Image Labeling Dataset. https://project.inria.fr/aerialimagelabeling/leaderboard/