Expanding Sparse LiDAR Depth and Guiding Stereo Matching for Robust Dense Depth Estimation

Zhenyu Xu, Yuehua Li, *Member, IEEE*, Shiqiang Zhu, and Yuxiang Sun, *Member, IEEE*

***Abstract*—Dense depth estimation is an important task for applications, such as object detection, 3-D reconstruction, etc. Stereo matching, as a popular method for dense depth estimation, has been faced with challenges when low textures, occlusions or domain gaps exist. Stereo-LiDAR fusion has recently become a promising way to deal with these challenges. However, due to the sparsity and uneven distribution of the LiDAR depth data, existing stereo-LiDAR fusion methods tend to ignore the data when their density is quite low or they largely differ from the depth predicted from stereo images. To provide a solution to this problem, we propose a stereo-LiDAR fusion method by first expanding the sparse LiDAR depth to a semi-dense depth with RGB image as reference. Then, based on the semi-dense depth, a varying-weight Gaussian guiding method is proposed to deal with the varying reliability of guiding signals. A multi-scale feature extraction and fusion method is further used to enhance the network, which shows superior performance over traditional sparse invariant convolution methods. Experimental results on different public datasets demonstrate our superior accuracy and robustness over the state of the arts.**

***Index Terms*—Computer vision for automation, sensor fusion, AI-based methods.**

I. Introduction

Dense depth estimation is an important task in robotic vision. It plays a crucial rolein various applications, such as object detection and 3-D reconstruction, etc. Stereo matching is one way to estimate dense depth. Traditional methods basically follow the procedure including matching-cost calculation, cost aggregation, disparity iteration and results refinement. Recently, deep learning [1] has been employed in stereo matching. Some end-to-end networks based on the encoder-decoder structure with 2-D convolution have been proposed [2]. Some recent state-of-the-art (SOTA) methods [3], [4] adopt the 3-D convolutional neural network (CNN). Although promising performance has been achieved by the existing methods, there are still some open challenges, for example, low textures and occlusions in the environments, as well as domain gaps especially when there are no sufficient training data. Although there are some solutions, such as domain translation [5], feature-layer normalization [6], multi-scale feature extraction [7], the performance is still not satisfactory.

As an accurate distance-measurement device, 3-D LiDAR provides another way to obtain depth data. It is capable of producing accurate depth in low-texture environments and performs well across different domains. However, it is still very expensive to obtain dense depth data with LiDARs, because the price significantly increases with the increasing number of light beams. Although LiDARs can only produce sparse depth measurements, the depth data are still valuable to dense depth estimation. Some works have adopted the idea of incorporating the sparse LiDAR depth into depth estimation to produce dense depth [8], [9]. But existing methods mainly focus on the regions with LiDAR points. For the regions without LiDAR points, the depth estimation performance is not satisfactory. To deal with the challenges in domain adaption and depth completion, stereo-LiDAR fusion has attracted great interests in recent years. For example, directly fusing the sparse LiDAR points in the input [10] or feature layer [11], [12], and directly using the sparse LiDAR points to guide the stereo matching [13]. These two methods have become the mainstream ways.

Inspired by [14], which tackles low density and imbalanced distribution problem by constructing a network to expand the sparse LiDAR data, we propose a much lighter and more flexible expansion scheme to expand sparse LiDAR points to a semi-dense depth maps using RGB information. Then, the cooperation among further data fusion, stereo-matching guiding and the proposed expansion scheme is explored. Instead of modifying sparse invariant convolution [15] to a multi-scale version [16], we combine the sparse expansion with normal multi-scale feature extraction [7], [17] to obtain the sparse features for further data fusion. A novel stereo-matching guiding method is also proposed to absorb the expanded data in the guidance rather than using the original sparse data only [13]. Finally, the dense depth can be obtained by cost aggregation through the commonly used cascaded 3D CNN [7], [18]. The experimental results demonstrate that the sparse expansion can not only improve the stereo-matching guiding but also boost the performance of sparse feature fusion. Even with small percentage of sparse inputs, the proposed network notably increases the robustness

Manuscript received 21 August 2022; accepted 10 January 2023. Date of publication 26 January 2023; date of current version 3 February 2023. This letter was recommended for publication by Associate Editor A. I. Comport and Editor C. C. Lerma upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant U21B6001, in part by the National Key Research and Development Program of China under Grant 2018AAA0102700, in part by the Ten Thousand Talents Program of Zhejiang Province under Grant 2019R51010, and in part by the Stable Support Project of State Administration of Science, Technology and Industry for National Defence Grant, PRC under Grant HTKJ2019KL502005. *(Corresponding author: Yuehua Li.)*

Zhenyu Xu is with the Research Center of Intelligent Robotics, Zhejiang Lab, Hangzhou 311121, China (e-mail: xuzhenyu@zhejianglab.com).

Yuehua Li is with the Research Center of Beijing, Zhejiang Lab, Hangzhou 311121, China (e-mail: liyh@zhejianglab.com).

Shiqiang Zhu is with Zhejiang Lab, Hangzhou 311121, China (e-mail: zhusq@zhejianglab.com).

Yuxiang Sun is with The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: sun.yuxiang@outlook.com).

Our code is available at: https://github.com/xzy-yqm/EG-Depth.

Digital Object Identifier 10.1109/LRA.2023.3240093

against domain shifts. The main contributions of this letter are summarized as follows:

1) We propose a novel light-weight sparse expansion scheme to flexibly use the RGB information and the corresponding sparse LiDAR points to generate a semi-dense depth map.
2) We propose a novel varying-weight Gaussian guiding method to exploit the expanded points together with original LiDAR points for cost volume aggregation guidance.
3) We are the first to combine the normal multi-scale feature extraction with sparse data expansion to deal with sparse feature extraction, which shows superior performance than sparse invariant convolution.

