Scene Completeness-Aware Lidar Depth Completion for Driving Scenario

Cho-Ying Wu* Ulrich Neumann*

***Abstract*— In this paper we propose Scene Completeness-Aware Depth Completion (SADC) to complete raw lidar scans into dense depth maps with fine whole scene structures. Recent sparse depth completion for lidar only focuses on the lower scenes and produce irregular estimations on the upper because existing datasets such as KITTI do not provide groundtruth for upper areas. These areas are considered less important because they are usually sky or trees and of less scene understanding interest. However, we argue that in several driving scenarios such as large trucks or cars with loads, objects could extend to upper parts of scenes, and thus depth maps with structured upper scene estimation are important for RGBD algorithms. SADC leverages stereo cameras, which have better scene completeness, and lidars, which are more precise, to perform sparse depth completion. To our knowledge, we are the first to focus on scene completeness of sparse depth completion. We validate our SADC on both depth estimate precision and scene-completeness on KITTI. Moreover, SADC only adds small extra computational cost upon base methods of stereo matching and lidar completion in terms of runtime and model size.**

I. INTRODUCTION

Autonomous driving usually adopts lidars as the main depth acquisition sensor due to its high precision and practicability on outdoor depth sensing. However, lidar scans are limited to number of scanlines and spatial resolutions, and thus they are sparse when aligned with images. Intrinsic sparsity makes it challenging for neural networks to extract effective features for various computer vision applications, such as semantic segmentation. Recently, research on lidar depth completion for autonomous driving tries to complete a sparse lidar depth map into a dense map [1]–[8] using KITTI Depth Completion Dataset [9]. However, for two reasons, their depth map processing or evaluations always crop out the upper side of maps.

First, these upper side areas are usually sky or trees of low scene understanding interest. Second, lidars are active sensors with limited scanlines and smaller vertical field-of-view than cameras. Thus, most lidar scans do not span the whole image height and are concentrated on the lower parts of images. For KITTI, topside 1/3 to 1/4 areas are unscanned by lidars. Also, KITTI's depth groundtruth is acquired by accumulating 3D point clouds with a 64-scanline lidar. Hence their groundtruth are also concentrated on lower parts of images. Both of KITTI's quantitative and qualitative evaluations focus only on the lower parts.

Nevertheless, upper scenes are especially important under several autonomous driving scenarios, such as a huge truck beside or just in front occupies a large area of the upper scene when close enough. Traffic signs or lights are important road structures extending to the upper parts. Although more and more research focuses on multi-modal learning from images and depth, *scene incompleteness issue is mostly ignored* for the following reasons. First, previous works on depth completion perform a standalone task without validating their completed depth maps on other tasks of scene understanding such as semantic segmentation. Second, not enough data of large objects extending to the upper scenes are collected, and thus the issue is generally omitted. However, autonomous driving needs to take care of all kinds of scenarios to prevent accidents and thus needs more attentions to scene completeness.

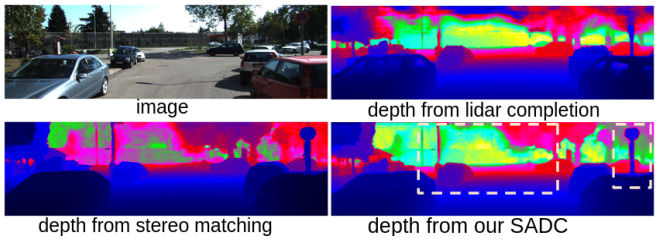

Fig. 1. **Comparison of depth from stereo matching network (PSMNet [10]), depth completion network (SSDC [4]), and our SADC.** Our results leverage advantages of both stereo cameras, which have more **structured upper scenes**, and lidars, which have **more precise depth measurements**.

In contrast, recent stereo matching algorithms by networks could produce dense depth estimation with more controlled upper scene structures. However, stereo matching is known for less reliable depth measurements for far range sensing and edge bleeding artifact [11], which produces distorted shapes.

In this work, to take advantages of both better upper scene controls from stereo matching, and more precise measurements on lower scenes from lidars, we propose to fuse depth information from these two modalities. To our knowledge, we are the first who focus on scene completeness issue of depth completion. We propose Scene-Aware Depth Completion (SADC) to fuse depth estimations from a stereo matching network and a lidar depth completion network. For sensor fusion, to analyze for each point which modality should be relied on more, we propose Attentional Point Confidence (APC) module to regress confidence maps for each modality and fuse multi-modal information. Later, we use a stacked hourglass network to refine estimations stage by stage with groundtruth. Output examples are in Fig. 1.

