# Fully convolutional and feedforward networks for the semantic segmentation of remotely sensed images

Martina Pastorino • Gabriele Moser • Sebastiano B. Serpico • Josiane Zerubia

# Outline





UNIVERSITÉ CÔTE D'AZUR Inria Università di Genova

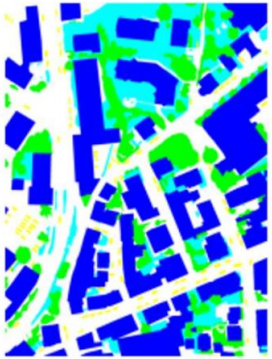
# 01

# Introduction

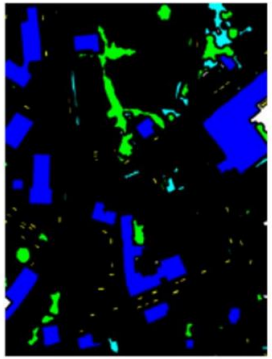
# The addressed problem

**The development of a supervised method for the semantic segmentation – or dense image classification – of remote sensing images at very high resolution (VHR) is a challenging problem**

- **Models based on deep learning (e.g., fully convolutional networks, FCNs) are capable to reach accurate classification results**
  - **however**, these models need **large datasets** which require **high efforts** in annotation and most often are **not available**
  - their **performances worsen** in case of **scarce ground truth (GT)**



*Original GT*



*GT used to train the model*

# The addressed problem

**FCNs extract information at different spatial resolutions**

- the use of the spatial details at the finest resolutions and the robustness to noise plus outliers at the coarsest has proven to improve the classification results
- stochastic models such as probabilistic graphical models (PGMs) are flexible and powerful approaches which can extract information from multiscale data for labeling purposes
- several approaches that combine DL and PGMs have been studied in order to improve the accuracy of the classification results by capturing the interactions between pixels

# Proposed solution

**The method in [1] was extended with the addition of fully connected neural networks (NNs) at different convolutional blocks of the FCN**

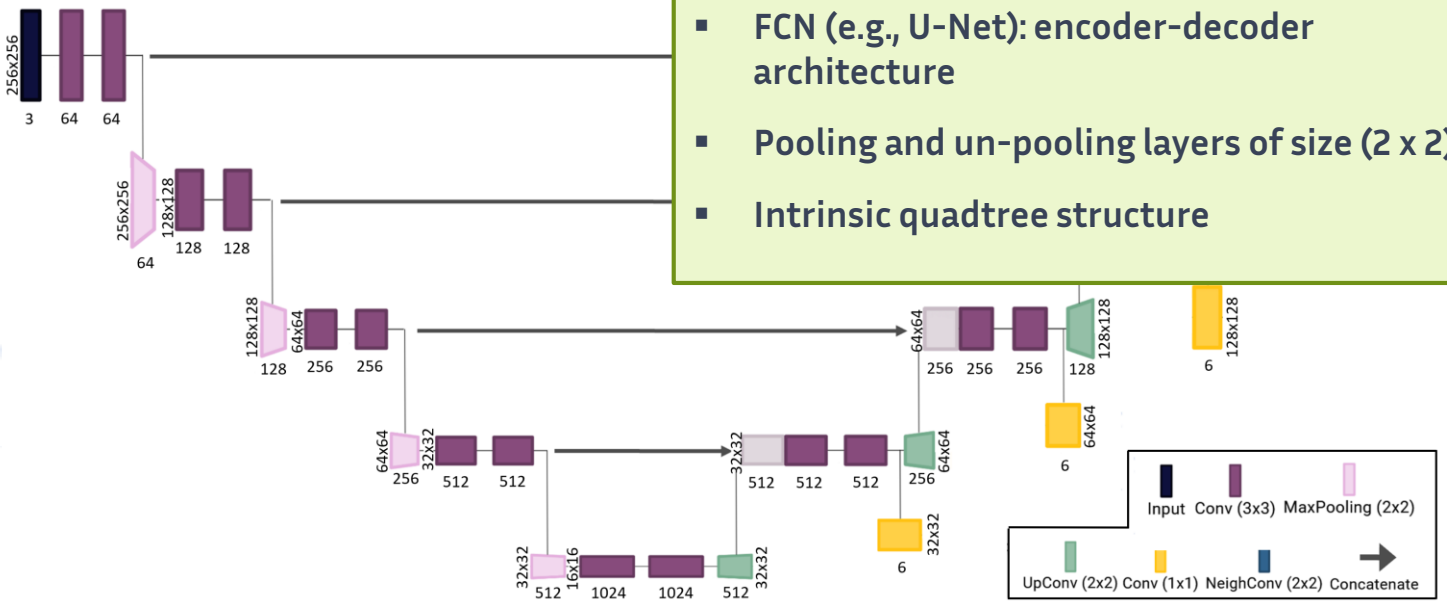
- **Architecture capable of integrating multiscale data without the need of ensemble learning techniques, thus making the approach end-to-end.**
- **Spatial-contextual information is favored in this framework, through**
  - a supplementary convolutional layer modeling the interactions between neighboring pixels at the same resolution.
  - an additional loss term which allows to integrate spatial information between neighboring pixels.

