Look Deeper into Depth: Monocular Depth Estimation with Semantic Booster and Attention-Driven Loss

Jianbo Jiao[1,2(✉)], Ying Cao[1], Yibing Song[3], and Rynson Lau[1]

[1] City University of Hong Kong, Kowloon, Hong Kong SAR
jianbjiao2-c@my.cityu.edu.hk, caoying59@gmail.com, dynamicstevenson@gmail.com
[2] University of Illinois at Urbana-Champaign, Urbana, USA
[3] Tencent AI Lab, Shenzhen, China
rynson.lau@cityu.edu.hk

Abstract. Monocular depth estimation benefits greatly from learning based techniques. By studying the training data, we observe that the per-pixel depth values in existing datasets typically exhibit a long-tailed distribution. However, most previous approaches treat all the regions in the training data equally regardless of the imbalanced depth distribution, which restricts the model performance particularly on distant depth regions. In this paper, we investigate the long tail property and delve deeper into the distant depth regions (*i.e.* the tail part) to propose an attention-driven loss for the network supervision. In addition, to better leverage the semantic information for monocular depth estimation, we propose a synergy network to automatically learn the information sharing strategies between the two tasks. With the proposed attention-driven loss and synergy network, the depth estimation and semantic labeling tasks can be mutually improved. Experiments on the challenging indoor dataset show that the proposed approach achieves state-of-the-art performance on both monocular depth estimation and semantic labeling tasks.

Keywords: Monocular depth · Semantic labeling · Attention loss

1 Introduction

Depth acquisition has been actively studied over the past decades with widespread applications in 3D modeling, scene understanding, depth-aware image synthesis, *etc.* However, traditional hardware or software based approaches are restricted by either environment or multi-view observations assumption. To overcome these limitations, there is a growing interest in predicting depth from a single image.

Monocular depth prediction is an ill-posed problem and inherently ambiguous. However, humans can well perceive depth from a single image, given that sufficient samples (*e.g.* the appearances of nearby/distant objects) have

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11219, pp. 55–71, 2018.
https://doi.org/10.1007/978-3-030-01267-0_4

been learned over lifetimes. With the success of deep learning techniques and available training data, the performance of monocular depth estimation has been greatly improved [5,53]. While existing methods measure depth estimation accuracy by vanilla loss functions (*e.g.* ℓ_1 or ℓ_2), they assume that all regions in the scene contribute equally without considering the depth data statistics. We have empirically found that the depth values in the indoor/outdoor scenes vary greatly across different regions and exhibit a long tail distribution (see Fig. 1). This is an inherent property of the nature that mainly caused by the perspective-effect during the depth acquisition process. Given such imbalanced data, loss functions that treat all regions equally will be dominated by the samples with small depth, leading the models to be "short-sighted" and not effective to predict the depth of distant regions.

Moreover, complement to the learned prior knowledge like perspective, semantic understanding of the scene (*e.g.* sky is faraway, wall is vertical) essentially benefits depth estimation. For example, knowing a cylinder-like object to be a pencil or a pole can help estimate its depth. Furthermore, depth information is also helpful to differentiate semantic labels, especially for different objects with similar appearances [4,11,41]. Estimating depth and semantics can thus be mutually beneficial. Unfortunately, there is a lack of strategy to efficiently propagate and share information across the two tasks.

In this work, we propose to address the above two challenges by presenting a deep network to predict depth as well as semantic labels from a single still image. A novel attention-driven loss with depth-aware objective is proposed to supervise the network training, which alleviates the data bias issue and guides the model to *look deeper* into the scene. In addition, in our synergy network architecture, we propose an information propagation strategy that performs in a dynamic routing fashion to better incorporate semantics into depth estimation. The strategy is achieved by a lateral sharing unit and a semi-dense skip-up connection, which allow information to propagate through internal representations across and within both tasks. Experimental results on the challenging indoor dataset show that, with the proposed loss and knowledge sharing strategy, the performance of monocular depth estimation is significantly improved and reaching state-of-the-art. Our contributions are summarized as follows:

- We propose a novel attention-driven loss to better supervise the network training on existing datasets with long tail distributions. It helps improve depth prediction performance especially for distant regions.
- We present a synergy network architecture that better propagates semantic information to depth prediction, via a proposed information propagation strategy for both inter- and intra-task knowledge sharing.
- Extensive experiments demonstrate the effectiveness of our method with state-of-the-art performance on both depth and semantics prediction tasks.

2 Related Work

Depth from Single Image. Early works on monocular depth estimation mainly leverage hand-crafted features. Saxena *et al.* [44] predict the monocu-

lar depth by a linear model on an over-segmented input image. Hoiem *et al.* [17] further group the superpixels into geometric meaningful labels and construct a 3D model accordingly. Later on, with large-scale RGB-D data available, data-driven approaches [21,22,27,28,30,35,43] become feasible. Eigen *et al.* [4,5] construct a multi-scale deep convolutional neural network (CNN) to produce dense depth maps. Some methods [24,29,34,51–53,56] try to increase the accuracy by including Conditional Random Fields (CRFs). Despite notable improvements, the model complexity increases as well. Other works [1,57] predict depth by exploring ordinal relationships. Data imbalance is reported in [28,43] while not explicitly addressed. Some other works [6,9,26,55] propose to supervise the network by a reconstruction loss from the other stereo or temporal view. While requiring no depth supervision, rectification and alignment are usually necessary, and they rely on multi-view images during training. Although remarkable performance has been achieved, the long tail property of depth data distribution has not yet been well-explored.

