Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks

Nicolas Audebert[1,2], Bertrand Le Saux[1], Sébastien Lefèvre[2]

[1] ONERA, *The French Aerospace Lab*, F-91761 Palaiseau, France
[2] Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France

Semantic segmentation of remote sensing images using deep networks

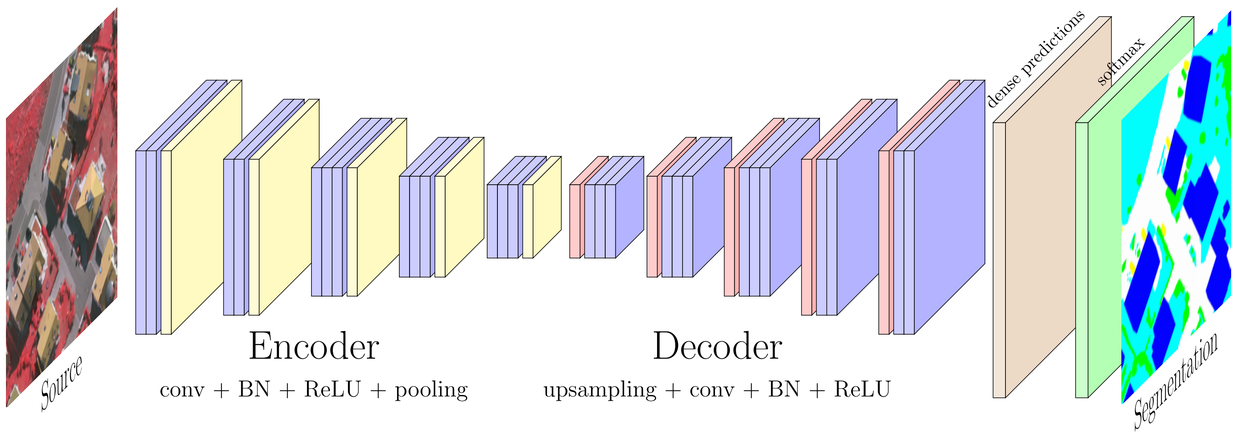


Figure 1: SegNet [1], Badrinarayanan et al. (2015)

Semantic segmentation of aerial images

1. Sliding window over the high resolution tile
2. Dense prediction using a Fully Convolutional Network
3. Agregation over the high resolution tile

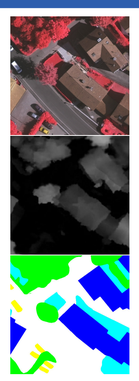
Implementation

- SegNet architecture [1] trained by Stochastic Gradient Descent

Challenges

- Data fusion : **how to fuse optical and Lidar data ?**
- Spatial context : **how to merge several context sizes ?**

ISPRS Vaihingen Dataset



ISPRS 2D Semantic Labeling Challenge (Vaihingen) [8]

- High resolution tiles (2300×2300px, 12.5 cm per pixel)
- Optical data: Infra-red/Red/Green (IRRG)
- Lidar data: Digital Surface Model (DSM)
- Dense ground truth with 6 classes

+ normalized DSM [3] (NDSM) and vegetation index (NDVI)

