2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Stereo-LiDAR Depth Estimation with Deformable Propagation and Learned Disparity-Depth Conversion

Ang Li[1], Anning Hu[1], Wei Xi[2,3], Wenxian Yu[1] and Danping Zou[1*]

***Abstract*— Accurate and dense depth estimation with stereo cameras and LiDAR is an important task for automatic driving and robotic perception. While sparse hints from LiDAR points have improved cost aggregation in stereo matching, their effectiveness is limited by the low density and non-uniform distribution. To address this issue, we propose a novel stereo-LiDAR depth estimation network with Semi-Dense hint Guidance, named SDG-Depth. Our network includes a deformable propagation module for generating a semi-dense hint map and a confidence map by propagating sparse hints using a learned deformable window. These maps then guide cost aggregation in stereo matching. To reduce the triangulation error in depth recovery from disparity, especially in distant regions, we introduce a disparity-depth conversion module. Our method is both accurate and efficient. The experimental results on benchmark tests show its superior performance. Our code is available at https://github.com/SJTU-ViSYS/SDG-Depth.**

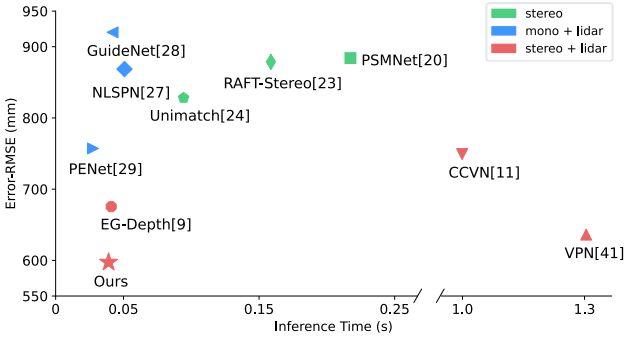

Fig. 1. Our network achieves the best trade off in accuracy and inference speed on the KITTI Completion dataset. Green, blue, and red marks represent results from stereo matching, monocular depth completion, and stereo-LiDAR fusion methods, respectively.

I. INTRODUCTION

Dense depth estimation is a fundamental task in autonomous driving, robotic navigation [1], and 3D reconstruction [2]. Stereo matching, a widely adopted technique for depth estimation, computes the dense disparity map between two rectified images. The disparity map is then converted into a depth map or a 3D point cloud through triangulation. Learning-based stereo matching methods have achieved impressive performances [3] in recent years. However, their performances still degrade in the case of severe illumination changes and textureless [4]. Recent studies show that the sparse depth from LiDAR [5][6][7] can be used as additional hints to guide stereo matching in challenging scenarios, where the depth hints are taken as additional inputs and processed by the neural network.

However, raw LiDAR depth data are sparse and non-uniformly distributed, making the neural network ignore them or produce noisy predictions. Therefore, the sparse depth usually needs to be expanded and spread to nearby pixels before further processing as demonstrated in prior work such as [8][9]. While achieving good results, their methods are based solely on local information, which may not work well in regions with occlusion or object boundaries where the depth is often discontinuous. In those areas, the expanded hints are usually over-smoothing or contain trailing effects, which leads to poor stereo matching results.

Furthermore, accurately recovering depth at a distance is challenging due to the quadratic growth of triangulation errors with distance. Existing methods [10][9][11] primarily focus on enhancing the accuracy of disparity predictions to compensate for triangulation errors. However, further improving disparity predictions typically comes at the cost of more complex networks and increased computational requirements, which may still result in large depth errors via triangulation even when only tiny disparity errors are present for distant points.

To address the aforementioned issues, we propose a novel stereo-LiDAR depth estimation network that adopts a sparse LiDAR deformable propagation module and a learned disparity-depth conversion module, as depicted in Figure 2. The propagation module computes propagation weights through local self-correlation, which integrates global context and local information. It then propagates this information within learned deformable windows to effectively expand depth hints across occluded and boundary areas. We utilize both the expanded disparity feature and RGB feature to construct a cost volume for stereo matching. Additionally, the expanded disparity guides cost aggregation during stereo matching. To obtain dense disparity, we perform cost aggregation using a commonly used coarse-to-fine 3D CNN [12][13]. To mitigate triangulation errors during disparity-to-depth conversion, we introduce a lightweight network that predicts adjustment residuals in both disparity and depth spaces based on high-frequency features.

