This CVPR2015 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Towards Unified Depth and Semantic Prediction from a Single Image

Peng Wang[1] Xiaohui Shen[2] Zhe Lin[2] Scott Cohen[2] Brian Price[2] Alan Yuille[1]
[1]University of California, Los Angeles [2]Adobe Research

Abstract

Depth estimation and semantic segmentation are two fundamental problems in image understanding. While the two tasks are strongly correlated and mutually beneficial, they are usually solved separately or sequentially. Motivated by the complementary properties of the two tasks, we propose a unified framework for joint depth and semantic prediction. Given an image, we first use a trained Convolutional Neural Network (CNN) to jointly predict a global layout composed of pixel-wise depth values and semantic labels. By allowing for interactions between the depth and semantic information, the joint network provides more accurate depth prediction than a state-of-the-art CNN trained solely for depth prediction [6]. To further obtain fine-level details, the image is decomposed into local segments for region-level depth and semantic prediction under the guidance of global layout. Utilizing the pixel-wise global prediction and region-wise local prediction, we formulate the inference problem in a two-layer Hierarchical Conditional Random Field (HCRF) to produce the final depth and semantic map. As demonstrated in the experiments, our approach effectively leverages the advantages of both tasks and provides the state-of-the-art results.

1. Introduction

Depth estimation and semantic segmentation from a single image are two fundamental yet challenging tasks in computer vision. While they address different aspects in scene understanding, there exist strong consistencies among the semantic and geometric properties of image regions. When the information from one task is available, it would provide valuable prior knowledge to guide the other one.

In the depth estimation literature, semantic information has long been used as a high-level guidance [14, 15, 23, 11, 29]. Certain semantic classes have strong geometric implications. For example, the ground is usually a horizontal plane in a canonical view, while the building facades are mostly vertical surfaces [14]. However, these approaches either assume the semantic labels are known [29], or perform semantic segmentation to generate the semantic labels [23]. Since the two tasks are performed sequentially, the errors in the predicted semantic labels are inevitably propagated to the depth results. On the other hand, in semantic segmentation, with the increasing availability of RGBD data from additional depth sensors, many methods use depth as another channel to regularize the segmentation [28, 31, 12] and have achieved much better performance than using RGB images alone.

Since the two tasks are mutually beneficial, extensive investigations have been done towards jointly solving them in videos [2, 8, 19, 34], in which 3D information can be easily obtained through structure from motion. However, the efforts in jointly tackling the two problems from a single image are preliminary [21], mostly because the inference of both tasks are more ill-posed in a single image. It is not trivial to formulate the joint inference problem, in which the two tasks could benefit each other. This paper is another step towards this direction. Unlike previous approaches [21], in which the consistency between the semantic and geometric property is limited to local segments or objects, we propose a unified framework to incorporate both global context from the whole image and local prediction from regions, through which the consistency between depth and semantic information is automatically learned through joint training.

Fig. 1 illustrates the framework of our approach. We formulate the joint inference problem in a two-layer Hierarchical Conditional Random Field (HCRF). The unary potentials in the bottom layer are pixel-wise depth values and semantic labels, which are predicted by a Convolutional Neural Network (CNN) trained globally from the whole image, while the unary potentials in the upper layer are region-wise depth and semantic maps, which come from another CNN-based regressor trained on local regions. The output of the global CNN, though coarse, provides very accurate global scale and semantic guidance, while the local regressor gives more details in depth and semantic boundaries. The mutual interactions between depth and semantic information are captured through the joint training of the CNNs, and are further enforced in the joint inference of HCRF.

We evaluated our method on the NYU v2 dataset [31] on both depth estimation and semantic segmentation. By inference using our joint global CNN, the depth prediction improves over the depth only CNN by an average 8% relative

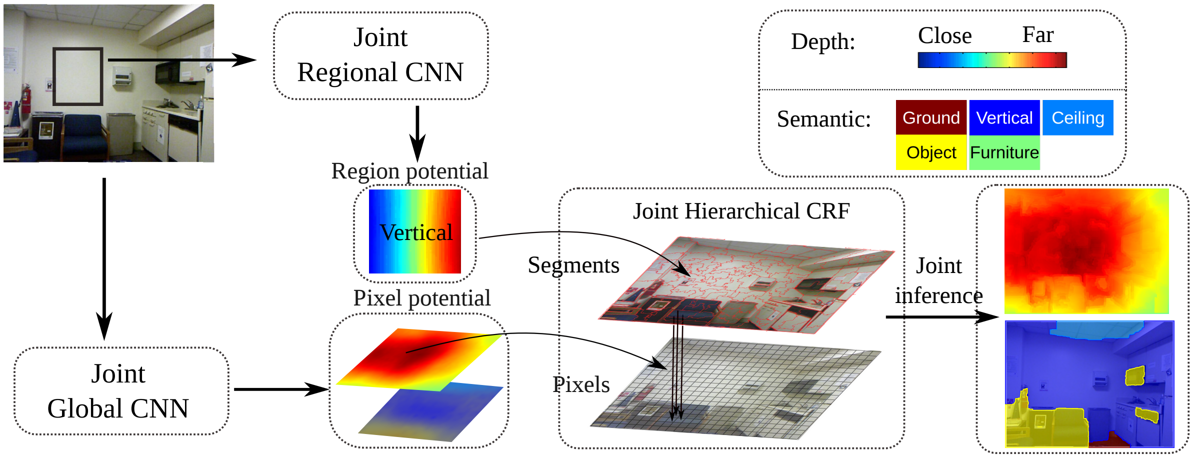

Figure 1. Framework of our approach for joint depth and semantic prediction. As described in Sec. 1, given an image, we obtain region-wise and pixel-wise potential from a regional and a global CNN respectively. The final results are jointly inferred through the Hierarchical CRF. We keep the color legend consistent in the paper.

gain, and also outperforms the state-of-the-art. After incorporating local predictions, the final depth maps produced by the HCRF are significantly improved in terms of visual quality, with much clearer structures and boundaries. Meanwhile in semantic segmentation, we further show that our joint approach outperforms R-CNN [10] that is currently known to be the most effective method for semantic segmentation, by 10% relatively in average IOU.

