Structure-Aware Residual Pyramid Network for Monocular Depth Estimation

Xiaotian Chen, **Xuejin Chen*** and **Zheng-Jun Zha**
National Engineering Laboratory for Brain-inspired Intelligence Technology and Application
University of Science and Technology of China
ustcxt@mail.ustc.edu.cn, {xjchen99, zhazj}@ustc.edu.cn

Abstract

Monocular depth estimation is an essential task for scene understanding. The underlying structure of objects and stuff in a complex scene is critical to recovering accurate and visually-pleasing depth maps. Global structure conveys scene layouts, while local structure reflects shape details. Recently developed approaches based on convolutional neural networks (CNNs) significantly improve the performance of depth estimation. However, few of them take into account multi-scale structures in complex scenes. In this paper, we propose a Structure-Aware Residual Pyramid Network (SARPN) to exploit multi-scale structures for accurate depth prediction. We propose a Residual Pyramid Decoder (RPD) which expresses global scene structure in upper levels to represent layouts, and local structure in lower levels to present shape details. At each level, we propose Residual Refinement Modules (RRM) that predict residual maps to progressively add finer structures on the coarser structure predicted at the upper level. In order to fully exploit multi-scale image features, an Adaptive Dense Feature Fusion (ADFF) module, which adaptively fuses effective features from all scales for inferring structures of each scale, is introduced. Experiment results on the challenging NYU-Depth v2 dataset demonstrate that our proposed approach achieves state-of-the-art performance in both qualitative and quantitative evaluation. The code is available at *https://github.com/Xt-Chen/SARPN*.

1 Introduction

Monocular depth estimation, which aims to predict the depth value of each pixel from a given RGB image, is crucial for understanding scene geometry, and can be applied to facilitate other vision tasks, such as semantic segmentation [Park *et al.*, 2017] and hand tracking [Qian *et al.*, 2014]. It is an ill-posed problem because of the inherent ambiguity due to perspective projection. Recently, CNN-based approaches have achieved significant success in monocular depth estimation [Laina *et al.*, 2016; Fu *et al.*, 2018; Xu *et al.*, 2018b; Hao *et al.*, 2018; Hu *et al.*, 2019]. To resolve the ambiguity, they typically employ an encoder-decoder architecture to implicitly fuse features that represents object appearance, geometry, semantics, spatial relations, etc. The encoder gradually extracts multi-scale features, and the decoder employs multi-stage upsampling as well as shortcut connections to restore object details in high-resolution predictions.

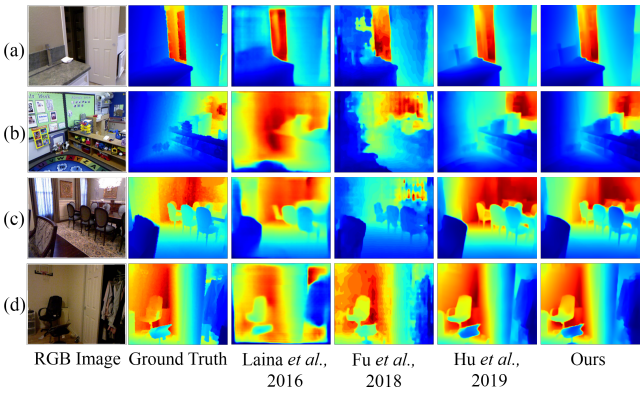

Figure 1: Problems in depth prediction: (a)(b) inaccurate depth values on large planar regions, such as walls. (c)(d) blurry boundaries and missing details (chair legs). Our approach simultaneously recovers large planar structures and object details.

Though a great improvement on average pixel-wise metrics has been made, the underlying structure of objects and stuff is not well preserved by current CNN-based methods. The problem becomes especially challenging when the size of objects and stuff varies widely in complex scenes. As Figure 1 shows, it is challenging for existing approaches to accurately recover the large-scale geometry (walls) and local details (boundaries and small parts) at the same time. This inaccurate inference at regions of diverse scales motivates us to fully exploit the hierarchical scene structure in depth prediction. Scene structure, depicting the organization and arrangement of multiple interrelated elements in a complex scene, varies widely according to the element type. The global structure represents the spatial arrangement of large-size elements such as walls, floors, and furniture objects. Local structure describes geometric details of objects and their parts. The natural hierarchy of scene structure provides essential con-

*Corresponding author

straints between the depth values of pixels in multiple scales. Although previous CNN-based techniques extract multi-scale image features and gradually fuse them to predict a depth map, the underlying hierarchical structure of the scene has not been taken into account.

