Spindle-Net: CNNs for Monocular Depth Inference with Dilation Kernel Method

Lei He*, Miao Yu† and Guanghui Wang‡
*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
*University of Chinese Academy of Sciences, Beijing 100049, China
†School of Electric and Information Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China
‡Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA

***Abstract*—Learning depth from a single image is an important issue in computer vision. To solve this problem, encoder-decoder architect is usually employed as a powerful architecture to learn the dense corresponding function. In this work, we propose a symmetrical Spindle network of the encoder-decoder to learn the fine-grained depth. Unlike traditional convolution neural network, we first boost up the feature maps from low-dimension space to a high-dimension space, then extract the features for monocular depth learning. In order to overcome limitation of the computer memory, a single image super-resolution technique is proposed to replace the boosting process by fusing local cues in edge direction. Given the super-resolution images, the monocular depth learning needs more global information than most architectures for pixel-wise predictions. To address this issue, dilation kernel method is proposed to enlarge the receptive field in each layer. For the task of the super-resolution, the proposed method achieves better performance than the state-of-the-art methods. Extensive experiments on the monocular depth inference demonstrate that the Spindle network could achieve comparable performance on the NYU and Make3D datasets, compared with the state-of-the-art algorithms. The proposed method reveals a new perspective to learn the depth from a single image, which shows a promising generality to other pixel-wise prediction problems.**

I. Introduction

Learning depth from a single image is one of the dense prediction problems, which assign a label to each and every single pixel in the image. Most of the pixel-wise inference tasks are built upon winning architectures of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), often initializing their networks with Alex, VGG, or ResNet.

However, the architectures of the Alex, VGG and ResNet are the extension of the digit-recognition network [18], which selects the invariant abstract features for high-level problems. For pixel-wise predictions, due to the limitation of the above architectures, transfer learning methods are usually taken to transfer the feature maps from high-level problems to pixel-wise predictions, which can be categorized as indirect methods. For these indirect methods, in order to remedy the limitations of the architectures, encoder-decoder networks are exploited to reconstruct predictions, as illustrated in monocular depth estimation [17], [31], [14]. For the sake of fine-grained predictions, the techniques of fusing middle-level features are utilized by the skip connection, and multi-scale side outputs are taken as supervised information.

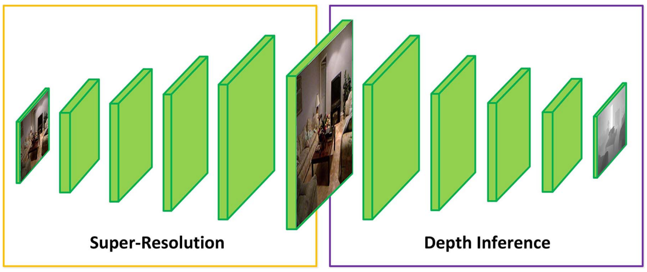

Fig. 1. Network architecture for our proposed Spindle network.

In this paper, we want to explore, from the architectures viewpoint, whether there exists a direct method to predict fine-grained depth without encoder-decoder. Inspired by the intuition, we propose a novel symmetrical structure of the encoder-decoder, Spindle-Net, to learn depth from a single image, as shown in Figure 1. The Spindle-Net consists of two modules, one is to embed the high-dimension feature maps, and the other is to extract feature maps for depth learning from a single image. Due to the limitation of computer memory, we replace the embedding part with single image super-resolution technique as a trade-off. Here, we take Laplacian pyramid super-resolution method to predict the image at $4\times$ resolution with two pyramid levels. In order to obtain a complete structure in high resolution space, we propose a direction sensitive algorithm (DsSRN) to fuse long-range local cues in edge direction. Through experiments on the datasets of monocular depth inference, the DsSRN achieves a better performance than the LapSRN [16] on quantitative and qualitative evaluations.

