University of Brasília - UnB
Institute of Exact Sciences
Department of Computer Science

Towards Complete 3D Indoor Scene Understanding from a Single Point-of-View

Current Stage of the Research Project
October/2020

Aloisio Dourado Neto

Supervisor
Prof. Dr. Teófilo Emidio de Campos

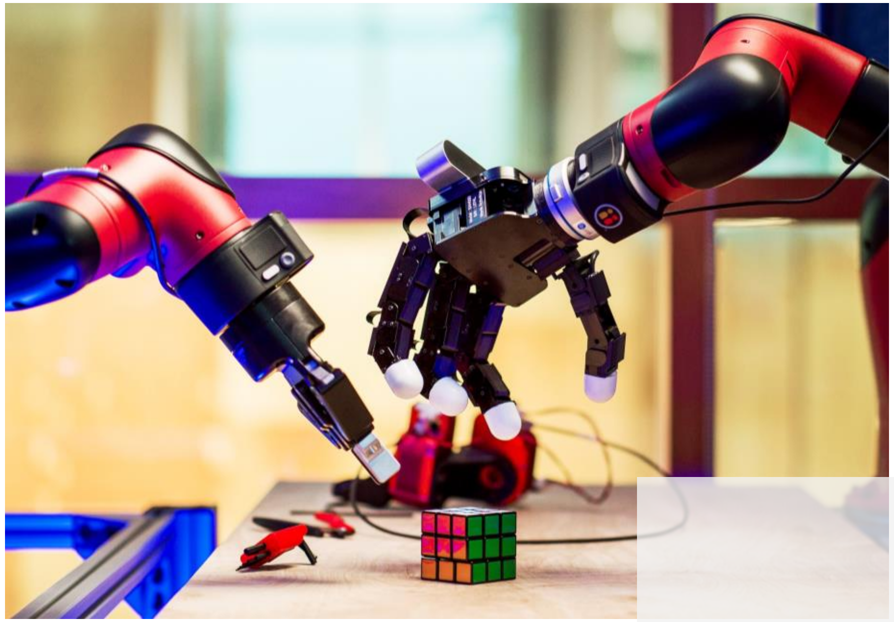
Presentation Outline

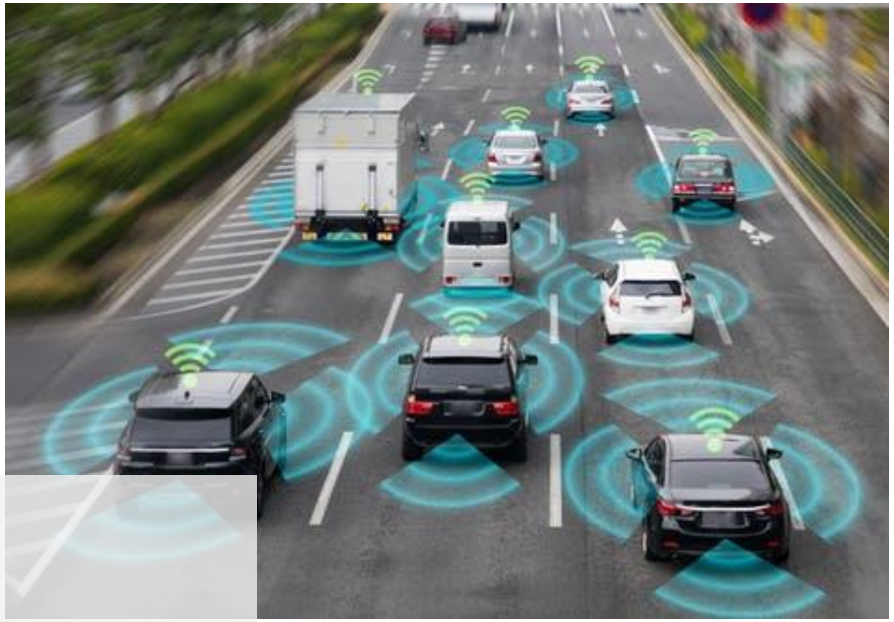
- Introduction
  - Motivation
  - The Semantic Scene Completion (SSC) task
  - Problem statement
- Previous works
- Concrete contributions, so far
  - Using 2D edges to improve detection of hard classes
  - Extending SSC to 360 degree
- Next steps

Introduction

Motivation

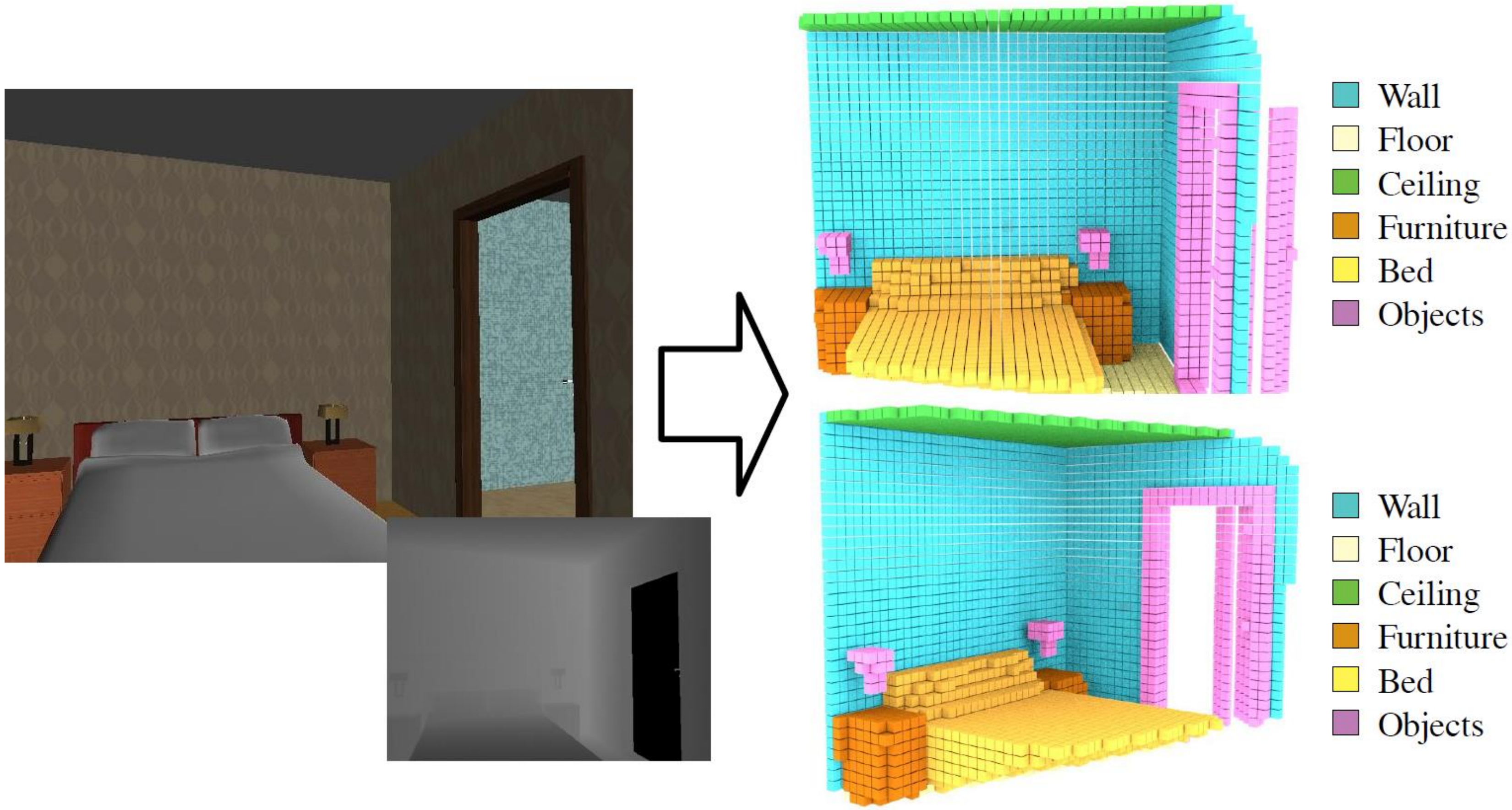
3D Scene Reasoning





Applications

Semantic Scene Completion

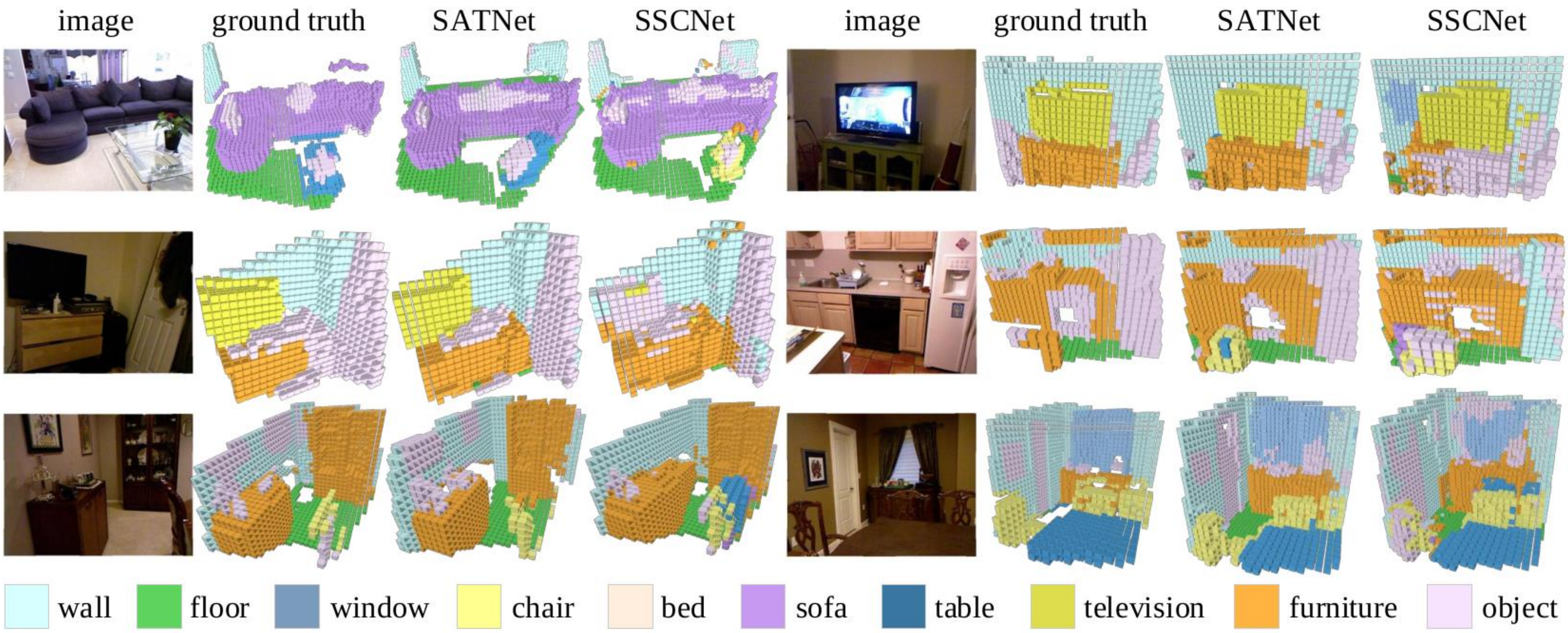


Introduced by Song *et al*.[107] in 2017

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

[107] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE. 2, 3, 4, 18, 45, 46, 47, 51, 52, 53, 64, 68, 70

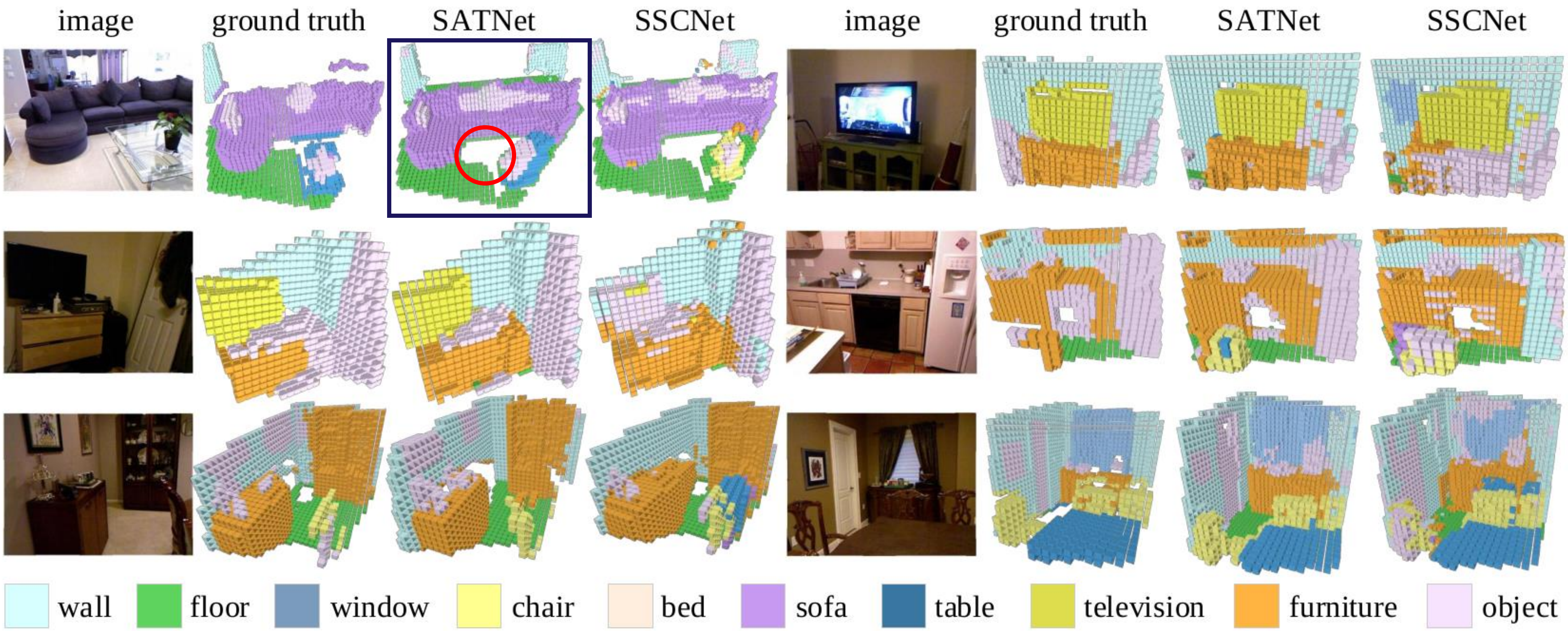
Problem Statement



Qualitative results on NYUv2 dataset from Liu *et al.* [70]

[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion. 2, 4, 45, 47, 52, 53, 58, 59

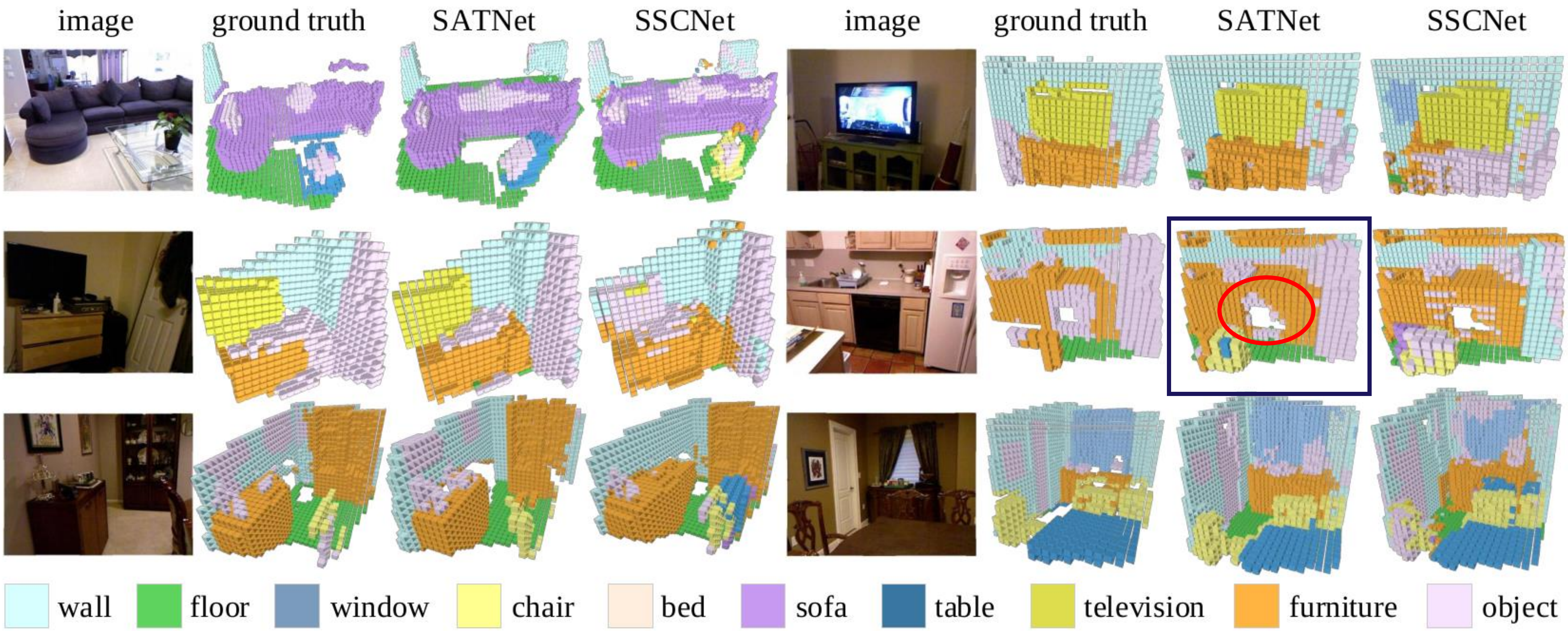
Problem Statement



Qualitative results on NYUv2 dataset from Liu *et al.* [70]

[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion. 2, 4, 45, 47, 52, 53, 58, 59

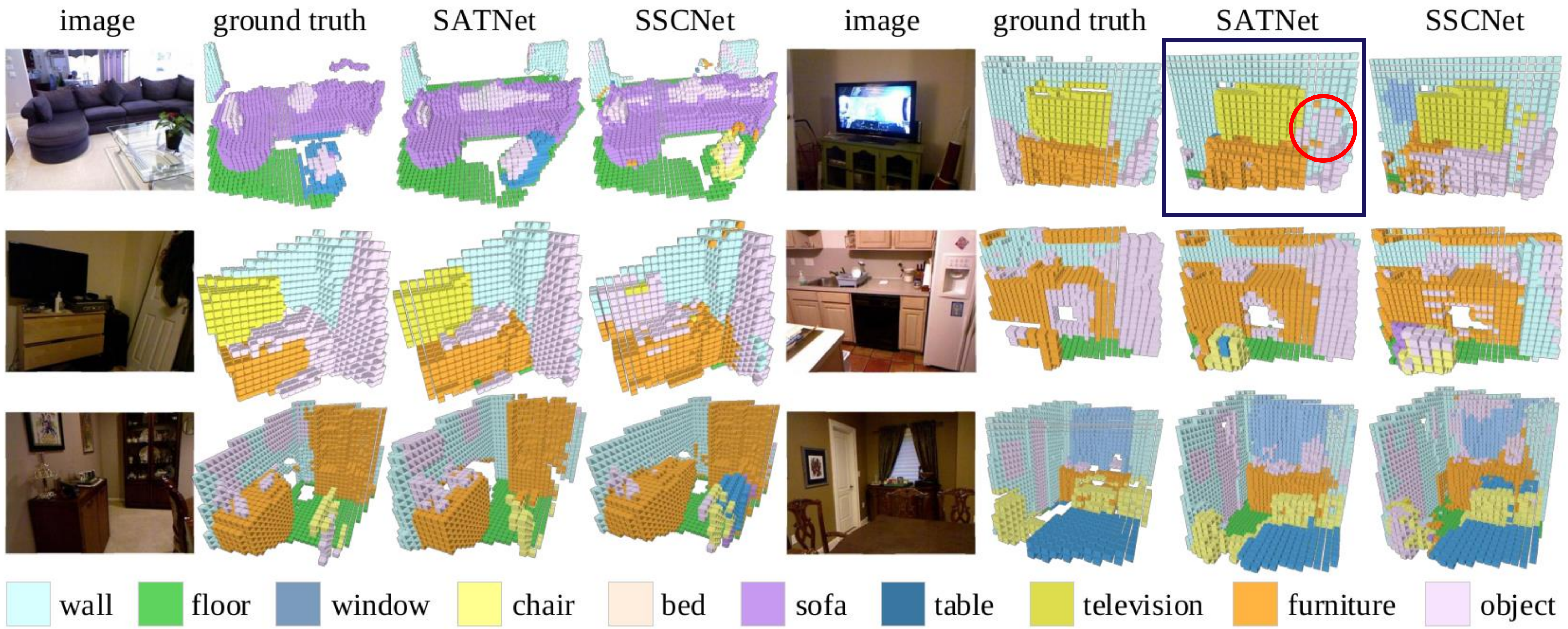
Problem Statement



Qualitative results on NYUv2 dataset from Liu *et al.* [70]

