Charge Own Job: Saliency Map and Visual Word Encoder for Image-Level Semantic Segmentation

Yuhui Guo, Xun Liang ✉, Hui Tang, Xiangping Zheng, Bo Wu, and Xuan Zhang

Renmin University of China, Beijing, China
{yhguo,xliang,huitang,xpzheng,wubochn,zhangxuanalex}@ruc.edu.cn

Abstract. Significant advances in weakly-supervised semantic segmentation (WSSS) methods with image-level labels have been made, but they have several key limitations: incomplete object regions, object boundary mismatch, and co-occurring pixels from non-target objects. To address these issues, we propose a novel joint learning framework, namely **S**aliency **M**ap and **V**isual **W**ord **E**ncoder (**SMVWE**), which employs two weak supervisions to generate the high-quality pseudo labels. Specifically, we develop a visual word encoder to encode the localization map into semantic words with a learnable codebook, making the network generate localization maps containing more semantic regions with the encoded fine-grained semantic words. Moreover, to obtain accurate object boundaries and eliminate co-occurring pixels, we design a saliency map selection mechanism with the pseudo-pixel feedback to separate the foreground from the background. During joint learning, we fully utilize the cooperation relationship between semantic word labels and saliency maps to generate high-quality pseudo-labels, thus remarkably improving the segmentation accuracy. Extensive experiments demonstrate that our proposed method better tackles above key challenges of WSSS and obtains the state-of-the-art performance on the PASCAL VOC 2012 segmentation benchmark.

Keywords: Weakly-supervised semantic segmentation · Saliency map · Visual word encoder · Pseudo labels.

1 Introduction

Semantic segmentation aims to predict pixel-wise classification results on images, which is one important and challenging task of computer vision. With the development of deep learning, a variety of Convolutional Neural Network (CNN) based semantic segmentation methods [7,8] have achieved promising successes. However, they require a large number of training images annotated with pixel-level labels, which is both expensive and time-consuming. Thus, various weakly supervised semantic segmentation (WSSS) methods have attracted increasing interest of researchers. Most existing WSSS studies adopt image-level labels as the weak supervision of the segmentation model, in which a segmentation network is trained on images with less comprehensive annotations that are cheaper to obtain than pixel-level labels. The image-level WSSS methods usually perform semantic segmentation through generated pseudo-labels as weak supervision. In general, using a classification network to generate class activation maps (CAM) [46] containing object localization maps, which can be as initial pseudo labels to achieve the semantic segmentation performance [35,4]. However, the classification network has the ability to classifity, which does not locate the integral extents of target objects, leading to the generated CAM that typically only cover the most discriminative parts of target objects. Thus, during the process of producing pseudo labels, WSSS will be confronted with the following key challenges: i) the extents of the target objects can not be covered completely [46], ii) the localization map is unable to obtain accurate object boundaries [22], and iii) the localization map contains co-occurring pixels between target objects and the background [23]. These three aspects in pseudo labels are important to the final semantic segmentation performance [35,4].

Recently, many WSSS methods have been proposed to focus on tackling these issues. According to different issues, existing methods can be divided into three categories. To address the incomplete object region issue of pseudo-labels, researchers utilize the pixel-affinity based strategy [2,1] or erasing strategy [10,22,25] to enlarge the receptive field and discover more discriminative parts for target objects. However, they only focus on the object coverage extents, and neglect that accurate object boundaries are benefit for semantic annotation. Thus, in order to address the object boundary mismatch issue, researchers propose to use the idea of explicitly exploring object boundaries from training images [13,9] to keep coincidence of segmentation and boundaries. Due to some co-occurring pixels exist in between the foreground and the background [11], these methods still lack of the clue to explore the correlation between the foreground and the background, thus they are unable to correctly separate the foreground from the background. In order to alleviate the co-occurring pixels issue between the foreground and the background, most existing WSSS methods use the saliency map [23,36,37,26,19,38,34,15] to induce processing the background, reducing the computation burden of the segmentation model and helping the segmentation model distinguish coincident pixels of non-target objects from a target object. However, these WSSS methods directly utilize the saliency maps from off-the-shelf saliency detection models as the clue of co-occurring pixels, which is easy to separate the foreground from the background, but such a way is not beneficial to that the segmentation model generates self-saliency maps, leading to a not end-to-end manner training process.

