Unsupervised High-Resolution Depth Learning From Videos With Dual Networks

Junsheng Zhou[1,*], Yuwang Wang[2], Kaihuai Qin[1], and Wenjun Zeng[2]

[1]Tsinghua University, Beijing, China
zhoujs17@mails.tsinghua.edu.cn, qkh-dcs@mail.tsinghua.edu.cn
[2]Microsoft Research, Beijing, China
yuwwan, wezeng@microsoft.com

Abstract

Unsupervised depth learning takes the appearance difference between a target view and a view synthesized from its adjacent frame as supervisory signal. Since the supervisory signal only comes from images themselves, the resolution of training data significantly impacts the performance. High-resolution images contain more fine-grained details and provide more accurate supervisory signal. However, due to the limitation of memory and computation power, the original images are typically down-sampled during training, which suffers heavy loss of details and disparity accuracy. In order to fully explore the information contained in high-resolution data, we propose a simple yet effective dual networks architecture, which can directly take high-resolution images as input and generate high-resolution and high-accuracy depth map efficiently. We also propose a Self-assembled Attention (SA-Attention) module to handle low-texture region. The evaluation on the benchmark KITTI and Make3D datasets demonstrates that our method achieves state-of-the-art results in the monocular depth estimation task.

1. Introduction

Estimating depth from RGB images has broad applications such as scene understanding, 3D modelling, robotics, autonomous driving, etc. However, depth estimation from a single image is a well-known ill-posed problem and has inherent scale ambiguity. Recently supervised deep learning methods [7, 30, 22] have achieved tremendous success in monocular depth estimation tasks. These methods use convolutional neural networks to predict depth from single RGB image input under the supervision of ground-truth depth obtained by laser scanners. But fully supervised learning methods still suffer from lack of large scale and diverse datasets. So unsupervised methods [11, 54] have been proposed to get rid of ground truth labeling and have attracted increasing interest in recent years.

The key idea of unsupervised learning from video is to simultaneously estimate depth of scenes and ego-motion of camera, and use the predicted depth and pose to synthesize target view from source view based on geometric constraint. Then the appearance difference between the synthesized target view and real target view is used as supervisory signal. This unsupervised method is similar to traditional structure-from-motion and stereo matching methods, and therefore suffers from the same problems like occlusion/disocclusion and non-texture region. Since the supervisory signal only comes from images themselves, the quality of images, e.g., the resolution, significantly impacts the performance. For example, disparity of distant object is always smaller than that of nearby object when the camera is moving forward. Since digital image is a discretized representation of the real world, its resolution limits the accuracy of the disparity. More specifically, it is difficult to distinguish whether the depth of an object is 40m or 80m when its disparity is less than one pixel. Actually we also observed that previous state-of-the-art methods that adopt down-sampling, are prone to produce large error on distant objects. In addition, due to the limitation of memory and computation power, previous methods usually take down-sampled images as training data and predict depth maps with the same size as input. Then the low-resolution results need to be upsampled to the original resolution, resulting in blurred border in this operation. Besides, due to the loss of fine-grained details, slim objects like traffic pole and tree are often neglected, which is a serious safety problem in

*Work done as an intern at MSRA.

practical application.

In this paper, we propose a simple yet effective dual (i.e., high and low resolution) networks architecture, which can efficiently leverage high-resolution data and directly generate high-resolution depth map. This new architecture effectively addresses the above problems and significantly improves the performance. The generated depth maps are fairly sharp and handle the slim objects and distant objects well.

As mentioned earlier, non-texture region is also an intractable problem for unsupervised depth learning. Supervisory signal is only derived from the image area that has varying intensity, meaning that the loss in the interior of low-texture region is always very scarce which may result in black hole in the depth map. Thus we propose a Self-assembled Attention module combined with dual networks to handle the non-texture problem. This module also leads to considerable improvement.

Empirical evaluation on the KITTI [9] and Make3D [38, 39] benchmarks demonstrates that our approach outperforms existing state-of-the-art methods for the monocular depth estimation task.

2. Related Work

Here we review the works which take single view image as input and estimate the depth of each pixel. These works are classified into supervised depth learning and unsupervised depth learning. We also introduce some super resolution and self-attention works that are related to our work.

