Single View Stereo Matching

单视角立体匹配：实际上是先预测右视图，再用立体匹配网络得到深度

Yue Luo[1] * Jimmy Ren[1] * Mude Lin[1] Jiahao Pang[1] Wenxiu Sun[1] Hongsheng Li[2] Liang Lin[1,3]

[1]SenseTime Research
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
[3]Sun Yat-sen University, China
[1]{luoyue,rensijie,linmude,pangjiahao,sunwenxiu,linliang}@sensetime.com
[2]hsli@ee.cuhk.edu.hk

Abstract

Previous monocular depth estimation methods take a single view and directly regress the expected results. Though recent advances are made by applying geometrically inspired loss functions during training, the inference procedure does not explicitly impose any geometrical constraint. Therefore these models purely rely on the quality of data and the effectiveness of learning to generalize. This either leads to suboptimal results or the demand of huge amount of expensive ground truth labelled data to generate reasonable results. In this paper, we show for the first time that the monocular depth estimation problem can be reformulated as two sub-problems, a view synthesis procedure followed by stereo matching, with two intriguing properties, namely i) geometrical constraints can be explicitly imposed during inference; ii) demand on labelled depth data can be greatly alleviated. We show that the whole pipeline can still be trained in an end-to-end fashion and this new formulation plays a critical role in advancing the performance. The resulting model outperforms all the previous monocular depth estimation methods as well as the stereo block matching method in the challenging KITTI dataset by only using a small number of real training data. The model also generalizes well to other monocular depth estimation benchmarks. We also discuss the implications and the advantages of solving monocular depth estimation using stereo methods.[1]

1. Introduction

Depth estimation is one of the fundamental problems in computer vision. It finds important applications in a large number of areas such as robotics, augmented reality, 3D reconstruction and self-driving car, etc. This problem is heavily studied in the literature and is mainly tackled with two types of technical methodologies namely active stereo vision such as structured light [33], time-of-flight [40], and passive stereo vision including stereo matching[17, 25], structure from motion [35], photometric stereo [5] and depth cue fusion [31], etc. Among passive stereo vision methods, stereo matching is arguably the most widely applicable technique because it is accurate and it poses little assumption to the sensors and the imaging procedure. Recent advances in this field show that the quality of stereo matching can be significantly improved by deep models trained with synthetic data and finetuned with limited amount real data [26, 28].

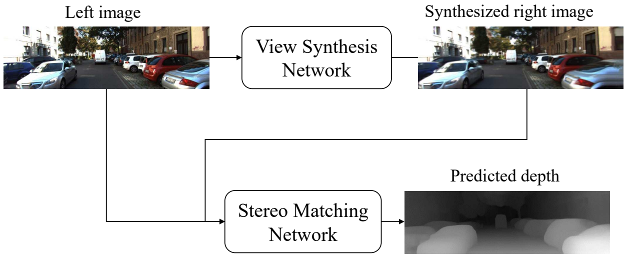

Figure 1: Pipeline of our approach on monocular depth estimation. We decompose the task into two parts: view synthesis and stereo matching. Both networks enforce the geometric reasoning capacity. With this new formulation, our approach is able to achieve state-of-the-art performance.

On the other hand, the applicability of monocular depth estimation is greatly limited by its accuracy though the single camera setting is much more preferred in practice in order to avoid calibration errors and synchronization problems occur to the stereo camera setting. Estimating depth from a single view is difficult because it is an ill-posed and geometrically ambiguous problem. Advancement of monocular depth estimation has recently been made by deep learning methods [4, 19, 20, 23] . However, comparing to the mentioned passive stereo vision methods which are

*Indicates equal contribution.

[1] Code is publicly available at https://github.com/lawy623/SVS.

grounded by geometric correctness, the formulation in the current state-of-the-art monocular method is problematic. The reasons are twofold. First, current deep learning approaches to this problem almost completely rely on the high-level semantic information and directly relate it to the absolute depth value. Because the operations in the network are general and do not have any prior knowledge on the function it needs to approximate, learning such semantic information is difficult even some special constraints are imposed in the loss function. Second, even the effective learning can be achieved, the relationship between scene understanding and depth needs to be established by a huge number of real data with ground truth depth. Such data is not only very expensive to obtain at scale, collecting high-quality dense labels is very difficult and time consuming if not entirely impossible. This significantly limits the potential of the current formulation.

In this paper, we take a novel perspective and show for the first time that monocular depth estimation problem can be formulated as a stereo matching problem in which the right view is automatically generated by a high-quality view synthesis network. The whole pipeline is shown in figure 1. The key insights here are that i) both view synthesis and stereo matching respect the underlying geometric principles; ii) both of them can be trained without using the expensive real depth data and thus generalize well; iii) the whole pipeline can be collectively trained in an end-to-end fashion that optimize the geometrically correct objectives. Our method shares a similar idea as revealed in the Spatial Transformation Network [12]. Although deep models can learn necessary transformations by themselves, it might be more beneficial for us to explicitly model such transformations. We discover that the resulting model is able to outperform all the previous methods in the challenging KITTI dataset [9] by only using a small number of real training data. The model also generalizes well to other monocular depth estimation datasets.

