S^3: Learnable Sparse Signal Superdensity for Guided Depth Estimation

Yu-Kai Huang Yueh-Cheng Liu Tsung-Han Wu Hung-Ting Su Yu-Cheng Chang
Tsung-Lin Tsou Yu-An Wang Winston H. Hsu
National Taiwan University

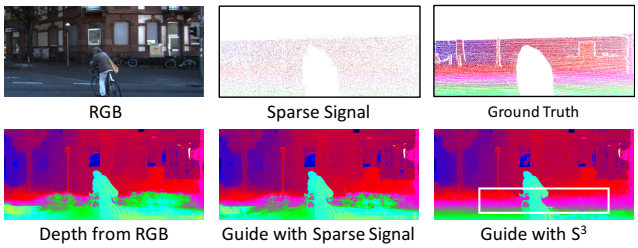
Abstract

Dense depth estimation plays a key role in multiple applications such as robotics, 3D reconstruction, and augmented reality. While sparse signal, e.g., LiDAR and Radar, has been leveraged as guidance for enhancing dense depth estimation, the improvement is limited due to its low density and imbalanced distribution. To maximize the utility from the sparse source, we propose Sparse Signal Superdensity *(S^3) technique, which expands the depth value from sparse cues while estimating the confidence of expanded region. The proposed S^3 can be applied to various guided depth estimation approaches and trained end-to-end at different stages, including input, cost volume and output. Extensive experiments demonstrate the effectiveness, robustness, and flexibility of the S^3 technique on LiDAR and Radar signal.*

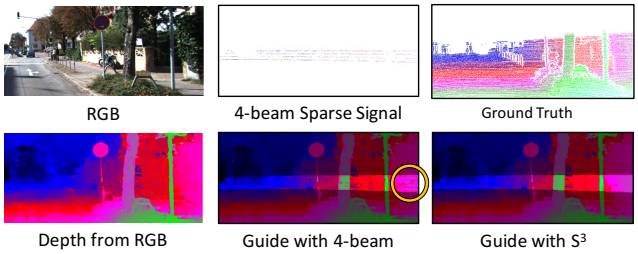
1. Introduction

Dense depth estimation is crucial in the field of 3D reconstruction [14], 3D object detection [44, 47], and robotic vision [25, 28]. Many works have proposed to estimate depth from RGB images or stereo pairs. Yet, the stereo estimation could be unreliable on homogeneous planes, large illumination changes, and repetitive textures [38, 43]; while monocular depth estimation is an ill-posed problem [11] and inherently ambiguous and unreliable [20, 24]. To attain a higher level of robustness and accuracy, modern solutions commonly leverage raw sparse signal, such as LiDAR [2, 34, 24] and Radar [5, 29], to improve depth estimation results or object detection for the challenging outdoor scenes, termed *guidance* in this paper.

Despite the success of those sparse-guidance methods, however, we still find two big problems with sparse signal. First, raw sparse signal can be ignored by networks when it is largely different from depth predicted with RGB (shown in Figure 1a). This situation stems from the low density property of the sparse signal, which is a common problem in many large-scale dataset. For example, KITTI dataset [12] wraps up an average density of 4.0% and



(a) Low Density Problem.



(b) Imbalanced Signal Distribution.

Figure 1: **Major problems of sparse depth signal.** (a) The network tends to ignore the low density hint if depth from RGB is hugely different from sparse signal. (b) Imbalanced signal distribution would make the guidance to be not equally distributed. We can observe the trace of the 4 scanning lines of LiDAR (yellow circle). Our proposed S^3 method successfully overcome the two problems. The noise in the top example is removed, and the guided result in the bottom example is smoother and closer to the ground truth. Best viewed in zoomed digital.

nuScenes dataset [4] has an average of less than 50 Radar points over a 900×1600 image. Actually, the guidance module tends to ignore the accurate but sparse signals when they strongly disagree with the original prediction.

Furthermore, imbalance guidance is also the main problem. As shown in Figure 1b, the algorithms only focus on the small region with high signal density while barely correct the low-density region between scanning lines and cause non-smoothing result. However, these low-density parts neither implicate less importance nor less confidence.

In fact, there could be important objects like cars at these parts, and the imbalanced guidance stems from the uneven signal distribution of sensing devices in space. For example, LiDAR signals are mostly localized on the scanning lines with the same polar angles in the spherical coordinate, and the azimuth resolution of Radar signals is poor [10, 37]. As a result, for previous methods that conduct experiments under the assumption of uniformly distributed signal can be unreliable for real-world imbalanced cases.

To tackle the critical *low density* and *imbalanced distribution* problems, we propose a novel framework, Sparse Signal Superdensity (S^3), to enhance the density and mitigate imbalanced sparse signal for guided depth estimation. S^3 consists of two components: (1) *sparse signal expansion* (2) *confidence weighting*. For *sparse signal expansion*, S^3 first estimates the expanded area for each sparse signal based on the RGB image, and then assigns appropriate depth value to the expanded region. For *confidence weighting*, S^3 measures the confidence of the assigned depth to control the amount of influence to the sparse-guidance methods. Our method effectively utilizes *confidence weighting* to increase the density of the sparse signal.

S^3 framework, implemented with a light-weight network, can be applied to existing sparse-guidance depth estimation methods. For instance, embedding it in existing depth networks and trained in an end-to-end fashion. Losses are developed to allow S^3 network to learn *sparse signal expansion* and *confidence weighting* from data either for pre-training purposes or training jointly with depth networks. We conduct qualitative experiments to show the effectiveness of S^3 network on LiDAR and Radar guidance methods. The experimental results show that using our proposed S^3 can solve the *low density* and *imbalanced distribution* problems. Our method can highly increase the utility of the sparse signal and make substantial improvements on four typical sparse-guidance schemes on KITTI [13, 27, 41] and nuScenes [4] dataset.