II. Related Work

A. Stereo Matching

The existing end-to-end stereo matching networks mainly adopt the encoder-decoder structure. They can be generally divided into 2-D CNN-based methods [2] and 3-D CNN-based methods [3]. The 3-D CNN-based methods, such as [4], [7], are leading the KITTI lead board. However, compared to 2-D convolution, 3-D convolution requires more memories and computation resources. By utilizing simplified cost aggregation [19] or coarse-to-fine cascaded convolution [18], the computation load can be reduced at the promise of depth estimation accuracy.

As aforementioned, stereo matching faces some problems, such as low textures, occlusions, and domain gaps. To deal with the former two problems, Rao et al. [20] incorporated global texture information into depth estimation, while edge-stereo [21] exploits the edge cues to enhance the performance. For the latter, input domain transformation [5], feature normalization [6] and multi-scale feature extraction [7], [18] are popular methods.

B. Depth Completion

Depth completion with 3-D LiDAR points is another way to recover dense depth data. Compared to depth prediction methods [22], [23] based on monocular images, depth completion networks [24], [25] could produce depth with higher accuracy, because the depth information of some sparse points is prior-known. With the available sparse depth information, Uhrig et al. [15] proposed sparse invariant convolution, which proved to be more effective than ordinary convolution.

The KITTI depth completion benchmark has been widely used for evaluating depth estimation performance with sparse LiDAR points as input. Performance increment can be observed when RGB images are introduced to the depth completion task [26]. Most existing depth completion works [24], [25] focus on robust depth estimation with both RGB images and sparse LiDAR points. With RGB information, experimental results from [11], [27] demonstrate that the ordinary convolution tends to produce more robust results than the sparse invariant one. From the results of [11], [27], it can be observed that these methods focus on regions with LiDAR points distributed. Extrapolation to regions without sparse points is still an open question.

Different from the existing works, our work regards the sparse LiDAR points as complementary information to enhance stereo matching in the way of fusion and guidance. Motivated by depth completion, the sparse LiDAR points are expanded to semi-dense maps with the guidance of RGB image for further processes. Besides, the feature extraction method is also studied to get better semi-dense depth features.

C. Stereo-LiDAR Fusion

With stereo images and sparse LiDAR points, stereo-LiDAR fusion networks are supposed to produce more robust and accurate dense depth estimation results. There are mainly two ways to incorporate sparse LiDAR points with stereo images: fusing the sparse points in different stages, and utilizing the sparse points to guide the cost aggregation of cost volumes.

The work [10] directly fuses the sparse LiDAR points in the input stage by projecting them to the image plane. The networks [12], [28] extract the sparse features in 2-D and 3-D, respectively. Similar to depth completion, Zhang et al. [11] used ordinary convolution for sparse feature extraction and got superior results than the sparsity-invariant one. In our work, we make a step further by adopting multi-scale feature extraction to obtain features with more diversity receptive fields.

For using sparse LiDAR points as guidance, Wang et al. [10] utilized the sparse information for conditional normalization of the cost volume. Paggi et al. [13] exploited it to conduct a Gaussian modulation for cost volume along the depth dimension. Huang et al. [14] introduced a learnable network to expand sparse LiDAR points in the height and width dimensions of images for further guidance, which inspires us to do the expansion before fusion and guidance. But different from the existing work [14] that constructs a network to learn the patterns of sparse LiDAR expansion, we propose a simple yet effective expansion method with the guidance of RGB images.

III. The Proposed Network

A. Network Overview

Fig. 1 displays the overview architecture of our proposed dense depth estimation network. It mainly consists of four parts: 1) the sparse expansion (SE) module, which expands the sparse LiDAR points with the guidance of an RGB image to a semi-dense depth map; 2) the multi-scale feature extraction and fusion (MFEF) module, in which the features of the input stereo images and the features of the semi-dense depth map are extracted respectively. The extracted features are then fused by concatenation at multiple scales; 3) varying-weight Gaussian guiding (VWGG) module, which generates varying Gaussian distributions depending on the reliability of the semi-dense depth map to modulate the 4-D cost volumes for guiding cost aggregation. The cost volume refers to the volume constructed by correlation and concatenation between features from the two images and the semi-dense depth map; 4) cascaded 3-D CNN is adopted along with the multi-scale features to get the final depth map.

B. Sparse Expansion (SE) Module

Depth completion works, such as [12], [28], have shown powerful capability to recover depth from monocular images and sparse LiDAR points. The work [14] implies that the expansion in the height and width dimension of sparse depth points is

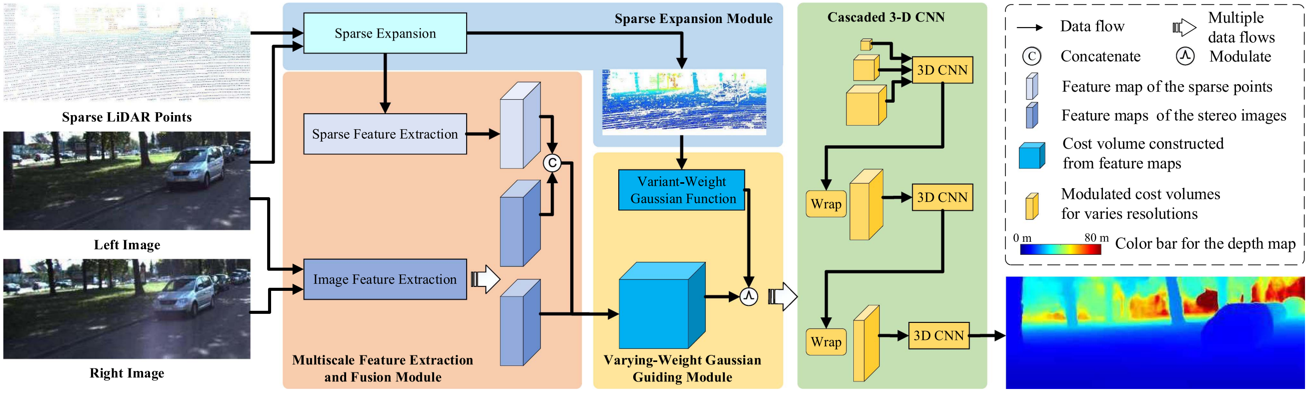
Sparse LiDAR Points
Left Image
Right Image
Sparse Expansion Module
Sparse Expansion
Sparse Feature Extraction
Image Feature Extraction
Multiscale Feature Extraction and Fusion Module
Variant-Weight Gaussian Function
Varying-Weight Gaussian Guiding Module
Cascaded 3-D CNN
3D CNN
Wrap
3D CNN
Wrap
3D CNN
Data flow
Multiple data flows
Concatenate
Modulate
Feature map of the sparse points
Feature maps of the stereo images
Cost volume constructed from feature maps
Modulated cost volumes for varies resolutions
0 m
80 m
Color bar for the depth map

