Learning Visual Words for Weakly-Supervised Semantic Segmentation

Lixiang Ru, Bo Du, Chen Wu

Institute of Artificial Intelligence, School of Computer Science, Wuhan University

{rulixiang,dubo,chen.wu}@whu.edu.cn

https://github.com/rulixiang/vwe

WUHAN UNIVERSITY

IJCAI 2021 MONTREAL

Abstract

Prevailing Weakly-Supervised Semantic Segmentation (WSSS) methods using image-level labels, *i.e.* predicting pixel-level labels with only image-level supervision, usually train a classification network and generate the Class Activation Maps (CAMs) from the network as the initial coarse labels. However, CAMs typically only consist of **partial discriminative object extents** and some **unexpected background regions**, which are attributed to the image-level supervision and the widely-used global average pooling (GAP), respectively.

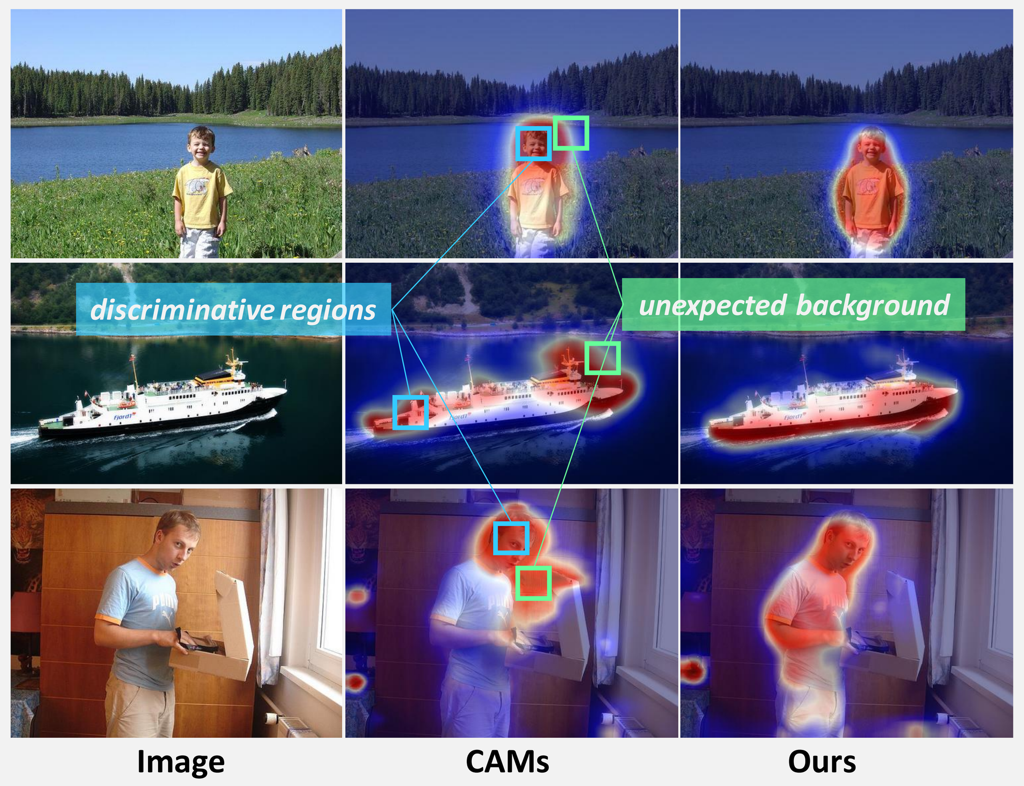


Figure: Illustration of the drawbacks of CAMs.

Contributions

The main contributions of this work are summarized as follows.

- We proposed to **learn and classify the local visual word labels**, which could enforce the network to discover more object extents and thus improve the quality of the generated pseudo pixel-level labels.
- We presented HSPP, a novel pooling method, which **averaged the local maximum and global average features** to alleviate the problem that the widely-used GAP and GMP can't estimate the objects accurately.
- We achieved **67.2%** and **67.3%** mIoU on the val and $test$ set of the PASCAL VOC 2012 dataset, which is the new state-of-the-art performance.

