C^{2}AM: Contrastive learning of Class-agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation

Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, Linlin Shen*
School of Computer Science & Software Engineering, Shenzhen University, China
Dept. of Computer Science, Wenzhou-Kean University, Wenzhou, China
Shenzhen Institute of Artificial Intelligence of Robotics of Society, Shenzhen, China
Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China
xiejinheng2020@email.szu.edu.cn, hxianxu@gmail.com, llshen@szu.edu.cn

A. Appendix

A.1. Training Details.

C^{2}AM in WSOL. The input images are resized to 256×256 and then augmented by random cropping to 224×224. Horizontal flipping is also used for training data augmentation. SGD is adopted as the default optimizer. The learning rate, momentum, and weight decay are set to 0.0001, 0.9, and 0.0001, respectively. The cosine annealing policy is applied to schedule learning rate. The default batch size for training on CUB-200-2011 and ImageNet-1K datasets are 16 and 256, respectively. The models are trained for 20 and 5 epochs on CUB-200-2011 and ImageNet-1K datasets. The default α is 0.05. Following [9], the input images are resized to 448×448 to extract class-agnostic bounding boxes as pseudo labels.

C^{2}AM in WSSS. Following the common protocol in previous studies, we use an augmented training set of 10,582 images to train C^{2}AM on PASCAL VOC2012 dataset. The input images are first resized to 512×512 and then augmented by random cropping to 448×448 and horizontal flipping. SGD is adopted as the default optimizer. The learning rate, momentum, and weight decay are set to 0.0001, 0.9, and 0.0001, respectively. The cosine annealing policy is applied to schedule learning rate and batch size is 128. The default α is set to 0.25. In inference, a single image is flipped and resized with four different scales, e.g., $\{0.5,, 1.0, 1.5, 2.0\}$. The final class-agnostic activation maps are created by adding the activation maps of those eight images and normalizing them to $[0, 1]$.

All models are implemented in PyTorch [7] and trained on NVIDIA A100 GPU with 40 GB memory or NVIDIA TESLA V100 with 32 GB memory.

Table 1. Semantic segmentation performance (mIoU(%)) using different batch size. ††: moco is adopted to initialize the backbone network $h(\cdot)$ of C^{2}AM.

Method \ Batch size	32	64	128	256
PSA $_{\text{CVPR'18}}$ [1] + C^{2}AM††	63.7	64.4	**65.5**	64.1
SC-CAM $_{\text{CVPR'20}}$ [2] + C^{2}AM††	64.0	64.9	**66.0**	64.8
SEAM $_{\text{CVPR'20}}$ [8] + C^{2}AM††	61.9	62.8	**63.9**	63.4
PuzzleCAM $_{\text{ICIP'21}}$ [3] + C^{2}AM††	63.5	64.3	**65.5**	64.6
AdvCAM $_{\text{CVPR'21}}$ [5] + C^{2}AM††	63.4	64.5	**65.4**	64.8

A.2. Details of Background Cues Generation

As shown in the second row of Figure 1, while the predicted foreground activation maps can completely cover the foreground object regions, the estimated boundary of objects is still coarse. To address this issue, we apply dense CRF [4] to the foreground activation maps to refine the estimated boundary. It can be obviously seen in the third row that, the quality of estimated boundary, e.g., the feet of the goat and the edge of the boat, has been greatly improved. The background cues (shown in the third row) can be obtained in this way. However, it can be observed that applying dense CRF also introduces noises. To solve this problem, we use the background cues as pseudo masks to further train a model for better background cue predictions. As shown in the fourth row, the prediction of background from a trained model is more smooth and correct than that in the third row. Training details can be found in the open repository[1] of [6].

A.3. Additional Analysis

Effects of batch size. We vary the training batch size from 32 to 256 and show that, with a default α, an appropriate batch size gives us an appropriate number of confident

