Unlocking the Potential of Ordinary Classifier: Class-specific Adversarial Erasing Framework for Weakly Supervised Semantic Segmentation

Table I: Quantitative comparison of the proposed frameworks with different ordinary classifiers backbones.

Backbone of Ordinary Classifier (mIoU)	Ours (mIoU)
ResNet38 (47.8%)	56.0%
ResNet101(46.8%)	52.4%
VGG16 (48.9%)	52.3%

A. Dependency on Ordinary Classifier

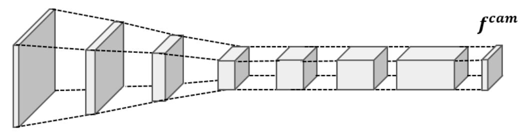
The proposed framework is designed to fully exploit the potential of the ordinary classifier. To demonstrate that our framework can utilize ordinary classifiers with various backbones, we perform experiments by replacing the backbone of the ordinary classifier from ResNet38 [11] to ResNet101 [4] or VGG16 [8]. In this experiment, while using the ordinary classifier with various backbones, we fix the backbone of the CGNet as ResNet38. Even though the number of parameters in ResNet38 is less than ResNet101, high resolution feature maps of ResNet38 are more suitable for the purpose of the proposed method which focuses on generating precise CAMs.

As shown in Table I, under the various backbone conditions, the proposed framework enables the CGNet to achieve significantly higher performance than the performance of the ordinary classifier.

B. Ablations Study on Masking Depth

As we mentioned in the main paper, we implement the ResNet38 [11] as backbone for both the CGNet and the ordinary classifier as many other previous works [1,2,7,9,12]. The architecture of the ResNet38 is shown in Fig. I with the intermediate feature maps and corresponding dimensions.

To the best of our knowledge, the masking methods in the AE scheme can be categorized into an image-level masking [6, 10] and a feature-level masking [5, 13]. In the proposed framework, we apply masking on the image-level rather than the feature-level. We experimentally verify the effectiveness of the image-level masking over the feature level masking within our framework. Figure II shows the comparison between the masking methods in both qualitative and quantitative manners. In the figure, masking at d_m means that the feature maps with the corresponding order



Feature maps	1	2	3	4	5	6	7	8
Resolution	$H \times W$	$\frac{H}{2}\times\frac{W}{2}$	$\frac{H}{4}\times\frac{W}{4}$	$\frac{H}{8}\times\frac{W}{8}$	$\frac{H}{8}\times\frac{W}{8}$	$\frac{H}{8}\times\frac{W}{8}$	$\frac{H}{8}\times\frac{W}{8}$	$\frac{H}{8}\times\frac{W}{8}$
Channels	64	128	256	512	1024	2048	4096	20

Figure I: The network architecture of ResNet38.

are masked. For example, when the masking depth $d_m = 3$, masking is applied on the feature maps which have the dimension of $\frac{H}{4} \times \frac{W}{4} \times 256$. For the fair comparison between two masking methods, all the hyperparameters in our framework except masking depth d_m are fixed.

As shown in Fig. II, the mIoU of the generated pseudo-label is significantly higher when the masking is done at the image-level rather than at the feature-level. In our view, in order to generate more precise CAMs, masking at the image-level is a more effective way to precisely erase the object from the image. This is because features can be entangled in the spatial domain by the receptive field rather than only representing the specific pixels at their corresponding coordinates. Therefore, even if a certain object is perfectly erased with a mask, the features near the object possibly contain the object-related information, which makes the features to be undesirably erased when using our framework. The image-level masking, on the other hand, literally "erases" only the object on image-level, and if the masking is perfect, then the network would not be able to find the object from the image.

As aforementioned, the feature-level masking leads the CGNet to overly erase the pixels near object boundaries while the image-level masking enables the CGNet to generate more precise CAMs. In conclusion, we experimentally verify that the image-level masking is superior to feature-level masking in both qualitative and quantitative manners.