To sum up, the contribution of this paper is three-fold:

1. We propose a unified framework for joint depth and semantic prediction from a single image. The consistency of the two tasks is learned through joint training, and enforced in different stages throughout the framework to boost the performance of both tasks.
2. We formulate the problem in a two-layer HCRF to enforce synergy between global and local predictions, where the global layouts are used to guide the local predictions and reduce local ambiguities, while the local results provide detailed region structures and boundaries.
3. Through extensive evaluation, we demonstrate that jointly addressing the two problems in our framework benefits both tasks, and achieves the state-of-the-art.

1.1. Related work

The literature of depth estimation and semantic segmentation is very rich when considering them as two independent tasks. Interestingly, though developed separately, the techniques used to solve the two tasks are quite similar. MRF-based approaches are common choices in semantic segmentation [1, 36], while they have also been explored in depth prediction [30, 14]. Data-driven approaches based on non-parametric transfer are another popular trend in both scene parsing [5, 32, 33, 35] and depth estimation [17, 24]. Recently, CNN have shown its effectiveness in both tasks. In [6], a two-level CNN is learned to directly predict the depth maps, which significantly outperforms the previous state-of-the-arts. Similar progress has also been achieved in semantic segmentation [3, 7, 10, 4]. Inspired by these work, we also use CNN to train our model for joint global and local prediction.

Noticing the correlations between the two problems, some methods try to use the information from one task to regularize the other. Nevertheless, the interaction between the depth and semantic information is mostly a one-way channel in previous work. Several methods try to getter better semantic segmentation results given RGB-D data [28, 31, 12, 13], while others take the predicted semantic labels to estimate depth [23, 15]. However, in order to solve one problem, these methods rely on either the ground-truth data, or an independent solution to the other problem. Their results therefore are heavily limited by the availability of the ground-truth data or the quality of the previous step.

While promising, the joint inference of these two tasks to to enforce consistency between them is an under-explored direction in the literature. In [11], the consistency between the geometric and semantic properties of segments are built, in which each semantic segment is also predicted to be one of the three geometric classes: horizontal, vertical, and sky. However, such a geometric classification is still too coarse to produce an accurate depth map, and too loose to constrain the semantic prediction. Moreover, the consistency between the two components is limited to local regions. Ladicky et.al [21] jointly train a canonical classifier considering both the loss from semantic and depth labels of the objects. However, they use local regions with hand-crafted features for prediction, which is only able to generate very coarse depth and semantic maps, with many local prediction distortions over large backgrounds. Unlike these methods, we capture the mutual information through joint training in a unified framework, which captures more synergy between semantic and depth prediction. In addition, from a global to local strategy, we achieve long range context to generate global reasonable results while maintaining segments boundary information. Finally, our trained CNNs provide robust estimation under the large appearance variation of images and segments. As a result, our model achieves better results both quantitatively and qualitatively.

2. Formulation

As shown in previous image segmentation work [1, 20], semantic inference should consider both short-range pixel-wise interactions and high-order context. Similarly, the consistency in depth and semantic prediction should also be enforced both globally and locally. To this end, instead of a standard pixel-wise Conditional Random Field, we propose a two-layer Hierarchical Conditional Random Field (HCRF) [1, 20] to formulate the joint depth and semantic prediction problem.

As shown in Fig.1, our HCRF is composed of two layers of nodes and edges. In the bottom layer, the nodes are the pixels in the image $\mathcal{I}$. For each pixel $i \in \mathcal{I}$, we would like to predict its depth value d_i and semantic label l_i. We use $\mathbf{x}_i = \{d_i, l_i\}$ to denote the inference output at pixel i. Meanwhile in the upper layer, we decompose image $\mathcal{I}$ to local segments, and use the segments to represent the nodes. Similarly, we would like to infer the depth and semantic labels $\mathbf{y}_s = \{d_s, l_s\}$ for each segment $s \in \mathcal{S}$, where $\mathcal{S}$ denotes the set of segments after decomposition. We use $\mathcal{R}_s$ to denote all the pixels inside segment s, and use $\mathcal{X}_s$ to denote the predicted labels of $\mathcal{R}_s$. Apparently there are three kinds of edges in the HCRF, the pair-wise edges between neighboring pixels, the edges between neighboring segments, and the edges connecting $\mathcal{R}_s$ and s. Given such a model, the energy for minimization is formulated as:

$$\min_{\mathcal{X}} \sum_{i\in\mathcal{I}} \psi_i(\mathbf{x}_i) + \lambda_{ie} \sum_{i,j\in\mathcal{I}} \psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j), \\ + \lambda_y \min_{\mathcal{Y}} \left(\sum_{s\in\mathcal{S}} \psi_s(\mathcal{X}_s, \mathbf{y}_s) + \lambda_{ce} \sum_{s,t\in\mathcal{S}} \psi_{s,t}(\mathbf{y}_s, \mathbf{y}_t) \right), \quad (1)$$

where $\psi_i(\mathbf{x}_i)$ is the pixel-level unary potential in the bottom layer, $\psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j)$ is the pair-wise edge potential between pixels, and $\psi_{s,t}(\mathbf{y}_s, \mathbf{y}_t)$ is the edge potential between segments in the upper layer. λ_y is a balancing parameter. In addition, the cross-layer potential term $\psi_s(\mathcal{X}_s, \mathbf{y}_s)$ usually could be further decomposed as:

$$\psi_s(\mathcal{X}_s, \mathbf{y}_s) = \phi_s(\mathbf{y}_s) + \sum_{i\in\mathcal{R}_s} \phi_s(\mathbf{y}_s, \mathbf{x}_i), \quad (2)$$

where $\phi_s(\mathbf{y}_s)$ is the unary potential of segments in the upper layer, and $\phi_s(\mathbf{y}_s, \mathbf{x}_i)$ is the edge potential between segment s and the pixel i inside segment s.