Our contributions can be summarized as follows.

- First, we discover that the monocular depth estimation problem can be effectively decoupled into two subproblems with geometrical soundness. It forms a new foundation in advancing the performance in this field.

- Second, we show that the whole pipeline can be trained end-to-end and it outperforms all the previous monocular methods by a large margin using a fraction of training data. Notably, this is the first monocular method to outperform the stereo blocking matching algorithm in terms of the overall accuracy.

2. Related Works

There exists a large body of literature on depth estimation from images, either using single view [30], stereo views [32], several overlapped images from different viewpoints [7], or temporal sequence [29]. For monocular depth estimation, Saxena *et al*. [30] propose one of the first supervised learning-based approaches to single image depth map prediction. They model depth prediction in a Markov random field and use multi-scale texture features that have been hand-crafted. Recently, deep learning has proven its ability in many computer vision tasks, including the single image depth estimation. Eigen *et al*. [4] propose the first CNN framework that predicts the depth in a coarse-to-fine manner. Laina *et al*. [19] employ a deeper ResNet [11] structure with an efficient up-sampling design and achieve a boosted performance. Liu *et al*. [23] also propose a deep structured learning approach that allows for training CNN features of unary and pairwise potentials in an end-to-end way. Chen *et al*. [1] provide a novel insight by incorporating pair-wise depth relation into CNN training. Compared with depth, these rankings on pixel level are much more easy to obtain. Further lines of research in supervised training of depth map prediction use the idea of depth transfer from example images [14, 15, 24], or combining semantic segmentation [3, 18, 21, 22, 36]. However, large amount of high-quality labels are in need to establish the transformation from image space to depth space. Such data are not easy to collect at scale in real life.

Recently, a small number of deep network based methods attempt to estimate depth in an unsupervised way. Garg *et al*. [8] first introduce the unsupervised method by only supervising on the image alignment loss. However, their loss is not fully differentiable so that they apply first Taylor expansion to linearize their loss for back-propagation. Godard *et al*. [10] also propose an unsupervised deep learning framework, and they employ a novel loss function to enforce consistency between the predicted depth maps from each camera view. Kuznietsov *et al*. [16] adopt a semi-supervised deep method to predict depths from single images. Sparse depth from LiDAR sensors is used for supervised learning, while a direct image alignment loss is integrated to produce photoconsistent dense depth maps in a stereo setup. Zhou *et al*. [38] jointly estimate depth and camera pose in an unsupervised manner.

Despite that those unsupervised methods reduce the demand of expensive depth ground truth, their mechanisms are still inherently problematic since they are attempting to regress a depth/disparity directly from a single image. The network architecture itself does not assume any geometric constraints and it acts like a black box. In our work, we propose a novel strategy to decompose this task into two separate procedures, namely synthesizing a corresponding right view followed by a stereo matching procedure. Such idea is similar to the Spatial Transformation Network [12], which learns a transformation within the network before conducting visual tasks like recognition.

To synthesize a novel view, DeepStereo [6] first proposes to render an unseen view by taking pixels from other views, and [39] predicts the appearance flow to reconstruct the target view. The Deep3D network of Xie *et al*. [37] addresses the problem of generating the corresponding right view from an input left image. Their method produces a distribution over all the possible disparities for each pixel, which is used to generate the right image.

Conducting stereo matching on the original left input and the synthetic right view is now a 1D matching problem. The vast majority of works on stereo matching focus on learning a matching function that searches the corresponding pixels on two images [17, 25]. Mayer *et al*. [26] introduce their fully convolutional DispNet to directly regress the disparity from the stereo pair. Later, Pang *et al*. [28] adopt a multi-scale residual network developed from DispNet and obtain refined results. These methods still rely on large amount labelled disparity as ground truth. Instead of using data from the real world, training on synthetic data [26] becomes a more feasible solution to these approaches.

3. Analysis and our approach

In this section, we demonstrate how we decompose the task of monocular depth estimation into two separate tasks. And we illustrate our model design for view synthesis and stereo matching separately.

3.1. Analysis of the whole pipeline

In our pipeline, we decompose the task of monocular depth estimation into two tasks, namely view synthesis and stereo matching. The whole pipeline is shown in figure 2. By tackling this problem using two separate steps, we find that both procedures obey primary geometric principles and they can be trained without expensive data supply. After that, these networks can be collectively trained in an end-to-end manner. We further hypothesize that, when the whole pipeline is trained end-to-end, both components will not degrade their capacity of constraining geometric correctness, and the performance of the whole pipeline will be promoted thanks to joint training. Therefore, we are desired to choose both methods that can explicitly model the geometric transformation in the network design.

The first stage is view synthesis. For a stereo pair, binocular views are rendered by well synchronized and calibrated cameras, resulting in the strong correspondence between pixels in the horizontal direction. Unlike previous warp-based methods that generally require an accurate estimation of the underlying geometry, Deep3D [37] proposes a new probabilistic scheme to transfer pixels from the original image. By this mean, it directly formulates the transformation from left image to right image using a differentiable selection layer. We adopt its design and develop our view synthesis network based on it. Other reconstruction plans [8, 10, 16] are also viable alternatives, but the choice of the specific view synthesis method is independent of the main insight of the paper.