In this paper, our goal is to overcome these challenges of WSSS with image-level labels by improving the performance of the localization map generated by the classification network. For this purpose, we propose a novel joint learning method for WSSS, namely saliency map and visual word encoder (SMVWE), to simultaneously learn semantic word labels and saliency maps. As shown in Figure 1, we design a visual word encoder to help the classification network learn the semantic word labels, leading to that the generated localization map could cover more integral semantic extents of target objects. Due to the image-level

WSSS task is unable to directly use the semantic word labels, we use an unsupervised way to generate their vector representations in each forward pass, i.e., each semantic word in a trainable codebook utilizes the manhattan distance to encode the feature maps from the classification network. In such a way, it alleviates the sparse object region problem, but does not separate their boundaries from the background effectively. Thus, we design a saliency map selection mechanism to address inaccurate object boundaries and co-occurring pixels among objects, where the saliency maps from off-the-shelf saliency detection models are used as pseudo-pixel feedback. Specifically, the classification network based on image-level labels performs semantic segmentation for L target object classes and one background class, thus generating L foreground localization maps and one background localization map to represent the saliency maps. To obtain accurate object boundaries and discard the co-occurring pixels, we compare our generated saliency maps with off-the-shelf groundtruth saliency maps by a saliency loss, producing more effective saliency maps to improve the quality of final pseudo labels. Moreover, we also use the multi-label classification losses containing the image-level label prediction and the semantic word label prediction, which combine with the saliency loss to optimize our proposed model, thus generating higher-quality pseudo-labels for training the semantic segmentation network.

In summary, our main contributions are three folds:

- We propose a novel joint learning framework for WSSS, namely saliency map and visual word encoder (SMVWE), which learns from pseudo-pixel feedback by combining two weak supervisions, thereby effectively preventing the localization map from producing wrong attention regions.
- We develop a visual word encoder to generate semantic word labels. By enforcing the classification network to learn the generated semantic word labels, more object extents could be discovered, thus alleviating the sparse object region problem.
- We design a saliency map selection mechanism to separate the foreground from the background, which could capture precise object boundaries and discard co-occurring pixels of non-target objects, remarkably improving the quality of pseudo-labels for training semantic segmentation networks.

2 Related Work

2.1 Weakly-Supervised Semantic Segmentation

Existing weakly-supervised semantic segmentation methods using image-level labels mainly focus on two types of algorithms, including single- and multi-stage methods. Single-stage methods [17,27,30,31] could achieve the semantic segmentation of images through a high-speed and simple end-to-end process. For example, RRM [43] proposes an end-to-end network to mine reliable and tiny regions and use them as ground-truth labels, then combining a dense energy loss to optimize the segmentation network. SSSS [3] adopts local consistency, semantic fidelity, and completeness as guidelines, proposing a segmentation-based network and a self-supervised training scheme to solve the sparse object region problem for WSSS. Though these methods are effective for semantic segmentation, they barely achieve high-quality pseudo-labels to improve the segmentation accuracy.

Moreover, existing multi-stage methods generally perform the following three steps: (i) generate an initial localization map to localize the target objects; (ii) improve the initial localization map as the pseudo labels; and (iii) using generated pseudo-labels to train the segmentation network. Recently, many approaches [19,23,34] are devoted to alleviate the incomplete object region problem during generating pseudo-labels process. For example, adversarial erasing methods [18,36] help the classification network learn non-salient regions features and expand activation maps through erasing the most discriminative part of CAMs. Instead of using the erasing scheme, SEAM [35] proposes the consistency regularization on generated CAMs from various transformed images, and designs a pixel correlation module to exploit the context appearance information, leading to further improvement on CAMs consistency for semantic segmentation. ScE [4] proposes to iteratively aggregate image features, helping the network learn non-salient object parts, hence improving the quality of the initial localization maps. To improve the network training, MCOF [34] mines common object features from the initial localization and expands object regions with the mined features, then using saliency maps to refine the object regions as supervision to train the segmentation network. Similarly, the DSRG approach [19] proposes to train a semantic segmentation network starting from the discriminative regions and progressively increase the pixel-level supervision using the seeded region growing strategy. Moreover, MCIS [32] proposes to learn the cross-image semantic relations to mine the comprehensive object pattern and uses the co-attention to exploit context from other related images, thus improving localization maps to benefit the semantic segmentation learning. In this work, we also focus on semantic segmentation with image-level supervision and aim to improve the quality of the initial pseudo labels.

2.2 Saliency Detection

Saliency detection (SD) methods generate the saliency map that separates the foreground objects from the background in an image, which is benefit for many computer vision tasks. Most existing WSSS [36,37,26,38,15] methods have greatly benefited from SD that exploits the saliency map as the background cues of pseudo-labels. For example, the MDC method [38] uses CAMs of a classification network with different dilated convolutional rates to find object regions, and uses saliency maps to find background regions for training a segmentation model. STC [37] trains an initial segmentation network using the saliency maps of simple images, and uses the image-level annotations as supervision information to improve the initial segmentation network. Moreover, some methods [5,40] integrate class-agnostic saliency priors into the attention mechanism and utilize class-specific attention cues as an additional supervision to boost the segmentation performance. SSNet [42] jointly solves WSSS and SD using a single network, and makes full use of segmentation cues from saliency annotations to improve

