University of Brasília - UnB

Institute of Exact Sciences
Department of Computer Science

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

Qualifying examination of the Ph.D. Program in Computer Science

Aloisio Dourado Neto

Supervisor
Prof. Dr. Teófilo Emidio de Campos

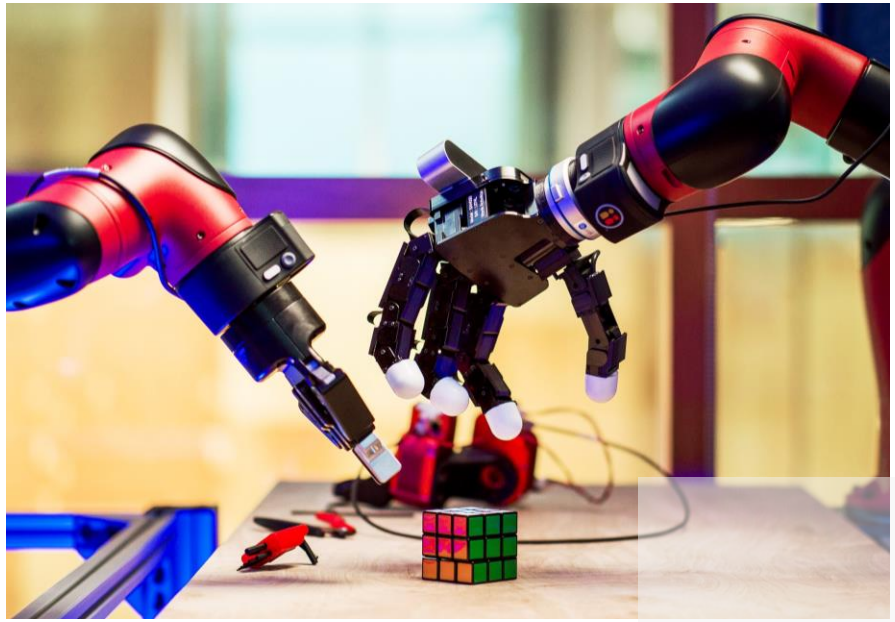
Presentation Outline

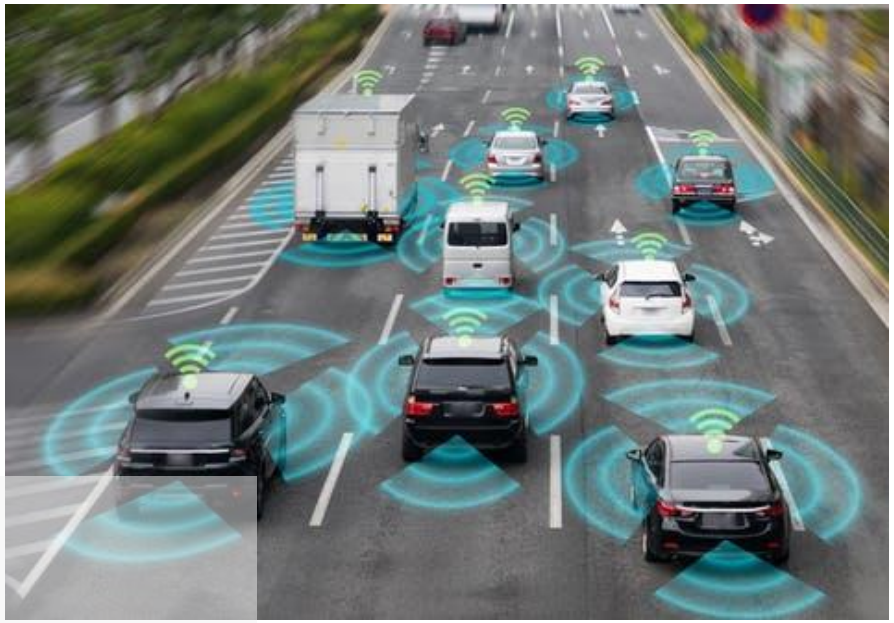
- Introduction
  - Motivation
  - Problem statement
  - Objectives
  - Contributions and publications
- Background
- Research steps
  - 2D domain: fully convolutional networks, domain adaptation and semantic segmentation
  - 3D domain: Using RGB Edges to improve Semantic Scene Completion from RGB-D Images
  - 360^0 3D: Extending Semantic Scene Completion for 360^0 Coverage
- Work plan

