Indoor Scene Structure Analysis for Single Image Depth Estimation

Wei Zhuo , Mathieu Salzmann , Xuming He , and Miaomiao Liu

Australian National University, Canberra, Australia
NICTA, Canberra, Australia
{wei.zhuo, mathieu.salzmann, xuming.he, miaomiao.liu}@nicta.com.au

Abstract

We tackle the problem of single image depth estimation, which, without additional knowledge, suffers from many ambiguities. Unlike previous approaches that only reason locally, we propose to exploit the global structure of the scene to estimate its depth. To this end, we introduce a hierarchical representation of the scene, which models local depth jointly with mid-level and global scene structures. We formulate single image depth estimation as inference in a graphical model whose edges let us encode the interactions within and across the different layers of our hierarchy. Our method therefore still produces detailed depth estimates, but also leverages higher-level information about the scene. We demonstrate the benefits of our approach over local depth estimation methods on standard indoor datasets.

1. Introduction

Without any prior information, estimating the depth of a scene from a single image is a highly ambiguous problem. Humans, however, can easily perceive depth from a static monocular input, thanks to the data and knowledge they accumulated over the years. Intuitively, this suggests that learning from existing image-depth pairs should make single image depth estimation a realistic, achievable goal.

This observation has been the motivation for several recent approaches to monocular depth estimation [25, 26, 20, 15, 21, 18, 5]. These methods, however, typically model depth only at a local scale. For instance, [18] predict the depth of each pixel individually. While, in contrast, [25, 26, 20, 15, 21] encode some higher-level information by modeling the relationships of neighboring superpixels, the resulting methods still lack reasoning about the global structure of the scene. This contradicts our intuition that humans exploit such higher-level scene structure to apprehend their environment.

Recovering the structure of a scene has nevertheless been studied in the past [13, 19, 9, 8, 12, 28, 7]. The resulting methods typically represent the scene of interest at a coarse scale. As a consequence, they fail to provide a detailed description of the scene. More importantly, while these methods indeed infer the scene structure, they do not yield an absolute depth estimate; typically, only normals are predicted by these techniques, which leaves at least a global scale ambiguity for depth.

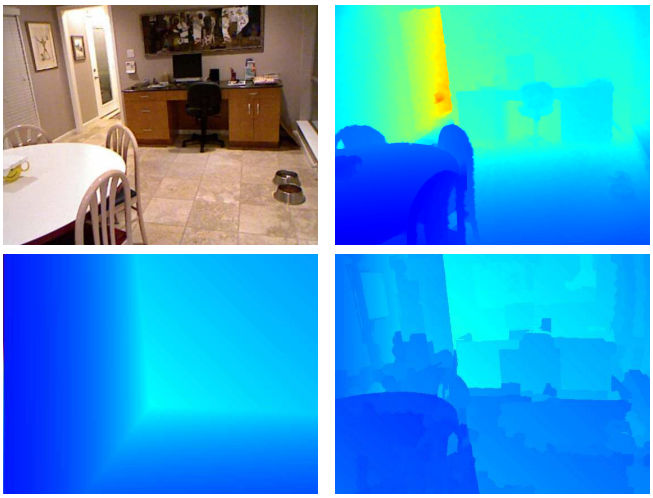

Figure 1. **Depth estimation from a single image:** (Top) Image and ground-truth depth map. (Bottom) Estimated layout and detailed depth map. Color indicates depth (red is far, blue is close).

In this paper, we propose to exploit high-level scene structure for detailed single image depth estimation. To this end, we introduce an approach that relies on a hierarchical representation of the scene depth encoding local, mid-level and global information. This lets us model the detailed depth of a scene while still benefiting from information about its global structure.

More specifically, our hierarchical representation of the scene depth consists of three layers: superpixels, regions and layout. The superpixels allow us to model the local

depth variations in the scene. In contrast, the regions and layout let us account for mid- and large-scale scene structures. We model the depth estimation problem with a Conditional Markov Random Field (CRF) with variables for each layer in our hierarchy. This CRF allows us to encode interactions within and across layers, and thus to effectively exploit local and global information jointly. As illustrated in Fig. 1, inference in our model therefore yields depth estimates ranging from coarse to fine levels of details.

We demonstrate the effectiveness of our method on two standard indoor datasets. Our experiments evidence the benefits of exploiting higher-level scene structure over local depth estimation methods.

2. Related Work

In contrast to classical multiview approaches to 3D scene reconstruction, single image depth estimation has gained popularity only recently. Nonetheless, in a few years, great progress has been made on this challenging task.

Due to the ambiguities inherent to the problem, existing methods rely on training data (i.e., image-depth pairs). In such a scenario, a natural approach is to learn regressors to predict local depth. This approach was employed in [20], where a specific regressor from image features to pixel depth was trained for each semantic class in the data. Following a related idea, [18] trained classifiers for specific semantic labels at some chosen canonical depths. These classifiers were then employed to predict pixel depth.

