Indoor Scene Structure Analysis for Single Image Depth Estimation

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

Presenter: Shahzor Ahmad

Background / Motivation

- **Problem**: depth from monocular vision is ill-posed
- **Motivation**: Humans perceive depth from static monocular images due to prior accumulated knowledge about scene structure.
- **Goal**: learn to predict depth given a set of training RGBD images.

Contribution

- Depth is modelled at local (super-pixels), mid (regions) and global (scene box layout) levels jointly.
- Prior art on depth estimation, by contrast, reasons locally – either at the pixel (e.g., [18]) or super-pixel (e.g., [21, 15]) level.
- Prior art on scene global layout (e.g., [12]) only yield a Manhattan box layout, or sparse surface normal (e.g., [19]), but not absolute, metric depth.

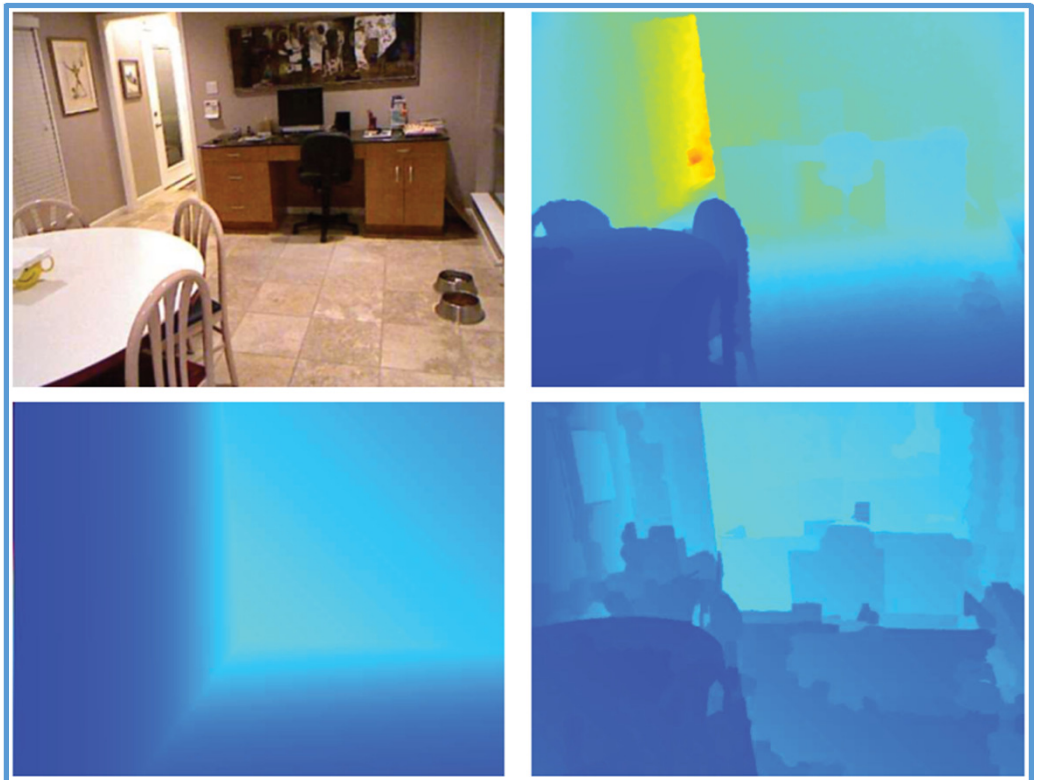

Figure 1. **Depth estimation from a single image:** (Top) Image and ground-truth depth map. (Bottom) Estimated layout and detailed depth map. Color indicates depth (red is far, blue is close).

The CRF Model (Structure-aware depth estimation)

Local super-pixels take on labels *Y* from a 4d discrete space, giving plane parameters (3 params) and centroid depth (1 param)

Mid-level region labels *R* (take on values from the same discrete space as *Y*)

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

Inference is performed using Distributed Convex Belief Propagation (DCBP) [27]

Global layout labels L (come from a space of quantized scales)

The Discrete Space of Super-pixel and Region Labels

- Discrete state space *S*, defined by quantizing the range of valid depths for super-pixel / region centroid into *V* values.
  - Range: 0.5 to 10 (meters?)
  - Steps = 0.5 => *V* = 20
- Super-pixels (resp., regions) are assumed to be planar (resp. close to planar).
- Manhattan world assumption is invoked, restricting plane normals to 3 possible dominant directions, computed from vanishing points using formulas from Lee et. al. CVPR'09 [19].