Supervised depth learning Supervised depth learning trains the model with RGB image and ground truth depth label where there is clear supervision for each pixel. These methods need ground truth labels captured with time of flight device [14] or RGBD cameras, which has the limitation of high cost or limited depth range. [39] uses Markov Random Field (MRF) to infer a set of plane parameters for the given scenes. [17] tries non-parametric sampling to estimate the depth by matching the given image with the images in depth dataset. Eigen et al. [7] are the first to employ CNN in learning depth and they also proposed to use two networks. But there exists significant difference between theirs and our architecture. Their motivation is that the coarse depth map generated by a coarse-scale network can be used as additional information and concatenated with RGB images as the input of another network. Both networks take the same low-resolution images as inputs and the resolution of outputs is smaller. Our dual networks can directly process high-resolution images and output high-resolution depth maps. [24] builds a deep fully convolutional network based on ResNet [13]. A significant gain is achieved by leveraging the powerful deep learning method. [45, 29] try to explore the structured information in the depth map by either the CRF or explicit plane models. [8] treats depth estimation as a classification problem to achieve robust inference. Other works find that depth estimation can be combined with other tasks and benefit from each other, *e.g.*, normal [6, 36], segmentation [28, 23] and optical flow [33]. Due to lack of large scale depth labels, [21] increases the performance of depth learning with synthetic data and GAN [12] loss. [27] uses multi-view Internet photo collections to generate training data. [46] uses sparse points whose depth are estimated from SLAM system as supervision.

Unsupervised depth learning There is no ground truth depth label for self-supervised depth learning. Instead the network training is supervised by multiview constraint, *e.g.*, multiview images captured by stereo [11, 26] or multiview cameras or video captured by monocular cameras [54]. An appearance loss is calculated by warping the source view to the target view using the inferred depth. [31] adds a consistency constraint between the estimated depth from different views. [49] combines depth learning with normal and edge extraction and achieves a better performance. [56] proposes to leverage optical flow to estimate occlusions and remove invalid regions in computing the loss. [2] utilizes synthetic data and style transfer [55] to collect more diverse training data. [26] uses stereo image pairs to recover the scale.

CNN for Super Resolution Deep-learning-based Super Resolution (SR) is also a popular research topic recently, which aims to recover a high-resolution image from low-resolution image. Dong et al. [5] was the first to apply deep learning method in SR. Huang et al. [15] extended self-similarity based SR and used the detected perspective geometry to guide the patch searching. Kim et al. [18] used VGG as backbone and achieved better performance. Shi et al. [40] proposed an efficient sub-pixel convolution layer which learns an array of upscaling filters to upscale the LR feature maps. The architecture of SR is also used in unsupervised depth learning by Pillai et al. [35]. But only using low-resolution images as input is not able to fully explore the valuable information contained in high-resolution images. Our dual networks are different from super resolution and able to process high-resolution data efficiently. Especially for unsupervised depth learning, the resolution of training data significantly influences the performance.

Self-Attention Self-attention mechanism [41, 3, 47, 25] calculates the response at a position in a sequence by attending to all positions within the same sequence. In neural network, the convolutional and recurrent operations can only process on local neighborhood at a time. [43] presents non-local operations to capture long-range dependencies in the whole input image. [52] learns to efficiently find global, long-range dependencies within internal representations of images. [32] uses a non-local mechanism for hard attention

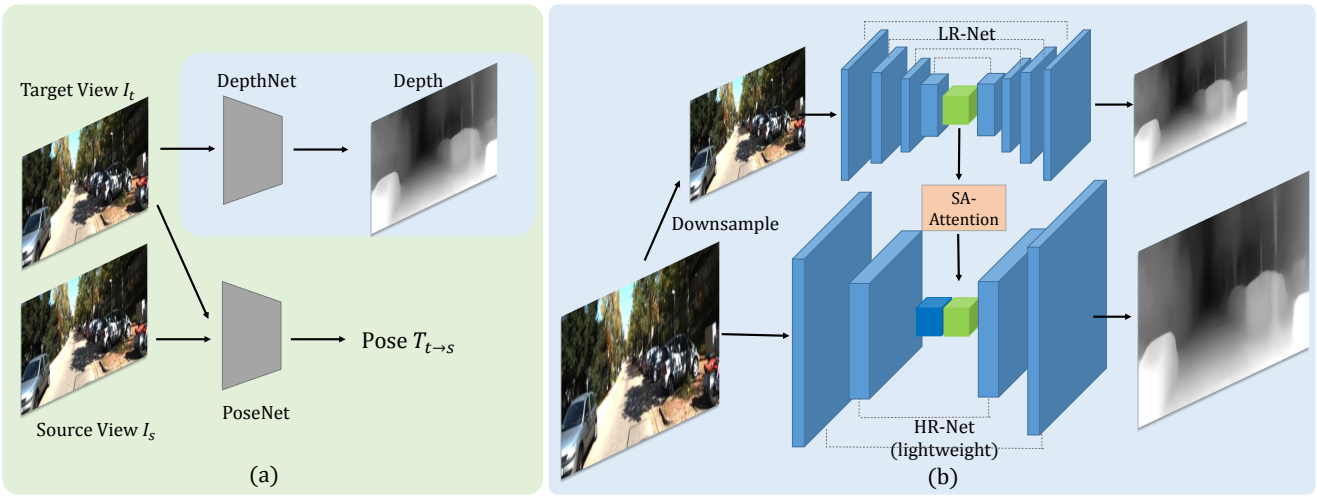

Figure 1. Overview of our method. (a) is the pipeline of unsupervised depth learning, which consists of depth network and pose network. (b) is the specific architecture of DepthNet in (a), which is composed of LR-Net, HR-Net and SA-Attention module. LR-Net takes low-resolution images as input and HR-Net takes high-resolution images as input. The deepest global features in LR-Net are fed to HR-Net. SA-Attention module refines the information flowing from LR-Net to HR-Net.

on visual question answering datasets. Our work aims to leverage self-attention as a mechanism to enhance supervisory signal in non-texture regions and refine the global features.

