Stereo-LiDAR Fusion by Semi-Global Matching with Discrete Disparity-Matching Cost and Semidensification

Yasuhiro Yao, Ryoichi Ishikawa, *Member, IEEE*, and Takeshi Oishi, *Member, IEEE*

***Abstract*—We present a real-time, non-learning depth estimation method that fuses Light Detection and Ranging (LiDAR) data with stereo camera input. Our approach comprises three key techniques: Semi-Global Matching (SGM) stereo with Discrete Disparity-matching Cost (DDC), semidensification of LiDAR disparity, and a consistency check that combines stereo images and LiDAR data. Each of these components is designed for parallelization on a GPU to realize real-time performance. When it was evaluated on the KITTI dataset, the proposed method achieved an error rate of 2.79%, outperforming the previous state-of-the-art real-time stereo-LiDAR fusion method, which had an error rate of 3.05%. Furthermore, we tested the proposed method in various scenarios, including different LiDAR point densities, varying weather conditions, and indoor environments, to demonstrate its high adaptability. We believe that the real-time and non-learning nature of our method makes it practical for applications in robotics and automation.**

***Index Terms*—Sensor Fusion, Computer Vision for Automation, Range Sensing**

I. Introduction

REAL-time depth estimation is crucial for a wide range of robotics and automation applications. Depth data are needed not only by autonomous vehicles but also by various mobile systems and robots to understand and navigate their environment. Standard methods for depth measurements include triangulation and time of flight (ToF). The widely used devices for these methods are stereo cameras, which rely on triangulation, and Light Detection And Ranging (LiDAR), which operates using ToF principles.

Both stereo cameras and LiDAR systems have distinct advantages and limitations. Stereo cameras deliver depth information with a high resolution that is equivalent to the resolution of the input images. However, they perform poorly on untextured surfaces, repetitive patterns, and low-light environments due to challenges in finding correspondences. In contrast, LiDAR provides more precise depth measurements and is robust against variations in lighting and surface texture. However, LiDAR data are sparse because LiDAR captures depth information only at specific points where the laser beam intersects the target scene.

Sensor fusion overcomes these drawbacks of sensor systems. In particular, a stereo-LiDAR fusion system can obtain a highly accurate, high-resolution depth map without being affected by the environment or by scenes [1]. As with other research topics, sensor fusion systems employ a learning-based [2]–[10] or non-learning-based [1], [11]–[13] strategy. For stereo-LiDAR fusion, the performance of learning-based methods can be domain-dependent, as training is conducted on a specific dataset. Meanwhile, non-learning methods have the advantage of being less dependent on specific datasets and domains.

However, the previous state-of-the-art real-time non-learning methods are not robust due to outlier-sensitive costs and direct use of LiDAR disparities, including misprojections [13]. Therefore, we propose a stereo-LiDAR fusion method, particularly a non-learning approach that operates in real time and achieves an accuracy comparable to that of learning-based methods. Fig. 1 is an overview of the proposed method. We employ Semi-Global Matching (SGM) [14] as the base stereo algorithm. The primary reason for the suboptimal accuracy of stereo-LiDAR fusion is that the integration between the stereo camera and LiDAR is insufficient. The proposed method addresses this issue by introducing the following three key approaches:

- Discrete Disparity-matching Cost (DDC) discretely evaluates sparse disparities in the SGM framework.
- Semidensification partially densifies sparse disparities to provide prior information to SGM using DDC.
- Stereo-LiDAR consistency check ensures consistency in the disparity estimation by leveraging three views from the stereo cameras and LiDAR.

In addition, we demonstrate that the proposed method surpasses previous state-of-the-art (SOTA) real-time stereo-LiDAR fusion techniques and exhibits strong adaptability across various domains.

Sec. II reviews related works to position the proposed methodology within the existing literature. In Sec. III, we provide a brief overview of stereo SGM to facilitate a better understanding of the proposed approach and the variable notation. Secs. IV-A, IV-B, and IV-C describe DDC, semidensification, and the stereo-LiDAR consistency check, respectively. We evaluate the performance of the proposed method in Sec. V and present the conclusions in Sec. VI.

II. Related Work

We consider the stereo system to be a parallel dense stereo setup, in which disparity maps are generated from a pair of images captured by two cameras aligned parallel to their image planes. Although modern deep learning methods estimate relative depth from a monocular image [15], stereo images

Manuscript received: October, 22, 2024; Revised February, 3, 2025; Accepted March, 3, 2025.

This paper was recommended for publication by Editor Cadena Lerma, Cesar upon evaluation of the Associate Editor and Reviewers' comments. This work was partially supported by JSPS KAKENHI Grant Numbers JP24K21173 and JP24H00351.

Y. Yao, R. Ishikawa, and T. Oishi were with the Institute of Industrial Science, University of Tokyo, Tokyo, 153-8505 Japan (e-mail: yao@cvl.iis.u-tokyo.ac.jp)

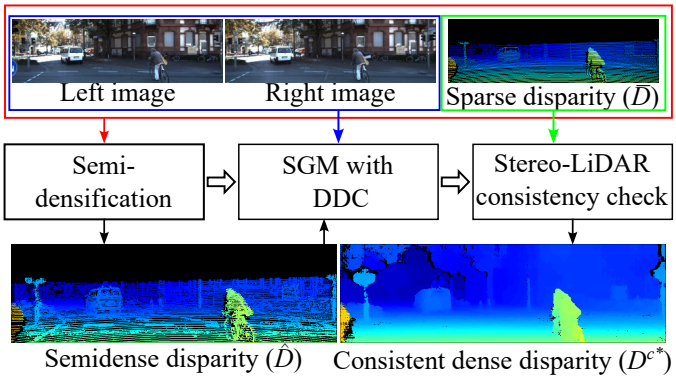


Fig. 1. Flow chart of the proposed method. The semidensification process takes stereo images and the sparse disparity map ($\bar{D}$) and outputs the semidense disparity map ($\hat{D}$). SGM with DDC takes stereo images with either a semidense disparity map ($\hat{D}$) or a sparse disparity map ($\bar{D}$) and outputs a dense disparity map. The stereo-LiDAR consistency check annotates invalid disparities based on the consistency of the three views to obtain a consistent dense disparity map (D^{c*}).

are still required to obtain depth at the real scale, which is our subject matter. In the following sections, we present a brief overview of related work on parallel dense stereo and stereo-LiDAR fusion methods.