With the super resolution image from the embedding SR network, our goal in next stage is to extract more global information than the most recent pixel-wise models. Currently, there exists two approaches to retrieve the global information, one is to increase the depth of the neural network, the other is to enlarge the receptive field in each convolution layer by dilation convolution [2], [3]. In this study, we build depth inference network on the ResNet [8] by removing the last three fully connected layers. Chen *et al.* [3] apply atrous convolution with a rate to replace consecutive striding. To address the degenerate problem of the atrous convolution, we propose a

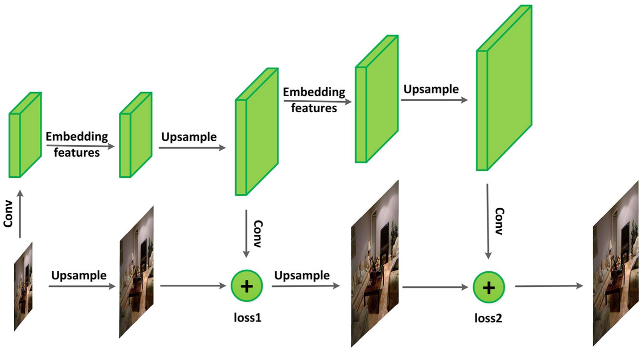

Fig. 2. The Laplacian pyramid framework for Super-Resolution.

novel approach, dilation kernel, to capture global information more robustly than the atrous convolution. In the end, our proposed methods achieve comparable performance, compared with the state-of-the-art approaches.

In summary, the contributions of the paper are three-fold. First, we propose a novel Spindle-Net to learn depth from a single image directly, revealing a new perspective to address the pixel-wise predictions. Second, a novel direction sensitive method is proposed to perceive long range cues along edge direction, which achieves competitive performance compared with the state-of-the-art methods in super-resolution. Finally, we propose enlarging receptive field approaches to learn depth from a single image and obtains comparable results. The proposed network together with all trained parameters will be available online.

II. Related Work

Learning depth from a single image has been extensively studied in the literature, from classic machine learning approaches, to deep convolutional neural networks.

To tackle this task, classic methods [10], [26], [27], [28], [29] usually make strong geometric assumptions that the scene structure consists of horizontal planes, vertical walls and superpixels, employing the Markov random field (MRF) to inference the depth by leveraging the handcrafted features. Non-parameter algorithms [11], [13] are another type of classical methods for learning the depth from a single image, relying on the assumption that the similarities between regions in the RGB images imply similar depth cues as well. After clustering the training dataset based on the global features (e.g. GIST [23], HOG), these methods first perform matching to search for the candidate RGB-D of the input RGB image in the feature space, then, the candidate pairs are warped and fused to obtain the final depth.

There exist two types of CNN-based approaches for the task of depth estimation in the related references: supervised learning approaches and unsupervised learning methods. One of the first supervised learning methods, proposed by Eigen *et al.* [5], addressed this issue by fusing the depths from the global network and refined network. Their work later was extended to use a multi-scale convolutional network to fully integrate the global and local information in a deeper neural network [4]. Other methods to obtain the fine-grained depth leveraged the representation of the neural network and the inference of the CRFs. Liu *et al.* [20] presented a deep convolutional neural field model based on fully convolutional networks and a novel superpixel pooling method, combining the strength of deep CNN and the continuous CRF into a unified CNNs framework. Laina *et al.* [17] built a neural network on ResNet, followed by designed up-sampling blocks to obtain high resolution depth. However, the middle-level features are not fused into the network to obtain detailed information of the depth.

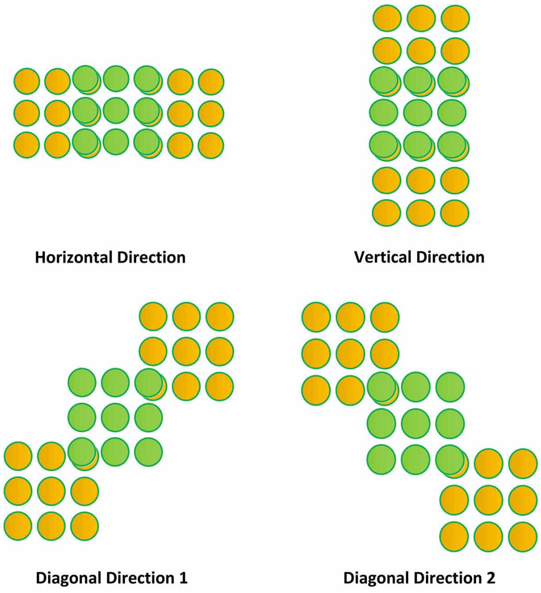

Fig. 3. Direction sensitive network for embedding long range features.

