A Comparative Study of Semantic Segmentation Models for Building Footprint Extraction Using Satellite Imagery

John Jewell*
AI Engineering and Technology
Vector Institute
Toronto, Canada
john.jewell@vectorinstitute.ai

Jinbiao Ning*
Advanced Algorithm Team
Thales Canada, Transportation Solutions
Toronto, Canada
ningjb17@gmail.com

Elham Ahmadi
Geolocation Intelligence Team
RBC
Toronto, Canada
elham.ahmadi@rbc.com

Shihao Ma
Department of Computer Science
University of Toronto
Toronto, Canada
rex.ma@mail.utoronto.ca

***Abstract*—As high resolution satellite imagery becomes increasingly available in both the public and private domains, a number of beneficial applications that leverage this data are enabled. Extraction of building footprints in satellite imagery is a core component of many downstream applications of satellite imagery such as humanitarian assistance and disaster response. This paper offers a comparative study of methods for building footprint extraction in satellite imagery. The focus is to explore state-of-the-art semantic segmentation models in computer vision using the SpaceNet 2 Building Detection Dataset. Four high-level approaches, and six total variants, are trained and evaluated including U-Net, UNet++, Fully Convolutional Networks (FCN) and DeepLabv3. The Intersection over Union (IoU) is used to quantify the segmentation performance on a held out test set. In our experiments, we found that DeepLabv3 with a Resnet-101 backbone is the most accurate approach to building footprint extraction out of the surveyed methods. In general, models that leverage pretraining achieve high accuracy and require minimal training. Conversely, models that do not leverage pretraining are inaccurate and require longer training regimes. In addition to conducting novel experiments and offering a thorough analysis of the results, this paper highlights future work that can be done to extend this comparative study.**

***Index Terms*—Semantic Segmentation, Building Extraction, SpaceNet Dataset, U-Net, UNet++, DeepLabv3, Fully Convolutional Networks**

I. Introduction

Current approaches to extracting building footprints in satellite imagery are primarily based on manual techniques. Advancing automated building footprint localization will play an important role in downstream uses of map data including humanitarian and disaster response [1]. Recent progress in core computer vision tasks, most notably semantic segmentation, present the opportunity to realize precise, automated building footprint localization.

Semantic segmentation is a subclass of image segmentation where pixels are grouped together based on their class [2]. It plays a critical role in a broad range of applications such as autonomous driving (e.g. self-driving cars or autonomous trains), geospatial analysis (e.g. building footprint extraction) and medical image segmentation (e.g. biomedical marker discovery). The goal of semantic segmentation is to label each pixel of an image with a class, effectively partitioning the pixels in the image into groups based on object type. Due to the high dimensional nature of both the input and the output space, semantic segmentation has traditionally been a very challenging task in computer vision [2]. Fortunately, recent supervised deep learning approaches have achieved robust semantic segmentation performance on a variety of challenging benchmarks [3]. These approaches use large datasets of images with corresponding pixel-wise labels to train neural networks by iteratively updating the parameters of the model to minimize a differentiable loss that characterizes the difference between predictions and labels. At inference, new samples are fed to the network and it produces a segmentation map with the same spatial resolution as the input image that encodes the label of each pixel.

This study seeks to explore cutting-edge semantic segmentation methods for building footprint extraction in satellite imagery. The Spacenet 2 Building Detection dataset [1] is used as the benchmark dataset to evaluate the approaches. The dataset contains over 10,000 high resolution satellite images with corresponding ground truth labels. The task involves segmenting building footprints from the background area. Our main contribution is to offer a comparative study of 4 different segmentation networks (i.e. U-Net, UNet++,

* The authors have contributed equally to this work, listed alphabetically.

DeepLabv3 and FCN), and six total architecture variants, for the task of building footprint localization. The study also implicitly assesses the suitability of transfer learning for the task of building footprint localization by featuring methods that leverage pretraining in addition to methods that are trained from scratch. To the best of our knowledge, this is the first work to benchmark this specific set of approaches on the Spacenet 2 Building Detection dataset with a consistent experimental setup.