After generating a high-quality novel view, our stereo matching network transforms the high-level scene understanding problem into a 1D matching problem, which results in less computational complexity. In order to better utilize the geometric relation between two views, we take the idea of 1D correlation employed in DispNetC[26]. We further adopt the DispFullNet structure mentioned in [28] to achieve full resolution prediction.

3.2. View synthesis network

Our view synthesis network is shown in the upper part of figure 2. We develop this network based on Deep3D [37] model. Here we briefly introduce the structure of it. At the very beginning, an input left image I_l is processed by a baseline network. We then upsample the features from different intermediate levels to the same resolution, in order to incorporate low-level features into final use. Those features are then summed up to further produce a probabilistic disparity map. After completing a selection operation, pixels on original I_l can be selectively mixed up to form a new pixel on the right image.

The operation of selection is the core component in this network. This module is also illustrated in figure 2. Denote I_l as the input left image, previous Depth Image-Based Rendering (DIBR) techniques choose to directly warp the left image based on estimated disparity into a corresponding right image. Suppose D is the predicted disparity aligned with the left image, the procedure can be formulated as

$$\widetilde{I}_r(i, j - D_{i,j}) = I_l(i, j), \qquad (i, j) \in \Omega_l, \tag{1}$$

where Ω_l is the image space of I_l and i, j refer to the row and column on I_l respectively. Though this function captures the geometric correspondence between images in a stereo setup, it requires an accurate disparity map to reconstruct the right view. At the same time, the function is not fully differentiable with respect to D which limits the opportunity of training by a deep neural network. The selection module, instead, formulates the reconstruction as a process of probabilistic summation. Denote $D \in \mathbb{R}^{W \times H \times C}$ as the probabilistic disparity result, where W and H are the width and height of left image and C indicates the number of possible disparity shifts, the reconstruction can then be formulated as

$$\widetilde{I}_r = \sum_d I_l^{(d)} D^d. \tag{2}$$

Here, $I_l^{(d)}(i, j) = I_l(i, j + d)$ is the shifted left image whose stride is predetermined by possible disparity values d. This operation sums up the stacked shifted input by learned weights and ensures the differentiability of the whole system.

To supervise the reconstruction quality, we do not propose any special loss function. We find that a simple L1 loss supervising on the reconstructed appearance is sufficient for the task of view synthesis:

$$L_{view} = \frac{1}{N} \sum_{i,j} \left| \widetilde{I}_r(i, j) - I_r(i, j) \right| \tag{3}$$

训练一个右视图合成网络，在用左右视图concat进入自编码网络得到深度图的视差图

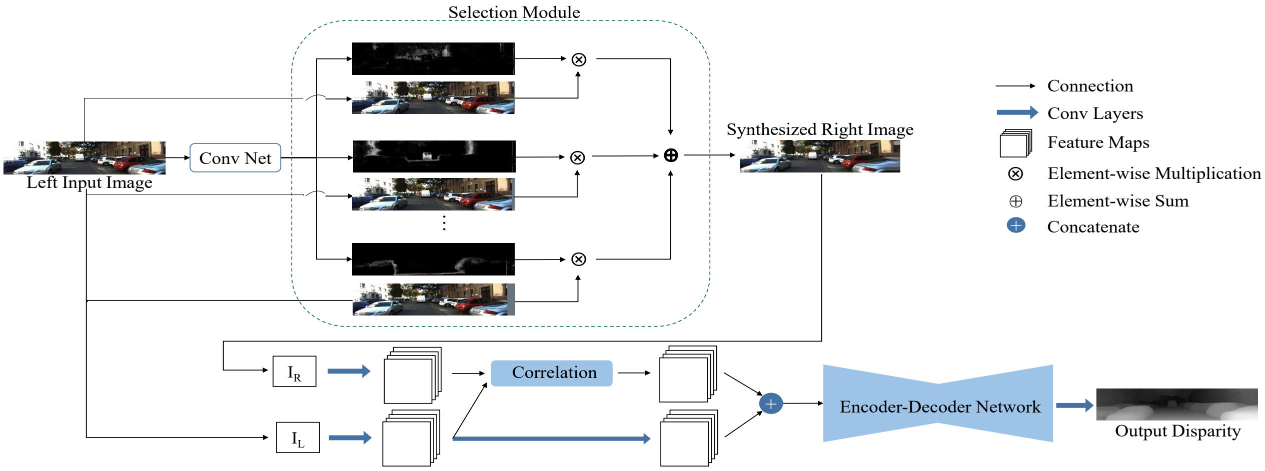

Figure 2: Details of our single view stereo matching network. Upper part is the view synthesis network. The input image is first processed by a CNN. It results in probabilistic disparity maps that help to reconstruct a synthetic right view by selectively taking pixels from nearby locations on the original left image. A stereo matching network, which is shown on the lower part of the figure, then takes both the original left image and synthetic right image to calculate an accurate disparity, which can be transformed into a corresponding depth map given the camera settings.

