IJCAI 2021 MONTREAL

Learning Visual Words for Weakly-Supervised Semantic Segmentation

Lixiang Ru, Bo Du, Chen Wu

WUHAN UNIVERSITY

Multi-step pipeline for WSSS with image-level labels

Image

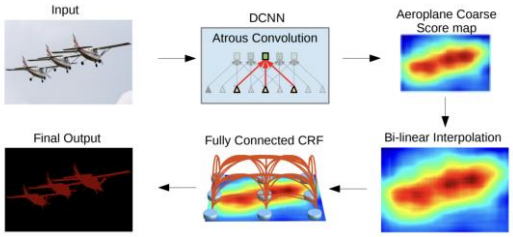
Classification Network

Class Activation Mapping [1]

CAMs

Refinement

Pseudo labels

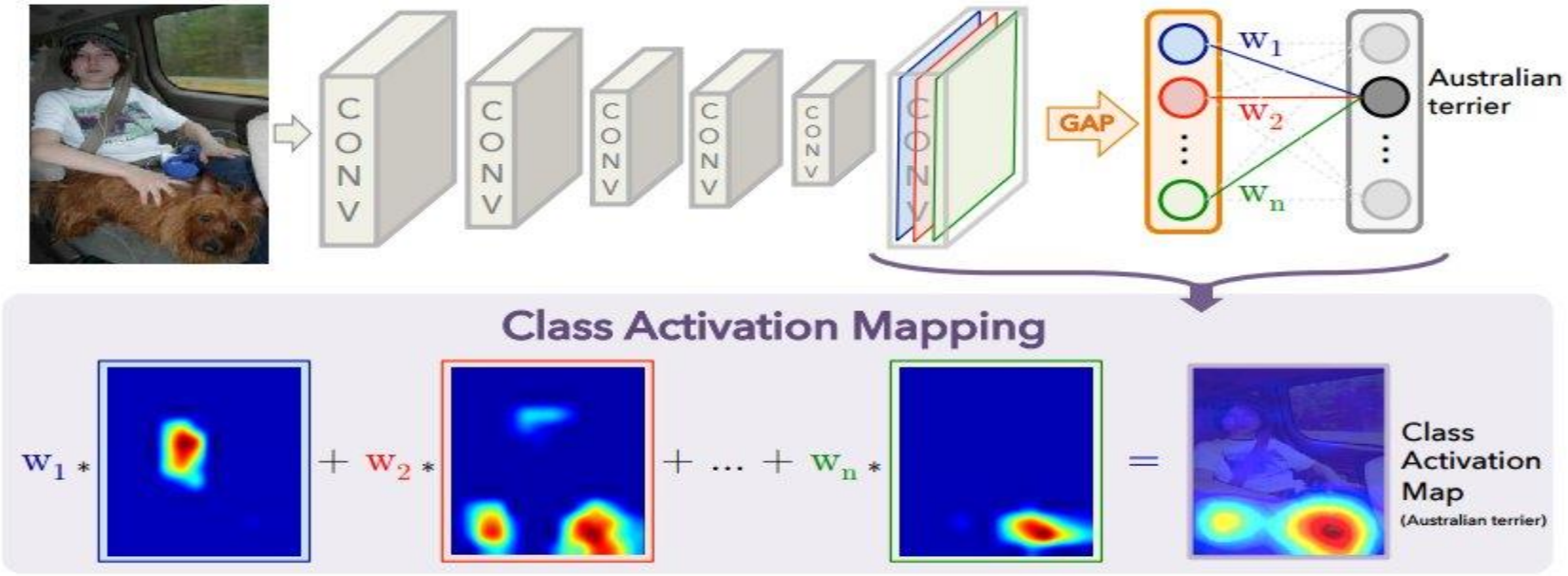


Segmentation model

[*Chen et al*. 2017]

[1] Zhou, Bolei, et al. "Learning deep features for discriminative localization." CVPR. 2016.