A. Parallel dense stereo

The stereo system estimates disparities by analyzing the local similarity between two images along their epipolar lines. A widely used method for finding the optimal solution is the energy minimization approach [14], [16], [17], which includes pixel-matching cost and smoothness terms in 2D space. In most cases, this minimization problem is considered NP-hard [16]. Although several methods, such as graph cuts [16] and belief propagation [17], have been proposed to solve this problem, these methods are computationally expensive.

In contrast, SGM [14] reduces computational costs by approximating the 2D smoothness constraint using a combination of multiple one-dimensional (1D) constraints. Currently, SGM is one of the most widely used stereo-matching methods due to its high performance. Several variants of SGM have also been developed [18]–[20]. Learning-based dense stereo techniques have recently been introduced [21], [22]; however, training a model to handle all potential and unforeseen scenarios remains a challenge. For this reason, we consider SGM to be a suitable foundational algorithm for stereo matching.

B. Stereo-LiDAR fusion

Non-learning stereo-LiDAR fusion has evolved and improved over the years. Badino et al. utilized LiDAR data to narrow the stereo matching search space and introduced predefined paths for dynamic programing [11]. Maddern et al. proposed a probabilistic model to fuse LiDAR and stereo disparities by combining the priors of individual sensors [12]. Yao et al. proposed a method for selecting, using belief propagation, appropriate depth values from LiDAR projections in the surrounding area [1] and then smoothing using total generalized variation [23]. Forkel et al. incorporated a LiDAR-based matching cost into SGM stereo to determine whether an estimated depth was similar to the LiDAR measurement of the depth [13].

Many recent studies have adopted learning-based approaches and have led to significant improvements in accuracy. Park et al. first developed a neural network (NN) that integrated LiDAR and stereo disparities [2], and they formulated the problem of uncalibrated sensor fusion in a unified deep learning framework [4]. Wang et al. employed a stereo-matching network with enhanced techniques rather than directly fusing estimated depths across LiDAR and stereo modalities [3]. Cheng et al. proposed a self-supervised method for training an NN to remove occluded LiDAR projections, enabling the inference of dense disparity maps [5]. Choe et al. introduced a geometry-aware network for long-range depth estimation [8]. Zhang et al. proposed a method for coupling depth cues in two modalities in a compact network architecture [9]. Meng et al. presented a real-time, NN-based approach for coarse depth prediction and subsequent depth refinement [7], [10].

Among the comparison methods, only [11], [12] and [13] meet the criteria for real-time processing and do not require learning. However, these approaches are less accurate than offline or learning-based methods. In contrast, the proposed method is a real-time, non-learning approach that achieves competitive accuracy with both offline and learning-based methods.

III. Preliminary: Stereo SGM

In this section, we present an overview of stereo SGM [14]. SGM utilizes a strategy that minimizes the cost of pixel-wise matching while applying smoothness constraints to estimate the disparity image. We define the matching cost for a pixel $\mathbf{p} \in \Omega$ at a possible disparity $d_\mathbf{p} \in \mathbb{N}$ as $C(\mathbf{p}, d_\mathbf{p})$, where $\Omega \subset \mathbb{N}^2$ is the set of pixel coordinates. Relying solely on matching costs may result in inconsistencies across the disparity map $D = \{d_\mathbf{p} \mid \mathbf{p} \in \Omega\}$. To address this problem, SGM introduces a smoothness term that penalizes significant changes in disparity between neighboring pixels as follows:

$$E(D) = \sum_{\mathbf{p}} \Bigg\{ C(\mathbf{p}, d_\mathbf{p}) + \sum_{\mathbf{q} \in N_\mathbf{p}} P_1 T\big[|d_\mathbf{p} - d_\mathbf{q}| = 1\big] + \sum_{\mathbf{q} \in N_\mathbf{p}} P_2 T\big[|d_\mathbf{p} - d_\mathbf{q}| > 1\big] \Bigg\}, \tag{1}$$

where $T[\cdot] = 1$ when $\cdot$ is true, and $T[\cdot] = 0$ otherwise. $N_\mathbf{p}$ represents the neighboring pixels of pixel $\mathbf{p}$. P_1 is a constant penalty applied to neighboring pixel $\mathbf{q} \in N_\mathbf{p}$ when there is a small change in disparity (i.e., by one pixel). P_2 is a constant penalty for large changes in disparity. We obtain the optimal disparity image $D^* = \{d^*_\mathbf{p} \mid \mathbf{p} \in \Omega\}$ by minimizing $E(D)$. Such global minimization in 2D is NP-complete for many discontinuity-preserving energies [16].

SGM divides the problem into several 1D paths in the image. We used vertical, horizontal, and diagonal paths (eight

paths in total) in this study. The cost $L_{\mathbf{r}}$ of each path $\mathbf{r}$ is calculated by the propagation along with the path $\mathbf{r}$ as:

$$\begin{aligned} L_{\mathbf{r}}(\mathbf{p}, d_{\mathbf{p}}) = \min \big\{ & L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d_{\mathbf{p}}), L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d_{\mathbf{p}}-1) + P_1, \\ & L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d_{\mathbf{p}}+1) + P_1, \min_i L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, i) + P_2 \big\} \\ & + C(\mathbf{p}, d_{\mathbf{p}}) - \min_k L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, k). \end{aligned} \tag{2}$$

We obtain the optimal disparity of a pixel $d^*_{\mathbf{p}}$ by minimizing the aggregate cost along different paths as follows:

$$d^*_{\mathbf{p}} = \arg\min_{d_{\mathbf{p}}} \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, d_{\mathbf{p}}). \tag{3}$$

Finally, a parabola is fitted to the optimal disparities between the pixel and its two neighbors to obtain their subpixel disparities.

