This manuscript was accepted to be presented at the IEEE International Geoscience and Remote Sensing Symposium 2020.

MAP-REPAIR: DEEP CADASTRE MAPS ALIGNMENT AND TEMPORAL INCONSISTENCIES FIX IN SATELLITE IMAGES

Stefano Zorzi[1], Ksenia Bittner[2], Friedrich Fraundorfer[1]

[1] Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria
(stefano.zorzi, fraundorfer)@icg.tugraz.at
[2] Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany
ksenia.bittner@dlr.de

ABSTRACT

In the fast developing countries it is hard to trace new buildings construction or old structures destruction and, as a result, to keep the up-to-date cadastre maps. Moreover, due to the complexity of urban regions or inconsistency of data used for cadastre maps extraction, the errors in form of misalignment is a common problem. In this work, we propose an end-to-end deep learning approach which is able to solve inconsistencies between the input intensity image and the available building footprints by correcting label noises and, at the same time, misalignments if needed. The obtained results demonstrate the robustness of the proposed method to even severely misaligned examples that makes it potentially suitable for real applications, like *OpenStreetMap* correction.

***Index Terms*—** deep learning, segmentation, building footprint, remote sensing, high-resolution aerial images, cadastre map alignment

1. INTRODUCTION

Semantic segmentation is still a challenging problem in Remote Sensing. Automatic detection and extraction of precise object outlines, such as human constructions, is in the interest of many cartographic and engineering applications. The most effective way to deal with this problem is the use of *Convolutional Neural Networks* trained in a supervised manner. Accurate ground truth annotations allows to achieve great detection and segmentation accuracies, however, these good annotations are hard to come by because they might be misaligned due to multiple causes e.g. human errors or imprecise digital terrain model. Furthermore, the maps may not be temporally synchronized with the satellite images failing to take into account variations in the constructions, i.e. new buildings may have been built or destroyed.

Several related works tackle this problem with different approaches. Good alignment performance are achieved in [1] by training a CNN to predict a displacement field between a map and an image. The same authors proposed in [2] a multi-rounds training scheme which ameliorates ground truth annotations at each round to fine-tune the model. More recently, a method that performs a sequential annotation adjustment using a combination of consistency and self-supervised losses has been published [3].

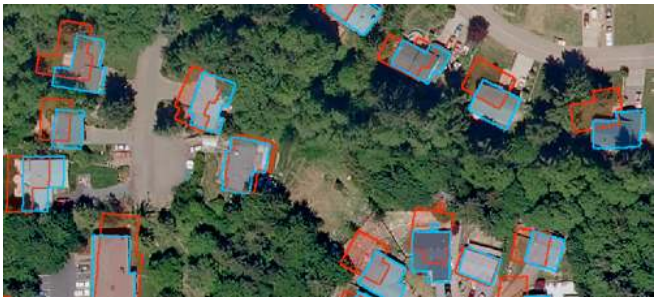

Fig. 1: *MapRepair* result. Misaligned annotations in red, corrected map in cyan.

In this paper we propose an end-to-end self-supervised deep learning method for the generation of aligned and temporally coherent cadastre annotation in satellite and airborne imagery. The aim of the method is to align in one single shot all the object instances present in the intensity image and, at the same time, detect obsolete footprints and segment constructions that lack annotations.

2. METHODOLOGY

Our goal is to train a deep neural network which can not only generate an aligned cadastre map, but can also remove obsolete footprints and detect new buildings. The overall network model is shown in Figure 2, and it is composed of two different branches. The first branch estimates and performs a projection for every building instance in order to produce a map perfectly registered with the intensity image. During this process, obsolete footprints are discarded. If a building does not have a footprint in the map, the second branch segments and regularizes the construction providing an accurate and visually pleasing building boundary. The results from two paths are then merged to produce the final corrected map.

