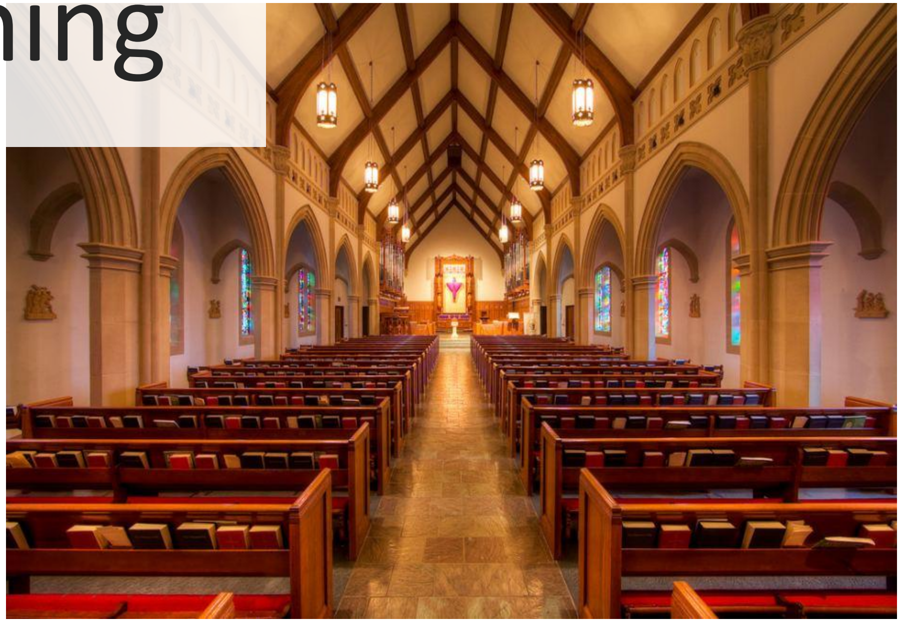
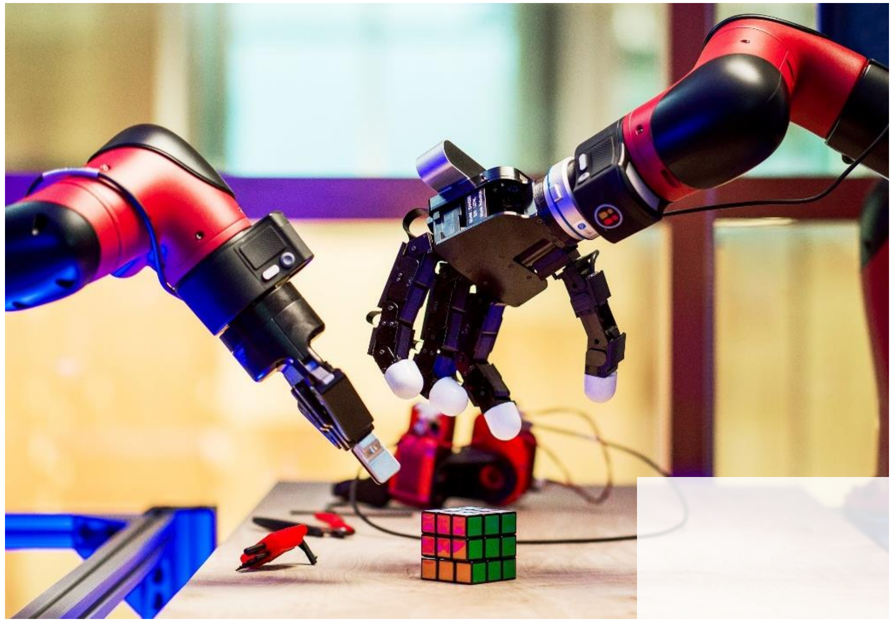
Semantic Scene Completion from a Single 360-Degree Image and Depth Map

