# Learning Visual Words for Weakly-Supervised Semantic Segmentation



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 sole imagelevel supervision and aggregation of global features, respectively.



Image

Figure: Illustration of the drawbacks of CAMs.

# Contributions

The main contributions of this work are summarized as follows.

- We propose 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 CAMs.
- We present HSPP, a novel pooling method, which incorporates the local maximum and global average features to alleviate the problem of aggregation of global features.
- ► We achieve 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

To encourage the network to discover more object extents, we propose a visual words learning module, which utilizes a codebook to encode the feature maps extracted by CNN. The encoded visual word labels are then used to supervise the training process of the classification network. We also propose a novel feature aggregation method, *i.e.* hybrid spatial pyramid pooling (HSPP), which incorporates 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 the extracted feature map  $F \in \mathbb{R}^{h \times w \times d}$ , we use the  $\cos$  distance to measure their similarity:

It's normalized row-wise using *softmax*:

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

# Loss Function

The overall loss of the proposed network is formulated as

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



Figure: Overview of our proposed network.

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

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

$$Y_i = \arg\max_i P_{ij} \,. \tag{3}$$

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



Figure: Illustration of the proposed HSPP.

The output of HSPP module is calculated by weighting the outputs of GAP and multi-scale max pooling,  $f^{hspp} = \frac{1}{\gamma + 3}$ 

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

$$\frac{1}{3} \left( \sum_{r \in \{1,2,4\}} f_r^{max} + \gamma f^{gap} \right),$$
(4)

# Quantitative Results

Method	Refinement	train	val	_						
PSA CVPB'2018		48 0	46.8		=	Baseline	VWI	E HSPP	train	val
IRNet CVPB'2019		48.3	-0.0		-	$\frac{2}{}$	••••		48.3	47.0
SC-CAM CVPB'2020	_	50.9	49.6			$\checkmark$	$\checkmark$		51.1	50.2
Ours		52.9	52.0			$\checkmark$		$\checkmark$	50.6	50.0
IRNet CVPR'2019		66.5	_			$\checkmark$	$\checkmark$	$\checkmark$	52.9	52.0
1Stage CVPR'2020	+ IRNet	66.9	65.3	•	=	$(h) \Delta h la$	tion	studies		r
Ours		67.7	65.7	•		oronose	nd m	othode	on PA	
(a) Evaluation a	nd compar	ison	of th	= e				ethous and anal	on r cot	JUCAL
denerated CAM	s in mlol l			Ŭ			un a		501.	
generated OAM				0	Dealtha		7			
				Sup	Backbo		$\frac{val}{200}$	test		
	WideRes	Net38		_	WideRe	esinet38	80.8	82.5		
	DeepLab	)		${\cal F}$	VGG16		69.8	-		
	DeepLab	v2			ResNet	101	76.3	77.6		
	AffinityNe	et cvpr	2018		WideRe	esNet38	61.7	63.7		
	IRNet cvi	PR'2019			ResNet	50	63.5	64.8		
	SSDD ICC	CV'2019			WideRe	esNet38	64.9	65.5		
	SC-CAM	CVPR'2	020		ResNet	101	66.1	65.9		
	SEAM cv	'PR'2020	)	${\mathcal I}$	WideRe	esNet38	64.5	65.7		
	BES ECC	/'2020			ResNet	101	65.7	66.6		
	MCIS ECO	CV'2020			ResNet	101	66.2	66.9		
	Ours w/o	CRF			ResNet	101	66.3	66.3		
	Ours w/ (	CRF		$\mathcal{I}$	ResNet	101	67.2	67.3		

