WEAKLY SUPERVISED BUILDING SEGMENTATION FROM AERIAL IMAGES

Muhammad Usman Rafique, Nathan Jacobs

University of Kentucky, Lexington, Kentucky
{usman.rafique, nathan.jacobs}@uky.edu

ABSTRACT

We propose a novel framework for weakly supervised semantic segmentation from aerial images. Instead of requiring labels for every pixel, our method only requires a bounding box for each building and leverages domain information to translate these into pixel-level predictions. We convert the bounding boxes into probabilistic masks, each represented using a bivariate Gaussian distribution. We propose a loss function that encompasses our domain knowledge that the bounding box is an upper bound for the object it contains. Combining these two elements significantly improves over many baseline methods. We show extensive results on a recent, large-scale dataset prepared by the United Nations Global Pulse and compare with several baselines.

Index Terms— Semantic segmentation, building detection, weakly supervised learning

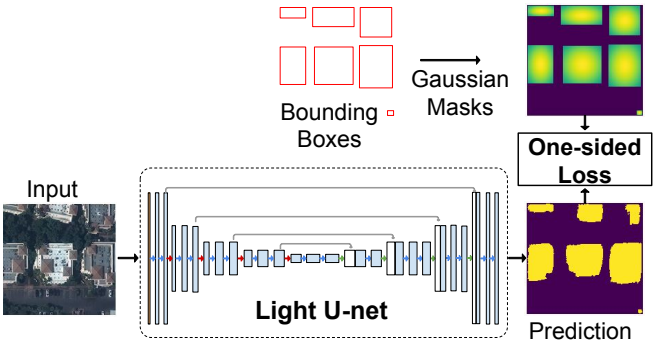


Fig. 1: Overview of the proposed approach. We propose a novel method of generating Gaussian masks from horizontal bounding boxes to train the weakly supervised segmentation network. Our proposed one-sided loss function leverages the domain knowledge.

1. INTRODUCTION

Overhead images are collected at an astonishingly high frequency by many organizations. This relatively new image source is a convenient tool for several applications, ranging from environmental monitoring [1], detecting marine animals [2], identification of vegetation [3], land cover classification [4], and forecasting commercial activity [5]. With the capability to capture large areas quickly, overhead images are an ideal source for disaster response. While recent advances in machine learning have led to rapidly improving image understanding systems, the process of manually annotating images for training of convolutional neural networks (CNN) is slow. We aim to address this issue by requiring low fewer annotations. Even though pixel-wise building segmentation networks provide rich information, they require complex, pixel-wise annotation, usually as polygons [6]. On the other hand, it is much easier to mark horizontal bounding boxes for objects of interest and to model the problem as object detection. But output of object detection (bounding box) is not as rich as the dense pixel-wise prediction of segmentation networks. In this work, we aim to bridge this gap by presenting a method to train pixel-wise segmentation network by utilizing only bounding box annotations.

A conventional method of dealing with limited annotations is to use fine-tuning [7], a form of transfer learning. Transfer learning has been used in overhead images. For example, Yan *et al.* [8] used fine-tuning to identify ice in the arctic ocean. However, there are three main drawbacks of fine-tuning: 1) as explained by Yosinski *et al.* [7], retraining a part of the network is not trivial and several factors including data and classes impact the final performance, 2) fine-tuning a pre-trained model requires *some* fully labeled data and hence increases time-to-deploy, and 3) recently, He *et al.* [9] have shown through extensive experiments that fine-tuning primarily speeds up training initially and does not significantly affect the final performance. Keeping this in mind, we propose an end-to-end weakly supervised training method, which requires less amount of label annotation.

For natural images, there are several weakly supervised methods for training per-pixel estimation with only bounding region supervision. Dai *et al.* [10] proposed a method to iteratively clean the bounding boxes to get better segmentation masks for training. Khoreva *et al.* [11] introduced a method to prepare better masks from bounding boxes using multiple iterations of several algorithms, such as MCG [12] and Grabcut [13]. Typically, weakly supervised approaches use multiple passes for every bounding box to *clean* the data

and solve a major problem of overlapping objects. In case of aerial images, this problem is virtually nonexistent: there is no overlap of bounding boxes in typical scenarios. For example, land cover segmentation has exclusively one label per region (in contrast to the possibility of overlapping image regions for different objects at different distances). Based on this fact, we are able to propose a simple training method that does not require iterative cleaning of bounding boxes.

