CNNs for Remote Sensing Semantic Segmentation

Ondřej Pešek, Martin Landa
Czech Technical University in Prague, Faculty of Civil Engineering

At: Institut plánování a rozvoje, Praha

7th May 2025

Table of Contents

Methodology

Architectures Examined

- FCN [1]
- U-Net [2]
- SegNet [3]
- DeepLabv3+ [4] with three backbone models [5]
  - ResNet-50
  - ResNet-101
  - ResNet-152

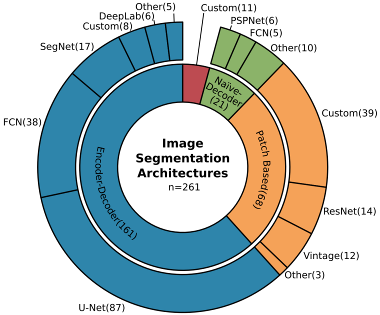


Figure 1: Overview of the most frequently used convolutional neural networks for semantic segmentation in the field of remote sensing. Source: [6]

Approaches Examined

- Different features of interest
  - Vegetation
  - Urban features
- Different approaches
  - Land cover classification
  - Land use classification
- Different spatial resolution
  - Sentinel-2 – 10 to 60 metres
  - Aerial imagery – 10 centimetres
- Different input bandsets
  - All bands provided in the data source
  - RGB
  - All bands enhanced by NDVI
- Different class number
  - Binary classification
  - Multi-class classification
- Different dataset sizes
- Compare with random forests

Use cases

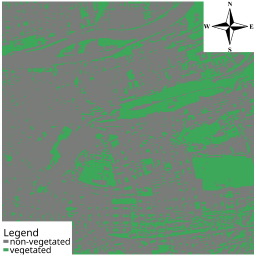
Use Cases — Urban Greenery

Use case definition:

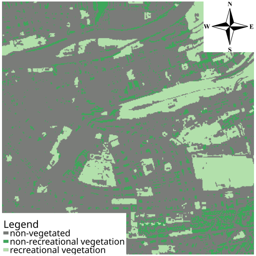
- Published in Remote Sensing Applications: Society and Environment journal: https://doi.org/10.1016/j.rsase.2024.101238
- Training dataset published on Zenodo server: https://zenodo.org/records/8413116
- Using Sentinel-2 data as input
- Using 10-metre imagery is rare for urban greenery detection
- The first study using CNNs for urban green land use classification
- Very limited training dataset
- Three levels
  - Level 1: Non-vegetated pixels, vegetation
  - Level 2: Non-vegetated pixels, non-recreational vegetation, recreational vegetation
  - Level 3: Non-vegetated pixels, non-recreational vegetation, low recreational vegetation, high recreational vegetation

Use Cases — Urban Greenery

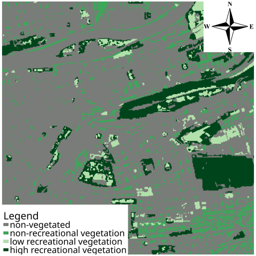
(a) Dlaždice Sentinel-2 z 03/06/2019



(b) Dataset — úroveň 1



(c) Dataset — úroveň 2



(d) Dataset — úroveň 3

Figure 2: Porovnání různých úrovní masek.

Use Cases — Urban Greenery

Table 1: Overall accuracy values in per cent for various configurations and architectures on the **level-1** nomenclature. The values were computed over the validation dataset. Rows abbreviations explained: dN% means dropout, N indicates the dropout ratio, and rnN indicates the ResNet depth in the DeepLabv3+ backbone model. Columns abbreviations explained: fb means the full-band images dataset, rgb means the dataset comprising the red, green, and blue bands, rgb_ndvi means the rgb dataset enhanced by the NDVI, and the _a suffix means that the dataset was augmented. The best result is highlighted in bold.

