Learning Local Displacements for Point Cloud Completion
Supplementary

Yida Wang[1], David Joseph Tan[2], Nassir Navab[1], Federico Tombari[1,2]
[1]Technische Universität München [2]Google Inc.

1. Supplementary materials

As we discussed in the paper, this document aims at showing the detailed parameters of our architectures and more comprehensive results for both object completion and semantic scene completion. It also includes additional qualitative results that compares different methods against the proposed.

1.1. Parameters in architectures

This work introduces two architectures to highlight the benefits of the proposed layers. We list the parameters set in every layer of our direct architecture in Table 1 and our transformer architecture in Table 2.

1.2. Object completion

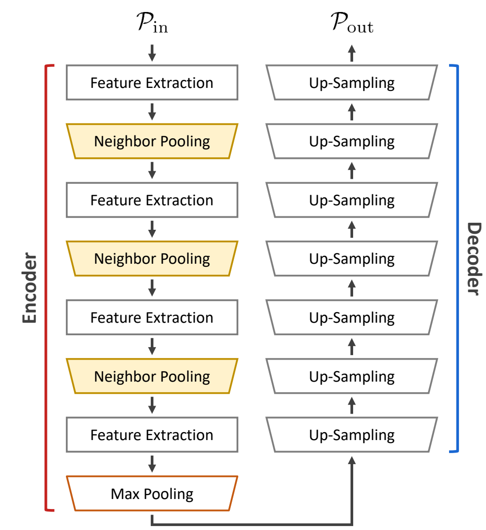
We exhibit a more detailed comparison on the object completion evaluation in Table 3, Table 4 and Table 5 for the Completion3D [14], PCN [26] and MVP [11] datasets, respectively. While we only show the average results in the paper, these tables show the per-category evaluation. Based on these results, our architectures are better in most categories when evaluating the Chamfer distance in Table 3 and Table 4; while, better in all categories when evaluating the F-Score in Table 5.

1.3. Semantic scene completion with voxels

Since most of the point cloud approaches only perform completion, we compared our semantic scene completion results to the voxel-based approaches in Table 6. In order to do this, we converted our high resolution point cloud to a lower resolution $60 \times 36 \times 60$ voxels. Table 6 shows the per-category comparison against the voxel-based approaches. Notably, although downsizing our point cloud introduces errors and difference (*e.g.* the objects in the point cloud are hollow while in the voxels are solid), we still achieve competitive IoU results.

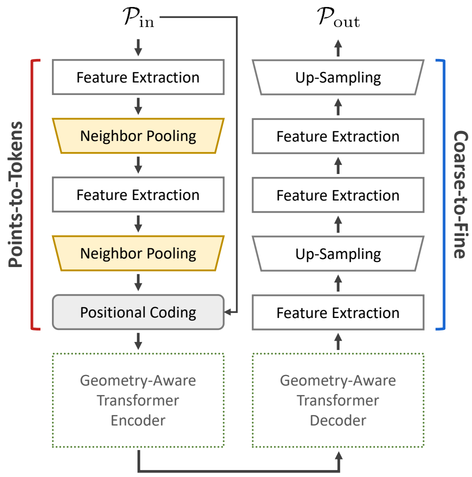
1.4. Semantic scene completion with point clouds

We illustrate the semantic scene completion results in Fig. 1, evaluated on CompleteScanNet [21]. Since there is no other point cloud completion approach that explicitly claim that they can reconstruct scenes, we utilize the architectures that were designed for object completion: PCN [26], MSN [8], PoinTr [25] and VRCNet [11]. Due to this, in Fig. 1, we perform the more complicated semantic completion while the other methods carry out the simpler



