Aerial Image Semantic Segmentation using Neural Search Network architecture

No Author Given

No Institute Given

Abstract. In remote sensing data analysis and computer vision, aerial image segmentation is a crucial research topic, which has many applications in environmental and urban planning. Recently, deep learning is using to tackle many computer vision problem, including aerial image segmentation. Results have shown that deep learning gains much higher accuracy than other methods on many benchmark data sets. In this work, we propose a neural network called NASNet-FCN, which based on Fully Convolutional Network - a frame work for solving semantic segmentation problem and image feature extractor derived from state-of-the-art object recognition network called Neural Search Network Architecture. Our networks are trained and judged by using benchmark dataset from ISPRS Vaihingen challenge. Results show that our methods achieved state-of-the-art accuracy with potential improvements.

1 Introduction

Semantic segmentation is a task of predicting dense-pixel maps from original input images. Each pixel is mapping with predefined classes such as car, tree, building. This is the fundamental research topic in remote sensing data analysis and has many applications in real-life for example urban planning, forest management and environmental modelling. Although having been extensively researched for about two decades, there is still no fully automated method used in practice. The main challenge of this task is the heterogeneous appearance and high intra-class variances of objects e.g buildings, streets and cars on very high resolution images [1].

In the past, [2–5] used the hand-crafted feature extracted from one pixel or a window of small size of aerial image as the input feature for classification algorithm e.g Support Vector Machine, Random Forest, AdaBoost to learn the non-linear decision boundary between classes. Other researches [6, 7] used unsupervised feature learning algorithm to create input feature for neural network learning on road detection task. The unsupervised feature learning algorithm is proved that it can learn filters similar like oriented edge detectors and Gabor wavelets and possibility choose the right filters for given task.

Recently, deep learning, especially deep convolutional neural network(CNN), has been used to tackle many problems in computer vision and boost the accuracy of these problems and achieved state-of-the-art results compared with other methods. For semantic segmentation task, Fully Convolutional Network(FCN)

[8] is the first works try to use CNN to build pixel-to-pixel prediction. FCN used object recognition neural network architectures as feature extraction step and the feature map is upsampled by using fractional stride convolution or deconvolution layer.

In this work, we build up our semantic segmentation network based on the state-of-the-art object recognition network called Neural Architecture Search Network(NASNet) [9] and FCN framework. We evaluate our model performance on the challenging ISPRS dataset [1] and compare to other state-of-the-art results. We also investigate the effect of stronger image feature extractor on the semantic segmentation results. To the best of our knowledge, our work is the first applying NASNet [9] to semantic segmentation task.

Section II will describe the research using CNN in object recognition and semantic segmentation of aerial image. In the following, section III we will explain our model in detail. The experiment results on the dataset of ISPRS challenge are demonstrated in section IV. Our paper ends with conclusion in section V.

2 Related Work

In this section, we will briefly review some works using CNNs for object recognition and semantic segmentation task.

Currently, deep convolution neural network has dominated the ImageNet Large Scale Visual Recognition Competition(ILSVRC) since 2012, when the eight-layer CNN named AlexNet [10] was proposed. In ILSVRC 2014 competition, the VGGNets [11] consists of 19-layers or 16-layers, having smaller filter size than the filter size of AlexNet [10] but still get the same effective receptive field as large kernel size in AlexNet. GoogleLeNet [12] has more layer than the two previous, but using less parameters than AlexNet [10] 12 times. The idea of adding more layers for increasing accuracy became a revolution when ResNet [13] - a 152-layer network with top 5-error better than human performance was introduced. ResNet [13] used residual connection to ease the training of the networks. Inspired by residual connection in ResNet [13], DenseNet [14] established connection with other layer in each dense block to take advantage of feature reuse and reduce vanishing gradient problem. Usually, each novel architecture as in [10, 12–14] requires a great amount of time to design architecture and do experiments. To ease the process of finding new neural network architecture, AutoML [15] has been developed and used by Google Brain to find the new CNN models achieving the highest accuracy in object recognition task.