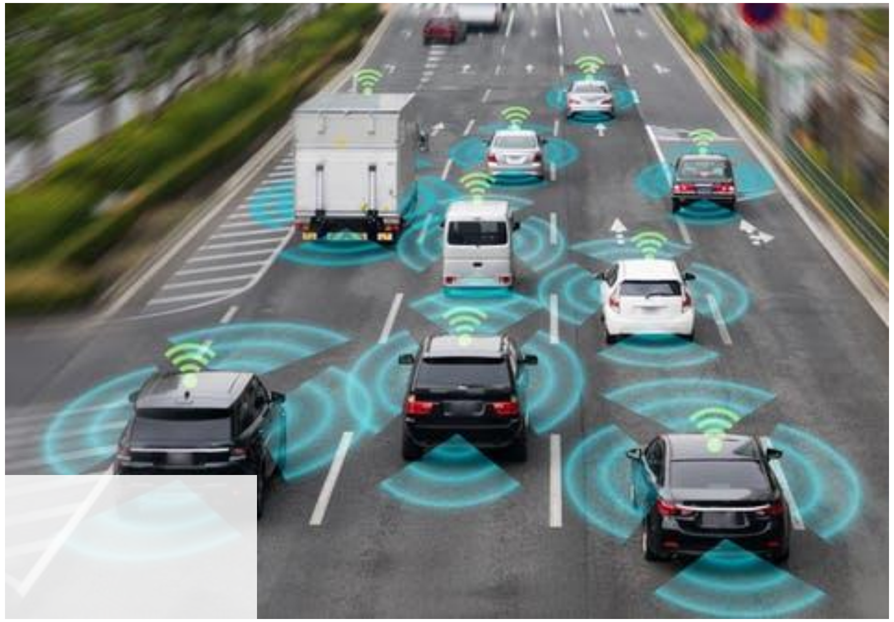
Aloisio Dourado, Teófilo Emidio de Campos
University of Brasilia
Brasilia, Brazil