How to generate CAMs



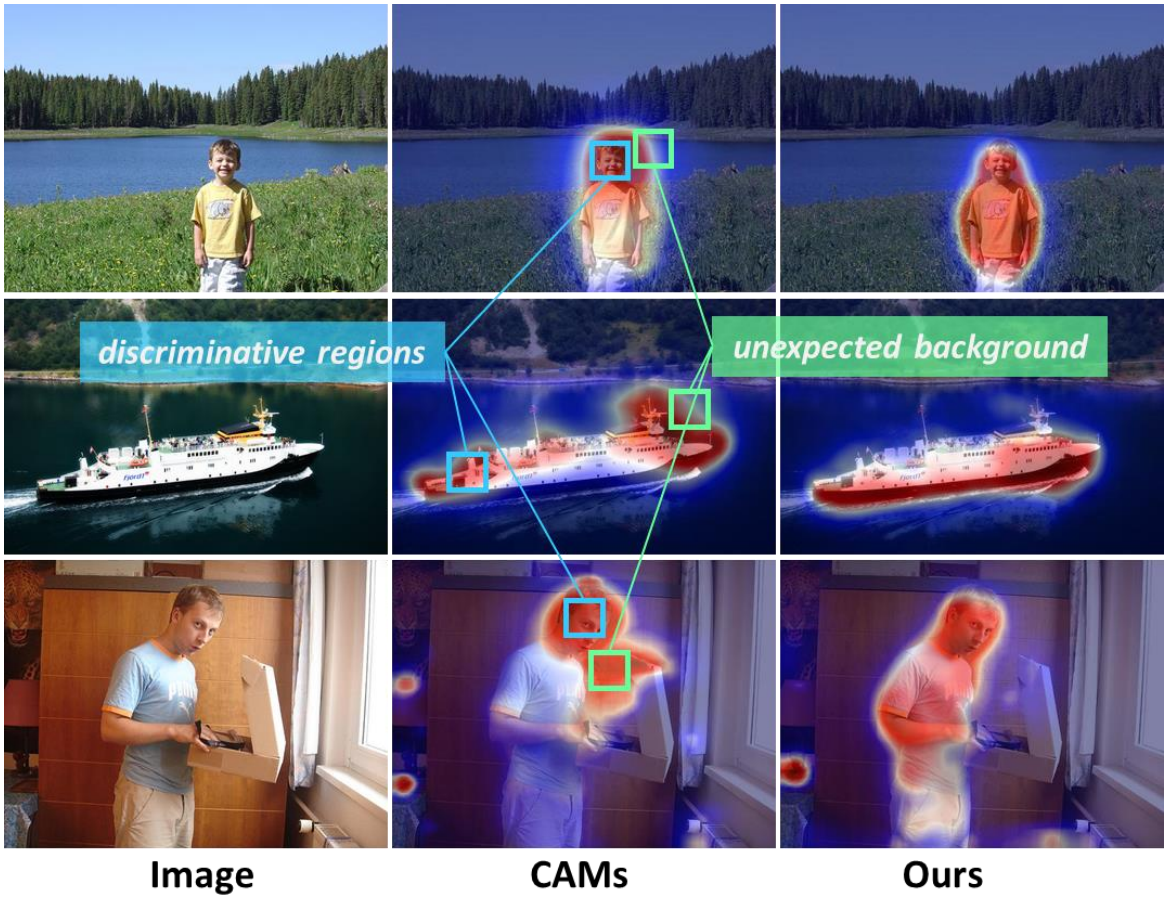
CAMs for class c are given by weighting each feature map with its contribution to class c.

[1] Zhou, Bolei, et al. "Learning deep features for discriminative localization." CVPR. 2016.