*Corresponding Author

[1]https://github.com/backseason/PoolNet

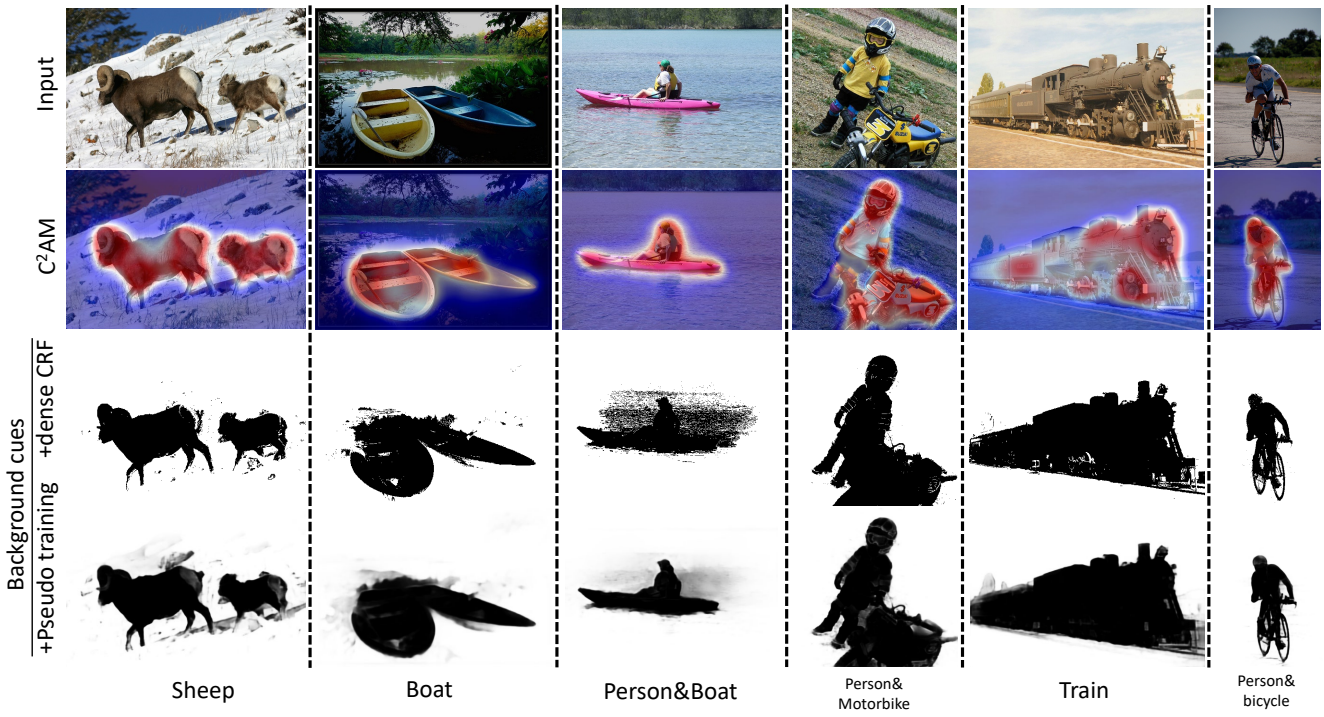

Figure 1. Illustration of the variation from class-agnostic maps to background cues. Best viewed in color.