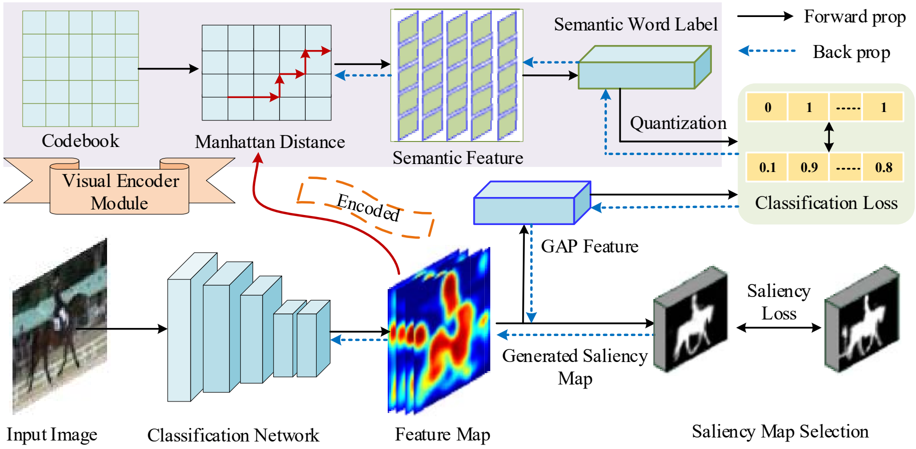

Fig. 1. Overview of the proposed method. We develop a visual encoder module to encode the feature map from the classification network into semantic words with a learnable codebook, covering more object regions. Moreover, we design a saliency map selection mechanism to separate the foreground from the background. The proposed model is jointly trained based on the classification loss and the saliency loss.

the segmentation performance. Different from these saliency-guided methods, our SMVWE method generates self-saliency maps using localization maps and utilizes off-the-shelf saliency maps as their pseudo-pixel feedback, while most existing methods directly use the off-the-shelf saliency map to guide the generation process of the pseudo labels, which is not benefit to tackle the co-occurring pixel problem.

3 Proposed Method

3.1 Motivation

Our SMVWE mainly focus on these two comprehensive information containing the target object location from the localization map and the boundary information from the saliency map. Firstly, we explore more fine-grained labels in the training procedure, namely semantic word labels, to supervise the classification network, making the network discover more semantic regions, thus the generated localization map could be more accurate for covering the object parts. Then, we employ the saliency map as pseudo-pixel feedback to the localization maps from both the foreground objects and the background. Next, we will explain how SMVWE can tackle the sparse object coverage, inaccurate object boundary and co-occurring pixel problems in image-level WSSS.

The image-level WSSS task is unable to directly use the semantic word labels, so we use an unsupervised way to generate their vector representations in each forward pass, i.e., each semantic word in a trainable codebook utilizes the manhattan distance to encode the feature maps from the classification network. In such a way, it alleviates the sparse object region problem, and improves the accuracy of the generated localization map.

To tackle the inaccurate object boundary and co-occurring pixel problems, we first use the $L+1$ localization maps encoded by the semantic word labels to generate the foreground object map and the background map, then these generated saliency map are evaluated by a saliency loss using off-the-shelf saliency maps, addressing the boundary mismatch and assigning the co-occurring pixels of non-target objects to the background. Thus, our method can better separate the foreground objects from the background.

Lastly, the objective function of SMVWE is formulated with three parts: two multi-label classification losses from semantic word labels and image-level labels respectively, and the saliency loss from the generation process of the saliency map. By jointly training the three objectives, we can combine the localization map encoded by semantic word labels with the saliency map to generate higher-quality pseudo labels.

3.2 Semantic Word Learning

The localization map generated from the classification network only covers the most discriminative extents of objects. The reason is that the goal of the classification network is essentially classification ability, not localization map generation. Thus, we propose a visual word encoder (VWE) module to enforce the classification network to cover integral object regions via the semantic word labels.

Due to only image-level labels in the WSSS task can be employed to annotate pixels in images, no extra labels are available. For this reason, we employ the codebook to encode the extracted convolutional feature map $\boldsymbol{M} \in R^{C\times H\times W}$ to specific semantic words, where C denotes the channels, W and H denote width and height, respectively. Then, the manhattan distance is used to measure the similarity between the pixel at position i in $\boldsymbol{M}$ and the j-th word in codebook $\boldsymbol{B} \in R^{N\times K}$, where N is the number of words and K is the feature dimension. The similarity matrix $\boldsymbol{D}$ can be formulated as below:

$$\boldsymbol{D}_{ij} = manhattan(\boldsymbol{M}_i, \boldsymbol{B}_j) = |\boldsymbol{M}_i - \boldsymbol{B}_j| \tag{1}$$

After obtained $\boldsymbol{D}$, we use $softmax$ to normalize row-wise, then computing the j-th word in codebook $\boldsymbol{B}$ represents the semantic probability of the i-th pixel in feature map $\boldsymbol{M}$.

$$\boldsymbol{P}_{ij} = softmax(\boldsymbol{D}_i) = \frac{exp(\boldsymbol{D}_{ij})}{\sum_{n=1}^{N} exp(\boldsymbol{D}_{in})} \tag{2}$$

The semantic word $\boldsymbol{Z}_i$ with the maximum probability will be denoted the semantic word label for $\boldsymbol{M}_i$, where the index of the maximum value in the i-th row of $\boldsymbol{P}_{ij}$ is denoted as:

$$\boldsymbol{Z}_i = argmax\boldsymbol{P}_{ij} \tag{3}$$

Then, we use a N-dimensional vector $\boldsymbol{z}^{word}$ to denote the semantic word label of the image $\boldsymbol{I}$, where $\boldsymbol{z}_j^{word} = 1$ if the j-th word is in $\boldsymbol{Z}$, and $\boldsymbol{z}_j^{word} = 0$, otherwise. $\boldsymbol{z}^{word}$ will make the classification network discover more semantic extents of target objects during the training procedure.