The Discrete Space of Super-pixel and Region Labels (Contd.)

- Lee et. al. CVPR'09 [19].
  - Ray

$$P = \lambda K^{-1} p, \qquad \lambda > 0$$

  - Normal direction of the three major axes given coordinates of three vanishing points (x_k, y_k) in image.

$$v_k = (x_k, y_k, 1)^T \quad \Leftrightarrow \quad V_k = \frac{K^{-1} v_k}{\| K^{-1} v_k \|_2}$$

- For any given super-pixel label y_p, or region label r_γ depth *d* and normal ***n*** at any pixel in the super-pixel (or region) may be computed.

Local Depth Estimation

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

$$Y = \{y_1, y_2, \ldots, y_{N_s}\}$$

$$\phi_p(y_p) = \frac{1}{N_p} \sum_{i=1}^{N_p} \left(d_p^i(y_p) - d_{r,p}^i\right)^2$$

Unary potential penalizes, for all super-pixels, per-pixel deviation of depth due to assigned labelling from predicted depth regressed on candidate depths (i.e., of corresponding super-pixels in nearest neighbour training images)

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

Pairwise term penalizes differences in per-pixel normal and depths in neighbouring super-pixels if they don't occluded each other

Exploiting Mid-Level Structures

$$E_m(Y,R) = \sum_{\gamma} \phi_\gamma(r_\gamma) + \sum_{\gamma,p} \phi_{\gamma,p}(r_\gamma, y_p), \qquad R = \{r_i, r_2, ..., r_{N_r}\}$$

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

3*V* dimensional histogram of depth-normal labels, computed through voting by super-pixels in nearest neighbour regions in training images to test region

Depth and normal computed under given labeling

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

Inter-layer (local and mid-level) coherence imposed to penalize differences between per-pixel depth predicted by region and super-pixels it contains

Incorporating Global Structures
(Spatial Box Layout, Hedau et. al. CVPR'09 [11])

Incorporating Global Structures (Contd.)

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

Porobability that pixel *i* belong to clutter; ensures super-pixel depths are not over-smoothened

***L* encodes scale of predicted box layout, takes values form a space of quantized scales**

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

Inter-layer (local and global) coherence imposed to penalize differences between per-pixel depth predicted by box layout faces and super-pixels it contains

Baselines

- **Depth Transfer (Karsch et. al., ECCV'12 [15]):** Inspired by the non-parametric SIFT flow / label transfer method of Liu et. al. '09, '11, for object recognition and scene parsing, and applied to *depth* transfer.
- **Semantic Depth (Ladicky et. al., CVPR'14 [18]):** Instead of learning per-pixel depth classifier, it learns classifiers to predict likelihood for a semantic class at an arbitrarily fixed canonical depth; jointly models semantic segmentation and depth estimation.
- **DC-Depth (Liu et. al., CVPR'14 [21]):** A discrete-continuous CRF models relationships between superpixels, where depth is represented by continuous variables, and relationships by discrete ones.
- **Deep Depth (Eigen et. al. arXiv'14 [5]):** Deep-learning based method utilizing 140K RGBD training images.

Datasets

- **NYU v2 (Silberman et. al. ECCV'12 [29]):**
  - 1449 RGB and corresponding in-painted depth images (795 train, 654 test)
  - 464 different indoor scenes across 26 scene classes, gathered from commercial and residential buildings in three US cities
  - Kinect camera intrinsics are given
  - Depth specified in meters
- **RMRC 2014 Challenge [1]:**
  - 4105 RGB-D training images
  - Since GT depth for test images is not provided, 114 train images were randomly obtained to form a test set

Evaluation Metrics – Depth [21, 18]

- average relative error (**rel**): $\frac{1}{N}\sum_{\mathbf{u}}\frac{|g_{\mathbf{u}}-d_{\mathbf{u}}|}{g_{\mathbf{u}}}$,
- average $\mathbf{log}_{10}$ error: $\frac{1}{N}\sum_{\mathbf{u}}|log_{10}g_{\mathbf{u}}-log_{10}d_{\mathbf{u}}|$,
- root mean squared error (**rms**): $\sqrt{\frac{1}{N}\sum_{\mathbf{u}}(g_{\mathbf{u}}-d_{\mathbf{u}})^2}$,

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

$$t = 1.25, 1.25^2, 1.25^3$$

Evaluation Metrics – Normals [6]

- Dense per-pixel scene normals were predicted using method of [6], and evaluated using the following metrics:
  - **Mean** angle difference between estimated & GT normals
  - **Medain** angle difference between estimated & GT normals
  - **% of pixels** with angular difference < threshold (11.25 ^{o}, 22.5 ^{o}, 30^{o})

Results – 1/6
(NYU v2 Quant.)

