(Supplementary materials) Small Objects Matters in Weakly-supervised Semantic Segmentation

Cheolhyun Mun*
Samsung Research
Seoul, Korea
cheolhyunmun@yonsei.ac.kr

Sanghuk Lee*
SOCAR AI Research
Seoul, Korea
li-xh16@yonsei.ac.kr

Youngjung Uh
Yonsei University
Seoul, Korea
yj.uh@yonsei.ac.kr

Junsuk Choe
Sogang University
Seoul, Korea
jschoe@sogang.ac.kr

Hyeran Byun
Yonsei University
Seoul, Korea
hrbyun@yonsei.ac.kr

We provide the following supplementary materials in this appendix:

- In Sec. 1, we illustrate the distribution of each dataset (i.e., PASCAL VOC, MS COCO, and PASCAL-B) and the procedure of building PASCAL-B dataset thoroughly.
- In Sec. 2, we briefly explain each models which used for evaluation.
- In Sec. 3, we describe the implementation detail of each method we use.
- In Sec. 4, we give a concise explanation of elastic weight consolidation [4].
- In Sec. 5, we demonstrate the effectiveness of our proposed metric, dataset, and loss function with fully-supervised methods.
- In Sec. 6, we provide the qualitative results of each method on three datasets: PASCAL VOC, MS COCO, and PASCAL-B.

1. Dataset details

1.1. Number of instances per class per size

Fig. 1 shows the per-class per-size distribution of validation set for each dataset in detail. As shown in Fig. 1(a), PASCAL VOC 2012 [5] suffers from an imbalance problem in terms of class and size of instances. In particular, it has too many instances for the person class (i.e., 15th class) compared to the other classes. Some classes even do not have small instances. For PASCAL VOC, large instances account for 50% of the total number of instances while small instances only take 18.2%.

Secondly, MS COCO [10] also has a serious class imbalance problem with some categories (Fig. 1(b)). Additionally, it has imbalanced distribution in terms of instance size though the amount is less than PASCAL VOC. As in Table 1, the number of small instances makes up about 43.7% of the total instances while that of large instances is only 24.3%.

Different from these two datasets, PASCAL-B is the more balanced dataset. Fig. 1 (c) illustrates that our dataset alleviates the problems of class and size imbalance. In other words, PASCAL-B does not have the case that a specific class has too many instances and it has similar number of instances for all sizes as shown in Table 1.

Instance size	PASCAL VOC	MS COCO	PASCAL-B
Large	1,668 (**49.0%**)	65,407 (24.3%)	1,283 (32.1%)
Medium	1,118 (32.8%)	86,469 (32.1%)	1,468 (**36.7%**)
Small	621 (18.2%)	117,789 (**43.7%**)	1,245 (31.2%)
Total	3,407	269,665	3,996

Table 1. The number of instances by size for each dataset.

1.2. Process of constructing new dataset

Firstly, we collected images from the LVIS [6] which includes at least one of 20 categories of the PASCAL VOC classes. However, since `potted plant` class does not exist in the LVIS dataset, we collected images with `potted plant` class from MS COCO [10]. Then, we converted the annotations which do not belong to the 20 categories of the PASCAL VOC dataset into background

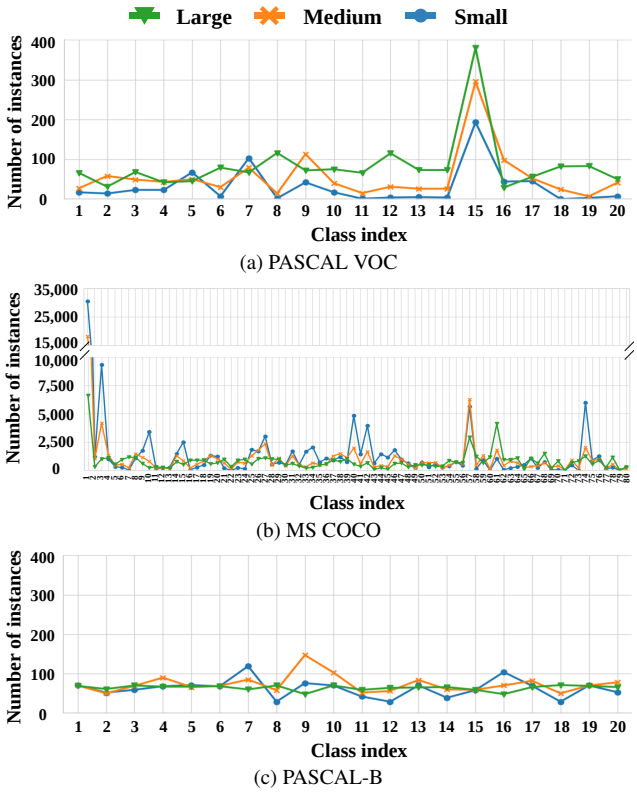

Figure 1. Dataset distribution. We plot the number of instances of each class by size.

class. After finishing the above process, 35,242 images remain. Among the remaining images, a few images have improper annotation as shown in Fig. 2. Therefore, two computer vision experts (authors of this paper) manually filtered out such images for two weeks and we had 15,263 images left. Finally, we randomly sampled images to ensure the balance over classes and object size distribution and constructed PASCAL-B which consists of 1,137 images with 20 classes. We give some sample images for the PASCAL-B dataset in Fig. 3.

2. Description for evaluated methods

We choose several methods with different weak-level supervision to validate the comprehensiveness of our metric and method.

Bounding box supervision: BBAM and BANA BBAM [8] utilizes the existing object detector Faster R-CNN [14] to highlight the regions where the detector concentrate on. They call these highlighted maps a bounding box attribution map. Then, they expand their bounding box attribution map by introducing a perturbation method. It distinguishes a small subset of the input image that leads to the same prediction as to the original image. Using perturbation methods, they try to diminish the useless information (i.e., background) for the detector.

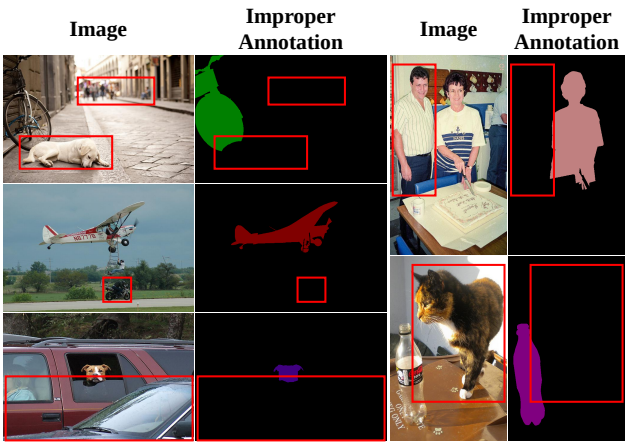

Figure 2. Example images with improper annotations. Red bounding boxes indicate missing annotations.

In BANA [12], Oh et al. find that the background regions around the bounding box are consistent. Based on the observation, they effectively distinguish the foreground and background regions in a bounding box by computing the cosine similarity between features in the bounding box and out of it. Additionally, they try to reduce the effect of noisy labels by utilizing the distances between CNN features and classifier weights.

Saliency supervision: EDAM, NS-ROM and RCA EDAM [17] separates the class-specific information from the whole activation map by applying L2-normalization along the channel dimension. Then it utilizes a self-attention mechanism to highlight similar regions among the series of class-specific activation maps. In the end, it enhances the results by using refined saliency maps with the threshold according to the value of the activation map.

NS-ROM [19] exploits the objects in non-salient regions. Therefore, they introduce a graph-based global reasoning unit to make the model learn global relations. Also, they filter out the background regions using saliency supervision, while capturing the objects outside the saliency map using class activation maps (CAMs). Finally, they enrich their pseudo masks by setting more ignore pixels to generate new pseudo masks after training the segmentation network. Then they train another segmentation network using new pseudo masks.

RCA [21] bridges the gap between image-level semantic information and pixel-level object regions by regional semantic contrast and aggregation. Regional semantic contrast leverages a memory bank to enforce the embedding of the pseudo region to get close to memory embedding of the same category while pushing away from other categories. Also, they utilize a non-parametric attention module called semantic aggregation. It aggregates memory representations

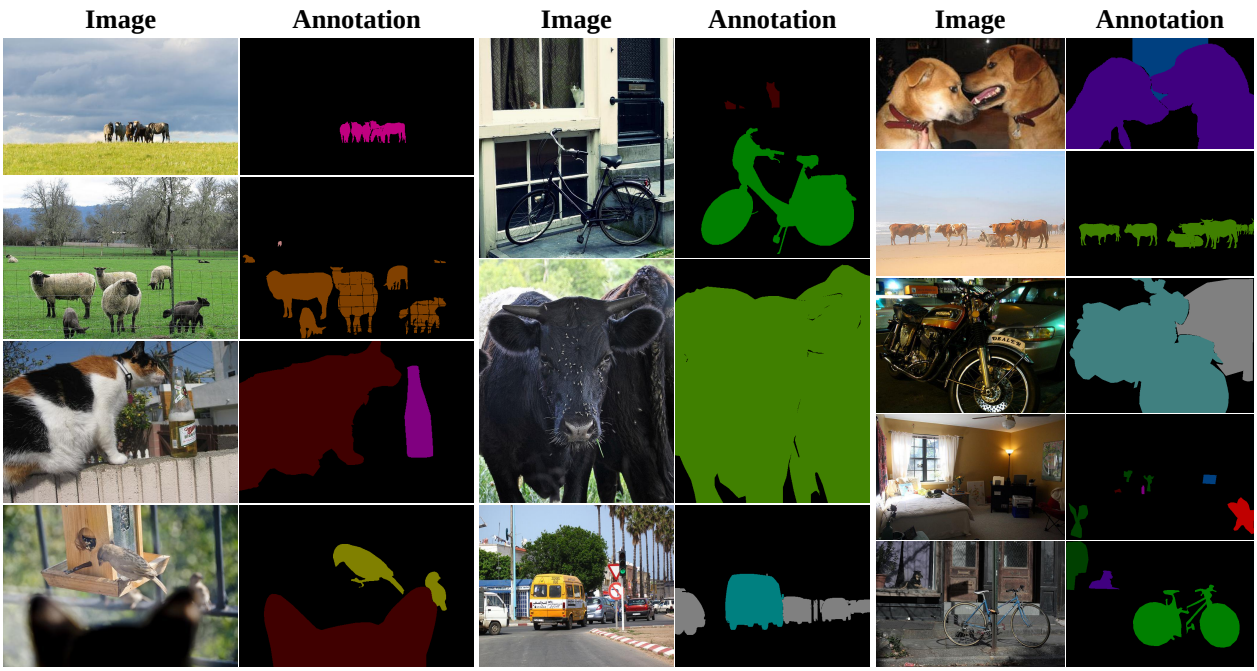

Figure 3. Sample image of PASCAL-B.

for each image and mines inter-image context to capture more informative dataset-level semantics.