3. Approach

In this section, we first review the nature of 3D geometry of unsupervised depth learning. Then we introduce the proposed dual networks architecture and SA-Attention module.

3.1. Framework

The fundamental idea behind unsupervised depth learning is multiview geometry constraint. Given two frames I_s and I_t with known camera intrinsics, once the relative pose $T_{t\to s}$ and the scene depth of I_t are estimated, we can synthesize I_t from I_s as:

$$p_{s\to t} = KT_{t\to s}D_t(p_t)K^{-1}p_t \tag{1}$$

where p_t and $p_{s\to t}$ denote the homogeneous coordinates of a pixel in I_t and the synthesized view $I_{s\to t}$ respectively, D_t denotes the depth of target view and K denotes the camera intrinsic matrix. Then the homogeneous coordinates $p_{s\to t}$ can be projected to the image plane in a fully differential manner [16] to obtain the synthesized image $I_{s\to t}$.

The entire pipeline consists of two main modules: the DepthNet and the PoseNet, which aim to estimate monocular depth and pose between nearby views respectively. The supervisory loss is from the appearances difference between the target view and synthesized views as:

$$L_{vs} = \sum_j |I_t(j) - I_{s\to t}(j)| \tag{2}$$

where j indexes the pixel coordinates.

3.2. Dual Network Architecture

Motivation As mentioned above, due to the special nature of unsupervised depth learning, the resolution of training images significantly impacts the training effect. The deficiencies of training with low-resolution images can be summarized as below:

1) Due to the decrease of resolution, the accuracy of disparity deteriorates which results in large error on distant objects.

2) Upsampling the low-resolution depth map to high-resolution blurs the border of objects.

3) Loss of fine-grained details makes the model prone to ignore slim objects like traffic pole and tree.

However, directly training with high-resolution images requires vast computation resource. On the other hand, reducing the scale of the model will also deteriorate the performance. To address this problem, we propose a dual networks architecture which can fully explore the rich information contained in high resolution data while avoiding expensive cost. The key idea is that low-resolution data already contain enough global semantic information, and at the same time the most valuable information of high-resolution data are the fine-grained details. It is difficult and unnecessary to train high-resolution data in a single network. This means that we can separate the process of feature extraction into two parts. The first part captures global semantic information and the second part fills in the details.

Design As shown in Figure 1, the dual networks architecture consists of three components: Low-Resolution Network (LR-Net), High-Resolution Network (HR-Net) and

SA-Attention module that links these two networks. LR-Net takes low-resolution (128×416) images as input and HR-Net takes high-resolution (384×1248) images as input. Both networks share similar encoder-decoder architecture with skip-connection and generate low-resolution and high-resolution depth maps respectively. Their supervisory signal comes from the photometric loss computed by predicted depth and pose. LR-Net is designed to extract the important global semantic features from low-resolution images so it contains more convolutional layers and more parameters. However the training of LR-Net is efficient due to its small input size.

HR-Net is designed as a lightweight and shallow model so it only adds a small amount of overhead while it directly processes high-resolution data. Due to its limited capacity, HR-Net is not able to generate plausible result by itself. This module is specifically used to extract the fine-grained details in high-resolution images which will be combined with global features passed by LR-Net to generate high-resolution and high-accuracy depth map. More specifically, the deepest features in LR-Net with smallest size is delivered to HR-Net and concatenated with the deepest features in HR-Net. Then the concatenated features are upsampled gradually in the decoder and the details are also filled in. More details are shown in supplementary material. In addition, since the photometric loss can be computed by high-resolution depth map, the supervisory signal is more accurate which improves the performance.

In practice, both networks adopt a multi-scale architecture and predict depth maps with $\frac{1}{1}, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ resolution relative to the input size. LR-Net is firstly trained to achieve a considerable performance. Then the parameters are fixed and used as global features extractor for training HR-Net. Quantitative results of these two networks are shown in the ablation study later.

3.3. SA-Attention Module

Although with the help of dual networks, our model already achieves state-of-the-art performance, there still exists some problem in this pipeline. Since the important global features are learned by LR-Net, the accuracy of the learned global features directly affects the prediction of HR-Net. In some cases where severe non-texture region exists, the global features of that region cause much deviation and it is difficult for HR-Net to rectify that error.

