Unsupervised Monocular Depth Estimation via Recursive Stereo Distillation

Xinchen Ye, Xin Fan, Mingliang Zhang, Rui Xu, and Wei Zhong

***Abstract*—Existing unsupervised monocular depth estimation methods resort to stereo image pairs instead of ground-truth depth maps as supervision to predict scene depth. Constrained by the type of monocular input in testing phase, they fail to fully exploit the stereo information through the network during training, leading to the unsatisfactory performance of depth estimation. Therefore, we propose a novel architecture which consists of a monocular network (Mono-Net) that infers depth maps from monocular inputs, and a stereo network (Stereo-Net) that further excavates the stereo information by taking stereo pairs as input. During training, the sophisticated Stereo-Net guides the learning of Mono-Net and devotes to enhance the performance of Mono-Net without changing its network structure and increasing its computational burden. Thus, monocular depth estimation with superior performance and fast runtime can be achieved in testing phase by only using the lightweight Mono-Net. For the proposed framework, our core idea lies in: 1) how to design the Stereo-Net so that it can accurately estimate depth maps by fully exploiting the stereo information; 2) how to use the sophisticated Stereo-Net to improve the performance of Mono-Net. To this end, we propose a recursive estimation and refinement strategy for Stereo-Net to boost its performance of depth estimation. Meanwhile, a multi-space knowledge distillation scheme is designed to help Mono-Net amalgamate the knowledge and master the expertise from Stereo-Net in a multi-scale fashion. Experiments demonstrate that our method achieves the superior performance of monocular depth estimation in comparison with other state-of-the-art methods.**

***Index Terms*—Unsupervised, Depth Estimation, Monocular, Stereo Distillation, Recursive**

I. Introduction

WITH the rapid development of multimedia technology, depth information has been served as a basic element in many applications, such as augmented reality, 3D movies, multimedia content understanding, view synthesis and 3D reconstruction [1], [2], [3], [4], [5]. An effective way to estimate depth information is to directly predict it from a single RGB image. However, it is an ill-posed problem because of the color-depth inconsistency. Recently, the employment of deep learning has brought significant advancements in monocular depth estimation. Although some supervised learning methods [6], [7] depending on color-depth training pairs have been widely used, the performance is difficult to be further improved due to the limited single-view inputs and the unavailable dense ground-truth depth for supervised training.

To overcome the above limitation, the unsupervised learning methods [8], [9], [10], [11], [12] have focused on predicting depth map from the models that are trained on stereo image pairs, without requiring any ground-truth depth information. However, constrained by the type of network input (i.e., monocular images) in testing phase, these methods fail to fully exploit the stereo training data through the network, since they can only take a single image as input, i.e., left image, but leverage the right image to assist the supervised training based on epipolar geometry. Otherwise, if the stereo image pairs are taken as input, the network will become a stereo network for stereo matching, and the testing process is unfeasible when the test example is merely the monocular image. Thus, how to take advantage of the stereo information to improve the performance of monocular depth estimation and simultaneously allow the network to test on monocular inputs, motivates us to design a novel network architecture that can solve both the above problems at the same time.

Therefore, we propose an unsupervised learning architecture to realize monocular depth estimation via recursive stereo distillation. As shown in Fig. 1, the whole framework consists of a monocular network (Mono-Net) and a sophisticated stereo network (Stereo-Net). Mono-Net aims to infer a coarse depth map from the single left image. Then, the right image together with the obtained coarse depth map is used to generate the synthetic left image. Due to the limitation of input type, Mono-Net can only extract the features from monocular images but cannot fully use the stereo correlation information from stereo image pairs through the network. To further improve the performance of Mono-Net, we propose to cascade a Stereo-Net at the end of the Mono-Net. Stereo-Net takes the stereo image pairs, the coarse depth map, and the error map between synthetic and real left images as input to regress a more accurate depth map. Note that, Stereo-Net can be regarded as a role of teacher in the whole framework, and the final performance of Mono-Net depends entirely on the 'teacher's own ability' and 'teaching ability' of Stereo-Net. Thus, the key insight of our method lies in the following two aspects, i.e., 1) Stereo-Net should be carefully designed so that it can accurately estimate depth maps by fully exploiting more useful information; 2) The sophisticated Stereo-Net should be properly used to guide the learning of Mono-Net without changing its network structure and increasing its computational burden.

Based on the above analysis, to improve the 'teacher's own ability' of Stereo-Net, we propose a recursive estimation

This work was supported by National Natural Science Foundation of China (NSFC) under Grant (No. 61702078, 61733002, 61772106 and 61906029).

X. Ye, X. Fan, R. Xu, and W. Zhong are with DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Liaoning, and Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China. (Corresponding author: Xin Fan, e-mail: xin.fan@ieee.org).

M. Zhang is with the School of Mathematics and Statistics, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China.

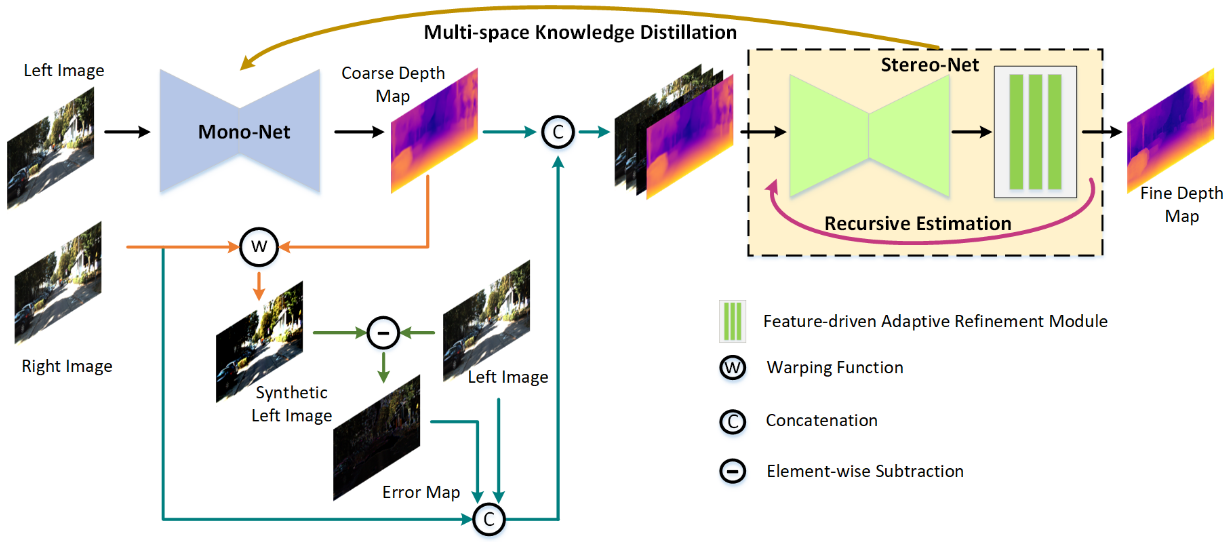


Fig. 1. Network overview. It includes a Mono-Net $\mathcal{M}$ and a Stereo-Net $\mathcal{S}$, where $\mathcal{M}$ is a lightweight network that takes a single image as input, while $\mathcal{S}$ takes stereo images pair as input. $\mathcal{S}$ contains a recursive estimation strategy and a feature-driven adaptive refinement module to further improve the accuracy of depth estimation. The multi-space knowledge distillation scheme is designed to distill knowledge from $\mathcal{S}$ and squeeze into $\mathcal{M}$.

strategy that takes the output from a previous iteration as input and iteratively estimates the depth map by reusing a single Stereo-Net with shared weights. Simultaneously, we introduce a feature-driven adaptive refinement module at the end of Stereo-Net to further alleviate the issues of outliers and blurred depth boundaries caused by the common regression problem. Besides, to improve the ‘teaching ability’ of Stereo-Net, we design a multi-space knowledge distillation scheme to help Mono-Net amalgamate the knowledge and master the expertise from Stereo-Net, and finally improve the performance of Mono-Net. The proposed scheme can distill the useful information of Stereo-Net from the aspects of output space, feature space, and long-range dependencies in a multi-scale fashion. In testing phase, the model can be flexibly chosen by using Mono-Net independently or ‘Mono-Net + Stereo-Net’ together according to the different types of test input (monocular image or stereo image pairs, respectively), with regard to the balance between prediction accuracy and computational complexity. Extensive experiments on public KITTI, Cityscapes and Make3D dataset exhibit our superior performance compared with other state-of-the-art methods.