To further numerically analyze scene completeness, we adopt Structural similarity index (SSIM) [12], and Multi-

*Authors are with Viterbi School of Engineering, University of Southern California.

Scale SSIM (MS-SSIM)[13] as a variant to measure structural similarity for completed depth maps.

To serve real-world settings, we show several examples of completed depth from our SADC, which have much better upper scene controls than previous depth completion works. Next, different from previous works, which treat depth completion as a standalone task, we also validate our scene completeness-aware recovered depth on semantic segmentation. We use the state of the art (SOTA) of outdoor RGB-D semantic segmentation, SSMA, to show that our recovered depth could help better scene understanding.

II. Related Work

Sparse Depth Completion Recent works of sparse depth completion focus on the lidar depth completion with real-world data from KITTI Depth Completion Benchmark. [1] adopt a sparsity-invariant convolution operation to upsample depth maps. [2] stacks sparse depth maps and images to form a 4-channel input to a ResNet-based depth completion network. SSDC [4] uses ego-pose coherence as constraints and adopt a photometric loss to regress depth. CSPN [14] adopts convolutional spatial propagation to enhance local information, but their time complexity is high. Other studies, such as Deep-Lidar [3] and PwP [15] adopt an extra surface normal regression module and leverage depth-normal constraint to help the depth regression. However, these methods either crop out the upper scene of depth maps or produce random structures on the upper areas, since KITTI only provides groundtruth for lower scenes. By contrast, we adopt another depth modality from stereo matching, which contains more thorough upper scene structures than lidars to address scene completeness issue. Also, these works focus on the depth completion as an independent task, and do not provide further studies on how their completed depth maps help scene understanding in computer vision.

In the sensor fusion context, recent work Park. [16] and CCV-Norm [17] also fuse stereo cameras and lidars. The former work concentrates on how lidar information could be used to enhance disparity estimations on KITTI Stereo Evaluation Benchmark. The latter, CCV-Norm, experiments on both KITTI Stereo Evaluation and Depth Completion. However, they also crop out upper scenes in their processing. Both of their work focus on precise depth estimations of lower scenes and do not evaluate scene completeness. Neither do they further study on applicability of acquired depth or disparity maps to other scene understanding tasks.

Stereo Matching Stereo matching is a fundamental problem in computer vision. Traditional works such as SGM and variants [18]–[21] match left/right frame features and output sparse disparity estimations. Recent stereo matching methods using neural networks could estimate dense disparity maps. PSMNet [10] applies 3D convolutions to cost volumes to directly regress disparities and attain the SOTA. The estimated dense disparities have more structured upper scenes than depth from lidar completion. However, stereo matching usually suffers from edge bleeding that estimated disparities bleed out from object contours and form distorted areas [11], [22]. Further, stereo matching-based methods are unreliable for long range sensing or in areas without textures. To compensate these issues of stereo matching, our SADC leverages both scene completeness of stereo matching networks and higher precision of lidar completion network, to further produce a precise and scene completeness-aware depth from the two modalities.

RGBD Semantic Segmentation Most works on the RGBD semantic segmentation focus on indoor scenes [23]–[27]. Depth sensing at indoor is generally easier than outdoor. Indoor depth acquisition usually exploits devices with lower resolution, smaller operating ranges, and higher depth density, such as Kinect [28]. Therefore, RGBD semantic segmentation at outdoors is arguably harder than indoors. Recently, SSMA [29] is the SOTA on RGBD semantic segmentation on outdoor scene. SSMA combines two AdapNet++ [29] branches and densely fuses information from images and depth encoders with a decoder to regress the depth map. We adopt SSMA and validate completed depth from our SADC on outdoor semantic segmentation.

III. Methods

Whole network design of our SADC is in Fig. 2. Our goal is to construct a network for sensor fusion, which takes advantages of depth from stereo matching with more structured upper scene, and depth from lidar completion with higher precision, to produce a both scene completeness-aware and precise depth map.

PSMNet and SSDC are adopted as our base methods for stereo matching and lidar completion respectively. We use the estimated depth maps from two modalities, D_{stereo} and D_{lidar}, as inputs to our SADC. SADC consists of two parts, multi-modal fusion and regression with a stacked hourglass network.

At the multi-modal fusion stage, we utilize early fusion strategy. Early fusion incorporates multi-modal information before an encoder stage and has the advantages of retaining finer local structures and neighborhood relationships. Opposed to early fusion, late fusion is usually adopted for multi-modal learning with modalities from different domains to capture higher-level semantics, such as fusing information of images and depth [5], [30]. Our SADC operates information fusion only in the *depth* domain, and thus early fusion of retaining local information and structures is more desirable.