To alleviate the non-texture problem, we propose a Self-assembled Attention module to refine the global features before they are fed to HR-Net as shown in Figure 1. The key idea is that two pixels with similar features, appearance or close spatial distance are more likely to have similar depth. SA-Attention module explicitly calculates the similarity between each position and all the other positions, and similar pixels are bundled together. In the SA module, we concatenate the input features, RGB values of resized input image and the pixels' coordinates together in feature channel. Then the concatenated feature is embedded in a low-dimensional space. Suppose the input feature is f_i, this operation can be done by a conv layer to obtain embedded feature f_i'. The assembled feature at the j position f_j'' can be written as a weighted sum of other similar features:

$$f_j'' = \sum_i w_{ij} \cdot f_i' \tag{3}$$

where i indexes the pixel coordinates and w_{ij} represents the similarity between feature at position i and j. There exist several choices to evaluate the similarity of two vectors, such as L1, L2 distance or cosine similarity. In practice, we use dot product of vectors for its simplicity:

$$w_{ij} = f_i' \cdot f_j' \tag{4}$$

Then the assembled feature f_j'' is passed to HR-Net. This module serves like a valve and refines the information flowing from LR-Net to HR-Net since it ensures pixels with strong similarity to have the similar depth features, which enhances the supervisory signal in non-texture regions. Its quantitative effect is shown in the ablation study later.

3.4. Loss

In this subsection we introduce the components of our training loss function.

Photometric Loss Following [11], we adopt a combination of L1 and the Structural Similarity (SSIM) [44] for appropriate assessment of the discrepancy between two images. In addition, per-pixel minimum trick proposed by [10] is also adopted. This trick is that we use three views to compute the photometric loss: one target view and two source views. So we obtain two error maps from two source views. Instead of averaging both error maps we calculate their minimum. This is an effective way to handle occlusion/disocclusion. So the final photometric loss is

$$L_{ph} = \sum_p \min_s (\alpha L_{SSIM} + (1-\alpha) \| I_t(p) - I_{s \to t}(p) \|_1) \tag{5}$$

where p indexes over pixel coordinates, s denotes the index of source views, α is set to 0.85 and L_{SSIM} represents

$$L_{SSIM} = \frac{1 - SSIM\left(I_t, I_{s \to t}\right)}{2} \tag{6}$$

Smoothness Loss Besides photometric loss, edge-aware depth smoothness loss is adopted which encourages the network to generate smooth prediction in continuous region while preserving sharp edge in discontinuous region:

$$L_{smooth} = \sum_p |\nabla D_t(p)| \cdot \left(e^{-|\nabla I_t(p)|}\right)^T \tag{7}$$

