NDC-Scene: Boost Monocular 3D Semantic Scene Completion in Normalized Device Coordinates Space – Supplementary Material –

Jiawei Yao[1*] Chuming Li[2,4*] Keqiang Sun[3*] Yingjie Cai[3] Hao Li[3]
Wanli Ouyang[4†] Hongsheng Li[3,4,5†]
[1] University of Washington [2] The University of Sydney
[3] CUHK-SenseTime Joint Laboratory [4] Shanghai AI Laboratory [5] CPII under InnoHK
jwyao@uw.edu, chli3951@uni.sydney.edu.au, wanli.ouyang@sydney.edu.au,
{kqsun@link, caiyingjie@link, haoli@link, hsli@ee}.cuhk.edu.hk

A. Architectures details

A.1. NDC-Scene

We follow [2] and exploit a pre-trained EfficientNetB7 [12] as the 2D image encoder.

Similar to [2, 3], we adopt DDR [7] as the basic block in the 3D branch of our dual decoder. The proposed dual decoder has four layers, each doubles the resolution and reduces the channel number by half and has a DDR block in the 3D branch and a ResNet [5] block in the 2D branch. The channel numbers of the final decoder layers for NYUv2 [11] and SemanticKITTI [1] are respectively 200 and 64.

We project all the four feature maps generated by the dual decoder to the target space, concatenate them and use a point-wise convolution to reduce the channel number to 200 and 64, respectively for NYUv2 and SemanticKITTI, which results in the input of the light-weight 3D UNet. The light-weight 3D UNet consists of two convolution layers, each with stride 2 to downscale the resolution by half and double the channel number, and two deconvolution layers, each doubles the scale and reduce the channel number by half.

Similar to [2], the final completion head consists of an ASPP module to aggregate features in multi-scales, followed by a point-wise 3D convolution to produce the classification logits.

A.2. NDC-FA

In NDC-FA, the dual-decoder is replaced with a 2D decoder, a FLoSP module and a 3D UNet. For fair comparison, the tree modules have the same structure as that in [2]. We detail the structure of NDC-FA in Fig. 1.

*These authors contribute equally to this work.
†Corresponding authors.

Method	Modality	IoU	mIoU
MonoScene [2]	2D	42.5	26.9
NDC-Scene(ours)	2D	44.2	29.0
LMSCNet [9]	2.5/3D	44.1	20.4
3DSketch [3]	2.5/3D	**71.3**	**41.1**
AICNet [6]	2.5/3D	43.8	23.8

(a) NYUv2 [11] (test set)

Method	Modality	IoU	mIoU
MonoScene [2]	2D	34.2	11.1
NDC-Scene(ours)	2D	37.2	12.7
LMSCNet [9]	2.5/3D	56.7	17.6
Local-DIFs [8]	2.5/3D	**57.7**	22.7
JS3C-Net [13]	2.5/3D	56.6	23.8
S3CNet [4]	2.5/3D	45.6	**29.5**

(b) SemanticKITTI [1] (hidden test set)

Table 1: **Quantitative comparsion** against 2.5D/3D input SSC baselines. NDC-Scene is even comparable to some 2.5/3D input methods on NYUv2 [11].

A.3. NDC-CI

In NDC-CI, the feature maps of the 3D branches in the proposed dual decoder are voxels in camera space S^R rather than S^N. Thus the voxels does not share the same 2D coordinates (x, y) with the 2D pixels, i.e., in the proposed DAA module, a 3D feature with position (x, y, d) does not has a corresponding 2D pixels (x, y) on the 2D feature map. For alignment, we perform bilinear interpolation on the 2D feature map to achieve the corresponding 2D feature on (x, y), as the input of the DAA module for the 3D feature on (x, y, d).

B. Additional results

B.1. Performance

B.1.1 Comparison against 2.5/3D-input baselines

We also compare NDC-Scene with several original SSC methods, i.e., requiring additional 3D input. Although this setting is not fair because we exploit RGB-only input, NDC-Scene still outperforms AICNet [6] and LMSCNet [9] in mIoU with an obvious gap (5.2, 8.6) and achieves comparable IoU on NYUv2 (Tab. 1a). 3DSketch [3], with

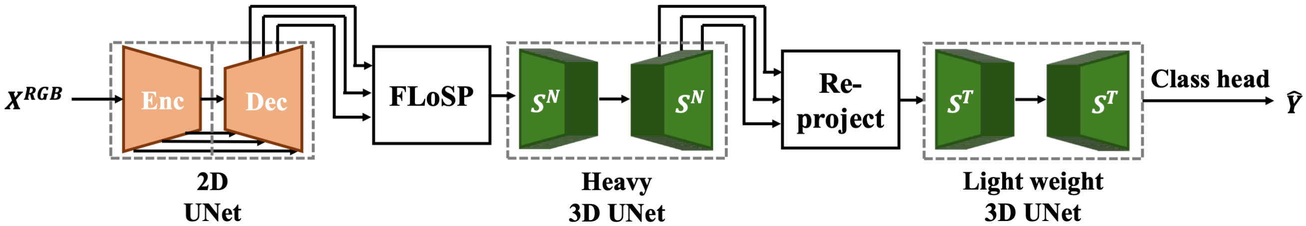

Figure 1: **NDC-FA**.

