Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Salient Object Detection Enhanced Pseudo-Labels For Weakly Supervised Semantic Segmentation⋆

Given-name1 Surname1[a,*], Given-name2 Surname2[a,1], Given-name3 Surname3[b], Given-name4 Surname4[b]

[a]*Affiliation 1, Address, City and Postal Code, Country*
[b]*Affiliation 2, Address, City and Postal Code, Country*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Keyword1, Keyword2, Keyword3

ABSTRACT

This paper addresses the limitations of generating pseudo-labels based on Class Activation Maps (CAM) in weakly supervised semantic segmentation tasks by proposing a novel salient object fusion framework. This framework complements CAM localization information by capturing complete contours and edge details of salient targets through the designed RGB-SOD network. We also designed a saliency object selector to dynamically balance the weights of CAM and SOD when generating single-class pseudo-labels, further improving the quality of pseudo-labels. Despite its simplicity, our method achieved competitive performances of 77.52% and 77.73% on the PASCAL VOC 2012 validation and test sets respectively, significantly enhancing the performance bottlenecks of existing methods. This work highlights the importance of effectively integrating complementary information to improve weakly supervised segmentation tasks.

1. Introduction

Semantic segmentation is a fundamental task in computer vision that aims to assign a class label to each pixel in an image. It has been widely applied in various fields, such as autonomous driving, medical imaging, and scene understanding. Recently, deep learning-based methods have achieved remarkable success in semantic segmentation tasks, especially fully supervised methods that require pixel-level annotations for training. However, obtaining pixel-level annotations is labor-intensive and time-consuming, which limits the scalability of fully supervised methods. In contrast, weakly supervised semantic segmentation methods only require image-level annotations, which are more accessible and easier to obtain. These methods have attracted increasing attention in recent years due to their practicality and efficiency.

Current weakly supervised semantic segmentation (WSSS) methods typically begin by training a classification network to generate Class Activation Maps (CAM) as initial pseudo-labels. These pseudo-labels are then refined in subsequent stages, ultimately leading to fully supervised training based on refined pseudo-labels. However, since CAMs activate not only the target objects but also contextual information that aids in class recognition, they often lead to activation confusion between target objects and non-target objects or things that frequently appear together. Consequently, CAM-based WSSS faces several key challenges: 1) the localization maps only capture a small portion of the target object, 2) they suffer from mismatches at the boundaries of objects, and 3) they almost fail to separate co-occurring pixels from the target objects (e.g., railways from trains).

To overcome these three challenges, we consider that saliency maps often have characteristics of complete target lo-

⋆This is an example for title footnote coding.
*Corresponding author: Tel.: +0-000-000-0000; fax: +0-000-000-0000;
e-mail: author3@author.com (Given-name3 Surname3)
[1]This is author footnote for second author.

calization and clear boundaries.In this paper, we first propose a new RGB-SOD network that generates saliency maps with better edge features and the ability to fully capture salient objects, complementing the excellent localization capability of CAMs.

However, how to combine the two presents another significant challenge. Saliency maps themselves do not have a concept of categories, so we initially applied the SOD to only one class of data. Then, based on the CAM and saliency map, we defined a feasibility of localization for a suitable target category, selected an appropriate threshold to obtain high-quality saliency maps that could serve as pseudo-labels. We found that these generated pseudo-labels are of extremely high quality and, when fused with pseudo-labels generated by traditional methods, showed significant improvement. From this perspective, we believe our method using two complementary streams of information can resolve the performance bottlenecks in WSSS.

2. Related Work

2.1. Weakly Supervised Semantic Segmentation

2.2. Salient Object Detection

Salient object detection refers to the use of image processing techniques and computer vision algorithms to locate the most "salient" areas in an image. Salient regions are those parts of the image that are eye-catching or important, such as the areas that the human eye would first focus on when viewing an image. Salience is a highly subjective perception and varies depending on the environment. Salient objects often differ across contexts and saliency is not easily captured by mathematical formulas.

Early work in the SOD field primarily relied on low-level visual features such as color, contrast, and texture. One of the most representative early models, proposed by Itti and others, simulated the lower visual characteristics of the human visual system, using a center-surround mechanism to predict areas of attention. With the development of deep learning technologies, recent research has shifted towards using deep neural networks to identify salient regions, leveraging high-level semantic information and achieving significant progress across multiple benchmark datasets.

Specifically, methods based on Convolutional Neural Networks (CNN) have become mainstream in SOD because they can learn more complex feature representations, thereby accurately detecting salient regions in various complex scenarios. Additionally, some studies have integrated attention mechanisms, allowing the networks to focus more on salient regions while ignoring irrelevant backgrounds.