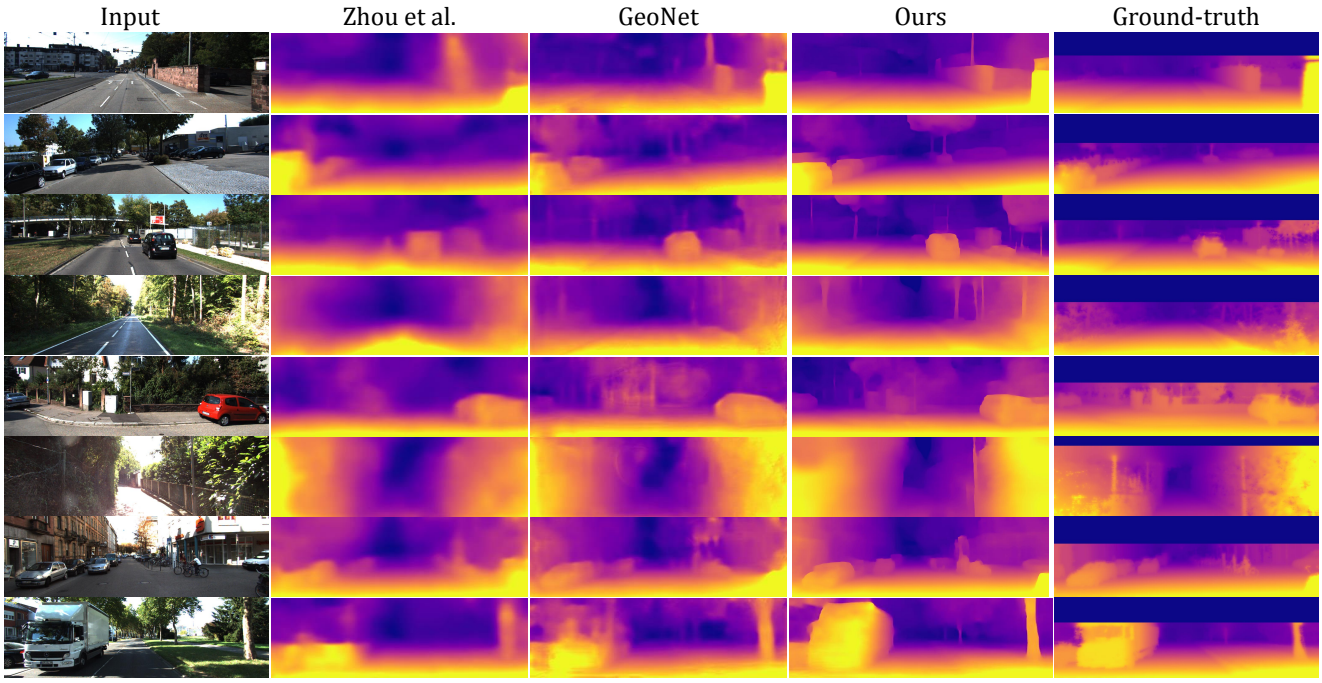

Figure 2. Qualitative comparison between Zhou *et al*. [54], GeoNet [50], ours and ground-truth (interpolated for visualization). In some cases, our results are more accurate than the depth maps obtained by laser scanner (*e.g*., the car window in the last row) since the laser can not handle transparent objects like glass window and returns the depth value behind the glass.

where ∇ is the vector differential operator, and T denotes the transpose of image gradient weighting.

Feature Reconstruction Loss Even though only photometric loss and smoothness loss are already sufficient to obtain a comparable result, this supervision is still not accurate. In practice, we adopt the feature reconstruction loss proposed by [51]. The homogeneous coordinates computed by Equation 1 can also be used to warp the last decoder layer's features which contain 16 channels in source view to the target view. Then L2 norm is computed with respect to the target view's features f_t and synthesized view's features $f_{s\to t}$. Again the per-pixel mimimum trick is also used here:

$$L_{fr} = \sum_{p} \min_{s} \|f_t(p) - f_{s\to t}(p)\|_2 \tag{8}$$

Then the total loss for training LR-Net is the combination of the above three losses:

$$L_{total} = \lambda_1 L_{ph} + \lambda_2 L_{smooth} + \lambda_3 L_{fr} \tag{9}$$

As for the training of HR-Net, we only use L_{ph} and L_{smooth} since the cost of computing Feature Reconstruction Loss on high-resolution feature maps is expensive.

4. Experiments

In this section, we evaluate our approach on KITTI dataset [9] and Make3D dataset [38, 39]. Camera pose evaluation, ablation study of each module and visualization of the results are also conducted. Finally the implementation details are clarified.

4.1. KITTI Depth

We evaluate our approach on KITTI 2015 [34] by the split of Eigen et al. [7]. As shown in Table 1, our results significantly outperform existing state-of-the-art methods. Qualitative comparisons are shown in Fig 2. Visually our depth predictions are sharper and unambiguous. Slim structures like traffic poles and trees are also handled well. In some cases, such as the glass windows on the cars, our model performs better than the measurement obtained by laser radar. In addition, the output sizes of previous methods are mostly 128×416 or 192×640, but our model's output size is 384×1248, which is close to KITTI's original size. The static frames are removed as in [54] during training since the supervisory signal comes from the disparity generated by the camera motion. Median scaling proposed by [7] is adopted to align the predictions with the ground-truth depth. All the settings during evaluation are the same as previous methods [54, 50].

4.2. Make3D

We directly evaluate our model trained by KITTI on Make3D dataset *without* any fine-tuning. We just resize the test images to 384×1248 resolution and feed them to the

Method	Dataset	Supervision	Error metric				Accuracy metric		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train set mean	-	-	0.361	4.826	8.102	0.377	0.638	0.804	0.894
Eigen et al. [7]	K	Depth	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Liu et al. [30]	K	Depth	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Godard et al. [11]	K	Stereo	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. [11]	K+CS	Stereo	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Kuznietsov et al. [22]	K	Depth+Stereo	0.113	0.741	4.621	0.189	0.862	0.960	0.986
DORN [8]	K	Depth	0.072	0.307	2.727	0.120	0.932	0.984	0.994
Yang et al. [48]	K+CS	Stereo	0.114	1.074	5.836	0.208	0.856	0.939	0.976
Casser et al. [4]	K	Instance Label	0.109	0.825	4.750	0.187	0.874	0.958	0.983
Zhou et al. [54]	K	-	0.183	1.595	6.709	0.270	0.734	0.902	0.959
GeoNet [50]	K	-	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DDVO [42]	K	-	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Godard et al.(V2) [10]	K	-	0.129	1.112	5.180	0.205	0.851	0.952	0.978
Yang et al. [48]	K	-	0.131	1.254	6.117	0.220	0.826	0.931	0.973
Ours	K	-	**0.121**	**0.837**	**4.945**	**0.197**	**0.853**	**0.955**	**0.982**

Table 1. Comparison to existing methods on KITTI 2015 [9] using the Eigen split [7]. Our method achieves state-of-the-art result.

