Supplementary Material for Paper
Data Augmented 3D Semantic Scene Completion with 2D Segmentation Priors

Aloisio Dourado, Frederico Guth and Teofilo de Campos
University of Brasilia
Campus Darcy Ribeiro. Asa Norte, Brasilia, DF - 70910-900, Brazil
aloisio.dourado.bh@gmail.com fredguth@fredguth.com t.decampos@oxfordalumni.org

1. Introduction

This supplementary material contains extra information and resources that complement the paper and enable the reproduction of all our results, as listed below.

- Additional information regarding the experiments presented in the main paper.
- Complete comparison tables including all previous Semantic Scene Completion (SSC) approaches that we are aware of.
- Additional figures with images for a qualitative evaluation of the results on the three evaluated datasets.
- Source code (see `README.md` file).

2. Additional experimental details about the semantic segmentation method

To produce the results presented in the main paper, we trained two different 2D semantic segmentation networks. The results of the *"depth + RGB"* models presented in the ablation study were produced using a simplified version of RefineNet [8] (single-mode) which architecture is shown in Figure 6. The results of the *"depth + RGB + surface normals"* models presented in the ablation study and in the result tables were produced using a bimodal 2D segmentation network. Its architecture was presented in the paper. For convenience, we show it here again in Figure 7.

The main difference between these two networks is the addition of surface normals as a second input mode and a corresponding second pre-trained ResNet-101 backbone. MMF modules are also added to combine the features from the two modes.

Figure 8 presents the learning curves of the two models regarding the fine-tuning stage of training. In that stage, the ResNet backbones weights are unfrozen. Note that the bimodal model takes a little longer to stabilize compared to the single-mode one. This is somewhat expected, since

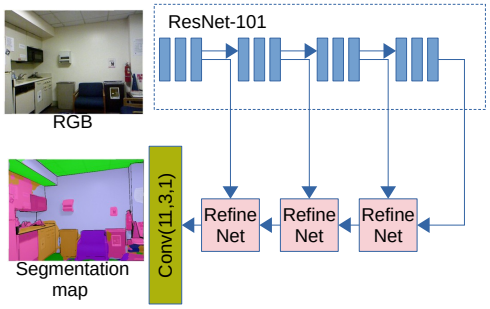

Figure 6: **2D single-mode segmentation network architecture.** This is a simplified version of RefineNet [8].

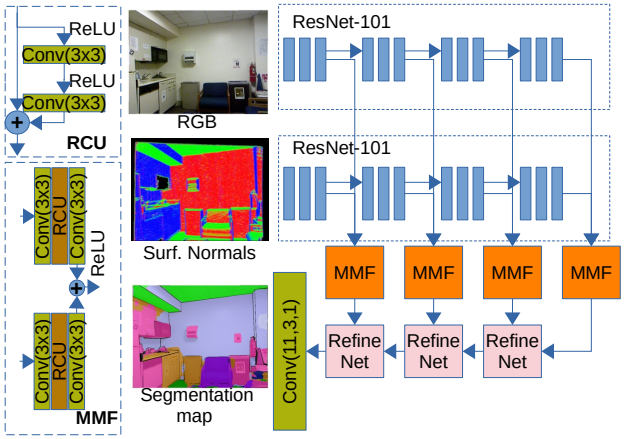

Figure 7: **2D bimodal segmentation network architecture.** This is in the main paper as Figure 3 and it is shown here again to facilitate comparison with the unimodal network of Figure 6. The Residual Convolution Unit (RCU) and the RefineNet module were first defined in [8]. Here, we use a simplified MMF block [10].

the ResNet-101 backbones are pre-trained on RGB images, and the surfaces normals represent a completely different domain. However, the bimodal model achieves a better vali-

dation final score, even though the train mIoU of the single-mode model is better. This indicates that adding surface normals as input helps reduce model overfitting.

