Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ23839

Weakly Supervised Semantic Segmentation with Deep Learning

Xinyan Xu

Shanghai Luhang Middle School, Shanghai, China
DorisXu0622@163.com

Abstract. Weakly supervised semantic segmentation aims to achieve segmentation performance comparable to fully supervised methods through low-cost annotation forms such as image level labels or bounding boxes. This article systematically reviews two types of weakly supervised learning methods based on image level labels and bounding box supervision. For image level label supervision, mainstream methods generate initial seed regions through Class Activation Mapping (CAM) and use pixel correlation expansion or iterative optimization strategies (such as erasure and adversarial training) to solve the problem of CAM only covering discriminative regions; Representative works such as SEC and AE-PSL improve segmentation integrity by introducing significance constraints or self - training mechanisms. For bounding box supervision, BoxSup et al. have demonstrated that instance segmentation frameworks such as CRF refinement or Mask RCNN can effectively utilize intra box coordinate information to generate high-quality pseudo labels, while recent work such as BBAM has explored intra box pixel level relationships through attention mechanisms. Furthermore, this article compares the efficiency performance trade-off between two types of supervised signals: image level label annotation has the lowest cost but relies on complex post-processing. Experiments have shown that hybrid supervised methods combining multi-stage self-training and cross modal consistency constraints, such as SDI-MTL, can significantly narrow the performance gap between weakly supervised and fully supervised methods. Future directions include exploring noise robust label propagation mechanisms and weakly supervised learning frameworks for cross task collaboration.

Keywords: Weakly supervised semantic segmentation, computer vision, deep learning.

1. Introduction

Semantic segmentation is a computer vision technique that is a pixel level classification task aimed at assigning a category label to each pixel in an image, in order to understand the content of the image and identify and classify different objects and scenes in the image at the pixel level [1]. It not only identifies objects in the image, but also provides specific category information of the objects. Unlike object detection which only focuses on finding specific objects and providing bounding boxes, semantic segmentation provides more detailed information. For example, in a street image, it can not only recognize "people" and "cars", but also accurately indicate the specific contours of each person and car. There are many typical models for semantic segmentation.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ23839

Weakly Supervised Semantic Segmentation is a branch of semantic segmentation task that utilizes incomplete, imprecise, or cheaper annotated data to train models. Its advantages over traditional semantic segmentation include significantly reducing annotation costs, as pixel level annotation requires professional tools and personnel, while weakly supervised training can be achieved using image classification label, graffiti, or bounding boxes, reducing annotation time by more than 90%. There are also extended application scenarios. Weak supervision can utilize existing coarse annotation data to achieve segmentation. In addition, it promotes the generalization ability of the models.

Therefore, weakly supervised semantic segmentation sacrifices a small amount of accuracy in exchange for a significant improvement in annotation efficiency [2]. Its core value lies in transferring the burden of manual annotation to algorithm design, promoting the practical implementation of semantic segmentation in fields. In the future, the performance of weakly supervised methods is expected to further approach the level of fully supervised methods.

The first model is Class Activation Maps, which is an important visualization technique in semantic segmentation and image classification. In semantic segmentation tasks, CAMs can generate rough category activation regions to assist weakly supervised segmentation or provide model interpretability. The core idea and basic principle of CAMs is to use Global Average Pooling and fully connected layers to map the weights back to the spatial positions of feature maps, generating category related heat maps. The second type is the Bounding Box. Bounding Box is not a direct output result, but it may play an auxiliary role in data annotation, model training, or post-processing. It can play a role in data annotation and preprocessing for weakly supervised learning. It can optimize post-processing and reduce false positives. Extract the minimum bounding box from the connected regions in the segmentation results, and filter out noise by combining the target size or position prior. Accelerated processing, in large images, first generate a Bounding Box using an object detection model, and then perform fine segmentation on the areas within the box.

2. Image-level label supervision

2.1. Overview

In semantic segmentation tasks, using classification labels for supervision is a weakly supervised learning method that trains models by relying only on image level category labels. The goal is to use classification labels to generate pixel level segmentation masks, reducing reliance on expensive manual annotation. The key challenge is how to transfer image level category information to pixel level, achieving region localization and segmentation. Improvements can be made through cross image relationships. It can also be improved through multimodal fusion, combined with CLIP's text image alignment knowledge. By combining classification labels with generative optimization methods, weakly supervised semantic segmentation gradually approaches fully supervised performance while reducing annotation costs. Its advantages include significantly reducing annotation cost, and can be combined with pre trained models to improve small data generalization. The disadvantage is that the segmentation accuracy is usually lower than that of fully supervised methods, and so on. There are many methods and models.

