Effective Building Extraction by Learning to Detect and Correct Erroneous Labels in Segmentation Mask

Praveer Singh, Nikos Komodakis

EFFECTIVE BUILDING EXTRACTION BY LEARNING TO DETECT AND CORRECT ERRONEOUS LABELS IN SEGMENTATION MASK

Praveer Singh and Nikos Komodakis

École des Ponts ParisTech & Université Paris Est, France
Praveer.Singh@enpc.fr

ABSTRACT

Semantic segmentation is pivotal for remote sensing image analysis. Although existing segmentation techniques perform well on similar landscape images, their generalization capability on an entirely different landscape is extremely poor. One of the primary reasons is that they partially or wholly, neglect the underlying relationship that exist in the joint space of input and output variables. Thus, effectively they lack to impose structure in their output predictions which is necessary for successful segmentation. In this paper, we address this problem and propose a novel solution by modeling the joint distribution of input-output variable which in turn enforces some structure in the initial segmentation mask. To this end, we first detect erroneous labels, in the form of *Error maps*, in the initial building masks. These Error maps are then used to correct the corresponding erroneous labels through a replacement technique. We evaluate our methodology on the benchmark *Inria Aerial Image Labeling* dataset, which is a large scale high resolution dataset for building footprint segmentation. In contrast to previous methods, our predicted segmentation masks are much closer to ground truth, owning to the fact that they are able to effectively correct both the large errors as well as the *blobby* effects. We lastly perform on par with other state-of-the-arts, validating the efficacy of our technique.

***Index Terms*—** Deep learning, high-resolution imagery, semantic segmentation, structured prediction, building footprint extraction.

1. INTRODUCTION

Recently, we have witnessed an explosion of petabytes of high-resolution remote sensing datasets such as Sentinel 1-5 [2], SpaceNet [3] and Inria Aerial Image Labeling dataset [1]. Till now, these datasets have been manually annotated by experts. However, with the availability of such an enormous amount of high-resolution data, it is truly herculean to label all of them by hand. This necessitates the automatic segmentation of these remote sensing images to quickly and effectively detect varied points of interest such as roads, buildings, forests in an image for tasks such as urban scene planning, green cover monitory and other emergency relief operations such as floods, forest fires and cyclones. Deep learning methods have recently shown significant improvements on automatic semantic labeling tasks for classical datasets such as Vaihingen or Potsdam. [4, 5] fused semantic maps from multiple sources through a residual correction technique. Whereas, [6] corrected shifts in OSM maps through an iterative refinement technique using Recurrent Neural networks. In addition, [7] used boundary detections to improve the semantic segmentation and report impressive performance on Vaihingen dataset.

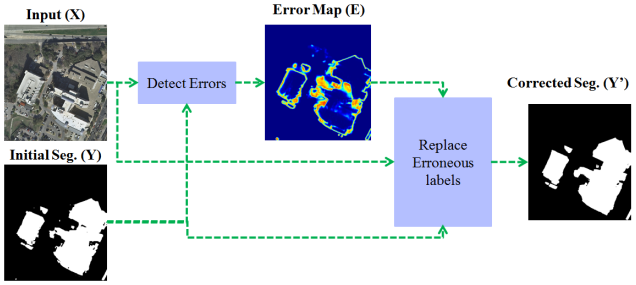

Fig. 1: The Network Architecture.

Interestingly, [1] showcased the drawback of models trained over such classical datasets, by highlighting their lack of generalization capability to other cities that are captured under different conditions. They, henceforth build a new dataset covering a much larger surface of the earth including both densely populated urban landscape as well as sparse alpine regions under varied illumination conditions. Subsequently, they split their train and test data such that they come from different cities. Finally, they also addressed problem of *blobby-predictions i.e.*, curvy edges of building masks. By fusing feature maps from different levels of convolution network, both [1] and [8] combine low level edge information with high level semantic rich information. For the same task, [9] uses a multi-task loss in order to approximate the distance transform and the semantic maps.

Lately, most of the aforementioned techniques have partially or fully neglected the idea of enforcing structure in the output label space. This mainly results in these blobby ef-

