NOISY SUPERVISION FOR CORRECTING MISALIGNED CADASTER MAPS WITHOUT PERFECT GROUND TRUTH DATA

Nicolas Girard[1], *Guillaume Charpiat*[2] *and Yuliya Tarabalka*[1,3]

[1]TITANE team, Inria, Université Côte d'Azur, France
[2]TAU team, Inria Saclay, LRI, Université Paris-Sud, France
[3]LuxCarta Technology
email: firstname.lastname@inria.fr

ABSTRACT

In machine learning the best performance on a certain task is achieved by fully supervised methods when perfect ground truth labels are available. However, labels are often noisy, especially in remote sensing where manually curated public datasets are rare. We study the multi-modal cadaster map alignment problem for which available annotations are misaligned polygons, resulting in noisy supervision. We subsequently set up a multiple-rounds training scheme which corrects the ground truth annotations at each round to better train the model at the next round. We show that it is possible to reduce the noise of the dataset by iteratively training a better alignment model to correct the annotation alignment.

***Index Terms*—** Noisy supervision, Multi-modal alignment, Ground truth annotation correction, Optical images

1. INTRODUCTION

One of the main tasks in remote sensing is semantic segmentation. The supervised approach needs good ground truth annotations most often in the form of class-labeled polygons outlining objects of the image. However, these good annotations are hard to come by because even if they exist (for example OpenStreetMap (OSM) annotations [1]), they can be misaligned due to human error, imprecise digital terrain model or simply a lack of precision of the original data (scanned cadaster maps from local authorities). Each object can be misaligned in a different way compared to surrounding objects and the misalignment can include complex deformations such as slight stretching and rotation.

The aim of this paper is to correct the alignment of noisy OSM annotations when only these annotations are available. Several related works tackle the noisy label problem. Some use special losses to explicitly model the label noise [2] which penalize erroneous outputs less if they could be due to label noise. Others perform simultaneous edge detection and alignment [3] which can handle small displacements in an unsupervised manner. The task of aligning OSM annotations has already been tackled in [4], using convolutional neural networks for building segmentation and a Markov Random Field for aligning buildings onto the building segmentation image. However, the neural network has to be trained on a small dataset of building image with corresponding good ground-truth annotations.

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

Fig. 1: Qualitative alignment results on a crop of an image of Bloomington from the Inria dataset. Red: initial OSM annotations; green: aligned annotations.

We propose in this paper to use the self-supervised multi-task multi-resolution deep learning method for aligning cadaster maps to images of [5] in the noisy supervision setting. The dataset used for training that method has misalignment noise and still the model learned to align. We will explore this interesting behavior and experiment a kind of unsupervised learning to correct noisy annotations with a model trained on these noisy annotations. See Fig. 1 for an example of results. It leverages the natural tendency of neural networks to be robust to a certain amount of noise and does not require any special loss function.

2. METHODOLOGY

We provide here a short description of the self-supervised multi-task multi-resolution deep learning alignment method

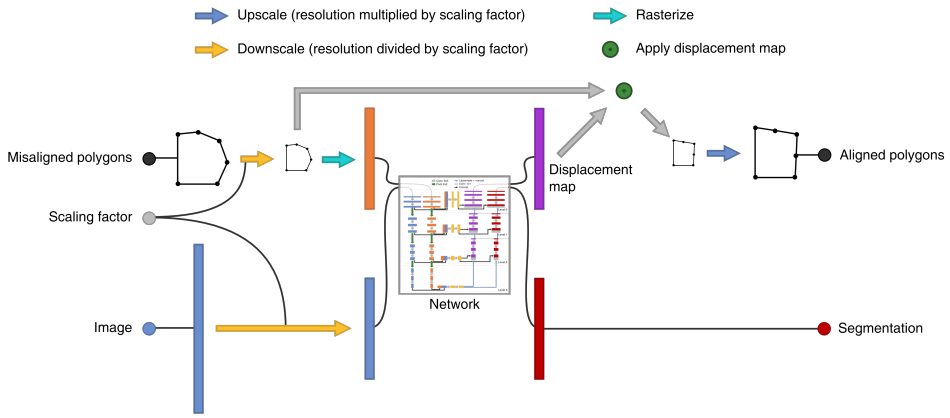

Fig. 2: One step of the base alignment method, applied repeatedly at increasing resolutions for the final alignment.

of [5] (referred to by "base alignment method" from now on), focusing on the most relevant parts. Its code is available here: https://github.com/Lydorn/mapalignment.