[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion. 2, 4, 45, 47, 52, 53, 58, 59

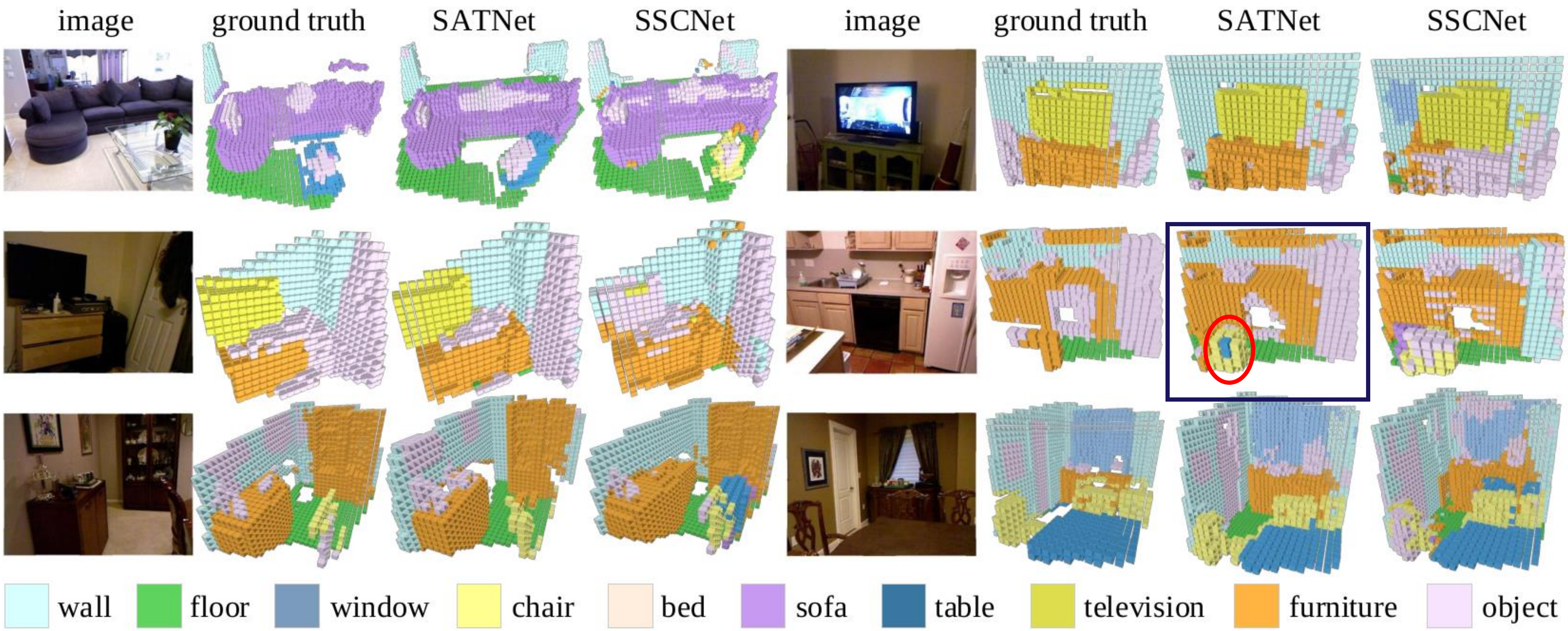
Problem Statement



Qualitative results on NYUv2 dataset from Liu *et al.* [70]

[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion. 2, 4, 45, 47, 52, 53, 58, 59

Problem Statement



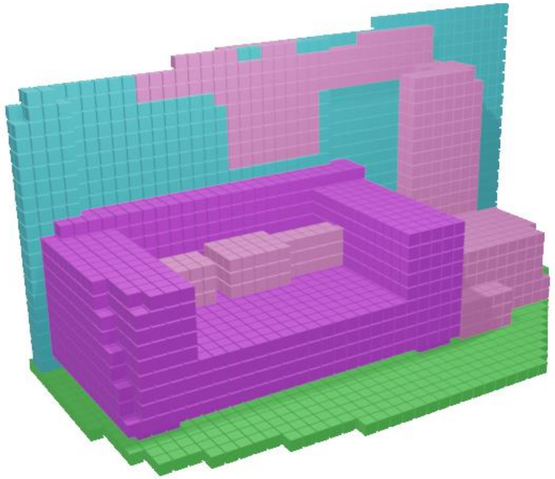
Qualitative results on NYUv2 dataset from Liu *et al.* [70]

[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion. 2, 4, 45, 47, 52, 53, 58, 59

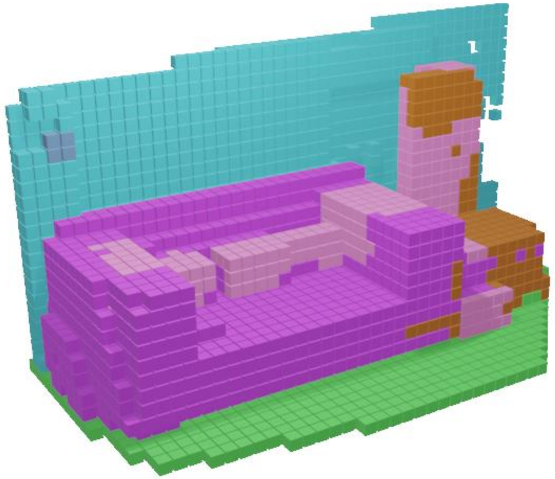
Problem Statement

- Two main deficiencies of current approaches:
  - the RGB part of the RGB-D image is not completely explored;
  - they are limited to the restricted FOV of depth sensors like Kinect

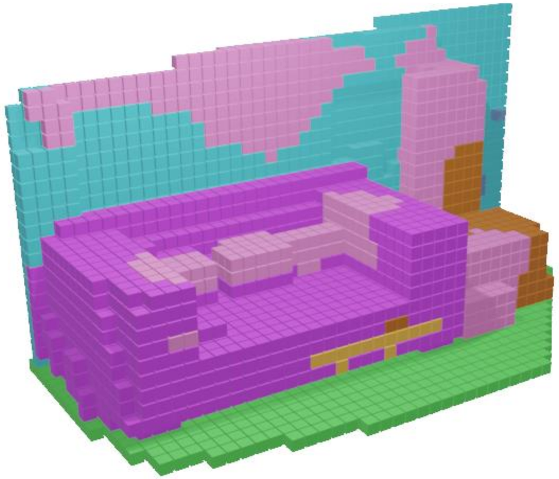
Improvements on regular SSC Datasets



Ground Truth



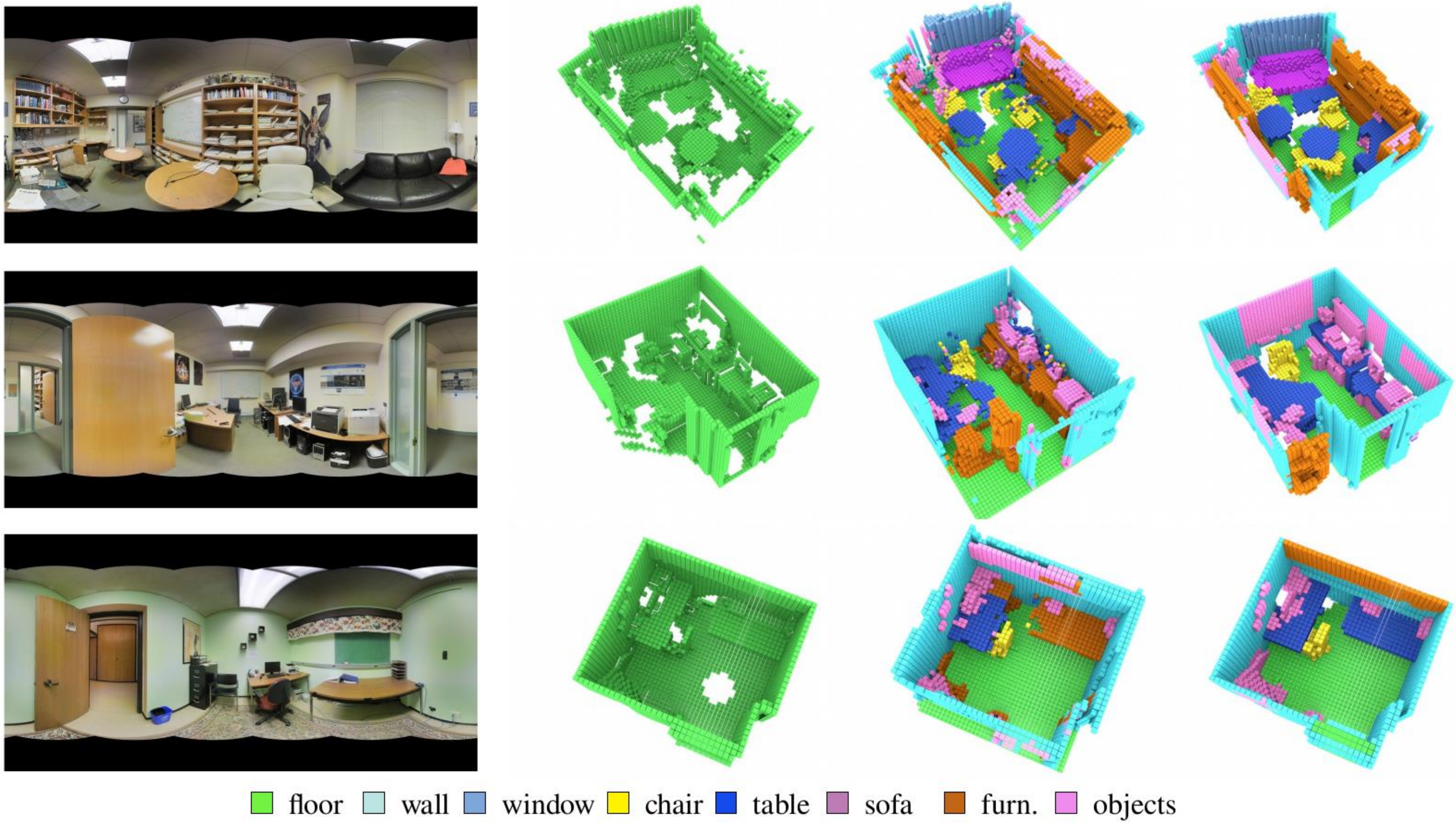
SSCNet



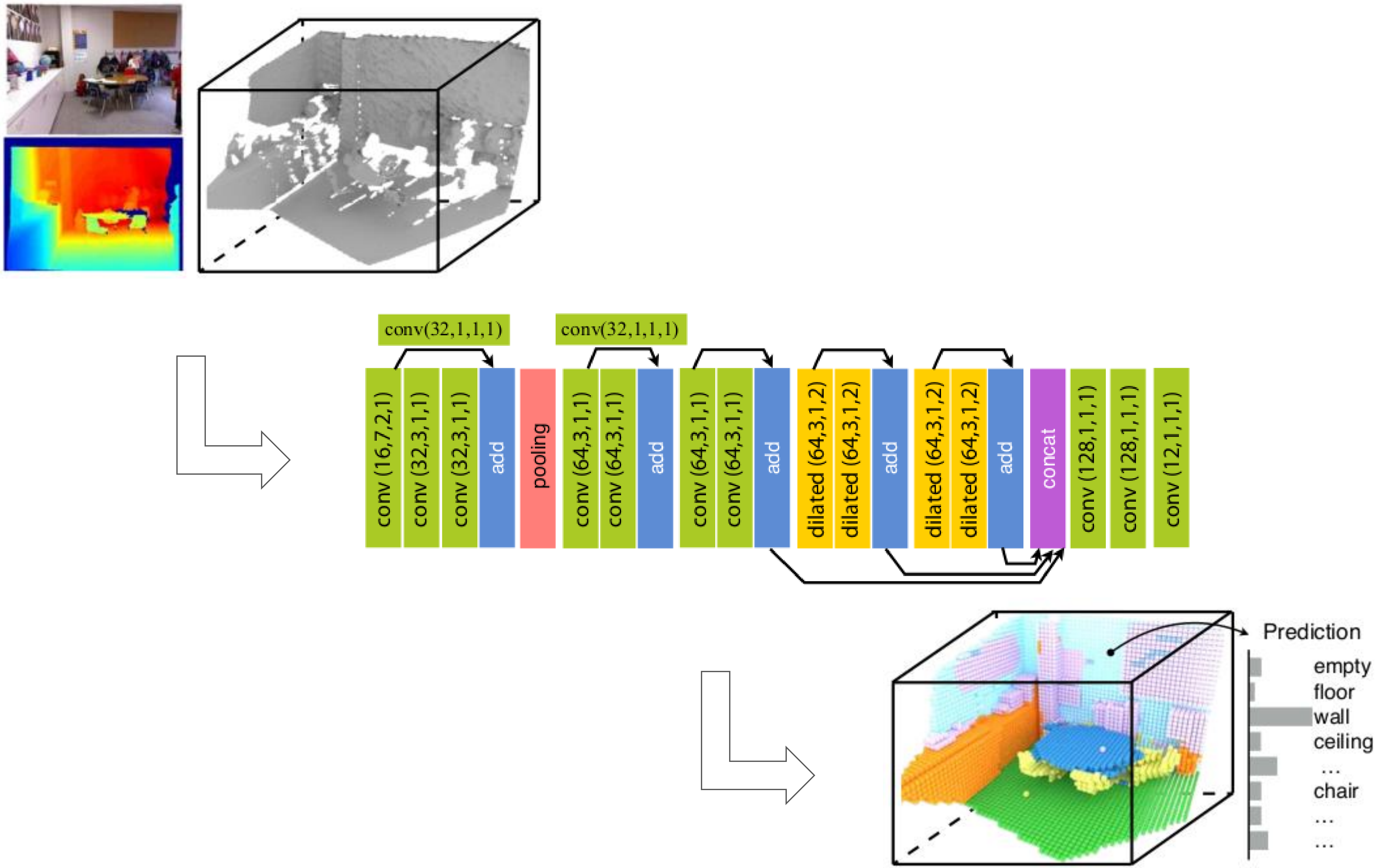
EdgeNet-MF

360 degree SSC

RGB Image | Input Volume | Predicted Volume | GT



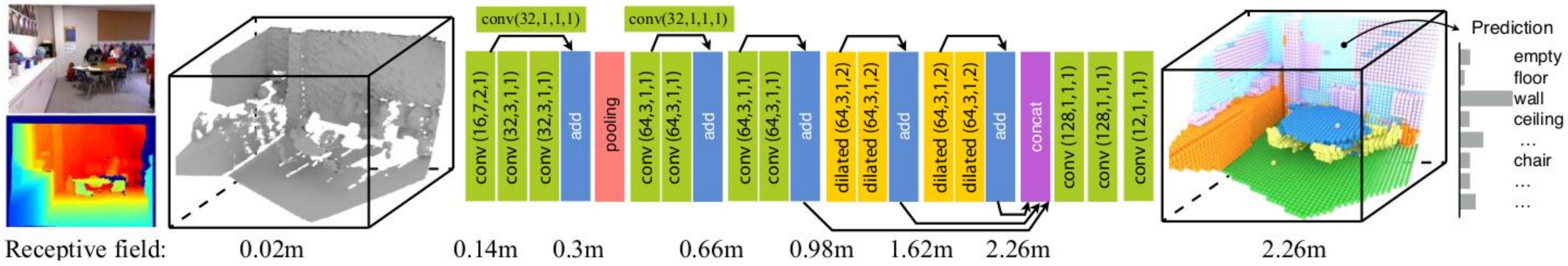
Previous Works

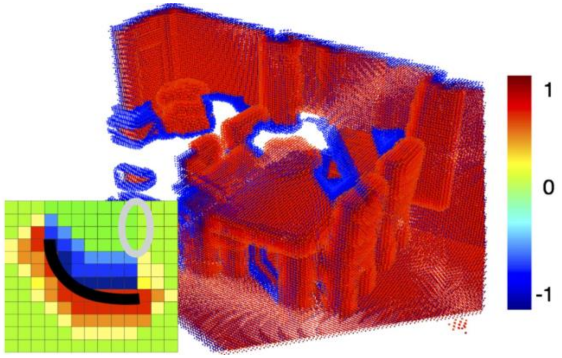


Previous Works

Depth maps only

- SSCNET: Song et al. [107]
  - Seminal paper
  - Proposed F-TSDF encoding
  - Dilated convolutions to favor the receptive field
  - Introduced SUNCG Dataset

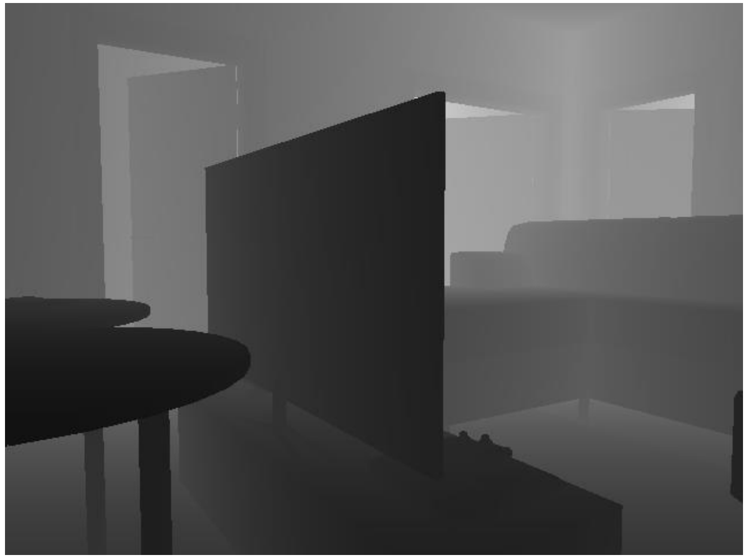




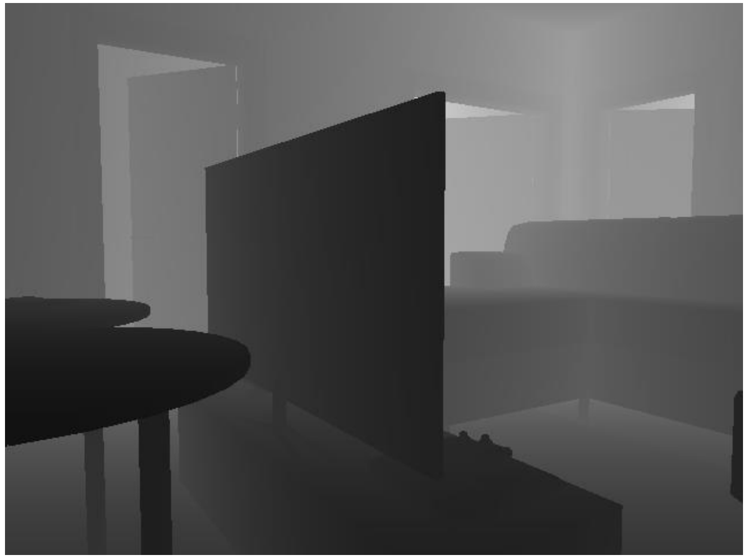
[107] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: Semantic Scene Completion from a Single Depth Image. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, pp. 190–198, Piscataway, NJ, July 2017. IEEE. 2, 3, 4, 18, 45, 46, 47, 51, 52, 53, 64, 68, 70

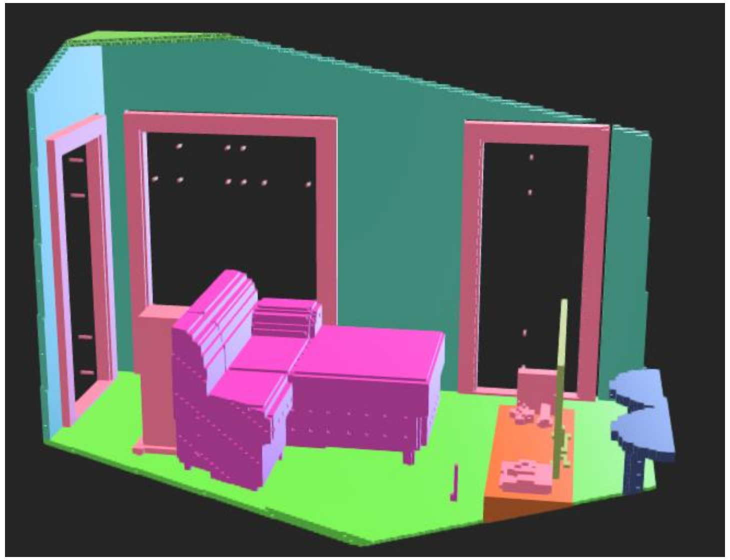
Why volumetric encoding is important?