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
DepthTransfer	0.374	0.134	1.12	49.81%	79.46%	93.75%
DC-Depth	0.335	0.127	1.06	51.55%	82.32%	95.00%
SemanticDepth	-	-	-	**54.22%**	82.90%	94.09%
Ours	**0.305**	**0.122**	**1.04**	52.50%	**83.77%**	**96.16%**

Method	mean	median	$\theta < 11.25$	$\theta < 22.5$	$\theta < 30$
DepthTransfer	**43.0**	**40.5**	6.9%	23.2%	34.9%
DC-Depth	45.7	42.2	19.7%	25.7%	35.4%
SemanticDepth	-	-	-	-	-
Ours	46.7	41.9	**21.1%**	**35.2%**	**41.7%**

Table 1. **NYUv2: Comparison of our approach with the baselines.**

Results – 2/6
(NYU v2 Qual.)

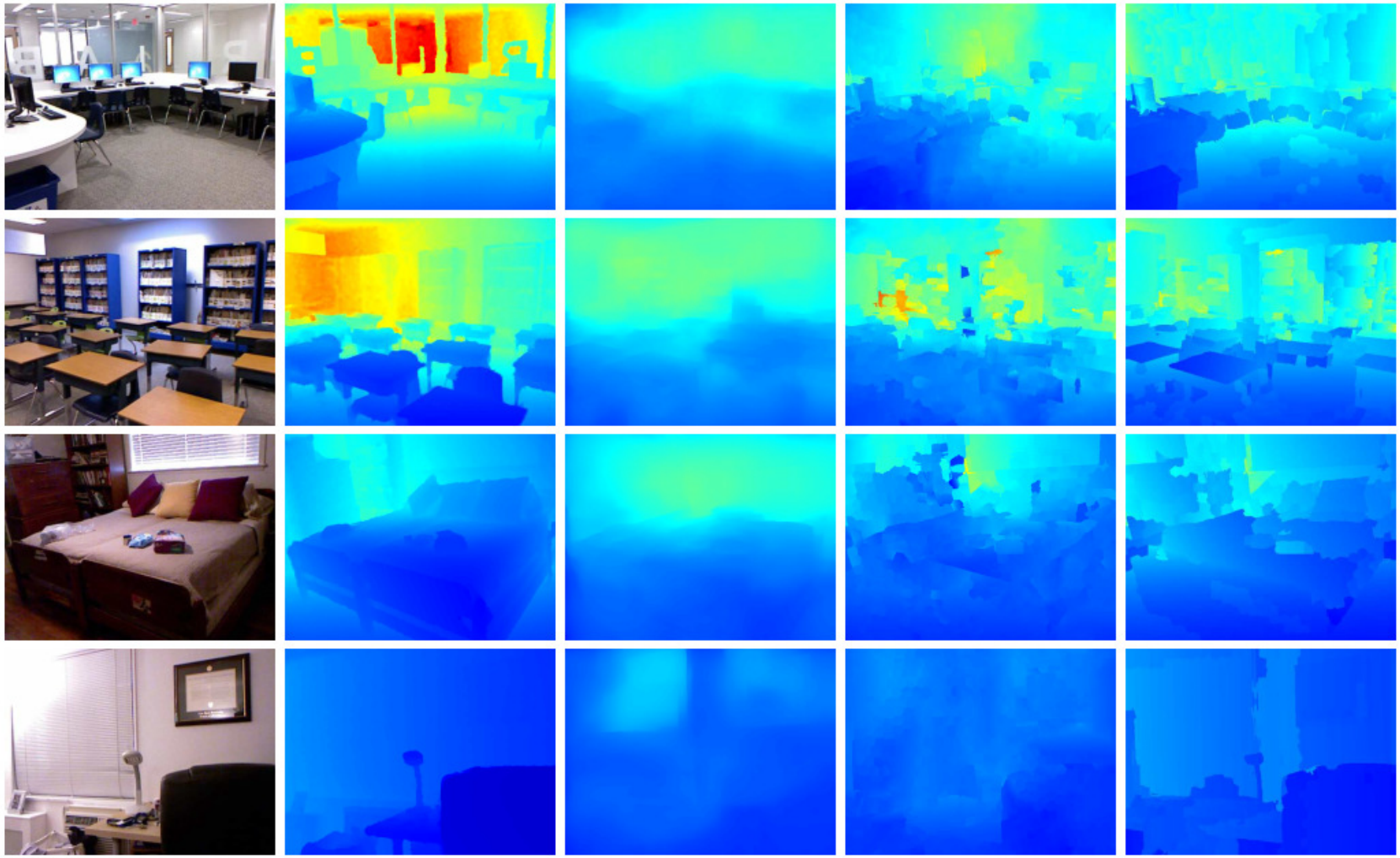

Figure 2. **NYUv2: Qualitative comparison.** Depth maps estimated by the different baselines and by our approach. Note that our approach typically avoids the oversmoothing of DepthTransfer, while better modeling the scene structure than DC-Depth.

Results – 3/6
(NYU v2 Ablation Study Quant.)

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.

Results – 4/6
(NYU v2 Ablation Studv Qual.)

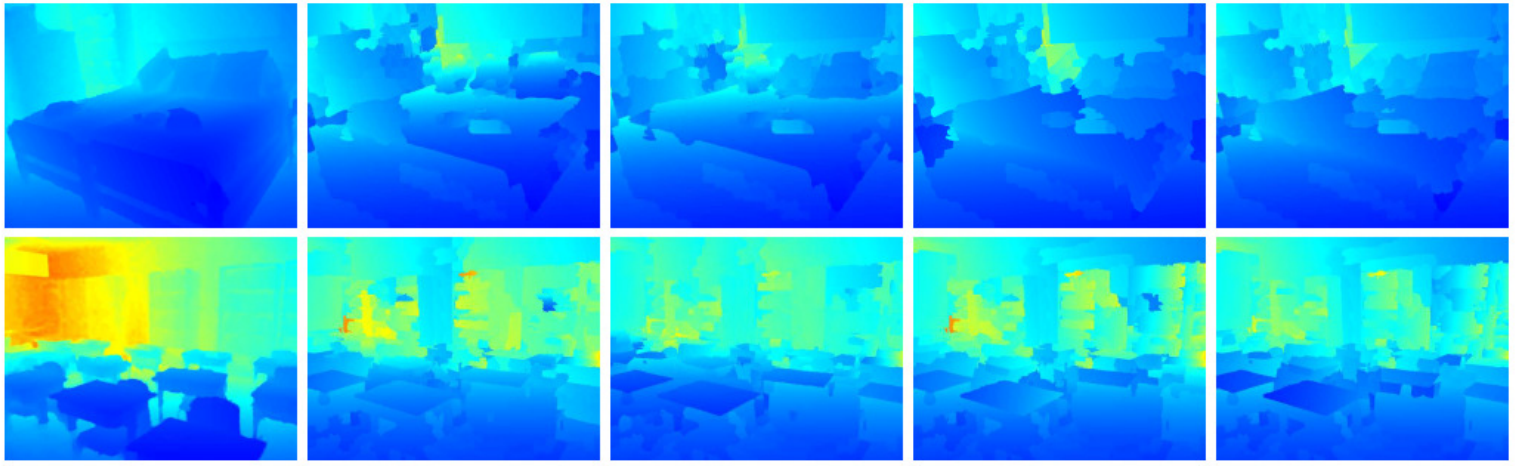

Figure 3. **NYUv2: Ablation study.** Depth maps obtained by the different components of our approach.