In this paper, we propose a novel method of generating probabilistic masks as well as a novel loss function that allow us to predict dense, pixel-wise predictions while requiring only bounding boxes for training. We show our results on the recently released, large-scale dataset of overhead images [14] by *humanity and inclusion* and UN Global Pulse for disaster response. Our main contributions are 1) proposed probabilistic masks using bivariate Gaussian distribution, and 2) a novel one-sided loss function that leverages domain knowledge, and 3) our results on recently released large-scale, real-world dataset that include several baselines and a surprising finding that a naïve baseline performs reasonably well.

2. OUR APPROACH

2.1. Probabilistic Masks

We use bounding boxes to generate a pixel-wise dense mask for every image. As shown in Section 4, generating binary masks from bounding boxes is not the optimal solution. Instead, we use a bivariate Gaussian distribution to model probabilistic masks. This is motivated by the fact that pixels near the center of a bounding box are more likely to be a building than those around the edges due to possible misalignment of building and the horizontal bounding box. Since we know that every box contains a building, and we are more certain of a pixel near the center being a building than pixels around the edges, we set the mean of the Gaussian distribution at the center of the bounding box. For a bounding box with dimensions $w \times h$, we use the following Gaussian distribution to sample value for all pixels within the bounding box:

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right] \tag{1}$$

where (x, y) are the pixel coordinates, (μ_x, μ_y) is the mean of the distribution. We treat x and y as independent and hence correlation is zero. The covariance matrix is $\Sigma = \text{diag}(\sigma_x^2, \sigma_y^2)$. We set center of the bounding box as mean of the distribution (μ_x, μ_y). We propose to represent $\sigma_x^2 =$ and σ_y^2 in the form

$$\sigma_x^2 = \frac{w^2}{s_G}, \quad \sigma_y^2 = \frac{h^2}{s_G}, \tag{2}$$

with a hyper-parameter scaling factor s_G. Because of parameterization in equation (2), mask values of points at edges are independent of the bounding box dimensions. Finally, we normalize the Gaussian mask by $p(x,y) \leftarrow p(x,y)/max(p(x,y))$. The normalized form has maximum value of one whereas the original bivariate Gaussian sums to 1 (giving much smaller values to individual pixels).

2.2. One-sided Loss Function

We propose a loss function that represents the knowledge of the building within each bounding box. While we know that building area is less than or equal to the *bounding* box area, we don't know the exact fraction of the horizontal bounding box that is occupied by the building. So, a loss function that tries the predicted building area to be *equal to* a fraction, say 0.8, of the bounding box is incomplete: it is possible for some cases that area of bounding box is very close to the building area (in case when a building occupies most of the box). To leverage this knowledge, we propose the following one-sided loss function

$$\mathcal{L}(O, O_{GT}) = k_1\delta(FP) + k_2\delta(FN) + \mathcal{L}_{\mathcal{A}} \tag{3}$$

$$\mathcal{L}_{\mathcal{A}} = k_3 \cdot \frac{max(s_A A_{GT} - A, 0)}{A_{GT}} \cdot (FN) \tag{4}$$

where O and O_{GT} are the network output and the ground truth (Gaussian) masks on the interval $[0, 1]$. $A = \sum O$ and $A_{GT} = \sum O_{GT}$ are the areas of output and ground-truth masks. s_A is a hyper-parameter within the interval $(0, 1]$. $FP = O \cdot (1 - O_{GT})$ and $FN = (1 - O) \cdot O_{GT}$ are false positive and false negative predictions, respectively, and k_1, k_2, and k_3 are hyper-parameters which control the scale of different loss terms. $\delta(x)$ can be any loss function, we use mean squared error with respect to $\mathbf{0}$: $\delta(x) = ||x - \mathbf{0}||^2$. False positives and false negatives, with respect to the bounding boxes are penalized, as shown in equation (3). Equation (4) states that if predicted area is greater than s_A (say 0.8) of the ground-truth area, there is no penalty. Hence, a penalty is applied only if the output area is less than a particular fraction (s_A) of the ground-truth area. Further, the area penalty $\mathrm{L}_A is normalized by area of bounding box and applied to false negatives o$

3. EXPERIMENTAL SETUP

We conducted all experiments using a variant of U-Net [15] with half the feature maps as compared to the original U-Net. We have released our code [1].