We evaluate the proposed network on various benchmark datasets, including KITTI depth completion [14], Virtual KITTI2 [15], and MS2 [16] dataset. The experimental results

[1] Shanghai Key Laboratory of Navigation and Location-based Service, Shanghai Jiao Tong University. [2] Intelligent Perception Institute, Midea Corporate Research Center. [3] Blue-Orange Lab, Midea Group. *Corresponding author: Danping Zou (dpzou@sjtu.edu.cn). This work was supported by National Key R&D Program of China (2022YFB3903801) and Midea Group's 3D Vision Project.

show our method achieves state-of-the-art accuracy and efficiency (see Figure 1). Our key contributions are highlighted as follows:

- We introduce a novel stereo-LiDAR fusion network for precise and efficient depth estimation, setting new benchmarks for accuracy and speed across various datasets.
- We design a sparse LiDAR deformable propagation module to effectively expand depth hints across occlusion and boundary areas. The module propagates sparse hints within learned varying-shaped windows, incorporating global context and local information.
- We develop a lightweight disparity-depth conversion module that enables precise depth recovery from disparity with low memory and computational requirements.

II. RELATED WORKS

A. Stereo Matching

Stereo matching aims to find a dense disparity map between two rectified images, facilitating scene depth recovery through triangulation. Learning-based stereo matching methods have made significant progress. Early approaches [17][18] adopted siamese networks to extract patch-wise features or predict matching costs. Follow-up methods [19][20] introduced 3D CNN to capture pixel-wise correspondence for more accurate predictions. To reduce the memory and computational costs of 3D CNN, many methods [21][13][22] have be proposed. Xu et al. [21] used deformable convolution to simplify cost aggregation, and Tankovich et al. [22] utilized coarse-to-fine convolution to reduce computation.

Though existing stereo-matching methods [21][23][24] have achieved promising results, they may exhibit decreased performance in challenging scenarios, such as textureless regions, occlusion, and distant objects. Recent research suggests that incorporating LiDAR points can enhance performance, highlighting the potential of using multi-modal information for dense depth estimation.

B. Monocular Depth Completion

Monocular depth completion aims to predict a dense depth map from a single image and sparse LiDAR points. Existing methods can be classified into two categories: spatial propagation methods [25][26][27][14] and fusion-based methods [28][29][30][31][6]. Spatial propagation methods typically involve learning an affinity matrix based on RGB images to propagate depth values. In contrast, fusion-based methods leverage the geometric information from multi-modal data, including single images and LiDAR points. These methods [28][29] typically employed two sub-networks to extract multi-layer features from RGB images and sparse depth, respectively. These features are then fused at different stages to generate depth maps. However, the significant sparsity and non-uniform distribution of LiADR points pose challenges, leading to performance degradation in regions with insufficient LiDAR points.

C. Stereo-LiDAR Fusion

Stereo-LiDAR fusion methods [10][32][33][34][35] produce more precise depth predictions by combining stereo images and sparse LiDAR points. There are primarily two ways to explore geometric cues from sparse LiDAR points and stereo images: fusing the two-modalities information at the feature level, and leveraging sparse points to guide cost aggregation in stereo matching. Early fusion-based methods [36][37] extracted features from sparse depth and stereo images and integrated multi-modal information by concatenating these features. Zhang et al. [10] constructed an attention map by incorporating image features and depth features for depth prediction.

In contrast to the aforementioned approaches, methods that employ sparse points as guidance explicitly leverage the metric geometric information of LiDAR points. Poggi et al. [7] constructed a Gaussian modulation to regulate the original cost volume. Based on this idea, Huang et al. [8] and Xu et al. [9] expanded sparse depth into semi-dense depth as guidance for cost volume. Despite achieving performance improvements, these methods either expand depth within fixed-shape windows or propagate depth based on the original RGB images, resulting in limited performance due to cross-boundary propagation and illumination variations.

Compared to existing stereo methods, we propose a novel and efficient stereo-LiDAR depth estimation network. Specifically, we design a learnable network that propagates sparse LiDAR points within deformable windows, incorporating both global context and local information, to produce semi-dense hints as guidance. Moreover, we develop a lightweight disparity-depth conversion module to accurately recover depth from disparity, leveraging high-frequency image information. Experiments show that our method achieves significantly better prediction accuracy and speed than existing methods.

III. METHOD

Figure 2 shows the overview architecture of our stereo-LiDAR depth estimation network. It mainly consists of four components: 1) a sparse disparity **D**eformable **P**ropagation (DP) module, which expands sparse hints to a semi-dense hint map along with its confidence through learned varying-shaped windows. The confidence map indicates the reliability of the propagated hint map; 2) a **C**onfidence-based **G**aussian (CG) module, which constructs a Gaussian distribution along the disparity using the expanded semi-dense disparity map and its confidence map to constrain the cost volume effectively; 3) a coarse-to-fine 3D CNN is employed to produce dense disparity from the modulated cost volumes; 4) a **D**isparity-**D**epth **C**onversion (DDC) module, which accurately recovers depth from disparity and reduces triangulation error based on high-frequency features.

