Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Semi-supervised Depth Estimation from Sparse Depth and a Single Image for Dense Map Construction

Xinghong Huang, Zhuang Dai, Xubin Lin, Chenjie Suo, Li He, Yisheng Guan and Hong Zhang

Biomimetic and Intelligent Robotics Lab (BIRL)
School of Electromechanical Engineering, Guangdong University of Technology
Guangzhou, Guangdong Province, China
{*huangxinghong95,dytrong1994*}*@gmail.com,*{*heli,ysguan*}*@gdut.edu.cn,*{*hzhang*}*@ualberta.ca*

***Abstract*— For robust navigation, the objective in visual SLAM is to create a dense map from a sparse input. Although there have been a significant number of endeavors on real-time mapping, the existing works for visual SLAM systems still fail to preserve adequate geometry details that are important for navigation. This paper estimates pixel-wise depth from a single image and a few depth points which are constructed from registered LiDAR or acquired from visual SLAM systems to construct a dense map. The main idea is to employ a set of new loss functions consisting of photometric reconstruction consistency (forward-backward consistency and left-right consistency), depth loss, nearby frame geometric consistency, and smoothness loss and propose a depth estimation network based on ResNet. The experimental results show that the proposed method is superior the state-of-the-art methods on both raw LiDAR scans dataset and semi-dense annotation dataset. Furthermore, the errors of the sparse depth produced by stereo ORB-SLAM2 are evaluated and this sparse depth and a single image are fed into the proposed model to further demonstrate the superiority of the proposed work.**

***Index Terms*— semi-supervised depth estimation, dense map construction**

I. INTRODUCTION

Simultaneous localization and mapping (SLAM) is a significant method for robotics and autonomous driving. Other than the input 2D images, the SLAM system needs to obtain the real depth information for robust 3D navigation. Generally, the real depth can be obtained by range sensors [1]–[3] or stereo camera [4], [5]. Among these methods, stereo ORB-SLAM2 [4] plays an important role in SLAM systems since it is robust and low-cost. Feature-based SLAM methods, such as ORB-SLAM, typically keep tracking hundreds of sparse landmarks. The existing sparse landmark methods are good at localization but may fail to provide a robust navigation. The main reason for the existence of spatial sparsity in landmarks is that the depth value of one landmark is calculated by triangulation of the matched keypoints. Therefore, the sparse SLAM system may surffer from a poor navigation performance due to the absence of detailed environment description. One of the most effective solutions for robust navigation is to use the dense depth map of squential frame to obtain the physical environment information.

In past years, there has been a strong interest and research effort on depth estimation using a single image only. Some researchers utilized the sparse depth acquired from a range sensor, along with a single image, to address the single image depth estimation problem, which may be ill-posed in some cases. Recently, the development of deep learning has improved depth estimation [6]–[13] and made it possible to construct a dense map using SLAM systems. However, these methods can only generate relatively high-quality depth from expensive range sensor (for example, LiDAR) [11], [13]. Therefore, there is still some works to be done before the actual application. Another effective approach is to obtain sparse depth and a single image from existing stereo SLAM systems. However, the sparse depth values exist a certain level of errors, which requires the network to have some anti-noise capability. Noise existing in sparse depth and the uncertainty of the number of sparse depths produced by SLAM systems limit the accuracy of the depth estimation to some extent.

In this paper, a depth estimation model is proposed takes sparse depth and a single image as input to predict the full-resolution depth image. The proposed work makes the following two contributions.

1) A new loss function consisting of photometric reconstruction consistency (forward-backward consistency and left-right consistency), depth loss, nearyby frame geometric consistency, and smoothness loss is proposed to optimize the ResNet-based model.
2) To the best of the authors' knowledge, the errors of sparse depth produced by the popular stereo ORB-SLAM2 system are evaluated for the first time in this work. In addition, this sparse depth and a single image are used to estimate the dense depth map to finish dense map construction. Furthermore, it is demonstrated the proposed model can be embedded into the SLAM systems.

II. RELATED WORK

Depth prediction from a single image. In the last

*Hong Zhang is also with the Department of Computing Science, University of Alberta, AB, Canada.