Our main contributions are summarized as follows. 1) This paper presents a novel framework that can break the limitations of simultaneously exploiting the stereo features and keeping the testing process feasible and efficient with monocular image. We develop a two-stage network architecture, which consists of an Mono-Net (student) that infers depth maps from monocular inputs, and an Stereo-Net (teacher) that further excavates the stereo information by taking stereo pairs as input. During training, Stereo-Net guides the learning of Mono-Net and devotes to enhance the performance of Mono-Net without changing its network structure and increasing its computational burden. Thus, monocular depth estimation with superior performance and fast runtime can be achieved in testing phase by only using the lightweight Mono-Net. 2) We have successfully solved two important problems raised from the proposed network architecture, i.e., how to improve the ‘teacher’s own ability’ (design of Stereo-Net) and its ‘teaching ability’ (design of the guided learning for Mono-Net), both of which can boost the performance of monocular depth estimation. For the ‘teacher’s own ability’, we design a light-weight network architecture of depth estimation for Stereo-Net, i.e., a pipeline of recursive estimation and adaptive refinement to provide more accurate depth inference without increasing the complexity. For the ‘teaching ability’, we propose a novel multi-space knowledge distillation scheme to help Mono-Net acquire knowledge from Stereo-Net through comprehensive consideration of both the pixel-wise difference and nonlocal correlation in the multi-scale feature alignment. Extensive experiments have shown that both the ‘teacher’s own ability’ and ‘teaching ability’ can bring significant improvement for the final performance.

II. Related Work

Monocular Depth Estimation. The CNN-based methods for monocular depth estimation mainly include supervised learning methods [6], [7], [13], [14] and unsupervised learning methods [8], [9], [10], [15], [16]. The supervised methods need large quantities of ground-truth depth data for training, which is undesirable in practical applications. To avoid this issue, the unsupervised methods are proposed to reformulate the depth estimation problem into the image reconstruction problem without any ground-truth depth data during training. They utilize a differentiable warping function [17] to obtain a synthetic image and then build the reconstruction (or photometric) loss [8] to measure the difference between the synthetic and real images. Wong *et al.* [18] tried to learn a sufficient feature representation by designing a two-branch decoder. Chen *et al.* [12] modeled the geometric structure of objects by integrating both depth and semantic information with shared decoder. Zhao *et al.* [19] attempted to transfer knowledge from synthetic dataset with ground-truth depth via domain adaptation technique for better monocular depth estimation in real dataset. However, these unsupervised monocular depth estimation methods have a common problem, i.e., they use one of the stereo image pairs as a supervisory signal, which fails to fully exploit the stereo information through the network, leading to the unsatisfactory performance in testing phase.

Recursive Strategy. The learning-based methods using a single network usually suffer from the degraded performance. To this end, some methods [20], [21], [22], [23] proposed to stack multiple networks such that the later network refines the output from the previous one in a coarse-to-fine manner. Ilg *et al.* [22] introduced a network cascade that consists of variants of FlowNet for optical flow estimation. Ummenhofer *et al.* [23] proposed an architecture which is composed of multiple stacked networks to jointly estimate depth, surface normal and optical flow. However, a drawback of the cascaded way is the increasing number of parameters and complex training process. In contrast, taking inspiration from classical energy-based methods [24], [25], [26] which iteratively estimate depth map by solving the optimization model, some methods [27], [28] proposed to refine the results recursively. Carreira *et al.* [27] proposed a self-correcting model which recursively transmits initial solution by feeding back error predictions for human pose estimation. Han *et al.* [28] explored a dual-state recursive network for image super-resolution, where the recursive signals are exchanged between low-resolution and high-resolution states. Motivated by these methods, we propose a recursive estimation strategy that constructs a more compact and effective network for depth estimation.

Distillation. The distillation technique has been vastly studied in recent years [29], [30], [31], [32], [33]. Hinton *et al.* [29] proposed to encourage the transfer of the knowledge from the cumbersome model to the lightweight model. Rusu *et al.* [30] proposed a policy distillation method to extract the policy of a reinforcement learning agent for deep Q-networks. Anil *et al.* [31] developed a cost-effective online distillation to speed up training process. Li *et al.* [34] observed that the true labels are noisy and show multimode characteristics, and then they developed a framework to learn from noisy labels. Radosavovic *et al.* [35] proposed a data distillation that collects predictions from multiple transformations of unlabeled data to automatically produce new training annotations. Phuong *et al.* [32] designed a multi-exit architecture based on the distillation to allow early exits to mimic later exits. Tung *et al.* [33] proposed a similarity-preserving distillation by computing pairwise similarity matrices from the output activation maps for both teacher and student networks. Inspired by these distillation techniques, we propose a multi-space knowledge distillation scheme by distilling useful information from Stereo-Net so as to help Mono-Net to infer an accurate depth map.

III. Proposed Method

The basic idea of the unsupervised learning methods is to predict depth map by using the synthetic view as the supervisory signal instead of requiring any ground-truth depth data. Given the training data of calibrated stereo image pairs $\{I_l^i, I_r^i\}_{i=1}^N$, where N is the number of training data. The left image I_l is first fed into the network to get the corresponding disparity map d_l. If the camera parameters are given (baseline b and focal length f), the disparity map d_l can be immediately converted into the depth map by the function $D_l = bf/d_l$ [1]. Then, the synthetic left image $\hat{I}_l$ is obtained by using the warping operation $f_w(\cdot)$:

$$\hat{I}_l = f_w(I_r, d_l),$$

where $f_w(\cdot)$ is fully differentiable to facilitate back-propagation in network training [17]. Finally, the reconstruction loss between the synthetic images $\{\hat{I}_l^i\}_{i=1}^N$ and the real images $\{I_l^i\}_{i=1}^N$ is established as follows:

$$L_{rec}^l = \frac{1}{N}\sum_{i=1}^{N} ||I_l^i - \hat{I}_l^i||, \tag{1}$$

In practice, we can also feed the right image into the network to obtain the synthetic right one based on the right disparity map d_r and the real left image I_l, and thus get the loss L_{rec}^r for the right view. For simplicity, we only present the reconstruction loss for left images.

A. Network Architecture

As shown in Fig. 1, our framework mainly contains a Mono-Net and a Stereo-Net. Mono-Net is a lightweight network that takes the left image I_l as input and generates the corresponding coarse depth map d_l^c. Then, the synthetic left image $\hat{I}_l^c$ is obtained through warping operation $f_w(\cdot)$ by using the coarse depth map d_l^c and the right image I_r, and subsequently the error map e_l^c between the synthetic and real left images is computed. In contrast, Stereo-Net is a sophisticated network that takes the stereo image pairs $\{I_l, I_r\}$, the coarse depth map d_l^c, and the error map between synthetic and real left image e_l^c as input, and recursively generates the fine depth map d_l^f. Through the multi-space knowledge distillation, Stereo-Net offers rich supervisory information to guide the learning of Mono-Net.

As shown in Fig. 2, both the Mono-Net $\mathcal{M}$ and Stereo-Net $\mathcal{S}$ are fully convolutional networks based on the encoder-decoder structure [36] with skip connection [37]. To further improve the 'teacher's own ability' of Stereo-Net $\mathcal{S}$, we integrate the atrous spatial pyramid pooling (ASPP) [38] into the intermediate layer to enlarge receptive field and extract multi-scale features, and propose to cascade a feature-driven adaptive refinement module at the end of the encoder-decoder structure in a recursive manner. On the other hand, to improve the 'teaching ability' of $\mathcal{S}$, we also design a novel knowledge distillation scheme to transfer knowledge from Stereo-Net $\mathcal{S}$ to Mono-Net $\mathcal{M}$ in the multi-scale and multi-space fashion.

[1] Here, we do not strictly distinguish the concept between the disparity map and depth map.