Natural Language Supervision: CLIM CLIM [18] is built upon Contrastive Language-Image Pre-training (CLIP) [13]. Firstly, it additionally defines background classes for each image. Then, CLIM utilizes initial CAM to generate foreground masked-out image I_F and background masked-out image I_B. Lastly, using CLIP, it calculates the cosine similarity between these images and corresponding text category labels. For I_F, the similarity with a ground-truth label is maximized to gradually expand the activations for the whole foreground objects, while the similarity with the corresponding background label is minimized to decouple the foreground from the background. On the other hand, for I_B, the similarity with a ground-truth label is minimized to recover more probable foreground contents.

Image supervision: IRN, CDA, AMN and RIB IRN [1] predicts a displacement of each pixel pointing to the centroid to get the class agnostic map based on the rough semantic segmentation map from CAMs. By incorporating CAMs with a class-agnostic map, it obtains instance-wise CAMs and refines the prediction map by the random-walk algorithm.

CDA [16] is proposed to tackle the co-occurrence context information problem for WSSS. It first cuts some simple object instances using predicted segmentation masks by the trained network. Then it augments original images by pasting the obtained instances, and re-train the network with those augmented images.

The authors of AMN [9] raise an issue that global thresholding for CAM can lead to low-quality pseudo mask. To address this problem, they introduce new training objectives which apply per-pixel classification and label conditioning. Per-pixel classification makes discriminative part be reduced while expanding the non-discriminative part. Additionally, label conditioning is used to decrease the activation of non-target classes.

In RIB [7], Lee et al. argue that CAMs focus on the discriminative part because of the information bottleneck problem. The information bottleneck problem is that the only information highly related to tasks remains when the information goes backward of a layer in the network. According to the other works related to information bottleneck theory, it becomes worse with double-sided saturating activation functions such as softmax. Inspired by this, they propose to fine-tune the model with a one-sided saturating function to alleviate information bottleneck while expanding CAMs with global non-discriminative region pooling.

3. Implementation detail

All the experiment results of baseline methods [1, 7–9, 12, 16–19, 21] are reproduced by the official code, and we strictly follow the hyper-parameter settings provided by each paper. For the MS COCO dataset, we refer to the settings of RIB [7]. We set $\tau = 5$ for L_{sw} and $\lambda = 500$ for L_{sb} in all cases. For balanced training with our loss function, we train the segmentation networks for 30k iterations

for the PASCAL VOC dataset. We use pixel-wise cross-entropy loss for the first 20k, 15k, and 25k iterations, then fine-tune them with L_{sb} until the end of training models for BANA [12], EDAM [17], and others [1, 9, 16, 18, 19, 21], respectively. For the MS COCO dataset, the number of training iterations is 100k. We train the segmentation network with pixel-wise cross-entropy loss for the first 40K iterations, then fine-tune the network with L_{sb} for the remaining iterations. Note that we do not change all the other hyper-parameters of each baseline model.

All the experiments were done by one GeForce RTX 3090 GPU for PASCAL VOC and two RTX 3090 GPUs for MS COCO, which take 11 hours and 53 hours, respectively.

4. Elastic weight consolidation

Elastic Weight Consolidation (EWC) [4] is a technique for continual learning problem which tries to make the model learn various tasks. EWC aims to find the optimal point for the model to be optimized with several tasks. To achieve this goal, EWC constrains the parameters of the model which have a high correlation with the past training data. In other words, EWC suppresses the change of parameters based on its importance for the previous task. The loss function for EWC is defined as:

$$L_{total} = L_{now} + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2, \qquad (1)$$

where λ controls the importance of the previous task. It means that as the value of λ gets larger, it suppresses the updates of parameters more. F_i shows the importance of i-th parameter for the previous task. It indicates the correlation of parameters with past training data. In [4], it utilizes the diagonal elements of the Fisher information matrix. Lastly, $(\theta_i - \theta^*_{A,i})$ is the change of parameter between present model ($i.e.$, θ_i) and previous model ($i.e.$, $\theta^*_{A,i}$).

5. Extension to fully-supervised methods

In main paper, we demonstrate our evaluation metric, dataset, and loss function for weakly-supervised methods. However, they also can be applied in a fully-supervised manner. Table 2 reports the accuracy of fully-supervised methods [2, 3, 11, 15, 20] in terms of mIoU and IA-mIoU. It shows the same tendency as the experiment results of weakly-supervised methods except that the performances are generally more increased than the weakly-supervised methods when using our loss function.

6. Qualitative result

We show the visualization of prediction maps for each method [1, 3, 7–9, 12, 16–19, 21] on three datasets: PASCAL VOC (from Fig. 4 to Fig. 13), MS COCO (from Fig. 14 to Fig. 16), and PASCAL-B (from Fig. 17 to Fig. 26). Each figure shows that models with our loss function catch the objects more clearly including small-sized ones since our loss aims to constrain the network to be trained in balance considering the size of instances.

Dataset	PASCAL VOC		
Method	mIoU	IA-mIoU	IA_S
FCN [11]	67.8 (+0.8)	59.8 (+4.9)	17.1 (+7.9)
PSP [20]	76.7 (+0.6)	65.2 (+5.4)	22.1 (+13.2)
DeepLabV1 [2]	76.9 (+2.0)	65.6 (+6.4)	18.9 (+13.8))
DeepLabV2 [3]	77.8 (+0.6)	65.8 (+3.7)	18.8 (+5.6)
Segmentor [15]	79.9 (+0.6)	69.5 (+5.2)	24.1 (+16.7))
Dataset	PASCAL B		
Method	mIoU	IA-mIoU	IA_S
FCN [11]	56.6 (+1.2)	40.3 (+5.0)	10.1 (+5.5)
PSP [20]	63.3 (+0.1)	42.4 (+4.9)	13.4 (+6.3)
DeepLabV1 [2]	65.7 (+1.3)	45.4 (+5.8)	13.3 (+7.1)
DeepLabV2 [3]	66.6 (+1.3)	46.2 (+3.2)	15.6 (+4.2)
Segmentor [15]	67.9 (−0.2)	45.9 (+4.9)	13.1 (+7.6)

Table 2. Experimental results of fully-supervised method for PASCAL VOC and PASCAL-B.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 3, 4

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 4

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 4

[4] Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017. 1, 4

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[7] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 4

[8] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of*

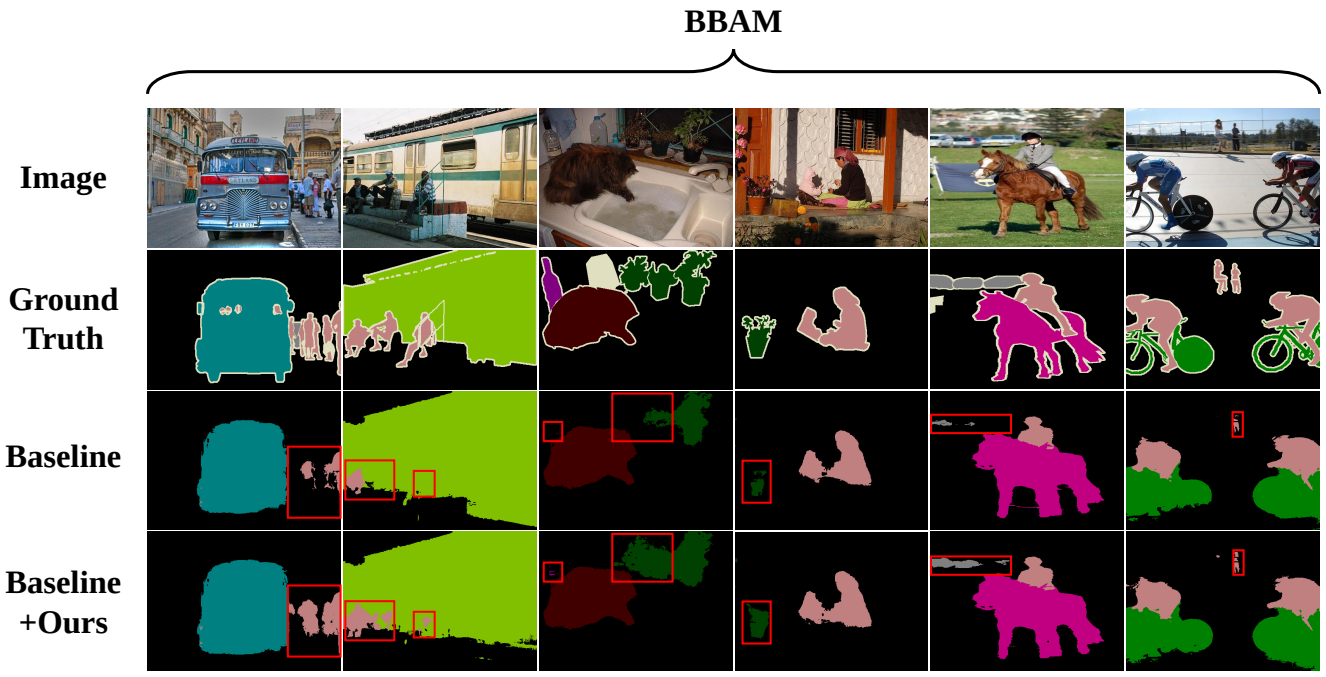

Figure 4. Visualization of BBAM on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

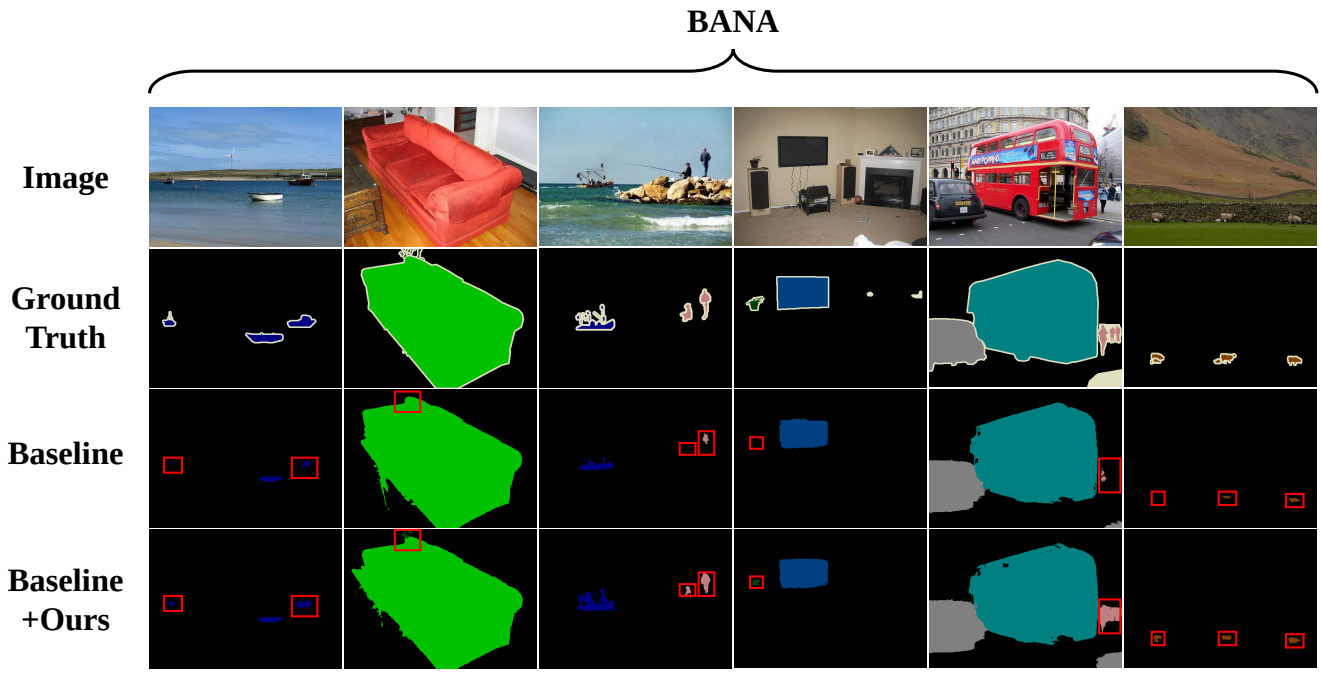

Figure 5. Visualization of BANA on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

the IEEE/CVF conference on computer vision and pattern recognition, pages 2643–2652, 2021. 2, 3, 4