Depth with Semantics. As depth and semantic labels share context information, some methods [3,4,11,42,46] take depth map as a guidance to improve the semantic segmentation performance. In [46], Silberman *et al.* propose the NYU RGBD dataset and use the combination of RGB and depth to improve the segmentation. Based on this dataset, some methods [3,11] take RGBD as input to perform semantic segmentation. Eigen and Fergus [4] design a deep CNN that takes RGB, depth, surface normal as input to predict the semantic labels. Owing to the power of CNN models, other methods [41,49,50] are proposed to better leverage depth for semantic labeling recently. While great performance has been demonstrated, the ground truth depth is indispensable for the labeling task. On the other hand, prior information encoded in the semantic labels can be leveraged to assist depth prediction. Instead of directly mapping from color image to depth, Liu *et al.* [33] first perform a semantic segmentation on the scene and then use the labels to guide depth prediction, in a sequential manner.

Joint Representation Sharing. Some recent works attempt to investigate the representation sharing between different tasks [16,19,20,27,38,39,51]. Ladicky *et al.* [27] propose a semantic depth classifier and analyze perspective geometry for image manipulation, whereas they rely on hand-crafted features locally. In [12], a traditional framework is presented for joint segmentation and 3D reconstruction. Wang *et al.* [51] use a CNN following by a hierarchical CRF to jointly predict semantic labels and depth. However, they only modify the last layer for prediction and rely on superpixels and CRF. A concurrent work [23] proposes a weighting strategy for multi-task losses. Misra *et al.* [38] propose a cross-stitch (CS) network for multi-task learning. While performs better than baselines, it may suffer from propagation interruption if the combination weights degenerates into 0. The two-parallel-CNN design also increases the number of parameters and learning complexity. Another sharing approach [18] applying dense connections between each layer in a CNN is proposed for recognition tasks. The fully-dense connections share all the information but increase memory consumption as well.

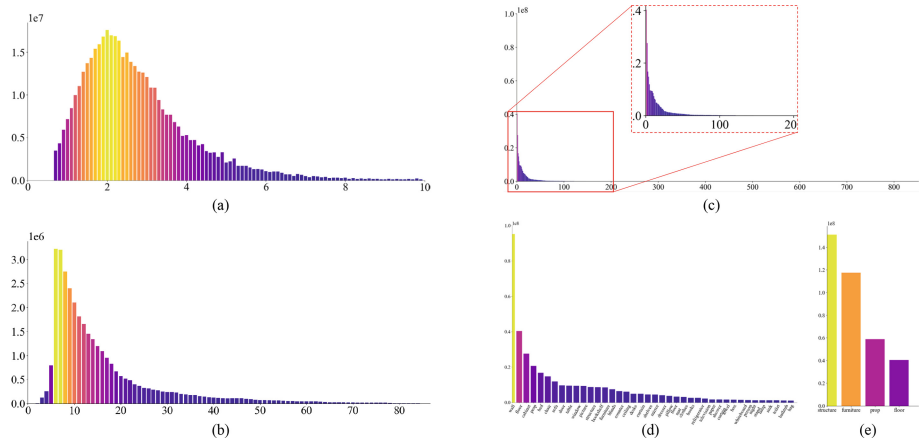

Fig. 1. Long tail distributed datasets on depth and semantic labels. Vertical axes indicate the number of pixels. (a) shows the depth value (horizontal axis, in meter) distribution of the NYUD v2 dataset [46], and (b) shows the distribution of the KITTI dataset [7]. (c) gives the semantic label distribution (label index as horizontal axis) of the NYUD v2, while (d, e) are the distributions of the mapped 40 [10] and 4 [46] categories from the 800+ categories in (c). Imbalanced long-tailed distribution can be observed in these datasets, even for semantic labels mapped to only four categories.

In our work, we jointly train semantic labeling and depth estimation in an end-to-end fashion, without complicated pre- or post-processing. We also propose to capture better synergy representations between the two tasks. Furthermore, we investigate the long-tail data distribution in existing datasets and propose an attention-driven loss to better supervise the network training.

3 Depth-Aware Synergy Network

3.1 Depth-Aware Objective

Most state-of-the-art monocular depth estimation methods make use of CNNs to enable accurate depth prediction. In these frameworks, the depth prediction is formulated as a regression problem, where ℓ_1 or ℓ_2 loss is usually used to minimize the pixel-wise distance between the predicted and ground truth depth maps based on the training data. When estimating monocular depth, we observe that a long tail distribution resides in both indoor (NYUD v2 [46]) and outdoor (KITTI [7]) depth datasets. As shown in Fig. 1(a), (b), the number of samples/pixels per depth value falls dramatically after a particular depth, with only a small depth range dominating a large number of pixels. This data imbalance problem shares similarity with that in object detection [32,45] but differs in nature. It is because the inherent natural property of perspective effect from the imaging process leads to the uneven distribution of depth pixels, which can not be eliminated by simply increasing training data. As a result, training deep

models on such datasets using the loss functions that treat all pixels equally as in previous works can be problematic. The easy samples with small depth pixel values can easily overwhelm the training while hard samples with large depth pixel values have very limited contribution, leading the models tend to predict smaller depth values.

