IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883979

FEATURE AND OUTPUT CONSISTENCY TRAINING FOR SEMI-SUPERVISED BUILDING FOOTPRINT GENERATION

Qingyu Li [1,2], *Yilei Shi* [3], *Xiao Xiang Zhu* [1,2]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
[3] Remote Sensing Technology, Technical University of Munich (TUM), Munich, Germany

ABSTRACT

Building footprint maps are important to urban planning and monitoring. However, most existing approaches that fall back on convolutional neural networks (CNNs), require massive annotated samples for network learning. In this research, we propose a novel semi-supervised network, which can help to deal with this issue by leveraging a large amount of unlabeled data. Considering that rich information is also encoded in feature maps, we propose to integrate the consistency of both features and outputs in the end-to-end network training of unlabeled samples on data perturbation, enabling to impose additional constraints. Experiments are conducted on Inria dataset. Our approach is much superior to the state-of-the-art methods in both quantitative and qualitative results.

Index Terms— semantic segmentation, semi-supervised, building, consistency training

1. INTRODUCTION

Building footprint generation is of great interest in the remote sensing community and involves a wide range of applications, e.g., disaster management and urban planning. High-resolution remote sensing imagery, which provides huge opportunities for meaningful geospatial target extraction at a large scale, becomes a fundamental data source for building footprint generation. Early methods focus on the design of hand-crafted features that can best depict buildings. Nonetheless, the empirical feature design is satisfactory only under specific requirements or on specific data and lacks good generalization capabilities. Nowadays, Convolutional Neural Networks (CNNs) have been widely used for the task of building footprint generation from remote sensing imagery [1] [2] [3], as they surpass conventional methods in terms of accuracy of efficiency. CNNs can directly learn hierarchical contextual features from the raw input and offer greater generalization capabilities. However, there remains a challenge for generating building footprint maps on a large scale — massive data need to be collected to promote the generalization performance of CNNs. However, the manual annotation of reference data is a very time-consuming and costly process.

Recently, several methodologies have taken advantage of semi-supervised learning to address this issue. Among them, consistency training-based approaches (e.g., CR [4] and PiCoCo [5]) not only are simple to implement but also require no additional weakly labeled examples. Consistency training-based methods exploit the teacher-student framework and encourage both the student model and teacher model to give consistent outputs for unlabeled inputs that are perturbed in various ways. By doing so, the generalization capability of the network can be improved. However, there is still a certain gap in performance between these two models when the outputs are not completely correct during training. Inspired by [6] that more discriminative contextual information can be captured by feature maps, we propose a new consistency loss that measures the discrepancy between both feature maps and outputs of student model and those of teacher model, offering a strong constraint to regularize the learning of the network.

2. METHODOLOGY

2.1. Overview

As shown in Fig. 1, the proposed framework is composed of a shared encoder E, a main decoder D, and an auxiliary decoder G. The segmentation network F is constituted as $F = E \circ D$ and is trained on the labeled set in a fully supervised manner. The auxiliary network $A = E \circ G$ is trained on the unlabeled examples by enforcing the consistency of both features and outputs between D and G. D takes as input the encoder's output $\mathbf{z}_{out}$, but G is fed with its perturbed version $\tilde{\mathbf{z}}_{out}$, in which the perturbation p is applied to the output of E. By doing so, the representation learning of E can be further improved by unlabeled examples, and subsequently, that of the segmentation network F.

2.2. Objective Function

In the proposed approach, labeled and unlabeled data are jointly trained by minimizing a global loss function L as

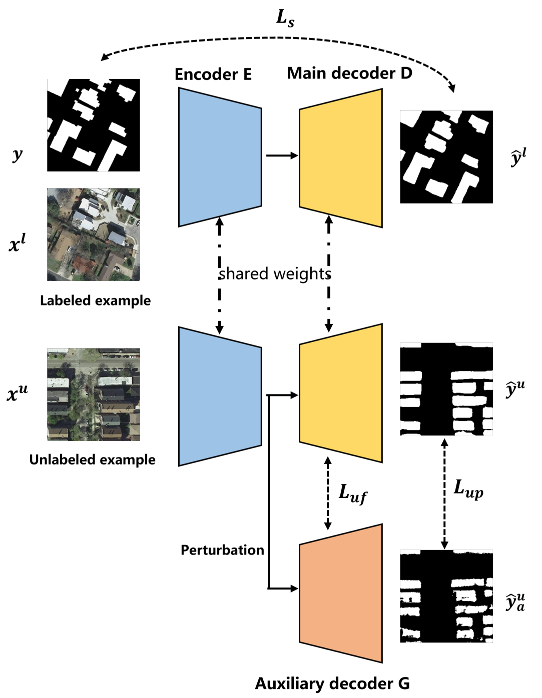

Fig. 1. Overview of the proposed approach.