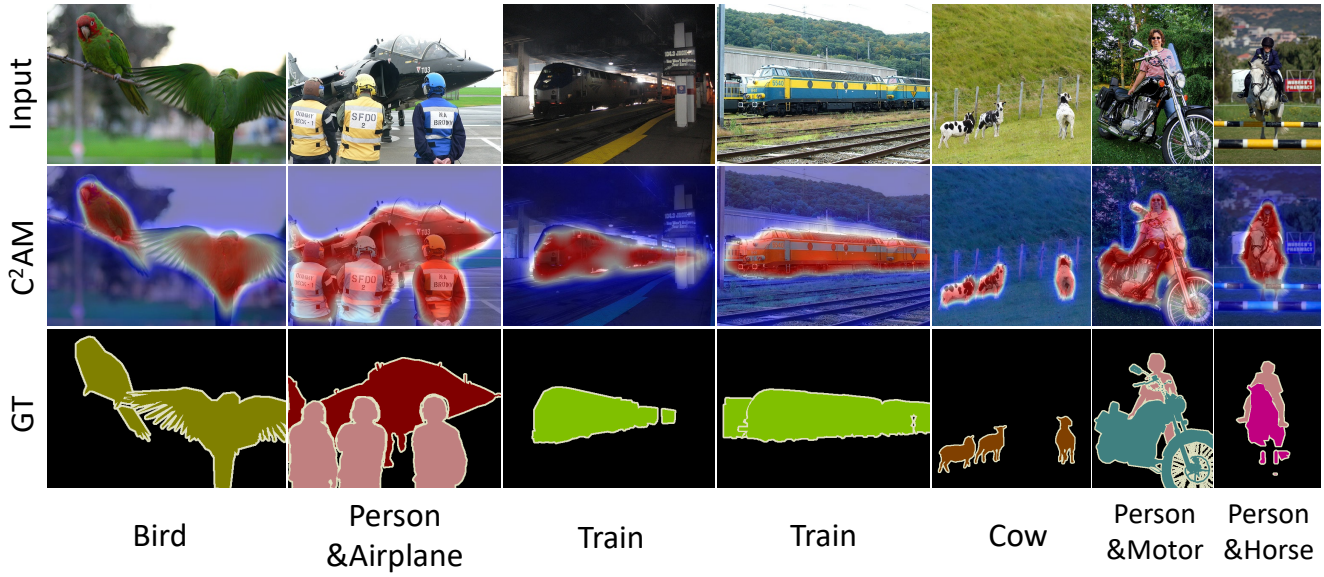

Figure 2. Visualization of the class-agnostic activation maps generated by C^2AM on PASCAL VOC2012 dataset. The images include classes of bird, person, airplane, train, cow, horse, and motor. Best viewed in color.

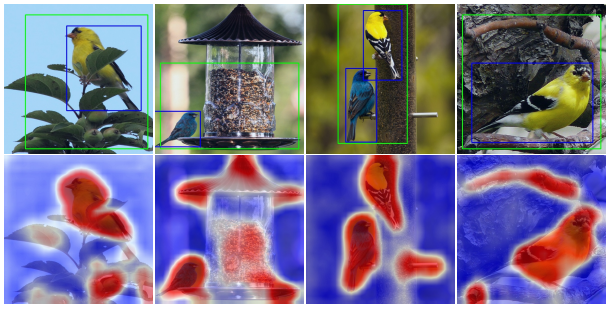

Figure 3. Failure cases in ImageNet-1K dataset.

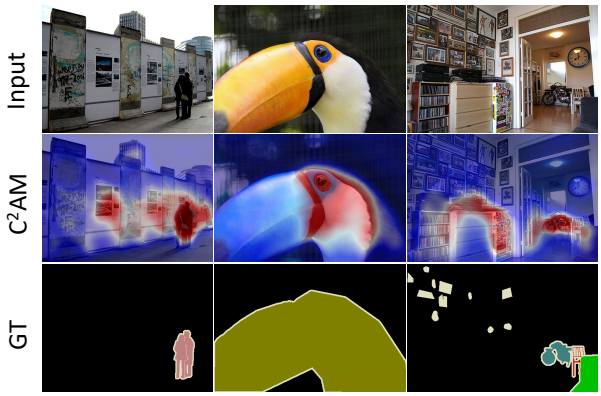

Figure 4. Failure cases in PASCAL VOC2012 dataset.

positive pairs, which leads to a better performance. The experiment is conducted on PASCAL VOC2012 dataset and the results are presented in Table 1. C^2AM trained with a batch size of 128 produced the best improvement over all of the WSSS methods in the table. When varying the batch size from 32 to 128, more samples are considered in the rank weighting to select more similar positive pairs for better contrastive learning. There is a minor drop in performance when the batch size is set to 256. It is because that, with a default α, e.g., 0.25, only enlarging the batch size will also include more dissimilar pairs with significant weights, which will affect the learning of network. Therefore, varying the batch size and α together may achieve better results. We leave it to the future works.

A.4. Visual Results

More visual results on ImageNet-1K and PASCAL VOC2012 datasets are presented in Figure 2 and 5, respectively. On ImageNet-1K dataset, it can be observed that, the foreground activation maps can successfully discriminate the foreground object regions, i.e., the full body of bird, fish, and turtle, etc, from the noisy backgrounds. On PASCAL VOC2012 dataset, the foreground activation maps also work well to separate multiple objects from the background regions. Besides, class closely related regions, e.g., railroad, are successfully discriminated as backgrounds, and excluded from the activated foreground regions, i.e., train.

A.5. Limitation

We also provide some failure cases in Figure 3 and 4. As shown in the first and fourth column of Figure 3, C^2AM successfully identifies the sky as background but falsely identifies the branches connected with the bird as foreground. In Figure 4, much more diverse background regions, e.g., the poster in the first column and the bookcase in the third column, are included in the foreground activation maps. These failure cases encourage us to improve our C^2AM in future works.

