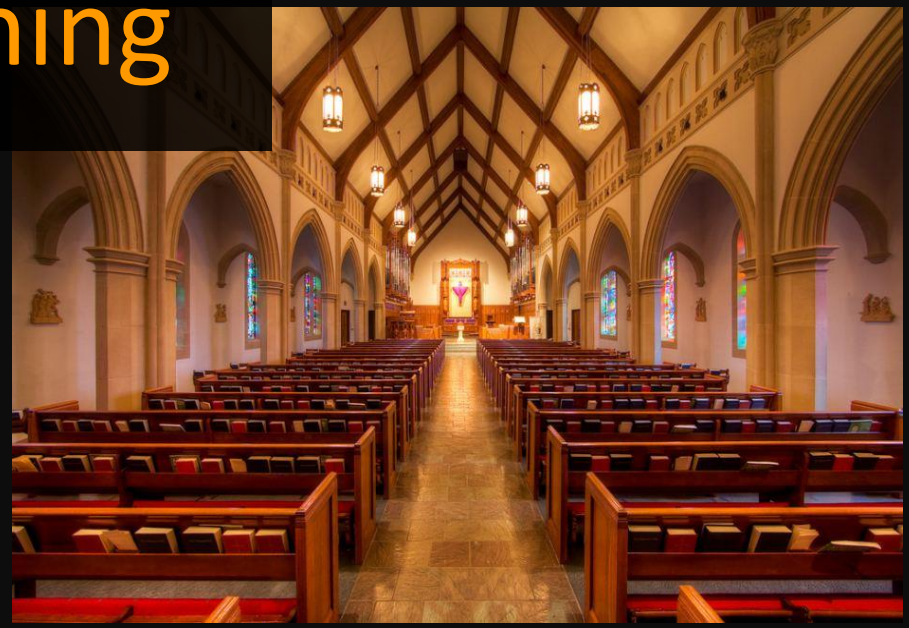
3D Indoor Semantic Scene Completion from a Single Point of View

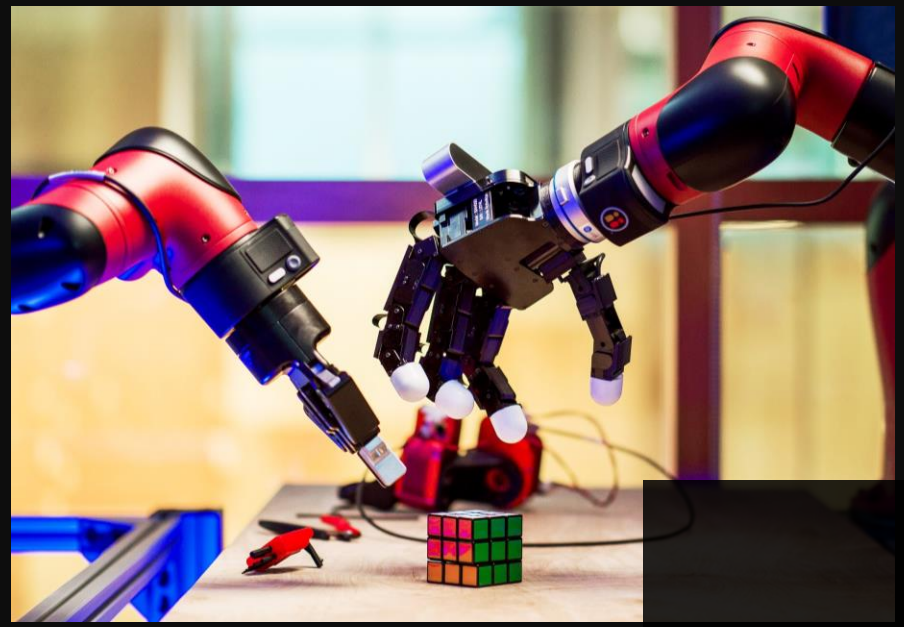
Aloisio Dourado
University of Brasilia
Brasilia, Brazil

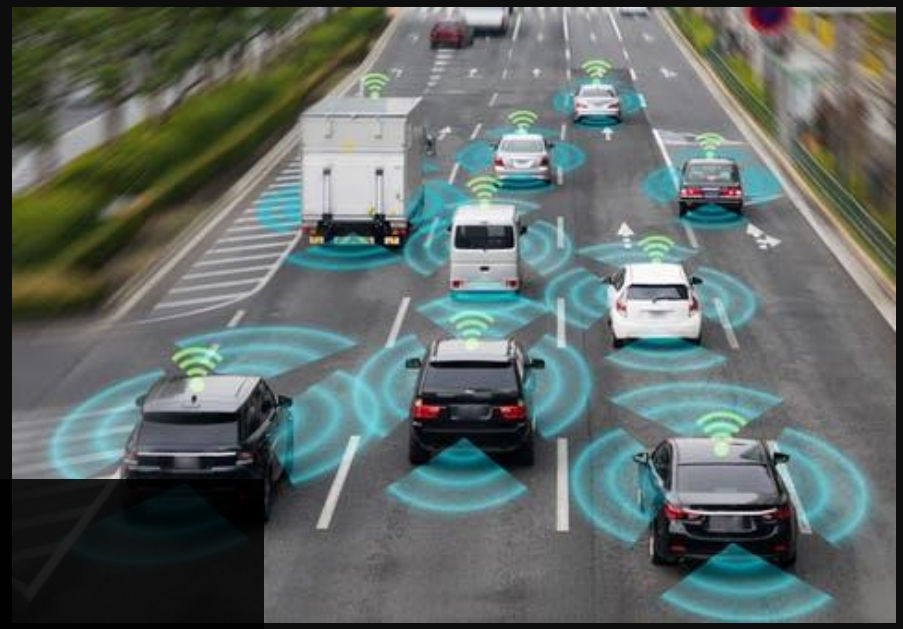
Presentation Outline

- Motivation
- Introduction to Semantic Scene Completion (SSC)
- Improving SSC from Regular RGB-D Images using Edges
- Extending SSC to a full 360^{o} scene coverage
- Conclusion
- Future Work