IV. Methodology

As shown in Fig. 1, our approach consists of three main components: semidensification, SGM with DDC, and a stereo-LiDAR consistency check. The semidensification process generates a partially densified disparity map from the sparse disparity map $\bar{D} = \left\{ \bar{d}_{\mathbf{p}} \in \mathbb{N} \cup \{\text{invalid}\} \mid \mathbf{p} \in \Omega \right\}$ using stereo images. We assume that D and $\bar{D}$ are geometrically aligned and have the same pixel coordinates. The DDC leverages a robust integration of stereo-LiDAR images to manage measurement noise and misprojection caused by occlusion and miscalibration. The stereo-LiDAR consistency check evaluates the consistency between the stereo images and LiDAR data.

A. Discrete Disparity-matching Cost

We propose a disparity-matching cost that considers the sparse disparity map derived from LiDAR measurements into the proposed SGM framework. This matching cost applies penalties based on different scenarios and takes on discrete values, similar to the strategy used in SGM. LiDAR-SGM [13] utilizes a quadratic cost; However, this approach tends to overpenalize when the disparity deviates significantly from prior disparity values and is less tolerant of misprojections in sparse disparity maps.

We define DDC by the pepenalties $(0, Q_1, Q_2)$ for the following three cases:

1) 0: no penalty if estimated disparity matches prior value; this preserves accurately measured data,
2) Q_1: small penalty when the estimated disparity slightly differs from the prior, thereby allowing for handling noise,
3) Q_2: fixed penalty for larger differences that accounts for misprojections and enables disparity estimation away from the prior,

under the condition that $Q_1 \leq Q_2$.

The baseline stereo-matching cost is the Hamming distance of census-transformed images [24], $H\colon \Omega \times \mathbb{N} \to \mathbb{R}$. By combining the stereo-matching cost and DDC, we derive the joint-matching cost $\bar{C}$ as follows:

$$\begin{aligned} \bar{C}(\mathbf{p}, d_{\mathbf{p}}) = (1-\alpha)\, H(\mathbf{p}, d_{\mathbf{p}}) + \alpha \Big\{ & Q_1 T\left[\left| d_{\mathbf{p}} - \bar{d}_{\mathbf{p}} \right| = 1 \right] \\ & + Q_2 T\left[\left| d_{\mathbf{p}} - \bar{d}_{\mathbf{p}} \right| > 1 \right] \Big\}, \end{aligned} \tag{4}$$

where α is a parameter that balances the contributions of the stereo and disparity-matching costs. The cost C in Eq. 2 is replaced with $\bar{C}$ in Eq. 4, and we find the optimal disparity by solving Eq. 3.

Note that the sparse disparity map used for the DDC computation originates from either the semidensification step or directly from the LiDAR disparity data.

B. Semidensification

Semidensification enhances the prior information extracted from the sparse disparity map to generate the semidense disparity map $\hat{D} = \left\{ \hat{d}_{\mathbf{p}} \in \mathbb{N} \cup \{\text{invalid}\} \mid \mathbf{p} \in \Omega \right\}$. By increasing the density of the sparse disparity map, DDC gains more impact since DDC is only applied for pixels where the sparse disparity value exists. In addition, this process eliminates misprojections to ensure DDC robustness.

The semidensification process fills the sparse disparity map at a pixel $\mathbf{p}$ with the semidense disparity $\hat{d}_{\mathbf{p}}$. This disparity, $\hat{d}_{\mathbf{p}}$, must minimize the cost of stereo matching within the disparities in $M_{\mathbf{p}}$, where $M_{\mathbf{p}}$ is defined as a window of size $(2r_s+1) \times (2r_s+1)$ centered at $\mathbf{p}$. In addition, the minimum matching cost must be less than the threshold value T_s. If no neighboring disparities meet these conditions, the sparse disparity value is used directly, and it remains invalid if no sparse data are available. Thus, the semidense disparity $\hat{d}_{\mathbf{p}}$ is calculated as follows:

$$\hat{d}_{\mathbf{p}} = \begin{cases} \arg\min_{\bar{d}_{\mathbf{q}} | \mathbf{q} \in M_{\mathbf{p}}} H\left(\mathbf{p}, \bar{d}_{\mathbf{q}}\right) & \text{if } \min H\left(\mathbf{p}, \bar{d}_{\mathbf{q}}\right) < T_s, \\ \bar{d}_{\mathbf{p}} & \text{otherwise.} \end{cases} \tag{5}$$

Note that this process may mitigate misprojections when the matching cost is high due to factors like occlusion by replacing the disparity value with a neighboring value that achieves a low matching cost. The semidense disparity $\hat{d}_{\mathbf{p}}$ is used as an alternative to the original sparse disparity $\bar{d}_{\mathbf{p}}$ when the matching cost is calculated (Eq. 4).

We refrain from applying spatial smoothing during the semidensification process because the addition of a smoothness term would lead to high computational costs. In addition, preserving the details of small or thin objects is more effective at this stage. Because spatial smoothing is applied later in the SGM process, the semidensification step focuses exclusively on the matching cost.

C. Stereo-LiDAR consistency check

The consistency check filters out disparity values that are not consistent across multiple views. In SGM [14], the consistency check uses two camera views. However, stereo-LiDAR fusion involves three views, two from the cameras and one from the LiDAR; thus, it is more efficient to incorporate the LiDAR view into the consistency check process than to rely solely on stereo camera views.

