Pattern Recognition 83 (2018) 328–339

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference

Bo Li, Yuchao Dai*, Mingyi He*

Shaanxi Key Lab of Information Acquisition and Processing, School of Electronics and Information, Northwestern Polytechnical University, Xi'an, 710129, China

ARTICLE INFO

Article history:
Received 2 October 2017
Revised 12 May 2018
Accepted 31 May 2018
Available online 6 June 2018

Keywords:
Monocular depth estimation
Deep convolutional neural network
Soft-weighted-sum-inference
Dilated convolution

ABSTRACT

Monocular depth estimation is very challenging in complex compositions depicting multiple objects of diverse scales. Albeit the recent great progress thanks to the deep convolutional neural networks, the state-of-the-art monocular depth estimation methods still fall short to handle such real-world challenging scenarios. In this paper, we propose a deep end-to-end learning framework to tackle these challenges, which learns the direct mapping from a color image to the corresponding depth map. First, we represent monocular depth estimation as a multi-category dense labeling task by contrast to the regression-based formulation. In this way, we could build upon the recent progress in dense labeling such as semantic segmentation. Second, we fuse different side-outputs from our front-end dilated convolutional neural network in a hierarchical way to exploit the multi-scale depth cues for monocular depth estimation, which is critical in achieving scale-aware depth estimation. Third, we propose to utilize soft-weighted-sum inference instead of the hard-max inference, transforming the discretized depth scores to continuous depth values. Thus, we reduce the influence of quantization error and improve the robustness of our method. Extensive experiments have been conducted on the Make3D, NYU v2, and KITTI datasets and superior performance have been achieved on NYU v2 and KITTI datasets compared with current state-of-the-art methods, which shows the superiority of our method. Furthermore, experiments on the NYU v2 dataset reveal that our classification based model is able to learn the probability distribution of depth.

1. Introduction

Depth estimation aims at predicting pixel-wise depth for a single or multiple images, which is an essential intermediate component toward 3D scene reconstruction and understanding. It has been shown that depth information can benefit tasks such as recognition [1,2], human-computer interaction [3], and 3D model reconstruction [4]. Traditional techniques have predominantly worked with multiple images to make the problem of depth prediction well-posed, which include *N*-view reconstruction, structure from motion (SfM) and simultaneous localization and mapping (SLAM).

However depth estimation from a monocular single viewpoint still lags far behind its multi-view counterpart. This is mainly due to the fact that the problem is ill-posed and inherently ambiguous: a single image on its own does not provide any depth cue explicitly (*i.e.*, given a color image of a scene, there are infinite number of 3D scene structures explaining the 2D measurements exactly). When specific scene dependent knowledge is available, depth estimation or 3D reconstruction from single images can be achieved by utilizing geometric assumptions such as the "Blocks World" model [5], the "Origami World" model [6], repetition of structures [7], or other special cues like shading [8]-defocus [9]. However, these cues typically work for images with specific structures or conditions, and may not be applied to general scenarios.

Recently, learning based monocular depth estimation methods that predicting scene geometry directly by learning from data, have gained popularity. Typically, such approaches recast the underlying depth estimation problem in a pixel-level scene labeling pipeline by exploiting the relationship between monocular image and depth. Fully convolutional neural network (CNN) has been proved to be an effective method to solve these kinds of problems. There has been considerable progress in applying deep convolutional neural network (CNN) to this problem and excellent performances have been achieved [11–19]. Deep convolutional neural network based methods have occupied the leading boards of monocular depth estimation benchmarking datasets such as the KITTI dataset [10] and the NYU v2 dataset [4].

Albeit the above success, the state-of-the-art monocular depth estimation methods still fall short to handle real-world challenging complex decompositions depicting multiple objects of diverse

* Corresponding authors: Mingyi He and Yuchao Dai.
E-mail address: myhe@nwpu.edu.cn (M. He).

https://doi.org/10.1016/j.patcog.2018.05.029
0031-3203/

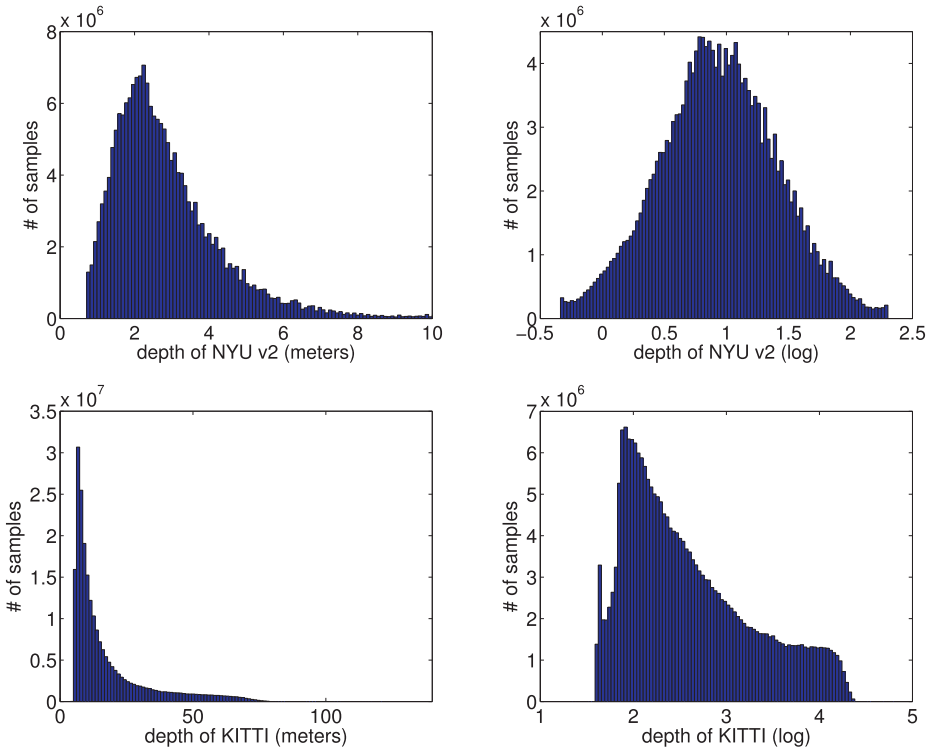

Fig. 1. Distribution of the depth values of the NYU v2 dataset [4] and the KITTI dataset [10] in the form of a histogram. We illustrate the depth distribution both in the original depth and in the log space.

scales due to the following difficulties: 1) The serious data imbalance problem. Due to the perspective effect which introduces a larger number of samples with small depth value than samples with large depth value. (In Fig. 1, we illustrate the distribution of the depth values in the form of histogram for the NYU v2 dataset and the KITTI dataset correspondingly. The imbalanced distribution of the depth can be well observed.); 2) There are more rapid changes in depth value compared with other dense predictions tasks such as semantic labeling and 3) As a small local patch could not provide sufficient depth cue, long-range context information is needed to handle the scale ambiguity in monocular depth estimation. Even though there have been various post-processing methods to refine the estimated depth map from the deep network [11–18,20], the bottleneck in improving monocular depth estimation is still the specially designed CNN architecture, which is highly desired. In this paper, we would like to advocate that by properly representing the problem and designing a new network model, our deep end-to-end learning based monocular depth estimation method outperforms current state-of-the-art methods (with or without post-processing).

Specifically, in this paper, we present a deep CNN based framework to tackle the above challenges, which learns the direct mapping from the color image to the corresponding depth map in an end-to-end manner. We recast monocular depth estimation as a multi-category dense labeling by contrast to the widely used regression formulation. Our network is based on the deep residual network [21], where dilated convolution and hierarchical fusion layers are designed to expand the receptive field and to fuse multi-scale depth cues. In order to reduce the influence of quantization error and to improve the robustness of our method, we propose to use a soft-weighted-sum inference. Extensive experimental results show that even though we train our network as a standard classification task with the multinomial logistic loss, our network is able to learn the probability distribution among different categories. The overall flowchart of our framework is illustrated in Fig. 2[1].