We propose a novel confidence regression module, Attentional Point Confidence (APC), to estimate the pixel-level confidence of lidars, $M_{lidar} \in [0, 1]^{H \times W}$, where H and W are height and width of inputs. APC decides for each pixel which modality is *more probable* to estimate *more reliable depth*. Previous works [3], [31] also use confidence maps for RGBD fusion without direct supervisions on confidence regression. However, for stereo cameras/lidars fusion, since we have priors that depth from stereo matching is more structured on upper scenes and depth from lidar scans is more precise, using a direct supervision on confidence regression could make the network regress better confidence maps of

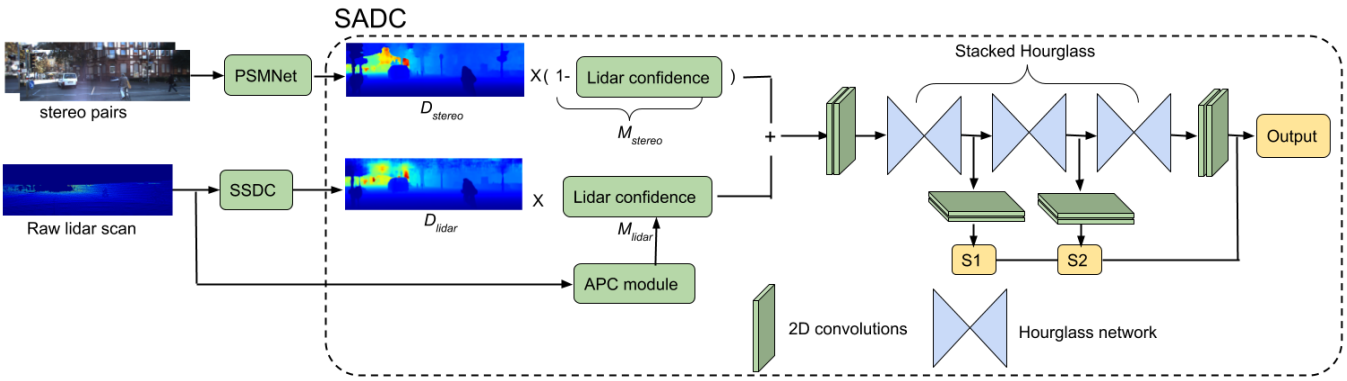

Fig. 2. **Network pipeline of our SADC.**

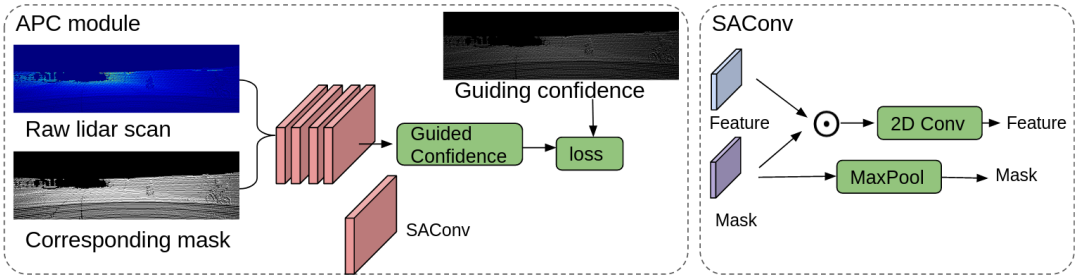

Fig. 3. **Structure of APC module and Sparisty Attentional Convolution (SAConv)** [5]. $\odot$ is for point-wise product.

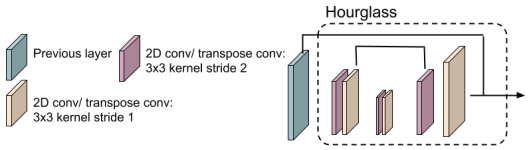

Fig. 4. **Structure of an hourglass network**.

maintaining and combining both advantages from stereo cameras and lidars.

We create a guiding confidence M_g from the raw lidar scans. Lidar measurements are comparatively precise and thus pixels at these positions in M_{lidar} should have higher confidence. We set their scores to 1. Next, depth of neighboring pixels are generally similar. We dilate the confidence map at each raw lidar measurement position of using a Gaussian distribution kernel. We choose hyperparameters of the dilation kernel, *i.e.* kernel size and variance, based on the point density of raw lidar scans. For KITTI, we find density along a scanline is 44.6% in the center and 30.6% near the left/right side. Thus, we use a 3×3 kernel and choose a variance which makes confidence scores drop to 0.5 with 1-pixel distance from the center. Example of M_g is shown at 3rd row of Fig. 9.

