An Extensible Multi-Sensor Fusion Framework for 3D Imaging

Talha Ahmad Siddiqui Rishi Madhok* Matthew O'Toole
Carnegie Mellon University
{tsiddiqu, rmadhok}@alumni.cmu.edu, mpotoole@cmu.edu

Abstract

*Many autonomous vehicles rely on an array of sensors for safe navigation, where each sensor captures different visual attributes from the surrounding environment. For example, a single conventional camera captures high-resolution images but no 3D information; a LiDAR provides excellent range information but poor spatial resolution; and a prototype single-photon LiDAR (SP-LiDAR) can provide a dense but noisy representation of the 3D scene. Although the outputs of these sensors vary dramatically (*e.g.*, 2D images, point clouds, 3D volumes), they all derive from the same 3D scene. We propose an extensible sensor fusion framework that (1) lifts the sensor output to volumetric representations of the 3D scene, (2) fuses these volumes together, and (3) processes the resulting volume with a deep neural network to generate a depth (or disparity) map. Although our framework can potentially extend to many types of sensors, we focus on fusing combinations of three imaging systems: monocular/stereo cameras, regular LiDARs, and SP-LiDARs. To train our neural network, we generate a synthetic dataset through CARLA that contains the individual measurements. We also conduct various fusion ablation experiments and evaluate the results of different sensor combinations.*

1. Introduction

An important challenge faced by the self-driving car industry is safety. When driving down a road, an autonomous vehicle needs to reliably "see" its surroundings in order to make safe decisions. Moreover, it is important to drive reliably in adverse weather conditions (rain, snow, or fog), operate at different times of the day under different lighting conditions (day or night), and detect pedestrians, bicyclists, or other cars in the presence of partial occluders.

Many self-driving cars are therefore equipped with a wide variety of sensors, such as cameras, LiDARs, RADARs, and IMUs, to perceive their 3D environment reliably. The high-resolution 2D images from conventional cameras can be used to identify cars, signs, or pedestrians on the road, but provide poor depth perception by themselves. LiDARs are ideal for detecting the 3D position of objects on the road, but provide limited spatial resolution.

New computational sensors are also on the horizon, such as the emergence of single-photon LiDARs (SP-LiDAR) [20]. A SP-LiDAR uses an extremely-sensitive photo-detector known as a single-photon avalanche diode (SPAD) [29], which can detect 3D points with far fewer

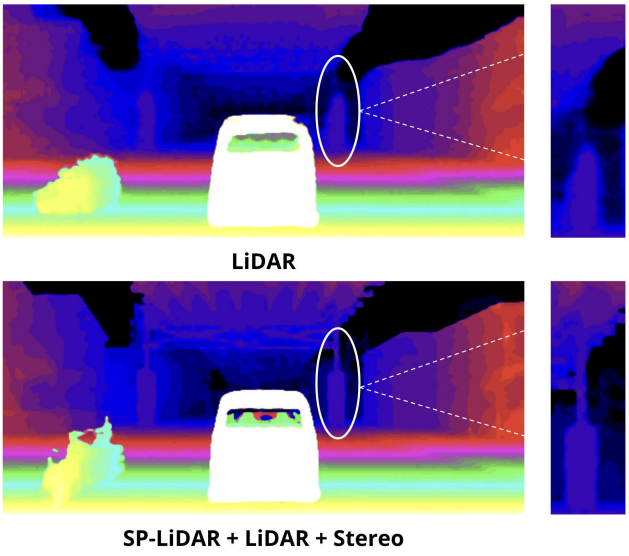

Figure 1: Disparity maps generated with our framework, using two sensor combinations: LiDAR only (top), and the fusion of a SP-LiDAR, a stereo camera, and a LiDAR (bottom). The highlighted insets on the right show the preservation of fine details when combining multiple sensors together. Note that we report all results in terms of disparity, because our network builds on PSMNet [4]: a stereo matching framework. All depth values (*e.g.*, measurements captured with the LiDARs) are therefore converted into disparity values.

*Now affiliated with Microsoft.

photons than a conventional multi-photon LiDAR. However, SP-LiDARs are also sensitive to ambient photons (*e.g.*, light emitted by our sun), resulting in both denser but noisier measurements. SP-LiDARs have been used for 3D imaging at $10\,\mathrm{km}$ range [17, 28], sensing through hazy environments (*e.g.*, fog or murky water) [22, 31], and even imaging objects hidden around corners [27].

Combining sensory data from multiple sources helps to overcome the limitations of any one sensor. In this work, we propose an extensible framework for fusing the output of a heterogeneous sensor array. Our approach involves lifting the sensors' measurements (*e.g.*, a 2D image, a collection of 3D points, or a noisy 3D volume) to a temporary 4D volume, referred to as a *cost volume*. Similar lifting operations have been used for 3D geometric reasoning tasks [33], where the volume can be interpreted as a voxelized representation of the scene. We process the resulting cost volume with a deep neural network to extract the disparity map of a scene, as shown in Figure 1. Note that we report results in terms of disparity, because the basis for our solution is a stereo matching framework called PSMNet [4].

