Stacked Encoder-Decoders for Accurate Semantic Segmentation of Very High Resolution Satellite Datasets

Maria Papadomanolaki, Maria Vakalopoulou, Nikos Paragios, Konstantinos Karantzalos

STACKED ENCODER-DECODERS FOR ACCURATE SEMANTIC SEGMENTATION OF VERY HIGH RESOLUTION SATELLITE DATASETS

Maria Papadomanolaki[a], *Maria Vakalopoulou*[b], *Nikos Paragios*[b], *Konstantinos Karantzalos*[a]

[a] Remote Sensing Laboratory, National Technical University of Athens, Greece
[b] Center for Visual Computing, CentraleSupélec, Inria, Université Paris-Saclay, France

ABSTRACT

Semantic segmentation is a mainstream method in several remote sensing applications based on very-high-resolution data, achieving recently remarkable performance by the use of deep learning and more specifically, pixel-wise dense classification models. In this paper, we exploit the use of a relatively deep architecture based on repetitive downscale-upscale processes that had been previously employed for human pose estimation. By integrating such a model, we are aiming to capture low-level details, such as small objects, object boundaries and edges. Experimental results and quantitative evaluation has been performed on the publicly available ISPRS (WGIII/4) benchmark dataset indicating the potential of the proposed approach.

***Index Terms*—** Car detection, Semantic segmentation, Fully convolutional networks

1. INTRODUCTION

Semantic segmentation is a well studied problem for the remote sensing community. Traditionally, the approaches in the literature include supervised techniques, implementing different classifiers such as support vector machines or random forests and using a big variety of adhoc features, depending on the application, semantic categories and datasets. Additionally, sophisticated mathematical models as Conditional Random Fields (CRF) or Markov Random Fields (MRF) were also used by semantic segmentation techniques to incorporate spatial relationships between objects [1, 2].

Currently, deep learning techniques and more specifically models which perform pixel-wise dense classification with fully convolutional networks (FCN), are holding the state-of-the-art results for semantic segmentation both in computer vision and remote sensing communities. Shelhamer *et.al.* [3] first proposed a FCN architecture for semantic segmentation problems, replacing the fully connected layers by convolutional layers with kernels that cover their entire input region. After this architecture a big variety of other architectures as [4, 5, 6] have been proposed and reported very high accuracies on a wide range of applications.

Very high resolution (VHR) remote sensing datasets, depicting the Earth's surface in very high spatial details are the ideal datasets for producing accurate semantic segmentation maps. Recently, a variety of very high resolution datasets have been made available and are used as benchmarks for a plethora of methods. The [7, 8, 2] are only some of the publicly available remote sensing datasets, used for semantic segmentation in urban environments. A number of deep learning frameworks based mainly on Convolutional Neural Networks have been proposed to tackle semantic segmentation tasks [9, 10]. However, due to signal downscaling processes the spatial resolution is reduced and critical edge/boundary information is lost resulting to noisy and blurry object boundaries [11, 12].

In order to address this challenge, in this paper we exploit a relatively deep architecture based on repetitive downscale-upscale processes which has been previously employed successfully for human pose estimation [6]. The idea is that by stacking and combining the results of a number of pixel-wise dense classification models we can transfer the learned features across different models and thus enrich the deployed feature space, maintaining small objects and edges.

The rest of the paper is organized as follows. In section 2 the exploited Stacked Hourglass Network is described. In section 3 all the implementation details and the tested dataset are presented while in section 4, quantitative and qualitative results are presented. Lastly, in section 5 a final conclusion is made.

2. METHODOLOGY

The model that is going to be described here was implemented in [6] for human pose estimation. It is based on encoder- decoder architectures and consists of multiple encoder-decoder parts that are successive and similar to each other.