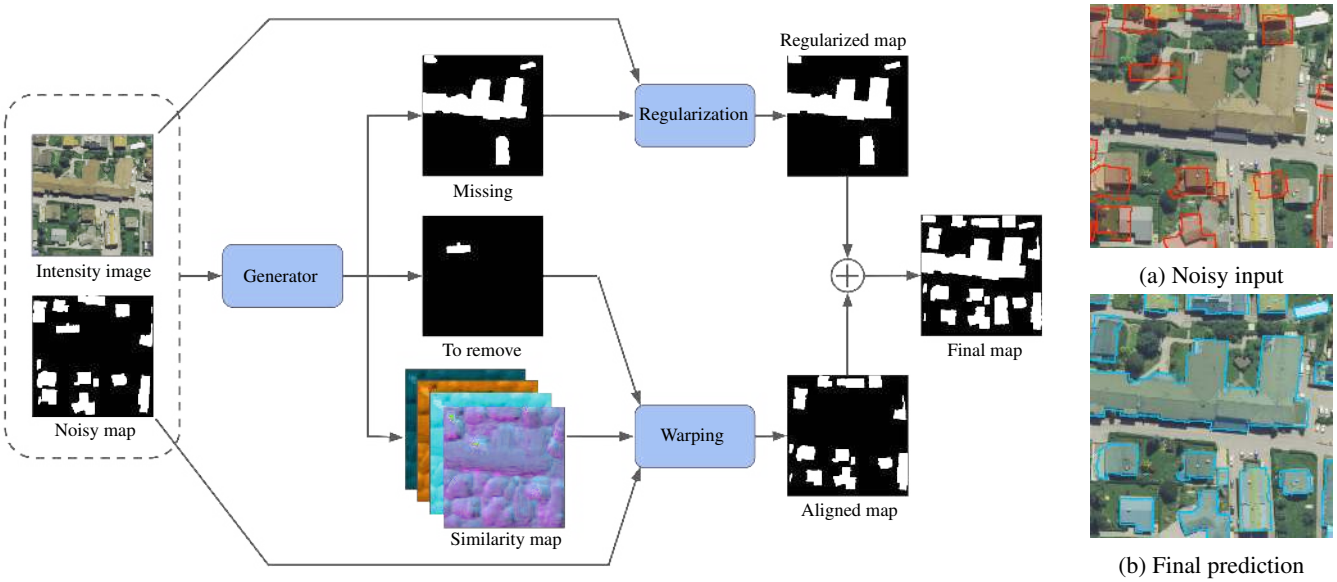

Fig. 2: Workflow of the proposed *MapRepair* method. The intensity image and the noisy annotations are given to a network that generates a transformation map and segments missing and obsolete footprints. The aligned cadastre layer is produced by a warping function while the segmentation is improved by a regularization network. The final refined annotations are obtained merging the results from the regularization branch and the warping branch. Noisy and refined annotations overlaid to the RGB image are shown on the right side.

2.1. Similarity map estimation and instance warping

In order to better align individual object instances to the image content, a generator network G is exploited to predict a similarity transformation map $T \in \mathbb{R}^{4\times H\times W}$ where the channels store translation (along x and y axis), rotation and scale values for each pixel location. The model receives as input the intensity image $I \in \mathbb{R}^{3\times H\times W}$ and a binary mask $y = \{0,1\}^{H\times W}$ which represents the noisy or misaligned annotations.

$$T = G(I, y) \tag{1}$$

A similarity transformation is then computed independently for every building averaging the values of the tensor T under the area described by the object instance. The transformation for the i-th instance can be written as:

$$t_i = \frac{1}{N_i} \sum_{p\in\omega_i} T_p \tag{2}$$

where N_i and ω_i are the number of points and the set of points of the instance i-th, respectively. The four values of t_i define a $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ similarity transformation.

The refined annotation for the i-th object instance $\hat{y}_i$ is expressed as the transformed version $\hat{y}_i = \text{warp}(y_i, t_i)$ of the noisy instance annotation y_i by the predicted transformation t_i.