below:

$$L = L_s + \lambda_u \cdot L_{cons} \; , \tag{1}$$

where L_s is a supervised loss on labeled data. λ_u is a weighting function to control the importance of a consistency loss term L_{cons}. Note that L_{cons} is not backpropagated through D, and D is trained only by labeled examples. By doing so, D is only trained on original input data. This is helpful from two aspects. On the one hand, it can avoid collapsing solutions. If L_{cons} is backpropagated through both main decoder D and auxiliary decoder G, main decoder D will collapse since L_{cons} will be minimized if predictions of both D and G are zeros. On the other hand, the method can be better adapted to the test stage since no perturbation is applied to test images.

For the labeled set, a supervised loss L_s is exploited to train the segmentation network F. In order to avoid overfitting, an annealed version of the bootstrapped Cross-Entropy loss [7] is chosen to compute the supervised loss L_s, and it is denoted as:

$$L_s = \frac{1}{|S_l|} \sum_{x_i^l, y_i \in S_l} \{F(x_i^l) < \eta\} \mathbf{H}(y_i, F(x_i)) \; , \tag{2}$$

where $F(x_i)$ is the output probability from F for a labeled example x_i, y_i is its ground reference label, and $\mathbf{H}(.,.)$ is the cross entropy-based loss. In order to avoid overfitting, L_s is computed only over the pixels with a probability less than the threshold η that serves as a ceiling to prevent over-training on easily labeled data [8]. Following [7], we gradually increase η from 0.5 to 0.9 during the beginning of training.

For an unlabeled example x_i^u, $\mathbf{z}_{out}$ is derived as the output from the shared encoder E. Afterward, the perturbation is applied to the output of the encoder E and the perturbed encoder's output $\tilde{\mathbf{z}}_{out}$ is generated. Finally, $\mathbf{z}_{out}$ and $\tilde{\mathbf{z}}_{out}$ are taken as input for D and G, respectively.

The training objective of the unlabeled set is to minimize a consistency loss L_{cons}, which is defined as:

$$L_{cons} = L_{up} + \omega_u \cdot L_{uf} \; , \tag{3}$$

where L_{uf} and L_{up} measure the discrepancy between the features and outputs of D and those of G, respectively. ω_u is a hyperparameter to introduce a weight to model the relative importance of two losses. More specifically, L_{up} is defined as:

$$L_{up} = \frac{1}{|S_u|} \sum_{x_i^u \in S_u} \mathbf{T}(D(\mathbf{z}_{out}), G(\tilde{\mathbf{z}}_{out})) \; , \tag{4}$$

with $\mathbf{T}(.,.)$ as mean squared error-based loss.

Note that the contribution of our approach is that a loss term L_{uf} is introduced into the proposed network by imposing the consistency on features between the main decoder and auxiliary decoder, which is able to harness the detailed information in the feature maps. Let $\phi_j(q)$ be the activations of the jth layer of the network ϕ when processing the input q. For D and G, $D_j(\mathbf{z}_{out})$ and $G_j(\tilde{\mathbf{z}}_{out})$ will be the corresponding feature maps at jth depth in the decoder. Here, j represents the position where upsampling operations are applied in the decoder. Then, L_{uf} is denoted as:

$$L_{uf} = \frac{1}{|S_u|} \sum_{x_i^u \in S_u} \sum_{j=1}^{J} \mathbf{T}(D_j(\mathbf{z}_{out}), G_j(\tilde{\mathbf{z}}_{out})) \; , \tag{5}$$

where J is the total number of depth in the decoder. In other words, J represents how many upsampling operations are applied in the decoder.