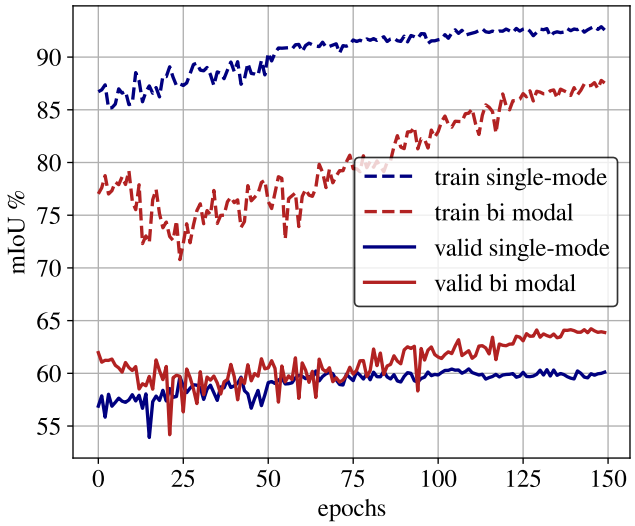

Figure 8: **Learning curves** of the fine-tune stage of 2D segmentation networks on NYDv2 (no pre-training on SUNCG).

In Table 5 we present the semantic segmentation results of the 2D models. As expected, the better learning curve of the bimodal network leads to better per class results.

3. Overfitting reduction using the proposed data augmentation approach

Figure 9 presents the training and validation learning curves of SPAwN on NYUDv2, with and without data augmentation. Note the inversion of the positions of the red and blue curves when data augmentation is used. Although the regular train curve (blue/dashed) reaches a higher score at the end of the training when compared to the data augmented curve (red/dashed), the final data augmented score in validation (red/solid) is higher than the regular curve (blue/solid). This indicates overfitting reduction due to data augmentation.

Also note that regular training starts overfitting around the 76th epoch, while the data augmented validation score keeps raising until the 118th epoch. This indicates that training can go on for more epochs to reach better results using our data augmented models.

4. Complete comparison tables

Tables 6, 7, and 8 complement Tables 2, 3 and 4 of the main paper (respectively) and compare our results to all the competing SSC solutions that we are aware of. Note the superiority of our method among all straight-forward

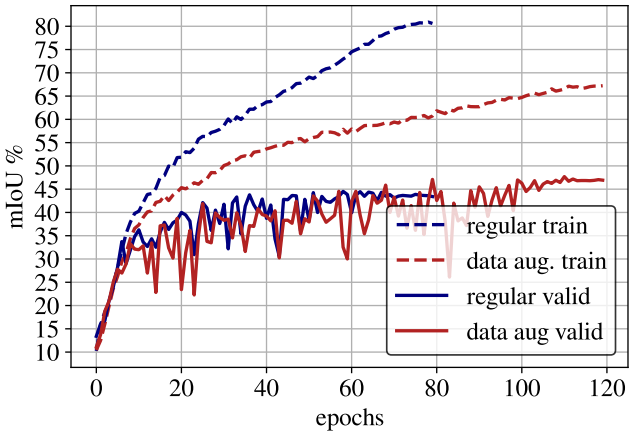

Figure 9: **Learning curves** of the training on NYUDv2 with and without data augmentation (no pre-training on SUNCG).

methods. It is worth mentioning that our data augmentation strategies can be applied with other methods, such as the SISNet, certainly leading to further improvement.

5. Additional Qualitative Analysis

Figures 10, 11, and 12 complement Figure 7 of the main paper by presenting additional results for a qualitative analysis on SUNCG, NYUDv2, and NYUCAD, respectively. Each row of the figures corresponds to one scene. From top to bottom, we present images of RGB image, depth map, surface normals, 2D predictions (obtained with our bimodal semantic segmentation network), projected visible surface, projected semantic priors, SSC predictions (obtained by our method), and 3D ground truth.

Figures 10 and 12 show that our method benefits from the excellent semantic segmentation results, guiding SPAwN to generate outstanding semantic scene completion results, filling the gaps left by simply projecting semantic priors to 3D. In SUNCG, both RGB and depth maps come from synthetic 3D models of the scenes, so the 2D segmentation results approach perfection. Segmentation results are also excellent in NYUCAD, even though the RGB images come from real scenes and there is a level of mismatch between depth maps and RGB textures (see the chairs and the bookshelf on the fourth column of Figure 12). The surface normals have certainly played an essential role in the quality of our bimodal segmentation CNN.