A. Deformable Propagation (DP) Module

Feature extraction with global awareness: Given an image $I \in \mathbb{R}^{H\times W\times 3}$ and the corresponding sparse disparity map $D \in \mathbb{R}^{H\times W}$, our objective is to propagate disparity

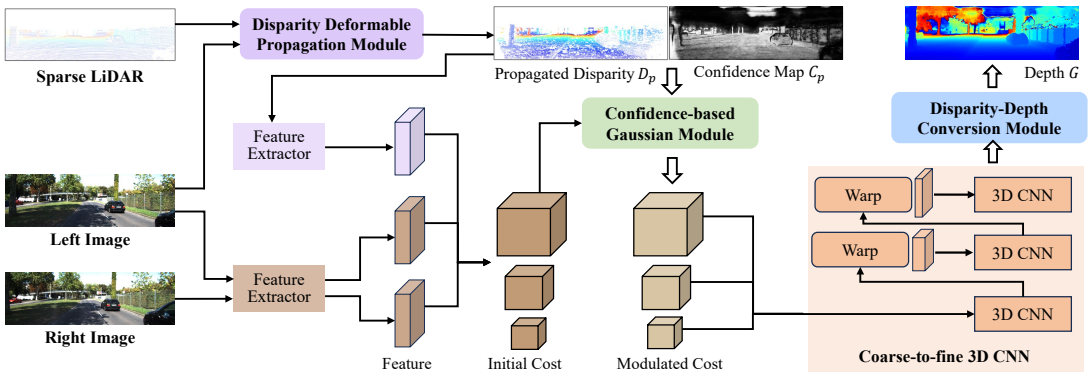

Fig. 2. The architecture of our proposed network. Firstly, the disparity **D**eformable **P**ropagation (DP) module propagates sparse LiDAR within varying-shaped windows to semi-dense disparity. Based on the generated disparity map and confidence map, the **C**onfidence-based **G**aussian (CG) module regulates the cost volume that is constructed from the features of stereo images and expanded disparity. Subsequently, dense disparity is obtained by employing coarse-to-fine 3D CNN on the regulated cost volume. Finally, the learned **D**isparity-**D**epth **C**onversion (DDC) module accurately recovers depth from the disparity of 3D CNN.

values to the surrounding area based on pixel correlations. We've observed that propagation using features extracted from a limited local region [8] often leads to noisy hints. To address this, as illustrated in Figure 3, we adopt a stacked encoder-decoder architecture to extract features, taking the left image and the left feature used for cost volume construction as inputs. This enables us to encode a larger receptive field of image information into the feature representation F_g, facilitating hint propagation with global awareness.

Deformable propagation: Sparse hint propagation within fixed-shape windows usually encounters challenges at object boundaries because depth is usually discontinuous in these regions. Assigning a single depth value across such areas will result in over-smoothness. Inspired by deformable convolution [38][39], we design deformable propagation with local self-correlation [40] to improve propagation across boundaries, as shown in Figure 3. The deformable propagation consists of three steps: 1) generating a learned 2D offset field; 2) computing propagation weights using local self-correlation based on the offset field; and 3) propagating sparse disparity within the deformable windows.

Given the extracted global-aware feature F_g, the learned 2D offset field O is produced with a convolution layer and a sigmoid operation.

$$O = Sigmoid(Conv(F) - 0.5) \times 2 \in \mathbb{R}^{H \times W \times P^2 \times 2} \quad (1)$$

where P is size of the propagation window, and P is set to be 9 in our experiments if not specified. O is rescaled to $[-1, 1]$ for stable computation.

The deformable propagation weight A is calculated based on the local self-correlation of the feature F_g, which is formulated as follows,

$$A = Softmax(F_g^* \psi(F_g, O)) \in \mathbb{R}^{H \times W \times P^2} \quad (2)$$

where F_g^* represents the feature reshaped from F_g; ψ denotes sampling operation within deformable windows generated based on the offset field.

Finally, the propagated disparity D_p and corresponding confidence map C_p can be formulated as follows,

$$D_p = A \cdot \psi(D, O) \in \mathbb{R}^{H \times W} \quad (3)$$

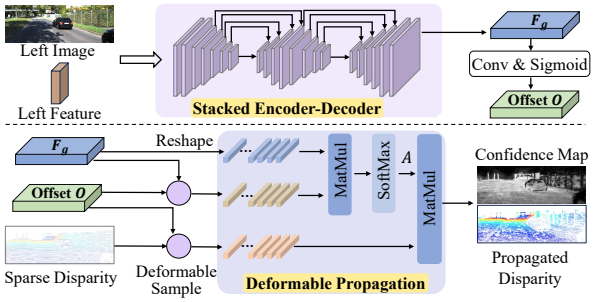

Fig. 3. Disparity deformable propagation module. The module computes the propagation weight by employing local self-correlation based on the learned 2D offset field and propagates sparse hints within the deformable windows.

$$C_p = A \cdot \psi(M_{sparse}, O) \in \mathbb{R}^{H \times W} \quad (4)$$

where $M_{sparse} \in \{0, 1\}$ represents the valid mask of the sparse disparity D. C_p indicates the reliability of the propagated depth D_p.