Semantic segmentation task usually uses the object recognition network as feature extractor part before learning the prediction map for each pixel in different ways. In [8], FCN proposed by Long used the feature map from final and intermediate layers to learn the upsample feature map. Meanwhile, [16, 17] created a symmetric encoder-decoder architecture and residual connection is added from feature extractor part to enhance detail features. Instead of upsampling small feature map to get the original size, [18–21] use dilated convolution [22] to keep the size of feature map unchanged and reduce the impact of field-of-view

problem in normal convolution kernel. Other works [23] use prediction map at multiple scale to learn the model parameters.

In the past, neural network applied to aerial image segmentation by using patch-based approach. Neural network is considered as a classification for pixel in the center of a fixed-size window of pixels extracted from original high resolution images. Recent works based on the method that has gained successful results in semantic segmentation. [24] modified the FCN architecture by using no downsampling layer and increasing the kernel and padding size of pooling layer in the VGG network to reduce the computational cost but still achieved the same accuracy compare to FCNs. Ensemble prediction is also employed as in [25, 26]. In [26], an ensemble of SegNet model is constructed by using multi - kernel convolution size at the last decoder layer to combine predictions at various scale, resulted in smoothing predictions. Additional channels such as Digital Surface Model (DSM), Normalized Difference Vegetation Index (NDVI), Normalize DSM (nDSM) are employed to the model to boost the accuracy. In [27], author proposed a method to increase the boundary-detection accuracy by using additional edge information extracted from boundary-detector network HED [28] to create edge-channel for image, achiving state-of-the-art results on ISPRS dataset.

3 Method

3.1 Neural Architecture Search Network (NASNet)

Finding a neural network architecture achieving the state-of-the-art results requires a great effort of designing and training from researchers. Different with other previous architectures, NASNet architecture is found automatically by AutoML system [15]. The main components of NASNet are two types of cell: Normal Cell and Reduction Cell. Reduction Cell will reduce the width and height of feature map by a half after forwarding the input feature map through. In contrast, Normal Cell will keep these two dimension the same as the input feature map. The general structure of NASNet is built by stacking N Normal Cell between Reduction Cell as in Fig 1. There are two types of NASNet model, NASNet-large with N equal to 6 aims to get maximum possible accuracy and NASNet-mobile with N equal to 4 focus on running on limited resources devices.

Each cell in Normal Cell or Reduction Cell is composed of a number of blocks. Each block is built from the set of popular operations in CNN models with various kernel size e.g: convolutions, max pooling, average pooling, dilated convolution, depth-wise separable convolutions. Finding best architecture for Normal Cell and Reduction Cell with 5 blocks is described details in [9]. The best structure of Normal Cell and Reduction Cell is described in Fig 3.

3.2 Fully Convolution Network(FCN)

FCN [8] is the pioneering work in applying CNN to semantic segmentation by taking advantage of succeed object recognition model. To satisfy the pixel-dense

prediction requirement of semantic segmentation, FCN replace the last fully-connected layer in the object recognition network by convolution layer. By using convolution layer with kernel size 1 by 1, it can squash the number of channels in the last convolution layer to the number of classes. The feature map then is up-sampled by using deconvolution layer or bilinear upsampling operation. We can get the desired dense-pixel feature map from the previous predicted map, but the predicted label are too coarse. In order to overcome these issue, the authors combine the output prediction map with lower layers by summing feature maps of the same size, thus allow models make the local prediction that respect global structure. Also, it requires smaller stride step during deconvolution process, thus improves the details and accuracy of dense prediction.

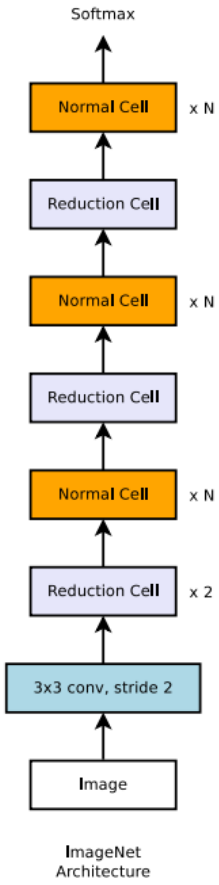

Fig. 1. NASNet general structure [9].