C. More Results and Qualitative Comparison

To show that our framework can produce precise CAMs, more results for PASCAL VOC 2012 not included in the main paper due to page limit are shown in Fig. III. With

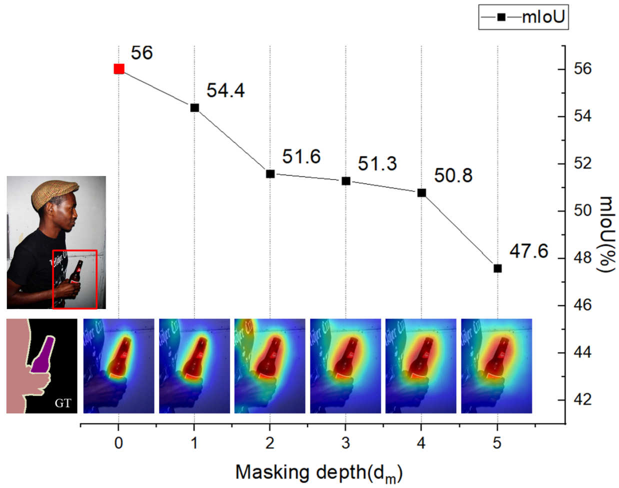


Figure II: Comparison between the image-level masking and the feature-level masking. Masking depth d_m with 0 indicates image-level masking. d_m more than 1 denotes feature-level masking.

those refined CAMs, we generate pseudo pixel-level labels by applying dense crf and AffinityNet [1]. With the synthesized pseudo-labels, we train the semantic segmentation network [3]. Also, more results of Deeplab trained by our pseudo-labels and comparison are shown in Fig. IV.

Qualitative comparison for MS-COCO dataset is also available in Fig V and Fig. VI, which shows the CAMs and results of Deeplab trained by pseudo-labels, respectively.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 1, 2, 4

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 1

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR*, 2015. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018. 1

[6] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 1

[7] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5208–5217, 2019. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1