Our main contributions can be summarized as :

- We propose a deep end-to-end learning framework to monocular depth estimation by recasting it as a multi-category classification task, where both dilated convolution and hierarchical feature fusion are used to learn the scale-aware depth cues.
- Our network is able to output the probability distribution among different depth labels. We propose a soft-weighted-sum inference, which could reduce the influence of quantization error and improve the robustness.
- Our method achieves the state-of-the-art performance on both indoor and outdoor benchmarking datasets, Make3D, NYU v2 and KITTI dataset.

2. Related work

In this section, we briefly review related works for monocular depth estimation, which can be roughly categorized as conventional non deep learning based methods and deep learning based methods.

Non deep learning based methods: Seminal work by Saxena et al. [22] tackles the problem with a multi-scale Markov Random Field (MRF) model, with the parameters of the model learned through supervised learning. Liu et al. [23] estimated the depth map from predicted semantic labels, achieving improved performance with a simpler MRF model. Ladicky et al. [24] showed that perspective geometry can be used to improve results and demonstrated how scene labeling and depth estimation can benefit each

[1] As a deep learning based method, the success of our method depends on the following 2 conditions/assumptions: 1) the consistency (correlation) between the training dataset and the testing dataset; and 2) the camera settings (focal length, installation angles, resolution and etc.) of the training datasets and the testing datasets are consistent or very similar.

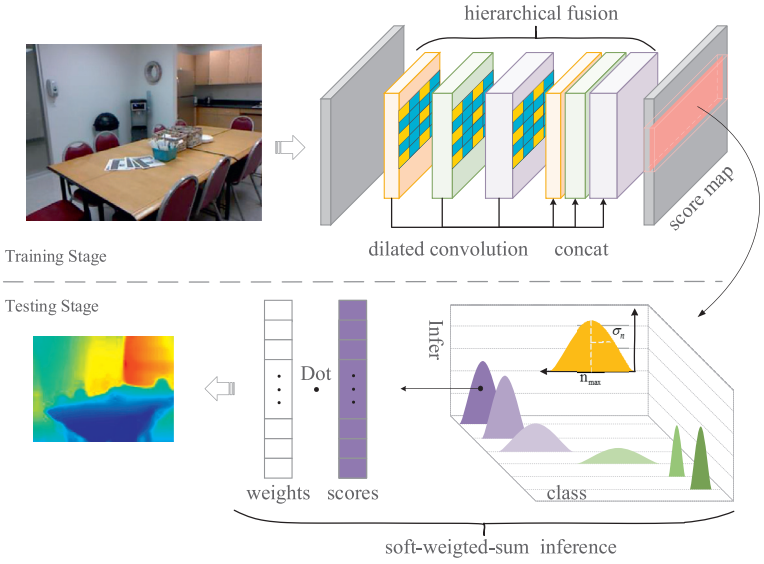

Fig. 2. Flowchart of our monocular depth estimation framework, which is built upon deep Residual network [21] and consists of dilated convolution and hierarchical feature fusion. Soft-weighted-sum inference is used to predict continuous depth values from the discrete depth labels. We also illustrate typical probability distribution of labels from the network, which shows that our classification based framework is able to learn the similarity between labels. Differences with previous works (hierarchical fusion, dilated convolution, and soft-weighted-sum inference) are highlighted.

other under a unified framework, where a pixel-wise classifier was proposed to jointly predict a semantic class and a depth label from a single image. Besides these parametric methods, other works such as [25,26] recast monocular depth estimation in a non-parametric fashion, where the whole depth map is inferred from candidate depth maps. Karsch et al. [25] proposed a non-parameter method, which is built on a pixel transfer framework. Liu et al. [26] proposed a discrete-continuous CRFs, which aims to avoid the over-smoothing and maintain occlusion boundaries. Anirban et al. [27] proposed a Neural Regression Forest model for this problem. Ji et al. [28] estimated the depth and semantic jointly with elastic CRF. These works provide important insights and cues for single image depth estimation problem, while most of them utilized hand-crafted features thus limited their performance especially for complex scenarios.

Deep learning based methods: Recently, monocular depth estimation has been greatly advanced thanks to deep convolutional neural network (CNN). Eigen et al. [29] presented a framework by training a large hierarchical deep CNN. However, partly due to the fully connected layers used in the network model, their network has to be trained with very large scale data. By contrast, Li et al. [11] proposed a patch-based CNN framework and a hierarchical-CRF model to post-process the raw estimated depth map, which significantly reduces the number of training images needed. Liu et al. [12] proposed a CRF-CNN joint training architecture, which could learn the parameters of the CRF and CNN jointly. Wang et al. [13] proposed a CNN architecture for joint semantic labeling and monocular depth prediction. Chen et al. [30] proposed an algorithm to estimate metric depth using annotations of relative depth.

Very recently, Laina et al. [16] proposed to use the Huber loss instead of the L_2 loss to deal with the long tail effect of the depth distribution. Cao et al. [15] demonstrated that formulating depth estimation as a classification task could achieve better results than regression with L_2 loss, while insufficient analysis is given for the success. In addition, different with our method, they used hard-max inference in the testing phase. Xu et al. [17] proposed a Multi-Scale Continuous CRFs to better extract the hierarchical information and improve the smoothness of the final results. Our hierarchical information fusion strategy is much simpler than [17], while we also achieve comparable results.

Besides the above methods using ground truth depth maps to supervise the network learning, there is another group of methods that use novel view synthesis to supervise the network learning by exploiting the availability of stereo images and image sequences [18,31–33] . Garg et al. [31] proposed to train a network for monocular depth estimation using an image reconstruction loss, where a Taylor approximation is performed to linearize the loss. Godard et al. [18] replaced the use of explicit depth data during training with easier-to-obtain binocular stereo footage, which enforces consistency between the disparities produced relative to both the left and right images, leading to improved performance and robustness compared to existing approaches. Along with this pipeline, Zhou et al. [32] presented an unsupervised learning framework for the task of monocular depth and camera motion estimation from unstructured video sequences based on image warping to evaluate the image error. Kuznietsov et al. [33] learned depth in a semi-supervised way, where sparse ground-truth depth and photoconsistency are jointly used. Ummenhofer et al. [34] trained a convolutional network end-to-end to compute depth and camera motion from successive, unconstrained image pairs, where the architecture is composed of multiple stacked encoder-decoder networks.

The key supervision signal for these “unsupervised” methods comes from the task of novel view synthesis: given one input view of a scene, synthesize a new image of the scene seen from a different camera pose. Essentially, pairs of rectified stereo images or consecutive image frames have already encoded the depth information implicitly.

Our work is also related to the works on FCN (fully convolutional network) based dense labeling. Long et al. [35] proposed the fully convolutional neural network for semantic segmentation, which has been widely used in other dense labeling problems. Hariharan et al. [36] presented that low-level CNN feature is better with respect to the boundary preservation and object location. Recently, Yu et al. [37] demonstrated that dilated convolution could expand the receptive field of the corresponding neuron while keeping the resolution of the feature map. Chen et al. [38] successfully utilized the dilated convolution on the semantic problem and show how to build them on the pre-trained CNN.