In our model, the potential terms introduced in Eqn.(1) and Eqn.(2) are defined as follows:

Unary potentials. As illustrated in Fig.1, the pixel-level potential $\psi_i(\mathbf{x}_i)$ is provided by a CNN trained globally on the whole image, which jointly predicts pixel-wise depth values and probabilities of semantic labels. The details of the global CNN training and prediction will be introduced in Section 3. Similarly, the segment-level potential $\phi_s(\mathbf{y}_s)$ in Eqn.(2) is generated by a CNN-based regressor trained on local regions, with details described in Section 4.

Edge potentials. For pixel-wise edge potentials, we only consider neighboring pixels, and define

$$\psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{1}\{l_i \neq l_j\}(\exp(-\mathrm{edge}(i,j)) + e_d(d_i, d_j)), \quad (3)$$

where $\mathbb{1}\{l_i \neq l_j\}$ is a switching function which enables penalizing when the semantic labels of i and j are different. $\mathrm{edge}(\cdot)$ is the output from a semantic edge detection method [22], and $e_d(d_i, d_j) = \exp(-[\|d_i - d_j\|_1 - t_d]_+)$, where $[x]_+ = \max\{x, 0\}$ represents the hinge loss. This term generally enforces pairwise smoothness, except when there is a strong semantic edge or possible depth discontinuity between i and j. The definition of $e_d(d_i, d_j)$ gives credit for assigning different labels when the depth difference is greater than a threshold t_d.

For the segment-wise edge potentials, we only consider neighboring segments as well. For each segment s, we calculate the mean and variance of the pixel RGB values inside the segment to get its local appearance feature $\mathbf{f}_s$. Meanwhile, for each pair of neighboring segments s and t, we calculate the geodesic distance between them $\mathrm{dist}_g(s,t)$ based on the semantic edge map produced by [22]. We then get the appearance-based distance between two segments:

$$\mathrm{dist}_a(s,t) = \mathrm{dist}_g(s,t) + \lambda_a \|\mathbf{f}_s - \mathbf{f}_t\|, \quad (4)$$

where λ_a is a balancing weight. We also define the depth-based distance between the two segments $\mathrm{dist}_d(s,t)$ to be the average of pixel-wise depth difference within the overlapping boundary areas of the two segments. Then the edge potential between two segments $\psi_{s,t}(\mathbf{y}_s, \mathbf{y}_t)$ is defined as:

$$\begin{aligned}\psi_{s,t}(\mathbf{y}_s, \mathbf{y}_t) \quad &= \mathbb{1}\{l_s \neq l_t\}(\exp(-\mathrm{dist}_a(s,t)) + e_d(s,t)), \\ &+ w(l_s, l_t)\mathrm{dist}_d(s,t), \quad (5)\end{aligned}$$

where $e_d(s,t) = \exp(-[\mathrm{dist}_d(s,t) - t_d]_+)$ has the similar functionality as in Eqn.(3) that allows different semantic labels if the depth change between the two segments is large. $w(l_s, l_t)$ is a smoothness weight matrix which is learned from the data, in which higher value of $w(l_s, l_t)$ requires a higher depth smoothness between segments s, t when their semantic labels l_s, l_t are consistent, and vice versa.

For the cross-layer edge potentials $\phi_s(\mathbf{y}_s, \mathbf{x}_i)$ between the segments and the pixels, we simply enforce consistency when the pixels are inside the segment, and have no constraints if the pixels do not belong to the segment.

Given the above definition, we see that the pixel-level unary potentials encode coarse global layout, while the

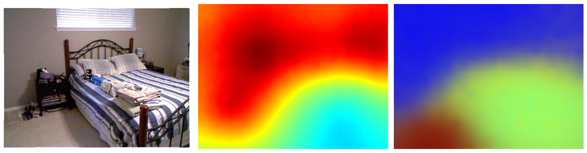

Figure 2. An example of the global network output. Middle: Depth map. Right: Semantic probability map.

segment-level unary potentials focus on local region details. The edge potentials incorporate the consistency between the depth and semantic labels. Therefore through joint inference, our model is able to better exploit the interactions between global and local predictions, as well as between depth and semantic information. We will describe the inference procedure in details in Section 5.

3. Joint Global Depth and Semantic Prediction

In this section, we describe how we train a CNN with the whole image as input to predict pixel-wise depth and semantic maps, which are used as the pixel-level unary potentials in our HCRF model.

CNN has shown its effectiveness in predicting not only discrete class labels [18] but also structured continuous maps. In [6], with the use of ground-truth depth data, a CNN is trained to directly predict a depth map using the whole image as input, which achieves global context. Inspired by this work, we extend it to a CNN that directly predicts pixel-wise depth values jointly with semantic labels from the whole image.

We follow the CNN structure in [6] in the earlier layers. However, in addition to the depth nodes in the final layer, we further introduce semantic nodes to predict the semantic labels. Formally, our loss function during the network training is composed of two parts:

$$\text{loss}(\mathcal{X}, \mathcal{X}^*) = \frac{1}{n}\sum_{i=1}^{n}(\log d_i - \log d_i^*)^2 + \lambda_l \frac{-1}{n}\sum_{i=1}^{n}\log(P(l_i^*)),$$

$$\text{and } \; P(l_i^*) = \exp(z_{i,l_i^*}) / \sum_{l_i} \exp(z_{i,l_i}), \tag{6}$$

where d_i and l_i are the predicted depth values and semantic labels, while d_i^* and l_i^* are the ground truth. z_{i,l_i} is the output of the semantic node corresponding to pixel i.