[10] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 1

[11] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 1

[12] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12765–12772, 2020. 1

[13] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 1

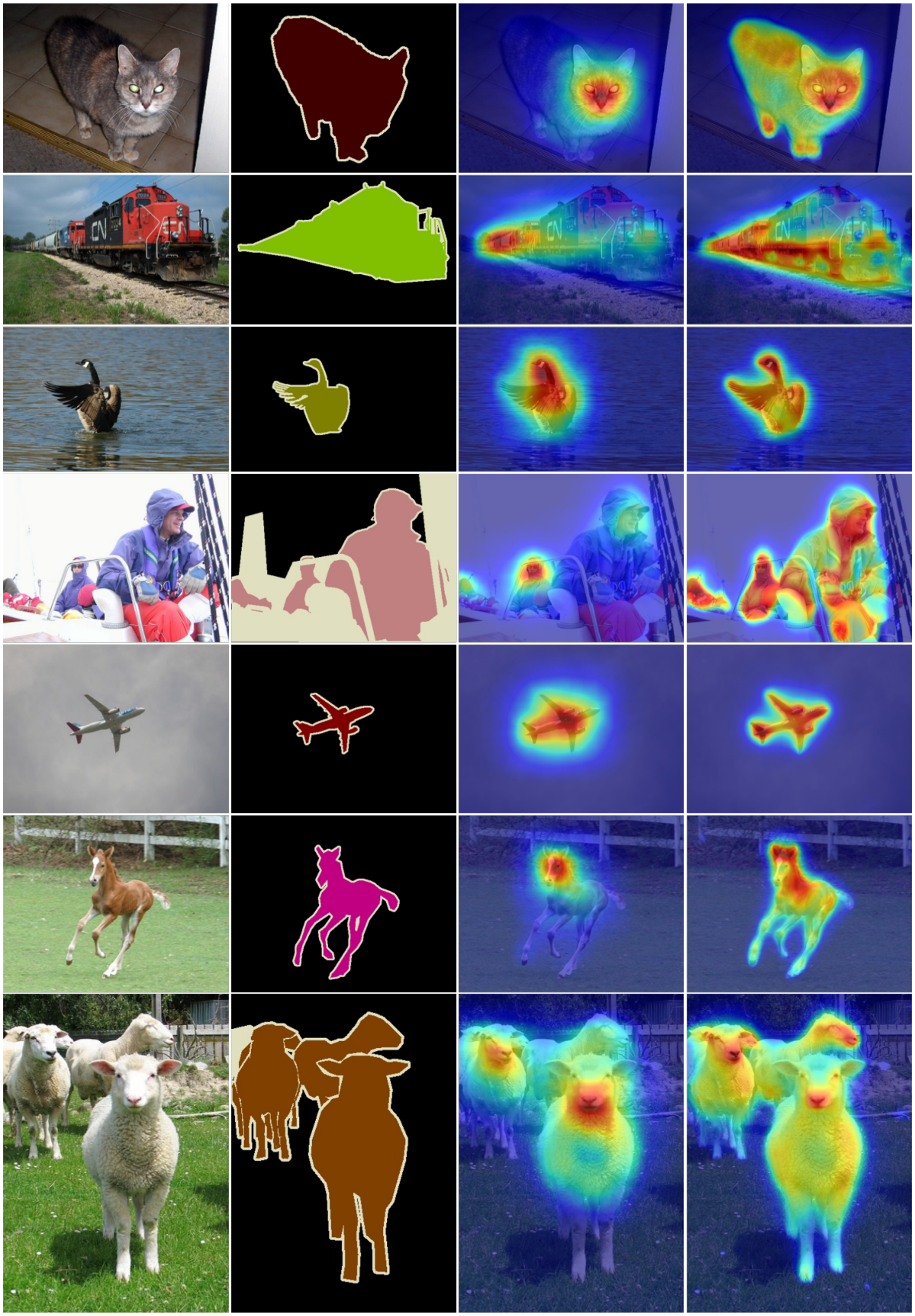


Figure III: Qualitative comparison of CAMs for PASCAL VOC 2012. From left to right: images, ground-truths, baseline CAMs with ResNet38 backbone, our CAMs.

