Pattern Recognition 83 (2018) 430–442

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# Deep hierarchical guidance and regularization learning for end-to-end depth estimation

Zhenyu Zhang, Chunyan Xu*, Jian Yang*, Ying Tai, Liang Chen

*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, PR China*

ARTICLE INFO

*Article history:*
Received 3 March 2017
Revised 1 May 2018
Accepted 13 May 2018

*Keywords:*
Depth estimation
Multi-regularization
Deep neural network

ABSTRACT

In this work, we propose a novel deep Hierarchical Guidance and Regularization (HGR) learning framework for end-to-end monocular depth estimation, which well integrates a hierarchical depth guidance network and a hierarchical regularization learning method for fine-grained depth prediction. The two properties in our proposed HGR framework can be summarized as: (1) the hierarchical depth guidance network automatically learns hierarchical depth representations by supervision guidance and multiple side conv-operations from the basic CNN, leveraging the learned hierarchical depth representations to progressively guide the upsampling and prediction process of upper deconv-layers; (2) the hierarchical regularization learning method integrates various-level information of depth maps, optimizing the network to predict depth maps with similar structure to ground truth. Comprehensive evaluations over three public benchmark datasets (including NYU Depth V2, KITTI and Make3D datasets) well demonstrate the state-of-the-art performance of our proposed depth estimation framework.



## 1. Introduction

Estimating depth from monocular RGB images is a fundamental problem in computer vision and many tasks can benefit from the depth information such as scene understanding [1], 3D modeling [2,3], robotics [4,5], action recognition [6], etc. The depth of a scene can be inferred from stereo cues [7,8]. However, when it comes to monocular scenes, depth estimation becomes actually an inherently ambiguous and ill-posed problem since a given image may correspond to an infinite number of possible world scenes [9]. To deal with the depth estimation problem, existing previous works often rely on strong priors and focus on the geometric knowledge [10–14]. Other works benefitting from RGB-D data show an improvement on dense depth maps estimation [15–18]. Additionally, some efforts [19,20] have been made to leverage the labeling information which also contributes to depth estimation tasks. However, these above works rely on hand-craft features, strong priors and pre-processing to solve the depth estimation task, which have no universality for different real-world scenes.

In recent years, deep learning methods have reached a big breakthrough on computer vision tasks [21] such as image classification [22–24], semantic segmentation and scene parsing [25–27] and pose estimation [28]. Efforts based on CNNs have also been successfully introduced to monocular depth estimation tasks [9,29–31], which significantly improve the estimation performance. However, these methods still rely on multi-stage processing pipelines, e.g., super-pixel and CRF, respectively trained multi-scale networks. Moreover, their predictions are not sufficiently high resolution. Motivated by these witnesses, a more straightforward method for high resolution pixel-wise prediction is necessary.

Previous works often employ feedforward networks [9,29,30,34] for the problem of depth estimation, which do not well consider multi-scale features of previous lower layers with abundant image detail information. The recent work [34] proposed a deeper network architecture for depth estimation and obtained good performance, however, the predictions of their method are too smooth and lack details, which is caused by the pooling operations. Eigen and Fergus [32] have adopted skip connections to introduce multi-scale information in process of estimating depth maps, but these operations also introduce much noise to the final prediction. As illustrated in Fig. 1, the predictions of [32] contain inaccurate geometric structures and coarse boundaries, especially at the location of the table lamp. The more recent work [33] introduced the detailed information using multi-Scale continuous CRFs with a sequential deep networks, but the predictions of their method also contain much noise. As illustrated in Fig. 1, the predictions of [33] obtains much obvious discontinuity

* Corresponding authors.
*E-mail addresses:* cyx@njust.edu.cn (C. Xu), csjyang@njust.edu.cn (J. Yang).

https://doi.org/10.1016/j.patcog.2018.05.016
0031-3203/

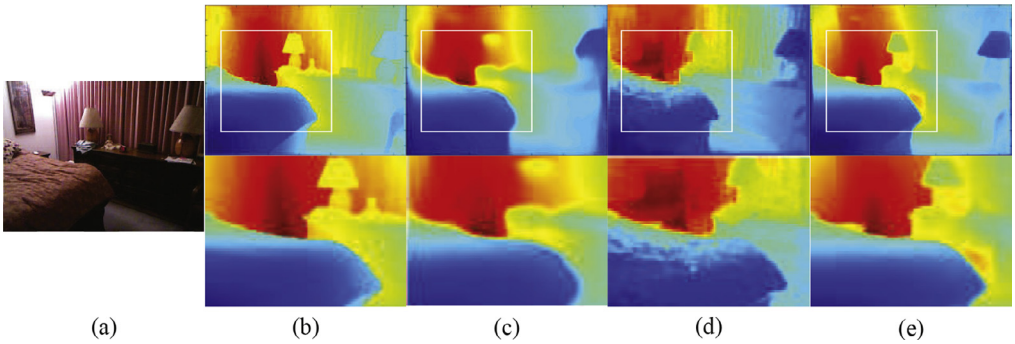


**Fig. 1.** Illustration of predicted depth maps with different methods. (a) input RGB image; (b) ground truth; (c) depth maps predicted by Eigen and Fergus [32]; (d) predictions of [33]; (e) predictions of our HGR framework. The predictions of our HGR framework are more fine-grained and detailed.