Each encoder- decoder part performs a symmetrical downscale- upscale process on the input patch making at the same time repetitive use of residual modules, as presented in [13]. A single residual module consists of 3 layers, each one performing a batch normalization, a ReLU activation function and a convolution. The convolutions use a filter size of 3x3 or 1x1 changing only the patch depth, leaving the dimensions unaltered. Each residual module is followed by a maxpooling layer that reduces the patch to its half using a filter size of 2 and a stride of 2. In total, there are 4 Residual-maxpooling combinations pooling down to a very low resolution.

Right after that, the encoder is followed by a symmetrical decoder that restores the patch dimensions using 4 Residual-upsampling combinations. Unlike other encoder-decoder approaches, this network does not make use of the common unpooling layers, but instead performs the upsampling using the nearest neighbour technique. A single encoder-decoder part of the whole network produces a heatmap, which is a vector containing the probabilities that this specific part produced for each image category. Due to the symmetrical downsampling and upsampling each part's shape resembles an hourglass.

The idea of continuous downsampling-upsampling procedures comes from the need to process image information across multiple scales. Each hourglass produces a heatmap on which the model can apply a loss function. The network as a whole, comprises of multiple stacked encoder-decoder parts depending on the dataset needs. In this way, this very deep architecture is able to produce more than one heatmap in a single forward pass. This approach can be very constructive, as the model redefines its parameters by repeatedly processing information not only in a local but also in a wider perspective.

3. DATASET AND IMPLEMENTATION DETAILS

The above analyzed model was applied on the dataset of ISPRS Vaihingen 2D Labeling Challenge which consists of very high resolution images that depict the city of Vaihingen and have 3 available channels (InfraRed, Red, Green). From the 16 images that are provided along with their groundtruth, we used 14 of them for training and 2 for validation.

For the training process, we used approximately 22000 patches which were feedforwarded to the model for 60 epochs. We employed the RMSprop optimization method, with a learning rate of 2.5e-3. The whole process lasted for about 24 hours on a single NVIDIA GeForce GTX TITAN with 12 GB of GPU memory.

As far as the architecture details are concerned, we use 4 successive hourglasses and a patch size of 128x128. In each hourglass, the patch dimensions are reduced down to 8x8 and are then upsampled again to the original dimensions of 128x128.

Training as well as testing datasets had been normalized before processed by the network via mean and standard deviation. All processes were implemented with the open source Torch deep learning platform [14].

Similarly, we trained SegNet for 90 epochs. The learning rate was set to 0.01 and optimization was performed by Stochastic Gradient Descent. The whole process lasted about 11 hours in the same computer system. Training and testing data were normalized in the same way, and all processes were implemented in Caffe deep learning platform [15].

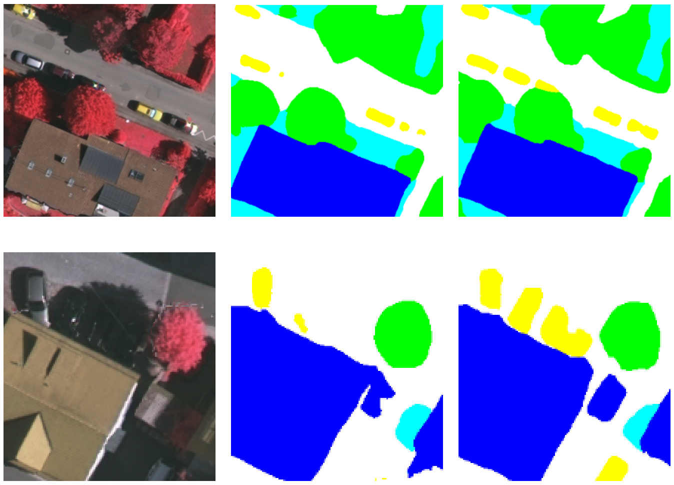


Fig. 1. Zoomed in regions from areas 12 (top line) and 4 (bottom line): (from left to right) the original image, the predictions of SegNet and Stacked Hourglass Networks are shown. *(White: Impervious Surfaces, Dark Blue: Buildings, Light Blue: Low Vegetation, Green: Trees, Yellow: Cars)*

4. RESULTS AND DISCUSSION

We tested the model on the 33 ground-truth-less images of the ISPRS dataset. The testing was implemented using a sliding window of 128x128 and a step equal to 64 along both rows and columns. Such an approach results in overlapping patches, contributing in this way to the production of more coherent predicted images without many visible patch boundaries. The results were compared to the SegNet [4] architecture which is based on a similar encoder-decoder approach, but lacks the repetitive use of downsampling- upsampling parts.