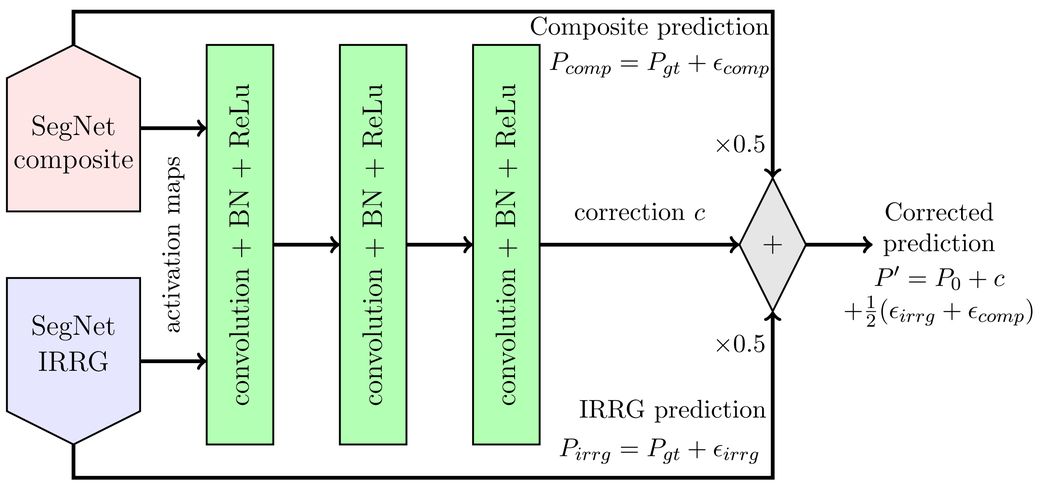
Data fusion with residual correction

Naive data fusion

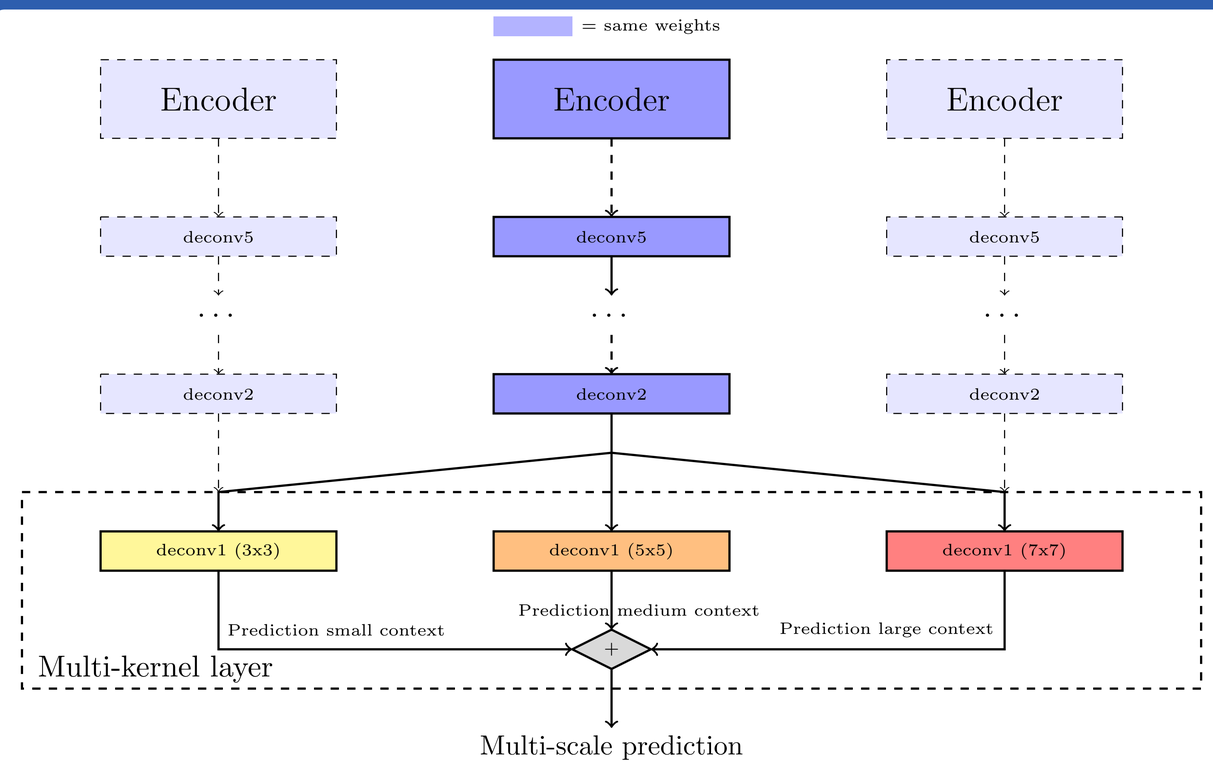
- **Dual stream** [2] **SegNet** trained on IRRG and composite (DSM/NDSM/NDVI)
- Fusion by averaging the prediction maps: **+0.4%** accuracy w.r.t IRRG only

Residual correction

- Residual correction based on a **dual-stream SegNet**
- Use a **residual network** to merge the predictions, inspired by **signal correction techniques** to improve noisy operations (e.g. averaging uncertain predictions)
- Fusion by residual correction: **+0.8%** accuracy compared to IRRG only



Multi-kernel convolution



Dotted layers are virtual layers sharing weights with the actual SegNet. Only the final layer has triple weights.