Hansung Kim, Adrian Hilton
CVSSP, University of Surrey
Surrey, UK

3D Scene Reasoning

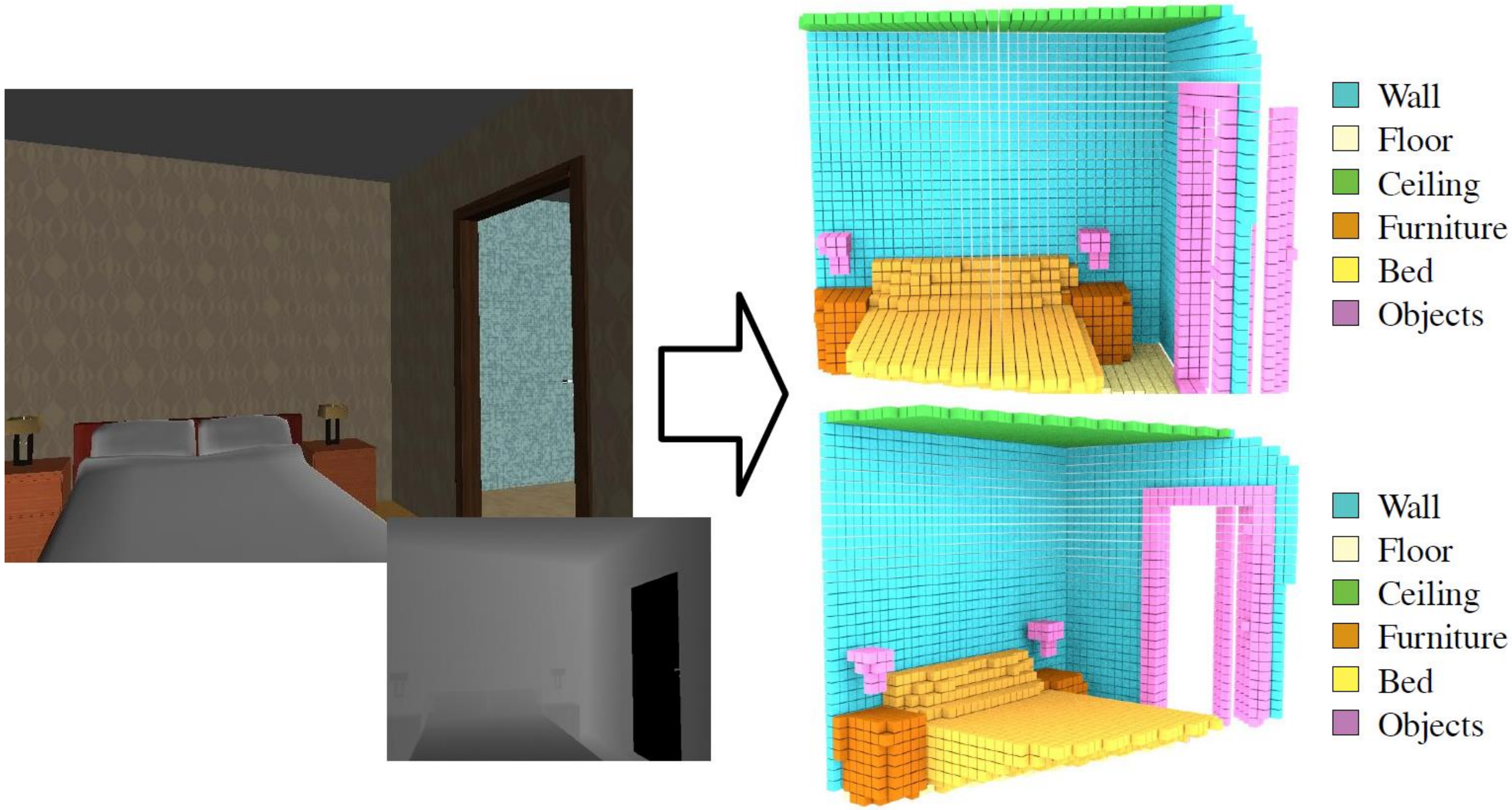






Applications

Semantic Scene Completion



Introduced by Song *et al.*[1] in 2017

Trained a 3D CNN that jointly deals with both completion and semantic segmentation

[1] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic Scene Completion from a Single Depth Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

Characteristics of current approaches

- Uses as input RGB-D (Microsoft® Kinect)
- Based on 3D CNNs
- Requires a large amount of data to train
- Trained on synthetic datasets (SUNCG)
- Fine-tuned on real data (NYU)
- Uses Flipped Truncated Signed Distance Function (F-TSDF)

Types of SSC Solutions

- Depth map only:
  - SSCNET: Song *et al*.[1]
  - Spatial Group Convolutions: Zhang *et al*.[2]
  - View-Volume Network : Guo and Tong[3]

Neglects the RGB channels from the input data

[1] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic Scene Completion from a Single Depth Image. In *CVPR,* 2017
[2] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, 2018
[3] Y. Guo and X. 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, pages 726–732, Stockholm, Sweden,July 2018

Types of SSC Solutions

- Depth maps plus RGB:
  - Guedes *et al.*[4]

Suffer from RGB data sparsity after projection to 3D

[4] A. B. S. Guedes, T. E. de Campos, and A. Hilton. Semantic scene completion combining colour and depth: preliminary experiments. CoRR, abs/1802.04735, 2018

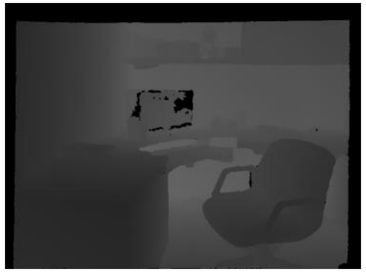
Types of SSC Solutions

- Depth map plus 2D segmentation:
  - Two stream 3D semantic scene completion: Garbade *et al.*[5]
  - TNetFusion: Liu *et al.*[6]

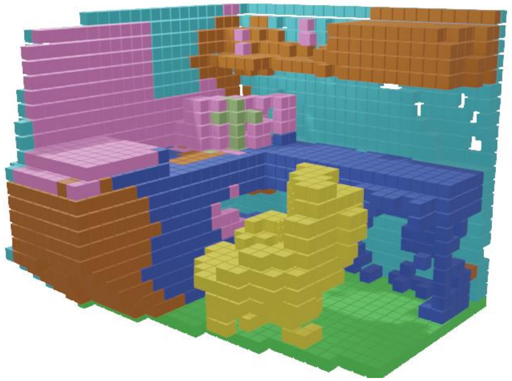
Requires a complex two step training procedure (2D CNN then 3D CNN)

[5] M. Garbade, J. Sawatzky, A. Richard, and J. Gall. Two stream 3D semantic scene completion. CoRR, abs/1804.03550, 2018
[6] S. Liu, Y. HU, Y. Zeng, Q. Tang, B. Jin, Y. Han, and X. Li. See and think: Disentangling semantic scene completion. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. CesaBianchi, and R. Garnett, editors,Conference on Neural Information Processing Systems (NeurIPS), pages 263–274. Curran Associates, Inc., 2018

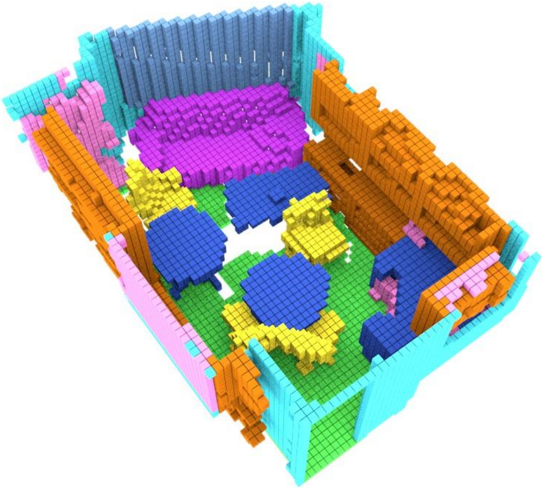
Current Semantic Scene Completion Limitations