If employing the histogram distributions of each semantic word generated by counting their frequencies to represent the feature map, it will lead to non-continuities and make the training process intractable [28]. Thus, we compute the soft frequency of the j-th word by accumulating the probabilities in $\boldsymbol{P}$:

$$\boldsymbol{e}_j^{word} = \frac{1}{H \cdot W} \sum_{i=1}^{H \cdot W} \boldsymbol{P}_{ij} \tag{4}$$

where $\boldsymbol{e}_j^{word}$ denotes the appearance frequency of the j-th word in $\boldsymbol{M}$. As shown in Figure 1, $\boldsymbol{e}^{word}$ will model the mapping relations between semantic words and image-level labels. Moreover, inspired by [28], we will set the codebook $\boldsymbol{B}$ as a trainable parameter, which makes it could be learned automatically via the back propagated gradients.

3.3 Saliency Map Feedback

In WSSS, utilizing the saliency map is a common practice to better provide the information of object boundaries. Different from existing methods that make full use of the off-the-shelf saliency map as a part of their feature maps, our method generates the saliency maps using the foreground localization map and the background localization map, where the off-the-shelf saliency map is only used as the pseudo-pixel feedback by a saliency loss.

First, generating a foreground map $\boldsymbol{F}_{fg} \in R^{H \times W}$ by aggregating the localization maps of target objects, and performing the inversion of a background map $\boldsymbol{F}_{bg} \in R^{H \times W}$ generated by the background localization map to represent the foreground map. Then, we use $\boldsymbol{F}_{fg}$ and $\boldsymbol{F}_{bg}$ to generate the saliency map $\boldsymbol{F}_s$.

$$\boldsymbol{F}_s = (1 - \mu)\boldsymbol{F}_{fg} + \mu(1 - \boldsymbol{F}_{bg}) \tag{5}$$

where $\mu \in [0, 1]$ is a hyper-parameter to adjust a weighted sum of the foreground map and the inversion of the background map.

Moreover, our method also addresses the saliency bias during generating the foreground map and the background map. Because the saliency detection model obtains the saliency map via different datasets, the saliency bias is inevitable. Thus, we introduce an overlapping ratio strategy [42] between the localization map and the saliency map to address this issue, i.e., the i-th localization map $\boldsymbol{F}_i$ is overlapped with the groundtruth saliency map $F_s^{'}$ more than $\delta\%$, which is classified as the foreground, otherwise the background. The foreground map and the background map are represented as follows:

$$\boldsymbol{F}_{fg} = \sum_{i=1}^{L} z_i \cdot \boldsymbol{F}_i \cdot \mathbf{1}[\phi(\boldsymbol{F}_i, \boldsymbol{F}_s^{'}) > \delta] \tag{6}$$

$$\boldsymbol{F}_{bg} = \sum_{i=1}^{L} z_i \cdot \boldsymbol{F}_i \cdot \mathbf{1}[\phi(\boldsymbol{F}_i, \boldsymbol{F}_s^{'}) \leq \delta] \tag{7}$$

where $z_i \in R^L$ is the binary image-level label and $\phi(\boldsymbol{F}_i, \boldsymbol{F}_s^{'})$ is used to compute the overlapping ratio between $\boldsymbol{F}_i$ and $\boldsymbol{F}_s^{'}$. We first use $\boldsymbol{C}_i$ and $\boldsymbol{C}_s$ to represented the binarized maps corresponding to $\boldsymbol{F}_i$ and $\boldsymbol{F}_s^{'}$ respectively. For example, at the pixel Q in $\boldsymbol{F}$, $\boldsymbol{C}_N(Q) = 1$ if $\boldsymbol{F}_N(Q) > 0.5$; $\boldsymbol{C}_N(Q) = 0$, otherwise. Then, using $\phi(\boldsymbol{F}_i, \boldsymbol{F}_s^{'}) = |\boldsymbol{C}_i \cap \boldsymbol{C}_s| / |\boldsymbol{C}_i|$ to compute the overlapping ratio $\delta\%$ between $\boldsymbol{F}_i$ and $\boldsymbol{F}_s^{'}$.