Method	Supervision		Abs Rel	Sq Rel	RMSE
	Depth	Pose			
Train set mean			0.893	13.98	12.27
Karsch et al. [17]	✓		0.428	5.079	8.389
Liu et al. [30]	✓		0.475	6.562	10.05
Laina et al. [24]	✓		0.204	1.840	5.683
Godard et al. [11]		✓	0.544	10.94	11.76
Zhou et al. [54]			0.383	5.321	10.47
DDVO [42]			0.387	4.720	8.090
Godard et al.(V2) [10]			0.361	4.170	7.821
Ours			**0.318**	**2.288**	**6.669**

Table 2. Evaluation on Make3D [39] dataset. Our result is obtained by the model trained on KITTI without any fine-tuning.

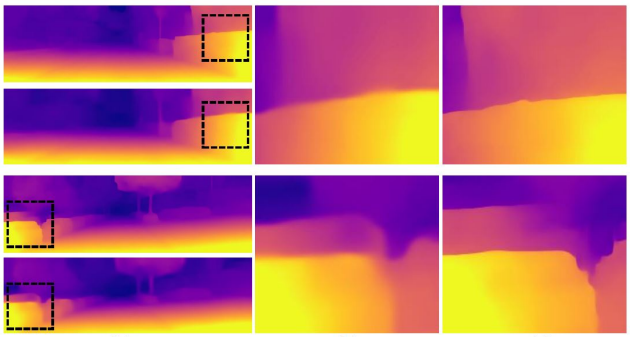

Figure 3. Comparison of the depth map at the object boundary. (a) Predicted depth maps with (top) and without HR-Net. (b) Zoomed patches without HR-Net. (c) Zoomed patches with HR-Net. Since HR-Net directly generates high-resolution depth map, the boundary is fairly sharp.

network. As shown in Table 2, our result outperforms existing state-of-the-art methods as well although it has not been trained on that dataset, which shows the generalization ability of our model.

4.3. KITTI Odometry

We have evaluated the performance of our method on KITTI odometry split. As same as previous methods, we use the 00-08 sequences for training and the 09-10 sequences for testing. Our PoseNet is the same as Zhou et al. [54]'s. As shown in Table. 8, the pose result of dual-network is better than LR-Net's while their PoseNets share the same architecture. The gain of the pose comes from more accurate depth prediction. "Training frames" means the number of input frames used by PoseNet during training and testing.

4.4. Ablation Study

Dual Networks As shown in Table 3, Figure 3 and Figure 5, both quantitative result and visual effect are improved noticeably with the help of HR-Net. The depth of traffic poles and trees are more accurate. Since the depth map is of high-resolution and need not to be upsampled, the border of objects is sharper than low-resolution result.

Distant Objects As mentioned before, the disparity of distant objects becomes more accurate since our HR-Net can fully explore the information of high-resolution images.

Method	Error metric				Accuracy metric		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline (LR-Net only)	0.132	0.929	5.208	0.209	0.833	0.946	0.979
LR-Net+HR-Net (w/o SA module)	0.123	0.881	5.016	0.198	0.851	0.955	0.982
LR-Net+HR-Net (w/ SA module)	0.121	0.837	4.945	0.197	0.853	0.955	0.982

Table 3. Evaluation of each component in our model on KITTI's eigen test split.

Method	Distance	Error metric				Accuracy metric		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
w/o HR-Net	$\leq 20m$	0.108	0.448	2.107	0.153	0.901	0.973	0.990
	$> 20m$	0.199	2.685	9.885	0.310	0.648	0.865	0.940
w/ HR-Net	$\leq 20m$	0.102	0.391	1.959	0.146	0.912	0.977	0.990
	$> 20m$	0.187	2.430	9.695	0.305	0.662	0.875	0.946

Table 4. Evaluation of the dual networks architecture on distant objects and nearby objects. The gain of distant objects (Abs Rel: 0.012) is higher than that of nearby objects (Abs Rel: 0.006).

	w/o HR-Net	w/ HR-Net
Score	20784.4	106987.2

Table 5. Evaluation of the depth maps' average sharpness on KITTI's eigen test split. The Tenengrad score coarsely represents the sharpness of an image. Higher score means sharper boundary.

Region	w/o SA	w/ SA
Road	0.085	0.080
Tree & grass	0.190	0.192

Table 6. Evaluation of SA module in regions with and without noticeable textures (i.e. tree & grass vs. road) respectively with Abs rel metric. Without SA means the global features are directly fed to the HR-Net. The gain mostly comes from low-texture regions (Abs rel: 0.005) compared with regions (Absrel: -0.002) with noticeable textures.