Why volumetric encoding is important?

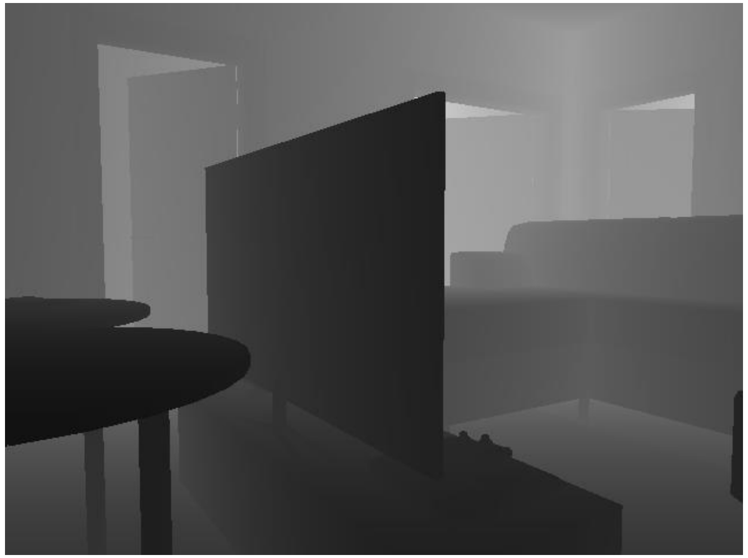


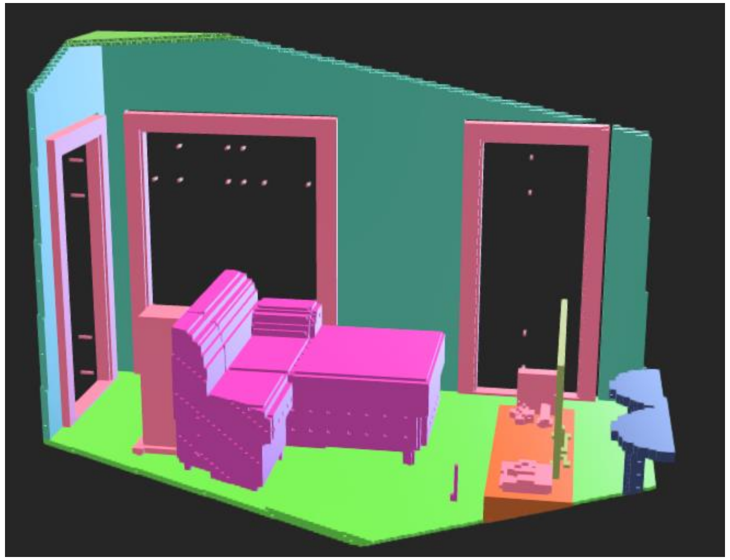
Why volumetric encoding is important?

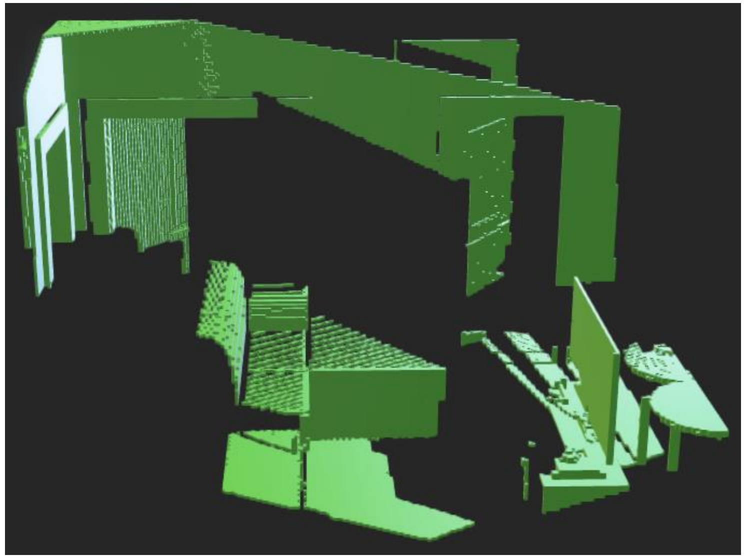




Why volumetric encoding is important?

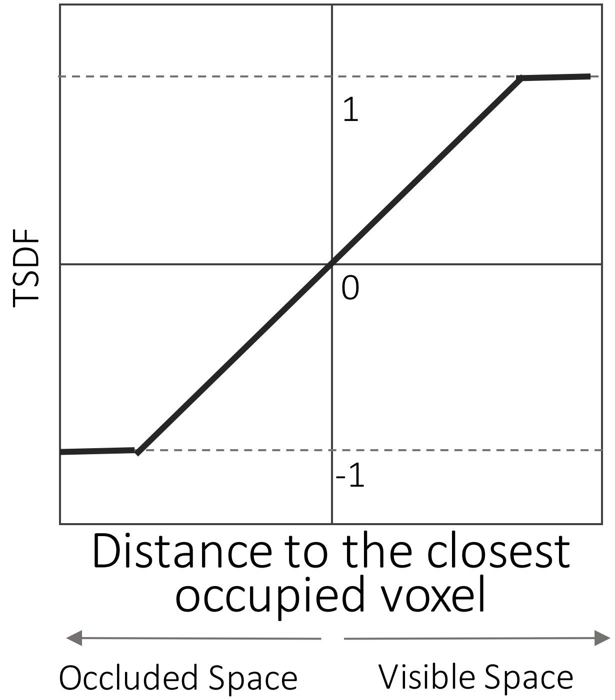


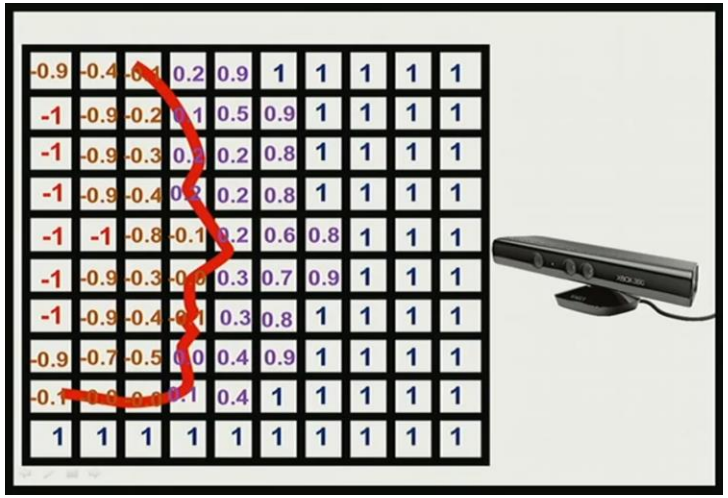


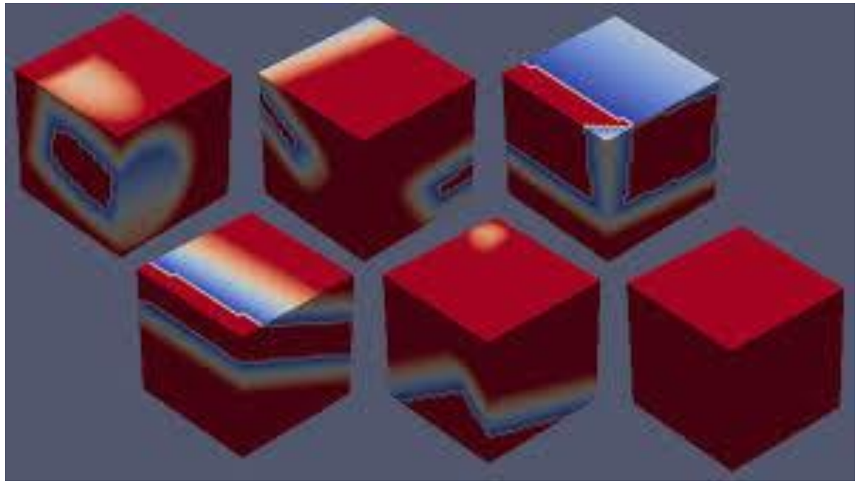


TSDF

- TSDF: Truncated Signed Distance Function



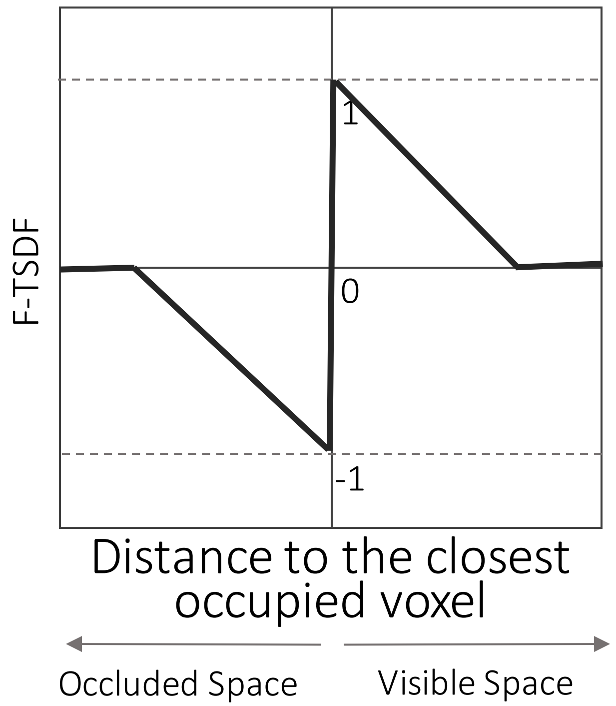


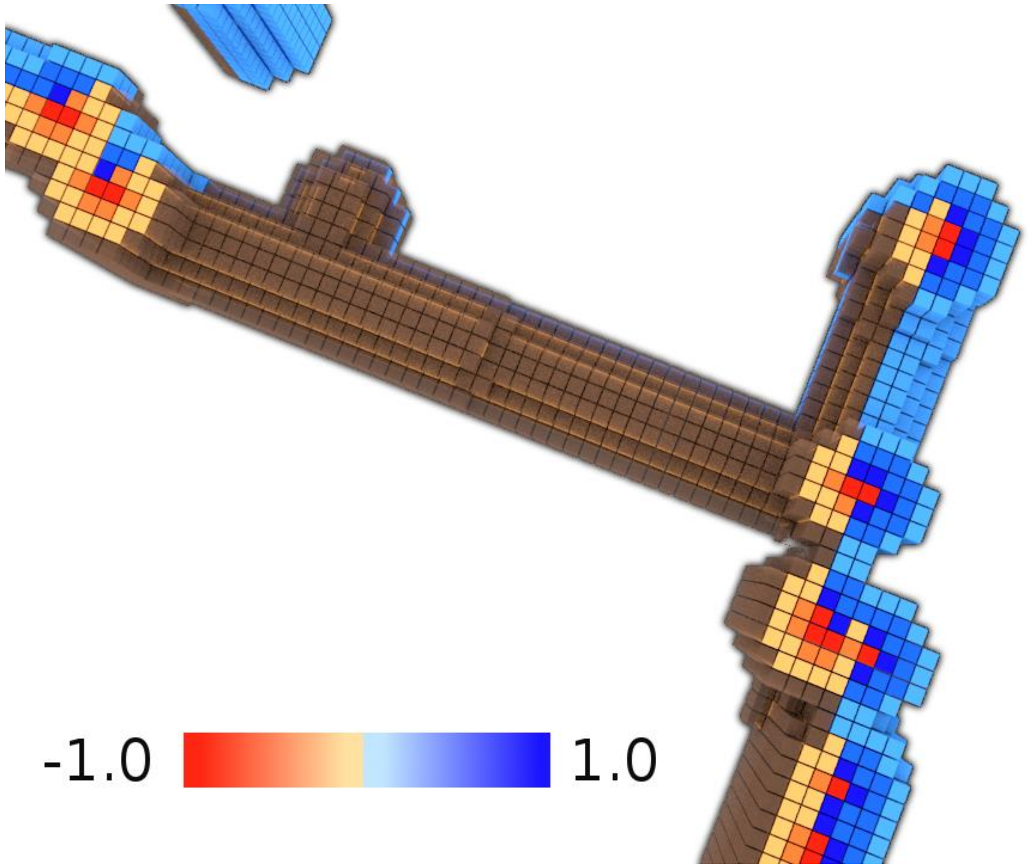


TSDF

F-TSDF

- F-TSDF: Flipped Truncated Signed Distance Function



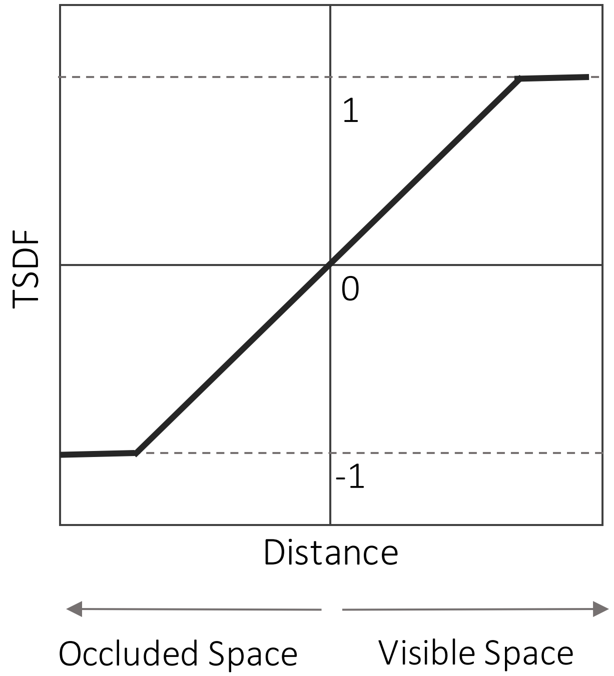


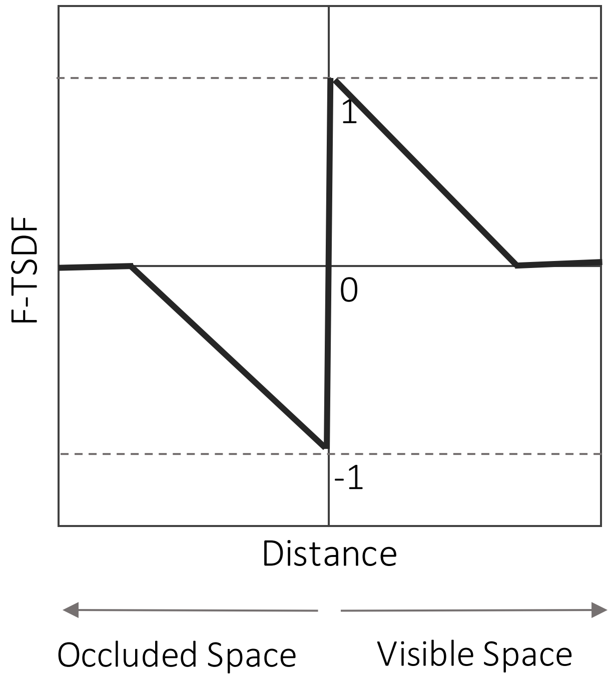
F-TSDF

F-TSDF = sign(TSDF) · (1-|TSDF|)

TSDF vs F-TSDF

- F-TSDF: Flipped Truncated Signed Distance Function



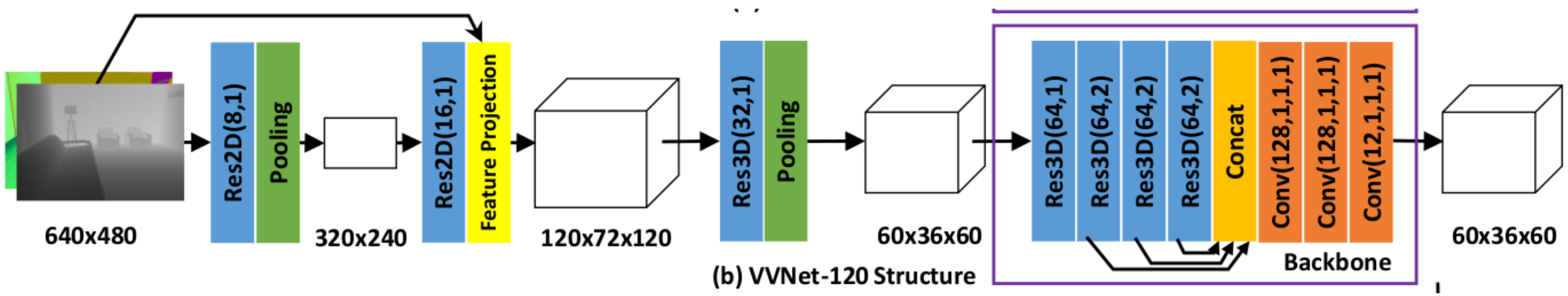


F-TSDF = sign(TSDF) · (1-|TSDF|)

Previous Works

Depth maps only

- Guo and Tong [40]:
  - 2D features projected to 3D

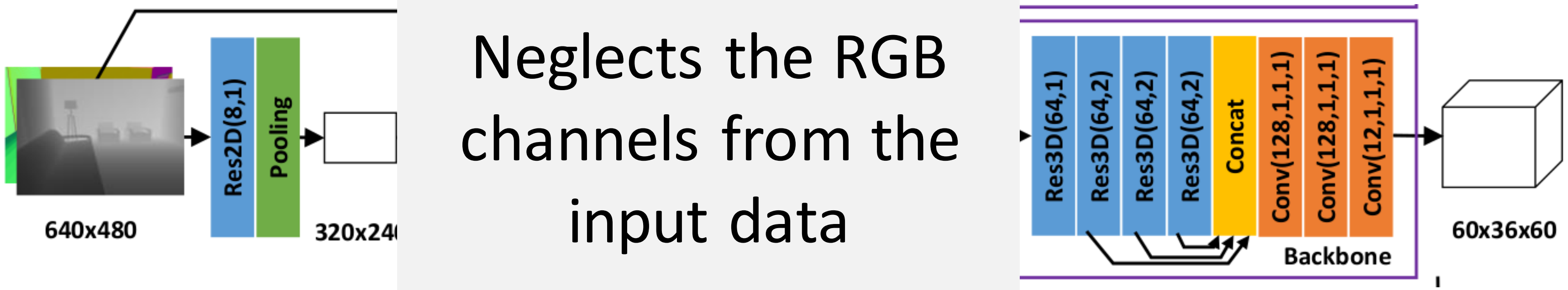


[40] Guo, Y. and Tong, X.: View-Volume Network for Semantic Scene Completion from a Single Depth Image. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 726–732, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization, ISBN 978-0-9992411-2-7. https://doi.org/10.24963/ijcai.2018/101. 2, 4, 18, 46, 52, 53

