2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Stereo-augmented Depth Completion from a Single RGB-LiDAR image

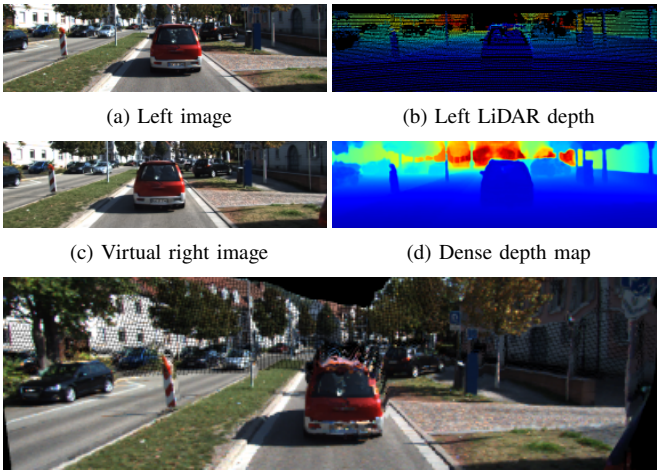
Keunhoon Choi[1†], Somi Jeong[1†], Youngjung Kim[2], Kwanghoon Sohn[1]

***Abstract*— Depth completion is an important task in computer vision and robotics applications, which aims at predicting accurate dense depth from a single RGB-LiDAR image. Convolutional neural networks (CNNs) have been widely used for depth completion to learn a mapping function from sparse to dense depth. However, recent methods do not exploit any 3D geometric cues during the inference stage and mainly rely on sophisticated CNN architectures. In this paper, we present a cascade and geometrically inspired learning framework for depth completion, consisting of three stages: view extrapolation, stereo matching, and depth refinement. The first stage extrapolates a virtual (right) view using a single RGB (left) and its LiDAR data. We then mimic the binocular stereo-matching, and as a result, explicitly encode geometric constraints during depth completion. This stage augments the final refinement process by providing additional geometric reasoning. We also introduce a distillation framework based on teacher-student strategy to effectively train our network. Knowledge from a teacher model privileged with real stereo pairs is transferred to the student through feature distillation. Experimental results on KITTI depth completion benchmark demonstrate that the proposed method is superior to state-of-the-art methods.**

I. Introduction

Perceiving dense and accurate depth information is a fundamental problem for various applications such as autonomous driving, robotic navigation, and 3D reconstruction. Consequently, active 3D sensors including structured-light, ToF, and LiDAR sensors are becoming very popular, which can measure depth information with errors of less than few centimeters. Among them, LiDAR is a dominating solution for outdoor environments thanks to its insensitivity to sunlight and long-range capability. However, it navigates a trade-off between resolution and cost, *i.e.*, the commercial LiDAR provides sparse depth measurements only (32-64 horizontal scan lines in the vertical direction), while the high-end device is prohibitively expensive. One promising attempt to address this trade-off is *depth completion*, where dense depth is estimated from a sparse LiDAR data.

With the success of deep learning, various depth completion approaches based on convolutional neural networks (CNNs) have been proposed. To reduce the uncertainty from highly sparse and irregular depth measurements, state-of-the-art methods have leveraged RGB image to guide the depth completion. In literature, the image-guided depth completion can be divided into two main categories: *single image-based* and *stereo-based* approach.



(a) Left image (b) Left LiDAR depth

(c) Virtual right image (d) Dense depth map

(e) 3D reconstruction

Fig. 1. Illustration of our depth completion process. Given a single (a) left image and (b) sparse LiDAR depth, a view extrapolation network generates (c) virtual right view image. Through the stereo matching and depth refinement networks, (d) dense depth map is estimated from the virtual stereo pair. We present the depth results as (e) 3D reconstruction.

The single image based approaches [1], [2], [3] denoted as *Mono-LiDAR*, take the sparse LiDAR data and its corresponding RGB image as inputs. Simply feeding these inputs may neglect 3D geometric constraints, which are essential for estimating the accurate depth. To alleviate this problem, some methods [4], [5], [6], [7], [8], [9] infer additional auxiliary information related to depth such as surface normal, confidence, and affinity to regularize the completion process and improve its accuracy. Despite their significant improvements, it is still a geometrically ambiguous problem due to the lack of direct prior knowledge related to the absolute depth values.

The stereo-based approaches [10], [11], [12], [13], [14] denoted as *Stereo-LiDAR*, exploit the LiDAR depth and stereo image pair. They estimate the accurate depth map by imposing distinct geometric constraints with the stereo cues. In practice, however, it is not preferred because calibration and synchronization problems often occur between stereo cameras or between camera and LiDAR sensors. Despite its high accuracy, these problems significantly limit the practical application of this formulation.

