Correcting misaligned buildings over aerial images by a Deep Learning multi-resolution approach

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

[1] Université Côte d'Azur, Inria, TITANE team
[2] TAU team, INRIA Saclay, LRI, Université Paris-Sud

nicolas.girard@inria.fr

Figure 1: Alignment problem

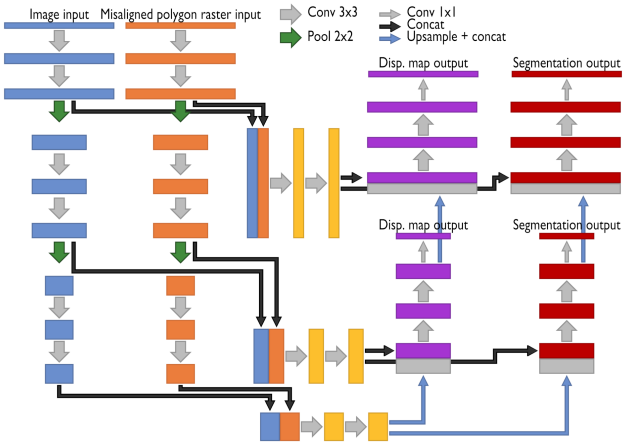

Figure 2: CNN model inspired by U-Net, here drawn with 2 pooling operations (the final model has 3 pooling operations which makes the network deeper).

A common problem in remote sensing is the spacial misalignment of various data sources. For example, groundtruth data of building rooftops often do not align with the buildings in images as in Fig.1. This can be due to:

- Different angles of capture which make the rooftops move (even on orthorectified images because the Digital Terrain Model is not precise and does not include buildings)
- Human error when annotating the buildings
- Lack of precision of the groundtruth data source

We propose a Deep Learning method that builds on [1]. Its primary objective is to compute a dense displacement map that aligns the building polygons on the image (for example building polygons from Open Street Map [2]). We added a second objective which is to output a segmentation of the buildings from the optical image (otherwise known as pixel-wise classification).
The main building block of the method is a modification of the widely-used U-Net [3] with 2 image inputs and 2 image outputs, as seen in Fig.2.
The total loss is a combination of 2 types of losses:

- Displacement map loss: is the mean squared error of the predicted displacement vectors for each pixel;
- Segmentation loss: is the cross entropy of the predicted class for each pixel.

The authors would like to thank ANR for funding the study.

The segmentation loss helps to train the network as the model has to learn where building are in order to predict the displacement map.
Additionally, each of those 2 losses combine intermediary losses at different resolutions stages inside the Convolutionnal Neural Network. As can be seen in Fig.2, there are 2 displacement map outputs (and 2 segmentation outputs). The first output is at the original resolution and the second output is at half the original resolution. As the final model has 3 pooling operations, it has a third intermediary loss at a quarter of the original resolution. These intermediary losses allow the gradients to propagate better in the network and help to start the training. After a certain number of iterations, only the losses at the original resolution are kept.
The full alignment method can deal with polygons misaligned by as much as 32 pixels. However training a model to learn a map with a maximum amplitude of 32 pixels is very hard and requires a big network because a misaligned vertex can be anywhere in a window of size $(2 * 32 + 1)^2 = 4225$ pixels. Instead, we use a multi-resolution approach. In the first stage the inputs are downsampled by a factor of 8, then fed to the network, then the polygons are aligned by the output displacement map and then upsampled by a factor of 8. We repeat the process with downsampling factors 4, 2 and then 1. Now the model at each stage has to predict a displacement with an amplitude up to 4 pixels. This way a misaligned vertex can be anywhere in a window of size $(2*4+1)^2 = 81$ pixels which is much less than before, the complexity of the task is greatly reduced. We thus have 4 different models trained in parallel with downsampled inputs and artificially generated displacements of maximum amplitude equal to 4 pixels.
We tested the method on a test image for which we know the true misalignment due to a change of angle of capture. See Fig.3 for a small crop of that image. Our method generalizes very well to cities the network has never seen before, such as this test image.

Figure 3: Green: groundtruth, red: misaligned, blue: aligned by the proposed method. When correctly aligned, the polygon has a cyan color.

References

[1] Armand Zampieri, Guillaume Charpiat and Yuliya Tarabalka, *Coarse to fine non-rigid registration: a chain of scale-specific neural networks for multimodal image alignment with application to remote sensing,* arXiv preprint arXiv:1802.09816, 2018.

[2] OpenStreetMap contributors, *Planet dump retrieved from https://planet.osm.org,* https://www.openstreetmap.org, 2017.

[3] Olaf Ronneberger, Philipp Fischer and Thomas Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation,* arXiv preprint arXiv:1505.04597, 2015.