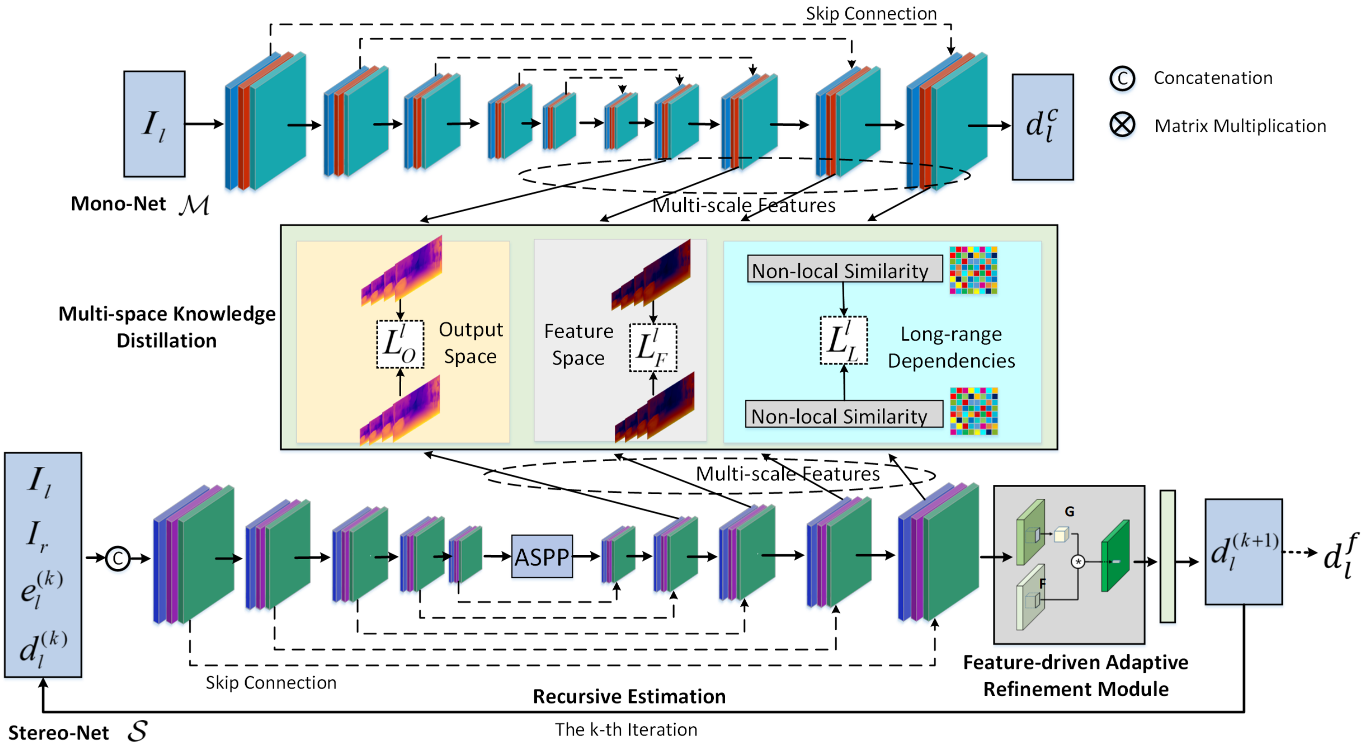


Fig. 2. Structures of Mono-Net $\mathcal{M}$, Stereo-Net $\mathcal{S}$, and the multi-space knowledge distillation scheme. We propose to cascade the feature-driven adaptive refinement module with $\mathcal{S}$ and update network weights in a recursive manner. The multi-space knowledge distillation scheme is designed to transfer knowledge from $\mathcal{S}$ to $\mathcal{M}$ in the aspects of output space, feature space and long-range dependencies based on multi-scale feature extraction.

B. Recursive Estimation and Refinement

The depth map predicted from CNN usually produces undesired artifacts and blurry results. The reason is that the network has stride convolutions and simple upsampling operations, leading to the loss of spatial details. Besides, typical regression models only output the mean values of possible depth values without the variances, which further degrades the depth map especially on depth boundaries.

To cope with this problem, we propose a recursive estimation and refinement strategy (shown in Fig.2). Unlike the cascaded network where each module has its own parameters, the proposed recursive network removes the repetitive architecture and iteratively updates the output by reusing the single Stereo-Net with shared weights. In particular, assuming k is the number of iteration, the proposed recursive optimization strategy is expressed as

$$d_l^{(k)} = \mathcal{S}(concat(I_l, I_r, e_l^{(k-1)}, d_l^{(k-1)})),\ k = 1, \dots, K, \quad (2)$$

where $d_l^{(k)}$ is the output depth map of $\mathcal{S}$ at k-th iteration, and $d_l^{(0)} = d_l^c$, the k-th error map $e_l^{(k)} = I_l - f_w(I_r, d_l^{(k)})$ and $e_l^{(0)} = e_l^c$. Compared with the cascaded network, the recursive estimation strategy can dramatically decrease the parameters and simplify the training procedure, simultaneously improve the performance of Stereo-Net. As shown in Fig.3, the recursive Stereo-Net is equivalent to unrolling several repetitive networks with shared weights to gradually refine the output depth map. Independent of the number of iterations, the memory usage of Stereo-Net is fixed, and equals to the capacity of just one single encoder-decoder model.

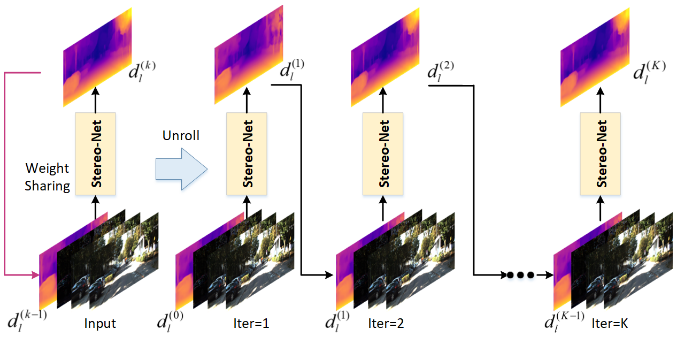


Fig. 3. The proposed recursive strategy for $\mathcal{S}$. This strategy indicated by the pink arrow can iteratively update output of $\mathcal{S}$ with weight sharing, which is equivalent to unrolling several repetitive networks.

The reconstruction loss for Stereo-Net $\mathcal{S}$ is reformulated as

$$L_{rec_s}^l = \frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} ||I_l^i - \hat{I}_l^{i,(k)}||, \quad (3)$$

where $\hat{I}_l^{i,(k)}$ is the synthetic left image for i-th sample at k-th iteration in $\mathcal{S}$, which is generated through the warping based on the depth map $d_l^{(k)}$.

In addition to the recursive estimation strategy, we also introduce a feature-driven adaptive refinement module [39] to further improve the accuracy of depth estimation at each iteration. Assuming that $F(x, y, c)$ is a feature at position

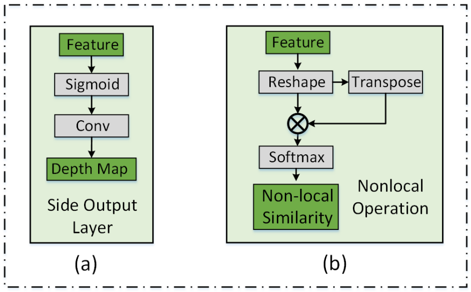


Fig. 4. The detailed structure of (a) side output layer and (b) long-range (non-local) operation.

(x, y) of channel c, $\mathcal{N}(x, y)$ is a neighborhood containing $p \times p$ pixels centered at (x, y) and $G_{(u,v)\in\mathcal{N}(x,y)}(u, v, c)$ represents the adaptive filter with the size of $p \times p$ applied on the position (x, y, c), in which the filter weights are adaptive to the features from Stereo-Net. The refined feature map $\widetilde{F}$ is expressed as

$$\begin{aligned}\widetilde{F}(x, y, c) &= G_{(u,v)\in\mathcal{N}(x,y)}(u, v, c) * F_{(u,v)\in\mathcal{N}(x,y)}(u, v, c), \\ &= \frac{\sum_{(u,v)\in\mathcal{N}(x,y)} \omega(u, v, c) F(u, v, c)}{\sum_{(u,v)\in\mathcal{N}(x,y)} \omega(u, v, c)}\end{aligned} \tag{4}$$

where $*$ denotes convolution operator, $\omega(u, v, c)$ is the weight in filter G, which measures the feature similarity between the positions (u, v) and (x, y) at channel c. We adopt the bilateral filter function to define the filter G, i.e.,

$$\omega(u,v,c) = \exp\left(-\frac{(x-u)^2+(y-v)^2}{2\sigma_1^2} - \frac{||F(x,y,c)-F(u,v,c)||^2}{2\sigma_2^2}\right) \tag{5}$$

where σ_1, σ_2 are smoothing hyper-parameters. The above module is end-to-end trainable and alleviates the issue of outliers and blurred depth boundaries.

After adopting the recursive estimation strategy and refinement module, $\mathcal{S}$ produces the fine depth map d_l^f, which is more accurate and capable of providing more plausible synthetic left image $\hat{I}_l^f$ as the new supervisory signal. Finally, we leverage the knowledge distillation technique [29] to encourage the coarse depth map d_l^c to imitate the fine depth map d_l^f, which will be introduced in the following.

C. Multi-Space Knowledge Distillation

From the perspective of knowledge distillation [29], Mono-Net can be treated as the student network while Stereo-Net as the teacher network. Our goal is to encourage the output from Mono-Net to approach the one from Stereo-Net by means of the knowledge distillation technique. As shown in Fig. 2, the proposed multi-space distillation method distills and exploits useful information in a multi-scale fashion from three aspects, i.e., output space, feature space and long-range dependencies.

Output space distillation. Multi-scale features in the decoder part are fed into our designed side output layer (Fig. 4(a)) to generate the depth map on each scale. Thus, the distillation loss for output space is designed to measure the difference between the coarse depth map $\{d_{l,s}^c\}_{s=1}^S$ and the fine depth map $\{d_{l,s}^f\}_{s=1}^S$ derived from multiple side output layers:

$$L_O^l = \frac{1}{NS}\sum_{i=1}^{N}\sum_{s=1}^{S} ||d_{l,s}^{c,i} - \mathcal{O}(d_{l,s}^{f,i})||, \tag{6}$$

where $d_{l,s}^{c,i}$ and $d_{l,s}^{f,i}$ are the coarse depth map and fine depth map at scale s from $\mathcal{M}$ and $\mathcal{S}$, respectively. $\mathcal{O}(\cdot)$ is the stop-gradient operation[2]. When computing the forward pass of the algorithm, $\mathcal{O}(\cdot)$ is the identity operation. For the backward pass, it becomes a null gradient [11].