[9] Minhyun Lee, Dongseob Kim, and Hyunjung Shim. Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

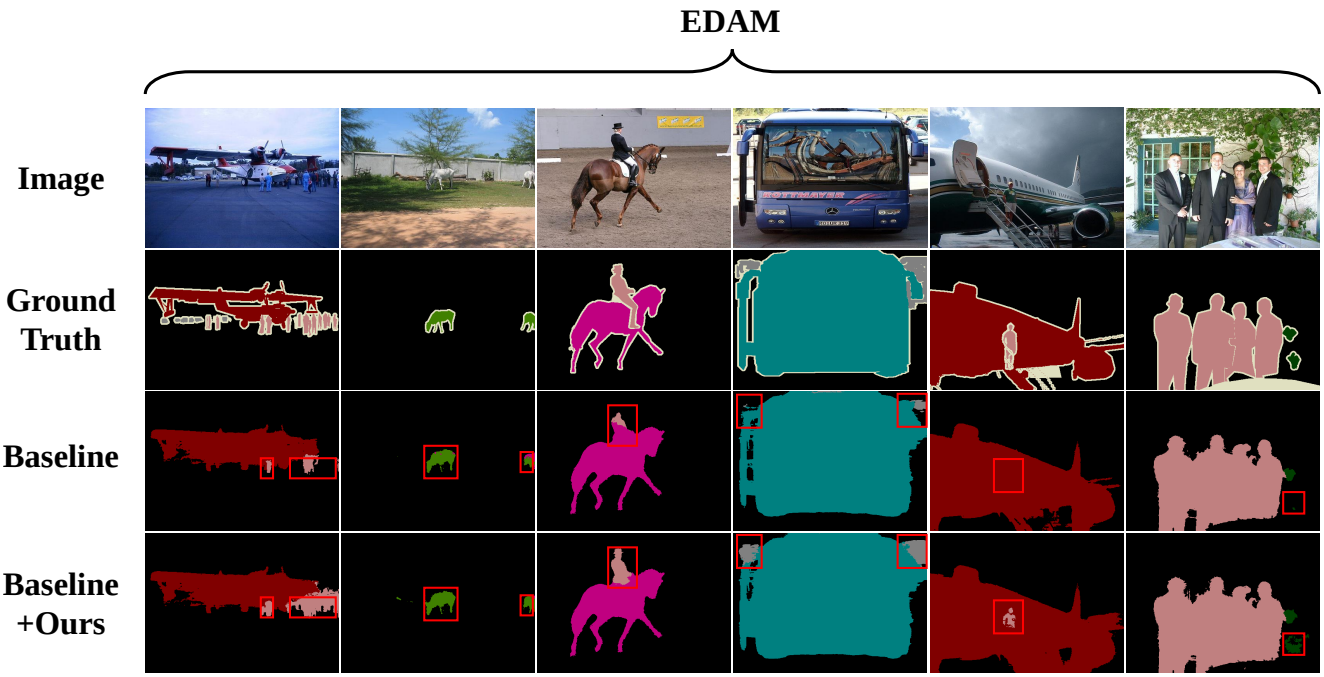

Figure 6. Visualization of EDAM on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

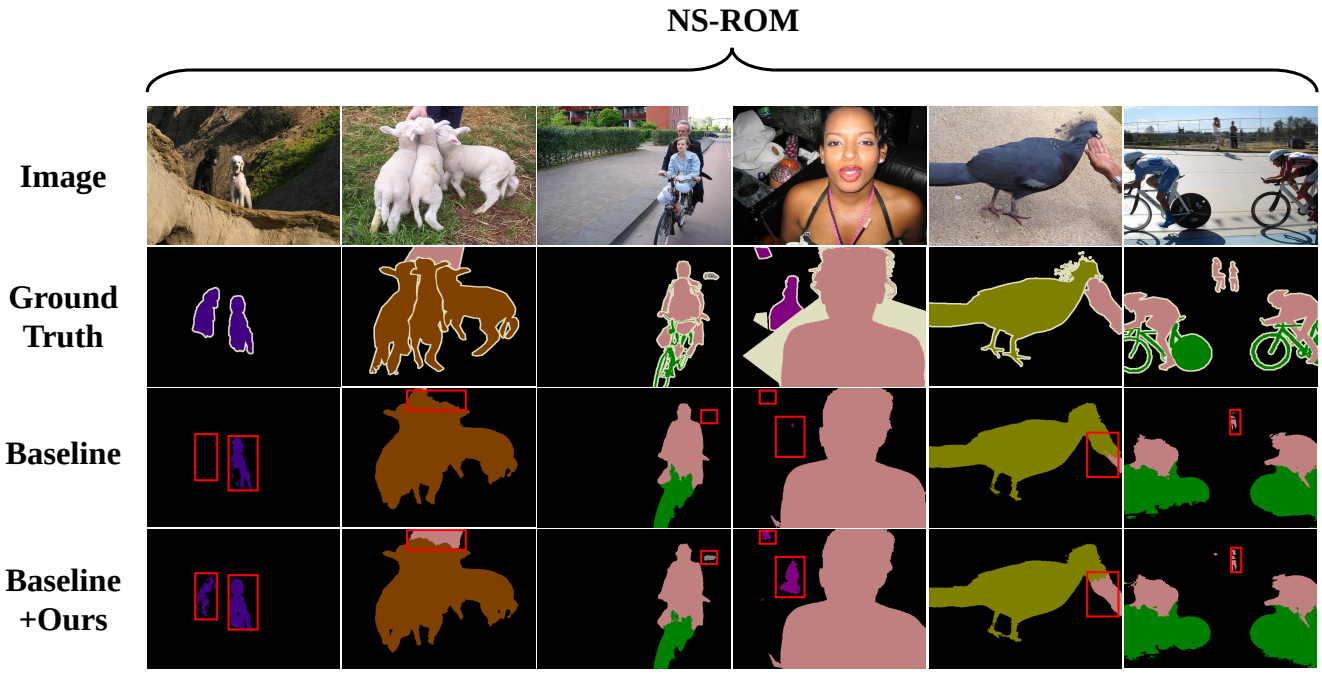

Figure 7. Visualization of NS-ROM on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

sion and Pattern Recognition, pages 4330–4339, 2022. 3, 4

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1

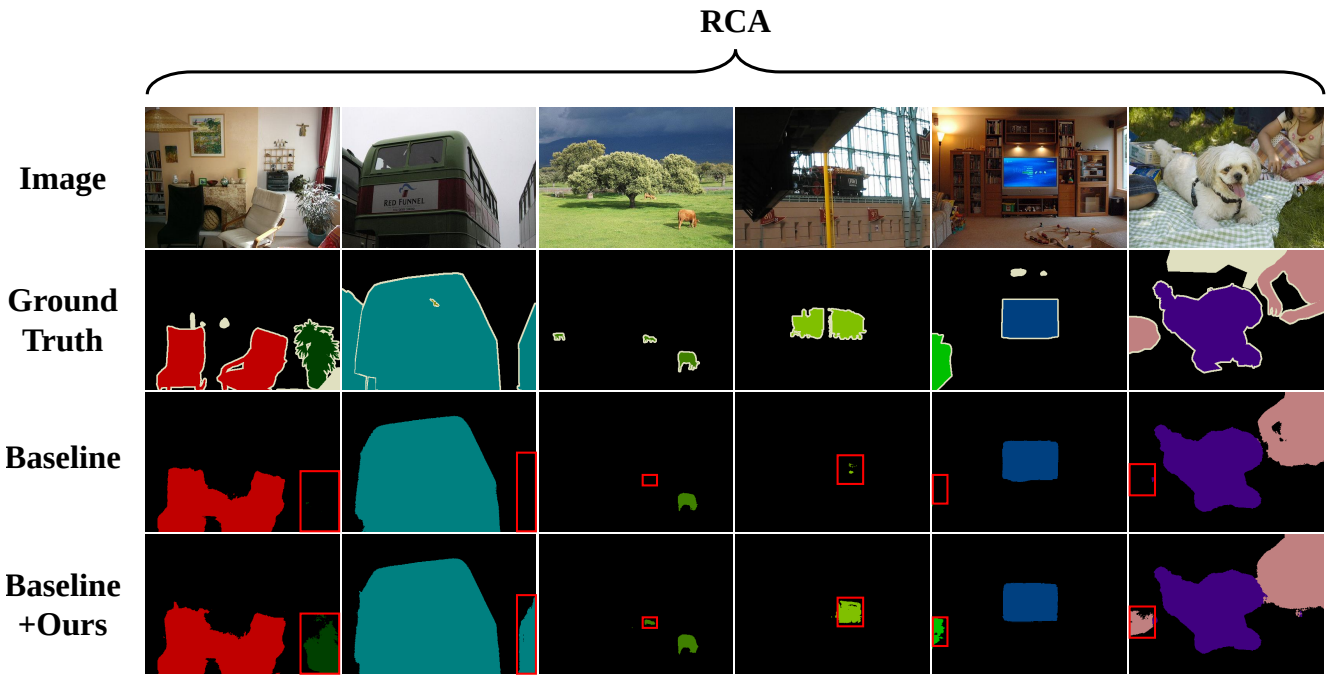

Figure 8. Visualization of RCA on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