Fig. 1. The architecture of our proposed network. It consists of four modules: sparse expansion (SE) module, multi-scale feature extraction and fusion (MFEF) module, varying-weight Gaussian guiding (VWGG) module, and cascaded 3-D CNN module. We initially expand the sparse LiDAR points with the guidance of an RGB image. Based on the semi-dense depth map, we use the MFEF module to obtain features from the LiDAR data and stereo images. After fusing the features, the cost volume is constructed and guided by the semi-dense depth map with the VWGG module. Finally, the depth map can be obtained after cost aggregation through the cascaded 3-D CNN. The figure is best viewed in color.

Algorithm 1. The Sparse Expansion Algorithm.

```
Input: RGB image $\mathbf{I}$ and LiDAR projected image $\mathbf{D}$
begin
    for $\mathbf{D}(x, y) > 0,\ x \in [0, W-1], y \in [0, H-1]$: do
        for $\mathbf{D}(x+\alpha, y+\beta) = 0,\ \alpha, \beta \in [-R, R]$: do
            if $0 \leq x+\alpha < W$ and $0 \leq y+\beta < H$: then
                $d_{rgb} \leftarrow \frac{1}{C}\sum_{c=0}^{C}|\mathbf{I}(x,y,c) - \mathbf{I}(x+\alpha, y+\beta, c)|$
                if $d_{rgb} < t$ : then
                    $\mathbf{D_{exp}}(x+\alpha, y+\beta) \leftarrow D(x, y)$
                end
            end
        end
    end
    for $x \in [0, W-1],\ y \in [0, H-1]$: do
        Initialize: $\mathbf{D_{all}}(x, y) = 0$
        if $\mathbf{D}(x, y) > 0$: then
            $\mathbf{D_{all}}(x, y) \leftarrow \mathbf{D}(x, y)$
        end
        if $\mathbf{D_{exp}}(x, y) > 0$: then
            $\mathbf{D_{all}}(x, y) \leftarrow \mathbf{D_{exp}}(x, y)$
        end
    end
end
```

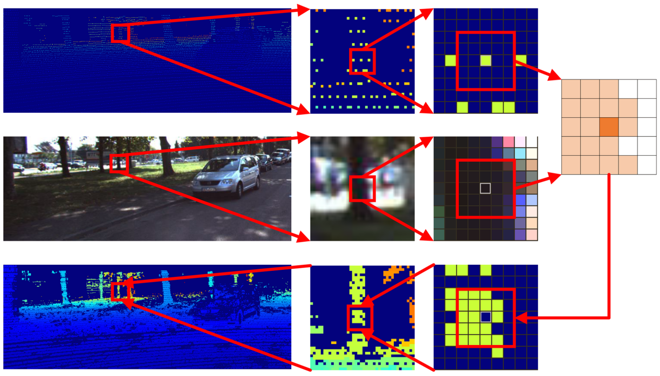

Fig. 2. The procedure of sparse expansion. The valid depth value of the projected LiDAR image is expanded to its invalid neighbourhood when the mean difference of the value on all RGB channels is less than a predefined threshold. The figure is best viewed in color.

beneficial for guiding stereo matching. This motivate us to find a simple yet effective way to do the expansion of sparse LiDAR points with monocular visual cues.

The procedure of sparse expansion is illustrated in Fig. 2. For simplicity, we project points in the LiDAR coordinate frame into the camera frame using the extrinsic and intrinsic parameters. Thus, the input of the sparse expansion module is an RGB image $\mathbf{I}$ and a LiDAR projected image $\mathbf{D}$, where $\mathbf{I}(x, y, c)$ represents the value of each channel $c \in [0, C-1]$ at the pixel location $(x, y), x \in [0, W-1], y \in [0, H-1]$, and $\mathbf{D}(x, y)$ represents the corresponding depth value. W, H is the width and the height of the image while C is the channel number ($C = 3$ for RGB image). We expand a valid sparse point, $\mathbf{D}(x, y) > 0$, to its invalid neighborhoods, $\mathbf{D}(x+\alpha, y+\beta) \leq 0, \alpha, \beta \in [-R, R]$, if the mean difference of RGB value between the valid sparse point and the neighborhood point is smaller than a given threshold t. R is the maximum neighborhood range, which equals 2 in Fig. 2. After the expansion of all valid sparse points, the expansion depth map, $\mathbf{D_{exp}}$, where $\mathbf{D_{exp}}(x, y)$ denotes the depth value at the pixel location $(x, y), x \in [0, W-1], y \in [0, H-1]$, can be obtained. Then, the final output semi-dense depth map $\mathbf{D_{all}}$ can be generated by combining the original sparse depth $\mathbf{D}$ and the expansion one $\mathbf{D_{exp}}$. Algorithm 1 describes the detailed expansion process.