Several methods have proposed to go beyond purely local depth estimation. For instance, [4] introduced an approach based on sparse coding to directly predict the depth of the entire scene. Similarly, [5] trained a deep network to predict the pixel depth of a whole image. However, to respect fine details, such global scene prediction methods require large amounts of data. By contrast, many techniques favor modeling the relationships between neighboring (super)pixels to encourage coherence across the image. With the exception of [15, 16, 17] that formulate depth recovery as a purely continuous optimization problem, such coherence is typically encoded in a graphical model. This approach was introduced by [25, 26] with relatively simple relationships between the superpixels. A simple smoothness term was also employed in [20] together with local geometric reasoning and the previously mentioned regression as data term. In [21], additional discrete variables were employed to model more complex superpixel relationships, thus yielding a higher-order discrete-continuous graphical model. Despite the reasoning about neighbor interactions, all the above-mentioned models fail to consider the global structure of the scene, which provides important cues for depth estimation.

Estimating the structure of a scene has itself been an active area of research in recent years. For instance, [31] modeled structure in a coarse manner as the absolute mean depth of the scene. To model more detailed structure, much work originated from the idea of geometric context introduced in [13]. This idea was extended to predict the layout of indoor scenes with a box model, thus relying on the Manhattan world assumption [12, 28]. Instead of a box model, [22] classified a scene into 15 geometry categories to represent its structure. A more accurate representation was proposed in [19], which produces sparse surface normals. Similarly, in [6], local normals were predicted, but by exploiting Exemplar SVMs on regions found to be discriminative. Recently, [7] improved such normal estimation by making use of a CRF and reasoning about normal discontinuities. One of the main drawbacks of these scene structure analysis methods is that they do not truly estimate depth, but only normals, thus leaving at least one global scale ambiguity, and often more since the relative ordering of surface regions with different normals may not be defined by the recovered structure.

Here, we propose to leverage high-level scene structure for detailed depth estimation. As such, while inspired by the work of [21], our formulation models a much more complete scene representation, which encompasses a hierarchy of local, mid-level and global cues. As evidenced by our results, single image depth estimation benefits from such higher-level reasoning.

3. Structure-Aware Depth Estimation

We now introduce our hierarchical model to perform single image depth estimation. As mentioned earlier, depth estimation is expressed as inference in a CRF, which allows us to encode relationships within and between the different layers in our hierarchy. To this end, let us denote by Y, R and L the variables that represent local depth, mid-level and global structures, respectively. Inference is achieved by maximizing the joint distribution of our CRF, or equivalently minimizing the energy

$$E(Y, R, L) = E_l(Y) + E_m(Y, R) + E_g(Y, L) \ , \qquad (1)$$

where each individual energy term corresponds to a particular layer in our model. In the remainder of this section, we describe these different terms in details.

3.1. Local Depth Estimation

To estimate detailed depth, our model relies on image superpixels. Each superpixel is represented as a plane in 3D, which translates the depth estimation problem into finding the best plane parameters for each superpixel. In particular, here, we encode each plane with the depth of its centroid and its normal direction.

More specifically, let $Y = \{y_1, y_2, ..., y_{N_s}\}$ be the set of discrete variables representing N_s superpixels in an image,

where each y_i can take values from a discrete state space $\mathcal{S}$. We define this state space by quantizing the range of valid depths for the superpixel centroid into V values, with the range determined from maximum and minimum depths of the training data. Furthermore, we make use of the Manhattan world assumption, and restrict the superpixels normal direction to 3 possible dominant directions, defined by the vanishing point estimation method of [24]. This lets us define the first energy term in Eq. 1 as

$$E_l(Y) = \sum_p \phi_p(y_p) + \sum_{p,q} \phi_{p,q}(y_p, y_q) \ , \qquad (2)$$

where ϕ_p is a unary potential encoding the cost of assigning label y_p to superpixel p, and $\phi_{p,q}(y_p, y_q)$ is a pairwise potential encouraging coherence across the superpixels.

The unary potential is based on the regression term in [21]. To this end, we first retrieve K candidate training images similar to the input image by nearest-neighbor search based on a combination of distances on GIST, PHOG and ObjectBank features (i.e., L_2 distance for GIST and ObjectBank, and χ^2-distance for PHOG). For each superpixel in the input image, we then compute the plane parameters of the corresponding area in each candidate, and use Gaussian Process (GP) regressors to predict the plane parameters of the superpixel of interest from these plane parameters (i.e., one regressor for each of the four plane parameters). The GP regressors rely on an RBF kernel, and were trained in a leave-one-image-out manner from the training data. Let $d^i_{r,p}$ be the depth of the i^{th} pixel of superpixel p, estimated from the regression results. We define our unary potential as

$$\phi_p(y_p) = \frac{1}{N_p} \sum_{i=1}^{N_p} \left(d^i_p(y_p) - d^i_{r,p} \right)^2 \ , \qquad (3)$$

where N_p is the number of pixels in superpixel p and d^i_p is the depth of pixel i in superpixel p for a particular state y_p.