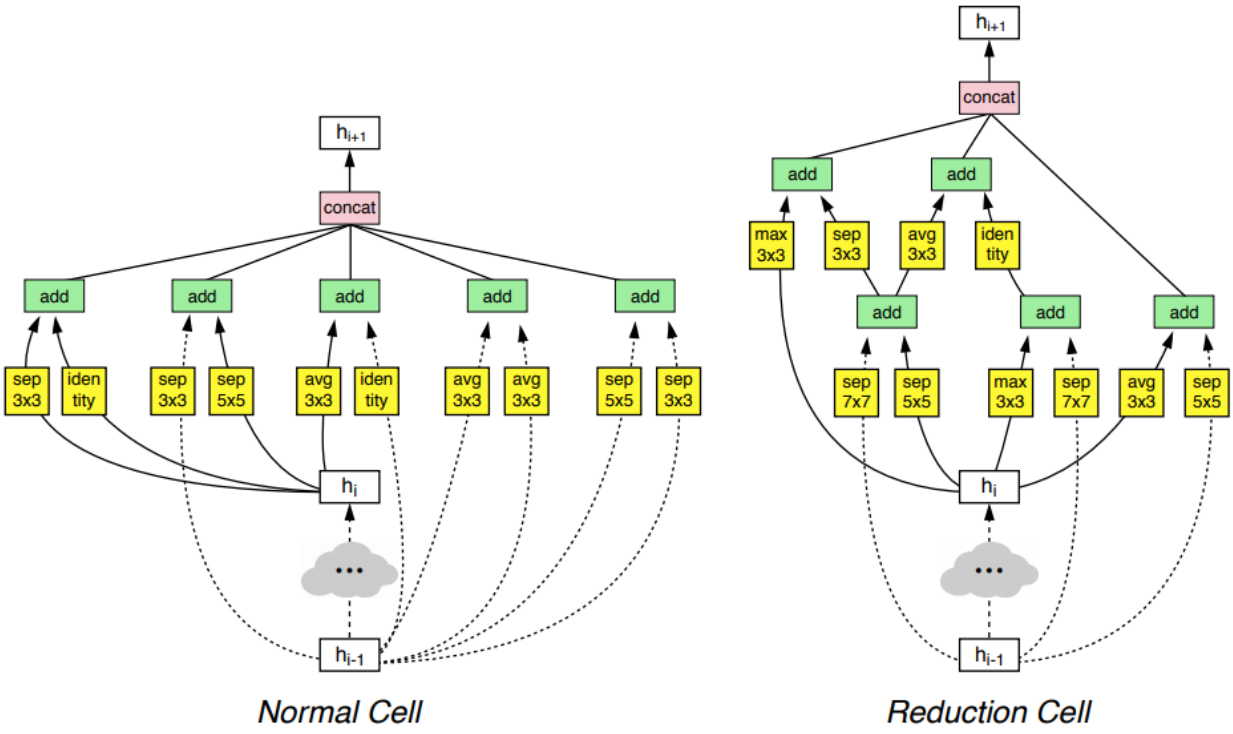

Fig. 2. Normal Cell and Reduction Cell architecture [9].

3.3 Our proposed model

The idea of FCN [8] can be applied to other CNN for object recognition to solve pixel-dense classification task. As in [8], authors experiment with three different architectures: AlexNet [10], GoogleLeNet [12] and VGG [11]. The VGG [11] model achieved the best performance over these three models. For the aerial image segmentation, there are some works using FCN framework and ResNet [13], VGG [11] as the feature extractor and achieved competitive results. But none of the works apply NASNet [9] model as feature extractor part and compare results with other methods using FCN idea. Also as stated in [9], the accuracy of Faster-RCNN [29] model for object detection and localization task is boosted when plugin NASNet model to the Faster-RCNN. With all of that reasons, we want to investigate the effect when using NASNet model for aerial image segmentation task.