Regular RGB-D Sensor




Panoramic Image from
Matterport Camera

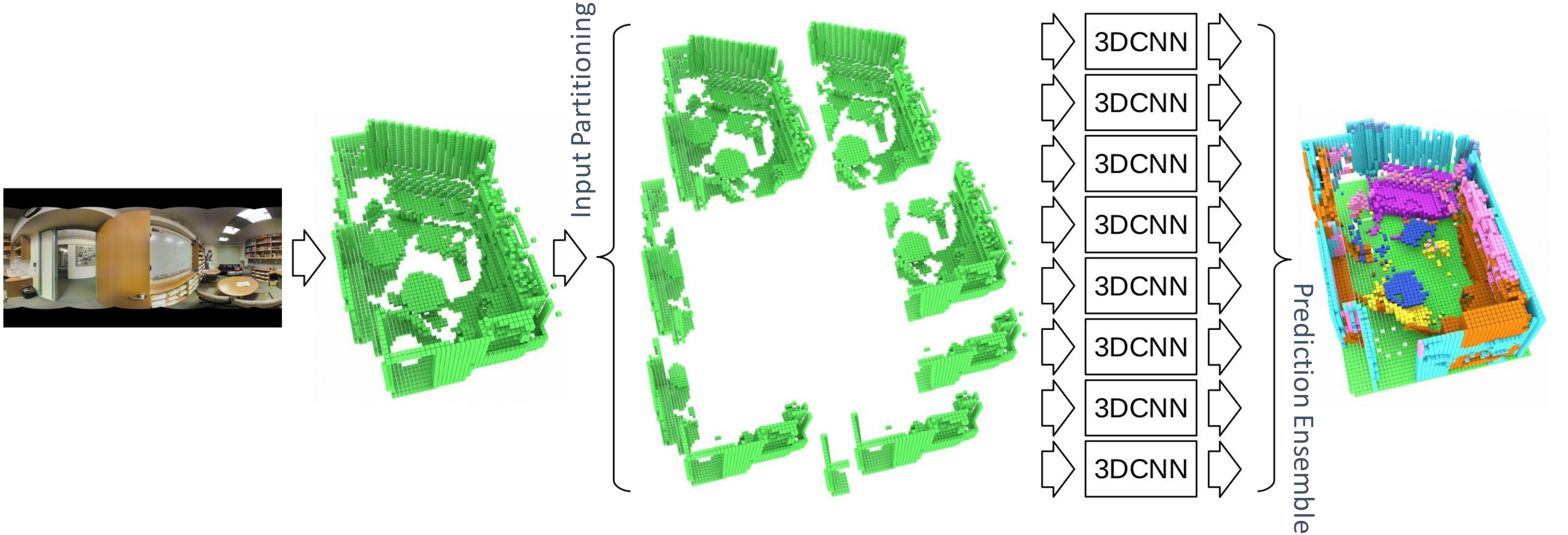


floor wall window chair table sofa furn. objects

Obstacles to 360^o Semantic Scene Completion

- The task has a high memory footprint
- Current 360^o datasets are not large enough or not diverse enough to train deep 3D CNNs