Chapter 1

Introduction

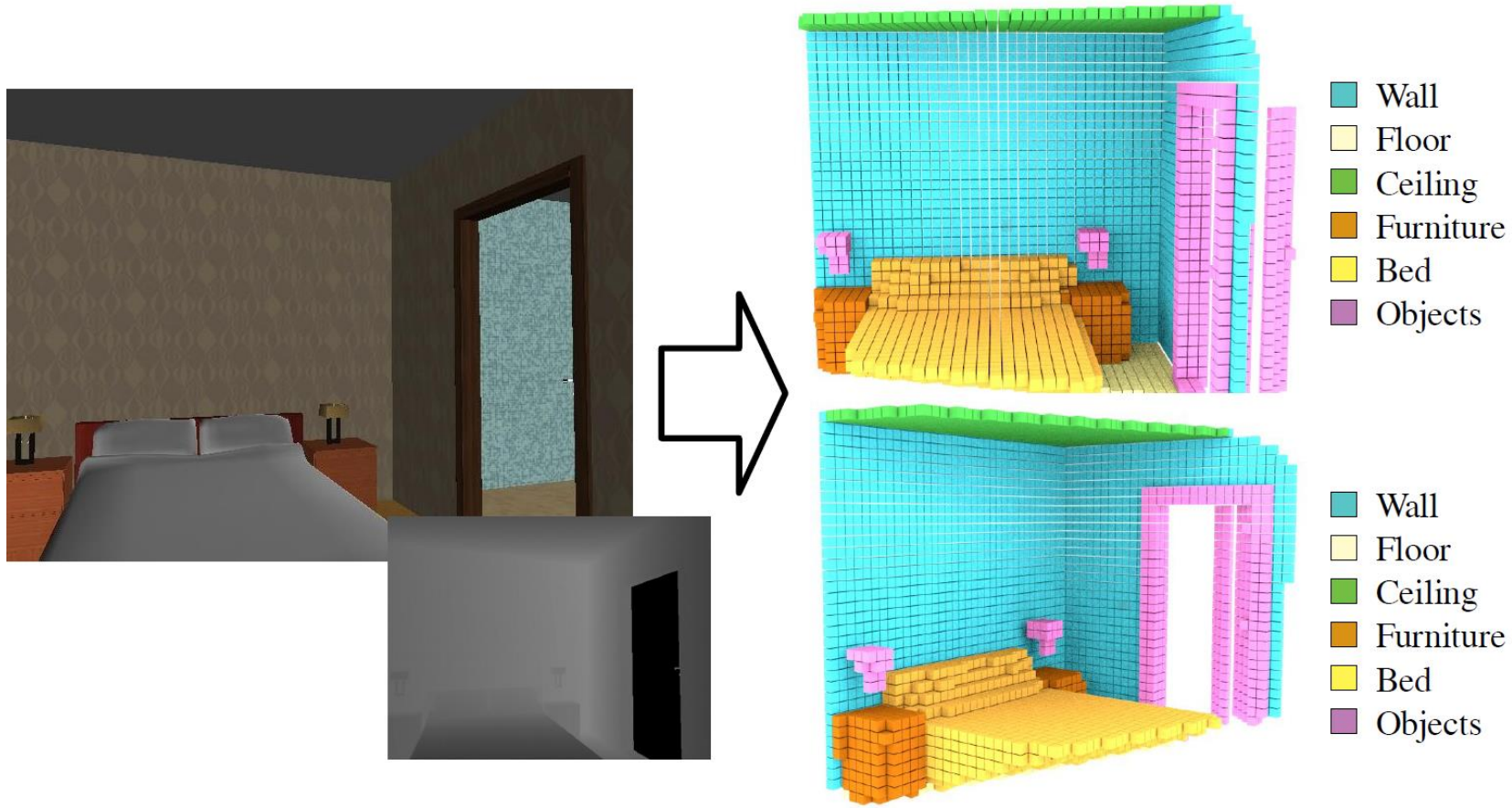
Reasoning about scenes in 3D is still an open field of study in Computer Vision. Despite great advances we have seen in the last few decades, there is still a lot of room for improvement.