Assume that we estimate the disparity $d^*_{\mathbf{p}}$ at a pixel $\mathbf{p}$ of the base image and the disparity of the matching pixel in the other image $d'_{\mathbf{q}_m}$. The pixel $\mathbf{q}_m$ is obtained by traversing the epipolar line on the matching image: $\mathbf{q}_m = e_{\text{bm}}(\mathbf{p}, d^*_{\mathbf{p}})$. If the corresponding disparities differ significantly, the disparity is set

TABLE I
VARIATIONS OF OUR METHOD IN THE EVALUATIONS

Name	Semi-densification	SGM with DDC	Stereo-LiDAR consistency check
DSGM		√	√
SDSGM	√	√	√

TABLE II
OUR PARAMETERS IN THE EVALUATIONS

Stage	Name	Value	Meaning
Semi-densification	T_s	2	Threshold for semidensification
	r_s	6	Window size for semidensification
SGM with DDC	P_1	10	Small SGM smoothness cost
	P_2	120	Large SGM smoothness cost
	Q_1	5	Small disparity-matching cost
	Q_2	160	Large disparity-matching cost
	α	0.7	Blending ratio of costs
Consistency-check	T_c	2	Threshold for consistency check
	r_c	20	Window size for consistency check

to invalid [14]. The proposed method includes a consistency check between the base camera and LiDAR. We consider that the disparity is consistent if the estimated disparity $d^*_{\mathbf{p}}$ matches at least one of the disparities in its neighboring pixels in the sparse disparity map $\bar{d}_{\mathbf{q}} \mid \mathbf{q} \in K_{\mathbf{p}}$. Here, $K_{\mathbf{p}}$ is defined as a window of size $(2r_c+1) \times (2r_c+1)$ centered at $\mathbf{p}$. $d^*_{\mathbf{p}}$ and $\bar{d}_{\mathbf{q}}$ are considered to be matched if $\left|d^*_{\mathbf{p}} - \bar{d}_{\mathbf{q}}\right| \leq T_c$, where T_c is a given threshold value. Finally, the integrated three-view consistency check can be expressed as follows:

$$d^{c*}_{\mathbf{p}} = \begin{cases} d^*_{\mathbf{p}} & \text{if } \left|d^*_{\mathbf{p}} - d'_{\mathbf{q}_m}\right| \leq 1, \mathbf{q}_m = e_{\mathrm{bm}}(\mathbf{p}, d^*_{\mathbf{p}}), \\ & \text{or if } \min_{\mathbf{q} \in K_{\mathbf{p}}} \left|d^*_{\mathbf{p}} - \bar{d}_{\mathbf{q}}\right| \leq T_c. \\ \text{invalid} & \text{otherwise.} \end{cases} \tag{6}$$

Note that we did not use the semidense disparity map during the consistency check because the propagated disparities may not satisfy the geometric relations between the sensors. The final output of our method is a consistent dense disparity map $D^{c*} = \left\{d^{c*}_{\mathbf{p}} \mid \mathbf{p} \in \Omega\right\}$ derived by Eq. 6.

V. EXPERIMENTS

We compared the proposed method to SOTA stereo-LiDAR fusion and non-learning stereo approaches by analyzing the contribution of each component and its robustness across different scenarios. Our evaluation covers the accuracy of and processing time required by the proposed method (Sec. V-B), the impact of semidensification, and the role of the consistency check (Sec. V-C). In addition, we tested the proposed method's robustness under various conditions, including different input densities (Sec. V-D), other weather conditions, and indoor scenes (Sec. V-E). We also evaluated the effect of varying parameters (Sec. V-F). Ablation studies highlight the differences in performance between the DSGM and SDSGM variations, as detailed in Table I, where the distinction between the variations is the application of semidensification. The parameters used in our evaluations other than Sec. V-F are described in Table II.

A. Implementation and dataset

The proposed method was integrated with an open-source SGM implementation in CUDA[1]. The comparative methods were run on our platform when the implementation was available. Otherwise, we referenced the results reported in the original studies. The platform used in these experiments was an NVIDIA Jetson Orin NX with 16GB of memory.

We utilized the KITTI 141 dataset, which is a subset of the KITTI stereo dataset [25]. The KITTI 141 was extracted by [12] and is one of the datasets most commonly used to benchmark stereo-LiDAR fusion methods. The dataset contains 141 sets of rectified stereo images, LiDAR point clouds captured by Velodyne HDL 64E, and corresponding ground truth dense disparity images (Sec. V-B and V-C). We also used 32- and 16-scan-line disparity maps created by vertically sampling the 64-scan-line original map to half and quarter densities based on scan angle (Sec. V-D). For the evaluation, we used the code provided with the KITTI benchmark[2].

To evaluate the adaptability to various scenes (Sec. V-E), we applied the method to the CARLA[3] and Middlebury[4] dataset. CARLA is a dataset of simulated outdoor scenes proposed in [13], including 500 sets of rectified stereo images and 64-scan-lines LiDAR data under two coupled weather and time conditions, named "ClearSunset" and "HardRainNoon". Middlebury is a dataset of indoor scenes by [26]. We utilized the 2021 mobile dataset, including 24 sets of rectified stereo images and a ground truth disparity map. We randomly sampled 1% of each ground truth map to obtain the corresponding sparse disparity map.

B. Overall performance

First, we compare the overall performance of the proposed method with the performances of existing approaches. In addition, to evaluate DDC compared with an outlier-robust cost, we implemented a variant of LiDAR-SGM [13] using Huber cost. The KITTI evaluation code provides two error rates: the covered error and the total error. The covered error measures the error only in regions where valid estimations exist, excluding invalid areas. In contrast, the total error calculates the error rate by filling invalid pixels through the background interpolation [14]. We used the error rate to represent the percentage of cases where the estimated value differed from the ground truth by three pixels or more. Because most LiDARs work at 10–20 Hz, we consider a method to be real-time if its processing time per frame is less than 100 ms.