C. *Multi-Scale Feature Extraction and Fusion (MFEF) Module*

The results from depth completion [27] and stereo-LiDAR fusion [11] demonstrate that ordinary convolution is superior to sparse invariant convolution [15] when the image and sparse depth data are fused. We make a step further to adopt multi-scale feature extraction [7] for both image and expanded sparse points feature extraction, as it has the potential to increase the receptive filed of the features to increase the robustness. Given a pair

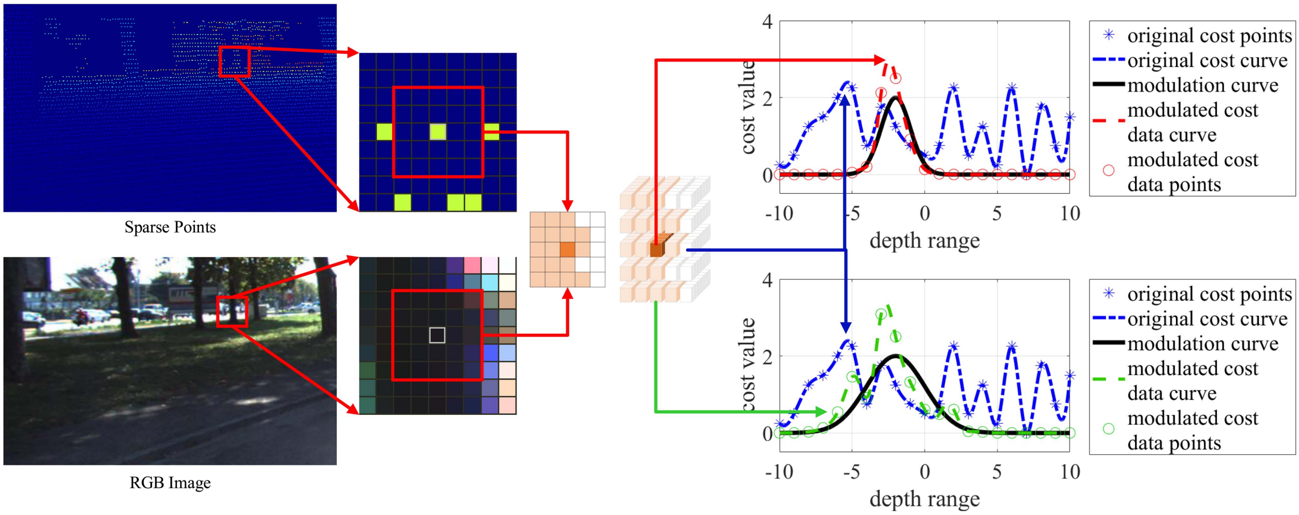

Fig. 3. The procedure of the varying-weight Gaussian guiding (VWGG) module based on the sparse expansion of the original LiDAR data. The data in semi-dense depth maps are divided into three types: the original LiDAR data, the expanded sparse data and the invalid data. We process the features in the constructed cost volume with corresponding coordinates (x, y) by modulating with a Gaussian function for the valid data while leaving the invalid data untouched. Parameters, such as the weight, variance of the Gaussian function, is made varying depending on the reliability of the sparse data. Given the same distribution of original cost volume (blue curve/data points), features of the invalid data keep unchanged (blue curve/data points), while those of the original LiDAR data, the expanded sparse data are modulated by varying Gaussian functions (black curves in right up, right down sub-figures) and result in the red, green modulated curves/data points respectively.

of stereo images and an expanded semi-dense depth image, a kind of UNet-like [7], [17] encoder-decoder structure is used for multi-scale feature extraction. Similar to the method in HSM-Net [17], the spatial pyramid pooling (SPP) is utilized to broaden the receptive filed of the lowest resolution. Specifically, the input is first fed into the encoder with five residual blocks, and then goes through the SPP module to the corresponding five decoder blocks. The average pooling size is set to $(H \times s) \times (W \times s)$, $s \in \{1/32, 1/64, 1/96, 1/128\}$.

D. Varying-Weight Gaussian Guiding (VWGG) Module

Guiding stereo matching by Gaussian modulation of the constructed cost volume along the depth dimension has been proposed in [13]. With the proposed SE module, Gaussian guiding can be expanded along the height and width of the image dimension. However, the reliability varies when the original sparse points are expanded to their neighborhoods. To solve this problem, we propose the VWGG module. Fig. 3 shows its procedure.

Different from GSM [13] that uses the same Gaussian distribution hyer-parameters for cost volume modulation, we use varying weight depending on the reliability of the valid sparse depth points. In general, the original sparse points are more reliable than the expanded ones, so higher weight and lower variance should be assigned to the Gaussian distribution of the original sparse points compared to the expanded ones. Besides the previous defined sparse depth image $\mathbf{D}$, $\mathbf{D_{exp}}$, we introduce the binary masks $\mathbf{M}$ and $\mathbf{M_{exp}}$, where $\mathbf{M}(x, y) = 1, \mathbf{M_{exp}}(x, y) = 1, x \in [0, W-1], y \in [0, H-1]$, specifying which elements of $\mathbf{D}$, $\mathbf{D_{exp}}$ are valid respectively. To guide the stereo matching, the sparse depth map $\mathbf{D}$, $\mathbf{D_{exp}}$ are converted to the sparse disparity map $\mathbf{D}'$, $\mathbf{D}'_{\mathbf{exp}}$, following GSM [13] by the equation $d = b \cdot f / z$, where $d = \mathbf{D}'(x, y)$ or $\mathbf{D}'_{\mathbf{exp}}(x, y)$ denotes the disparity value, $z = \mathbf{D}(x, y)$ or $\mathbf{D_{exp}}(x, y)$ represents the depth value, the focal length f and baseline b are the setup used to acquire stereo images.