To sum up, our contributions are highlighted as follows,

- The first work to point out the defective properties of the sparse signal and the subsequent influence to the depth estimation results.
- The novel and general framework Sparse Signal Superdensity (S^3) enhances the density of sparse signal, mitigates the imbalanced distribution problem, and provides extra confidence cues for depth estimation.
- S^3 largely increases the robustness and accuracy on depth estimation tasks using sparse signals, e.g., LiDAR and Radar.

2. Related Work

In this section, we will introduce guided depth estimation approaches and review related ideas about signal expansion.

Guided Mono Estimation. Previous works guide monocular depth estimation networks with external active sensors to address the technically ill-posed problem [11] and improve performance [50, 17, 24, 23, 39, 52, 41] known as Depth Completion. Cheng *et al*. [8] fuse the sparse depth as input and propagate the information to the surrounding pixels. Cadena *et al*. [3] concatenate the features of the cross-modality data to learn an auto-encoder for completing the partial or noisy depth. Ma and Karaman [24] fuse different modalities in the first convolution layer to generate high-resolution depth. The methods aim at completing the depth from sparse depth signal and an image.

Guided Stereo Estimation. Previous works guide stereo matching results with external sparse signal for better predicted results [21, 2, 30, 38, 9]. Stereo matching leverages epipolar geometry to match pixels across image pairs and produce disparity [51], which can be transformed to depth by triangulation. PSMNet [6] and GANet [48] are renowned stereo backbones. Poggi *et al*. [32] propose guided techniques on cost volume to alleviate the domain shift. Yet, their method assumes sparse signal to be uniformly distributed, which does not consider imbalanced signal problem. You *et al*. [47] propose a graph-based depth correction algorithm to refine the stereo results in 3D domain with cheap LiDAR sensors. Nonetheless, their algorithm design does not take the imbalanced signal issue into account. Wang *et al*. [43] propose input fusion and regularize batch normalization conditioning on LiDAR signal. The above methods utilize the raw sparse signal for guidance or correction, which puts little emphasis on the inherent problems of the sparse signal mentioned.

Signal Expansion. The expansion idea has shown in tasks like superpixel segmentation [1, 42, 46, 36], depth completion, and depth sampling [15, 22, 45]. Superpixel aggregates pixels with similar semantics, but they do not imply similar depth values. Depth completion and depth sampling complete the sparse depth, but most of the previous works do not measure the confidence of the expanded depth and rely on heavily computational resources.

Shivakumar *et al*. [38] propose *promotion* of the depth signal to the neighboring pixels in the cost volume to improve depth estimation. The incentive to promote the sparse signal is close to our application on cost volume. However, their methods are only applicable to Semi Global Matching [16] algorithm. Furthermore, there are lots of hand-

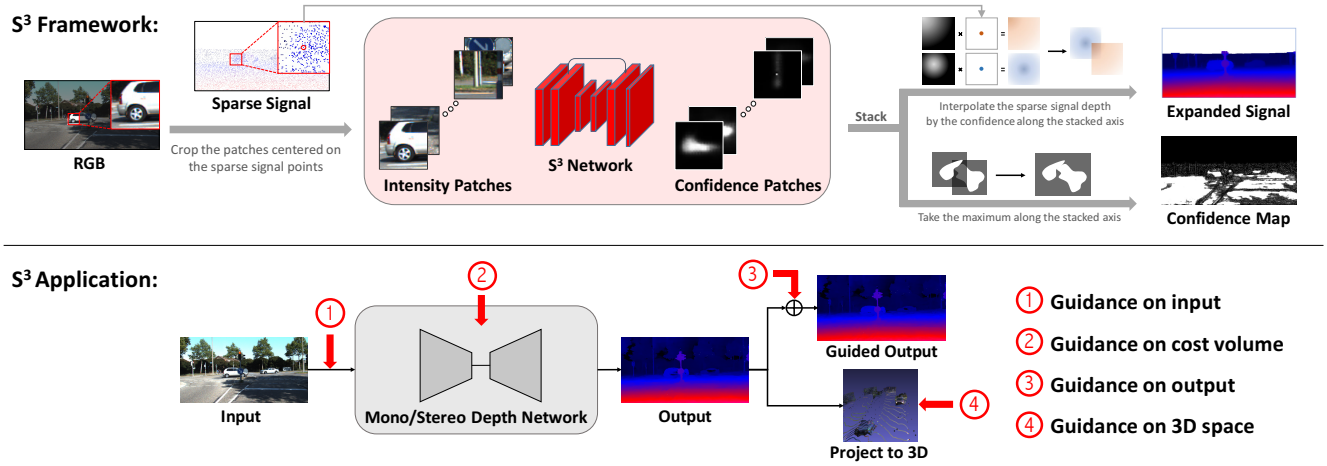


Figure 2: **Sparse Signal Superdensity (S^3) overview.** The top pipeline illustrates the details of S^3 framework to expand sparse signal and generate the final expanded depth and confidence map (Section 3.2). The bottom demonstrates the application of our module to guide on different stages of depth estimation (Section 4).

tuned hyper-parameters and assumptions, like promotion with Gaussian, which may not hold for real data.

3. Method

3.1. Intuition of Sparse Signal Superdensity

To solve the issues of *low density* and *imbalanced distribution*, we propose expanding the sparse cues to the neighbor region. Our idea is that neighboring pixels with similar color intensities belong to the same image structure or object and thus have similar depth values.