Architecture	fb	fb_a	rgb	rgb_a	rgb_ndvi	rgb_ndvi_a
FCN_d00	72.8	79.6	70.7	79.4	70.1	79.8
FCN_d50	72.2	79.3	69.7	78.2	61.7	78.8
U-Net_d00	91.8	**94.8**	88.8	93.7	92.5	94.7
U-Net_d50	93.6	**94.8**	92.6	93.6	92.2	94.3
SegNet_d00	51.7	89.2	50.4	85.8	78.2	91.3
SegNet_d50	82.2	87.7	80.6	82.7	68.7	88.8
DeepLabV3+_rn50_d00	75.5	82.3	68.4	77.5	68.5	83.1
DeepLabV3+_rn50_d50	80.2	83.6	44.8	79.8	67.7	78.0
DeepLabV3+_rn101_d00	75.0	81.8	57.4	79.6	64.5	83.0
DeepLabV3+_rn101_d50	74.3	84.8	72.2	78.1	65.2	76.0
DeepLabV3+_rn152_d00	76.4	82.1	68.6	78.1	65.9	82.8
DeepLabV3+_rn152_d50	74.4	84.8	71.4	78.3	66.4	76.0
random_forests	92.7	92.7	90.1	90.1	93.2	93.2

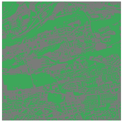
Use Cases — Urban Greenery

Table 2: Overall accuracy values in per cent for various configurations and architectures on the **level-2** nomenclature. The values were computed over the validation dataset. Rows abbreviations explained: dN% means dropout, N indicates the dropout ratio, and rnN indicates the ResNet depth in the DeepLabv3+ backbone model. Columns abbreviations explained: fb means the full-band images dataset, rgb means the dataset comprising the red, green, and blue bands, rgb_ndvi means the rgb dataset enhanced by the NDVI, and the _a suffix means that the dataset was augmented. The best result is highlighted in bold.

Architecture	fb	fb_a	rgb	rgb_a	rgb_ndvi	rgb_ndvi_a
FCN_d00	64.8	75.6	64.9	75.0	64.9	75.3
FCN_d50	64.5	74.1	62.2	76.1	59.3	74.7
U-Net_d00	87.1	**90.0**	85.6	88.9	86.0	89.6
U-Net_d50	82.4	89.9	86.1	89.0	80.4	86.8
SegNet_d00	72.8	83.2	67.0	80.0	77.7	85.7
SegNet_d50	65.0	71.4	58.2	79.2	51.6	77.8
DeepLabV3+_rn50_d00	67.6	78.8	62.4	73.8	74.6	79.5
DeepLabV3+_rn50_d50	73.3	78.2	64.1	68.8	74.9	73.7
DeepLabV3+_rn101_d00	65.1	75.0	61.0	72.3	76.9	76.7
DeepLabV3+_rn101_d50	71.5	76.7	66.0	56.7	73.4	73.8
DeepLabV3+_rn152_d00	66.2	75.4	59.1	73.1	76.6	76.4
DeepLabV3+_rn152_d50	64.4	72.8	65.9	60.8	74.1	73.3
random_forests	84.3	84.3	78.6	78.6	80.6	80.6

Use Cases — Urban Greenery

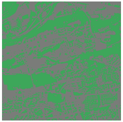
(a) A patch of a Sentinel-2 image tile.



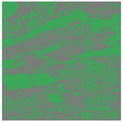
(b) Ground-truth mask.