Feature space distillation. Unlike the distillation in output space, the feature space distillation transfers the knowledge of feature representations from Stereo-Net $\mathcal{S}$ to Mono-Net $\mathcal{M}$. Its loss encourages $\mathcal{M}$ and $\mathcal{S}$ to have the similar perceptual information in feature space, which is defined as follows:

$$L_F^l = \frac{1}{NS}\sum_{i=1}^{N}\sum_{s=1}^{S} ||F_{l,s}^{\mathcal{M},i} - \mathcal{O}(F_{l,s}^{\mathcal{S},i})||, \tag{7}$$

where $F_{l,s}^{\mathcal{M},i}$ and $F_{l,s}^{\mathcal{S},i}$ are the decoder features for i-th sample at scale s in $\mathcal{M}$ and $\mathcal{S}$, respectively.

Long-range (non-local) dependencies distillation. Through our observation, pixels with similar appearances have more chances of belonging to the same object and often have close depth values. Long-range dependencies (LRD) between neighboring pixels can provide the information of nonlocal correlation, which is essential for depth estimation [40]. Therefore, we introduce the non-local operation [40] to capture long-range dependencies by computing pairwise similarity between any two positions. Assuming the dimension of feature F is $h \times w \times c$, the reshape function $\{\mathbb{R} : h \times w \times c \rightarrow hw \times c\}$ recasts F as $\mathbb{R}(F)$ with the dimension of $hw \times c$. The non-local similarity matrix M is defined as

$$M = \mathbb{S}(\mathbb{R}(F) \otimes \mathbb{R}^T(F)), \tag{8}$$

where $\mathbb{S}$ is the softmax operation, $\otimes$ is the matrix multiplication and T is the transpose operator. Fig. 4(b) shows the detailed structure of low-range (nonlocal) operation. The distillation loss for guiding Mono-Net $\mathcal{M}$ to learn the long-range dependencies from Stereo-Net $\mathcal{S}$ can be formulated as follow:

$$L_L^l = \frac{1}{NS}\sum_{i=1}^{N}\sum_{s=1}^{S} ||M_{l,s}^{\mathcal{M},i} - \mathcal{O}(M_{l,s}^{\mathcal{S},i})||. \tag{9}$$

The final loss $L_{distill}^l$ for multi-space knowledge distillation is expressed as

$$L_{distill}^l = L_O^l + \rho_1 L_F^l + \rho_2 L_L^l, \tag{10}$$

where ρ_1 and ρ_2 is the adjustment parameter.

Our multi-space knowledge distillation loss has some advantages: 1) It provides a set of supervisory signals for the whole framework. The loss does not depend on any ground-truth data, which can be used as an unsupervised way. 2) It is an asymmetric loss, which encourages the learning flow from

[2] https://www.tensorflow.org/api_docs/python/tf/stop_gradient.

Mono-Net to Stereo-Net. 3) When distilling the knowledge from Stereo-Net to Mono-Net, we just need to add this loss to the overall training loss and thus can improve the performance of Mono-Net without changing its network structure and complexity, which is very easy to be implemented.

D. Total Loss

The overall training loss includes three parts: loss for Mono-Net, loss for Stereo-Net, and distillation loss, i.e.,

$$L = L_M + L_S + \alpha_0 L_{distill}, \tag{11}$$

where α_0 is the adjustment parameter. Note that L can be expressed as $L = L^l + L^r$ for both left and right views. For simplicity, we only give the loss function about the left image. The loss of Mono-Net L_M is defined as:

$$L_M^l = \alpha_1 L_{photo}^l + \alpha_2 L_{smooth}^l, \tag{12}$$

where α_1, α_2 are the weighting parameters, L_{photo} and L_{smooth} are the photometric and smoothness losses, respectively.

Photometric loss. Following [9], we adopt a combination of the reconstruction loss and the structural similarity (SSIM) [41] to measure the difference between real and synthetic images more accurately:

$$L_{photo}^l = \gamma \frac{1}{N} \sum_{i=1}^{N} \frac{1 - SSIM(I_l^i - \hat{I}_l^i)}{2} + (1-\gamma) L_{rec}^l, \tag{13}$$

where γ is the adjustment parameter.

Smoothness loss. Since discontinuity of depth usually appears where strong image gradients are presented, we introduce a second-order edge-aware smoothness loss to enforce discontinuity and local smoothness within a depth map:

$$L_{smooth}^l = \frac{1}{N} \sum_{i=1}^{N} ||\nabla_x^2 d_l^{c,i}|| e^{-\|\nabla_x^2 I_l^i\|} + ||\nabla_y^2 d_l^{c,i}|| e^{-\|\nabla_y^2 I_l^i\|}, \tag{14}$$

where ∇ is the differential operator. Note that, the Stereo-Net loss L_S has a similar form with the Mono-Net loss L_M, just replacing L_{rec}^l with $L_{rec_s}^l$ in L_{photo}, and $d_l^{c,i}$ with $d_l^{f,i}$ in L_{smooth}, respectively.

IV. Discussion

Note that, Pilzer *et al.* [11] and our method both use the distillation mechanism motivated by Hinton *et al.* [29] to transfer knowledge from teacher to student network. However, there are some obvious differences between these two methods:

1) Both methods utilize the errors between the synthesized and real views as auxiliary to help improve the performance of teacher network, but are different approaches. [11] exploits the stereo information by means of the cycle-(in)consistency between the monocular input and corresponding virtual views during training. In contrast, ours (Stereo-Net) takes stereo image pairs, coarse depth maps and error maps between synthesized and real views as input, and fully exploits stereo information through the network to effectively extract stereo features. Meanwhile, we design the recursive estimation and adaptive refinement strategy to improve the performance and reduce the complexity of our teacher network. Therefore, our framework is more accurate and lightweight than [11] that uses two bulky networks (backward, inconsistency-aware) to realize its functionality.

2) We design a more effective and comprehensive distillation scheme to improve the 'teaching ability' of teacher network. [11] only considers the distillation in the final depth output and multi-scale feature space, leading to inadequate 'teaching ability' and limited performance improvement. In contrast, we consider the distillation additionally by side-outputting the depth maps from each scale and transfer the multi-scale depth output information to our student network. Besides, we introduce the distillation of LRD in our scheme. LRD is to compute the pair-wise similarity between any two positions within a feature map, and taking LRD into consideration contributes to transfer richer information in the distillation. In the ablation study of V-B, we have verified the effectiveness of each component for the proposed multi-space knowledge distillation scheme.

3) The mode of test input is different between [11] and our method. In the testing phase, when there are only monocular input images available, the method of [11] can keep all the networks (student and teacher) unaltered, while our method needs to discard the Stereo-Net and operate only with the Mono-Net. When there are stereo image pairs available in testing phase, the stereo input is unfeasible for [11] and our method allows stereo input mode by choosing the whole network ('Mono-Net + Stereo-Net') to achieve more accurate performance for monocular depth estimation. However, considering the accuracy-efficiency trade-offs in practice, student network is the final choice for both methods, and the results in V-A demonstrate that our method for student network outperforms [11] both in numerical and visual experiments.

V. Experiments

Training Dataset. We evaluate the proposed method on KITTI [46], Cityscapes [47] and Make3D [24] datasets. KITTI [46] contains sparse 3D laser measurements taken from a Velodyne laser sensor for outdoor scenes. The splitting modes for KITTI dataset include Eigen split [6] and KITTI split [9]. The Eigen split [6] has 22600 stereo image pairs for training and 697 stereo image pairs for testing, while the KITTI split [9] has 29000 stereo image pairs for training and 200 pairs for testing. Cityscapes [47] is collected from a moving vehicle using stereo camera, and we select 22973 stereo pairs as training dataset. Make3D [24] consists of 400 images as training set and 134 images as testing set. Since Make3D only includes RGB-D pairs without stereo images, it cannot be used for training, and thus we only test the generalization on Make3D through the well-trained KITTI model.

Training Details. For Mono-Net and Stereo-Net, we adopt the ResNet-50 [48] as the backbone in the encoder and flip the encoder as the decoder by replacing the downsampling layer with deconvolution layers. The feature-driven adaptive refinement module has five successive adaptive filters with the

TABLE I

RESULTS ON KITTI USING THE EIGEN SPLIT [6]. THE LIGHT GREY, DARK GREY AND WHITE BACKGROUNDS IN THE TABLE MEAN THAT THE TRAINING MANNERS ARE SUPERVISED, SEMI-SUPERVISED AND UNSUPERVISED RESPECTIVELY. FOR THE SUPERVISED AND SEMI-SUPERVISED SETTINGS, THE BEST RESULTS ARE IN ITALICS AND BOLD-FACE WITH UNDERLINE RESPECTIVELY. FOR THE UNSUPERVISED SETTING, THE BEST RESULTS ARE MARKED AS BOLD-FACE.