Drawbacks of CAMs

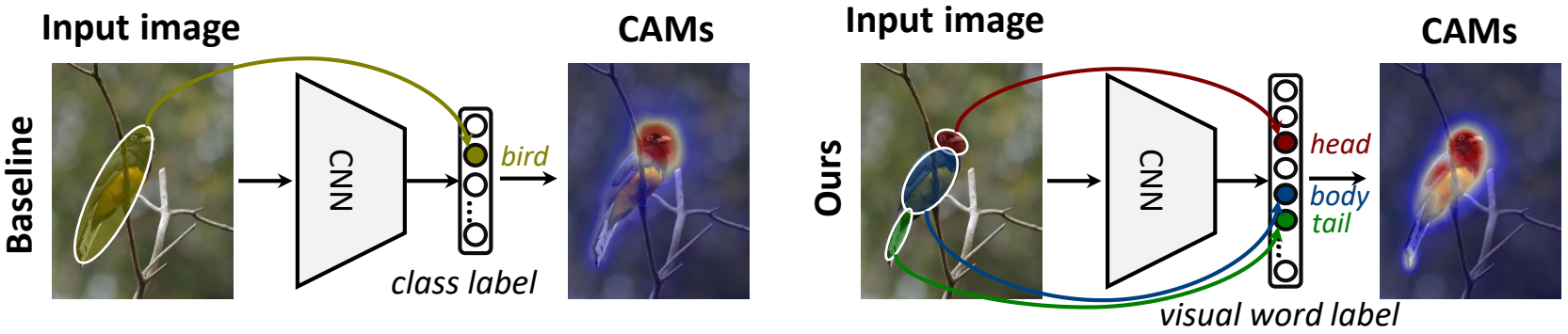
CAMs have 2 typical drawbacks

- usually only discover **partial discriminative regions**; ⟵ **Sole image-level supervision**
- often include some **undesired background**. ⟵ **Aggregation of global information**