Our framework is able to fuse different sensor combinations, including monocular cameras, stereo cameras, LiDARs, and SP-LiDARs. Because all these sensor combinations do not exist in current available datasets, we simulate our data using CARLA [10], an open source simulated environment that supports development of autonomous urban driving systems. We also experiment with different signal-to-background noise ratios (SBR) for SP-LiDARs, and find that fusion with SP-LiDAR produces good results even when subjected to very poor SBR conditions.

The key contributions of our work include

- an extensible fusion framework for heterogeneous sensor arrays that lift measurements into a common volumetric representation;
- a new simulated dataset that contains measurements for stereo cameras, LiDARs, and SP-LiDARs; and
- an evaluation of sensor fusion on the task of disparity estimation using different combinations of sensors, including a monocular camera, stereo camera, LiDAR, and SP-LiDAR.

2. Related Work

2.1. 3D Sensors used in Autonomous Vehicles

Light Detection and Ranging (LiDAR) systems [12] are commonly used in autonomous vehicles. These systems work by firing a pulse of light at an object, measuring the time required for the light to return in response, and using this response to infer the 3D position and reflectivity of the object. The photodectors used in conventional LiDARs may require upwards of hundreds or thousands of photons to measure a single 3D point. The LiDAR repeats this process multiple times while scanning the environment to produce a 3D point cloud. These point clouds tend to be sparse however, especially for objects far from the sensor; see Figure 3 for an example.

A single-photon LiDAR (SP-LiDAR) uses a single-photon sensor (*e.g.*, a single photon avalanche diode, or SPAD [29]) to produce measurements from individual photons. SP-LiDARs are therefore much more efficient than conventional multi-photon LiDARs, and can produce higher resolution scans and detect objects at longer distances [17, 34, 35]. Unfortunately, SP-LiDARs are also far more sensitive to the ambient light present in an environment (*e.g.*, sunlight), resulting in many spurious 3D points. Recovering 3D shape thus requires censoring the noisy photon present in the measurements [20].

In this work, we focus on fusing both current and emerging LiDAR systems with mono or stereo camera systems. We aim to combine the best features of all available sensors, by leveraging the high-spatial resolution of regular cameras and 3D information recovered from LiDARs.

2.2. Sensor Fusion for Vision tasks

Besides 3D sensing, many works aim to solve higher-level vision tasks, such 3D object detection and segmentation. Prior approaches have focused on using a single imaging modality, such as monocular cameras [6], stereo cameras [7], and LiDARs [24, 41]. More recently, pseudo-LiDAR based approaches [40, 37] have also shown significant improvements for 3D object detection; these methods convert the depth maps from a stereo camera into 3D point cloud, and process this point cloud directly to solve the detection task.

Fusion architectures have also been proposed for 3D object detection, combining information from multiple imaging modalities [11]. Fusion generally occurs either in 2D space or 3D space, and typically focuses on fusing camera and LiDAR information. For 2D fusion, LiDAR data is processed in either range view (LiDAR's native view) [23] or Bird's Eye View (BEV) [18, 19, 32, 38] and fused with RGB images. Cheng *et al*. [9] employ fusion in 3D space where the sensor data is used to create a volumetric representation and given as input to the fusion network. Road segmentation tasks have also been performed using LiDAR and camera fusion [3, 8, 21]. Other combinations of sensors, *e.g.*, RADARs and cameras, have been used recently for 3D object detection [25, 26]. However, fusion with emerging sensor technologies, like SP-LiDARs, have not received as much attention.

2.3. Disparity Estimation

Unsupervised learning based approaches, such as [9], perform depth estimation using a LiDAR and stereo cam-

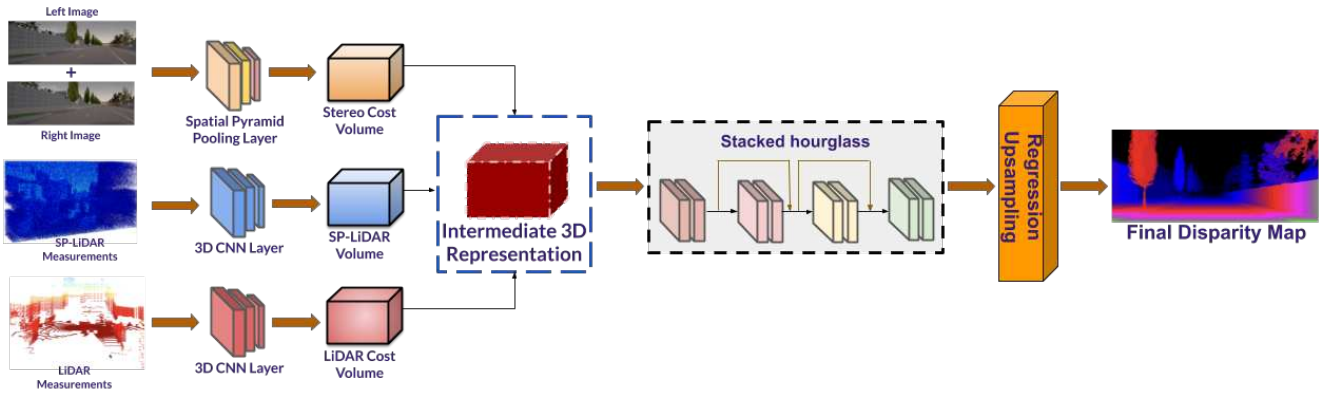

Figure 2: Overview of our proposed model. We take as input raw sensor measurements from three sensors: a stereo camera system, SP-LiDAR, and LiDAR. We first calculate a cost volume representation for each of the three sensory inputs (Section 3.1). We then normalize these cost volumes and fuse them by addition (Section 3.2). Finally, we pass the fused volume through a series of 3D CNNs that regress towards a disparity map of the scene.