Based on the above observations, we propose to guide the network to pay more attentions to the distant depth regions during training and adaptively adjust the backpropagation flow accordingly. The proposed depth-aware objective is formulated as:

$$L_{DA} = \frac{1}{N}\sum_{i=1}^{N}(\alpha_D + \lambda_D)\cdot \ell(d_i, d_i^{GT}), \tag{1}$$

where i is the pixel index, N is the number of pixels in the depth map. d_i and d_i^{GT} are the predicted depth value and ground truth value respectively. $\ell(\cdot)$ is a distance metric can be ℓ_1, ℓ_2, *etc.* α_D is a depth-aware attention term that guides the network to focus more on distant hard depth regions to reduce the data distribution bias. Therefore, the gradients during backpropagation weight more on minority distant regions with respect to vast nearby regions. In this way, α_D should be positively correlated to the depth and can be defined as a linear function with respect to the ground truth depth.

To avoid gradient vanishing at the beginning of training and avoid cutting off of learning for nearby regions, a regularization term λ_D is introduced along with the attention term as:

$$\lambda_D = 1 - \frac{\min(\log(d_i), \log(d_i^{GT}))}{\max(\log(d_i), \log(d_i^{GT}))}, \tag{2}$$

which describes the learning state during training. If the network at current state predicts pixel i close to the ground truth, the regularization term λ_D approaches 0. When the network does not accurately predicts the value, λ_D approaches 1. As a result, even for very near ($\alpha_D \to 0$) regions that are not accurately predicted, the gradients can still be backpropagated, which approaches the original ℓ loss function. In this way, Eq. 2 ensures the stableness during training. Our depth-aware objective guides the network to adaptively focus on different regions and automatically adjusts the strength/attention for each training sample, thus ensures the optimization direction of the model to be comparatively balanced. In sum, while L_{DA} preserves the focus on nearby pixel samples, it enables the network to put more attentions on the distant ones during training.

3.2 Network Architecture

The proposed synergy network is a multi-task deep CNN that mainly consists of four parts: the depth prediction sub-network, semantic labeling sub-network, knowledge sharing unit/connection, and the attention-driven loss. An overview architecture is shown in Fig. 2. The input RGB image is passed through a backbone encoder (*e.g.* VGG [47], ResNet [14]) to convert the color space into a high-dimension feature space. Following the backbone are two sub-networks

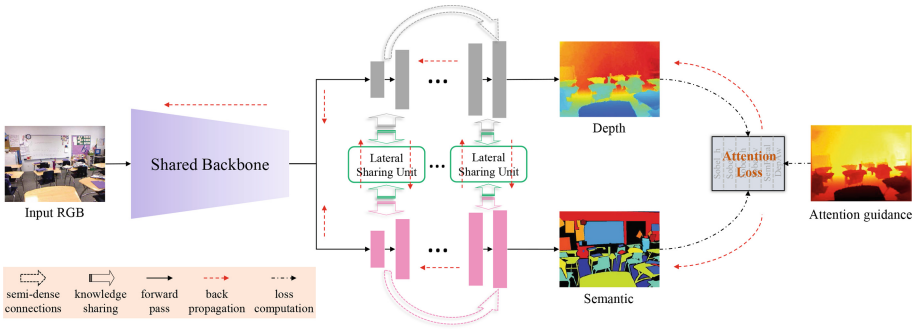

Fig. 2. Overview of the proposed network architecture. A single RGB image is fed into the shared backbone encoder network (purple), and then decoupled to the depth prediction (grey) and semantic labeling (pink) sub-networks. Knowledge between the two sub-networks is shared through lateral sharing units (details in Fig. 3 left) for both inference and backpropagation, together with internal sharing by semi-dense up-skip connections (Fig. 3 right). The training is supervised by an attention loss (Sect. 3.3). (Color figure online)

reconstructing the depth and semantic labels from the shared high-dimension feature. Knowledge sharing between these two tasks is achieved by a Lateral Sharing Unit (LSU), which is proposed to automatically learn the propagation flow during the training process and results in an optimum structure at test time. Besides, knowledge sharing is also performed internally at each sub-network through the proposed semi-dense up-skip connections (SUC). Finally, the whole training process is supervised by an attention-driven loss which consists of the proposed depth-aware and other attention-based loss terms.

