SEMANTIC SEGMENTATION USING DEEP NEURAL NETWORKS FOR SAR AND OPTICAL IMAGE PAIRS

*Wei Yao[1], *Dimitrios Marmanis[1,2], Mihai Datcu[1]

[1]Department of Photogrammetry & Image Analysis, IMF, German Aerospace Center (DLR), Germany
[2]Department Photogrammetry & Remote Sensing, Technische Universitaet Muenchen (TUM), Germany

ABSTRACT

Semantic segmentation for synthetic aperture radar (SAR) imagery is a rarely touched area, due to the specific image characteristics of SAR images. In this research, we propose a dataset which consists of three data sources: TerraSAR-X images, Google Earth images and OpenStreetMap data, with the purpose of performing SAR and optical image semantic segmentation. By using fully convolutional networks and deep residual networks with pre-trained weights, we investigate the accuracy and mean IOU values of semantic segmentation for both SAR and optical image patches. The best segmentation accuracy results for SAR and optical data are around 74% and 82%. Moreover, we study SAR models by combining multiple data sources: Google Earth images and OpenStreetMap data.

Index Terms— Deep learning, Semantic segmentation, TerraSAR-X, Google Earth, OpenStreetMap

1. INTRODUCTION

In remote sensing area, semantic segmentation has been always a challenging task, meanwhile, it's also a critical step for various applications. There are already a few researches focus on extracting different object from optical satellite-borne or air-borne data [1]; however, very rare cases have been studied for SAR images, or only for very simple applications [2].

In computer science area, the booming development of deep learning methods have shown great power in image information mining from big dataset. Currently, most for computer vision applications. For our remote sensing applications, it's reasonable to assume that, deep learning methods can be one of the ever powerful algorithms that have the potential to beat the other on-the-shelf algorithms (SVM, random forest, etc.).

Hence in this research, we investigate the potential of deep learning methods for a large amount of data, and present here our preliminary semantic segmentation results for high resolution SAR and optical images, based on deep learning methods. Specifically, our dataset contains over 6000 image patches with a size of 200x200 pixels, which are labeled by four categories: building, natural, landuse and water. We are particularly interested in the extraction of buildings. We plan to experiment on different deep learning models, and modify the networks architecture as well, in order to see their effects on getting more accurate segmentation results. Moreover, with multiple sources of knowledge, we study the characteristics of SAR models.

We choose the well-known fully convolutional networks (FCN), together with 50 layer deep residual networks and Atrous convolution networks for our experiments.

In conclusion, the following points show our main contributions:

- We introduce a heterogeneous dataset which consists of TerraSAR-X imagery, Google Earth optical imagery, OpenStreetMap data for pixel-wise semantic segmentation with four categories.
- We build optical models and SAR models, based on deep learning Fully Convolutional Networks (FCN) and Deep Residual Networks learning scheme.
- We study SAR models (i.e., feature maps) by combining multiple data sources: Google Earth images and OpenStreetMap data.

2. METHODOLOGY

2.1. Dataset Description

Our Dataset consists of three data sources, which are TerraSAR-X GEC products, OpenStreetMap data, optical Google Earth images, with the same resolution of 2.9 meters. It covers 15 cities of North Rhine-Westphalia (NRW), Germany. The TerraSAR-X GEC products are with a ground resolution of 2.9 meters, and the incident angles are between 20 and 45 degree on various shooting dates and orbits. As the geocoded coordinates are provided by GEC products, we use open source data from OpenStreetMap to build the corresponding ground truth. This largely reduces the human labeling effort, however, there is also shortcomings, that quite an amount of pixels are without specific map information, due to the lack of geographical information from the open source data. Thanks to the error tolerant ability of deep learning methods,

[*] Authors have contributed equally in this work.

we can still use the majority data of our dataset. Besides, the corresponding optical data were from Google Earth. All three data sources are processed to the same resolution, i.e., with the same image size. Figure 1 shows an example of image patches from three heterogeneous data sources. For OpenStreetMap patch, black color stands for buildings, blue color stands for landuse, red color stands for natural, green color stands for water which is not shown in this example.