3.3. Stereo matching network

There exists a large body of literature tackling the problem of stereo matching. Recent advancements are achieved by deep learning models. Not only because deep networks help to effectively find out similar pixel pairs, research also show that these networks can be trained on a large amount of synthetic data and they can still generalize well on real images [26]. In our pipeline, we select the state-of-the-art DispNetC [26] structure as the desired network for the stereo matching task. We further follow the modifications made in [28] to adopt a DipFulNet structure for full-resolution output. The structure of this method can be seen in the lower part of figure 2. We briefly illustrate the method here, and the detailed settings can be found in their papers.

After processed by several convolutional operations, 1D correlation will be calculated based on resulted features. This correlation layer is found very useful in the stereo matching problem since it explicitly encodes the geometric relationship into the model design, and the horizontal correlation is indeed an effective cue for finding the most similar pairs. The features will be further concatenated with higher-level features from the left image I_l. An encoder-decoder network further processes the concatenated features and produces disparity at different scales. These intermediate and final results will be supervised by ground truth disparity using L1 loss.

3.4. End-to-end training of the whole pipeline

These two networks can be combined for joint training once being trained to obtain the ability of geometric reasoning for the task of view synthesis and stereo matching separately. End-to-end training of the whole pipeline can thus be performed to enforce the collaboration of these two sub-networks.

4. Experiments

In this section, we present our experiments and results. Our method achieves state-of-the-art monocular depth estimation result on the widely used KITTI dataset [9]. We discover and show the key insights of this method and prove the correctness of our methodology. We also make the first attempt to run our single view approach on the challenging KITTI Stereo 2015 benchmark [27].

4.1. Dataset and Evaluation Metrics

We evaluate our approach on the publicly available KITTI benchmark [9]. In order to fairly compare with other methods on monocular depth estimation, we use the raw sequences of KITTI and employ the split scheme proposed by Eigen *et al.* [4]. This split results in a test set with 697 images. Remaining data is used for training and validation. Overall we have 22600 stereo pairs for training our view synthesis network. Except for stereo image pairs, the dataset also contains sparse 3D laser measurements taken from a Velodyne laser sensor. They can be projected onto image space and served as the depth labels. Parameters of the stereo setup and the camera intrinsics are also provided, therefore we can transfer depth into disparity as ground truth during end-to-end training and recover the depth from disparity during inference.

Evaluation metrics are as follows and they indicate the error and performance on predicted monocular depth.

$$\text{ARD} = \frac{1}{N}\sum_{i=1}^{N}\left|Dep_i - Dep_i^{g.t.}\right|/Dep_i^{g.t.}$$

$$\text{SRD} = \frac{1}{N}\sum_{i=1}^{N}\||Dep_i - Dep_i^{g.t.}\|^2/Dep_i^{g.t.}$$

RMSE = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|Dep_i - Dep_i^{g.t.}\|^2}$

RMSE(log) = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|log(Dep_i) - log(Dep_i)^{g.t.}\|^2}$

Accuracy = % $Dep_i : max(\frac{Dep_i}{Dep_i^{g.t.}}, \frac{Dep_i^{g.t.}}{Dep_i}) = \delta < thr$

Here N is the number of pixels that are not empty on the depth ground truth.

To compare with other works in a consistent manner, we only evaluate on a cropped region proposed by Eigen *et al*. [4]. Also, previous methods restrict the depth distance in different ranges for evaluation, we provide our result using both the cap of 0-80m (following Eigen *et al*. [4]) and 1-50m (following Garg *et al*. [8]). This requires to discard the pixels on which the depth is outside the proposed range.

4.2. Implementation Details

The training of the model is divided into two stages. First we train the two networks used for different purposes separately. In the second stage, we combine the two parts and further finetune the whole pipeline in an end-to-end fashion. The training is conducted using caffe framework [13].

In the first stage, networks are trained separately. For the training of view synthesis network, 22600 stereo pairs from KITTI are taken into use. We select VGG16 as the baseline network and initialize the weights of it using the model pre-trained from ImageNet [34]. All other weights are initialized following the same scheme in [37]. Compared with original deep3D model [37], we make some modifications to make it suitable for view synthesis task on KITTI dataset. First, the size of input is larger and is selected to be 640 × 192. It retains the aspect ratio of original KITTI images. Second, one more convolution layer is employed before deconvolution at each branch. Third, since the disparity ranges differently in KITTI and 3D movie dataset, we change the possible disparity range. A 65-channel probabilistic map representing possible disparity from 0 to 64 now becomes the final features. Last, to accommodate larger inputs and the deeper network structure, we decrease the batch size as 2, and we remove the origin BatchNorm layers in the deep3D model. The model is trained for 200K iterations with initial learning rate equals to 0.0002. For the training of DispFullNet used for stereo matching, we follow the training scheme specified in [28]. The model is trained mainly on the synthetic FlyingThings3D dataset [26] and optional finetuned on the KITTI stereo training set [27]. This KITTI stereo training set contains 200 stereo pairs with relatively high-quality disparity labels, and it has not overlap with the test data from KITTI Eigen test set. The detailed settings can be found in Pang's paper *et al*. [28].