Sparse data are intrinsically hard for CNN to extract effective features. In APC, We utilize Sparsity-Attentional Convolution (SAConv) [5], to extract features from sparse lidar maps. SAConv attends on feature extraction of each nonzero point with an extra mask to keep track of visibility. After regressing M_{lidar}, we calculate the confidence loss as $L_c = \|M_{lidar} - M_g\|_2^2$. Structures of APC and SAConv are illustrated in Fig. 3. In APC, we use 4 layers SAConv of 3×3 kernels with stride 1. The channel size is 64 between input and confidence output.

After getting M_{lidar}, the confidence for stereo cameras is by $M_{stereo} = 1 - M_{lidar}$. Then, the fused depth is computed by

$$D_f = D_{stereo} \times M_{stereo} + D_{lidar} \times M_{lidar}. \quad (1)$$

The second stage is depth regression. We use stacked hourglass network [32] with dense connections for regressing depth stage by stage. Our stacked hourglass network consists of 3 cascaded encoder-decoder structures and has the advantage of refining depth maps stage by stage, compared with mostly used single encoder-decoder of FCN-like structure in other depth completion works [4] [2] [5] [30]. Structure of a single hourglass is illustrated in Fig. 4. The stacked hourglass produces 3 stage outputs ($S1$, $S2$, and $S3$). We further use skip connection and densely connect each corresponding layer of these hourglasses and also pass the regressed depth to every subsequent stage to enhance information flow. Finer depth is regressed at later stages. At inference time $S3$ is the final depth output. Note that ReLU [33] and batch normalization [34] are adopted after each convolution in stacked hourglass and APC. The network channel size before the stacked hourglass is 32 and increase to 64 after.

We use groundtruth, D_{gt}, to directly supervise the regression and calculate loss terms for each stage output. The corresponding mean square error losses are computed as follows.

$$L_i = \|D_{gt} - Si\|_2^2, \forall i \in [1, 3]. \quad (2)$$

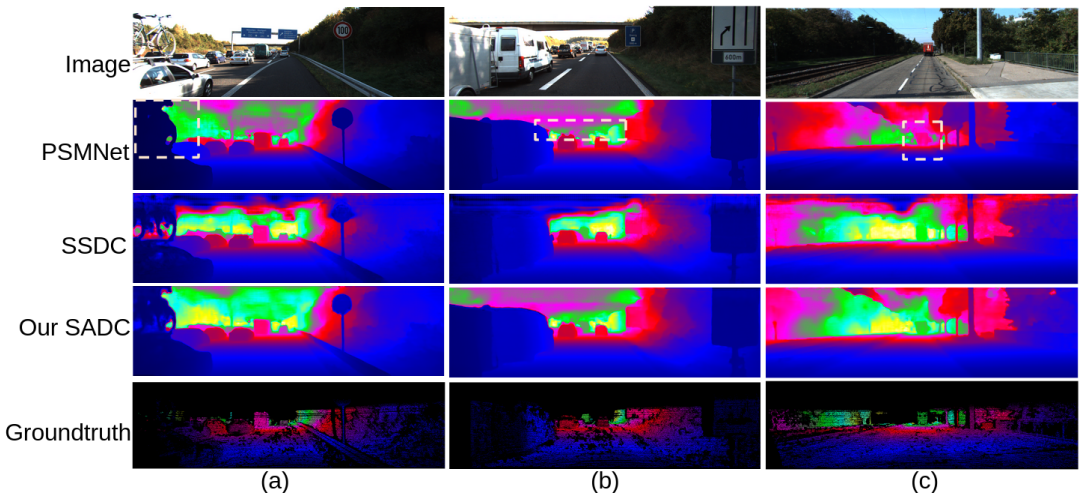

Fig. 5. **Qualitative results of PSMNet(depth from stereo matching), SSDC (depth from lidar completion), and our SADC on KITTI Depth Completion validation set.** We show driving scenarios of large trucks beside and cars with loads. Vehicle structures extend to upper scenes. SSDC fails to regress upper structures. Shape distortion of PSMNet could be seen in highlights (a) Bicycle contour. (b) Bridge structure bleeds into the background and produces irregular estimations. (c) Truck at a distance shows a distorted shape and imprecise ranges (in dark red) compared with groundtruth (in green).

The total loss is $L_1 + L_2 + L_3 + L_c$. Note that D_{gt} from KITTI Depth Completion does not contain points on the upper scenes, and thus the number of network parameter should be limited to prevent overfitting, which would cause networks to regress fine depth structures on the lower scenes but regress irregular shapes on the upper.

