Towards Unified Depth and Semantic Prediction from a Single Image

Peng Wang[1], Xiaohui Shen[2], Zhe Lin[2], Scott Cohen[2], Brian Price[2], Alan Yuille[1]
[1]University of California, Los Angeles. [2]Adobe Research.

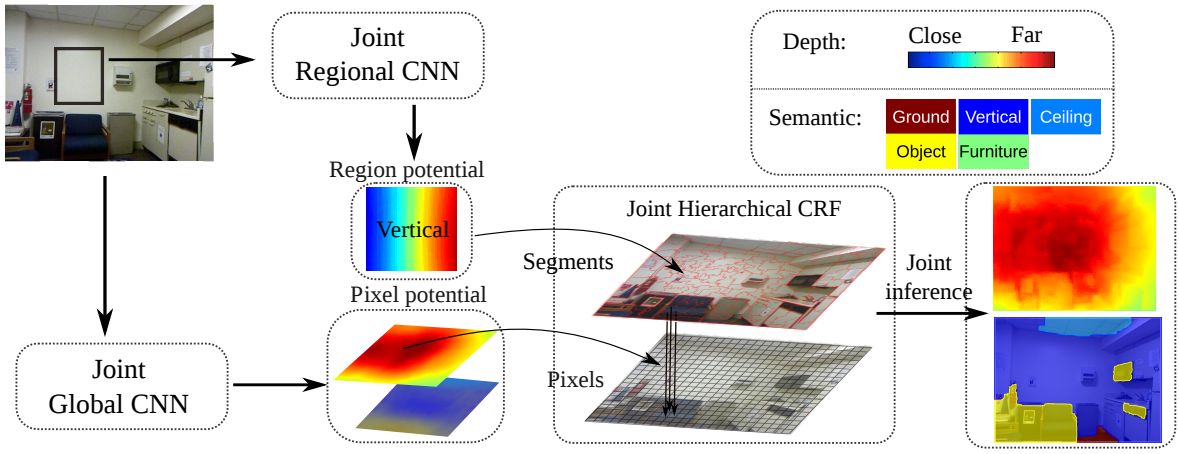

Figure 1: Framework of our approach for joint depth and semantic prediction. Given an image, we obtain region-wise and pixel-wise potential from a regional and a global CNN respectively. The final results are jointly inferred through the Hierarchical CRF.

Depth estimation and semantic segmentation from a single image are two fundamental yet challenging tasks in computer vision. While the two tasks are strongly correlated and mutually beneficial, they are usually solved separately or sequentially. Motivated by the complementary properties of the two tasks, in this paper we propose a unified framework for joint depth and semantic prediction. In addition, our approach also combines global and local information, which achieves long range context to avoid local ambiguity, and use local prediction to maintain the details in the image. Through extensive evaluations, we demonstrate our framework effectively leverages the benefits of the two tasks and achieves the state-of-the-art results.

Fig. 1 illustrates the framework of our approach. We formulate the joint inference problem in a two-layer Hierarchical Conditional Random Field (HCRF). The unary potentials in the bottom layer are pixel-wise depth values and semantic labels, which are predicted by a joint Convolutional Neural Network (CNN) trained using end-to-end strategy. The unary potentials in the upper layer are region-wise depth and semantic maps, which are computed from a joint regional CNN-based regressor trained on local segments. The output of the global CNN, though coarse, provides very accurate global scale and semantic guidance, while the local regressor gives more details in depth and semantic boundaries. In the Joint HCRF, we induce the pairwise terms performing local smoothness through local appearance similarity and interaction between depth and semantic labels such as enforcing depth smoothness for semantic labels like ground and ceiling, while encouraging semantic dissimilarity when there is depth discontinuity. In this framework, the mutual interactions between depth and semantic information are captured through the joint training of the CNNs, and further enforced in the joint inference of HCRF.

We evaluated our method on the NYU v2 dataset [4] on both depth estimation and semantic segmentation. By inference using our joint global CNN, the depth prediction improves over the depth only CNN by an average 8% relative gain, and also outperforms the state-of-the-art. After incorporating local predictions, the final depth maps produced by the HCRF are significantly improved in terms of visual quality, with much clearer structures and boundaries. Meanwhile in semantic segmentation, our joint HCRF approach outperforms R-CNN [2] that is currently known to be the most effective method for semantic segmentation, by 10% relatively in average IOU. In Fig. 2, at first row, we show a comparison result between ours and other two state-of-the-art methods, i.e. DC Depth [3] and DCNN [1]. DC Depth uses small local segments which suffers from local distortions due to lack of global cues. DCNN does not have the constraint from semantic, thus the prediction may be negatively influenced by appearance variation, e.g. the two black couches in the image. At second row, we show a qualitative result predicted from joint global CNN to the joint HCRF that demonstrates the effectiveness of both CNN and HCRF.

[1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*. 2014.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[3] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *CVPR*, 2014.

[4] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV (5)*, pages 746–760, 2012.

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

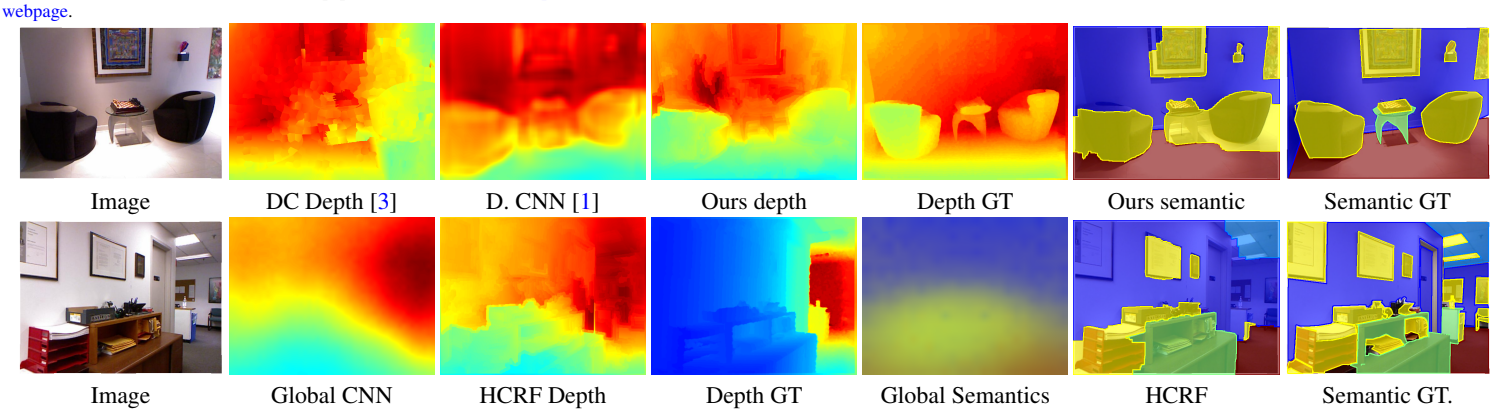

Figure 2: First row: a comparison result between ours and other state-of-the-art methods. Second row: an example illustrating the prediction of our joint global CNN and joint HCRF.