In this paper, we introduce a Structure-Aware Residual Pyramid Network (SARPN) to fully exploit scene structures in multiple scales for depth prediction. A Residual Pyramid Decoder (RPD) is proposed to predict multi-scale depth maps in a coarse-to-fine manner. Depth maps in upper levels in the pyramid represent the global scene structure, while depth maps in lower levels capture more local structures of objects or parts. To convey the global structure and constrain the generation of finer details, we proposed a residual refinement module to predict residual depth maps, which progressively add details on the scene structure on a larger scale. In order to fuse multi-scale features extracted from the input image for residual prediction, we propose an Adaptive Dense Feature Fusion (ADFF) module to adaptively select more effective features for each scale. Integrating the residual pyramid decoder and adaptive dense feature fusion module, our method simultaneously preserves the hierarchical scene structures and produces accurate depth estimation for both large-size shapes and fine details of small object parts, as Figure 1 shows. Our contributions are summarized as follows:

- We propose a Structure-Aware Residual Pyramid Network (SARPN), which takes the underlying scene structure in multiple scales into account for accurate depth prediction.
- Our Adaptive Dense Feature Fusion (ADFF) module adaptively selects features from all scales to predict residual depths at different structure scales.
- The proposed method achieves state-of-the-art performance on the challenging NYUD v2 dataset. More importantly, the visual quality of recovered depth maps is significantly improved.

2 Related Work

In recent years, CNNs have become the most successful techniques for various visual tasks, and were firstly used for monocular depth estimation [Eigen *et al.*, 2014] in a multiple scale scheme. Later on, fully convolutional network (FCN) was proposed for semantic segmentation [Long *et al.*, 2015] and has been widely used in many dense prediction tasks, including depth estimation.

When FCN-based architecture was first adopted for depth estimation, the resolution and accuracy ware largely improved by using ResNet to extract features and up-projection blocks [Laina *et al.*, 2016]. In order to improve the quality of depth estimation for local details, many strategies have been introduced. Applying conditional random field as post-processing [Li *et al.*, 2015] or integrating it in CNNs [Xu *et al.*, 2017] largely improves the prediction quality for small objects. Later, an attention model is integrated to improve the estimation performance [Xu *et al.*, 2018b]. Multi-scale architecture becomes a common solution to avoid the loss of local details caused by spatial pooling and convolutions [Fu *et al.*, 2018]. Instead of multi-scale network structure, dilated convolution is used to extract multi-scale features for depth estimation [Hao *et al.*, 2018]. Hu et al. [2019] proposed an effective multi-scale feature fusion module to produce clear object boundaries. Although these methods have achieved remarkable results by fusing multi-scale features, they still face the problem of inaccurate prediction for complex scenes of which the structure varies largely in scales, from large room layout to fine object details.

In order to better restore structure details, a few methods design new loss functions to explicitly constrain scene geometry. Zheng et al. [2018] proposed an order-sensitive softmax loss to constrain global layouts. Similarly, Fu et al. [2018] used an ordinary regression loss. With respect to clear boundaries and details, a loss function is designed by combining depth, surface normal and gradient in a local neighborhood of depth maps [Hu *et al.*, 2019].

Due to the strong correlation between many visual tasks, such as depth estimation, semantic segmentation, and normal estimation, many approaches employ a joint task learning framework. A multi-scale CNN was designed to simultaneously perform semantic segmentation, depth estimation, and normal estimation [Eigen and Fergus, 2015]. A set of intermediate auxiliary tasks are utilized to guide the final depth estimation and semantic segmentation [Xu *et al.*, 2018a]. Zhang et al. proposed a novel joint task-recursive learning method to recursively refine the results of depth estimation and semantic segmentation [Zhang *et al.*, 2018]. A synergy network is proposed to automatically learn information sharing strategy between depth estimation and semantic segmentation [Jiao *et al.*, 2018]. Moreover, based on the observed long-tail distribution of depth values, an attention-driven loss is also designed to improve the accuracy [Jiao *et al.*, 2018].