Since the training data with ground truth semantic labels are very limited compared with raw RGB-D data, we first train the network to only predict depth values using RGB-D training data (i.e., drop the semantic nodes in the final layer), and then fine-tune the network with added semantic nodes using the RGB-D data with available semantic labels.

Once trained, given an input image, the network will predict a depth map and a probability map of each pixel belonging to a semantic label. Since it is trained globally, the predicted maps are quite coarse (Fig.2). Nevertheless, they provide very accurate global scale and semantic layout, which helps avoid prediction errors caused by local appearance ambiguities. Moreover, as will be shown in the experiments, the joint-prediction network after fine-tuning provides more accurate depth maps than the network trained to predict depth alone, which demonstrates that semantic information can regularize the CNN that benefit depth prediction.

We use d'_i to denote the depth value at pixel i predicted by the global CNN, and use $P(l_i)$ to denote the predicted probabilities of semantic labels at pixel i. The pixel-wise unary term in Eqn.(1) can be written as:

$$\psi_i(\mathbf{x}_i) = -log(P(l_i)) + \lambda_i \|d_i - d'_i\|_1. \tag{7}$$

4. Joint Local Depth and Semantic Prediction

While the depth and semantic maps predicted by the global CNN accurately capture the scene layout, they still lack details in local regions. Therefore in order to recover scene structures and object boundaries, we decompose the image into segments by over-segmentation [25], and predict the semantic label and depth map for each segment. The predicted results are then used as the segment-level unary potentials in our HCRF to complement the global results.

The training and prediction of depth and semantic labels in local segments are not as straightforward as in the global inference. First we need to find a proper way to represent the depth and semantic labels inside the segment, i.e., $\mathbf{y}_s = \{d_s, l_s\}$ in Sec. 2. For semantic labels, we use the majority of the pixel-wise semantic labels to represent the segment label l_s, which is a generally valid assumption. However for depth, it is too coarse to use a single depth value to represent d_s. Meanwhile, when cropping out the local segment from the image, the global scale information is lost, and it is difficult to tell its absolute depth values by looking at the segment alone. A more feasible task would be to predict a relative depth trend inside the segment. Therefore we transform the absolute depth map of the segment to a normalized relative depth map by subtracting the absolute depth value at the segment center d_c and re-scaling it to have range [0,1]. Given the normalized depth map, the depth value at center d_c and the scale change sc, we can exactly recover the absolute depth values of each pixel in the segment $d_i = d_n * sc + d_c$, where d_n are the relative depth values in the normalized depth map. Therefore, in the local prediction stage, we would like to estimate the normalized depth map of the segment, while $[d_c, sc]$ are two unknown variables that we would infer in the HCRF.

4.1. Normalized Joint Templates

Even if we normalize the depth map of the segment, it is still difficult to train a regressor from the image to the map. This is because the depth of local segments is highly ambiguous when solely judging from its local appearance.

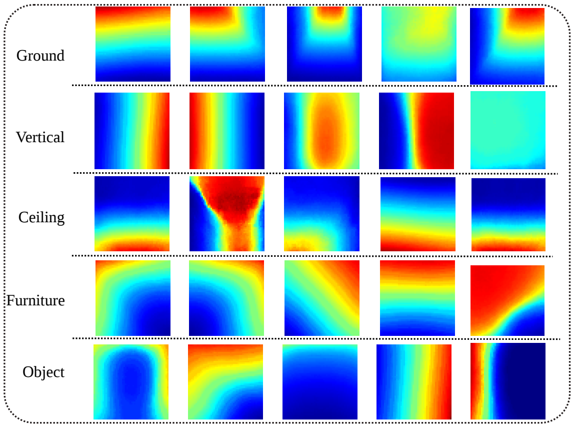

Figure 3. Examples of joint semantic and depth templates for local segments. The normalized depth maps in each row are associated with their corresponding semantic labels.

Nevertheless, the patterns in the depth maps of local segments are less diverse, e.g. often like a plane or a corner, and therefore could be captured by a limited number of templates. Thus, we formulate the local depth estimation problem as a prediction of the composition from a set of normalized templates, which largely constrains the learning space.

To generate the templates, we use both the semantic and depth ground truth to ensure consistency. To avoid the dominance of segments from a large semantic class, we first cluster the segments according to their semantic labels. For the segments with the same semantic label, we cluster their normalized depth maps using L_1 distance metric to generate a set of templates. Fig. 3 illustrates a subset of our joint templates, which provides meaningful patterns like a plane, a corner or a curved surface.

4.2. Joint Template Regression

Given a segment s and a set of templates $\mathbf{T}_j$, we would like to learn the affinities of this segment to the templates. The affinity during training is defined as:

$$\begin{aligned} a(s, \mathbf{T}_j) &= \mathbb{1}\{l_s = l_{\mathbf{T}_j}\}\mathrm{S}_d(s, \mathbf{T}_j)/\max_k \mathrm{S}_d(s, \mathbf{T}_k) \\ \mathrm{S}_d(s, \mathbf{T}_j) &= \exp(-\|\mathbf{d}_s - \mathbf{d}_{\mathbf{T}_j}\|_1). \end{aligned} \tag{8}$$

where $\mathbf{d}$ denotes the values in the normalized depth maps. Intuitively, when the semantic labels are different, the affinity of the segment to the template is zero, otherwise it is determined by the similarity of their normalized depth maps. We use CNN as the local training model as well, which takes the warped bounding box of the segments as input, with loss function defined as the sum of sigmoid cross entropy loss over the affinities, i.e.:

$$l(\mathbf{a}_s, \mathbf{a}_s^*) = \frac{-1}{N_t}\sum_{i=1}^{N_t}(a_i \log a_i^* + (1 - a_i)\log(1 - a_i^*)),$$

where $\mathbf{a}_s = [a_1, \cdots, a_{N_t}]$ are the affinities of segment s to all the templates. Based on the loss defined, our local CNN is learned through fine-tuning the global CNN in Section 3. After regression, we choose top N (N=2 in experiments) templates with the highest affinities and aggregate their normalized depth values as well as the semantic labels to the segment with their affinities as weights. The averaged results are the prediction of the normalized depth map and the probability of semantic labels of that segment.