*The work in this paper is partially supported by the Natural Science Foundation of China (Grant No. 61673125, 61703115), the Frontier and Key Technology Innovation Special Funds of Guangdong Province (Grant No. 2016B090910003), and the Program of Foshan Innovation Team of Science and Technology (Grant No. 2015IT100072).

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

few years, several single image depth estimation methods have been proposed. Saxena et al. [14] proposed the first supervised learning-based approach for single image depth map prediction. They used the markov random field and hand-crafted multi-scale texture features to accomplish the depth prediction. Karsch et al. [15] tackled the depth estimation using non-parametric methods, which retrieves the depth value of the corresponding pixel from the established database.

In recent years, deep learning has been successfully applied to learn image features to estimate dense depth. Eigen et al. [6] proposed a convolutional neural network (CNN) architecture with a coarse-to-fine strategy to predict the depth. Meanwhile, they also introduced surface normal and semantic labels to improve the accuracy of the estimated dense depth in the same year [7]. [8] proposed a ResNet-based encoder-decoder convolutional architecture, encompassing residual learning, to model the ambiguous mapping between monocular images and depth maps. Recently, Godard et al. [16] proposed an image alignment loss in convolutional encoder-decoder architecture but additionally enforce left-right consistency of the predicted disparities in the stereo pair. Kuznietsov et al. [10] used sparse ground-truth depth for supervised learning and enforced the deep network to produce photo consistent dense depth maps in a stereo setup using a direct image alignment loss. Pilzer et al. [17] proposed a deep generative network that learned to predict the disparity map between two image views in a calibrated stereo camera setting.

Depth prediction from a single image and sparse depth. Despite over a decade of research effort devoted to depth prediction using a single image only, the accuracy and reliability of the proposed methods are still far from being practical. It is natural to improve the depth prediction by using the sparse depth as the input of the model with a single image. In a recent study, Ma et al. [19] developed a deep regression model to learn a direct mapping from LiDAR and monocular camera to dense depth. Eldesokey et al. [18] introduced an algebraically-constrained convolution layer for CNNs with sparse input and also proposed the novel strategies for determining the confidence from the convolution operation and propagation of consecutive layers. Although, the above-mentioned methods have achieved good results, the sparse depth input is far denser than the stereo visual SLAM systems can provide, making it difficult to apply in stereo ORB-SLAM2.

However, the depth estimation from the sparse depth and a single image becomes much more challenging when the input sparse depth image has much lower density. In order to resolve this problem, many studies have attempted to perform depth prediction by using low-density sparse depth and a single image. Liao et al. [12] introduced 2D planar observation from the remaining laser range finder to provide an additional reference depth signal as input and obtained higher accuracy and lower errors compared to the single image only depth prediction. Cadena et al. [11] proposed to use sparse depth on extracted FAST corner features as part of the input to the system to produce a low-resolution depth prediction. Ma et al. [13] proposed to use a deep regression network to learn directly from a single image and sparse depth uniformly sampled from the LiDAR data, and also explored the impact of a number of depth samples on prediction accuracy. Although, the sparse depth map density of these methods conforms to the number of feature-based SLAM system landmarks, their accuracy still needs to be further improved.

Several recent studies have investigated the superior performance of ResNet [20] on depth estimation [8], [19], [21]. This paper also adopts the ResNet as the main architecture of encoder and successively extracts low-resolution high-dimensional features from the sparse depth and a single image. The decoder upsamples the output of the encoder to produce the dense depth using up-convolution operation. Then the photometric reconstruction consistency, raw LiDAR scans or semi-dense annotation as supervision and geometric consistency are used to train the regression network. The superior performance of the proposed depth estimation approach is validated thorough experiments.

The rest of this paper is organized as follows. We will describe our proposed model for depth estimation in Section III. The propsed model will be evaluated experimentaly in its estimated dense depth error and accuracy in Section IV. Conclusions will be drawn and future work outlined in Section V.

III. Proposed Method

In this section, the proposed framework (shown in Fig. 1) for learning dense depth from sparse depth and a single image (see Fig. 2 for an example) using adjacent frames is presented. The loss terms of the improved framework are described in detail below.