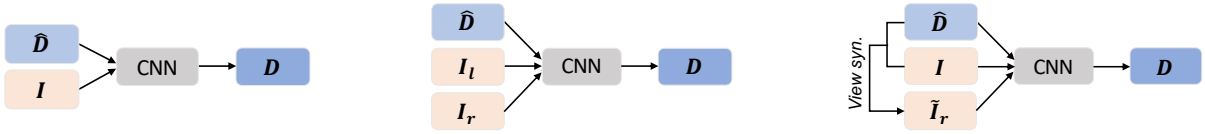
To tackle the aforementioned problems, in this paper, we present a novel stereo-augmented depth completion network (SDCNet) that infers high-precision depth maps from a single RGB and its LiDAR data by infusing stereo knowledge. The key insight is that we can formulate stereo image pairs by extrapolating a virtual (right) view using RGB and LiDAR

This research was supported by the Agency for Defense Development under the grant UD2000008RD. (*Corresponding author: Kwanghoon Sohn*)

[1]K. Choi, S. Jeong, and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, Korea. E-mail: {keunhoonchoi,somijeong,khsohn}@yonsei.ac.kr

[2]Y. Kim is with Agency for Defense Development (ADD), Daejeon 34060, Korea. E-mail: read12300@add.re.kr

[†]These authors contribute equally.



(a) Mono-LiDAR (b) Stereo-LiDAR (c) Our method

Fig. 2. Comparison of image-guided depth completion model. (a) Mono-LiDAR model takes a sparse LiDAR data $\hat{D}$ and a single image I, and (b) Stereo-LiDAR model takes $\hat{D}$ and stereo image pair I_l and I_r to produce the dense depth map D. (c) Our model takes the same input as the mono-LiDAR model and infers the depth output in the same way as the stereo-LiDAR model by synthesizing the right image $\tilde{I}_r$ based on $\hat{D}$ and I.

data. By this means, we can mimic the stereo setup, thus explicitly imposing the geometric constraints with the stereo cue. The network architecture is divided into three parts: (i) *view extrapolation* to generate the right view image based on the given LiDAR depth and image; (ii) *stereo matching* to infer a stereo depth value by capturing the geometric correspondence between the virtual stereo representations; and (iii) *depth refinement* to correct less confident value in the stereo depth based on inputs.

To train our network effectively, we introduce a distillation strategy [15], which consists of *teacher* and *student* models. The teacher model is trained to estimate the stereo depth using real stereo images and LiDAR data, where the real stereo pairs are defined as the privileged information. Then, the knowledge from the teacher is transferred to the student network, which is our full configuration. The student network exploits the initialized weights from the teacher model and tries to mimic the output of the teacher network. By doing so, the student network not only yields a high-precision depth from a single RGB-LiDAR data via pseudo binocular stereo-matching but also generates a high-quality right view image. We conduct various experiments on KITTI depth completion benchmark [16] to demonstrate that our proposed method outperforms the state-of-the-art methods. Figure 1 shows the obtained results over the process.

II. Related Work

A. Mono-LiDAR Depth Completion

To compensate for the ambiguity arising from sparse LiDAR measurements, the mono-LiDAR depth completion exploits a single image additionally to achieve rich information, as depicted in Figure 2 (a). Ma *et al*. [1] introduced a deep regression model that takes the concatenated depth and image as input and infers the dense depth, and extended it in a self-supervised training framework with a sequence of images [2]. Chen *et al*. [3] designed a 2D-3D fuse block to take advantage of 2D and 3D feature representations of image and depth. To address the problem of lack of direct prior knowledge of depth values, there have been several attempts to exploit depth-related auxiliary information. Some methods [4], [5], [6] modeled the geometric constraints between the predicted surface normal and depth by using the locally linear orthogonality. Others [7], [8], [9] introduced the depth refinement approaches by learning local/non-local affinities. Compared to these approaches that exploit the inferred depth-related information, our method improves performance by directly providing the geometric constraints through the virtual stereo pair as shown in Figure 2 (c).

B. Stereo-LiDAR Depth Completion

The stereo camera and LiDAR sensor are widely used passive and active depth sensors. The depth from stereo cameras is dense but error-prone, while the depth from the LiDAR sensor is accurate but sparse. The stereo-LiDAR depth completion takes advantage of their complementary characteristics to estimate the dense depth, as shown in Figure 2 (b). Park *et al*. [10] attempted to estimate the high-precision disparity map by fusing sparse LiDAR data into the coarse disparity map estimated from SGM [17], and extended it to exploit uncalibrated stereo and LiDAR data [13]. Wang *et al*. [12] employed a conditional cost volume normalization (CCVNorm) to regularize features of stereo RGB-LiDAR inputs. To reduce the difficulty to collect ground-truth depth maps, unsupervised [11] and self-supervised [14] stereo-LiDAR depth completion frameworks were proposed. Compared to the above methods, our method is more practical and simple since only a single RGB-LiDAR pair is required.