Mathematical modeling. Given two images I and J of same size $H \times W$, but of different modalities, e.g. with I an RGB image (picture from a satellite) and J a binary image (cadaster, indicating for each pixel whether it belongs to a building or not), the alignment problem aims at finding a deformation, i.e. a 2D vector field $\mathbf{f}$ defined on the discrete image domain $[1, H] \times [1, W]$, such that the warped second image $J \circ (\mathrm{Id} + \mathbf{f})$ is well registered with the first image I. To do this, in a machine learning setting, we consider triplets $(I, J, \mathbf{f}_{\mathrm{gt}})$ consisting of two images together with the associated ground-truth deformation $\mathbf{f}_{\mathrm{gt}}$. Image pairs (I, J) are given as inputs, and the model's estimated deformation $\hat{\mathbf{f}}$ is optimized to be close to the ground truth deformation $\mathbf{f}_{\mathrm{gt}}$.

Displacement map cost function. The displacement map loss function is the mean squared error between the predicted displacement map $\hat{\mathbf{f}}$ and the ground truth displacement map $\mathbf{f}_{\mathrm{gt}}$. The actual loss used by the base alignment method is a little more complex but for the purpose of this paper we can consider the simplified loss:

$$L^{\mathrm{disp}}(\hat{\mathbf{f}}) = \sum_{\mathbf{x} \in [1,H] \times [1,W]} \left\| \hat{\mathbf{f}}(\mathbf{x}) - \mathbf{f}_{\mathrm{gt}}(\mathbf{x}) \right\|_2^2 \quad (1)$$

Model. The neural network used by the base alignment method is a transformed U-Net [6] with 2 image inputs and 2 image-like outputs for the displacement map and the segmentation image, see Fig. 2 for a schema of the model. The segmentation output is only used during training, having its own cross-entropy loss function. The input image I has 3 channels, with real values normalized to $[-1, 1]$, standing for RGB. The input misaligned polygon raster J also has 3 channels, with Boolean values in $\{0, 1\}$, corresponding to polygon interior, edge and vertices. The output displacement map has 2 channels with real values in $[-4$ px, 4 px$]$, standing for the x and y components of the displacement vector. The model uses a multi-resolution approach by applying a neural network at increasing resolutions, iteratively aligning polygons from a coarse to fine scale. The scales used are $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ and 1. Thus displacements of up to 32 px can be handled.

3. EXPERIMENTAL SETUP

Dataset. The model is trained on a building cadaster dataset consisting of the two available following ones: Inria Aerial Image Labeling Dataset [7], and "Aerial imagery object identification dataset for building and road detection, and building height estimation" [8]. The resulting dataset has 386 images (with the majority being 5000×5000 px) of 16 cities from Europe and USA. Each image has in average a few thousand buildings. The building footprints were pulled from OSM for all images. these polygon annotations are inconsistent across images alignment-wise (see Fig. 1 and 3). Some are perfect, and some are misaligned by up to 30 px. However the base alignment method [5] assumes perfect annotations in its formulation.

Self-supervised training. The model needs varied ground truth labels (displacements) in order to learn, while the dataset is assumed to be made of aligned image pairs only ($\mathbf{f} = 0$). The dataset is thus enhanced by adding random deformations in the form of 2D Gaussian random fields for each coordinate with a maximum absolute displacement of 32 px. The polygon annotations A are then inversely displaced by the generated displacements, to compute the misaligned polygons, which are then rasterized. We obtain training triplets of the form $(I, J, \mathbf{f})$ with $J = rast(A \circ (\mathrm{Id} + \mathbf{f})^{-1})$. For the multi-resolution pipeline, 4 different models are trained independently with downscaling factors 8, 4, 2 and 1 (one per resolution). For clarification, labels are the ground truth displacements at each pixel, i.e. a 2D vector, and annotations are the polygons outlining objects.

Multiple-rounds training. We train the base alignment

method with the same hyper-parameters as [5], which were selected to avoid overfitting. We perform multiple rounds of training on the whole dataset to achieve our goal of aligning the whole dataset. It consist of iteratively alternating between training the alignment model on the available dataset (see Alg. 1) and correcting the alignment of the training dataset to provide a better ground truth for the next training round. The multiple-rounds training is explained in Alg. 2.