The unsupervised learning methods for depth estimation from a single image achieved significant progress, where the inferred monocular depth is taken as a intermediate result for computing the reconstruction loss between two images with large portion of overlap. By exploiting the epipolar geometry constraints, Garg *et al.* [6] first inferred the monocular depth through photometric consistency on the stereo images. Godard *et al.* [7] further enforced the consistency between the disparity produced relative to both the left and right images. Furthermore, Zhou *et al.* [31] simultaneously learned the monocular depth and camera pose, which extends the experimental images from stereo pairs to video sequences. These deep learning methods are mostly built on the encoder-decoder based neural network.

Recently, in order to remove the ambiguity between the scene depth and the focal length, He *et al.* [9] proposed a novel deep neural network to infer the fine-grained monocular depth from both the fixed- and varying-focal-length datasets. The extensive experiments demonstrate that the embedding focal length is able to improve the depth learning accuracy from single images.

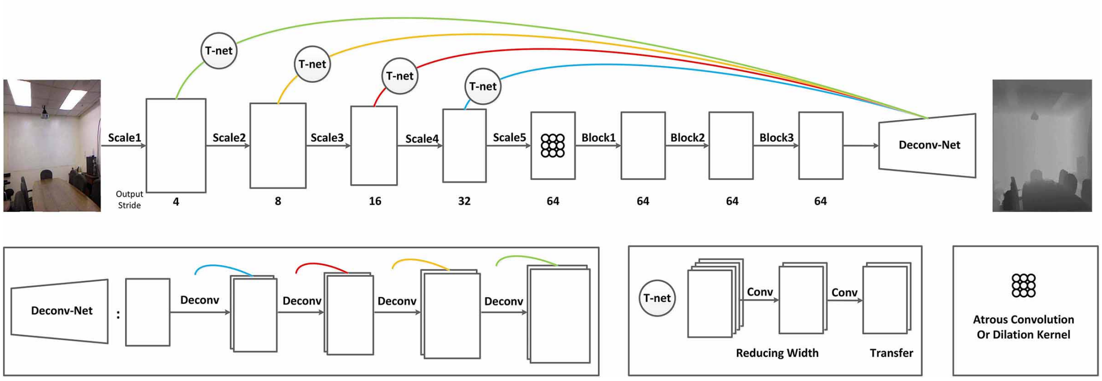

Fig. 4. The framework for depth inference.

III. Spindle-Net

The proposed Spindle-Net is conceptually natural and intuitive. As shown in Figure 1, the structure consists of two modules, one is used to boost up feature maps from low-dimension space to a high-dimension space; and the other is used to extract features for the task of learning depth from a single image. In the following, we will introduce the key principle and elements of the Spindle-Net in details.

A. *Embedding SR*

The spatial resolution of the feature maps is important for inferring fine-grained depth, including other pixel-wise predictions. First, we extract high-dimension spatial features, rather than decreasing the resolution by striding convolution and pooling operations. By considering the limitations of the memory and load in computer, the width of the feature map should not be too large. For the task of the monocular depth inference, we set the channel number of the feature map to 3, and exploiting the super-resolution technique to simplify the process for embedding the high-dimension spatial features.

We utilize a deep neural network to learn single image super-resolution based on the Laplacian pyramid framework, like the LapSRN [16], taking a low resolution (LR) image as input and progressively predicts residual images through on each pyramid level, as shown in Figure 2. In order to enlarge the receptive field of the high frequency features, Lai *et al.* [16] employed many recursive blocks in every level. Although parameters sharing across blocks and local skip connection is explored, the recursive convolutions in each block greatly increases the structural and training complexities. Due to the same parameters in each convolution, the receptive field of recursive convolutions is computed as following.

$$RF = (K_{size} - 1)(N - 1) + K_{size} \tag{1}$$

where K_{size} is the size of the convolution kernel, N is the number of the convolution, and RF is the receptive field in the N-th layer.