[1] M. Pastorino, G. Moser, S. B. Serpico, and J. Zerubia, "Semantic segmentation of remote sensing images through fully convolutional neural networks and hierarchical probabilistic graphical models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

# 02

# Proposed framework

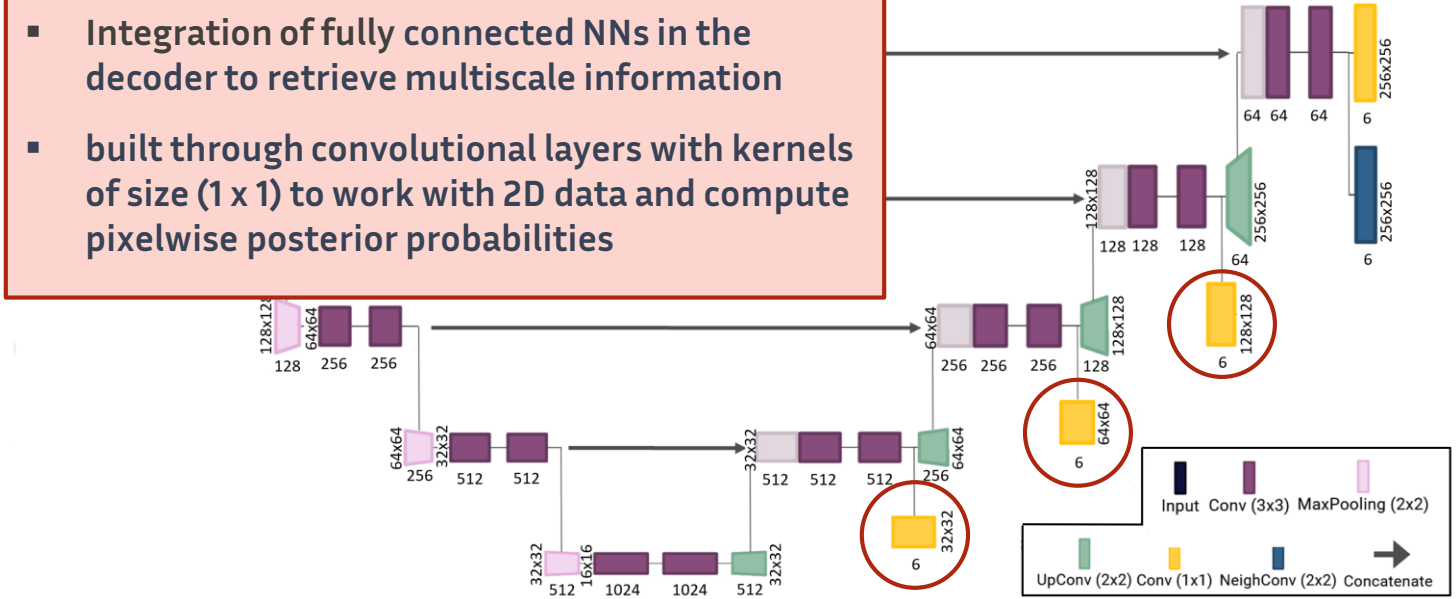
# Overall architecture




- FCN (e.g., U-Net): encoder-decoder architecture
- Pooling and un-pooling layers of size (2 x 2)
- Intrinsic quadtree structure

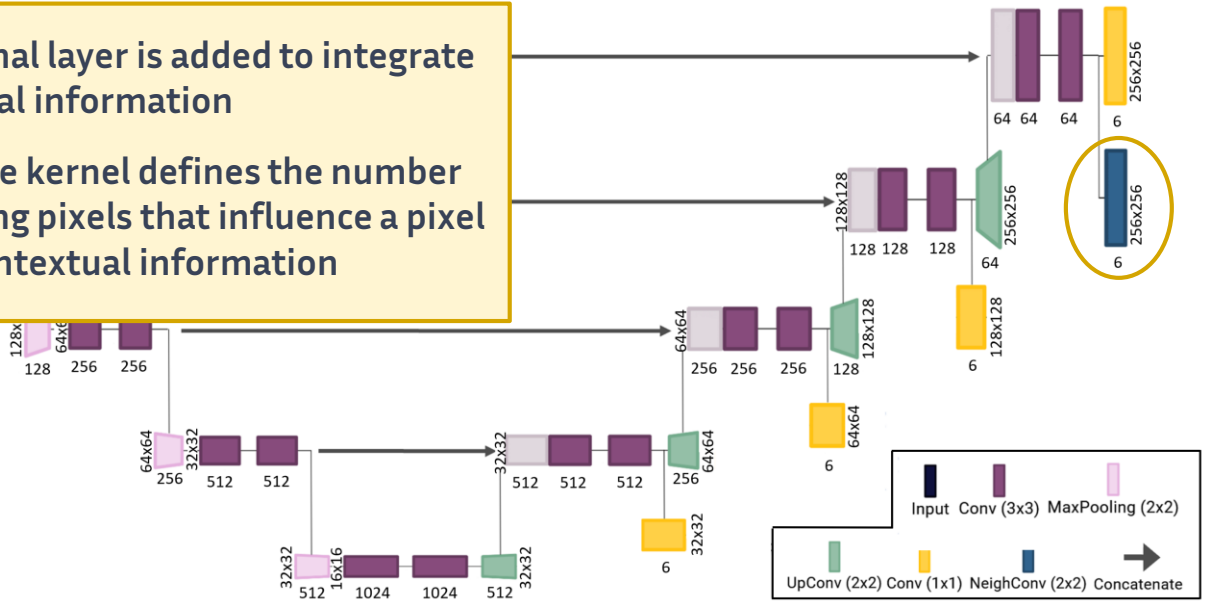
# Overall architecture

- Integration of fully connected NNs in the decoder to retrieve multiscale information
- built through convolutional layers with kernels of size (1 x 1) to work with 2D data and compute pixelwise posterior probabilities



# Overall architecture

- a convolutional layer is added to integrate further spatial information
- the size of the kernel defines the number of neighboring pixels that influence a pixel → spatial-contextual information






UNIVERSITÉ CÔTE D'AZUR Inria Università di Genova

# Loss function

weighting factor for imbalanced training data

$$\mathcal{L}_u = -\frac{1}{L}\sum_{l=1}^{L}\sum_{k=0}^{M-1}\frac{\widehat{P}_{max}}{\widehat{P}_k}\sum_{i\in S^l} t_{ik}\log(p_{ik}), p_{ik} = \frac{\exp(z_{ik})}{\sum_{n=0}^{M-1}\exp(z_{in})}$$

$$\mathcal{L}_p = -\sum_{a=1}^{W_L}\sum_{b=1}^{H_L}\sum_{k=0}^{M-1} t_{ik}\log\left(\sum_{j=0}^{D-1}\sum_{q=0}^{D-1} h_{j,q}\cdot p_{(a-j,b-q),k}\right)$$