A. Nearby frame transformation

The accuracy of the transformation between frame t and frame $t+1$ is significant to the proposed system. Unlike [24] that used the network to estimate the transformation between sequence frames, the Perspective-n-Point (PnP) problem [25] is solved to estimate the relative transformation $T_{t->t+1}$ and $T_{t+1->t}$ followed by [19]. Note that in priciple, $\|t_{t->t+1}\|_2$ and $\|t_{t+1->t}\|_2$ should be equal. Thus the translation error can be formulate as:

$$e_r = \|t_{t->t+1}\|_2 - \|t_{t+1->t}\|_2 \tag{1}$$

where $t_{t->t+1}$ and $t_{t+1->t}$ are the translation of $T_{t->t+1}$ and $T_{t+1->t}$, respectively. If $e_r > \lambda_t$, the calculation of this transformation high probability is failure. In this case, the frame $t+1$ is replaced by frame t and the transformation is set to identity matrix. where λ_t is scalar.

B. Loss Terms

In this section, the total loss Ls is defined as:

$$Ls = \lambda_d E_d + \lambda_r E_r + \lambda_g E_g + \lambda_s E_s \tag{2}$$

where λ_i are scalars and the E terms are defined below:

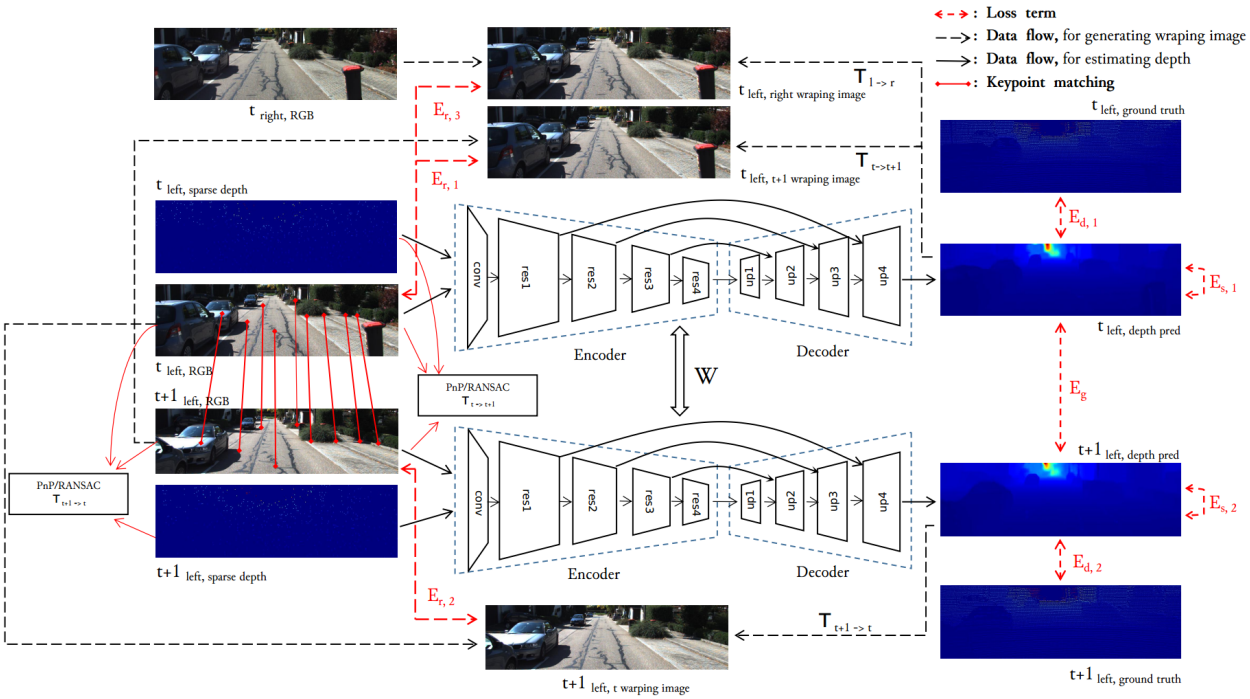


Fig. 1: Summary of the proposed framework. The framework contains four loss terms: E_r, E_d, E_g and E_s. As for the structure of the network, the encoder consists of several ResNet blocks, and the decoder uses the up-convolution operation. Meanwhile, the skip connection is also used to fuse the low-level and high-level features.