IV. Experiments

A. Sparse Depth Completion

Dataset. We evaluate spare depth completion on KITTI Depth Completion Benchmark. This dataset contains 42K stereo pairs and lidar scans as training data and 3.4K frames for validation. Since the image sizes differ slightly, as [4], we uniformly bottom crop the size to 352×1216. Data augmentations for training are utilized as follows. (1) Scaling by a factor $s \in [1, 1.5]$. (2) Random rotation by a degree $r \in [-5°, 5°]$. (3) Horizontal flip with probability 0.5. The validation set is used for evaluation. Inputs to our SADC are generated by PSMNet [10] and SSDC [4]. We use their released code and best pretrained weights on KITTI.

Error Metrics. We follow error metrics the same as most previous works. (1) RMSE: root mean square error; (2) Rel: mean absolute relative error; (3) δ_i: percentage of predicted pixels where the relative error is within 1.25^i. Formally,

$$\delta_i = \frac{|\{\hat{d} : max(\frac{\hat{d}}{d}, \frac{d}{\hat{d}}) < 1.25^i\}|}{|\{d\}|}, \qquad (3)$$

where $|.|$ denotes the cardinality of a set. $\hat{d}$ and d are prediction and associated groundtruth. Most studies adopt $i = 1, 2, 3$.

Structural Similarity Metrics. We propose to introduce several metrics for evaluating structural similarity of recovered depth maps. Reference is image intensity. (1) SSIM: Structural Similarity Index (2) MS-SSIM: Multiscale Structural Similarity. These metrics are widely used in image quality assessment focusing on the structural similarity to the reference. They compute local mean, standard deviation, and cross-covariance of a pair to show their local and global similarity.

TABLE I

Comparison on KITTI Depth Completion validation set

Methods	RMSE	Rel	$\delta 1$	$\delta 2$	$\delta 3$
PSMNet	2.4107	0.1296	98.6	99.8	99.9
SSDC	1.0438	**0.0191**	99.3	99.8	99.9
SADC	**1.0096**	0.0226	**99.5**	**99.9**	**100.0**

TABLE II

Structural Similarity Comparison on KITTI Depth Completion validation set

Methods	SSIM	MS-SSIM
PSMNet	0.2510	0.2513
SSDC	0.2472	0.2487
SADC	**0.2579**	**0.2590**

Results. We follow the steps in PSMNet and SSDC to predict depth maps of scenes. The quantitative comparison of depth error on KITTI Depth Completion val set is in Table I. Note that the numerical results only evaluate depth estimations on the lower scenes. Qualitative comparison is shown in Fig. 5. From both numerical and visual results, although PSMNet produces more structured upper scenes than SSDC, the depth estimation error is larger on the lower. By contrast, while SSDC has smaller numerical error, it produces irregular and unstructured depth estimations on the upper scenes. Our SADC combines the advantages of both stereo matching and depth completion and produces both scene completeness-aware and precise depth estimations.

We next quantitatively evaluate scene completeness using SSIM and MS-SSIM. We use image intensity as reference to eliminate color difference and retain spatial structures. Larger values represent higher structural similarity. The result is shown in Table II. We also visualize local SSIM index maps in Fig. 7. Our SADC has both the best SSIM and MS-SSIM. From the index maps, one can see SADC has better structures on the upper side than SSDC, and could prevent shape distortions as seen in PSMNet.

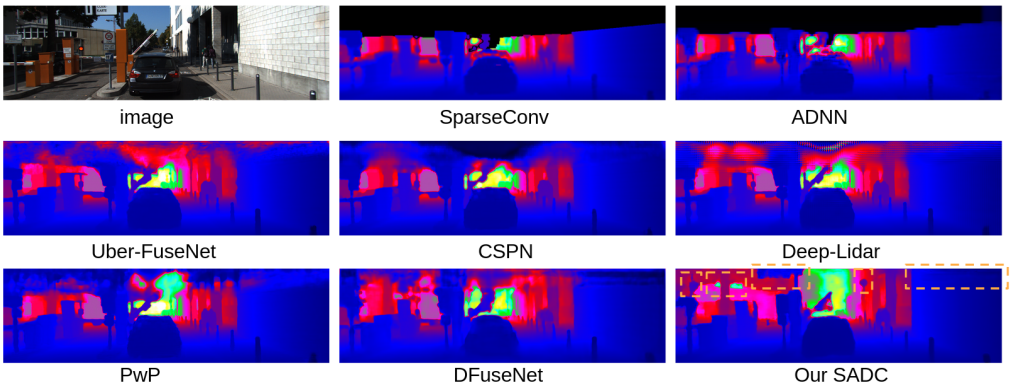

Fig. 6. **Comparison on KITTI Depth Completion *test* set.** Results of other works are directly from KITTI website. ADNN and SparseConv straightly crop out fields where no grountruth points exist and show null areas.