In Figure 11, the first column shows that specular surfaces give poor surface normals, and saturation corrupts RGB information. These specular surfaces have perturbed the segmentation priors and generated a poor 3D prediction for the whiteboard. On the other hand, the fourth column of that figure shows that our method has correctly identified the blackboard surface as “objects”, even though that region

training set	model	2D RGB-D semantic segmentation (IoU, in percentages)											
		ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
NYUDv2	Single-mode	64.1	83.8	75.7	62.5	**75.0**	62.8	58.3	38.9	52.2	57.1	54.4	60.6
	Bimodal	**73.3**	**89.2**	**76.7**	**62.9**	63.4	**67.6**	**62.1**	**40.3**	**56.7**	**58.7**	**55.5**	**64.2**
SUNCG → NYUDv2	Single-mode	-	-	-	-	-	-	-	-	-	-	-	-
	Bimodal	**74.6**	**90.6**	**77.4**	**64.7**	**64.2**	**72.0**	**62.9**	**43.8**	**54.5**	**58.7**	**56.7**	**65.5**

Table 5: **Semantic segmentation results of 2D models on NYUDv2 test set**. For each model we show per class segmentation IoU and the average score. The bimodal network is superior in average and in most of the classes. We did not test fine-tuning from SUNCG in single-mode setup.

model	pipeline type	scene completion			semantic scene completion (IoU, in percentages)											
		prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SISNet-BiSeNet [1]	iterative	93.3	96.1	89.9	85.2	90.0	83.7	80.8	60.0	83.5	80.8	68.6	77.3	86.7	70.1	78.8
SISNet-DeepLabv3 [1]		92.6	96.3	89.3	85.4	90.6	82.6	80.9	62.9	84.5	82.6	71.6	72.6	85.6	69.7	79.0
SSCNet[11]	straight-forward	76.3	95.2	73.5	96.3	84.9	56.8	28.2	21.3	56.0	52.7	33.7	10.9	44.3	25.4	46.4
TNetFuse[9]		53.9	95.2	52.6	60.6	57.3	53.2	52.7	27.4	46.8	53.3	28.6	41.1	44.1	29.0	44.9
DCRF[14]		-	-	-	95.4	84.3	57.7	24.5	28.2	63.4	55.3	34.5	19.6	45.8	28.7	48.8
ForkNet[12]		-	-	-	95.0	85.9	73.2	54.5	46.0	81.3	74.2	42.8	31.9	63.1	49.3	63.6
VVNet[6]		90.8		91.7	84.0	61.0	54.8	49.3	83.0	75.5	55.1	43.5	68.8	57.7	66.7	
EdgeNet[3]		93.3	90.6	85.1	97.2	95.3	78.2	57.5,	51.4	80.7	74.1	54.5	52.6	70.3	60.1	70.2
ESSC[13]		92.6	90.4	84.5	96.6	83.7	74.9	59.0	55.1	83.3	78.0	61.5	47.4	73.5	62.9	70.5
CCPNet[15]		**98.2**	**96.8**	**91.4**	99.2	89.3	76.2	63.3	58.2	**86.1**	**82.6**	65.6	53.2	**76.8**	65.2	74.2
SPAwN (ours)		91.9	88.7	82.3	**99.3**	**96.1**	**84.4**	**75.1**	**59.2**	81.5	78.1	**67.3**	**80.1**	76.3	**70.4**	**78.9**

Table 6: **Results on SUNCG test set**. Our SPAwN semantic scene completion overall results with regular training (not augmented) surpasses by far all known previous solutions on SUNCG synthetic images with straightforward pipeline and gets close to much more complex SISNet models (complementing Table 2 of the main paper).

was incorrectly labeled as "wall" in the ground truth.

6. Source code and models

To enable the reproduction of all our results, including rendering of resulting images for qualitative analysis, we provide all source code developed for this paper, along with pretrained model weights of the most important results.

