Supplementary Material: 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

1. S^3 Framework Tradeoff

We discuss the tradeoff between the performance and overhead by dividing the key factors into (1) patch size (2) sample rate, and (3) model size.

For (1) the patch size cropped by the center of sparse signals, doubling the size would quadruple the tensor memory and inference time, and the performance would improve but converge till the sparse cues are effective enough for a local structure.

For (2), the sample rate is the % of the sparse signals chosen for expansion by S^3, other sparse signals remain the same. Higher sample rate would cover and overlap more expanded region without extra memory but increase the computational cost linearly. We find that 25% sample rate can effectively reduce the inference time without hurting much performance.

For (3) the model size (altered by number of channels and convs), reducing the model size effectively reduces the memory usage and inference time, but suffers performance drop larger than (2) the sample rate.

Here we highlight the flexibility to apply our S^3 framework. If a user prefers real-time usage, then he or she should use a small sample rate. If a user prefers to reduce the memory usage, then he or she should use a small patch size. And if a user wants to reach state-of-the-art performance, then he or she should maximize the sample rate and model size.

2. More Guidance on Cost Volume

2.1. Guidance on Batch Normalization

Wang *et al*. [4] propose to add guidance to the batch normalization in the cost volume. They leverage Conditional Batch Normalization (CBN) operation to predict the feature-wise affine transformation in dependence on the condition of sparse LiDAR signal L^s. In particular, the CBN can be written in the following given a mini-batch of data indexed i and cost volume features $F \in \mathbb{R}^{C\times H\times W\times D}$,

$$F^{CCVNorm}_{i,c,h,w,d} = \gamma_{i,c,h,w,d}\frac{F_{i,c,h,w,d} - \mathbb{E}_\beta[F_{\cdot,c,\cdot,\cdot,\cdot}]}{\sqrt{Var_\beta[F_{\cdot,c,\cdot,\cdot,\cdot}] + \epsilon}} + \beta_{i,c,h,w,d} \tag{1}$$

$$\gamma_{i,c,h,w,d} = \begin{cases} \phi^g(d)g_c(L^s_{i,h,w}) + \psi^g(d), & \text{if } L^s_{i,h,w} \text{ is valid} \\ \overline{g}_{c,d}, & \text{otherwise} \end{cases} \tag{2}$$

$$\beta_{i,c,h,w,d} = \begin{cases} \phi^h(d)h_c(L^s_{i,h,w}) + \psi^h(d), & \text{if } L^s_{i,h,w} \text{ is valid} \\ \overline{h}_{c,d}, & \text{otherwise} \end{cases} . \tag{3}$$

The γ and β parameters are conditioned on the sparse source $L^s_{i,h,w}$ if it is valid, otherwise the parameters are reduced to the unconditional ones. g_c and h_c compute the intermediate representations of the sparse signal. ϕ and ψ modulate the final normalization parameters based on the intermediate representations. More details are presented in the original paper.

The following we demonstrate how our proposed S^3 module is applied to the conditional batch normalization. With the expanded disparity L^{exp} and confidence L^{cnf} from S^3, we improve the batch normalization process as

$$\gamma^{Ours}_{i,c,h,w,d} = \begin{cases} L^{cnf}_{i,c,h} \cdot \Big(\phi^g(d)g_c(L^{exp}_{i,h,w}) + \psi^g(d)\Big) + \\ \quad (1 - L^{cnf}_{i,c,h}) \cdot \overline{g}_{c,d}, \text{ if } L^{exp}_{i,h,w} \text{ is valid} \\ \overline{g}_{c,d}, \text{ otherwise} \end{cases} \tag{4}$$

$$\beta^{Ours}_{i,c,h,w,d} = \begin{cases} L^{cnf}_{i,c,h} \cdot \Big(\phi^h(d)h_c(L^{exp}_{i,h,w}) + \psi^h(d)\Big) + \\ \quad (1 - L^{cnf}_{i,c,h}) \cdot \overline{h}_{c,d}, \text{ if } L^{exp}_{i,h,w} \text{ is valid} \\ \overline{h}_{c,d}, \text{ otherwise} \end{cases} . \tag{5}$$

The intuitive explanation for γ^{Ours} is that we interpolate the valid value and invalid one of γ in Equation 2 by the expanded confidence L^{cnf} if L^{exp} is valid.

Method	iRMSE ↓	iMAE ↓	RMSE ↓	MAE ↓
Wang *et al*. [4]	**1.40**	0.81	0.7493	0.2525
+ Ours	1.54	**0.79**	**0.7037**	**0.2396**

Table 1: **More Results of Guidance on Cost Volume.** The experiment shows the results when applying our S^3 to the batch normalization of the cost volume.

2.2. Experiments on Batch Normalization

We follow the training protocols and implementation details in the original paper to conduct our experiments. We apply both the input and cost volume guidance with S^3 following their proposed model. In Table 1, we present the results on KITTI Depth Completion dataset [3]. We find that the performance gain is smaller than the one in Table 2 of the main paper. We contribute it to the amount of training data, where KITTI Stereo contains about 200 pairs of data while KITTI Depth Completion is hundred times larger. Ideally, it is more likely to have large performance gains for small datasets, which highlights our framework is useful when small amount of data is available in hand.

3. Details about the Confidence of S^3

3.1. Confidence Aggregation

The main paper mentions that we use *maximum* operation to aggregate confidence patches into the final confidence map in Equation 2 ($C'(i',j') = \max_{k \in S_k} C_k(i',j')$). The following we discuss why choosing the *maximum* operation. Suppose a pixel coordinate (i',j') without sparse signals ($(i',j') \neq (i_k,j_k), \forall k \in S_k$) and (i',j') is expanded by three nearby sparse signal sources with depth (d_1,d_2,d_3) and confidence (c_1,c_2,c_3), we consider two alternative aggregation operations (1) averaging the confidence and (2) interpolation with the confidence itself.