Comparing to SegNet, Stacked Hourglass Networks managed to locate cars with much more detail. In Figure 1, a zoomed region of area 12 and area 4 of the ISPRS testing images is presented as an example. One can observe that the exploited architecture can detect cars even if they are in shadowed places with a very high accuracy. More specifically, for this specific testing area, the overall pixel-wise car accuracy

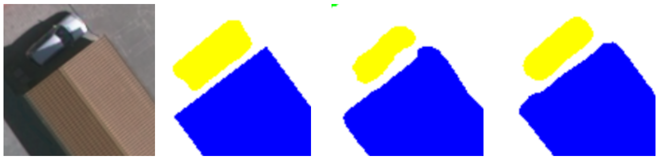
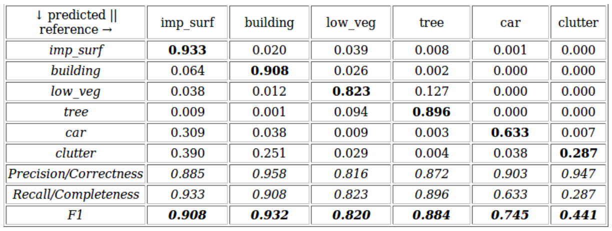


Fig. 2. Results from zoomed area of validation image depicting area 15. (from left to right) the original image, the ground truth, the predictions of SegNet and Stacked Hourglass Networks are shown.



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.933**	0.020	0.039	0.008	0.001	0.000
building	0.064	**0.908**	0.026	0.002	0.000	0.000
low_veg	0.038	0.012	**0.823**	0.127	0.000	0.000
tree	0.009	0.001	0.094	**0.896**	0.000	0.000
car	0.309	0.038	0.009	0.003	**0.633**	0.007
clutter	0.390	0.251	0.029	0.004	0.038	**0.287**
Precision/Correctness	*0.885*	*0.958*	*0.816*	*0.872*	*0.903*	*0.947*
Recall/Completeness	*0.933*	*0.908*	*0.823*	*0.896*	*0.633*	*0.287*
F1	***0.908***	***0.932***	***0.820***	***0.884***	***0.745***	***0.441***

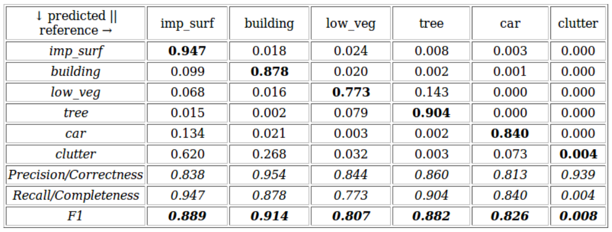
Table 1. Confusion matrix for the SegNet architecture.

that was extracted from the confusion matrix was 96%, outperforming the accuracy of SegNet which was 83%. In addition, the hourglass model performs a more detailed detection on the building boundaries. Continuing with the bottom pictures of Figure 1, one can notice the difference between the two architectures on area 4 of ISPRS testing images. On the one hand, SegNet has merged the two neighboring building areas while on the other hand, hourglass model has separated them as desired.

In Figure 2, there is another characteristic example of Stacked Hourglass Networks' better performance regarding boundaries as both buildings and cars have more accurate shapes. Here, one can also compare the results with the ground truth, as the zoomed region was extracted from one of the validation images. In particular, building boundaries tend to converge more accurately to the buildings' ground truth shape which is square. Moreover, the car prediction is more correct and complete comparing to SegNet's inaccurate curved shapes.