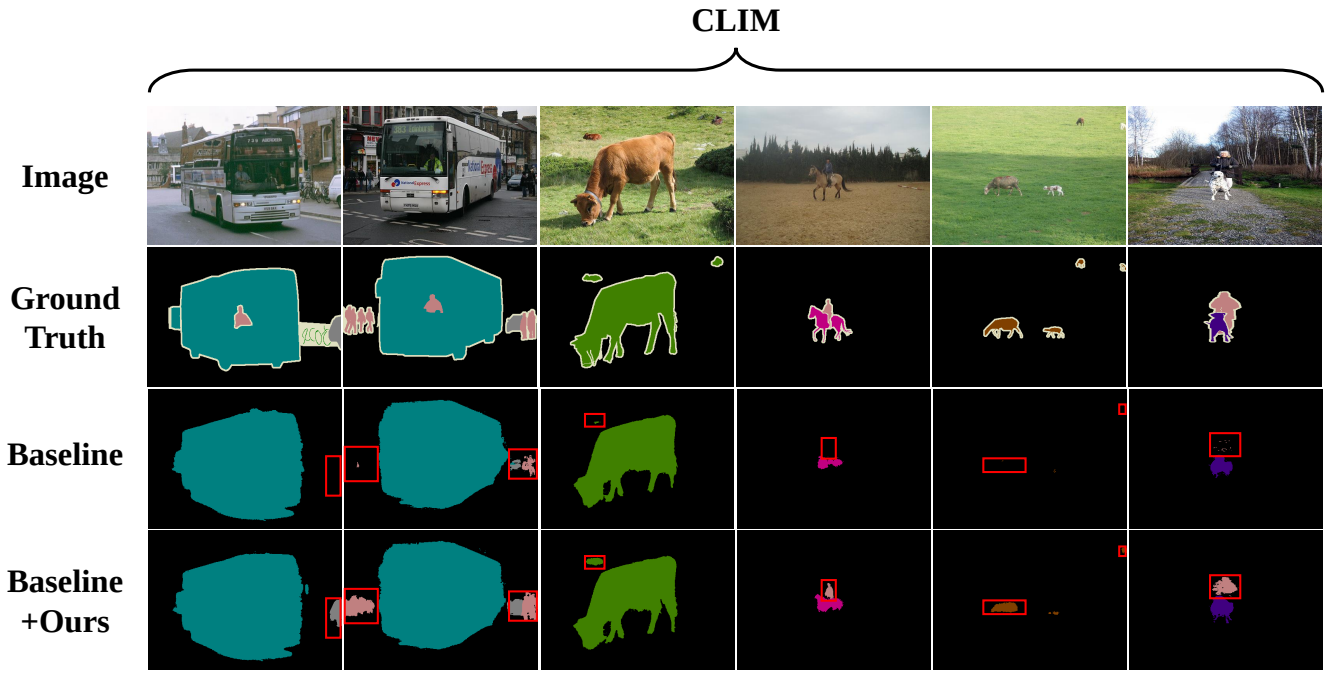

Figure 9. Visualization of CLIM on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 4

[12] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings*

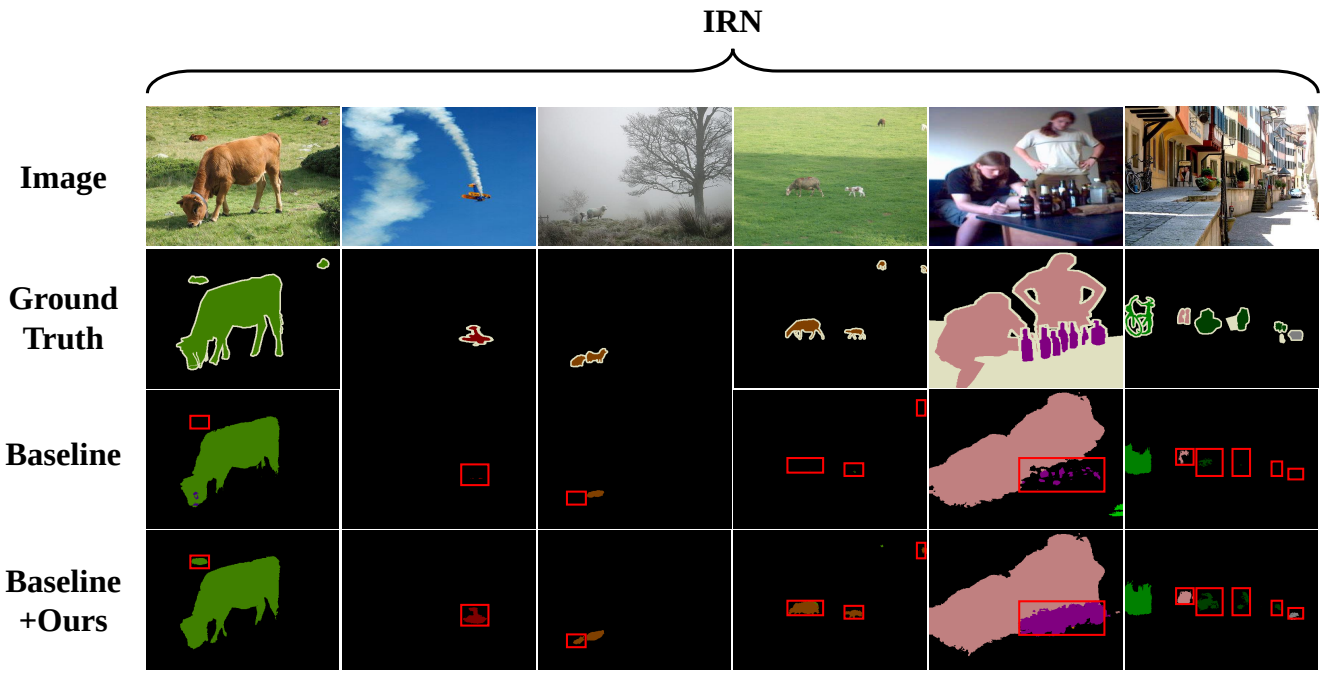

Figure 10. Visualization of IRN on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

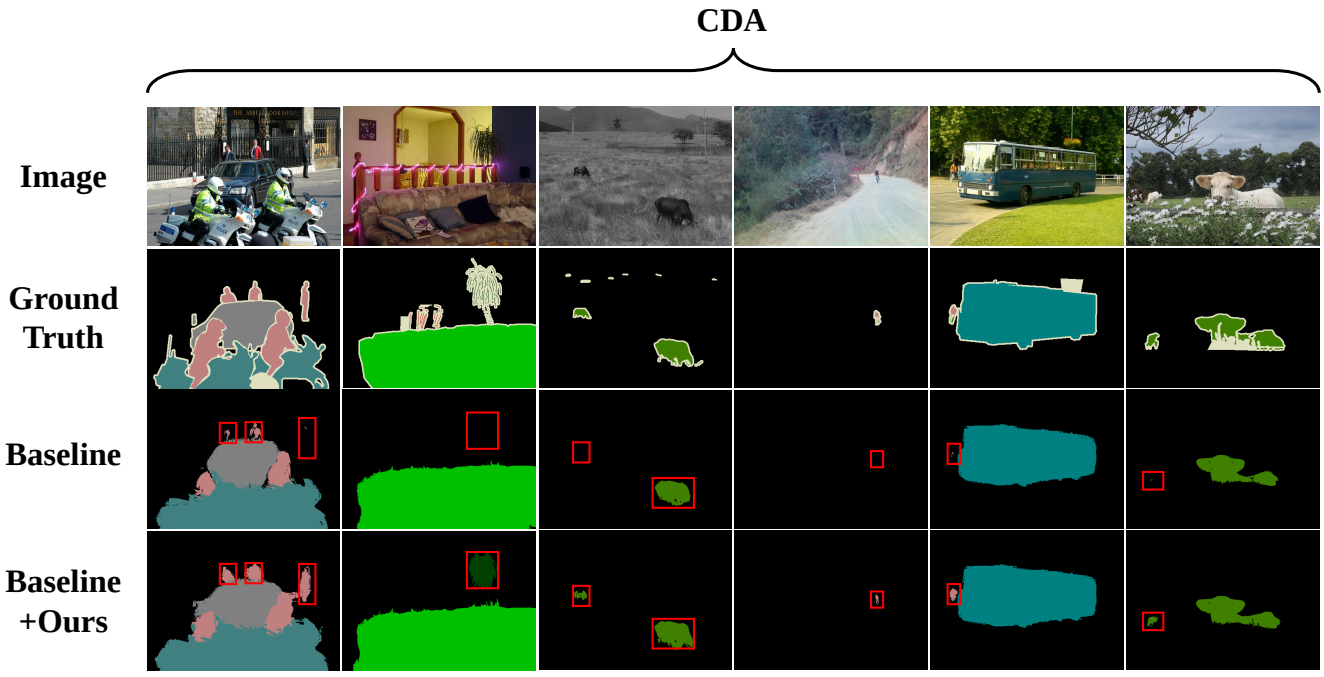

Figure 11. Visualization of CDA on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6913–6922, 2021. 2, 3, 4

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual

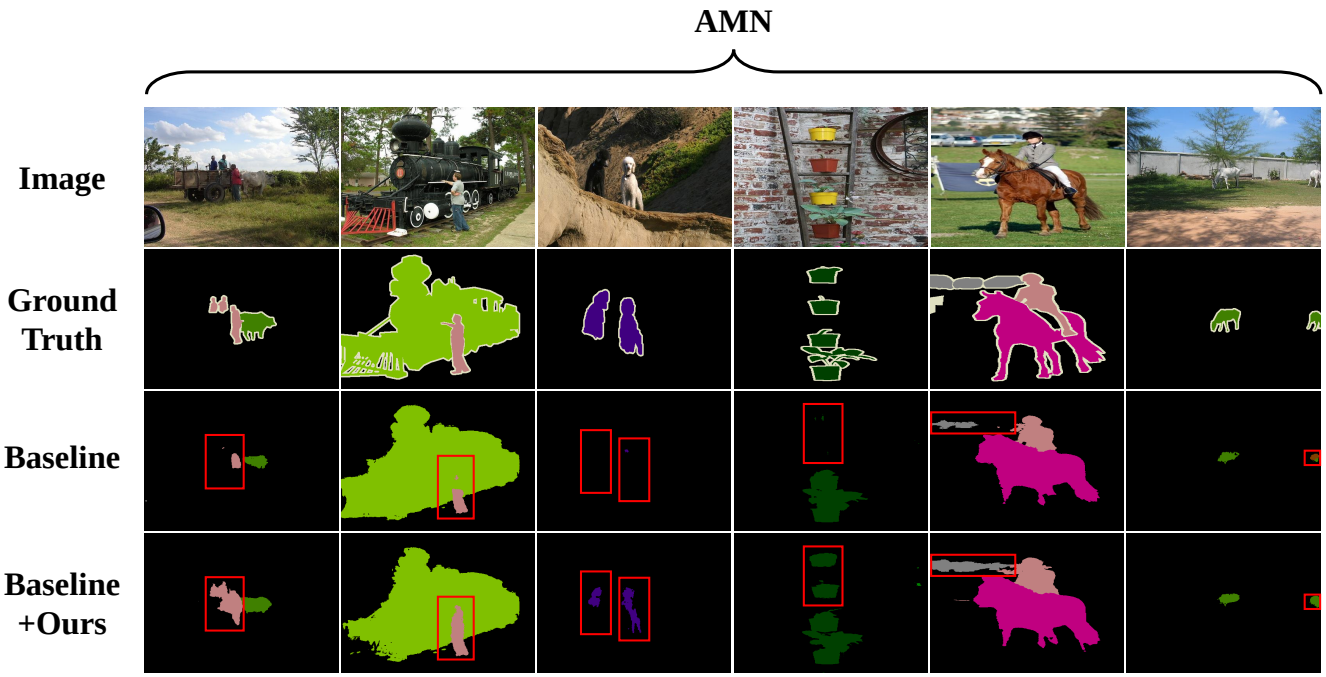

Figure 12. Visualization of AMN on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