The depth and semantic ambiguity caused by local segment appearance is still a problem in template regression. Therefore we use three techniques to further reduce ambiguity. First, the output of the global CNN in Sec. 3 gives us a very good global layout to regularize the local prediction. Second, masking out the background outside a segment as in R-CNN [10],could reduce confusions when two segments share a same bounding box. Therefore, for a segment, we take the fc_6 layer output of the local CNN both from its bounding box and masked region, and concatenate it with the global prediction within the corresponding bounding box to form our feature vector. We train a Support Vector Regressor upon that feature to predict a segment's affinities to the templates. Third, the ambiguity of prediction will decrease when the segments are larger. Therefore instead of performing the regression on small segments produced by over-segmentation, we cluster them to generate multi-scale large segments (30, 50, 100 segments in three scales respectively). Consider that a small segment s is covered by a larger segment s_L, we can map the depth and semantic predictions of s_L back to segment s. The final depth and semantic prediction of s is a weighted averaging of the results from multiple s_L covering s. The details of getting larger segments is in our supplementary material.

Fig. 4 gives two predicted examples from the learned model. We can see our depth prediction is robust to image variations and does not depend on particular structures, and has the potential to overcome the difficulties met in traditional line and vanishing point detection methods [30].

Given the normalized depth map, as mentioned earlier, we can represent the depth values in the segment using two parameters: center depth d_c and scale factor sc. We hereby define the segment-level unary potential $\phi_s(\mathbf{y}_s)$ as:

$$\phi_s(\mathbf{y}_s) = -\log(P(l_s)) + \lambda_d(\|d_c - d_{gc}\|_1 + \|sc - sc_g\|_1) \tag{9}$$

where $P(l_s)$ is the predicted probability of semantic labels on segment s. d_{gc} is the absolute depth from the global depth prediction at the segment center, and sc_g is the depth scale from the global prediction within the segments bounding box. Intuitively, we want d_c and sc to be close to the one predicted by global CNN, which can also be regarded as the message passed from the pixel-level potential. Once d_c and sc are inferred, we can combine them with the normalized depth map to get the absolute depth for each pixel in this segment, which can be used to calculate the edge potentials between the segments in Eqn.(5), as well as to enforce the

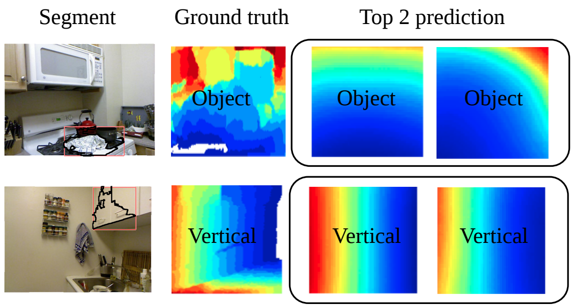

Figure 4. Illustration of local prediction results from two difficult segments (located in the red box). Our prediction is robust to the complex scenario or even blurred cases.

consistency between the global pixel-wise prediction and local segment-level prediction.

5. Joint HCRF Inference

To do inference over the Joint HCRF, direct inference over the joint space of semantic label and depth through loopy belief propagation (LBP) [26] costs a long time for convergence. We consider a more efficient alternating optimization strategy by minimizing one when fixing the other.

Semantic inference given the depth. Given the estimated depth, we first perform LBP to infer the semantic labels in the segment level, and then pass the predictions of local segments to their covering pixels. We then infer the labels in the pixel level, which can be solved through MAP.

Depth inference given the semantic label. Similarly, we first infer the depth variables in the segment-level, namely, the center depth d_c and the scale factor sc. The inference of continuous depth variables are impractical for LBP. Thus, we quantize the center depth d_c of a segment to be a set of discrete offsets (in our experiment, we set it to be 20 uniformly distributed values within range $[-r_d, r_d]$) from the respective value predicted in the global model, and the scale sc to be a shift of respective global scale (10 intervals within $[-r_s, r_s]$ and then truncate the values within the range $[min_{sc}, max_{sc}]$). Theoretically, our quantization follows the same spirit of particle belief propagation [27].

In our experiments, our global predictions are already very good. Therefore we use the global prediction as our initialization, and perform 1 iteration by first estimating semantic labels and then predicting depth. It already produces the state-of-the-art results, and more iterations brings very little improvement in our experiments. To further accelerate the algorithm, we use graph cut to efficiently solve pixel-wise semantic labeling. In pixel-level depth inference, we find the smoothness term makes little difference in the final solution. Thus, the depth inference is reduced to a linear combine of global prediction and local prediction considering the weight λ_y in Eqn.(1) which is very easy to learn through maximum likelihood using the ground truth depth.

6. Experiments

Data. We evaluate our method on the NYU v2 dataset [31] which contains images taken by Kinect camera in 464 indoor scenes. We use the official train/test split, using 249 scenes for training the global depth prediction. After evening the distribution (1200 images per scene), the total number of depth images are 200K. The joint depth and semantic label set contains 1449 images, and it is partitioned into 795 training images and 654 testing images. Due to the limited number of data, we also use images from the NYU v1 dataset that are not overlapped with the 654 testing images for training. There are 894 annotated semantic labels in the dataset. In order to better ensuring consistency between depth maps and semantic labels with limited data, we mapped the semantic labels into 5 categories conveying strong geometric properties, i.e. $\{Ground, Vertical, Ceiling, Furnitures, Objects\}$. Fig. 5 illustrates our mapped labels. When train the global CNN, we do the data augmentation similar to the method in [6], which gives us 2 million depth images for training.

