Designing Deep Networks for Surface Normal Estimation

Xiaolong Wang[1], David F. Fouhey[1], Abhinav Gupta[1],
[1]Robotics Institute, Carnegie Mellon University

Input Image Surface Normal (Output) Input Image Surface Normal (Output)

Figure 1: Sample results. Notice our method predicts not only the coarse structure correctly, but also many of the details (e.g., the legs of the coffee table). On the right, the chair surface and legs and even the top of the shopping bags are captured correctly. Normal legend: blue → X; green → Y; red → Z.

Convolutional neural networks (CNNs) have shown incredible promise in learning representations for visual tasks such as scene classification and object detection. But their performance tasks such as 3D scene understanding has been not as extensively studied. In this paper, we want to explore the effectiveness of CNNs on the task of predicting surface orientation, or surface normals, from a single image. One could treat this as per-pixel regression and directly apply a CNN. However, decades of research has shown that the output space of this task is governed by powerful physical constraints.

In this paper, we demonstrate how to incorporate insights about 3D representation and reasoning into a deep learning framework for surface normal prediction. While CNNs have been particularly successful for learning image representations, we believe their design can benefit from past research in 3D scene understanding. Thus, rather use a standard feed-forward network, our network structure (see Fig. 2), incorporates the following insights:

1. **Global and local:** we include complementary global and local networks. The global network predicts coarse layout from the whole image and the local network operates in a sliding-window fashion, considering local evidence. Their competing predictions are fed into a final fusion network that arbitrates between the two. This fusion network can be seen as a form of learned reasoning.

2. **Man-made constraints:** our global network is also trained to predict a box-layout, and we estimate vanishing points in the input images. The box-layout prediction as well as the coarse predictions snapped to the vanishing points are passed on to the fusion network.

3. **Local structure:** our local network is trained to predict classic line-labeling categories, and its predictions are fed to the fusion network.

Surface normal prediction is a structured problem; we reduce it to a classification problem with the coding scheme introduced in [3].

We validate our method on the NYUv2 dataset [4], training with the provided raw video data and adopting the per-pixel evaluation protocol introduced by [1]. We show some qualitative results in Fig. 1: our method not only captures the coarse structure but also many of the fine details. We show additional results in Fig. 3: notice how the seats of chairs, and many items on dressers are predicted correctly and that there are crisp boundaries between surfaces with different orientations (e.g., between the tops and sides of counters and tables).

We report results in Table 1 using [3]'s ground-truth; our method substantially out-performs the state-of-the art by as much as 9%. Our full experiments show that: (1) our approach outperforms the state-of-the-art by a large margin; (2) our full model with constraints yields a 7% improvement over a standard feed-forward CNN; and (3) our physical constraints give us a 5% boost in the most strict evaluation metric.

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

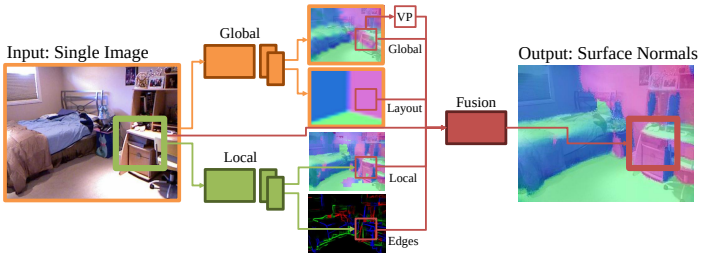

Figure 2: **Approach Overview.** We process the image with global and local CNNs while including past insights into 3D representation. A fusion CNN considers conflicting evidence to yield a final prediction, enabling the soft application of constraints like the Manhattan-world assumption.

Table 1: Our approach compared with the state-of-the-art. We improve on the state-of-the-art in every error metric.

	Ours	UNFOLD [2]	Discr. [3]	3DP (MW) [1]
Median	**14.8**	17.9	23.1	19.2
Mean	**26.9**	35.2	33.5	36.3
% < 11.25°	**42.0**	40.5	27.7	39.2
% < 30°	**68.2**	58.9	58.7	57.8

Input Image Output Input Image Output

Figure 3: Sample results of our method. Notice the details predicted in the output, including items on dressers and crisp boundaries between surfaces.

[1] David F. Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3D primitives for single image understanding. In *ICCV*, 2013.

[2] David F. Fouhey, Abhinav Gupta, and Martial Hebert. Unfolding an indoor origami world. In *ECCV*, 2014.

[3] Lubor Ladický, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*, 2014.

[4] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.