For each pixel with coordinates (x, y) in corresponding sparse image such that element is valid, we process the features by modulation with a Gaussian function centered on the disparity $d = \mathbf{D}'(x, y)$ or $d = \mathbf{D}'_{\mathbf{exp}}(x, y)$. On the other hand, each point with $\mathbf{M}(x, y) = 0$ or $\mathbf{M_{exp}}(x, y) = 0$ is left untouched. Given the original cost volume $\mathcal{V} \in \mathbb{R}^{W \times H \times D \times F}$, where D is the max disparity, F is the feature length of the constructed cost volume, the guided cost volume $\mathcal{G} \in \mathbb{R}^{W \times H \times D \times F}$, can be obtained by multiplying modulation function in two steps regardless of the value of $\mathbf{M}(x, y)$, $\mathbf{M_{exp}}(x, y)$ as:

$$M_1 = 1 - \mathbf{M}(x, y) \cdot \left(1 - k_1 \cdot e^{-\frac{(d - \mathbf{D}'(x,y))^2}{2c_1^2}}\right), \tag{1}$$

$$M_2 = 1 - \mathbf{M_{exp}}(x, y) \cdot \left(1 - k_2 \cdot e^{-\frac{(d - \mathbf{D}'_{\mathbf{exp}}(x,y))^2}{2c_2^2}}\right), \tag{2}$$

$$\mathcal{G}(x, y, d) = M_2 \cdot M_1 \cdot \mathcal{V}(x, y, d), \tag{3}$$

where k_1, k_2 determine the weight of the Gaussian for original sparse points, expanded sparse points, respectively. c_1, c_2 represent the variance of respective original sparse points and expanded sparse points, respectively. Depending on the reliability of sparse points, weights $k_1 \geq k_2$ and $c_1 \leq c_2$ are applied for guiding the stereo matching while the best setting is obtained by experiments.

E. Cascaded 3-D CNN

Cost aggregation should be done to obtain the final depth map, which is commonly implemented in the hourglass form of 2-D [2], [19] or 3-D CNN [3], [7]. Recent works [4], [20] have shown the superior performance of 3-D CNN over 2-D CNN, even though 3-D CNN always consumes more memories and computing resources. To deal with the high computation burden of 3-D CNN, coarse-to-fine cascaded 3-D CNN [7], [18] has been proposed to improve the efficiency of cost aggregation. Instead of uniform sampling a predefined range [18] to get the disparity search range in next stage, the disparity uncertainty estimation-based method [7] is adopted in our network. The network is trained at different scales with the multi-scale loss:

$$\mathcal{L} = \mu_1 \mathcal{L}_1 + \mu_2 \mathcal{L}_2 + \mu_3 \mathcal{L}_3, \tag{4}$$

where $\mathcal{L}_1$ is the loss of on the finest level and $\mathcal{L}_3$ is the loss on the coarsest level. The weight for each loss is set fixed as 2, 1, 0.5 according to the results in [7].

IV. Experimental Results and Discussions

A. Datasets

1) Sceneflow: This is the large synthesis dataset for stereo matching and scene flow. We use this dataset for training our network before evaluating its robustness in other datasets without fine-tuning.

2) Middlebury: This is an indoor stereo dataset with full, half, and quarter resolutions. This dataset is used for the robustness evaluation with the model pretrained on the SceneFlow dataset.

3) KITTI Stereo Evaluation 2012 and 2015: The two are the stereo datasets collected by a car in the real world. These datasets are also used for the robustness evaluation with the model trained on the SceneFlow dataset.

4) KITTI Depth Completion: This dataset is released for depth completion and depth prediction with stereo images, sparse depth maps and semi-dense ground-truth depth maps in the real world. It consists of 42,949 pairs of training images and 3,426 pairs of validation images. Since the test set does not contain stereo images, we split the validation set into two sub-sets, 2,426 pairs of stereo images for testing and 1,000 for validation, which is commonly adopted in LiDAR stereo fusion works [10], [11], [12]. This dataset is used for training and evaluating from scratch.

B. Implementation Details

Our implementation is based on Pytorch. The Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) optimizer is employed for training the whole network in an end-to-end way. The smooth L-1 loss function is used during training and all the intermediate outputs are included in the loss calculation. We train the network from scratch on the SceneFlow dataset and KITTI depth completion dataset, respectively. During training, the images from the KITTI depth completion dataset are cropped down to 256×1216 patch first on the bottom of images, and then randomly cropped to 256×512 patches into the network, while the images from the SceneFlow dataset are randomly and directly cropped to 256×512 patches into the network. For both datasets, the leaning rates start at 10^{-3} and reduces to half in epoch 12, 16, and 18. A total of 20 epochs are run on two NVIDIA V100 GPUs with a batch size of 8.

During training, data augmentation is applied to improve network robustness. Different chromatic augmentation [17] and asymmetric masking [7] are used for stereo images. Random masking, which randomly replaces a rectangular region in the projected LiDAR image with zeros, is applied only in the robustness evaluation to raise the robustness in face of nonuniform distribution of LiDAR data.

C. Evaluation on KITTI Depth Completion

Qualitative analysis is first done by comparing our network with the recent open-source works, such as stereo matching network CF-NET [7], stereo-LiDAR fusion network GSM [13], and depth completion network PE-NET [9]. As shown in Fig. 4, depth maps and errors in two different cases are visualized. Our proposed network is able to produce high-precision depth estimation, while the network CF-NET [7] fails in distant regions. Comparing the depth maps between PE-NET [9] and ours, we can see that ours outperform the depth completion network PE-NET [9], especially in the top region (e.g., the sky area) without sparse LiDAR points. Compared to the stereo-LiDAR fusion network GSM [13], higher-precision depth estimation results can be obtained by our network as seen from the errors, and more detailed recovered depth results can be observed in our depth maps, especially in the tree regions.

Then quantitative analysis is performed by comparing our proposed network with classic stereo matching methods [2], [3], top-performing depth completion works [8], [24] and SOTA stereo-LiDAR fusion networks [10], [11], [12]. We adopt the metrics used in the official KITTI depth completion benchmark [15], namely, RMSE, MAE, IRMSE, and IMAE. These metrics have been widely used in depth completion [8], [24] and stereo-LiDAR fusion works [10], [11], [12].