Method	Dataset	Error Metric (lower is better)				Accuracy Metric (higher is better)		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [6]	K	0.203	1.548	6.307	0.246	0.702	0.890	0.958
Liu *et al.* [7]	K	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Kundu *et al.* [13]	K	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Xu *et al.* [14]	K	*0.132*	*0.911*	-	*0.162*	*0.804*	*0.945*	*0.981*
Luo *et al.* [42]	K	**0.102**	**0.700**	**4.681**	0.200	**0.872**	**0.954**	0.978
Tosi *et al.* [43]	K	0.111	0.867	4.714	**0.199**	0.864	**0.954**	**0.979**
Godard *et al.* [9]	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhan *et al.* [10]	K	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Zhao *et al.* [19]	K	0.158	1.151	5.285	0.238	0.811	0.934	0.970
Chen *et al.* [12]	K	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Pilzer *et al.* [11]	K	0.142	1.230	5.785	0.239	0.795	0.924	0.968
Wong *et al.* [18]	K	0.133	1.126	5.515	0.231	0.826	0.934	0.969
Puscas *et al.* [44]	K	0.135	1.181	5.582	0.235	0.828	0.933	0.967
Godard *et al.* [45]	K	0.130	1.144	5.485	0.232	0.831	0.932	0.968
Ours	K	**0.105**	**0.842**	**4.810**	**0.196**	**0.861**	**0.947**	**0.978**
Godard *et al.* [9]	CS+K	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Wong *et al.* [18]	CS+K	0.118	0.996	5.134	0.215	0.849	0.945	0.975
Ours	CS+K	**0.104**	**0.815**	**4.616**	**0.193**	**0.865**	**0.951**	**0.979**

size of 3×3, followed by a side output layer to generate the fine depth map. In the training procedure, we first train Mono-Net and then fix Mono-Net to train Stereo-Net. Subsequently, the Mono-Net and Stereo-Net are jointly fine-tuned. Finally, we add the multi-space knowledge distillation scheme to train additional several epochs to improve the performance of Mono-Net. We set the batch size as 4 and adopt the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-4}$. For KITTI and Cityscapes dataset, the initial learning rates are 10^{-4} and 10^{-5} respectively and are downgraded by half at epoch 12, 18, 24, and 30. For the adjustment parameters, we set $\alpha_0 = 1$, $\alpha_1 = 1$, $\alpha_2 = 0.1$ and $\rho_1 = 0.01$, $\rho_2 = 0.85$. We set $\gamma = 0.85$, scale level $S = 4$ (without the minimum scale) and smoothing hyper-parameters $\sigma_1 = \sigma_2 = 10$. The iteration number k for Stereo-Net is 2 and the atrous rates for ASPP [38] are 1, 6, 8, 12. Note that, for the long-range dependencies distillation, we only compute the non-local similarity matrix M at the lower two scales, i.e. $s = 3$ and $s = 4$, since the computation will be dramatically increased with the increase of dimensions. The input resolution is 512×256. The whole framework is implemented by Tensorflow framework with GTX 1080Ti GPU acceleration. We adopt the measures used in [6] for quantitative evaluation.

A. Performance Comparison

Objective comparison. Table I lists the comparison results with other state-of-the-art methods. Note that, all the compared methods take single images as input in testing phase. Training settings can be classified into two parts, i.e., 1) directly training on KITTI dataset using stereo pairs (denoted as K), and 2) first pretraining on Cityscapes dataset and then fine-tuning on KITTI dataset (CS+K).

Compared with the supervised methods (light grey region), the proposed method 'Ours' (Mono-Net) achieves surprisingly comparable performance and obtains better numerical results on most metrics except for *RMSE log* and $\delta < 1.25^3$. Especially, 'Ours' exceeds the methods [6], [7], [13] in all metrics with a large margin. Note that [42], [43] have exploited stereo information and depth labels during training, thus we categorize them into semi-supervised learning methods (dark grey region). Compared with [42], [43], 'Ours' still maintains comparable performance in terms of both error and accuracy metrics. For the unsupervised setting (white region), 'Ours' outperforms other methods in all metrics. Note that both methods ([45] and 'Ours') are trained from scratch without any pretrained model for fair comparison. Our method achieves better numerical results in all metrics. The second best unsupervised method [12] leverages the extra semantic information as supervision to assist depth estimation, which is still inferior to our method.

For the training setting CS+K, we choose the methods that are also evaluated on CS+K to make the comparison. First, for all the methods and all the metrics, training on CS+K can further improve the performance of depth estimation compared with that on K alone. Besides, 'Ours' exceeds the two state-of-the-art methods under all the metrics, which demonstrate the effectiveness of our method.

Subjective comparison. Fig. 5 shows the corresponding visual results. The ground-truth depth map is interpolated from sparse measurements for visualization purpose. In Fig. 5(c), the state-of-the-art supervised method [14] exhibits comparable qualitative results with 'Ours', but it is still inferior to 'Ours', e.g. the second column picture, and suffers from blurred boundaries in some regions, e.g. the tram and car in the third and fourth column picture. Compared with the unsupervised methods, it is shown that the methods [9], [18] produce black holes for the glass areas in the third column, and the methods [9] fail to restore the reasonable depth values for the texture-less areas in the fourth column. The method in [11] obtains blurred depth maps and loses some depth information for the truck in

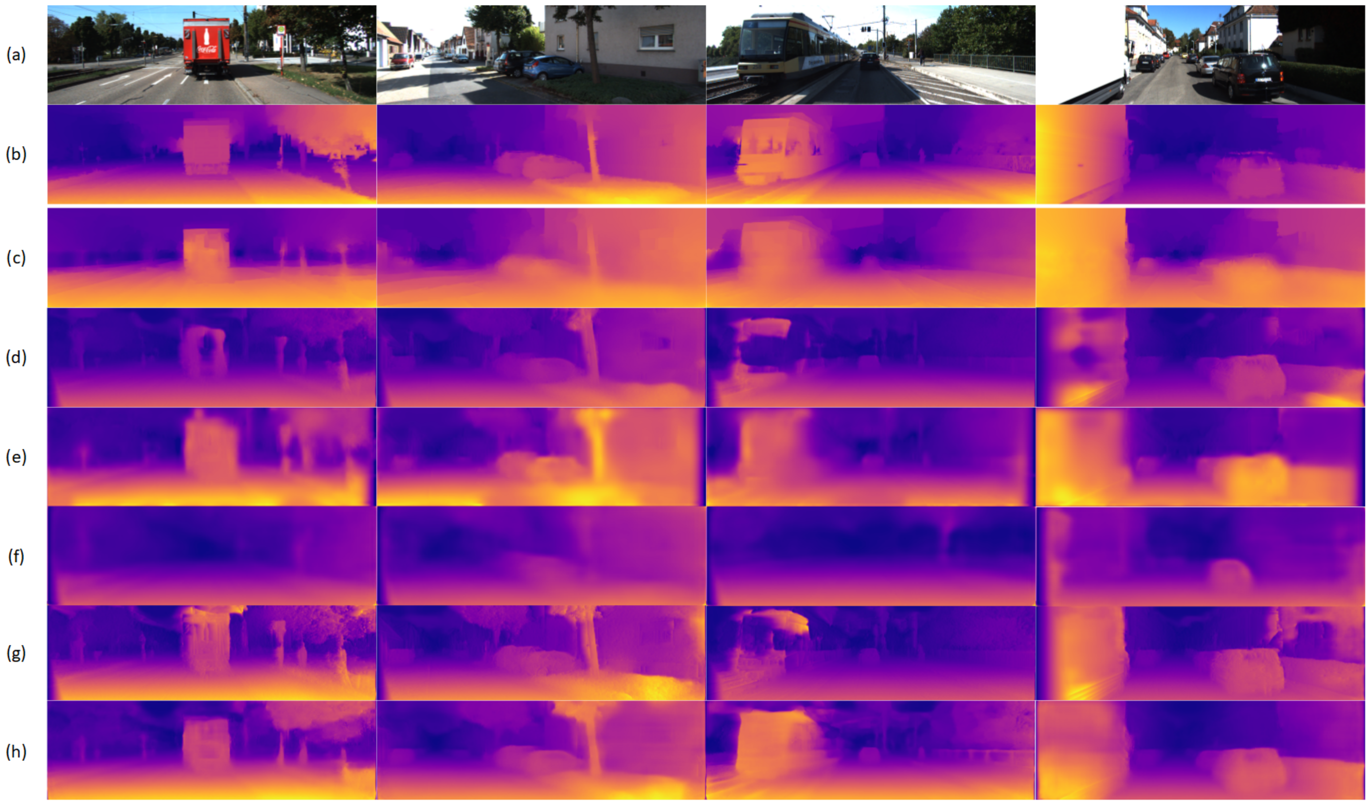


Fig. 5. Qualitative comparison with different methods on KITTI dataset [46]. (a) Color image, (b) Ground-truth, (c) Xu *et al.* [14], (d) Godard *et al.* [9], (e) Zhan *et al.* [10], (f) Pilzer *et al.* [11], (g) Wong *et al.* [18], (h) Ours.