Intuitively, the ad-hoc method is to expand points by color thresholds inspired by cross-based support window method [49]. To be specific, let I, G and G_{exp} be the color intensity map, sparse signal map and expanded map. Given a central pixel (i, j) (the coordinate of the source point), we greedily expand from the central value $G(i, j)$ to its neighbor pixels (i', j') and fill in the expanded pixels $G_{exp}(i', j')$ with $G(i, j)$ as shown in Figure 3. The expansion stops until the maximum of color intensity differences is larger than a threshold or the expansion size reaches the limit.

Although the expanded map G_{exp} can substitute the sparse G to perform any guidance techniques in depth estimation, the expanded points may provide false guidance to the estimating process, especially for occlusions or pixels across object boundary. As a result, instead of applying the same level of guidance to all pixels, we provide a confidence map C to measure the reliability of the expanded value in G_{exp} and the level of guidance to apply for depth estimation.

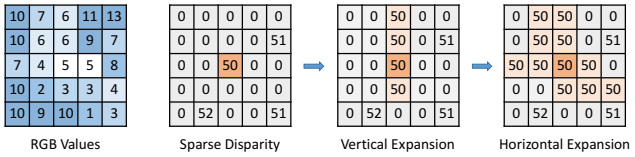


Figure 3: **Intuition for *sparse signal expansion* by ad-hoc method.** Sparse depth map (right) is expanded according to RGB image (left) presented in one channel here. Zero in the sparse map means no signal. The example expands the center signal according to difference of color intensity with threshold $= 2$.

3.2. Learnable Sparse Signal Superdensity

We propose leveraging a neural network to learn how to expand sparse signals and the corresponding confidence with the concept of *sparse signal expansion* and *confidence weighting* from Section 3.1. We expand each sparse signal to a patch by a S^3 network and aggregate all the expanded patches to form the final output.

To be specific, we predict how confident the sparse depth $G(i, j)$ can expand from the center pixel (i, j) to the neighboring pixel (i', j') with S^3 network. We set the expansion space to be a square patch of size $2L + 1$ for each sparse signal, where $|(i, j) - (i', j')| \leq L$. The input of the S^3 network is a crop of the intensity map $I(i - L : i + L, j - L : j + L)$. The output is a confidence patch of the same size and saved in $C_k(i - L : i + L, j - L : j + L) \in [0, 1]$, where k is the index of k'th sparse depth signal and $C_k = 0$ for other pixels out of the patch. Then, we aggregate the confidence patches to be the expanded depth map G_{exp} by

the following interpolation equation:

$$G_{exp}(i',j') = \frac{1}{|S_k|} \sum_{k \in S_k} C_k(i',j') \cdot G(i_k, j_k), \quad (1)$$

where (i_k, j_k) is the pixel coordinate of the k'th sparse signal and S_k is the set of indices of the sparse signal. The operation means that a pixel with no signal from depth sensors is assigned with an interpolated depth value from its nearby sparse signal values. Consequently, the more confident S^3 network considers the source signal to be, the more likely the assigned depth value is to be. Finally, we aggregate the confidence maps by taking the maximum among the confidence patches.

$$C(i',j') = \max_{k \in S_k} C_k(i',j'). \quad (2)$$

Note that $C(i',j') = 0$ if (i',j') has no expanded signal. $G_{exp}(i',j') = G(i',j')$ and $C(i',j') = 1$ if $(i',j') = (i_k, j_k)$ for a $k \in S_k$.

We formulate a general method to learn S^3 network along with any depth backbone. Here, the confidence value can act as the weights between the guided depth G_{exp} and the original estimated depth from monocular estimation or stereo matching D. That is,

$$D_{out} = G_{exp} \cdot C + D \cdot (1 - C). \quad (3)$$

With the depth ground truth D^*, the supervised loss on the output depth D_{out} can be formed as $L_{sup} = \|D^* - D_{out}\|$. We also supervise G_{exp} with D^* and add regularization

$$L_{S^3} = \lambda_1 \cdot C \cdot \|D^* - G_{exp}\| + \lambda_2 \cdot \|C\|. \quad (4)$$

The first term in Equation 4 means the more confident the expanded depth is, the more accurate it should be. The second term prevents excessive confidence for pretraining. In practice, the gradient of C of the first term is detached, otherwise, $C = 0$ can be a bad local minimum. The model is trained end-to-end so that the expansion process is learned from data. The main difference between having and not having S^3 is that S^3 increases the density of the sparse signal by providing an additional confidence map to tell the subsequent depth estimation algorithms how reliable the expanded depth is.

4. Application of S^3

S^3 network can learn to expand different modality data, including the most widely used LiDAR and Radar. Furthermore, S^3 works on both depth and disparity representation, allowing users to use our module in various applications. For instance, disparity is preferred for robotic tasks due to the need to provide higher accuracy in the nearby region [43].

Many works have proposed signal-guidance schemes to enhance depth estimated from RGB as addressed in Section 1 and 2. These methods can be divided into three categories: (1) Guidance on Input and Output (2) Guidance on Cost Volume (3) Guidance on 3D Space. We will introduce how to apply our module for each type of methods (overview in Figure 2) in the following.

4.1. Guidance on Input and Output

For guidance on input, the most intuitive way is to concatenating these external sparse signal as one of the input to the neural network. This strategy is widely used in dense depth estimation domain for either monocular [50, 23, 24] or stereo [43] depth estimation. For these approaches, we can simply replace the original raw sparse signal as our expanded signal along with the confidence map.