The deformable propagation is performed at a resolution of $1/4$, ensuring high efficiency.

B. Confidence-based Gaussian Module

As demonstrated in [7][8], modulating the cost volume with Gaussian distributions derived from sparse LiDAR points can significantly enhance stereo matching performance. However, the quality of the propagated depth D_p varies in different regions, and erroneous propagation may bias stereo matching. Therefore, we extend the Gaussian modulation introduced in [7] to a confidence-based Gaussian modulation, which adapts to the varying reliability of the propagated depth.

Let the original cost volume constructed from the left feature, semi-dense disparity feature and shift right feature be $CV \in \mathbb{R}^{H \times W \times D_{max} \times L}$, where D_{max} represents the max disparity and L is the number of feature channels; $D_p \in \mathbb{R}^{H \times W}$ denotes the propagated semi-dense disparity. Then the Gaussian modulation can be described as:

$$CV'(x, y, d) = f \cdot CV(x, y, d) \quad (5)$$

$$f = 1 - M(x,y) + M(x,y) \cdot k \cdot C_p(x,y) \cdot e^{-\frac{(d-D_p(x,y))^2}{2\omega^2}} \quad (6)$$

where f represents the pixel-varying modulation weights; $C_p(x,y)$ describing the confidence at pixel (x,y); $M \in \{0,1\}$ is a valid mask of $C_p > \rho$ to exclude unreliable propagated depth; k and ω are constants to adjust the height and width of the Gaussian distribution; ρ, k and ω are set to 0.4, 2 and 8 in our implementation, respectively; $d \in \{0, 1, \cdots, D_{max} - 1\}$; In this function, when unreliable propagation occurs, C takes on a very small value and the mask M will be 0, leading the CV' to be the original cost volume, and excluding the erroneous guidance.

C. Coarse-to-fine 3D CNN

While Gaussian modulation is applied, the cost volume still exhibits noise that hampers accurate matching. Cost aggregation with 3D CNN is leveraged to incorporate extensive contextual information for precise dense disparity estimation. To alleviate the computational burden, a coarse-to-fine 3D CNN [12][13] is employed in our network, where the generated multi-scale disparity maps are used for training losses. Furthermore, we incorporate the searching range adjustment based on disparity uncertainty [13] into our network to further enhance efficiency. As shown in Figure 1, our network achieves high inference speed through the coarse-to-fine and adaptive searching range strategies.

D. Disparity-Depth Conversion (DDC) Module

Due to the triangulation error growing quadratically with the distance, recovering scene depth accurately from disparity is challenging, especially in distant regions. To address this challenge, a lightweight disparity-depth conversion module is used for compensating the triangulation error as shown in Figure 4. The network predicts two pixel-wise residuals : δ_1 and δ_2, which are used for compensating disparity and depth errors respectively.

$$G = \frac{b \cdot f}{D_s + \delta_1} + \delta_2 \quad \delta_1, \delta_2 \in \mathbb{R}^{H \times W} \quad (7)$$

Here $G \in \mathbb{R}^{H \times W}$ is the converted depth map; b and f are the baseline and focal length of the stereo camera, respectively; D_s is the disparity map from stereo matching.

To enhance edge awareness, we leverage the high-frequency information [41] in the original images to enable the network to acquire error correction capabilities while preserving details. We first warp the right image to the left viewpoint using the disparity D_s obtained from stereo matching, and then compute an error map by comparing the warped right image with the left image. Subsequently, a network with a stacked U-Net-like architecture takes four inputs : the left image, the error map, the disparity D_s, and the derived depth G_s, and produces the two residuals δ_1 and δ_2 in Equation 7. To ensure stable computation, δ_1 and δ_2 are rescaled to $[-0.2, 0.2]$ and $[-0.6, 0.6]$, respectively. The entire process can be formulated as follows:

$$\delta_1, \delta_2 = Net(I_l, I_l - Warp(I_r, D_s), D_s, G_s) \quad (8)$$

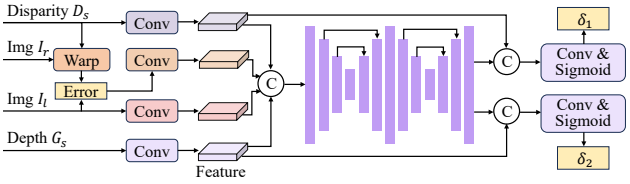

Fig. 4. Disparity-depth conversion module. The module generates pixel-wise residuals δ_1 and δ_2 in both the disparity and depth space, based on high-frequency features.