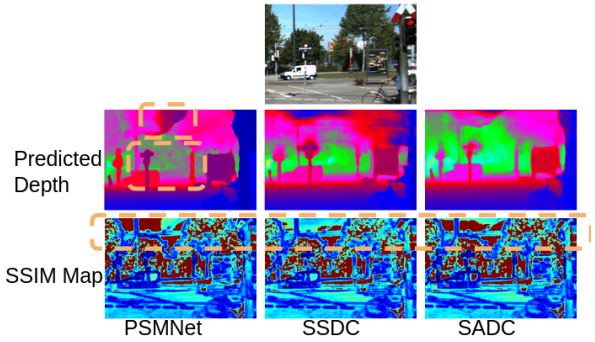

Fig. 7. **SSIM index map comparison.** We show the estimated depth and SSIM map in false color. For better visualization, we trim values larger than the global mean and show them in brown. Areas with *darker blue* represent *lower values and lower similarity*. Orange box highlights the main difference. SSDC's result is with unstructured upper scenes, and PSMNet's shows a distorted structure and imprecise long-range estimations.

We further compare with other depth completion methods on KITTI Depth Completion *test* set. However, the test set only provides images and lidar scans without stereo pairs. We find few corresponding stereo data from other KITTI benchmark datasets and compare with others using their sample results published on the KITTI website. SparseConv [1], ADNN [35], Uber-FuseNet [36], CSPN [14], Deep-Lidar [3], PwP [15], and DFuseNet [37] are included as comparisons and shown in Fig. 6.

Our SADC is the only work which successfully reconstructs the upper scene structures among the methods in comparison. Note that the lower side depth regression could be further improved if using a lidar completion network with more scene priors such as surface normals. In this work, we emphasize scene completeness of depth completion.

B. Network Structure Study

1) Confidence Map: We first analyze different confidence map generation strategies described as follows.

(1) Point Dilation: Dilate each raw lidar point with a 3x3 kernel 3 times. All reachable pixels are with confidence score 1.0 to construct the map M_{lidar} for lidar modality. $M_{stereo} = 1 - M_{lidar}$.

(2) Self-Attention: Use self-attention mechanism with a softmax layer to construct self-guided confidence map for stereo and lidar modalities. We plot this structure in Fig. 8.

(3) Guided Dilation: As described in SADC.

Samples of acquired confidence are shown in Fig. 9. The numerical comparison is shown in Table III. One could observe that direct point dilation without learning performs the worst. Self-Attention mechanism in deep learning lets a

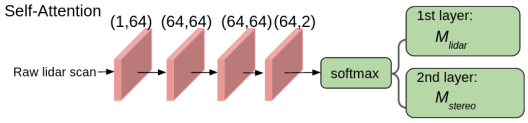

Fig. 8. **Self-Attention structure we compared with.** Parentheses denote input/output channel size.

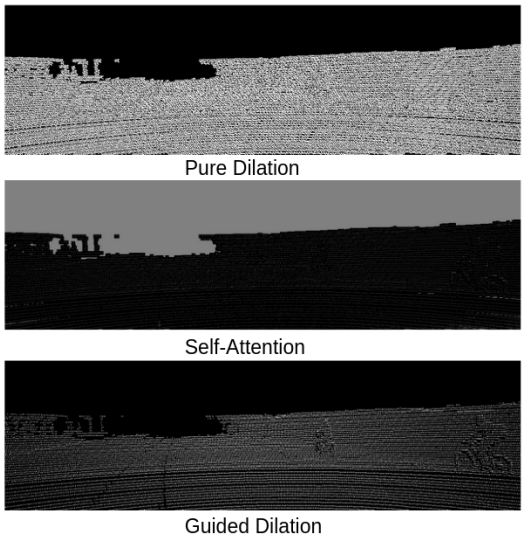

Fig. 9. **Confidence maps generated from 3 strategies we study.** Map from self-attention shows confidence of 0.5 on the upper scene since there is no raw lidar measurements, *i.e.* no cues to be more or less confident in lidar modality. Its confidence attends at positions of raw lidar scans.

TABLE III

COMPARISON OF DIFFERENT METHODS FOR GENERATING CONFIDENCE MAPS.