Our model follows the same design of FCN-8s as in [8]. By excluding the fully connected layer from the network and add the deconvolution layer with kernel

size and stride equal to 4 and 2 respectively, we can double the feature map size. After that, we fuse the upsampled feature map with all of the output in Normal Cell of NASNet having the same size to encourage finer details in prediction map. The process of doubling feature map size and fusing them is continued in the same fashion as described before. For the last upsampling operation, we use the deconvolution layer with kernel size equal to 16 and stride equal to 8 to produce 8 times upsample feature map. Now the feature map has the same width and height as with the original input. The predicted feature map is then passed through a softmax function, resulting prediction probability vector for each pixels. The loss function is calculated by using average sum cross-entropy across dense-pixel of ground truth and the probability map. Back-propagation algorithm is employed for optimizing the loss function as usual. In this work, we use two verisons of NASNet: NASNet-large and NASNet-mobile to create semantic segmentation models called NASNet-large-FCN and NASNet-mobile-FCN respectively. The next section will describe experiments and results in details.

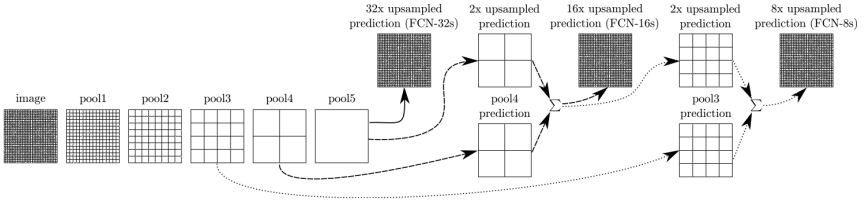

Fig. 3. Original FCN model based on VGG [8].

4 Experiment

4.1 Dataset

Our models are evaluated using ISPRS Vahingen 2D semantic segmentation dataset as in [1]. There are 33 patches, each of them contains very high resolution (more than 5 million pixels) true ortho photo (TOP) and Digital Surface Model (DSM) data with 9-cm ground sampling distance. Each TOP image contain three information in three channels : infrared, red, green (IRRG). The goal of these challenges is to label each pixel of the image with one of six classes: building, low vegetation, tree, car, clutter and impervious surface. There are 16 images with groundtruth provided for training purpose out of all 33 patches. Illustration of data can be found in Fig 4. This task is challenging because of complex appearance of objects e.g buildings, streets, trees and cars in very high-resolution data.

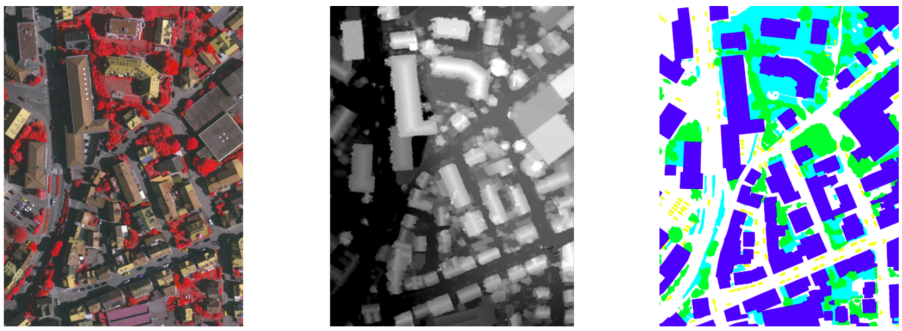

Fig. 4. Example of the dataset. (a) IRRG image, (b) DSM image, (c) Groundtruth image

4.2 Environment setup

For training and prediction, we used a cluster with 2 x Intel(R) Xeon(R) CPU E5-2697 v4@ 2.30GHz, 64GB of RAM and Nvidia Tesla K40m GPU. The operating system is Cent-OS 7. For implementation, we use Python 2.7, Keras [30] framework with Tensorflow [31] backend.