Our approach

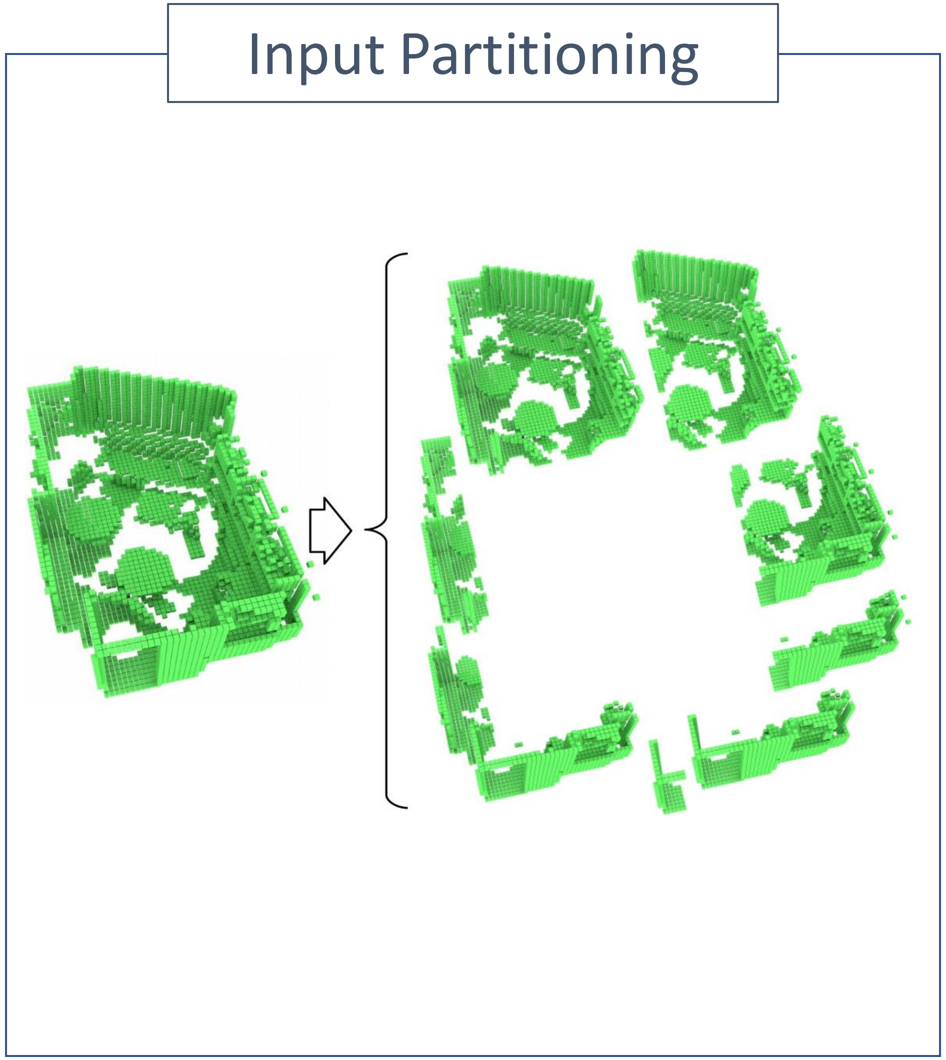


This approach allows to use existing large and diverse RGB-D datasets for training.

The 3DCNN is trained using SUNCG and fine-tuned in NYUDV2

Our approach

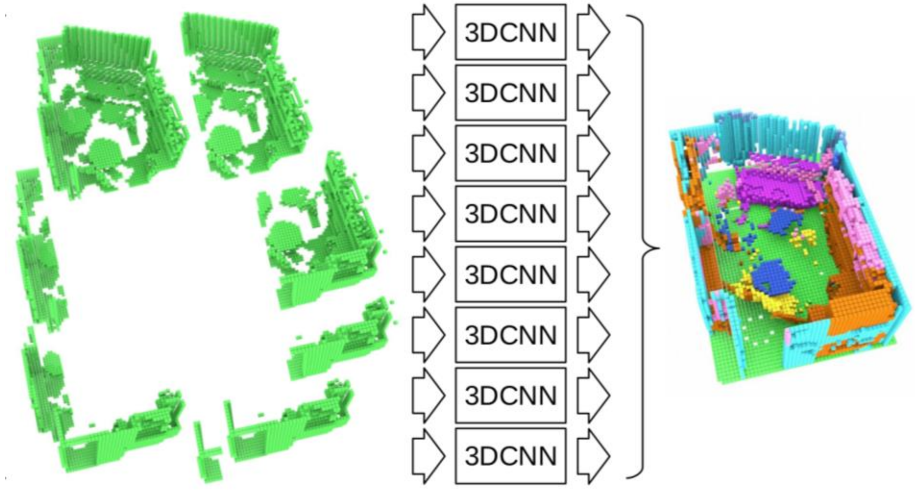
- Input volume:
  - 480 x 144 x 480 voxels
  - Voxel size: 0.02m
  - coverage: 9.6 x 2.8 x 9.6 m
- 8 partitions, emulating the field of view of a standard RGB-D sensor
- The partitions are taken from the sensor position, using a 45º step
- We move the point-of-view 1.7m back from the original sensor position, to get more overlapped coverage