3D Scene Reasoning

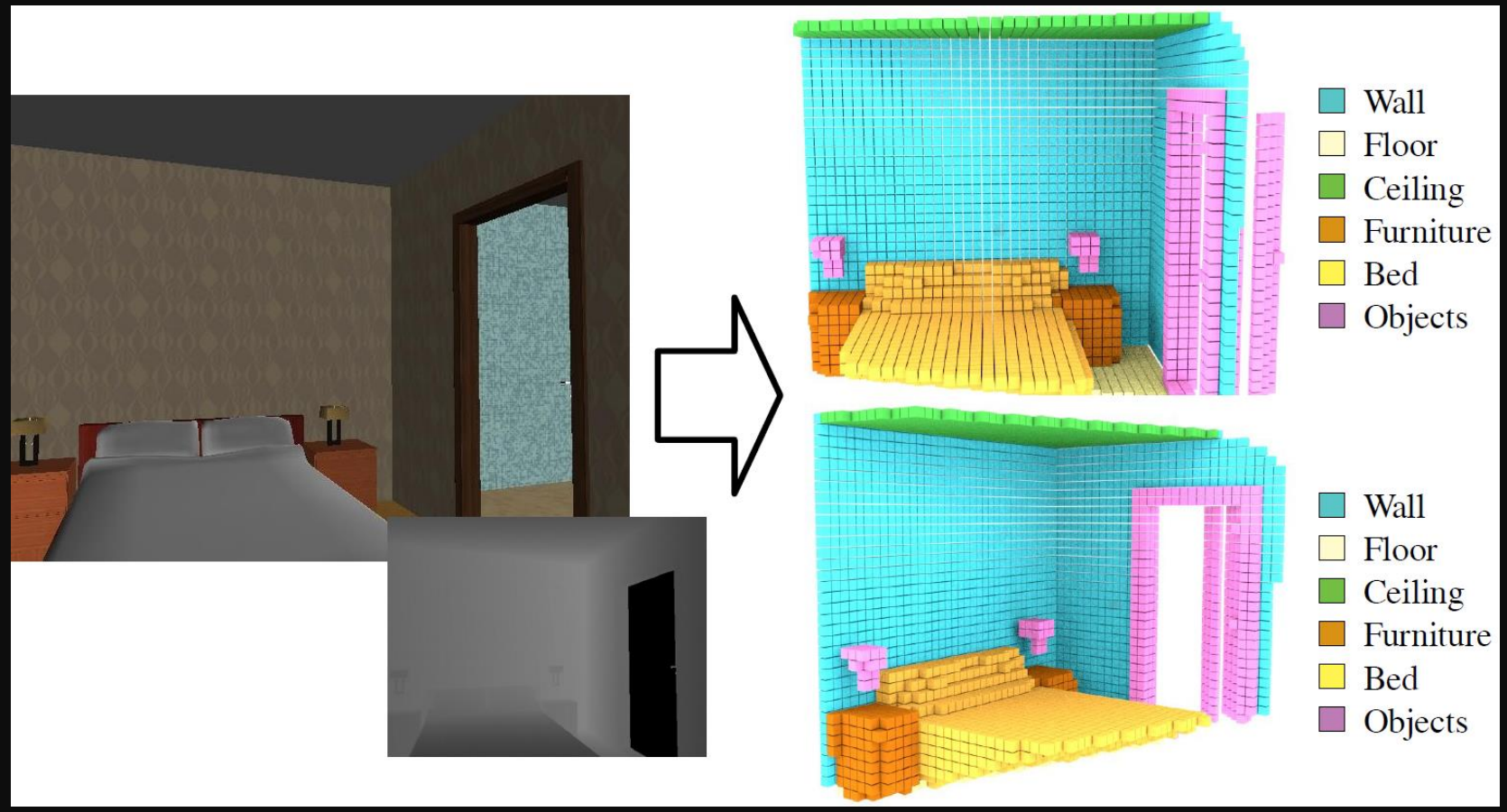






Applications

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

Previous works

- Depth map only:
  - SSCNET: Song *et al*.[1]
    - Typical contracting only 3D CNN with dilated convolutions
    - Depth map encoded with F-TSDF
    - Weighted softmax loss
    - Train on SUNCG, Fine tune on NYU
  - Spatial Group Convolutions: Zhang *et al*.[2]
    - U-Shaped 3D CNN
    - SGC used in the decoding branch
    - No fine tune
  - View-Volume Network : Guo and Tong[3]
    - Applyied 2D convolutions to the depth maps then projected resulting features to 3D using a projection layer
    - Used a proprietary evaluation protocol

[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

Previous works

- Depth maps plus RGB:
  - Guedes *et al*.[4]
    - 3 channels of RGB data projected to 3D
    - Same architecture as SSCNET
    - no significant improvement
    - Some implementation details:
      - Customized Caffe environment, locked to an old version: hard to setup
      - All job done inside Caffe layers: GPU overload, time consuming
      - Memory intensive: batch size = 1