2.2.Class activation maps

Standard CAM: In semantic segmentation Class Activation Maps (CAM) is a weakly supervised localization technique based on classification networks, which can generate pixel level rough region

DOI: 10.54254/2755-2721/2025.TJ23839

response maps using image level category labels [3]. The key assumption is that the feature map of the last convolutional layer in the classification network contains spatial semantic information. It has structural dependencies, such as having to include a GAP layer. Its limitation is that it cannot be directly applied to GAP free networks such as partial transformers. The application process of CAM in semantic segmentation includes generating pseudo labels, optimizing pseudo labels, and training segmentation models. The second method is pseudo label optimization, CRF (Conditional Random Field), diffusion method and adversarial erasure. The third method is to train segmentation models. Firstly, the input is the original image plus optimized pseudo labels. The advantages of CAM include low annotation cost, Interpretability and flexibility. The disadvantage is rough positioning. The forefront improvement direction is multimodal fusion, self training iteration, optimizing the model through high confidence pseudo label iteration [4] and application scenarios include medical imaging. CAM and its variants provide fundamental tools for weakly supervised semantic segmentation, and subsequent research focuses on improving localization integrity and reducing noise.

The improvement methods include Grad CAM and XGrad CAM. The second type is LayerCAM, whose core idea is to use weighted combinations of multi-layer feature maps to solve the problem of low resolution or incomplete semantics in a single layer. Its characteristics include multi-scale fusion and detail preservation. Improvement methods include adaptive weighting and non local fusion. The comparison between these two is that Grad CAM has a single dependency layer, while LayerCAM is a multi-layer fusion [5]. Grad CAM has a low resolution, while LayerCAM has a high resolution. Grad CAM is suitable for coarse positioning of large objects, while LayerCAM is suitable for small object/fine structure positioning. There are three general directions for improvement. The first is to enhance localization integrity. The second approach is to reduce noise, such as confidence filtering. The third approach is to combine new architecture. Finally, it is recommended to use different models in different scenarios. By flexibly selecting and optimizing CAM methods, better results can be achieved in areas such as weakly supervised segmentation and interpretability analysis.

3. Bounding-box level supervision

3.1. Overview

The Bounding Box model is one of the fundamental technologies used in computer vision for object detection and semantic segmentation. In semantic segmentation tasks, Bounding Box is usually used as a preprocessing step or auxiliary information to help the model more accurately locate and segment specific objects in the image. Its core functions include target positioning, area division, and multi-objective processing. In semantic segmentation, it can be used in the preprocessing stage, such as first detecting Bounding Boxes and then finely segmenting the areas within the boxes. It can also be used for attention mechanisms, such as using Bounding Box information to guide the network to focus on specific areas. There is also post-processing optimization, combined with Bounding Box information to optimize the accuracy of edge segmentation.

Its advantages are high computational efficiency, suitability for real-time applications, providing object level global information, and easy integration with other visual tasks. There are also limitations, such as the difficulty of accurately matching irregular object shapes with rectangular boxes, and the detection effect on small or dense objects.

DOI: 10.54254/2755-2721/2025.TJ23839

3.2. Box-to-segmentation techniques

Bounding Box is usually not the direct core output, but it plays an important role in related tasks such as object detection or instance segmentation, and may be used in conjunction with semantic segmentation. It is a rectangular box surrounding the target object in the image, usually represented by coordinates such as the top left and bottom right corner coordinates, or the center point coordinates plus width and height [6]. It is mainly used to locate the position of objects, but cannot provide pixel level fine segmentation information. It is different from semantic segmentation, which assigns category labels to each pixel and outputs pixel level masks without relying on Bounding Boxes. Bounding Boxes only mark the rough position and range of objects and do not distinguish the pixel categories inside the object (such as its specific shape). In complex tasks such as instance segmentation, it is possible to first use Bounding Boxes to detect object positions, and then perform semantic segmentation on the regions within each box.

Bounding Box has many functions in semantic segmentation. Although semantic segmentation directly outputs pixel masks, Bounding Box may still be used in the following scenarios: firstly, data annotation: when annotating semantic segmentation datasets, Bounding Box may be used to quickly locate the target, and then refine the annotation mask. Next is preprocessing: cropping the Bounding Box area as input for semantic segmentation to reduce computational complexity. There is also post-processing: generating a Bounding Box for the segmentation results, which is used for target size analysis or visualization. The related technologies of Bounding Box include Mask R-CNN and Two Stage. The former is a classic instance segmentation model that first detects objects through Bounding Box and then predicts masks within each Box [7]. The latter first generates candidate Bounding Boxes and then performs semantic segmentation on each Box. In pure semantic segmentation tasks, Bounding Box is not necessary, but in multi task learning or complex scenarios (such as instance segmentation, object detection+segmentation joint tasks), it can assist in locating and optimizing segmentation results [8]. Understanding the differences and connections between the two can help in selecting suitable models and application scenarios.