Lateral Sharing Unit. We empirically explore different information sharing structures, which reveals that different multi-task networks result in diverse performance and the knowledge sharing strategy is hard to tune manually. In our synergy network, we propose a bi-directional *Lateral Sharing Unit* (LSU) to automatically learn the sharing strategy in a dynamic routing fashion. Information sharing is achieved for both forward pass and backpropagation. Between every two up-conv layers in the network, we add such LSU to share residual knowledge/representations from the other task, in addition to the intra-task propagation. Different from hand-tuned structures, our LSU is able to acquire additional fractional sharing from inter and intra-task layers. Specifically, the structure of LSU is illustrated in Fig. 3 left, which provides fully-sharing routes

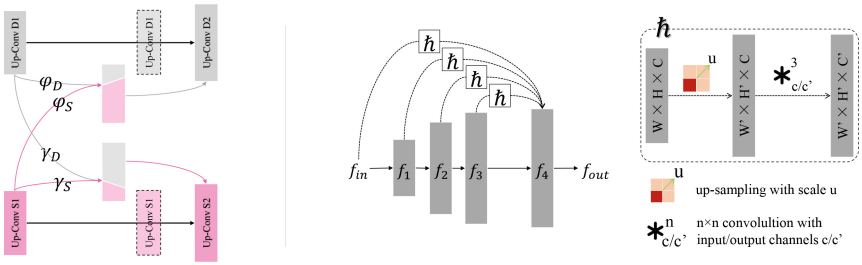

Fig. 3. Left: Structure of the proposed lateral sharing unit at every two consecutive up-conv layers D1 and D2, with identity mappings (black links). **Right:** Structure of the proposed semi-dense up-skip connections; dotted lines indicate up-skip connections, with operator $\hbar$ (bilinear up-sampling with convolution) shown on the right.

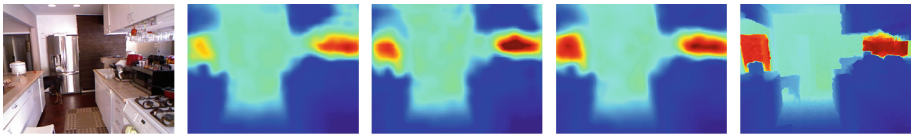

Fig. 4. Illustration on the effectiveness of LSU. All depth maps are with the same scale.

between the two tasks. Suppose the feature maps generated by current up-conv layers are D1 and S1. Then the feature representation for sharing can be formed as,

$$\begin{cases} LSU_{D2} = D1 + (\varphi_D \cdot D1 + \varphi_S \cdot S1) \\ LSU_{S2} = S1 + (\gamma_D \cdot D1 + \gamma_S \cdot S1) \end{cases}, \tag{3}$$

where φ_D, γ_D are the weighted parameters for feature D1, and φ_S, γ_S for feature S1. The sharing representations LSU_{D2} and LSU_{S2} are propagated to the subsequent up-conv layers. Note all the parameters in LSU are learnt during training, resulting in dynamic sharing route between every two up-conv layers. Although all LSUs share same internal structure, their parameters are not tied, allowing for a more flexible sharing. We propose to add identity mappings in addition to the combined sharing. With identity mappings, the intra-task information propagation is ensured, avoiding the risk of "propagation interruption" or feature pollution. Such residual-like structure (identity connection [15] associated with the residual sharing) also benefits efficient backpropagation of gradients. In addition, our LSU is applied between consecutive up-conv layers, instead of the encoding backbone. In this way, much fewer combination parameters and network parameters need to learn. An example illustrates the effectiveness of our LSU is shown in Fig. 4. We can see that when incorporating LSU, semantics is propagated to the depth thus improve its accuracy (the top-right cabinet). Whereas if without the identity mapping, artifacts may also be introduced by the semantic propagation (bottom-right cabinet). With identity mapping, less artifacts and higher accuracy can be achieved (the fourth column).

Semi-dense Up-skip Connections. In order to perform better intra-task knowledge sharing and preserve long-term memory, we introduce the *Semi-dense Up-skip Connections* (SUCs) between up-conv layers, as shown in Fig. 2 and detailed in Fig. 3 right. Denote f_{in} and f_{out} as the input and output features of the decoder, the output features of each up-conv layer as f_i. In addition to the short-term memory from preceding single up-conv layer, we add skip connections to propagate long-term memory. Therefore, our SUC is formulated as,

$$f_{out} = \hbar(f_{in}) + \sum_{i=1}^{n} \hbar(f_i), \tag{4}$$

where n is the number of up-conv layers ($n = 4$ in our work), and $\hbar$ denotes an up-resize operation in order to match the size of feature in the last up-conv layer. We also tried the concatenation of features which performs slightly worse than the summation. Our SUC is performed in a semi-dense manner between adjacent up-conv layers, instead of fully-dense in the encoder. In this way, the memory consumption is reduced to a large extent without performance sacrifice according to our experiment. In addition, with long- short-term connections the features from different up-conv steps are able to fuse in a coarse-to-fine multi-scale fashion, which incorporates both global and local information.

3.3 Attention-Driven Loss

Depth-Aware Loss. As defined in Sect. 3.1, during training, we use depth-aware loss term (Eq. 1) to supervise the depth prediction task. Specially, we set the attention term $\alpha_D = d^{GTn}$ where d^{GTn} is the normalized ground truth depth (attention guidance in Fig. 2) over whole range. The distance metric ℓ is set as reverse smooth L_1-norm [8,28] due to its robustness.