Previous Works

Depth maps only

- Guo and Tong [40]:
  - 2D features projected to 3D



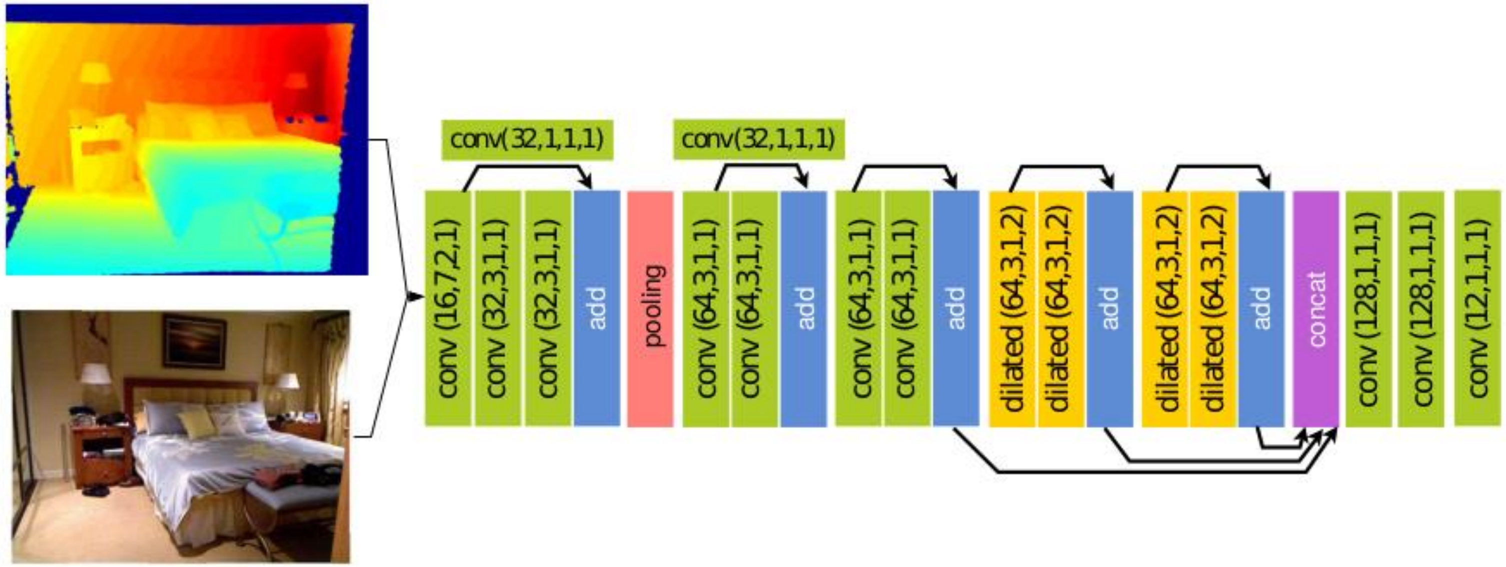
Neglects the RGB channels from the input data

[40] Guo, Y. and Tong, X.: View-Volume Network for Semantic Scene Completion from a Single Depth Image. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 726–732, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization, ISBN 978-0-9992411-2-7. https://doi.org/10.24963/ijcai.2018/101. 2, 4, 18, 46, 52, 53

Previous Works

Depth maps plus RGB

- Guedes *et al*.[38]

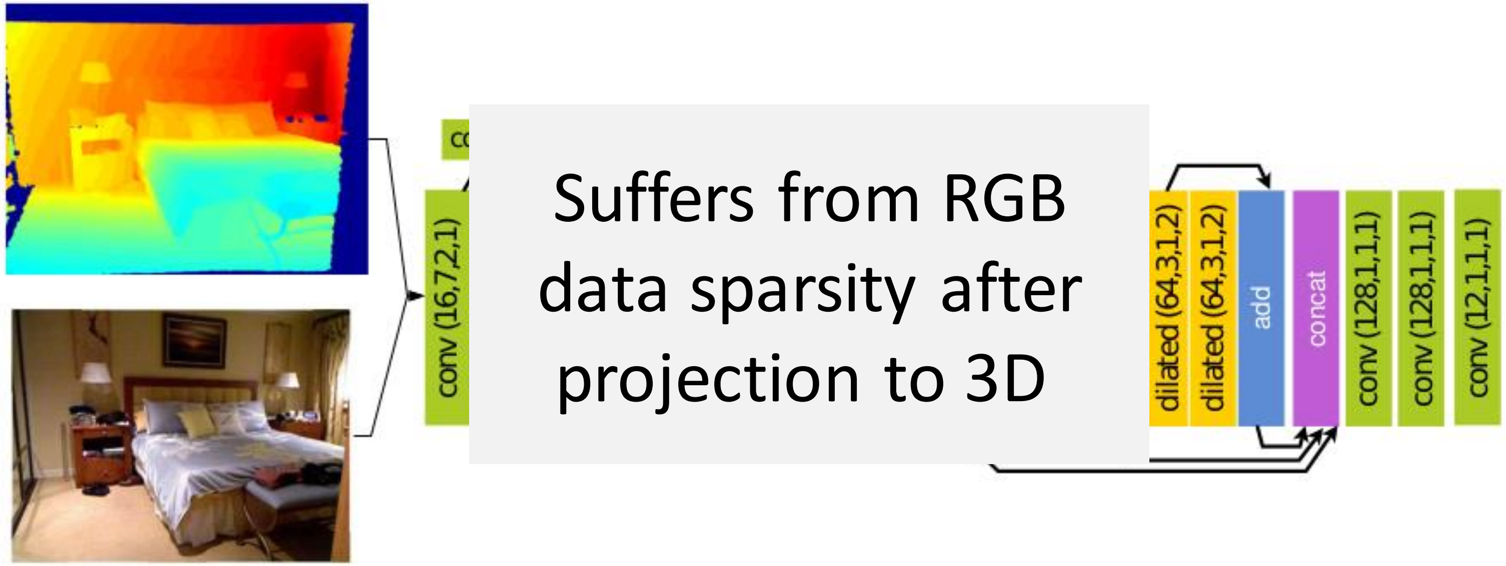


[38] Guedes, A.B.S., de Campos, T.E., and Hilton, A.: Semantic scene completion combining colour and depth: preliminary experiments. In ICCV workshop on 3D Reconstruction Meets Semantics (3DRMS), Venice, Italy, October 2017. Event webpage: http://trimbot2020.webhosting.rug.nl/events/events-2017/3drms/. Also published at arXiv:1802.04735. 4, 45, 46, 47, 52, 53

Previous Works

Depth maps plus RGB

- Guedes *et al.*[38]



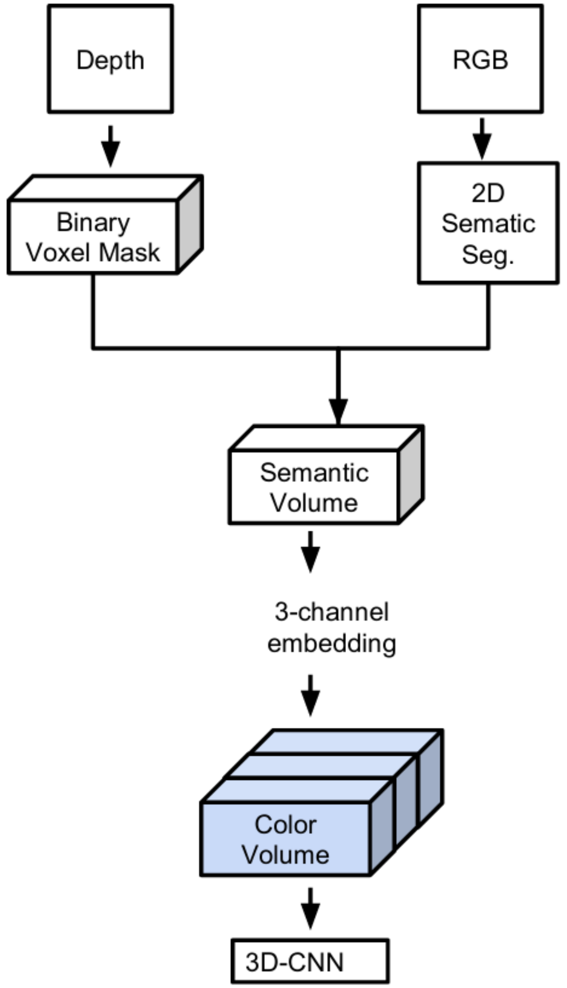
Suffers from RGB data sparsity after projection to 3D

[38] Guedes, A.B.S., de Campos, T.E., and Hilton, A.: Semantic scene completion combining colour and depth: preliminary experiments. In ICCV workshop on 3D Reconstruction Meets Semantics (3DRMS), Venice, Italy, October 2017. Event webpage: http://trimbot2020.webhosting.rug.nl/events/events-2017/3drms/. Also published at arXiv:1802.04735. 4, 45, 46, 47, 52, 53

Previous Works

Depth map plus 2D segmentation

- Two stream 3D semantic scene completion: Garbade *et al*.[36]

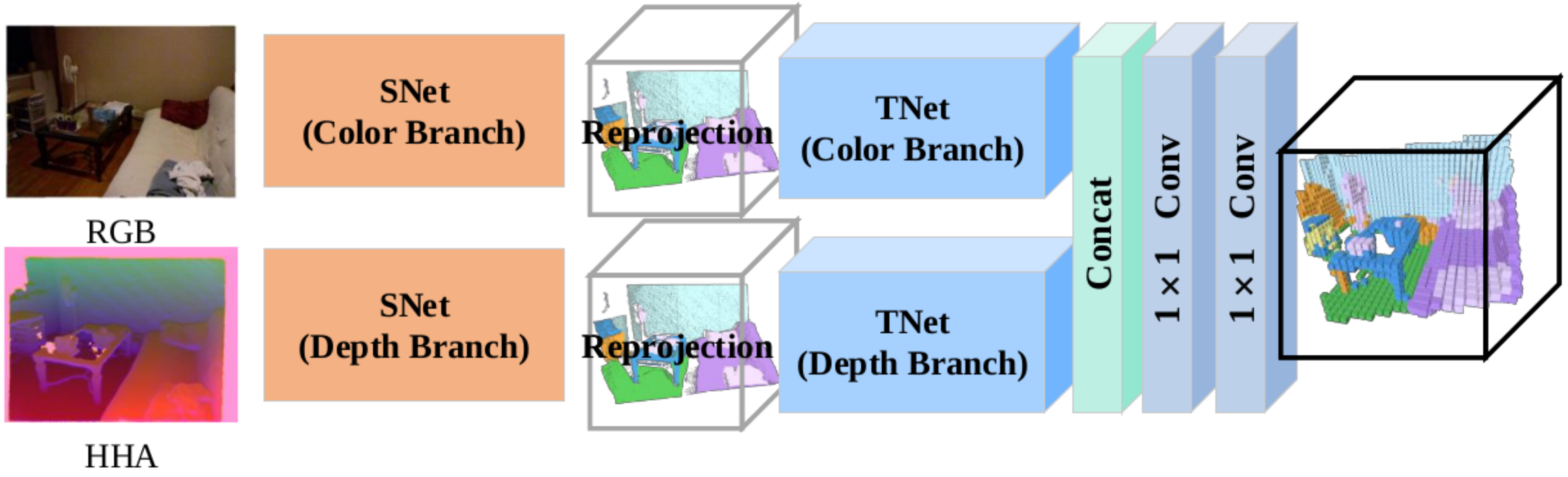


[36] Garbade, M., Sawatzky, J., Richard, A., and Gall, J.: Two stream 3D semantic scene completion. Tech. Rep. arXiv:1804.03550, Cornell University Library, 2018. http://arxiv.org/abs/1804.03550. 4, 45, 47, 52, 53

Previous Works

Depth map plus 2D segmentation

- TNetFusion: Liu *et al.*[70]

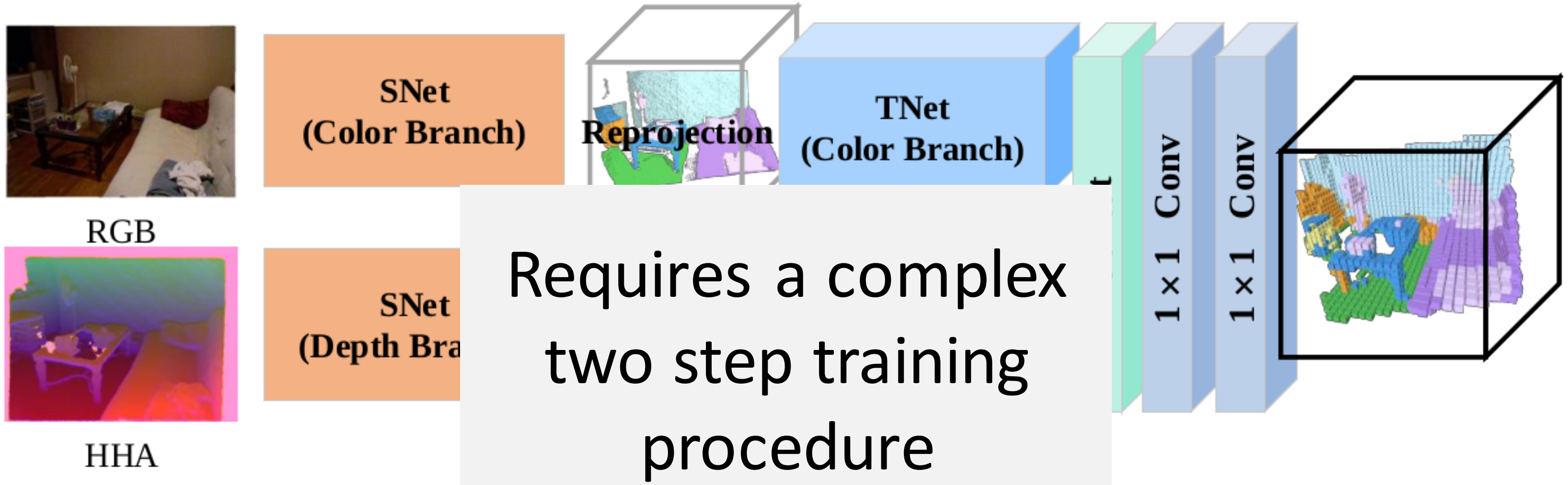


[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion. 2, 4, 45, 47, 52, 53, 58, 59

Previous Works

Depth map plus 2D segmentation

- TNetFusion: Liu *et al.*[70]

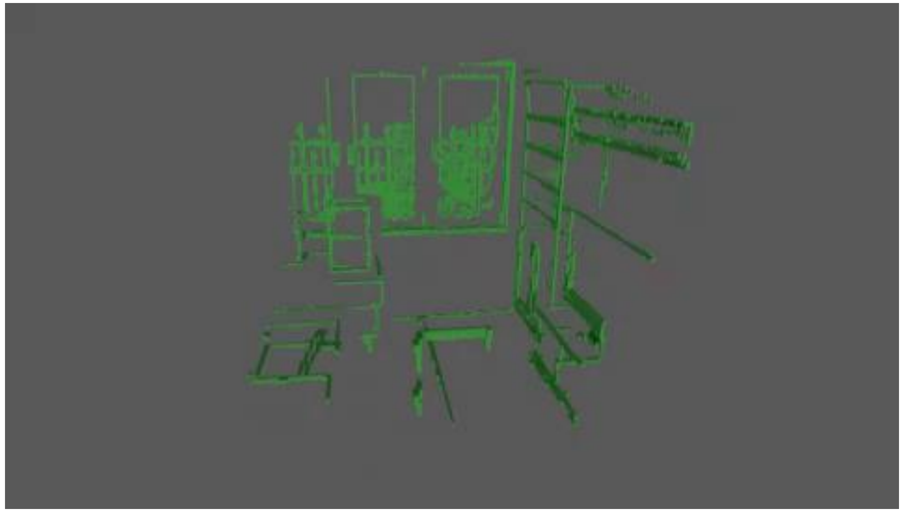


Requires a complex two step training procedure

[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: See and think: Disentangling semantic scene completion. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): Procedings of Conference on Neural Information Processing Systems 31 (NIPS), pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion. 2, 4, 45, 47, 52, 53, 58, 59

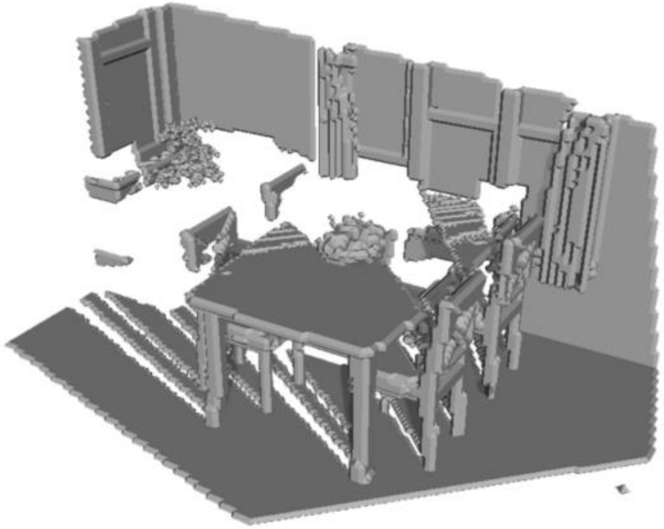
Using RGB Edges to improve Semantic Scene Completion from RGB-D Images

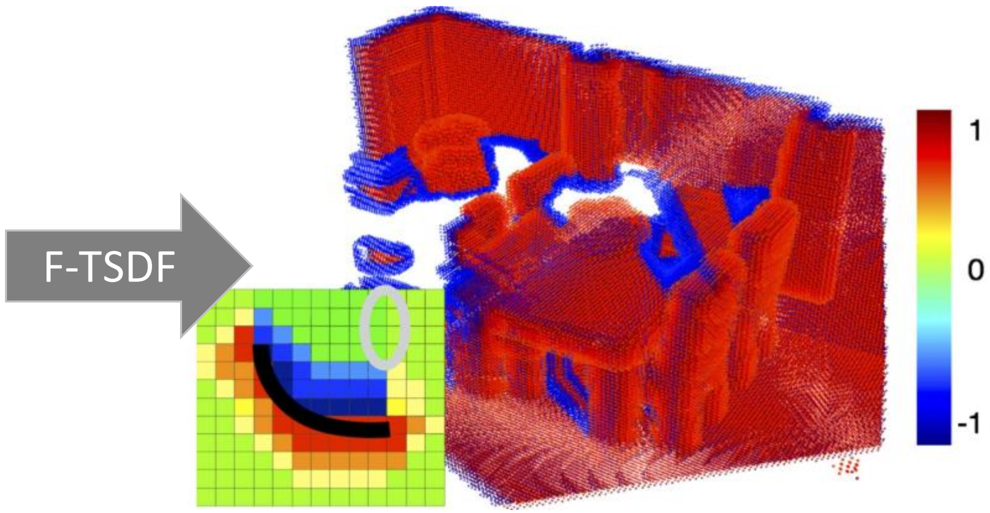
EdgeNet



F-TSDF and the RGB Volume

- It is possible to apply F-TSDF to the occupancy volume

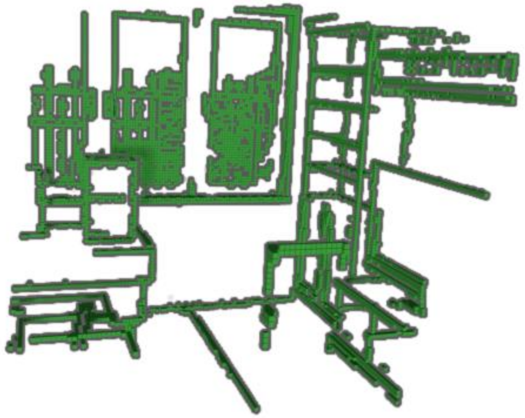


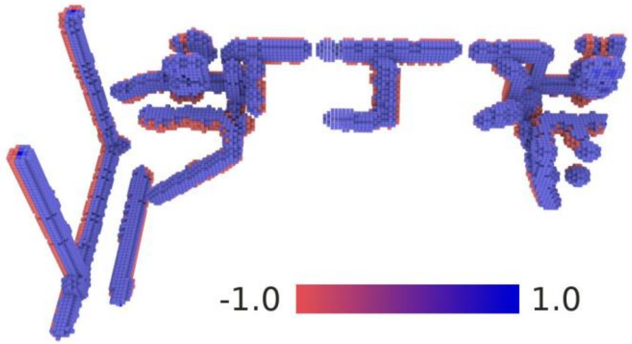


- However, RGB data is not binary!

