Convolutional Feature Masking for Joint Object and Stuff Segmentation

Jifeng Dai, Kaiming He, Jian Sun
Microsoft Research

The topic of semantic segmentation has witnessed considerable progress due to the powerful features learned by convolutional neural networks [5]. The current leading approaches for semantic segmentation exploit shape information by extracting CNN features from masked image regions. This strategy introduces artificial boundaries on the images and may impact the quality of the extracted features. Besides, the operations on the raw image domain require to compute thousands of networks on a single image, which is time-consuming.

In this paper, we propose to exploit shape information via masking convolutional features. The proposal segments (*e.g.*, super-pixels) are treated as masks on the convolutional feature maps. The CNN features of segments are directly masked out from these maps and used to train classifiers for recognition. We further propose a joint method to handle objects and "stuff" (*e.g.*, grass, sky, water) in the same framework. State-of-the-art results are demonstrated on the challenging PASCAL VOC benchmarks, with a compelling computational speed.

Convolutional Feature Masking

The Convolutional Feature Masking (CFM) layer is a layer used after the final convolutional layer. We apply all convolutional layers on the image to compute the convolutional feature maps. This operation is only performed once on the entire image. We also obtain the candidate segments (like superpixels from Selective Search [7] or MCG [1]), which are binary masks on the raw images. We project these binary masks to the domain of the convolutional feature maps. These masks are then applied on the convolutional feature maps. We call the resulting features as *segment features*. Figure 1 shows an illustration.

With CFM layer, we fine-tune the network using a pipeline similar to the SPP-Net pipeline [4] for object detection. Because the convolutional feature maps are only computed once, our method is orders of magnitude faster than R-CNN-based methods [2, 3] for semantic segmentation.

Joint Object and Stuff Semantic Segmentation

The semantic categories in natural images can be roughly divided into *objects* and *stuff*. Objects have consistent shapes and each instance is countable, while stuff has consistent colors or textures and exhibits as arbitrary shapes, *e.g.*, grass, sky, and water. So unlike an object, a stuff region is not appropriate to be represented as a rectangular region or a bounding box.

We show a generalization of our framework to address this issue involving stuff. We can simultaneously handle objects and stuff by a single solution. Our generalization is to modify the underlying probabilistic distributions of the samples during training. Instead of treating the samples equally, our training will bias toward the proposals that can cover the stuff as compact as possible. Fig. 2 shows our joint object and stuff segmentation results using the CFM framework.

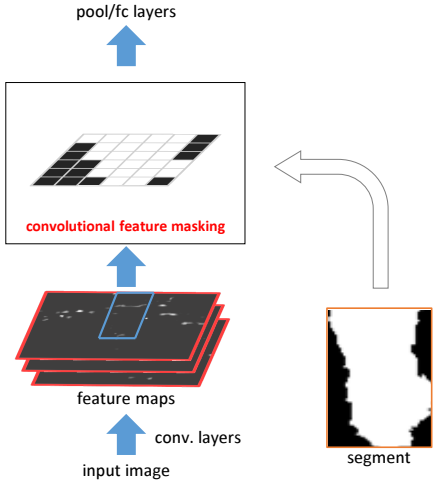

Figure 1: An illustration of the CFM layer.

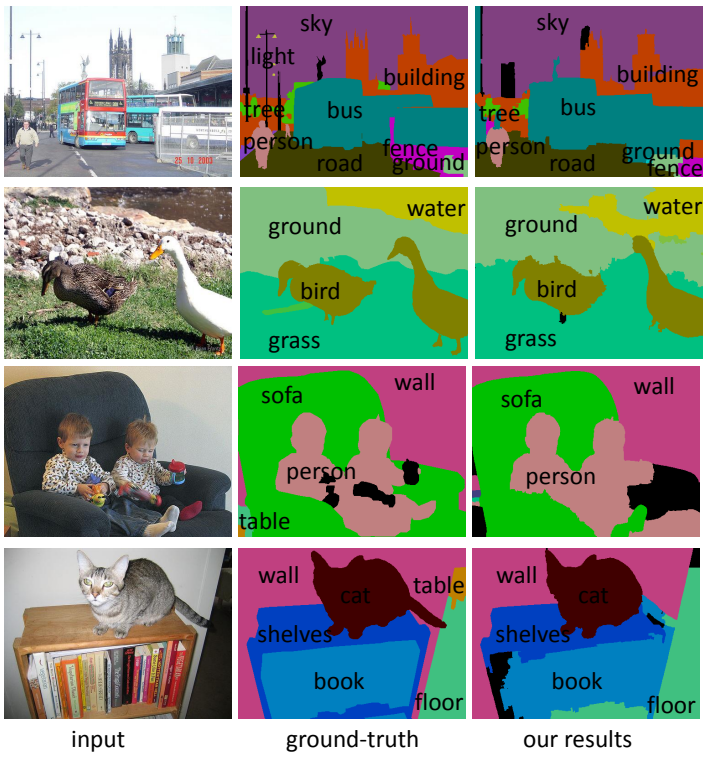

Figure 2: Some results of CFM for **joint object and stuff segmentation**. The images are from the enriched PASCAL VOC 2010 val set [6].

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. CVPR, 2014.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[3] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*. 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *arXiv preprint arXiv:1406.4729*, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[6] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*. 2014.

[7] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.