The source code is available here: https://cic.unb.br/~teodecampos/aloisio/. Detailed instructions can be found in the `README.md` file.

model	pipeline type	train	scene compl.			semantic scene completion (IoU, in percentages)											
			prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SISNet-BiSeNet[1]	iterative	NYU	90.7	84.6	77.8	53.9	93.2	51.3	38.0	38.7	65.0	56.3	37.8	25.9	51.3	36.0	49.8
SISNet-DLabv3[1]			92.1	83.8	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
SSCNet[11]	straight-forward	NYU	57.0	**94.5**	55.1	15.1	94.7	24.4	0.0	12.6	32.1	35.0	13.0	7.8	27.1	10.1	24.7
ESSCNet[13]			71.9	71.9	56.2	17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0.0	33.4	11.8	26.7
EdgeNet[3]			76.0	68.3	65.1	17.9	94.0	27.8	2.1	9.5	51.8	44.3	9.4	3.6	32.5	12.7	27.8
DDRNet[7]			71.5	80.8	61.0	21.1	92.2	33.5	6.8	14.8	48.3	42.3	13.2	13.9	35.3	13.2	30.4
DCRF[14]			-	-	-	18.1	92.6	27.1	10.8	18.8	54.3	47.9	17.1	15.1	34.7	13.0	31.8
TS3D[4]			-	-	60.0	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
CCPNet[15]			**91.3**	92.6	**82.4**	23.5	**96.3**	35.7	20.2	25.8	61.4	56.1	18.1	28.1	37.8	20.1	38.5
SketchAware[2]			85.0	81.6	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
SPAwN			82.0	74.2	63.8	36.3	94.0	38.3	26.1	33.7	61.2	54.8	25.1	35.0	43.5	29.6	43.4
SPAwN+DA			80.8	77.8	65.7	41.5	94.2	38.0	**30.4**	40.3	69.6	57.2	29.4	**41.4**	48.8	34.1	47.7
SPAwN+DA+TTDA			82.3	77.2	66.2	41.5	94.3	38.2	30.3	**41.0**	**70.6**	**57.7**	**29.7**	40.9	**49.2**	**34.6**	**48.0**
SSCNet[11]	straight-forward	SUNCG + NYU	59.3	92.9	56.6	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	13.9	31.1	12.6	30.5
CSSCNet[5]			62.5	82.3	54.3	-	-	-	-	-	-	-	-	-	-	-	30.5
DCRF[14]			-	-	-	18.1	92.6	27.1	10.8	18.8	54.3	47.9	17.1	15.1	34.7	13.0	31.8
VVNet[6]			**86.4**	92.0	**80.3**	19.3	94.8	28.0	12.2	19.6	57.0	50.5	17.6	11.9	35.6	15.3	32.9
TNetFuse[9]			67.3	85.8	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[12]			-	-	-	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[15]			78.8	**94.3**	67.1	25.5	**98.5**	38.8	27.1	27.3	64.8	**58.4**	21.5	30.1	38.4	23.8	41.3
SPAwN			77.6	82.6	66.7	**47.3**	93.4	**41.3**	28.9	41.6	**69.5**	57.1	**33.1**	30.9	50.9	35.0	48.1
SPAwN+DA			79.8	80.8	67.1	44.1	94.0	39.9	31.5	41.6	67.4	57.3	32.5	42.8	52.5	35.0	49.0
SPAwN+DA+TTDA			81.2	80.4	67.8	44.2	94.2	40.9	**33.5**	**42.5**	69.3	**58.4**	32.4	**44.3**	**53.4**	**36.3**	**49.9**

Table 7: **Results on NYUDv2 test set** . The column "train" indicates datasets used for training the models. SUNCG + NYU means trained on SUNCG and fine-tuned on NYUDv2. In both scenarios, our SPAwN semantic scene completion overall surpasses all known previous solutions with straightforward pipeline and gets close to much more complex SISNet models (complementing Table 3 of the main paper).