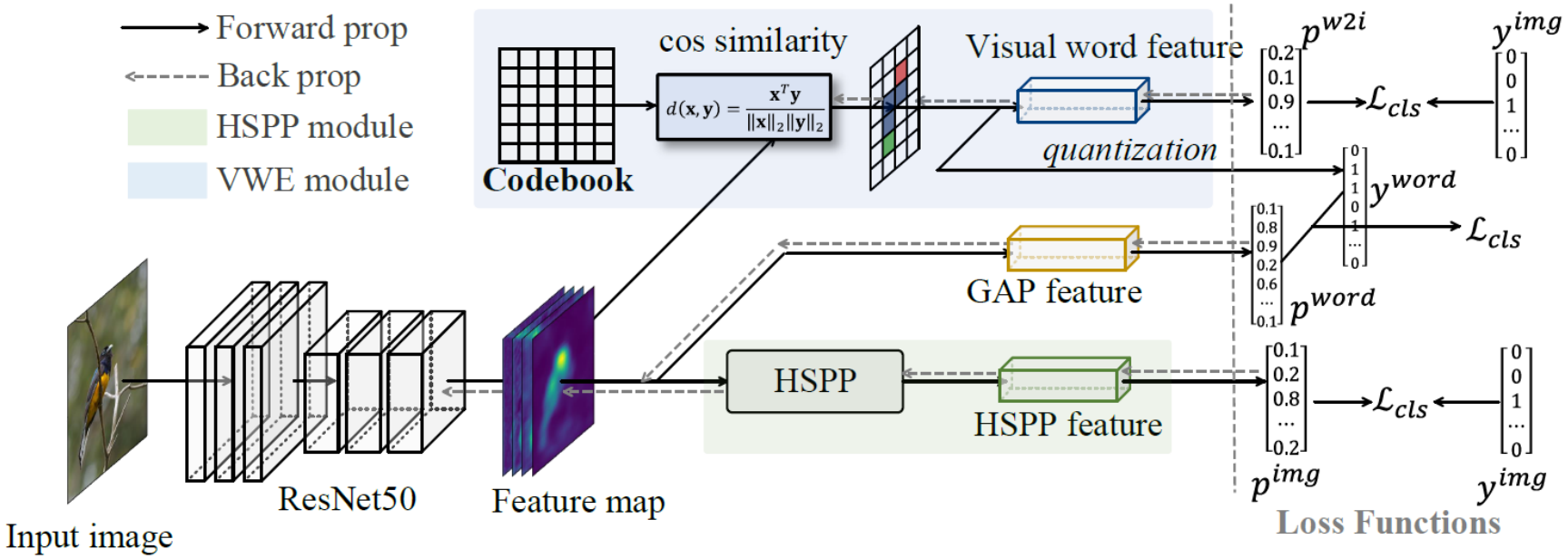
Motivation

- By enforcing the network to classify the **image-level label** and **visual word labels**, more object extents could be discovered.



- Aggregating the **global average** and **local maximums** of feature maps as output, more object extents and fewer background regions are preserved.

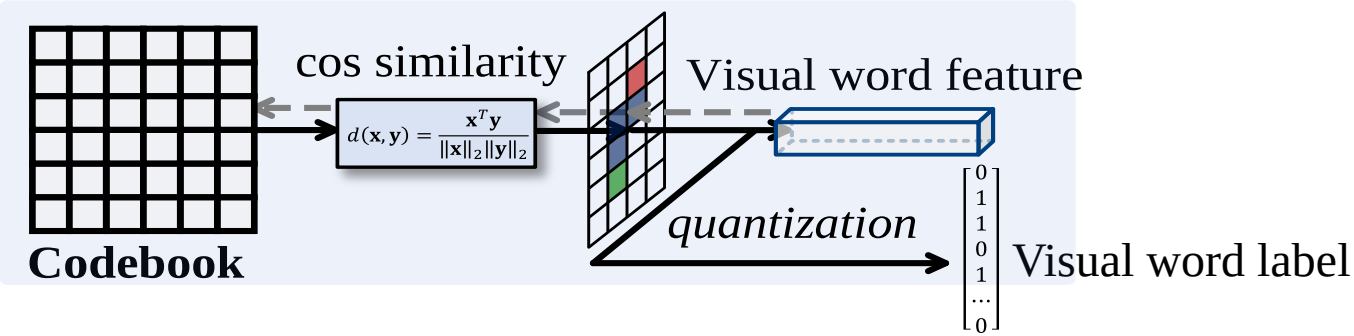
Overview



Our classification network for inferring CAMs includes:

- a CNN backbone to **extract convolutional feature maps**;
- a Visual Word Encoder module (VWE) to **encode visual word labels**;
- a hybrid spatial pyramid pooling (HSPP) module to **aggregate beneficial object information**.

Visual Words Encoder



- Given visual word codebook C and the extracted feature map F, visual word probability:

$$P_{ij} = \frac{\exp(S_{ij})}{\sum_{n=1}^{k}\exp(s_{in})}, with\ S_{ij} = \frac{F_i^T C_j}{\|F_i\|_2 \left\|C_j\right\|_2}.$$

- Visual word label is given as the word with the maximum probability:

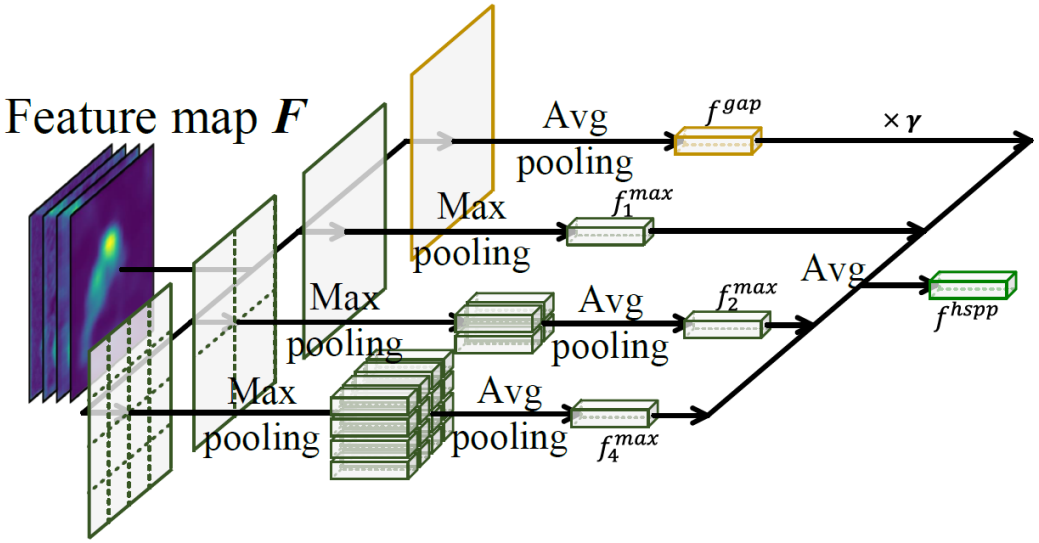
$$Y_i = argmax_j\ P_{ij}.$$

- For the input image X, its visual word label is a k-dimensional vector y^{word}, where $y_i^{word} = 1$ if the j-th word exists in Y and $y_i^{word} = 0$ otherwise.

- To learn the codebook from back-propagated gradients, we use the visual word frequency to predict the image-level label:

$$f_j^{word} = \frac{1}{hw}\sum_{i=1}^{hw} P_{ij}\,.$$

Hybrid spatial pyramid pooling



- Multi-scale local max-pooling to preserve local discriminative features:

$$f_r^{max} = \frac{1}{r^2}\sum_{i=1}^{r}\sum_{j=1}^{r} F_{i,j,:}^{max}, r = \{1,2,4\}.$$

- Global average pooling to ensure object completeness:

$$f^{gap} = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} F_{i,j,:}$$

- The outputs of hybrid spatial pyramid pooling (HSPP) is the weighted sum of them:

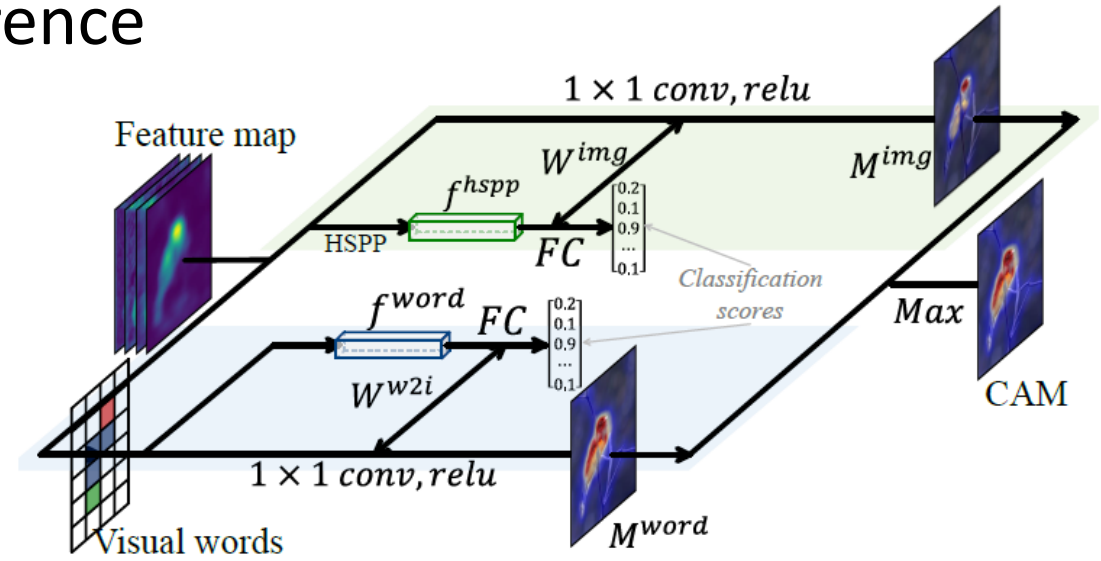
$$f^{hspp} = \frac{1}{\gamma+3}\left(\sum_{r\in\{1,2,4\}} f_r^{max} + \gamma f^{gap}\right)$$