Table III shows the quantitative evaluation results. Among the non-learning methods, the proposed method achieves sufficient coverage and the lowest error rate for both the covered and total errors. In addition, the proposed method outperforms methods that are not real-time. By comparing the results of DSGM with LiDAR-SGM [13] and its variant, we found DDC was more effective than the quadratic and Huber costs. We consider this because DDC's discreteness managed the

[1]https://github.com/fixstars/libSGM
[2]https://www.cvlibs.net/datasets/kitti/eval_scene_flow.php
[3]https://www.mucar3.de/icra2023-lidar-sgm
[4]https://vision.middlebury.edu/stereo/data/scenes2021/

TABLE III
DISPARITY ESTIMATION RESULTS ON KITTI 141

Method	Input	Non-learning	Realtime	Time [ms]	GPU platform	Coverage [%]	Covered error [%]	Total error [%]
SGM [14]	Stereo	√	√	37	Jetson Orin NX	93.0	3.73	6.00
JointEst. [18]	Stereo	√		1947	-	99.9	4.32	4.33
SSM-TGV [1]	Stereo + LiDAR	√		250	Jetson Orin NX	100.0	3.32	3.32
Probabilistic [12]	Stereo + LiDAR	√	√	(24)	AMD R9 295x2	(99.6)	(5.91)	-
LiDAR-SGM [13]	Stereo + LiDAR	√	√	(24)	GTX 1050 Ti	(97.5)	(3.87)	-
LiDAR-SGM (Huber)	Stereo + LiDAR	√	√	39	Jetson Orin NX	93.7	3.50	4.93
DSGM (Ours)	Stereo + LiDAR	√	√	40	Jetson Orin NX	99.0	2.81	3.15
SDSGM (Ours)	Stereo + LiDAR	√	√	50	Jetson Orin NX	99.6	**2.61**	**2.79**
CNN [2]	Stereo + LiDAR		√	(45)	Titan X	(99.8)	(4.84)	-
Fastfusion [10]	Stereo + LiDAR		√	(49)	Titan Xp	100.0	3.05	3.05
CCVNorm [3]	Stereo + LiDAR			(1011)	GTX 1080 Ti	(100.0)	(3.35)	(3.35)
LSNet [5]	Stereo + LiDAR			3284	Jetson Orin NX	100.0	**2.17**	**2.17**

The **bests among non-learning or realtime methods** are indicated in red. The **bests among all** are indicated in blue.
Values in brackets were obtained from the cited papers.

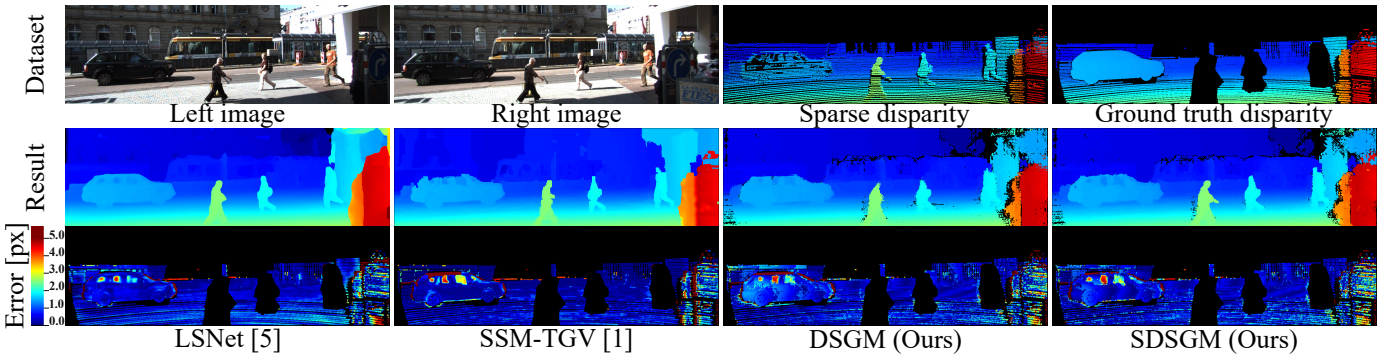


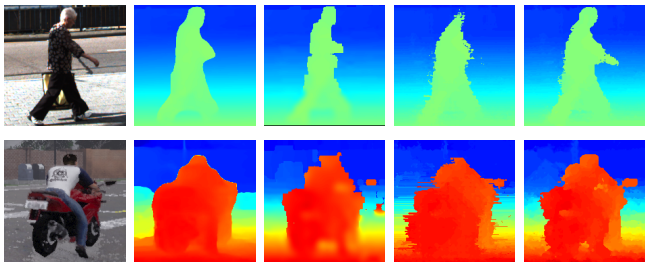
Fig. 2. Dataset and results of KITTI 141 evaluation. Overall, offline learning-based LSNet [5] showed the least error, as seen on the car in the error maps. SSM-TGV [1] showed a more significant error than our methods, as seen on the car roof in the error map.

TABLE IV
SEMIDENSIFICATION EFFECTS ON KITTI 141

	Coverage [%]	Covered error [%]
Input sparse disparity map	7.3	4.66
Semidense disparity map	28.6	**4.15**

misprojection and noise in the sparse disparity map more efficiently than other costs. Due to the high coverage rate of learning-based methods, directly comparing the covered error rates does not constitute a fair comparison. However, the total error rate of the proposed method is lower than that of FastFusion [10]. Although LSNet [5] achieves the lowest overall error rate, it requires approximately 60 times more computational time than does the proposed approach.

Figure 2 shows the example dataset and visual results, highlighting the qualitative evaluations. The figure shows the best methods of non-learning and learning for comparison. As in Fig. 3, SDSGM successfully recovered the detailed silhouette of the pedestrian. Conventional methods often oversmooth thin or small objects due to their strong constraints on smoothness. In contrast, the semidensification effectively generates the proper prior information for these objects because it does not enforce smoothness terms.

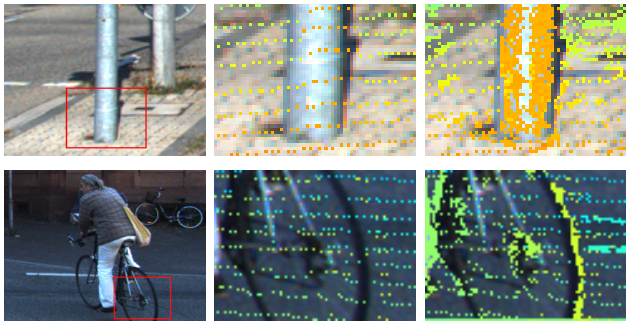


(a) Images (b) LSNet [5] (c) SSM-TGV [1] (d) DSGM (Ours) (e) SDSGM (Ours)

Fig. 3. Resulting SDSGM silhouettes are finer than those of the other methods (Upper: KITTI and Lower: CARLA HardRainNoon). A background interpolation [14] was used to carry out the visual comparison. The effect is visible at (Upper) the arm and (Lower) the side mirror areas.