E. Loss Function

The proposed network is trained end-to-end in a supervised manner. We formulate the loss function by using propagated disparity map D_p, disparity maps obtained from different scales of 3D CNN described in Section III-C, and the depth map G generated from the disparity-depth conversion (DDC) module, as follows,

$$L = a\mathcal{L}_1(D_p, D_{gt}) + b_i \sum_{i=1}^{3} L^i_{disparity} + \lambda L_{depth}(G, G_{gt}) \quad (9)$$

where $\mathcal{L}_1$ represent the L-1 loss function; $L^i_{disparity}$ denotes the L-1 loss based on multi-scale disparity maps of 3D CNN; L_{depth} represents L-1 and L-2 loss of the final depth map G. In our experiments, a is set to 0.5, $b_i, i \in \{1, 2, 3\}$, are set to 0.5, 1.0, and 2.0, and λ is set to 0.7, respectively.

IV. EXPERIMENTS

A. Datasets

We conduct experiments on three benchmark datasets, including **KITTI depth completion** [14], **Virtual KITTI2** [15] and **MS2** [16] datasets. KITTI depth completion dataset [14] is a large-scale outdoor driving scenarios dataset. It contains stereo image pairs and raw noise sparse LiDAR points collected by a Velodyne LiDAR. The semi-dense ground-truth depth is obtained by [14], as shown in Figure 5. The dataset provides 42949 frames for training and 3426 frames for validation with the image size of 375×1242.

Virtual KITTI2 datasets [15] is a synthetic dataset and provides dense ground-truth depth maps. The dataset contains five scenes, and we follow [42] to use "Scene01" and "Scene02" for network training, and the remaining scenes are used for testing. In total, there are 680 frames for training and 1446 frames for testing. As in [42], we randomly sample points from the dense depth map, resulting in sparse depth maps with a density of 5% which is close to the average density of the sparse depth of KITTI depth completion dataset.

MS2 [16] is a large-scale multi-spectral stereo dataset collected in the real world, which provides stereo images, raw LiDAR points, and semi-dense ground-truth depth. Due to the repetitive scenes in the original dataset, we select four splits for training ("2021-08-06-11-23-45", "2021-08-13-16-14-48", "2021-08-13-16-31-10", "2021-08-13-17-06-04") and one split for validation ("2021-08-13-16-08-46"). In total, there are 10120 frames for training and 1272 frames for validation. The image size is 384×1224.

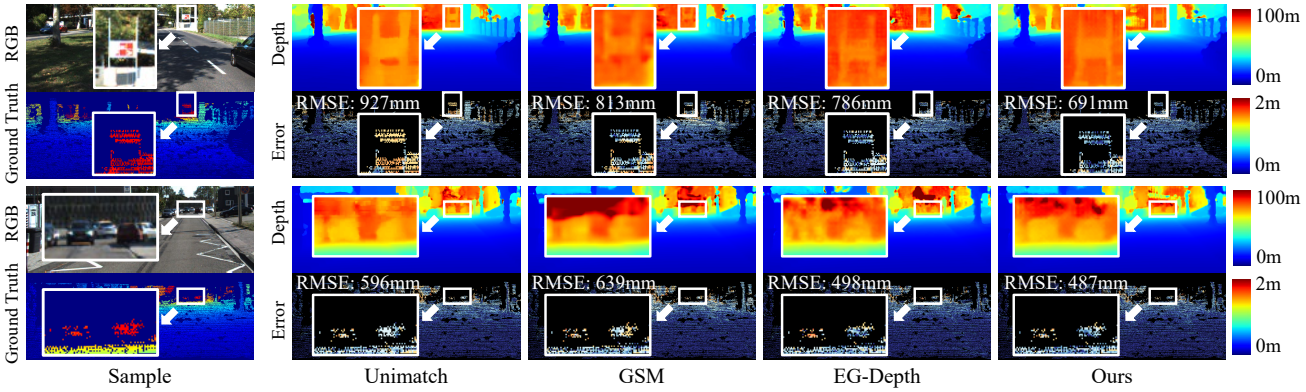

Fig. 5. Qualitative results on KITTI depth completion dataset [14]. Our network produces more accurate predictions with smaller depth errors (in blue) and more regular object shapes in distant regions, compared to other state-of-the-art stereo and stereo-LiDAR methods.

B. Implementation Details

We implement our network with PyTroch [43] and conduct training on NVIDIA RTX 4090 GPUs. Across all datasets, we train our network using the Adam optimizer with parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) and a batch size of 4. During training, images are cropped to 256×512. For the KITTI completion dataset, the network is trained from scratch for 25 epochs and the learning rate starts at 10^{-3} and reduces to half at epochs 14, 17, 19, and 24. For Virtual KITTI2 dataset, we fine-tune the network using weights pre-trained on KITTI completion dataset for 5000 steps, with a learning rate of 3×10^{-4}. For MS2 dataset, we train the network from scratch for 30 epochs with a constant learning rate of 10^{-3}. During validation, full-sized original images are fed into the network.