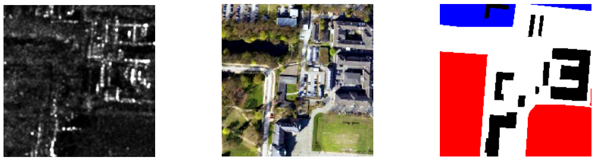

(a) TerraSAR-X Patch (b) Google Earth Patch (c) OpenStreetMap Patch

Fig. 1. Example of image patches from heterogeneous data sources.

2.2. Fully Convolutional Networks & Atrous Convolution Networks & Deep Residual Networks

Fully convolutional networks (FCN) are a kind of deep-learning neural networks which change the last fully connected layers of classification networks to fully convolutional networks. In such context, the networks are adjusted to solve a ”pixel in, pixel out”, namely segmentation problem [3]. They have shown great success in a number of computer vision and aerial remote sensing applications.

Atrous convolution, also known as dilated convolution, is a shorthand for convolution with upsampled filters. This idea have been used before in the context of DCNNs. In practice, atrous convolution computes feature maps more densely, in order to recover full resolution feature maps. Compared to regular convolution, atrous convolution allows us to effectively enlarge the view field of filters without increasing the number of parameters or the amount of computation [4].

Deep Residual Networks are a kind of very deep neural networks with many layers that have obtained impressive results recently. They have an intriguing ”connection skipping” mechanism which enables the inputs of a lower layer available to a node in a higher layer [5].

2.2.1. Networks Explanation

For our experiments, we use the ”FCN-ResNet50-32s” and ”Atrous-ResNet50-16s” pre-trained models, specifically, the deep residual network is with 50 layers and strides of 16 pixels or 32 pixels for optical and SAR models. Caffe framework [6] and Keras [7] which is based on Tensorflow are used to implement our deep-learning algorithms. By experimenting with different deep learning models, we will analyze their results from the quantitative and visualizing perspectives in the next chapter.

3. RESULTS AND ANALYSIS

In this section, results for the fully connected neural networks, deep residual neural networks are presented.

3.1. Quantitative Results

For our experiments, we have used two quantitative results to evaluate our models: accuracy and mean IOU. The mean Intersection-Over-Union (mean IOU) is a common evaluation metric for image semantic segmentation. It computes the IOU for each semantic class, then computes the average over classes. The IOU is defined as:

$$\mathsf{IOU} = \frac{\mathsf{true\ positive}}{\mathsf{true\ positive} + \mathsf{false\ positive} + \mathsf{false\ negative}}. \tag{1}$$

Then a confusion matrix is obtained based on the predictions, and mean IOU is calculated from it.

Table 1. Segmentation accuracies for different data sources and models.

Data Model	TerraSAR-X	Google Earth
Atrous-ResNet50-16s	0.740	0.829
FCN-ResNet50-32s	0.647	0.827

Table 1 describes the segmentation accuracy values by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. Generally, Google Earth image patches get higher accuracies, with around 82% correct segmentation. The best segmentation accuracy for SAR image patches is 74%. It’s interesting to notice that, regarding different models, there is a big difference for TerraSAR-X image patches, while almost no change for Google Earth image patches. This means the Atrous convolution matters a lot to TerraSAR-X data.

Table 2. Segmentation mean IOUs for different data sources and models.

Data Model	TerraSAR-X	Google Earth
Atrous-ResNet50-16s	0.300	0.437
FCN-ResNet50-32s	0.260	0.422

Table 2 describes the segmentation mean IOU values by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. The mean IOU values are much lower than segmentation accuracy results. Like accuracy, Google Earth image patches get better values. But regarding each data source, there is not much significant difference between models.

3.2. Visualized Analysis

Generally, results from Atrous model are better from FCN model. It takes around 250 epochs to get converged for the optical imagery, but much more epochs for SAR imagery. Hence, for building, landuse, natural and water categories, Google Earth patch results with 250 epochs, TerraSAR-X patch results with both 750 epochs and 1000 epochs are shown below.