The results are illustrated in Table I. From the results, we have several observations: 1) The stereo networks [2], [3] perform well on the inverse depth metric iRMSE, and the depth completion works [8], [24] obtain outstanding results on depth metrics (i.e. MAE, RMSE). The reason may be that the stereo networks are trained to predict the disparity while the depth completion networks are trained to estimate the depth. Similar results can be observed for the stereo-LiDAR fusion network. For example, CCVN [10] that is trained to predict the disparity ranks 1st in terms of IRMSE and 2nd in terms of IMAE, VolumPropagation [12] that predicts the depth directly ranks 1st in terms of RMSE; 2) Even with more information, the network is not guaranteed to obtain better performance on all the metrics. For example, the performance of Listereo [11] with extra stereo information is inferior to [8], [24]. With more LiDAR data, the stereo-LiDAR fusion network VolumPropagation [12] fails to outperform the stereo networks [2], [3] in terms of IRMSE; 3) The robustness of SOTA stereo-LiDAR fusion networks [10], [11], [12] on different metrics is far from satisfactory. The reason may be that the depth metrics (i.e. MAE, RMSE) tend to be sensitive to far distances and the inverse depth metrics is vulnerable to short distances.

For our proposed network, we first train it to predict disparity and then convert the disparity to depth map like [2], [3], [10]. This results in the tendency of better performance on inverse depth metrics than depth metrics. By properly integrating the stereo images and LiDAR data, our network succeeds in outperforming the stereo networks [2], [3] in terms of all the metrics.

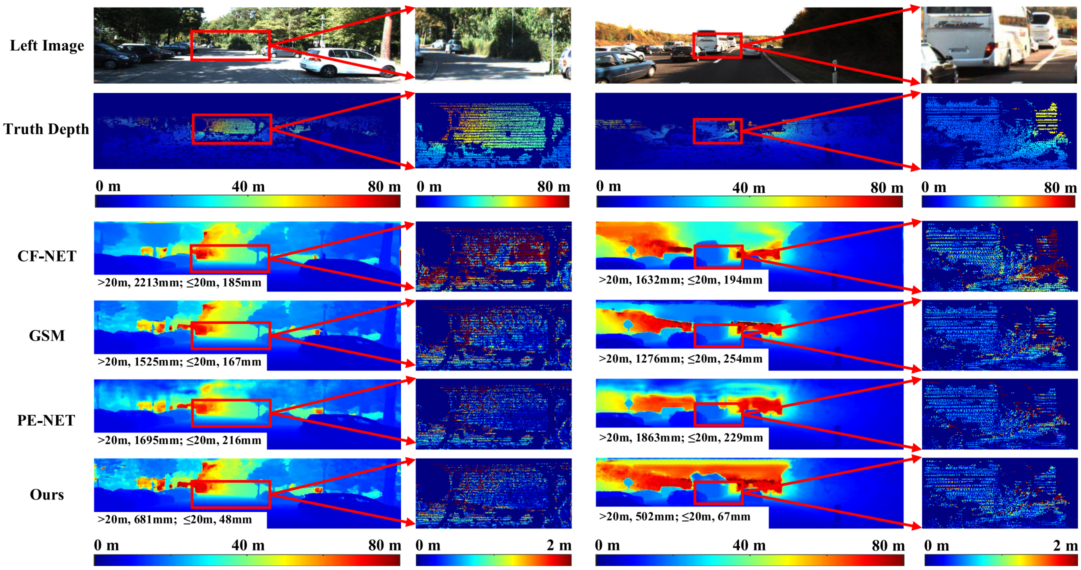
Left Image
Truth Depth
0 m
40 m
80 m
0 m
80 m
CF-NET
>20m, 2213mm; ≤20m, 185mm
>20m, 1632mm; ≤20m, 194mm
GSM
>20m, 1525mm; ≤20m, 167mm
>20m, 1276mm; ≤20m, 254mm
PE-NET
>20m, 1695mm; ≤20m, 216mm
>20m, 1863mm; ≤20m, 229mm
Ours
>20m, 681mm; ≤20m, 48mm
>20m, 502mm; ≤20m, 67mm
0 m
40 m
80 m
0 m
2 m

Fig. 4. Qualitative results on the KITTI depth completion dataset. The top two rows show the left image and the ground-truth depth maps with the selected region zoomed out. The bottom four rows show the results of CF-NET [7], stereo-LiDAR fusion network GSM [13], depth completion network PE-NET [9] and ours. We visualize the estimated depth maps and part of error maps for detailed analysis. We also include the depth metric RMSE in the condition of distance >20 meters and ≤20 meters. The figure is best viewed in color.

TABLE I
Quantitative Comparative Results on the KITTI Depth Completion Dataset

method	Modality	RMSE(mm)	MAE(mm)	IRMSE(1/km)	IMAE(1/km)
GC-NET [3]	Stereo	1031.4	405.40	1.681	1.036
PSM-NET [2]	Stereo	884.0	332.00	1.649	0.999
PE-NET [9]	Mono+LiDAR	730.08	210.55	2.17	0.94
semAtt-NET [8]	Mono+LiDAR	709.4	205.49	2.030	0.900
DySPN [24]	Mono+ LiDAR	709.1	**192.71**	1.880	0.820
Listereo [11]	Stereo+LiDAR	832.2	283.91	2.190	1.100
CCVN [10]	Stereo+LiDAR	749.3	252.50	**1.397**	*0.807*
VolumPropagation [12]	Stereo+LiDAR	**636.2**	205.10	1.872	0.987
Ours	Stereo+LiDAR	*675.5*	*197.16*	*1.600*	**0.787**

The best results are highlighted in bold font, and the second best results are highlighted in italic font.

Compared to top-performing depth completion works [8], [24] and SOTA stereo-LiDAR fusion networks [10], [11], [12], our proposed network ranks 1st in terms of IMAE and 2nd in terms of all the other metrics, which shows that it is robust on different ranges. This is probably due to the expansion of the guided LiDAR data and corresponding multi-scale features for fusion. Particularly, we do observe the slight performance gap with VolumPropagation [12] in terms of RMSE, and with CCVN [10] in terms of IRMSE. This may be caused by the expansion errors of the over-exposed image regions, such as glass under sunlight, which presents similar RGB appearance but quite different depth values.