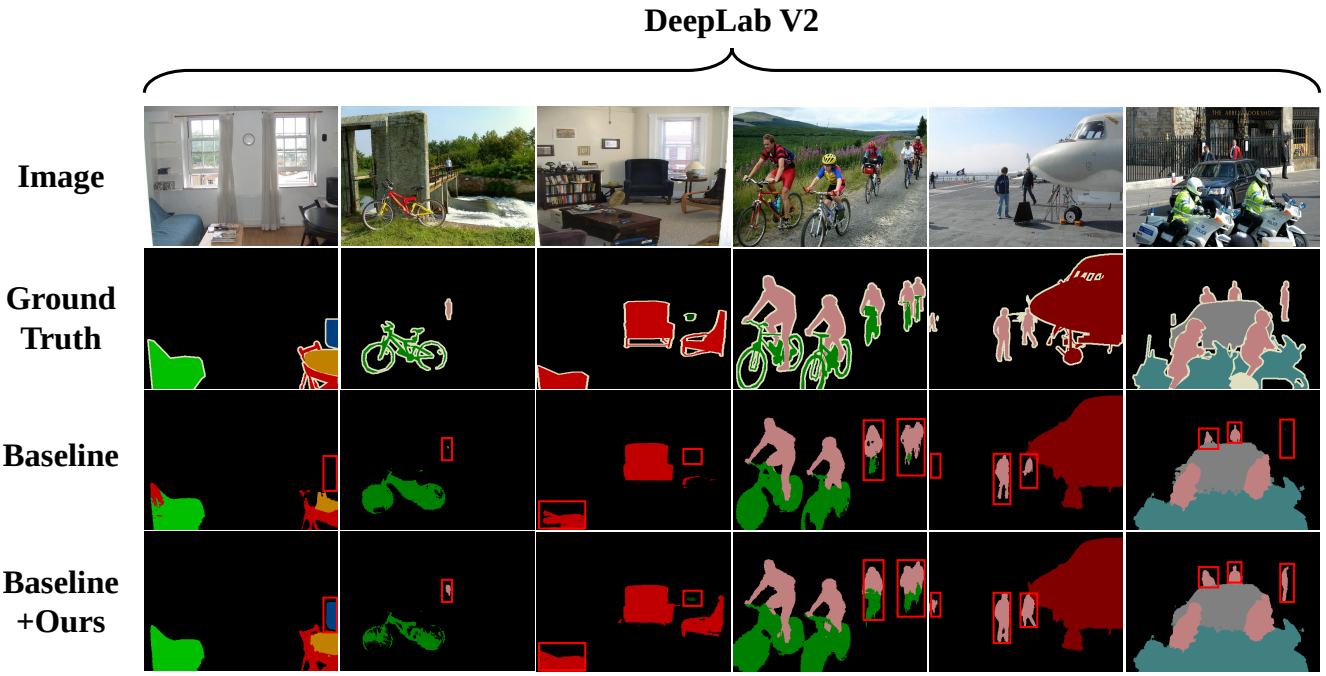

Figure 13. Visualization of DeepLab V2 on PASCAL VOC. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 2

[15] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia

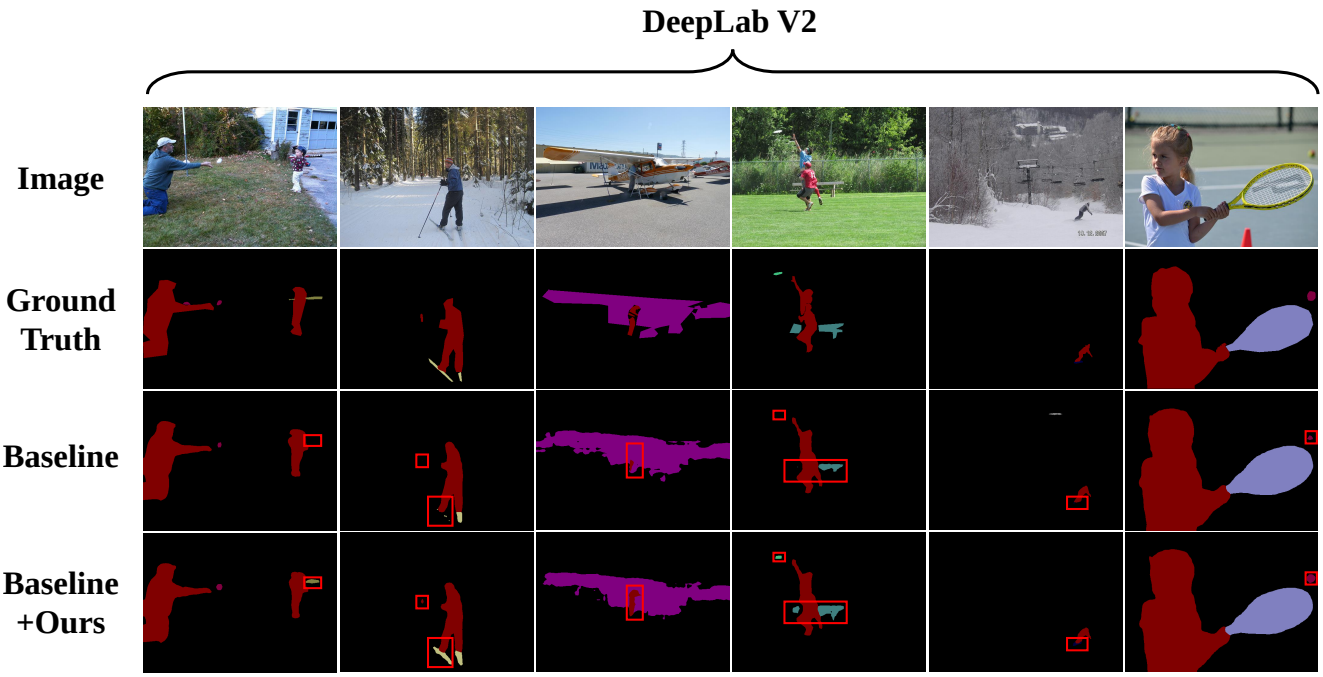

Figure 14. Visualization of DeepLab V2 on MS COCO. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

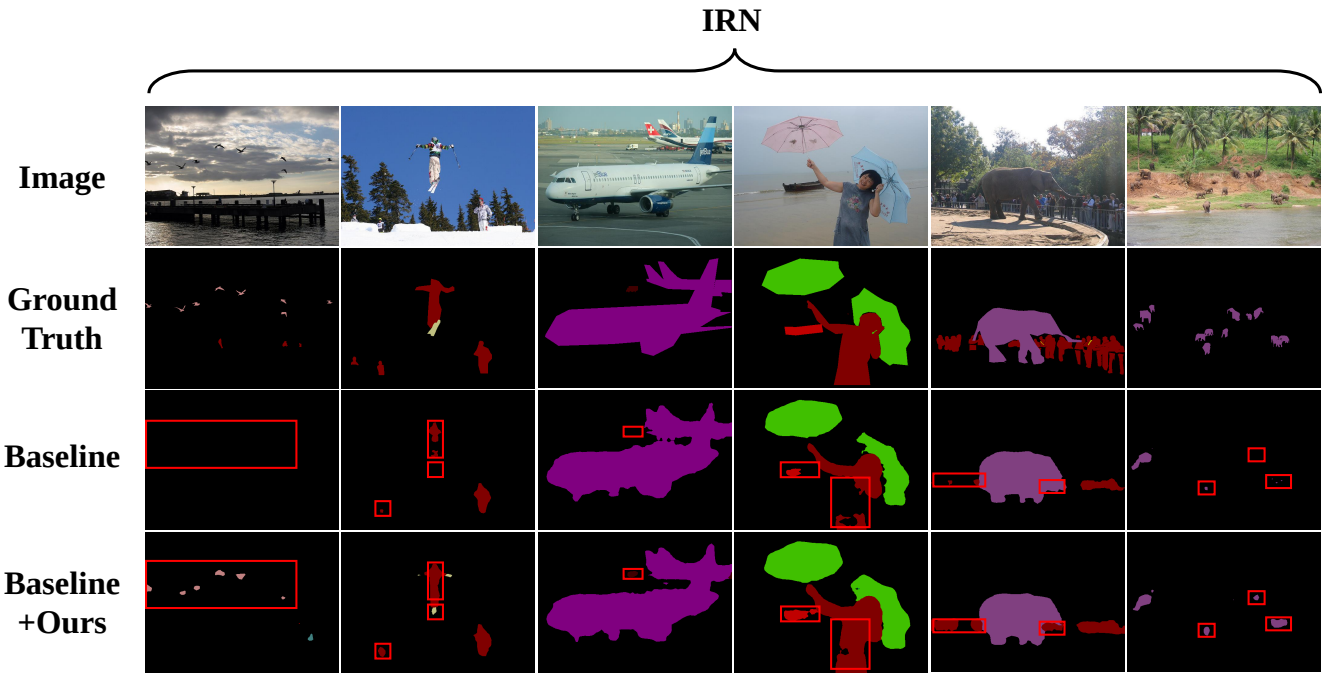

Figure 15. Visualization of IRN on MS COCO. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 4

[16] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. *arXiv preprint arXiv:2103.01795*,