3. EXPERIMENT

The effectiveness of the proposed method is validated on the Inria dataset [9], which is a benchmark dataset consisting of 360 large-scale aerial images, in which each image is of the size of 5000 × 5000 and has three bands (i.e., red, green, blue) at a spatial resolution of 0.3 m/pixel. The ground reference building masks of this dataset are only publicly released for five cities (Austin, Chicago, Kitsap County, Western Tyrol, and Vienna). All aerial images and ground-truth building masks are cut into small patches with the size of 256 × 256 pixels. Data split in the Inria dataset is according to the setup in [9]. More specifically, for each city, images with ids 1-5

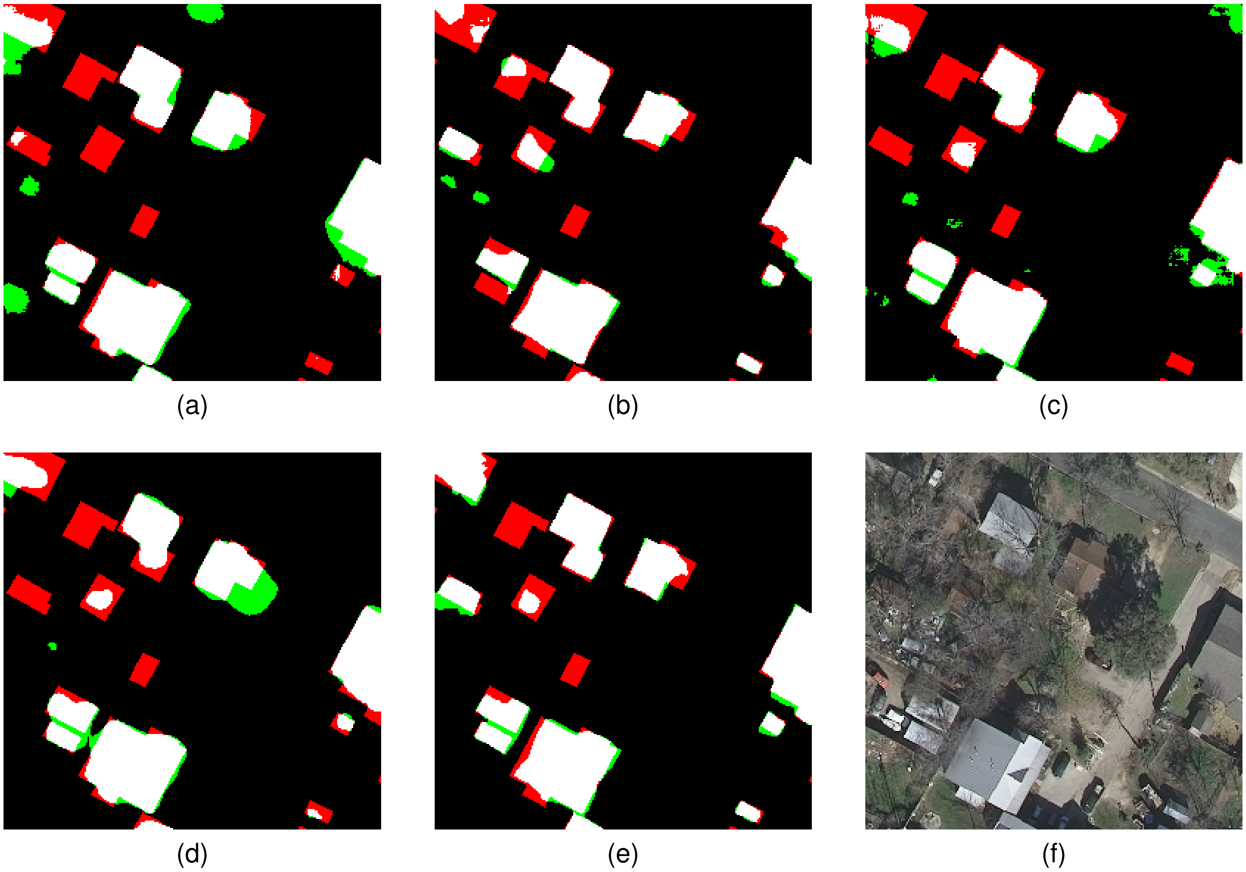

Fig. 2. Results obtained from (a) SL, (b) CCT [7], (c) CR [4], (d) PiCoCo [5], and (e) proposed method. In this experiment, the ratio of labeled data to unlabeled data is 1:10 (3600 labeled, 36252 unlabeled). (f) is aerial imagery from the Inria dataset (spatial resolution: 0.3 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

are used for validation, and the remaining 31 images are for training. Afterward, we randomly split the training data into two parts, which are labeled set and unlabeled set, and the pixel-level annotations are excluded in the unlabeled set. Under the semi-supervised setting, the ratios of labeled data to unlabeled data are set as 1:10. The statistics are derived from the validation set.