The pairwise term $\phi_{p,q}$ relies on an occlusion classifier trained of the features of [14]. Given the predicted occlusion label o_{pq} for the boundary between two neighboring superpixels p and q, this potential is expressed as

$$\phi_{p,q}(y_p, y_q) = w_l \cdot \begin{cases} 0 & \text{if } o_{pq} = 1 \\ g_{pq} \|\mathbf{n}_p(y_p) - \mathbf{n}_q(y_q)\|^2 + \\ \frac{1}{N_{pq}} \sum_{j=1}^{N_{pq}} (d^j_p(y_p) - d^j_q(y_q))^2 & \text{if } o_{pq} = 0 \end{cases} \qquad (4)$$

where N_{pq} is the number of pixels shared by superpixels p and q, $\mathbf{n}_p(y_p)$ is the normal corresponding to a particular state y_p, and g_{pq} is a weight based on the image gradient on the boundary of the superpixels, i.e., $g_{pq} = \exp(-\mu_{pq}/\sigma)$ with μ_{pq} the mean gradient on the boundary.

While inspired by [21], the energy described above includes at most pairwise terms, and therefore allows us to perform inference more efficiently. Importantly, however, this energy still reasons at a local level. Next, we present our approach to incorporating higher-level scene structures via the additional terms in Eq. 1.

3.2. Exploiting Mid-level Structures

The superpixels employed above are typically quite small and therefore encode little information about the scene. As a consequence, not only do they encode little structure, but one also cannot reliably exploit their appearance to help depth prediction. Thus only their location in the candidate images retrieved using global image descriptors is utilized in the previous model. To better exploit appearance and encode more information about the scene structure, here we propose to make use of larger regions.

To this end, let $R = \{r_i, r_2, ..., r_{N_r}\}$ be the set of discrete variables representing N_r regions extracted from the input image, where each r_i can be assigned a value from the same state space $\mathcal{S}$ as the superpixel variables $\{y_p\}$. We define the second term in Eq. 1 as

$$E_m(Y, R) = \sum_\gamma \phi_\gamma(r_\gamma) + \sum_{\gamma,p} \phi_{\gamma,p}(r_\gamma, y_p) \ , \qquad (5)$$

where ϕ_γ is a unary potential on the region variables, and $\phi_{\gamma,p}$ a pairwise potential accounting for the interactions of the regions and the superpixels.

Since our regions are much larger than our superpixels, their appearance is also more discriminative. Therefore, we follow a feature-based nonparametric approach inspired by [30] to define the unary term ϕ_γ. In particular, we first retrieve K_r candidate training images by nearest neighbor search using image-level GIST, PHOG and ObjectBank features. Here, we select the K_r images based on their best rank after nearest-neighbor search on each feature type individually. We found this strategy to be more reliable than combining the features for large retrieval sets. For each region in the input image, we compute region-level features [1] and retrieve K_c nearest-neighbor regions from the candidate image pools for each feature type, after pruning the ones that are too distant from and with too dissimilar sizes to the query region. Each superpixel in each retrieved region then votes for a centroid depth and a normal orientation in a V-dimensional and a 3-dimensional histogram, respectively. Let us denote by $P_d(d)$ and $P_n(\mathbf{n})$ the resulting normalized histograms and $P_{dn}(d, \mathbf{n})$ the bin-wise product of $P_d(d)$ and $P_n(\mathbf{n})$ (i.e., a 3V-dimensional histogram). We express the unary term in Eq. 5 as

$$\phi_\gamma(r_\gamma) = w_m \cdot \left(\max \left(P_{dn}(d(r_\gamma), \mathbf{n}(r_\gamma)) \right) - P_{dn}(d(r_\gamma), \mathbf{n}(r_\gamma)) \right) ,$$

[1] We used the same 20 features as in [30].

where $d(r_\gamma)$ is the centroid depth corresponding to the state r_γ, and similarly for the normal direction.

The pairwise term in Eq. 5 penalizes inconsistencies between the depth predicted for a region and the depth predicted for the superpixels it covers. For each superpixel in a region, this term is defined as

$$\phi_{\gamma,p}(r_\gamma, y_p) = \frac{w_{m,l}}{N_p} \sum_{i=1}^{N_p} \left(d_p^i(y_p) - d_\gamma^i(r_\gamma)\right)^2 \qquad (6)$$

where N_p is the number of pixels in superpixel p, and, with a slight abuse of notation w.r.t. index i, $d_p^i(y_p)$ and $d_\gamma^i(r_\gamma)$ represent the depth of the i^{th} pixel in superpixel p and of its corresponding pixel in region γ.

Note that the energy in this layer of our model can be thought of as encoding longer range connections between the superpixels. Importantly, however, the resulting model remains pairwise.

3.2.1 Extracting Regions

Here, we briefly describe our strategy to extract the regions acting as mid-level structures in the potentials described above. Our goal is to obtain regions that are preferably (close to) planar, of relatively uniform appearance and as large as possible. To this end, we rely on the gPb+Segmentation framework of [3].

Since our training data consists of RGB-D images, we can directly employ the RGB-D extension of the gPb+Segmentation framework, introduced recently by [10, 23]. At test time, however, we only have access to RGB images. An easy way around this problem would be to directly employ the original method of [3]. Unfortunately, the resulting regions are either highly non-planar, or too small, both of which make them ill-suited for our purpose.