1) Depth Supervision Loss E_d: The supervised loss term measures the difference between the ground truth GT and predicted depth D on the set of pixels with known depth Ω of ground truth.

$$E_d = \sum_{k\in\{t,t+1\}} \lambda_{d,k} M_k \sum_{i,j\in\Omega_k} \left\| D^k_{i,j} - GT^k_{ij} \right\|_1 \tag{3}$$

where Ω_t and Ω_{t+1} are the corresponding pixels of t and $t+1$ images where the ground truth is available, respectively. M_t and M_{t+1} are the binary matrix for the t and $t+1$ images where the ground truth is available, respectively. $\lambda_{d,t}$ and $\lambda_{d,t+1}$ represent the weights for the t and $t+1$ images, respectively.

2) Photometric Reconstruction Consistency Loss E_r: The photometric reconstruction consistency loss is used not only between t and $t+1$ images (forward-backward consistency) but also between the left and the right images at time t (left-right consistency). Given the relative transformation $T_{t\to t+1}$ and the predicted depth D_t, the $t+1$ image I_{t+1} can easily be inversed warping to t image I_t. Let p_l be the homogeneous coordinates of a pixel in the reference view, the projected coordinates are obtained as:

$$p_{r,t} = KT_{l\to r} D_{l,t} K^{-1} p_{l,t} \tag{4}$$

$$p_{l,t+1} = KT_{t\to t+1} D_{l,t} K^{-1} p_{l,t} \tag{5}$$

$$p_{l,t} = KT_{t+1\to t} D_{l,t+1} K^{-1} p_{l,t+1} \tag{6}$$

where $p_{r,t}$, $p_{l,t+1}$ and $p_{l,t}$ are the projected coordinates on $I_{r,t}$, $I_{l,t+1}$ and $I_{l,t}$, respectively. K is the camera intrinsic parameter. Then the new frames, $\hat{I}_{r,t}$, $\hat{I}_{l,t+1}$ and $\hat{I}_{l,t}$ can be synthesized using inversely warped proposed by [30]. Multi-scale strategy [19] is used to establish the photometric reconstruction loss

$$\begin{aligned} E_r = {} & \lambda_{r,1} \sum_{s\in S} \frac{1}{s} M_{l,t} \left\| \hat{I}_{l,t} - I_t \right\|_1 \\ & + \lambda_{r,2} \sum_{s\in S} \frac{1}{s} M_{l,t+1} \left\| \hat{I}_{l,t+1} - I_{t+1} \right\|_1 \\ & + \lambda_{r,3} \sum_{s\in S} \frac{1}{s} M_{r,t} \left\| \hat{I}_{r,t} - I_t \right\|_1 \end{aligned} \tag{7}$$

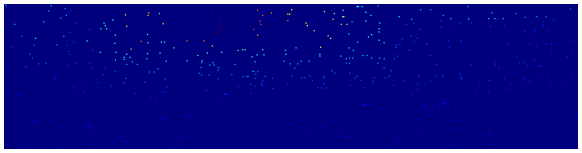
where S is the set of all scaling factors, $\lambda_{r,1}$, $\lambda_{r,2}$ and $\lambda_{r,3}$ are scalar values and $M_{l,t}$, $M_{l,t+1}$ and $M_{l,t+1}$ are binary matrices for the corresponding image where the ground truth is available, respectively.

3) Nearby frame geometric consistency Loss E_g: In principle, the prediction depths D_t and D_{t+1} predicted from images I_t and I_{t+1}, respectively, should have some relationship. From Eq. 5, the projected coordinates from current view to nearby view can be obtained. Using the projected coordinates, the new depth map $\hat{D}_{t+1}$ from D_{t+1} can be generated as photometric reconstruction. Note that in order to establish the connection between D_t and $\hat{D}_{t+1}$, the point cloud is transformed from t image coordinate system to $t+1$ image coordinate system

$$P_{t\to t+1} = T_{t\to t+1} D_t K^{-1} p_t \tag{8}$$

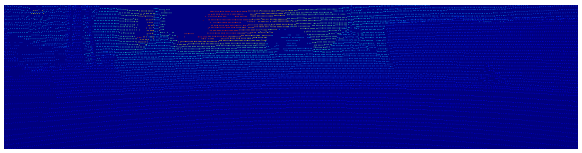
Then the depth of z coordinate of $P_{t\to t+1}$ is denoted as $\hat{D}_{t\to t+1}$. In principle, $\hat{D}_{t\to t+1}$ and $\hat{D}_{t+1}$ should be equal. Thus,