3 Methodology

Our network consists of three main parts: an encoder for multi-scale feature extraction, an adaptive dense feature fusion module, and a residual pyramid decoder, as Figure 2 shows. We first introduce the network architecture in Sec. 3.1. The residual pyramid decoder and adaptive dense feature fusion module are explained in Sec. 3.2 and 3.3, respectively.

3.1 Structure-Aware Residual Pyramid Network

Our approach begins with an encoder which extracts multi-scale features $\{\mathbf{F}_{ex}^i\}_{i=1}^L$ from the input image, where $\mathbf{F}_{ex}^i$ indicates the feature maps extracted at the i-th level. L is the number of layers in our network. Following the state-of-the-art approach [Hu *et al.*, 2019], we use SENet [Hu *et al.*, 2018] as the backbone of our encoder. It extracts more effective features by re-weighting features of different channels. Given an input image with size $W \times H$, the size of these feature maps are respectively $[\frac{W}{2^i}, \frac{H}{2^i}]$, and they carry both high-level semantic information and low-level detail information. Then, these multi-scale feature maps are simultaneously fed to our dense feature fusion module to produce a Fused Feature Pyramid (FFP). These feature maps in FFP are represented by $\{\mathbf{F}_{fs}^i\}_{i=1}^L$, where $\mathbf{F}_{fs}^i$ indicates the fused feature maps at the i-th level of the pyramid of fused features.

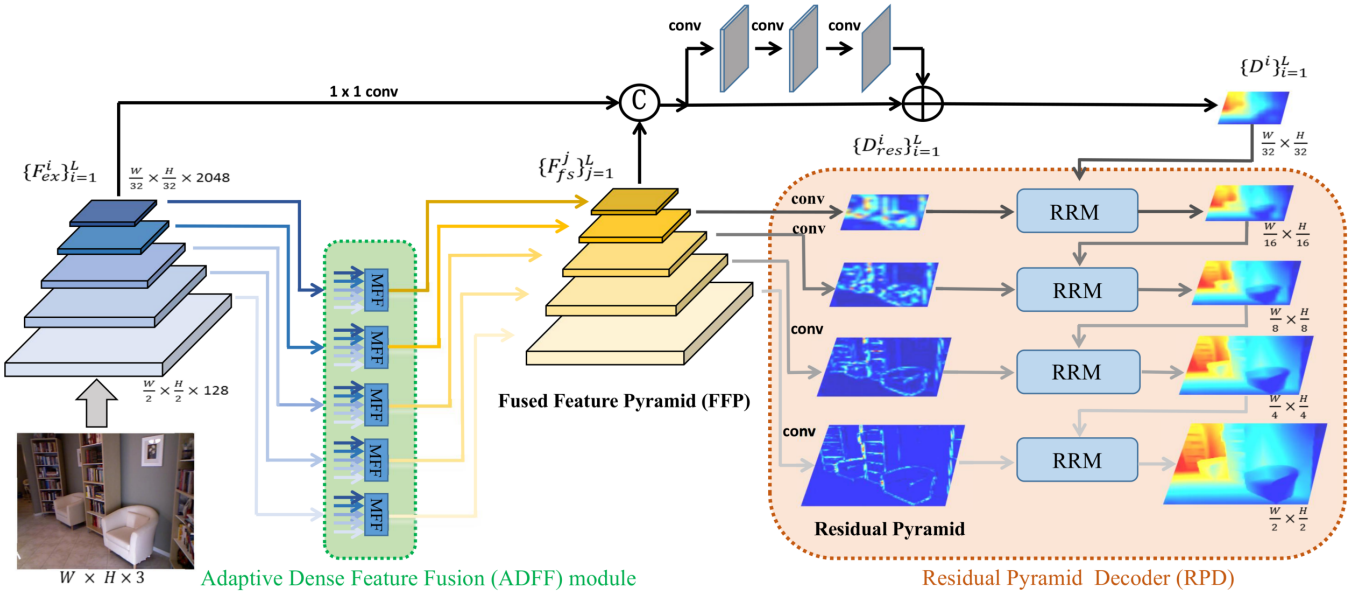

Figure 2: The network architecture. Our Structure-Aware Residual Pyramid Network consists of an encoder which extracts multi-scale visual features, a Residual Pyramid Decoder (RPD) which progressively infers depth maps in a coarse-to-fine manner, and an Adaptive Dense Feature Fusion (ADFF) module for dense feature fusion. The residual pyramid effectively adds structure details in each level based on the scene layout predicted at a coarser level.

