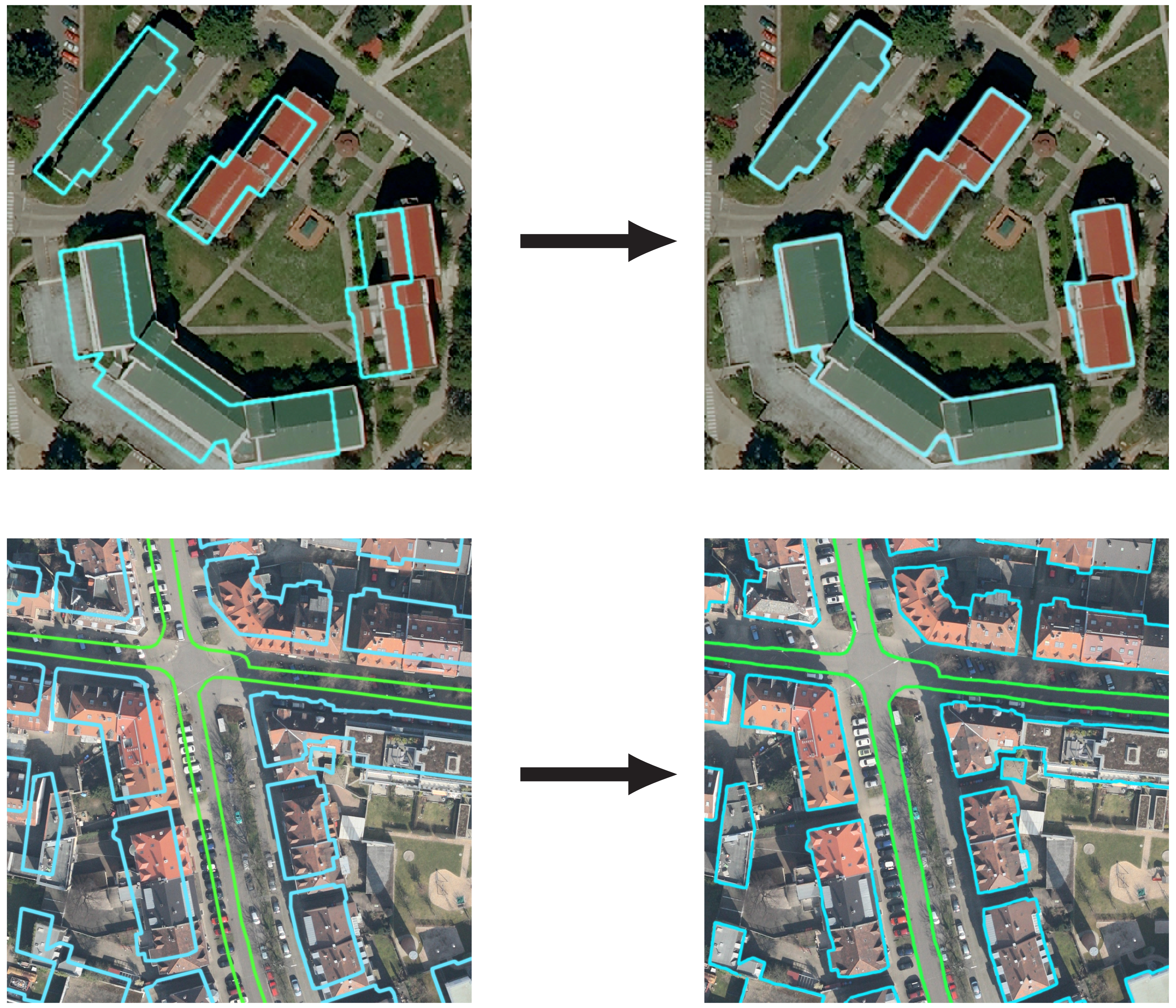
Multimodal image alignment through a multiscale chain of neural networks, with application to remote sensing