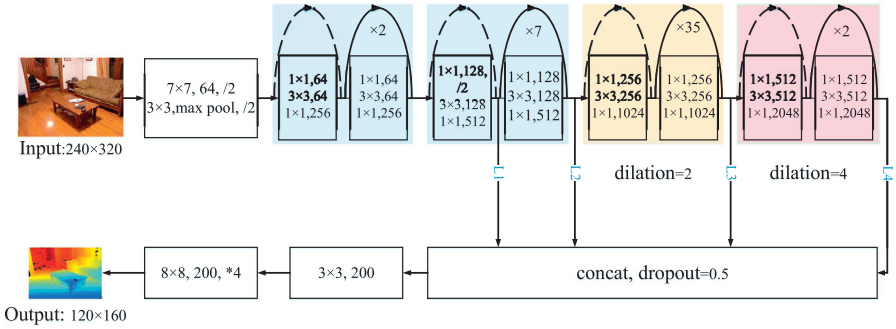

Fig. 3. Illustration of our network architecture. The detail of the basic residual block could be referred to [21]. $\times n$ means the block repeats n times. We present all the hyper-parameters of convolution and pooling layers. All the convolution layers are followed by batch normalization layer except for the last one. /2 means that the layer's stride is 2. *4 means the deconv layer's stride is 4. Dilation shows the dilated ratio of the corresponding parts. $L1, \ldots L4$ are our skip connection layers.

3. Our framework

Targeting at handling the real-world challenges with the current state-of-the-art methods, we propose a deep end-to-end learning framework for monocular depth estimation, which learns the direct mapping from a color image to the corresponding depth map. Our framework for monocular depth estimation consists of two stages: model training with classification loss and inference with soft-weighted-sum. First, by recasting monocular depth estimation as multi-class labeling, we design a hierarchical fusion dilated CNN to learn the mapping from an RGB image to the corresponding depth score map directly. Our network architecture hierarchically fuses multi-scale depth features to achieve scale-aware monocular depth estimation. Second, we propose a soft-weighted-sum inference by contrast to the hard-max inference, which transfers the discretized depth scores to continuous depth values. In this way, we could reduce the influence of quantization error and improve the robustness.

3.1. Network architecture

Our CNN architecture is illustrated in Fig. 3, in which the weights are initialized from a pre-trained 152 layers deep residual CNN (ResNet) [21]. Different from existing deep network [39,40], ResNet [21] explicitly learns the residual functions with reference to the layer inputs, which makes it easier to optimize with higher accuracy from considerably increased network depth. ResNet [21] was originally designed for image classification. In this work, we re-purpose it to make it suitable to our depth estimation task by

- Removing all the fully connected layers. In this way, we greatly reduce the number of model parameters as most of the parameters are in the fully connected layers [14]. Although preserving the fully connected layers is beneficial to extract long-range context information, our experiments show that it is unnecessary in our network thanks to dilated convolution.
- Taking advantage of the dilated convolution [37]. Dilated convolution could expand the receptive field of the neuron without increasing the number of model parameters. Furthermore, with the dilated convolution, we could remove some pooling layers without decreasing the size of the receptive field of correspondent neurons. In addition, we could keep the resolution of the feature map and final results, *i.e.*, the output resolution has been increased.
- Hierarchical fusion. We concatenate intermediate feature maps with the final feature map directly. This skip connection design could benefit the multi-scale feature fusion and boundary preserving.

Dilated convolution: Recently, dilated convolution [37] has been successfully utilized in deep convolutional neural network, which could expand the field of perception without increasing the number of model parameters. Specially, let $F : \mathcal{Z}^2 \to \mathcal{R}$ be a discrete function. Let $\Omega_r = [r, r]^2 \cap \mathcal{Z}^2$ and let $k : \Omega_r \to \mathcal{R}$ be a discrete filter of size $(2r+1)^2$. The discrete convolution filter $*$ can be expressed as

$$(F * k)(\mathbf{p}) = \sum_{\mathbf{s}+\mathbf{t}=\mathbf{p}} F(\mathbf{s})k(\mathbf{t}). \tag{1}$$

We now generalize this operator. Let l be a dilation factor and let $*_l$ be defined as

$$(F *_l k)(\mathbf{p}) = \sum_{\mathbf{s}+l\mathbf{t}=\mathbf{p}} F(\mathbf{s})k(\mathbf{t}). \tag{2}$$

We refer to $*_l$ as a dilated convolution or an l-dilated convolution. The conventional discrete convolution $*$ is simply the 1-dilated convolution. An illustration of dilated convolution could be found in Fig. 4.

Hierarchical fusion: As the CNN is of hierarchical structure, which means high-level neurons have larger receptive field and more abstract features, while the low-level neurons have smaller receptive field and more detail information. Thus, combining multi-scale information for pixel-level prediction tasks have received considerable interests.

We propose to concatenate the high-level feature map and the intermediate feature map. The skip connection structure benefits both the multi-scale fusion and boundary preserving. In our network, the $L1$, $L2$, $L3$, $L4$ layers are of the same size, we concatenate them directly.

In conclusion, we briefly summarize our final network design. Typically, the pre-trained residual network consists of of 4 parts. We remove the max-pooling layer in the last 2 parts and expand the corresponding convolution kernel with dilation 2 and 4 respectively. Then, a concatenation layer is added to fuse the hierarchical multi-scale information from layers $L1 - L4$. The last two layers of our network are convolution layer and deconvolution layer. The parameters setting is presented in Fig. 3.

3.2. Soft-weighted-sum inference

As illustrated in Fig. 1, the depth values follow an imbalance distribution due to the perspective effect. Transforming the depth to the log space could significantly reduce the imbalance effect. Therefore, we reformulate monocular depth estimation as a classification task by equally discretizing the depth value in the log space. Specifically,

$$l = \text{round}((\log(d) - \log(d_{\min}))/q), \tag{3}$$

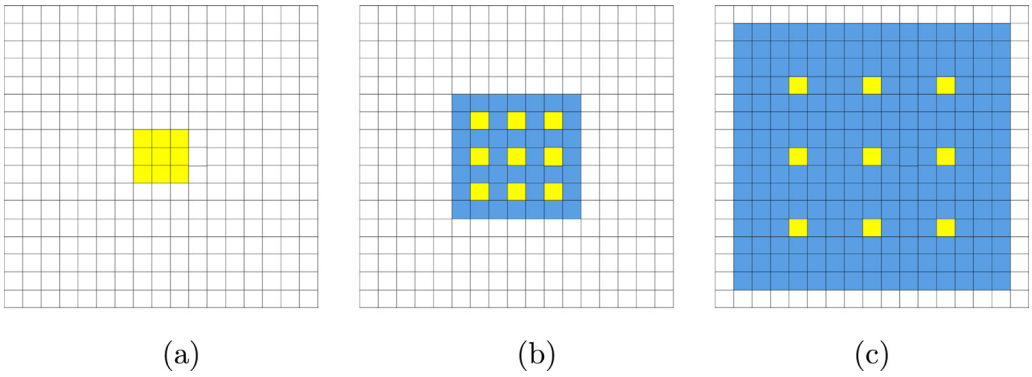

Fig. 4. Systematic dilation supports the exponential expansion of the receptive field without loss of resolution or coverage. (a), (b), (c) are 1-dilated, 2-dilated, 4-dilated convolution respectively. And the corresponding receptive fields are 3×3, 7×7, and 15×15. The receptive field grows exponentially while the number of parameters is fixed.

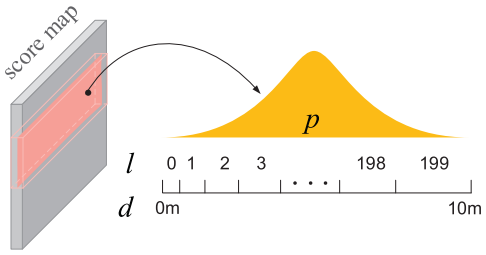

Fig. 5. Illustration of our discretizing method, which maps a continuous depth value to a discrete label. The inverse mapping is defined between the discrete depth label and the continuous depth value.In the figure, we illustrate the non-linear mapping between the actual depth and the discrete depth label.

where l is the quantized label, d is the continuous depth value, d_{min} is the minimum depth value in the dataset or set to be a small value like 0.1. q is the width of the quantization bin. An illustration is presented in Fig. 5.