Layers	Parameters		
Feature Extraction	$s = 10$	$D_{\text{out}} = 16$	
Neighbor Pooling	$\tau = 8$		
Feature Extraction	$s = 10$	$D_{\text{out}} = 64$	
Neighbor Pooling	$\tau = 4$		
Feature Extraction	$s = 10$	$D_{\text{out}} = 64$	
Neighbor Pooling	$\tau = 4$		
Feature Extraction	$s = 10$	$D_{\text{out}} = 64$	
Max Pooling	–		
Up-Sampling	$s = 5$	$N_{\text{up}} = 2$	$D_{\text{out}} = 256$
Up-Sampling	$s = 10$	$N_{\text{up}} = 8$	$D_{\text{out}} = 64$
Up-Sampling	$s = 10$	$N_{\text{up}} = 4$	$D_{\text{out}} = 64$
Up-Sampling	$s = 10$	$N_{\text{up}} = 4$	$D_{\text{out}} = 32$
Up-Sampling	$s = 10$	$N_{\text{up}} = 8$	$D_{\text{out}} = 3$
Up-Sampling	$s = 10$	$N_{\text{up}} = 1$	$D_{\text{out}} = 3$
Up-Sampling	$s = 10$	$N_{\text{up}} = 8$	$D_{\text{out}} = 3$

Table 1. Parameters in each layer of our *direct* architecture.



Layers	Parameters		
Feature Extraction	$s = 10$	$D_{\text{out}} = 16$	
Neighbor Pooling	$\tau = 4$		
Feature Extraction	$s = 10$	$D_{\text{out}} = 64$	
Neighbor Pooling	$\tau = 4$		
Positional Coding	–		
Transformer	*Similar to [25]*		
Feature Extraction	$s = 10$	$D_{\text{out}} = 64$	
Up-Sampling	$s = 10$	$N_{\text{up}} = 8$	$D_{\text{out}} = 3$
Feature Extraction	$s = 10$	$D_{\text{out}} = 64$	
Feature Extraction	$s = 10$	$D_{\text{out}} = 64$	
Up-Sampling	$s = 10$	$N_{\text{up}} = 8$	$D_{\text{out}} = 3$

Table 2. Parameters in each layer of our *transformer* architecture.

completion task.

We observe from the other methods [8, 11, 25, 26] that their results show a high level of noise such that the objects in the scenes are no longer comprehensible. In comparison, our results have significantly less noise and produce reconstructions that are very similar to the ground truth. Moreover, a particular attention is given to PoinTr [25] since we derived our transformer architecture from them. Comparing our results against [25], our reconstructions are significantly more accurate. This in effect demonstrate the important contribution of our proposed layers to our transformer architecture.

References

[1] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2021. 5

[2] 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. 5

[3] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017. 4

[4] Andreas Geiger and Chaohui Wang. Joint 3d object and layout inference from a single rgb-d image. In *German Conference on Pattern Recognition*, pages 183–195. Springer, 2015. 5

[5] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 4

[6] Yuxiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 2018. 5

[7] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 1417–1424, 2013. 5

[8] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11596–11603, 2020. 1, 2, 4

[9] Shice Liu, YU HU, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 263–274. Curran Associates, Inc., 2018. 5

[10] Liang Pan. Ecg: Edge-aware point cloud completion with graph convolution. *IEEE Robotics and Automation Letters*, 5(3):4392–4398, 2020. 4

[11] Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu. Variational relational point completion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8524–8533, 2021. 1, 2, 3, 4