In this paper, we present the RGB-SOD network, developed against this backdrop, aimed at further advancing SOD technology, especially in tasks involving weakly supervised semantic segmentation. By precisely capturing the boundaries of salient regions, it supports the generation of high-quality pseudo-labels.

3. Proposed Method

3.1. Prerequisites

We first introduce the method for generating attention maps. Given an input image I, let y be the image-level label. The output features F of the last convolutional layer have C channels, corresponding to the number of classes. Following the last convolutional layer is a global average pooling layer, where feature F is pooled into a vector f of size C. We compute the classification loss by applying the sigmoid cross-entropy loss function, which is defined as follows:

$$L_{ce} = -\frac{1}{C}\sum_{c=1}^{C}\left(y^c \log\left(\sigma\left(f^c\right)\right) + \left(1-y^c\right)\log\left(1-\sigma\left(f^c\right)\right)\right)$$

where σ is the sigmoid function. Attention maps can be generated from the output of the last convolutional layer. For a given class c, the attention map A^c is derived from channel c of F, and can be expressed as:

$$A^c = \frac{\mathrm{ReLU}\left(F^c\right)}{\max\left(\mathrm{ReLU}\left(F^c\right)\right)}$$

3.2. Salient Object Detection Net

We have constructed the above RGB-SOD network, which is one of the core components of our framework, responsible for salient object detection. The network consists of two main parts: the RGB encoder and decoder.

RGB Encoder: This includes four sequentially arranged Swin Transformer blocks (Swin Block1 to Block4), which process the input RGB image in succession, gradually extracting and refining features. Each Swin block is connected to a corresponding edge-aware module, labeled R1 to R4. Specifically, the R4 block is designed as RGB-Edge Aware to enhance the network's perception of image edges.

Decoder: Comprised of four Pixel Attention Modules (PATM), it aims to process and integrate features from various stages of the encoder. The decoder upsamples the feature maps through a transpose convolution layer (TCONV1) and connects via concatenation operations with the PATM modules. The PATM modules focus on preserving the spatial details of salient regions and enhancing the clarity of boundaries.

The decoder ultimately produces a predicted saliency map (Pred Sal-Map) that aligns with the ground truth saliency map (GT Sal-Map) and is guided by two supervisory signals: saliency supervision and edge supervision, ensuring the accuracy and effectiveness of model training. Additionally, the NAMLAB edge module is integrated into the network, further optimizing edge features and improving the recognition of RGB edges.

Overall, the RGB-SOD network leverages the advantages of saliency detection in terms of locating integrity and edge clarity, complemented by the localization information provided by the CAM. Such a design enables the network to more precisely capture the complete contours of target objects, handle the separation issues of co-occurring objects, significantly improving the quality of pseudo-labels, and providing more reliable labeling information for weakly supervised semantic segmentation tasks.

3.3. Overall Framework

We propose a salient object fusion framework that spans two parallel paths based on Class Activation Mapping (CAM) and Salient Object Detection (SOD), aimed at generating high-quality pseudo-labels for weakly supervised semantic segmentation tasks. The specific process is as follows: Given an input single-class image and its category label, we first obtain the CAM attention map through the classification model to preliminarily estimate the rough location of the target object, serving as seed areas. Simultaneously, we input the single-class image into a specially designed SOD model to capture the complete contours and fine boundary details of the target object using its saliency detection capability, generating a Salient Object map.

Next, a critical component is the Salient Object Selector module, which reasonably filters and selects the SOD map, picking high-quality Salient Object Pseudo-Masks that have high consistency (α^*) in localization with the CAM attention map.

In the pseudo-label generation stage, the CAM attention map is considered an initial seed and is expanded using an Expansion strategy to generate preliminary rough pseudo-labels. Then, we integrate the optimized Salient Object Pseudo-Masks, merging the two complementary types of information through a supervised optimization strategy to finally produce high-precision Pseudo-Masks pseudo-labels.

Finally, we integrate the high-quality Pseudo-Masks pseudo-labels into the Pseudo Ground-Truth dataset, serving as a supervisory signal to train the target semantic segmentation model, significantly enhancing segmentation performance.

The core innovation of this framework is the clever integration of CAM and SOD, two complementary approaches to object localization. It designs a series of modules that coordinate and merge the two approaches at different stages, generating high-quality supervisory signals, overcoming the limitations of single-path based approaches, and effectively optimizing weakly supervised semantic segmentation.

3.4. Salient Object Selector

In this study, we explore a strategy for enhancing the accuracy of pseudo-labels for single-class images using Class Activation Mapping (CAM) and Salient Object Detection (SOD) information, termed the Salient Pseudo-Labeling Strategy. The core of this strategy lies in selecting SOD images that align with CAM localization as pseudo-labels to generate high-quality labels for segmentation model training.