With the quantized label, we train our network with the multinomial logistic loss.

$$L(\theta) = -\left[\sum_{i=1}^{N}\sum_{k=1}^{K} 1\{y^{(i)} = k\} \log \frac{\exp(\theta^{(k)T}x^{(i)})}{\sum_{i=1}^{K}\exp(\theta^{(j)T}x^{(i)})}\right], \tag{4}$$

where N is the number of training samples, $\exp(\theta^{(k)T}x^{(i)})$ is the probability of label k of sample i, and k is the ground truth label.

In the testing stage, we propose to use the soft-weighted-sum inference. It is worth noting that, this method transforms the predicted score to the continuous depth value in a natural way. Specifically:

$$\hat{d} = \exp\{\mathbf{w}^T\mathbf{p}\}, w_i = \log(d_{\min}) + q \cdot i, \tag{5}$$

where $\mathbf{w}$ is the weight vector of depth bins. $\mathbf{p}$ is the output score. In our experiments, we set the number of bins to 200 for the NYU v2 dataset, Make3D dataset and 50 for the KITTI dataset. An illustration is presented in Fig. 5.

3.3. Data augmentation

According to the work of [12–18,29,31], proper data augmentation is important in improving the final depth estimation performance. In this work, we augment our dataset by 4 times for the NYU v2 and the KITTI dataset, and 10 times for the Make3D dataset. The augmentation methods we utilized include:

- Color: Color channels are multiplied by a factor $c \in [0.9, 1.1]$ randomly.
- Scale: We scale the input image by a factor of $s \in [1.3, 1.5]$ randomly and crop the center patch of images to match the network input size.[2]
- Left-Right flips: We flip the input color image left and right horizontally.
- Rotation: We rotate the input image randomly by a factor of $r \in [-5, 5]^{\circ}$.

3.4. Implementation details

Before proceeding to the experimental results, we give implementation details of our method. Our implementation is based on the efficient CNN toolbox: Caffe [41] with an NVIDIA Tesla Titan X GPU.

The proposed network is trained by using stochastic gradient decent with a batch size of 2 (This size is too small, thus we average the gradient of 8 iterations for one back-propagation), the momentum of 0.9, and weight decay of 0.0004. Weights are initialized by the pre-trained model from ResNet [21].The network is trained with iterations of 60k by a fixed learning rate 0.001 in the first 40k iterations, then divided by 10 every 10k iterations. It takes about 48 hours for our training for the NYU v2 and the KITTI dataset. Our method takes about 0.5 seconds on average to predict an input map of resolution 320×240.

4. Experimental results

In this section, we report our experimental results on monocular depth estimation for both outdoor and indoor scenes. We used the NYU v2 dataset [4], the KITTI dataset [10], and the Make3D dataset [22], which have been widely used in the previous works [22–24,26,29]. We compared our method with the state-of-the-art methods published recently.

For quantitative evaluation, we report errors obtained with the following metrics, which have been extensively used in [22–24,26,29]:

Threshold: % of d_i s.t. $\max\left(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\right) = \delta < thr$

Rel: $\frac{1}{|T|}\sum_{d\in T}|\hat{d} - d|/d$

$\log_{10}$: $\frac{1}{|T|}\sum_{d\in T}|\log_{10}\hat{d} - \log_{10} d|$

Rms: $\sqrt{\frac{1}{|T|}\sum_{d\in T}\|\hat{d} - d\|^2}$

Rms(log): $\sqrt{\frac{1}{|T|}\sum_{d\in T}\|\log\hat{d} - \log d\|^2}$

Rel(sqr): $\frac{1}{|T|}\sum_{d\in T}\|\hat{d} - d\|^2/d$

Rms(sc-inv): $\sqrt{\frac{1}{2|T|}\sum_{d\in T}\left(\log\hat{d} - \log d + \frac{1}{|T|}\sum_{d\in T}\left(\log d - \log\hat{d}\right)\right)^2}$

[2] The similar scaling factors have been widely used in related works. Upsampling the images too much will result in severe image distortion, While upsampling the images too little will not result in enough variation.

Table 1
Monocular depth estimation results on the NYU v2 dataset. All the results of competing methods are quoted directly from the reported papers.

Method	Train num	Accuracy (higher is better)			Error (lower is better)					
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Rel	Rel(sqr)	log10	Rms	Rms(log)	Rms(sc-inv)
Karsch et al. [25]	–	–	–	–	0.350	–	0.131	1.2	–	–
Ladicky et al. [24]	725	54.2%	82.9%	94.1%	–	–	–	–	–	–
Li et al. [11]	795	62.1%	88.6%	96.8%	0.232	–	0.094	0.821	–	–
Liu et al. [12]	795	65.0%	90.6%	97.6%	0.213	–	0.087	0.759	–	–
Wang et al. [13]	200k	60.5%	89.0%	97.0%	0.220	0.210	0.094	0.745	0.262	–
Eigen et al. [14]	120K	76.9%	95.0%	98.8%	0.158	0.121	–	0.641	0.214	0.171
Cao et al. [15]	120k	81.9%	96.5%	**99.2%**	0.141	–	0.060	**0.540**	–	–
Laina et al. [16]	12k	81.1%	95.3%	98.8%	0.127	–	0.055	0.573	–	–
Xu et al. [17]	95k	81.1%	95.4%	98.7%	**0.121**	–	**0.052**	0.586	–	–
Ours	12k	**83.2%**	**96.5%**	98.9%	0.134	**0.095**	0.056	**0.540**	**0.187**	**0.132**

where d is the ground truth depth, $\hat{d}$ is the estimated depth, and T denotes the set of all points in the images. Rms(sc-inv) denotes the scale-invariant root mean squared error. The scale-invariant mean squared error was defined in [29].

4.1. NYU v2 dataset

The NYU v2 dataset [4] contains around 240k RGB-depth image pairs, of which comes from 464 scenes, captured with a Microsoft Kinect. The official split consists of 249 training and 215 testing scenes. We equally sampled frames out of each training sequence, resulting in approximately 12k unique images. After offline augmentations, our dataset comprises of approximately 48k RGB-D image pairs. We fill in the invalid pixels of the raw depth map with the "colorization" method, which is provided in the toolbox of NYU v2 dataset [4].

The original image resolution is 480×640. We downsampled the images to 240×320 as our network input. The resolution of our network output is 120×160, which is half of the input size. For this dataset, we quantize the depth value into 200 bins.

In Table 1, we compared our method with the recent published state-of-the-art methods [14–17]. Clearly, our method achieves the best performance under most of the error metrics. It is worth noting that reference [14], [15] and [17] utilized much more samples than us. In addition, references [15] utilized CRF as post processing method, while we have not applied any post-processing.

In Fig. 6, we provide a qualitative comparison of our method with [16] and [29]. (We compare with these methods as they represent the state-of-the-art methods). From Fig. 6, it is clear to observe that our results are of high visual quality, although we have not applied any post-processing.

4.2. KITTI dataset

The KITTI dataset [10] consists of a large number of outdoor street scene images of the resolution 376×1242. We utilized the "Eigen" training/testing split, which consists of 22,600 training images and 697 testing images. We fill in the invalid pixels of the raw depth maps with the "colorization" method, which is provided in the toolbox associated with the NYU v2 dataset [4]. For the error calculation, we only consider the lower crop of the image of dimension 256×1242. While in the training phase, we input the entire image to the network for more context information.