[12] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 4

[13] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene comple-

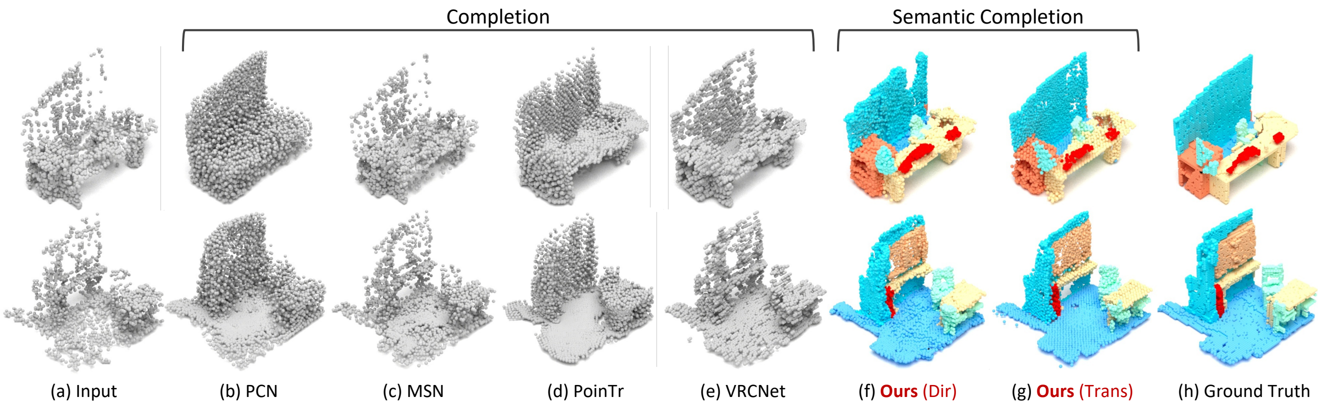


Figure 1. Semantic scene completion results on the CompleteScanNet [21] dataset

Output Resolution = 2,048, L2 metric, Completion3D [14] benchmark

Method	plane	cabinet	car	chair	lamp	sofa	table	vessel	*Avg.*
FoldingNet [24]	12.83	23.01	14.88	25.69	21.79	21.31	20.71	11.51	19.07
PointSetVoting [27]	6.88	21.18	15.78	22.54	18.78	28.39	19.96	11.16	18.18
AtlasNet [5]	10.36	23.40	13.40	24.16	20.24	20.82	17.52	11.62	17.77
PCN [26]	9.79	22.70	12.43	25.14	22.72	20.26	20.27	11.73	18.22
TopNet [14]	7.32	18.77	12.88	19.82	14.60	16.29	14.89	8.82	14.25
SA-Net [19]	5.27	14.45	7.78	13.67	13.53	14.22	11.75	8.84	11.22
SoftPoolNet [18]	6.39	17.26	8.72	13.16	10.78	14.95	11.01	6.26	11.07
GRNet [23]	6.13	16.90	8.27	12.23	10.22	14.93	10.08	5.86	10.64
PMP-Net [20]	3.99	14.70	8.55	10.21	9.27	12.43	8.51	5.77	9.23
CRN [15]	3.38	13.17	8.31	10.62	10.00	12.86	9.16	5.80	9.21
SCRN [16]	3.35	12.81	7.78	9.88	10.12	12.95	9.77	6.10	9.13
VRCNet [11]	3.94	10.93	6.44	9.32	8.32	11.35	8.60	5.78	8.12
ASFM-Net [22]	**2.38**	9.68	5.84	**7.47**	7.11	**9.65**	**6.25**	4.84	6.68
Ours (direct)	3.52	12.72	7.37	9.21	8.57	11.66	8.77	4.97	8.35
–without $\mathcal{L}_{\text{order}}$	3.64	12.83	7.48	9.34	8.70	11.79	8.88	5.07	8.47
Ours (transformer)	2.41	**9.54**	**4.99**	7.89	**6.89**	9.92	7.20	**4.29**	**6.64**
–without $\mathcal{L}_{\text{order}}$	2.48	9.62	5.10	7.99	7.01	10.04	7.29	4.39	6.74

Table 3. Evaluation on the object completion on Completion3D [14] benchmark based on the Chamfer distance trained with L2 distance (multiplied by 10^4) with the output resolution of 2,048.

tion from a single depth image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 5

[14] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 383–392, 2019. 1, 3, 4

[15] Xiaogang Wang, Marcelo H. Ang Jr. , and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 4

[16] Xiaogang Wang, Marcelo H Ang Jr, and Gim Hee Lee. A self-supervised cascaded refinement network for point cloud completion. *arXiv preprint arXiv:2010.08719*, 2020. 3, 4

[17] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8608–8617, 2019. 4, 5

[18] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Softpoolnet: Shape descriptor for point cloud completion and classification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 70–85, Cham, 2020. Springer International Publishing. 3, 4

[19] Xin Wen, Tianyang Li, Zhizhong Han, and Yu-Shen Liu. Point cloud completion by skip-attention network with hierarchical folding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,