[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

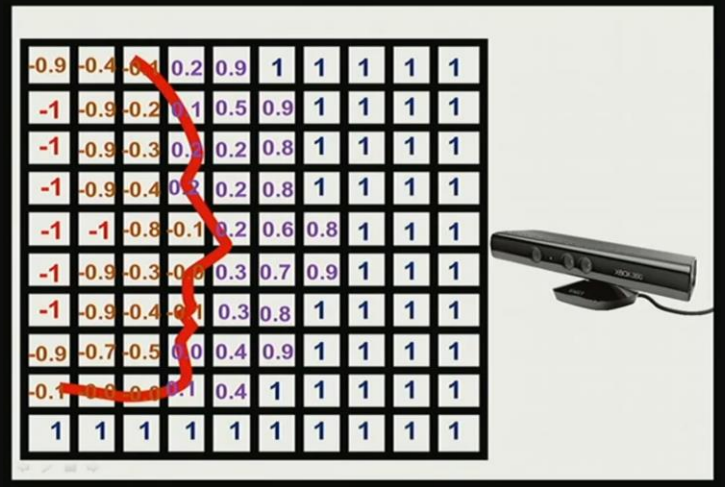
Previous works

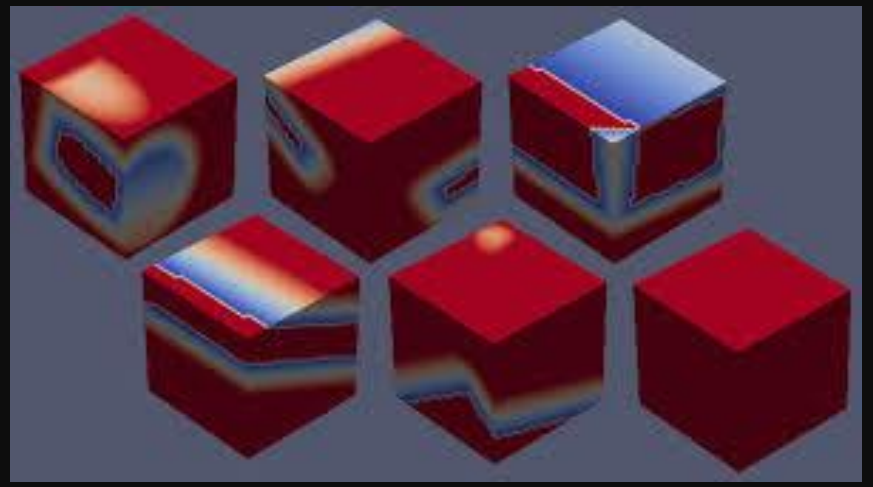
- Depth map plus 2D segmentation:
  - Two stream 3D semantic scene completion: Garbade *et al*.[5]
    - 2D pretrained segmentation CNN and a Fully Connected CRF to generate a 2D segmentation map from RGB
    - Predicted 2D labels are projected to 3D and fused to the 3D branch (no one-hot-encoding)
  - TNetFusion: Liu *et al.*[6]
    - depth maps and RGB information as input of an encoder-decoder 2D segmentation CNN
    - ResNet-101 as the encoder branch
    - Generated features are projected to 3D and fused to the 3D stream

[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

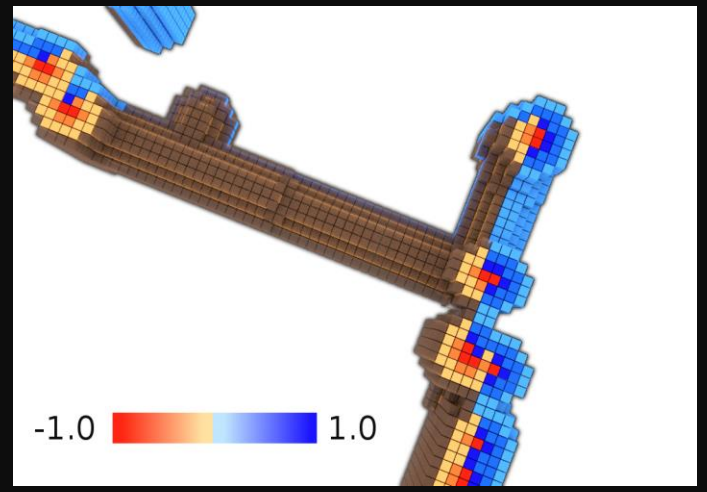
One note about previous works

- When projecting 2D data to 3D, resulting volume is sparse
- Song *et al.* has shown that using F-TSDF to generate dense 3D input volumes improves results
  - It is easy to apply F-TSDF to occupancy volume because it is binary
  - RGB data is not binary!





TSDF

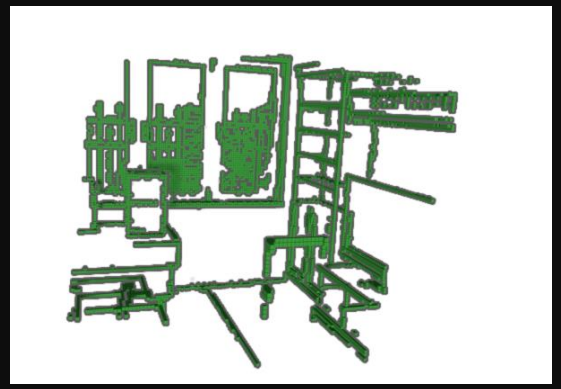


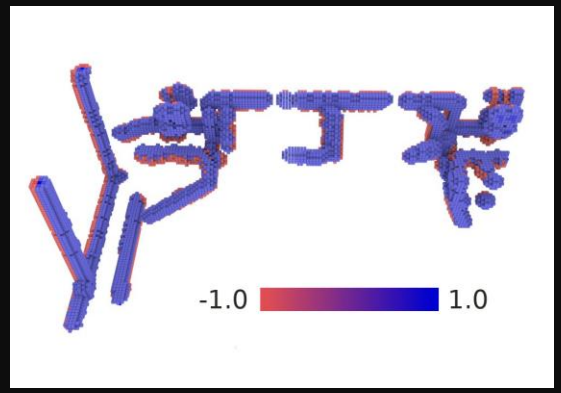
F-TSDF

$$\text{F-TSDF} = \text{sign}(\text{TSDF}) \cdot (1 - |\text{TSDF}|)$$

Our approach: EdgeNet

- We extract information from RGB data using image edges:
  - Easy to get: Canny Edge Detector[7]
  - It is possible to apply F-TSDF to image edges (binary data)