D. Ablation Study

To verify the effectiveness of each proposed module in our network, various ablation studies have been performed. Table II displays the results.

1) Ablation on the SE Module: To evaluate the performance of our proposed SE module, we keep the Gaussian guiding weight fixed, that is $k_2 = k_1 = 10$, $c_2 = c_1 = 1$. From Table II, our proposed SE module achieves better performance in terms of almost all the metrics [15] than that without SE. Further exploration on the expansion range and expansion threshold demonstrates that the performance tends to increase with larger expansion range, while higher threshold leads to inferior RMSE and iRMSE but superior MAE and iMAE. This is consistent with our intuition that direct expansion without considering the reliability may bring smooth but high-variance results. For the final model, we select the max range $R = 2$ and threshold $t = 255$ as the setting in the SE module to reserve the smooth results and deal with high-variance problem by cooperating with the VWGG module.

The computation complexity of the SE module is quite low with floating point operations (FLOPs) equaling to $W * H * (1 + (C + 8)4R^2)$, the inference time of which is 12 ms for a single thread on NVIDIA TESLA V100. By contrast, the expansion network [14] needs 0.14 ms $* 1216 * 256 * 5\% = 2.179$ s for a sparse image on KITTI depth completion dataset with resolution of 1216*256 and 5% density.

2) Ablation on the VWGG Module: The VWGG module is proposed to balance the accuracy and completion performance. The best modulation parameters [13] $k_1 = 10$, $c_1 = 1$

TABLE II

ABLATION STUDY OF THE PROPOSED NETWORK ON THE KITTI DEPTH COMPLETION DATASET. ROUND R, t: THE EXPANSION TO THE NEIGHBOURHOODS IN THE ROUND FORMAT WITH MAX RANGE R AND THRESHOLD t. k_2, c_2 IS THE MODULATION HYPER-PARAMETERS FOR THE EXPANDED POINTS

Modules	Method	Metrics			
		RMSE(mm)	MAE(mm)	IRMSE(1/km)	IMAE(1/km)
SE	no expansion	735.54	227.81	1.6732	0.8543
	round 1 10	**693.47**	210.52	1.6507	0.8454
	round 1 255	758.40	197.78	1.6388	0.8067
	round 2 10	694.63	204.05	**1.6159**	0.7917
	round 2 255	697.70	**195.50**	1.6724	**0.7896**
VWGG	no VWGG	737.84	252.51	1.6756	0.8754
	$k_2 = 10, c_2 = 1$	697.70	**195.50**	1.6724	0.7896
	$k_2 = 8, c_2 = 2$	683.60	196.85	1.6483	0.7948
	$k_2 = 2, c_2 = 8$	**675.50**	197.16	**1.6000**	0.7872
	$k_2 = 1, c_2 = 10$	690.07	198.41	1.6028	**0.7867**
MFEF	no feature fusion	696.00	200.70	1.6393	0.7983
	sparse invariant convolution	708.60	199.88	1.6482	**0.7826**
	multiscale convolution	**675.50**	**197.16**	**1.6000**	0.7872

The best results are marked bold for each module and each metric.

TABLE III

ABLATION RESULTS ON THE KITTI DEPTH COMPLETION DATASET FOR PROPOSED MODULES

method	RMSE (mm)	MAE (mm)	IRMSE (1/km)	IMAE (1/km)
GSM	764.66	231.09	1.6935	0.8284
MFEF+SE	737.84	252.51	1.6756	0.8757
GSM+MFEF+SE	684.70	200.63	1.6218	0.7886
VWGG+SE	692.47	199.77	1.6355	0.7882
VWGG+MFEF+SE	675.50	197.16	1.6000	0.7872

TABLE IV

ROBUSTNESS EVALUATION OF THE PROPOSED NETWORK AGAINST DOMAIN SHIFT

method	KITTI2012 D1_all(%)	KITTI2015 D1_all(%)	Middlebury bad 2.0 (%)
PSM-NET [2]	15.1	16.3	39.5
CasStereo [18]	11.8	11.9	40.6
GA-NET [4]	10.1	11.7	32.2
CF-NET [7]	4.7	6.5	21.8
GSM [13] + 3% sparse	2.41	3.06	3.84
GSM [13] + 5% sparse	1.99	2.46	3.36
GSM [13] + 10% sparse	1.50	1.85	3.12
GSM [13] + 15% sparse	1.25	1.56	3.10
Ours + 3% sparse	2.3	3.9	2.7
Ours + 5% sparse	1.2	1.8	2.5
Ours + 10% sparse	0.7	0.9	2.2
Ours + 15% sparse	0.6	0.7	2.1

are adopted in our proposed network for the original LiDAR points. The max range $R = 2$ and threshold $t = 255$ is selected for the SE module for further cooperating with the VWGG module. The ablation results for the VWGG module is given in Table II. Without the VWGG module, relative high performance degradation can be observed, which shows the value of expanded data. Obviously, with the proposed module, which allocates relative low weight k_2 but high variance c_2 for modulating the expanded points, we can greatly improve our network on RMSE and gain better results on iRMSE or iMAE at a slight loss of MAE. More ablation results compared with GSM [13] are given in Table III, which demonstrate the superior of our proposed module on all metrics.

3) Ablation on the MFEF Module: The MFEF module can bring notable performance gain in terms of all metrics when compared to the network without feature fusion. We also compare our proposed module with the widely used sparse invariant convolution-based method [15]. As shown in Table II, the proposed module can achieve better performance in terms of most metrics and comparable results in terms of iMAE. Results in Table III imply that the MFEF module can help improve the stereo-matching guiding of either our proposed method or GSM [13].