In order to further enlarge the receptive field and preserve structure, we propose a direction sensitive network (DsSRN) to replace the embedding features unit, which consists of four parallel blocks, as shown in Figure 3. Each block implements three convolutions with overlap simultaneously, which is distributed horizontally, vertically, and diagonally, respectively. Such a parallel convolution network creates a receptive field as

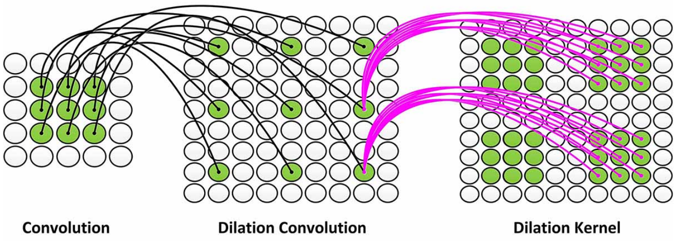

Fig. 5. The Neural Network Architectures of Convolution, Dilation Convolution and Dilation Kernel.

$$RF = (3K_{size} - 2\gamma - 1)(N - 1) + 3K_{size} - 2\gamma \tag{2}$$

where γ is the overlap rate between two neighbor kernels in each block.

In the following experiments, we set the overlap rate to 2. Obviously, the kernel direction network is able to extract more global residual image than the above network. Together with the neck structure to reduce the channel number of the feature maps, the proposed network has approximately same number of parameters as the feature embedding part in the LapSRN.

B. *Depth Inference*

For the second part of the Spindle-Net, we first design modules based on going deeper network [3]. As shown in Figure 4, we duplicate several copies of the last ResNet block, and arrange them in cascade. We utilize ResNet [8] to learn global information, then progressively extracts fine-grained depth by de-convolution techniques. During each de-convolution phase, the side outputs of the ResNet are combined with the corresponding feature maps. By this way, the Spindle-Net is able to predict monocular fine-grained depth under the guidance of the global information. To incorporate long range information, we propose atrous convolution and dilation kernel method only acting on the last layer of scale 5.

Then, we illustrate dilation kernel method in details. Consider two-dimensional signals, for each location $\boldsymbol{i}$ on the output $\boldsymbol{y}$ and a filter $\boldsymbol{w}$, dilation convolution [2], [3] is applied over the input feature map x as

$$\boldsymbol{y}[\boldsymbol{i}] = \sum_{\boldsymbol{k}} \boldsymbol{x}[\boldsymbol{i} + r \cdot \boldsymbol{k}]\boldsymbol{w}[\boldsymbol{k}] \tag{3}$$

where the atrous rate r corresponds to the stride with which we sample the input signal, which is equivalent to convolving the input $\boldsymbol{x}$ with upsampled filters produced by inserting $r-1$ zeros between two consecutive filter values along each spatial dimension. Standard convolution is a special case for the rate $r = 1$.

By the same formulation, the output $\boldsymbol{y}$ of the dilation kernel method is as following:

$$\boldsymbol{y}[\boldsymbol{i}] = \sum_{\boldsymbol{k}} \sum_{\boldsymbol{j}} \boldsymbol{x}[\boldsymbol{i} + r \cdot \boldsymbol{k} + \boldsymbol{j}]\boldsymbol{w}[\boldsymbol{k} + \boldsymbol{j}] \tag{4}$$

where $\boldsymbol{j}$ is the element-wise location of the sub-convolution in the dilation kernel.

In principle, dilation kernel is an intuitive extension of the dilation convolution, by replacing each element of dilation convolution with convolution kernel, which is critical for robust extraction of the global information, as shown in Figure 5. Obviously, the dilation convolution is a special case of the dilation kernel when the elements of sub-convolution is reduced to one.

C. Loss Function

Super-Resolution. Let $\boldsymbol{x_{LR}}$ be the input LR image and θ be the set of network parameters to be optimized. Our goal is to learn a mapping function f to generate an HR image $\boldsymbol{x_{HR}} = f(\boldsymbol{x_{LR}}, \theta)$ that approaches the ground truth HR image $\boldsymbol{x_{GT}}$. By utilizing the bicubic downsampling method, $\boldsymbol{x_{LR}}$ is resized from the ground truth $\boldsymbol{x_{GT}}$ at each level. Like in Lai *et al.* [15], [16], we use the Charbonnier penalty function [1] as loss function.

Depth-Inference. We formulate the task of depth prediction from monocular RGB input as the problem of learning a non-linear mapping $D = f(I, \theta)$ from the image I to the output depth D. The parameters θ of the proposed network are learned through minimizing the loss function defined on the prediction and the ground truth. Following Laina *et al.* [17], we take the following BerHu loss as the error function by integrating the advantages of both the L2 norm and L1 norm, resulting in accelerated optimization and detailed structure.

$$B(y - \overline{y}) = \begin{cases} |y - \overline{y}| & |y - \overline{y}| < c \\ \frac{(y-\overline{y})^2 + c^2}{2c} & |y - \overline{y}| > c \end{cases} \tag{5}$$

where $c = 0.05max_i(|y_i - \overline{y}_i|)$, and i indexes the pixels in the current batch.