In the decoder part, different from the previous methods that directly predict a depth map by sequentially upsampling feature maps [Laina *et al.*, 2016; Hu *et al.*, 2019], our residual pyramid progressively predicts multiple depth maps in a coarse-to-fine manner. The depth map at the top level with size $\frac{W}{32} \times \frac{H}{32}$ is predicted first as the initial scene layout. We utilize a 1×1 convolution operation to reduce the channel number of the feature maps $\mathbf{F}^L_{ex}$ to the same as the channel number of feature maps $\mathbf{F}^L_{fs}$ of fused feature pyramid and concatenate them together. A residual block is used to predict a depth map $\mathbf{D}^L$ in size of $[\frac{W}{2^L}, \frac{H}{2^L}]$ from the concatenated feature maps. Then we gradually refine the depth prediction by our proposed residual pyramid decoder.

3.2 Residual Pyramid Decoder

Our residual pyramid decoder predicts depth maps of multiple scales in order to restore the hierarchical scene structures in a coarse-to-fine manner. As shown in Figure 2, the depth maps in lower resolutions depicts more global scene layout, while the depth maps in higher resolutions contain more structure details. In each level of the pyramid decoder, we predict a residual map instead of a dense depth map from fused image features in FFP. The residual map and the depth map predicted at the upper level are integrated together to produce a refined depth map in the current scale using our Residual Refinement Module (RRM). The components of each RRM are shown in Figure 3. The depth map $\mathbf{D}^{i+1}$ predicted at the upper scale is upsampled to the current scale by bilinear interpolation. A residual depth map $\mathbf{D}^i_{res}$ is generated by utilizing the fused features $\mathbf{F}^i_{fs}$. After adding the residual map and the upsampled depth map, a residual block, which contains three convolutional layers, is employed to refine the prediction and outputs a depth map $\mathbf{D}^i$ at the i-th scale. This residual architecture induces our network to effectively represent the structure details at each scale and hierarchically refine scene structures. Meanwhile, the global scene layout is well preserved by our residual pyramid decoder.

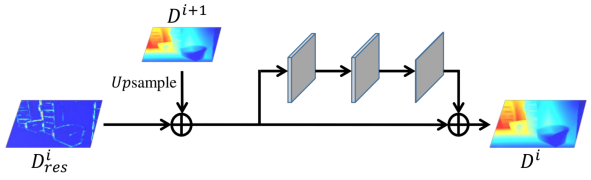

Figure 3: A Residual Refinement Module (RRM) for the i-th level.

3.3 Adaptive Dense Feature Fusion

In general, due to pooling operations and convolution operations with strides in CNNs, a large amount of low-level visual features are lost. As a result, it is difficult for the decoder to recover the lost low-level structure details. However, both low-level features and high-level features are critical for predicting residual maps in all layers, because the residual maps convey additional details on a global scene structure, as the residual pyramid illustrates in Figure 2. In order to provide sufficient information for the prediction of a residual map in each level, we propose an Adaptive Dense Feature Fusion (ADFF) module. This dense fusion module consists of L Multi-scale Feature Fusion (MFF) modules to predict L fused feature maps, which compose a fused feature pyramid for residual prediction.

In each layer, the MFF adaptively selects eligible features from all feature scales when predicting the depth map for