4.3 Implementation details

From the very high-resolution image, we randomly extract 1000 square regions of 224 by 224 from each TOP images. The extracted data is divided to two set: training set and validation set. Eighty-percent of data is used for training, the others is used for validation process. Model is trained from scratch, using Adam [32] optimization algorithm with fixed learning of 0.0001, batch size of 10, and stop learning when model starts overfit. We do not put much effort to find the best hyper-parameter set due to computation resources limitation. For the prediction phase, we slide a square window of 224 by 224 through the test image and generate prediction map for each window image. We used overlap prediction to smooth the prediction map at boundary of extracted windows. The python code for training process can be found at `https://github.com/thinh2/NASNet_FCN`.

4.4 Results

The pixel-dense prediction maps are judged by the benchmark organizers. The competition website public the results and methods each team used in their submissions. For evaluation metric, the F1 score for each classes overall test set is derived, together with the overall accuracy as described in [1].

Table 1. Results on ISPRS Vaihingen dataset.

Submission	Imp suf	Building	Low veg	Tree	Car	Overall Acc
RIT_L8	89.6%	92.2%	81.6%	88.6%	76.0%	87.8%
ADL_3	89.5%	93.2%	82.3%	88.2%	63.3%	88.0%
BKHN_9	90.7%	94.4%	81.8%	88.3%	80.9%	88.8%
DLR_9	92.4%	95.2%	83.9%	89.9%	81.2%	90.3%
BKHN_4	92.7%	95.1%	84.7%	89.8%	86.6%	90.7%
VNU1	**89.2%**	**92.6%**	**80.1%**	**88.0%**	**74.6%**	**87.5%**
VNU2	**89.8%**	**92.0%**	**81.3%**	**88.2%**	**67.7%**	**87.8%**
VNU4	**91.2%**	**93.6%**	**81.5%**	**88.5%**	**77.7%**	**89.0%**

Table 1 shows our prediction results over ISPRS Vaihingen dataset and some selected results from challenge website. We have three models named VNU1, VNU2, and VNU4 which used the original FCN-8s model [8], NASNet-mobile-FCN and NASNet-large-FCN respectively. As we can see, with deeper layers for learning image features, NASNet-large-FCN achieved higher accuracy than the two others. Meanwhile, the NASNet-mobile-FCN achieved the same accuracy with original FCN-8s model, but using fewer parameters (5 millions vs 134 millions). The NASNet-mobile-FCN model also surpass the performance of RIT_L8, which used an ensemble of FCN and random forest with the hand-designed feature. The NASNet-large-FCN model achieved slightly higher accuracy than the BKHN_9 and ADL_3 results, which used Fully Convolutional DenseNet [33] and patch-based prediction [34].

Compare with other methods achieved state-of-the-art results, our methods do not outperform the top accuracy e.g DLR_9 and BKHN_4. Both of these methods used the additional data channel and ensemble learning, achieved greater than 1% compare with our current best overall accuracy prediction. While DLR_9 uses edge information extracted from original image as additional input channel for learning and prediction, BKHN_4 's method uses height information from nDSM and DSM data for learning ensemble of the FCN models. Our models do not use the height data, it leads to some mis-classified region where buildings are covered by shadow and is classified as clutter class. Examples of our prediction and other team prediction can be found in Fig 5.

5 Conclusions

In this work, we employed the state-of-the-art object recognition model NASNet and FCN framework to tackle the aerial image segmentation problem. Experiment results show that the NASNet model boost the accuracy of original FCN-8s models and achieved state-of-the-art results with potential improvments. In the future, we will apply NASNet to other semantic segmentation framework and use post-processing technique e.g Conditional Random Field to boost the performance of our models.