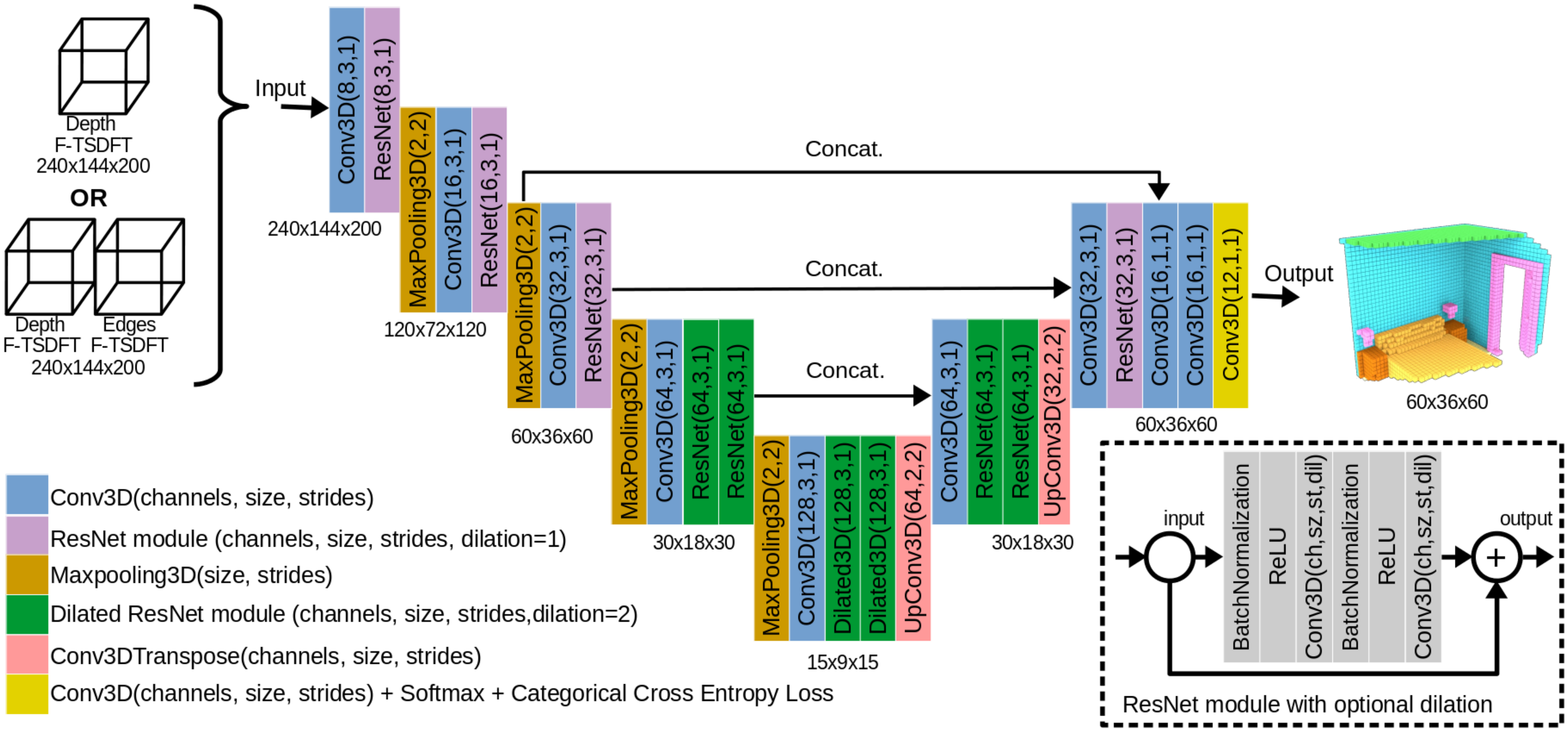
Our approach

- Each partition of the input is processed by our CNN, generating 8 predicted volumes
- Overlapping areas are ensembled using the sum rule
- Each predicted partition size is 60 x 36 x 60
- The resulting ensembled volume size is 120 x 36 x 120