For guidance after the output of the depth prediction network, a naive way is to add the accurate but sparse signal to the predicted depth. Similar schemes are used by Chen *et al*. [7], called shortcut connection in the paper, and You *et al*. [47], who ignores the sparse signals largely different from stereo results to avoid numerical error and add those signals back to the corrected depth. We modify the naive method by interpolation with Equation 3 so that more pixels are guided with the expanded G_{exp} and confidence C.

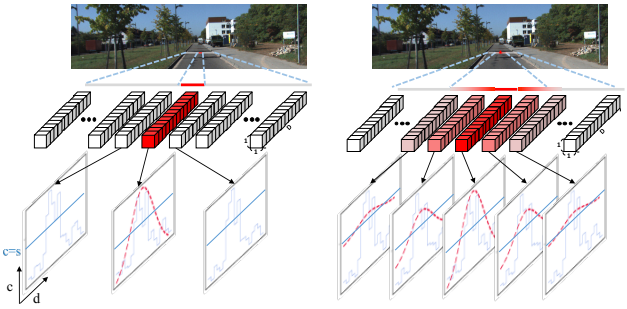
4.2. Guidance on Cost Volume

Many practices have tried to modify the cost volume, an intermediate representation of matching relationships between pixels, either guidance with external cues [32, 40, 38] or confidence measure [33] in the field of stereo matching. The cost volume in the stereo network consists of 3D features with geometric and contextual information that allows the subsequent convolution to regress the disparity probability [18, 6, 48]. Here, we take Guided Stereo Matching (GSM) [32] as an example to explain how S^3 framework is applied to cost volume. Another example, CCVNorm [43], is presented in the supplementary materials.

GSM [32] peaks the correlated features of the cost volume suggested from the sparse signal with Gaussian function to provide guidance to the network. Specifically, let $G \in \mathbb{R}^{H \times W}$ be external sparse but accurate data, V specifies a binary mask whether G has signal on pixel coordinate (i,j), and the cost volume is $CV \in \mathbb{R}^{H \times W \times D_{max} \times F}$, where D_{max} is the max disparity and F is the feature number. Given the pixel coordinate (i,j) and disparity value $G(i,j)$ from external cue G, they apply Gaussian function

$$f^{GSM}(i,j,d) = h \cdot e^{-\frac{(d-G(i,j))^2}{2w^2}} \quad (5)$$

on the features $CV(i,j,d) \leftarrow ((1 - V(i,j)) \cdot 1 + V(i,j) \cdot f^{GSM}(i,j,d)) \cdot CV(i,j,d)$ of the cost volume, where h and w are hyper-parameters to control the height and width of



(a) GSM (b) GSM + Ours

Figure 4: **Application of S^3 on cost volume.** We show the slice of the cost volume along a horizontal line. The d-axis denotes the disparity value and c-axis is the cost value. Given a guiding point (red), (a) GSM [32] guides the features of the point on cost volume. (b) We expand the disparity hint to its neighbors and guide more features with transformed Gaussian based on confidence.

the Gaussian, $\forall d \in \{0, 1, \cdots, D_{max} - 1\}$. The function f^{GSM} enlarges the feature values having positive relation to sparse cues, while suppressing others.

We propose fusing the expanded disparity map G_{exp} and the correspondent confidence map C in a novel approach:

$$f^{Ours}(i, j, d) = C \cdot \left(h \cdot e^{-\frac{(d - G_{exp}(i,j))^2}{2w^2}} \right) + s. \quad (6)$$

The shift range s preserves the minimum feature value when $(d - G_{exp}(i, j))^2$ is large or $C = 0$. When s is positive, value in cost volume $CV(i, j, d)$ will not be suppressed to zero so that the gradient of network would not be blocked during back-propagating. s can be a learnable parameter for training. The confidence value C acts as a switch to control how much guidance should be applied according to the expanded guidance G_{exp}.

The largest difference between our approach and others are learnable and confidence-based expansion, which is visualized in Figure 4. Additionally, GSM is a subset of ours. Lastly, our module is flexible to apply to other guidance-based approaches like CCVNorm [43] on cost volume.

4.3. Guidance on 3D Space

In addition to using sparse signal information on input or cost volume, performing sparse signal guidance on 3D space is an intuitive alternative. Take Graph-based Depth Correction (GDC) algorithm proposed by You *et al.* [47] as an example, the algorithm first projects the dense depth estimated from monocular or stereo network to 3D space. Then, it forms a neighborhood-relation graph considering depth value via k-nearest neighbor.

$$W = \arg\min_{W} \|Z - WZ\|_2^2, \quad (7)$$

where Z denotes the depth vector, and W denotes the edge weight between two points. Given the sparse 3D point cloud data, it then corrects the projected points with the relation graph in an optimization manner.

$$Z' = \arg\min_{Z'} \|Z' - WZ'\|^2, \quad (8)$$

where $Z'_{1:n} = G$. The first n points are set to their correct depth value from the hint of the sparse signals, and the algorithm corrects the rest of points $Z'_{n+1:}$ by minimizing the reconstruction loss. The algorithm corrects the neighbors of the sparse signal points via the relation built from W, and the neighbors of the neighbors would also be corrected. The algorithm would propagate the correct depth value via the graph relation for the sparse signals in the long run.