We adopt the standard metrics described in the official KITTI depth completion benchmark [14] to evaluate the quality of the estimated depth maps, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and their inverse ones iRMSE and iMAE.

C. Benchmark Evaluation

We evaluate our network on the three aforementioned benchmarks [14][15][16], comparing it with state-of-the-art stereo, monocular depth completion, and stereo-lidar fusion methods. The quantitative results are presented in Table I for KITTI depth completion dataset, Table II for MS2 and Virtual KITTI2 datasets. The results demonstrate that our method achieves superior performance in most metrics with a higher inference speed as shown in Table I. For example, our method achieves an RMSE of 623.2, representing a **21.5%** reduction compared to the sparse guidance method GSM [7] with an RMSE of 793.4. Compared to EG-Depth [9] with a similar speed to ours, our method achieves an RMSE of 623.2, which indicates a **7.7%** reduction compared to EG-Depth with an RMSE of 675.5. The results on Virtual KITTI2 and MS2, as shown in Table II, also demonstrate a significant improvement in the prediction of our method.

To intuitively compare different methods, we show the visual results of different methods on KITTI depth completion dataset in Figure 5. It can be observed that our method produces more accurate predictions in distant regions, whereas the other two stereo-lidar methods generate predictions with larger depth errors. Besides, our method generates predictions with a more regular object shape for the cars at a distance (the second sample in Figure 5), while other methods failed. Additionally, Figure 6 showcases visual results on Virtual KITTI2, highlighting our method's ability to produce depth predictions with sharper edges and more regular object shapes. In summary, our approach generates higher-quality predictions in distant and object edge regions with a lower RMSE. More quantitative and qualitative results are provided in the supplementary material.

TABLE I
QUANTITATIVE RESULTS ON KITTI DEPTH COMPLETION VALIDATE DATASET.

Method	Inputs	RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)	FPS (Hz)	Memory (MB)
GC-Net [19]	S	1031.4	405.40	1.681	1.036	2.4	8073
PSMNet [20]	S	884.0	332.00	1.649	0.999	4.6	4465
RaftStereo[23]	S	878.8	301.80	1.751	0.954	6.3	7839
Unimatch [24]	S	828.2	283.19	1.666	0.923	10.6	5970
GuideNet [28]	M+L	920.2	232.21	2.318	0.946	23.8	2139
NLSPN [27]	M+L	868.6	236.70	2.614	1.026	19.7	**1939**
PENet [29]	M+L	757.2	209.00	2.220	0.920	**36.3**	2571
Listereo [37]	S+L	832.2	283.91	2.190	1.100	-	-
GSM [7]	S+L	793.4	271.48	1.531	0.864	4.4	4473
CCVN [11]	S+L	749.3	252.50	**1.397**	0.807	1.0	8260
S3 [8]	S+L	703.7	239.60	1.540	0.790	4.3	6300
SLFNet [10]	S+L	641.1	**197.00**	1.773	0.876	-	-
VPN [42]	S+L	636.2	205.10	1.872	0.987	0.7	-
EG-Depth [9]	S+L	675.5	197.16	1.600	0.787	24.4	5513
Ours	S+L	**623.2**	197.55	1.519	**0.772**	25.6	5700

*"S", "M+L" and "S+L" represent stereo camera, monocular camera with LiDAR, and stereo camera with LiDAR, respectively. **Bold** and underline refer to the best and second-best results, respectively.*

D. Evaluation on Different Ranges

We evaluate the performance at different distances to provide a more comprehensive view of the enhancements achieved by our method, as demonstrated in Table III. Our approach exhibits superior accuracy in each region compared to other competitive methods, especially in the distant region. For example, in the distant regions of $20-100m$, our method achieves a 6% reduction in RMSE compared to the second-best method EG-Depth.

TABLE II

QUANTITATIVE RESULTS ON MS2 (REAL-WORLD) [16] AND VIRTUAL KITTI2 (SYNTHETIC) [15] DATASETS.

Dataset	Method	RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)
MS2	Unimatch [24]	1706.72	946.80	5.173	2.519
	GSM [7]	1434.48	815.21	3.494	1.882
	EG-Depth [9]	1251.01	696.00	3.284	1.763
	Ours	**1056.54**	**604.95**	**3.209**	**1.736**
Virtual KITTI2	Unimatch [24]	3730.59	1091.4	7.291	2.163
	CCVN [11]	3726.83	915.6	8.814	2.456
	GSM [7]	3510.12	966.8	7.059	1.609
	VPN [42]	3217.16	712.0	7.168	2.694
	EG-Depth [9]	3184.22	815.9	4.302	**1.072**
	SLFNet [10]	2843.16	**696.2**	6.794	2.007
	Ours	**2821.44**	776.8	**4.224**	1.105

TABLE III

EVALUATION ON DIFFERENT RANGES ON KITTI DEPTH COMPLETION.