C. Ablation studies

1) Semidensification: We verified the impact of semidensification in enhancing sparse input disparities in SGM with DDC. Table IV compares the original sparse disparity maps with our semidense disparity maps on the KITTI 141 dataset in terms of accuracy and density. The results indicate that semidensification improves both the coverage and error rates. These improvements are visually demonstrated in Fig. 4, where semidensification simultaneously removes misprojection and densifies the disparity map.



(a) Image (b) Sparse map (c) Semidense map

Fig. 4. Top-row images: Misprojections in (b) the sparse map appear as disparities of the road (yellow dots) on the pole. Misprojections decreased in (c) the semidense disparity. Bottom-row images: (b) A few LiDAR disparities are projected on thin objects, such as the bicycle wheels. (c) The semidense map contains disparity values for such objects.

TABLE V
CONSISTENCY CHECK EFFECTS ON KITTI 141

Input	Coverage [%]	Covered error [%]	Total error [%]
Stereo	97.1	**1.95**	3.12
Camera-LiDAR	99.4	2.57	2.81
Stereo-LiDAR	99.6	2.61	**2.79**

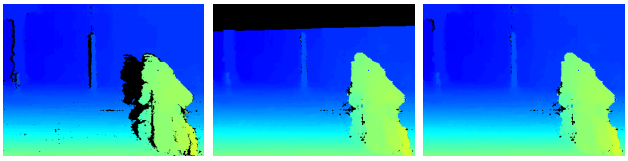
2) Stereo-LiDAR consistency check: We evaluated the proposed stereo-LiDAR consistency check and compared it with the conventional stereo consistency check [14] and the camera-LiDAR consistency check. Here, the camera-LiDAR check only checks the consistency between the base camera and LiDAR. The quantitative and qualitative results are shown in Table V and Fig. 5, in which all disparity maps have been generated by SDSGM prior to the consistency checks. Since the stereo and camera-LiDAR consistency checks label more invalid pixels than the stereo-LiDAR check, they reduce coverage and improve the covered error. Meanwhile, the stereo-LiDAR check achieved the most coverage and the least total error.

D. Robustness to LiDAR density

We applied the proposed and comparison methods to KITTI 141 with 32 and 16 LiDAR scan lines and obtained the results in Table VI. The proposed method outperformed a previous non-learning offline SOTA approach [1] for maps having both 32 and 16 scan lines. In addition, with maps having 32 scan lines, the proposed method achieved a total error of 3.27% and outperformed the result of [1] with maps having 64 scan lines, which was 3.32% (refer to Table III).

E. Adaptability to various datasets

To assess the performance of the proposed method across different scenarios, we evaluated the method by CARLA ClearSunset, CARLA HardRainNoon, and Middlebury datasets. For adaptability evaluation purposes, the parameters used in the compared methods and ours were the same as those used in the KITTI experiments. Figure 6 and 7 presents the visual results, and Table VII provides the quantitative results. The proposed SDSGM achieved the best



(a) Stereo (b) Camera-LiDAR (c) Stereo-LiDAR

Fig. 5. Qualitative evaluations of the consistency checks. (a) The stereo check labels the area blocked in the second camera as invalid. (b) Camera-LiDAR check labels the area outside the LiDAR field of view as invalid. (c) Stereo-LiDAR check has the most valid pixels.

TABLE VI
RESULTS ON DIFFERENT LIDAR DENSITIES

Scan lines	Method	Time [ms]	Cove-rage[%]	Covered error[%]	Total error[%]
32	SSM-TGV [1]	241	100.00	4.00	4.00
	DSGM (Ours)	39	96.82	3.04	3.77
	SDSGM (Ours)	51	98.99	**2.92**	**3.27**
16	SSM-TGV [1]	239	100.00	5.14	5.14
	DSGM (Ours)	37	94.79	3.32	4.43
	SDSGM (Ours)	51	98.11	**3.34**	**3.91**

performance on the ClearSunset and Middlebury datasets. LSNet, which utilized the same parameters and NN weights as were described in Sec. V-B, showed a significantly higher error rate than did the non-learning methods. This demonstrates the advantage of non-learning methods over learning-based approaches in terms of adaptability without fine-tuning.

For the HardRainNoon dataset, DSGM outperformed SDSGM, suggesting that semidensification does not improve the accuracy in this scenario. In contrast, when using a proper parameter ($T_s = 9$), semidensification enhances accuracy, as in the row indicated with † in Table VII. Hard rain scenes present a challenge to stereo vision because rain introduces noise into the image, which increases the stereo-matching cost, even for correct disparity estimations. As a result, T_s should be higher than in other scenarios to accommodate the larger stereo-matching cost.

F. Parameter Study

We evaluated the effect of parameters across all datasets used above. We used the same values of P_1 and P_2 as the original stereo SGM implementation. For other parameters we introduced, we evaluated the effect on the total error by varying them. The graphs and discussions in Fig. 8 highlight the evaluation. The effects of each parameter were relatively independent, so each graph in Fig. 8 focuses on one parameter for clarity of illustration. Although parameter tuning is possible for each dataset, we found that the parameters in Table II are well-balanced for different scenarios.