(a) Single image



(b) Sparse depth with 500 available depth values

(c) Dense depth prediction form sparse depth and a single image



(d) Raw LiDAR scans

Fig. 2: Example of prediction using the proposed depth estimation method on KITTI with size 912×228 of input and output image. From top to bottom: (a) Single image; (b) Sparse depth with 500 available depth values; (c) Dense dense prediction from sparse depth and a single image; (d) Raw LiDAR scans.

the nearby frame consistency loss E_g can be defined as:

$$E_g = M_g \left\| \hat{D}_{t \to t+1} - \hat{D}_{t+1} \right\|_1 \tag{9}$$

where M_g is the binary mask calculated from the ground truth

$$M_g = \left\| \hat{D}^{gt}_{t \to t+1} - \hat{D}^{gt}_{t+1} \right\|_1 < \lambda_{g,1} \tag{10}$$

where $\lambda_{g,1}$ is the scalar value. Note that in the experiment, the $T_{t \to t+1}$ is not accurate, so the ground truth is used to obtain the $\hat{D}^{gt}_{t \to t+1}$ and $\hat{D}^{gt}_{t+1}$ like $\hat{D}_{t \to t+1}$ and $\hat{D}_{t+1}$, respectively, in order to create the binary mask and better establish the loss function.

4) Smoothness Loss E_s: As suggested in [19], the smoothness loss term is a regularization term that encourages the depth map to be locally smooth. Therefore, the smoothness regularization term in the proposed framework is defined as:

$$E_s = \lambda_{s,t} \left\| \nabla^2 D_t \right\|_1 + \lambda_{s,t+1} \left\| \nabla^2 D_{t+1} \right\|_1 \tag{11}$$

where D_t and D_{t+1} are the predicted depths from image I_t and I_{t+1}, respectively. λ_{st} and λ_{st+1} are the weights of the smoothness regularization at time t and $t+1$, respectively.

IV. EXPERIMENTAL EVALUATION

In order to demonstrate the performance of the proposed method, experimental assessments were performed on KITTI dataset [26] followed by [16]. For the generation of sparse depth, the sampling strategy like [13] was used. See Figure 2 for an example. Practically, the errors in the sparse depth calculated by SLAM systems will affect the accuracy of depth estimation inescapably. Therefore, in the experiment, not only the sparse depth from the ground truth was sampled but also the sparse depth from stereo ORB-SLAM2 system was used as input, to evaluate the results of the proposed framework. Note that in the experiment, the raw LiDAR scans dataset was used as the ground truth as shown in Table I following previous researches [10]–[13], [16]. However, in recent researches, Aleotti et al. [27] and Ali et al. [28] mentioned that when LiDAR points are projected into the camera space, artifacts results around objects that are occluded in the image but not from the LiDAR point of view. Fortunately, Uhrig et al. [29] provided preprocessed annotated depth maps in terms of this issue.

In the rest of the experiments, the proposed method is evaluated based on the raw LiDAR scans and the sparse depth is sampled from the ground truth. Since the semi-dense annotated dataset has higher accuracy and denser depth than the raw LiDAR scans data. Thus for the sparse depth from stereo ORB-SLAM2 system, the semi-dense annotation datasets are used as ground truth. Note that in the proposed framework, two sizes of input and output images are used, they are 1216×352 used by [18], [19] and 912×228 used by [13], respectively. In order to better integrate the our proposed framework into the SLAM systems, the depth accuracy of 3D landmarks of stereo ORB-SLAM2 is evaluated based on the semi-annotated dataset as shown in Table II. Finally, we feed the sparse depth of 3D landmarks and the left camera image are fed into the proposed framework and the competitive methods, the results are shown in Table III.