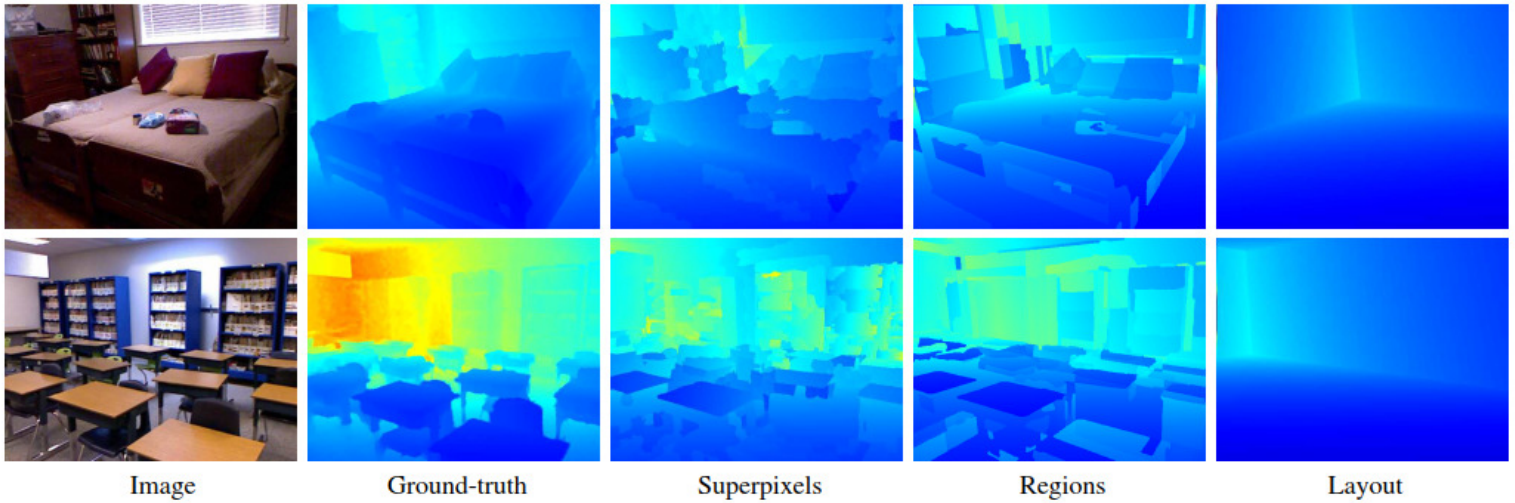

Figure 4. **NYUv2: Depth of the different layers in our model.** We show the depth maps estimated by our final model, corresponding to the variables associated with each layer in our hierarchy.

Results – 5/6
(RMRC Ablation Study Quant.)

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.**

Results – 6/6
(RMRC Ablation Study Qual.)

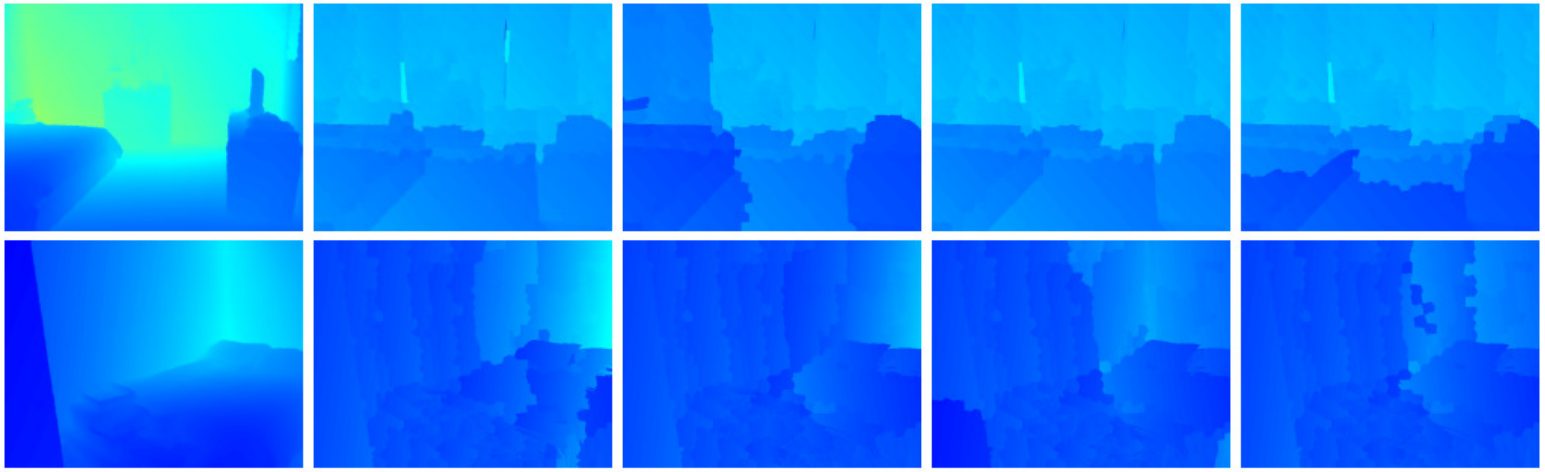

Figure 5. **RMRC Indoor: Ablation study.** Depth maps obtained by the different components of our approach.

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.

Thank You!