We improve the algorithm with the expanded depth G_{exp} and confidence C in the following approach. Suppose there are n_e expanded points and m points to be corrected, we first built the graph in Equation 7, and then minimize the reconstruction error considering the confidence.

$$Z' = \arg\min_{Z'} \|(C'G_{exp} + (I - C')Z') - W(C'G_{exp} + (I - C')Z')\|^2. \quad (9)$$

Here $C' \in \mathbb{R}^{(n+n_e+m)\times(n+n_e+m)}$ is a diagonal matrix, where $C'_{kk} = 1$ for $k \in \{1, \cdots, n\}$, $C'_{kk} = C$ for $k \in \{n+1, \cdots, n+n_e\}$, and $C'_{kk} = 0$, otherwise. The modification differs from Equation 8 in that $Z'_{n+1:n+n_e}$ is interpolated to the suggested value G_{exp} with confidence C. For C close to 0, the influence of the guidance value is negligible. For C close to 1, the guidance value is as confident as the one from sparse signal. Such modification not only allows more points to be corrected by the algorithm, but also takes the magnitude of guidance into consideration.

5. Experiment

5.1. Experimental Setting

Dataset. We use SceneFlow [26], KITTI Stereo 2012 [13], and 2015 [27] to conduct experiments for LiDAR sparse signal, and NuScenes v1.0 dataset [4] for Radar sparse signal. SceneFlow [26] dataset is a large-scale synthetic stereo dataset mainly for pretraining purpose. KITTI Stereo 2012 [13] and KITTI Stereo 2015 [27] datasets contain stereo and LiDAR data with an application to autonomous driving. Due to no dense depth ground truth provided on NuScenes, we accumulate consecutive frames of LiDAR signals (5 before and 5 after the frame of interest) for evaluation as KITTI dataset did [13].

The sparse signal for KITTI Stereo datasets is obtained according to the original paper. For Guided Stereo Matching (GSM) [32] experiments, we sub-sample 15% of pixels from the semi-dense disparity maps. For Graph-based

Depth Correction (GDC) [47] experiments, we obtain the 4-beam LiDAR signal by slicing point clouds into separate lines by an elevation step of $0.4°$.

Training Protocol. For GSM [32], we pretrain on SceneFlow, fine-tune on the training set of KITTI Stereo 2012, and test on the training set of KITTI Stereo 2015, following the protocols in the original paper. We also fine-tune on KITTI Stereo 2015, and test on KITTI Stereo 2012. For GDC [47], we use the officially released SceneFlow pretraining from PSMNet [6] and fine-tune on the training sets of KITTI Stereo 2012 and 2015, and test on 2015 and 2012, respectively. For monocular depth estimation on nuScenes dataset, the network is trained supervisedly with L1 loss on LiDAR signal and guided with two algorithms: (1) Guidance on Output in Section 4.1 (2) GDC in Section 4.3.

Implementation Detail. We implement the proposed methods with PyTorch [31] framework. The architecture of S^3 network is a light-weight version of U-Net [35] structure with patch size 32 with the last Sigmoid layer to normalize the confidence map. The number of parameters for S^3 network is 0.7M and only takes 11% of the depth network like PSMNet [6]. The inference time of the module is 0.14ms per patch for a single thread on one NVIDIA TESTLA V100 GPU with batch size 512, which can be sped up by parallelism of patch operations. S^3 network is pretrained on SceneFlow for 8000 iterations end-to-end with PSMNet [6] optimized with Adam [19] and 0.001 learning rate. Following previous works [6, 48], we randomly crop 256 by 512 for training and pad to full resolution for testing on SceneFlow and KITTI datasets. For nuScenes, we rescale input images to 288 by 512 and train sparse-to-dense [23] monocular backbone from scratch for 35k iterations. Then, the depth is guided by S^3 network pretrained from SceneFlow.

Evaluation Metric. We follow standard metrics to evaluate the results. For disparity maps, we use average pixel error (Avg) and n-pixel error rate ($> n$). The "Avg" is defined as $\frac{1}{N}\sum|D_{\text{pred}} - D_{\text{gt}}|$, where N denotes the number of pixels included in valid ground truth disparity map. The "$> n$" represents the percentage of disparity error that is greater than n. We evaluate depth maps with root mean squared (RMS) error, mean absolute relative error (REL), and δ_i. The δ_i means the percentage of the relative error within a threshold of 1.25^i. Except for δ_i, the other metrics are the smaller the better.

5.2. Guidance Experiment

5.2.1 Guidance on Input and Output.

In Table 1, even though our input guidance simply concatenating the superdensity as input, our approach can still improve upon the guided results with PSMNet. On the other hand, we contribute the huge gain of our output guidance to the density of the sparse signal, since the only difference is that more pixels are guided by expanded signal. Also, the improvement strengthens our idea that neighboring pixels of the sparse signal have similar depth and are able to be modeled with confidence by the center depth value.

Model	Avg Disp Error ↓	$> n$ Disp Error Rate (%) ↓				
		> 1	> 2	> 3	> 4	> 5
In	0.891	22.72	6.12	3.02	2.09	1.63
In + Ours	**0.851**	**21.93**	**5.98**	**2.77**	**1.78**	**1.34**
Out	0.935	26.37	8.29	3.98	2.59	1.94
Out + Ours	**0.418**	**8.90**	**1.97**	**1.05**	**0.73**	**0.55**

Table 1: **Guidance on Input (In) and Output (Out) Experiments on KITTI Stereo 2015.** (Section 5.2.1)

Dataset	Model	Avg	> 2	> 3	> 4	> 5
KITTI 2015	GANet [48]	1.949	20.72	12.43	8.78	6.73
	+ GSM	1.698	15.84	9.30	6.68	5.25
	+ GSM + Ours	**1.027**	**6.65**	**2.86**	**1.92**	**1.51**
KITTI 2015 (ft)	PSMNet [6]	1.200	6.34	3.12	2.18	1.75
	+ GSM	0.763	2.74	1.83	1.51	1.34
	+ GSM + Ours	**0.443**	**1.65**	**0.96**	**0.71**	**0.57**
KITTI 2012	GANet [48]	1.640	17.41	11.32	8.28	6.45
	+ GSM	1.370	12.26	7.90	5.92	4.74
	+ GSM + Ours	**0.836**	**4.70**	**2.27**	**1.54**	**1.18**
KITTI 2012 (ft)	PSMNet [6]	1.010	7.19	4.77	3.65	2.96
	+ GSM	0.526	2.68	1.76	1.34	1.10
	+ GSM + Ours	**0.342**	**1.37**	**0.86**	**0.65**	**0.52**