A. Evaluation Metric

In the experiment, the standard metrics used by previous researchers [13], [16]–[19] are used. Particularly, root mean squared error (RMSE), mean absolute error (MAE), root mean squared logarithmic error (RMSE log), absolute relative difference (Abs Rel) and the percentage of depths (δ) within a certain threshold distance to its ground truth are used. Note that in the evaluation of the sparse depth calculated from stereo ORB-SLAM2, only two metrices use RMSE and MAE are used due to the sparsity of sparse depth.

B. Implementation Details

The proposed network is trained with pretrain weight of ResNet34 using PyTorch [31]. Adam optimizer [32] with initial learning rate of $lr = 10^{-5}$, and use the learning rate decrease strategy used by [19] for a total of 15 epochs are used. The hyperparameters for loss are chosen as $\lambda_d = 1.0$, $\lambda_r = \lambda_g = \lambda_s = 0.1$, $\lambda_{d,t} = \lambda_{d,t+1} = 0.5$, $\lambda_{r,1} = \lambda_{r,2} = 0.3$, $\lambda_{r,3} = 0.4$, $\lambda_t = \lambda_{g,1} = 0.08$ and $\lambda_{s,t} = \lambda_{s,t+1} = 0.5$.

C. Results

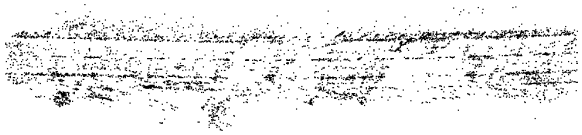
In this section, the quantitative comparison with the state-of-the-art methods is performed using either the raw LiDAR

TABLE I: Comparison with state-of-the-art methods on KITTI dataset. The projected raw LiDAR scans dataset is used as the ground truth as used by the previous researches. *Sample Num* means the number of the available depth values of sparse depth. The bolder represents the best performance among six depth prediction methods in terms of each metric.

Method		Errors (Lower is better)				Accuracy (Higher is better)		
Model	Sample Num	RMSE[m]	MAE[m]	RMSE log	Abs Rel	$\delta 1$	$\delta 2$	$\delta 3$
full-MAE [11]	650	7.140	-	0.179	-	0.709	0.888	0.956
Liao et al. [12]	250	4.500	-	0.113	-	0.874	0.960	0.984
DRN [13]	500	3.387	1.199	0.062	0.047	0.939	0.974	0.988
NConv-CNN-L1 [18]	500	3.295	1.055	0.025	0.062	0.942	0.971	0.986
Sparse-to-dense [19]	500	3.252	1.140	0.027	0.065	0.941	0.972	0.986
Ours-1216x352	500	3.141	1.021	0.024	0.059	0.943	0.974	0.987
Ours-912x228	500	**3.050**	**0.937**	**0.020**	**0.049**	**0.954**	**0.979**	**0.989**

TABLE II: Quantitative evaluation of landmarks produced by stereo ORB-SLAM2 on the KITTI dataset. In this Table, 3000 keypoints are extracted for each frame and the landmarks are collected after the local tracking module of stereo ORB-SLAM2. The semi-dense annotated dataset is used as the ground truth to evaluate the collected landmarks. Due to the sparsity of collecting landmarks, the two metrices RMSE and MAE are used to accomplish the quantitative evaluation. *Avg. Num* means the average number of the collecting landmarks of each frame. *Crop size* means the size of the frame to suit different depth estimation algorithms.

Dataset			Errors (Lower is better)	
Subdataset	Crop size	Avg. Num	RMSE[m]	MAE[m]
Train dataset	1216x352	515	2.564	0.975
	912x228	573	1.693	0.693
Test dataset	1216x352	501	2.635	0.966
	912x228	475	1.769	0.720
Whole dataset	1216x352	513	2.571	0.974
	912x228	563	1.707	0.696



(a) Sparse map produced by stereo ORB-SLAM2

(b) Dense map using the proposed depth estimation method

(c) Dense map using semi-dense annotated dataset (ground truth)

Fig. 3: Example of a dense map built from left to right on KITTI dataset, mainly a road and tree shadows (top view). From top to bottom: (a) Sparse map produced by stereo ORB-SLAM2; (b) Dense map using proposed depth estimation method from sparse depth and a single image. (c) Dense map using semi-dense annotated dataset.

scans dataset or the semi-dense annotated dataset for training and testing, as shown in Tables I and III, respectively. It can be observed from Tables I and III that the proposed method outperforms the existing state-of-the-art methods in terms of most of the metrics.