For (1) average the confidence $C(i',j') = \frac{c_1+c_2+c_3}{3}$, suppose the ground truth $D^*(i',j') = 50$, $(d_1,d_2,d_3) = (50,100,100)$, $(c_1,c_2,c_3) = (1,0.001,0.001)$. The nearby depth 100 is apparently not similar to the depth 50, so the estimated values for c_2 and c_3 are reasonable to be close to zero. Nonetheless, the values of c_2 and c_3 lower the averaged confidence to about 0.33, which does not make sense. This case particularly happens to the occlusions or object edges.

For (2) interpolation with confidence itself $C(i',j') = \frac{c_1 \cdot c_1 + c_2 \cdot c_2 + c_3 \cdot c_3}{c_1+c_2+c_3}$, suppose two cases: (a) $(d_1^a,d_2^a,d_3^a) = (50,50,50)$, $(c_1^a,c_2^a,c_3^a) = (1,0.01,0.01)$ and (b) $(d_1^b,d_2^b,d_3^b) = (50,50,50)$, $(c_1^b,c_2^b,c_3^b) = (1,0.9,0.9)$. The expectation of the final confidence for case (b) should be greater than or at least equal to the final confidence for case (a), since the expanded signals in case (b) vote for higher confidence values. However, the interpolated confidence for case (b) is about 0.94, while the one for case (a) is about 0.98, which is opposite to the expectation.

The above two counterexamples explain why neither *averaging* nor *interpolation* operations are used. Our proposed *maximum* operation can deal with the two cases to some degree. We look forward to some interesting and effective approaches to aggregate the confidence patches.

Method	% of pixel improved				Avg Error
	> 2 d	> 1 d	> 0.5 d	> 0 d	
GSM	2.4	6.2	14.6	88.3	1.370
GSM + Ours	8.2	15.2	27.5	96.9	**1.125**
GDC	0.3	0.9	2.4	14.7	0.950
GDC + Ours	1.1	2.7	5.9	21.0	**0.904**

Table 2: **Impact of Sparse Signals.** With our proposed method, the same depth correction algorithm can influence more depth pixels and achieve better performance. The "% of pixel improved $> n$ d" denotes the percentage of pixel improved for more than n disparity value owing to the depth fusion method GSM [2] and GDC [5].

3.2. Discussions on Confidence Map

An insightful comment from one of the reviewers is that the confidence maps along the stacked axis may relate to the slanted surfaces. Suppose there are three sparse depth pixels lying on the same slanted surface (e.g., road), and a neighboring pixel on the surface is interpolated by the three pixels with confidence predicted from S^3 network, the four pixels should form a slanted surface by projecting them to the 3D space with the intrinsic matrix. To this end, one could develop geometric constraints on the confidence from S^3 network via the projection matrix and the assumption that neighboring points fall on the same surface. In addition, one could leverage normal visualizations to help distinguish a good confidence prediction if the slanted assumption holds. We appreciate the idea and are open to have future discussions.

4. Impact of Sparse Signal Superdensity

Our analysis about the impact of sparse signal focuses on the following questions: (1) *How much improvement comes from sparse signal guidance?* (2) *How many more pixels are further improved due to the proposed S^3 method?* and (3) *are further improved pixels easy or hard cases?* In Table 2, relatively less pixels are largely improved by comparing the "> 2 d" and "> 0 d" columns. Furthermore, with our method, more pixels are guided and thus average pixel error is lower. Finally, our method shows about 4 times of improvement on "> 2 d", which is much larger than "> 0 d". This highlights that our S^3 can improve more on hard cases.

(a) Guided with sparse signal.

(b) Guided with expanded signal.

Figure 1: **Impact of expansion.** Applying our expanded signal of S^3 on GSM [2] can improve more depth points with the same source of sparse signal. Red points represents the pixels improved for more than 5, 2, and 0.5 disparity value from left column to right, respectively. More depth points are guided with our method by comparing the two top and bottom sub-figures. Best viewed in color.

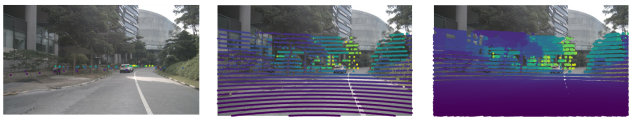

Figure 2: **Visualization of sparse signals (Radar and LiDAR) on nuScenes dataset [1].** Images from left to right are Radar, LiDAR, and the depth ground truth accumulated from 11 nearby frames. The Radar and LiDAR points (visually enhanced) are extremely sparse and imbalanced.

Here we visualize an example in Figure 1 to show the guided pixels (red) with (b) and without (a) our method. The region improved with the sparse signal is also improved with our method, since the expanded results of S^3 also contains the sparse signal. In addition, the improved and further improved region is mostly the homogeneous surface (e.g. plane road) where the stereo matching algorithm fails to find visual cues and produce accurate matches. Our method works on the homogeneous surface because the sparse signal gives the depth hint for S^3 module to estimate the slanted information about the surface.

5. More Visualization

We visualize the LiDAR and Radar signals with *low density* and *imbalanced distribution* problems in Figure 2. The elevation degree of the Radar sensor is poor so the points are mostly located at the horizontal vision line. Also, filtering operation is applied to the Radar point cloud to reduce the noise. As a result, the Radar signal is extremely sparse and imbalanced. The LiDAR signal is sparse and mostly located on the scanning lines.

References

[1] 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

[2] 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, 3

[3] 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

[4] 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

[5] 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. 2