3.4 Jointly Learning of Pseudo Label Generation

Our method generates the pseudo labels by two comprehensive information, i.e., semantic word encoding and saliency map, they respectively focus on different issues in WSSS task. To tackle sparse object region problem, we train the classification network on the localization map $\boldsymbol{M}$ through predicting the semantic word label $\boldsymbol{z}^{word}$, where the global average pooling is used to compute the semantic word score $\boldsymbol{s}^{word} = conv(\boldsymbol{f}^{gap}, \boldsymbol{W}^{word})$, and $\boldsymbol{W}^{word}$ denotes the weight matrix. We use the multi-label soft margin loss [29] to compute the classification loss for semantic words as follows:

$$L_{cls}(\boldsymbol{s}^{word}, \boldsymbol{z}^{word}) = \frac{1}{L} \sum_{i=1}^{L} [\boldsymbol{z}_i^{word} log \frac{exp(\boldsymbol{s}_i^{word})}{1 + exp(\boldsymbol{s}_i^{word})} + (1 - \boldsymbol{z}_i^{word}) log \frac{1}{1 + exp(\boldsymbol{s}_i^{word})}] \tag{8}$$

where $\boldsymbol{z}^{word}$ is obtained by Eq.3, L is the number of image classes.

To model the mapping relations between semantic words and image classes, we use an 1×1 conv layer with weight matrix $\boldsymbol{W}^{w2i}$ to transfer the semantic word frequency $\boldsymbol{e}^{word}$ into the class probability space, where the predicted score and the ground-truth image label are denoted by $\boldsymbol{p}^{w2i}$ and $\boldsymbol{z}^{img}$, respectively. Thus, the loss function $L_{cls}(\boldsymbol{p}^{w2i}, \boldsymbol{z}^{img})$ is formulated as the same form as Eq.8.

Then, we utilize the saliency map to tackle inaccurate object boundaries and co-occurring pixels, where the average pixel-level distance between the ground-truth saliency map $\boldsymbol{F}_s^{'}$ and the generated saliency map $\boldsymbol{F}_s$ is employed to calculate the saliency loss.

$$L_{sal} = \frac{1}{H \cdot W} \left\| \boldsymbol{F}_s^{'} - \boldsymbol{F}_s \right\|^2 \tag{9}$$

where $\boldsymbol{F}_s^{'}$ is obtained from the off-the-shelf saliency detection model PFAN [45] trained on DUTS dataset [33].

The overall loss of our proposed method is finally represented as the sum of the aforementioned loss terms.

$$L_{total} = L_{cls}(\boldsymbol{s}^{word}, \boldsymbol{z}^{word}) + L_{cls}(\boldsymbol{s}^{w2i}, \boldsymbol{z}^{img}) + L_{sal} \tag{10}$$

where L_{sal} mainly focuses on updating the parameters of L target object classes and one background class, while L_{cls} only evaluates the label prediction for L target object classes, excluding the background class.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Criteria. Following previous works [21,42], we evaluate the proposed method on the PASCAL VOC 2012 semantic segmentation benchmark [12]. PASCAL VOC 2012 consists of 21 classes, i.e., 20 foreground objects and the background. Following the common practice in semantic segmentation, we use the augmented training set with 10,582 images [16], validation set with 1,449 images and testing set with 1,456 images. For all experiments, the mean Intersection-over-Union (mIoU) is used as the evaluation criteria.

Implementation Details. The ResNet38 [39] is employed as the backbone network to extract feature maps. The classification network is trained via the SGD optimizer with a batch size of 4. Besides, we set the initial learning rate to 0.01 and decrease the learning rate every iteration with a polynomial decay strategy. The number of semantic words is set to 256. The images are randomly rescaled to 448×448. For the segmentation networks, we adopt DeepLab-LargeFOV (V1) [6] and DeepLab-ASPP (V2) [7], where VGG16 and ResNet101 are their backbone networks, i.e., VGG16 based DeepLab-V1 and DeepLab-V2, and ResNet101 based DeepLab-V1 and DeepLab-V2.

4.2 Ablation Study and Analysis

To validate the effectiveness of our proposed method, we conduct several experiments to analyze the effect of each component in the proposed method. For all experiments in this section, we adopt the DeepLab-V1 with VGG-16 as the segmentation network and measure the mIoU on the VOC 2012 validation set.

Dealing with Sparse Object Region

To validate whether the proposed VWE can cover more object regions in the input images reasonably, we compute the mIoU of the semantic word labels on the PASCAL VOC 2012 validation set. As shown in Table 1, it shows that the codebook can distinguish different semantic words reasonably and the proposed VWE can work effectively for encoding different objects of an image. Compared with existing methods, our VWE module can obtain higher performance on most objects for semantic segmentation, and brings an improvement of 0.9%

Table 1. Comparison with representative methods on the sparse object region problem. The best three results are in red, blue and green, respectively.