(c) FCN — D 0%. OA 72.8%

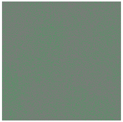
(d) FCN — D 50%. OA 72.2%



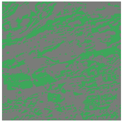
(e) U-Net — D 0%. OA 91.8%



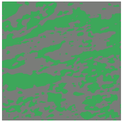
(f) U-Net — D 50%. OA 93.6%



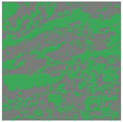
(g) SegNet — D 0%. OA 51.7%



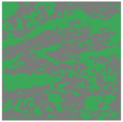
(h) SegNet — D 50%. OA 82.2%



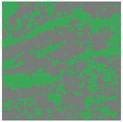
(i) DeeplabV3+ with ResNet-50 — D 0%. OA 75.5%



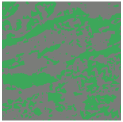
(j) DeeplabV3+ with ResNet-50 — D 50%. OA 80.2%



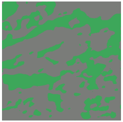
(k) DeeplabV3+ with ResNet-101 — D 0%. OA 75.0%



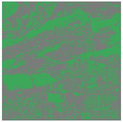
(l) DeeplabV3+ with ResNet-101 — D 50%. OA 74.3%



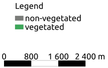
(m) DeeplabV3+ with ResNet-152 — D 0%. OA 76.4%



(n) DeeplabV3+ with ResNet-152 — D 50%. OA 74.4%



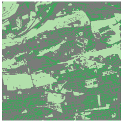
(o) Random forests. OA 92.7%



(p) Legend.

Use Cases — Urban Greenery

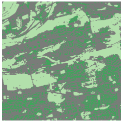
(a) A patch of a Sentinel-2 image tile.



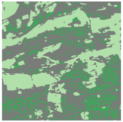
(b) Ground-truth mask.

(c) FCN — D 0%. OA 64.8%

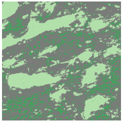
(d) FCN — D 50%. OA 64.5%



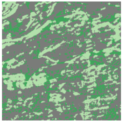
(e) U-Net — D 0%. OA 87.1%



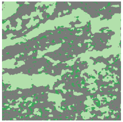
(f) U-Net — D 50%. OA 82.4%



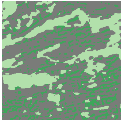
(g) SegNet — D 0%. OA 72.8%



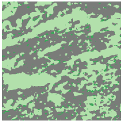
(h) SegNet — D 50%. OA 65.0%



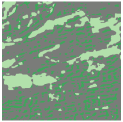
(i) DeeplabV3+ with ResNet-50 — D 0%. OA 67.6%



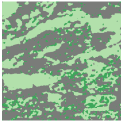
(j) DeeplabV3+ with ResNet-50 — D 50%. OA 73.3%



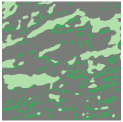
(k) DeeplabV3+ with ResNet-101 — D 0%. OA 65.1%



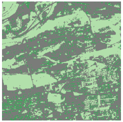
(l) DeeplabV3+ with ResNet-101 — D 50%. OA 71.5%



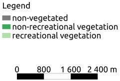
(m) DeeplabV3+ with ResNet-152 — D 0%. OA 66.2%



(n) DeeplabV3+ with ResNet-152 — D 50%. OA 64.4%



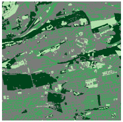
(o) Random forests. OA %84.3



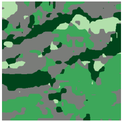
(p) Legend.

Use Cases — Urban Greenery

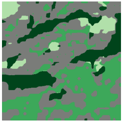
(a) A patch of a Sentinel-2 image tile.



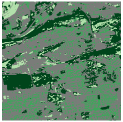
(b) Ground-truth mask.