VI. CONCLUSION

Stereo-LiDAR fusion is a technology that enhances depth estimation by combining stereo matching with LiDAR data. We focus on real-time, non-learning stereo-LiDAR fusion because it can be applied across various domains without the need for additional network training. The proposed method

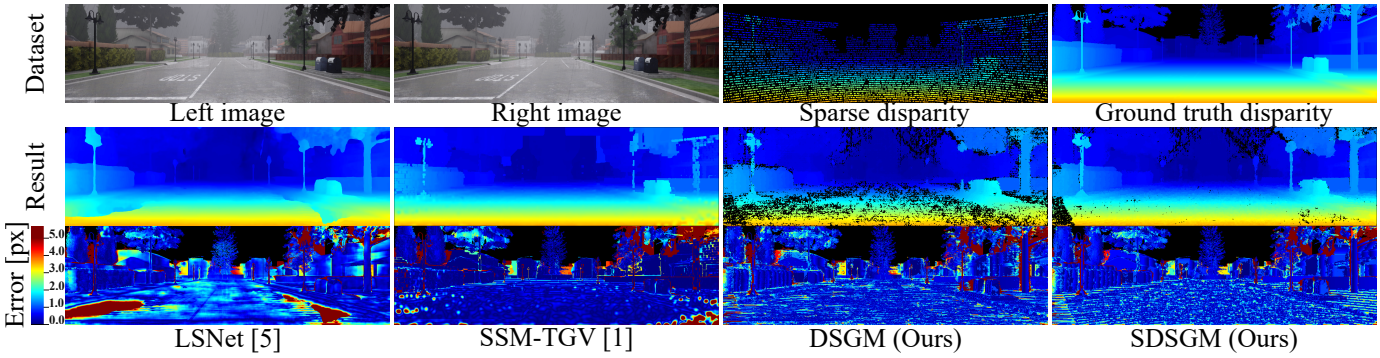


Fig. 6. CARLA HardRainNoon dataset and results. The displayed SDSGM result is obtained using a proper semidensification threshold († in Table VII). LSNet [5] showed a significantly larger error than non-learning methods. SSM-TGV [1] showed the significant error at the right lowest corner of the image. Overall, our methods showed more minor errors than the compared methods.

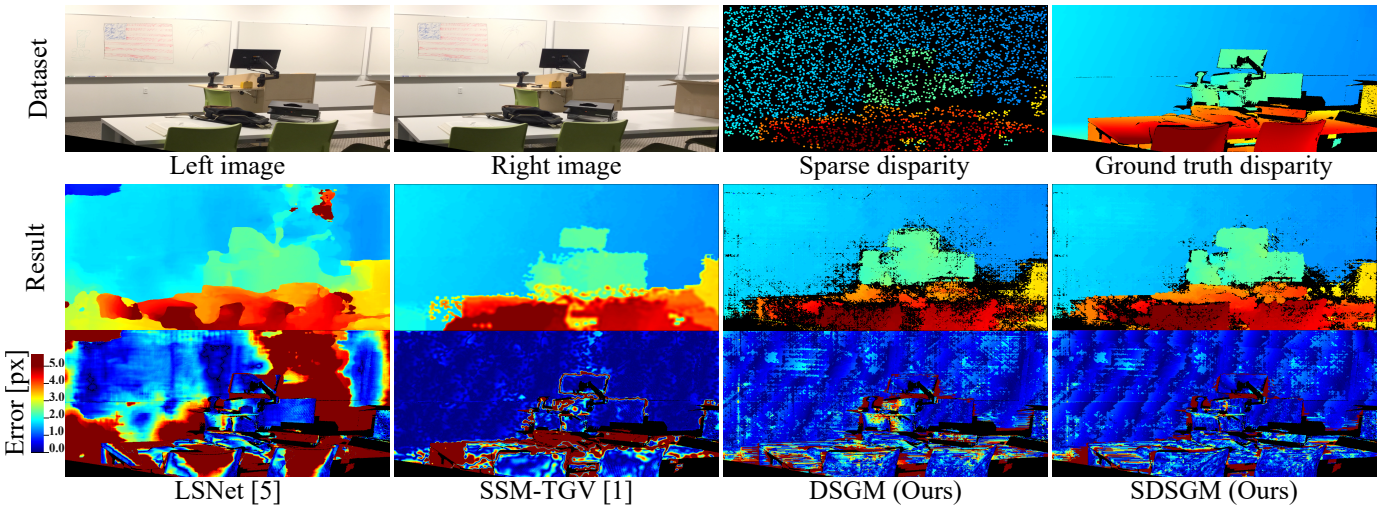


Fig. 7. Middlebury dataset and results. LSNet [5] showed a significantly larger error than non-learning methods. SSM-TGV [1] has a significant error on the desk board. Overall, our methods showed more minor errors than the compared methods.

integrates SGM with DDC, semidensification, and a stereo-LiDAR consistency check. We demonstrate that the proposed method achieves improved performance over previous real-time stereo-LiDAR fusion methods in terms of error rate and demonstrates strong adaptability across different domains.

REFERENCES

[1] Y. Yao, R. Ishikawa, S. Ando, K. Kurata, N. Ito, J. Shimamura, and T. Oishi, "Non-learning stereo-aided depth completion under mis-projection via selective stereo matching," *IEEE Access*, vol. 9, pp. 136 674–136 686, 2021.

[2] 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.

[3] 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.

[4] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation using uncalibrated lidar and stereo fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 321–335, 2019.

[5] 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.

[6] 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.

[7] H. Meng, C. Zhong, J. Gu, and G. Chen, "A gpu-accelerated deep stereo-lidar fusion for real-time high-precision dense depth sensing," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021, pp. 523–528.

[8] 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.

[9] 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.

[10] H. Meng, C. Li, C. Zhong, J. Gu, G. Chen, and A. Knoll, "Fastfusion: Deep stereo-lidar fusion for real-time high-precision dense depth sensing," *Journal of Field Robotics*, vol. 40, no. 7, pp. 1804–1816, 2023.

[11] H. Badino, D. Huber, T. Kanade *et al.*, "Integrating lidar into stereo for fast and improved disparity computation," in *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*. IEEE, 2011, pp. 405–412.

[12] W. Maddern and P. Newman, "Real-time probabilistic fusion of sparse 3d lidar and dense stereo," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2181–2188.

[13] B. Forkel and H.-J. Wuensche, "Lidar-sgm: Semi-global matching on lidar point clouds and their cost-based fusion into stereo matching," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2841–2847.