IV. Experiments

To demonstrate the effectiveness and evaluate the performance of the proposed Spindle-Net, we carry out comprehensive experiments on two publically available datasets: NYU v2 [22], Make3D [27]. In the following subsections, we report the details of our implementation and the evaluation results.

A. Experimental Setup

Datasets. The **NYU Depth v2** [22] consists of 464 scenes, captured using Microsoft Kinect. Followed by the official split, the training dataset is composed of 249 scenes with the 795 pair-wise images, and the testing dataset includes 215 scenes with 654 pair-wise images. In addition, the raw dataset contains 407,024 new unlabeled frames. For data augmentation, we sample equally-spaced frames out of each raw training sequence, and further align the RGB-D pairs by virtue of the provided toolbox, resulting in approximately 4k RGB-D images. Then, the sampled raw images and 795 pair-wise images are online augmented by Eigen *et al.* [5]. The input images and the corresponding depths are simultaneously transformed using small scaling, color transformations and flips with a chance of 0.5. Due to the hardware limitation, we down-sample the original frames from the size 640×480 pixels to 120×90 as the input to the network, 240×180 and 480×360 as supervisory information for super-resolution.

The **Make3D** dataset [27] contains 400 training images and 134 testing images of outdoor scenes, generated from a custom 3D laser scanner. While the depth map resolution of the ground truth is only 305×55, not matching the corresponding original RGB images with 1704×2272 pixels, we resize all RGB-D images to 345×460 by preserving the aspect ratio of the original images. Due to the neural network architecture and hardware limitations, we subsample the resolution of the RGB-D images to 80×112 as the input to the network, 160×224 and 320×448 as supervisory information for super-resolution.

Evaluation Metrics. For quantitative evaluation, we report errors obtained with the following extensively adopted error metrics.

- Average relative error: $\mathbf{rel} = \frac{1}{N}\sum_{y_i \in |N|} \frac{|y_i - y_i^*|}{y_i^*}$
- Root mean squared error: $\mathbf{rms} = \sqrt{\frac{1}{N}\sum_{y_i \in |N|} |y_i - y_i^*|^2}$
- Average log_{10} error: $\mathbf{log}_{10} = \frac{1}{N}\sum_{y_i \in |N|} |log_{10}(y_i) - log_{10}(y_i^*)|$
- Accuracy with threshold t: percentage (%) of y_i subject to $max(\frac{y_i^*}{y_i}, \frac{y_i}{y_i^*}) = \delta < t(t \in [1.25, 1.25^2, 1.25^3])$

where y_i is the estimated depth, y_i^* denotes the corresponding ground truth, and N is the total number of valid pixels in all images of the validation set.

Implementation Details. We use TensorFlow deep learning framework to implement the proposed network, and train the network on a single NVIDIA GeForce GTX TITAN with 12GB memory. The objective function is optimized using the Adam method [12]. During the initialization stage, weight layers of the RseNet are initialized using the corresponding model pre-trained on the ILSVRC [25] dataset for image

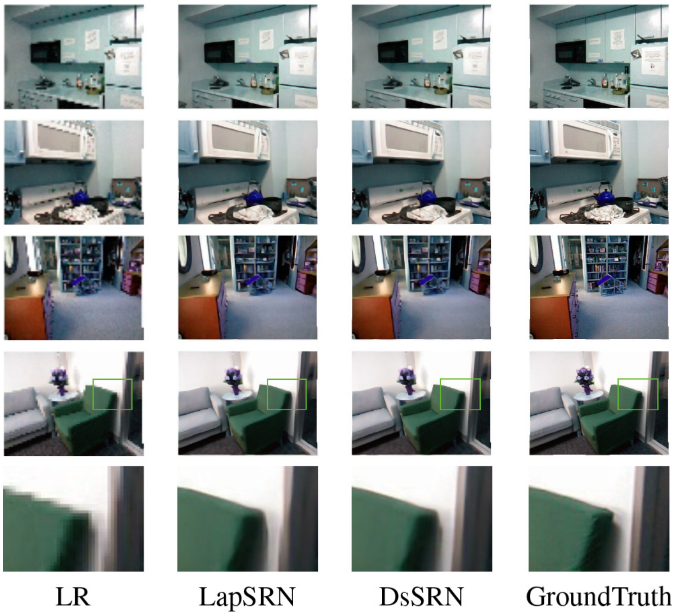

Fig. 6. Visual comparison for 4× SR on the NYU dataset. Each image in last row is a local zoom, corresponding to the green area in the fourth row.