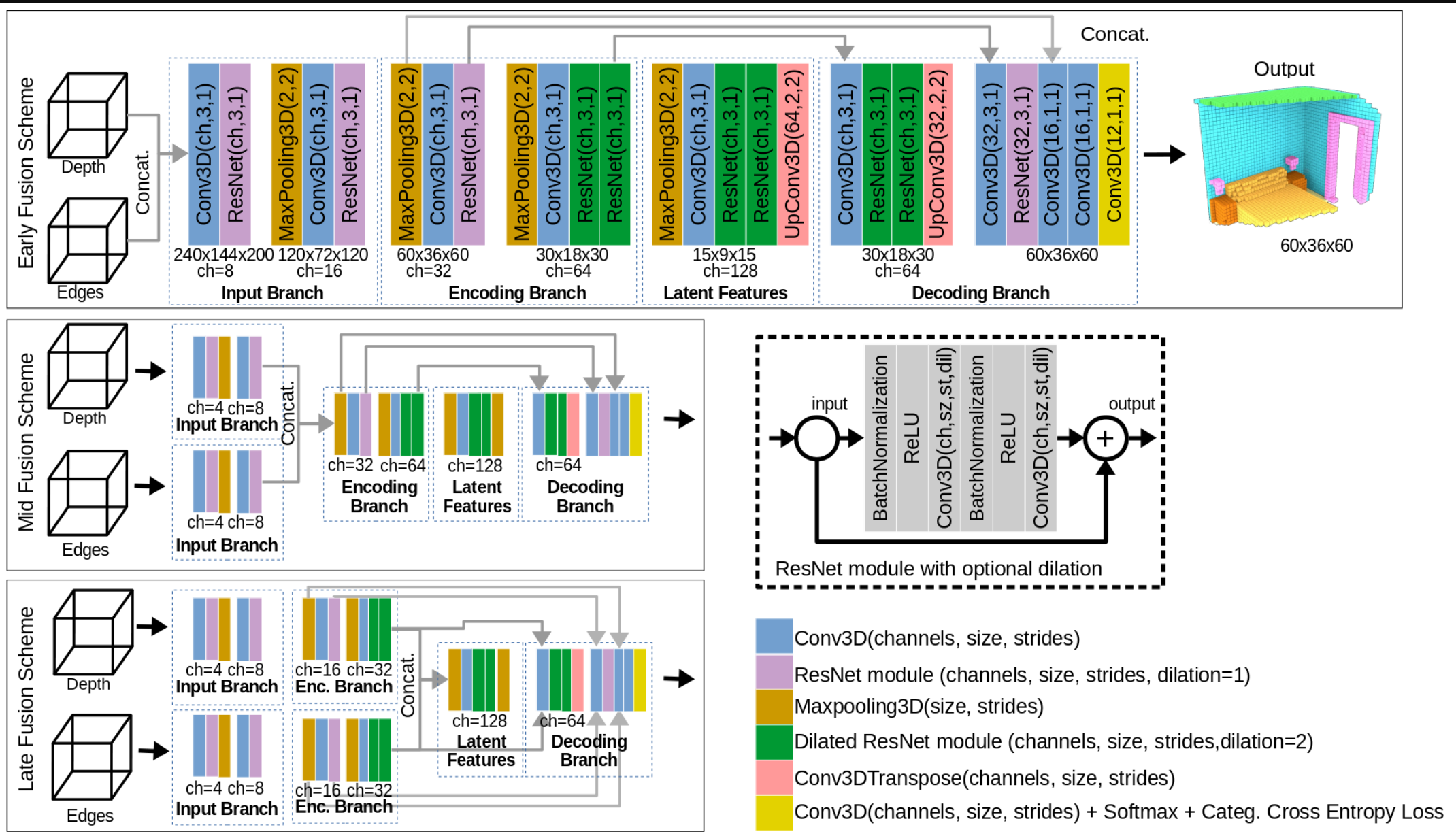


[7] J. Canny. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(6):679–698, Nov 1986

Our approach

- Preparing data takes time...
  - SUNCG dataset contains more than 130K 3D scenes
  - To train a single model, each scene is processed several times, according to the number of epochs
  - F-TSDF calculation is computationally intensive
- So, we preprocess all data, only once!
  - F-TSDF is calculated using portable c++ cuda code
  - We provide a software interface between cuda and python
  - Preprocessing code is independent from the deep learning framework
  - Preprocessing also includes calculating an occupancy grid for data balancing:
    - occupied voxels inside the room and FOV
    - non occupied voxels inside the room and FOV
    - all other voxels

Network architecture



Training protocol

- Data balancing:
  - for each training batch, we randomly initialize a tensor *rand* with ones and zeroes using the ratio $r=(2\sum occu/\sum occl)$
  - The final weight tensor is $w= occu + occl \cdot rand$

- Weighted Categorical Cross Entropy Loss:
  - $Lcce(p, y) =-\sum(w \cdot y \cdot log\ p)$

- Train on SUNCG, fine tune on NYU v2
- One Cycle Learning [8]
- SGD optimizer
- Batch size = 3
- training time:
  - SUNCG: from 7 days to 4 days
  - NYU: from 30 hours to 6 hours

[8] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. CoRR , abs/1803.09820, 2018.

Datasets

- SUNCG
  - 130K+ synthetic 3D scenes rendered from 45K+ human generated house models
  - Camera poses are brute force generated, then randomly select according NYU pose distribution
  - Only depth maps were provided, we generated RGB data from the house models
  - Default train/test split is provided
- NYUv2
  - Real 3D scenes captured with Kinect
  - 795 scenes for training and 654 for testing

Results

input	model	scene completion			semantic scene completion (IoU, in percentages)											
		prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
d	SSCNet[24]	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
	SSCNet*	92.7	89.7	83.8	97.0	94.6	74.3	51.1	43.7	78.2	70.9	49.5	45.2	61.0	51.3	65.2
	DCRF [25]	–	–	–	95.4	84.3	57.7	24.5	28.2	63.4	55.3	34.5	19.6	45.8	28.7	48.8
	VVNetR-120 [9]	90.8	91.7	84.0	**98.4**	87.0	61.0	54.8	49.3	83.0	**75.5**	55.1	43.5	68.8	57.7	66.7
	EdgeNet-D	93.1	90.4	84.8	97.2	94.4	78.4	56.1	50.4	80.5	73.8	54.5	49.8	69.5	59.2	69.5
d+s	SNetFuse[14]	56.7	91.7	53.9	65.5	60.7	50.3	56.4	26.1	47.3	43.7	30.6	37.2	44.9	30.0	44.8
	TNetFuse[14]	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
d+e	SSCNet-E	92.8	89.6	83.8	97.0	94.5	74.6	51.8	43.9	77.0	70.8	49.3	49.2	62.1	52.0	65.7
	EdgeNet-EF(Ours)	**93.7**	90.3	**85.1**	97.2	94.9	**78.6**	57.4	49.5	80.5	74.4	**55.8**	51.9	70.1	**62.5**	**70.3**
	EdgeNet-MF(Ours)	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
	EdgeNet-LF(Ours)	93.0	89.6	83.9	97.0	94.6	76.4	52.0	44.6	79.8	71.5	48.9	48.3	66.1	55.9	66.8