Our Approach: EdgeNet

- We extract information from RGB data using Canny Edge detector before F-TSDF

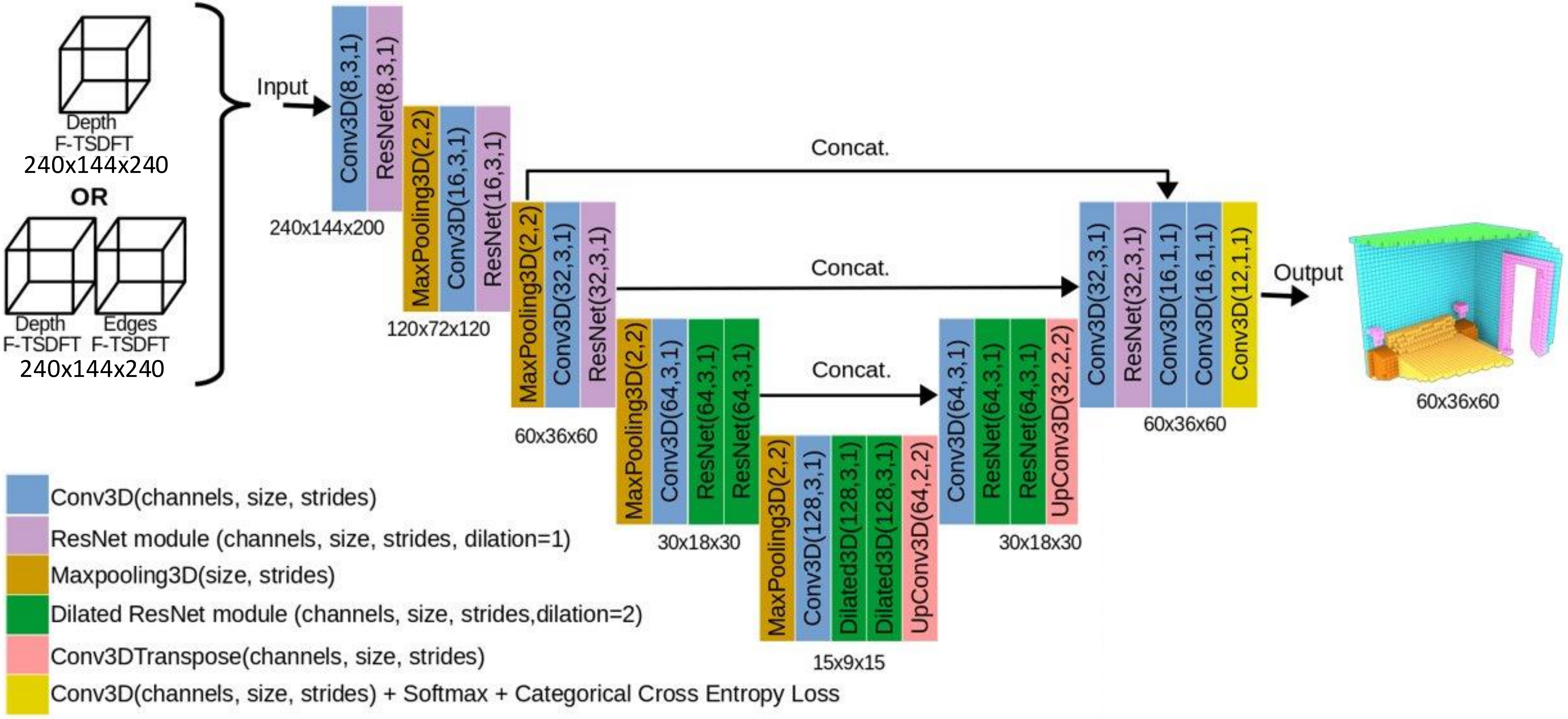




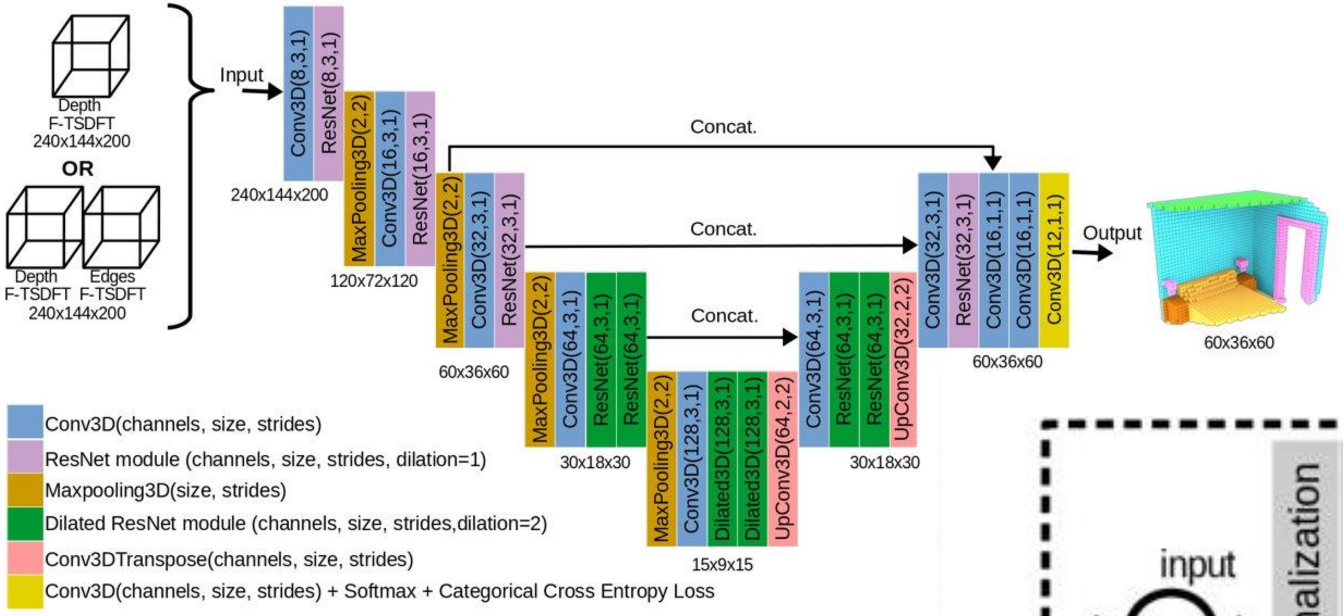
Our implementation

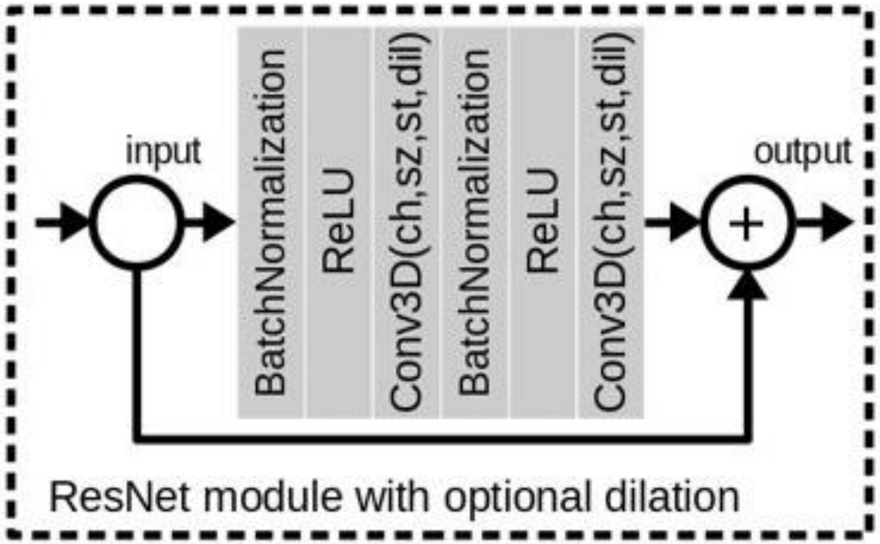
- Offline F-TSDF calculation using portable C++ CUDA code
- We provide a software interface between CUDA and Python
- Preprocessing code is independent from the deep learning framework

Network Architecture

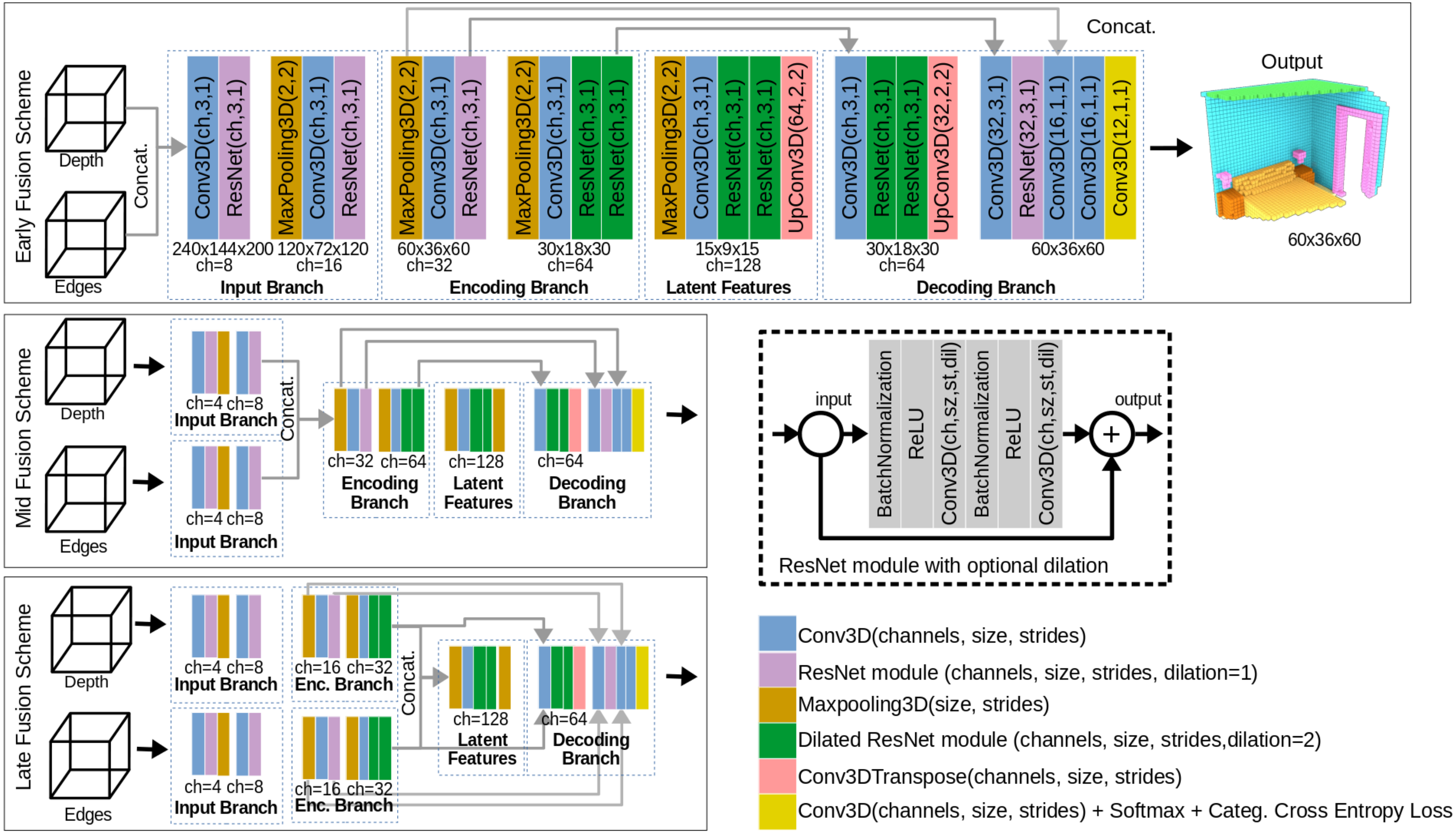


Network Architecture

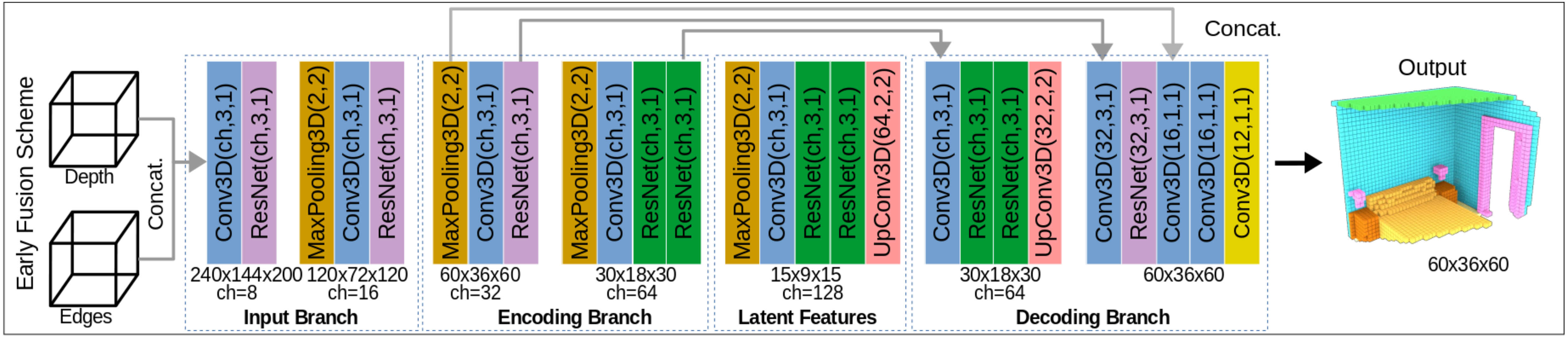




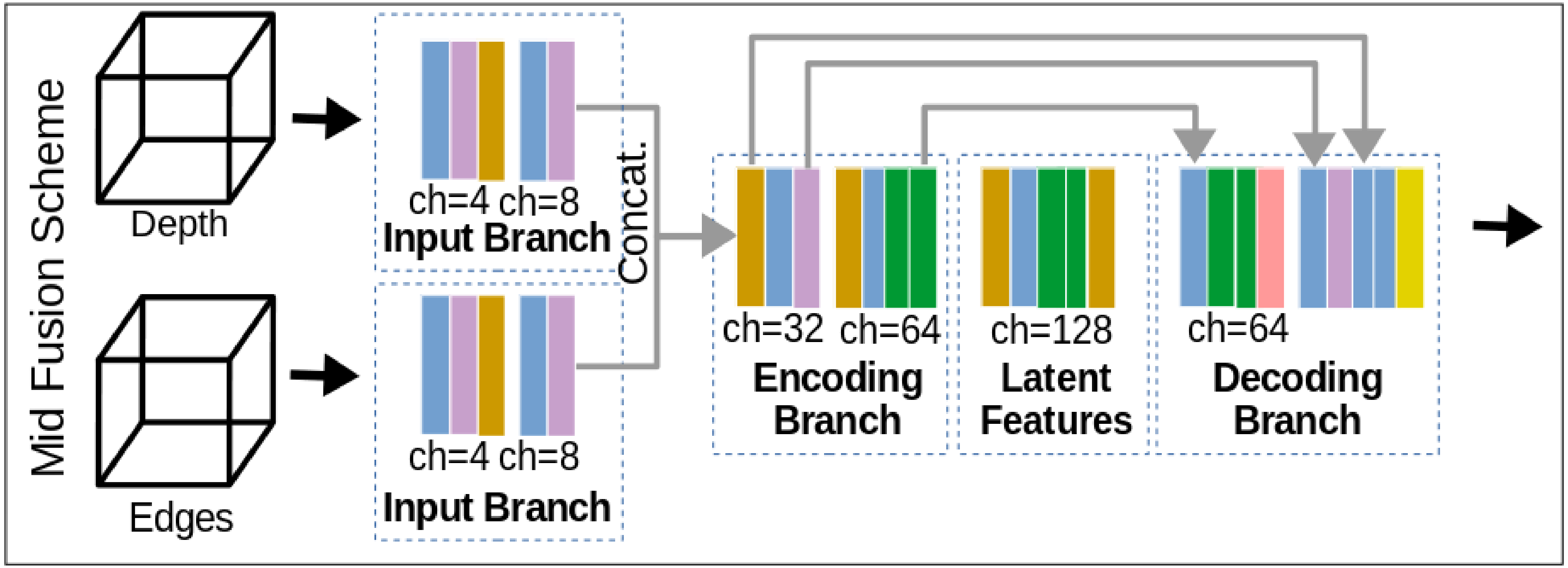
Network Architecture - Fusion Schemes



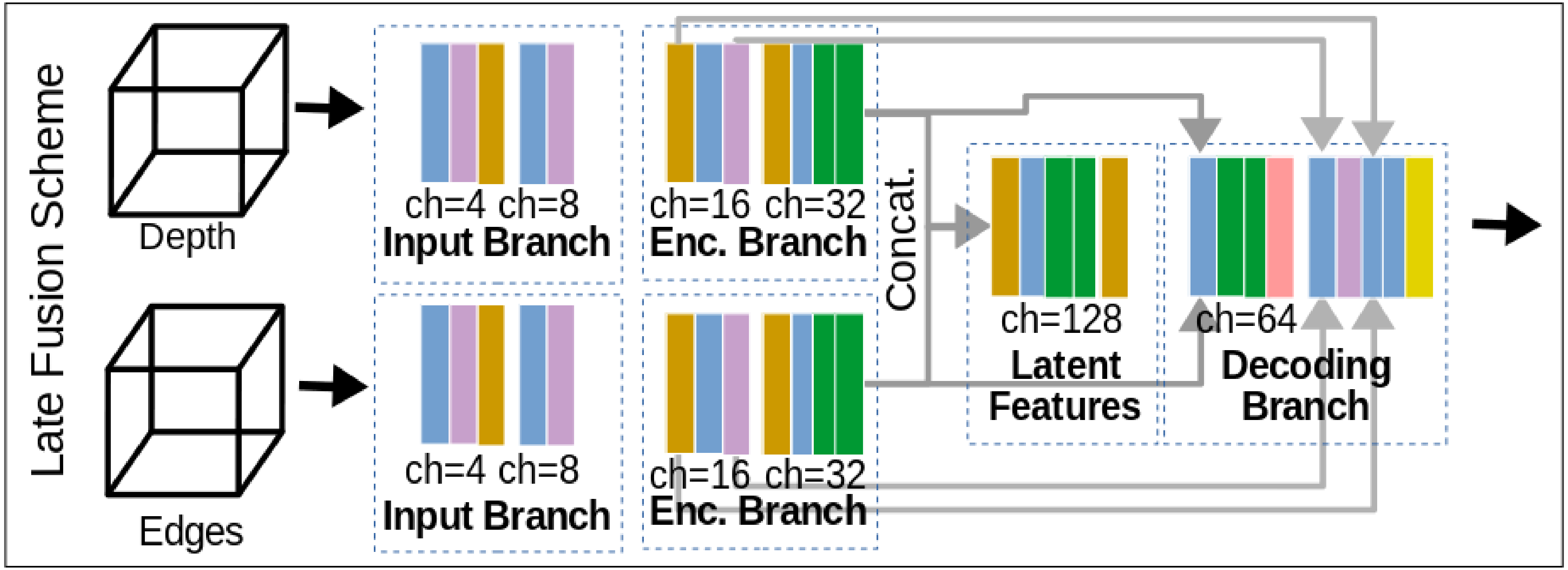
Network Architecture - Fusion Schemes



Network Architecture - Fusion Schemes



Network Architecture - Fusion Schemes



Datasets

- SUNCG*

(a) SUNCG dataset (b) 3D Scene (c) Synthetic depth and volumetric ground truth

- NYUDv2**

*Song *et al*.[107]

**Silberman *et al*.[102]

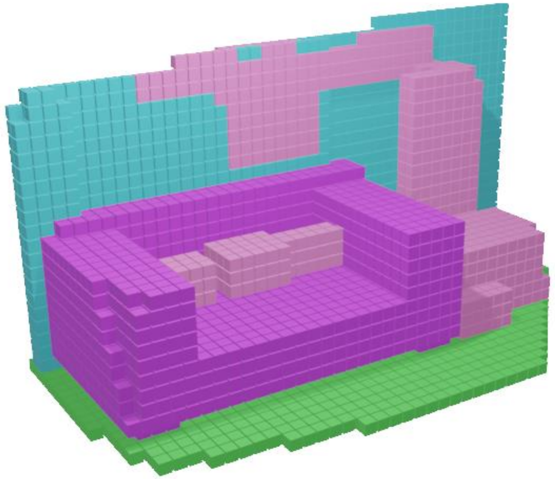
Training Time

- Ours
  - SUNCG: 4 days
  - NYU: 6 hours
- SSCNET
  - SUNCG: 7 days
  - NYU: 30 hours

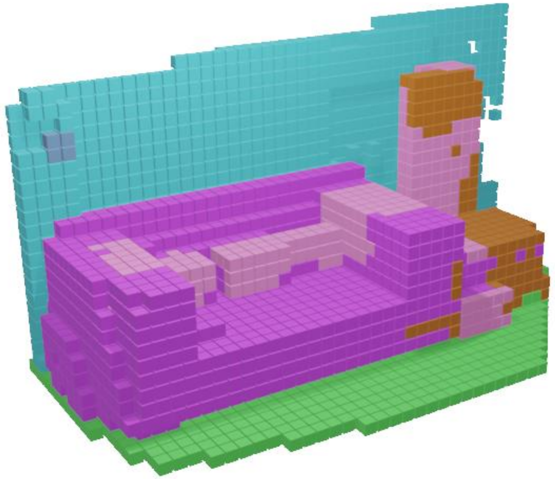
Quantitative Results

- New state-of-the-art result on SUNCG
- All new aspects of our solution contributed to the improvement
- Middle Fusion and Late Fusion schemes presented similar results on SUNCG
- Middle Fusion presented better results on NYUDV2