classification. The weights of other added network are assigned by sampling a Gaussian with zero mean and 0.01 variance, and the learning rate is set at 0.0001. Finally, our model is trained with a batch size of 8 for about 40 epochs. The first 20 epochs are trained for the super-resolution, and the rest 20 epochs are used to fine tune the depth inference network with fixing the super-resolution parameters.

B. SR evaluation

First, we compare the proposed DsSRN method with the state-of-the-art LapSRN [15] on the NYU dataset [22] and Make3D dataset [27]. We evaluate the SR results with three widely used image quality metrics: PSNR, SSIM, IFC, and compare performance on $4\times$ SR.

TABLE I
QUANTITATIVE EVALUATION OF THE SR ALGORITHM.

Methods	NYU			Make3D		
	PSNR	SSIM	IFC	PSNR	SSIM	IFC
LapSRN [15]	22.22	0.601	1.962	23.33	**0.573**	1.392
DsSRN	**22.65**	**0.615**	**2.078**	**23.38**	0.571	**1.471**

The quantitative results of the NYU dataste are reported in Table I. Our DsSRN achieves better performance than the LapSRN method, which demonstrates our proposed method is not only able to learn more robust and similar structural super-resolution image, but also correlated well with human perception of image super-resolution. For the Make3D dataset, our proposed approach performs better than the LapSRN method in terms of the PSNR and IFC metrics, but a slightly weak on the SSIM metric, as shown in Table I. This is because there exist mess trees with clutter leaves in the Make3D dataset, which lacks of structural information on the feature representations, compared to the indoor images.

TABLE II
DEPTH RECONSTRUCTION ERRORS ON THE NYU DEPTH DATASET.

Method	Error			Accuracy		
	rel	**rms**	**$\log_{10}$**	1.25	1.25^2	1.25^3
Karsch *et al.* [11]	0.374	1.12	0.134	0.447	0.745	0.897
Liu *et al.* [21]	0.335	1.06	0.127	-	-	-
Li *et al.* [19]	0.232	0.821	0.094	-	-	-
Liu *et al.* [20]	0.230	0.824	0.095	0.614	0.883	0.975
Wang *et al.* [30]	0.220	0.745	0.094	0.605	0.890	0.970
Eigen *et al.*[5]	0.215	0.907	-	0.611	0.887	0.971
R. and T. [24]	0.187	0.744	0.078	-	-	-
E. and F. [4]	0.158	0.641	-	0.769	0.950	0.988
Lai [17]	0.129	0.583	0.056	0.801	0.950	0.986
Ours-AC	0.295	0.600	0.112	0.488	0.877	0.967
Ours-DK	0.260	0.587	0.105	0.547	0.885	0.967

In order to show the realistic of the super-resolution image, we make visual comparisons on the NYU dataset for $4\times$ SR in Figure 6. Our method (DsSRN) accurately reconstructs high-quality HR images, approaching to ground truth in visualization. In addition, the proposed method can reconstruct parallel straight lines more accurate than the LapSRN, benefiting from large local perception in edge direction.

C. Depth analysis

We then experiment with monocular depth inference module with atrous convolution and dilation kernel method respectively. In the first series of experiments we focus on the NYU v2 dataset [22]. The results of our comparisons are reported in Table II. It is evident that the model with the dilation kernel achieves better performance than the one with the atrous convolution.

The proposed method is compared with the state-of-the-art approaches. Here, the results of other algorithms are from the reports of the original papers. The comparative results of the proposed approaches and baselines are also reported in Table II. The performance of our proposed method is comparable with most state-of-the-art approaches. However, the proposed method is slightly weak than the approach [17] with the best performance. The main reason is that the input image of the proposed method is generated from the low resolution (LR) images, other than the original high-resolution images used by the state-of-the-art methods. The LR image may lack detailed structure, and as a result, its super-resolution image is not able to reconstruct more accurate depth than the HR image. If all approaches are inputted with the same LR images, the proposed approach will definitely outperform others due to the super-resolution module. We will study which resolution is more suitable for the input in the stage of the super-resolution. Furthermore, the experiments have shown weakness about the strategy by embedding high-resolution features with the single image super-resolution, we will investigate other methods to boost up feature maps in the future, such as exploiting semantic information.