The original image resolution is 376×1240. We downsampled the images to 188×620 as our network input. The resolution of our network output is 94×310, which is half of the input size. For this dataset, we quantize the depth value into 50 bins.

We compared our method with the recent published state-of-the-art methods [15,18,29,31,32] with two different caps for the depth range (0–80 m and 0–50 m). As demonstrated in Table 2, our method achieves the best performance for most of the error metrics.

In Fig. 7, we provide a qualitative comparison of our method with [14] and [18]. (We compare with these methods as they represent the state-of-the-art methods.). From Fig. 7, it is clear to observe that our results are of high visual quality, although we have not applied any post-processing.

4.3. Make3d dataset

The Make3D [22] dataset consists of 534 images with corresponding depth maps. There are 400 training images and 134 testing images. All images were resized to 460×345 pixels. It is worth noting that this dataset was published many years ago, the resolution and distance range of the depth image is rather limited (only 55×355). Furthermore, it contains noise in the locations of glass window, boundary of objects *etc*. These limitations have some influence on the training stage and the resulting error metrics. Therefore we report errors based on two different criteria as presented in [26]: ($\mathbf{C}_1$) Errors are computed in the regions with ground-truth depth less than 70; ($\mathbf{C}_2$) Errors are computed in the entire image pixel, whose depth value is between (0,81]. We compare our method with the state-of-the-art methods, such as [16] and [17] *etc*. As illustrated in Table 3, Note that our method clearly outperforms most of the competing methods expect the most recent methods [16] and [17]. Laina et al. [16] utilized the Huber loss, which is more effective in handling noisy and outlying measurements. Xu et al. [17] is built upon hierarchical CRF, which is able to fit the noisy measurements.

Furthermore, we present a qualitative comparison of the depth estimation (in Fig. 8) with some methods on representative images from the Make3D dataset, which further demonstrates the superior performance of our method.

5. Performance analysis

In this section, we present more analysis of our model, where the experiments are conducted on the NYU v2 and KITTI dataset. First, we present a component analysis of our network architecture design, *i.e.*, the contribution of each component. Then, we analyze the choice of the number of depth bins. Lastly, we analyze the distribution of our network output, and the necessity of our soft-weighted-sum inference strategy.

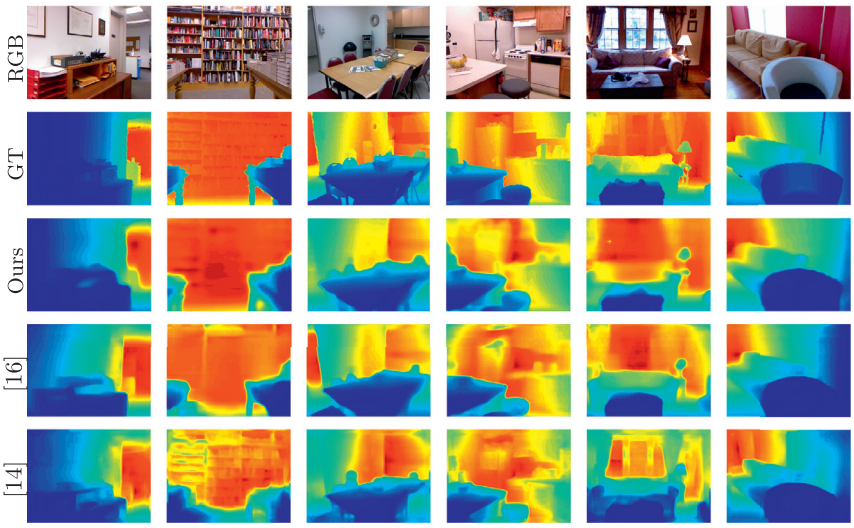

Fig. 6. Qualitative comparison of the estimated depth map on the NYU v2 dataset with our method and some state-of-the-art methods. Color indicates depth (red is far, blue is close). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Monocular depth estimation results on the KITTI dataset. All the results of the competing methods are quoted directly from the reported papers. K means the KITTI dataset. For a fair comparison, all the methods utilize the standard "Eigen split".

Method	Dataset	Cap	Accuracy (higher is better)			Error (lower is better)					
			$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Rel	Rel(sqr)	log10	Rms	Rms(log)	Rms(sc-inv)
Liu et al. [12]	K	80 m	65.6%	88.1%	95.8%	0.217	–	0.092	7.046	–	–
Eigen et al. [29] Coarse	K	80 m	67.9%	89.7%	96.7%	0.194	1.531	–	7.216	0.273	0.248
Eigen et al. [29] Fine	K	80 m	69.2%	89.9%	96.7%	0.190	1.515	–	7.156	0.270	0.246
Zhou et al. [32]	K	80 m	67.8%	88.5%	95.7%	0.208	1.768	–	6.856	0.283	–
Godard et al. [18]	K	80 m	80.3%	92.2%	96.4%	0.148	1.344	–	5.927	0.247	–
Cao et al. [15]	K	80 m	**88.7%**	96.3%	98.2%	0.115	–	–	4.712	0.198	–
Ours	K	80 m	86.8%	**96.7%**	**99.0%**	**0.104**	**0.697**	**0.046**	**4.513**	**0.164**	**0.173**
Garg et al. [31]	K	50 m	74.0%	90.4%	96.2%	0.169	1.080	–	5.104	0.273	–
Zhou et al. [32]	K	50 m	69.6%	90.0%	96.6%	0.201	1.391	–	5.181	0.264	–
Godard et al. [18]	K	50 m	81.8%	93.1%	96.9%	0.140	0.976	–	4.471	0.232	–
Cao et al. [15]	K	50 m	**89.8%**	96.6%	98.4%	0.107	–	–	**3.605**	0.187	–
Ours	K	50 m	87.9%	**96.9%**	**99.1%**	**0.101**	**0.586**	**0.044**	3.624	**0.157**	**0.167**

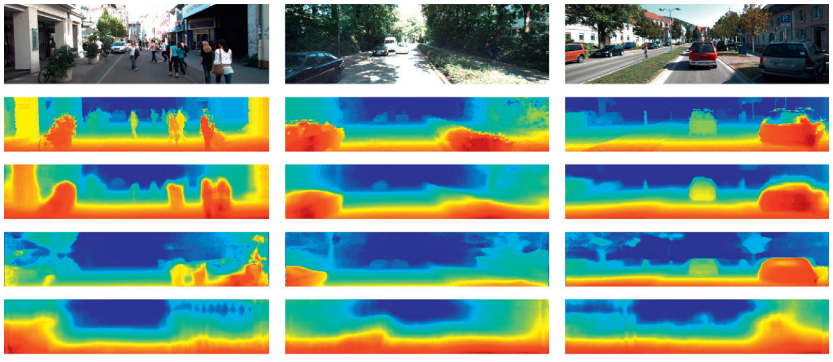

Fig. 7. Qualitative comparison of the depth map estimated on KITTI dataset. Color indicates depth (red is close, blue is far). Here, to better illustrate the details in the depth map, we use different coloring strategies for the NYU V2 dataset and the Make 3D dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1. Effect of architecture design

In order to explore the effectiveness of our hierarchical fusion dilated CNN, we conduct the following component analyze experiments. First, we train our network as a regression problem with the L_2 normal loss, instead of the classification. Second, we utilize the normal convolution kernel instead of the dilated convolution kernel in our network framework. Third, we analyze the role of the skip connection structure. At last, we use the different depth of ResNet (ResNet50, ResNet101, ResNet152) as our back-bone network. The corresponding experimental results are presented in Table 4. As we can see, recasting monocular depth estimation as a classification problem really improve the performance, and the dilated convolution could improve the performance significantly. Hierarchical fusion also improves the performance when compared with the baseline. When comparing the performance under different depth of ResNet, the deepest ResNet152 gives the best performance.