C. Generalized Distillation

The generalized distillation [15] is a learning strategy that combines transferring knowledge from teacher model to student model [18] and learning from privileged information [19]. The privileged information indicates the available information only in the training phase. The teacher network can enrich the feature representation from the privileged information and transfer more informative knowledge to the student network. Especially, some monocular depth estimation methods adopt the distillation to train the compact depth model [20], optimize multi-task student network from multi-teacher networks [21], and distill knowledge from stereo depth network to train monocular depth network [22]. Our method differs from these works in that we utilize stereo images as the privileged information and transfer the knowledge obtained from stereo images to train our method effectively.

III. Proposed Method

We aim at inferring accurate dense depth maps D from a single RGB image I_l and sparse LiDAR depth $\hat{D}$. To provide direct geometric reasoning, we propose a Stereo-augmented Depth Completion Network (SDCNet) consisting of *view extrapolation*, *stereo matching*, and *depth refinement* networks, as illustrated in Figure 3. We design a teacher-student strategy taking advantage of the privileged information. The following section details how to train SDCNet based on the novel learning strategies as depicted in Figure 4.

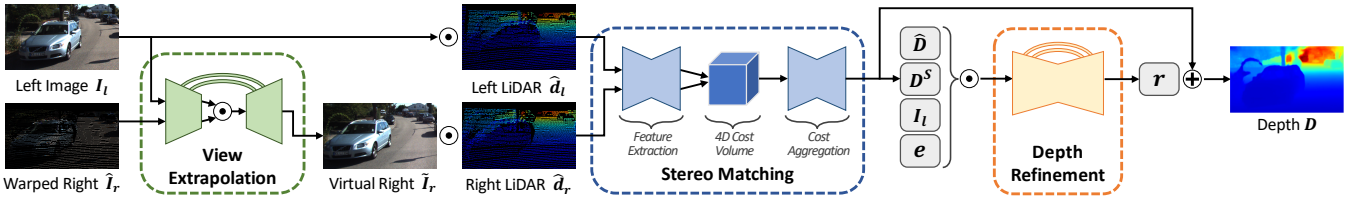


Fig. 3. Details of SDCNet. We first generate a virtual right image $\tilde{I}_r$ using the left image I_l and the sparse right image $\hat{I}_r$ in the view extrapolation network. The stereo depth map D^S is estimated using the virtual stereo RGB-LiDAR pair $(I_l \odot \hat{d}_l, \tilde{I}_r \odot \hat{d}_r)$ in the stereo matching network. Lastly, with the concatenated inputs $(\hat{D}, D^S, I_l, e)$, the depth refinement network infers the residual signals r to refine D^S. Note that $\odot$ and $\oplus$ indicate the concatenation and element-wise summation operators, respectively.

A. Teacher Model

We first train the teacher model that consists of stereo matching network using a real stereo RGB-LiDAR pair, as depicted in Figure 4 (a). By considering the real stereo data as the privileged information, it can estimate the accurate inverse depth output (disparity) by finding the corresponding points between the left and right data. Inspired by CCVNorm [12], we form the left and right sparse LiDAR disparity inputs $(\hat{d}_l, \hat{d}_r)$ as follows. We first project the LiDAR points $\hat{D}$ to the left and right image (I_l, I_r) coordinates based on the triangulation, and then convert them into the disparity maps $(\hat{d}_l, \hat{d}_r)$. Same as the conventional stereo matching methods [23], [24], it consists of feature extraction and cost aggregation steps. The concatenated RGB-LiDAR stereo pair is forwarded to the feature extraction network and a 4D cost volume is created by calculating the correlation between the extracted left and right features horizontally. To aggregate the cost volume, we employ the cost aggregation network with a semi-global guided aggregation (SGA) layer proposed in GA-Net [23], and finally obtain a disparity d^T from the privileged real stereo pair.

To train the teacher model, we use a disparity loss $\mathcal{L}_{dis}$ to minimize the pixel-wise difference between d^T and a semi-dense ground-truth disparity map d_{gt}, expressed as

$$\mathcal{L}_{disp}(d^T, d_{gt}) = \left\| \mathbb{I}_{d_{gt}>0} \cdot (d^T - d_{gt}) \right\|_1 . \tag{1}$$

Here, we consider only valid pixels of d_{gt} given that the missing pixels may negatively affect overall performance. It encourages the teacher model to encode the geometric constraints from the real stereo inputs. Once the stereo matching network is trained, its weights are served for initializing the student. By this mean, it can effectively transfer the valuable knowledge gained from the real stereo pairs to the student.