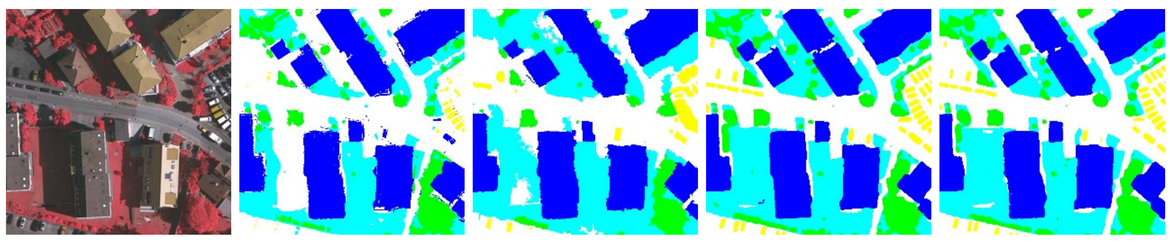
Split the final convolution into **several kernels** with different sizes to average several models working on **multiple neighbourhood sizes**

⇒ model averaging with **shared weights** at **several context sizes**

- Multi-scale: **+0.5%** accuracy compared to standard SegNet

Results

Method	imp surf	building	low veg	tree	car	Accuracy
FCN ("UZ_1")	89.2%	92.5%	81.6%	86.9%	57.3%	87.3%
FCN [4]	89.8%	92.1%	80.4%	88.2%	82.0%	87.6%
CNN + RF + CRF [6]	89.5%	93.2%	82.3%	88.2%	63.3%	88.0%
FCN [5]	90.3%	92.3%	82.5%	89.5%	76.3%	88.5%
FCN + RF + CRF ("DST_2")	90.5%	93.7%	83.4%	89.2%	72.6%	89.1%
SegNet++	**91.5**%	94.3%	82.7%	89.3%	**85.7**%	89.4%
Segnet++ w/ fusion	91.0%	**94.5**%	**84.4**%	**89.9**%	77.8%	**89.8**%



(a) IRRG data (b) "SVL"[3] (c) RF + CRF[7] (d) Segnet++ (e) Segnet++ w/ fusion

(white: roads, blue: buildings, cyan: low vegetation, green: trees, yellow: cars)

Conclusion

- **Dual stream SegNet** improves classification accuracy of Earth Observation images thanks to **fusion of optical and Lidar** data by **residual correction**
- **Multi-kernel** improves ensemble averaging by agregating predictions **over several local neighbourhoods**
- **Do-it-yourself** with our code and pre-trained models:

`https://github.com/nshaud/DeepNetsForEO`

References

[1] V. Badrinarayanan, Alex Kendall, and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *arXiv preprint arXiv:1511.00561* (2015).

[2] A. Eitel et al. "Multimodal deep learning for robust RGB-D object recognition". In: *Proceedings of the International Conference on Intelligent Robots and Systems*. IEEE, 2015.

[3] Markus Gerke. *Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*. Tech. rep. International Institute for Geo-Information Science and Earth Observation, 2015.

[4] G. Lin et al. "Efficient piecewise training of deep structured models for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[5] D. Marmanis et al. "Semantic Segmentation of Aerial Images with an Ensemble of CNNs". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* 3 (2016).

[6] S. Paisitkriangkrai et al. "Effective semantic pixel labelling with convolutional networks and Conditional Random Fields". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015.

[7] N. T. Quang et al. "An Efficient Framework for Pixel-wise Building Segmentation from Aerial Images". In: *Proceedings of the Sixth International Symposium on Information and Communication Technology*. ACM, 2015, p. 43.

[8] F. Rottensteiner et al. "The ISPRS benchmark on urban object classification and 3D building reconstruction". In: *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci* 1 (2012).

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF). Nicolas Audebert's work is supported by the Total-ONERA research project NAOMI. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-13-JS02-0005-01 (Asterix project).