The rest of the paper is organized as follows. We start with a brief introduction of the related work in semantic segmentation and building localization in Section II. The methods and SpaceNet dataset are presented in Section III. In Section IV, the implementation details and experiment results with a detailed discussion are introduced. At the end of this paper, the conclusions and future work are given.

II. Related Work

A. Semantic Segmentation

With the advent of deep learning, recently proposed methods for semantic segmentation have shown impressive performance on a variety of benchmark datasets [4]. As a seminal work in this line of research, Fully Convolutional Networks (FCN) build on the success of Deep Convolutional Neural Networks (DCNN) for image classification by efficiently making dense predictions for per-pixel tasks by avoiding the use of fully-connected layers [5]. Feature maps from intermediate layers of a backbone network are up-sampled to the dimensions of the desired output and combined to generate the predicted semantic map. SegNet [6] extended FCN by using a symmetric encoder-decoder architecture. The decoder progressively up-samples and refines the low dimensional features generated by the encoder to yield the predicted semantic map. U-Net [3] proposes a similar symmetric structure but also includes skip connections between encoder and decoder layers at the same level of spatial resolution hierarchy. U-Net++ [7] extends U-Net to include skip connections between encoder and decoder layers at multiple levels of the spatial resolution hierarchy.

In a separate line of work, the original DeepLab [8] leverages dilated convolutions to avoid having to excessively downsample the size of features maps generated by the encoder. The output features of the encoder are up-sampled with bi-linear interpolation and fed to a Fully Connected Conditional Random Field (CRF) to iteratively refine the predicted semantic maps. In the second iteration [9], the DeepLab architecture is optimized to handle objects at multiple scales with the inclusion of Atrous Spatial Pyramid Pooling (ASPP). DeepLabv3 [10] introduced a novel encoder-decoder with dilated, depth-wise separable convolutions to capture sharper object boundaries without the use of a CRF for post-processing. Most Recently, Chen et al [11] substituted the Resnet [12] backbone with an Aligned Inception [13] based architecture and introduced a more elaborate up-sampling scheme in the decoder.

B. Semantic Segmentation on Building Extraction

Inspired by the impressive performance of semantic segmentation models, significant effort has been made to transfer the success of deep learning based semantic segmentation methods to building footprint extraction [14]. As a seminal approach to semantic segmentation, FCN have been explored for building footprint localization. Sang et al. [15] investigated fully residual convolutional neural networks for aerial image segmentation, utilizing FCN with a Resnet backbone [12] and additional upsampling skip connections.

Furthermore, U-Net based semantic segmentation models have been extensively studied [16], [17], [18], [19] for building footprint extraction. A multi-constraint fully convolutional network (MC–FCN) model using U-Net as the basic structure of a semantic segmentation model was proposed to perform end-to-end building segmentation in Wu et al [16]. A U-Net-based semantic segmentation method was also explored for the extraction of building footprints from high-resolution multispectral satellite images in Li et al [17]. A fusion solution for an ensemble of U-net models was proposed to extract building contours from the segmentation of aerial images [18]. In order to capture objects of different scales in the deep features, Li et al. [19] proposed a novel aerial image segmentation method adapting U-Net [3] based on convolutional neural networks and inserted a group of cascaded dilated convolutions at the bottom of U-Net which had different dilation rates.

Similarly, DeepLab based semantic segmentation models have attracted great attention in recent research due to their exceptional performance. In [20], a semantic segmentation network modified from DeepLabV3 was applied to urban-scale building footprint extraction using RGB satellite imagery. An improved algorithm based on DeepLabv3+ was developed for semantic segmentation of remote sensing images in [21].

III. Approach

A. Method

The four approaches to semantic segmentation that were explored include: U-Net [3], U-Net++ [7], Fully Convolutional Networks (FCN) [5] and DeepLabv3 [10]. For both FCN and DeepLabv3, two variants of the architecture with different backbones (Resnet-50 and Resnet-100) are included. Thus, in total, six approaches are benchmarked on the task of building footprint extraction in aerial images. The following section offers a brief description of each high level approach.