Network training

- To reduce background information, we use the output feature of HSPP f^{hspp} to predict the image-level label;
- To encourage more object extents to be discovered, we also use the output feature to predict the encoded visual word labels;
- To learn the codebook for visual words encoding, we use the visual word features to predict the image-level label.
- The overall loss function is:

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

CAMs Inference



- We follow the original way to infer CAMs from the trained network:

$$M_c^{img} = \sum_{i=1}^{d} \left(W_{i,c}^{img} F_{:,:,i} \right)$$

- We also use the encoded visual word maps and learned mapping relation between visual words and image-level label to generate complementary information:

$$M_c^{word} = \sum_{i=1}^{d} \left(W_{i,c}^{w2i} P_{:,:,i} \right)$$

- The final CAMs is $M_c = \max(M_c^{img}, M_c^{word})$.

Experimental settings

Dataset

- Augmented PASCAL VOC 2012 dataset, with 10582 train, 1449 val, and 1456 test images.

Classification Network

- Backbone: ResNet50 pre-trained on ImageNet.
- Trained for 6 epochs with batch size of 16.

Refinement and Segmentation Network

- **Refinement** : IRNet [1] with default settings.
- **Segmentation Network**: DeepLabV2 [2] with ResNet101 as backbone and default settings.

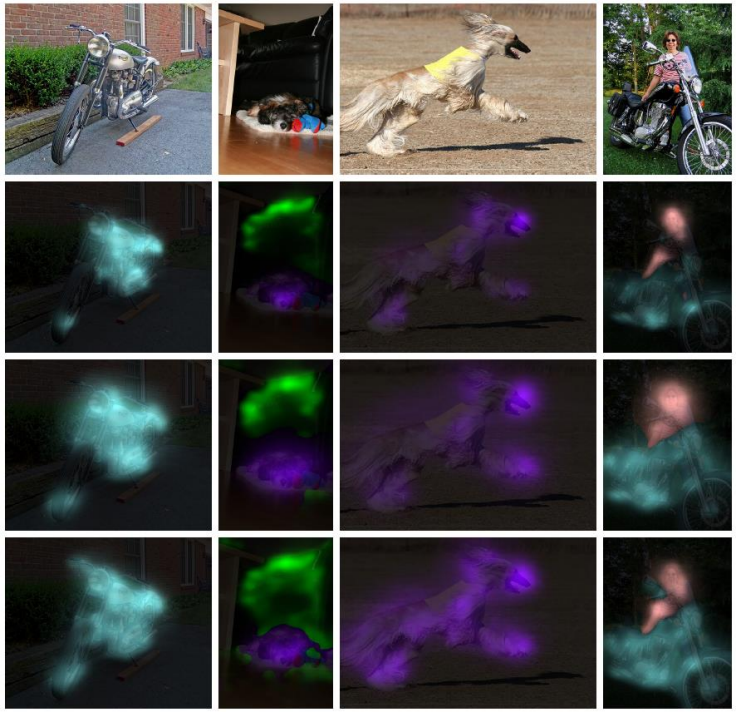
[1] Ahn et al. "Weakly supervised learning of instance segmentation with inter-pixel relations." CVPR. 2019.

[2] Chen et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE TPAMI, 2018.

Evaluation & Visualization of CAMs

Table 2: Evaluation and comparison of the generated pseudo labels in mIoU. The best results are highlighted in **bold**.