model	pipeline type	train	scene compl.			semantic scene completion (IoU, in percentages)											
			prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SISNet-BiSeNet[1]	iterative	NYUCAD	94.2	91.3	86.5	65.6	94.4	67.1	45.2	57.2	75.5	66.4	50.9	31.1	62.5	42.9	59.9
SISNet-DLabv3[1]			94.1	91.2	86.3	63.4	94.4	67.2	52.4	59.2	77.9	71.1	58.1	46.2	65.8	48.8	63.5
DCRF[14]	straight-forward	NYUCAD	-	-	-	35.5	92.6	52.4	10.7	40.0	60.0	62.5	34.0	9.4	49.2	26.5	43.0
TS3D[4]			80.2	94.4	76.5	34.4	93.6	47.7	31.8	32.2	65.2	54.2	30.7	32.5	50,1	30.7	45.7
DDRNet[7]			88.7	88.5	79.4	54.1	91.5	56.4	14.9	37.0	55.7	51.0	28.8	9.2	44.1	27.8	42.8
CCPNet[15]			**91.3**	**92.6**	82.4	56.2	96.6	58.7	35.1	44.8	68.6	65.3	37.6	35.5	53.1	35.2	53.2
SketchAware[2]			90.6	92.2	**84.2**	59.7	94.3	**64.3**	32.6	51.7	72.0	68.7	45.9	19.0	60.5	38.5	55.2
SPAwN			83.7	87.2	74.6	64.0	94.6	61.4	33.3	63.1	80.4	**72.8**	47.6	**44.0**	**64.0**	42.7	60.7
SPAwN+DA			82.9	88.0	74.5	65.2	**94.7**	60.9	36.4	69.1	82.0	72.1	48.3	41.4	63.4	43.9	61.6
SPAwN+DA+TTDA			84.5	87.8	75.6	**65.3**	**94.7**	61.9	**36.9**	**69.6**	**82.2**	**72.8**	**49.1**	43.6	63.4	**44.4**	**62.2**
SSCNet[11]	straight-forward	NYUCAD + SUNCG	75.4	**96.3**	73.2	32.5	92.6	40.2	8.9	40.0	60.0	62.5	34.0	9.4	49.2	26.5	40.0
CCPNet[15]			**93.4**	91.2	**85.1**	58.1	**95.1**	60.5	36.8	47.2	69.3	67.7	39.8	37.6	55.4	37.6	55.0
SPAwN			87.7	88.4	78.7	69.9	94.9	67.6	35.0	**68.8**	**82.8**	76.0	53.2	42.4	64.0	45.8	63.7
SPAwN+DA			84.8	90.0	77.6	76.1	94.9	67.2	37.8	67.2	81.7	76.8	55.7	49.9	65.3	46.1	65.3
SPAwN+DA+TTDA			86.3	90.1	78.9	**77.6**	95.0	**68.0**	**38.1**	67.9	82.2	**77.1**	**56.8**	**50.0**	**65.7**	**46.5**	**65.9**

Table 8: **Results on NYUDCAD**. Our SPAwN models hold the best and second-best overall results on both training scenarios, when compared to previous straight-forward solutions. When fine-tuned from SUNCG, SPAwN surpasses both SISNet models, which are much more complex than ours (complementing Table 4 of the main paper).