Method	REL	RMS	log 10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ladicky et al. [2014]	-	-	-	0.542	0.829	0.941
Li et al. [2015]	0.232	0.821	0.094	0.621	0.886	0.968
Eigen et al. [2014]	0.215	0.907	-	0.611	0.887	0.971
Laina et al. [2016]	0.127	0.573	0.055	0.811	0.953	0.988
Xu et al. [2017]	0.121	0.586	0.052	0.811	0.954	0.987
Xu et al. [2018b]	0.125	0.593	0.057	0.806	0.952	0.986
Hao et al. [2018]	0.127	0.555	0.053	0.841	0.966	0.991
Fu et al. [2018]	0.115	*0.509*	0.051	0.828	0.965	0.992
Qi et al. [2018]	0.128	0.569	0.057	0.834	0.960	0.990
Jiao et al. [2018]	0.126	**0.416**	0.050	0.868	0.973	0.993
Hu et al. [2019]	0.115	0.530	0.050	0.866	0.975	0.993
Our Baseline	0.123	0.547	0.052	0.854	0.969	0.992
Our Baseline + RPD	0.115	0.528	0.050	0.871	0.975	0.993
Ours: Baseline + RPD + ADFF	**0.111**	0.514	**0.048**	**0.878**	**0.977**	**0.994**
Eigen and Fergus [2015]*	0.158	0.641	-	0.769	0.950	0.988
Xu et al. [2018a]*	0.120	0.582	0.055	0.817	0.954	0.987
Zhang et al. [2018]*	0.144	*0.501*	-	0.815	0.962	0.992
Jiao et al. [2018]*	*0.098*	*0.329*	*0.040*	*0.917*	*0.983*	*0.996*

Table 1: Comparisons with state-of-the-art depth estimation approaches on NYUD v2 Dataset. Note that joint task learning is employed in the methods marked by *. The best results on each metric among the single-task approaches are marked in bold type. The results better than ours are marked in italics.

each individual scale. We follow the detailed implementation of MFF proposed in [Hu *et al.*, 2019]. The L feature maps $\{\mathbf{F}^i_{ex}\}_{i=1,\dots,L}$ are first resized to the resolution of current scale using bilinear interpolation and refined with a residual refine block. The refined feature maps are concatenated and fed into a conv-layer to reduce the number of channels.

3.4 Loss Function

In order to train our residual pyramid network for predicting accurate depth maps while preserving scene structures in various scales, we compute the difference between the predicted depth map $\mathbf{D}^i$ and the ground-truth $\mathbf{G}^i$ at each scale and combine the losses of all scales together. For each scale, we follow the definition of the loss function proposed in [Hu *et al.*, 2019]. It consists of three terms, l_{depth} considering the pixel-wise difference between the predicted depth $\mathbf{D}^l$ and the ground truth G^l, l_{grad} which penalizes errors round edges, and l_{normal} to further improve fine details. Combing all the L scales, our loss function for the entire network is formulated as

$$L = \sum_{i=1}^{L} l^i_{depth} + l^i_{grad} + l^i_{normal}. \quad (1)$$

4 Experiments

To demonstrate the effectiveness of the proposed approach, we evaluate our approach on the challenging NYUD v2 dataset [Silberman *et al.*, 2012]. We compare our approach with a couple of state-of-the-art approaches and show the superiority of the proposed method on both quantitative and qualitative evaluations.

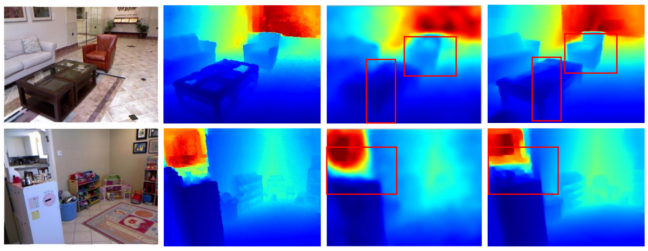

Figure 4: Comparison with [Jiao et al., 2018]. The depth maps predicted by our method preserve much more accurate depth around object boundaries and keep finer structures, as highlighted in the boxes.

4.1 Experimental Setup

The NYU-Depth v2 dataset [Silberman *et al.*, 2012] contains 464 video sequences of indoor scenes captured with Microsoft Kinect. 654 aligned RGB-Depth pairs are provided for testing depth estimation methods for indoor scenes. All images have a resolution of 640 × 480. To training our network, we use the training dataset which contains 50K RGBD images, select and then augment in the same way as [Hu *et al.*, 2019]. Each image is downsampled to 320 × 240 using bilinear interpolation, and then center-cropped to 304 × 228. The predicted depth maps are in a resolution of 152 × 114. For testing, the predicted depth maps are upsampled to match the size of the corresponding ground truth using bilinear interpolation.