B. Student Model

The goal of SDCNet is to leverage the 3D geometric cues from augmented stereo pairs with only a single RGB-LiDAR data. To train it effectively, we build the student model that exploits the distilled knowledge from the teacher, as shown in Figure 4 (b). It consists of the view extrapolation, stereo matching, and depth refinement networks. The view extrapolation network focuses on extrapolating the right view $\tilde{I}_r$ using I_l and $\hat{D}$ to mimic the stereo setup. The stereo matching and depth refinement networks estimate the accurate depth map from the virtual stereo pairs. Here, the stereo matching network has the same architecture as the

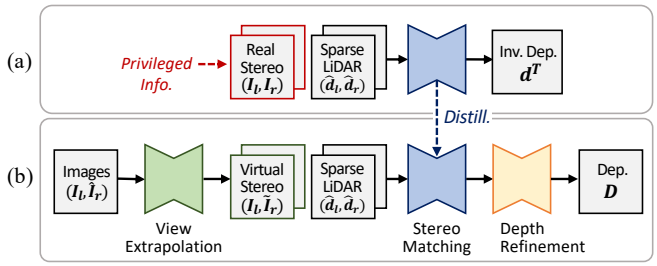


Fig. 4. Overall distillation scheme. (a) The teacher model is trained to estimate accurate depth using real stereo image and sparse LiDAR data. (b) The student model utilizes the distilled stereo knowledge from the teacher to generate well-synthesized virtual right image. Based on the virtual stereo pair, it estimates the depth by infusing the stereo knowledge.

teacher model, and its weights are initialized with the weights of the teacher. By this mean, it allows the teacher's depth estimation capacity to be transferred to the student.

Unlike other view extrapolation methods [27], [28] that generate the right view image based on the inherent depth information, we treat it as an image inpainting problem [29], whose goal is to fill the missing pixels of the image. Concretely, we generate the sparse right image $\hat{I}_r$ by moving the pixels in I_l along the valid inverse depth $\hat{d}_l$. It is expressed as $\hat{I}_r(i, j - \hat{d}_l(i,j)) = I_l(i,j)$, where i, j refer to the row and column on the image. The view extrapolation network has an encoder-decoder architecture with skip-connections [30]. The encoder takes I_l and $\hat{I}_r$ as inputs, and the extracted features of I_l and $\hat{I}_r$ are concatenated. The concatenated features are passed to the decoder through skip-connections, and then the decoder outputs the virtual right image $\tilde{I}_r$. As will be seen in our experiments, the sparse pixels in $\hat{I}_r$ are properly propagated to adjacent pixels to fill the blank pixels, and as a result, $\tilde{I}_r$ can be replaced with I_r.

Using the augmented stereo images $(I_l, \tilde{I}_r)$ and the sparse LiDAR inputs $(\hat{d}_l, \hat{d}_r)$, the stereo disparity map d^S is inferred from the stereo matching network, in which the initial weights come from the teacher. Then, d^S is converted to the depth value D^S based on the given camera parameters. In that D^S is predicted from the synthesized stereo images, it may contain inaccurate values. To alleviate this problem, we adopt the depth refinement network that estimates the residual depth values of D^S, which is inspired by Pang *et al.* [24]. Concretely, we warp $\tilde{I}_r$ according to D^S and obtain the synthesized left image $\tilde{I}_l$. Here, we use bilinear interpolation [31] as the warping operation. The concatenation of $\hat{D}$, D^S, I_l, and absolute error $e = |I_l - \tilde{I}_l|$ is forwarded into the depth refinement network, and outputs the corresponding

TABLE I

QUANTITATIVE EVALUATION ON KITTI DEPTH COMPLETION TEST DATASET. [25]

Method	RMSE ↓	MAE ↓	iRMSE ↓	iMAE ↓	Run-time [s]
CSPN [7]	1019.64	279.46	2.93	1.15	1
NConv [26]	829.98	233.26	2.60	1.03	0.02
S2D [2]	814.73	249.95	2.80	1.21	0.08
PwP [5]	777.05	235.17	2.42	1.13	0.1
DeepLiDAR [6]	758.38	226.50	2.56	1.15	0.07
FuseNet [3]	752.88	221.19	2.34	1.14	0.09
CSPN++ [8]	**743.69**	**209.28**	2.07	0.90	0.2
NLSPN [9]	**741.68**	**199.59**	**1.99**	**0.84**	0.22
Ours	752.36	238.43	**2.04**	**0.82**	0.34

residual depth values r. Finally, we obtain the high-precision depth map D by summing the residual r and D^S.