era fusion architecture with noisy LiDAR points. Recently, a lot of progress has been made in disparity estimation using stereo image pairs [13, 14, 30, 39]. We formulate our problem of fusing multiple sensors for the task of disparity estimation. Specifically, we use PSMNet [4] as a base model which is used for disparity estimation using a stereo image pair. The model architecture first captures global contextual information using spatial pyramid pooling layers, and produces a cost volume. The 3D convolution layers then regularize this cost volume using stacked multiple hourglass networks.

We extend the PSMNet architecture to add fusion branches for SP-LiDAR and a regular LiDAR, generating cost volumes for each imaging modalities. Recently, Lindell *et al*. [20] showcased that fusing information from a SP-LiDAR and a high-resolution camera image significantly improves depth estimation, even with low signal-to-background noise ratio. However, their current architecture does not handle information from stereo image pairs. We therefore take inspiration from both [4] and [20] to propose a novel fusion architecture that forms cost volume representation for a stereo camera system, SP-LiDAR, and LiDAR.

3. Proposed Approach

Our objective is to fuse sensor data obtained from multiple sources, and output a disparity map of the scene. The core idea of our approach is to convert all sensor data into volumetric representations of the scene, referred to as cost volumes. Our 4D cost volumes (height $\times$ width $\times$ disparity $\times$ features) encode features over 3D space; specifically, we discretize 3D space according to the spatial resolution of our camera (height $= 256$, width $= 512$) across a range of disparity values for a given baseline (disparity $= 192$). We hypothesize that fusion can be done efficiently by first lifting sensor data into this common representation of 3D space. Moreover, this lifting operation can be extended to other sensors as well, provided that there exists a logical mapping of sensor data to its corresponding cost volume.

Figure 2 provides an overview of our proposed approach. The input consists of a pair of 2D images from a stereo camera system, a sparse 3D point cloud obtained by a LiDAR, and a dense but noisy volume given by a SP-LiDAR; see Figure 3 for visualizations of the input. First, we map the sensor data to cost volumes. These volumes capture complementary information of the same 3D scene. Second, we pass each of these individual volumes through 3D CNNs that learn respective features associated with each of these volumes. Third, we normalize each of these volumes and fuse them by addition. Normalization converts the cost volume to the same scale, thereby making fusion by addition robust. Fourth, we pass the normalized-fused intermediate cost volume to a stacked hourglass architecture [4] and regress towards a 2D disparity map.

In the remainder of this section, we describe the cost volume construction and sensor fusion network in more detail.

3.1. Cost Volume Constructions

Mono- and Stereo-Camera Systems We assume a stereo camera system captures left and right RGB images with a spatial resolution of 512×256 and a 90° horizontal field of view. Given a camera baseline B and focal length f, the disparity image d is

$$d = \frac{Bf}{z} \tag{1}$$

where z is the corresponding depth map.

We follow the approach used in PSMNet [4] to construct our initial cost volume from this stereo image pair. Left and right images are first passed through a weight-sharing CNN

followed by spatial pyramid pooling layers to capture contextual information. The corresponding feature maps are then concatenated across every disparity level to generate a 4D cost volume CV_C, with dimension (height $\times$ width $\times$ disparity $\times$ features).

We can also construct a 4D cost volume for a monocular camera, where only the left image is used as an input to the network. The cost volume is constructed by replicating the values in the image along the disparity dimension, and passing the result through a 3D CNN network to generate the 4D cost volume.

Note that all disparity maps and cost volumes are computed with respect to the left camera.

LiDAR Systems We assume a 64 channel LiDAR captures a point cloud for a standard full 360° scan. Our objective is to lift the sparse LiDAR point cloud into a 4D cost volume CV_L, where the entries in both CV_C and CV_L represent the same points in 3D space. This will ensure that the cost volumes are spatially compatible.

Let's assume that the LiDAR and the left camera of the stereo system share the same center of projection. We trim the point cloud data to a 90° field of view to match our left camera. The depth measurements obtained from LiDAR are then converted to the disparity domain by using Equation (1). This produces a sparse 3D volume, where a voxel with a value of 1 indicates the position of a 3D LiDAR point. Finally, we process the result with a 3D CNN network to form our 4D cost volume CV_L.

SP-LiDAR Systems We follow the same approach described by Lindell *et al*. [20] to generate dense and noisy SP-LiDAR volumes, except that we discretize our volumes with respect to disparity levels instead of depth values.

In order to evaluate the robustness of our fusion strategy, we explored three different signal-to-background ratios (SBR) between the signal and ambient photons: **SBR 0.052**, **SBR 0.0052** and **SBR 0.00052**.[1] In each of these experiments, we simulate SP-LiDAR data using just 1 signal photon per pixel on average; as a result, the number of 3D points captured with a SP-LiDAR is higher than a regular LiDAR. We then add a number of random ambient photons according to our target SBR value. For example, at a SBR of 0.0052, there are 192 ambient photons (an average of 1 ambient photon per histogram bin).