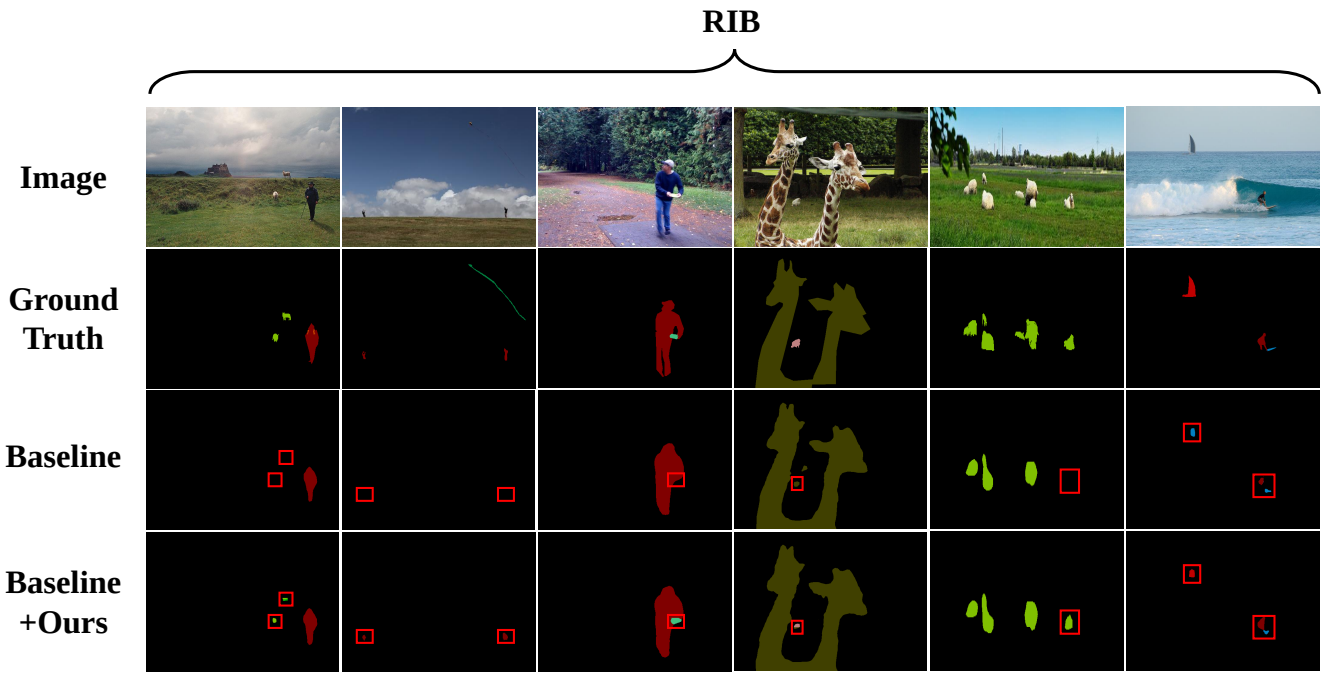

Figure 16. Visualization of RIB on MS COCO. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

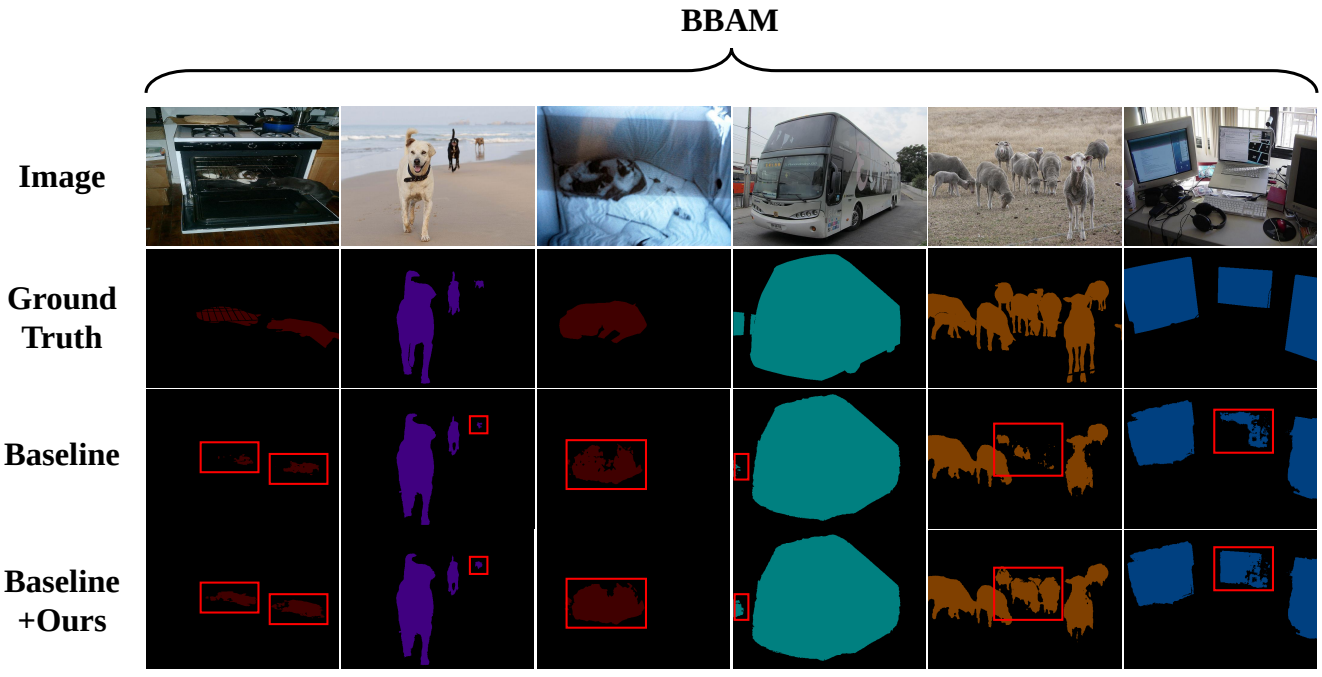

Figure 17. Visualization of BBAM on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

2021. 3, 4

[17] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages

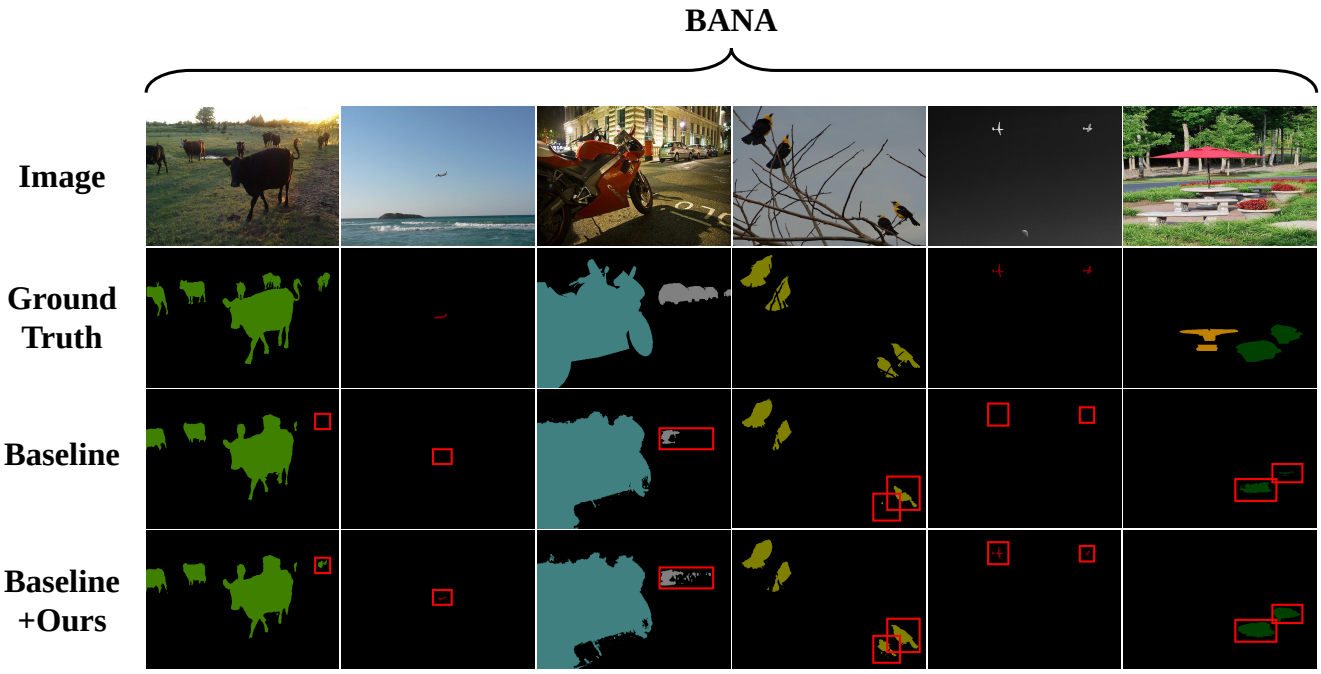

Figure 18. Visualization of BANA on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

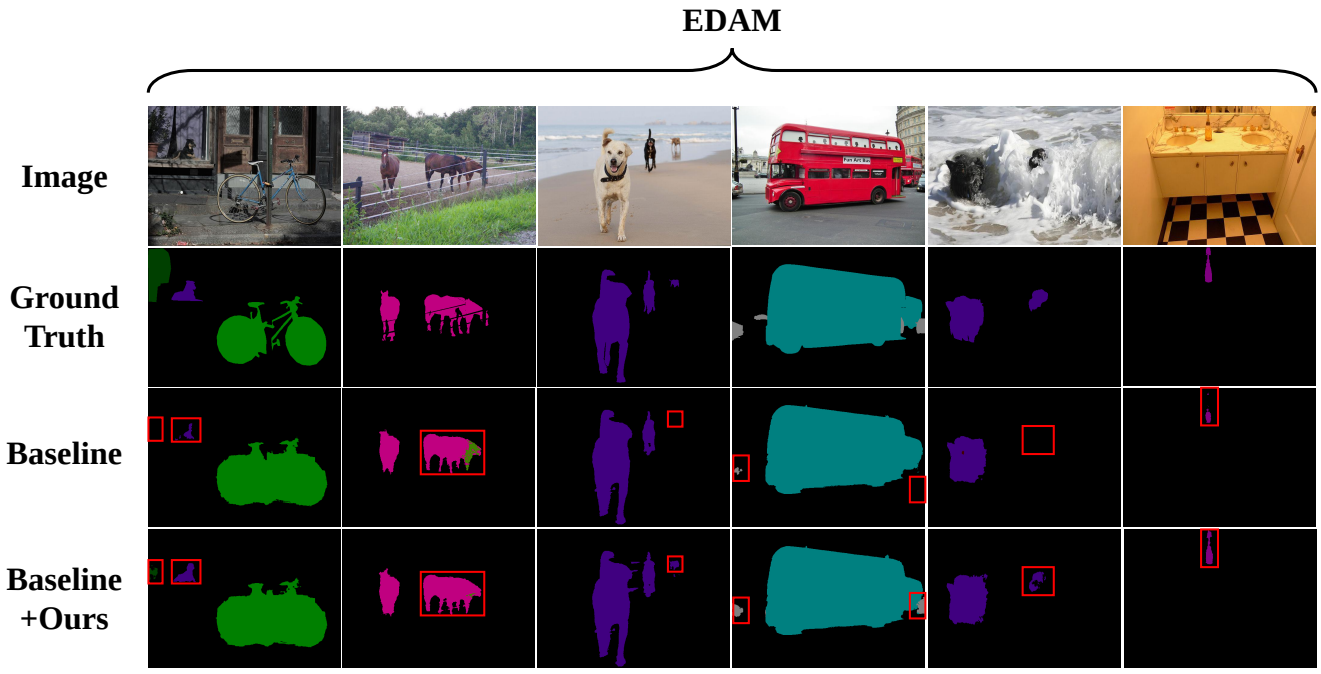

Figure 19. Visualization of EDAM on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