U-Net: U-Net is an encoder-decoder architecture for semantic segmentation. The encoder consists of a contracting path to capture context and the decoder consists of an expanding path that enables precise localization [3]. Skip connections copy feature maps from the encoder to the decoder layers at the same level of the spatial resolution hierarchy. This enables the flow of high level information that may be lost in the low dimensional output of the encoder [3].

U-Net++: U-Net++ is an encoder-decoder architecture for semantic segmentation that builds on U-Net by linking the encoder and decoder through a series of nested, dense skip

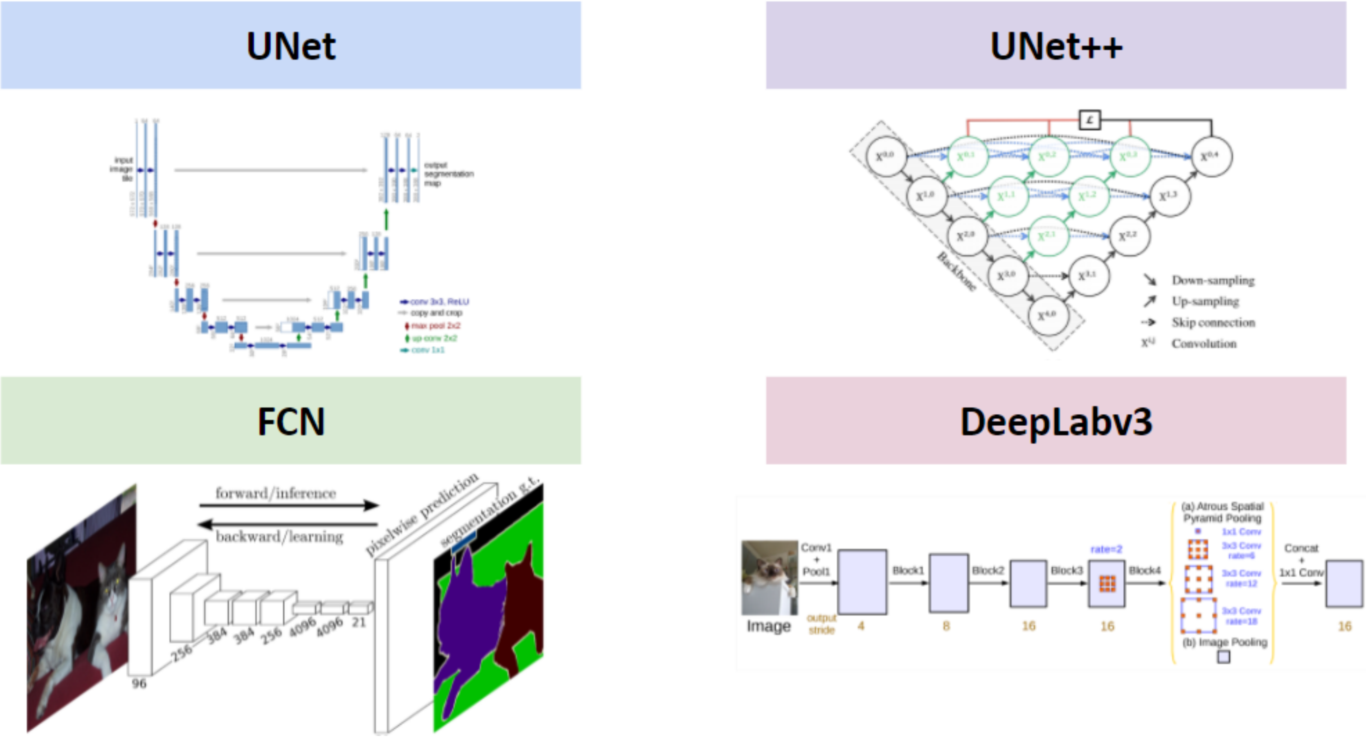

Fig. 1. A visualization of the model architectures for UNet [3], UNet++ [7], FCN [5] and DeepLabv3 [10].

pathways. The re-designed skip pathways aim to reduce the semantic gap between the feature maps of the encoder and decoder sub-networks [7]. When compared with the U-Net architecture, U-Net++ not only has direct or skipped connections between down-sampling layers and up-sampling layers but also convolutional connections, which can pass more features into the up-sampling layers.