The predicted aligned annotations $\hat{y}$ for the binary image y is then calculated as the combination of the single instance transformations and can be expressed as:

$$\hat{y} = \sum_{i=1}^{M} \text{warp}(y_i, t_i) \tag{3}$$

where y_i represents the i-th instance of the noisy binary mask and M is the number of object instances in the sample image.

The loss function used to train the model to generate the similarity transformation map is a combination of the mean squared error and the mean absolute error between the predicted binary annotations $\hat{y}$ and the ground truth annotations.

2.2. Segmentations and regularization

Maps may not be temporally synchronized with the satellite or airborne data, failing to take into account variations in human constructions, i.e. removed or new buildings.

In order to solve this problem, the model G also predicts two segmentation masks: the first represents footprints of buildings that lack of annotation in y, while the second shows the annotations that must be removed because obsolete.

The missing footprints predicted by G have round corners and an irregular shape due to the lack of geometric constraints during the prediction. In order to ameliorate the segmentation result we post-process the result with the regularization model proposed in [4]. This network for footprint refinement is capable of generating regular and visually pleasing building boundaries without losing segmentation accuracy.

The segmentation of the obsolete annotations is instead used by the warping function to filter out-of-date or wrong instances.

During training the ground truth of both the missing instances and the obsolete instances is known and binary cross-entropy losses are computed for these two segmentation branches.

The generator network G is used both for the alignment task and for the detection task, therefore it is trained using a linear combination of the alignment losses and the segmentation losses.

2.3. Network models

The convolutional neural network used as generator G is a recurrent residual U-Net [5] modified to produce three outputs: two segmentation masks and the similarity transformation map. The network we adopted is a simple but yet precise segmentation model which guarantees high building segmentation accuracy. The input image has 4 channels, since it is the concatenation of the intensity image I and the noisy annotation mask y. The outputs have values that ranges in $[0, 1]$ for the segmentation masks and in $[-1, 1]$ for the similarity transformation map since we use $\mathrm{sigmoid}$ and $\tanh$ activation functions, respectively.

The annotation instances are warped using a *Spatial Transformer Network* [6] that ensures to have differentiable warping operations and allows gradient flow during back-propagation. The warping function performs scale and rotation with respect to the barycenter of the selected annotation instance. It is noted that the generator G does not receive any information about the separation in instances and about the location of the barycenter of the buildings present in the input mask. The network, in fact, learns to identify building instances and understands the transformation rules during training.

The regularization network used to refine the segmentation is pre-trained and it is only used during inference.

3. EXPERIMENTAL SETUP

3.1. Dataset

The generator network G and the regularization model are trained in the Inria Aerial Image Labeling Dataset [7] composed of 180 images (5000×5000 px resolution) of 5 cities from US to Europe. Two of these images are used as test set. During training we consider the annotation masks provided in the dataset as ground truth even if some of these images contain misalignments.

3.2. Self-supervised training

The network must receive misaligned and incorrect annotations in order to learn. Since the dataset is assumed to be made of aligned image pairs some synthetic misalignments and errors must be introduced to alter the ground truth images. The noise is therefore enhanced by introducing global and instance random translations, rotations and scales. Random translations have a maximum absolute value of 64 pixels, while random rotations ranges between $-30°$ and $30°$. In order to create the ground truth for the segmentation branches some footprints have also been randomly removed and some others have been injected in the annotation masks.

Fig. 3: Alignment result in kitsap36. Synthetic misaligned annotations on the left. *MapRepair* prediction on the right.

Fig. 4: Alignment result in bloomington22. Noisy OSM annotations are overlaid in red. *MapRepair* prediction is in cyan. Removed annotations are yellow and segmented buildings are pink.