3.1. Dataset and Evaluation Metric

We use the recently introduced mapping challenge dataset [14]. The dataset contains 280 741 training and 60 317 validation

[1] http://github.com/UkyVision/weakly-supervised-segmentation

RGB images of size 300×300. The original dataset has been prepared for instance segmentation of buildings from satellite images. For proof of concept of our proposed approach, we use horizontal rectangular bounding boxes for training. To evaluate with ground-truth labels, we convert all instances annotations to prepare a dense, binary mask, as typically used for evaluation of semantic segmentation networks. For quantitative results, we use Jaccard index, also known as intersection over union (IoU) of the building pixels, a common metric for segmentation tasks.

3.2. Implementation Details

We implemented the proposed method in Keras. We used Adam optimizer [16] with learning rate of $5e^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The network was trained from scratch for 52 638 iterations (3 epochs) with batch size 16.

For scaling of Gaussian masks, we use scale factor $s_G = 2.5$ - this gives normalized $p(x, y)$ of 0.75 at midpoints of all edges of the bounding box: $(\mu_x \pm w/2, 0)$ and $(0, \mu_y \pm h/2)$. We use $k_1 = 1$, $k_2 = 0.8$, and $k_3 = 0.3$. For area scaling, we use $s_A = 0.8$, implying that there is no penalty if predicted area within a box is more than 80% of the bounding box. We used non-exhaustive grid search to estimate these hyper-parameters and we find that the final outcome is not very sensitive to these, except higher values of s_A which lead to degrading performance.In a nutshell, we have lower penalty for false negatives ($k_2 = 0.8$) than false positive ($k_1 = 1$) but we add extra penalty on false negatives ($k_3 = 0.3$) if the predicted area is less than ground-truth area.

3.3. Baseline Methods

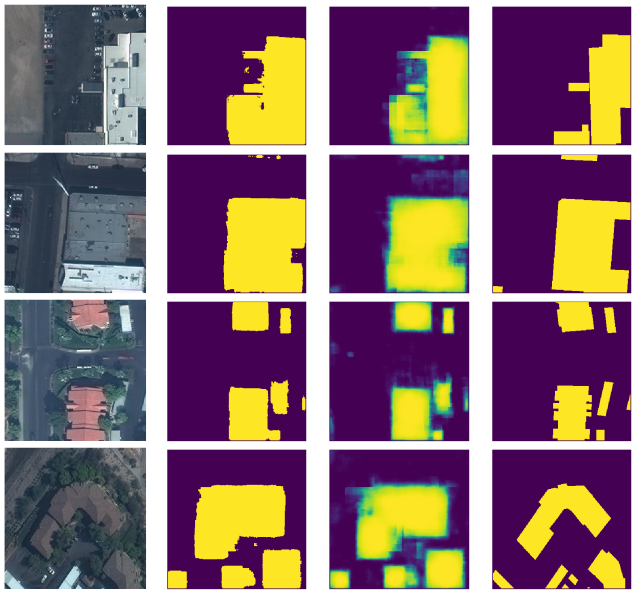
We use a *naïve* baseline in which we convert bounding boxes to binary masks and trained with cross-entropy loss. Secondly, we consider *naïve + one-sided* in which we use binary masks but we use the proposed one-sided loss function in equation (3). Motivated by the recent work [11], we also use Grabcut [13] in several settings. First, we used *oracle + grabcut* in which the bounding boxes are provided at the test time and the unsupervised background subtraction method is used to segment the region within bounding boxes. We also use a *oracle + grabcut2* in which it is indicated that inner 30% is building and the grabcut algorithm only needs to segment the remaining region of the bounding box. As a reference, we provide results of *fully supervised* segmentation as well.