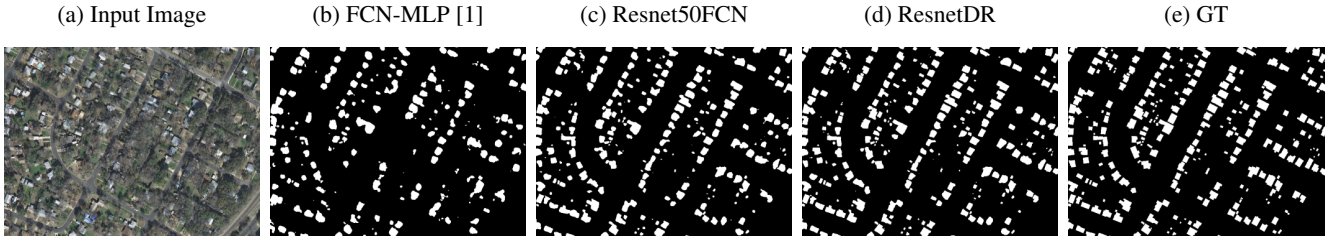

Fig. 2: Segmentation masks for different techniques along with the final ground truth (GT). WHile Resnet50FCN drastically improves over FCN-MLP [1], our method (ResnetDR) further corrects the fuzzy or "blobby" effects of Resnet50FCN (caused due to naive up-sampling) by simply enforcing the underlying structure of building footprint shapes.

fects due to naive up-sampling during the segmentation masks prediction. On the contrary, a few of those that do enforce structure, either do it through a novice human assumption about the structure of the output label space (in the form of hand engineered CRF pairwise potentials as in [6]). Or they rely on semantic maps from other sources to refine the initial predicted labels [5, 4]. However, these refinements through residual correction can only correct small errors (such as on the boundaries) while leaving major segmentation errors intact.

Built on these observations, we argue that we need to learn a joint structure of input and output variables that can effectively predict the replacements for major segmentation errors and thus, rectify them. Inspired from [10], we henceforth, propose a joint input-output model which segments the building footprints in high resolution imagery. While [10] addresses a continuous variable assignment task (disparity estimation), we solve here a binary classification problem. Our proposed method is buildup of two steps. Firstly, our model predicts an error map E based upon the input image X and initial segmentation Y. We then update the labels in regions of high error probabilities with a new label prediction which in turn relies on X, Y and E for its decision. Through the error map E, we learn a joint distribution between input and output variables which further helps in enforcing structure in the final label prediction. As shown in Figure 2, our model refines the blobby effect by learning the underlying structure of the building footprints which is enforced using the error maps. Finally, this leads our method to perform on par with existing state-of-the-arts on Inria Image Labeling dataset [1].

2. PROPOSED FRAMEWORK AND METHODOLOGY

Let's assume our initial input image to be $X = x_{i=1}^{C\times H\times W}$, where C, H and W are the channel, height and width of X respectively. Similarly, $Y = y_{i=1}^{1\times H\times W}$, be the initial segmentation map. Our technique aims to model a joint relationship between input X and output variables (Y) to rectify and produce a much more refined version of Y. This can be formulated as $Y' = F(X, Y)$ where Y' denote the updated segmentation mask after replacing erroneous labels with new labels.

As shown in Figure 1, our proposed methodology comprises of two major steps. First, we predict errors (E) occurring in the initial building segmentation mask Y with the help of input image X. Next, we utilize these error maps to correct those erroneous labels in Y, whose error probability is large enough (in E) by replacing them with predicted labels. In the following subsections, we explain these two steps in detail.

2.1. Error Detection

The error detection component (F_e) computes the probability map (E) to detect the erroneous labels in the initial segmentation mask Y, stated as:

$$E = F_e(X, Y). \tag{1}$$

In other words, F_e learns from the joint input output space of X & Y to predict an error probability score map where each pixel has a value between 0 and 1. Specifically, it predicts whether or not a particular label y_i is erroneous, if so, y_i gets replaced with a correct label in the next step. F_e can be easily formulated as a deep neural network which requires as such no explicit auxiliary loss and can be learnt under a single umbrella of one loss between corrected segmentation and ground truth segmentation.

2.2. Erroneous Label Replacement

The updated label Y' is a convex combination of the initial segmentation mask Y and updates from the replacement component denoted by F_r. It is given as:

$$Y' = E \odot F_r(X, Y, E) + (1 - E) \odot Y, \tag{2}$$

where $\odot$ represents element-wise product. The error map E generated from F_e acts as a gateway to restrict F_r so as to just focus on the erroneous labels of Y and replace them with the predicted ones. Similar to F_e, F_r can also be modeled using any deep learning architecture. If we restrict the E probability