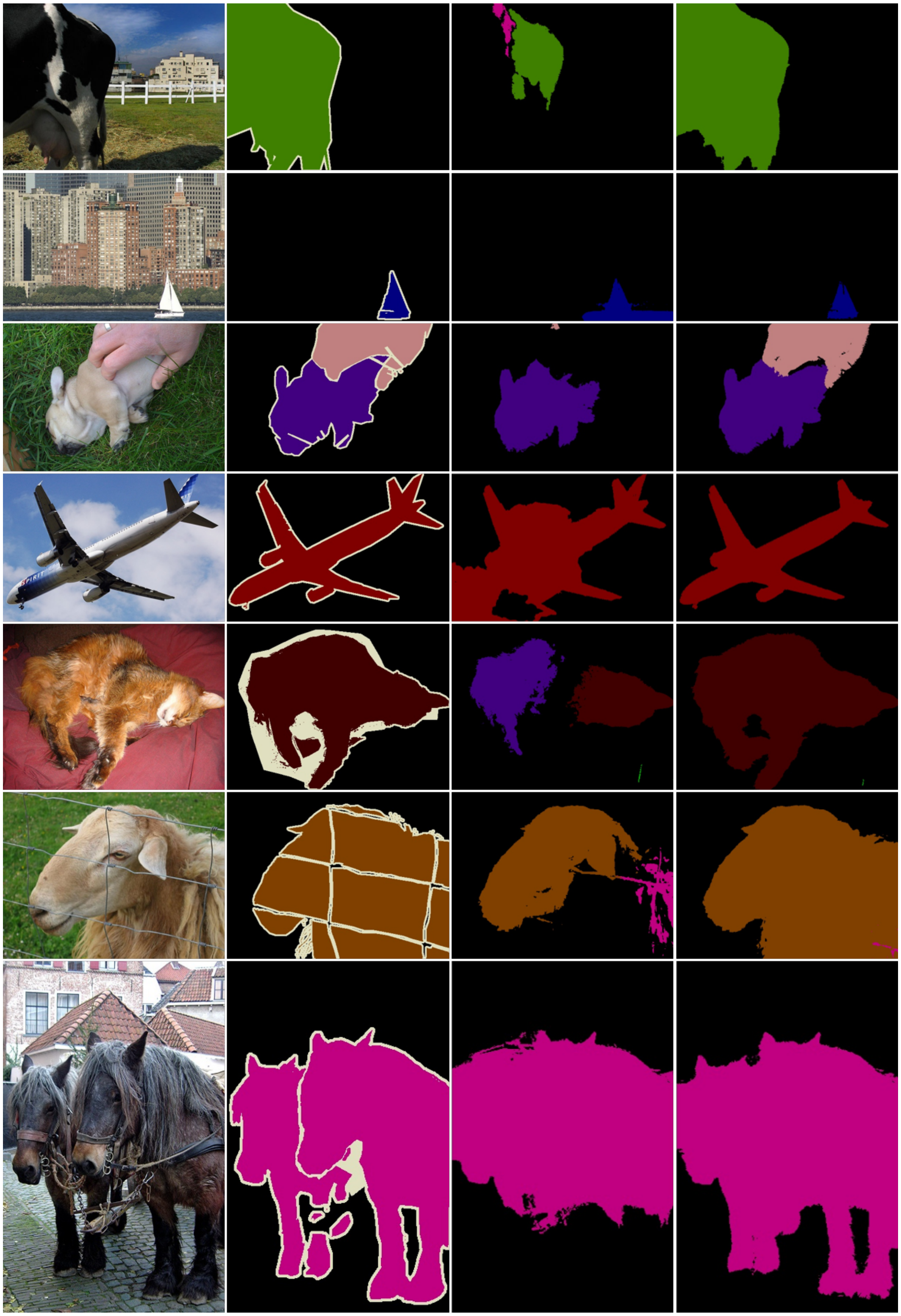


Figure IV: Qualitative comparison of semantic segmentation maps for PASCAL VOC 2012. From left to right: images, ground-truths, Deeplab trained with the baseline [1], Deeplab trained by ours.