Method Overview

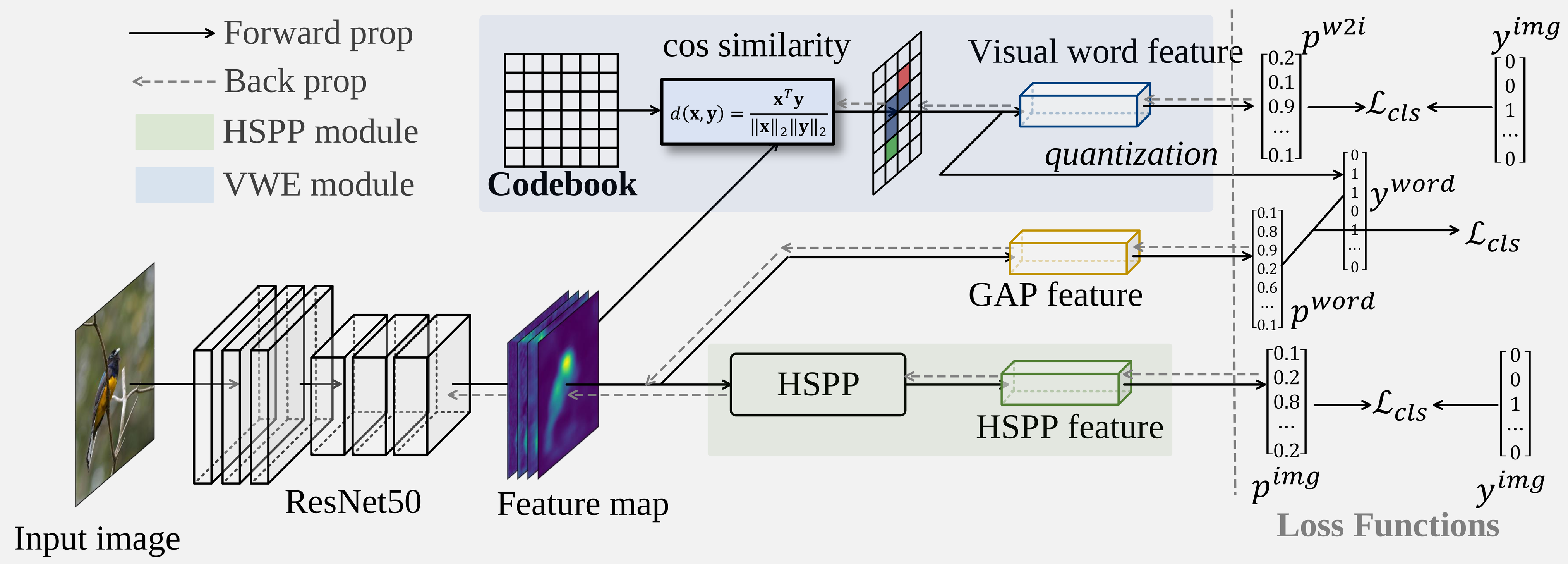


Figure: Overview of our proposed network.

To encourage the network to discover more object extents, we proposed a visual words learning module, which utilized a codebook to encode the feature maps extracted by CNN. The encoded visual word labels were then used to supervise the training process of the classification network. We also proposed a novel feature aggregation method, *i.e.* hybrid spatial pyramid pooling (HSPP), which incorporated GMP to reduce background information and GAP to ensure object completeness in the generated CAMs

Visual Words Learning

Given codebook $C \in \mathbb{R}^{k\times d}$ and feature map $F \in \mathbb{R}^{h\times w\times d}$, we use the $\cos$ distance to measure the their similarity:

$$S_{ij} = \frac{F_i^\top C_j}{||F_i||_2 ||C_j||_2}. \quad (1)$$

It's normalized row-wise using $softmax$ function:

$$P_{ij} = \frac{\exp(S_{ij})}{\sum_{n=1}^{k}\exp(S_{in})}. \quad (2)$$

The visual word label Y_i is the index of the maximum value in the i-th row of P_{ij}

$$Y_i = \arg\max_j P_{ij}\,. \quad (3)$$

The visual word labels are given as a k-dimensional vector y^{word}, where $y_j^{word} = 1$ if the j-th word is in Y, and $y_j^{word} = 0$ otherwise.