Depth Range	Method	RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)
0−20m	Unimatch [24]	268.0	115.3	1.706	0.991
	GSM [7]	237.6	105.9	1.643	0.912
	EG-Depth [9]	228.1	96.5	1.607	0.869
	Ours	**227.4**	**95.8**	**1.596**	**0.853**
20−100m	Unimatch [24]	1711.3	890.8	1.458	0.716
	GSM [7]	1663.9	875.2	1.353	0.695
	EG-Depth [9]	1365.3	588.5	1.200	0.498
	Ours	**1284.0**	**569.9**	**1.146**	**0.495**

E. Ablation

To verify the effectiveness of each module in our network, we conduct various ablation studies on KITTI depth completion dataset.

Ablation on key components: To assess the individual contributions of each module, including deformable propagation (DP), confidence-based Gaussian (CG), and disparity-depth conversion (DDC) modules, we remove each module from the entire network and train each model from scratch. The results are shown in Table IV. The removal of any single module leads to a reduction in network performance. By combining all the proposed modules, our network achieves the best performance. Additionally, it can be observed that although the disparity maps produced by all models are of similar quality (e.g. EPE, D1), the models with the DDC module, such as (a), (b), and (d), predict better depth results compared to (c) without the DDC module. This highlights the substantial contribution of the DDC module, as triangulation introduces errors in the conversion, in contrast, the proposed DDC module enables accurate depth recovery from the disparity obtained from stereo matching, as intended.

Ablation on deformable propagation (DP) module: The DP module is designed to enhance sparse hints propagation by expanding hints within deformable windows. We conduct ablation experiments on this module, involving propagation within fixed-shape windows and propagation within deformable windows of different sizes. The results are presented in Table V. It can be seen that propagation within deformable windows produces superior performance. Besides, the performance of the network decreases when the window size is too large, which is likely due to the network capturing too much contextual information that may be irrelevant to the current pixel, resulting in decreased propagation quality.

TABLE IV

ABLATION ON KEY MODULES ON KITTI DEPTH COMPLETION.

	Method	Depth map RMSE (mm)	Depth map MAE (mm)	Depth map iRMSE (1/km)	Depth map iMAE (1/km)	Disparity map EPE (pix.)	Disparity map D1 (%)
(a)	w/o DP	688.8	218.71	1.589	0.801	0.31	0.15
(b)	w/o CG	673.2	219.70	1.629	0.815	0.31	0.16
(c)	w/o DDC	710.2	213.66	1.540	0.794	0.30	0.15
(d)	Ours Full	**623.2**	**197.55**	**1.519**	**0.772**	**0.29**	**0.14**

TABLE V

ABLATION STUDY ON DEFORMABLE PROPAGATION (DP) AND CONFIDENCE-BASED GAUSSIAN (CG) MODULES.

Modules	Method	RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)
DP Module	w/o offset	649.3	217.52	1.563	0.819
	with offset	**623.2**	**197.55**	**1.519**	**0.772**
	$size = 7$	646.0	211.64	1.555	0.814
	$size = 9$	**623.2**	**197.55**	**1.519**	**0.772**
	$size = 11$	635.7	200.44	1.537	0.777
CG Module	w/o confidence	638.9	206.72	1.588	0.811
	with confidence	**623.2**	**197.55**	**1.519**	**0.772**
	$k = 1, \omega = 8$	644.2	206.77	1.556	0.786
	$k = 2, \omega = 8$	**623.2**	**197.55**	**1.519**	**0.772**
	$k = 8, \omega = 2$	642.8	213.69	1.573	0.818
	$k = 8, \omega = 1$	653.7	214.82	1.588	0.819

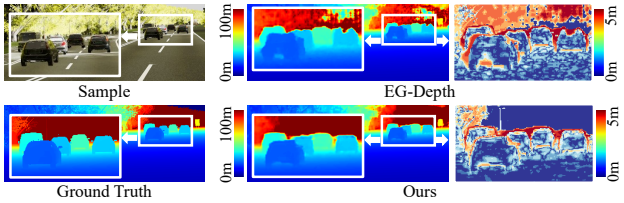

Fig. 6. Qualitative results on Virtual KITTI2 dataset [15]. Our network produces more accurate predictions with sharper edges in distant regions, compared to another state-of-the-art stereo-LiDAR method (EG-Depth) [9].

Ablation on confidence-based Gaussian (CG) module: We also adjust the hyper-parameter, including height k and width ω of CG module to achieve optimal performance. As shown in Table V, the optimal parameters of $k = 2$ and $\omega = 8$ are adopted for the best performance.