4. Discussion

In computer vision tasks, Bounding Box and Image level Labels are two different granularity annotation methods. Bounding Box provides richer spatial information than Image level Labels, making it significantly advantageous in many tasks. The following are the main advantages of Bounding Box that Image level Labels do not have. Firstly, it has spatial positioning capability. Bounding Box can clearly label the position and range of objects in the image (rectangular area), which can be directly used for tasks such as object detection and instance segmentation. Image level labels only provide the overall category of the image, but cannot determine where the object is or how many instances it has. For example, in autonomous driving, it is necessary to know the specific location of pedestrians or vehicles (Bounding Box), and only knowing the "Image level Label" is not enough [9]. The second is whether it supports more complex tasks. The third is to reduce background interference. The fourth is to quantify the size and proportion of objects. The fifth question is whether it supports multi-instance differentiation. Bounding Box can annotate multiple instances of the same category (such as multiple cats in an image). Image level labels can only indicate the presence of cats, but cannot distinguish between quantity or individuals. The sixth is efficient weakly supervised learning [10]. The seventh one is that the model has stronger interpretability.

DOI: 10.54254/2755-2721/2025.TJ23839

The core advantage of Bounding Box lies in the richness of spatial information, while Image level Labels are more suitable for low-cost, coarse-grained tasks. Both can be combined according to actual needs (for example, pre training with Image level Labels first, and then fine-tuning with Bounding Box). Therefore, when precise object localization is required (such as autonomous driving, medical imaging), tasks depend on object size or quantity (such as industrial counting), and data is sufficient and annotation resources permit, Bounding Box can be selected. When only classification is needed (such as automatic labeling of photo albums) or annotation costs are limited (weakly supervised learning), image level labels can be selected.

5. Conclusion

This article systematically studies weakly supervised semantic segmentation methods based on image level labels and bounding box supervision, and explores the performance efficiency trade-off under different supervised signals. Experiments have shown that image level labels have become one of the most practical weakly supervised settings due to their extremely low annotation cost, but they are susceptible to insufficient coverage of discriminative regions due to the quality of the initial seed in the class activation map (CAM). In recent years, some works have significantly alleviated this problem by introducing self-training, cross image consistency constraints, or saliency priors (such as EPS, CDA, etc.). In contrast, bounding box supervision provides stronger spatial constraints and can generate more accurate pseudo labels (such as Box2Mask, BBTP, etc.), but its annotation cost is slightly higher and requires the design of robust intra box pixel mining mechanisms (such as CRF optimization or attention modeling). Further， analysis reveals that there is complementarity between the two types of supervisory signals. Image level labels are more suitable for coarse-grained semantic localization, while bounding boxes have more advantages in fine-grained object boundary segmentation. Future research directions include hybrid supervision strategies, exploring how to efficiently combine image level labels and bounding boxes to approximate fully supervised performance with minimal annotation cost; Noise robust learning, designing more stable pseudo label generation and correction mechanisms to reduce the accumulation of initial noise errors; And cross task transfer, utilizing bounding box information from object detection or instance segmentation tasks to enhance weakly supervised learning of semantic segmentation. The experiments in this study have verified the enormous potential of weakly supervised methods in reducing annotation dependencies, but further exploration is needed to narrow the gap with fully supervised performance, especially in complex scenarios or multi class tasks.

References

[1] Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., & Agrawal, A. (2018). Context encoding for semantic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 7151-7160).

[2] Ahn, J., & Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4981-4990).

[3] Chen, Z., & Sun, Q. (2023). Extracting class activation maps from non-discriminative features as well. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3135-3144).

[4] Zou, Y., Yu, Z., Liu, X., Kumar, B. V. K., & Wang, J. (2019). Confidence regularized self-training. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5982-5991).

[5] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 325-341).

DOI: 10.54254/2755-2721/2025.TJ23839

[6] Mousavian, A., Anguelov, D., Flynn, J., & Kosecka, J. (2017). 3d bounding box estimation using deep learning and geometry. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 7074-7082).
[7] He, Y., Zhu, C., Wang, J., Savvides, M., & Zhang, X. (2019). Bounding box regression with uncertainty for accurate object detection. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 2888-2897).
[8] Dai, J., He, K., & Sun, J. (2015). Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1635-1643).
[9] Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., & Ruan, X. (2017). Learning to detect salient objects with image-level supervision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 136-145).
[10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).