Table 2: **Experiments on GSM [32].** "ft" refers to fine-tuning on another KITTI Stereo dataset. (Section 5.2.2)

5.2.2 Guidance on Cost Volume

In Table 2, applying our method in Section 4.2 on GSM can boost a large gap of performance. In the visualization results of Figure 5, GSM does not correct much depth pixel from the stereo output, but it does when applying S^3. This tells that the network tends to ignore sparse signal when the density is not high enough, which consents to the *low density* problem and our motivation of solution. Note that we use GANet [48] as the backbone for no fine-tuning cases because we fail to reproduce GSM results on PSMNet [6].

5.2.3 Guidance on 3D Space

In Table 3, the results show consistent improvement when applying our method in Section 4.3. The performance gain of GDC is smaller than GSM because the number of points of 4-beam LiDAR is less than the sub-sampled one from

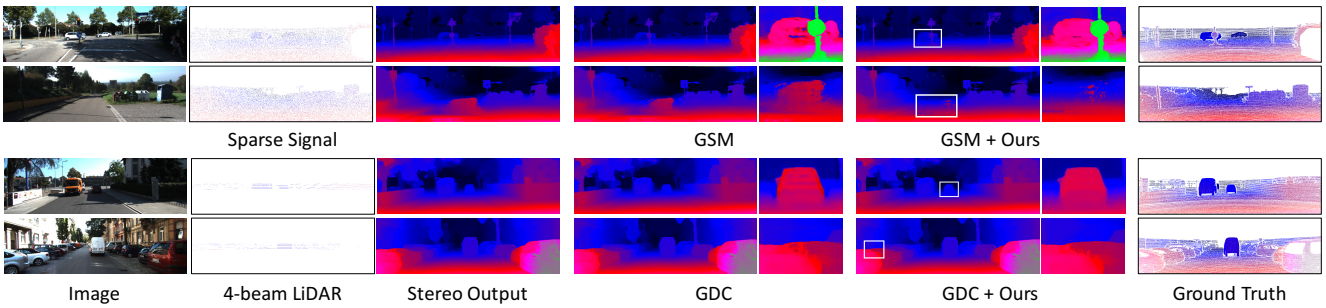


Figure 5: **Visualization on KITTI Stereo Datasets with Methods GSM [32] and GDC [47].** We show the original depth color map and the zoomed one (visually enhanced) to compare results with (5th column) and without (4th column) our method, which is best viewed in zoomed digital and color. The first row shows that our S^3 can fix the unreliable matches on the distant cars which is the low density region. The second row demonstrates that the noise from domain shift cannot completely be removed without our method. The third row illustrates that S^3 reduces the *imbalanced signal distribution* problem, which the scanning lines of LiDAR are obvious in the results of GDC [47]. The last example shows that the edge of cars are better preserved with our method.

Model	Fine-tune	KITTI Stereo 2012						KITTI Stereo 2015					
		Avg	> 1	> 2	> 3	> 4	> 5	Avg	> 1	> 2	> 3	> 4	> 5
PSMNet [6]		8.156	89.54	78.83	68.04	57.71	48.21	8.568	86.32	73.02	60.07	48.66	38.97
+ GDC		7.995	84.56	74.82	65.14	55.60	46.66	8.566	83.87	71.25	58.94	47.85	38.32
+ GDC + Ours		7.776	80.32	71.27	62.34	53.45	45.01	8.479	81.84	69.60	57.78	47.03	37.74
PSMNet [6]	✓	1.039	17.82	7.37	4.82	3.66	2.96	1.028	23.58	6.75	3.46	2.44	1.96
+ GDC	✓	0.950	15.65	6.75	4.46	3.41	2.77	0.952	21.08	6.06	3.19	2.27	1.82
+ GDC + Ours	✓	**0.904**	**14.53**	**6.31**	**4.20**	**3.22**	**2.62**	**0.915**	**20.07**	**5.76**	**3.05**	**2.17**	**1.75**

Table 3: **Experiments on GDC Algorithm Proposed in Pseudo-LiDAR++ [47].** (Section 5.2.3)

GSM. The visualization in the fourth row of Figure 5 illustrates the *imbalanced signal distribution* problem is reduced with our method. The results are presented in the disparity domain, since the Pseudo-LiDAR point cloud [44] originates from stereo matching. Also, we evaluate on the task of depth estimation instead of object detection because the focus of this paper is to improve depth estimation results.

5.3. Radar Guidance

We test the effectiveness of our module for Radar signal on nuScenes [4] dataset, which is one of the first datasets containing Camera, Radar, and LiDAR in diverse scenes and weather conditions. We choose guidance on output and guidance on 3D (GDC [47]) to improve the prediction of monocular depth estimation shown in Table 4. The improvement of "GDC + Ours" on LiDAR modality is significant compared to Table 3 because the LiDAR source here is 32-beam instead of 4-beam. The improvement from Radar modality is minor compared to LiDAR because the number of Radar point cloud is extremely sparse due to small elevation degree. However, with the help of S^3, the performance gain can be amplified. The experiment demonstrates the success of our proposed S^3 framework on both Radar and LiDAR sparse signals.