Joint Gradient Loss. In order to better preserve details on local structure and surface regions, we propose to set constraints on gradients and introduce the gradient loss layers with kernels set as the Sobel detector in both horizontal (∇_h) and vertical (∇_v) directions,

$$L_g(d, d^{GT}) = \frac{1}{N} \sum_{i=1}^{N} \left|\nabla_h d_i - \nabla_h d_i^{GT}\right| + \left|\nabla_v d_i - \nabla_v d_i^{GT}\right|. \tag{5}$$

In addition, the semantic information is also taken into consideration as a joint gradient loss term, by substituting the semantic segmentation result s for d^{GT} as: $L_g(d, s)$. Then the joint gradient loss term is formulated as $L_{JG} = L_g(d, d^{GT}) + L_g(d, s)$.

Semantic Focal Loss. As shown in Fig. 1 (c–e), the category distribution also belongs to a long-tailed one, even mapping to much fewer number (*e.g.* 40 or 4) of categories. Such imbalanced distribution not only influences the semantic

labeling task but also impacts the depth prediction through LSUs and back-propagation. Inspired by the Focal Loss [32] proposed for object detection, we propose to guide the network to pay more attention to the hard tailed categories and set the loss term as,

$$L_{semF}(l, l^{GT}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} l_{i,k}^{GT}\alpha_k(1-l_{i,k})^{\gamma}\log(l_{i,k}), \tag{6}$$

where l_i is the label prediction at pixel i and k is the category index. α_k and γ are the balancing weight and focusing parameter to modulate the loss attention.

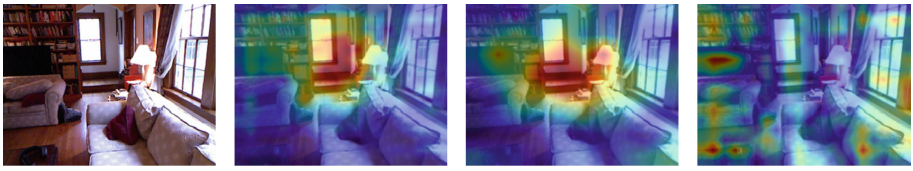

Fig. 5. Network attention visualization. Given an input RGB image, the spatial attention of the network is shown as an overlay to the input.

The above loss terms/layers consist the proposed attention-driven loss as in Fig. 2, which is defined as,

$$L_{attention} = L_{DA} + L_{JG} + L_{semF}. \tag{7}$$

3.4 Attention Visualization

In order to better illustrate the proposed attention-driven loss, we visualize the learned attention of the network, *i.e.* which region the network focuses more on. Following [54], here we use the spatial attention map to show the network attention. The attention maps of the network on monocular depth estimation in shown in Fig. 5 (second column) as heat-map, where red indicates high values. Note that the attention map here is different from the attention guidance in Fig. 2, although they share the similar high-level meaning. Here the attention map is represented by the aggregation of the feature activations from the first up-conv layer. In addition to the depth estimation, the attention maps of the shared backbone and semantic labeling are also presented for a thorough understanding of the network attention distribution in Fig. 5.

From the visualization we can see the network mainly focuses on distant regions when performing monocular depth estimation. On the other hand, the shared backbone focuses on a larger region around the distant area, indicating a more general attention on the whole scene while still driven by the distance. For the attention of semantic labeling, besides the dominant categories, some "tailed"

categories also receive high attention, *e.g.* television, books, bag, *etc.* The above attention visualization results provide a better understanding of the network focus and validate the mechanism of the proposed attention-driven approach.

4 Experiments

In this section, we evaluate the proposed approach on monocular depth estimation, and compare to state-of-the-art methods. Performance on semantic labeling is also presented to show the benefits of knowledge sharing.

4.1 Experimental Setup

Dataset and Evaluation Metrics. We use the NYU Depth v2 (NYUD2) dataset [46] for our evaluation, which consists of 464 different indoor scenes with 894 different object categories (distributions shown in Fig. 1). We follow the standard train/test split with 795 aligned (RGB, depth) pairs for training, and 654 pairs for testing, as adopted in [35,53,56]. Besides, each of the standard splits images is manually annotated with semantic labels. In our experiment, we map the semantic labels into 4 and 40 categories, according to [46] and [10], respectively. We perform data augmentation on the training samples by random in-plane rotation ($[-5°, +5°]$), translation, horizontal flips, color (multiply with RGB value $\in [0.8, 1.2]^3$) and contrast (multiply with value $\in [0.5, 2.0]$) shift.

We quantitatively evaluate the performance of monocular depth prediction using the metrics of: mean absolute relative error (rel), mean $\log_{10}$ error ($\log 10$), root mean squared error (rms), rms(log), and the accuracy under threshold ($\delta < 1.25^i, i = 1, 2, 3$), following previous works [4,9,28,51].

Implementation Details. We implement our proposed deep model on a single Nvidia Tesla K80 GPU, using the PyTorch [40] framework. In our final model, the ResNet-50 [14] pre-trained on ImageNet is taken as our shared backbone network, by removing the last classification layers. The structure of decoder layers are set following state-of-the-art designs [28,53]. All the other parameters in the depth decoder, semantic decoder, SUCs, and LSUs are randomly initialized by the strategy in [13] and trained from scratch. We train our model with a batch size of 12 using the Adam solver [25] with parameters $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$. α, γ are set with reference to [32]. The images are first down-sampled to half size with invalid borders cropped, and at the end up-sampled to the original size using techniques similar to previous works [4,30,35]. We first freeze the semantic branch with all the LSUs, and train the rest model for depth prediction with a learning rate of 10^{-3}. Then freeze the depth branch and train the rest with learning rate of 10^{-5} on backbone and 10^{-3} on semantic branch. Finally, the whole model is trained end-to-end with initial learning rate 10^{-4} for backbone and 10^{-2} for others. The learning rate is decreased by 10 times every 20 epochs.