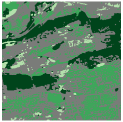
(c) FCN — D 0%. OA 59.1%



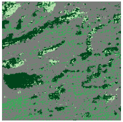
(d) FCN — D 50%. OA 58.5%



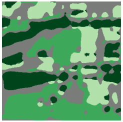
(e) U-Net — D 0%. OA 80.6%



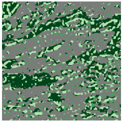
(f) U-Net — D 50%. OA 73.9%



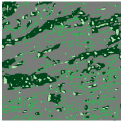
(g) SegNet — D 0%. OA 60.6%



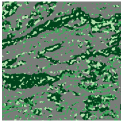
(h) SegNet — D 50%. OA 45.8%



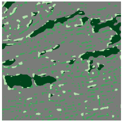
(i) DeeplabV3+ with ResNet-50 — D 0%. OA 59.1%



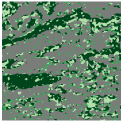
(j) DeeplabV3+ with ResNet-50 — D 50%. OA 65.9%



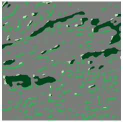
(k) DeeplabV3+ with ResNet-101 — D 0%. OA 58.7%



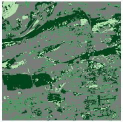
(l) DeeplabV3+ with ResNet-101 — D 50%. OA 60.0%



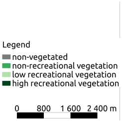
(m) DeeplabV3+ with ResNet-152 — D 0%. OA 57.3%



(n) DeeplabV3+ with ResNet-152 — D 50%. OA 60.7%



(o) Random forests. OA 80.9%

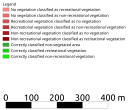


(p) Legend.

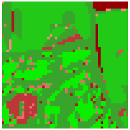
Use Cases — Urban Greenery

(a) Bing Maps aerial photo covering the sample area.

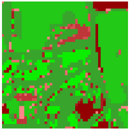
(b) Ground-truth mask.



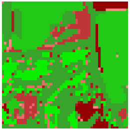
(c) Legend.



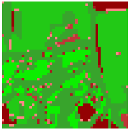
(d) After a training using full-band input.



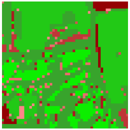
(e) Using only RGB.



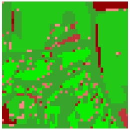
(f) Using full-band input enhanced with NDVI.



(g) Using augmented full-band input.



(h) Using only augmented RGB.



(i) Using augmented full-band input enhanced with NDVI.

Use Cases — Urban Greenery

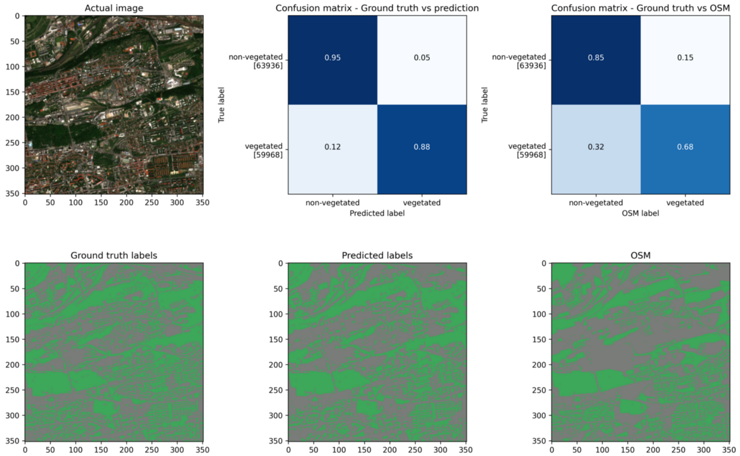


Figure 7: Comparison of OSM data and results from the level-1 detection of U-Net without dropout layers trained on the full-band dataset. The validation patch.