Photons detected by a SP-LiDAR are modeled using a Poissonian process [29]. Hence, in our SP-LiDAR simulation, we introduce Poisson noise to both the signal and the ambient light level to replicate the single-photon characteristics. Figure 4 shows responses across different disparity values for a few SBR values. Figure 4(a) represents the scenario when there is no noise or ambient light, and hence the position of the peak represents the disparity value (*i.e*., the object's distance from the SP-LiDAR). Figure 4(b-d) represents the same signal but with Poisson noise for SBR values 0.05, 0.005, and 0.0005 respectively. An increase in the number of ambient photons produces a noisier SP-LiDAR volume, hence making it more difficult to extract the correct disparity values.

[1] We abbreviate these values as SBR 0.05, 0.005, and 0.0005 for the rest of this paper. Note that these values represent noisier measurements than those evaluated by Lindell *et al*. [20], where SBR was 1.0, 0.1, and 0.04.

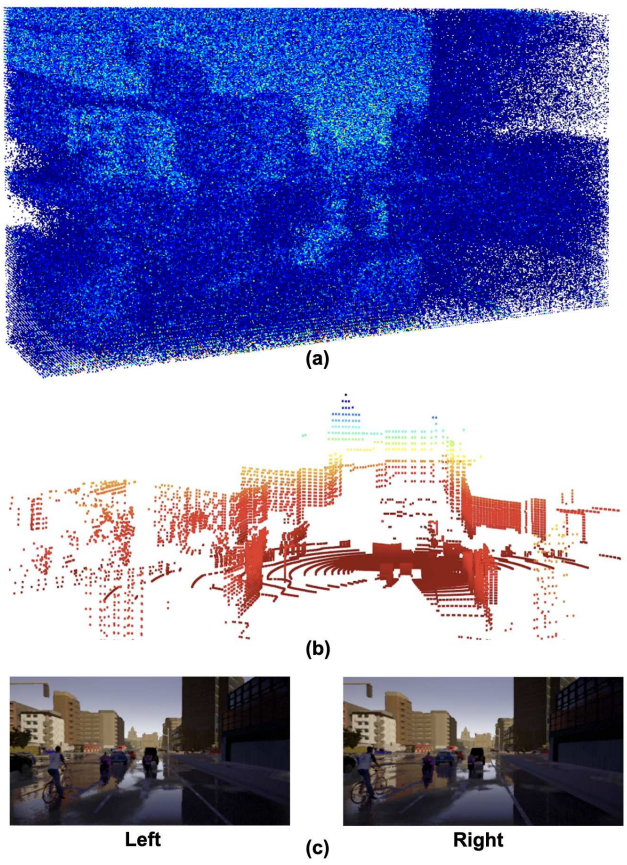

Figure 3: Inputs to our proposed model. **(a)** Dense and noisy volume representing photons detected with a SP-LiDAR, using an SBR of 0.005. **(b)** Sparse point cloud from a conventional LiDAR. **(c)** Left and right camera images of the scene.

Our cost volume construction for SP-LiDARs is similar to that of LiDARs. For every detected photon, we increment by 1 the value of the corresponding voxel. This results in a dense and noisy volume. This volume is once again passed through 3D CNNs to generate the 4D feature volume CV_S, sharing the same dimensions as the other cost volumes.

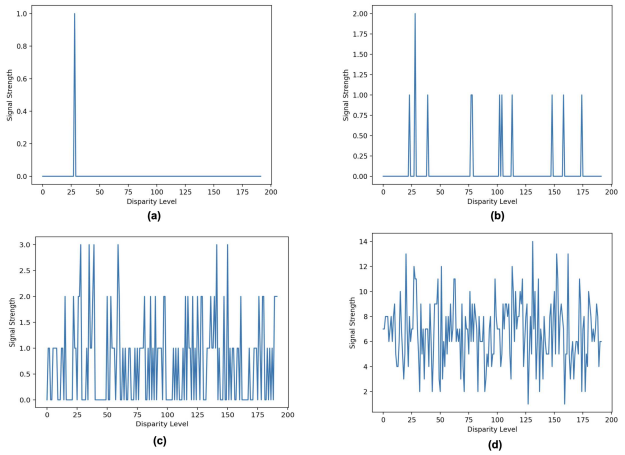

Figure 4: Visualization of one sample of SP-LiDAR data, *i.e.*, representing the signal received by a specific pixel. **(a)** When no noise is present, the object reflects light back at a specific time, which can be converted into a disparity value. **(b-d)** In practice, the same signal shown in (a) is corrupted by ambient photons, which arrive at random times. Here, for the ground truth signal shown in (a), we illustrate three different signal-to-background noise ratios: (b) SBR 0.05, (c) SBR 0.005, and (d) SBR 0.0005.

3.2. Sensor Fusion Network

The last step of our pipeline is to combine and process the individual cost volumes. Each of the cost volumes obtained have the dimension (height $\times$ width $\times$ disparity $\times$ features), where the feature dimension has length 64. The individual cost volumes are now instance normalized and then added together to output a fusion vector (CV_F) of the same size as these individual cost volumes.

$$CV_F = Norm(CV_C) + Norm(CV_S) + Norm(CV_L) \quad (2)$$

The fused cost volume CV_F is then passed to the stacked hourglass 3D CNN architecture as described in [4]. This is followed by an upsampling layer to output a 3D vector of dimension (height $\times$ width $\times$ disparity) via bilinear interpolation. Finally, as described in [14], we regress to the disparity map of size (height $\times$ width).