FCN: FCN maps arbitrary-sized input images to predicted semantic maps using solely convolutional layers [5]. In-network up-sampling layers are leveraged to make pixel-wise predictions by increasing the spatial resolution of the features generated by the backbone of the network to the height and width of the output. Once up-sampled, semantic information from low resolution feature maps is combined with appearance information from high resolution feature maps to produce precise segmentations. Both an FCN with a Resnet-50 backbone (FCN-50) and a Resnet-101 backbone (FCN-101) are benchmarked in the experiments section. The backbones are pretrained using the COCO train2017 semantic segmentation dataset [22] and fine-tuned for the building footprint extraction task.

DeepLabv3: DeepLabv3 is an encoder-decoder architecture for semantic segmentation that leverages dilated convolutional filters to increase the receptive field of the network and prevent excessive down-sampling [10]. A Spatial Pyramid Pooling module is used to capture context at multiple resolutions which is helpful in localizing objects of different sizes. Standard convolutional layers are factored into depth-wise separable convolutions followed by point-wise convolutions. This dramatically reduces the floating point operations per convolutional layer while maintaining network expressiveness. Both variations of DeepLabv3, with a Resnet-50 backbone (DLV3-50), and a Resnet-101 backbone (DLV3-101) are benchmarked in the experiments section. The backbones are pre-trained using the COCO train2017 semantic segmentation dataset [22] and fine-tuned for the building footprint extraction task.

B. Dataset

In order to benchmark the aforementioned approaches to building footprint extraction in satellite images, the SpaceNet Building Detection V2 dataset [1] is used. This dataset contains high resolution satellite imagery and corresponding labels that specify the location of building footprints. The dataset includes 302,701 Building Labels from across 10,593 multi-spectral satellite images of Las Vegas, Paris, Shanghai and Khartoum. The labels are binary and indicate whether each pixel is building or background, as can be seen in Figure 2.

IV. Experiment

A. Implementation Details

The experiments were implemented in Python using the PyTorch Framework and conducted on 4 NVIDIA Telsa P100 GPU devices. The architecture for each approach is consistent with that specified in the original papers [3], [5], [7], [10]. Each method is trained for 30 Epochs using the ADAM optimizer [23] with a learning rate of 2e-4. Random seeds are used to strive for consistency in evaluation and reproducibility of

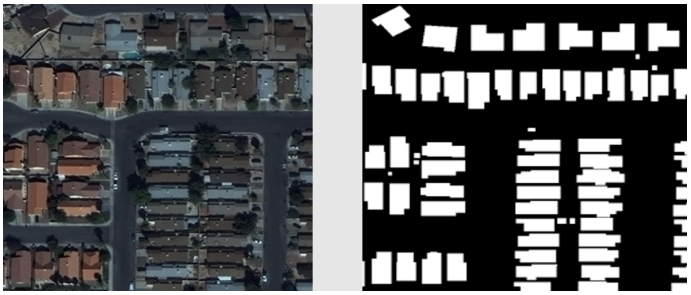

Fig. 2. An example of images (left) and labels (right) in the SpaceNet Building Detection V2. [1]

the experiments. Additional details about the implementation can be found in III.

B. Experimental Setup

The dataset is divided into training (80%), validating (10%) and testing (10%) sets. Images are resized from 650x650 to 384x384 using bi-cubic interpolation and normalized using the mean and standard deviation of the Imagenet dataset [24]. The proposed semantic segmentation models are trained on the training set, while the validating set is used to determine a stopping criteria. Lastly, the trained model is evaluated on the testing set. Intersection over Union (IoU) is the metric used to evaluate the model performance and measures the overlap between the labels of the prediction and ground truth. IoU ranges from 0 to 1 where 1 denotes perfect and complete overlap.