To address this issue, we propose to compute the probability of a boundary by combining two different sources of information. First, we rely on the standard gPb algorithm applied to our RGB input image. As a second source of information, we make use of the estimated scene geometry in the form of the orientation map of [19]. Orientation maps assign one major normal direction to the pixels in an image. Unfortunately, these maps are sparse (i.e., not all pixels are assigned an orientation). Furthermore, for our purpose, we would not want to have all pixels with the same orientation to belong to the same region, since they could potentially belong to different surfaces. Therefore, we compute the connected components of the orientation maps, and assign a label to each pixel indicating the component it belongs to. We then apply the gPb algorithm with brightness features only to the resulting label image.

Let us denote by gPb_{rgb} and gPb_g the boundary probabilities obtained from the RGB image and the geometry image, respectively. The combined boundary probability of a pixel at location (u, v) for boundary orientation θ is then given by

$$\mathrm{gPb}_c(u, v, \theta) = (1 - \alpha)\mathrm{gPb}_{rgb}(u, v, \theta) + \alpha\mathrm{gPb}_g(u, v, \theta),$$

where, in practice, we use $\alpha = 0.5$. To obtain the final regions, we then apply the OWT-UCM method of [3] with a threshold of 0.1 on this combined boundary map. We found this combination of RGB and geometry cues to yield large, planar and uniform regions, well-suited for our approach.

3.3. Incorporating Global Structure

As a final layer in our representation, we aim to reason about the global structure of the scene, which neither the superpixels, nor the regions are able to model. To this end, we make use of the layout estimation method of [11]. This method models the geometry of an indoor scene as a box made of five surfaces (i.e., left/middle/right wall, ceiling and floor), with an additional prediction of the probability of each pixel to belong to clutter. Note, however, that the output of this method is not truly a 3D representation, in the sense that the global scale of the box is not determined.

To make use of such global structure, let us denote by L the discrete variable encoding the scale of the predicted layout, which can take value in a state space $\mathcal{L}$ representing quantized scales. The energy for the last layer in our model can be written as

$$E_g(Y, L) = \sum_p \phi_{L,p}(L, y_p) \ , \qquad (7)$$

and thus consists of a single pairwise potential that encourages coherence between the superpixels and the layout. In particular, we define this potential as

$$\phi_{L,p}(L, y_p) = \frac{w_g}{N_p} \sum_{i=1}^{N_p} (1 - P_c^i) \cdot (d_p^i(y_p) - d_L^i(L))^2 \ , \qquad (8)$$

where P_c^i denotes the probability of pixel i belonging to clutter, and, with a similar slight abuse of notation w.r.t. index i as before, $d_p^i(y_p)$ and $d_L^i(L)$ represent the depth of the i^{th} pixel in superpixel p and of its corresponding pixel in the layout. Importantly, the use of the clutter probability prevents us from oversmoothing the depth predicted by the superpixels.

Since this energy term is pairwise, so is our entire model. In our experiments, we make use of the Distributed Convex Belief Propagation (DCBP) method of [27] to perform inference in our CRF. Note that the inference results yield not only a detailed depth estimate coming from the superpixels, but also an estimate of the region depths, as well as a full 3D layout of the scene.

4. Experimental Evaluation

We evaluated our approach on two publicly available datasets: the NYUv2 depth dataset [29] and the RMRC Indoor dataset [1]. These two datasets both contain images collected from a wide variety of indoor scenes. For NYUv2, we compare our results with the state-of-art single image depth estimation methods. In particular, we consider the following three baselines:

1. **DepthTransfer [15].** This method predicts depth by transferring depth maps from similar images in the training set. These depth maps are then merged by a continuous optimization strategy that encourages smoothness across the image.

2. **DC-Depth [21].** This technique makes use of a high-order discrete-continuous CRF to estimate depth, where complex relationships between the neighboring superpixels can be encoded via discrete variables.

3. **SemanticDepth [18].** This method learns a pixel-wise classifier for each semantic class in the dataset at canonical depth. Note that this method therefore makes use of an additional source of information in the form of semantic pixel labels. Note also that it was trained on a different training/test partition from the one provided with the dataset. Therefore comparison against their results is to take with a pinch of salt.

For the sake of completeness, we also report the results of the DeepDepth method of [5]. Note, however, that this method relies on a much larger training set consisting of the 120K raw images of the NYUv2 dataset and thus should not be considered as a true baseline.

In addition to the comparison with these methods, we also perform an ablation study where we provide the results of our local model (Section 3.1), our model with mid-level structures (Sections 3.1 and 3.2), and our model with global structure, but no mid-level structures (Sections 3.1 and 3.3). We refer to these models as **Ours-local**, **Ours-mid** and **Ours-global-only**, respectively. Our complete model will be referred to as **Ours**.

For our quantitative evaluation, we report the following three standard metrics: average relative error (**rel**), average $\mathbf{log}_{10}$ error, and root mean squared error (**rms**). We also report the metrics used in [18], defined as

$$\% \text{ correct} : \left(\frac{1}{N} \sum_{u=1}^{N} [[\max(\frac{d_u}{g_u}, \frac{g_u}{d_u}) = \delta < t]] \right) \cdot 100 \ ,$$

with $t = 1.25, 1.25^2, 1.25^3$, and where g_u is the ground-truth depth at pixel u, d_u is the corresponding estimated depth, N is the total number of pixels in all the images, and $[[\cdot]]$ denotes the indicator function. Furthermore, even though normal estimation is not the main target of our approach, we report the five normal error metrics used in [6]: the **mean** and **median** angle difference between the estimated normals and the ground-truth ones, and the percentage of pixels whose angle difference w.r.t. ground-truth is below a threshold (i.e., $\theta < 11.25$, 22.5 and 30 degrees). To evaluate these metrics, the scene normals were estimated from the predicted depth maps by the method of [6].

