ContrastMask: Contrastive Learning to Segment Every Thing Supplementary Material

Xuehui Wang[1†], Kai Zhao[2], Ruixin Zhang[2], Shouhong Ding[2], Yan Wang[3], Wei Shen[1(✉)]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Youtu Lab, Tencent [3]Shanghai Key Lab of Multidimensional Information Processing, ECNU
{wangxuehui, wei.shen}@sjtu.edu.cn; {ruixinzhang, ericshding}@tencent.com;
kz@kaizhao.net; ywang@cee.ecnu.edu.cn

In this supplementary material, we provide more discussions about our method as well as some new results, including both qualitative and quantitative. Here, we simplify contrastive learning/positive/negative as CL/pos/neg, respectively.

1. Possible application scenarios

Our method is a kind of label-efficient learning method, which is known as the key to extending AI algorithms to real-world applications. We give two examples for reference. 1) In autonomous driving, when encountering an unknown scene where new objects are not involved in existing training set, we can annotate these objects quickly with only box annotations and then combine them with existing training set. Then, a more accurate and robust model can be retrained. 2) Suppose that a shopping mall wants to build an automatic alert system to monitor whether the fire passage is blocked by some objects. This requires training a scene parsing model. There exist familiar objects (person, box) that have abundant annotations in public datasets or internal datasets of a company but also unusual objects (wheelbarrow, rubbish bin) that do not have mask annotations. We can apply our partially-supervised method here to reduce annotation burden.

2. Relation to a teacher-student model

One may argue that our method looks like a teacher-student model, and thus the low quality of CAM may lead to negative impacts on the segmentation performance. But we think this is a kind of misunderstanding. First, we use CAM as a prior to form query-pos/-neg pairs for CL, rather than a teacher supervision. Note that as an unsupervised/weakly-supervised learning framework, CL requires some priors, even though the priors are not always correct, to determine pos/neg keys, *e.g.*, instance discrimination (MoCo [2], CVPR20) and color consistency (DenseCL [4], CVPR21). Second, although sometimes CAMs are incorrect, the influence is limited (less than 1.4 mAP compared with using GTs for *novel* as the prior). This is benefited from that we employ two strategies to enhance the robustness of the CL head: 1) We adopt the query-sharing strategy which forms a query based on both *base* and *novel* data in a batch. 2) We only consider high confidence areas of CAMs as the aforementioned prior, which diminishes the impact of errors.

3. Visualization results

We first visualize the CAMs and corresponding pseudo masks generated by our method of some examples in Fig. 1. These pseudo masks are used as the prior to obtain the foreground and background queries and sample keys for *novel* categories.

We also provide more visualizations of the segmentation results produced by our method in Fig. 2. We regard *nonvoc* as *base* and *voc* as *novel* categories. Note that *novel* categories only have box-level annotations. We successfully obtain good segmentation results on *novel* categories.

4. Updated quantitative results

When preparing code releasing, we find a minor bug that a configuration file of MMDetection [1] changes slightly and we omit this by accident, which could cause a label-leaked problem for a few settings. Thus, we fix this bug and update the quantitative results regarding "$nonvoc \rightarrow voc$" using ResNet-50 [3] and ResNet-101 [3] as backbones under $3\times$ schedule, which are shown in Table 1 of our paper. We sincerely apologize for this. Fortunately, after updating the results, our method still outperforms others with an obvious improvement, achieving SOTA performance on all settings, which does not influence our previous statement and conclusion. Besides, we also provide stronger results by using ResNeXt [5] backbones under the $3\times$ schedule to show the potential of our method.

[†]Work done during an internship at Youtu Lab, Tencent.
[✉]Corresponding Author.