Method	**bkg**	**aero**	**bike**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**	**table**
AffinityNet [2]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2
MCOF [34]	87.0	78.4	29.4	68.0	44.0	67.3	80.3	74.1	82.2	21.1	70.7	28.2
SSNet [42]	90.0	77.4	37.5	80.7	61.6	67.9	81.8	69.0	83.7	13.6	79.4	23.3
SEAM [35]	88.8	68.5	33.3	85.7	40.4	67.3	78.9	76.3	81.9	29.1	75.5	48.1
CIAN [14]	88.2	79.5	32.6	75.7	56.8	72.1	85.3	72.9	81.7	27.6	73.3	39.8
Ours (VWE)	89.2	75.7	31.1	82.4	66.1	61.7	87.5	77.8	82.8	32.3	81.4	34.5
Ours (SMVWE)	90.8	77.9	31.6	89.4	56.9	57.8	86.4	77.9	82.9	32.3	76.9	52.5

Method	**dog**	**horse**	**mbk**	**person**	**plant**	**sheep**	**sofa**	**train**	**tv**	**mIoU**
AffinityNet [2]	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	58.4
MCOF [34]	73.2	71.5	67.2	53.0	47.7	74.5	32.4	71.0	45.8	60.3
SSNet [42]	78.0	75.3	71.4	68.1	35.2	78.2	32.5	75.5	48.0	63.3
SEAM [35]	79.9	73.8	71.4	75.2	48.9	79.8	40.9	58.2	53.0	64.5
CIAN [14]	76.4	77.0	74.9	66.8	46.6	81.0	29.1	60.4	53.3	64.3
Ours (VWE)	77.4	77.6	76.7	75.1	51.2	78.7	42.7	71.8	59.6	**65.4**
Ours (SMVWE)	80.7	80.3	81.8	74.3	44.5	80.7	54.7	68.8	60.5	**67.5**

Table 2. Comparison with representative methods on the inaccurate object boundary problem using the SBD set of the VOC 2012 validation set.

Method	**Recall (%)**	**Precision (%)**	**F1-score (%)**
CAM [46]	22.3	35.8	27.5
SEAM [35]	40.2	45.0	42.5
BES [9]	45.5	46.4	45.9
Our SMVWE	**62.3**	**76.5**	**69.4**

(65.4% vs 64.5%) compared to the state-of-the-art method [35]. Thus, under the supervision of the generated semantic word labels, our proposed method can cover more object extents, which effectively addresses the sparse object-region problem and improves the performance of the localization map.

Dealing with Inaccurate Boundary and Co-occurring Pixel

Inaccurate boundary problem. To evaluate the boundary quality of pseudo-labels, our method compares with representative methods [9,35,46] by using the SBD set of the VOC 2012 validation set, where the SBD set containing boundary annotations is benefit to test the boundary quality of pseudo labels through the Laplacian edge detector [9]. As shown in Table 2, we use the evaluation metrics of recall, precision, and F1-score to demonstrate that our method remarkably

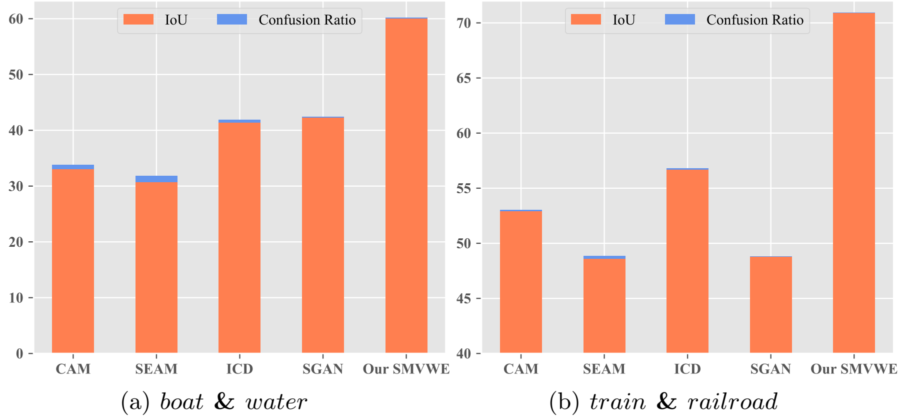

Fig. 2. Comparison with representative methods on the co-occurring pixel problem. The lower confusion ratio denotes the better, and the higher IoU denotes the better.

outperforms other methods. Figure 3 shows our some visualization results, which validate that our method works well on tackling the object boundary mismatch problem.

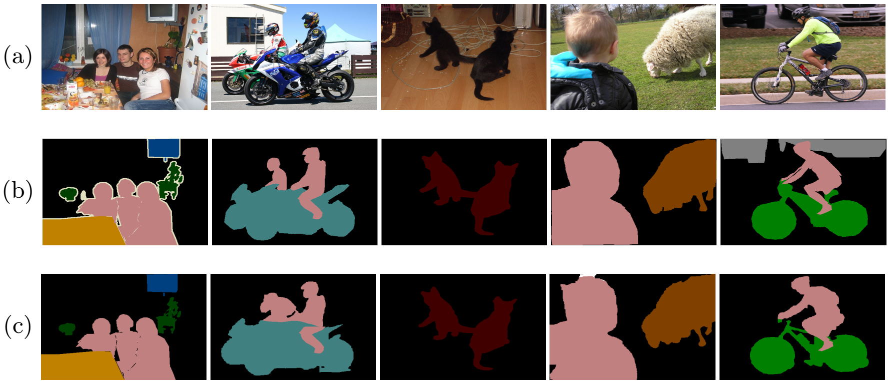

Fig. 3. Qualitative segmentation results on PASCAL VOC 2012 validation set. (a) Original images, (b) groundtruth and (c) our SMVWE. Segmentation results are predicted by ResNet101 based DeepLab-V2 segmentation network.