and noise at the location of bed, walls and table lamp. The main problem above is that the semantic information contained in image details is much different from the depth information, which we call a "semantic gap". For example, in Fig. 1, the regions of bed and curtain in the RGB image contain much texture information because of rivel, and the pedestal of the table lamps has similar color compared with the curtain. However, in the depth map, the depths of bed and curtain are smooth and continuous without much texture, but the pedestal has very different depths compared with the curtain. As a result, the texture and detailed information of RGB images may not be always proper and useful for depth estimation because of such semantic gap. As the semantic gap is large, the successful method [35] in super-resolution or other approaches in semantic segmentation [36,37] and edge detection [38] may not be suitable for depth estimation task. Inspired by these witnesses, we argue that a feedforward network without lower feature guidance may be difficult to better predict depth maps with fine-grained details, and meanwhile the lower feature guidance may not be suitable to directly used in depth estimation tasks. We propose a novel hierarchical network architecture which is significantly suitable for depth estimation or other dense map prediction task. This proposed network architecture integrates hierarchical features to progressively refine the estimation process and guide the network to produce fine-grained prediction. Further, it solves the 'semantic gap' problem between image texture and depth information, improving the performance of the final prediction. As illustrated in Fig. 1, the predictions of our approach contain fine-grained details and match the geometric structure exactly. To our knowledge, this effort is first been made in depth estimation task.

While estimating depth maps, the depth structure information is of great importance. For example, depth values are similar on the positions of an object, but possibly containing huge variations on the surroundings of its boundaries. Previous works attempt to directly extract this depth structure information from RGB images by super-pixel CRF models [29–31] or using hand-craft feature (HSV color histogram) based local image patch correlation [39]. However, these methods only utilize pixel-level depth information, failing to predict accurate depth maps on the textured images containing many objects. In the work of [9], Eigen *et al.* present a scale invariant loss which leverages the depth relations of different pixels to optimize their network. This is a more direct way to use depth information, but actually scale invariant loss introduces no region-level depth structure and its advantages are not demonstrated experimentally. Due to these limitations of considering less depth structure information, the above approaches are difficult to improve the performance.

In this paper, we propose a novel deep Hierarchical Guidance and Regularization (HGR) framework for depth estimation, which well integrates multi-scale depth semantic features and various-level information of depth maps to guide the estimation processing. Given an input monocular RGB image, our method can predict a correspondingly-sized depth map in an end-to-end way, as illustrated in Fig. 2. The basic feedforward network mainly associates to multiple convolutional and deconvolutional operations, and the architecture fashion is based on ResNet [40]. To resolve the texture loss caused by pooling operations, we develop a hierarchical depth guidance and regularization strategy to utilize multi-level details. First, refining networks are build on side of conv-net of each scale, predicting multi-scale depth maps guided by supervisions. In this way, the features of these refining networks are highly correlative with depth information. Second, we combine the depth semantic features with upper deconv-features of the corresponding scale. This concatenation operation provides hierarchical depth guidance for the estimation process, progressively resolve the ambiguity and smoothness by hierarchical depth details. Finally, a multi-regularized network learning method is applied to optimize our network, which leverage hierarchical structure information from depth maps to guide the network optimization. The whole framework is trained in an end-to-end way, predicting depth maps without any pre-and post-processing operations.

To sum up, the major contributions of this work can be summarized as follows:

- We propose a novel hierarchical guidance network architecture for depth estimation, which well leverages hierarchical lower depth semantic features to progressively guide the predicting processing of deconv-upper layers. Then our method can well consider the semantic difference between images texture and depth details and predict depth maps with more fine details.
- A novel multi-regularized network learning strategy is introduced to optimize the parameters of our depth estimation network. This learning method employs various-level information of depth maps, which contributes to high quality estimation. We demonstrate theoretically and experimentally why it is suitable for this task.
- Depth maps are predicted end-to-end by our framework, without any pre-and post-processing operation. We demonstrate that the proposed method obtains state-of-the-art depth estimation performance over three public benchmark datasets (including NYU Depth V2, KITTI and Make3D datasets), especially in outdoor scene datasets.