4.2 Experimental Results

Architecture Analysis. We first compare different settings of the network architecture: depth-only branch, *i.e.* ResNet with up-convs; with the SUC; with our proposed depth-aware loss (L_{DA}); adding semantic branch with and without LSUs. To better illustrate the effectiveness of the proposed knowledge sharing strategy, we also include the CS structure [38] (substitutes LSU) for comparison. Our final method with the attention-driven loss is compared to these baselines. In this analysis the semantic labels are mapped to 4 categories. The comparison results are shown in Table 1, where we can see the performance is continuously improved by incorporating each term. Specifically, after introducing the proposed depth-aware loss, performance among all the metrics are improved by a large margin. We note the CS structure do benefits representation sharing while our LSU performs slightly better. The synergy boosting from semantic labeling task also benefits a lot to the depth estimation. To summarize, the attention-driven loss contributes most to the performance, with secondary contributions of knowledge sharing from semantic labeling.

Table 1. Architecture analysis. Results are shown on NYUD2 dataset with 4-category mapped as semantic labeling task

Method	Lower is better				Higher is better		
	rel	log10	rms	rms (log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
depth	0.157	0.062	0.642	0.208	0.763	0.943	0.985
+SUC	0.147	0.057	0.572	0.192	0.797	0.951	0.987
+SUC+L_{DA}	0.126	0.050	0.416	0.154	0.868	0.973	0.993
+SUC+L_{DA}+sem.	0.112	0.045	0.367	0.140	0.896	0.978	0.994
+SUC+L_{DA}+sem.+CS	0.110	0.044	0.363	0.138	0.898	0.979	0.995
+SUC+L_{DA}+sem.+LSU	0.105	0.042	0.351	0.133	0.906	0.980	0.995
Proposed	0.100	0.040	0.333	0.127	0.915	0.983	0.996

Table 2. Analysis on robustness to data "tail". Study performed on NYUD2 with 4-category mapped semantic labels

Depth range	Lower is better				Higher is better		
	rel	log10	rms	rms (log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
$\leq$4 m	0.105	0.042	0.300	0.130	0.908	0.981	0.995
$\leq$6 m	0.101	0.041	0.326	0.127	0.915	0.983	0.996
$\leq$8 m	0.100	0.040	0.326	0.127	0.915	0.983	0.996
All	0.100	0.040	0.333	0.127	0.915	0.983	0.996

Robustness to "Tail". In order to validate the robustness of the proposed approach to long-tailed data, we perform an ablation study on the tailed part of the data. Specifically, we divide the depth range of the test data into four parts by cutting corresponding tails by 2 m for each (*i.e.*, $\leq$4 m, 6 m, 8 m, 10 m). Then we evaluate our method on these depth ranges as shown in Table 2. From the table we can see that even our attention-driven loss supervises the network to focus more on distant depth, it performs well on shorter-tailed data and consistently among different ranges, which indicates the proposed attention loss is able to adaptively vary according to the data distribution. In addition, our method also achieves state of the art even on nearby depth.

Table 3. Comparison with state-of-the-art methods on NYUD2 dataset. The last two rows show the proposed approach with 4 and 40 semantic categories, respectively

Method	Lower is better				Higher is better		
	rel	log10	rms	rms (log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Karsch *et al.* [21]	0.349	0.131	1.214	-	0.447	0.745	0.897
Ladicky *et al.* [27]	-	-	-	-	0.542	0.829	0.941
Liu *et al.* [36]	0.335	0.127	1.06	-	-	-	-
Zhuo *et al.* [56]	0.305	0.122	1.04	-	0.525	0.838	0.962
Li *et al.* [29]	0.232	0.094	0.821	-	0.621	0.886	0.968
Liu *et al.* [34]	0.230	0.095	0.824	-	0.614	0.883	0.975
Eigen *et al.* [5]	0.215	-	0.907	0.285	0.611	0.887	0.971
Roy and Todorovic [43]	0.187	0.078	0.744	-	-	-	-
Eigen and Fergus [4]	0.158	-	0.641	0.214	0.769	0.950	0.988
Laina *et al.* [28]	0.127	0.055	0.573	0.195	0.811	0.953	0.988
Xu *et al.* [53]	0.121	0.052	0.586	-	0.811	0.954	0.987
Li *et al.* [30]	0.143	0.063	0.635	-	0.788	0.958	0.991
Wang *et al.* [51]	0.220	0.094	0.745	0.262	0.605	0.890	0.970
Mousavian *et al.* [39]	0.200	-	0.816	0.314	0.568	0.856	0.956
Jafari *et al.* [19]	0.157	0.068	0.673	0.216	0.762	0.948	0.988
Laina *et al.* [28]+sem.	0.122	0.052	0.525	0.184	0.813	0.958	0.989
Proposed-4c	*0.100*	**0.040**	*0.333*	*0.127*	*0.915*	**0.983**	**0.996**
Proposed-40c	**0.098**	**0.040**	**0.329**	**0.125**	**0.917**	**0.983**	**0.996**