C. Results

The IoU of each method on the test set is reported in Figure I. DLV3-101 achieves the best performance with an IoU of 0.7734 followed closely by DLV3-50, FCN-50 and FCN-101. U-Net and U-Net++ perform comparatively worse with an IoU of 0.5644 and 0.6554, respectively. The performance gap can be attributed to the fact that FCN-50, FCN-101, DLV3-50 and DLV3-100 benefit from pre-training whereas U-Net and U-Net++ do not. This performance gap is also apparent in Figure 4 which shows the train and validation loss of each method across epochs. Methods that leverage pretraining are able to achieve better performance on both the train and validation set from the onset of training. The validation loss begins to plateau after only a few epochs which suggest that training is finished and should be early stopped to prevent over-fitting. Alternatively, U-Net and U-Net++ have train and validation losses that consistently decrease over the course of training. This highlights the fact that models that leverage pretraining converge to the optimal set of parameters faster, in addition to offering better performance.

Qualitative results are available in Figure 3, which shows an example input image, ground truth label and predicted semantic map for each method. The prediction quality of the methods parallels the quantitative results but performance is impressive across the board. The methods are able to generate precise semantic maps in scenes densely populated with building footprints. Additionally, predicted semantic maps in scenes that are sparsely populated with building footprints are robust to false positives, even in cases where roadways, parking lots or other structures are present.

A preliminary analysis of the importance of model architecture conditioned on pretraining yields interesting results. The performance among methods that leverage pretraining is similar, even across different architectures and backbones. Conversely, when considering the performance among methods that do not leverage pretraining, U-Net++ vastly outperforms U-Net. Although this warrants further experiments to validate, one hypothesis is that model architecture becomes less relevant as the amount of pretraining increases.

Model	IoU
U-Net	0.5644
U-Net++	0.6554
FCN-50	0.7455
FCN-101	0.7472
DLV3-50	0.7612
DLV3-101	0.7734

TABLE I
IoU SCORE ON TEST SET FOR EACH APPROACH

D. Future Work

There is a considerable amount of future work that can be done to extend this research. One such area is exploring optimal pre-processing, data augmentation and post-processing schemes for semantic segmentation models within the context of building feature extraction in satellite imagery. Additionally, a wider set of approaches and datasets can be included in the experiments to yield more complete and robust results.

Conclusion

In this study, we trained and evaluated several state-of-the-art semantic segmentation models using the SpaceNet dataset, including U-Net, UNet++, FCN and DeepLabv3. Our results showed that DeepLabv3 with a Resnet-101 backbone is the most accurate approach to building footprint extraction among the models we explored. Models that leverage pretraining (i.e. FCN-50, FCN-101, DLV3-50 and DLV3-101) achieve higher accuracy and require minimal training compared to models without pretraining (i.e. U-Net and UNet++). This study implies that it is suitable to apply transfer learning for the task of building footprint extraction using satellite imagery.

Acknowledgment

The authors would like to thank Vector Institute for making this collaboration possible and providing academic support and computing infrastructure during all phases of this work. We would also like to thank Sedef Akinli Kocak, Shayaan Mehdi and Mazen Al-Rifai for their tremendous support as part of the project management team.