Hybrid Spatial Pyramid Pooling

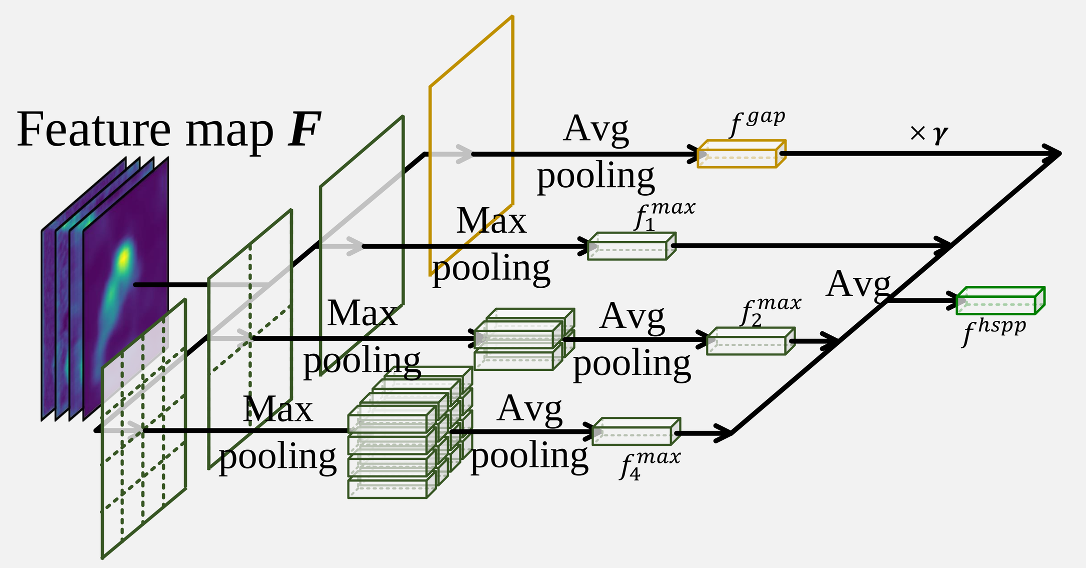


Figure: Illustration of the proposed HSPP.

The output of HSPP module is calculated by weighting the outputs of GAP and multi-scale max pooling,

$$f^{hspp} = \frac{1}{\gamma+3}\Big(\sum_{r\in\{1,2,4\}} f_r^{max} + \gamma f^{gap}\Big), \quad (4)$$

Loss Function

The overall loss of the proposed network is finally formulated as the sum of the aforementioned loss terms.

$$\mathcal{L} = \underbrace{\mathcal{L}_{cls}(p^{img}, y^{img})}_{\text{Learn image label}} + \underbrace{\mathcal{L}_{cls}(p^{word}, y^{word})}_{\text{Learn visual word label}} + \underbrace{\mathcal{L}_{cls}(p^{w2i}, y^{img})}_{\text{Learn image label with visual words}} \quad (5)$$

Quantitative Results

Method	Refinement	$train$	val
PSA CVPR'2018	-	48.0	46.8
IRNet CVPR'2019		48.3	-
SC-CAM CVPR'2020		50.9	49.6
Ours		**52.9**	**52.0**
IRNet CVPR'2019	+ IRNet	66.5	-
1Stage CVPR'2020		66.9	65.3
Ours		**67.7**	**65.7**

(a) Evaluation and comparison of the generated CAMs in mIoU.

Baseline	VWE	HSPP	$train$	val
✓			48.3	47.0
✓	✓		51.1	50.2
✓		✓	50.6	50.0
✓	✓	✓	**52.9**	**52.0**

(b) Ablation studies of our proposed methods on PASCAL VOC $train$ and val set.

	Sup	Backbone	val	$test$
WideResNet38	$\mathcal{F}$	WideResNet38	80.8	82.5
DeepLab		VGG16	69.8	-
DeepLabv2		ResNet101	76.3	77.6
AffinityNet CVPR'2018	$\mathcal{I}$	WideResNet38	61.7	63.7
IRNet CVPR'2019		ResNet50	63.5	64.8
SSDD ICCV'2019		WideResNet38	64.9	65.5
SC-CAM CVPR'2020		ResNet101	66.1	65.9
SEAM CVPR'2020		WideResNet38	64.5	65.7
BES ECCV'2020		ResNet101	65.7	66.6
MCIS ECCV'2020		ResNet101	66.2	66.9
Ours w/o CRF	$\mathcal{I}$	ResNet101	66.3	66.3
Ours w/ CRF		ResNet101	**67.2**	**67.3**

(c) Evaluation of the semantic segmentation results.