In our experiments, the superpixels were computed using SLIC [2]. For each test image, we retrieve $K = 7$ candidates from the training images to obtain the input to the superpixel regression model. For the regions, we retrieve $K_r = 250$ candidate images. For each query region and each local features, we then obtain $K_c = 30$ candidate regions, after pruning the candidate regions whose centroid is at a distance $d > 100$ pixels from the query region centroid, and whose area ratio ($r_{area} = 2(area_a - area_b)/(area_a + area_b)$) with the query region is smaller than 0.2. When building the histogram of normal orientations, we only take into account the superpixels whose angle difference is less than 45 degree w.r.t. at least one of the three dominant normal directions in the query image. This allows us to discard the candidates that have an orientation too different from the scene in the query image.

The states of our superpixel and region variables were obtained by quantizing the depth from 0.5 to 10 by steps of 0.5 (i.e., $V = 20$). In conjunction with the 3 normal orientations, this yields 60 states for each variable. In practice, to speed up inference, we restrict the states to the 20 values with highest probability P_{dn} in the $3V$-dimensional histogram built for the region unary potential. Note that we found this to come at very little loss of accuracy in the final results. In this setting, and given the result of gPb, estimating the depth of an image containing roughly 650 superpixels takes about 2 minutes.

The parameters of our CRF (weights of the potentials) were obtained by validation on a set of 69 images taken among the training data. To this end, we followed a strategy where the potentials were incrementally added to the energy after the previous weights were determined. Note that we did not fine-tuned the weights, but mostly found the right order of magnitude of each potential among the values $\{0.1, 1, 10, 100, 1000\}$.

NYUv2:
The NYUv2 depth dataset contains 1449 pairs of aligned RGB and depth images, partitioned into 795 training images and 654 test images. These images were acquired in a variety of real-world indoor scenes. Each image was cropped to 427×561 pixels. In our evaluation, we make use of a mask in each image that only considers the ground-truth pixels with non-zero depth.

The results of our approach and of the baselines are shown in Table 1. In terms of depth accuracy, we outper-

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	mean	median	$\theta < 11.25$	$\theta < 22.5$	$\theta < 30$
DepthTransfer	0.374	0.134	1.12	49.81%	79.46%	93.75%	**43.0**	**40.5**	6.9%	23.2%	34.9%
DC-Depth	0.335	0.127	1.06	51.55%	82.32%	95.00%	45.7	42.2	19.7%	25.7%	35.4%
SemanticDepth	-	-	-	**54.22%**	82.90%	94.09%	-	-	-	-	-
Ours	**0.305**	**0.122**	**1.04**	52.50%	**83.77%**	**96.16%**	46.7	41.9	**21.1%**	**35.2%**	**41.7%**

Table 1. **NYUv2: Comparison of our approach with the baselines.** In terms of depth accuracy, we outperform the two baselines (DepthTransfer and DC-Depth) working under the same settings as us. Furthermore, we outperform the SemanticDepth approach on two out of three thresholds, despite the fact that we do not make use of any pixel label information. Recall, however, that SemanticDepth employed a different training/test partition. In terms of normal accuracy, we outperform the baselines on three out of the five metrics.