Armand Zampieri[1] Guillaume Charpiat[2] Nicolas Girard[1] Yuliya Tarabalka[1]

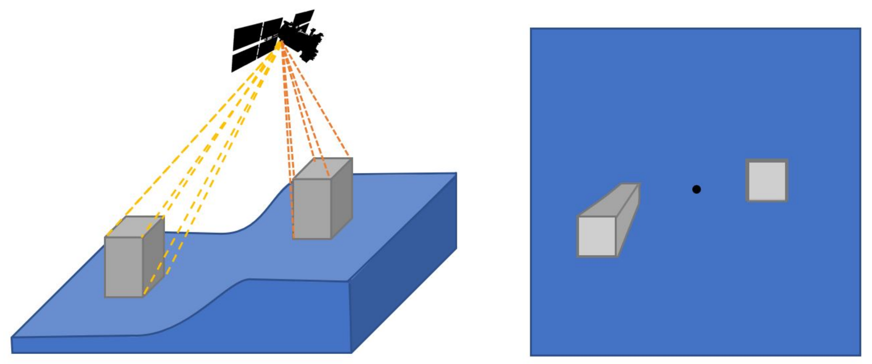
armand.zampieri@gmail.com [1] TITANE team, INRIA, Université Côte d'Azur. [2] TAU team, INRIA Saclay, LRI, Université Paris-Sud. firstname.lastname@inria.fr

Introduction

- **Goal: multi-modal alignment**



- **Data**: – **optical images:** RGB/multi-spectral pictures from satellite / airplane
  – **binary cadaster maps:** building rooftops, roads, etc. as polygons
- **Why?**: used as ground truth when **training segmentation algorithms** such as [1]) but actually often **not aligned**, because of:
  – **different angles of capture**, making rooftops move (even on orthorectified images, as Digital Terrain Model is not precise and does not include buildings)
  – **human error** when annotating buildings
  – **lack of precision** of the groundtruth data



Framework

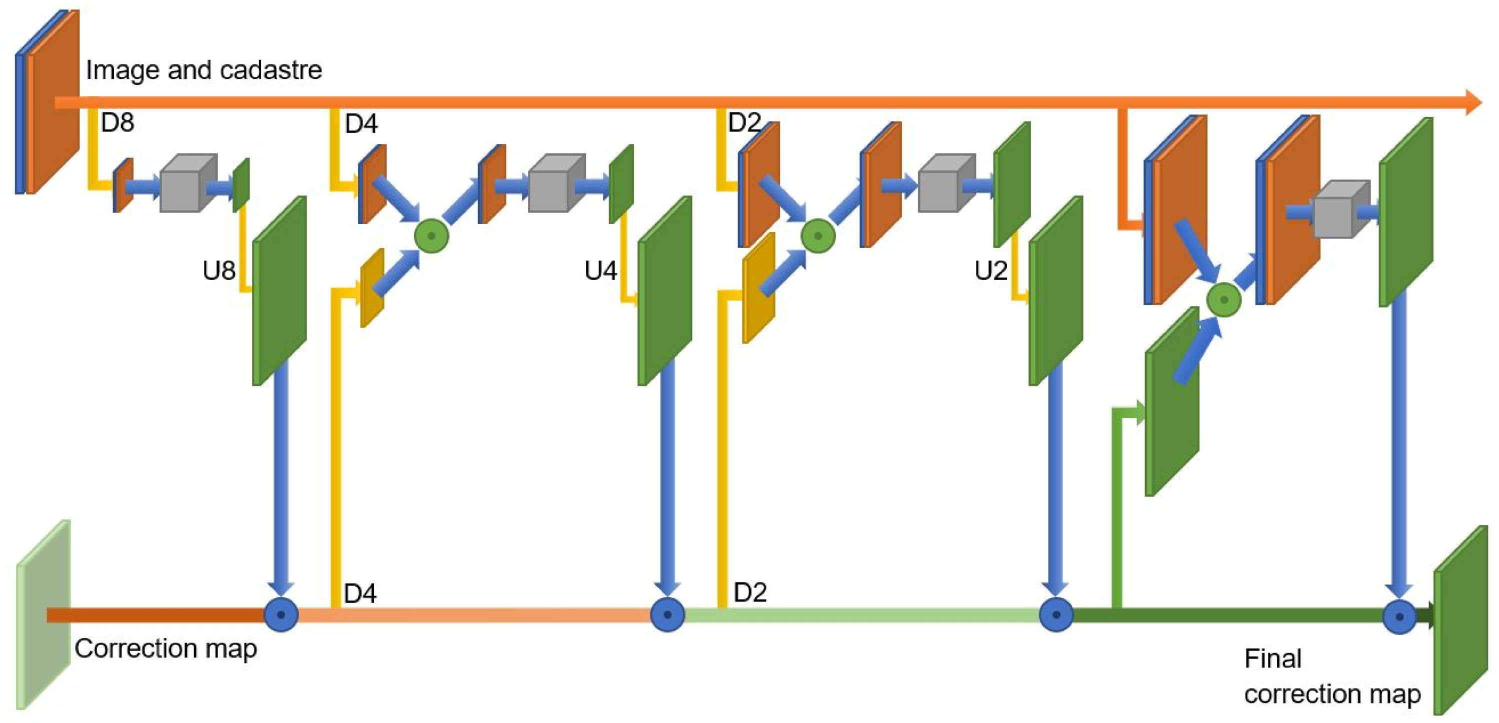
- **Inputs:** optical image and polygon raster of misaligned buildings
- **Output: displacement map** v that aligns the building polygons to the image
- **Loss:** $\sum_{\text{pixels } p} w_p \, \|v(p) - v_{GT}(p)\|^2$ with weights depending on pixel class / type
- **Difficulties: many small objects, multimodality**, varied classes, shadows, trees...
  ⇒ do not try to learn common descriptors to be matched later
  ⇒ but predict directly the displacement