Algorithm 1: Alignment training [5]

Input: Images $\mathcal{I} = \{I, ...\}$ and corresponding annotations $\mathcal{A} = \{A, ...\}$

Build dataset with random deformations: $\mathcal{D} = \{(I, J_{rand}, \mathbf{f}_{rand}), ...\}$ with $J_{rand} = rast(A \circ (\mathrm{Id} + \mathbf{f}_{rand})^{-1})$;

Train multi-resolution model M to perform this mapping: $(I, J_{rand}) \mapsto \mathbf{f}_{rand}$;

Output: Trained model M

Algorithm 2: Multiple-rounds training

Input: Original annotations $\mathcal{A}_0$, number of rounds R

for $r = 1$ **to** R **do**

1. Get model M_r using Alg. 1 with input $\mathcal{A} = \mathcal{A}_{r-1}$;
2. Apply M_r on the original annotations $\mathcal{A}_0$: $\mathcal{A}_r = M_r(\mathcal{A}_0)$;

Output: Aligned annotations A_R

Ablation studies. To justify the design choices of the multiple-rounds training, we performed ablation studies. The first ablation study (AS1) changes the second step of Alg. 2 by applying the model on the previous corrected annotations instead of the original annotations: $\mathcal{A}_r = M_r(\mathcal{A}_{r-1})$ in order to test whether it is better to iteratively align annotations. The second ablation study (AS2) trains the model only once on the original annotations, and applies it R times to iteratively align the annotations (as in AS1). This is implemented by additionally replacing step 1 of Alg. 2 by $M_r = M_1$ for $r > 1$ and leaving it as is for $r = 1$. This is meant to test the usefulness of re-training.

Robustness to noise. In an additional experiment (Noisier) we misaligned all original annotations further with random zero-mean displacements up to 16 px. We then applied our alignment method for correcting these noisier annotations to study its robustness to more noise.

4. RESULTS

As annotations of our dataset are noisy they cannot be used as ground truth to measure quantitative results. We can first visualize qualitative results in Fig. 3. In order to measure the effectiveness of the multiple rounds training, we manually

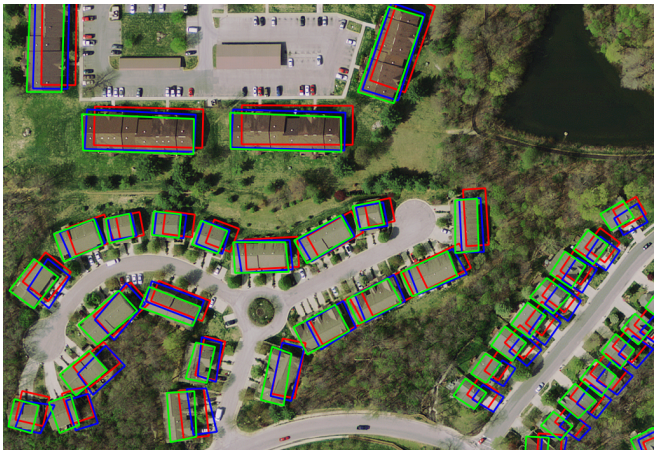

Fig. 3: Qualitative alignment results on a crop of bloomington22 from the Inria dataset. Red: initial dataset annotations; blue: aligned annotations round 1; green: aligned annotations round 2.

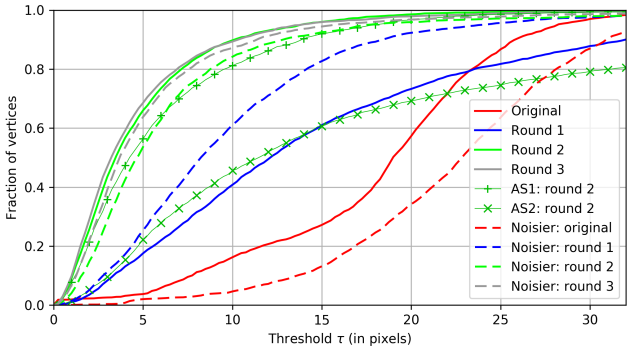

Fig. 4: Accuracy cumulative distributions measured with the manually-aligned annotations of bloomington22 from the Inria dataset.

aligned annotations for one 5000×5000 px image (771 buildings) to get a good ground-truth. We chose the bloomington22 image because it has severe misalignment. To measure the accuracy of an alignment, for any threshold τ we compute the fraction of vertices whose ground truth point distance is less than τ. In other words, we compute the Euclidean distance in pixels between ground truth vertices and aligned vertices, and plot the cumulative distribution of these distances in Fig. 4 (higher is better) for all experiments and rounds.