Method	Refinement	*train*	*val*
AffinityNet 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**



Semantic Segmentation Results

	Sup	Backbone	*val*	*test*
WideResNet38	$\mathcal{F}$	WideResNet38	80.8	82.5
DeepLabv2		ResNet101	76.3	77.6
BoxSup ICCV'2015	$\mathcal{B}$	VGG16	50.7	51.7
BCM CVPR'2019		VGG16	66.8	-
ScribbleSup CVPR'2016	$\mathcal{S}$	VGG16	63.1	-
SEC ECCV'2016	$\mathcal{I}$	VGG16	50.7	51.7
AffinityNet CVPR'2018		WideResNet38	61.7	63.7
DSRG CVPR'2018		ResNet101	61.4	63.2
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**

Ablation & Analysis

Table 1: Ablation studies of our proposed methods on the *train* and *val* set. Baseline: ResNet50. VWE: Visual Word Encoder. HSPP: Hybrid Spatial Pyramid Pooling. The best results are highlighted in **bold**.

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

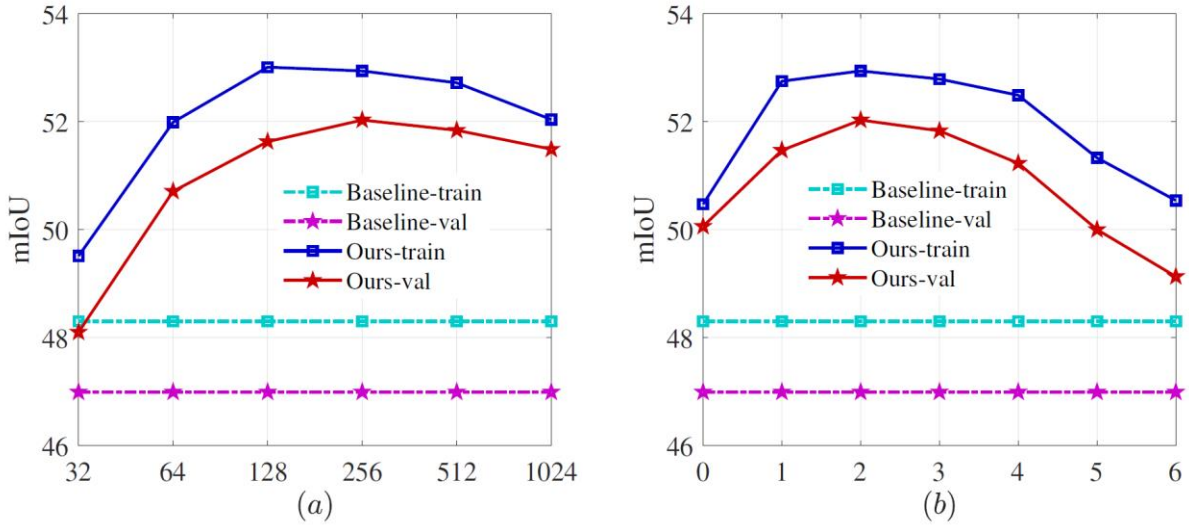


Figure 9: Impact of (a) the number of visual words k, (b) the weight factor γ in Eq 7.

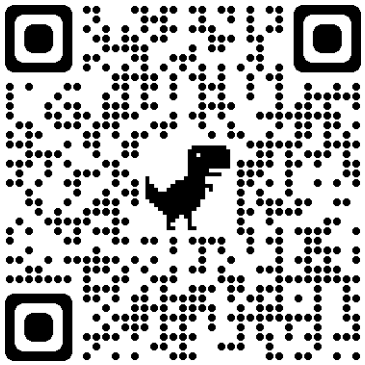
Visualization of Codebook

Figure 5: Samples of the learned words. In each row, images with green frame denote the dominant samples from this category, while images with red frame denote wrong words.

WUHAN UNIVERSITY

IJCAI 2021 MONTREAL

Thank you



Code available at *https://github.com/rulixiang/vwe*.