For a more thorough quantitative evaluation, Tables 1 and 2 outline how the employed architectures act to the whole testing ISPRS images. The two confusion matrices that are presented were produced via submission of our predicted ground-truth-less testing images to the ISPRS Test Project regarding Vaihingen 2D Labeling Challenge. In general, SegNet has achieved a little higher accuracies, as proved by the F1 scores that describe how well the performance was. Precision values are also a bit higher for SegNet in all image categories which means that Stacked Hourglass Networks have produced more inaccurate building predictions. More specifically, the hourglass architecture resulted in more inaccuracies in the interior of the building boundaries as it sometimes confused roof windows with 'Impervious Surfaces' (Figure 3). Nevertheless, this architecture was more efficient when dealing with shapes and boundaries, leading to higher recalls in some classes. More specifically, having managed to produce more exact building borderline predictions, 'Impervious Surfaces' are detected in a more complete way resulting in a higher recall comparing to SegNet. The same idea applies for the higher recalls of 'Cars' and 'Trees'.

5. CONCLUSION



↓ predicted || reference →	imp_surf	building	low_veg	tree	car	clutter
imp_surf	**0.947**	0.018	0.024	0.008	0.003	0.000
building	0.099	**0.878**	0.020	0.002	0.001	0.000
low_veg	0.068	0.016	**0.773**	0.143	0.000	0.000
tree	0.015	0.002	0.079	**0.904**	0.000	0.000
car	0.134	0.021	0.003	0.002	**0.840**	0.000
clutter	0.620	0.268	0.032	0.003	0.073	**0.004**
Precision/Correctness	*0.838*	*0.954*	*0.844*	*0.860*	*0.813*	*0.939*
Recall/Completeness	*0.947*	*0.878*	*0.773*	*0.904*	*0.840*	*0.004*
F1	***0.889***	***0.914***	***0.807***	***0.882***	***0.826***	***0.008***

Table 2. Confusion matrix for the Stacked Hourglass Network.

In this paper, we tested a relatively deep architecture based on repetitive downscale-upscale processes that had been previously employed for human pose estimation. Our purpose was to conduct experiments and observe the model's behaviour when dealing with very high resolution remote sensing data. The produced results were compared with SegNet and the accuracies of the two architectures were found to be very much alike. Although SegNet achieves higher accuracies, the hourglass architecture results in more correct shapes and more accurate borderlines between classes especially in the case of small objects such as cars. Using post processing techniques as CRFs the proposed accuracies of the tested model can be further ameliorated. Finally, in the future we are planning to further evaluate the exploited architecture for the semantic instance segmentation problem.

6. REFERENCES

[1] M. Vakalopoulou, N. Bus, K. Karantzalosa, and N. Paragios, "Integrating edge/boundary priors with classification scores for building detection in very high resolution data," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2017, pp. 3309–3312.

[2] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2015, pp. 1–9.

[3] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, April 2017.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE PAMI*, 2017.

[5] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng, *Stacked Hourglass Networks for Human Pose Estimation*, pp. 483–499, Springer International Publishing, Cham, 2016.

[7] Franz Rottensteiner, Gunho Sohn, Markus Gerke, Jan Dirk Wegner, Uwe Breitkopf, and Jaewook Jung, "Results of the ISPRS benchmark on urban object detection and 3d building reconstruction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 93, no. 0, pp. 256 – 271, 2014.