TABLE III
DEPTH RECONSTRUCTION ERRORS ON THE MAKE3D DEPTH DATASET.

Method	Error (lower is better)		
	rel	**rms**	**$\log_{10}$**
Karsch *et al.* [11]	0.355	9.20	0.127
Liu *et al.* [21]	0.335	9.49	0.137
Liu *et al.* [20]	0.314	8.60	0.119
Li *et al.* [19]	0.278	7.19	0.092
Roy and Todorovic [24]	0.260	12.40	0.119
Lai [17]	0.176	4.46	0.072
Ours-AC	0.226	3.40	0.074
Ours-DK	0.217	6.79	0.080

In addition, the proposed model is also evaluated on the Make3D dataset [27]. Following [4], [17], the error metrics are computed on the regions with ground truth depth maps less than 70m. The experimental comparisons are reported in Table III. Our proposed method achieves outstanding performance on root mean squared (rms) metric, while shows comparable performance on other two metrics. Please note that proposed approach is working with low resolution images, while others are using full resolution. In this dataset, the atrous convolution method performs better than the dilation kernel approach. We think that the size of sub-convolution in dilation kernel and its location in the network should be taken into consideration. In the future, in order to improve the performance of the dilation kernel, the size, shape and location of the proposed dilation kernel should be explored in the hierarchical network architecture for pixel-wise predictions.

V. CONCLUSION

In this paper, we have proposed a deep convolutional network, Spindle-Net, to learn depth from a single image. The idea provides a new perspective to address the pixel-wise predictions. We have presented a comprehensive evaluation on various design choices. During the super-resolution phase, in order to perceive long range cues in edge direction, we propose a novel direction sensitive method, which achieves outstanding performance. Furthermore, we extended the atrous convolution and proposed a novel dilation kernel approach to robustly capture the global information. We have shown the promising results of the proposed Spindle-Net in the context of learning depth from single images. The network design is novel and has great potential for other pixel-wise predictions. The proposed network with trained parameters will be available on the author's website.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China under the grant No (61333015, 61421004, 61772444, 61573351), and by the Open Projects Program of National Laboratory of Pattern Recognition under the grant No 201700015.

REFERENCES

[1] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *IJCV*, 61(3):211–231, 2005.

[2] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, PP(99):1–1, 2016.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *CVPR*, pages 2650–2658, 2015.

[5] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, pages 2366–2374, 2014.

[6] R. Garg, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756. Springer, 2016.

[7] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. July 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] L. He, G. Wang, and Z. Hu. Learning depth from single images with deep neural network embedding focal length. *arXiv preprint arXiv:1803.10039*, 2018.

[10] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM transactions on graphics (TOG)*, 24(3):577–584, 2005.

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

[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] J. Konrad, M. Wang, and P. Ishwar. 2d-to-3d image conversion by learning depth from examples. In *CVPR*, pages 16–22. IEEE, 2012.

[14] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. *arXiv preprint arXiv:1702.02706*, 2017.

[15] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. July 2017.

[16] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *arXiv preprint arXiv:1710.01992*, 2017.

[17] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248. IEEE, 2016.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[19] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, pages 1119–1127, 2015.

[20] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015.

[21] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *CVPR*, pages 716–723, 2014.

[22] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[23] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[24] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, pages 5506–5514, 2016.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[26] A. Saxena, S. H. Chung, and A. Y. Ng. 3-d depth reconstruction from a single still image. *IJCV*, 76(1):53–69, 2008.

[27] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 31(5):824–840, 2009.

[28] G. Wang, H. T. Tsui, Z. Hu, and F. Wu. Camera calibration and 3d reconstruction from a single view based on scene constraints. *Image and Vision Computing*, 23(3):311–323, 2005.

[29] G. Wang, H. T. Tsui, and Q. M. J. Wu. What can we learn about the scene structure from three orthogonal vanishing points in images. *Pattern Recognition Letters*, 30(3):192–202, 2009.

[30] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, pages 2800–2809, 2015.

[31] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. July 2017.