Prediction Ensemble

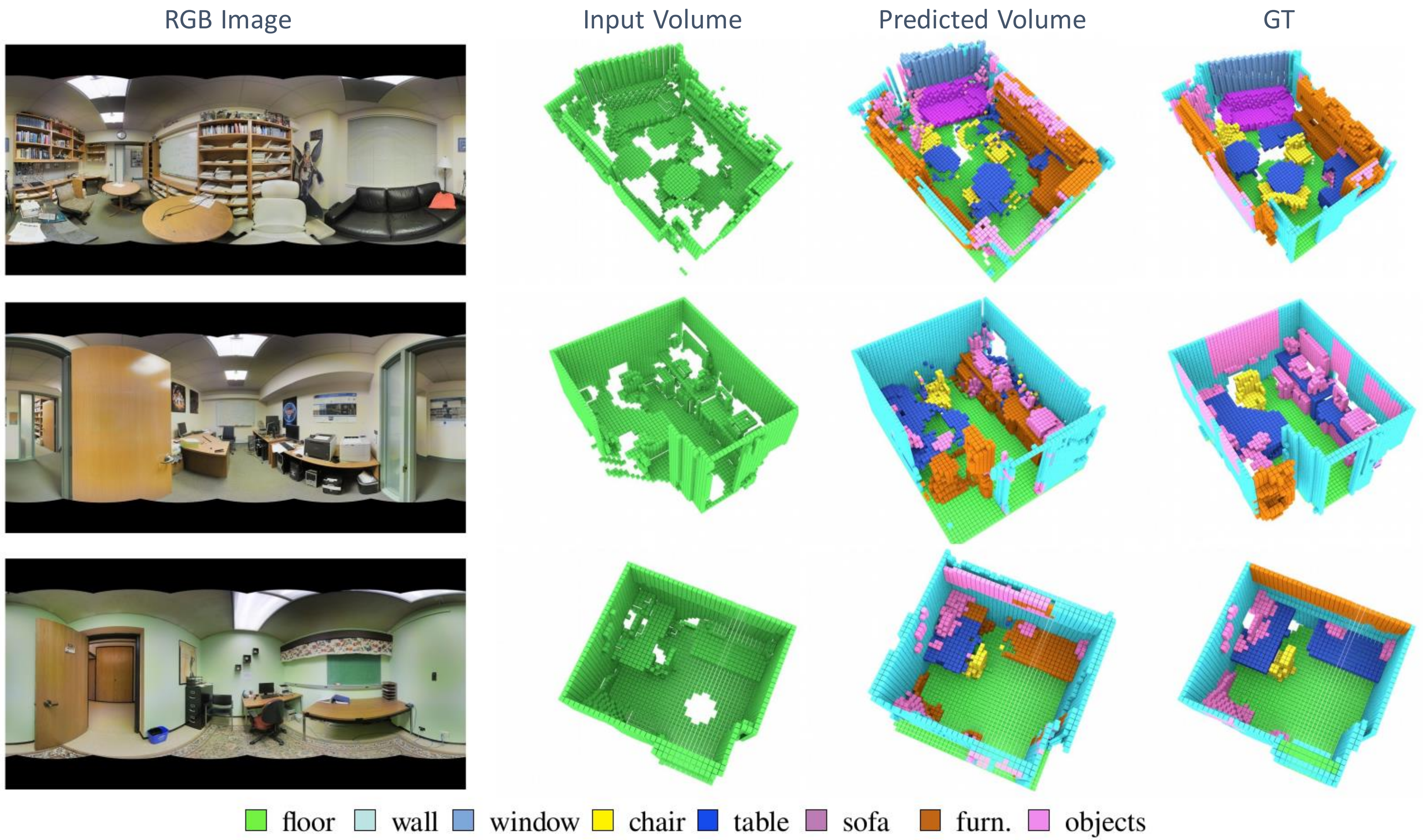


Our Network: EdgeNet[8]



[8] Dourado, A., de Campos, T. E., Kim, H., and Hilton, A. (2019). EdgeNet: Semantic scene completion from RGB-D images. Technical Report arXiv:1908.02893, Cornell University Library. http://arxiv.org/abs/1908.02893.

Results on Stanford 2D-3DS Dataset



Results on Stanford 2D-3DS Dataset

evaluation dataset	model	scene coverage	semantic scene completion (IoU, in percentages)											
			ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
NYU v2 RGB-D	SSCNet	partial	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	**13.9**	31.1	12.6	30.5
	SGC		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		**23.6**	**95.0**	28.6	**12.6**	13.1	**57.7**	**51.1**	16.4	9.6	**37.5**	13.4	32.6
Stanford 2D-3D-S	**Ours**	full (360°)	15.6	92.8	**50.6**	6.6	**26.7**	-	35.4	**33.6**	-	32.2	**15.4**	**34.3**