We adopt four loss functions to force the student to learn from the teacher's knowledge; disparity distillation loss $\mathcal{L}_{distill}$, KL divergence loss $\mathcal{L}_{KL}$, image reconstruction loss $\mathcal{L}_{im}$, and depth refinement loss $\mathcal{L}_{ref}$. By reducing the difference between outputs from the teacher and student model, the student can mimic the teacher model. In other words, the outputs of the teacher are regarded as soft labels, providing informative cues for the student.

Specifically, the disparity distillation loss $\mathcal{L}_{distill}$ aims to make the disparity from the real and virtual stereo image pairs identical. The estimated disparity of the teacher d^T may inevitably contain inaccurate depth values. To prevent such unreliable values from distilling to the student, we utilize the confidence map [32], [33], whose value $C^T(p) \in [0,1]$ indicates the reliable degree of d^T. Based on the confidence threshold τ, we handle the distilled values, expressed as

$$\mathcal{L}_{distill}(d^S, d^T) = \left\| \mathbb{I}_{C^T \geq \tau} \cdot (d^S - d^T) \right\|_1, \tag{2}$$

where d^S and d^T indicate the estimated disparity maps from the student and teacher model using the virtual and real stereo data. The KL divergence loss $\mathcal{L}_{KL}$ enforces the feature representations of I_r and $\tilde{I}_r$ close, defined as

$$\mathcal{L}_{KL}(F^S, F^T) = \sum F^T \log(F^T / F^S), \tag{3}$$

where F^S and F^T indicate the feature representations of $\tilde{I}_r$ and I_r, which are extracted using the pretrained VGG-16 network [34]. These losses force $\tilde{I}_r$ to play the same role as I_r, thus facilitating to produce the precise right image.

We additionally exploit the ground-truth data I_r and D_{gt} to train the student model. For the view extrapolation network, we introduce image reconstruction loss $\mathcal{L}_{im}$ to reduce the photometric difference between $\tilde{I}_r$ and I_r. It consists of a combination of L1 loss and a structured similarity (SSIM) loss [35], defined as

$$\mathcal{L}_{im}(\tilde{I}_r, I_r) = \alpha \frac{1 - \mathrm{SSIM}\left(\tilde{I}_r - I_r\right)}{2} + (1-\alpha) \left\| \tilde{I}_r - I_r \right\|_1, \tag{4}$$

where α controls a relative importance between them. For the depth refinement network, we adopt the depth refinement loss $\mathcal{L}_{ref}$, which aims to minimize the difference between the refined depth D and the semi-dense ground-truth depth D_{gt}. Same as Equation 1, it is expressed as

$$\mathcal{L}_{ref}(D, D_{gt}) = \left\| \mathbb{I}_{D_{gt}>0} \cdot (D - D_{gt}) \right\|_1. \tag{5}$$

IV. EXPERIMENTS

A. Experimental Settings

1) Dataset: We used KITTI depth completion benchmark [16] to train and evaluate our method. To be specific, we split images into 42,949 sets for training, 1,000 sets for validation, and 1,000 sets for test. Note that each training and validation set consists of RGB stereo images and synchronized LiDAR data with its semi-dense ground-truth data, and test set consists of only single RGB image and LiDAR data. To ignore regions without LiDAR data, the input images are cropped to 1216×256 for training and test.

2) Evaluation metrics: Following the KITTI benchmark, we utilize four metrics for the quantitative evaluation: Root Mean Square Error(RMSE), Mean Absolute Error(MAE) for depth in millimeter unit and inverse of them (iRMSE, iMAE) in inverse kilometer unit which are related with disparity.

- $RMSE(mm) = \sqrt{\frac{1}{\mathcal{V}} \sum_{v \in \mathcal{V}} \left| (D - D_{gt}) \right|^2}$
- $MAE(mm) = \frac{1}{\mathcal{V}} \sum_{v \in \mathcal{V}} \left| (D - D_{gt}) \right|$
- $iRMSE(1/km) = \sqrt{\frac{1}{\mathcal{V}} \sum_{v \in \mathcal{V}} \left| (1/D - 1/D_{gt}) \right|^2}$
- $iMAE(1/km) = \frac{1}{\mathcal{V}} \sum_{v \in \mathcal{V}} \left| (1/D - 1/D_{gt}) \right|$

where $v \in \mathcal{V}$ indicates valid ground-truth pixels.

3) Implementation details: Our network was implemented in Pytorch [36] and trained with NVIDIA RTX GPU with 24G of RAM. The weights of networks are initialized by a zero-mean Gaussian random initialization, and the Adam optimizer [37] was employed for optimization, where $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Additionally, the initial learning rate is 10^{-4} and is halved at every 10 epochs, and the batch size is 1. We set the parameters as $\alpha = 0.85$ and $\tau = 0.15$ empirically. We trained the teacher model with real stereo data for 10 epochs and the student model for 30 epochs. Here, the weights of the stereo matching network in the student model are initialized with the weights of the teacher, and its learning rate is set 0.1 times lower than other networks for distilling the knowledge effectively.