In the second stage, two networks with pre-trained weights are now trained end-to-end. A small number of data from the KITTI Eigen training set with ground truth disparity labels will be taken to finetune the whole pipeline. Since the input to the stereo matching network has a larger dimension, upsample is performed inside the network to enlarge the synthetic right view resulted from the first stage.

Data augmentation is optionally done in both stages. The input will be randomly resized to a dimension slightly greater than the desired input size. And then it will be cropped into the desired size and fed into the network. The color intensity will also multiply a factor between 0.8 to 1.2.

4.3. Depth Estimation by Stereo Matching method

First, the evaluation of depth estimation of the stereo matching network given perfect right images is presented. The result is shown in the Table 1, denoted as "Stereo_gt_right". The stereo matching network clearly outperforms state-of-the-art methods for single image depth estimation, even the stereo matching network is mainly trained on rendered dataset [26].

The intuition here is that predicting depth from stereo images has a much higher accuracy than predicting depth by any of the previous monocular depth methods. This means we are able to achieve much higher performance if we can provide a sophisticated view synthesis module.

4.4. Comparisions with state-of-the-art methods

Next, results on the KITTI Eigen split dataset are reported when right images are predicted by our view synthesis network. Results are compared to six recent baseline methods as showed in Table 1, [4, 24] are supervised methods, [16] is a semi-supervised method, and [10, 38, 8] are unsupervised methods. Our proposed method is also a semi-supervised method.

Result without end-to-end finetuning: After the training of both networks converged, we directly feed the right image synthesized by the view synthesis network to the stereo matching network to predict the depth for the given left images. The result is reported in Table 1.

As one can see, even without finetuning the whole network in KITTI dataset, our method performs better than the unsupervised method [10], and gets comparable performance with the state-of-the-art semi-supervised method [16]. The performance achieved by our method demonstrates that decoupling the problem of monocular depth estimation into two separate sub-problems is simple yet effective by explicitly enforcing geometrics constraints, which is critical for estimating depth from images.

Result with end-to-end finetuning: We further finetune the whole system with a small amount of training data from KITTI Eigen split training set, *i.e*. 700 training samples. The left, right images and the depth images are used as training samples to our proposed method.

The results are reported in Table 1, as one can see, our method outperforms all compared methods, with ARD metric reduced by **17.5%** compared with Godard *et al*. [10] and **16.8%** compared with Kuznietsov *et al*. [16] at the cap of 80 m. Our proposed method performs the best for almost all metrics. It shows that end-to-end training further optimizes the collaboration of these two sub-networks and it

Approach	cap	ARD	SRD	RMSE	RMSE(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
		lower is better				higher is better		
Stereo_gt_right	$0-80$ m	0.062	0.424	3.677	0.164	0.939	0.968	0.981
Eigen *et al.* [4]	$0-80$ m	0.215	1.515	7.156	0.270	0.692	0.899	0.967
Liu *et al.* [24]	$0-80$ m	0.217	1.841	6.986	0.289	0.647	0.882	0.961
Zhou *et al.* [38]	$0-80$ m	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Godard *et al.* [10]	$0-80$ m	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Kuznietsov *et al.* [16]	$0-80$ m	0.113	0.741	4.621	0.189	0.862	0.960	**0.986**
Ours, w/o end-to-end finetuned	$0-80$ m	0.102	0.700	4.681	0.200	0.872	0.954	0.978
Ours	$0-80$ m	**0.094**	**0.626**	**4.252**	**0.177**	**0.891**	**0.965**	0.984
Stereo_gt_right	$1-50$ m	0.058	0.316	2.675	0.152	0.947	0.971	0.983
Zhou *et al.* [38]	$1-50$ m	0.190	1.436	4.975	0.258	0.735	0.915	0.968
Garg *et al.* [8]	$1-50$ m	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard *et al.* [10]	$1-50$ m	0.108	0.657	3.729	0.194	0.873	0.954	0.979
Kuznietsov *et al.* [16]	$1-50$ m	0.108	0.595	3.518	0.179	0.875	0.964	**0.988**
Ours, w/o end-to-end finetuned	$1-50$ m	0.097	0.539	3.503	0.187	0.885	0.960	0.981
Ours	$1-50$ m	**0.090**	**0.499**	**3.266**	**0.167**	**0.902**	**0.968**	0.986

Table 1: Quantitative results of our method and approaches reported in the literature on the test set of the KITTI Raw dataset used by Eigen *et al.* [4] for different caps on ground-truth and/or predicted depth. Best results are shown in bold. Our proposed method achieves improvement over all compared state-of-the-art approaches.