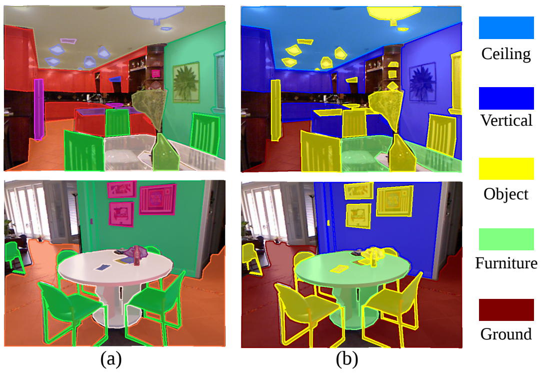

Figure 5. We map the detailed semantic classes in (a) to five main semantic classes in (b).

Implementation details. The structure of the global CNN is the same as the one in [6], and the resolution for semantic output is 20×26, yielding 3120 additional output nodes. We use caffe [16] for our network implementation. For inference over our graphical model, we use the LBP tool provided by Meltzer[1].

For the parameters balancing unary and edge potentials, in Eqn.(1), $\lambda_y = 4$, which is learned through ML as stated in Sec. 5. $\lambda_{ie} = 3, \lambda_{ce} = 2$, which are learned through cross-validation. For the parameters balancing the semantic and depth, we adjust them to make their numerical ranges comparable. Specifically, $\lambda_l = 0.05$ in Eqn.(6), $\lambda_i = \lambda_d = 10$ in Eqn.(7) and Eqn. (9), $\lambda_a = 0.1$ in Eqn.(4). For the threshold t_d, we set it to be $0.2m$.

In Sec. 4.1, when clustering the templates, the numbers in five semantic class are $[40, 40, 40, 60, 60]$ respectively. We keep $C = 0.3$ when learning the SVR. To balance different features in SVR, we normalize each feature with L_2 norm, and concatenate all the features and weight each type of feature based on its relative feature length, i.e. $w_i = \sum_j L_j / L_i$ where L_i is the length of feature type i.

[1] http://www.cs.huji.ac.il/ talyam/inference.html

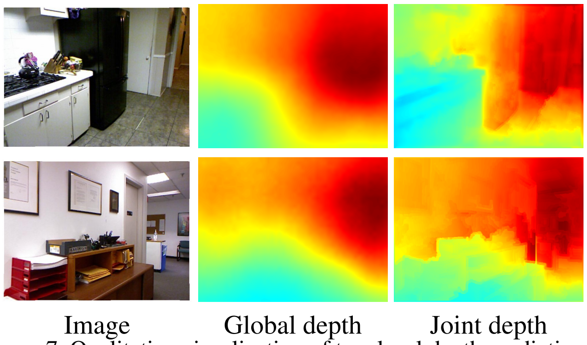

Figure 7. Qualitative visualization of two level depth prediction.

To infer the depth in Sec. 5, we set $r_d = 0.5m$, $r_c = 0.25m$ and $[min_{sc}, max_{sc}] = [0.05m, 0.5m]$. In addition, the learned weight matrix of $w(s_c, s_k)$ in Eqn.(5) is attached in the supplementary material. By our matlab implementation, it takes about 4 days to learn our models, and the testing time for our algorithm is around $40s$ for a 480×640 under a desktop with $3.4GHz$ processor and a K-40 GPU.

6.1. Quantitative results

Depth estimation To evaluate the depth prediction, we take various available metrics from the previous work [24, 6] to measure different aspects of the depth results. Formally, given the predicted absolute depth of a pixel $d_{\mathbf{x}}$ and the ground truth $d^*_{\mathbf{x}}$, the evaluation metrics are: (1) Abs relative difference(Rel): $\frac{1}{N}\sum_{\mathbf{x}} \frac{|d_{\mathbf{x}} - d^*_{\mathbf{x}}|}{d^*_{\mathbf{x}}}$; (2) Square relative difference(Rel(sqr)): $\frac{1}{N}\sum_{\mathbf{x}} \frac{|d_{\mathbf{x}} - d^*_{\mathbf{x}}|^2}{d^*_{\mathbf{x}}}$; (3) Average log_{10} error: $\frac{1}{N}\sum_{\mathbf{x}} |\log_{10}(d_{\mathbf{x}}) - \log_{10}(d^*_{\mathbf{x}})|$; (4) RMSE (linear): $\sqrt{\frac{1}{N}\sum_{\mathbf{x}} |d_{\mathbf{x}} - d^*_{\mathbf{x}}|^2}$; (5) RMSE (log): $\sqrt{\frac{1}{N}\sum_{\mathbf{x}} |log(d_{\mathbf{x}}) - log(d^*_{\mathbf{x}})|^2}$; (6) Threshold: $\%$ of $d_{\mathbf{x}}$ s.t. $\max(\frac{d_{\mathbf{x}}}{d^*_{\mathbf{x}}}) < thr$, where $thr \in \{1.25, 1.25^2, 1.25^3\}$.

We compare our results with five most recent methods, i.e. Make3D [30], Depth Transfer [17], DC Depth [24], Canonical Depth [21] and Depth CNN [6]. We follow the test setting exactly as that in Depth CNN[2] [6].