16765–16774, 2021. 2, 3, 4

[18] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4483–4492, June 2022. 3, 4

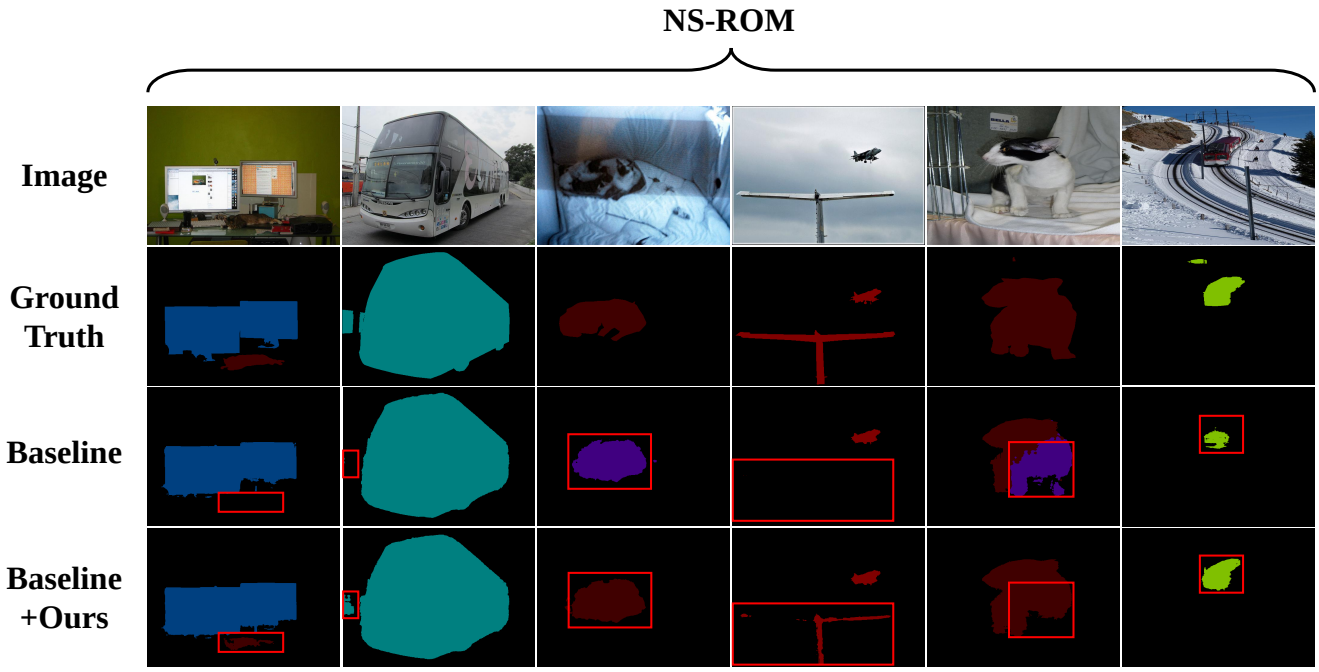

Figure 20. Visualization of NS-ROM on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

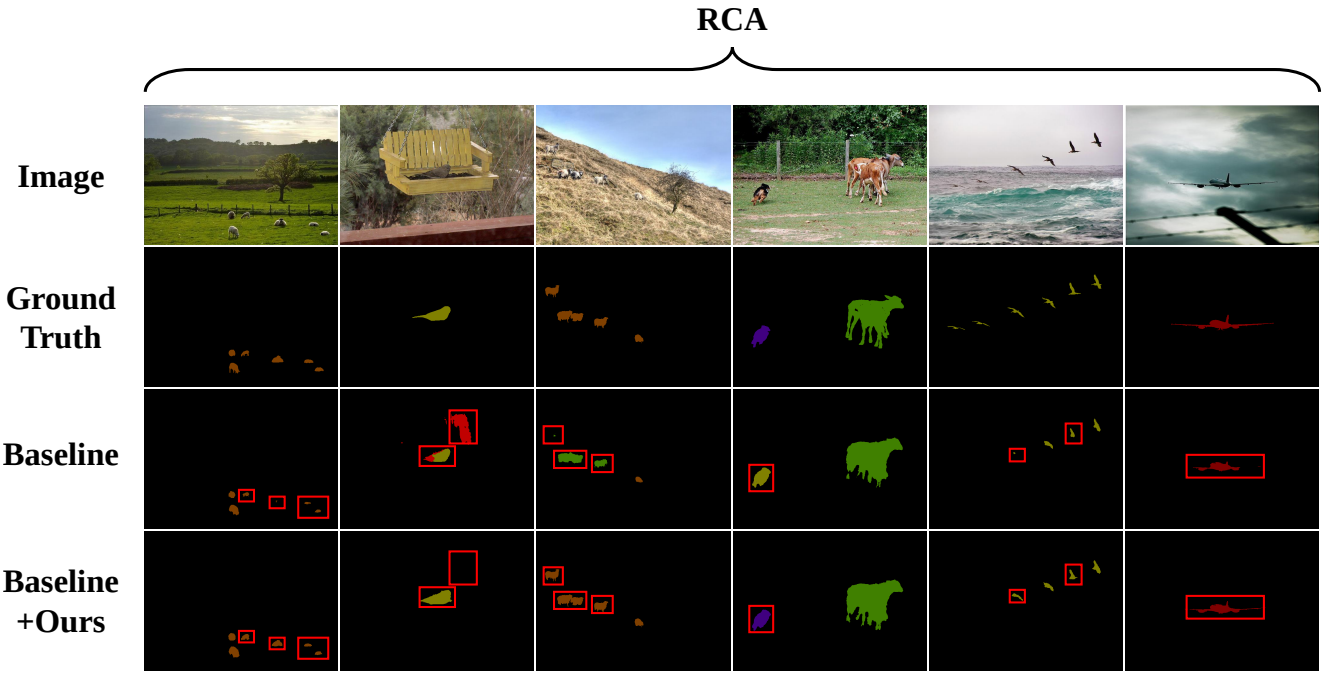

Figure 21. Visualization of RCA on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

[19] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2623–2632, 2021. 2, 3, 4

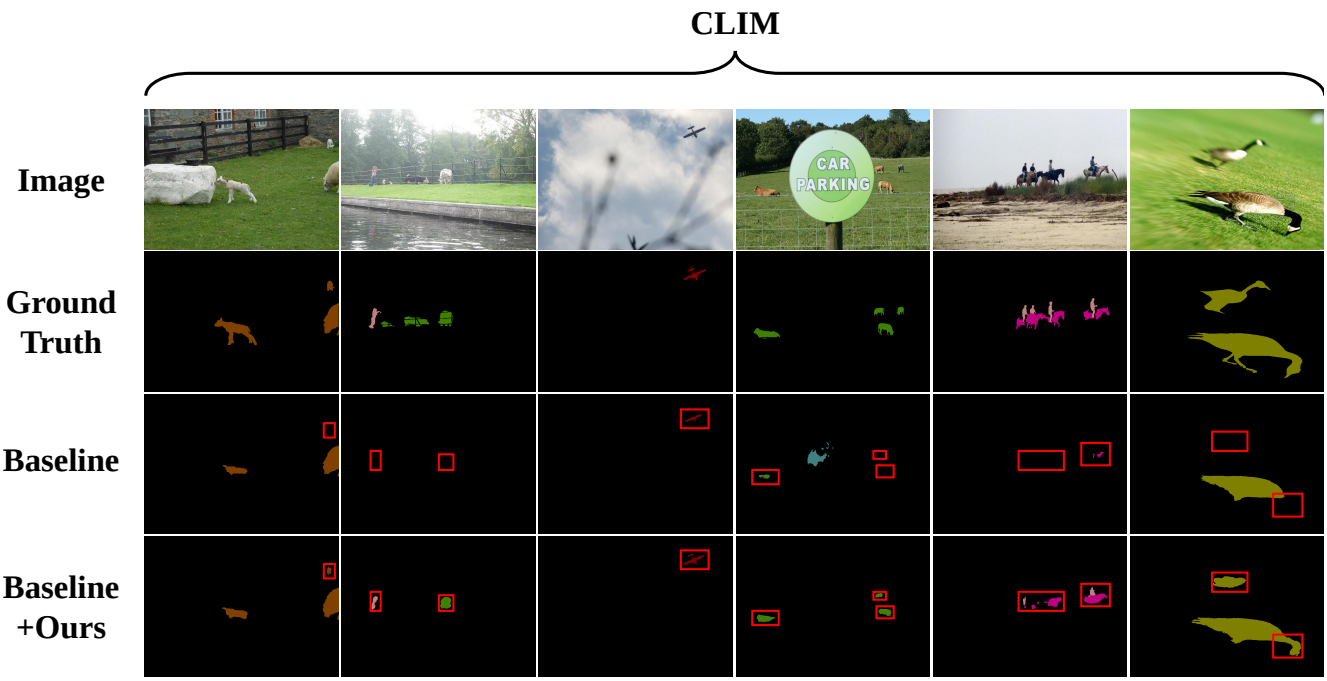

Figure 22. Visualization of CLIM on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