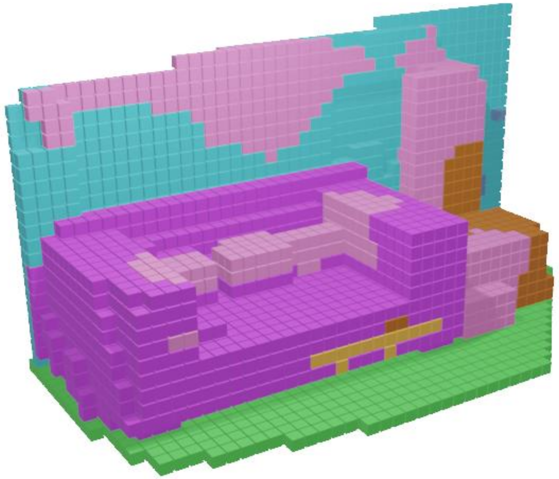
Qualitative Results



Ground Truth

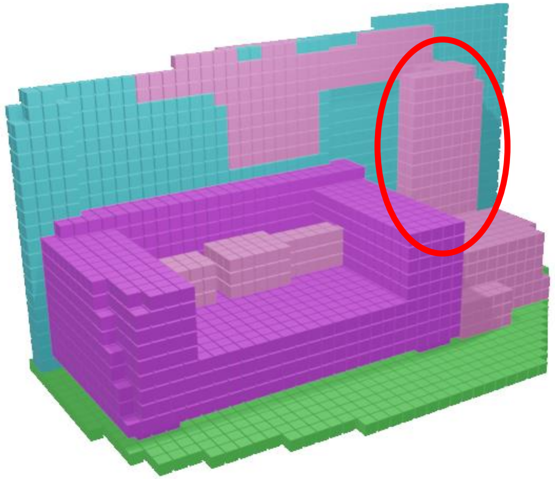


SSCNet

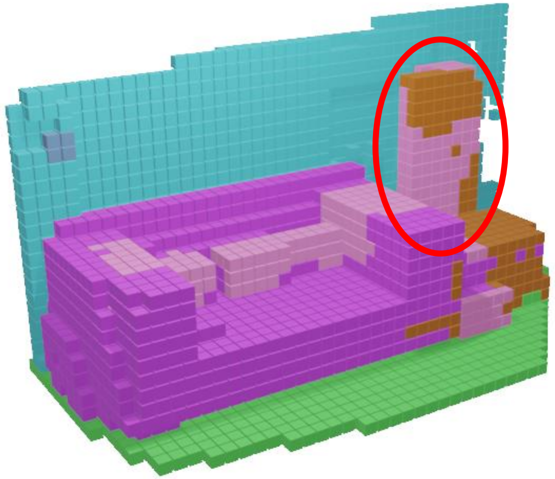


EdgeNet-MF

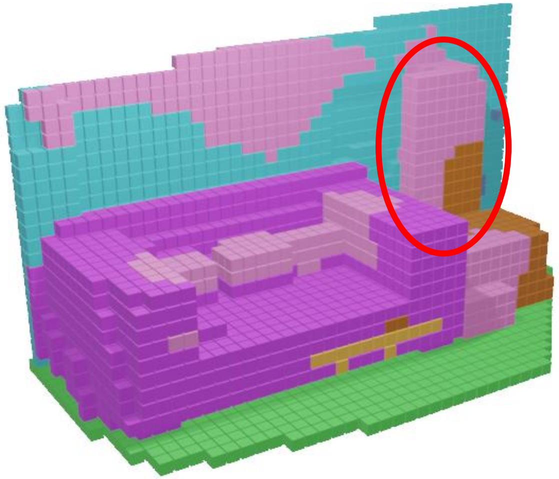
Qualitative Results



Ground Truth



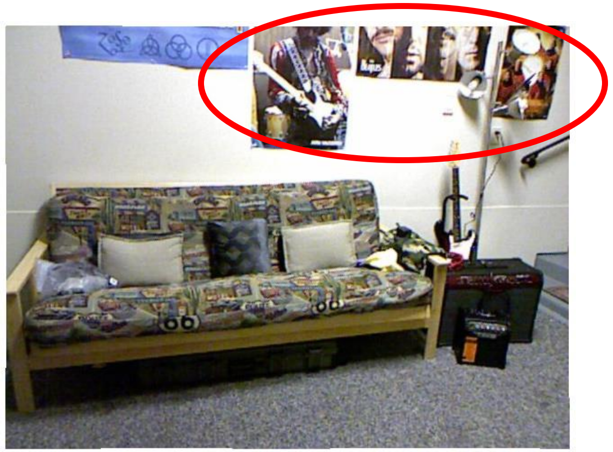
SSCNet

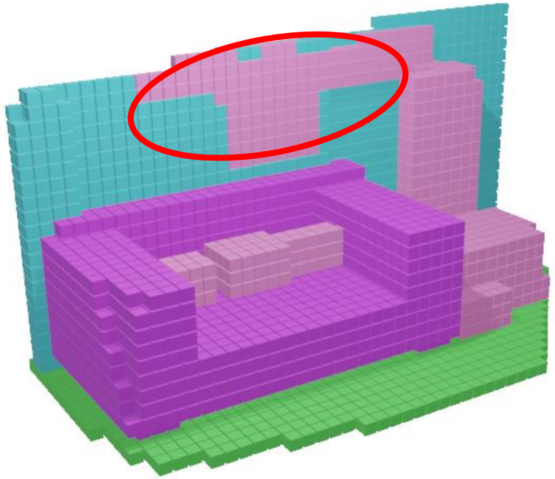


EdgeNet-MF

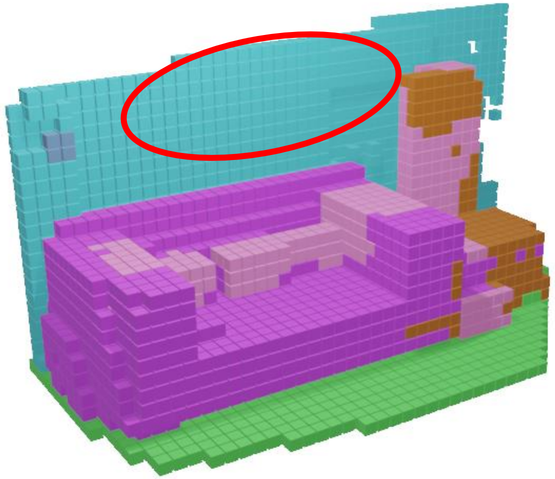
Higher overall accuracy

Qualitative Results

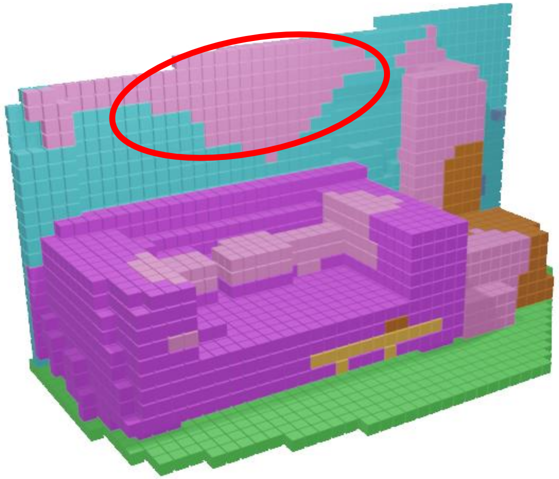




Ground Truth



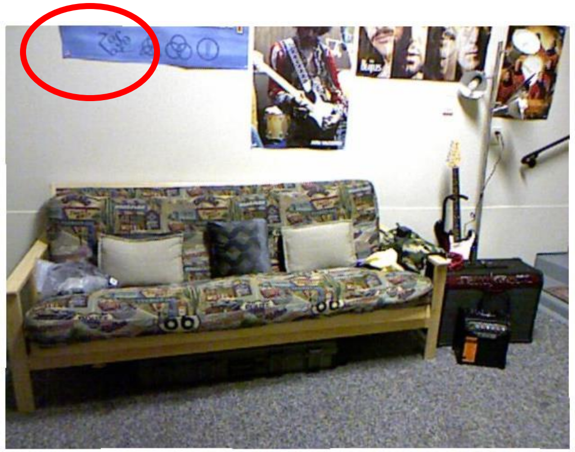
SSCNet

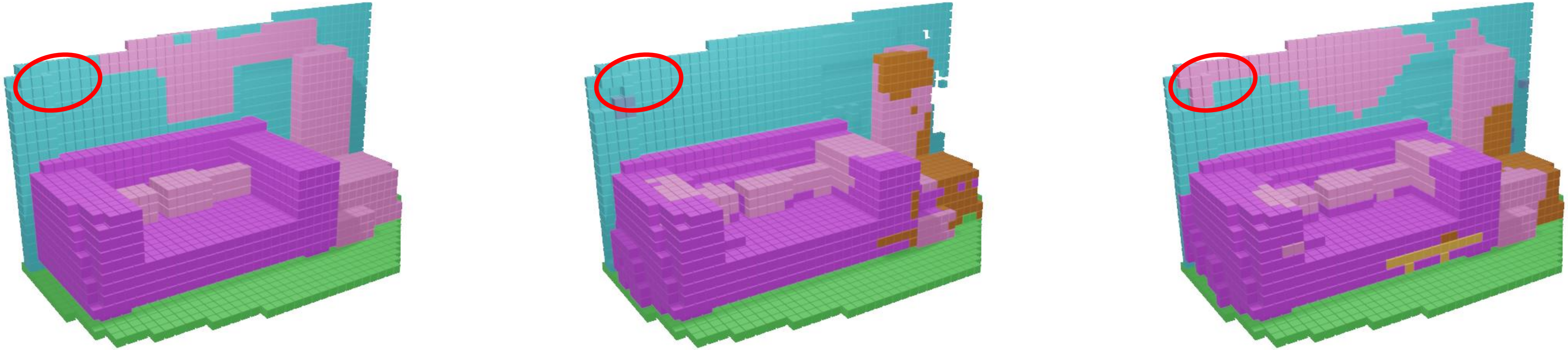


EdgeNet-MF

Hard-to-detect classes

Qualitative Results





Ground Truth SSCNet EdgeNet-MF

NYU Ground Truth errors

Conclusions

- A new end-to-end network architecture
- A new RGB enconding strategy
- Visually perceptible improvements
- Improvement over the state-of-the-art result on SUNCG
- We surpased other end-to-end approaches on NYUv2
- An efficient and lightweight training pipeline for the task

Publication

EdgeNet: Sematic Scene Completion from a Single RGB-D Image

EdgeNet: Semantic Scene Completion from a Single RGB-D Image

Aloisio Dourado, Teofilo Emidio de Campos
University of Brasilia
Brasilia, Brazil
aloisio.dourado.bh@gmail.com, t.decampos@st-annes.oxon.org

Hansung Kim, Adrian Hilton
University of Surrey
Surrey, UK
{h.kim, a.hilton}@surrey.ac.uk

***Abstract*—Semantic scene completion is the task of predicting a complete 3D representation of volumetric occupancy with corresponding semantic labels for a scene from a single point of view. In this paper, we present EdgeNet, a new end-to-end neural network architecture that fuses information from depth and RGB, explicitly representing RGB edges in 3D space. Previous works on this task used either depth-only or depth with colour by projecting 2D semantic labels generated by a 2D segmentation network into the 3D volume, requiring a two step training process. Our EdgeNet representation encodes colour information in 3D space using edge detection and flipped truncated signed distance, which improves semantic completion scores especially in hard to detect classes. We achieved state-of-the-art scores on both synthetic and real datasets with a simpler and a more computationally efficient training pipeline than competing approaches.**

I. Introduction

The ability of reasoning about scenes in 3D is a natural task for humans, but remains a challenging problem in Computer Vision [1]. Knowing the complete 3D geometry of a scene and the semantic labels of each 3D voxel has many practical applications, like robotics and autonomous navigation in indoor environments, surveillance, assistive computing and augmented reality.

Currently available low cost RGB-D sensors generate data form a single viewing position and cannot handle occlusion among objects in the scene. For instance, in the scene depicted on the left part of Figure 1, parts of the wall, floor and furniture are occluded by the bed. There is also self-occlusion: the interior of the bed, its sides and its rear surfaces are hidden by the visible surface.

Given a partial 3D scene model acquired from a single RGB-D image, the goal of scene completion is to generate a complete 3D volumetric representation where each voxel is labelled as occupied by some object or free space. For occupied voxels, the goal of *semantic* scene completion is to assign a label that indicates to which class of object it belongs, as illustrated on the right part of Figure 1.

Before 2018, most of the work on scene reasoning only partially addressees this problem. A number of approaches only infer labels of the visible surfaces [2], [3], [4], while others only consider completing the occluded part of the scene, without semantic labelling [5]. Another line of work focuses on single objects, without the scene context [6].

The term semantic scene completion was introduced by Song *et al.* [7], who showed that scene completion and semantic labelling are intertwined and training a CNN to jointly deals with both tasks can lead to better results. Their approach only uses depth information, ignoring all information from RGB channels. Colour information is expected to be useful to distinguish objects that approximately share the same plane in the 3D space, and thus, are hard to be distinguished using only depth. Examples of such instances are flat objects attached to the wall, such as posters, paintings and flat TVs. Some types of closed doors and windows are also problematic for depth-only approaches.

Recent research also explored colour information from on RGB-D images to improve semantic scene completion scores. Some methods project colour information to 3D in a naive way, leading to a problem of data sparsity in the voxelised data that is fed to the 3D CNN [8], while others uses RGB information to train a 2D segmentation network and then project generated features to 3D, requiring a complex two step training process [9], [10].

Our work focuses on enhancing semantic scene segmentation scores using information from both depth and colour of RGB-D images in an end-to-end manner. In order to address the RGB data sparsity issue, we introduce a new strategy for encoding information extracted from RGB image in 3D space. We also present a new end-to-end 3D CNN architecture to combine and represent the features from colour and depth. Comprehensive experiments are conducted to evaluate the main aspects of the proposed solution. Results show that our fusion approach can enhance results of depth-only solutions and that EdgeNet achieves equivalent performance to current state-of-the-art fusion approach, with a much simpler training protocol.

To summarise, our main contributions are:

- EdgeNet, a new end-to-end CNN architecture that fuses depth, RGB edge information to achieve state-of-the-art performance in semantic scene completion with a much simpler approach;
- a new 3D volumetric edge representation using flipped signed-distance functions which improves performance and unifies data agregation for semantic scene completion from RGBD;

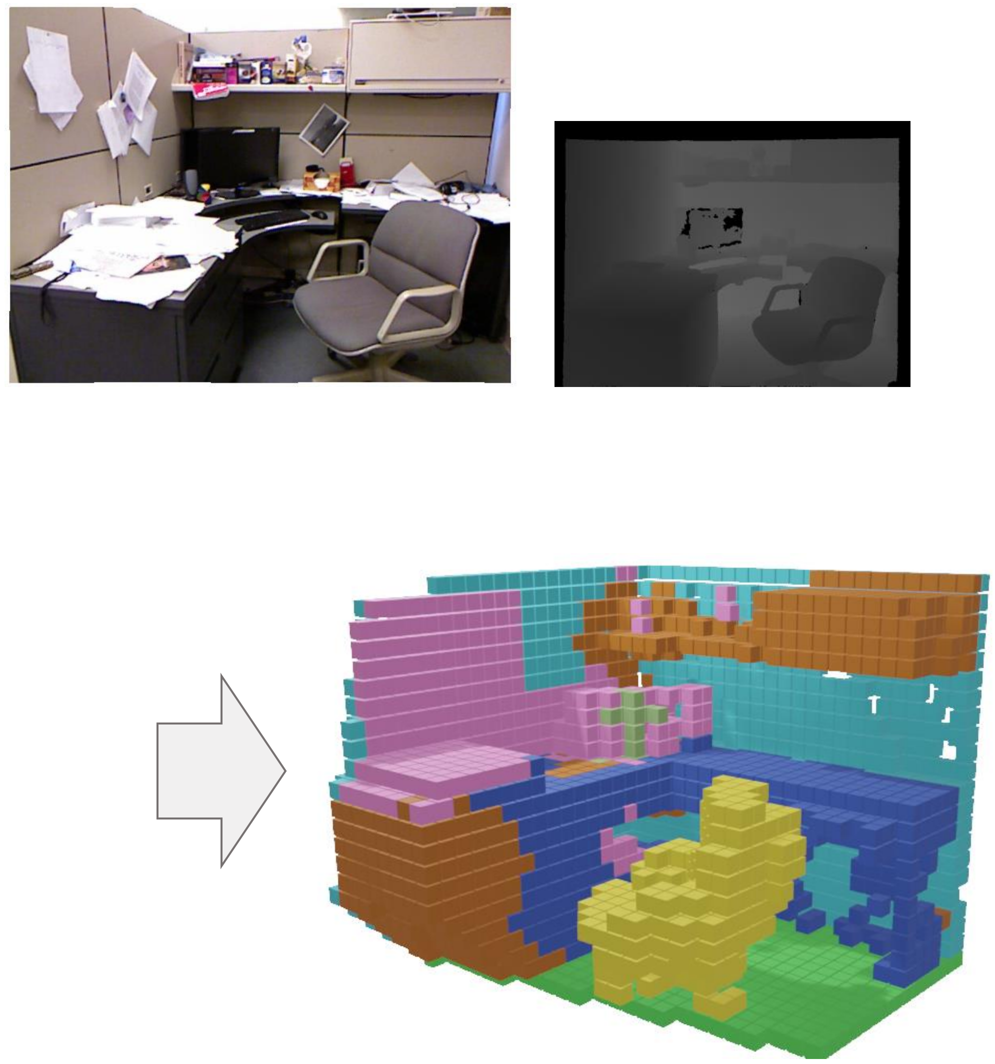
*Accepted for publication in the proceedings of the 25th International Conference on Pattern Recognition (ICPR2020) (Capes Qualis A2)

[29] Dourado, A., de Campos, T.E., Kim, H., and Hilton, A.: EdgeNet: Semantic scene completion from RGB-D images. Tech. Rep. arXiv:1908.02893, Cornell University Library, 2019. http://arxiv.org/abs/1908.02893. 6, 44, 68