E. Robustness Evaluation

The generalization capability across different domains is also evaluated. We verify the robustness of the proposed network by training only on the SceneFlow dataset and testing the performance on the other three real-world datasets without fine-tuning. The metric D1_all [29] is adopted for KITTI 2012/2015 dataset while the metric bad 2.0 [7] is applied for Middlebury dataset. It can be seen from Table IV that with only 3% of sparse points, our proposed network can boost generalization performance than the start-of-art stereo matching networks. Since the ground-truth depth of the KITTI dataset is sparse itself, the exact percentage of the sparse points with respect to the whole image is indeed much lower than 3%. Extra test results on the Middlebury dataset also demonstrate that the performance in terms of the metric bad 2.0 [7] deteriorates to 16% when only 1% sparse points are available. Compared with the other stereo LiDAR fusion method GSM [13], our proposed network is superior even though the performance gain decreases when the density of the sparse input goes down.

Besides, the robustness to different resolution of LiDAR data is evaluated in Fig. 5. As the sparse density goes down from 5% to 1%, the performance of our network keeps outperforming GSM [13], but the performance gain continuously decreases. When the random masking is applied, the performance at low density can be improved at the cost of gain loss at high density.

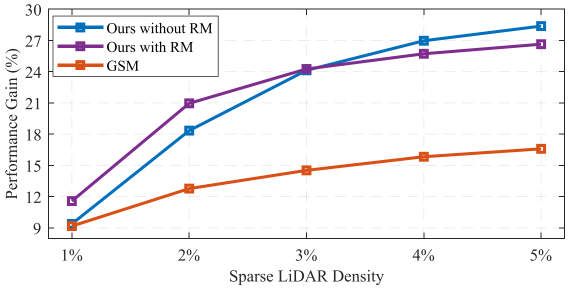

Fig. 5. The performance gain with different resolutions of LiDAR data. RM is short for random masking.

V. Conclusion and Future Work

In this letter, we proposed a novel stereo-LiDAR fusion network for robust dense depth estimation. By first expanding the sparse LiDAR with the corresponding RGB image, a semi-dense depth map can be obtained as a basis for sparse feature extraction and stereo matching guiding. According to the reliability of the sparse input, the varying-weight Gaussian guiding protocol is then proposed for improving the guiding around the sparse neighborhood, and multi-scale feature extraction and fusion is applied for enhancing the feature receptive field. Experimental results show that our proposed network performs well in terms of all the metrics and is robust to different domains. We plan to incorporate semantic information in the future to improve the expansion performance when faced with abnormal conditions, such as over-exposure.

References

[1] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, 2016.

[2] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.

[3] A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 66–75.

[4] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 185–194.

[5] R. Liu, C. Yang, W. Sun, X. Wang, and H. Li, "StereoGAN: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12757–12766.

[6] X. Song, G. Yang, X. Zhu, H. Zhou, Z. Wang, and J. Shi, "AdaStereo: A simple and efficient approach for adaptive stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10328–10337.

[7] Z. Shen, Y. Dai, and Z. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13906–13915.

[8] D. Nazir, M. Liwicki, D. Stricker, and M. Z. Afzal, "SemAttNet: Towards attention-based semantic aware guided depth completion," *IEEE Access*, vol. 10, pp. 120781–120791, 2022, doi: 10.1109/ACCESS.2022.3214316.

[9] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "PENet: Towards precise and efficient image guided depth completion," in *Proc. IEEE Int. Conf. Robot. Automat. IEEE*, 2021, pp. 13656–13662.

[10] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 5895–5902.

[11] J. Zhang, M. S. Ramanagopal, R. Vasudevan, and M. Johnson-Roberson, "LiStereo: Generate dense depth maps from LiDAR and stereo imagery," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 7829–7836.

[12] J. Choe, K. Joo, T. Imtiaz, and I. S. Kweon, "Volumetric propagation network: Stereo-LiDAR fusion for long-range depth estimation," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 4672–4679, Jul. 2021.

[13] M. Poggi, D. Pallotti, F. Tosi, and S. Mattoccia, "Guided stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 979–988.

[14] Y.-K. Huang et al., "S3: Learnable sparse signal superdensity for guided depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16706–16716.

[15] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity Invariant CNNs," in *Proc. Int. Conf. 3D Vis*, 2017, pp. 11–20.

[16] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, "HMS-Net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion," *IEEE Trans. Image Process.*, vol. 29, pp. 3429–3441, 2020.

[17] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5515–5524.

[18] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2495–2504.

[19] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1959–1968.

[20] Z. Rao, M. He, Y. Dai, Z. Zhu, B. Li, and R. He, "NLCA-Net: A non-local context attention network for stereo matching," *APSIPA Trans. Signal Inf. Process.*, vol. 9, 2020, Art. no. e18.

[21] X. Song, X. Zhao, L. Fang, H. Hu, and Y. Yu, "EdgeStereo: An effective multi-task learning network for stereo matching and edge detection," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 910–930, 2020.

[22] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "Neural window fully-connected CRFs for monocular depth estimation," in *Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognit.*, 2022, pp. 3916–3925.

[23] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "VIP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3997–4008.

[24] Y. Lin, T. Cheng, Q. Zhong, W. Zhou, and H. Yang, "Dynamic spatial propagation network for depth completion," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1638–1646.

[25] Z. Yan, K. Wang, X. Li, Z. Zhang, J. Li, and J. Yang, "RigNet: Repetitive image guided network for depth completion," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 214–230.

[26] F. Ma and S. Karaman, "Sparse-to-Dense: Depth prediction from sparse depth samples and a single image," in *Proc. IEEE Int. Conf. Robot. Automat*, 2018, pp. 4796–4803.

[27] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi, "Sparse and dense data with CNNs: Depth completion and semantic segmentation," in *Proc. IEEE Int. Conf. 3D Vis*, 2018, pp. 52–60.

[28] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation with the 3D LiDAR and stereo fusion," in *Proc. IEEE Int. Conf. Robot. Automat*, 2018, pp. 2156–2163.

[29] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.