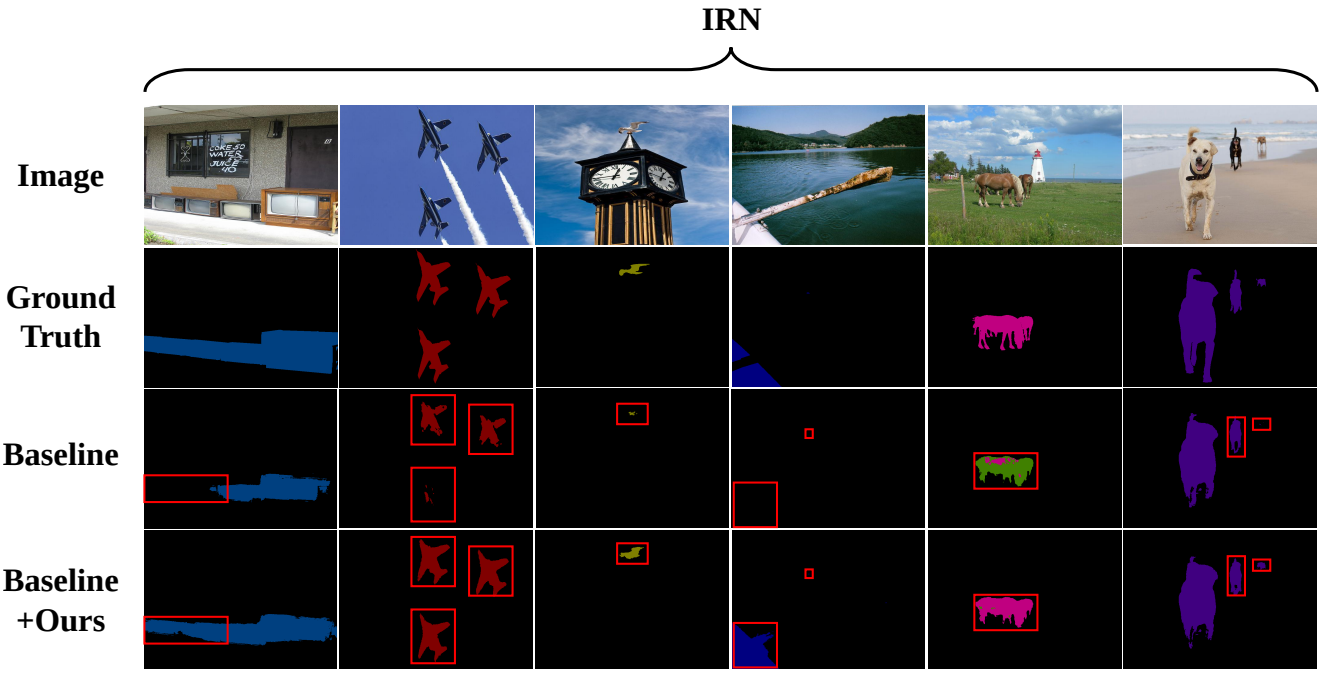

Figure 23. Visualization of IRN on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

[20] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 4

[21] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly su-

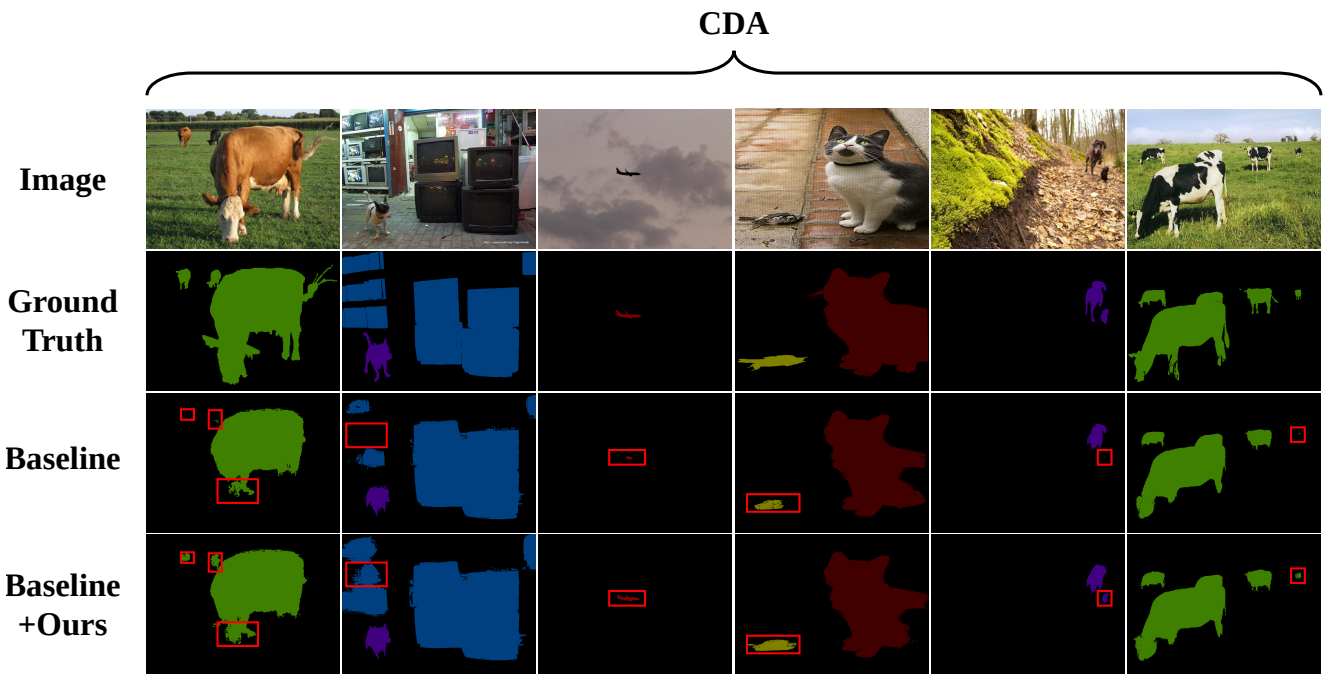

Figure 24. Visualization of CDA on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

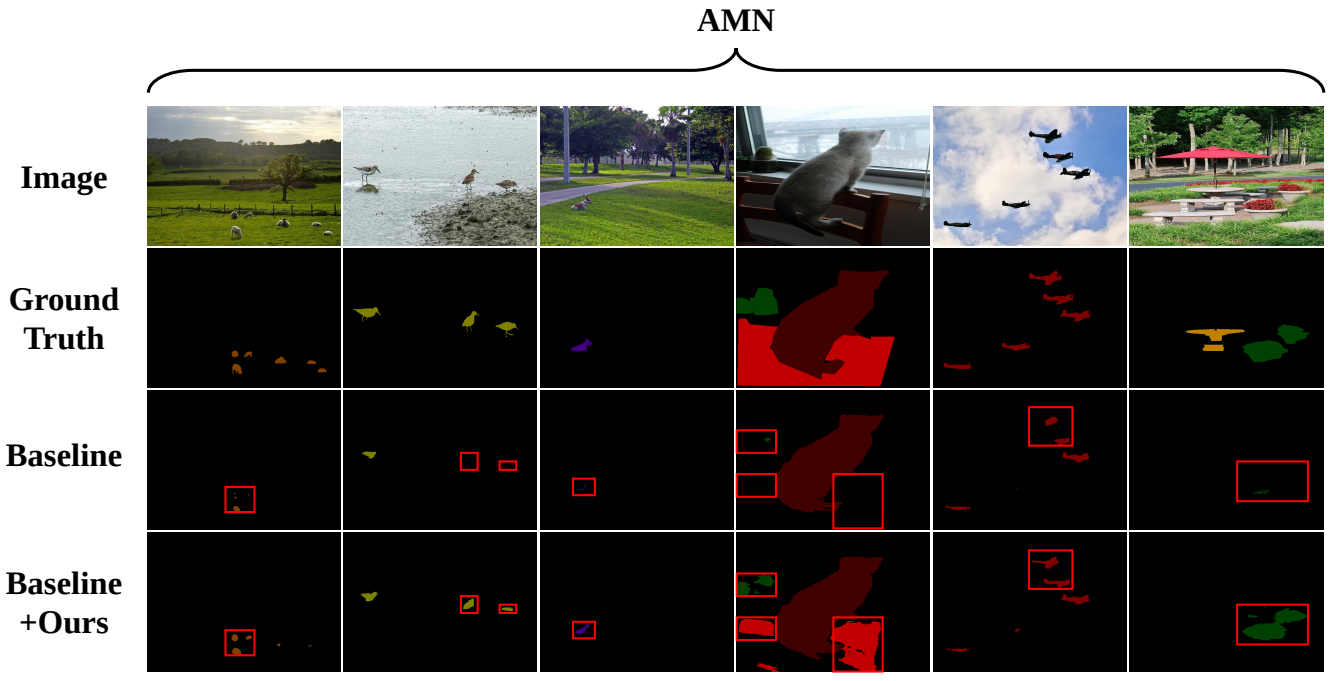

Figure 25. Visualization of AMN on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.

pervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022. 2, 3, 4

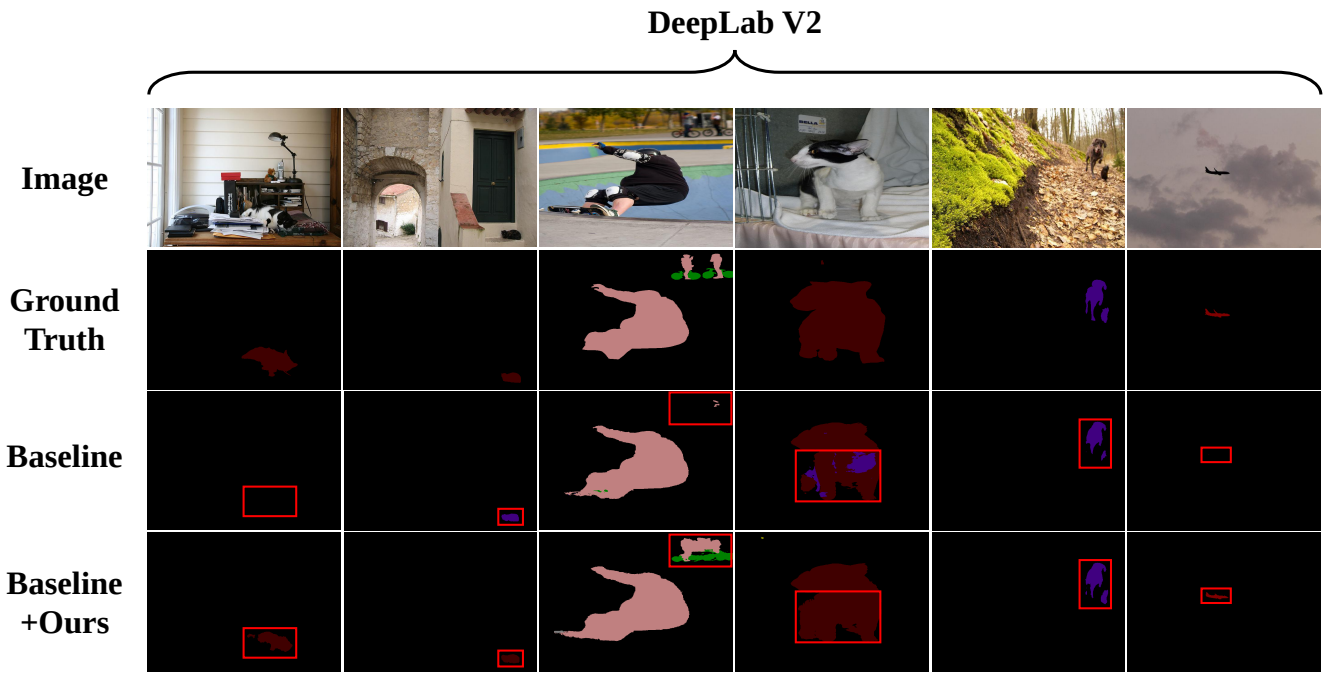

Figure 26. Visualization of DeepLab V2 on PASCAL-B. Our loss function successfully fine-tunes baseline model to improve the ability of capturing objects including small-sized ones which is expressed by red bounding boxes.