Approach	FT VSN	FT SMN	cap	ARD	SRD	RMSE	RMSE(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
				lower is better				higher is better		
Finetune-0	✗	✗	$0-80$ m	0.102	0.700	4.681	0.200	0.872	0.954	0.978
Fintuned_synthesis_200	✓	✗	$0-80$ m	0.100	0.682	4.515	0.195	0.879	0.957	0.979
Fintuned_synthesis_700	✓	✗	$0-80$ m	0.099	0.672	4.593	0.194	0.879	0.957	0.979
Finetuned_stereo_gt_right_0	✗	✗	$0-80$ m	0.062	0.424	3.677	0.164	0.939	0.968	0.981
Finetuned_stereo_gt_right_200	✗	✓	$0-80$ m	0.065	0.452	3.844	0.168	0.933	0.967	0.981
Finetuned_stereo_gt_right_700	✗	✓	$0-80$ m	0.053	0.382	3.400	0.144	0.947	0.975	0.986
Finetune-200	✓	✓	$0-80$ m	0.100	0.670	4.437	0.192	0.882	0.958	0.979
Finetune-500	✓	✓	$0-80$ m	0.094	0.635	4.275	0.179	0.889	0.964	0.984
Finetune-700	✓	✓	$0-80$ m	0.094	0.626	4.252	0.177	0.891	0.965	0.984

Table 2: Quantitative results of different variants of our proposed method on the test set of the KITTI Raw dataset used by Eigen *et al.* [4] at the cap of 80m. "FT VSN" denotes whether the view synthesis network has been finetuned in an end-to-end fashion, while "FT SMN" denotes whether the stereo matching network has been finetuned in an end-to-end fashion. Top three rows: comparisons of different view synthesis network settings. Middle three rows: comparisons of different stereo matching network settings. Bottom three rows: empirical comparisons by different number of training samples. The number in the method names means the number of samples to finetune the network.

leads to the state-of-the-art result. Qualitative comparisons are shown in Figure 3. Our proposed method also achieves much more visually accurate estimations than the compared methods.

4.5. Analyzing the function of two sub-networks after end-to-end training

In this section, we analyze the function of two sub-networks after end-to-end training. If the end-to-end training breaks the origin functionality of the two sub-networks but the overall performance increases, the whole network would be overfitted to the KITTI dataset, which will make it hard to generalize to other datasets or scenes. To examine the function of two sub-networks, we conduct the following two groups of experiments.

Analyzing function of view synthesis sub-network: We replaced the stereo matching sub-network in the finetuned network with the one before finetuneing. Since pre-trained stereo matching sub-network is only pre-trained to complete the stereo matching task using real left-right pairs, if after replacing, the whole network could still get good performance in the task of single image depth estimation, the origin functionality of the view synthesis network after the finetuning process could still be retained.

The results are reported in top three rows of Table 2, denoted as "Finetuned_synthesis_K", where K represents the number of training samples. As one can see from Table 2, the results by "Finetuned_synthesis_K" outperform the method without finetune. From another perspective, the average PSNR between synthesized views and ground truth views in test set increases from 21.29dB to 21.32dB after finetuning. The preservation of functionality may be due to the reason that during the finetuning process, the stereo matching sub-network acts as another loss to bet-

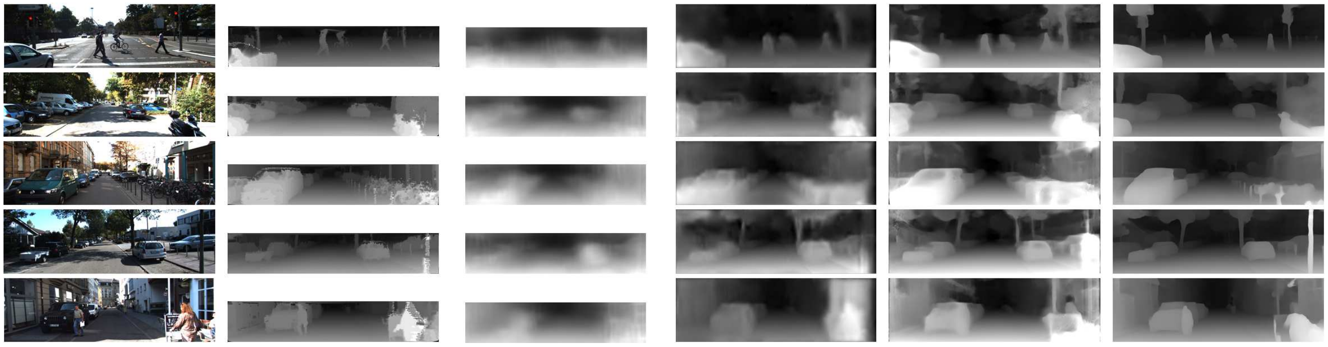
(a) Input (b) Ground-truth (c) Eigen *et al*. [4] (d) Garg *et al*. [8] (e) Godard *et al*. [10] (f) Ours

Figure 3: Qualitative results on the KITTI Eigen test set. Sparse ground-truth labels have been interpolated for visualization. Note that the prediction of our method can better separate the background and foreground or different entities close to each other. Also, our results are crisper and neater. In addition, we are doing better on the objects such as trees, poles, traffic sign and pedestrians, whose depth are generally hard to be inferred accurately.

ter constrain the view synthesis sub-network to generate geometric-reasonable right images.

Experiment	cap	ARD	SRD	RMSE	RMSE(log)
Our_Best	0 − 80 m	0.094	0.626	4.252	0.177
Kuznietsov [16]	0 − 80 m	0.113	0.741	4.621	0.189
Prob_Disp	0 − 80 m	0.212	2.075	6.314	0.294
NoKitti200_BF	0 − 80 m	0.119	0.969	5.079	0.207
NoKitti200_AF	0 − 80 m	0.101	0.673	4.425	0.176

Table 3: Additional experimental results. Upper part is our best result and the previous state-of-the-art result. Middle part shows the result directly calculated from the probabilistic disparity map obtained in our view synthesis network. Lower part shows the results before and after finetuning without 200 high-quality KITTI labels.

