International Journal of Computer Vision (2023) 131:3252–3271
https://doi.org/10.1007/s11263-023-01862-2

Toward Practical Weakly Supervised Semantic Segmentation via Point-Level Supervision

Junsong Fan[1,2] · **Zhaoxiang Zhang**[1,2,3]

Received: 2 September 2022 / Accepted: 24 July 2023 / Published online: 12 August 2023

Abstract
Weakly supervised semantic segmentation (WSSS) aims to reduce the cost of collecting dense pixel-level annotations for segmentation models by adopting weak labels to train. Although WSSS methods have achieved great success, recent approaches mainly concern the image-level label-based WSSS, which is limited to object-centric datasets instead of more challenging practical datasets that contain many co-occurrent classes. In comparison, point-level labels could provide some spatial information to address the class co-occurrent confusion problem. Meanwhile, it only requires an additional click when recognizing the targets, which is of negligible annotation overhead. Thus, we choose to study utilizing point labels for the general-purpose WSSS. The main difficulty of utilizing point-level labels is bridging the gap between the sparse point-level labels and the dense pixel-level predictions. To alleviate this problem, we propose a superpixel augmented pseudo-mask generation strategy and a class-aware contrastive learning approach, which manages to recover reliable dense constraints and apply them both to the segmentation models' final prediction and the intermediate features. Diagnostic experiments on the challenging Pascal VOC, Cityscapes, and the ADE20k datasets demonstrate that our approach can efficiently and effectively compensate for the sparse point-level labels and achieve cutting-edge performance on the point-based segmentation problems.

Keywords Weakly supervised learning · Semantic segmentation · Deep learning

1 Introduction

Semantic segmentation is one of the essential tasks for scene understanding, which parses the scenes into regions with semantic classes. In the era of deep learning, this field is dominated by FCN-based models (Chen et al., 2014, 2018a, 2017, 2018b; Zhao et al., 2017; Yuan & Wang, 2018; Fu et al., 2018), which solve the problem with a per-pixel classification paradigm with fully convolutional networks. However, these deep models usually require dense pixel-level labels for training, which is very costly to obtain. For example, it is reported that the annotation cost is 1.5 h/image on the Cityscapes dataset (Cordts et al., 2016). Such difficulty hinders collecting enough training data to apply the deep models in practical scenarios.

To alleviate the data scarcity problem, researchers proposed weakly supervised semantic segmentation (WSSS). The WSSS aims to utilize only coarsely labeled data, i.e., weak labels, to train semantic segmentation models and still requires the trained models to predict accurate dense pixel-level segmentation results for testing. Typical weak labels include image-level labels (Wei et al., 2017a; Ahn & Kwak, 2018; Fan et al., 2018; Huang et al., 2018; Lee et al., 2019; Ahn & Kwak, 2018; Wang et al., 2018; Zeng et al., 2019; Fan et al., 2020c, a, b; Jiang et al., 2019), scribbles (Vernaza & Chandraker, 2017; Lin et al., 2016; Tang et al., 2018), bounding boxes (Dai et al., 2015; Song et al., 2019; Khoreva et al., 2017; Li et al., 2018), and sparse points (Qian et al., 2019; Bearman et al., 2016), etc. Compared to the dense

Communicated by Rynson W.H. Lau.

✉ Zhaoxiang Zhang
zhaoxiang.zhang@ia.ac.cn

Junsong Fan
junsong.fan@ia.ac.cn

[1] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences (HKISI_CAS), Hong Kong 999077, China

[2] Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China

[3] University of Chinese Academy of Sciences (UCAS), Beijing 100190, China