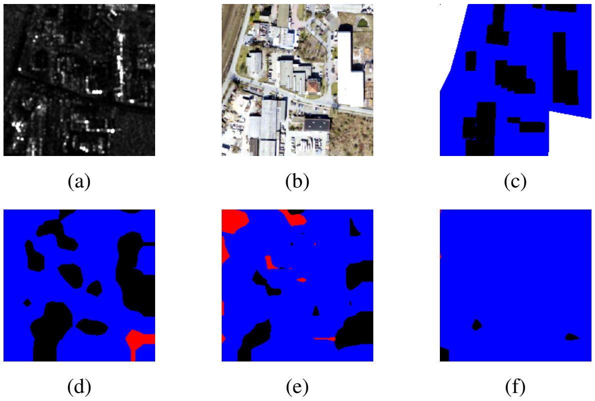


Fig. 2. Building example of Google Earth Patch Result and TerraSAR-X Patch Result. (a) TerraSAR-X Patch. (b) Google Earth Patch. (c) OpenStreetMap Patch. (d) Google Earth Patch Result with 250 Epochs. (e) TerraSAR-X Patch Result with 750 Epochs. (f) TerraSAR-X Patch Result with 1000 Epochs.

Figure 2 mainly shows the building segmentation results by using the pre-trained Atrous residual network for SAR and optical image patches, with different number of epochs. TerraSAR-X results got the strong backscatter locations which show as bright spots areas in the image patches. As the number of iterations increased, the segmented building areas almost disappeared. This means the network didn't model the strong backscatter locations together with their shadows, which was caused by the special imaging mechanism of SAR imagery. However, Google Earth result obtained from Atrous model detected building more precisely.

Figure 3 mainly shows the landuse segmentation results by using the pre-trained Atrous residual network for SAR and optical image patches, with different number of epochs. In case of landuse class, SAR results looked better than optical results. This maybe due to the strong backscatter spots within the landuse area. Figure 3 (d) retained some part of natural area, which indicated a combination of both SAR and optical data might bring a better result. Also due to the small-scale of buildings, comparing to Figure 2, they were difficult to be correctly segmented.

Figure 4 mainly shows the natural segmentation results by using the pre-trained Atrous residual network for SAR and optical image patches, with different number of epochs. The natural category was relatively easy to be segmented, as the strong texture features it showed. For those middle-scale

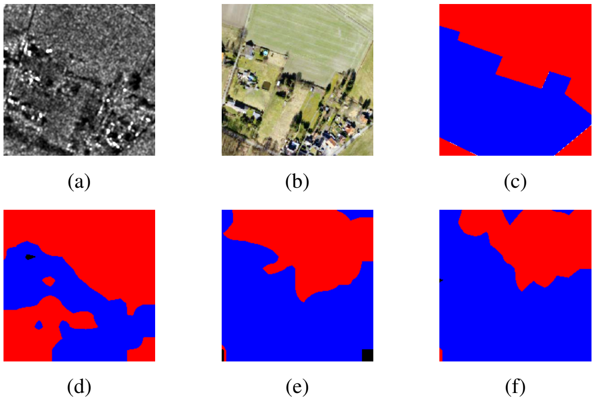


Fig. 3. Landuse example of Google Earth Patch Result and TerraSAR-X Patch Result. (a) TerraSAR-X Patch. (b) Google Earth Patch. (c) OpenStreetMap Patch. (d) Google Earth Patch Result with 250 Epochs. (e) TerraSAR-X Patch Result with 750 Epochs. (f) TerraSAR-X Patch Result with 1000 Epochs.

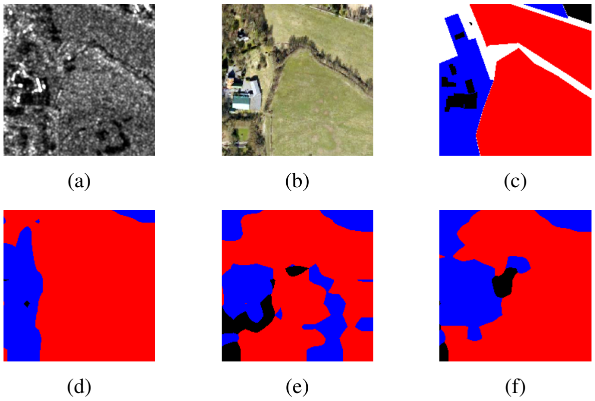
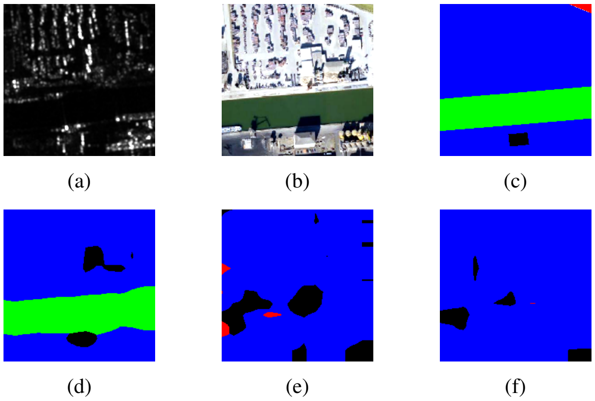