train		_					
<b>48</b> 0	46 P	3	Baseline	e VW	E HSPP	train	val
40.0 40.0						48.3	47 0
- 500 406		$\checkmark$	$\checkmark$		51.1	50.2	
52.9	52 C	)	$\checkmark$	v	$\checkmark$	50.6	50.0
66.5	-		$\checkmark$	$\checkmark$	$\checkmark$	52.9	52.0
66.9	65.3	3		otion	otudioo		r
67.7	65.7	7			SUUIES		
ricon (	nf th		propose	ea m	elliods		ISCAL
15011 (	ח וו	E	VOC tr	ain a	and val	set.	
		Sup	Backbone	val	test		
WideResNet38			WideResNet38	80.8	82.5		
DeepLab		$\mathcal{F}$	VGG16	69.8	-		
DeepLabv2			ResNet101	76.3	77.6		
AffinityNet CVPR'2018			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		$\mathcal{I}$	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		au	ResNet101	66.3	66.3		
Ours w/ CRF		1.					
	train 48.0 48.3 50.9 52.9 66.5 66.9 67.7 iSON ( 67.7 iSON ( 67.7 iSON ( 67.7 iSON ( 67.7	train val 48.0 46.8 48.3 - 50.9 49.6 52.9 52.0 66.5 - 66.9 65.3 67.7 65.7 67.7 65.7 ison of th SNet38 ov2 et CVPR'2018 PR'2019 CVPR'2020 V2020 CV'2020 CV'2020 CV'2020 CV'2020 CV'2020	train       val         48.0       46.8         48.3       -         50.9       49.6         52.9       52.0         66.5       -         66.9       65.3         67.7       65.7         ison of the       Jup         Net38       J         PR'2019       J         CVPR'2020       I         PR'2020       I         V/2020       I         CRF       J	$train val$ Baseline         48.0       46.8       Baseline         48.3       - $\checkmark$ 50.9       49.6 $\checkmark$ 52.9       52.0 $\checkmark$ 66.5       - $\checkmark$ 66.9       65.3       (b) Abla         67.7       65.7       propose         ison of the       Sup       Backbone         SNet38 $\mathcal{F}$ VGG16         v2       Prizo18       WideResNet38         PR'2019 $\mathcal{I}$ WideResNet38         CVPR'2020 $\mathcal{I}$ WideResNet38         PR'2020 $\mathcal{I}$ ResNet101         PR'2020 $\mathcal{I}$ WideResNet38         PR'2020 $\mathcal{I}$ ResNet101         PR'2020 $\mathcal{I}$ ResNet101         PR'2020 $\mathcal{I}$ <td>train       val         48.0       46.8         48.3       -         50.9       49.6         52.9       52.0         66.5       <math>\checkmark</math>         66.9       65.3         67.7       65.7         ison of the       Sup         SNet38       <math>\mathcal{F}</math>         VGG16       69.8         <math>\mathcal{F}</math>       VGG16         Not2       -         PR'2019       ResNet101         CVPR'2018       WideResNet38         VVCVR'2020       <math>\mathcal{I}</math>         VideResNet38       64.9         CVPR'2020       <math>\mathcal{I}</math>         VideResNet38       64.5         ResNet101       65.7         ResNet101       66.2         CR</td> <td><math>train val</math>       Baseline VWE HSPP         48.0       46.8       Baseline VWE HSPP         48.3       -       <math>\checkmark</math>         50.9       49.6       <math>\checkmark</math> <math>\checkmark</math>         52.9       52.0       <math>\checkmark</math> <math>\checkmark</math>         66.5       -       <math>\checkmark</math> <math>\checkmark</math>         66.9       65.3       (b) Ablation studies       proposed methods         67.7       65.7       proposed methods       VOC train and val         ison of the       WideResNet38       80.8       82.5         Net38       <math>\mathcal{F}</math>       VGG16       69.8       -         Net2       ResNet101       76.3       77.6         et cvPr2018       WideResNet38       61.7       63.7         PR2019       ResNet50       63.5       64.8         CVPR2020       <math>\mathcal{I}</math>       WideResNet38       64.9       65.5         CVPR2020       <math>\mathcal{I}</math>       WideResNet38       64.5       65.7         (2020       ResNet101       65.7       66.6       69.9         CRF       <math>\tau</math>       ResNet101       66.3       66.3</td> <td>train       val         48.0       46.8         48.3       -         50.9       49.6         52.9       52.0         <math>\checkmark</math> <math>\checkmark</math>         66.5       -         66.9       65.3         66.9       65.3         (b)       Ablation studies of ou         proposed methods on PA         VOC train and val set.         Son of the       VideResNet38         Sup       Backbone         V2       F         ResNet101       76.3         77.6       F         VQC train and val set.</td>	train       val         48.0       46.8         48.3       -         50.9       49.6         52.9       52.0         66.5 $\checkmark$ 66.9       65.3         67.7       65.7         ison of the       Sup         SNet38 $\mathcal{F}$ VGG16       69.8 $\mathcal{F}$ VGG16         Not2       -         PR'2019       ResNet101         CVPR'2018       WideResNet38         VVCVR'2020 $\mathcal{I}$ VideResNet38       64.9         CVPR'2020 $\mathcal{I}$ VideResNet38       64.5         ResNet101       65.7         ResNet101       66.2         CR	$train val$ Baseline VWE HSPP         48.0       46.8       Baseline VWE HSPP         48.3       - $\checkmark$ 50.9       49.6 $\checkmark$ $\checkmark$ 52.9       52.0 $\checkmark$ $\checkmark$ 66.5       - $\checkmark$ $\checkmark$ 66.9       65.3       (b) Ablation studies       proposed methods         67.7       65.7       proposed methods       VOC train and val         ison of the       WideResNet38       80.8       82.5         Net38 $\mathcal{F}$ VGG16       69.8       -         Net2       ResNet101       76.3       77.6         et cvPr2018       WideResNet38       61.7       63.7         PR2019       ResNet50       63.5       64.8         CVPR2020 $\mathcal{I}$ WideResNet38       64.9       65.5         CVPR2020 $\mathcal{I}$ WideResNet38       64.5       65.7         (2020       ResNet101       65.7       66.6       69.9         CRF $\tau$ ResNet101       66.3       66.3	train       val         48.0       46.8         48.3       -         50.9       49.6         52.9       52.0 $\checkmark$ $\checkmark$ 66.5       -         66.9       65.3         66.9       65.3         (b)       Ablation studies of ou         proposed methods on PA         VOC train and val set.         Son of the       VideResNet38         Sup       Backbone         V2       F         ResNet101       76.3         77.6       F         VQC train and val set.

(c) Evaluation of the semantic segmentation results.

# **Qualitative Results**



semantic segmentation masks of the PASCAL VOC val dataset.

# Visual Words in Codebook



(5)



Figure: *Left*: Visualization results of the generated CAM. *Right*: The predicted

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.