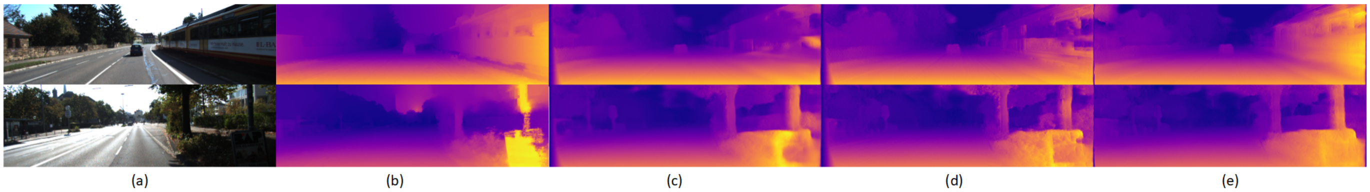


Fig. 6. Visual comparison on KITTI dataset [46] for CS+K. (a) Color image, (b) Ground-truth, (c) Godard *et al.* [9], (d) Wong *et al.* [18], (e) Ours.

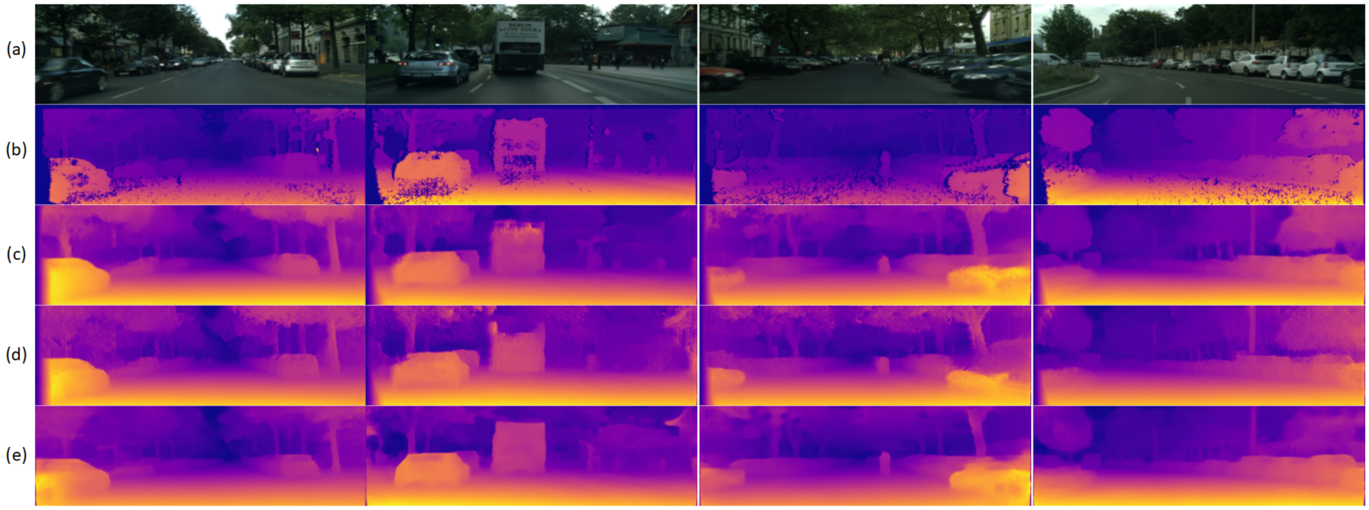


Fig. 7. Visual comparison on Cityscapes dataset [47] for CS+K. (a) Color image, (b) Ground-truth, (c) Godard *et al.* [9], (d) Wong *et al.* [18], (e) Ours.

TABLE II

OBJECTIVE COMPARISON WITH PILZER *et al.* [11] ON KITTI DATASET [46]. SINCE BOTH [11] AND OURS HAVE THE STUDENT AND TEACHER NETWORKS, WE SEPARATELY COMPARE THE PERFORMANCE UNDER EACH CASE (TOP PART FOR THE STUDENT NETWORKS, BOTTOM PART FOR THE TEACHER NETWORKS).

Method	Dataset	Error Metric (lower is better)				Accuracy Metric (higher is better)		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Pilzer *et al.* [11]	K	0.142	1.230	5.785	0.239	0.795	0.924	0.968
Ours	K	**0.105**	**0.842**	**4.810**	**0.196**	**0.861**	**0.947**	**0.978**
Pilzer *et al.* [11] (teacher)	K	0.098	0.830	4.656	0.202	0.882	0.948	0.973
Ours (stereo) (teacher)	K	**0.083**	**0.697**	**4.175**	**0.179**	**0.908**	**0.960**	**0.983**

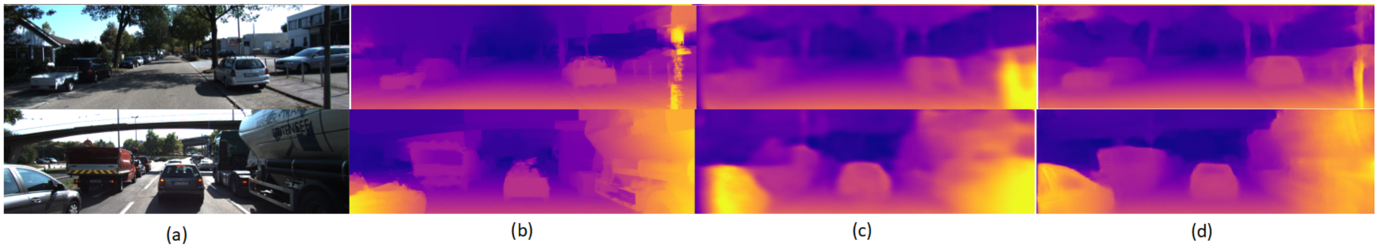


Fig. 8. Visual comparison of the performance of teacher networks on KITTI dataset [46]. (a) Color image, (b) Ground-truth, (c) Pilzer *et al.* [11] (teacher), (d) Ours (stereo) (teacher). Note that the performance comparison of student networks for both methods has given in Fig. 5.

the first column and tram in the third column. The compared method [10] provides blurred depth maps and is insufficient to recover the details, especially for slim and distant objects such as poles and trees. In contrast, the proposed method is capable of preserving sharp boundaries at objects and restoring more accurate depth values, which demonstrates its effectiveness for monocular depth estimation.

Fig. 6 and Fig. 7 show the visual results on KITTI [46] and Cityscapes [47] dataset for CS+K. In Fig. 6, both Godard *et al.* [9] and Wong *et al.* [18] provide more accurate depth maps than these trained on K alone for KITTI [46] dataset. However, [9], [18] cannot restore reasonable depth values in some regions, such as the tram in the first row. In contrast, our method has better visual results. For Cityscapes [47] dataset, we select 1525 stereo pairs as testing set, and directly give the ground-truth depth maps without interpolation. In Fig. 7, it can be seen that the method [9] does not provide good qualitative results in some areas, such as the bus in the second column and the cars in the bottom right corner of the third column. The method [18] has improved performance, but is still inferior to 'Ours', since it introduces some noise artifacts for all pictures.

Comparison with Pilzer *et al.* [11]. As shown in Table II and previous Fig. 5, our student network ('Ours') shows superior numerical and visual performances to [11]. To compare the performance between both the teacher networks, we also list the numerical results of 'Pilzer *et al.* [11] (teacher)' and 'Ours (stereo) (teacher)' in the bottom part of Table II. Both the teacher networks obtain obvious improvement than student networks and particularly our teacher network achieves better performance in all metrics than 'Pilzer *et al.* [11] (teacher)'. In addition, we also provide the comparison results of two teacher networks in Fig. 8. The teacher network of 'Pilzer *et al.* [11]' exhibits blurred depth boundaries and loses some details. In contrast, our method can provide more sharp boundaries and reasonable depth values, demonstrating our excellent 'teacher's own ability'. We have an obvious improvement against [11] under the evaluation of either the finally-used student networks or the teacher networks, e.g., nearly 17% and 26% improvements on *RMSE* and *Abs Rel* for the student networks, and about 10% and 15% improvement on average for teacher.

B. Ablation Study

Ablation study for Stereo-Net. We first investigate the influence of the proposed recursive estimation and feature-driven adaptive refinement module on the Stereo-Net. In testing phase, the inputs are stereo image pairs instead of monocular images. 'Stereo baseline' is the basic DispNet architecture [49] combined with the ASPP module, and the models with recursive estimation are denoted as 'Recursive iter=n' respectively ('Recursive iter=1' is equivalent to 'Stereo baseline'). As shown in Table. III, 'Recursive iter=2' significantly improves the results in all metrics compared with the 'Stereo baseline'. When increasing the iteration number to 3, the performance tends to saturation. Considering the trade-off between accuracy and complexity, we select $n = 2$ in our experiments. Meanwhile, adding the feature-driven adaptive refinement module to 'Stereo basline' ('Adaptive refine') also significantly boosts the performance in all metrics. Finally, our complete Stereo-Net 'Ours (stereo) (teacher)' exhibits the outstanding performance, e.g., the *Sq Rel* is decreased to 0.697 and the accuracy for $\delta < 1.25^3$ achieves 98.3%.