Analyzing function of stereo matching sub-network: In order to validate the function of stereo matching sub-network after end-to-end training, we test the stereo matching performance of the finetuned stereo matching sub-network by providing the true left and right image as inputs to predict the depth.

The results are provided in the middle three rows of Table 2, denoted as "Finetuned_stereo_gt_right_K". As shown in Table 2, "Finetuned_stereo_gt_right_200" performs slightly worse than "Finetuned_stereo_gt_right_0", this may be due to the reason that the finetuning process has forced the stereo matching sub-network to better fit on the imperfect synthesized right images. However, "Finetuned_stereo_gt_right_700" outperforms the pretrained stereo matching sub-network. The high performance of stereo matching results clearly demonstrates the stereo matching network still maintains its functionality after end-to-end finetuned.

Combining the above two experiment groups, we could conclude that after end-to-end training, the two submodules collaborate more effectively while preserving their individual functionalities. This may imply that our proposed method could generalized well to other datasets. Some qualitative results on Cityscape dataset [2] and Make3D dataset [31] are shown in Figure 5, which are estimated by our method finetuned in KITTI dataset. The results demonstrate the generalization ability of our proposed method on unseen scenes.

4.6. Primitive disparity obtained in the view synthesis network

Our view synthesis network produces a primitive disparity in order to do the rendering. The middle part in table 3 shows the estimation accuracy calculated from this probabilistic disparity map. We can see the result is much inferior to the final result of our proposed method. It shows our approach indeed makes a great improvement over the primitive disparity.

4.7. Analyzing the effect of training sample number

To study the effectiveness of our proposed method, we also evaluate our proposed method finetuned by different numbers of samples, *i.e*., 0, 200, 500, 700, named as "Finetune-K". Note that, when K equals to 0, finetuning is not performed on the whole network.

The results are reported in the bottom three rows of Table 2. As one can see from the results, more end-to-end finetuning samples could achieve higher performance, and our proposed method could outperform previous state-of-the-art methods by a clear margin only using 700 samples to finetune the whole network.

4.8. Use of 200 high-quality KITTI labels

As described before, we use 200 high-quality KITTI labels to optionally finetune the stereo matching network. In the lower part of table 3, we present the result without these labels before and after finetune(_BF&_AF). We can see that without seeing any real disparity from KITTI, our method already gets promising results. After finetuning without those high-quality labels, our method still beats the current

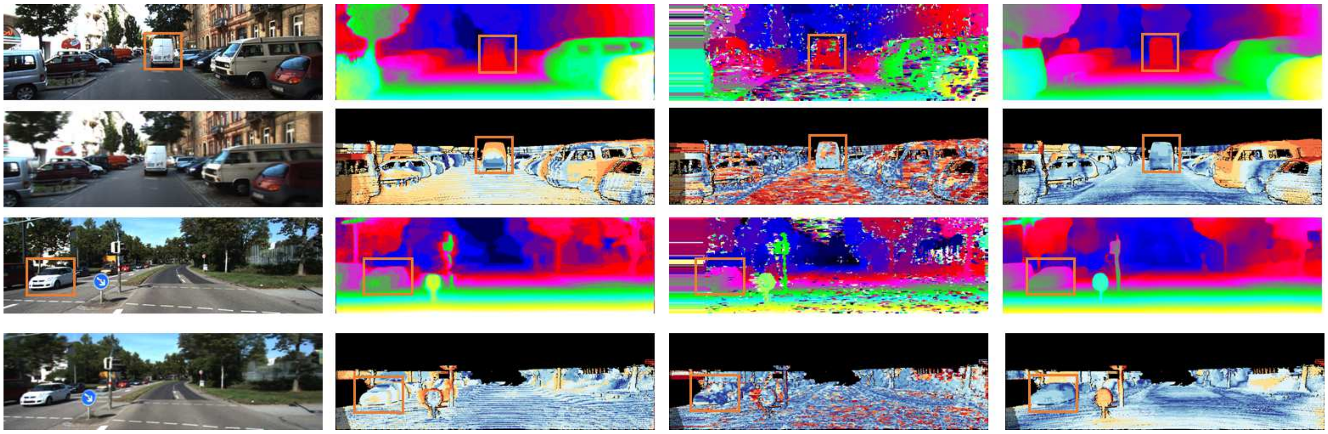
(a) Input image (b) Godard *et al*. [10] (c) OCV-BM (d) Ours

Figure 4: Empirical study on the qualitative comparisons on KITTI 2015 Stereo test set. The figures from left to right correspond to the input left images, estimated disparity maps or error maps by Godard *et al*. [10], block matching, and our method respectively. And the second and fourth rows are the error maps while the estimated disparity maps are plotted above each error maps, the synthesized right views are also presented in the first column. The error map uses the log-color scale described in [27], depicting correct estimates in blue and wrong estimates in red color tones. Best view in color.