B. Comparison with State-of-the-Art Methods

We evaluate our method with state-of-the-art depth completion methods on the KITTI depth completion quantitatively and qualitatively. We compare our method only with those published in the paper.

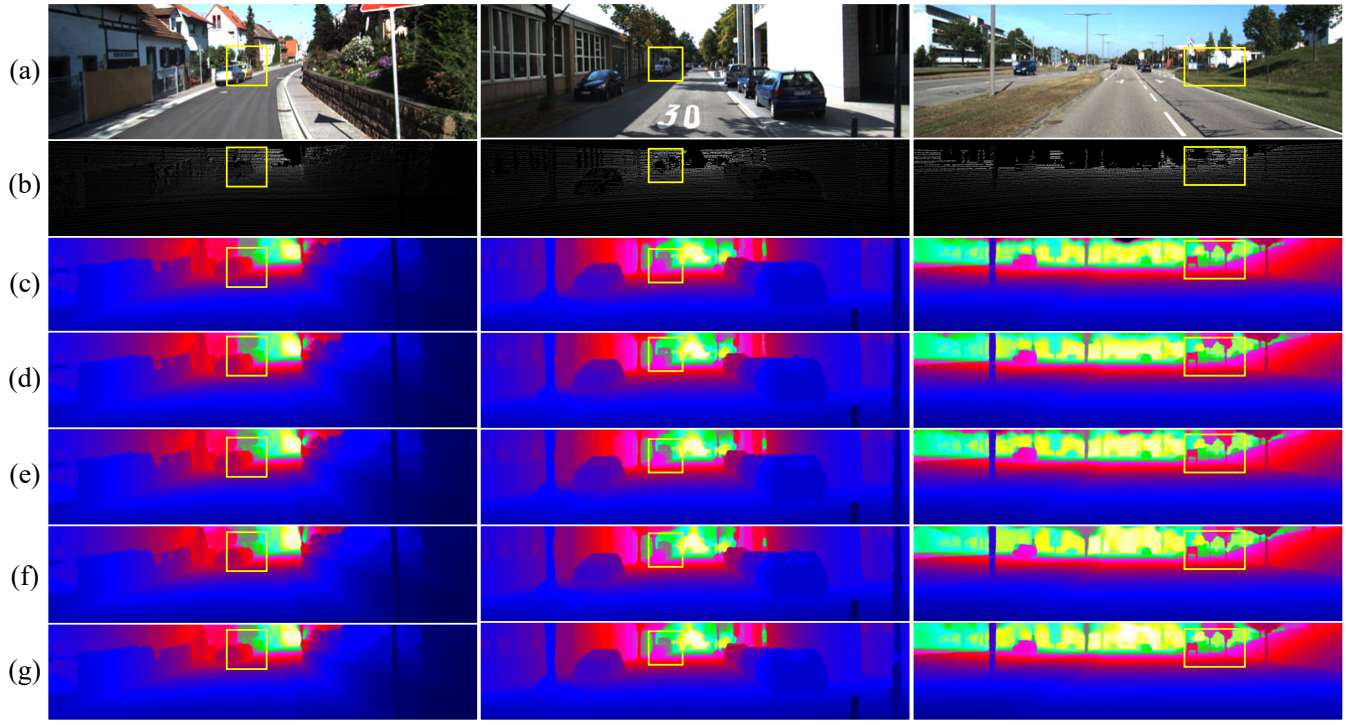


Fig. 5. Qualitative comparison with state-of-the-art methods on KITTI depth completion test set. (a) RGB image, (b) Sparse LiDAR, (c) S2D [2], (d) DeepLiDAR [6], (e) CSPN++ [8], (f) NLSPN [9], (g) Ours.

TABLE II

ABLATION STUDY FOR VIEW EXTRAPOLATION IN VALIDATION DATASET.

Method	PSNR ↑	SSIM ↑
w/o Distill.	30.7487	0.8660
w/ Distill.	32.8779	0.9128

TABLE III

ABLATION STUDY FOR DEPTH COMPLETION IN VALIDATION DATASET.