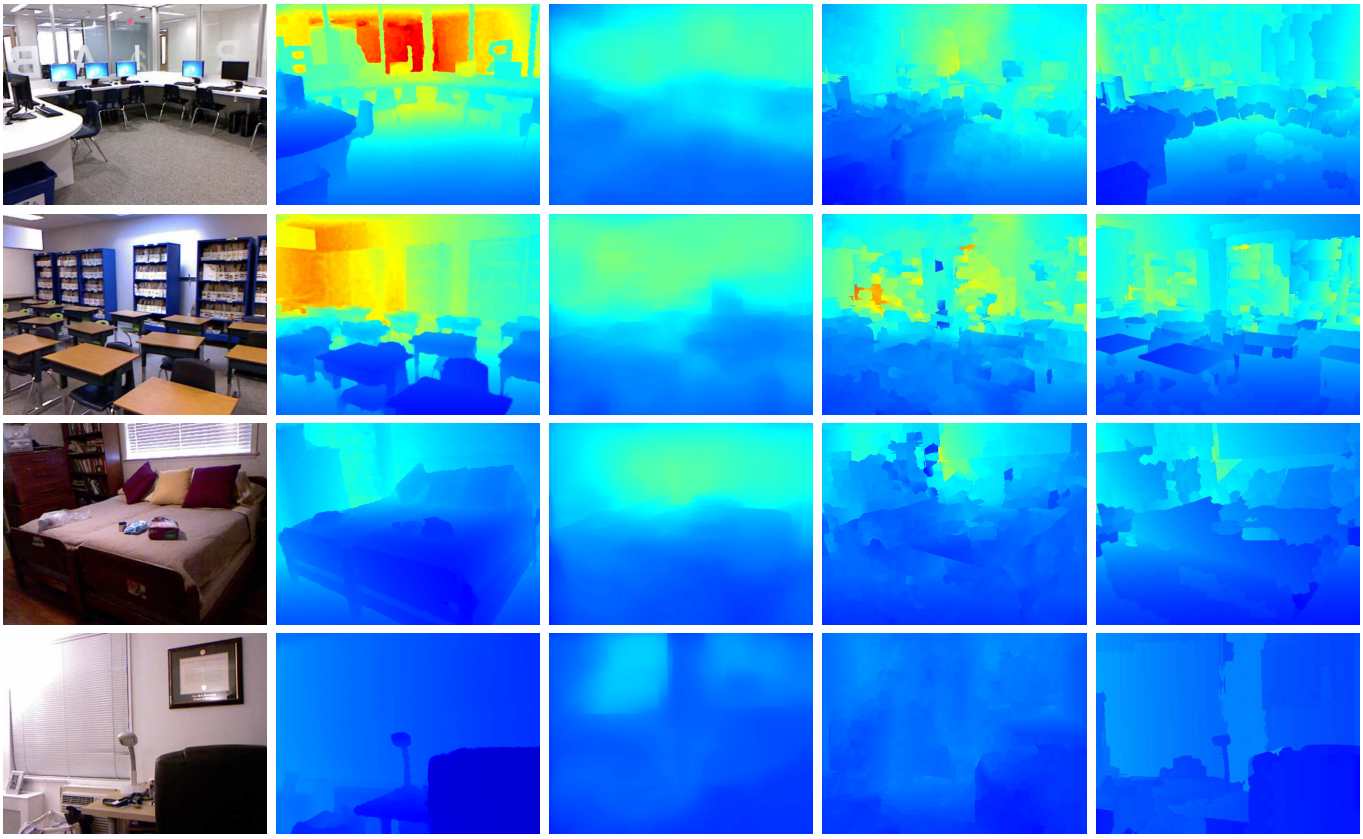

Figure 2. **NYUv2: Qualitative comparison.** Depth maps estimated by the different baselines and by our approach. Note that our approach typically avoids the oversmoothing of DepthTransfer, while better modeling the scene structure than DC-Depth.

form DepthTransfer and DC-Depth on all error metrics, and SemanticDepth on two out of three threshold values, despite the fact that it exploits the additional knowledge of pixel labels during training. The results of DeepDepth [5] on the depth metrics are as follows. rel: 0.215; rms: 0.9; $\delta < 1.25$: 61.10%; $\delta < 1.25^2$: 88.70%; $\delta < 1.25^3$: 97.10%. While they are more accurate, recall that DeepDepth relies on a much larger training set. In terms of normal accuracy, we outperform the baselines on three out of the five metrics. Fig. 2 provides a qualitative comparison of the depth maps recovered by the different approaches on several images. Altogether, these results confirm that the use of mid-level and global structure is beneficial.

In Table 2, we provide an analysis of the different parts of our model. The analysis evidences the fact that each layer contributes to improving the final accuracy. It also reveals that the mid-level structures seem to yield the major contribution. In Fig. 3, we provide a qualitative comparison of the depth maps obtained by the different components of our approach. Although not obvious at this scale, we observed that, while the mid-level structures help spatial depth coherence, they still respect the discontinuities in the image. Furthermore, the global structure yields more accurate depth ordering in the entire scene.

In addition to the depth of the superpixels, our model also predicts depth from the regions and from the global

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours-local	0.334	0.128	1.05	50.35%	82.31%	95.44%
Ours-mid	0.312	0.123	**1.03**	52.08%	**83.92%**	96.13%
Ours-global-only	0.325	0.128	1.07	50.38%	82.06%	95.35%
Ours	**0.305**	**0.122**	1.04	**52.50%**	83.77%	**96.16%**

Table 2. **NYU v2: Ablation study.** We evaluate the influence of the different components of our model. These results confirm that each parts of our model contributes to the final results, with a strong influence of the mid-level structures.