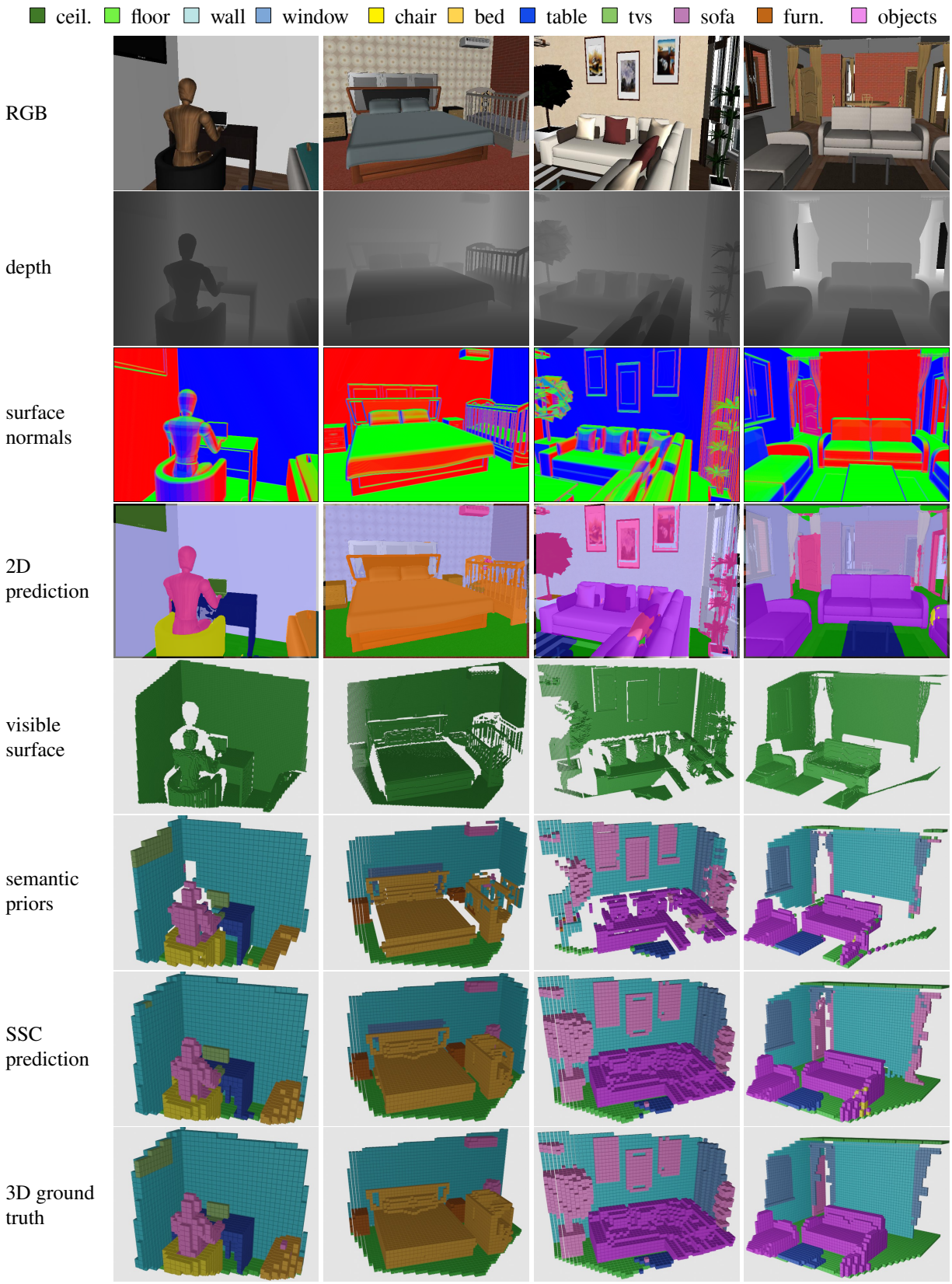

Figure 10: **Qualitative results on SUNCG.** From top to bottom: RGB, depth, surface normals, 2D segmentation with our bimodal CNN, visible surface, 3D priors, prediction with our SPAwN architecture, ground truth. (Best viewed in color.)