In the experiment, the sparse depth produced by stereo ORB-SLAM2 is evaluated in order to better integrate the proposed method to the existing SLAM systems and further demonstrate the performance of the proposed framework. It should be noted that the sparse depth refers to the z-coordinates of 3D landmarks of each frame. For the quantitative results of sparse depth in Table II, although there exsit some errors in the accuracy of sparse depth from stereo ORB-SLAM2, its accuracy is basically consistent with the localization accuracy of stereo ORB-SLAM2 in a certain range. Furthermore, using the sparse depth from stereo ORB-SLAM2 and left camera image as inputs, the proposed method and the state-of-the-art methods are evaluated. It can be observed from Table III that the proposed method still show its superior performance compared to the state-of-the-art methods in terms of most of the metrics. For instance, although the proposed method with size 1216×352 only better than Sparse-to-dense by 0.029 m in terms of RMSE, the proposed method outperforms the Sparse-to-dense by 0.197 m and 0.012 in terms of MAE and Abs Rel, respectively. Finally, the proposed method is integrated into stereo ORB-SLAM2 to obtain a dense map and tested on the KITTI dataset. It can be observed from Fig. 3 that the sparse map produced by stereo ORB-SLAM2 system is unable to preserve adequate geometry details of the surroundings. Integrating the depth estimation algorithm into stereo ORB-SLAM2 system can produce dense map that is visually close to the ground truth, which further demonstrates the effectiveness of the proposed method.

V. CONCLUSION

In this paper, a CNN is proposed for depth estimation from a sparse depth and a single image. Specifically, the photomet-

TABLE III: Comparison with state-of-the-art methods on the KITTI dataset. The semi-dense annotated dataset is used as the ground truth. A left camera image and the sparse depth of landmarks produced by stereo ORB-SLAM2 system are used as inputs to estimate the dense depth end-to-end. The bolder represents the best performance in terms of each metric.

Model	Errors (Lower is better)				Accuracy (Higher is better)		
	RMSE[m]	MAE[m]	RMSE log	Abs Rel	$\delta 1$	$\delta 2$	$\delta 3$
DRN [13]	3.293	1.403	0.029	0.067	0.945	0.983	**0.994**
NConv-CNN-L1 [18]	7.412	4.232	0.104	0.241	0.620	0.844	0.933
Sparse-to-dense [19]	3.195	1.456	0.034	0.074	0.945	0.982	0.992
Ours-1216x352	3.166	1.259	0.028	0.062	0.946	0.981	0.992
Ours-912x228	**3.084**	**1.191**	**0.024**	**0.054**	**0.951**	**0.985**	**0.994**

ric reconstruction consistency loss, the depth loss, the nearby frame geometric consistency loss and the smoothness loss are used to train the ResNet-based network. The experimental results either on raw LiDAR scans or semi-annotated datasets demonstrate the superiority of the proposed method over the-state-of-the-art methods. In addition, other than taking the LiDAR as sparse depth source, the proposed method is also tested with stereo ORB-SLAM2 as depth input. The obtained results also confirm the superiority of the proposed method. Finally, the proposed method is integrated to stereo ORB-SLAM2 and dense map is obtained which preserves adequate geometry details of the surroundings. The additional time cost introducing by the integration of the proposed method with stereo ORB-SLAM2 is limited and can be further significantly reduced by running on a GPU. The future work will include valuating the dense map quantitatively and further improving the accuracy of the pose of the stereo ORB-SLAM2 system using the dense map in real time.

REFERENCES

[1] J. Zhang and S. Singh, "LOAM: Lidar odometry and mapping in real-time," in Robotics: Science and Systems Conference, 2014.

[2] Henry, Peter, et al. RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. Experimental Robotics. 2014:647-663.

[3] Hess, Wolfgang , et al. "Real-Time Loop Closure in 2D LIDAR SLAM." 2016 IEEE International Conference on Robotics and Automation (ICRA) IEEE, 2016.