Output Resolution = 16,384, L1 metric, PCN [26] dataset

Method	plane	cabinet	car	chair	lamp	sofa	table	vessel	*Avg.*
3D-EPN [3]	13.16	21.80	20.31	18.81	25.75	21.09	21.72	18.54	20.15
ForkNet [17]	9.08	14.22	11.65	12.18	17.24	14.22	11.51	12.66	12.85
PointNet++ [12]	10.30	14.74	12.19	15.78	17.62	16.18	11.68	13.52	14.00
FoldingNet [24]	9.49	15.80	12.61	15.55	16.41	15.97	13.65	14.99	14.31
AtlasNet [5]	6.37	11.94	10.11	12.06	12.37	12.99	10.33	10.61	10.85
TopNet [14]	7.61	13.31	10.90	13.82	14.44	14.78	11.22	11.12	12.15
PCN [26]	5.50	10.63	8.70	11.00	11.34	11.68	8.59	9.67	9.64
MSN [8]	5.60	11.96	10.78	10.62	10.71	11.90	8.70	9.49	9.97
SoftPoolNet [18]	6.93	10.91	9.78	9.56	8.59	11.22	8.51	8.14	9.20
GRNet [23]	6.45	10.37	9.45	9.41	7.96	10.51	8.44	8.04	8.83
PMP-Net [20]	5.65	11.24	9.64	9.51	**6.95**	10.83	8.72	7.25	8.73
CRN [15]	4.79	9.97	8.31	9.49	8.94	10.69	7.81	8.05	8.51
SCRN [16]	4.80	9.94	9.31	8.78	8.66	9.74	7.20	7.91	8.29
PoinTr [25]	4.75	10.47	8.68	9.39	7.75	10.93	7.78	7.29	8.38
Ours (direct)	5.34	**9.20**	**8.26**	8.96	9.40	**10.46**	7.54	8.56	8.47
–*without* $\mathcal{L}_{\text{order}}$	5.47	9.34	8.37	9.09	9.54	10.59	7.69	8.66	8.59
Ours (transformer)	**4.43**	10.03	8.28	**8.96**	7.29	10.55	**7.31**	**6.85**	**7.96**
–*without* $\mathcal{L}_{\text{order}}$	4.56	10.17	8.42	9.10	7.41	10.66	7.41	6.96	8.09

Table 4. Evaluation on the object completion on PCN [26] dataset based on the Chamfer distance trained with L1 distance (multiplied by 10^3) with the output resolution of 16,384.

Output Resolution = 16,384, F-Score@1%, MVP [11] dataset

Method	plane	cabinet	car	chair	lamp	sofa	table	vessel	*Avg.*
TopNet [14]	0.789	0.621	0.612	0.443	0.387	0.506	0.639	0.609	0.576
PCN [26]	0.816	0.614	0.686	0.517	0.455	0.552	0.646	0.628	0.614
MSN [8]	0.879	0.692	0.693	0.599	0.604	0.627	0.730	0.696	0.690
SoftPoolNet [18]	0.843	0.568	0.636	0.623	0.698	0.568	0.680	0.71	0.666
GRNet [23]	0.853	0.578	0.646	0.635	0.710	0.580	0.690	0.723	0.677
ECG [10]	0.906	0.680	0.716	0.683	0.734	0.651	0.766	0.753	0.736
NSFA [29]	0.903	0.694	0.721	0.737	0.783	0.705	0.817	0.799	0.770
CRN [15]	0.898	0.688	0.725	0.670	0.681	0.641	0.748	0.742	0.724
VRCNet [11]	0.928	0.721	0.756	0.743	0.789	0.696	0.813	0.800	0.781
PoinTr [25]	0.888	0.681	0.716	0.703	0.749	0.656	0.773	0.760	0.741
Ours (direct)	0.926	0.738	0.766	0.783	0.837	0.709	0.829	0.821	0.801
–without $\mathcal{L}_{\text{order}}$	0.910	0.750	0.741	0.734	0.835	0.715	0.839	0.783	0.788
Ours (transformer)	**0.942**	**0.753**	**0.780**	**0.799**	**0.851**	**0.725**	**0.844**	**0.836**	**0.816**
–without $\mathcal{L}_{\text{order}}$	0.922	0.731	0.759	0.776	0.831	0.703	0.824	0.813	0.795

Table 5. Evaluation on the object completion on MVP [11] dataset based on the F-Score@1% trained with L2 Chamfer distance and the output resolution of 16,384.