[14] H. Hirschmuller, "Stereo processing by semiglobal matching and mu-

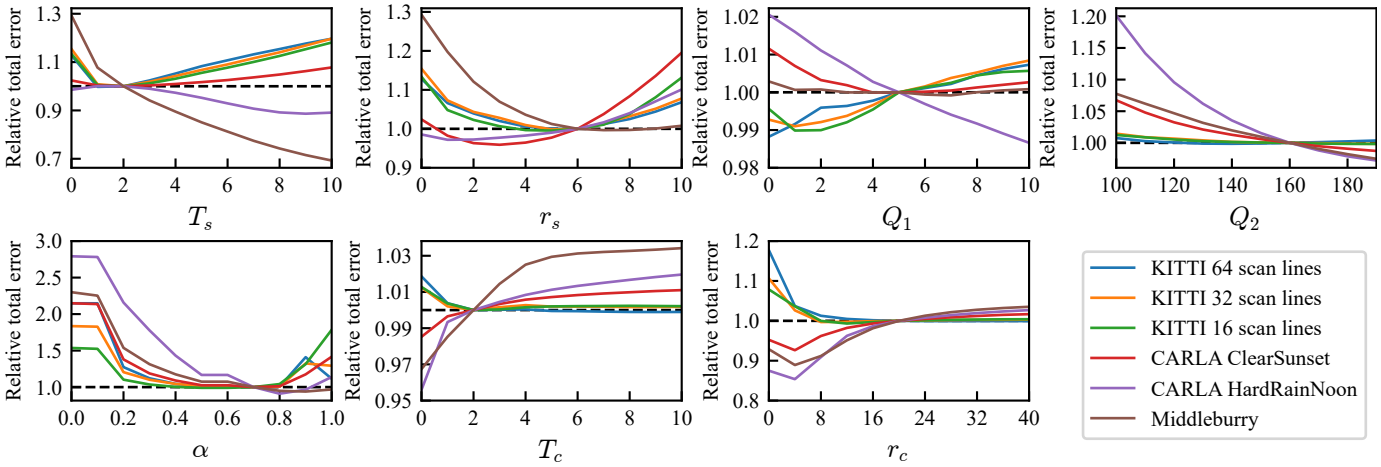


Fig. 8. Parameter study results. In a graph, the X-axis is the varied parameter, and the Y-axis is the total error relative to the total error using the parameters in Table II. The parameter-wise observations follow. Large T_s and r_s, which densify the semidense disparity maps, improved the results for Middlebury. We consider this due to the less coverage of Middlebury sparse disparity maps (1 %) than other datasets. Large Q_1 and Q_2, which strongly impose the sparse disparity matching constraints, performed well for CARLA and Middlebury. We consider this to be the case because the sparse data in these artificial datasets (created through computer graphics or sampling) is more reliable than in real datasets. Similarly, small T_c and r_c, which lead to the consistency check to the stereo-only like, improved results for CARLA and Middlebury. The reason for this is considered as follows. Since the LiDAR is located at the same position as the left camera in these artificial datasets, the LiDAR consistency check is biased to keep more foreground disparities and less background disparities. In contrast, for the real dataset of KITTI, we see the stereo-LiDAR consistency check improved the results.

TABLE VII
RESULTS ON VARIOUS DATASETS

Dataset	Method	Time [ms]	Coverage [%]	Covered error[%]	Total error [%]
CARLA Clear Sunset (Outdoor)	SGM [14]	33	69.5	11.20	21.58
	JointEst. [18]	1824	99.9	10.21	10.26
	SSM-TGV [1]	239	99.4	12.97	13.10
	LSGM [13]	-	-	-	(12.95)
	DSGM (Ours)	35	80.9	**7.34**	10.29
	SDSGM (Ours)	49	87.5	7.40	**10.05**
	LSNet [5]	3123	100.0	17.26	17.26
CARLA Hard Rain Noon (Outdoor)	SGM [14]	34	63.6	18.77	34.45
	JointEst. [18]	1834	99.6	23.39	23.58
	SSM-TGV [1]	239	99.4	13.35	13.49
	LSGM [13]	-	-	-	(14.03)
	DSGM (Ours)	35	74.0	**8.94**	**12.16**
	SDSGM (Ours)	52	77.0	9.30	12.34
	†*SDSGM (Ours)*	*54*	*92.6*	*8.23*	*10.94*
	LSNet [5]	3124	100.0	28.69	28.69
Middleburry (Indoor)	SGM [14]	38	81.3	10.00	16.78
	JointEst. [18]	1423	99.9	13.50	13.45
	SSM-TGV [1]	217	99.7	15.80	15.80
	DSGM (Ours)	43	92.8	7.55	9.43
	SDSGM (Ours)	59	95.8	**6.17**	**7.30**
	LSNet [5]	4621	100.0	10.46	10.46

† A reference using proper semidensification parameter ($T_s = 9$).
Values in brackets were obtained from the cited papers.

tual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.
[15] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.
[16] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.
[17] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International journal of computer vision*, vol. 70, pp. 41–54, 2006.
[18] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *European Conference on Computer Vision*. Springer, 2014, pp. 756–771.
[19] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. C. Moure, and A. M. López, "Embedded real-time stereo estimation via semi-global matching on the gpu," *Procedia Computer Science*, vol. 80, pp. 143–153, 2016.
[20] J. Kallwies, T. Engler, B. Forkel, and H.-J. Wuensche, "Triple-sgm: stereo processing using semi-global matching with cost fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 192–200.
[21] H. Wang, R. Fan, P. Cai, and M. Liu, "Pvstereo: Pyramid voting module for end-to-end self-supervised stereo matching," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4353–4360, 2021.
[22] X. Guo, J. Lu, C. Zhang, Y. Wang, Y. Duan, T. Yang, Z. Zhu, and L. Chen, "Openstereo: A comprehensive benchmark for stereo matching and strong baseline," *arXiv preprint arXiv:2312.00343*, 2023.
[23] Y. Yao, M. Roxas, R. Ishikawa, S. Ando, J. Shimamura, and T. Oishi, "Discontinuous and smooth depth completion with binary anisotropic diffusion tensor," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5128–5135, 2020.
[24] R. Spangenberg, T. Langner, and R. Rojas, "Weighted semi-global matching and center-symmetric census transform for robust driver assistance," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2013, pp. 34–41.
[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.
[26] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*. Springer, 2014, pp. 31–42.