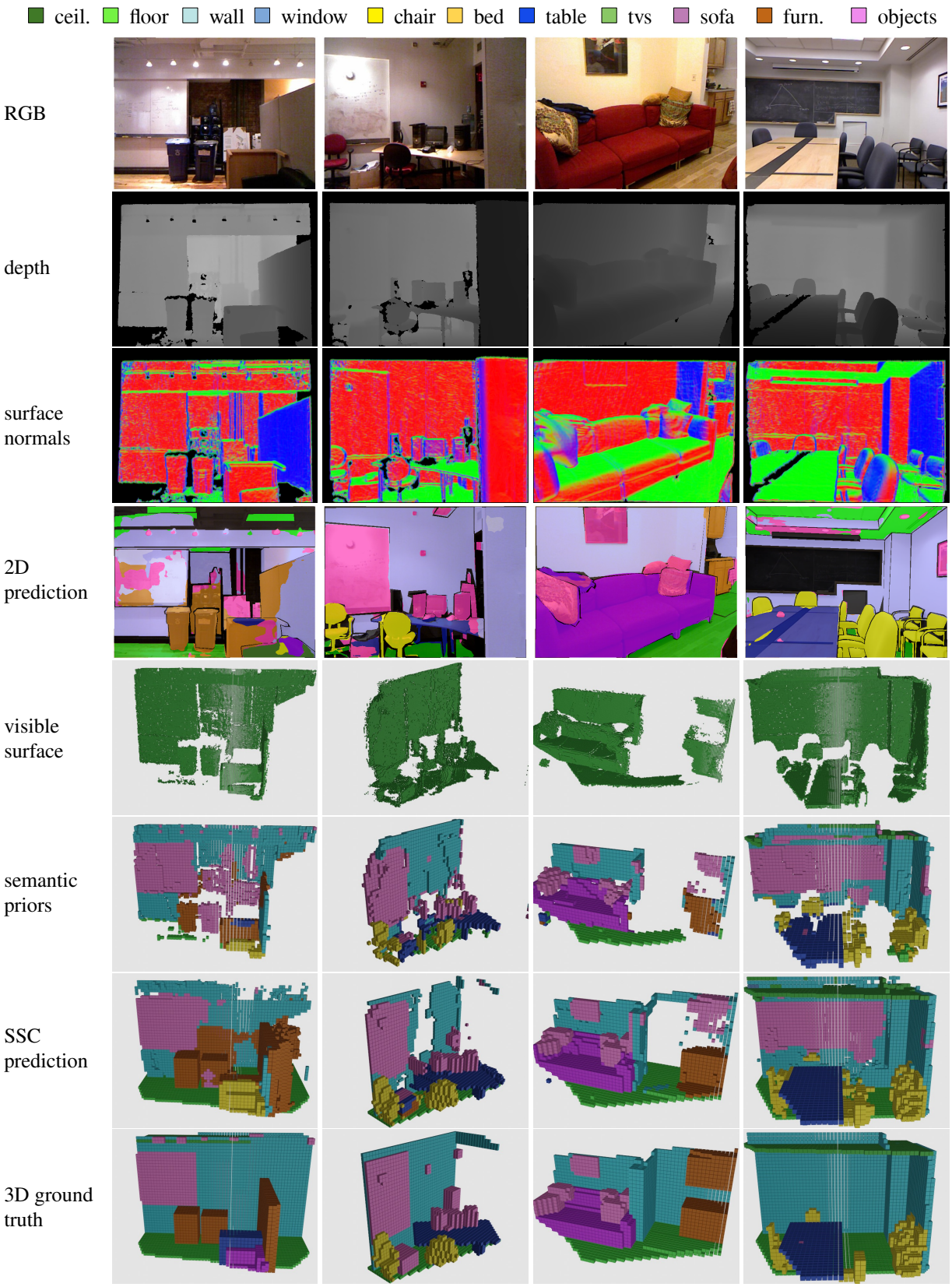

Figure 11: **Qualitative results on NYUDv2.** From top to bottom: RGB, depth, surface normals, 2D segmentation with our bimodal CNN, visible surface, 3D priors, prediction with SPAwN+S3P, Ground Truth. (Best viewed in color).