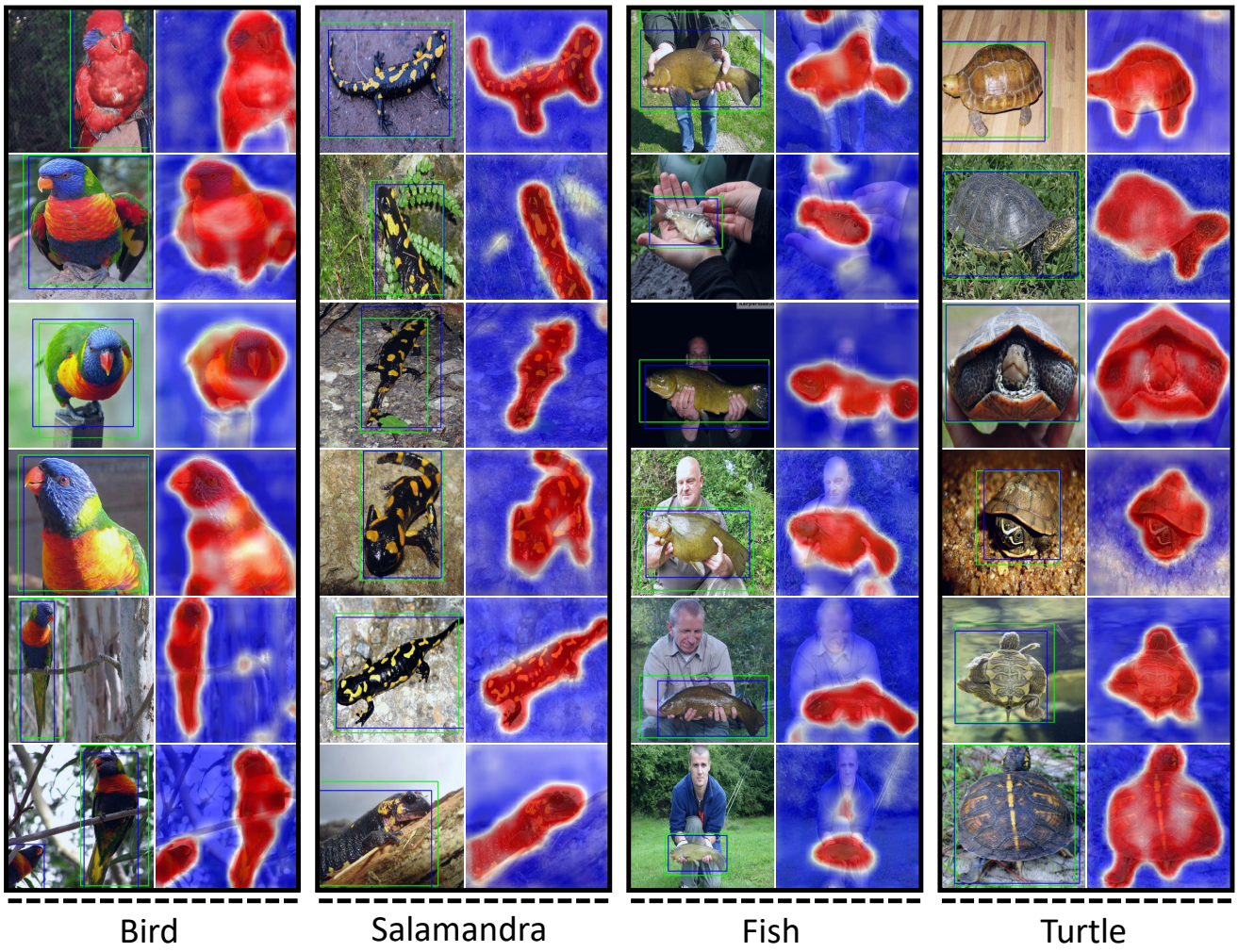

Figure 5. Visualization of the class-agnostic activation maps generated by C^2AM and the extracted class-agnostic bounding boxes on ImageNet-1K validation set. The images include classes of bird, salamandra, fish, and turtle. The ground truth and extracted bounding boxes are in blue and green, respectively. Best viewed in color.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, pages 4981–4990, 2018. 1

[2] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *CVPR*, pages 8988–8997, 2020. 1

[3] Sanghyun Jo and In-Jae Yu. Puzzle-cam: Improved localization via matching partial and full features. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 639–643, 2021. 1

[4] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011. 1

[5] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *CVPR*, pages 4071–4080, 2021. 1

[6] Jiangjiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019. 1

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, pages 8026–8037, 2019. 1

[8] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, pages 12275–12284, 2020. 1

[9] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *CVPR*, pages 13457–13466, 2020. 1