Method	RMSE ↓	MAE ↓	iRMSE ↓	iMAE ↓
Teacher	749.36	252.58	1.40	0.81
w/o Distill.	832.00	243.78	2.83	1.15
w/ Distill.	772.26	260.75	1.52	0.83

1) Quantitative Evaluation: Table I shows the quantitative results of the KITTI depth completion benchmark dataset, which are arranged in descending order by the RMSE metric. Among those methods in Table I, the methods of explicitly considering geometric constraints [8], [9] show stronger results than those that do not ([2], [3]). For similar reasons, our method can achieve good performance because it directly utilizes the inherent stereo geometric cues from the virtual stereo image pairs. Although it is not the highest performance in RMSE metric, our method perform well in iRMSE or iMAE. Contrast to RMSE and MAE, inverse metric is a metric that represents the error of disparity, the reciprocal of depth, so it represents the close distance error better. From this point of view, it can be seen that our method has greater far-field error compared to other SOTA methods in RMSE, but better near-field performance. It is probably because our method uses the stereo-matching manner to estimate disparity. Although the refinement network correct error of distance, it still exploit D^S which is estimated from the stereo-matching network. It seems that better results can be obtained if LiDAR data ($\hat{D}$) are leveraged more actively in the final depth refinement stage by using Spatial Propagation Network [7] used in SOTA method [8], [9] together with our method.

2) Qualitative Evaluation: Figure 5 shows some visual comparison results of predicted depth maps with several state-of-the-art methods. Our results are shown in the last row. We highlighted challenging areas with yellow boxes. Compared to the other methods (Figure 5(c)-(f)), our method (Figure 5(g)) refines those boundary areas well based on the stereo geometrical constraints.

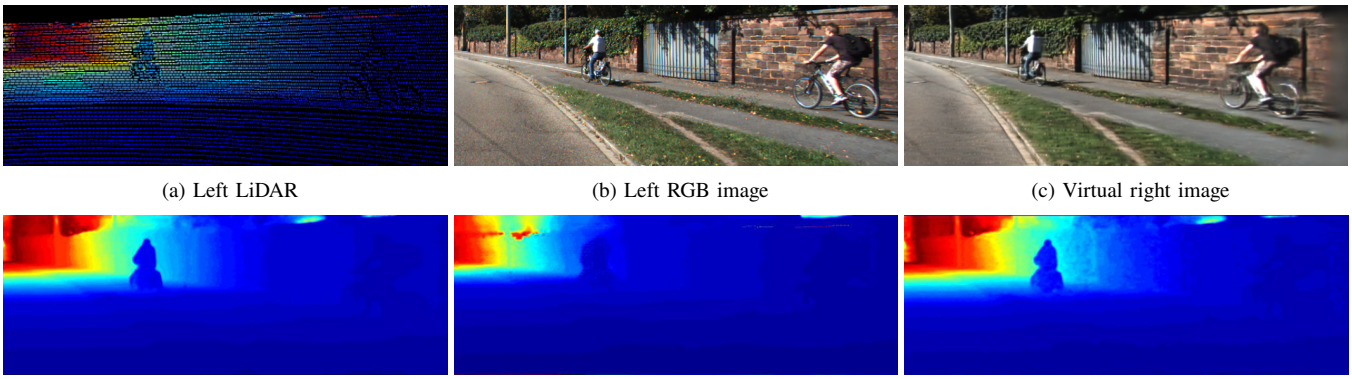
C. Ablation Study

To validate the learning components within our method, we perform the ablation study about view extrapolation and depth completion.

1) View Extrapolation: We train the view extrapolation network using only image reconstruction loss $\mathcal{L}_{im}$ as 'w/o Distill.' and train it using our distillation strategy as 'w/ Distill.'. As shown in Figure 6, the results of the extrapolated virtual view recover the details of the right image properly. Even though a lot of information is missing in $\hat{I}_r$, the valid pixels in $\hat{I}_r$ serves as a hint on how to propagate it to adjacent pixels. Moreover, the receptive field of the view extrapolation network is quite large, allowing the features missing from $\hat{I}_r$ to be passed by replacing it with the features in I_l. By doing so, it helps to synthesize the plausible $\tilde{I}_r$ by filling the proper pixel values in missing regions. When it comes to fine details, the results of 'w/o Distill.' tend to be smooth and lose

(a) Ground-truth right view (b) w/o Distill. (c) w/ Distill. (d) View Details

Fig. 6. Ablation study for view extrapolation.



(a) Left LiDAR (b) Left RGB image (c) Virtual right image

(d) Teacher (e) w/o Distill. (f) w/ Distill.

Fig. 7. Ablation study for depth completion.

high-frequency components. Compared to it, the results of 'w/ Distill.' show more natural results. This is because the network of 'w/ Distill.' is optimized to generate the same depth map with real stereo pairs, so it can recover more accurate and realistic images.

We utilized two error metrics to evaluate the effectiveness of distillation in our view extrapolation network quantitatively; Peak Signal-to-Noise Ratio (PSNR) and and Structural Similarity (SSIM) [35]. The quantitative comparisons were shown in Table II. It demonstrates that our teacher-student learning strategy is effective to learn the proper view extrapolation network, which facilitates the mono-LiDAR depth completion problem as stereo-LiDAR depth completion.