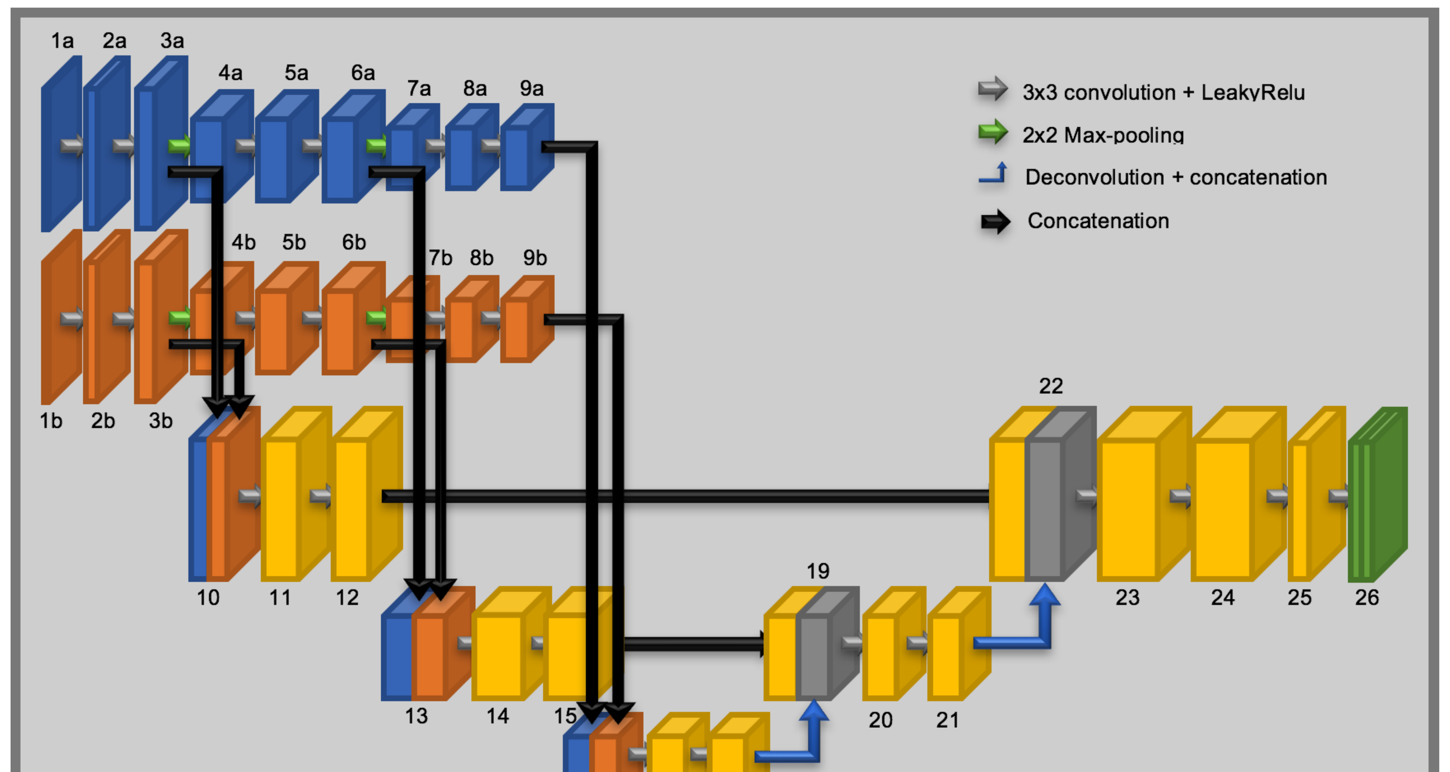
Dealing with scale

- **Large displacements** ⇒ difficult problem, network **optimization gets stuck!**
- **Trick: sufficiently zoomed-out images are perfectly registered**
  ⇒ **zoom in again progressively** and correct **small displacements** that appear
  ⇒ **multi-scale!**
- **Solution: Fully Convolutional** Neural Network applied **iteratively at increasing resolutions**, expecting small displacements at each scale
  ⇒ chain of double-U-nets

Complete pipeline

Coarse-to-fine displacement map estimation





at each scale (grey blocks)