In order to validate the superiority of our methods, we make a comparison with other competitors, including Supervised Learning (SL), CCT [7], CR [4] and PiCoCo [5]. SL is regarded as the baseline method that is only trained with labeled data. CCT [7], CR [4] and PiCoCo [5] are the state-of-the-art consistency training-based semantic segmentation methods where labeled and unlabeled data are jointly trained. The hyperparameters λ_u and ω_u are set as 0.6 and 0.2 for our method, respectively. All methods exploit Efficient-UNet [10] as the backbone and are implemented in a Pytorch framework on an NVIDIA Tesla with 16 GB of memory. All methods are trained by an optimizer of Adam with a learning rate of 0.1, and the training batch size of all models is set as 4.

4. RESULTS

The performance of all models is evaluated by F1-Score and Intersection Over Union (IoU). Table 1 and Fig 2 present the quantitative and qualitative results of all methods. Our proposed approach outperforms both supervised and semi-supervised methods in terms of F1 score and IoU. Notable, our method can effectively avoid more false alarms than other methods. This suggests that the proposed method has a better capability of utilizing unlabeled data to improve network performance.

5. CONCLUSION

In this paper, we have proposed a novel semi-supervised network that generates building footprints based on feature and

Table 1. Accuracy indices of different methods derived from the validation set of Inria dataset.

Method	F1-Score	IoU
SL	77.87 %	64.12 %
CCT [7]	83.00 %	70.93 %
CR [4]	78.27 %	64.30 %
PiCoCo [5]	80.91 %	67.94 %
Proposed method	**83.74 %**	**72.03 %**

output consistency training. We evaluate our approach on Inria dataset. Experimental results have demonstrated that our method is more competitive when compared with the state-of-the-art supervised and semi-supervised semantic segmentation methods. Notable that our method can offer more satisfactory building footprints, where omission errors can be alleviated to a large extent. In this regard, other works that only have limited labeled samples will benefit from the proposed approach.

6. ACKNOWLEDGEMENTS

This work is a part of the project ”Investigation of building cases using AI” funded by Bavarian State Ministry of Finance and Regional Identity (StMFH) and the Bavarian Agency for Digitization, High-Speed Internet and Surveying. The work is jointly supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*), by the Helmholtz Association through the Framework of Helmholtz AI (grant number: ZT-I-PF-5-01) - Local Unit “Munich Unit @Aeronautics, Space and Transport (MASTr)” and Helmholtz Excellent Professorship “Data Science in Earth Observation - Big Data Fusion for Urban Research”(grant number: W2-W3-100), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab ”AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” (grant number: 01DD20001) and by German Federal Ministry of Economics and Technology in the framework of the ”national center of excellence ML4Earth” (grant number: 50EE2201C).

7. REFERENCES

[1] Yilei Shi, Qingyu Li, and Xiao Xiang Zhu, “Building segmentation through a gated graph convolutional neural network with deep structured feature embedding,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 184–197, 2020.

[2] Qingyu Li, Yilei Shi, Xin Huang, and Xiao Xiang Zhu, “Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf),” *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[3] Qingyu Li, Lichao Mou, Yuansheng Hua, Yilei Shi, and Xiao Xiang Zhu, “Building footprint generation through convolutional neural networks with attraction field representation,” *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[4] Jiaxin Wang, Chris HQ Ding, Sibao Chen, Chenggang He, and Bin Luo, “Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label,” *Remote Sensing*, vol. 12, no. 21, pp. 3603, 2020.

[5] Jian Kang, Zhirui Wang, Ruoxin Zhu, Xian Sun, Ruben Fernandez-Beltran, and Antonio Plaza, “Picoco: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10548–10559, 2021.

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 694–711.

[7] Yassine Ouali, Céline Hudelot, and Myriam Tami, “Semi-supervised semantic segmentation with cross-consistency training,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12674–12684.

[8] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le, “Unsupervised data augmentation for consistency training,” *arXiv preprint arXiv:1904.12848*, 2019.

[9] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez, “Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark,” in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 3226–3229.

[10] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar, “Eff-unet: A novel architecture for semantic segmentation in unstructured environment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 358–359.