Guide	Modal	+Ours	Rel ↓	RMS ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
None	-		0.161	6.79	79.71	92.05	96.15
Out	Radar		0.161	6.79	79.71	92.05	96.15
Out	Radar	✓	0.161	6.77	79.80	92.10	96.17
GDC	Radar		0.161	6.79	79.71	92.06	96.15
GDC	Radar	✓	**0.160**	**6.76**	**79.96**	**92.13**	**96.17**
Out	LiDAR		0.154	6.63	80.36	92.41	96.38
Out	LiDAR	✓	0.090	4.59	89.63	95.74	98.05
GDC	LiDAR		0.150	6.62	80.60	92.42	96.37
GDC	LiDAR	✓	**0.055**	**3.64**	**95.97**	**97.87**	**98.79**

Table 4: **Experiments of Radar Signal on NuScenes [4] Dataset.** "Out" means guidance on output and GDC is graph-based depth correction [47]. We demonstrate the ability of our method to gain improvement even on extremely sparse Radar signal. (Section 5.3)

5.4. Ablation Study

Effectiveness of Each Component. We decompose our module with the expansion part and the confidence part.

Model	Avg	> 1	> 2	> 3	> 4	> 5
No Correction	1.010	16.87	7.19	4.77	3.65	2.96
+ Sparse Signal	0.526	6.45	2.68	1.76	1.34	1.10
+ Expansion	0.383	4.90	1.90	1.19	0.88	0.71
+ Confidence	**0.342**	**3.83**	**1.37**	**0.86**	**0.65**	**0.52**

Table 5: **Ablation Study of GSM [32] on KITTI 2012.** The best combination is to add both Expansion and Confidence on Sparse Signal. "No Correction" refers to the raw stereo output. (Section 5.4)

Expansion Model	Avg Error	Avg Error (Fine-tune)
No Expansion	1.370	0.526
Ad-hoc Method	1.155	0.582
SLIC [1]	1.027	0.489
Ours	**0.836**	**0.342**

Table 6: **Ablation Study of Different Expansion Methods on KITTI 2012 Applied with GSM [32].** (Section 5.4)

In Table 5, the main improvement comes from the expansion design, which realizes our arguments that expanding the sparse signal before guidance can improve. When considering the confidence of the expanded signal, S^3 network is allowed to learn the fine-grained magnitude of influence to the guidance and bring better results.

Sparsity Expansion. We discuss on how to expand the sparse signal in Table 6. Two baseline models closely related to the idea of expansion are chosen for the experiment: (1) The ad-hoc method mentioned in Section 3.1. (2) A superpixel algorithm, SLIC [1], which iteratively clusters the neighbor pixels based on color and distance. Confidence weighting is applied to the baselines by considering the inverse distance of the expanded point to the source point, i.e., expanded depth closer to the source has higher confidence.

In Table 6, performing expansion on the sparse signal is better than no expansion for no fine-tuning case. This tells that increasing the density of the external signal can help reduce the domain shift problem, where a network is initially trained on a synthetic dataset and tested on real imagery when real data is insufficient. This also meets the goal of improving the overall accuracy without retraining mentioned in GSM [32].

For fine-tuning case, simple expansion by color thresholds, like ad-hoc expansion, is worse than no expansion. This implies the stereo network can learn to leverage the sparse signal better than simple expansion techniques. Nevertheless, our proposed S^3 can jointly learn with the depth network to achieve better results.

The assumption of the confidence weighting for baseline methods may not hold all the time. The expansion of baselines can enlarge the guided field, but it would also provide false guidance to disparity discontinuous areas, where disparity changes sharply. The ablation study results demonstrate the learnable confidence weighting can avoid the ill assumption and improve performance.

Robustness. We also test the robustness of S^3 by sampling different density of the external signal in Figure 6. Surprisingly, our method with merely 0.28% of sparse data

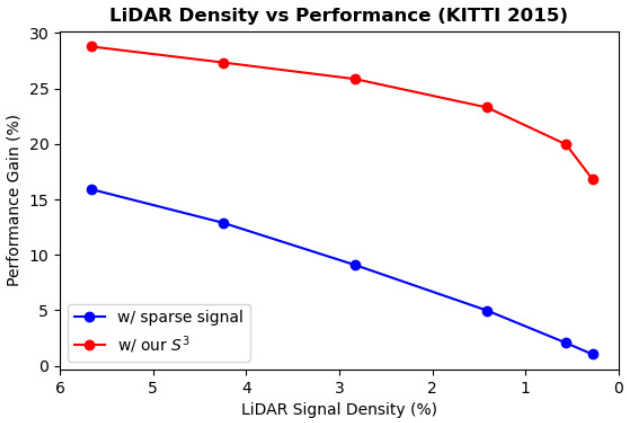


Figure 6: **Density vs Performance.** The figure stresses the robustness of S^3 for extremely low signal density.

beats GSM with 20 times denser, which strongly supports the idea to increase density of sparse data for guidance. In addition, our prediction suffers little performance drop until the external cue is extremely sparse, which emphasizes the robustness of S^3 to work under extreme environment.

6. Conclusion

In the paper, we propose S^3 framework to improve depth estimation results by considering the defective property of sparse signals. Our idea is deployable to existing sparse-guidance methods. Extensive experiments show consistent improvement among guidance approaches, and strengthen the idea that expansion on sparse signal can solve *low density* and *imbalanced distribution* problem. Our S^3 framework could become an important reference for future exploration on sparse-guidance methods.