We introduce a key parameter, α, aimed at quantifying the accuracy of SOD images in target class localization. Specifically, α characterizes their relationship and consistency in target localization by calculating the Intersection over Union (IoU) between CAM and SOD images under certain threshold conditions. This not only reveals the interaction between CAM and SOD but also allows us to refine the pseudo-label generation process by dynamically adjusting the threshold. The formula is expressed as follows:

$$\alpha = \mathrm{IoU}(\mathrm{P_{CAM}}, \mathrm{P_S})_\tau$$

Algorithm 1 SOD Enhanced Pseudo-Labels for single class images

Require: Pseudo labels $P_1, \ldots, P_L$, Salient maps $S_0, \ldots, S_L$, threshold $t_0 = 0.4$, threshold $t_1 = 0.2$
Ensure: Enhanced pseudo-labels $\hat{P}_0, \ldots, \hat{P}_L$

```
procedure SEP(P_0, ..., P_L, S_0, ..., S_L)
    for k = 1 to L do
        P̂_k ← P_k
        if S_k == 0_{H×W} then
            continue
        end if
        o_S = Intersect(S_k, P_k)/union(S_k, P_k)
        if object is silent and o_S > t_0 then
            P̂_k ← S_k
        end if
        if object is not silent and o_S > t_0 + t_1 then
            P̂_k ← S_k
        end if
    end for
end procedure
```

Furthermore, we also consider the inherent salience of the object itself. For this purpose, we introduce a norm metric in the selector to represent the salience intensity of individual objects, which is crucial for identifying objects with inherently weak salience, such as chairs, sofas, etc.

To integrate the object's inherent salience, we increase the norm of each class on the salience map as a measure of its salience. By combining this norm and the judgment parameter α, we obtain a new α^*:

$$\alpha^* = \left(\mathrm{IoU}(\mathrm{P_{CAM}}, \mathrm{P_S}) + \frac{\|S^c\|_2}{\max(\|S\|_2)}\right)_\tau$$

Here, $\mathrm{P_{CAM}}$ and $\mathrm{P_S}$ represent the pseudo-labels generated by CAM and SOD technologies, respectively, and τ is a preset threshold used to adjust the selection relationship between CAM and SOD images. With this method, α^* becomes a dynamically adjusted parameter, balancing the influence of CAM and SOD in the final pseudo-label generation. The final single-class pseudo-label, $\mathrm{P_{final}}$, is given by:

$$\mathrm{P_{final}} = (1 - \alpha^*) \cdot \mathrm{P_{CAM}} + \alpha^* \cdot \mathrm{P_S}$$

When the overlap between CAM and SOD is significant, indicating high consistency in their target region localization, the information from SOD gains more weight in the pseudo-label generation process; conversely, when their overlap is less, indicating that SOD is more independent in target localization, its information dominates in the pseudo-labels. Through this method, we not only enhance the quality of pseudo-labels but also improve the accuracy and reliability of label data during the segmentation model training process.

Table 1. Comparison of Various Methods on Segmentation Performance

Method	Public	Backbone	Sup.	Val	Test
BCM	CVPR19	ResNet101	I+B	70.2	-
BBAM	CVPR21	ResNet101		73.7	73.7
EPS	CVPR21	ResNet101	I+S	71.0	71.8
L2G	CVPR22	ResNet101		72.1	71.7
SEAM	CVPR20	ResNet38	I	64.5	65.7
AdvCAM	CVPR21	ResNet101		68.1	68.0
OC-CSE	ICCV21	ResNet38		68.4	68.2
CPN	ICCV21	ResNet38		67.8	68.5
VWE	IJCV22	ResNet101		70.6	76.7
CLIMS	CVPR22	ResNet101		70.4	70.0
MCTformer	CVPR22	ResNet38		71.9	71.6
SIPE	CVPR22	ResNet101		68.8	69.7
W-OoD	CVPR22	ResNet38		70.7	70.1
AMN	CVPR22	ResNet101		69.5	69.6
ViT-PCM	ECCV22	ResNet101		70.3	70.9
Yoon et al.	ECCV22	ResNet38		70.9	71.7
ToCo	CVPR23	ViT-B		69.8	70.5
CLIP-ES	CVPR23	ResNet101		73.8	73.9
ClusterCAM	IEEE Access24	DeiT-Se		70.3	70.7
SFC	AAAI24	ResNet101		71.2	72.5
(Ours)	-	ResNet101	I+S	74.9	74.59
(Ours)	-	ResNeSt101		76.55	77.06
(Ours)	-	ResNeSt269		77.52	77.73

Table 2. Performance metrics at different alpha values

Note: Here is some explanatory text about the table. This can describe what the data means or provide any relevant context.