Running time

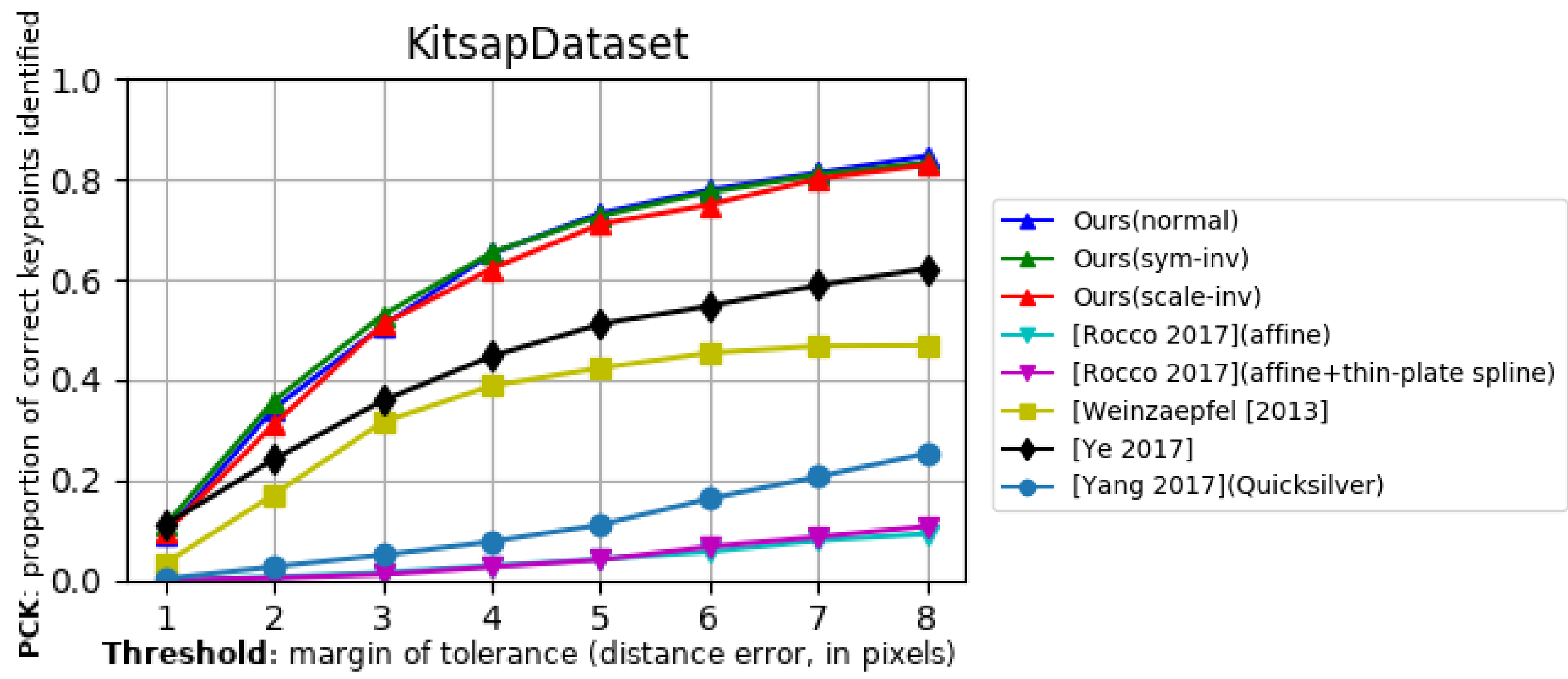
Method	Time	CPU	GPU
Ours	**80s**	2Ghz	GTX 1080Ti
[2]	238s	2.7Ghz	Q.M2000M
[4]	784s	2.7Ghz	Q.M2000M
[6]	9550s	3.5Ghz	GT 960M

for a 5000×5000 px image (NB: us=dense)

Code

`https://github.com/Lydorn/cadaster-alignement`

Quantitative Results



- **Accuracy measure**: proportion of points p for which $\|v(p) - v_{GT}(p)\| \leq$ given distance error threshold
- **Ground truth**: generated by applying random deformations to carefully-picked well-registered OpenStreetMap data
- **Three variations**:
  – normal: chain of four networks trained at different scales (2^s)
  – symmetry-invariant: averaged over 8 input transformations (mirroring/rotation)
  – scale-invariant: one network trained for one scale and applied to every scale
  ⇒ similar performance, far above other approaches (twice more precise)
- Performs well on other problems also (multiclass/roads alignment, stereovision...)

Conclusion

- Very effective at aligning buildings over aerial images, generalizes very well on different problems or data types.
- No strong scale dependence
- Later work: building height estimation

References

[1] E. Maggiori, Y. Tarabalka, G. Charpiat and P. Alliez. Can semantic labeling methods generalize to any city? the Inria aerial image labeling benchmark. IGARSS 2017.

[2] I. Rocco, R. Arandjelovic and J. Sivic. Convolutional neural network architecture for geometric matching. CVPR 2017.

[3] O. Ronneberger, P. Fischer and T. Brox. U-net: Convolutional networks for biomedical image segmentation. MICCAI 2015.

[4] P. Weinzaepfel, J. Revaud, Z. Harchaoui and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. ICCV 2013.

[5] X. Yang, R. Kwitt and M. Niethammer. Quicksilver: Fast predictive image registration - a deep learning approach. NeuroImage 2017.

[6] Y. Ye, J. Shan, L. Bruzzone and L. Shen. Robust registration of multimodal remote sensing images based on structural similarity. TGRS 2017.