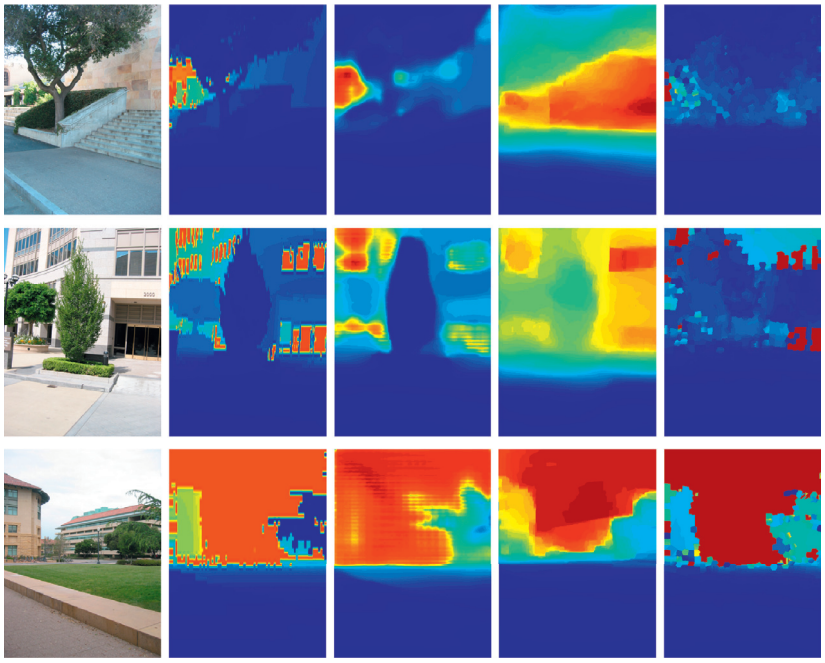

Fig. 8. Qualitative comparison of the depth map estimated on the Make3D dataset. Color indicates depth (red is far, blue is close). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Depth estimation results on the Make3D dataset.

Method	Error (C1) (lower is better)			Error (C2) (lower is better)		
	Rel	log10	Rms	Rel	log10	Rms
Karsch et al. [25]	0.355	0.127	9.2	0.361	0.148	15.1
Liu et al. [26]	0.335	0.137	9.49	0.338	0.134	12.6
Li et al. [11]	0.278	0.092	7.19	0.279	0.102	10.27
Liu et al. [12]	0.287	0.109	7.36	0.287	0.122	14.09
Ji et al. [28]	0.335	0.147	–	–	–	–
Godard et al. [18]	0.443	0.143	8.860	–	–	–
Laina et al. [16]	**0.176**	0.072	**4.46**	–	–	–
Xu et al. [17]	0.184	**0.065**	**4.38**	**0.198**	–	**8.56**
Ours	0.199	0.069	4.7	0.219	0.089	11.7

5.2. Effect of number of depth bins

To analyze the role of discretization of depth values, we vary the number of depth bins from 20 to 1000. Under the same training setting, we present the results in the form of "classification accuracy", which is the fraction of correctly predicted pixel labels over all pixels and the relative errors are reported in Table 5. With the increase of the number of bins, the "classification accuracy" drops dramatically. This trend presents that: At present, the network cannot distinguish the very detailed distance variation even if we train it with the very detailed "label". In other words, "the resolution of depth perception" of the network is limited.

Table 5
Component evaluation for different number of depth bins on the NYU v2 and the KITTI dataset.

Number of bins	20	50	100	200	500	1000
NYU v2						
Classification accuracy	90	67	41	25	12	7
Rel	0.209	0.179	0.138	0.134	0.137	0.139
KITTI						
Classification accuracy	92	71	56	32	26	11
Rel	0.144	0.104	0.105	0.107	0.102	0.104

Table 4
Component analysis for our classification based CNN architecture design on the NYU v2 dataset and the KITTI dataset.

Dataset	Method	Accuracy (%)			Error		
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Rel	log10	Rms
NYU v2	Regression	78.3%	95.6%	98.9%	0.147	0.069	0.624
	No dilation	78.2%	94.6%	98.5%	0.149	0.062	0.617
	No concat layer	82.5%	96.4%	98.9%	0.137	0.059	0.547
	Ours(ResNet50)	80.8%	95.7%	98.5%	0.147	0.067	0.601
	Ours(ResNet101)	82.0%	96.0%	98.9%	0.139	0.058	0.545
	Ours(ResNet152)	**83.2%**	**96.5%**	**98.9%**	**0.134**	**0.056**	**0.540**
KITTI	Regression	81.9%	93.7%	97.5%	0.171	0.068	5.279
	No dilation	83.1%	94.7%	98.5%	0.132	0.061	5.235
	No concat layer	85.9%	95.9%	98.8%	0.113	0.055	4.671
	Ours(ResNet50)	83.3%	95.6%	98.5%	0.128	0.060	5.325
	Ours(ResNet101)	85.7%	96.5%	98.9%	0.106	0.049	4.528
	Ours(ResNet152)	**86.8%**	**96.7%**	**99.0%**	**0.104**	**0.046**	**4.513**

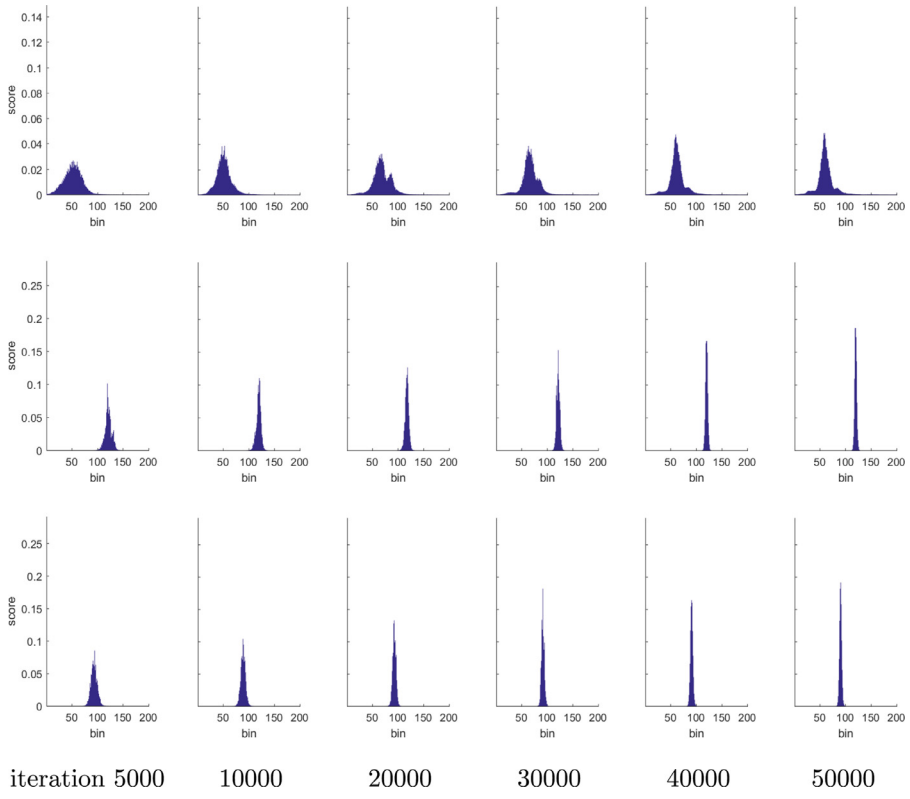

Fig. 9. Typical score distribution variation of our network output with respect to iterations. The points are randomly selected from the NYU v2 dataset.

While according to Table 5, we observe that the performance of monocular depth estimation is robust if the number of discretization bins is large enough. However, after an inflection point, the performance will keep stable. As we can see from Table 5, the relative error "Rel" has slight fluctuation when the number of bins is larger than 100.