Method	Public	Sup.	mIoU (%)
PSA	CVPR 2018	I	58.4
ICD	CVPR 2020	I	62.2
SubCat	CVPR 2020	I	63.4
SEAM	CVPR 2020	I	63.6
A^2GNN	TPAMI 2021	I	65.3
QA_CLIMS	ACM MM 2023	I	71.8
L2G	CVPR 2022	I+S	69.8
ESEPM (baseline)	-	I	72.5
(ours)	-	I+S	73.8

4. Experiments

4.1. Experimental Setup

4.2. Comparisons with State-of-the-art Methods

We trained our segmentation network using the conventional ResNeSt101+DeepLabV2 setup for comparison, but to further enhance network performance, we also employed the ResNeSt architecture. This architecture generally improves the learned feature representations, thereby enhancing performance in image classification, object detection, instance segmentation, and semantic segmentation.

In Table 1, we present the performance of our final trained network. Compared to other state-of-the-art (SOTA) methods, our pseudo-labels show significant improvements in performance on both the validation and test sets under the same training settings. You can see our results using ResNeSt269+DeepLabV3+ on the VOC server at the following links: http://host.robots.ox.ac.uk:8080/anonymous/HYF7A0.html and http://host.robots.ox.ac.uk:8080/anonymous/ZA3UVZ.html.

4.3. Ablation Study

4.3.1. the alpha parameter for the saliency object selector

4.3.2. the universality of the saliency object selector on pseudo-labels generated by different algorithms

We compared the changes in segmentation performance before and after the application of the saliency object selector under different baseline pseudo-label generation strategies. We

Table 3. Performance metrics at different alpha values

Note: Here is some explanatory text about the table. This can describe what the data means or provide any relevant context.

Alpha	Acc	Acc_class	mIoU	fwIoU
0	94.12%	85.30%	77.57%	89.10%
0.05	94.35%	85.89%	78.35%	89.48%
0.1	94.41%	86.18%	78.66%	89.58%
0.2	94.49%	86.63%	79.04%	89.75%
0.3	94.54%	87.13%	79.44%	89.89%
0.4	94.44%	87.36%	79.45%	89.74%
0.5	94.25%	87.39%	79.15%	89.46%
0.6	94.10%	87.72%	79.08%	89.29%
0.7	93.86%	87.71%	78.83%	88.92%
0.8	93.61%	87.43%	78.33%	88.51%
0.9	93.52%	87.35%	78.15%	88.36%
1	93.51%	87.33%	78.13%	88.35%

Table 4. Comparison

Note: Here is some explanatory text about the table. This can describe what the data means or provide any relevant context.

Method	SOD	Acc	Acc_class	mIoU	fwIoU
RS+EPM	×	93.51%	87.33%	78.13%	88.35%
RS+EPM	✓	93.86%	87.71%	78.83%	88.92%
QA-CLIMS	×	93.92%	86.46%	77.88%	88.72%
QA-CLIMS	✓	94.34%	87.09%	78.91%	89.47%
RCA	×	90.52%	68.30%	64.72%	82.46%
RCA	✓	91.91%	72.07%	68.03%	84.95%
L2G	×	93.36%	87.90%	76.86%	87.91%
L2G	✓	93.76%	88.13%	77.69%	88.59%
SEAM	×	85.40%	78.77%	60.60%	75.62%
SEAM	✓	87.28%	80.01%	63.38%	78.47%

examined strategies including RS+EPM, QA-CLIMS, RCA, L2G, and SEAM. This table shows the improvement in pseudo-label quality by the saliency object selector, in terms of accuracy, class accuracy, mean Intersection over Union (mIoU), and frequency-weighted IoU. These results demonstrate the good universality of our proposed saliency object selector, which can effectively enhance the quality of pseudo-labels generated by different baseline methods.

5. Conclusion

Acknowledgments

Acknowledgments should be inserted at the end of the paper, before the references, not as a footnote to the title. Use the unnumbered Acknowledgements Head style for the Acknowledgments heading.

References

Please ensure that every reference cited in the text is also present in the reference list (and vice versa).

Reference style

Text: All citations in the text should refer to:

1. Single author: the author's name (without initials, unless there is ambiguity) and the year of publication;
2. Two authors: both authors' names and the year of publication;
3. Three or more authors: first author's name followed by 'et al.' and the year of publication.

Citations may be made directly (or parenthetically). Groups of references should be listed first alphabetically, then chronologically.

References

Supplementary Material

Supplementary material that may be helpful in the review process should be prepared and provided as a separate electronic file. That file can then be transformed into PDF format and submitted along with the manuscript and graphic files to the appropriate editorial office.