We implement the proposed model using PyTorch [Paszke *et al.*, 2017]. The encoder, SENet, is initialized by a model pretrained on ImageNet [Deng *et al.*, 2009]. The other lay-

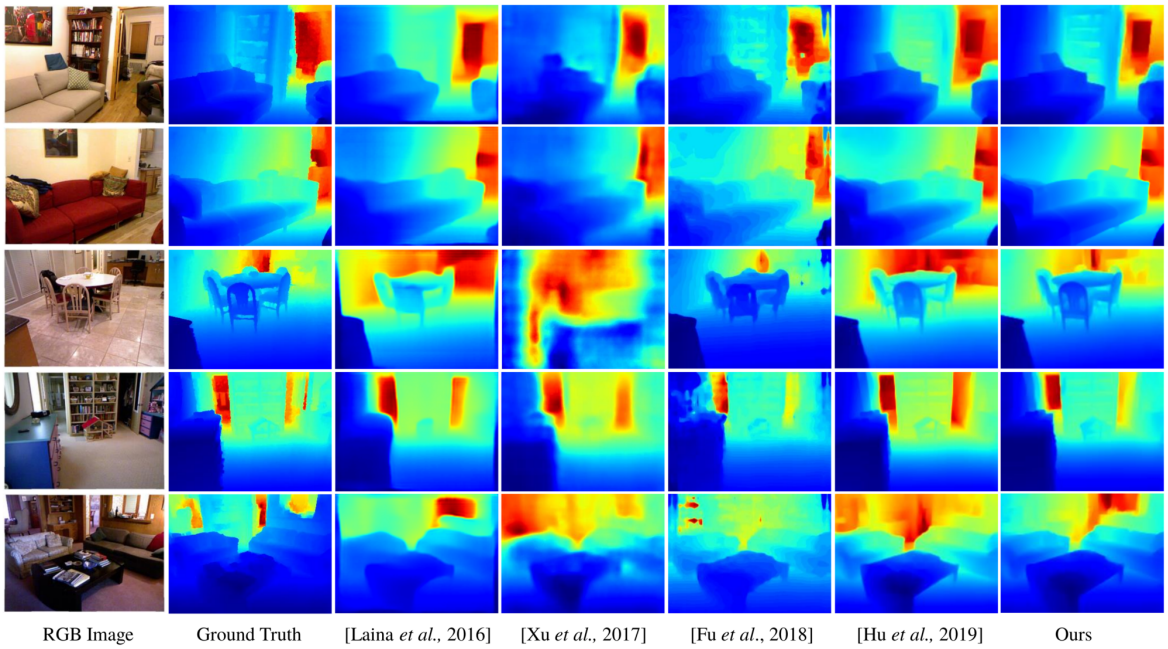

Figure 5: Qualitative results on the NYUD2 dataset.

ers in our network are randomly initialized. We use a step learning rate decay policy with Adam optimizer, and starting from an initial learning rate of $l_{init} = 10^{-4}$. It is reduced to 10% every 5 epochs. We use $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay as 10^{-4}. The proposed network was trained for 20 epochs with a batch size of 6.

4.2 Performance Comparison

Quantitative Evaluation

Following previous studies, we adopt four metrics including average relative error (REL), root mean squared error (RMS), mean $\log 10$ error (log 10), and accuracy with three thresholds, to quantitatively evaluate our depth estimation performance. Table 1 shows the results of our SARPN and recent approaches. Among the approaches of single task learning, our approach performs the best on REL, log 10 error, and accuracy with three thresholds. We are in the third position with respect to RMS. We speculate that the methods [Fu *et al.*, 2018; Jiao *et al.*, 2018] pay more attention to the absolute pixel-wise accuracy when designing their networks and loss functions, ignoring fine structures of target scenes. As a result, these methods achieve higher performance in RMS, but performs worse on the REL metric and other metrics.

We also compare our method with four approaches that employ joint task learning [Eigen and Fergus, 2015; Xu *et al.*, 2018a; Zhang *et al.*, 2018; Jiao *et al.*, 2018]. The results demonstrated that our method outperforms three methods and achieves comparative performance with [Jiao *et al.*, 2018], even they use a large number of extra labels for semantic segmentation during the training process. Moreover, the depth maps produced by [Jiao *et al.*, 2018] present very blurry object boundaries and miss geometric details. We compare the predicted depth maps in Figure 4 to demonstrate the capability of our method on restoring clear object boundaries and finer details.