Tab. 1 shows the quantitative results from all the algorithms. Our final algorithm, i.e. Joint HCRF, outperforms the state-of-the art Depth CNN [6] with a noticeable margin. The results of our Global Depth CNN are comparable to the one produced by [6]. We think the difference is mostly because we use a geometric preserving cropping for data augmentation (described in our supplementary material), yielding improvements on the metrics of Rel and RMSE. However, we did not use the scale invariance loss and do pre-training on imagenet as [6], which might lead to dropping of the δ metric. By fine-tuning the network to jointly predict depth and semantic labels, the joint global CNN is better than the depth-only CNN in 7 out of 8 metrics. It shows that the semantic labels regularized the depth prediction through the CNN training, which benefits the depth estimation. By enforcing the global and local consistency in our joint HCRF, although the quantitative results are slightly better than the global joint CNN, in Fig. 7 and Fig. 3 in our supplementary material, we show that it provides a significant improvement in visual quality both in semantic segmentation and depth estimation. The results from HCRF have sharper transitions at the surface boundaries and align to local details. The same phenomenon is also mentioned in [6]. Thus a better metric to measure the visual quality is worth investigating in the future work.

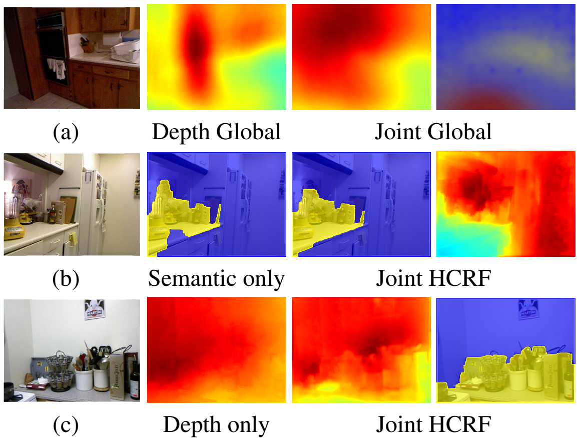

Figure 8. Examples showing the intuitions behind joint prediction.

Semantic prediction. To evaluate the semantic segmentation, we take the both the popularly used Intersection Over Union (IOU) and pixel accuracy percentage as evaluation metrics. We take the state-of-the-art segmentation method R-CNN [10] for comparison to show the effectiveness of our joint prediction. To obtain R-CNN results, we use the author's code[3], and follow the exactly same training strategy for the segmentation stated in their paper. For a fair comparison, we apply our trained model for region-wise features, and apply the same CRF as we did for local superpixels without considering the depth information.

Tab. 2 shows the compared results, and our joint estimation provides the best performance. As shown in the second row, adding only the semantic guidance from global CNN improves the performance about 2.5%, which shows the benefits of the interaction between global guidance and local prediction. By adding depth information into the framework, the accuracy is further improved, which proves the complementary of the depth and semantic information. We also tried to use a global jointly trained CNN to directly predict the semantic labels. However, such a global prediction

[2] For the results of Make3D, Canonical Depth and Depth CNN, we copy the results that reported in [6]. However, we find the setting of DC Depth is different in terms of evaluated image size. Thus, we asked the author for their results for a fair comparison. For Depth Transfer, we downloaded their code <http://kevinkarsch.com/?p=323>, and retrained the model to generate all the results.

[3] https://github.com/rbgirshick/rcnn

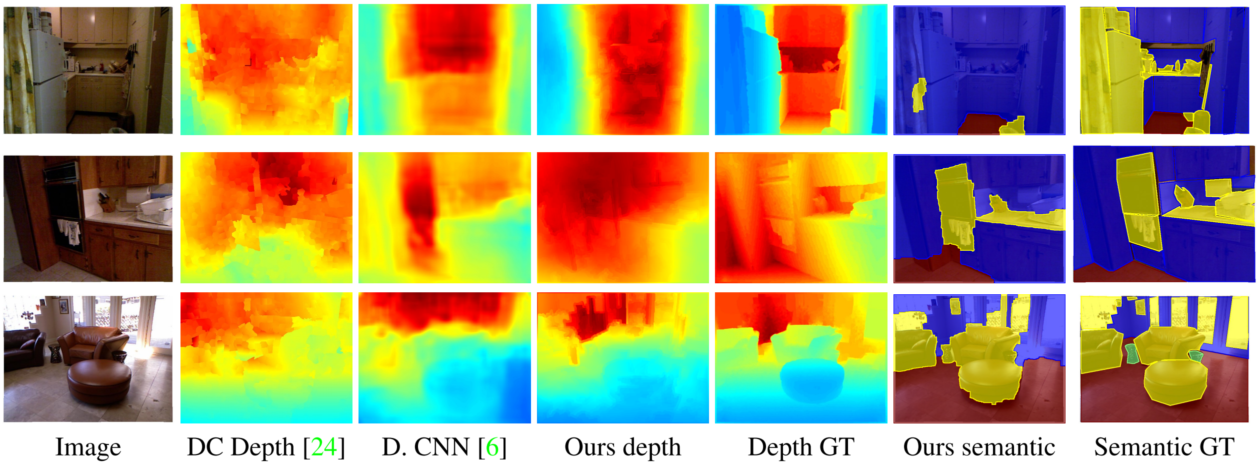

Figure 6. Qualitative comparison with other approaches. Depth maps are normalized by their respective max depth (Best viewed in color).

	Lower is better					Higher is better		
Criteria	Rel	Rel(sqr)	Log_{10}	RMSE(linear)	RMSE(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Make 3D [30]	0.349	0.492	-	1.214	0.409	0.447	0.745	0.897
Depth Transfer [17]	0.350	0.539	0.134	1.1	0.378	0.460	0.742	0.893
DC Depth [24]	0.335	0.442	0.127	1.06	0.362	0.475	0.770	0.911
Canonical Depth [21]	-	-	-	-	-	0.542	0.829	0.940
Depth CNN Coarse [6]	0.228	0.223	-	0.871	0.283	**0.618**	**0.891**	0.969
Depth CNN Fine [6]	0.215	0.212	-	0.907	0.285	0.611	0.887	0.971
Global CNN - Depth only	**0.207**	0.216	0.104	0.823	0.284	0.550	0.861	0.969
Global CNN - Joint	0.226	**0.208**	0.095	0.750	0.266	0.593	0.889	**0.976**
Joint HCRF	0.220	0.210	**0.094**	**0.745**	**0.262**	0.605	0.890	0.970

Table 1. Quantitative comparison between our method and other state-of-the-art baseline on the NYU v2 dataset.