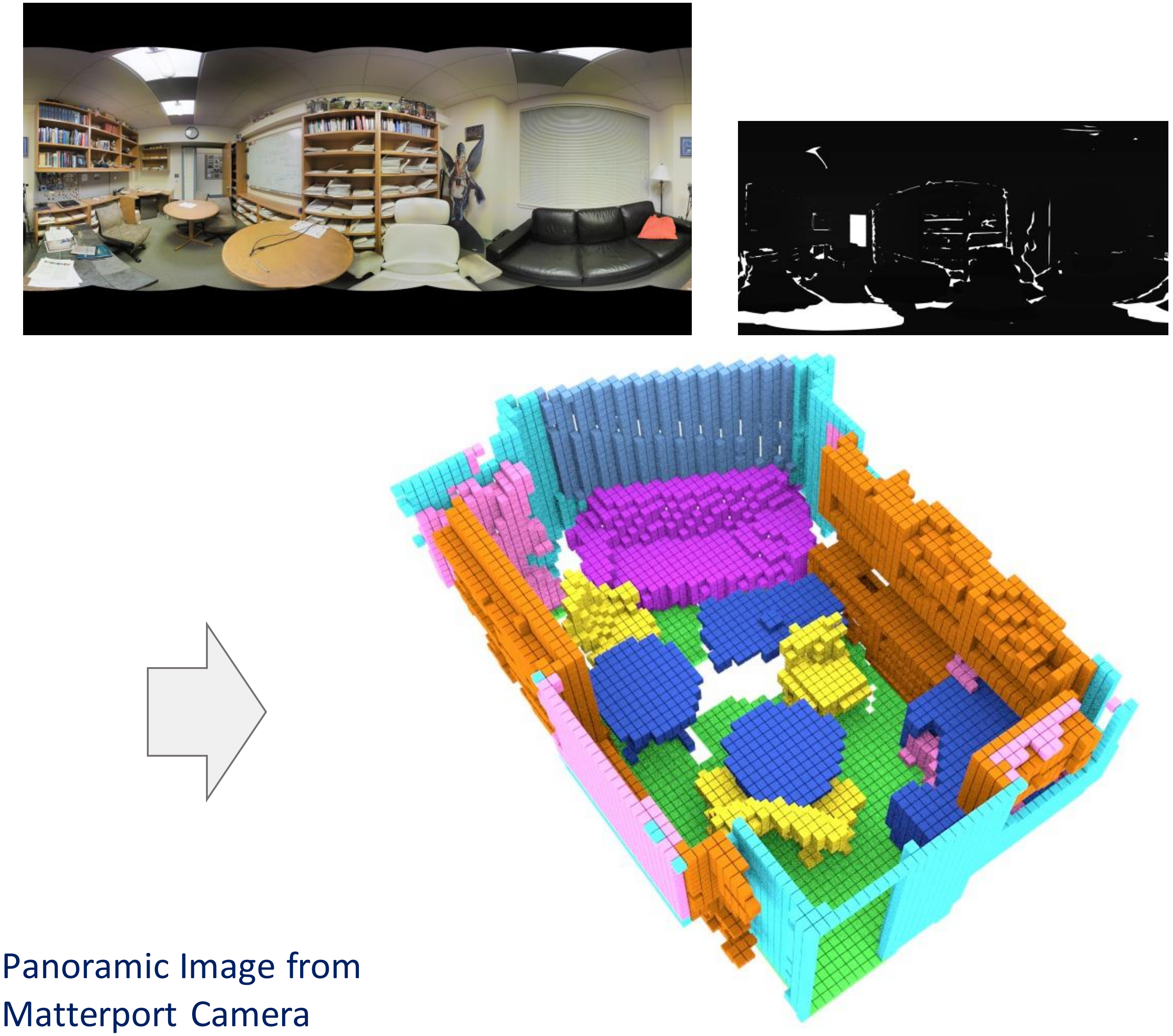
360° SCC

Extending Semantic Scene Completion for 360° Coverage

Current Semantic Scene Completion Limitations

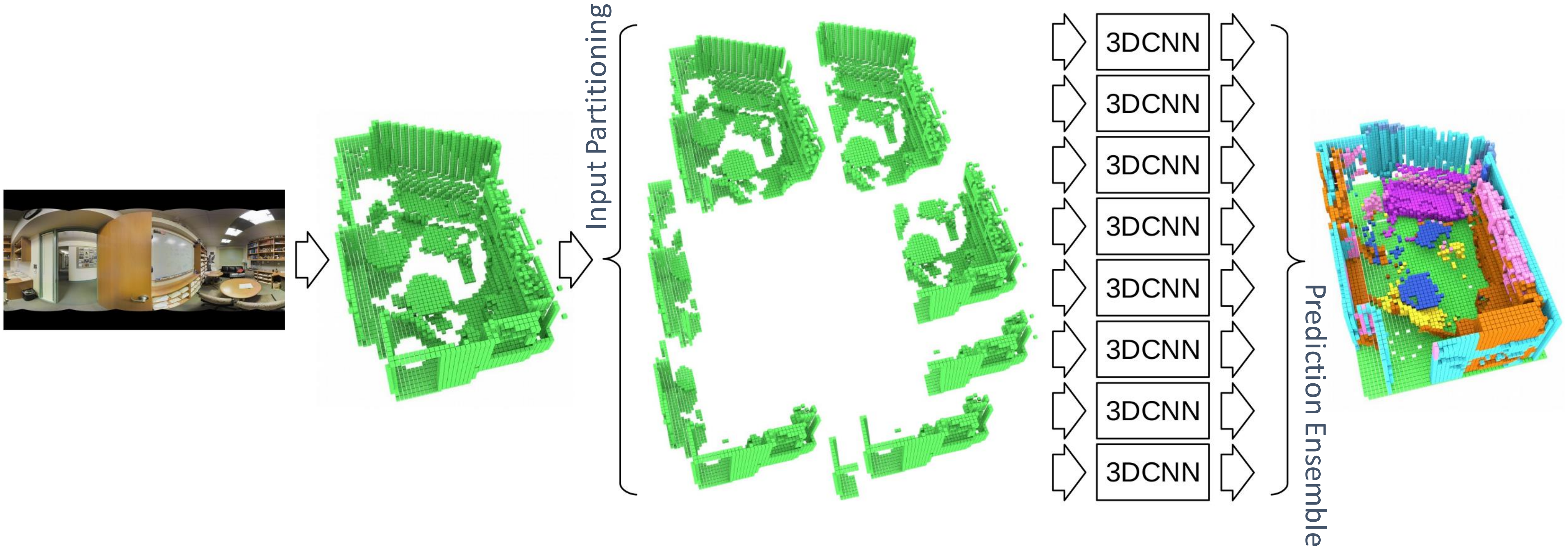


Regular RGB-D Sensor



Panoramic Image from
Matterport Camera

Our approach

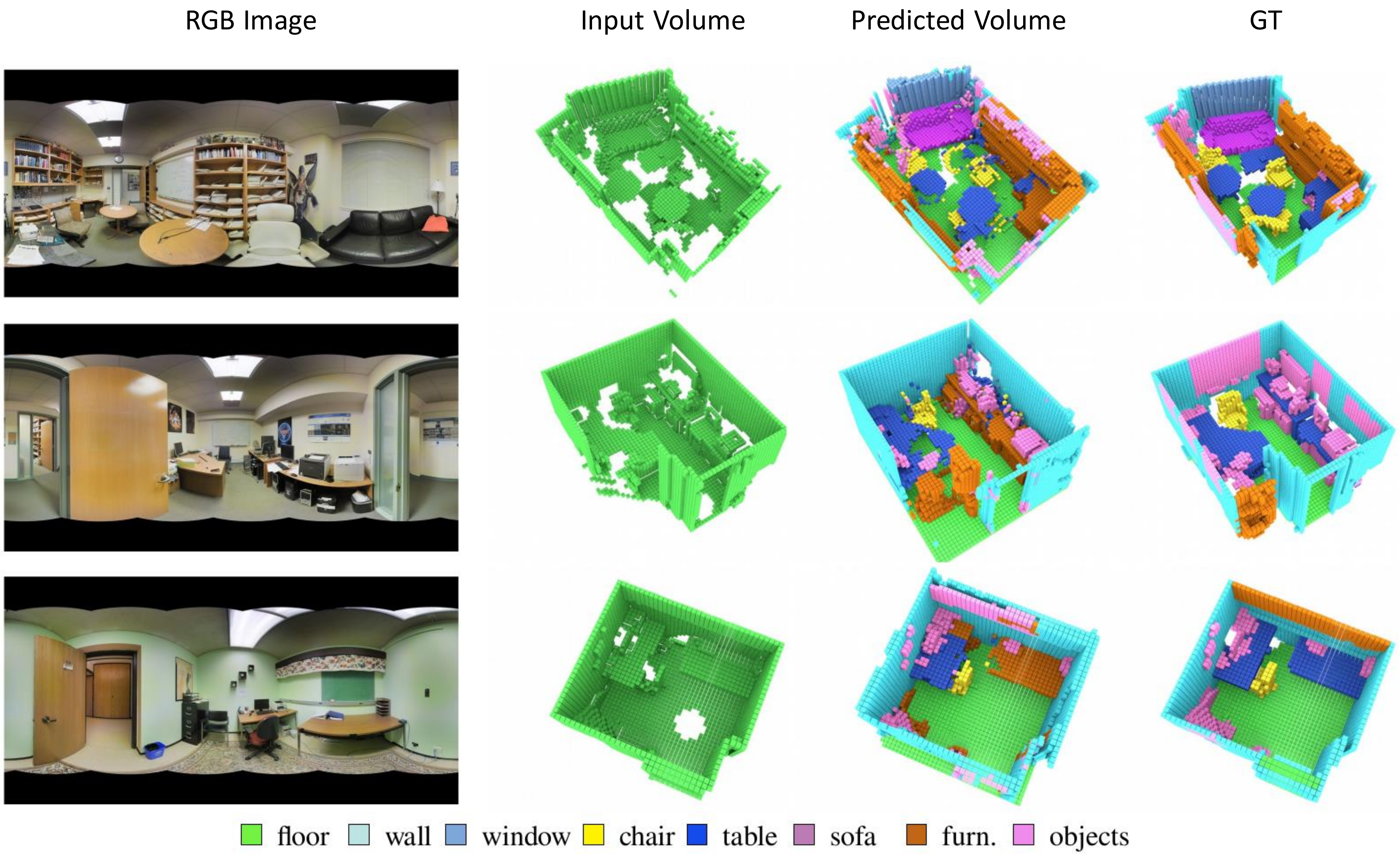


The 3DCNN is trained using SUNCG and fine-tuned in NYUDV2

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

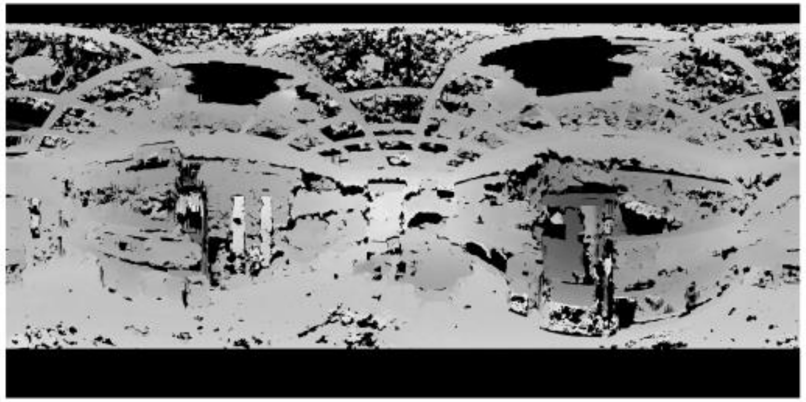
GT Pred.

Results on Stanford 2D-3DS Dataset



Experiments on Spherical Stereo Images

- Stereo capture using commercial 360^{O} cameras is one realistic approach to 360^{O} SSC
- faster compared to Matterport scanning
- depth estimation is subject to errors due to occlusions between two camera views and correspondence matching errors

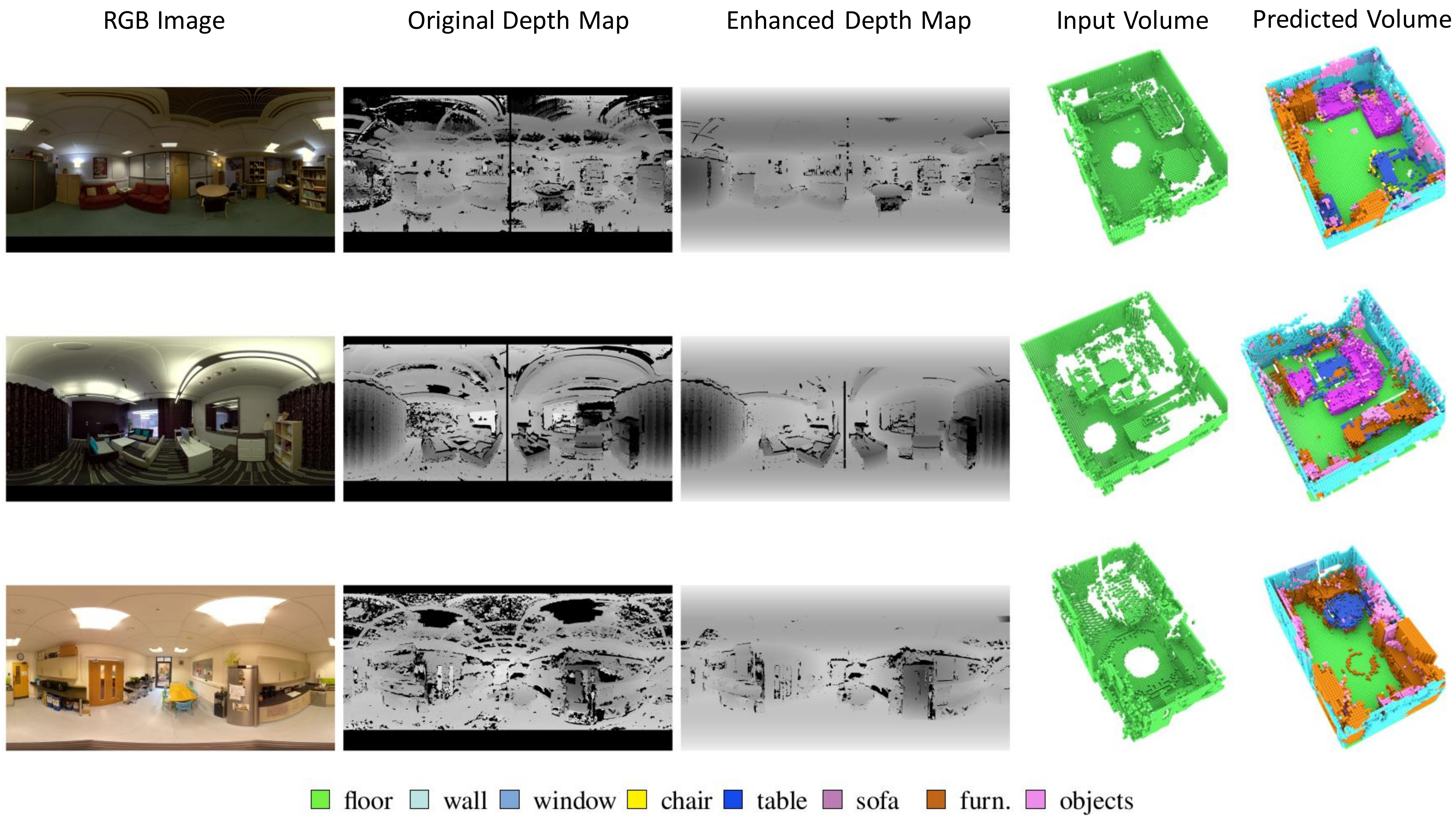


Our approach

- Vertical stereo setup
- Dense stereo matching with spherical stereo geometry [56]
- Depth map enhancement procedure:
  - Align the scene (Manhattan principle)
  - Apply Canny Edge Detector
  - RANSAC to fit a plane over coherent regions with similar colors

[56] Kim, H. and Hilton, A.: Block world reconstruction from spherical stereo image pairs. Computer Vision and Image Understanding (CVIU), 139(C):104–121, Oct. 2015, ISSN 1077-3142. http://dx.doi.org/10.1016/j.cviu.2015.04.001. 17, 69

Results on Spherical Images



Conclusions

- We introduced the 360^o Semantic Scene Completion
- Works with high-end sensors or off-the-shelf 360^o cameras
- Segmentation accuracy equivalent to limited view solutions
- High levels of completion of occluded regions

Publication

Sematic Scene Completion from a Single 360^{O} Image and Depth Map

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

Aloisio Dourado[1,a], Hansung Kim[2,b], Teofilo E. de Campos[1,c] and Adrian Hilton[2,d]
[1] *University of Brasilia, Brasilia, Brazil*
[2] *CVSSP, University of Surrey, Surrey, U.K.*

Keywords: Semantic Scene Completion, 360-Degree Scene Reconstruction, Scene Understanding, 360-Degree Stereo Images.

Abstract: We present a method for Semantic Scene Completion (SSC) of complete indoor scenes from a single 360° RGB image and corresponding depth map using a Deep Convolution Neural Network that takes advantage of existing datasets of synthetic and real RGB-D images for training. Recent works on SSC only perform occupancy prediction of small regions of the room covered by the field-of-view of the sensor in use, which implies the need of multiple images to cover the whole scene, being an inappropriate method for dynamic scenes. Our approach uses only a single 360° image with its corresponding depth map to infer the occupancy and semantic labels of the whole room. Using one single image is important to allow predictions with no previous knowledge of the scene and enable extension to dynamic scene applications. We evaluated our method on two 360° image datasets: a high-quality 360° RGB-D dataset gathered with a Matterport sensor and low-quality 360° RGB-D images generated with a pair of commercial 360° cameras and stereo matching. The experiments showed that the proposed pipeline performs SSC not only with Matterport cameras but also with more affordable 360° cameras, which adds a great number of potential applications, including immersive spatial audio reproduction, augmented reality, assistive computing and robotics.

1 INTRODUCTION

Automatic understanding of the complete 3D geometry of a indoor scene and the semantics of each occupied 3D voxel is one of essential problems for many applications, such as robotics, surveillance, assistive computing, augmented reality, immersive spatial audio reproduction and others. After years as an active research field, this still remains a formidable challenge in computer vision. Great advances in scene understanding have been observed in the past few years due to the large scale production of inexpensive depth sensors, such as Microsoft Kinect. Public RGB-D datasets have been created and widely used for many 3D tasks, including prediction of unobserved voxels (Firman et al., 2016), segmentation of visible surface (Silberman and Fergus, 2011; Ren et al., 2012; Qi et al., 2017b; Gupta et al., 2013), object detection (Shrivastava and Mulam, 2013) and single object completion (Nguyen et al., 2016).

In 2017, a new line of work was introduced, focusing on the complete understanding of the scene: Semantic Scene Completion (SSC) (Song et al., 2017). SSC is the joint prediction of occupation and semantic labels of visible and occluded regions of the scene. The works in this area are mostly based on the use of Convolution Neural Networks (CNNs) trained on both synthetic and real RGB-D data (Garbade et al., 2018; Guedes et al., 2017; Zhang et al., 2018a; Zhang et al., 2018b; Liu et al., 2018). However, due to the limited field-of-view (FOV) of RGB-D sensors, those methods only predict semantic labels for a small part of the room and at least four images are required to understand the whole scene.

This scenario recently started to change with the use of more advanced technology for large-scale 3D scanning, such as Light Detection and Ranging (LIDAR) sensor and Matterport cameras. LIDAR is one of the most accurate depth ranging devices using a light pulse signal but it acquires only a point cloud set without colour or connectivity. Some recent LIDAR devices provide coloured 3D structure by map-

a https://orcid.org/0000-0002-5037-7178
b https://orcid.org/0000-0003-4907-0491
c https://orcid.org/0000-0001-6172-0229
d https://orcid.org/0000-0003-4223-238X

36

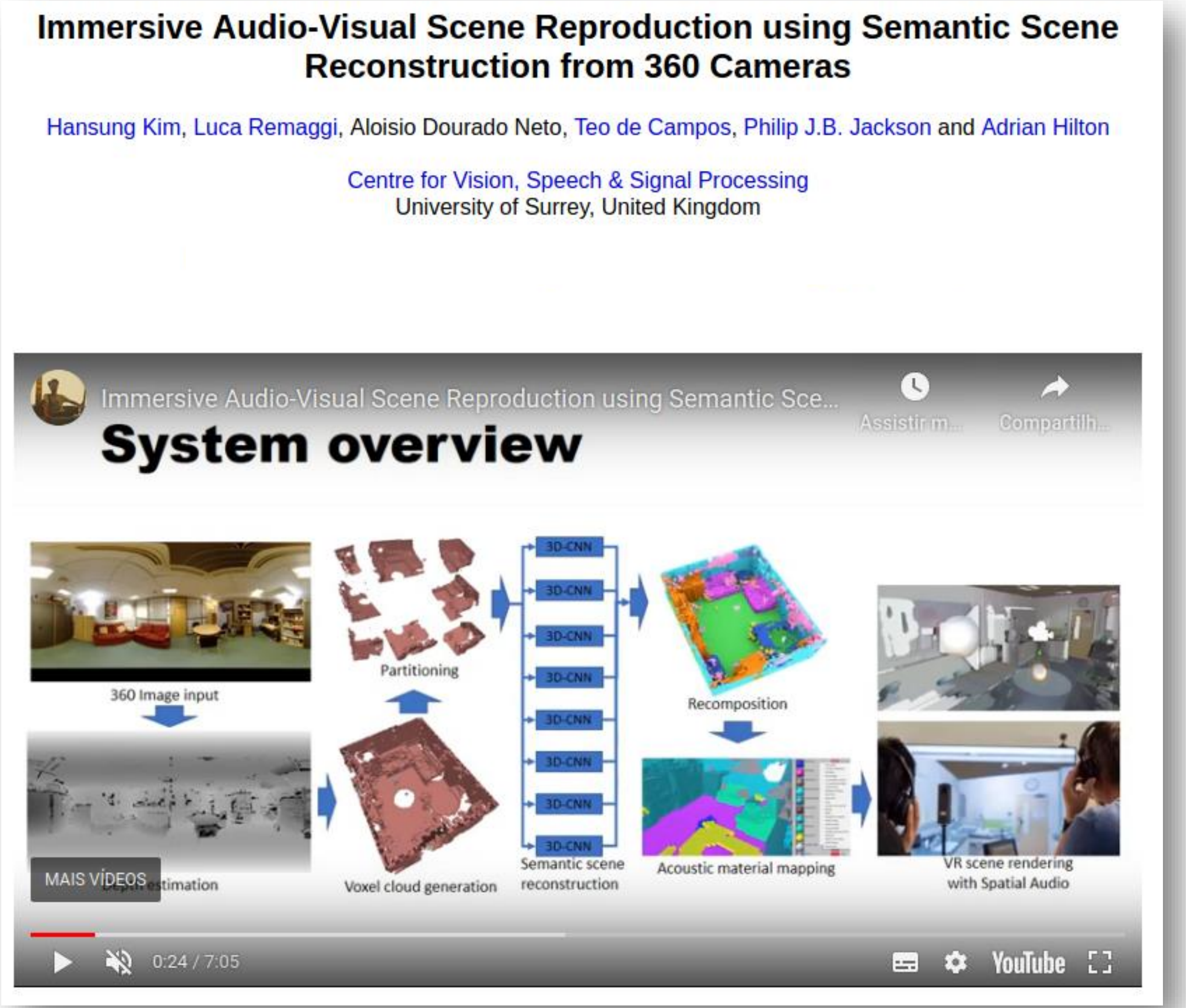
Dourado, A., Kim, H., E. de Campos, T. and Hilton, A.
Semantic Scene Completion from a Single 360-Degree Image and Depth Map.
DOI: 10.5220/0008877700360046
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)*, pages 36-46
ISBN: 978-989-758-402-2