Thres	Method	Prec	Recall	F1
0.25	[Laina *et al.*, 2016]	0.489	0.435	0.454
	[Xu *et al.*, 2018a]	0.516	0.400	0.436
	[Fu *et al.*, 2018]	0.320	**0.583**	0.402
	[Hu *et al.*, 2019]	0.644	0.508	0.562
	Ours	**0.645**	0.520	**0.570**
0.5	[Laina *et al.*, 2016]	0.536	0.422	0.463
	[Xu *et al.*, 2018a]	0.600	0.366	0.439
	[Fu *et al.*, 2018]	0.316	0.473	0.412
	[Hu *et al.*, 2019]	**0.668**	0.505	0.568
	Ours	0.663	**0.523**	**0.578**
1.0	[Laina *et al.*, 2016]	0.670	0.479	0.548
	[Xu *et al.*, 2018a]	**0.794**	0.407	0.525
	[Fu *et al.*, 2018]	0.483	0.512	0.485
	[Hu *et al.*, 2019]	0.759	0.540	0.623
	Ours	0.749	**0.554**	**0.630**

Table 2: Accuracy of recovered edge pixels in depth maps under different thresholds.

We also analyze the contribution of each component in our proposed network. We use a simple UNet-like architecture as our baseline, where SENet [Hu *et al.*, 2018] is employed as the backbone of our encoder. The decoder in our baseline employs a multi-stage upsampling scheme to recover a depth map. A variant (baseline+RPD) is implemented by adding the proposed RPD on the baseline model. As shown in Table 1, the performance is gradually improved by incorporating RPD and ADFF. More specifically, after adding the proposed RPD, performance among all the metrics are improved by a large margin from the baseline, while REL decreases

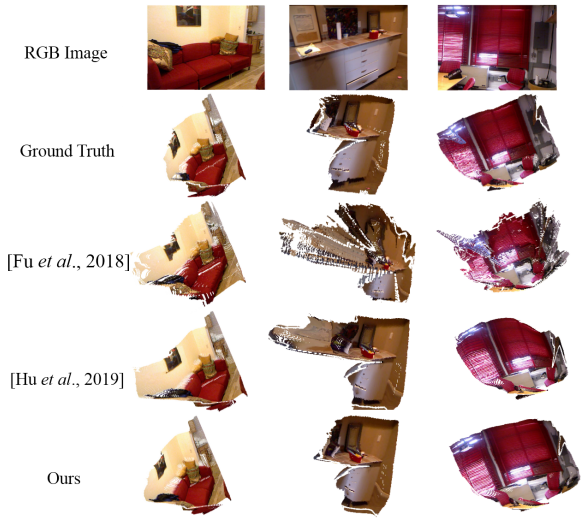

Figure 6: 3D projection from predicted depth maps. Our method better preserves the scene structure of various scales, especially the flat shape of large planar regions.

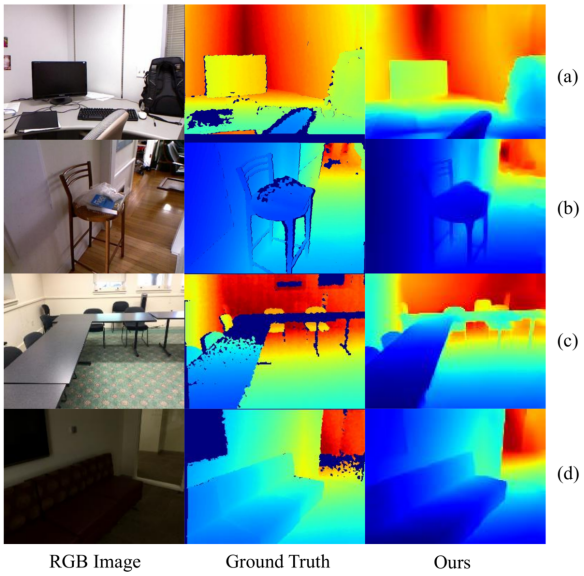

Figure 7: More results by applying our model on SUN-RGBD dataset (a)(b) and ScanNet dataset (c)(d).

by 6.5%, RMS decreases by 3.5%, log 10 error decreases by 3.8%. After adding the ADFF module, the performance is further improved, while REL decreases by 3.5%, RMS decreases by 2.7% and log 10 error decreases by 4%.