Here we respectively evaluate our result on distant objects and nearby objects. We consider the pixels with more than 20m ground-truth depth as distant objects and others as nearby objects. The quantitative result is shown in Table 4 and error map samples are shown in Figure 4. Both of them clearly show that our dual networks effectively reduce the error produced by distant objects.

Objects' boundary Qualitative comparison of objects' boundary is shown in Figure 3. To quantitatively evaluate the sharpness of the depth maps generated by HR-Net, we use Tenengrad measure [20] to coarsely compare the sharpness of the results with and without HR-Net (shown in Table 5). The definition of Tenengrad function is written as:

$$T = \sum_x \sum_y |G(x,y)|, (G(x,y) > t) \tag{10}$$

$$G(x,y) = \sqrt{G_x^2(x,y) + G_y^2(x,y)} \tag{11}$$

where T denotes Tenengrad gradient, G_x, G_y denote the horizontal and vertical gradients of depth map obtained by Sobel filters, t denotes the threshold and x, y denote pixel coordinates.

SA-Attention Module To verify SA module for low-texture region, we use segmentation model PSPNet [53] pretrained on Cityscapes to extract the semantic labels of KITTI's test split. Then we evaluate the performance of SA module in regions with and without noticeable textures (i.e. tree & grass vs. road) respectively as shown in Table 6. The SA module has much higher gain in low-texture regions compared with regions with noticeable textures.

Overhead Comparison Table 7 shows the results of training with different DepthNet encoders. Suppose the input size is 384×1248, our dual-network has more parameters and lower GFLOPs than ResNet18, which means more efficient. And it also performs better than other two backbones. The lower part also reveals that ImageNet pretraining is important.

4.5. Implementation Details

Our model is implemented with the Tensorflow [1] framework. Network architectures are shown in supplementary material. We first train the LR-Net for 40 epochs. Then we fix the parameters of LR-Net and continue to train the HR-Net for another 40 epochs. We use mini-batch size of 4 at both stages and optimize the network with Adam [19], where $\beta_1 = 0.9, \beta_2 = 0.999$. The settings of learning rate at both stages are the same as below:

$$lr = \begin{cases} 0.0002 & epochs \leq 10 \\ 0.0001 & 10 < epochs \leq 20 \\ 0.00005 & 20 < epochs \end{cases} \tag{12}$$

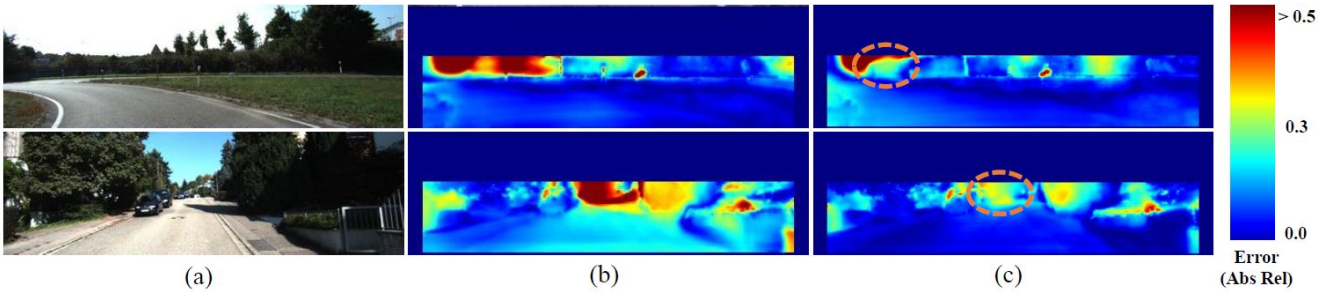

Figure 4. Evaluation of our dual networks' effect on distant objects. (a) Input images. (b) Error maps without HR-Net. (c) Error maps with HR-Net. The error induced by distant objects (orange circles) is improved by the dual networks visibly.

BackBones	Params(M)	GFLOPs	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
MobileNet V2 [37]	5.37	17.02	0.147	0.998	5.185	0.810	0.940	0.976
ResNet18	16.43	33.11	0.138	0.968	5.281	0.823	0.947	0.980
Dual-network(ResNet50)	34.16	25.80	0.121	0.837	4.945	0.853	0.955	0.982
w/ pt	-	-	0.121	0.837	4.945	0.853	0.955	0.982
w/o pt	-	-	0.135	0.973	5.235	0.823	0.947	0.980

Table 7. Comparison of our dual-network with other lightweight backbones when training with images of 384×1248 resolution. Params and GFLOPs denote the number of parameters and GFLOPs of the DepthNet (including encoder and decoder). w/ pt and w/o pt are the results of dual-network with and without ImageNet pretraining.