As shown in Fig. 9, the error map e_l is calculated between the synthetic and real left images, in which lower errors are marked as blue regions while higher errors as red regions. 'Stereo baseline' presents the results with higher errors. When adding the recursive estimation or adaptive refinement module to 'Stereo baseline', the errors are decreased. 'Ours (stereo) (teacher)' provides more reasonable and prominent error maps, implying that our Stereo-Net is able to generate more accurate depth map.

TABLE III
ABLATION STUDY FOR STEREO-NET. RESULTS ARE EVALUATED ACCORDING TO EIGEN SPLIT.

Method	Error Metric (lower is better)				Accuracy Metric (higher is better)		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Stereo baseline	0.091	1.182	4.402	0.190	0.896	0.957	0.976
Recursive iter=2	0.084	0.717	4.253	0.180	0.903	0.960	0.977
Recursive iter=3	0.083	0.723	4.247	0.181	0.903	0.960	0.978
Adaptive refine	0.086	0.778	4.222	0.184	0.905	0.959	0.977
Ours (stereo) (teacher)	**0.083**	**0.697**	**4.175**	**0.179**	**0.908**	**0.960**	**0.983**

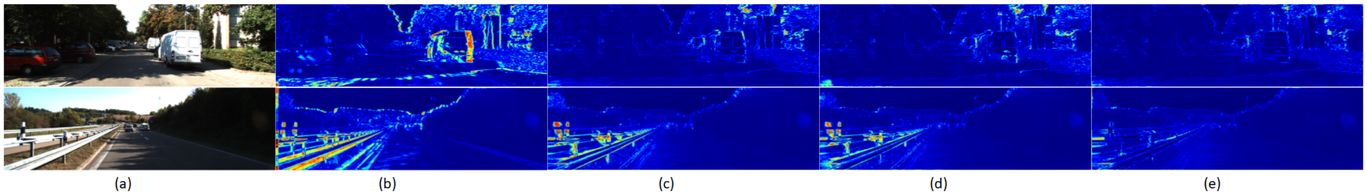


Fig. 9. Visualization of error maps e_l for different components of Stereo-Net on KITTI dataset [46]. (a) Color image, (b) Stereo baseline, (c) Recursive iter=2, (d) Adaptive refine, (e) Ours (stereo) (teacher).

TABLE IV
ABLATION STUDY FOR THE DISTILLATION SCHEME. RESULTS ARE EVALUATED ACCORDING TO KITTI SPLIT.

Method	Error Metric (lower is better)				Accuracy Metric (higher is better)		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Our baseline	0.116	1.249	5.829	0.206	0.847	0.945	0.975
Ours w/o distill	0.113	1.190	5.607	0.202	0.851	0.946	0.977
Output distill	0.107	1.053	5.385	0.188	0.863	0.952	0.980
Feature distill	0.105	1.039	5.386	0.188	0.862	0.952	0.981
Long-range distill	0.109	1.067	5.448	0.193	0.857	0.949	0.978
Ours	**0.093**	**1.015**	**5.043**	**0.170**	**0.889**	**0.964**	**0.985**
Ours single	0.107	1.032	5.435	0.193	0.861	0.949	0.980
Ours	**0.093**	**1.015**	**5.043**	**0.170**	**0.889**	**0.964**	**0.985**

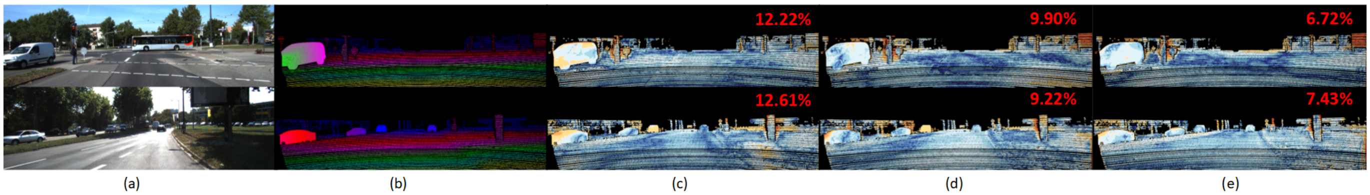


Fig. 10. Qualitative results of error maps on KITTI dataset [46]. (a) Color image, (b) Sparse ground-truth, (c) Our baseline, (d) Ours w/o distill, (e) Ours. The error rate is shown in each case.

Ablation study for multi-space knowledge distillation. The proposed distillation scheme can amalgamate knowledge from Stereo-Net to Mono-Net in the aspects of output space, feature space and long-range dependencies, and we conduct the ablation study for each aspect. 'Our baseline' is the single Mono-Net, while 'Ours w/o distill' is the 'Mono-Net + Stereo-Net' but without the proposed distillation scheme during training. The methods with only output space, feature space or long-range dependencies distillation are denoted as 'Output distill', 'Feature distill' and 'Long-range distill' respectively, while 'Ours' integrates the multi-space knowledge distillations together. As shown in Table. IV, 'Ours w/o distill' improves the performance of depth estimation in all metrics compared with 'Ours baseline' thanks to the sophisticated Stereo-Net. When adding different distillation schemes to 'Ours w/o distill', the Mono-Net has obvious improvements in all metrics, indicating that each component contributes to the knowledge transfer from Stereo-Net to Mono-Net. The best performance is achieved by 'Ours', which demonstrates the effectiveness of our multi-space knowledge distillation scheme. To verify the effectiveness of multi-scale distillation, we also apply the ablation experiment by distilling the multi-space knowledge only at the highest scale, i.e., $s = 1$ ('Ours single'), and compare with 'Ours' that distills the knowledge in a multi-scale fashion. The results of 'Ours single' are inferior to that of 'Ours' in all metrics, and thus it is necessary to exploit the multi-scale distillation to enhance the quality of depth estimation for Mono-Net.

Fig. 10 displays the visual results of error maps between sparse ground-truth and estimated disparities. For the disparity error maps, blue and red regions are lower and higher errors respectively. Compared with 'Ours baseline', 'Ours w/o distill'

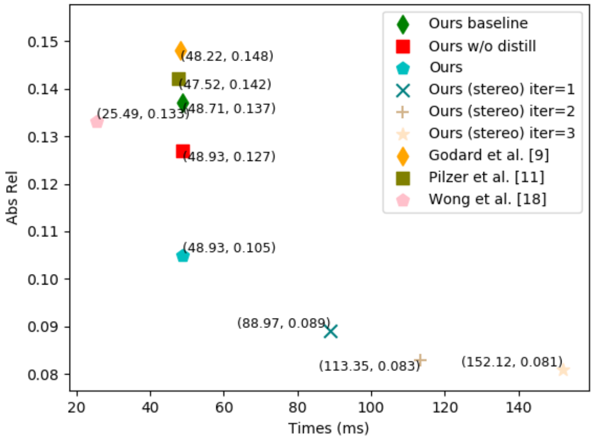


Fig. 11. Runtime vs Performance.

TABLE V
QUANTITATIVE RESULTS ON MAKE3D DATASET [24], INCLUDING 134 TESTING IMAGES.

Method	Error Metric (lower is better)			
	Abs Rel	Sq Rel	RMSE	RMSE log
Godard *et al.* [9]	0.544	10.948	11.765	0.193
Wong *et al.* [18]	0.465	8.420	10.730	0.215
Ours	**0.440**	**7.780**	**10.295**	**0.190**

has more blue regions, which means the closer to the ground-truth. The error rates are decreased from 12.22% and 12.61% to 9.90% and 9.22% respectively, demonstrating that it is beneficial to adopt the architecture 'Mono-Net + Stereo-Net', even without considering the proposed distillation scheme. 'Ours' yields more accurate results especially in car regions by adding the multi-space knowledge distillation scheme, and the error rates fall to 6.72% and 7.43% respectively.

C. *Runtime vs Performance*

In testing phase, the inputs can be monocular images or stereo image pairs by choosing Mono-Net alone or 'Mono-Net + Stereo-Net', i.e., 'Ours' and 'Ours (stereo) (teacher)' respectively. In Fig. 11, we conduct the analysis of runtime and performance for the proposed method and other unsupervised methods. We have download all the source codes of the other methods from their homepage and run these source codes on our hardware GTX 1080Ti GPU for fair comparison. Compared with these methods, 'Ours' produces lower errors in *Abs Rel*, despite taking a little more time. It can be seen that the method 'Ours (stereo) (teacher)' achieves the best performance in *Abs Rel* and with the increase of iteration number, the performance is improved but the running times are accordingly increased.

