Recurrent Scene Parsing with Perspective Understanding in the Loop

Shu Kong, Charless Fowlkes
Department of Computer Science
University of California, Irvine
{skong2,fowlkes}@ics.uci.edu

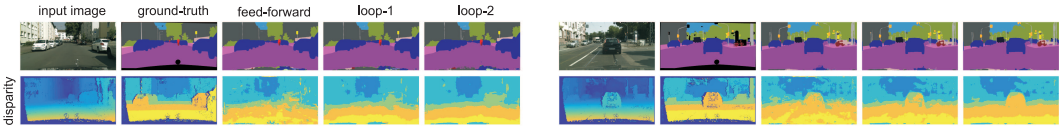

Figure 1: Visualization of two validation images from Cityscapes with the segmentation output and the predicted quantized disparity at each iteration of the recurrent loop. We depict "ground-truth" continuous and quantized disparity beside the input image. Our monocular disparity estimate makes predictions for reflective surfaces where stereo fails and recurrent iteration further improves estimates, particularly for featureless areas such as the pavement.

An intrinsic challenge of parsing rich scenes is understanding object layout relative to the camera. Roughly speaking, the scales of the objects in the image frame are inversely proportional to the distance to the camera. Humans easily recognize objects even when they range over many octaves of spatial resolution, e.g., the cars near the camera in urban scene can appear a dozen times larger than those at distance. The huge range and arbitrary scale at which objects appear poses difficulties for machine image understanding. Although individual local features (e.g., in a deep neural network) can exhibit some degree of scale-invariance, it is not obvious they handle the range scale variation that exists in images.

In this paper, we investigate how cues to perspective geometry conveyed by image content, estimated from stereo disparity, or measured directly via specialized sensors, might be exploited to improve recognition and scene understanding. We propose a depth gating module that adaptively selects pooling field sizes over higher-level feature activation layers in a convolutional neural network (CNN), as shown in Fig. 2. Adaptive pooling works with a more abstract notion of scale than standard multiscale image pyramids that operate on input pixels. This mechanism allows spatially varying processing over the visual field which can capture context for semantic segmentation that is not too large or small, but "just right", maintaining details for objects at distance while simultaneously using much larger receptive fields for objects closer near the camera. The gating architecture is trained with a loss that encourages selection of target pooling scales derived from "ground-truth" stereo disparity but at test time makes accurate inferences about scene depth using only monocular cues.

Furthermore, inspired by studies of human visual processing (e.g., [2]) that suggest dynamic allocation of computation depending on the task and image content (background clutter, occlusion, object scale), we propose embedding gated pooling inside a recurrent refinement module that takes initial estimates of high-level scene semantics as a top-down signal to reprocess feed-forward representations and refine the final scene segmentation, as shown in Fig. 3. This provides a simple implementation of "Biased Competition Theory" [1] which allows top-down feedback to suppress irrelevant stimuli or incorrect interpretations, an effect we observe qualitatively in our recurrent model near object boundaries and in cluttered regions with many small objects.

We train this recurrent adaptive pooling CNN architecture end-to-end and show that it matches state-of-the-art segmentation performance on three large-scale datasets, as shown in Table 1, using a model which, thanks to recurrent computation, is substantially more compact than many existing approaches, as shown in Table 1. Moreover, the monocular depth estimates produced by our gating

Southern California Machine Learning Symposium 2017, University of Southern California.

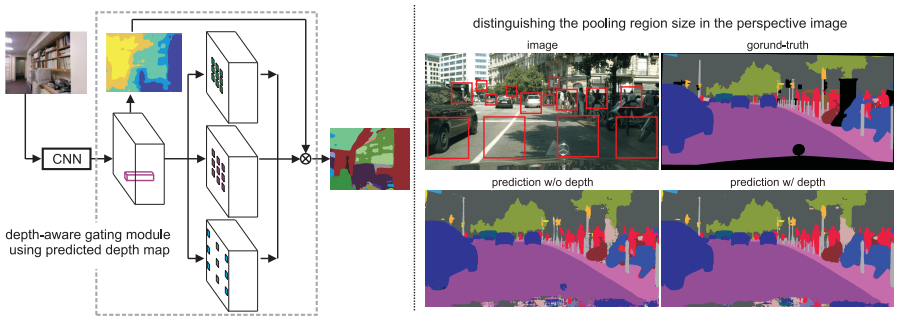

Figure 2: Left: depth-aware gating module using the predicted depth map. Right: an example image with ground-truth, segmentation result with and without the depth gating module. Rectangles overlayed on the image indicate pooling field sizes which are adapted based on the local depth estimate. We quantize the depth map into five discrete scales in our experiments. Using depth-gated pooling yields more accurate segment label predictions by avoiding pooling across small multiple distant objects while simultaneously allowing using sufficiently large pooling fields for nearby objects.

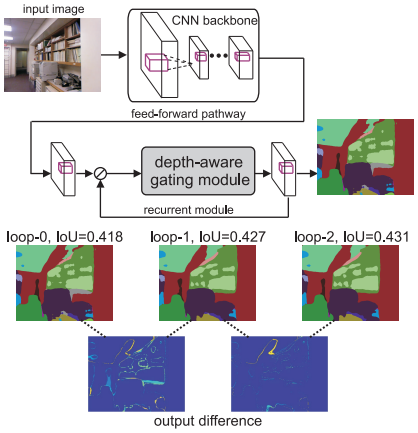

Figure 3: The input to our recurrent module is the concatenation (denoted by ⊘) of the feature map from an intermediate layer of the feed-forward pathway with the prior recurrent prediction. Our recurrent module utilizes depth-aware gating which carries out both depth regression and quantized prediction. Updated depth predictions at each iteration gate pooling fields used for semantic segmentation. This recurrent update of depth estimation increases the flexibility and representation power of our system yielding improved segmentation. We illustrate the prediction prior to, and after two recurrent iterations for a particular image. We also visualize the difference in predictions between consecutive iterations and note the gains in accuracy at each iteration as measured by average intersection-over-union (IoU) benchmark performance.

channel yield state-of-the-art performance on the NYU-depth-v2 benchmark. Fig. 1 shows some results of segmentation and depth estimation.

Table 1: Performance of semantic segmentation on different datasets. Results marked by * are evaluated by the dataset server on test set. Note that we train our models based on ResNet50 architecture on NYU-depth-v2 and SUN-RGBD datasets, and ResNet101 on the Cityscapes dataset.

	NYU-depth-v2		SUN-RGBD		Cityscapes
	IoU	pixel acc.	IoU	pixel acc.	IoU
baseline	0.406	0.703	0.402	0.776	0.738
w/ gt-depth	0.413	0.708	0.422	0.787	0.753
w/ pred-depth	0.418	0.711	0.423	0.789	0.759
loop1 w/o depth	0.419	0.706	0.432	0.793	0.762
loop1 w/ gt-depth	0.425	0.711	0.439	0.798	0.769
loop1 w/ pred-depth	0.427	0.712	0.440	0.798	0.772
loop2	0.431	0.713	0.443	0.799	0.776
loop2 (test-aug)	0.445	0.721	0.451	0.803	0.791 / 0.782*
Context [3]	0.406	0.700	0.423	0.784	- / 0.716*

References

[1] Beck, D. M. and Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision research*, **49**(10), 1154–1165.

[2] Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, **17**(3), 455–462.

[3] Lin, G., Shen, C., van den Hengel, A., and Reid, I. (2016). Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*.