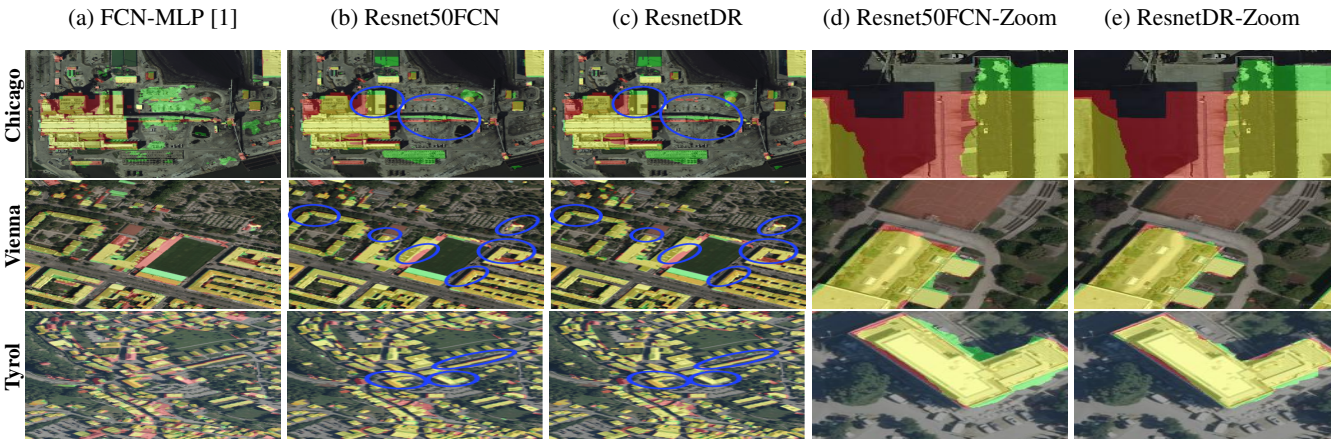

Fig. 3: *Qualitative Results*. Red segments are ground truth, green are predictions while yellow represents overlap of ground truth and predictions. The blue circles highlight regions with significant changes. We also showcase zoom-in for one of these regions as comparison.

maps to 0 and 1, the forward pass of Replacement happens as:

$$Y' = \begin{cases} Y, & \text{if } F_e(X,Y) = 0 \\ F_r(X,Y,E), & \text{if } F_e(X,Y) = 1. \end{cases} \quad (3)$$

This shows that only erroneous labels are being replaced while non-erroneous labels remain intact. For any end loss function L between corrected segmentation mask Y' and ground truth labels, its gradient becomes:

$$\frac{\mathrm{d}L}{\mathrm{d}F_r(.)} = \begin{cases} 0, & \text{if } F_e(X,Y) = 0 \\ \frac{\mathrm{d}L}{\mathrm{d}Y'}, & \text{if } F_e(X,Y) = 1. \end{cases} \quad (4)$$

In a way, the gradients update F_r only for those regions where erroneous labels are found in Y, thus restricting F_r to pay attention and predict replacements only for these particular regions. Additionally, passing Error maps to the F_r component, helps it to rely on the correct labels to predict replacements for the new erroneous labels. Altogether, F_r improves and makes correction of these erroneous labels, by jointly reflecting upon the Error maps, the input X and the initial segmentation Y.

3. TRAINING AND IMPLEMENTATION DETAILS

3.1. Dataset

We evaluated our method on the Inria Aerial Image Labeling Dataset [1] which consists of Airborne imagery of urban settlement over the United States and Austria. The entire dataset consists of two classes namely, building and not building. All the images are of size 5000 $\times$ 5000 and have a resolution of 30 cm with RGB bands. First 5 images from each class were chosen for validation, while the rest were used for training. For testing, we upload our test results to the project webpage.

3.2. Network Architecture

We initially train a Fully Convolutional Network [11] (FCN) adapted to Resnet-50 [12] architecture to generate our initial segmentation mask Y. This model, which we name as Resnet50FCN, is trained to reconstruct ground truth segmentation masks with a given input image X and ground truth labels. We henceforth treat Resnet50FCN as our *baseline*.

For detection, F_e is implemented by using 5 convolutional layers (except last one, each is followed by batch-norm and Relu). The last conv. layer is followed by a soft-max, thus yielding us E between 0 and 1. To follow the input image size, we add an up-sampling layer on top of the Error map E.

For replacement, F_r is implemented through compression block (compresses to 1/64 of the input resolution) and decompression block (decompresses to 1/4 of input resolution). These are essentially residual blocks with parameterized skip connection between symmetric layers in decompression and compression blocks respectively.

For additional implementation details of detection and replacement modules, we refer the reader to Section 3.2 of [10].