Given the ground truth disparity d, the predicted disparity $\hat{d}$, and the number of labeled pixels M, we minimize a smooth L_1 loss function [4]

$$L(d, \hat{d}) = \frac{1}{M} \sum_{i=1}^{M} S_{L_1}(d_i - \hat{d}_i), \quad (3)$$

where

$$S_{L_1}(y) = \begin{cases} 0.5y^2, & \text{if } |y| < 1, \\ |y| - 0.5, & \text{otherwise.} \end{cases}$$

4. Experiments

Our proposed fusion algorithm is evaluated on data simulated with CARLA [10]. We also perform ablation studies with different sensor inputs to the fusion network and evaluate the performance of each sensor combination. We describe our dataset and its properties, followed by experimental details. We then showcase our results, both quantitatively and qualitatively, and finally discuss the implications of our proposed fusion approach.

Dataset	SP-LiDAR	LiDAR	Stereo
nuScenes [2]	✗	✓	✗
Argoverse [5]	✗	✓	✓
Lyft Level 5 [15]	✗	✓	✗
Waymo Dataset [1]	✗	✓	✗
Ours (CARLA)	✓	✓	✓

Table 1: Different datasets and available sensors.

4.1. CARLA Dataset

Recently, many real world datasets have been released by various self driving car companies, a summary of which is presented in Table 1. However, none of these datasets capture SP-LiDAR measurements and only Argoverse [5] provides stereo data. Moreover, our task is to fuse sensor information to compute a disparity map of the scene, which requires ground truth depth values to train our network; this information is not directly available in real-world datasets.

We therefore collect our own data using CARLA, an open source simulator designed to develop and test algorithms for autonomous vehicle driving systems [10]. CARLA gives access to different virtual driving routes where we can place other actors like vehicles, pedestrians, and cyclists. For our data collection, we select one such vehicle to drive in the autopilot mode around different driving routes. We mount 4 different sensors on top of the vehicle to capture scenes: two cameras to capture the stereo pair, one LiDAR, and one depth sensor to capture the ground truth disparity values. We simulate SP-LiDAR measurements from the ground truth disparity values, as discussed in Section 3.1.

Figure 5 shows sample frames captured in CARLA, highlighting scene diversity. To make our dataset robust and challenging, we collected data under different ambient conditions and varying road traffic. Stereo images have a resolution of 256×512, and the LiDAR has 64 channels and detects objects up to a range of $100\,\mathrm{m}$. In all, we captured 10,000 frames (8,000 for training and 2,000 for testing).

4.2. Experiment Details

Our proposed model is implemented in PyTorch, building on top of the current PSMNet architecture. Our network

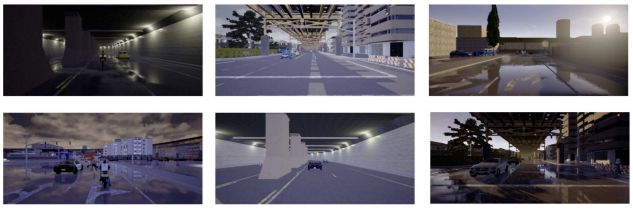

Figure 5: Sample frames captured in CARLA. We captured scenes under different ambient and traffic conditions.

is trained end-to-end with the Adam optimizer [16] with a learning rate of 0.001. The maximum disparity value d is 192. To compare our fusion performance with stereo matching networks, we use as our metric the average root mean square error (RMSE) between the predicted and ground truth disparity values. We also report the percentage of pixels that have disparity error greater than three pixels (>3px) and one pixel (>1px) respectively. Our model is trained from randomly initialized weights for each of the ablation experiments (detailed in Section 4.3). Training time for the network is approximately 24 hours on two NVIDIA RTX 2080Ti GPUs with a batch size of 4.

4.3. Ablation Experiments and Results

We divide our ablation studies into three sections and list both quantitative and qualitative results. In Section 4.3.1, we discuss the performance of the system when only one sensor is given as input to the network. In Section 4.3.2, we perform monocular + LiDAR and monocular + SP-LiDAR fusion, and compare the performance of the two LiDARs. Finally, in Section 4.3.3, we fuse stereo + LiDAR and stereo + SP-LiDAR, and compare the results with monocular fusion. We also discuss the results of combining all available sensors: stereo + SP-LiDAR + LiDAR.

4.3.1 Unimodal Experiments

In our first set of ablation experiments, we compute the disparity map by utilizing only one modality at a time. This is done by weighting the other input branches to zero in the fusion network. Table 2 summarizes the results of different single sensor experiments quantitatively and Figure 6 provides a qualitative analysis.

We observe that having just the LiDAR modality or a single camera as independent inputs produce poor results. The stereo camera and SP-LiDAR perform better, with low average RMSE and low pixel error rates. For example, in Figure 6, we observe that finer details like the inside of the vehicle are captured more accurately for both SP-LiDAR and the stereo camera. With a LiDAR or a monocular camera, we observe a fair amount of noise, especially for distant regions in the scene. We also compute the metrics for SP-LiDAR with decreasing SBR values and observe that, although the performance degrades slightly, SP-LiDAR performs better than regular LiDAR even under poor ambient light conditions.