V. CONCLUSION

We present a novel and efficient stereo-lidar depth estimation network. Sparse LiDAR is first adaptively propagated within deformable windows, resulting in a semi-dense disparity map and its corresponding confidence map. Subsequently, to address the variable reliability of propagated disparity, a confidence-based Gaussian module utilizes the semi-dense disparity and confidence map as inputs to guide cost aggregation. Finally, a lightweight module is employed to accurately recover depth from disparity obtained from coarse-to-fine 3D CNN. Comprehensive experiments are conducted in both real-world and synthetic datasets. The results demonstrate the superior performance of our method. Future work includes producing globally consistent scene depth and acquiring real-world datasets with high-precision, high-density ground truth for quantitative evaluation.

REFERENCES

[1] L. Nalpantidis and A. Gasteratos, "Stereo vision for robotic applications in the presence of non-ideal lighting conditions," *Image and Vision Computing*, vol. 28, no. 6, pp. 940–951, 2010.

[2] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *2011 IEEE intelligent vehicles symposium (IV)*. Ieee, 2011, pp. 963–968.

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

[4] S. S. Shivakumar, K. Mohta, B. Pfrommer, V. Kumar, and C. J. Taylor, "Real time dense depth estimation by fusing stereo with sparse depth measurements," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6482–6488.

[5] T. A. Siddiqui, R. Madhok, and M. O'Toole, "An extensible multi-sensor fusion framework for 3d imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1008–1009.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3313–3322.

[7] M. Poggi, D. Pallotti, F. Tosi, and S. Mattoccia, "Guided stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 979–988.

[8] Y.-K. Huang, Y.-C. Liu, T.-H. Wu, H.-T. Su, Y.-C. Chang, T.-L. Tsou, Y.-A. Wang, and W. H. Hsu, "S3: Learnable sparse signal superdensity for guided depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 706–16 716.

[9] Z. Xu, Y. Li, S. Zhu, and Y. Sun, "Expanding sparse lidar depth and guiding stereo matching for robust dense depth estimation," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1479–1486, 2023.

[10] Y. Zhang, L. Wang, K. Li, Z. Fu, and Y. Guo, "Slfnet: a stereo and lidar fusion network for depth completion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 605–10 612, 2022.

[11] 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 *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5895–5902.

[12] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2495–2504.

[13] Z. Shen, Y. Dai, and Z. Rao, "Cfnet: Cascade and fused cost volume for robust stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 906–13 915.

[14] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 international conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[15] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[16] U. Shin, J. Park, and I. S. Kweon, "Deep depth estimation from thermal image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1043–1053.

[17] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4353–4361.

[18] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[19] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 66–75.

[20] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418.

[21] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1959–1968.

[22] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 362–14 372.

[23] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[24] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[25] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 103–119.

[26] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. So Kweon, "Non-local spatial propagation network for depth completion," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 120–136.

[28] J. Tang, F.-P. Tian, W. Feng, J. Li, and P. Tan, "Learning guided convolutional network for depth completion," *IEEE Transactions on Image Processing*, vol. 30, pp. 1116–1129, 2020.

[29] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "Penet: Towards precise and efficient image guided depth completion," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 656–13 662.

[30] L. Liu, X. Song, J. Sun, X. Lyu, L. Li, Y. Liu, and L. Zhang, "Mff-net: Towards efficient monocular depth completion with multi-modal feature fusion," *IEEE Robotics and Automation Letters*, 2023.

[31] L. Teixeira, M. R. Oswald, M. Pollefeys, and M. Chli, "Aerial single-view depth completion with image-guided uncertainty estimation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1055–1062, 2020.

[32] S. S. Shivakumar, T. Nguyen, I. D. Miller, S. W. Chen, V. Kumar, and C. J. Taylor, "Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 13–20.

[33] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, "Noise-aware unsupervised deep lidar-stereo fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6339–6348.

[34] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving," *arXiv preprint arXiv:1906.06310*, 2019.

[35] N.-A.-M. Mai, P. Duthon, L. Khoudour, A. Crouzil, and S. Velastin, "Sparse lidar and stereo fusion (sls-fusion) for depth estimation and 3d object detection," 2021.

[36] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation with the 3d lidar and stereo fusion," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2156–2163.

[37] J. Zhang, M. S. Ramanagopal, R. Vasudevan, and M. Johnson-Roberson, "Listereo: Generate dense depth maps from lidar and stereo imagery," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7829–7836.

[38] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[39] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308–9316.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[41] H. Zhao, H. Zhou, Y. Zhang, Y. Zhao, Y. Yang, and T. Ouyang, "Eai-stereo: Error aware iterative network for stereo matching," in

Proceedings of the Asian Conference on Computer Vision, 2022, pp. 315–332.

[42] J. Choe, K. Joo, T. Imtiaz, and I. S. Kweon, "Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4672–4679, 2021.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.