Qualitative Results

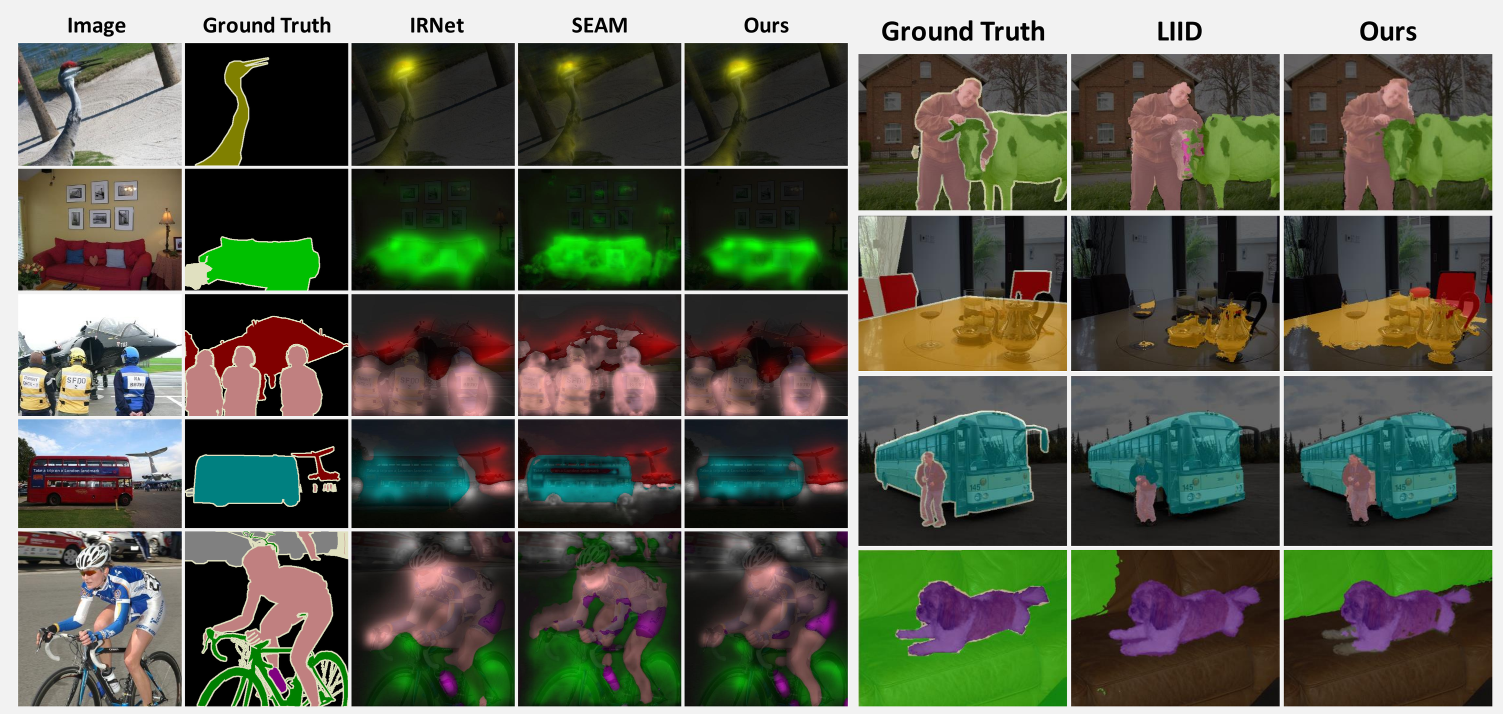
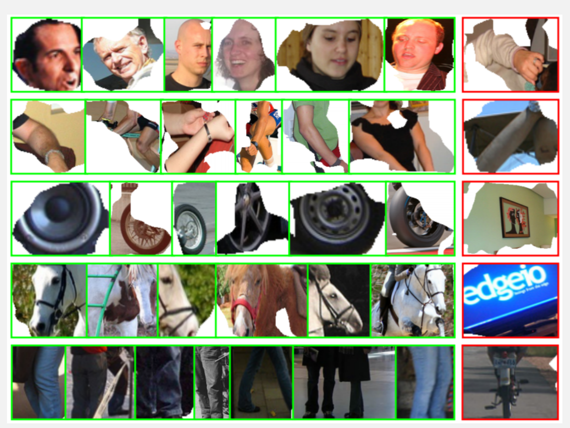


Figure: *Left*: Visualization results of the generated CAM. *Right*: The predicted semantic segmentation masks of the PASCAL VOC val dataset.

Visual Words in Codebook



This figure showed that the codebook could satisfactorily distinguish different visual words. We also observed that different parts of a visual object could be effectively encoded. For example, the visual words in Row 1, Row 2, and Row 5 could be roughly interpreted as $head$, arm, and leg of $person$, respectively.