Experiment/Metric	RMSE	>3px	>1px
Monocular	2.021	4.1	13.9
LiDAR	3.081	7.7	18.0
Stereo	1.719	3.0	10.2
SP-LiDAR (SBR 0.05)	**1.576**	**2.3**	**6.9**
SP-LiDAR (SBR 0.005)	1.610	2.3	7.4
SP-LiDAR (SBR 0.0005)	1.701	2.6	9.0

Table 2: Ablation experiments with single sensor input. Our framework performs well for both SP-LiDAR and stereo camera.

4.3.2 Monocular Fusion Experiments

For the next set of experiments, we fuse a monocular camera with either a LiDAR or SP-LiDAR. The quantitative results are presented in Table 3.

We observe that fusing LiDAR with a monocular camera slightly improves all the metrics as opposed to using just one sensor at a time, thus confirming that the fusion network is leveraging complementary information from both sensors. This can also be verified qualitatively in Figure 7; finer geometric details, such as the shape of the vehicle, are estimated more accurately.

Quantitatively, the fusion results of the SP-LiDAR + monocular camera are better than LiDAR + monocular camera fusion. Even with low SBR values, SP-LiDAR fusion is more effective than LiDAR-based fusion. However, we do not observe a significant improvement with monocular + SP-LiDAR fusion when compared to just SP-LiDAR. We hypothesize that SP-LiDAR data captures both range and spatial information fairly well, as can be seen in the volume constructed in Figure 3; we notice little benefit in fusing these measurements with a regular image in our framework. Note that this experiment is similar to Lindell *et al*. [20] work, which does improve depth reconstructions by fusing a high-resolution image and the output of a SP-LiDAR.

4.3.3 Stereo Fusion Experiments

Finally, we fuse information from a stereo camera with the LiDARs. Quantitative results are presented in Table 4. The table shows that fusing stereo information with both LiDAR and SP-LiDAR data significantly improves all the metrics.

SP-LiDAR and stereo fusion with a SBR of 0.05 achieves the best results. In Figure 8, under noisier settings (SBR of 0.005), we observe that the tree trunk is not accurately captured with any one sensor but the fusion of SP-

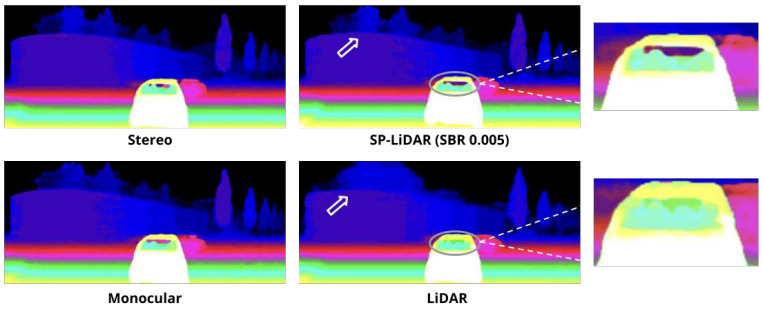

Figure 6: Qualitative results for the unimodal experiments. The SP-LiDAR and stereo camera capture finer geometric details like the interior of a vehicle (see highlighted inset), when compared to the LiDAR and monocular camera. LiDAR also produces noisy depth maps for distant regions in the scene as opposed to SP-LiDAR (highlighted by the arrow).

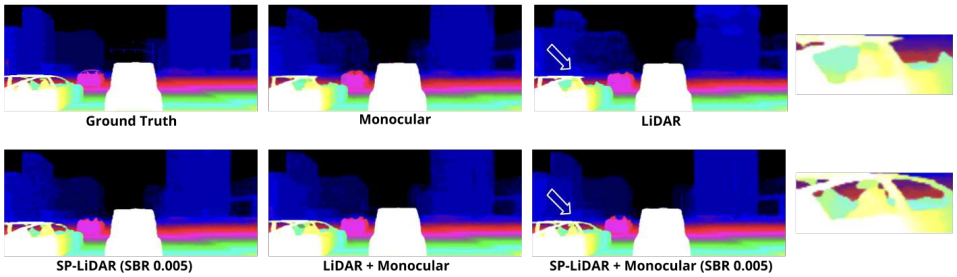

Figure 7: Qualitative fusion results of the two LiDARs with a monocular camera. Fusion captures finer details such as the contour of the vehicle (see highlighted inset).

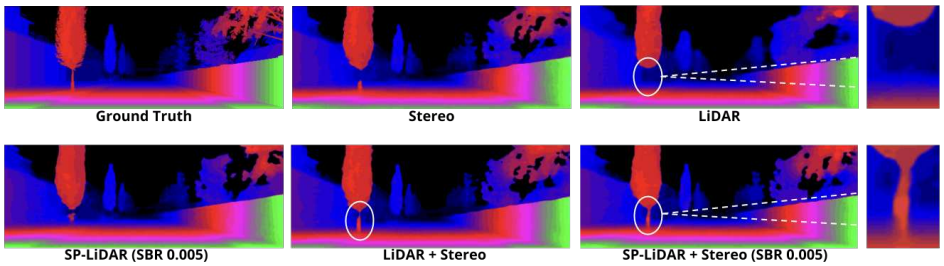

Figure 8: Qualitative fusion results of the two LiDARs with a stereo camera. Fine details like the trunk of the tree (see highlighted inset) are captured accurately with sensor fusion.