3.3. Training

Using an L1 loss between ground truth and predicted output Y', we optimize using an adam solver [13], with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate starting from 10^{-3} is decreases to 3×10^{-4} at 12, 10^{-4} at 18, 3×10^{-5} at 24 and finally 10^5 at 28 epochs. We continue training until 32 epochs. Each epoch consists of 500 batch iterations and each batch consists of 16 training samples where each sample is of size 1024 $\times$ 1024.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1. Quantitative Results

We show the quantitative results in Table 1 and Table 2 on both the validation and test dataset respectively, where we

Method		Austin	Chicago	Kitsap Co.	West Tyrol	Vienna	Overall
FCN-MLP [1]	IoU	61.20	61.30	51.50	57.95	72.13	64.67
	Acc.	94.20	93.43	98.92	96.66	91.87	94.42
Multi-Task [9]	IOU	76.76	67.06	73.30	66.91	76.68	73.00
	Acc.	93.21	99.25	97.84	91.71	96.61	95.73
Resnet50FCN	IoU	88.12	81.21	83.62	88.15	87.07	86.46
	Acc.	97.02	93.04	99.3	98.28	94.65	96.46
ResnetDR	IOU	89.42	83.56	84.57	89.07	88.30	**87.90**
	Acc.	97.37	94.03	99.34	98.42	95.12	**96.87**

Table 1: Evaluation results on validation set

Method		Bellin.	Blooming.	Inns.	S.Fo	Tyrol-E	Overall
FCN-MLP [1]	IoU	56.11	50.40	61.03	61.38	62.51	59.31
	Acc.	95.37	95.27	95.37	87.00	96.61	93.93
Resnet50FCN	IoU	63.34	63.20	76.07	74.91	77.58	72.07
	Acc.	95.90	96.61	97.18	91.67	98.01	95.78
ResnetDR	IOU	64.27	65.85	77.10	75.86	78.68	**73.30**
	Acc.	95.99	96.also52	97.30	92.01	98.11	**95.99**

Table 2: Evaluation results on test set.

compare our method (ResnetDR) with the baseline (Resnet 50FCN) and other previous best performing results namely, MLP [1] and Multi-Task Loss [9]. In both tables, we report the Intersection over Union (IoU) and Accuracy score (Acc.) of correct pixels in the segmentation mask [1]. As shown in Table 1, we outperformed the previous best method [9] on the validation set by a margin of 14.90% on IoU. While on our own baseline Resnet50FCN, we improve by a margin of 1.44%. On the test set too, we outperformed [1] by 14% and our own baseline by 1.22%.

4.2. Qualitative Results

For qualitative analysis, we report our results in Figure 3 where green patch represents predictions for each model, red the ground truth while yellow represents the overlay of ground truth and predictions. We observe that while a major improvement is seen from FCN-MLP to Resnet50FCN, it still is not able to perfectly correct the *blobby* effects. ResnetDR improves upon these blobby effects by not only refining the boundaries of the segmentations but also regularizing them to better reflect structure of building footprints. For *e.g.*, in Figure 3, in the Chicago case, we note that for the refinery (zoom-in shown in d & e), ResnetDR yields a more structured output in the form of parallel prediction edges. Similarly, for Vienna and Tyrol, predictions shrinking-in / protruding-out from the roofs of the houses for Resnet50FCN are constrained in case of ResnetDR to follow the edges of the roof.

All these cases prove that our technique learns the underlying structure of building footprints in the form of regularized and well-defined shapes with straight edges. Simultaneously, it learns to predict error maps that somewhat refine these initial segmentation masks in order to enforce strict rules that govern building footprint structures.

[1]Due to the unavailability of the results on test-set by [9], we haven't reported them in Table 2

5. CONCLUSIONS

We present a novel technique for structured semantic segmentation of high resolution satellite imagery. To this end, we propose to learn the joint space of input-output variables. Subsequently, our method enforces structure in the form of predicting a much more regularized building footprint and hence, resolves to a large extent the problem of blobby effects as reported in the past methods. We compare our technique with the state-of-the-art methods on the Inria aerial image labeling dataset, where we perform on par with others.

6. REFERENCES

[1] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre 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.

[2] European Space Agency (ESA), "Sentinel," `https://sentinel.esa.int/web/sentinel/home`, 2014.

[3] Digital Globe CosmiQWorks and NVIDIA, "Spacenet," `http://explore.digitalglobe.com/spacenet`, 2016.

[4] N. Audebert, B. Le Saux, and S. Lefvrey, "Fusion of heterogeneous data in convolutional networks for urban semantic labeling," in *2017 Joint Urban Remote Sensing Event (JURSE)*, March 2017, pp. 1–4.

[5] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre, "Joint Learning from Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps," in *EARTHVISION 2017 IEEE/ISPRS CVPR Workshop. Large Scale Computer Vision for Remote Sensing Imagery*, Honolulu, United States, July 2017.

[6] E. Maggiori, G. Charpiat, Y. Tarabalka, and P. Alliez, "Recurrent neural networks to correct satellite image classification maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 4962–4971, Sept 2017.

[7] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," vol. 135, 11 2016.

[8] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7092–7103, Dec 2017.

[9] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," *arXiv preprint arXiv:1709.05932*, 2017.

[10] S. Gidaris and N. Komodakis, "Detect, replace, refine: Deep structured prediction for pixel wise labeling," in *CVPR*, 2017.

[11] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, April 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.