4. RESULTS

The method has been evaluated in two Inria images: kitsap36 and bloomington22. The two images have a resolution of 5000×5000 pixels and in order to evaluate the full image we split them into 448×448 patches. Each patch is individually processed by the network and a 64 pixels border is discarded due to lack of context information that can lead to

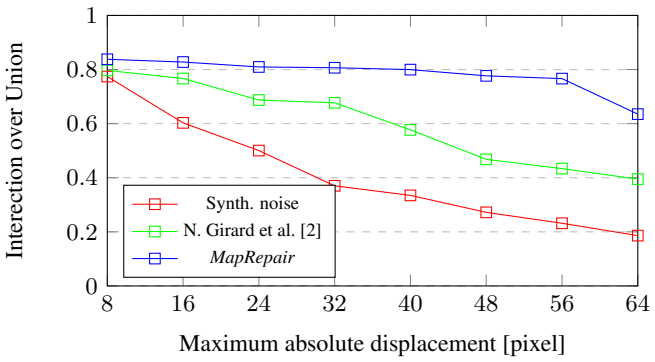

Fig. 5: IoU measured with the manually aligned annotations of kitsap36 from the Inria dataset. The plot shows the score of the synthetically misaligned annotations (red) and the score achieved after the correction (blue). The annotations are misaligned with gradually increasing random displacements and with random rotations and scales.

Table 1: Results in bloomington22 using OSM annotations.

	Alignment		Align & detect	
	IoU	Acc	IoU	Acc
Misaligned OSM	0.5235	0.9372	0.4739	0.9234
N. Girard et al. [2]	0.8302	0.9813	0.7369	0.9674
MapRepair align	0.8281	0.9812	0.7341	0.9673
MapRepair full	-	-	0.7914	0.9740

the generation of aligned annotations with errors and artifacts.

The kitsap36 image contains 1252 building instances having a wide range of shapes and sizes. The ground truth provided by the dataset contains several misalignments that are manually corrected in order to evaluate the algorithm prediction. In this image *MapRepair* correcs the original misaligned ground truth increasing the Intersection over Union (IoU) score from 0.71 to 0.82. Several experiments with synthetic misalignments are conducted in the same test image showing the robustness of the method to heavy annotation displacements. Building annotations are randomly rotated between $-30°$ and $30°$ and translated by increasing absolute displacements from 8 to 64 pixels.

The results in Figure 5 show that all the synthetic annotations aligned by *MapRepair* achieve IoU scores around 0.8. The best performance in reached with a maximum absolute displacement of 56 pixels where the network improves the IoU score from 0.23 to 0.77 (Figure 3). The efficiency starts dropping with an annotations misalignment of 64 pixels.

Bloomington22 is an image of the test-set of the Inria dataset, therefore the ground truth is not provided. For this region OSM provides 771 building footprints, most of them with severe misalignments. Furthermore, several construction do not have an OSM annotation. In order to measure the effectiveness of the correction we manually aligned the footprints and we annotated the unlabelled buildings. The quantitative and qualitative results in this image are shown in Table 1 and Figure 4, respectively.

5. CONCLUSIONS

We presented *MapRepair*, an approach for cadastre map refinement in satellite images composed of a multi-purpose neural network trained in a self-supervised manner. The model is capable of generating an aligned cadastre mask predicting a similarity transformation map and warping each object instance independently. Furthermore, it solves temporal synchronization errors removing unused footprints or segmenting new buildings in the scene. *MapRepair* achieves comparable or even higher alignment performance with respect to state-of-the-art methods, dealing effectively with heavily distorted annotations.

References

[1] N. Girard, G. Charpiat, and Y. Tarabalka, "Aligning and updating cadaster maps with aerial images by multi-task, multi-resolution deep learning," in *Asian Conference on Computer Vision*, Springer, 2018, pp. 675–690.

[2] ——, "Noisy supervision for correcting misaligned cadaster maps without perfect ground truth data," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2019, pp. 10 103–10 106.

[3] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Autocorrect: Deep inductive alignment of noisy geometric annotations," *ArXiv preprint arXiv:1908.05263*, 2019.

[4] S. Zorzi and F. Fraundorfer, "Regularization of building boundaries in satellite images using adversarial and regularized losses," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5140–5143.

[5] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *ArXiv preprint arXiv:1802.06955*, 2018.

[6] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[7] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2017.