Method	Ground	Vertical	Ceiling	Furniture	Object	Mean IOU	Pix acc.
R-CNN [10] CRF	57.837	64.062	16.513	17.8	45.536	40.349	68.312
Semantic HCRF	61.840	**66.344**	15.977	**26.291**	43.121	42.715	69.351
Joint HCRF	**63.791**	66.154	**20.033**	25.399	**45.624**	**44.200**	**70.287**

Table 2. Quantitative comparison between our method and R-CNN [10] on image segmentation task of NYU v2 dataset.

only achieves 30.5% in mean IOU, which is considerably lower than the results of our HCRF. The segmentation from the joint global CNN is very blurry, while HCRF provides much clearer boundaries.

6.2. Qualitative results

In Fig. 6, we further visually show the depth comparison results between our method, DC Depth [24] and DCNN [6], and the segmentation comparing with the ground truth. In Fig. 6, we can see that DC Depth uses small local segments which suffers from local distortions due to lack of global cues. DCNN does not have the constraint from semantic, thus the prediction may be negatively influenced by appearance variation, e.g. the refrigerator in the second image, and the reflection on the ground at right-bottom of the third image. In our case, our approach jointly considers both the global prediction and local details, and leverages the benefit from depth and semantic prediction, and therefore achieves more consistent depth changes with the ground truth.

In Fig. 7, we show that comparing with global depth output, the joint output provides more detailed structures in the scene, yielding visually more satisfied results. In addition, in Fig. 8, we illustrate the intuition behind the joint information of depth and semantic labels by doing experiments of removing one from the model and test the other. In Fig. 8(a), for global prediction, by adding the semantic constraint, the distortion of depth CNN prediction is fixed because of the smoothness constraint enforced by the "vertical" label. In Fig. 8(b), by considering local depth change and depth discontinuity, the model is able to handle the appearance confusion in semantic segmentation. In Fig. 8(c), for fine-level depth estimation, by adding semantic segments, the depth map are better aligned with object boundaries.

7. Conclusion

We propose a unified approach to jointly estimate depth and semantic labels from a single image. We formulate the problem in a hierarchical CRF which embeds the potential from a global CNN and a local regional CNN. Through joint inference, our algorithm achieves promising results in both depth and semantic estimation. In future work, we will extend to outdoor scenarios such as the KITTI dataset [9].

Acknowledgements. This work is supported by NSF CCF-1317376 and ARO 62250-CS, and partially done when the first author was an intern at Adobe

References

[1] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. Gonzlez. Harmony potentials - fusing global and local scale for semantic image segmentation. In *IJCV*, pages 83–102, 2012. 2, 3

[2] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pages 44–57, 2008. 1

[3] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV (7)*, pages 430–443, 2012. 2

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 2

[5] D. Eigen and R. Fergus. Nonparametric image parsing using adaptive neighbor sets. In *CVPR*, pages 2799–2806, 2012. 2

[6] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*. 2014. 1, 2, 4, 6, 7, 8

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. In *TPAMI*, pages 1915–1929, 2013. 2

[8] A. Flint, D. W. Murray, and I. Reid. Manhattan scene understanding using monocular, stereo, and 3d features. In *ICCV*, pages 2228–2235, 2011. 1

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 9

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 5, 7, 8

[11] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, pages 1–8, 2009. 1, 2

[12] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*. 2014. 1, 2

[13] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *CVPR*, pages 97–104, 2013. 2

[14] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. In *IJCV*, pages 151–172, 2007. 1, 2

[15] D. Hoiem, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from an image. In *IJCV*, pages 328–346, 2011. 1, 2

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[17] K. Karsch, C. Liu, and S. B. Kang. Depthtransfer: Depth extraction from video using non-parametric sampling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014. 2, 7, 8

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012. 4

[19] A. Kundu, Y. Li, F. Daellert, F. Li, and J. M. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, 2014. 1

[20] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, pages 739–746, 2009. 3

[21] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. June 2014. 1, 2, 7, 8

[22] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Efficient closed-form solution to generalized boundary detection. In *ECCV (4)*, pages 516–529, 2012. 3

[23] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, pages 1253–1260, 2010. 1, 2

[24] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *CVPR*, June 2014. 2, 7, 8

[25] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *CVPR*, pages 2097–2104, 2011. 4

[26] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *CoRR*, 2013. 6

[27] J. Peng, T. Hazan, D. Mcallester, and R. Urtasun. Convex max-product algorithms for continuous mrfs with applications to protein folding. In *ICML*, 2011. 6

[28] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, pages 2759–2766, 2012. 1, 2

[29] B. C. Russell and A. Torralba. Building a database of 3d scenes from user annotations. In *CVPR*, 2009. 1

[30] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. In *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 824–840, 2009. 2, 5, 7, 8

[31] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV (5)*, pages 746–760, 2012. 1, 2, 6

[32] G. Singh and J. Kosecka. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *CVPR*, pages 3151–3157, 2013. 2

[33] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. In *IJCV*, pages 329–349, 2013. 2

[34] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, 2014. 1

[35] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014. 2

[36] S. Zheng, M.-M. Cheng, J. Warrell, P. Sturgess, V. Vineet, C. Rother, and P. H. S. Torr. Dense semantic image segmentation with objects and attributes. In *CVPR*, 2014. 2