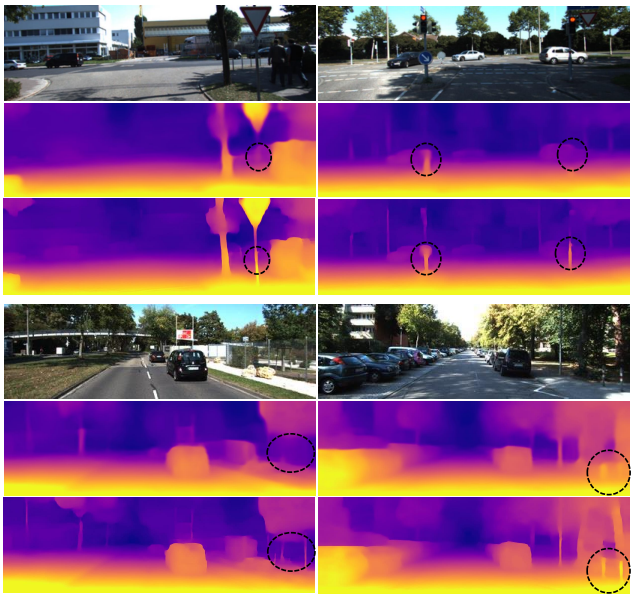

Figure 5. Comparison of slim objects' prediction. Top: original images. Middle: prediction without HR-Net. Bottom: prediction with HR-Net. The prediction of the traffic poles is improved significantly.

Zero padding before convolution is replaced by reflection padding as in [10]. All the predicted depth are normalized as in [42] in order to avoid the shrinking prediction during training. We observed that previous implementations generally down-sample the images beforehand for the convenience of subsequent training. Then data augmentations (e.g., random scaling and random cropping) are applied. However, down-sampling first then random cropping later will further lose the fine-grained details. We just simply rearrange the order, i.e., applying data augmentations first and down-sampling later. This trick improves the baseline's performance.

Method	Sequence 09	Sequence 10	#
ORB-Slam(full)	0.014 ± 0.008	0.012 ± 0.011	-
ORB-Slam(short)	0.064 ± 0.141	0.064 ± 0.130	-
Zhou et al. [54]	0.021 ± 0.017	0.020 ± 0.015	5
GeoNet [50]	$\mathbf{0.012 \pm 0.007}$	$\mathbf{0.012 \pm 0.009}$	5
DF-Net [56]	0.017 ± 0.007	0.015 ± 0.009	5
DDVO [42]	0.045 ± 0.108	0.033 ± 0.074	3
Ours(LR-Net)	0.017 ± 0.008	0.016 ± 0.009	3
Ours(LR-Net+HR-Net)	0.015 ± 0.007	0.015 ± 0.009	3

Table 8. Evaluation results of the average absolute trajectory error and standard deviation in meters. # denotes the number of training frames.

5. Conclusion

In this paper, we propose a dual networks architecture, which consists of LR-Net and HR-Net. This architecture is able to fully explore the fine-grained details contained in high-resolution images and directly generates high-resolution depth map. The proposed techniques in this paper can also be applied in other resolution-sensitive tasks like unsupervised flow learning, etc. However, there still exist some intractable problems in our approach like dynamic cars, pedestrians, bicycles and reflective objects, which are left to be addressed in the future work.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, 2016.

[2] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *CVPR*, 2018.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised learning of depth and ego-motion: A structured approach. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *arXiv preprint arXiv:1806.01260*, 2018.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] Felix Heide, Matthias B Hullin, James Gregson, and Wolfgang Heidrich. Low-budget transient imaging using photonic mixer devices. *ACM Transactions on Graphics (ToG)*, 2013.

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[17] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *PAMI*, 2014.

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Eric P Krotkov. *Active computer vision by cooperative focus and stereo*. Springer Science & Business Media, 2012.

[21] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. *arXiv preprint arXiv:1803.01599*, 2018.

[22] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, 2017.

[23] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *CVPR*, 2014.

[24] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV)*, 2016.

[25] Hongyang Li, Yu Liu, Wanli Ouyang, and Xiaogang Wang. Zoom out-and-in network with map attention decision for region proposal and object detection. *International Journal of Computer Vision*, 127:225–238, 2018.

[26] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *ICRA*, 2018.

[27] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

[28] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010.

[29] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *CVPR*, 2018.

[30] Miaomiao Liu, Mathieu Salzmann, and Xuming He. Discrete-continuous depth estimation from a single image. In *CVPR*, 2014.

[31] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018.

[32] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. Learning visual question answering by bootstrapping hard attention. In *ECCV*, 2018.

[33] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[34] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on ISA*, 2015.

[35] Sudeep Pillai, Rares Ambrus, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. *arXiv preprint arXiv:1810.01849*, 2018.

[36] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018.

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[38] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *NIPS*, 2006.

[39] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009.

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[42] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018.

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

[45] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017.

[46] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *ECCV*, 2018.

[47] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[48] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In *ECCV*, 2018.

[49] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *CVPR*, 2018.

[50] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.

[51] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

[52] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[53] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[54] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

[56] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018.