In order to prove the effectiveness of our method on preserving object details, we also compute edge accuracy to measure the quality of recovered edge details, same as [Hu *et al.*, 2019]. Precision, Recall, and F1 score are computed according to edge pixels in the ground truth map. From Table 2, we can see that our F1 score surpasses all other methods under three different thresholds. This indicates that our method restores the most structure details.

Qualitative Evaluation

We compare a series of depth maps predicted by our method and other state-of-the-art methods [Laina *et al.*, 2016; Xu *et al.*, 2017; Fu *et al.*, 2018; Hu *et al.*, 2019] in Figure 5. It can be seen that the depth maps predicted by our method are visually better than other methods. Scene structures are well preserved in different scales, especially for large planar regions and object details. For example, our method predicts accurate geometric details for the bookshelf in the first row, the chair in the third row, and the sofa in the fifth row. For large planar regions (the upper-left wall in the second row, and the floor of the third), our method also generates better results.

To better illustrate the capability of our method on preserving scene structure of large planar regions, we project the predicted depth maps as 3D point clouds and render them in novel views. As Figure 6 shows, our reprojected results are the closest to ground truth. In particular, the large wall regions recovered by our method are much more flat, while other methods suffer from severe distortions.

Model Generalization

In addition to the NYUD v2 dataset, we further explore the generalization ability of our proposed network on other datasets. We test our network, which is trained on the NYUD v2 dataset only, on ScanNet dataset [Dai *et al.*, 2017] and SUN-RGBD dataset [Song *et al.*, 2015], which contain more diverse RGBD data. As shown in Figure 7, even the data distribution of these two datasets and NYU Depth v2 is greatly different, our method could recover structures in various scales, including smooth large planar regions and object details. Moreover, our method also fills holes in the ground truth depth map automatically while maintains the scene structure.

5 Conclusion

In this paper, we propose a Structure-Aware Residual Pyramid Network for accurate monocular depth estimation. A residual pyramid decoder is introduced to predict multi-scale depth maps, which takes the underlying hierarchical scene structures into account. The residual pyramid induces our network to progressively add finer structures at a specific scale while preserving the coarser layout predicted at the upper level. Meanwhile, by using the proposed adaptive dense feature fusion module, the image features from all scales are adaptively fused when predicting the residual depth map for each scale. Experiment results demonstrate that our method achieves state-of-the-art performance in both quantitative and qualitative evaluation.

Acknowledgements

This work was supported by the National Key Research & Development Plan of China under Grant 2018YFC0307905, the National Natural Science Foundation of China (NSFC) under Grants 61632006, 61622211, and 61620106009, the Priority Research Program of Chinese Academy of Sciences under Grant XDB06040900, as well as the Fundamental Research Funds for the Central Universities under Grant WK3490000003 and WK2100100030.

References

[Dai *et al.*, 2017] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[Eigen and Fergus, 2015] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[Eigen *et al.*, 2014] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[Fu *et al.*, 2018] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[Hao *et al.*, 2018] Zhixiang Hao, Yu Li, Shaodi You, and Feng Lu. Detail preserving depth estimation from a single image using attention guided networks. In *3DV*, pages 304–313. IEEE, 2018.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[Hu *et al.*, 2019] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *IEEE Winter Conference on Applications of Computer Vision*, 2019.

[Jiao *et al.*, 2018] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *European Conference on Computer Vision*, pages 53–69, 2018.

[Ladicky *et al.*, 2014] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[Laina *et al.*, 2016] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248. IEEE, 2016.

[Li *et al.*, 2015] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[Park *et al.*, 2017] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *IEEE International Conference on Computer Vision*, pages 4980–4989, 2017.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[Qi *et al.*, 2018] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.

[Qian *et al.*, 2014] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1106–1113, 2014.

[Silberman *et al.*, 2012] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, pages 746–760, 2012.

[Song *et al.*, 2015] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

[Xu *et al.*, 2017] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5354–5362, 2017.

[Xu *et al.*, 2018a] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[Xu *et al.*, 2018b] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3925, 2018.

[Zhang *et al.*, 2018] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *European Conference on Computer Vision*, 2018.

[Zheng *et al.*, 2018] Kecheng Zheng, Zheng-Jun Zha, Yang Cao, Xuejin Chen, and Feng Wu. LA-Net: Layout-aware dense network for monocular depth estimation. In *ACM Multimedia Conference on Multimedia Conference*, pages 1381–1388, 2018.