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_p$$

$M$ number of classes

$L$ number of resolutions

$H_L, W_L$ height, width of the pixel grid

$p_{ik}$ estimated output probability that pixel $i$ belongs to class $k$

$(z_{i1}, \dots, z_{ij})$ prediction vector

$\widehat{P}_k$ prior probability of the $k$ −th class

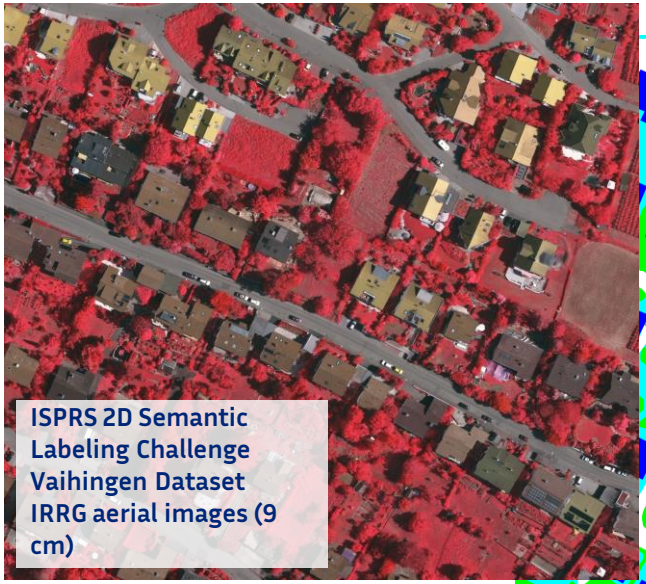
$h_{j,q}$ weight of the kernel of size $D \times D$