Ablation study on SUNCG test set

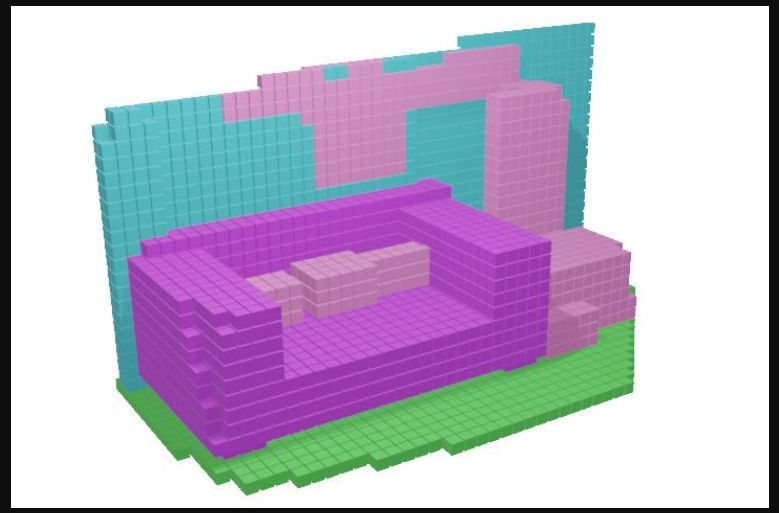
Results

train	input	model	scene completion			semantic scene completion (IoU, in percentages)											
			prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SUNCG	d	SSCNet[24]	55.6	91.9	53.2	5.8	81.8	19.6	5.4	12.9	34.4	26	13.6	6.1	9.4	7.4	20.2
	d+e	EdgeNet-EF(Ours)	**61.9**	80.0	**53.6**	9.1	**92.9**	18.3	5.7	15.8	40.4	30.7	9.2	3.3	13.7	11.6	22.8
		EdgeNet-MF(Ours)	60.7	80.3	52.8	**11.0**	92.3	**20.5**	7.2	**16.3**	42.8	**32.8**	**10.5**	**6.0**	**15.7**	**11.8**	**24.3**
		EdgeNet-LF(Ours)	59.9	**80.5**	52.3	3.2	87.1	19.9	**8.6**	15.4	**43.5**	32.3	8.8	4.3	13.7	10.0	22.4
NYU	d	SSCNet[24]	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
	d+e	EdgeNet-EF(Ours)	**78.1**	65.1	55.1	**21.8**	**95.0**	27.3	**8.4**	6.8	**53.1**	38.6	7.5	0.0	30.4	**13.3**	27.5
		EdgeNet-MF(Ours)	76.0	68.3	**56.1**	17.9	94.0	**27.8**	2.1	9.5	51.8	**44.3**	9.4	3.6	**32.5**	12.7	**27.8**
		EdgeNet-LF(Ours)	75.5	67.5	55.4	19.8	94.9	24.4	5.7	7.2	50.3	38.8	10.0	0.0	33.2	12.2	27.0
SUNCG + NYU	d	SSCNet[24]	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
		DCRF[25]	-	-	-	18.1	92.6	27.1	10.8	18.8	54.3	47.9	17.1	15.1	34.7	13.0	31.8
		VVNetR-120[9]	69.8	83.1	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
	d+c	Guedes *et al.* [7]	-	-	56.6	-	-	-	-	-	-	-	-	-	-	-	30.5
	d+s	Garbade *et al.* *[6]	69.5	82.7	**60.7**	12.9	92.5	25.3	20.1	16.1	56.3	43.4	17.2	10.4	33.0	14.3	31.0
		SNetFuse[14]	67.6	**85.9**	**60.7**	22.2	91.0	28.6	**18.2**	19.2	56.2	51.2	16.2	12.2	37.0	17.4	33.6
		TNetFuse[14]	67.3	85.8	**60.7**	17.3	92.1	28.0	16.6	19.3	**57.5**	**53.8**	**17.7**	**18.5**	**38.4**	**18.9**	**34.4**
	d+e	EdgeNet-EF(Ours)	77.0	70.0	57.9	16.3	**95.0**	27.9	14.2	17.9	55.4	50.8	16.5	6.8	37.3	15.3	32.1
		EdgeNet-MF(Ours)	**79.1**	66.6	56.7	**22.4**	**95.0**	**29.7**	15.5	**20.9**	54.1	53.0	15.6	14.9	35.0	14.8	33.7
		EdgeNet-LF(Ours)	77.6	69.5	57.9	20.6	94.9	29.5	9.8	18.1	56.2	50.5	11.4	5.2	35.9	15.3	31.6

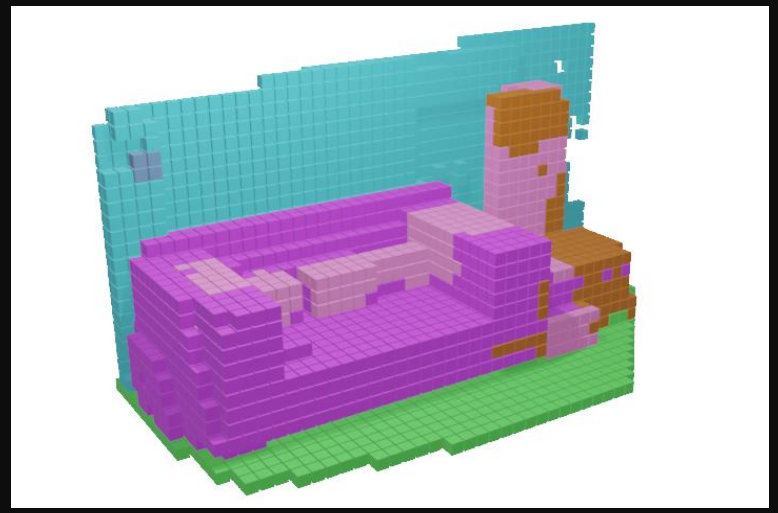
Semantic scene completion results on NYU test set