5. DISCUSSION AND CONCLUSION

After the first round of training, the annotations are on average better aligned than the original annotations, but in some cases the polygons are pushed into the wrong direction, resulting in poorer accuracy for some threshold levels (see the blue curve sometimes under the red curve in Fig. 4). However after the second round of training, the annotation alignment has been

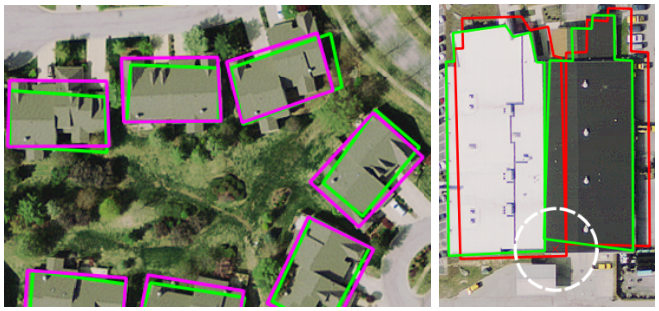

Fig. 5: **Left**: ambiguity of the perfect ground truth annotations. **Right**: alignment failure case. Magenta: manually aligned annotations; red: original dataset annotations; green: aligned annotations round 2.

significantly improved upon the first round (error divided by more than 3 for any quantile, cf. green curve compared to blue curve). The 3rd round does not bring any significant improvement in this case.

Note that a perfect alignment score cannot be expected, because of the ambiguity of the "perfect" ground truth. Indeed, when manually aligning bloomington22's annotations, we observed that the majority of buildings are annotated by a coarse polygon that does not outline the building precisely. Best aligning such a coarse polygon to a real, more complex building becomes an ill-posed problem, with multiple equally-good solutions, which creates ground truth ambiguity. See Fig. 5 (left) for an illustration of this problem, especially the building on the top-right. Fig. 5 (right) shows an example of a mistake of our approach. The left building was successfully aligned (through a slight vertical and horizontal squashing), but the adjoining building on the right was not, because the model only learned smooth displacement maps. A more well-designed displacement map generation allowing discontinuities could solve such problems.

The first ablation study shows the importance of aligning the original annotations in the second step of Alg. 2 as it achieves better accuracy. Indeed the aligned annotations after round 1 can be worse than the original annotations (see some blue polygons of Fig. 3 and the blue curve of Fig. 4), and consequently more difficult to align. The second ablation study shows that the re-training step in round 2 is very important, as skipping it does not improve the alignment compared to round 1.

An explanation of how this method is able to align misaligned annotations by training on these misaligned annotations could be that the dataset contains enough perfect ground truth annotations to steer the gradient descent in the right direction, while being mildly affected by noisy labels (even if the noise is not zero-mean) if overfitting is avoided. However the last experiment (Noisier) invalidates this explanation because in that case the fraction of well-aligned ground truth is negligible and still the model was able to align noisier annotations virtually as well as it did original annotations (it however needs a 3rd round to do so). Our current tentative explanation is that ground truth labels have a zero-mean noise (without bias). For the alignment task, the network tries to minimize the average error it makes. As such it tends to predict the mean value of the labels when it cannot do better. This is the case if the label noise is independent of the input, and if overfitting noisy labels is avoided. The network will learn the mean alignment, which corresponds to the underlying perfect ground truth. This explanation is further supported by a recent work on image restoration without clean data [9], where noisy images are de-noised by training a de-noising network on noisy images only.

In conclusion, even noisy/misaligned annotations are useful. Our model can be iteratively trained on them and align these annotations through a multiple-round training scheme.

6. REFERENCES

[1] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," 2017.

[2] Volodymyr Mnih and Geoffrey Hinton, "Learning to label aerial images from noisy data," in *ICML*, 2012.

[3] Zhiding Yu, Weiyang Liu, Yang Zou, Chen Feng, Srikumar Ramalingam, B. V. K. Vijaya Kumar, and Jan Kautz, "Simultaneous edge alignment and learning," in *ECCV*, Sept 2018.

[4] J. E. Vargas-Muoz, D. Marcos, S. Lobry, J. A. dos Santos, A. X. Falco, and D. Tuia, "Correcting misaligned rural building annotations in open street map using convolutional neural networks evidence," in *IGARSS*, 2018.

[5] Nicolas Girard, Guillaume Charpiat, and Yuliya Tarabalka, "Aligning and updating cadaster maps with aerial images by multi-task, multi-resolution deep learning," in *ACCV*, Dec 2018.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, 2015.

[7] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez, "Can semantic labeling methods generalize to any city? the Inria aerial image labeling benchmark," in *IGARSS*, 2017.

[8] Kyle Bradbury, Benjamin Brigman, Leslie Collins, Timothy Johnson, Sebastian Lin, Richard Newell, Sophia Park, Sunith Suresh, Hoel Wiesner, and Yue Xi, "Aerial imagery object identification dataset for building and road detection, and building height estimation," July 2016.

[9] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila, "Noise2noise: Learning image restoration without clean data," *CoRR*, 2018.