4. RESULTS

Quantitative results, shown in Table 1, reveal several interesting observations. First, the *oracle + grabcut* methods perform much worse, even though these methods know the bounding boxes even at test time. *oracle + grabcut* and *oracle + grabcut2* get IoU of 34.8% and 60.7% respectively. Second, the *naïve* baseline performs reasonably well by getting 70.25% IoU. Usefulness of our proposed one-sided loss function is highlighted by superior performance of *naïve + one-sided* method that achieves 72.54% IoU. Finally, we show that *ours-full* gets the best IoU of 74.34% among all the weakly supervised methods.

Method	Supervision	Loss	IoU(%)
Supervised	Full Masks	Cross-entropy	79.27
Oracle + grabcut	Bounding box	-	34.8
Oracle + grabcut2	Bounding box	-	60.7
Naïve	Bounding box	Cross-entropy	70.25
Naïve + one-sided	Bounding box	Proposed (3)	72.54
Ours full	Gaussian masks	Proposed (3)	74.34

Table 1: Quantitative results.



(a) (b) (c) (d)

Fig. 2: Qualitative results. We show input images (a), binarized output masks using threshold of 0.5 (b), raw output masks (c), and the ground-truth segmentation masks (d).

Qualitative results are shown in Figure 2. Currently, the limitation of our method is that prediction is sometimes bigger than true mask, as shown in last row of Figure 2. The prediction can be bigger than true masks because the network is trained on bounding boxes which are often bigger than buildings. However, we can see that around the edges, prediction have lower values (in greener shade) highlighting

the semantic understanding of the network as well as leaving room for further improvement by marking such regions with low scores.

5. CONCLUSION

We presented weakly supervised method for building segmentation. We show that representing uncertainty of objects within bounding boxes through a Gaussian probabilistic masks gives better results. The proposed custom loss function boosts performance by leveraging domain knowledge. Because objects are non-overlapping in aerial images, we show that our simple method gives good results without requiring multiple iterations over each image. A limitation of our work is that the output masks are sometimes bigger than true segmentation masks. There are several areas for future work: extending to the multi-class case and reducing the over-segmentation problem.

Acknowledgements

We gratefully acknowledge the support of NSF CAREER (IIS-1553116).

6. REFERENCES

[1] Ni-Bin Chang, Kaixu Bai, Sanaz Imen, Chi-Farn Chen, and Wei Gao, "Multisensor satellite image fusion and networking for all-weather environmental monitoring," *IEEE Systems Journal*, vol. 12, no. 2, pp. 1341–1357, 2018.

[2] Ana S Aniceto, Martin Biuw, Ulf Lindstrøm, Stian A Solbø, Fredrik Broms, and JoLynn Carroll, "Monitoring marine mammals using unmanned aerial vehicles: quantifying detection certainty," *Ecosphere*, vol. 9, no. 3, 2018.

[3] Temuulen T Sankey, Jason McVay, Tyson L Swetnam, Mitchel P McClaran, Philip Heilman, and Mary Nichols, "Uav hyperspectral and lidar data and their fusion for arid and semi-arid land vegetation monitoring," *Remote Sensing in Ecology and Conservation*, vol. 4, no. 1, pp. 20–33, 2018.

[4] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[5] Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang, "Urban perception of commercial activeness from satellite images and streetscapes," in *Companion Proceedings of the The Web Conference*, 2018, pp. 647–654.

[6] Guangming Wu, Xiaowei Shao, Zhiling Guo, Qi Chen, Wei Yuan, Xiaodan Shi, Yongwei Xu, and Ryosuke Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sensing*, vol. 10, no. 3, pp. 407, 2018.

[7] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[8] Yiming Yan, Zhichao Tan, and Nan Su, "Sea-ice scene classification using aerial images in arctic based on transfer learning," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018.

[9] Kaiming He, Ross Girshick, and Piotr Dollár, "Rethinking imagenet pre-training," *arXiv preprint arXiv:1811.08883*, 2018.

[10] Jifeng Dai, Kaiming He, and Jian Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2015.

[11] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2.

[12] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 128–140, 2017.

[13] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*. ACM, 2004, vol. 23, pp. 309–314.

[14] "crowdai mapping challenge," www.crowdai.org/challenges/mapping-challenge, 2018.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.