Co-occurring pixel problem. To measure the ability of our method on addressing the co-occurring pixels problem, we compare the performance of our method with representative methods (i.e., CAM [46], SEAM [35], ICD [13], SGAN [40]) by IoU and confusion ratio evaluation criteria, where the lower confusion ratio denotes the better, and the higher IoU denotes the better. The

Table 3. Performance comparisons of our method with state-of-the-art WSSS methods on PASCAL VOC 2012 dataset. All results are based on VGG16. *S* means the saliency map is used for existing methods and ours.

Methods	*S*	**val (%)**	**test (%)**
Segmentation Network : DeepLab-V1 (VGG-16)			
GAIN [25]	✓	55.3	56.8
MCOF [34]	✓	60.3	59.6
AffinityNet [2]	✗	58.4	60.5
SeeNet [18]	✓	61.1	60.8
OAA [20]	✓	63.1	62.8
RRM [43]	✗	60.7	61.0
ICD [13]	✓	64.0	63.9
BES [9]	✗	60.1	61.1
DRS[21]	✓	63.5	64.5
NSRM[41]	✓	65.5	65.3
Ours (SMVWE)	✓	67.5	67.2
Segmentation Network : DeepLab-V2 (VGG-16)			
DSRG [19]	✓	59.0	60.4
FickleNet [24]	✓	61.2	61.9
Split & Merge[44]	✓	63.7.	64.5
SGAN [40]	✓	64.2	65.0
Ours (SMVWE)	✓	68.2	68.1

IoU measures how much the target classes are predicted correctly, and the confusion ratio measures how much the co-occurring non-target class is incorrectly predicted as the target class.

As shown in Figure 2, we use two co-occurring pairs, i.e. *boat* with *water*, *train* with *railroad*, to compare our method with existing methods. Our method markedly outperforms other methods on the IoU evaluation criteria. Moreover, compared to other methods, only SGAN [40] method has a same lower confusion ratio with ours. For the following reasons, CAM [46] only captures the most discriminative region of target objects; SEAM [35] and ICD [13] both ignore the co-occurring pixels between target objects and non-target objects, while our method proposes a semantic word labels to discover more object regions, and designs a saliency map selection mechanism to obtain accurate object boundaries and discard the co-occurring pixels of non-target objects. Thus, our method generates higher-quality pseudo labels to perform the semantic segmentation task.

4.3 Comparison with State-of-the-arts

We compare our SMVWE method with state-of-the-art WSSS methods using only image-level labels. As shown in Table 4, our method remarkably outperforms other methods on the same VGG16 backbone. Noting that our performance improvement does not rely on a larger network structure and is superior to other

Table 4. Performance comparisons of our method with state-of-the-art WSSS methods on PASCAL VOC 2012 dataset. All results are based on ResNet101. S means the saliency map is used for existing methods and ours.

Methods	S	val (%)	test (%)
Segmentation Network : DeepLab-V1 (ResNet-101)			
MCOF [34]	✓	60.3	61.2
SeeNet [18]	✓	63.1	62.8
AffinityNet [2]	✗	61.7	63.7
FickleNet [24]	✓	64.9	65.3
OAA [20]	✓	65.2	65.2
RRM [43]	✗	66.3	65.5
ICD [13]	✓	67.8	68.0
DRS[21]	✓	66.5	67.5
Ours (SMVWE)	✓	70.1	69.6
Segmentation Network : DeepLab-V2 (ResNet-101)			
DSRG [19]	✓	61.4	63.2
BES [9]	✗	65.7	66.6
SGAN [40]	✓	67.1	67.2
DRS[21]	✓	70.4	70.7
Ours (SMVWE)	✓	71.3	71.5

existing methods based on a more powerful backbone (i.e. ResNet101 in Table 5). Because our method mainly relies on the cooperation of visual word encoder and saliency map selection strategy, which generates better pseudo labels for the semantic segmentation task. As shown in Table 5, our method achieves a new state-of-the-art performance (71.3% on validation set and 71.5% on test set) with the ResNet101 based DeepLab-V2 segmentation network. Figure 3 visualizes our semantic segmentation results on the validation set. These results show that our method can obtain more integral object regions and accurate object boundaries, and discard co-occurring pixels between target objects and the background.

5 Conclusion

In this paper, we proposed a saliency map and visual word encoder (SMVWE) method for image-level semantic segmentation. Particularly, we explored more fine-grained semantic word labels to supervise the classification network, making the generated localization map could cover more integral object regions. Moreover, we designed a saliency map selection mechanism to separate the foreground from the background, where the saliency maps were used as pseudo-pixel feedback. By joint learning of visual word encoder and saliency map feedback, our SMVWE successfully tackles the sparse object regions, boundary mismatch and co-occurring pixels problems, thus producing higher-quality pseudo labels for WSSS task. Extensive experiments demonstrate the superiority of our proposed method, and achieve the state-of-the-art performance using only image-level labels.