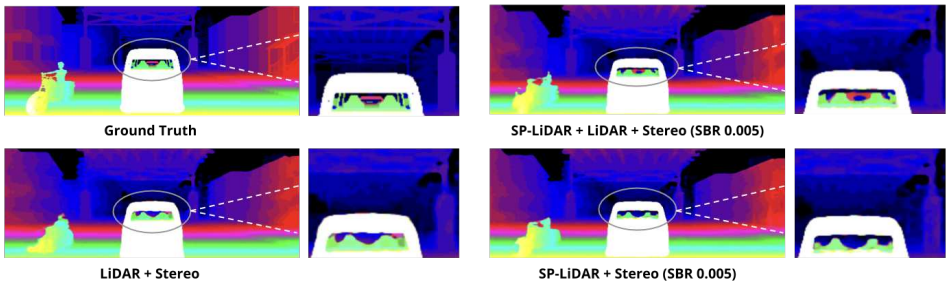

Figure 9: Qualitative fusion results for all the three sensors combined: stereo + SP-LiDAR + LiDAR. The overall predicted disparity map with three sensor fusion captures very intricate details and produces fewer artifacts when compared to any two sensor fusion combination.

Experiment/Metric	RMSE	>3px	>1px
LiDAR + Mono	2.017	3.7	12.3
SP-LiDAR + Mono (SBR 0.05)	**1.489**	**2.0**	**6.6**
SP-LiDAR + Mono (SBR 0.005)	1.670	2.3	7.1
SP-LiDAR + Mono (SBR 0.0005)	1.684	2.6	9.0

Table 3: Ablation fusion experiments with monocular camera and different LiDARs, based on Lindell *et al*. [20].

Experiment/Metric	RMSE	>3px	>1px
LiDAR + Stereo	1.645	2.6	8.8
SP-LiDAR + Stereo (SBR 0.05)	**1.341**	**1.6**	**5.2**
SP-LiDAR + Stereo (SBR 0.005)	1.398	1.7	5.5
SP-LiDAR + Stereo (SBR 0.0005)	1.570	2.1	7.3
SP-LiDAR + LiDAR + Stereo (SBR 0.005)	1.453	1.9	5.3

Table 4: Ablation fusion experiments with stereo pair and different LiDARs.

LiDAR and stereo information significantly improves performance. Even at the lowest SBR value of 0.0005, the fusion results shown in Table 4 achieves significant improvements when compared to individual sensors, demonstrating the effectiveness of our proposed fusion network.

To test the robustness of our proposed fusion architecture, we also fuse all three input modalities, *i.e*., stereo + SP-LiDAR + LiDAR using a SBR of 0.005. Qualitatively, we observe that three sensor fusion significantly improves the disparity maps. For example, in Figure 9, we find that the base of the overhead bridge is captured accurately in the three sensor fusion, when compared to two sensor fusion. However, quantitatively, we do not find a significant improvement in the metrics, with only the >1-pixel error improving slightly. This may be attributed to how the LiDAR is simulated in CARLA, where the SP-LiDAR measurements do not benefit from the inclusion of sparse LiDAR measurements.

4.4. Further Discussion

CARLA's simulation of a standard LiDAR generates a sparse point cloud, as shown in Figure 3, whereas the SP-LiDAR simulation produces dense, noisy measurements. We observe that denser data produces much better disparity maps despite the additional noise, and is effective in fusion experiments. In all our fusion experiments, a SP-LiDAR always produces the more accurate disparity map when compared to a regular LiDAR.

RGB information comes from two sources: a monocular camera or a stereo camera. We show in our proposed model that the cost volume constructed for a stereo camera provides more reliable fusion results when compared to the monocular case. Lindell *et al*. [20] showcased that fusing a high-resolution camera with a SP-LiDAR improves depth estimation results. Our proposed stereo + SP-LiDAR fusion further improves results compared to monocular fusion, significantly reducing disparity pixel error rates.

We perform various ablation experiments with different sensor fusion combinations. We also observe that the sensor combination SP-LiDAR + stereo produces equally good results as the three sensor fusion combination. This may be attributed to sparse LiDAR measurements not providing much information beyond what is already captured with a SP-LiDAR and stereo camera.

5. Conclusion

In this work, we present a fusion framework that takes measurements from a heterogeneous sensor array, lifts them to a shared 4D cost volume representing the surrounding 3D environment, and processes the result to obtain a high-quality disparity map of the scene. We estimate these disparity maps by fusing LiDARs, monocular cameras, stereo cameras, and SP-LiDARs, and we conduct ablation experiments to evaluate the results of different sensor combinations. We also present a new simulated dataset that includes measurements for all the sensors discussed in this paper.

In our experiments, multi-sensor fusion significantly improves the depth prediction when compared to working with a single sensor. We also observe that dense SP-LiDAR measurements produce more accurate depth maps when compared to sparse LiDAR point clouds, and stereo cameras provide useful information for fusion over monocular cameras. These observations are of critical importance when designing the sensing capabilities of any self-driving car.