Methods	RMSE	Rel	$\delta 1$	$\delta 2$	$\delta 3$
Point Dilation	1.4133	0.0412	99.0	99.7	99.9
Self-Attention	1.0384	0.0254	**99.5**	99.8	99.9
Guided Dilation	**1.0096**	**0.0226**	**99.5**	**99.9**	**100.0**

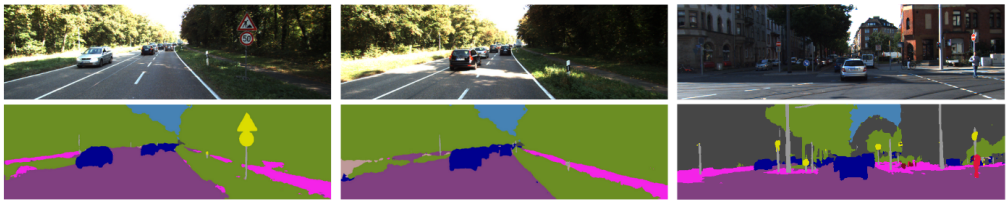

Fig. 10. **Semantic Segmentation results of SSMA with depth from our SADC on KITTI Semantic Segmentation dataset.**

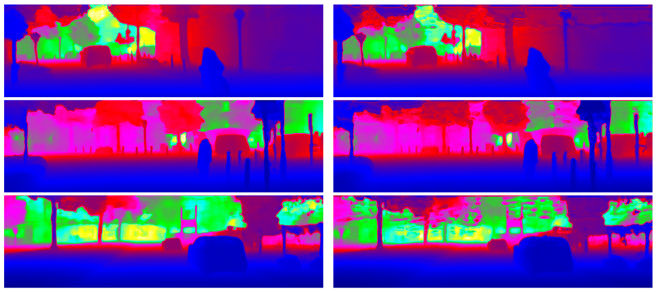

Fig. 11. **Comparison of SADC and using a deeper backbone structure.** Results of deeper backbone overfit to the lower scenes and shows unconstrained and irregular structures on the upper.

TABLE IV

NETWORK SIZE AND EFFICIENCY COMPARISON. INFERENCE RUNTIME IS REPORTED IN SECONDS PER FRAME.

Methods	parameter #	runtime(s)
PSMNet	52.4M	0.41
SSDC	26.1M	0.08
SADC	**1.2M**	**0.01**

network explore and regress confidence maps from its task's loss functions without direct supervisions. Guided Dilation in SADC, which performs self-attention to regress confidence maps and imposes a direct supervision under a Gaussian-dilated guiding map, could constrain the network to regress confidence maps with better performance.

2) Network Size and Efficiency: We also compare network size and inference runtime in Table IV. The result shows that our SADC only brings slight computational cost upon PSMNet and SSDC, but SADC combines the merits of stereo cameras and lidars.

Next, we try another network design using a deeper structure. We add one more convolution layer after each convolution operation in each hourglass, and double the channel size in the hourglass bottleneck. The number of parameter rises from 1.2M to 10.3M. We show the result comparison between SADC and the deeper structure version in Fig. 11.

C. Multi-modal Semantic Segmentation

We validate our SADC on outdoor semantic segmentation. KITTI Semantic Segmentation dataset contains 200 images without lidar information. We perform dataset registration to match data from KITTI Semantic Segmentation to KITTI Raw dataset, which contains all KITTI's publicly released data. Only 142/200 images have associated lidar scans (the rest are non-public). We separate the available data into 121 and 21 as our training and validation subset. Although Cityscapes dataset [38] has more training images on semantic segmentation, they adopt stereo cameras as the only depth acquisition device.

TABLE V

COMPARISON ON KITTI SEMANTIC SEGMENTATION DATASET.

Methods	mIoU
SDNet	51.15
SegStereo	59.12
SSMA(RGB)	54.76
SSMA(RGB+Depth from PSMNet)	61.51
SSMA(RGB+Depth from SSDC)	61.18
SSMA(RGB+Depth from SADC)	**61.57**

Most works of RGB-D semantic segmentation only validate on indoor scenes with finer depth acquisition and lower resolution of IR depth sensors such as Kinect. SSMA [29] is current SOTA work on outdoor RGB-D semantic segmentation. We follow SSMA's setting and use their Cityscapes pretrains to perform fine-tuning on KITTI. We follow most semantic segmentation works and adopt mean intersection over union (mIoU) as our metric. We also compare with two RGBD outdoor semantic segmentation methods SDNet [39] and SegStereo [40]. The quantitative and qualitative results are in Table V and Fig. 10. From them, SSMA with depth from SADC performs the best. Although the available training data are limited from KITTI, with the help of Cityscapes pretrained weights, we could still obtain visually reasonable semantic segmentation results.