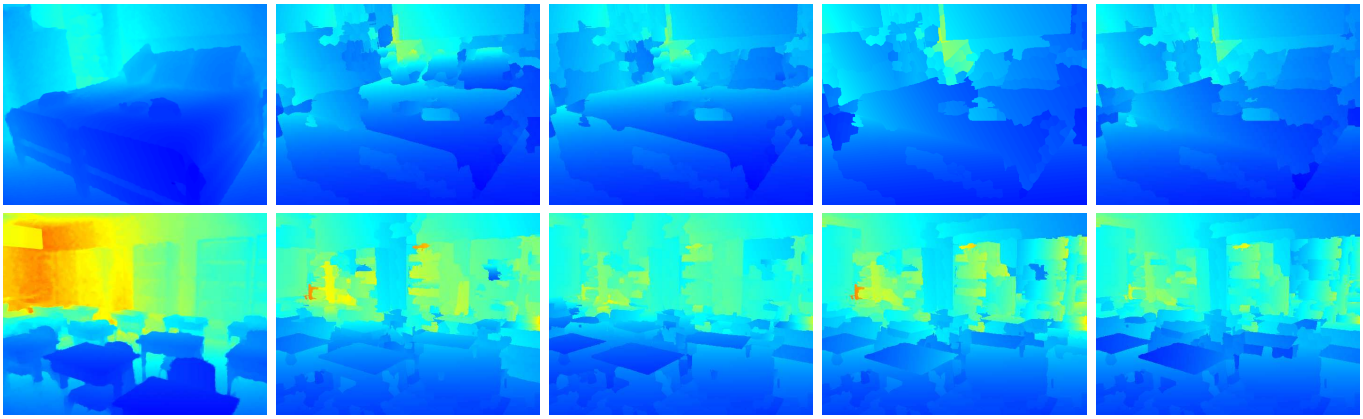

Figure 3. **NYUv2: Ablation study.** Depth maps obtained by the different components of our approach.

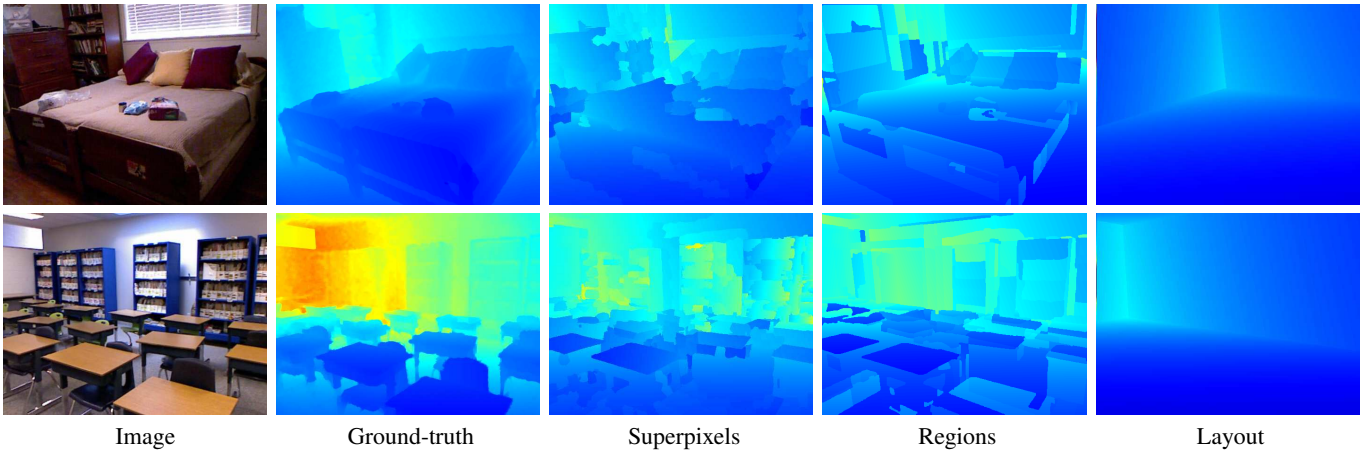

Figure 4. **NYUv2: Depth of the different layers in our model.** We show the depth maps estimated by our final model, corresponding to the variables associated with each layer in our hierarchy.

layout (although in a much coarser manner for the latter). Some resulting depth maps are depicted in Fig. 4.

RMRC Indoor:
We then evaluated our approach on the RMRC Indoor dataset [1]. Since this dataset does not provide ground-truth depth for the test images, and since our goal was to evaluate the different components of our model, we only employed the 4105 training images, from which we randomly sampled 114 images to form a test set with ground-truth. In this experiment, we used the same parameters as for NYUv2. In Table 3, we provide the various error metrics for the different parts of our model. As for NYUv2, we can see that each part contributes to the final results. However, here, the influence of the mid-level structures seems to be even larger than on NYUv2. To provide the reader with a rough idea of how our results compare to other methods, note that on the test data of the RMRC Challenge [1], the best reported relative depth errors were 0.33 for [5] and 0.39 for the second best approach by Baig and Torresani. In Figs. 5 and 6, we show the depth maps of the different components of our approach and the depth maps predicted by the variables in the different layers of our final model, respectively.

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours-local	0.440	0.167	1.24	39.38%	72.41%	89.83%
Ours-mid	0.395	**0.159**	**1.22**	**41.25%**	**74.29%**	**90.75%**
Ours-global-only	0.423	0.167	1.26	38.64%	71.09%	88.76%
Ours	**0.379**	**0.159**	**1.22**	40.67%	73.67%	90.01%

Table 3. **RMRC Indoor: Ablation study.** We compare the different components of our approach. As with NYUv2, we observe that all the parts of our model contribute the its final result, with a large contribution from the mid-level structures.