Use Cases — Cloud Detection

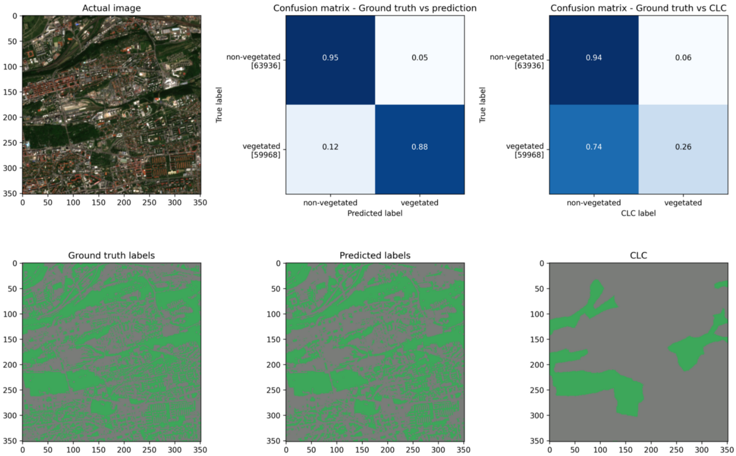


Figure 8: Porovnání dat CORINE Land Cover a výsledků vyprodukovaných architekturou U-Net bez výpadkových vrstev trénované na datasetu úrovně 1 využívajícím všech obrazových kanálů.

Use Cases — Urban Greenery

Conclusions:

- CNNs reached good resuls
  - Vegetation detection: Highest overall accuracy over 95%
  - Vegetation land use classifciation: Highest overall accuracy over 90%
- CNNs can handle even the land use task
- Results reached using CNNs are more accurate than available open datasets
- Extra bands and NDVI improve results
- Dropout did not improve results
- Data augmentation improved results
- The best results reached by U-Net

Use Cases — Road Surface Classification

Use case definition:

- Submitted for publishing, waiting for the second round of peer reviews... still
- Training dataset published on Zenodo server: `https://zenodo.org/records/10602515`
- Using 10-cm aerial data as input
- The first study using CNNs to distinguish between modular and compact road surface on remotely sensed data

Use Cases — Road Surface Classification

Table 3: Overall accuracy values in per cent for various configurations and architectures. The values were computed over the validation dataset. Rows abbreviations explained: `dN%` means dropout, `N` indicates the dropout ratio, and `rnN` indicates the ResNet depth in the DeepLabv3+ backbone model. Columns abbreviations explained: `fb` means the full-band images dataset, `rgb` means the dataset comprising the red, green, and blue bands, `rgb_ndvi` means the `rgb` dataset enhanced by the NDVI, and the `_a` suffix means that the dataset was augmented. The best result is highlighted in bold.

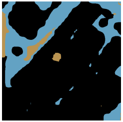
Architecture	fb	fb_a	rgb	rgb_a	rgb_ndvi	rgb_ndvi_a
FCN_d00	89.5	89.7	89.4	87.0	89.2	90.2
FCN_d50	71.4	83.2	80.9	67.8	74.2	73.0
U-Net_d00	**91.1**	89.2	87.6	90.3	90.9	89.6
U-Net_d50	77.6	67.7	69.1	71.2	79.8	66.8
SegNet_d00	90.3	88.1	86.8	87.8	89.6	88.8
SegNet_d50	66.7	83.7	80.1	75.4	79.2	84.6
DeepLabV3+_rn50_d00	81.2	79.7	81.3	86.9	75.4	84.7
DeepLabV3+_rn50_d50	67.2	69.3	70.9	81.9	66.5	63.1
DeepLabV3+_rn101_d00	75.9	82.6	80.0	84.9	83.9	86.2
DeepLabV3+_rn101_d50	69.0	65.7	83.2	81.7	62.6	48.1
DeepLabV3+_rn152_d00	77.4	85.9	84.4	83.2	80.8	74.0
DeepLabV3+_rn152_d50	75.3	69.8	79.3	82.1	63.8	70.4
random_forests	59.4	59.4	65.5	66.5	60.5	60.5

Use Cases — Road Surface Classification

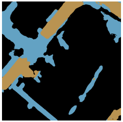
(a) A patch of the aerial imagery.

(b) Ground-truth mask.