Method	D1-bg	D1-fg	D1-all
Godard *et al*. [10]	27.00	28.24	27.21
OCV-BM	**24.29**	30.13	25.27
Ours	25.18	**20.77**	**24.44**

Table 4: Quantitative results on the test set of the KITTI 2015 Stereo Benchmark [27]. Best results are shown in bold. The number is the percentage of erroneous pixels, and a pixel is considered to be correctly estimated if the disparity is within 3px compared to the ground-truth disparity. Our method has already surpassed the stereo matching method, *i.e*. Block Matching method.

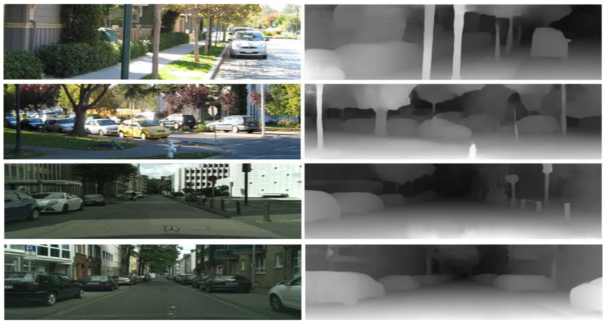
Figure 5: Qualitative results on Make3D dataset [31] (top two rows) and Cityscapes dataset [2] (bottom two rows).

state-of-the-art method. These high-quality labels, in fact, increase the capacity of the model to a certain extent, but without them, our method still makes an improvement under the same condition.

4.9. Comparison with stereo matching method

In this section, the comparisons with the proposed approach for depth estimation from single images and stereo matching method from stereo images are presented. The results are summarized in Table 4. As one can see, our method is the first single image depth estimation approach that surpasses the traditional stereo matching method, *i.e*. block matching method denoted as "OCV-BM" in the table. Exemplar visual results are shown in Fig. 4. Because the block matching method directly using low-level image feature to search the matched pixels in the left and right images, the disparity maps predicted by the block matching method are usually noised, which greatly degrades its performance, but the results are still geometrically correct. The geometric reasoning capacity is built in our network and high-level image feature is processed in the deep learning network, these two reasons enable our method to outperform the stereo matching method. Due to the miss of explicit geometric constraints in Godard *et al*. [10], its method gets sub-optimal results. Better performance of our method can be seen from the box regions in the figure.

5. Conclusion

In this work, we propose a novel perspective to tackle the problem of monocular depth estimation. We show for the first time that this problem can be decomposed into two problems, namely a view synthesis problem and a stereo matching problem. We explicitly encode the geometric transformation within both networks to better tackle the problems individually. Collectively training the whole pipeline results in an overall boost and we prove that both networks are able to preserve their original functionality after end-to-end training. Without using a large amount of expensive ground truth labels, we outperform all previous methods on a monocular depth estimation benchmark. Remarkably, we are the first to outperform the stereo blocking matching algorithm on a stereo matching benchmark using a monocular method.

References

[1] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *NIPS*, 2016. 2

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 7, 8

[3] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 2

[4] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 1, 2, 4, 5, 6, 7

[5] C. H. Esteban, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *TPAMI*, 30(3):548–554, 2008. 1

[6] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. *arXiv preprint arXiv:1506.06825*, 2015. 2

[7] Y. Furukawa, C. Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[8] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 2, 3, 5, 6, 7

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 4

[10] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. 2017. 2, 3, 5, 6, 7, 8

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[12] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In *NIPS*. 2015. 2

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *arxiv. 1408.5093*, 2014. 5

[14] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *ECCV*, 2012. 2

[15] J. Konrad, M. Wang, and P. Ishwar. 2d-to-3d image conversion by learning depth from examples. In *CVPRW*, 2012. 2

[16] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. 2017. 2, 3, 5, 6, 7

[17] L. Ladický, C. Häne, and M. Pollefeys. Learning the matching function. *arXiv preprint arXiv:1502.00652*, 2015. 1, 3

[18] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *CVPR*, 2014. 2

[19] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 1, 2

[20] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, 2015. 1

[21] C. Li, A. Kowdle, A. Saxena, and T. Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *NIPS*, 2010. 2

[22] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010. 2

[23] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015. 1, 2

[24] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *CVPR*, 2014. 2, 5, 6

[25] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016. 1, 3

[26] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 1, 3, 4, 5

[27] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 4, 5, 8

[28] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. *arXiv preprint arXiv:1708.09204*, 2017. 1, 3, 4, 5

[29] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun. Dense monocular depth estimation in complex dynamic scenes. In *CVPR*, 2016. 2

[30] A. Saxena, S. H. Chung, and A. Y. Ng. 3-d depth reconstruction from a single still image. *IJCV*, 76(1):53–69, 2008. 2

[31] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 30(5):824–840, 2009. 1, 7, 8

[32] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 2

[33] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003. 1

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[35] P. F. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *ECCV*, 1996. 1

[36] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, 2015. 2

[37] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*, 2016. 3, 5

[38] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 5, 6

[39] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *ECCV*, 2016. 2

[40] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *TPAMI*, 33(7):1400–1414, 2011. 1