D. *Generalization*

To test the cross-dataset generalization ability, we also apply our model trained on KITTI to Make3D dataset [24], and compare with other methods which also evaluate the generalization on Make3D. As shown in Table V, 'Ours' achieves the best performance in all metrics, demonstrating the effectiveness of the proposed method in cross-dataset generalization. Fig. 12 shows the corresponding qualitative results. It can be seen that [9] obtains unsatisfied visual results and [18] produces depth maps with noisy artifacts. In contrast, the proposed method can preserve fine details of slim objects and provide more reasonable depth values.

VI. CONCLUSION

We propose a novel architecture that consists of an Mono-Net to infer a coarse depth map from monocular input, and an Stereo-Net to further excavate the stereo information by taking the coarse depth map and stereo pairs as input. The recursive estimation and refinement strategy are used to enhance the ability of Stereo-Net in order to guide the learning of Mono-Net, while the multi-space knowledge distillation is proposed to help Mono-Net infer an accurate depth map without changing its architecture. Experimental results show that our method has superior performance.

REFERENCES

[1] D. Liu, C. Long, H. Zhang, H. Yu, X. Dong, and C. Xiao, "Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes," in *IEEE CVPR*, 2020, pp. 8136–8145.

[2] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, "Pose estimation and segmentation of multiple people in stereoscopic movies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1643–1655, 2015.

[3] M. Redi, N. OHare, R. Schifanella, M. Trevisiol, and A. Jaimes, "6 seconds of sound and vision: Creativity in micro-videos," in *IEEE CVPR*, 2014, pp. 4272–4279.

[4] Y. Zhang, D. Zou, J. S. Ren, Z. Jiang, and X. Chen, "Structure-preserving stereoscopic view synthesis with multi-scale adversarial correlation matching," in *IEEE CVPR*, 2019, pp. 5853–5862.

[5] J. Mahmud, T. Price, A. Bapat, and J. Frahm, "Boundary-aware 3d building reconstruction from a single overhead image," in *IEEE CVPR*, 2020, pp. 438–448.

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

[7] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *IEEE CVPR*, 2015, pp. 5162–5170.

[8] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*. Springer, 2016, pp. 740–756.

[9] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE CVPR*, no. 1–6, 2017.

[10] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *IEEE CVPR*, 2018.

[11] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *IEEE CVPR*, 2019.

[12] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *IEEE CVPR*, 2019, pp. 2624–2632.

[13] J. Nath Kundu, P. Krishna Uppala, A. Pahuja, and R. Venkatesh Babu, "Adadepth: Unsupervised content congruent adaptation for depth estimation," in *IEEE CVPR*, 2018, pp. 2656–2665.

[14] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *IEEE CVPR*, 2017.

[15] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *IEEE CVPR*, 2017.

[16] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *IEEE CVPR*, 2018.

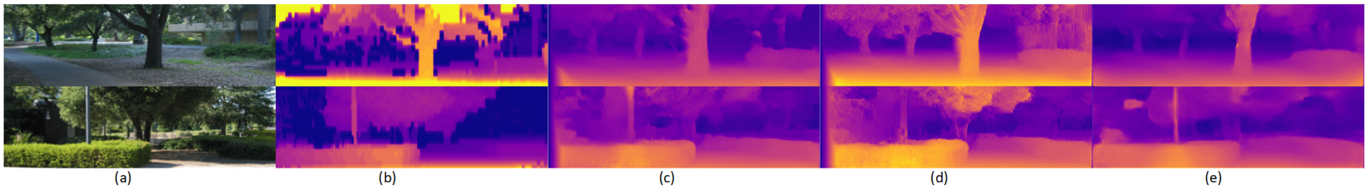


Fig. 12. Qualitative comparison on Make3D [46]. (a) Color image, (b) Ground-truth, (c) Godard *et al.* [9], (d) Wong *et al.* [18], (e) Ours.

[17] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *NIPS*, 2015, pp. 2017–2025.

[18] A. Wong and S. Soatto, "Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction," in *IEEE CVPR*, 2019.

[19] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *IEEE CVPR*, 2019.

[20] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. V. Gool, "Weakly supervised cascaded convolutional networks," in *IEEE CVPR*, 2017, pp. 5131–5139.

[21] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *ECCV*. Springer, 2016, pp. 614–630.

[22] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE CVPR*, 2017, pp. 1647–1655.

[23] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *IEEE CVPR*, 2017.

[24] A. Saxena, S. H. Chung, and A. Y. Ng, "3-D depth reconstruction from a single still image," *International journal of computer vision*, vol. 76, no. 1, pp. 53–69, 2008.

[25] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in *IEEE CVPR*, 2014, pp. 716–723.

[26] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *IEEE CVPR*, 2014, pp. 89–96.

[27] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *IEEE CVPR*, 2015.

[28] H. Wei, S. Chang, L. Ding, Y. Mo, and M. Witbrock, "Image super-resolution via dual-state recurrent networks," in *IEEE CVPR*, 2018.

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[30] A. A. Rusu, S. Gomez Colmenarejo, C. Gulcehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, and R. Hadsell, "Policy distillation," in *ICLR*, 2016.

[31] R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," in *ICLR*, 2018.

[32] M. Phuong and C. H. Lampert, "Distillation-based training for multi-exit architectures," in *IEEE ICCV*, 2019, pp. 1355–1364.

[33] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *IEEE ICCV*, 2019, pp. 1365–1374.

[34] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *IEEE ICCV*, 2017.

[35] I. Radosavovic, P. Dollar, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *IEEE CVPR*, 2018.

[36] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE CVPR*, 2015, pp. 1520–1528.

[37] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv:1606.00915*, 2016.

[39] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *IEEE CVPR*, 2018, pp. 8981–8989.

[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE CVPR*, 2018.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[42] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, "Single view stereo matching," in *IEEE CVPR*, 2018.

[43] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *IEEE CVPR*, 2019, pp. 9791–9801.

[44] M. M. Puscas, D. Xu, A. Pilzer, and N. Sebe, "Structured coupled generative adversarial networks for unsupervised monocular depth estimation," in *IEEE International Conference on 3D Vision (3DV)*, 2019.

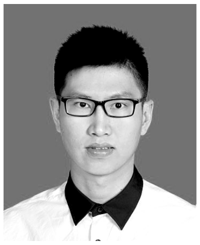
[45] C. Godard, O. M. Aodha, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *IEEE ICCV*, 2019.

[46] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[47] 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 *IEEE CVPR*, 2016, pp. 3213–3223.

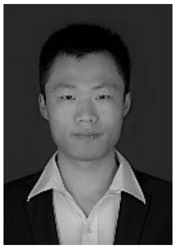
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[49] N. Mayer, E. Ilg, P. Hausser, 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 *IEEE CVPR*, 2016, pp. 4040–4048.



Xinchen Ye (M'17) received the B.E. degree and Ph.D. degree from the Tianjin University, Tianjin, China, in 2012 and 2016, respectively. He was with the Signal Processing Laboratory, EPFL, Lausanne, Switzerland in 2015 under the Grant of the Swiss federal government. He has been a Faculty Member of Dalian University of Technology, Dalian, Liaoning, China, since 2016, where he is currently an Associate Professor with the DUT-RU International School of Information Science and Engineering. His current research interests include image/video processing and 3D imaging. As a co-author, he received the Platinum Best Paper Award in the IEEE ICME 2017. He won the Rising Star Award in 2018 ACM Turing Celebration Conference-China (ACM TURC 2018).

Xin Fan received the B.E. and Ph.D. degrees in information and communication engineering from Xian Jiaotong University, Xian, China, in 1998 and 2004, respectively. He was with Oklahoma State University, Stillwater, from 2006 to 2007, as a post-doctoral research Fellow. He joined the School of Software, Dalian University of Technology, Dalian, China, in 2009. His current research interests include computational geometry and machine learning.



Mingliang Zhang received the B.E. degree and Ph.D. degree in School of Mathematical Sciences from Qufu Normal University, China, in 2014, and School of Mathematical Sciences from Dalian University of Technology, China, in 2020, respectively. He has been a Faculty Member of Mathematics and Statistics, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. His research interests include optimization, image processing and machine learning.

Rui Xu received the Ph.D. degree in 2007 from the graduate school of science and engineering, Ritsumeikan University, Japan. He worked in the digital technology research center, Sanyo Electric Co., Ltd., Japan, from 2008 to 2010. He worked as a senior researcher successively in Yamaguchi University and Ritsumeikan University from 2010 to 2015. Since December 2015, he served as an associate professor at Dalian University of Technology. His research fields include intelligent computing in medical images and computer vision.

Wei Zhong received the M.S. and Ph.D. degrees from Waseda University, Tokyo, Japan, in 2010 and 2014, respectively. He served as a Chief Designer and Director of the Institute of Image Processing Technology in the State Key Laboratory of Digital Multimedia Technology, Hisense Group, Qingdao, China, from 2014 to 2018. He is currently an Associate Professor in the International School of Information Science and Engineering, Dalian University of Technology, Dalian, China. His research interests include computer vision and VLSI design automation.