We believe that our proposed fusion architecture can be extended to support other sensors as well, including existing RADAR technologies used for automotive applications and novel computational imaging systems that provide unique 3D sensing capabilities (*e.g*., programmable light curtains [36]). Importantly, we believe this fusion framework can also be used to address higher-level vision tasks such as 3D object detection and segmentation, and is an important step towards achieving full driving automation.

Acknowledgments

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used in this work.

References

[1] Waymo open dataset: An autonomous driving dataset, 2019.

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[3] Luca Caltagirone, Mauro Bellone, Lennart Svensson, and Mattias Wahde. Lidar–camera fusion for road detection using fully convolutional neural networks. *Robotics and Autonomous Systems*, 111:125–131, 2019.

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019.

[6] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1259–1272, 2017.

[8] Zhe Chen, Jing Zhang, and Dacheng Tao. Progressive lidar adaptation for road detection. *IEEE/CAA Journal of Automatica Sinica*, 6(3):693–702, 2019.

[9] Xuelian Cheng, Yiran Zhong, Yuchao Dai, Pan Ji, and Hongdong Li. Noise-aware unsupervised deep lidar-stereo fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6339–6348, 2019.

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*, 2017.

[11] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Fabian Duffhauss, Claudius Glaeser, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *arXiv preprint arXiv:1902.07830*, 2019.

[12] Jeff Hecht. Lidar for self-driving cars. *Optics and Photonics News*, 29(1):26–33, 2018.

[13] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3195–3204, 2019.

[14] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[15] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Lyft level 5 av dataset 2019. urlhttps://level5.lyft.com/dataset/, 2019.

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

[17] Zheng-Ping Li, Xin Huang, Yuan Cao, Bin Wang, Yu-Huai Li, Weijie Jin, Chao Yu, Jun Zhang, Qiang Zhang, Cheng-Zhi Peng, et al. Single-photon computational 3d imaging at 45 km. *arXiv preprint arXiv:1904.10341*, 2019.

[18] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3D object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 7345–7353, 2019.

[19] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018.

[20] David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. Single-Photon 3D Imaging with Deep Sensor Fusion. *ACM Trans. Graph. (SIGGRAPH)*, 37(4):1–12, 2018.

[21] Xin Lv, Ziyi Liu, Jingmin Xin, and Nanning Zheng. A novel approach for detecting road based on two-stream fusion fully convolutional network. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1464–1469. IEEE, 2018.

[22] Aurora Maccarone, Aongus McCarthy, Abderrahim Halimi, Rachael Tobin, Andy M Wallace, Yvan Petillot, Steve McLaughlin, and Gerald S Buller. Depth imaging in highly scattering underwater environments using time-correlated single-photon counting. In *Emerging Imaging and Sensing Technologies*. International Society for Optics and Photonics, 2016.

[23] Gregory P Meyer, Jake Charland, Darshan Hegde, Ankit Laddha, and Carlos Vallespi-Gonzalez. Sensor fusion for joint 3d object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[24] Gregory P Meyer, Ankit Laddha, Eric Kee, Carlos Vallespi-Gonzalez, and Carl K Wellington. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12677–12686, 2019.

[25] Michael Meyer and Georg Kuschk. Deep learning based 3d object detection for automotive radar and camera. In *2019 16th European Radar Conference (EuRAD)*, pages 133–136. IEEE, 2019.

[26] Ramin Nabati and Hairong Qi. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3093–3097. IEEE, 2019.

[27] Matthew O'Toole, David B Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, 2018.

[28] Agata M Pawlikowska, Abderrahim Halimi, Robert A Lamb, and Gerald S Buller. Single-photon three-dimensional imaging at up to 10 kilometers range. *Optics express*, 25(10):11919–11931, 2017.

[29] Adithya K Pediredla, Aswin C Sankaranarayanan, Mauro Buttafava, Alberto Tosi, and Ashok Veeraraghavan. Signal processing based pile-up compensation for gated single-photon avalanche diodes. *arXiv preprint arXiv:1806.07437*, 2018.

[30] Matteo Poggi, Davide Pallotti, Fabio Tosi, and Stefano Mattoccia. Guided stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 979–988, 2019.

[31] Guy Satat, Matthew Tancik, and Ramesh Raskar. Towards photography through realistic fog. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2018.

[32] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. MVX-Net: Multimodal voxelnet for 3D object detection. In *IEEE Int. Conf. Robotics and Automation*, 2019.

[33] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.

[34] Anu Swatantran, Hao Tang, Terence Barrett, Phil DeCola, and Ralph Dubayah. Rapid, high-resolution forest structure and terrain mapping over large areas using single photon lidar. *Scientific reports*, 6:28277, 2016.

[35] Julian Tachella, Yoann Altmann, Ximing Ren, Aongus McCarthy, Gerald S Buller, Stephen Mclaughlin, and Jean-Yves Tourneret. Bayesian 3d reconstruction of complex scenes from single-photon lidar data. *SIAM Journal on Imaging Sciences*, 12(1):521–550, 2019.

[36] Jian Wang, Joseph Bartels, William Whittaker, Aswin C Sankaranarayanan, and Srinivasa G Narasimhan. Programmable triangulation light curtains. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018.

[37] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019.

[38] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*. IEEE, 2019.

[39] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019.

[40] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019.

[41] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.