2) Depth Completion: To understand the contribution of each component, we conduct the ablation study for depth completion as follows. We indicate the obtained depth from the teacher network using the real stereo data as 'Teacher', which can be seen as stereo-LiDAR completion. Using the mono-LiDAR data, we train our SDCNet from scratch without any teacher model as 'w/o Distill.', and our network with the distilled teacher's knowledge as 'w/ Distill.'. As shown in Table III, 'Teacher' outperforms since it takes advantage of the real stereo data, which can explicitly encode geometric constraints. Therefore, it can be regarded as the upper bound performance for our method. Compared to 'w/o Distill.', the transferred knowledge from the teacher is valuable for encoding geometric reasoning, thus improving the capacity of the student. It demonstrates the propriety of the proposed distillation scheme.

Figure 7 shows the estimated depth results of our ablation study. Compared to 'Teacher', 'w/o Distill.' shows the poor performance. It may occur due to the mis-synthesized right image and the lack of the teacher's knowledge. On the other hands, 'w/ Distill.' has comparable performance to 'Teacher' without real stereo data. It demonstrates that our distillation strategy has a positive effect on the entire network.

V. Conclusion

We present a novel mono-LiDAR depth completion method by leveraging conventional stereo knowledge. The key idea is to mimic the binocular stereo setup that explicitly encodes the geometric constraint. To this end, we adopt a novel teacher-student distillation strategy to generate the realistic right view image. The teacher network is trained using real stereo images as privileged information, and transfer its knowledge to the student network. By doing so, it enables the student network to learn not only view extrapolation but also depth completion effectively. We demonstrate the effectiveness of our methods in various experiments. We show for the first time that the mono-LiDAR depth completion can be solved as the stereo-LiDAR depth completion. The direction for further study is to reduce the computational complexity and enhance the depth quality by exploiting the sequential data.

REFERENCES

[1] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *ICRA*, 2018.

[2] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *ICRA*, 2019.

[3] Y. Chen, B. Yang, M. Liang, and R. Urtasun, "Learning joint 2d-3d representations for depth completion," in *ICCV*, 2019.

[4] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," in *CVPR*, 2018.

[5] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, "Depth completion from sparse lidar data with depth-normal constraints," in *ICCV*, 2019.

[6] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *CVPR*, 2019.

[7] X. Cheng and R. Wang, Peng aand Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *ECCV*, 2018.

[8] X. Cheng, P. Wang, C. Guan, and R. Yang, "CSPN++: Learning context and resource aware convolutional spatial propagation networks for depth completion." in *AAAI*, 2020, pp. 10 615–10 622.

[9] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. S. Kweon, "Non-local spatial propagation network for depth completion," in *ECCV*, 2020.

[10] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation with the 3d lidar and stereo fusion," in *ICRA*, 2018.

[11] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, "Noise-aware unsupervised deep lidar-stereo fusion," in *CVPR*, 2019.

[12] T.-H. Wang, H.-N. Hu, C. H. Lin, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization," in *IROS*, 2019.

[13] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation using uncalibrated lidar and stereo fusion," *IEEE TITS*, vol. 21, no. 1, pp. 321–335, 2020.

[14] J. Zhang, M. S. Ramanagopal, R. Vasudevan, and M. Johnson-Roberson, "Listereo: Generate dense depth maps from lidar and stereo imagery," in *ICRA*, 2020.

[15] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," in *ICLR*, 2016.

[16] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[17] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE TPAMI*, vol. 30, no. 2, pp. 328–341, 2007.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[19] V. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer." *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2023–2049, 2015.

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

[21] J. Ye, Y. Ji, X. Wang, K. Ou, D. Tao, and M. Song, "Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more," in *CVPR*, 2019.

[22] J. Cho, D. Min, Y. Kim, and K. Sohn, "A large rgb-d dataset for semi-supervised monocular depth estimation," *arXiv preprint arXiv:1904.10230*, 2019.

[23] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *CVPR*, 2019.

[24] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *ICCV*, 2017.

[25] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *3DV*, 2017.

[26] A. Eldesokey, M. Felsberg, and F. S. Khan, "Confidence propagation through cnns for guided sparse depth regression," *IEEE TPAMI*, 2019.

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

[28] J. Lee, H. Jung, Y. Kim, and K. Sohn, "Automatic 2d-to-3d conversion using multi-scale deep neural network," in *ICIP*, 2017.

[29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[31] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *NeurIPS*, 2015.

[32] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure." in *BMVC*, 2016.

[33] S. Kim, D. Min, S. Kim, and K. Sohn, "Unified confidence estimation networks for robust stereo matching," *IEEE TIP*, vol. 28, no. 3, pp. 1299–1313, 2018.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[36] 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," in *NeurIPS Workshop*, 2017.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.