Method	SSC Input	SC IoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-veh. (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (%)	person (0.07%)	bicyclist (0.07%)	motorcyclist. (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)	SSC mIoU
LMSCNet^{rgb} [9]	Occ	31.38	46.70	19.50	13.50	3.10	10.30	14.30	0.30	0.00	0.00	0.00	10.80	0.00	10.40	0.00	0.00	0.00	5.40	0.00	0.00	7.07
3DSketch^{rgb} [3]	RGB & TSDF	26.85	37.70	19.80	0.00	0.00	12.10	17.10	0.00	0.00	0.00	0.00	12.10	0.00	16.10	0.00	0.00	0.00	3.40	0.00	0.00	6.23
AICNet^{rgb} [6]	RGB & Depth	23.93	39.30	18.30	19.80	1.60	9.60	15.30	0.70	0.00	0.00	0.00	9.60	1.90	13.50	0.00	0.00	0.00	5.00	0.10	0.00	7.09
MonoScene [2]	RGB	34.16	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10	11.08
NDC-Scene(ours)	RGB	**36.19**	**58.12**	**28.05**	**25.31**	**6.53**	**14.90**	**19.13**	**4.77**	**1.93**	**2.07**	**6.69**	**17.94**	**3.49**	**25.01**	**3.44**	**2.77**	**1.64**	**12.85**	**4.43**	**2.96**	**12.58**

Table 2: **Quantitative comparsion** against RGB-inferred baselines and the state-of-the-art monocular SSC method on SemanticKITTI [1] (hidden test set).

a TSDF-based 3D input, outperforms ours in both mIoU and IoU, implying the effectiveness of TSDF for SSC, as analyzed in [10]. On the contrary, all the baselines are clearly better than NDC-Scene in both metrics on SemanticKITTI (Tab. 1b). An important reason is that outdoor SemanticKITTI contain much more detailed and irregular objects, which relies more on the depth information accuracy to achieve a qualified surface geometry.

B.1.2 Quantitative performance on SemanticKITTI (hidden test set)

The performance of NDC-Scene compared with the RGB-inferred baselines on the hidden test set of SemanticKITTI is in Tab. 1. We still outperform all the baselines by an obvious gap of +2.03 in IoU and +1.50 in mIoU.

B.1.3 Qualitative performance

Additional qualitative results are also included in Fig. 3 (NYUv2) and Fig. 2 (SemanticKITTI). In NYUv2, compared to other SSC baselines, NDC-Scene shows a significant improvement in completing instance-level object shapes (*e.g*. furniture, row 6; sofa and objects, row 2) and semantics (*e.g*. table, row 2; sofa, row 5). In SemanticKITTI, NDC-Scene has better performance than AICNet^{rgb} [6] and 3DSketch^{rgb} [3] and is comparable with MonoScene [2]. Our outputs reconstruct better scene layout shapes (*e.g*. vegetation, terrain and building), which are easily noticeable in rows 1-11. It also infers thin objects more accurately, *e.g*. pole (row 10).

	Ours(SemanticKITTI) IoU ↑	Ours(SemanticKITTI) mIoU ↑	MonoScene(SemanticKITTI) IoU ↑	MonoScene(SemanticKITTI) mIoU ↑
$\theta = 0°$	**37.24**	**12.70**	**37.21**	**11.50**
$\theta = 5°$	35,87 (-1.37)	12.55 (-0.15)	33.45 (-3.76)	10.20 (-1.30)
$\theta = 10°$	33.28 (-3.96)	11.28 (-1.42)	29.53 (-7.68)	8.65 (-2.85)
$\theta = 15°$	31.25 (-5.99)	10.21 (-2.49)	26.42(-10.79)	7.89 (-3.61)

Table 3: **Ablation study** for the robustness to pose ambiguity on SemanticKITTI [11].