(c) FCN — D 0%. OA 89.5%

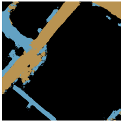


(d) FCN — D 50%. OA 71.4%

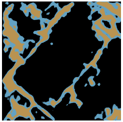


(e) U-Net — D 0%. OA 91.1%

(f) U-Net — D 50%. OA 77.6%

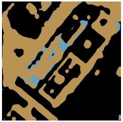


(g) SegNet — D 0%. OA 90.3%

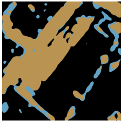


(h) SegNet — D 50%. OA 66.7%

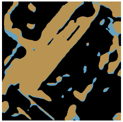
(i) DeeplabV3+ with ResNet-50 — D 0%. OA 81.2%



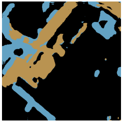
(j) DeeplabV3+ with ResNet-50 — D 50%. OA 67.2%



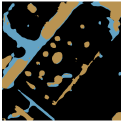
(k) DeeplabV3+ with ResNet-101 — D 0%. OA 75.9%



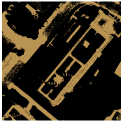
(l) DeeplabV3+ with ResNet-101 — D 50%. OA 69.0%



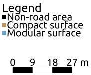
(m) DeeplabV3+ with ResNet-152 — D 0%. OA 77.4%



(n) DeeplabV3+ with ResNet-152 — D 50%. OA 75.3%



(o) Random forests. OA 59.4%

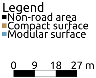


(p) Legend.

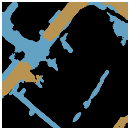
Use Cases — Road Surface Classification

(a) A patch of the aerial imagery.

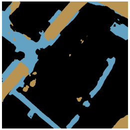
(b) Ground-truth mask.



(c) Legend.

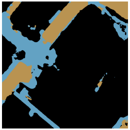


(d) After a training using full-band input.



(e) Using only RGB.

(f) Using full-band input enhanced with NDVI.



(g) Using augmented full-band input.

(h) Using only augmented RGB.

(i) Using augmented full-band input enhanced with NDVI.

Use Cases — Road Surface Classification

Conclusions:

- CNNs can deal with the task
- Extra bands improved results, NDVI did not
- Dropout did not improve results
- Ambivalent effect of simple data augmentation
- The best results reached by U-Net

Conclusions

- U-Net was always the best performing of the tested architectures
- In all the presented use cases, U-Net overperformed existing approaches and open datasets
- FCN and DeepLabv3+ are training data-greedier than the other architectures
- The number of classes did not have any influence on the performance of the CNN
- Extra bands and NDVI (usually) improve the performance but even the results reached on pure RGB outperform existing methods
- Thanks to its context consciousness, CNNs can deal even with challenging tasks as land use or clouds over desert detection

Conclusions — Desired Extensions

- More use cases
- Research on the relationship between batch normalisation and dropout
- Deeper research on various data augmentation techniques
- More architectures
- Make GIS addons out of the accompanying source code

References I

[1] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE conference on Computer Vision and Pattern Recognition CVPR*, pages 3431–3440, 2015.

[2] O. RONNEBERGER, P. FISCHER, and T. BROX. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI*, pages 234–241, 2015.

[3] 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*, 39(12):2481–2495, 2017.

[4] L. CHEN, Y. ZHU, G. PAPANDREOU, F. SCHROFF, and H. ADAM. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision ECCV*, pages 833–851, 2018.

[5] K. HE, X. ZHANG, S. REN, and J. SUN. Deep residual learning for image recognition. pages 770–778, 2016.

[6] T. HOESER and C. KUENZER. Object detection and image segmentation with deep learning on earth observation data: A review-part ii: Applications. *Remote Sensing*, 11(10), 2020.

[7] G. E. HINTON, N. SRISTAVA, A. KRIZHEVSKY, I. SUTSKEVER, and R. R. SALAKHUTDINOV. Improving neural networks by preventing co-adaptation of feature detectors, 2012. URL `http://arxiv.org/abs/1207.0580`.

Thank you for your attention.