Comparison with State-of-the-Art. We also compare other state-of-the-art methods with the proposed approach. Here we directly use the reported results in their original papers. The comparison results on NYUD2 is shown in Table 3. For our approach, we consider two sharing settings with the semantic labeling task: sharing information from 4 mapped categories, and 40 mapped categories, as shown in the last two rows. From the results in Table 3 we can see,

our approach performs favorably against other state-of-the-art methods. Note that [19,39,51] also utilize the semantic labeling information in a joint prediction manner, which perform not as well as ours. We also include a state-of-the-art method [28] accompanied a semantic labeling branch for better understanding of the semantic booster. The improvement over [28] favorably validates the effectiveness of adding semantic task, while information sharing is still underexplored. Another observation is that using more categories benefits the depth prediction, since it provides more semantic information of the objects in the scene.

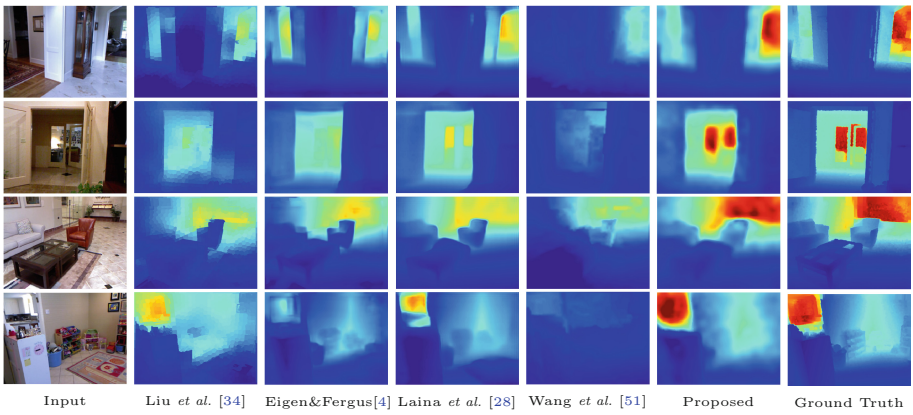

Fig. 6. Qualitative results on the NYUD2 dataset. Our method predicts more accurate depth compared to other state-of-the-art methods, especially on distant regions. Depth maps are in the same range with ground truth. Warm color indicates large depth.

Table 4. Evaluation of semantic labeling on the NYUD2-40

Method	Input	Pix. acc.	Mean acc.	IoU
FCN [37]	RGB-D	65.4	46.1	34.0
Eigen & Fergus [4]	RGB-D	65.6	45.1	34.1
Mousavian *et al.* [39]	RGB	68.6	52.3	39.2
RefineNet [31]	RGB	73.6	58.9	46.5
3DGNN [41]	RGB-D	-	55.7	43.1
Baseline	RGB	69.0	50.5	39.9
Without depth	RGB	75.7	55.7	48.9
Proposed	RGB	**81.1**	**62.2**	**50.9**

In addition to the quantitative comparison, some qualitative results are also presented in Fig. 6. All the depth maps are shown in the same range with the ground truth for better comparison. As we can see in the figure, the proposed

method predicts more accurate depth values compared to other methods. For instance, the large-depth (red) regions in these examples, and the wall region in the last example. Furthermore, semantic prior also benefits the depth prediction, *e.g.* the floor mat in the last example should have similar depth to the floor instead of floating. This again validates the effectiveness of the proposed approach, which focuses more on hard distant depth and object semantic meaning.

Semantic Labeling. Although the semantic labeling task is incorporated to perform knowledge sharing and boost the depth prediction task, the proposed network infers a semantic segmentation map as well. Here we evaluate whether the depth prediction task benefits semantic labeling, by three metrics in percentage (%): pixel accuracy, mean accuracy, Intersection over Union (IoU). We set the model without depth branch and L_{semF} as a baseline, and the model with L_{semF} (without depth) for comparison. Other semantic segmentation methods are also included for comparison (with their reported performance). The results on NYUD2 dataset with mapped 40 categories are shown in Table 4. As the table shows, our inferred semantic result achieves state-of-the-art performance as well. We note that without the depth information, our model still performs favorably against [4] and [37] which take RGB-D as input. This validates the effectiveness of the proposed SUC and L_{semF} to some extent. We also compare with [19,51] which mapped the raw data to 5 categories, different from the standard 4-category. After fine-tuning our 4-category model on their data, we achieve a result of (87.11, 66.77) on (pix.acc., IoU), with respect to (70.29, 44.20) from [51] and (73.04, 54.27) from [19].

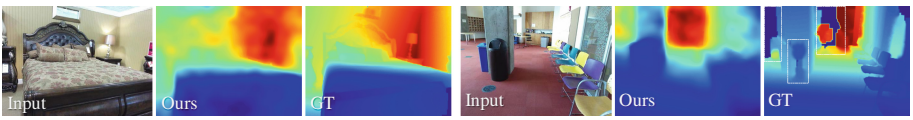

Fig. 7. Results on SUN. Some regions (white boxes) are difficult even to capture GT.