Fig. 4. Natural example of Google Earth Patch Result and TerraSAR-X Patch Result. (a) TerraSAR-X Patch. (b) Google Earth Patch. (c) OpenStreetMap Patch. (d) Google Earth Patch Result with 250 Epochs. (e) TerraSAR-X Patch Result with 750 Epochs. (f) TerraSAR-X Patch Result with 1000 Epochs.

buildings, the networks failed to correctly segment them for both data sources.


(a) (b) (c)

(d) (e) (f)

Fig. 5. Water example of Google Earth Patch Result and TerraSAR-X Patch Result. (a) TerraSAR-X Patch. (b) Google Earth Patch. (c) OpenStreetMap Patch. (d) Google Earth Patch Result with 250 Epochs. (e) TerraSAR-X Patch Result with 750 Epochs. (f) TerraSAR-X Patch Result with 1000 Epochs.

Figure 5 mainly shows the water segmentation results by using the pre-trained Atrous residual network for SAR and optical image patches, with different number of epochs. For water class, optical results showed almost perfect segmentation; meanwhile SAR results obtained nothing, even with some misclassification for buildings.

At this point, we can conclude that to extract building from single frame TerraSAR-X data is not easy with the deep neural networks used in the paper.

3.3. Conclusions and Outlook

Comparing the state of the art semantic segmentation of earth observation data [8], [9], the results we obtained are not yet good enough, however, our contribution lies on analyzing both SAR and optical models and segmentation results simultaneously. Regarding the quantitative results, there is still a large space to improve. These are supposed to achieve by changing networks models, architectures and fine-tuning model parameters. Moreover, at the moment our results are separate for SAR and optical data. Inspired by the visualized analysis, it would be very interesting to see the results of a combination of both data sources.

Hence, here are our outlook to the future work: Since we are facing a big data scenario, the segmentation results will be benefited by increasing the size of our dataset. Under such context, more detailed categories could be considered, for example, natural category can be split up into forest, grass, farm, etc. Then we can study the potential of our networks models to extract more detailed information.

Furthermore, we will investigate the behaviors of layers (i.e., add layers, skip layers, connect to lower layer, etc.), and their impacts on increasing the semantic segmentation accuracy for our dataset. Then we will adjust the networks architecture to combine multiple sources of knowledge, with the purpose of obtaining better SAR models by training optical and map information together.

4. REFERENCES

[1] D. Marmanis, J.D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNNs," in *ISPRS Annals of the Photogrammetry, remote sensing and Spatial Information Sciences*, Prague, Czech Republic, July 2016, vol. III-3.

[2] W. Yao, S.Y. Cui, H. Nies, and O. Loffeld, "Classification of land cover types in TerraSAR-X images using Copula and speckle statistics," in *Proceedings of the 10th European conference on Synthetic Aperture Radar (EUSAR 2014)*, 2014, pp. 743–746.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transaction of Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, pp. 640–651, 2017.

[4] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *Computing research repository (CoRR)*, vol. abs/1706.05587, 2017.

[5] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Deep residual learning for image recognition," *Computing Research Repository (CoRR)*, vol. abs/1512.03385, 2015.

[6] Y.Q. Jia, E. Shelhamer, J. Donahue, and S. Karayev, "Caffe: Convolutional architecture for fast feature embedding," *ACM multimedia 2014 open source software competition*, 2014.

[7] "Keras: The python deep learning library," 2017.

[8] A. Lagrange, Beaupere A. Le Saux, B., A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: from learning explicit features to convolutional networks," in *IEEE International Geosciences and Remote Sensing Symposium (IGARSS)*, 2015, pp. 4173–4176.

[9] N. Audebert, Le Saux B., and Lefévre S., "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," *Computing research repository (CoRR)*, vol. abs/1609.06846, 2016.