Qualitative results

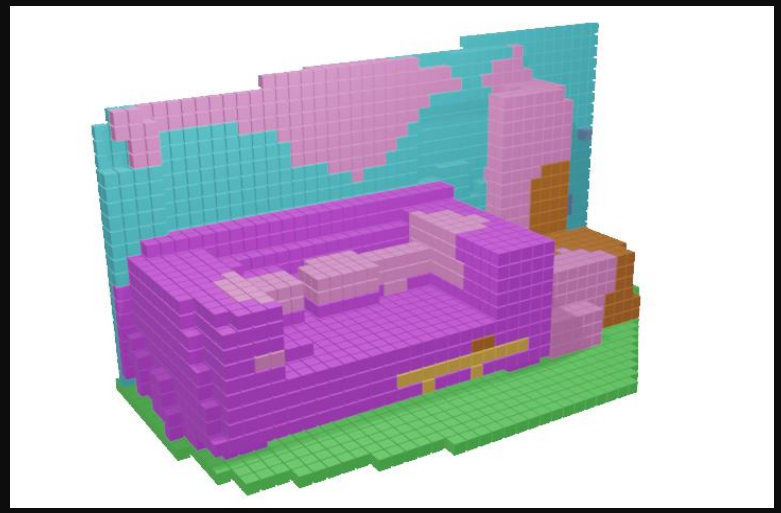
Image



Ground Truth

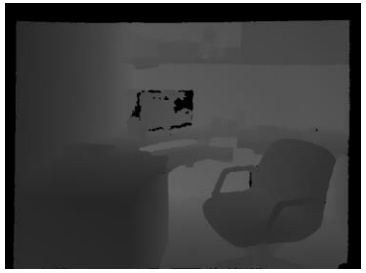


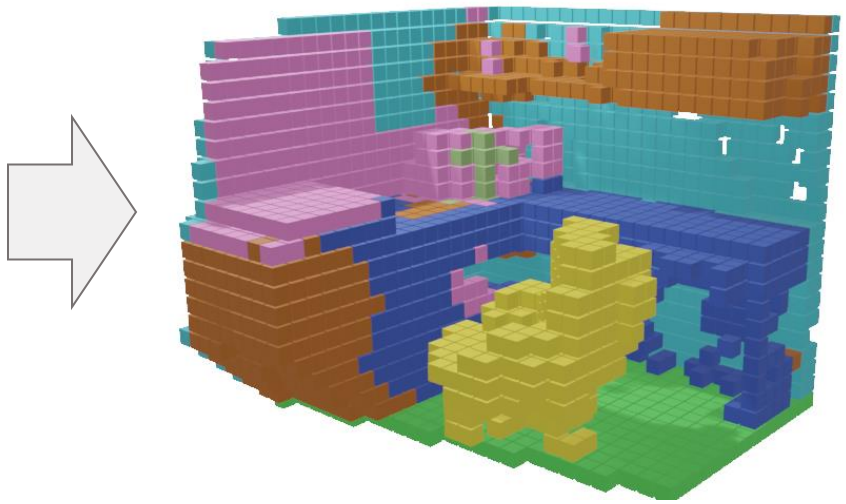
SSCNet



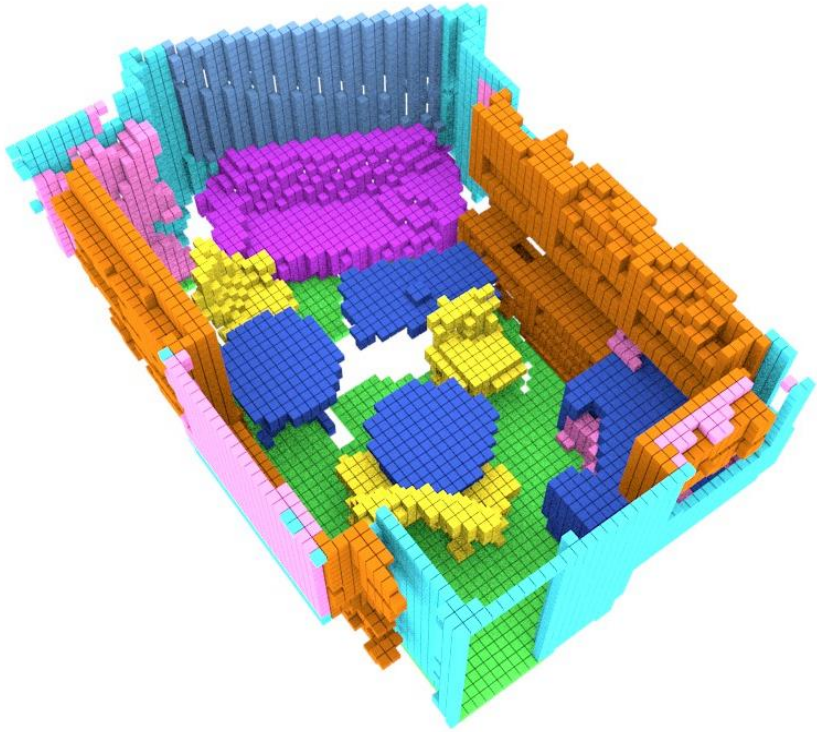
EdgeNet-MF

Current Semantic Scene Completion Limitations





Regular RGB-D Sensor

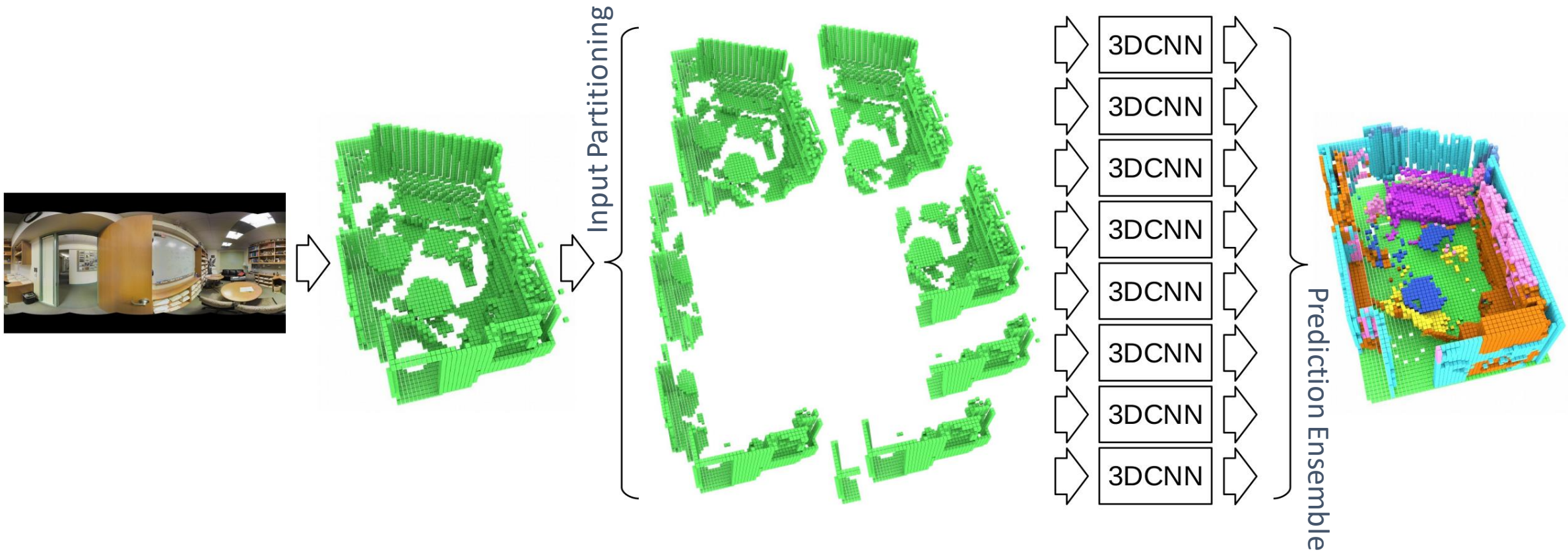


Panoramic Image from Matterport Camera

Obstacles to 360^{o} Semantic Scene Completion

- The task is highly memory consuming – a naïve full coverage approach may not be trainable with currently available GPUs
- Current 360^{o} datasets are not large enough or not diverse enough to train deep 3D CNNs