Acknowledgement

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 110-2634-F-002-026 and FIH Mobile Limited. We benefit from NVIDIA DGX-1 AI Supercomputer and are grateful to the National Center for High-performance Computing.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 2, 8

[2] Talha Ahmad Siddiqui, Rishi Madhok, and Matthew 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*, pages 1008–1009, 2020. 1, 2

[3] Cesar Cadena, Anthony R Dick, and Ian D Reid. Multimodal auto-encoders as joint estimators for robotics scene understanding. In *Robotics: Science and Systems*, volume 5, page 1, 2016. 2

[4] 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. 1, 2, 5, 7

[5] Simon Chadwick, Will Maddetn, and Paul Newman. Distant vehicle detection using radar and vision. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8311–8317, 2019. 1

[6] 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. 2, 4, 6, 7

[7] Yun Chen, Bin Yang, Ming Liang, and Raquel Urtasun. Learning joint 2d-3d representations for depth completion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10023–10032, 2019. 4

[8] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[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. 2

[10] Liam Daniel, Andrew Stove, Edward Hoare, Dominic Phippen, Mike Cherniakov, Bernie Mulgrew, and Marina Gashinova. Application of doppler beam sharpening for azimuth refinement in prospective low-thz automotive radars. *IET Radar, Sonar & Navigation*, 12(10):1121–1130, 2018. 2

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1, 2

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2, 5

[14] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 963–968, 2011. 1

[15] Simon Hawe, Martin Kleinsteuber, and Klaus Diepold. Dense disparity maps from sparse disparity measurements. In *2011 International Conference on Computer Vision*, pages 2126–2133, 2011. 2

[16] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814, 2005. 2

[17] Yu-Kai Huang, Tsung-Han Wu, Yueh-Cheng Liu, and Winston H. Hsu. Indoor depth completion with boundary consistency and self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[18] 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. 4

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019. 1

[21] Chi Li and Zhiguo Cao. Lidar-stereo: Dense depth estimation from sparse lidar and stereo images. In *Proceedings of the 2020 5th International Conference on Multimedia Systems and Signal Processing*, pages 11–15, 2020. 2

[22] Lee-Kang Liu, Stanley H Chan, and Truong Q Nguyen. Depth reconstruction from sparse samples: Representation, algorithm, and sampling. *IEEE Transactions on Image Processing*, 24(6):1983–1996, 2015. 2

[23] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3288–3295, 2019. 2, 4, 6

[24] Fangchang Mal and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8, 2018. 1, 2, 4

[25] Suresh B Marapane and Mohan M Trivedi. Region-based stereo analysis for robotic applications. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1447–1464, 1989. 1

[26] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas 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*, pages 4040–4048, 2016. 5

[27] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 5

[28] Lazaros Nalpantidis and Antonios Gasteratos. Stereo vision for robotic applications in the presence of non-ideal lighting conditions. *Image and Vision Computing*, 28(6):940–951, 2010. 1

[29] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–7, 2019. 1

[30] Kihong Park, Seungryong Kim, and Kwanghoon Sohn. High-precision depth estimation with the 3d lidar and stereo fusion. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2156–2163, 2018. 2

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 6

[32] 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. 2, 4, 5, 6, 7, 8

[33] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Quantitative evaluation of confidence measures in a machine learning world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5228–5237, 2017. 4

[34] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019. 1

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. 6

[36] Antonio Rubio, LongLong Yu, Edgar Simo-Serra, and Francesc Moreno-Noguer. Bass: boundary-aware superpixel segmentation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2824–2829, 2016. 2

[37] Marcel Sheeny, Andrew Wallace, and Sen Wang. 300 ghz radar object recognition based on deep neural networks and transfer learning. *IET Radar, Sonar & Navigation*, 14(10):1483–1493, 2020. 2

[38] Shreyas S Shivakumar, Kartik Mohta, Bernd Pfrommer, Vijay Kumar, and Camillo J Taylor. Real time dense depth estimation by fusing stereo with sparse depth measurements. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6482–6488, 2019. 1, 2, 4

[39] Shreyas S Shivakumar, Ty Nguyen, Ian D Miller, Steven W Chen, Vijay Kumar, and Camillo 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)*, pages 13–20, 2019. 2

[40] Aristotle Spyropoulos, Nikos Komodakis, and Philippos Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1621–1628, 2014. 4

[41] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20, 2017. 2

[42] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *European conference on computer vision*, pages 13–26, 2012. 2

[43] Tsun-Hsuan Wang, Hou-Ning Hu, Chieh Hubert Lin, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min 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)*, pages 5895–5902, 2019. 1, 2, 4, 5

[44] 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. 1, 7

[45] Adam Wolff, Shachar Praisler, Ilya Tcenov, and Guy Gilboa. Super-pixel sampler: a data-driven approach for depth sampling and reconstruction. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2588–2594, 2020. 2

[46] Jian Yao, Marko Boben, Sanja Fidler, and Raquel Urtasun. Real-time coarse-to-fine topologically preserving segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2947–2955, 2015. 2

[47] 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. In *ICLR*, 2020. 1, 2, 4, 5, 6, 7

[48] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 2, 4, 6

[49] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. Cross-based local stereo matching using orthogonal integral images. *IEEE transactions on circuits and systems for video technology*, 19(7):1073–1079, 2009. 3

[50] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018. 2, 4

[51] Zhengyou Zhang. Determining the epipolar geometry and its uncertainty: A review. *International journal of computer vision*, 27(2):161–195, 1998. 2

[52] Yiqi Zhong, Cho-Ying Wu, Suya You, and Ulrich Neumann. Deep rgb-d canonical correlation analysis for sparse depth completion. In *Advances in Neural Information Processing Systems*, pages 5331–5341, 2019. 2