5.3. *Effect of soft-weighted-sum inference*

One important contribution of this work is the proposed soft-weighted-sum inference. Here, we would like to elaborate the necessity and the effectiveness of it.

Firstly, we give the probability distribution variation of randomly selected positions along the training images in Fig. 9. The most interesting thing is that: In the training stage, we utilize the multinomial logistic loss, which means we do not specifically discriminate the distance between the "nearby" and "further" classes. While the probability distribution is rather clustered. More interestingly, the probability distribution roughly follows the Gaussian distribution, which means it is symmetric. At last, as the training goes on, the distribution of probability is becoming more concentrated, but always maintains symmetry similar to that of the Gaussian distribution.

Secondly, we use the hard-max inference and give the confusion matrix in Fig. 10. The confusion matrix presents a kind of diagonal dominant and symmetric structure, which means most of the error prediction occurs in nearby classes.

More discussions: These statistic results show that: Even though the model cannot distinguish the detailed depths well, it still learns the correct "concept" of depth as the non-zero predicted score is centralized, symmetry and around the right label. Considering these results, we propose the soft-weighted-sum inference instead of the hard-max inference.

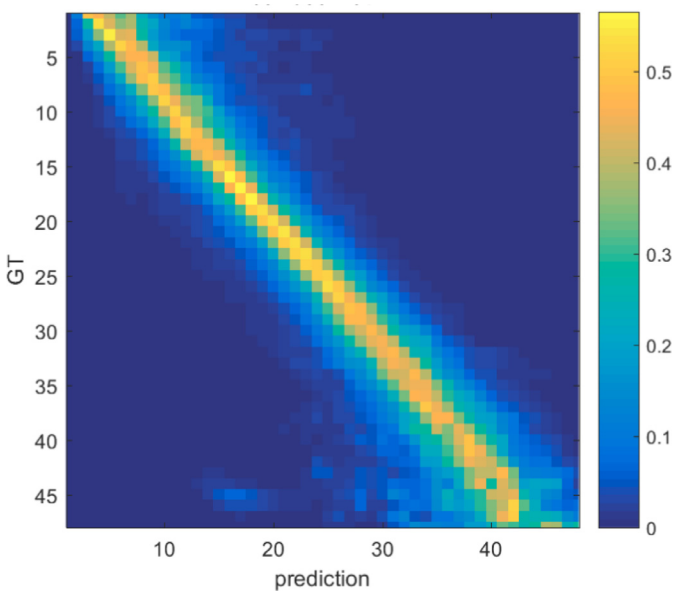

Fig. 10. Confusion matrix on the NYU v2 dataset. Here, we merge the 200 bins to 50 for better illustration. The larger class number means larger depth value.

In order to evaluate the effectiveness of our soft-weighted-sum inference, we give the contrast experiment in Table 6. The Table 6 shows that our soft-weighted-sum inference achieves better quantitative results.

In addition, we also give the quality comparison of the hard-max inference and our soft-weighted-sum inference in Fig. 11, which illustrate that our soft-weighted-sum method could get smoother results and keep the boundary well.

Table 6
Component analysis for our soft-weighted sum inference on the NYU v2 and KITTI dataset.

Dataset	Method	Accuracy (%)			Error		
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Rel	log10	Rms
NYU v2	Hard-max	82.4%	96.3%	98.9%	0.140	0.058	0.553
	Soft-weighted-sum	**83.2%**	**96.5%**	**98.9%**	**0.134**	**0.056**	**0.540**
KITTI	Hard-max	85.0%	96.5%	98.9%	0.112	0.052	4.718
	Soft-weighted-sum	**86.8%**	**96.7%**	**99.0%**	**0.104**	**0.046**	**4.513**

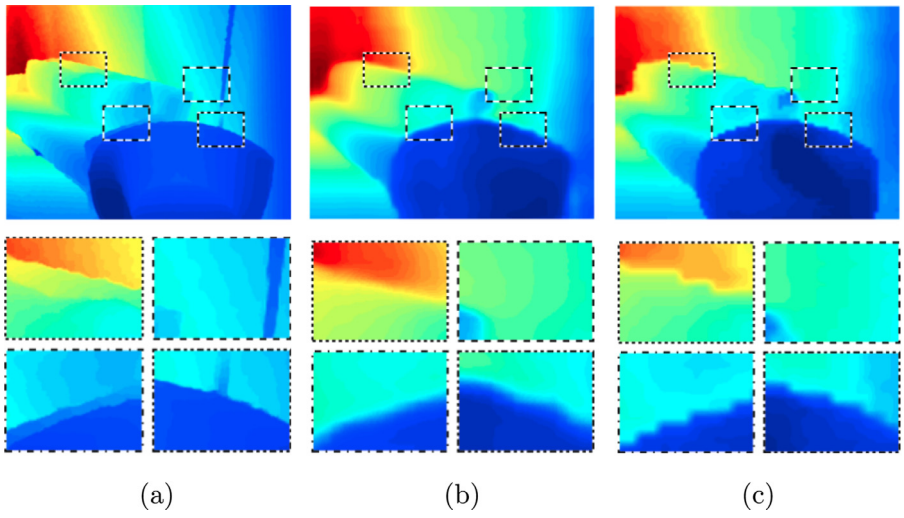

Fig. 11. Qualitative comparison between our soft-weighted-sum inference and hard-max inference on the NYU v2 dataset, (a) are the ground truth depth map. (b) are the results of our soft-weighted-sum inference. (c) are the results of hard-max inference.

Table 7
Performance evaluation for different data augmentation methods on the NYU v2 dataset.

Method	Accuracy (%)			Error		
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Rel	log10	Rms
Original	71.2%	93.3%	97.2%	0.199	0.103	0.805
Color	72.3%	93.5%	97.9%	0.189	0.097	0.793
Flip	73.2%	94.5%	97.5%	0.183	0.089	0.697
Scale	73.5%	94.6%	97.9%	0.181	0.085	0.610
Angle	72.2%	93.5%	97.1%	0.181	0.095	0.740

5.4. Effect of data augmentation

In this paper, we have used 4 different data augmentation methods: color, scale, flips, and rotation. All these methods have been widely used in related works such as [12–18,29,31]. The basic principle of data augmentation method is to increase the diversity of the training samples such that the learned network owns better generalization ability and reduces the influence of over-fitting. In the seminal work of [29], the authors have emphasized the importance of these data augmentation methods.

To analyze the contribution of each data augmentation method, we perform corresponding ablation studies. The corresponding results are reported in Table 7. We train our model on the standard NYU V2 training set with 795 images for fast comparison. As for the testing sets, we use standard 654 testing images. For each data augmentation method, we augment the original dataset by 4 times respectively. The results in Table 7 demonstrate that all of these 4 data augmentation methods improve the monocular depth estimation performance.

6. Conclusions

In this paper, we have proposed a deep end-to-end classification based framework to monocular depth estimation. By using both dilated convolution and hierarchical fusion of multi-scale features, our framework is able to deal with the real world difficulties in multi-scale depth estimation. Extensive experiments on both indoor and outdoor benchmarking datasets show the superiority of our method compared with the current state-of-the-art methods. More importantly, experiments also demonstrate that our model is able to learn a probability distribution among different depth labels, which inspires the proposed soft-weighted-sum inference.

Acknowledgments

This work was supported in part by Natural Science Foundation of China (61420106007, 61671387) and Australian Research Council (ARC) grant no. DE140100180.