Experiments on Spherical Stereo Images

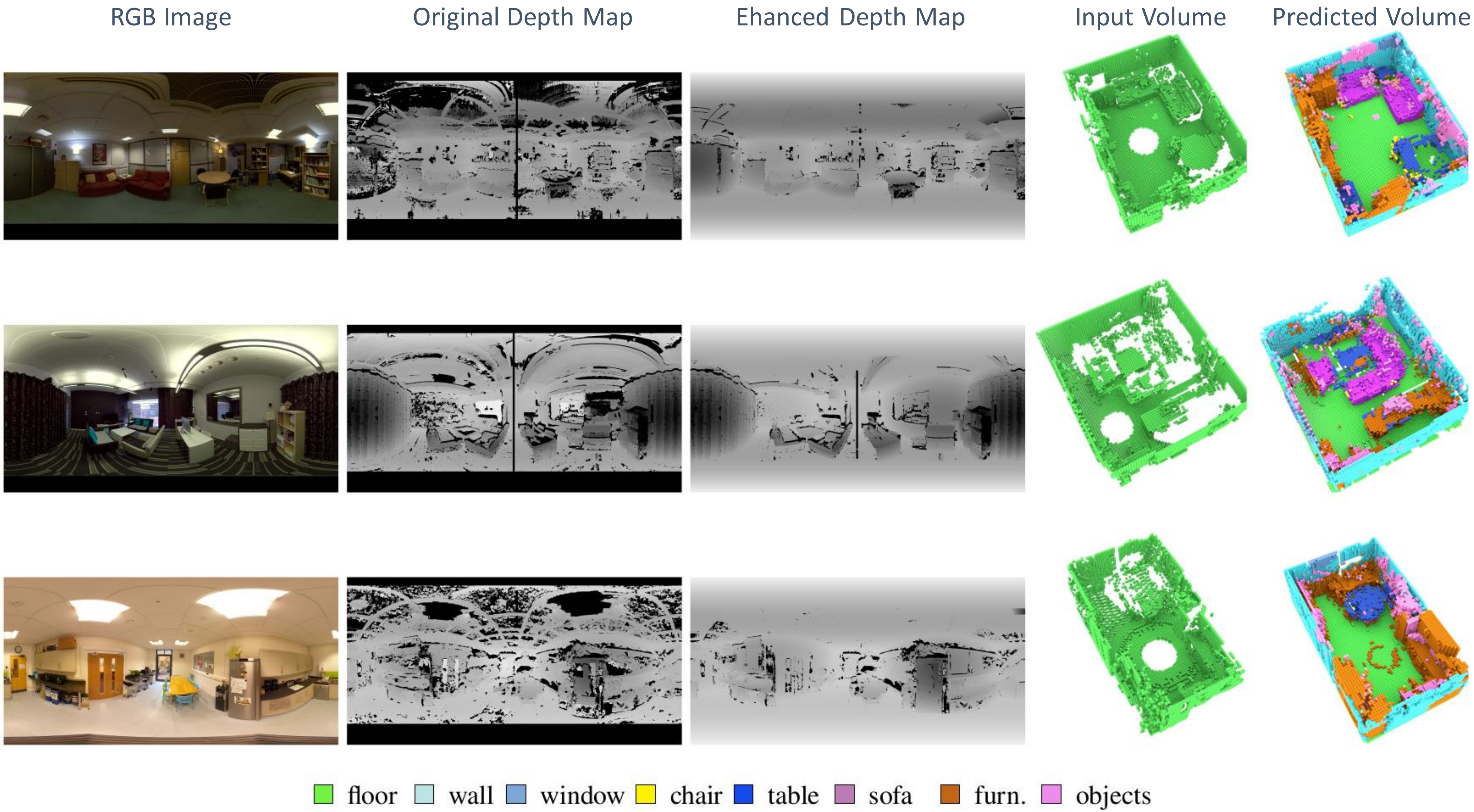
- Stereo capture using commercial 360^o cameras is one realistic approach to 360^o SSC
- The capture processes is faster compared to Matterport scaning
- However, depth estimation is subject to errors due to occlusions between two camera views and correspondence matching errors

Experiments on Spherical Stereo Images

- The scenes are captured as a vertical stereo image pair
- Dense stereo matching with spherical stereo geometry [7] is used to recover depth information
- We proposed a depth map enhancement procedure:
  - Align the scene using the Manhattan principle
  - Apply Canny Edge Detector to extract the most reliable depth estimations
  - Use RANSAC to fit a plane over coherent regions with similar colours

[7] Kim, H. and Hilton, A. (2015). Block world reconstruction from spherical stereo image pairs. Computer Vision and Image Understanding (CVIU), 139(C):104–121.

Results on Spherical Images



Conclusions

- This paper introduced the task of Semantic Scene Completion from a pair of 360º image and depth map.
- Our method predicts 3D voxel occupancy and its semantic labels for a whole scene from a single point of view
- The method can be applied to various range of images acquired from high-end sensors like Matterport to off-the-shelf 360º cameras
- Our method was evaluated the publicly available Stanford 2D-3D-Semantics dataset and a collection of 360º stereo images gathered with off-the-shelf spherical cameras.
- Qualitative analysis shows high levels of completion of occluded regions on both Matteport and spherical images.

Acknowledgements

- The authors would like to thank:
  - FAPDF(fap.df.gov.br)
  - CNPq grant PQ 314154/2018-3(cnpq.br)
  - PSRC Audio-Visual Media Platform Grant EP/P022529/1
  - TCU(tcu.gov.br)

Thank you!