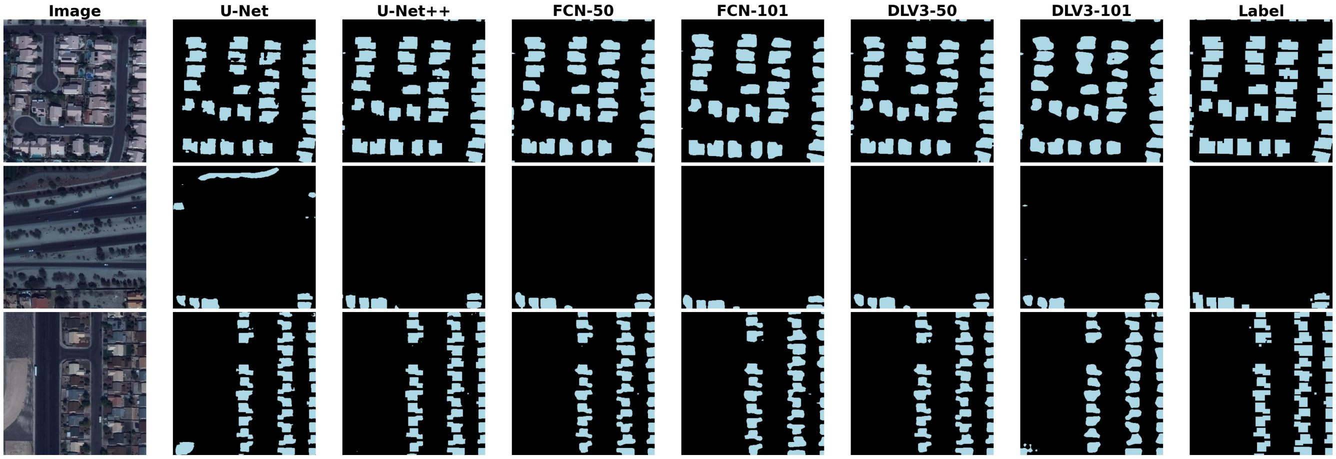

Fig. 3. A visualization of the predictions generated by each approach along with the input image (far left) and ground truth label (far right).

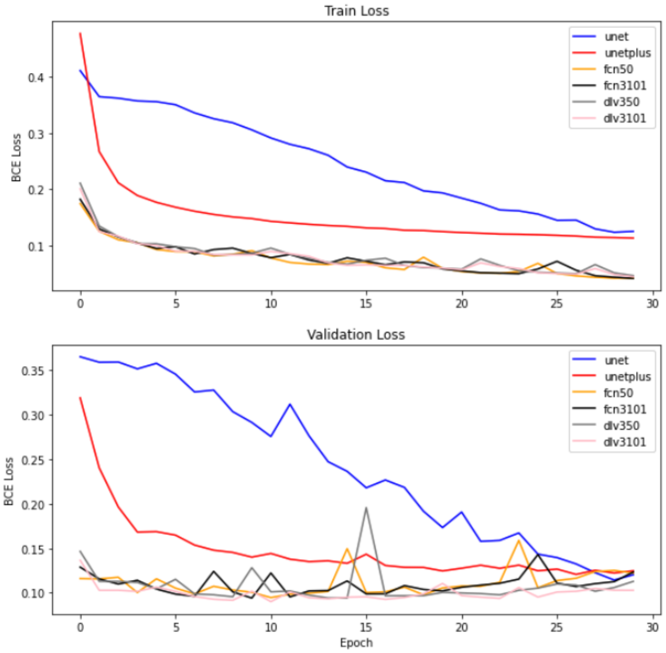

Fig. 4. Binary cross entropy loss for training set (top) and validation set (bottom) across epochs.

REFERENCES

[1] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[2] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[4] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[7] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[14] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417, 2021.

[15] Dinh Viet Sang and Nguyen Duc Minh. Fully residual convolutional neural networks for aerial image segmentation. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, pages 289–296, 2018.

[16] Guangming Wu, Xiaowei Shao, Zhiling Guo, Qi Chen, Wei Yuan, Xiaodan Shi, Yongwei Xu, and Ryosuke Shibasaki. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing*, 10(3):407, 2018.

[17] Weijia Li, Conghui He, Jiarui Fang, Juepeng Zheng, Haohuan Fu, and Le Yu. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data. *Remote Sensing*, 11(4):403, 2019.

[18] Remi Delassus and Romain Giot. Cnns fusion for building detection in aerial images for the building detection challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 242–246, 2018.

[19] Xiang Li, Yuchen Jiang, Hu Peng, and Shen Yin. An aerial image segmentation approach based on enhanced multi-scale convolutional neural network. In *2019 IEEE international conference on industrial cyber physical systems (ICPS)*, pages 47–52. IEEE, 2019.

[20] Aatif Jiwani, Shubhrakanti Ganguly, Chao Ding, Nan Zhou, and David M Chan. A semantic segmentation network for urban-scale building footprint extraction using rgb satellite imagery. *arXiv preprint arXiv:2104.01263*, 2021.

[21] Jiaqi Liu, Zhili Wang, and Kangxin Cheng. An improved algorithm for semantic segmentation of remote sensing images based on deeplabv3+. In *Proceedings of the 5th international conference on communication and information processing*, pages 124–128, 2019.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.