Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs

Bo Li[1,2], Chunhua Shen[2,3], Yuchao Dai[4], Anton van den Hengel[2,3], Mingyi He[1]
[1] Northwestern Polytechnical University, China; [2] University of Adelaide, Australia; [3] Australian Centre for Robotic Vision; [4] Australian National University

Predicting the depth (or surface normal) of a scene from single monocular color images is a challenging task. This paper tackles this challenging and essentially under-determined problem by regression on deep convolutional neural network (DCNN) features, combined with a post-processing refining step using conditional random fields (CRF). Our framework works at two levels, super-pixel level and pixel level. First, we design a DCNN model to learn the mapping from multi-scale image patches to depth or surface normal values at the super-pixel level. Second, the estimated super-pixel depth or surface normal is refined to the pixel level by exploiting various potentials on the depth or surface normal map, which includes a data term, a smoothness term among super-pixels and an auto-regression term characterizing the local structure of the estimation map. The inference problem can be efficiently solved because it admits a closed-form solution. Experiments on the Make3D and NYU Depth V2 datasets show competitive results compared with recent state-of-the-art methods.

1 Multi-scale regression via transferring network

To encode the depth and surface normal, we use deep network and formulate the estimation as a regression problem, where the relationship between image patch and its corresponding depth or surface normal is estimated through a new deep network built upon the network architecture of Krizhevsky et al [3].

Our deep network architecture is illustrated in Fig. 1. The first part, namely "shared weights", is initialized by the pre-trained net of [2] and kept unchanged during training. The input of our network is the multi-scale patch blocks. We add an concatenation layer to fuse the multi-scale information.

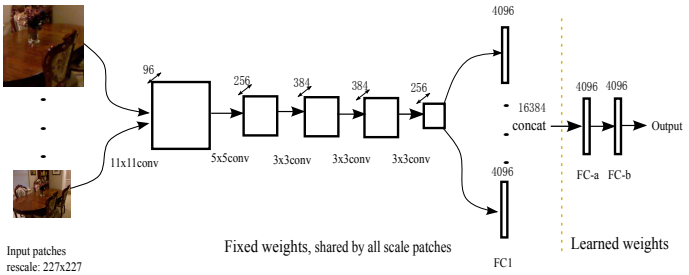

Figure 1: architecture of our CNN model

For both task, the Euclidean loss function is used,

$$\mathbf{E} = \frac{1}{2N}\sum_{i=1}^{N} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2, \tag{1}$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ are the ground truth and regressed value respectively. In our paper, $\mathbf{x}_i$ could be depth value or surface normal.

2 Refining via Hierarchical CRF

Our refining cost function 2 consists of three terms, namely the data term, super-pixel smoothness term and auto-regression term at pixel level. As for the normal vector refining, we transfer the normal vector to the spherical coordinate (θ,ϕ) and refine them respectively. Here, let $\mathcal{X}$ be the set of variable which could be the depth or surface normal map.

$$\mathbf{E}(\mathcal{X}) = \sum_{i\in\mathcal{S}} \phi_i(x_i) + \sum_{(i,j)\in\mathcal{E}_\mathcal{S}} \phi_{ij}(x_i,x_j) + \sum_{\mathcal{C}\in\mathcal{P}} \phi_\mathcal{C}(\mathbf{x}_\mathcal{C}) \tag{2}$$

where $\mathcal{E}_\mathcal{S}$ denotes the set of pairs of super-pixels that share a common boundary and $\mathcal{P}$ is the set of patches(size of 3×3) designed on the pixel level.

Potential 1: Unary term

$$\phi_i(x_i) = (x_i - \overline{x_i})^2, \tag{3}$$

where $\overline{x_i}$ denotes the depth regression result from our multi-scale deep network, this term is defined at the super-pixel level, measuring the quadratic distance between the estimated depth x_i and regressed depth $\overline{x_i}$.

Potential 2: Smoothness at super-pixel level

$$\phi_{ij}(x_i,x_j) = w_1\left(\frac{x_i - x_j}{\lambda_{ij}}\right)^2, \tag{4}$$

The quadratic distance is weighted by λ_{ij}, the color difference between connected super-pixels in LUV color space [1].

Potential 3: Auto-regression model

Here we use the auto-regression model to characterize the local correlation structure in the depth map. The key hypothesis of the auto-regression model is that neighboring pixels with similar intensities should have similar depth or angle [4, 5]. The auto-regression potential can be expressed as:

$$\phi_\mathcal{C}(\mathbf{x}_\mathcal{C}) = w_2\left(x_u - \sum_{r\in\mathcal{C}/u} \alpha_{ur} x_r\right)^2. \tag{5}$$

where $\mathcal{C}$ is the neighbourhood of pixel u and α_{ur} denotes the model auto-regression coefficient for pixel r in the neighbourhood $\mathcal{C}$.

A closed form solution Once the parameters in our Hierarchical CRF are determined, the MAP solution can be obtained in closed form solution. For convenience of expression, we express the energy function Eq. (2) in matrix form:

$$\mathbf{E}(\mathbf{x}) = \|\mathbf{H}\mathbf{x} - \overline{\mathbf{x}}\|_2^2 + w_1\|\mathbf{Q}\mathbf{H}\mathbf{x}\|_2^2 + w_2\|\mathbf{A}\mathbf{x}\|_2^2, \tag{6}$$

As the energy function is quadratic with respect to $\mathbf{x}$, a closed-form solution can be derived algebraically:

$$\mathbf{x}_{map} = (\mathbf{H}^\top\mathbf{H} + w_1\mathbf{H}^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{H} + w_2\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{H}^\top\overline{\mathbf{x}}. \tag{7}$$

[1] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[2] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[4] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM Transactions on Graphics (TOG)*, 2004.

[5] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*. 2012.

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