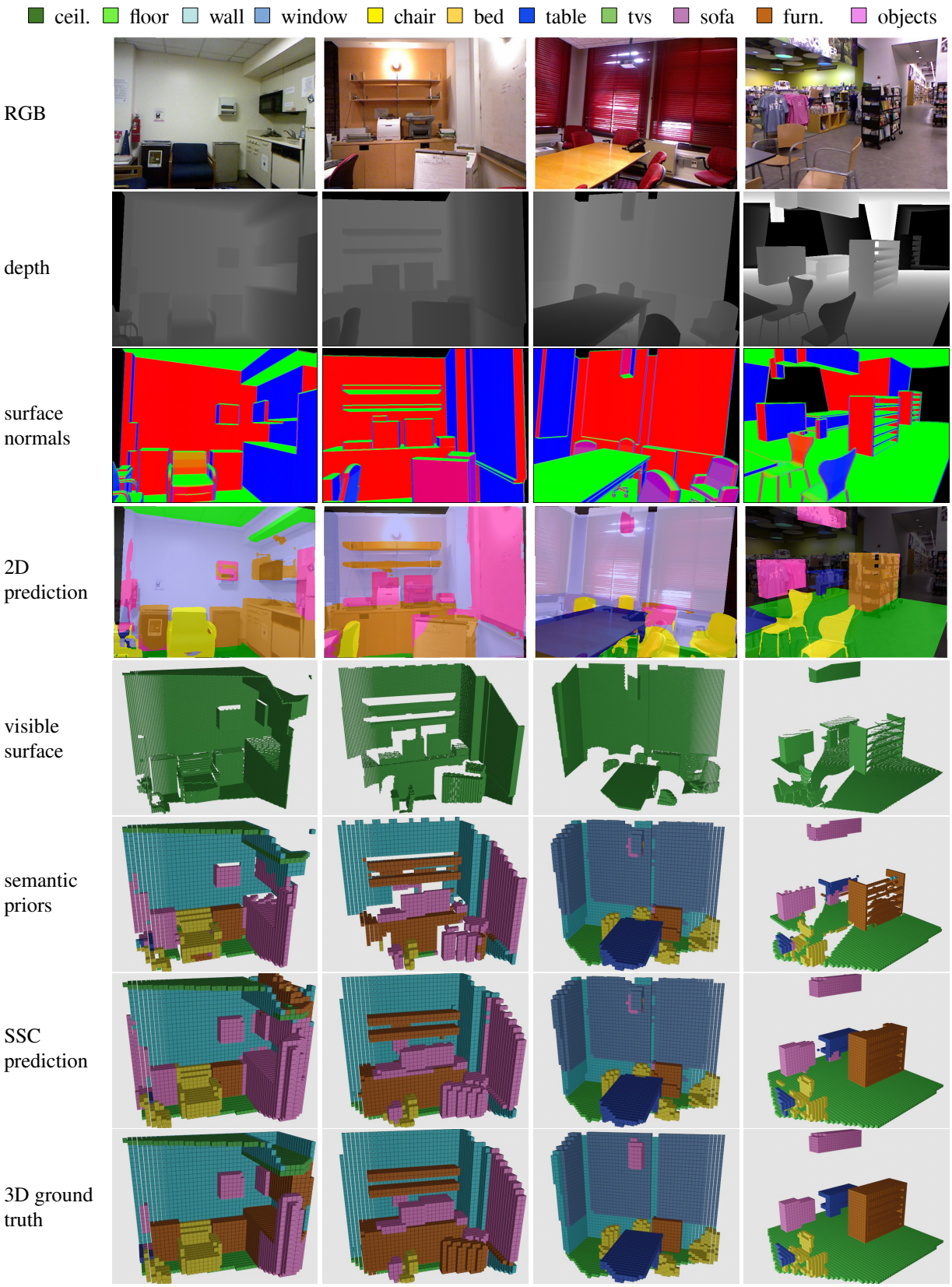

Figure 12: **Qualitative results on NYUCAD.** From top to bottom: RGB, depth, surface normals, 2D segmentation with our bimodal CNN, visible surface, 3D priors, prediction with our SPAwN+S3P, Ground Truth. (Best viewed in color.)

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 (CVPR)*, pages 324–333, June 2021. 3, 4, 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 (CVPR)*, June 2020. 4, 5

[3] Aloisio Dourado, Teofilo Emidio de Campos, Hansung Kim, and Adrian Hilton. EdgeNet: Semantic Scene Completion from a single RGB-D image. In *International Conference on Pattern Recognition (ICPR)*, 2020. 3, 4

[4] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two Stream 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 4, 5

[5] Andre Bernardes Soares Guedes, Teófilo Emídio de Campos, and Adrian Hilton. Semantic Scene Completion combining colour and depth: preliminary experiments. *CoRR arXiv*, 1802.04735, 2018. 4

[6] Yuxiao Guo and Xin Tong. View-Volume Network for Semantic Scene Completion from a Single Depth Image. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 726–732, Stockholm, Sweden, July 2018. 3, 4

[7] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. RGB-D based dimensional decomposition residual network for 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4, 5

[8] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[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, *Conference on Neural Information Processing Systems (NeurIPS)*, pages 263–274. Curran Associates, Inc., 2018. 3, 4

[10] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[11] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic Scene Completion from a Single Depth Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4, 5

[12] Y. Wang, D. J. Tan, N. Navab, and F. Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8607–8616, 2019. 3, 4

[13] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient Semantic Scene Completion network with spatial group convolution. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3, 4

[14] Liang Zhang, Le Wang, Xiangdong Zhang, Peiyi Shen, Mohammed Bennamoun, Guangming Zhu, Syed Afaq Ali Shah, and Juan Song. Semantic scene completion with dense CRF from a single depth image. *Neurocomputing*, 318:182–195, Nov. 2018. 3, 4, 5

[15] 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 (ICCV)*, October 2019. 3, 4, 5