References

[1] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, D. Kim, Robust human activity recognition from depth video using spatiotemporal multi-fused features, Pattern Recognit. 61 (2017) 295–308.
[2] H. Pan, S.I. Olsen, Y. Zhu, Feature representation of rgb-d images using joint spatial-depth feature pooling, Pattern Recognit. Lett. 80 (2016) 239–248.
[3] L. Chen, H. Wei, J. Ferryman, A survey of human motion analysis using depth imagery, Pattern Recognit. Lett. 34 (15) (2013) 1995–2006.
[4] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[5] A. Gupta, A. Efros, M. Hebert, Blocks world revisited: Image understanding using qualitative geometry and mechanics, in: Proceedings of the European Conference on Computer Vision, Springer, 2010, pp. 482–496.
[6] D. Fouhey, A. Gupta, M. Hebert, Unfolding an indoor origami world, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 687–702.

[7] C. Wu, J.-M. Frahm, M. Pollefeys, Repetition-based dense single-view reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3113–3120.
[8] R. Zhang, P.-S. Tsai, J.E. Cryer, M. Shah, Shape-from-shading: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 21 (8) (1999) 690–706.
[9] D. Ziou, F. Deschenes, Depth from defocus estimation in spatial domain, Comp. Vis. Image Underst. 81 (2) (2001) 143–165.
[10] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.
[11] B. Li, C. Shen, Y. Dai, A. van den Hengel, M. He, Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1119–1127.
[12] F. Liu, C. Shen, G. Lin, I. Reid, Learning depth from single monocular images using deep convolutional neural fields, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2016) 2024–2039.
[13] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A.L. Yuille, Towards unified depth and semantic prediction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2800–2809.
[14] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.
[15] Y. Cao, Z. Wu, C. Shen, Estimating depth from monocular images as classification using deep fully convolutional residual networks, IEEE Trans. Circuits Syst. Video Technol. (2017), doi:10.1109/TCSVT.2017.2740321.
[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, in: Proceedings of the Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 239–248.
[17] D. Xu, E. Ricci, W. Ouyang, X. Wang, N. Sebe, Multi-scale continuous crfs as sequential deep networks for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5354–5362.
[18] C. Godard, O.M. Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270–279.
[19] Y. Wang, J. Liu, Y. Li, J. Fu, M. Xu, H. Lu, Hierarchically supervised deconvolutional network for semantic video segmentation, Pattern Recognit. 64 (2017) 437–445.
[20] S. Mei, J. Ji, J. Hou, X. Li, Q. Du, Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks, IEEE Trans. on Geoscience and Remote Sensing 55 (8) (2017) 4520–4533.
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[22] A. Saxena, M. Sun, A.Y. Ng, Make3d: learning 3d scene structure from a single still image, IEEE Trans. Pattern Anal. Mach. Intell. 31 (5) (2009) 824–840.
[23] B. Liu, S. Gould, D. Koller, Single image depth estimation from predicted semantic labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1253–1260.
[24] L. Ladicky, J. Shi, M. Pollefeys, Pulling things out of perspective, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 89–96.
[25] K. Karsch, C. Liu, S.B. Kang, Depth transfer: depth extraction from video using non-parametric sampling, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2144–2158.
[26] M. Liu, M. Salzmann, X. He, Discrete-continuous depth estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 716–723.
[27] A. Roy, S. Todorovic, Monocular depth estimation using neural regression forest, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5506–5514.
[28] R. Ji, L. Cao, Y. Wang, Joint depth and semantic inference from a single image via elastic conditional random field, Pattern Recognit. 59 (2016) 268–281.
[29] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2366–2374.
[30] W. Chen, Z. Fu, D. Yang, J. Deng, Single-image depth perception in the wild, in: Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 730–738.
[31] R. Garg, G. Carneiro, I. Reid, Unsupervised cnn for single view depth estimation: Geometry to the rescue, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 740–756.
[32] T. Zhou, M. Brown, N. Snavely, D.G. Lowe, Unsupervised learning of depth and ego-motion from video, in: Proceedings of the European Conference on Computer Vision, 2017, pp. 1851–1860.
[33] Y. Kuznietsov, J. Stückler, B. Leibe, Semi-supervised deep learning for monocular depth map prediction, in: Proceedings of the European Conference on Computer Vision, 2017, pp. 6647–6655.
[34] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, T. Brox, Demon: Depth and motion network for learning monocular stereo, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5038–5047.
[35] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.
[36] B. Hariharan, P. Arbelaez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.
[37] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, in: Proceedings of the ICLR, 2016, pp. 1–10.
[38] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017) 1.
[39] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[40] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the ICLR, 2015.
[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 675–678.

Bo Li is currently a senior Ph.D. student in the School of Electronics and Information, Northwestern Polytechnical University (NPU), China. He received the B.E. degree in Electronic and Information Engineering from Northwestern Polytechnical University, Xian, China, in 2011. During 2013–2015, He was a visiting student in the University of Adelaide, Australia. His research interest mainly focuses on the deep CNN and its applications in the Computer Vision field like single image depth estimation, hyperspectral image classification, and skeleton based video recognition etc. He has published some papers on the CVPR, ICIP, ICME, IEEE TMM, MTA, etc.

Yuchao Dai is currently a Professor with School of Electronics and Information at the Northwestern Polytechnical University (NPU). He received the B.E. degree, M.E degree and Ph.D. degree all in signal and information processing from Northwestern Polytechnical University, Xian, China, in 2005, 2008 and 2012, respectively. He was an ARC DECRA Fellow with the Research School of Engineering at the Australian National University, Canberra, Australia from 2014 to 2017 and a Research Fellow with the Research School of Computer Science at the Australian National University, Canberra, Australia from 2012 to 2014. His research interests include structure from motion, multi-view geometry, low-level computer vision, deep learning, compressive sensing and optimization. He won the Best Paper Award in IEEE CVPR 2012, the DSTO Best Fundamental Contribution to Image Processing Paper Prize at DICTA 2014, the Best Algorithm Prize in NRSFM Challenge at CVPR 2017, the Best Student Paper Prize at DICTA 2017 and the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017.

Mingyi He (M06, SM16) received the B.Eng. and M.S. degrees in electronic engineering and signal processing from Northwestern Polytechnical University (NPU), Xian, China, in 1982 and 1985, respectively, and the Ph.D. degree in signal and information processing from Xidian University, Xian, China, in 1994. Since 1985, he has been with the School of Electronics and Information, NPU, where he has been a Full Professor since 1996. He is the Founder and Director of Shaanxi Key Laboratory and International Research Center for Information Acquisition and Processing, and the Director and Chief Scientist of the Center for Earth Observation Research, NPU. He had been a visiting scholar at Adelaide University, Adelaide, S.A., Australia, and Visiting Professor with Sydney University, Sydney, N.S.W., Australia, and Adelaide University. His research interests focus on advanced machine vision and intelligent processing, including signal and image processing, computer vision, hyperspectral remote sensing, 3-D information acquisition and processing, neural network, and deep learning. Dr. Mingyi He was a recipient of 11 national and provincial scientific prizes and 2 teaching achievement prizes in China. He was the corecipient of the 2012 CVPR best paper award, the 2017 APSIPA ASC best deep/machine learning paper award, the 2017 DICTA best student paper award, etc. He was also a recipient of the government lifelong subsidy from the State Council of China in 1993 and 2017 Baosteel Outstanding Teacher Award. He has acted as a General Chair or TPC (Co)Chair, and Area Chair for several national and international conferences. He has been a member of the Advisory Committee of National Council for Higher Education on Electronics and Information in China, a member of Chinese Lunar Exploration Expert Group, the Vice-President of Shaanxi Institute of Electronics, and the Vice-Director of the Spectral Imaging Earth Observation Committee of China Committee of International Society of Digital Earth. He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and a SIPTM (Signal and Information Processing Theory and Methods) committee member of the Asia-Pacific Signal and Information Processing Association.