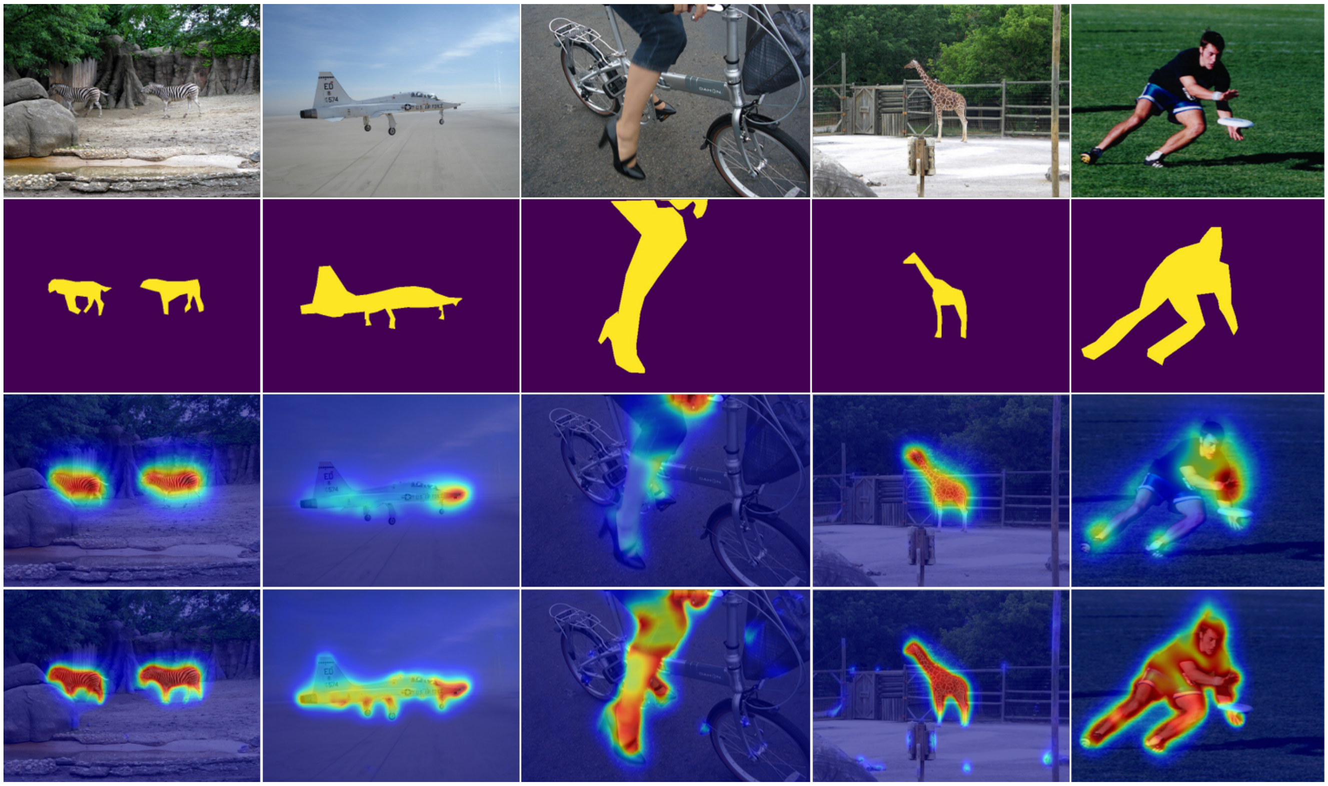


Figure V: Qualitative comparison of CAMs for MS-COCO. From top to bottom: images, ground-truth masks, baseline CAMs with ResNet38 backbone, our CAMs.

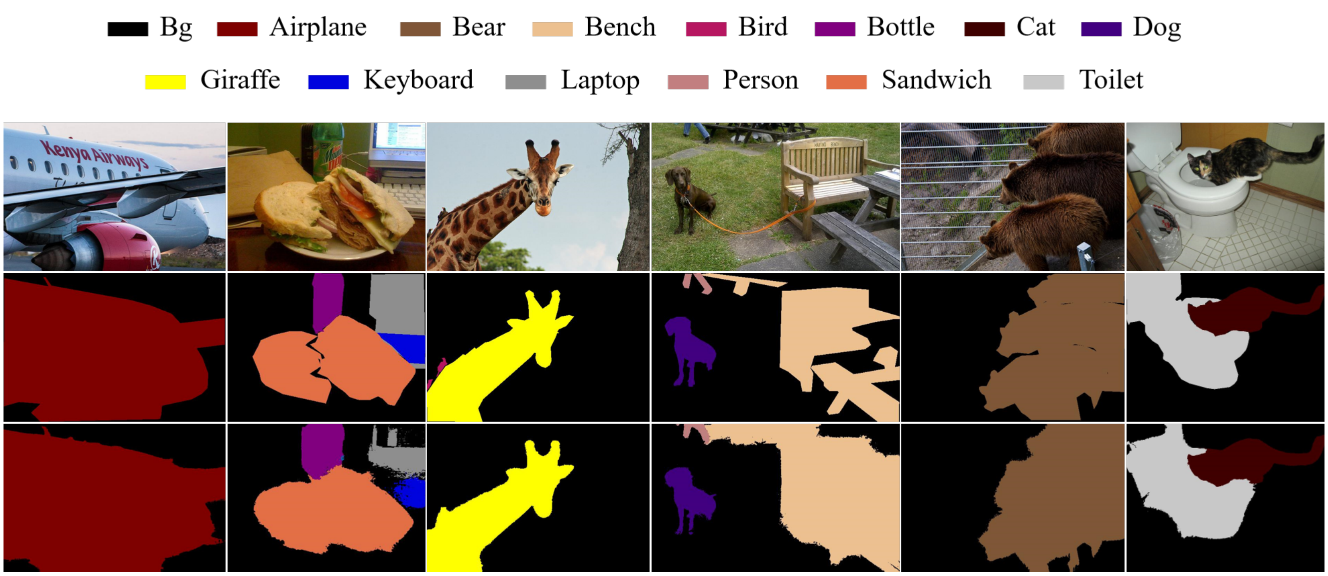


Figure VI: Result of semantic segmentation maps for MS-COCO. From top to bottom: images, semantic segmentation maps from ground-truth masks, results of Deeplab trained by ours.