# 03

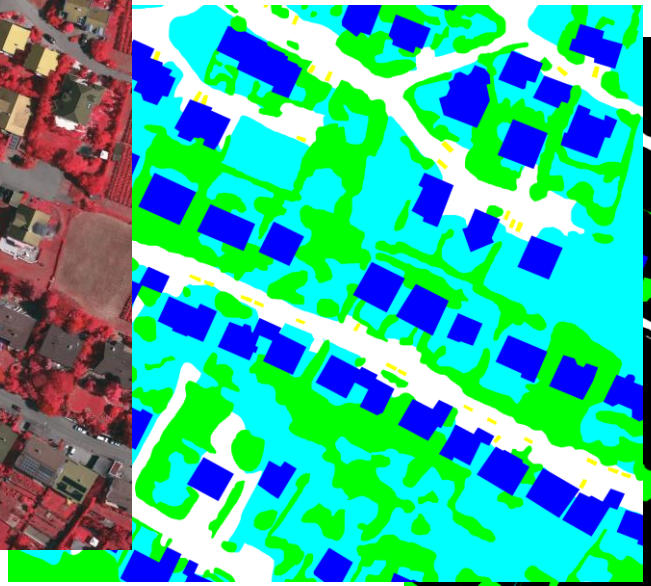
# Experimental results

# Dataset

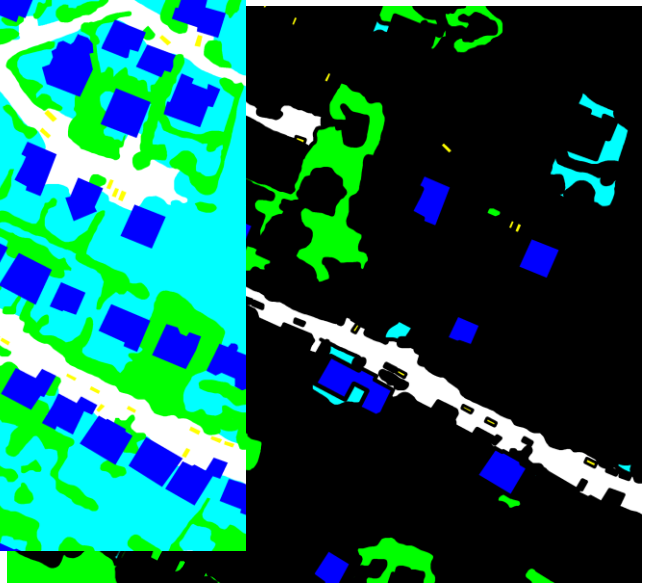
| | | |
|---|---|---|
| Buildings | Roads | Vegetation |
| Trees | Cars | Clutter |
| Unlabeled | | |




*Aerial image*



*Original GT*



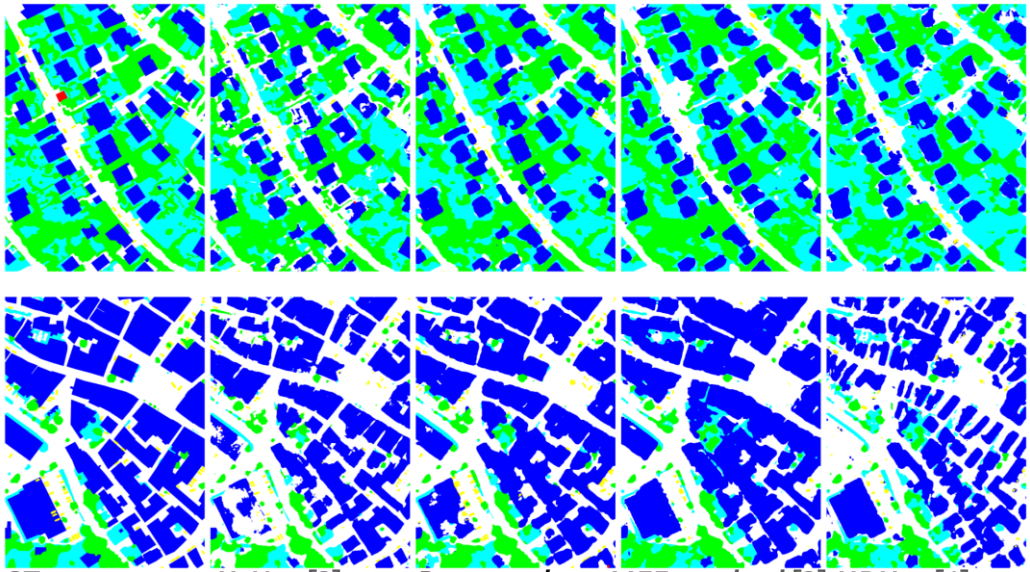
*Scarce GT used to train the model*

https://www2.isprs.org/commissions/comm2/wg4/benchmark/



UNIVERSITÉ CÔTE D'AZUR Inria Università di Genova

# Results




[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput. Ass. Interv.*, ser. LNCS, vol. 9351. Springer, pp. 234–241, 2015.
[3] S. Liu, C. He, H. Bai, Y. Zhang, and J. Cheng, "Light-weight attention semantic segmentation network for high-resolution remote sensing images," in *2020 IGARSS*, pp. 2595–2598, 2020.
[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 CVPR*, pp. 5686–5696, 2019.