B.2. Ablation

For completeness, we also validate the robustness of NDC-Scene to the camera pose on SemanticKITTI in Tab. 3. Similar to that on NYUv2, the performance degradation of NDC-Scene is much slower than MonoScene [2], i.e., NDC-Scene generalize better to different choices of camera pose.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 1, 2, 3

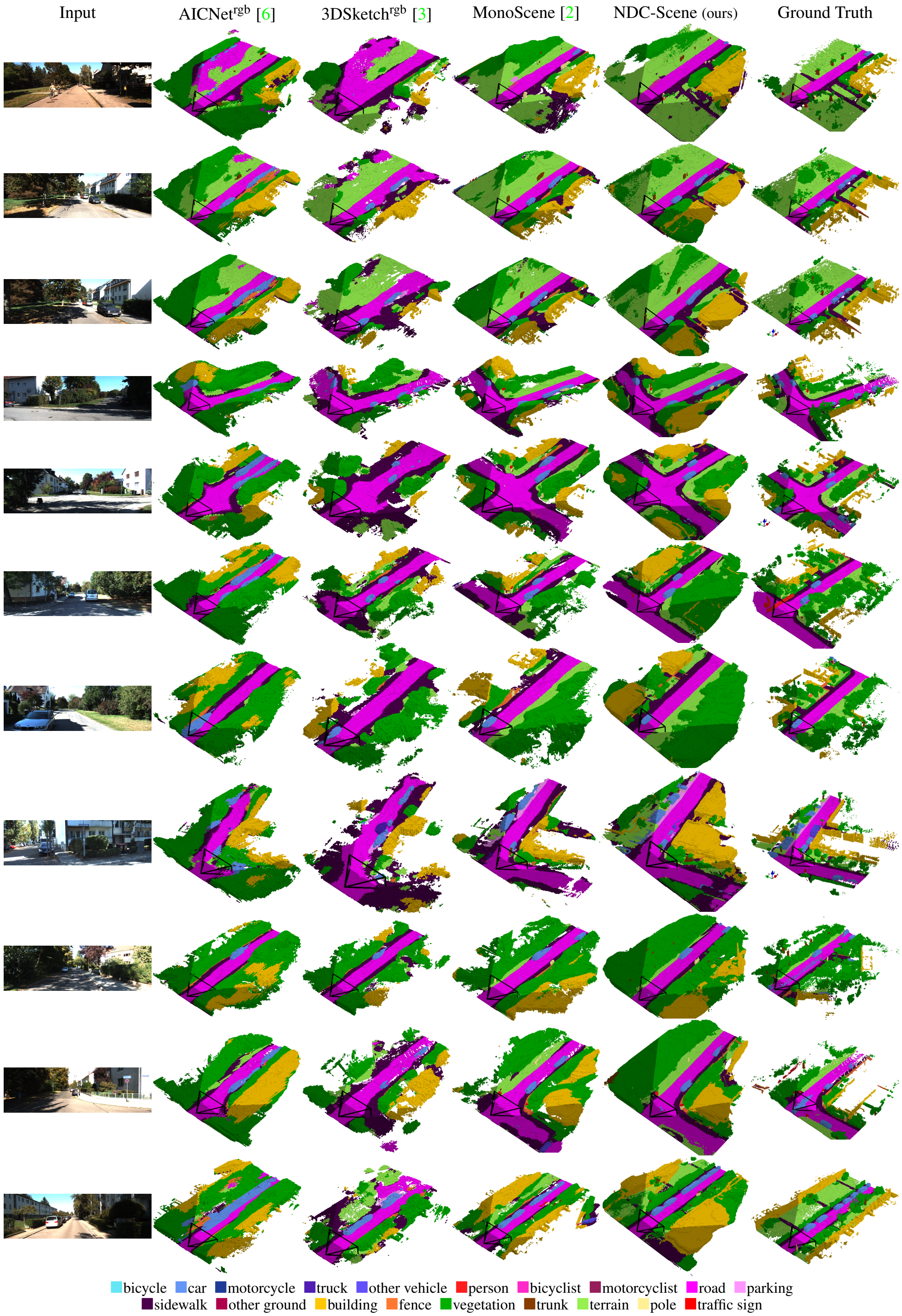

Figure 2: **Additional qualitative results on SemanticKITTI [1] (validation set).** From left to right: (a) RGB input, (b) results of AICNet^{rgb} [6] (c) results of 3DSketch^{rgb} [3] (d) results of MonoScene [2] (e) ours results. NDC-Scene achieve higher voxel-level accuracy and better semantic predictions on both datasets compared with existing SSC baselines.

Figure 3: **Additional qualitative results on NYUv2 [11].** From left to right: (a) RGB input, (b) results of AICNet^{rgb} [6] (c) results of 3DSketch^{rgb} [3] (d) results of MonoScene [2] (e) ours results. NDC-Scene achieve higher voxel-level accuracy and better semantic predictions on both datasets compared with existing SSC baselines.

[2] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2, 3, 4

[3] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4193–4202, 2020. 1, 2, 3, 4

[4] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Conference on Robot Learning*, pages 2148–2161. PMLR, 2021. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3351–3359, 2020. 1, 2, 3, 4

[7] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2019. 1

[8] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7205–7218, 2021. 1

[9] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 1, 2

[10] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d semantic scene completion: A survey. *International Journal of Computer Vision*, 130(8):1978–2005, 2022. 2

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012. 1, 2, 4

[12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[13] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3101–3109, 2021. 1