Our approach



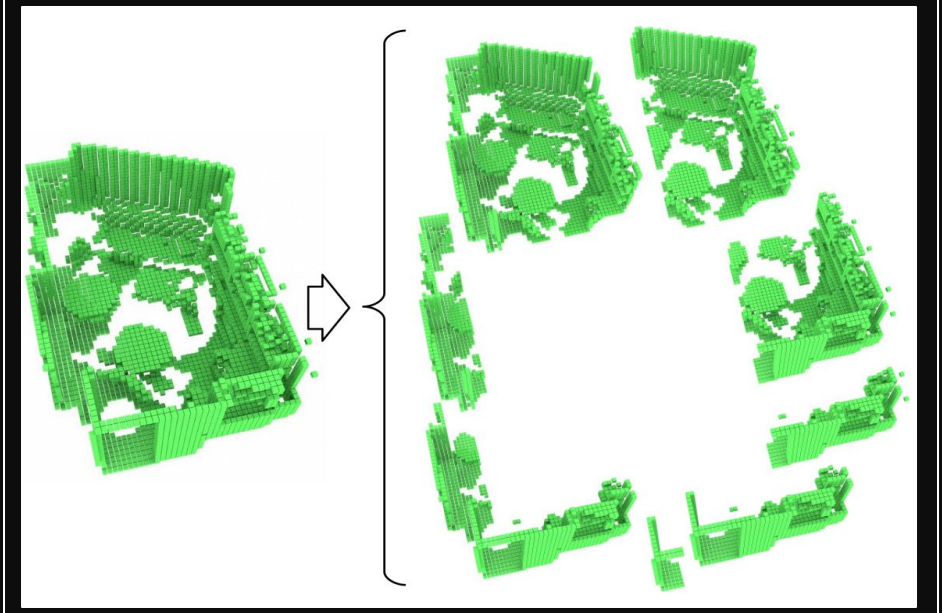
The 3DCNN is trained using SUNCG and fine-tuned in NYUDV2

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

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 45° step
- We move the point-of-view 1.7m back from the original sensor position, to get more overlapped coverage