V. CONCLUSIONS

Our SADC combines the advantages of scene completeness from stereo matching and higher precision from lidars to perform sparse lidar depth completion. Our APC module effectively operates sensor fusion to maintain upper scene structures from stereo matching and more precise measurements from lidars. SSIM and MS-SSIM are proposed to visually and numerically evaluate scene completeness. We show that under several scenarios, object of interest could extend to upper scenes and other lidar completion works, though have a good control on lower scenes, could not recover upper scene structures. We are the first lidar depth completion work attending on scene completeness and successfully recover the upper scene structures. We further study how lidar completion could help scene understanding. SOTA work of SSMA for outdoor semantic segmentation is adopted and numerical and visual results are shown to validate the scene completeness-aware depth maps from our SADC.

REFERENCES

[1] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger, "Sparsity invariant cnns," in *IEEE International Conference on 3D Vision (3DV)*, 2017, pp. 11–20.

[2] Fangchang Ma and Sertac Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[3] 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," *arXiv preprint arXiv:1812.00488*, 2018.

[4] 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)*. IEEE, 2019, pp. 3288–3295.

[5] 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 (NeurIPS)*, 2019, pp. 5332–5342.

[6] Byeong-Uk Lee, Hae-Gon Jeon, Sunghoon Im, and In So Kweon, "Depth completion with deep geometry and context guidance," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[7] Tsun-Hsuan Wang, Fu-En Wang, Juan-Ting Lin, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun, "Plug-and-play: Improve depth estimation via sparse data propagation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[8] Saif Imran, Yunfei Long, Xiaoming Liu, and Daniel Morris, "Depth coefficients for depth completion," in *CVPR*, 2019.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[10] Jia-Ren Chang and Yong-Sheng Chen, "Pyramid stereo matching network," in *CVPR*, 2018, pp. 5410–5418.

[11] Qiaosong Wang, Zhan Yu, Christopher Rasmussen, and Jingyi Yu, "Stereo vision–based depth of field rendering on a mobile device," *Journal of Electronic Imaging*, vol. 23, no. 2, pp. 023009, 2014.

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.

[13] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.

[14] Xinjing Cheng, Peng Wang, and Ruigang Yang, "Learning depth with convolutional spatial propagation network," *European Conference on Computer Vision (ECCV)*, 2018.

[15] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li, "Depth completion from sparse lidar data with depth-normal constraints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2811–2820.

[16] 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)*. IEEE, 2018, pp. 2156–2163.

[17] 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," *arXiv preprint arXiv:1904.02917*, 2019.

[18] Heiko Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 30, no. 2, pp. 328–341, 2007.

[19] Heiko Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, vol. 2, pp. 807–814.

[20] Stefan K Gehrig, Felix Eberli, and Thomas Meyer, "A real-time low-power stereo vision engine using semi-global matching," in *International Conference on Computer Vision Systems (ICCV)*. Springer, 2009, pp. 134–143.

[21] Simon Hermann and Reinhard Klette, "Iterative semi-global matching for robust driver assistance systems," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2012, pp. 465–478.

[22] David A Forsyth and Jean Ponce, "Computer vision: A modern approach," 2003.

[23] Weiyue Wang and Ulrich Neumann, "Depth-aware cnn for rgb-d segmentation," in *ECCV*, 2018, pp. 135–150.

[24] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *ICCV*, 2017, pp. 5199–5208.

[25] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *ACCV*, 2016, pp. 213–228.

[26] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[27] Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang, "Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 664–679.

[28] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 746–760.

[29] Abhinav Valada, Rohit Mohan, and Wolfram Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision (IJCV)*, jul 2019, Special Issue: Deep Learning for Robotic Vision.

[30] Maximilian Jaritz, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi, "Sparse and dense data with cnns: Depth completion and semantic segmentation," *IEEE International Conference on 3D Vision (3DV)*, 2018.

[31] Wouter Van Gansbeke, Davy Neven, Bert De Brabandere, and Luc Van Gool, "Sparse and noisy lidar completion with rgb guidance and uncertainty," in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–6.

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision (ECCV)*. Springer, 2016, pp. 483–499.

[33] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[34] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[35] Nathaniel Chodosh, Chaoyang Wang, and Simon Lucey, "Deep convolutional compressed sensing for lidar depth completion," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2018, pp. 499–513.

[36] 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 (ICCV)*, 2019, pp. 10023–10032.

[37] Shreyas S. Shivakumar, Ty Nguyen, Steven W. Chen, and Camillo J. Taylor, "Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion," *arXiv preprint arXiv:1902.00761*, 2019.

[38] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[39] Matthias Ochs, Adrian Kretz, and Rudolf Mester, "Sdnet: Semantically guided depth estimation network," in *German Conference on Pattern Recognition (GCPR)*. Springer, 2019, pp. 288–302.

[40] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 636–651.