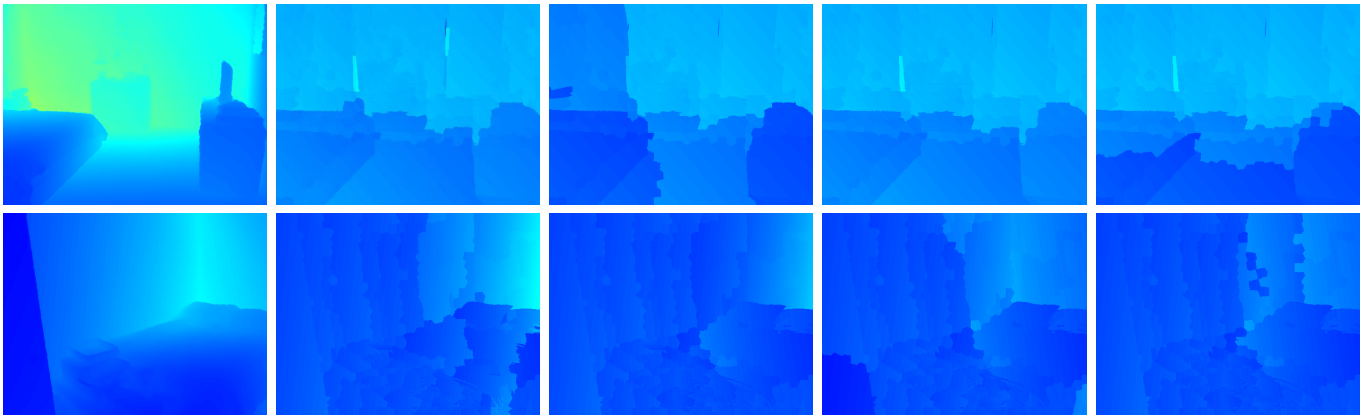

Figure 5. **RMRC Indoor: Ablation study.** Depth maps obtained by the different components of our approach.

Figure 6. **RMRC Indoor: Depth of the different layers in our model.** We show the depth maps estimated by our final model, corresponding to the variables associated with each layer in our hierarchy.

5. Conclusion

We have introduced an single image depth estimation approach that exploits the structure of the scene at different levels of details. Our experiments have demonstrated the benefits of such a structure-aware approach over local depth prediction methods. In particular, our evaluation has evidenced the fact that the mid-level structures, i.e., the regions, provided the largest contribution to the final accuracy of our model. In the future, we intend to investigate this phenomenon in more details, and study if this can be leveraged to introduce better potentials in our model. Furthermore, we plan to incorporate the use of semantic labels in our depth prediction framework.

6. Acknowledgements

The first author is supported by the China Scholarship Council. NICTA is funded by the Australian Government as represented by Department of Broadband, Communications and the Digital Economy, as well as by the ARC through the ICT Centre of Excellence program.

References

[1] RMRC challenge 2014. http://cs.nyu.edu/~silberman/rmrc2014/, June 2014. 5, 7

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Suesstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI*, 2012. 5

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011. 4

[4] M. H. Baig, V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan. Im2depth: Scalable exemplar based depth transfer. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 145–152. IEEE, 2014. 2

[5] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014. 1, 2, 5, 6, 7

[6] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3d primitives for single image understanding. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3392–3399. IEEE, 2013. 2, 5

[7] D. F. Fouhey, A. Gupta, and M. Hebert. Unfolding an indoor origami world. In *Computer Vision–ECCV 2014*, pages 687–702. Springer, 2014. 1, 2

[8] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Computer Vision–ECCV 2010*, pages 482–496. Springer, 2010. 1

[9] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *Advances in Neural Information Processing Systems*, pages 1288–1296, 2010. 1

[10] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*, 2014. 4

[11] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, pages 1849–1856. IEEE, 2009. 4

[12] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *Computer Vision–ECCV 2010*, pages 224–237. Springer, 2010. 1, 2

[13] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007. 1, 2

[14] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from a single image. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 3

[15] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *Computer Vision–ECCV 2012*, pages 775–788. Springer, 2012. 1, 2, 5

[16] J. Konrad, G.Brown, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee. Automatic 2d-to-3d image conversion using 3d examples from the internet. In *SPIE Stereoscopic Displays and Applications*, 2012. 2

[17] J. Konrad, M. Wang, and P. Ishwar. 2d-to-3d image conversion by learning depth from examples. In *3DCINE*, 2012. 2

[18] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 89–96. IEEE, 2014. 1, 2, 5

[19] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2136–2143. IEEE, 2009. 1, 2, 4

[20] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010. 1, 2

[21] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 716–723. IEEE, 2014. 1, 2, 3, 5

[22] V. Nedovic, A. W. Smeulders, A. Redert, and J.-M. Geusebroek. Stages as models of scene geometry. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1673–1687, 2010. 2

[23] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2759–2766. IEEE, 2012. 4

[24] C. Rother. A new approach to vanishing point detection in architectural environments. *Image and Vision Computing*, 20(9):647–655, 2002. 3

[25] A. Saxena, S. H. Chung, and A. Y. Ng. 3-d depth reconstruction from a single still image. *IJCV*, 2007. 1, 2

[26] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009. 1, 2

[27] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011. 4

[28] A. G. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 353–360. IEEE, 2013. 1, 2

[29] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[30] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *Computer Vision–ECCV 2010*, pages 352–365. Springer, 2010. 3

[31] A. Torralba and A. Oliva. Depth estimation from image structure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(9):1226–1238, 2002. 2