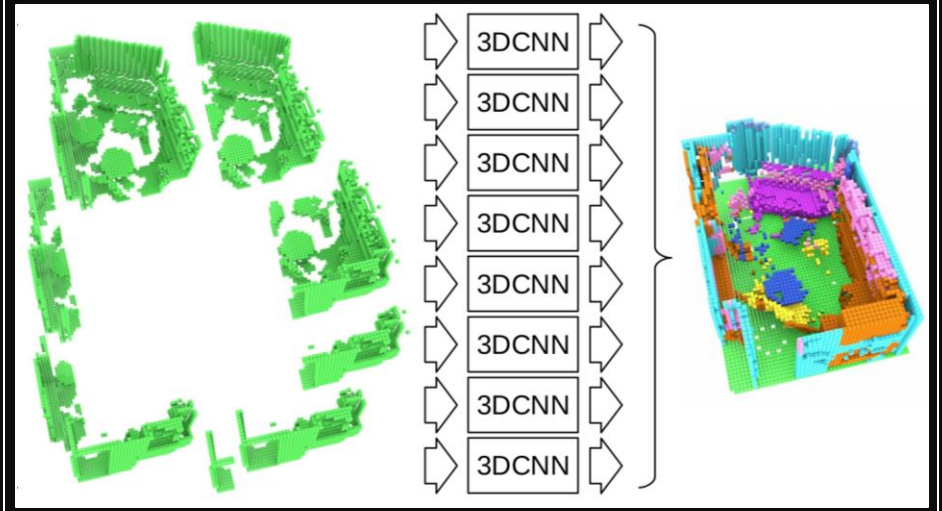
Input Partitioning



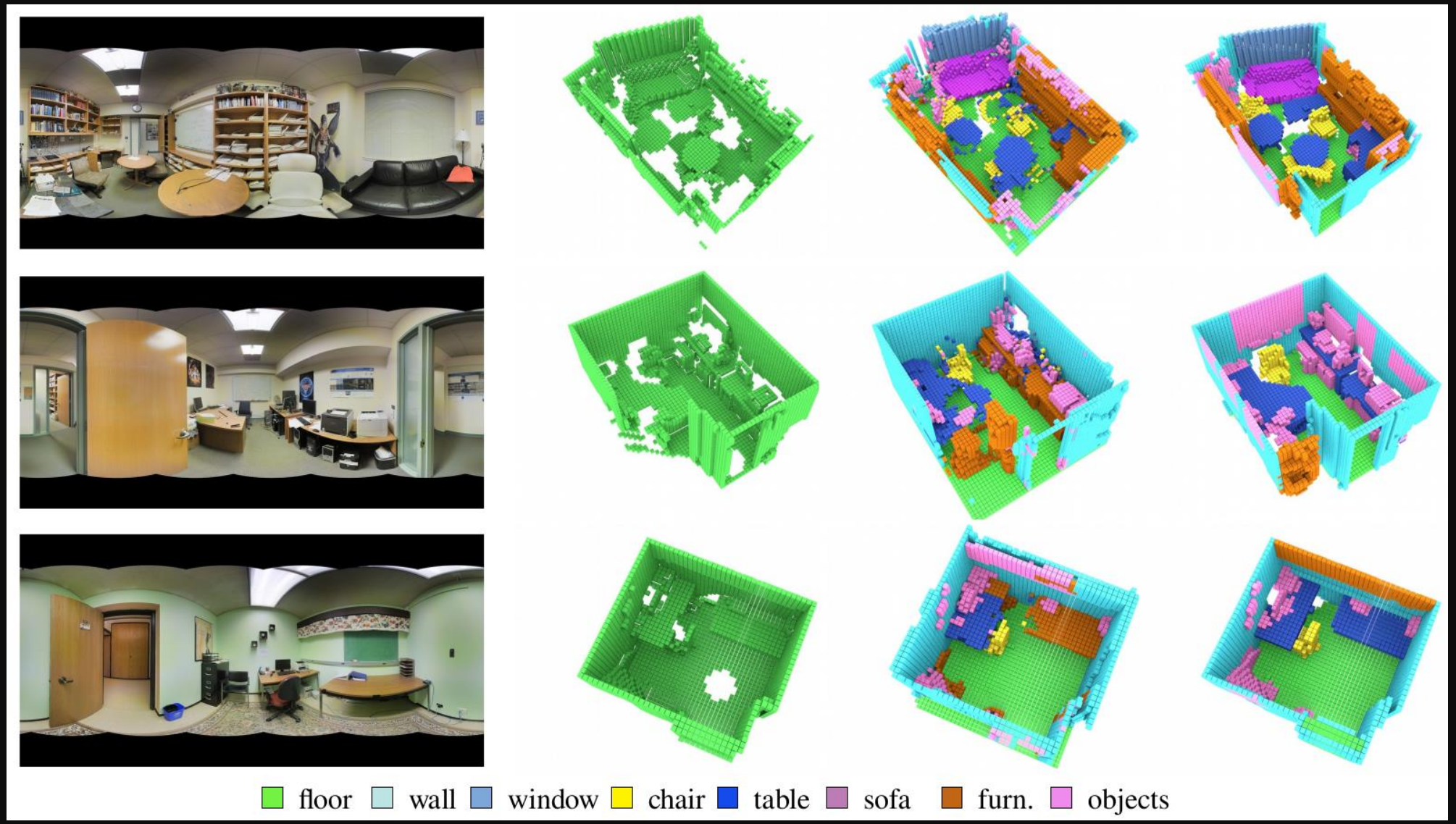
Our approach

- 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



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

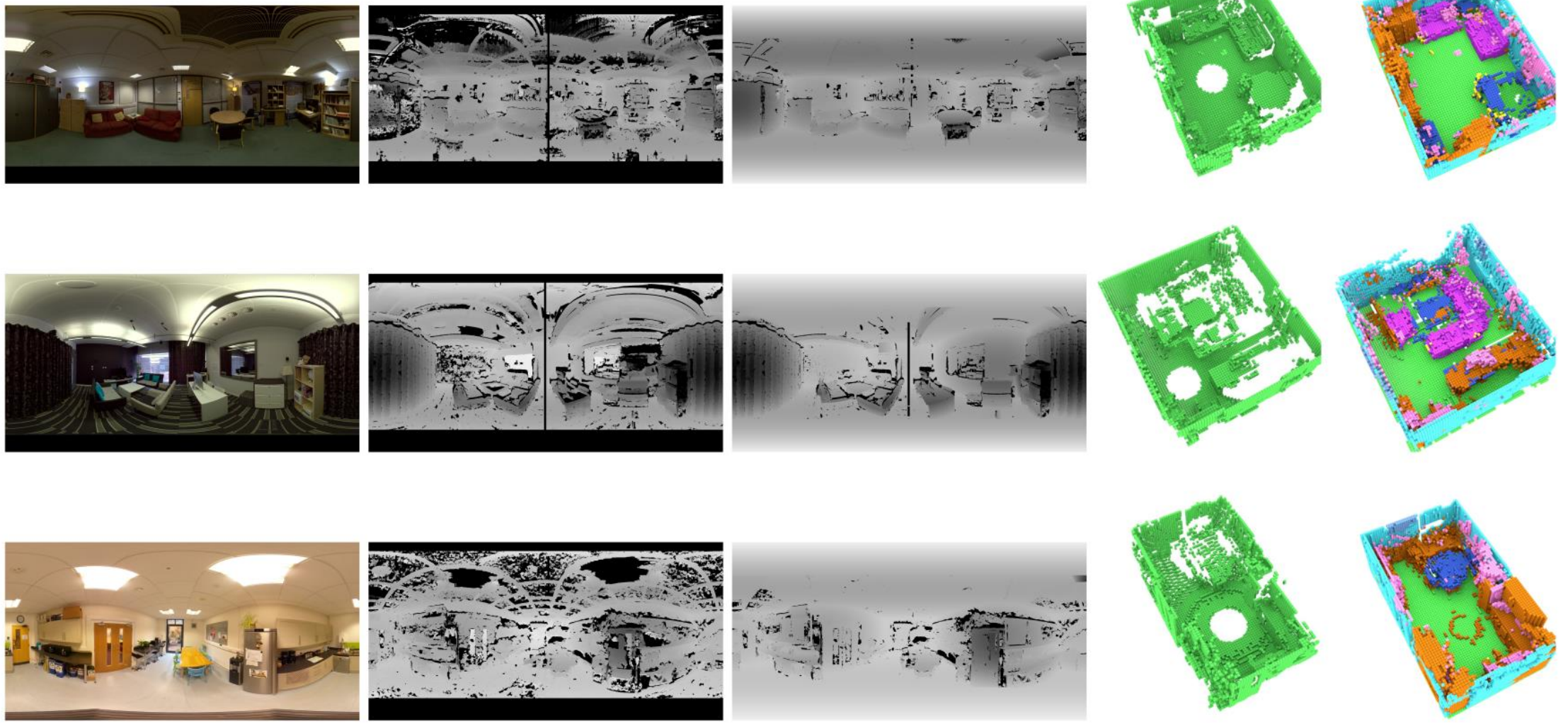
Experiments on Spherical Stereo Images

- The scenes are captured as a vertical stereo image pair
- Dense stereo matching with spherical stereo geometry [7] is used to recover depth information
- 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

RGB Image | Original Depth Map | Enhanced Depth Map | Input Volume | Predicted Volume



Conclusions

- The paper presented a new approach to fuse depth and colour into a CNN for semantic scene completion: the use of F-TSDF encoded 3D projected edges extracted from RGB images
- We presented a new end-to-end network architecture capable of properly aggregating edges and depth
- Experiments showed that both aggregating edges and the new proposed architecture have positive impact on semantic scene completion
- Qualitative results show visually perceptible improvements in 3D label inferences
- We have achieved improvement over the state-of-the-art result on the NYUv2 dataset, or end-to-end approaches
- We developed a faster lightweight training pipeline for the task

Conclusions

- We also 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
- Evaluated on the publicly available Stanford 2D-3D-Semantics dataset and on 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.