June 2020. 3

[20] Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. Pmp-net: Point cloud completion by learning multi-step point moving paths. *arXiv preprint arXiv:2012.03408*, 2020. 3, 4

[21] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. *arXiv preprint arXiv:2010.13662*, 2020. 1, 3

[22] Yaqi Xia, Yan Xia, Wei Li, Rui Song, Kailang Cao, and Uwe Stilla. Asfm-net: Asymmetrical siamese feature matching network for point completion. *arXiv preprint arXiv:2104.09587*, 2021. 3

[23] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao,

Method	res.	whole	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs	*Avg.*
Lin *et al.* [7]	60	36.4	0.0	11.7	13.3	14.1	9.4	29.0	24.0	6.0	7.0	16.2	1.1	12.0
Geiger and Wang [4]	60	44.4	10.2	62.5	19.1	5.8	8.5	40.6	27.7	7.0	6.0	22.6	5.9	19.6
SSCNet [13]	60	55.1	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	13.9	31.1	12.6	30.5
VVNet [6]	60	61.1	19.3	94.8	28.0	12.2	19.6	57.0	50.5	17.6	11.9	35.6	15.3	32.9
SaTNet [9]	60	60.6	17.3	92.1	28.0	16.6	19.3	57.5	53.8	17.7	18.5	38.4	18.9	34.4
ForkNet [17]	80	37.1	36.2	93.8	29.2	18.9	17.7	61.6	52.9	23.3	19.5	45.4	20.0	37.1
CCPNet [28]	240	63.5	23.5	96.3	35.7	20.2	25.8	61.4	56.1	18.1	28.1	37.8	20.1	38.5
SketchSSC [2]	60	71.3	43.1	93.6	40.5	24.3	30.0	57.1	49.3	29.2	14.3	42.5	28.6	41.1
SISNet [1]	60	**78.2**	**54.7**	93.8	**53.2**	**41.9**	**43.6**	**66.2**	**61.4**	**38.1**	**29.8**	**53.9**	**40.3**	**52.4**
Ours (direct)	60	63.7	38.1	97.1	37.0	15.5	18.7	55.2	54.9	29.6	21.4	49.2	23.7	40.0
–with $\gamma = 1$ *in* $\mathcal{L}_{\text{semantic}}$	60	58.2	35.1	94.3	34.0	12.7	15.8	52.3	52.0	26.7	18.4	46.3	20.9	37.2
Ours (transformer)	60	66.1	40.4	**98.6**	39.6	18.1	21.2	57.5	57.0	31.9	23.5	51.3	26.4	42.4
–with $\gamma = 1$ *in* $\mathcal{L}_{\text{semantic}}$	60	63.4	36.6	95.0	36.6	14.8	18.1	53.9	53.4	28.8	20.1	47.8	22.5	38.9

Table 6. Semantic completion on NYU dataset. The value in res. (x) is the output volumetric resolution which is $x \times 0.6x \times x$.

Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 365–381, Cham, 2020. Springer International Publishing. 3, 4

[24] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018. 3, 4

[25] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12498–12507, 2021. 1, 2, 4

[26] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018. 1, 2, 3, 4

[27] Junming Zhang, Weijia Chen, Yuping Wang, Ram Vasudevan, and Matthew Johnson-Roberson. Point set voting for partial point cloud analysis. *arXiv preprint arXiv:2007.04537*, 2020. 3

[28] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7801–7810, 2019. 5

[29] Wenxiao Zhang, Qingan Yan, and Chunxia Xiao. Detail preserved point cloud completion via separated feature aggregation. *arXiv preprint arXiv:2007.02374*, 2020. 4