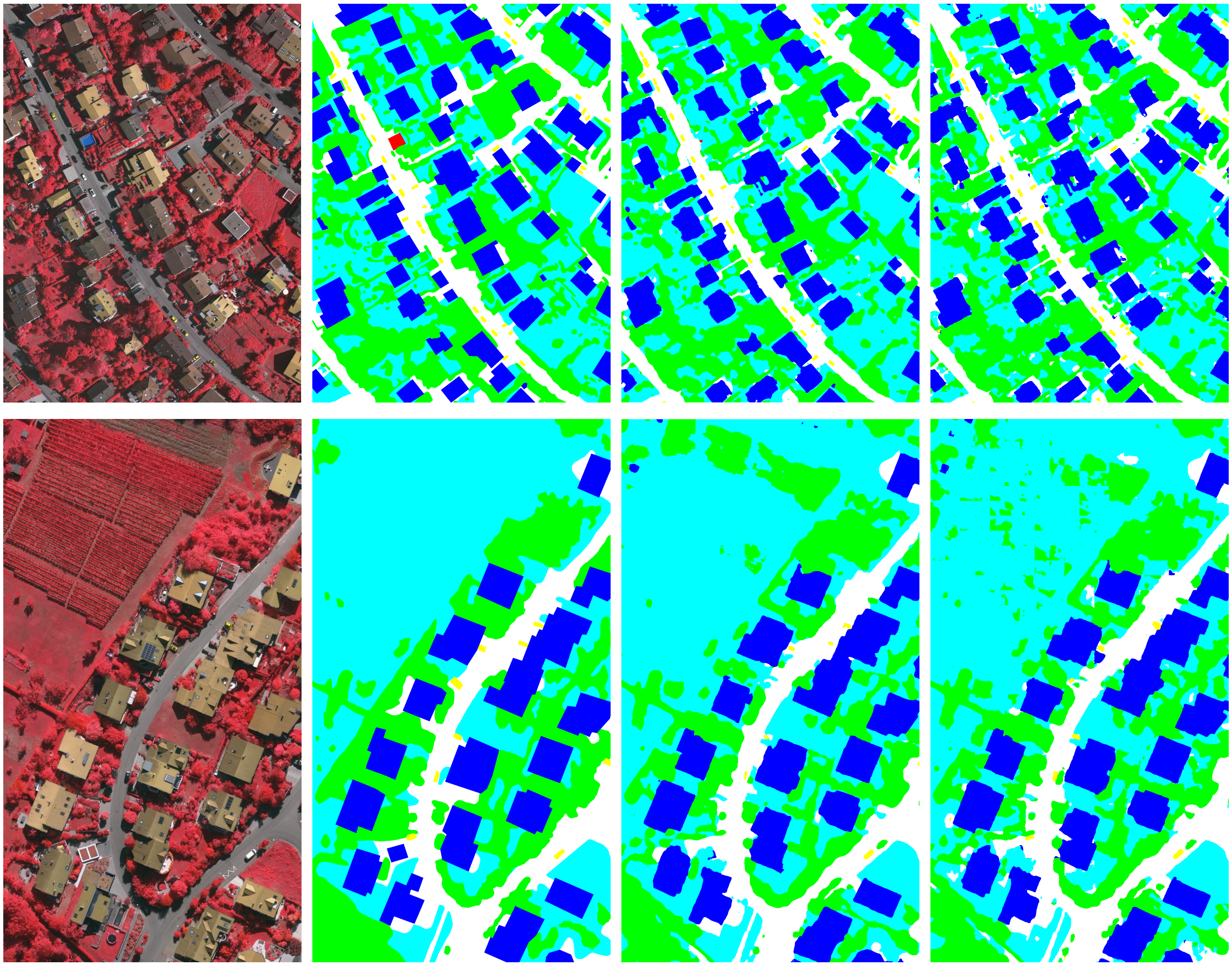


Fig. 3. Images from the areas 15 (top line) and 17 (bottom line). From left to right: Satellite image depicting area 15 and area 17 of ISPRS dataset, the corresponding ground truth, the SegNet predictions, the Stacked Hourglass Networks predictions.

[8] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2017, pp. 3226–3229.

[9] 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.

[10] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, Feb 2017.

[11] 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. PP, no. 99, pp. 1–1, 2017.

[12] L. C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4545–4554.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

[15] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.