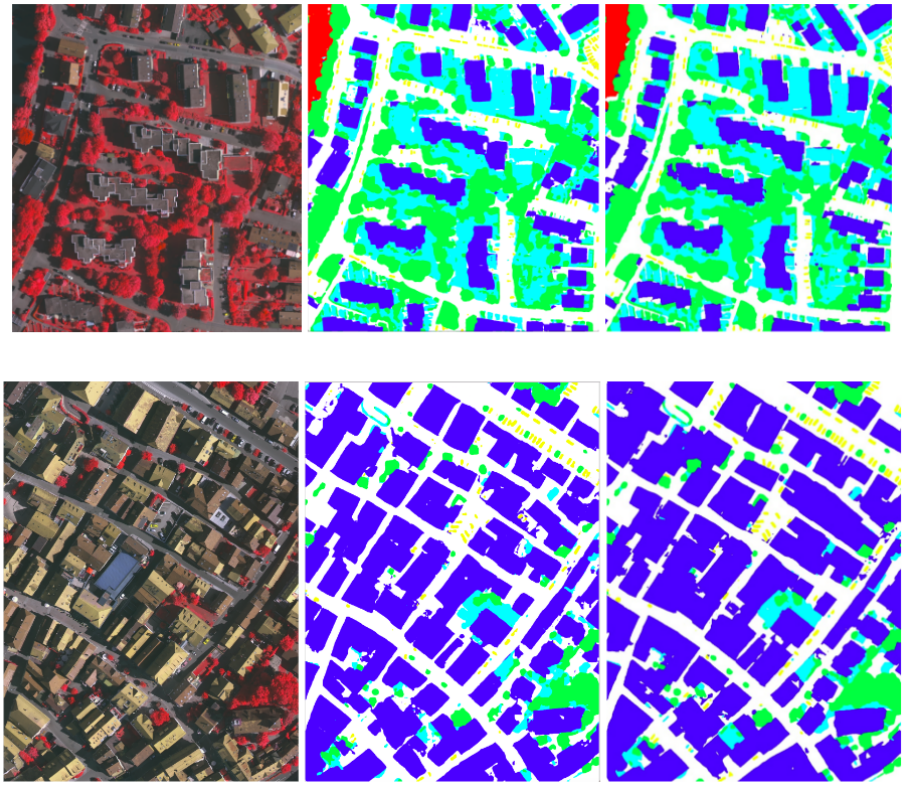

Fig. 5. Some prediction results from test set.Left: IRRG image, Middle: Our NASNet-large-FCN prediction, Right: BKHN_4 predictions.

References

1. M. Cramer, "The dgpf-test on digital airborne camera evaluation–overview and test design," *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2010, no. 2, pp. 73–82, 2010.
2. J. Porway, Q. Wang, and S. C. Zhu, "A hierarchical and contextual model for aerial image parsing," *International journal of computer vision*, vol. 88, no. 2, pp. 254–283, 2010.
3. P. Dollar, Z. Tu, and S. Belongie, "Supervised learning of edges and object boundaries," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 1964–1971, IEEE, 2006.
4. T. T. Nguyen, H. Grabner, H. Bischof, and B. Gruber, "On-line boosting for car detection from aerial images," in *Research, Innovation and Vision for the Future, 2007 IEEE International Conference on*, pp. 87–95, IEEE, 2007.
5. S. Kluckner and H. Bischof, "Semantic classification by covariance descriptors within a randomized forest," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pp. 665–672, IEEE, 2009.
6. V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *European Conference on Computer Vision*, pp. 210–223, Springer, 2010.
7. R. Rigamonti, E. Türetken, G. González Serrano, P. Fua, and V. Lepetit, "Filter learning for linear structure segmentation," tech. rep., 2011.
8. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
9. B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *arXiv preprint arXiv:1707.07012*, 2017.
10. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
11. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
12. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, *et al.*, "Going deeper with convolutions," Cvpr, 2015.
13. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
14. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, p. 3, 2017.
15. B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.
16. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
17. H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528, 2015.
18. L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018.

19. L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.
20. L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014.
21. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
22. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
23. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, 2017.
24. J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.
25. N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017.
26. N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Asian Conference on Computer Vision*, pp. 180–196, Springer, 2016.
27. D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.
28. S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.
29. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.
30. F. Chollet *et al.*, "Keras," 2015.
31. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning.," in *OSDI*, vol. 16, pp. 265–283, 2016.
32. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
33. S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pp. 1175–1183, IEEE, 2017.
34. S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, pp. 36–43, IEEE, 2015.