Generalization Analysis. In addition to the NYUD2 dataset, we further explore the generalization ability of our model to other indoor and outdoor scenes. Performance on another indoor dataset SUN-RGBD [48] is shown in Fig. 7, where *Ours* are predicted by our original model without finetuning on SUN. The results show that even SUN differs from NYU in data distribution, our model could predict plausible results. For outdoor scenes, we fine-tune the indoor model on 200 standard training images (with sparse depth and semantic labels) from the KITTI dataset [7]. The performance is (RMSE, RMSElog, $\delta < 1.25, \delta < 1.25^2$, $\delta < 1.25^3$) = (5.110, 0.215, 0.843, 0.950, 0.981), following the evaluation setups in [9,26]. We also evaluate on the Cityscapes dataset [2],

following the setups in [23]. The (Mean Error, RMSE) on the converted disparity is (2.11, 4.92), in comparison to (2.92, 5.88) for [23]. The above evaluations reveal that despite the difference in distribution and scene structures, our model is shown to have the generalization ability to other datasets.

5 Conclusions

We have introduced an attention-driven learning approach for monocular depth estimation, which also predicts corresponding accurate semantic labels. In order to predict accurate depth information for the whole scene, we delve into the *deeper* part of the scene and propose a novel attention-driven loss that supervises the training in an attention-driven manner. We have also presented a sharing strategy with LSU and SUC, to better propagate both inter- and intra-task knowledge. Experimental results on NYUD2 dataset showed that the proposed method performs favorably against state-of-the-arts, especially on hard distant regions. We have also shown the generality of our model to other datasets/scenes.

Acknowledgments. This work is partially supported by the Hong Kong PhD Fellowship Scheme (HKPFS) from the RGC of Hong Kong.

References

1. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NIPS (2016)
2. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
3. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information (2013). arXiv preprint arXiv:1301.3572
4. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS (2014)
6. Garg, R., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: geometry to the rescue. In: ECCV (2016)
7. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: CVPR (2012)
8. Girshick, R.: Fast r-cnn. In: ICCV (2015)
9. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
10. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from rgb-d images. In: CVPR (2013)
11. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV (2014)
12. Hane, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3d scene reconstruction and class segmentation. In: CVPR (2013)
13. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: ICCV (2015)

14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016)
16. He, S., Jiao, J., Zhang, X., Han, G., Lau, R.W.: Delving into salient object subitizing and detections. In: ICCV (2017)
17. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. ACM TOG **24**(3), 577–584 (2005)
18. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
19. Jafari, O.H., Groth, O., Kirillov, A., Yang, M.Y., Rother, C.: Analyzing modular cnn architectures for joint depth prediction and semantic segmentation. In: ICRA (2017)
20. Jiao, J., Yang, Q., He, S., Gu, S., Zhang, L., Lau, R.W.: Joint image denoising and disparity estimation via stereo structure pca and noise-tolerant cost. IJCV **124**(2), 204–222 (2017)
21. Karsch, K., Liu, C., Kang, S.B.: Depth transfer: depth extraction from video using non-parametric sampling. IEEE TPAMI **36**(11), 2144–2158 (2014)
22. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS (2017)
23. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR (2018)
24. Kim, S., Park, K., Sohn, K., Lin, S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: ECCV (2016)
25. Kingma, D., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
26. Kuznietsov, Y., Stückler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: CVPR (2017)
27. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: CVPR (2014)
28. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3D Vision (3DV) (2016)
29. Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: CVPR (2015)
30. Li, J., Klein, R., Yao, A.: A two-streamed network for estimating fine-scaled depth maps from single rgb images. In: ICCV (2017)
31. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: CVPR (2017)
32. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
33. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: CVPR (2010)
34. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: CVPR (2015)
35. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE TPAMI **38**(10), 2024–2039 (2016)
36. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: CVPR (2014)

37. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
38. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: CVPR (2016)
39. Mousavian, A., Pirsiavash, H., Košecká, J.: Joint semantic segmentation and depth estimation with deep convolutional networks. In: 3D Vision (3DV) (2016)
40. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
41. Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R.: 3d graph neural networks for rgbd semantic segmentation. In: ICCV (2017)
42. Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: features and algorithms. In: CVPR (2012)
43. Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: CVPR (2016)
44. Saxena, A., Sun, M., Ng, A.Y.: Make3d: learning 3d scene structure from a single still image. IEEE TPAMI **31**(5), 824–840 (2009)
45. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: CVPR (2016)
46. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
48. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: a rgb-d scene understanding benchmark suite. In: CVPR (2015)
49. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR (2017)
50. Del Giorno, A., Bagnell, J.A., Hebert, M.: A discriminative framework for anomaly detection in large videos. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 334–349. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_21
51. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: CVPR (2015)
52. Wang, P., Shen, X., Russell, B., Cohen, S., Price, B., Yuille, A.L.: Surge: surface regularized geometry estimation from a single image. In: NIPS (2016)
53. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: CVPR (2017)
54. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
55. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
56. Zhuo, W., Salzmann, M., He, X., Liu, M.: Indoor scene structure analysis for single image depth estimation. In: CVPR (2015)
57. Zoran, D., Isola, P., Krishnan, D., Freeman, W.T.: Learning ordinal relationships for mid-level vision. In: ICCV (2015)