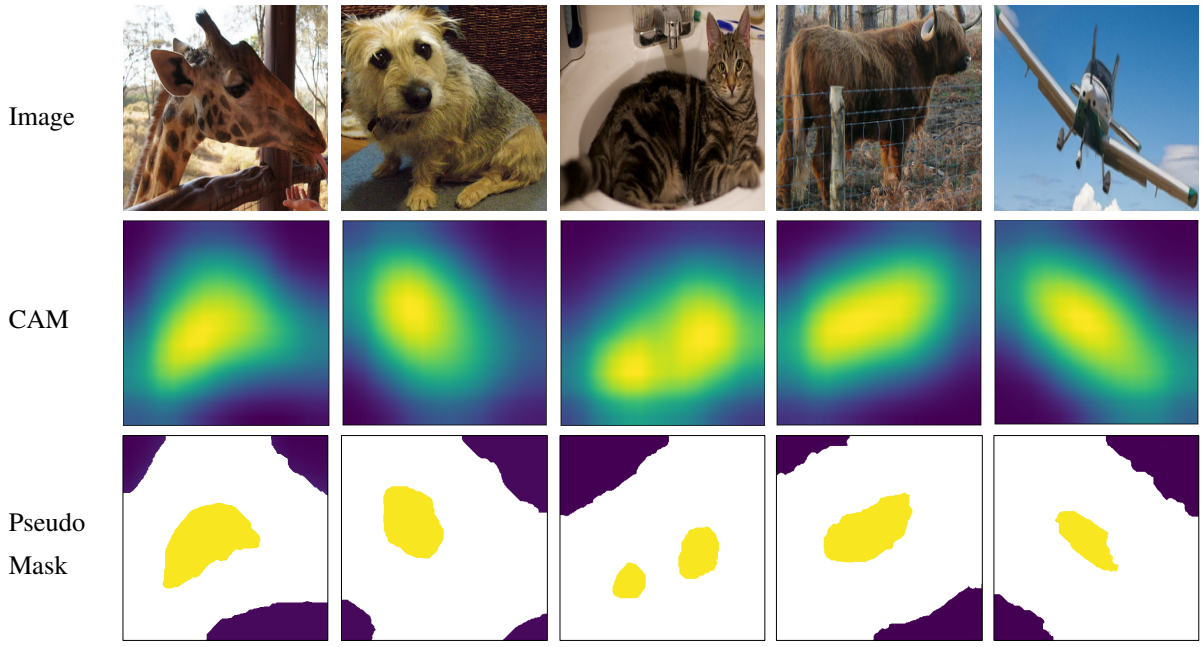

Figure 1. The visualization of CAMs and pseudo masks for some examples. The yellow and dark purple areas in pseudo masks denote for the foreground and background partitions, respectively.

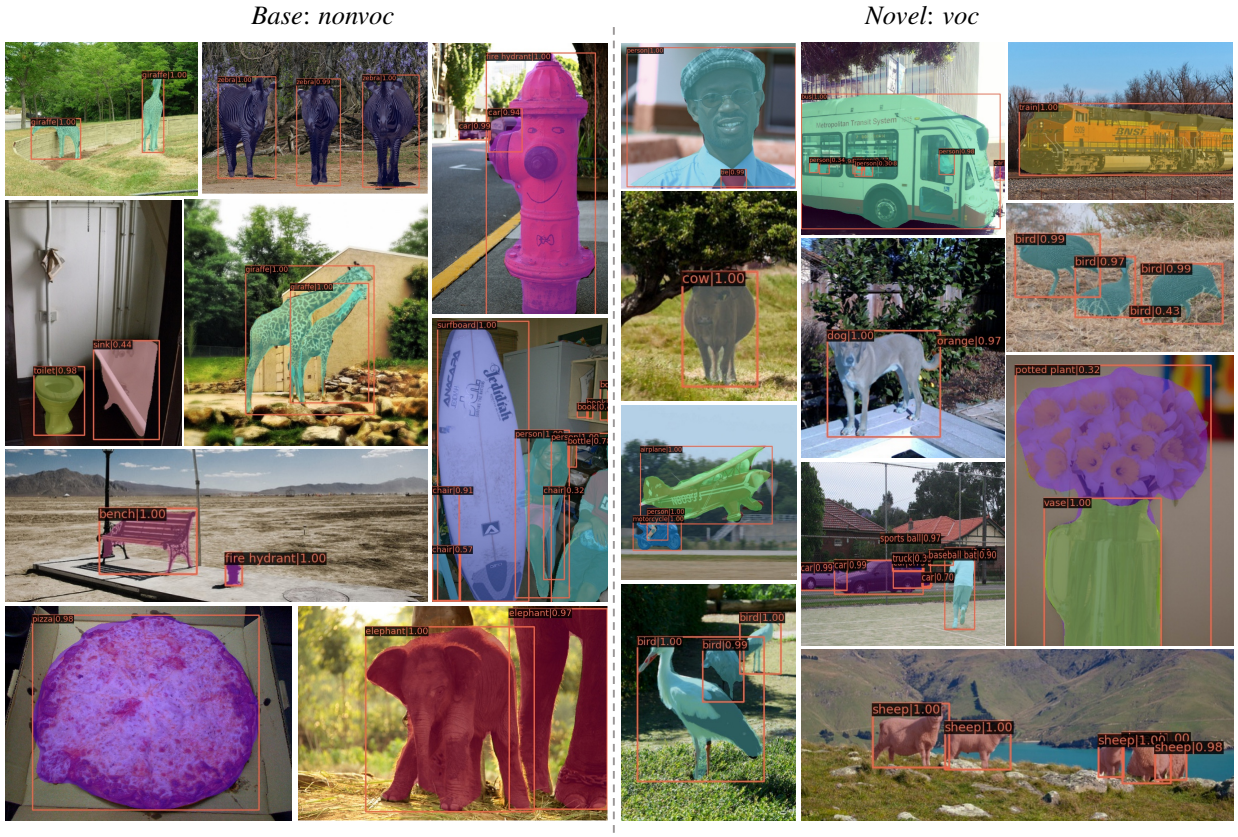

Figure 2. The visualization of our segmentation results on *base* and *novel* categories.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1

[3] 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 (CVPR)*, June 2016. 1

[4] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[5] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1