# Results (full images)

| | Architecture | buildings | impervious surf. | low vegetation | trees | cars | OA | recall | prec | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Full GT | U-Net [2] | 0.86 | 0.91 | 0.79 | 0.90 | **0.86** | 0.89 | 0.87 | 0.85 | 0.86 |
| | MFF method [3] | **0.96** | **0.93** | 0.71 | 0.89 | 0.61 | 0.89 | 0.82 | **0.88** | 0.85 |
| | HRNet [4] | 0.89 | 0.89 | 0.50 | 0.89 | 0.81 | 0.79 | 0.80 | 0.81 | 0.81 |
| | **Proposed method** | **0.96** | 0.91 | **0.80** | **0.91** | 0.81 | **0.91** | **0.88** | 0.87 | **0.88** |
| Scarce GT | U-Net [2] | 0.87 | **0.93** | 0.64 | **0.87** | 0.76 | 0.82 | 0.81 | 0.84 | 0.83 |
| | MFF method [3] | **0.94** | 0.82 | 0.65 | 0.85 | 0.60 | 0.81 | 0.78 | 0.83 | 0.80 |
| | HRNet [4] | 0.84 | 0.75 | **0.82** | 0.69 | 0.49 | 0.77 | 0.72 | 0.79 | 0.75 |
| | **Proposed method** | **0.94** | 0.78 | 0.80 | **0.87** | **0.90** | **0.86** | **0.86** | **0.85** | **0.86** |

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput. Ass. Interv.*, ser. LNCS, vol. 9351. Springer, pp. 234–241, 2015.
[3] S. Liu, C. He, H. Bai, Y. Zhang, and J. Cheng, "Light-weight attention semantic segmentation network for high-resolution remote sensing images," in *2020 IGARSS*, pp. 2595–2598, 2020.
[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 CVPR*, pp. 5686–5696, 2019.

# Results (cropped images)

| | Architecture | buildings | impervious surf. | low vegetation | trees | cars | OA | recall | prec | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Full GT | FCN + PGM [1] | 0.84 | 0.81 | 0.68 | 0.92 | **0.86** | 0.81 | 0.82 | 0.72 | 0.77 |
| | U-Net [2] | **0.92** | 0.83 | 0.71 | 0.92 | 0.74 | 0.85 | **0.83** | 0.84 | 0.83 |
| | MFF method [3] | 0.91 | **0.97** | 0.61 | 0.88 | 0.64 | 0.85 | 0.80 | 0.82 | 0.81 |
| | HRNet [4] | 0.89 | 0.81 | 0.42 | 0.92 | 0.63 | 0.77 | 0.73 | 0.76 | 0.74 |
| | **Proposed method** | **0.92** | 0.85 | **0.73** | **0.93** | 0.73 | **0.86** | **0.83** | **0.85** | **0.84** |
| Scarce GT | FCN + PGM [1] | 0.94 | 0.68 | 0.49 | 0.86 | **0.74** | 0.76 | 0.74 | 0.75 | 0.75 |
| | U-Net [2] | **0.96** | 0.65 | 0.47 | **0.89** | 0.48 | 0.76 | 0.69 | **0.81** | 0.75 |
| | MFF method [3] | 0.91 | 0.75 | 0.44 | 0.87 | 0.51 | 0.76 | 0.70 | 0.74 | 0.72 |
| | HRNet [4] | 0.84 | 0.77 | 0.68 | 0.80 | 0.29 | 0.77 | 0.68 | 0.74 | 0.71 |
| | **Proposed method** | 0.90 | **0.79** | **0.70** | **0.89** | 0.68 | **0.82** | **0.79** | **0.81** | **0.80** |

# 04

# Conclusions and perspectives

# Conclusions and perspectives

- Novel semantic segmentation method based on FCN, fully connected NNs, and a spatial loss function.
- Addresses a major challenge of deep learning for semantic segmentation of VHR RS images in case of scarce GT
- The advantages of the proposed techniques are progressively more relevant as the training set is farther from the ideal densely-labeled case and closer to real-world annotations

## Future work

- extension of this methodology to the multisensor case, with optical and radar images acquired by different missions and therefore with different spatial resolutions, frequencies, and bands.
- study of the effect on classification of the incorporation of different amounts of long-range spatial information (adaptive kernel)
- integration of the proposed method with transfer learning to work with datasets related to other real world applications, such as natural disaster management

Thank you very much for your attention!

UNIVERSITÉ CÔTE D'AZUR Inria Università di Genova