Acknowledgements This work was supported by the National Natural Science Foundation of China (62072463, 71531012), and the National Social Science Foundation of China (18ZDA309).

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: CVPR. pp. 2209–2218 (2019)
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: CVPR. pp. 4981–4990 (2018)
3. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: CVPR. pp. 4252–4261 (2020)
4. Chang, Y., Wang, Q., Hung, W., Piramuthu, R., Tsai, Y., Yang, M.: Weakly-supervised semantic segmentation via sub-category exploration. In: CVPR. pp. 8988–8997 (2020)
5. Chaudhry, A., Dokania, P.K., Torr, P.H.S.: Discovering class-specific pixels for weakly-supervised semantic segmentation. In: BMVC (2017)
6. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR (2015)
7. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
8. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. vol. 11211, pp. 833–851. Springer (2018)
9. Chen, L., Wu, W., Fu, C., Han, X., Zhang, Y.: Weakly supervised semantic segmentation with boundary exploration. In: ECCV. vol. 12371, pp. 347–362 (2020)
10. Choe, J., Lee, S., Shim, H.: Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **43**(12), 4256–4271 (2021)
11. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: CVPR. pp. 3130–3139 (2020)
12. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
13. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: CVPR. pp. 4282–4291 (2020)
14. Fan, J., Zhang, Z., Tan, T., Song, C., Xiao, J.: CIAN: cross-image affinity net for weakly supervised semantic segmentation. In: AAAI. pp. 10762–10769 (2020)
15. Fan, R., Hou, Q., Cheng, M., Yu, G., Martin, R.R., Hu, S.: Associating inter-image salient instances for weakly supervised semantic segmentation. In: ECCV. vol. 11213, pp. 371–388 (2018)
16. Hariharan, B., Arbelaez, P., Bourdev, L.D., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV. pp. 991–998 (2011)
17. Hong, S., Oh, J., Lee, H., Han, B.: Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In: CVPR. pp. 3204–3212 (2016)

18. Hou, Q., Jiang, P., Wei, Y., Cheng, M.: Self-erasing network for integral object attention. In: NeurIPS. pp. 547–557 (2018)
19. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: CVPR. pp. 7014–7023 (2018)
20. Jiang, P., Hou, Q., Cao, Y., Cheng, M., Wei, Y., Xiong, H.: Integral object mining via online attention accumulation. In: ICCV. pp. 2070–2079 (2019)
21. Kim, B., Han, S., Kim, J.: Discriminative region suppression for weakly-supervised semantic segmentation. In: AAAI. pp. 1754–1761 (2021)
22. Kim, D., Cho, D., Yoo, D.: Two-phase learning for weakly supervised object localization. In: ICCV. pp. 3554–3563 (2017)
23. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: ECCV. vol. 9908, pp. 695–711 (2016)
24. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: CVPR. pp. 5267–5276 (2019)
25. Li, K., Wu, Z., Peng, K., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: CVPR. pp. 9215–9223 (2018)
26. Oh, S.J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. In: CVPR. pp. 5038–5047 (2017)
27. Papandreou, G., Chen, L., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: ICCV. pp. 1742–1750 (2015)
28. Passalis, N., Tefas, A.: Learning bag-of-features pooling for deep convolutional neural networks. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 5766–5774 (2017)
29. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS. pp. 8024–8035 (2019)
30. Pinheiro, P.H.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR. pp. 1713–1721 (2015)
31. Roy, A., Todorovic, S.: Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In: CVPR. pp. 7282–7291 (2017)
32. Sun, G., Wang, W., Dai, J., Gool, L.V.: Mining cross-image semantics for weakly supervised semantic segmentation. In: ECCV. vol. 12347, pp. 347–365 (2020)
33. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR. pp. 3796–3805 (2017)
34. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: CVPR. pp. 1354–1362 (2018)
35. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: CVPR. pp. 12272–12281 (2020)
36. Wei, Y., Feng, J., Liang, X., Cheng, M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: CVPR. pp. 6488–6496 (2017)
37. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M., Feng, J., Zhao, Y., Yan, S.: STC: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2314–2320 (2017)

38. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In: CVPR. pp. 7268–7277 (2018)
39. Wu, Z., Shen, C., van den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognit. **90**, 119–133 (2019)
40. Yao, Q., Gong, X.: Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. IEEE Access **8**, 14413–14423 (2020)
41. Yao, Y., Chen, T., Xie, G., Zhang, C., Shen, F., Wu, Q., Tang, Z., Zhang, J.: Non-salient region object mining for weakly supervised semantic segmentation. In: CVPR. pp. 2623–2632 (2021)
42. Yu, Z., Zhuge, Y., Lu, H., Zhang, L.: Joint learning of saliency detection and weakly supervised semantic segmentation. In: ICCV. pp. 7222–7232 (2019)
43. Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K.: Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: AAAI. pp. 12765–12772 (2020)
44. Zhang, T., Lin, G., Liu, W., Cai, J., Kot, A.C.: Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In: ECCV. vol. 12367, pp. 663–679 (2020)
45. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: CVPR. pp. 3085–3094 (2019)
46. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)