*Published in the proceedings of the 15th International Conference on Computer Vision Theory and Applications (VISAPP2020) (Qualis A1)

[31] Dourado, A., Kim, H., de Campos, T.E., and Hilton, A.: Semantic scene completion from a single 360-degree image and depth map. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020), vol. 5: VISAPP, pp. 36–46. 7, 61

Application Paper

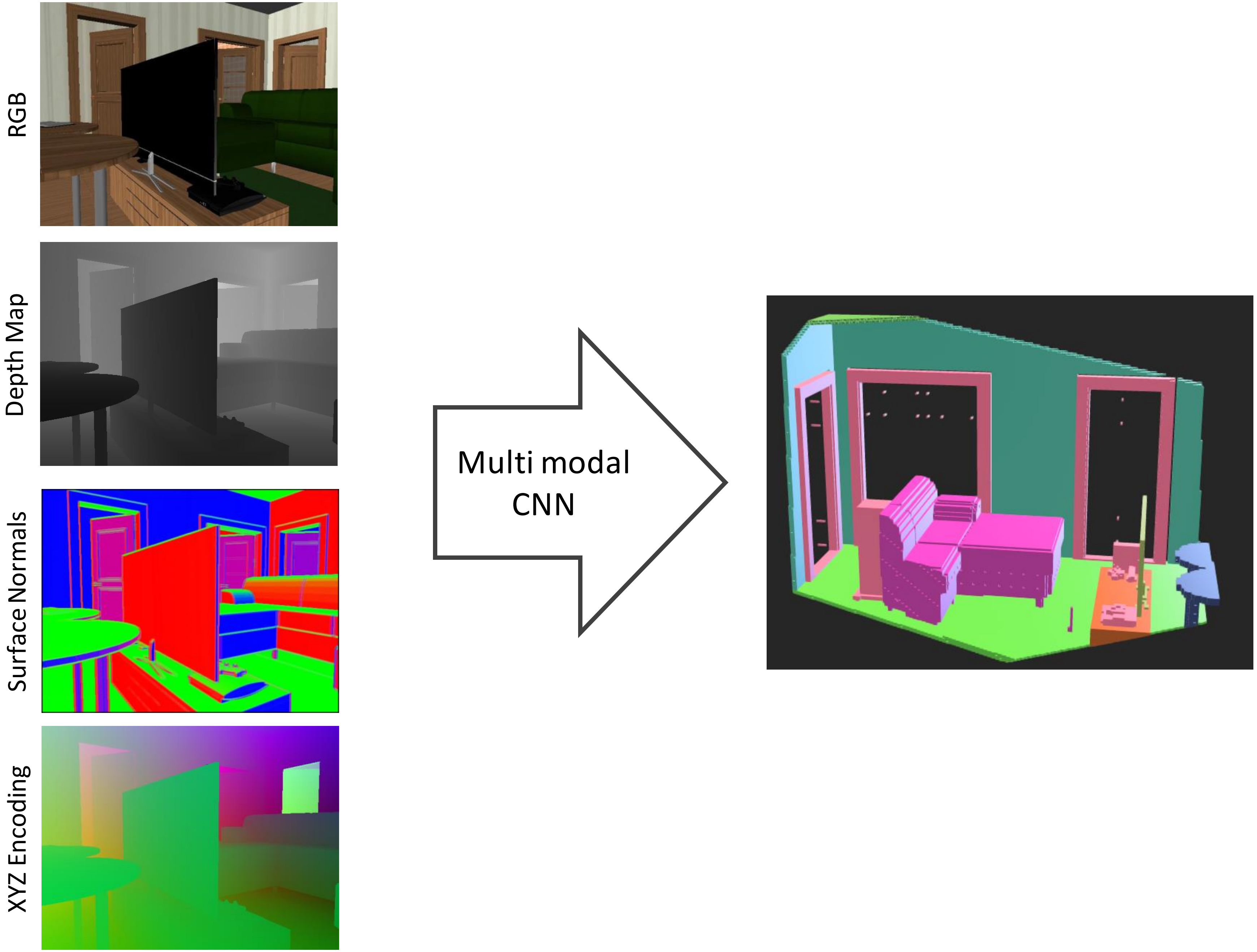
Immersive Audio-Visual Scene Reproduction using Semantic Scene Reconstruction from 360^o Cameras



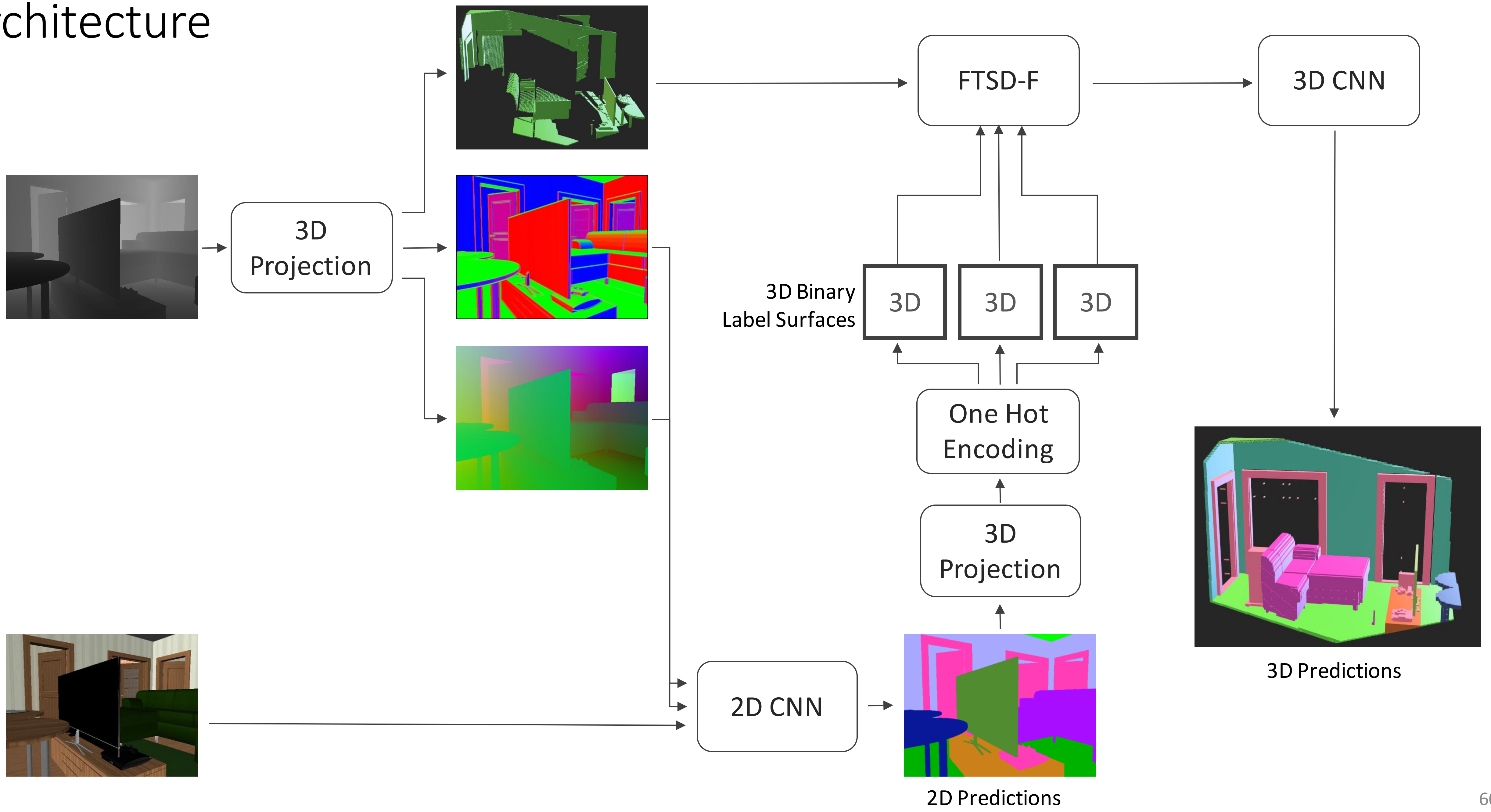
https://www.cvssp.org/hkim/paper/CVST2020/

Next Steps

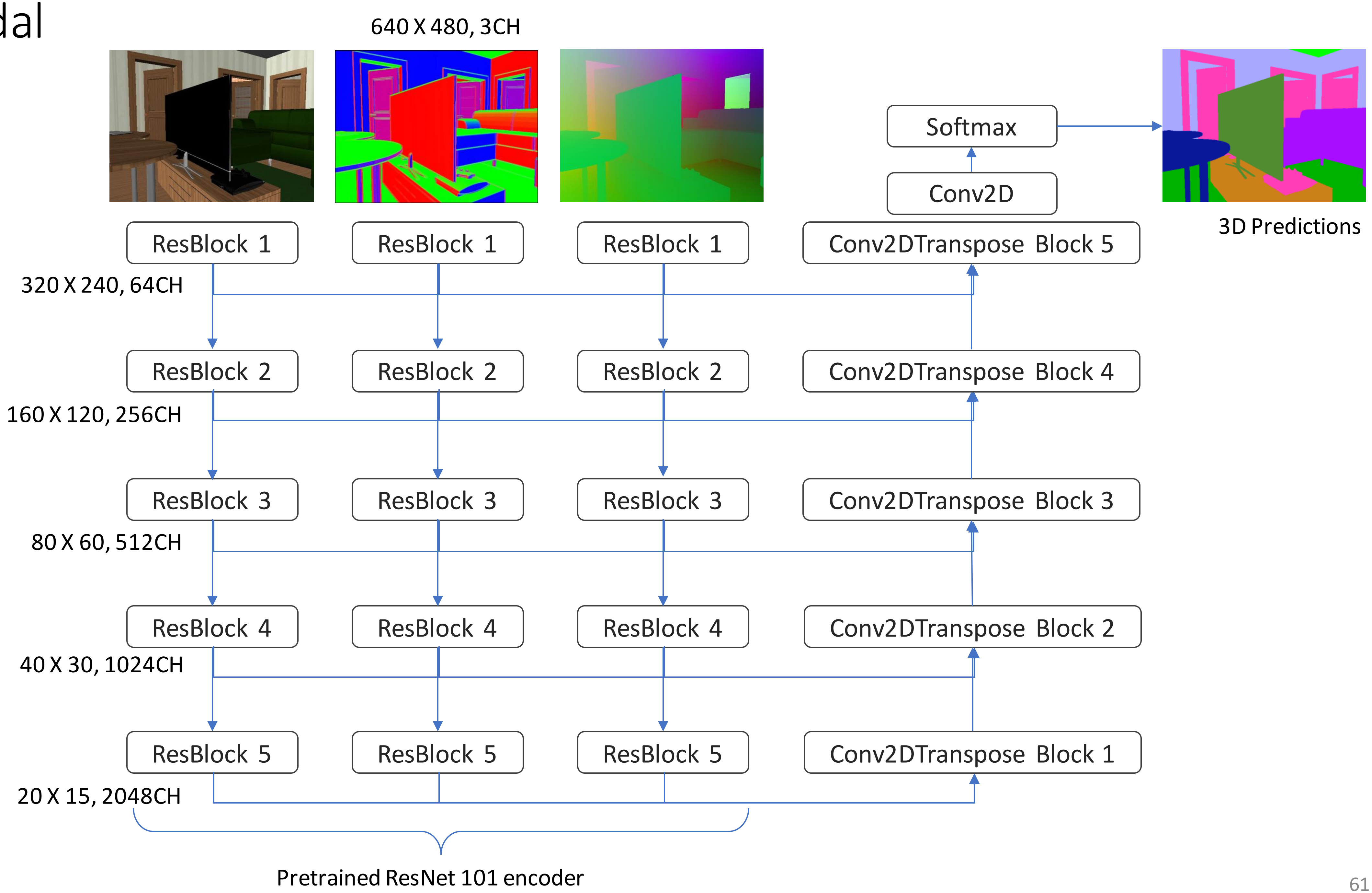
Multi modal Semantic Scene Completion



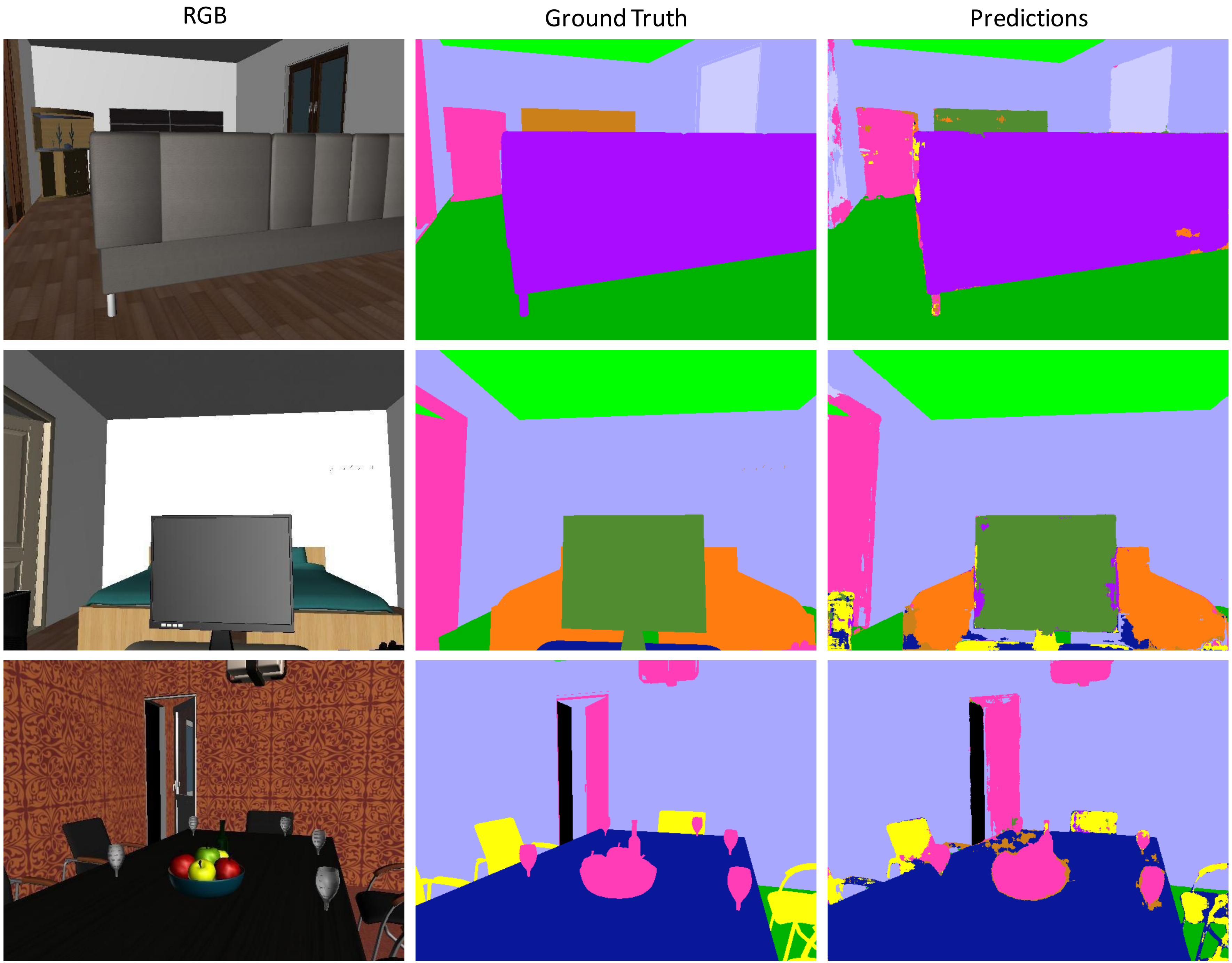
Multi modal architecture



2D multimodal network architecture



Qualitative results, so far



Thank you!

Results – ablation study on SUNCG

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

Results – ablation study on SUNCG

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

Effect of our efficient training pipeline

Results – ablation study on SUNCG

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

Effect of our u-shaped architecture, with 3D dilated residial modules

Results – ablation study on SUNCG

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

Effect of adding edges

Results – ablation study on SUNCG

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

Effect of adding edges

Results on NYU-DV2

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

Effect of different fusion strategies

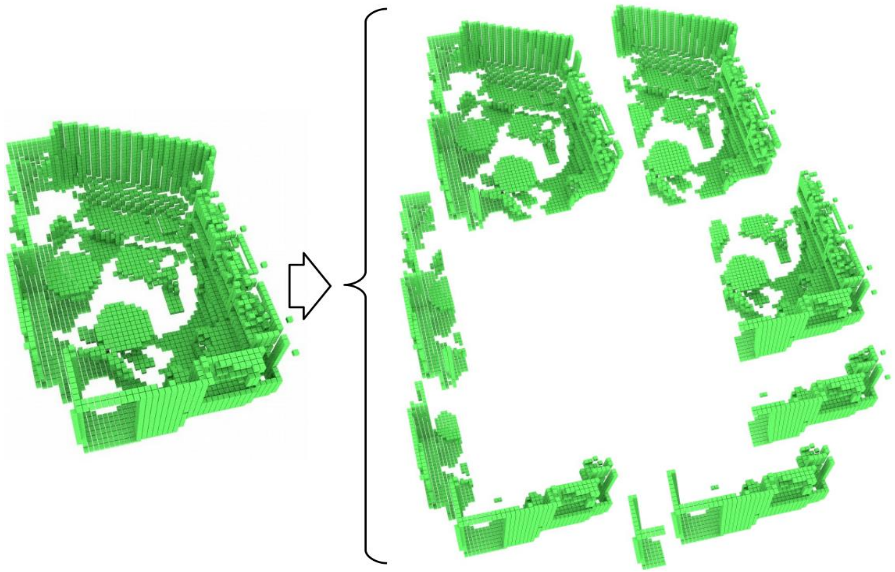
Results on NYU-DV2

train	input	model	scene completion prec.	rec.	IoU	semantic scene completion (IoU, in percentages) 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

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^{o} step
- We move the point-of-view 1.7m back from the original sensor position, to get more overlapped coverage

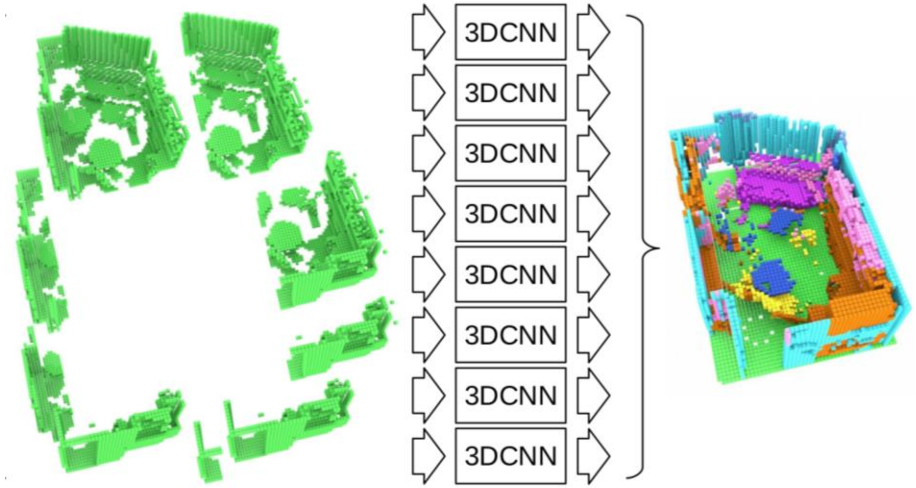
Input Partitioning



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

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**