[4] Mur-Artal, Raul , and J. D. Tardos . "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras." IEEE Transactions on Robotics (2017):1-8.

[5] Engel, Jakob , V. Koltun , and D. Cremers . "Direct Sparse Odometry." IEEE Transactions on Pattern Analysis and Machine Intelligence (2017):1-1.

[6] Eigen, David, C. Puhrsch, and R. Fergus. "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network." International Conference on Neural Information Processing Systems 2014.

[7] Eigen, David , and R. Fergus . "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture." (2014).

[8] Laina, Iro , et al. "Deeper Depth Prediction with Fully Convolutional Residual Networks." (2016).

[9] Zhou, Lipu , et al. "Unsupervised Learning of Monocular Depth Estimation with Bundle Adjustment, Super-Resolution and Clip Loss." (2018).

[10] Kuznietsov, Yevhen , Stckler, Jörg, and B. Leibe . "Semi-Supervised Deep Learning for Monocular Depth Map Prediction." (2017).

[11] C. Cadena, A. R. Dick, and I. D. Reid, "Multi-modal auto-encoders as joint estimators for robotics scene understanding." in Robotics: Science and Systems, 2016.

[12] Liao, Yiyi , et al. "Parse Geometry from a Line: Monocular Depth Estimation with Partial Laser Observation." (2016).

[13] Ma, Fangchang , and S. Karaman . " [IEEE 2018 IEEE International Conference on Robotics and Automation (ICRA) - Brisbane, Australia (2018.5.21-2018.5.25)] 2018 IEEE International Conference on Robotics and Automation (ICRA) - Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image." (2018):1-8.

[14] Saxena, Ashutosh , S. H. Chung , and A. Y. Ng . "3-D Depth Reconstruction from a Single Still Image." International Journal of Computer Vision 76.1(2008):53-69.

[15] Karsch, Kevin , C. Liu , and S. B. Kang . "Depth Extraction from Video Using Non-parametric Sampling." European Conference on Computer Vision Springer, Berlin, Heidelberg, 2012.

[16] Godard, Clément, O. Mac Aodha , and G. J. Brostow . "Unsupervised Monocular Depth Estimation with Left-Right Consistency." (2016).

[17] Pilzer, Andrea , et al. "Unsupervised Adversarial Depth Estimation Using Cycled Generative Networks." 2018 International Conference on 3D Vision (3DV) 2018.

[18] Eldesokey, Abdelrahman , M. Felsberg , and F. S. Khan . "Propagating Confidences through CNNs for Sparse Data Regression." (2018).

[19] Ma, Fangchang , G. V. Cavalheiro , and S. Karaman . "Self-supervised Sparse-to-Dense: Self-supervised Depth Completion from LiDAR and Monocular Camera." (2018).

[20] He, Kaiming , et al. "Deep Residual Learning for Image Recognition." (2015).

[21] Zhan, Huangying , et al. "Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction." (2018).

[22] Long, Jonathan, E. Shelhamer, and T. Darrell . "Fully Convolutional Networks for Semantic Segmentation." IEEE Transactions on Pattern Analysis & Machine Intelligence 39.4(2014):640-651.

[23] Garg, Ravi , et al. "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue." (2016).

[24] Zhou, Tinghui , et al. "Unsupervised Learning of Depth and Ego-Motion from Video." (2017).

[25] Lepetit, Vincent , F. Moreno-Noguer , and P. Fua . "EPnP: An AccurateO(n) Solution to the PnP Problem." International Journal of Computer Vision 81.2(2009):155-166.

[26] Geiger, A. , et al. "Vision meets robotics: The KITTI dataset." The International Journal of Robotics Research 32.11(2013):1231-1237.

[27] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In 15th European Conference on Computer Vision (ECCV) Workshops, 2018.

[28] https://arxiv.org/abs/1905.07542?context=cs

[29] Uhrig, Jonas , et al. " [IEEE 2017 International Conference on 3D Vision (3DV) - Qingdao (2017.10.10-2017.10.12)] 2017 International Conference on 3D Vision (3DV) - Sparsity Invariant CNNs." (2017):11-20.

[30] Jaderberg, Max , et al. "Spatial Transformer Networks." (2015).

[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

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