Applications

Semantic Scene Completion

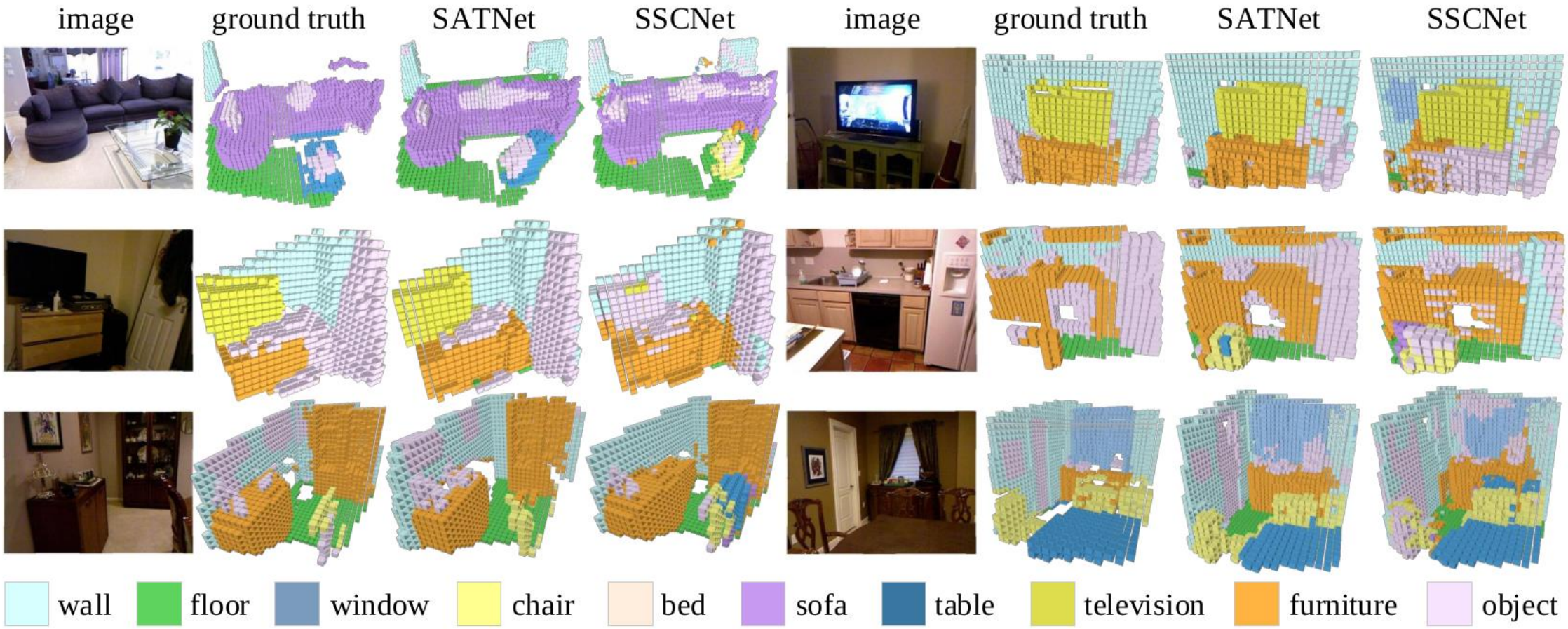


Introduced by Song *et al.*[107] in 2017

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

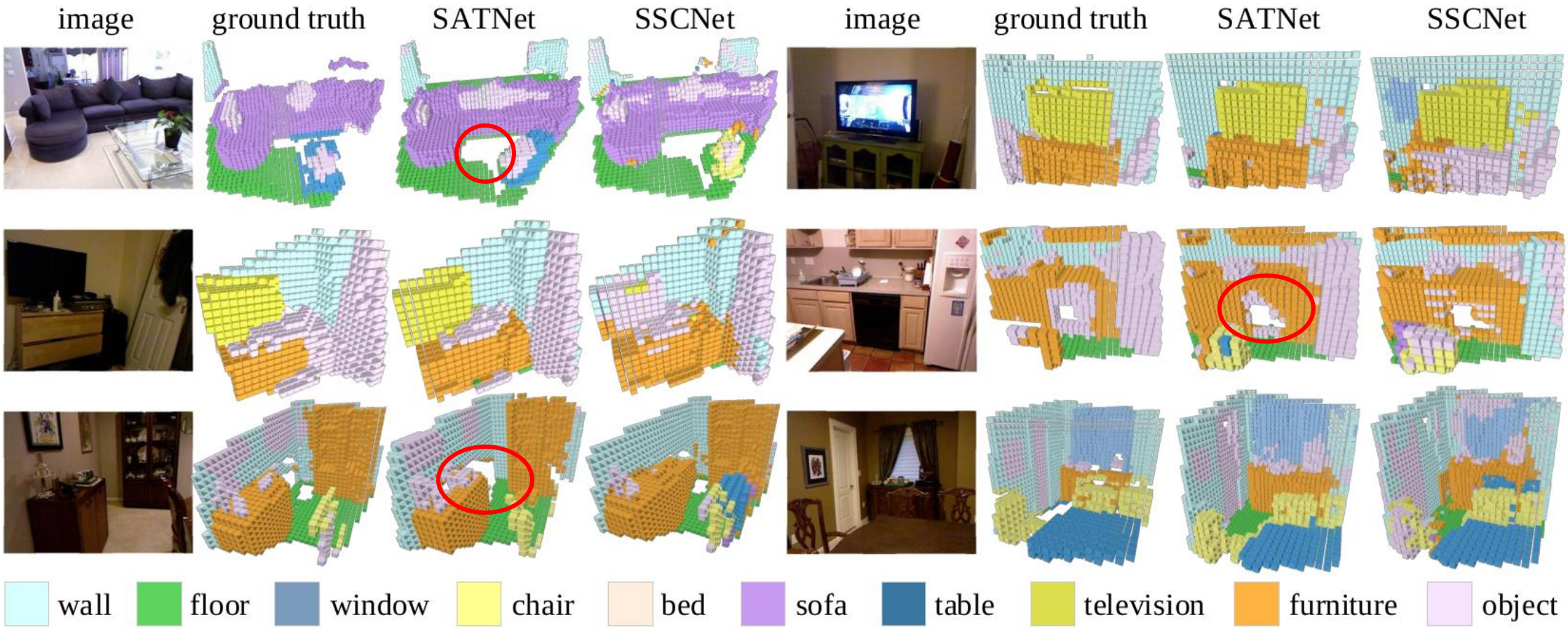
Problem Statement



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

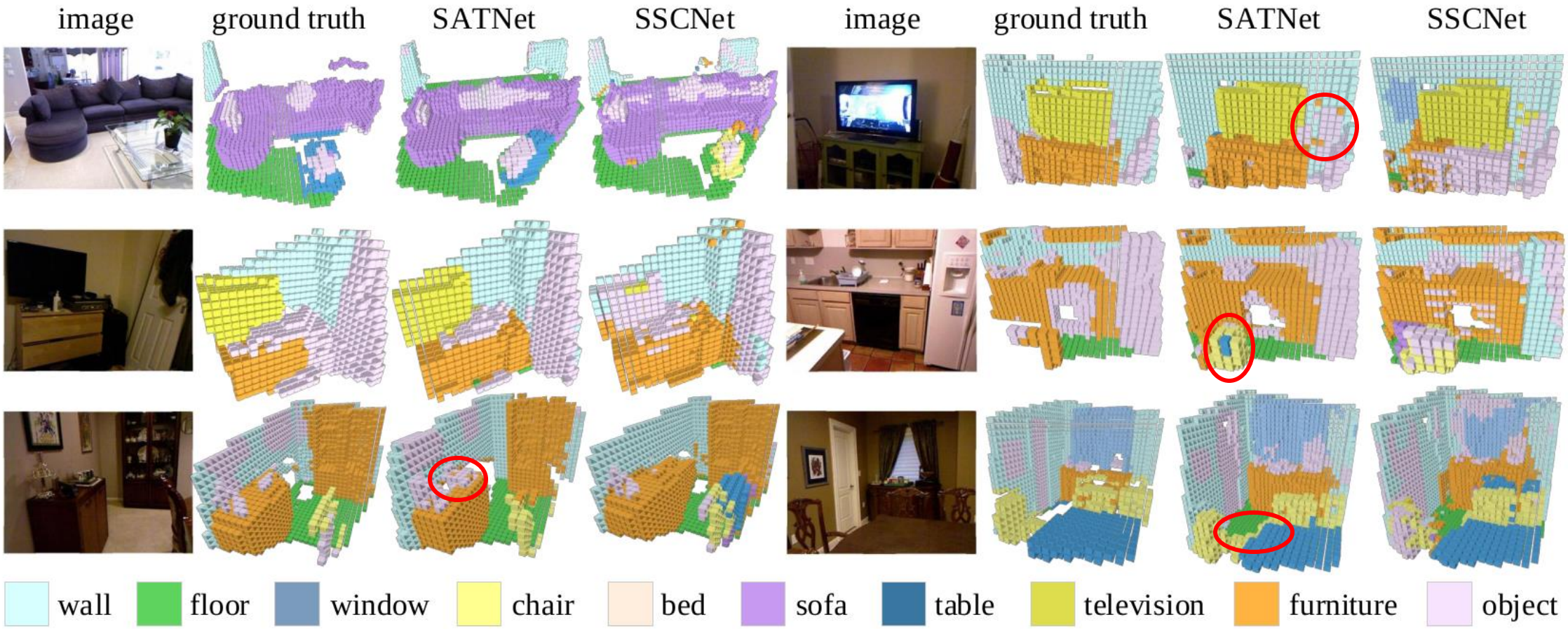
Problem Statement



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

- 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

Objectives

New tools and models that could push SSC solutions towards a complete understating of the whole indoor scene

2D Domain

- to assess the benefits of domain adaptation techniques in the context of image segmentation

3D Domain

- to propose and evaluate a new SSC model that uses the RGB information present n RGB-D images

360º 3D

- to propose and evaluate a solution to perform 360º SSC

Contributions

Domain adaptation in the context of image segmentation

Paper - Domain Adaptation for Holistic Skin Detection

Domain Adaptation for Holistic Skin Detection

Aloisio Dourado
Department of Computer Science, University of Brasilia
Brasilia, DF, 70910-900, Brazil
aloisio.dourado.bh@gmail.com

Frederico Guth
fredguth@fredguth.com

Teofilo de Campos
t.decampos@oxfordalumni.org
http://cic.unb.br/~teodecampos/

Li Weigang
weigang@unb.br
http://cic.unb.br/~weigang/

Human skin detection in images is a widely studied topic of Computer Vision for which it is commonly accepted that analysis of pixel color or local patches may suffice. This is because skin regions appear to be relatively uniform and many argue that there is a small chromatic variation among different samples. However, we found that there are strong biases in the datasets commonly used to train or tune skin detection methods. Furthermore, the lack of contextual information may hinder the performance of local approaches. In this paper we present a comprehensive evaluation of holistic and local Convolutional Neural Network (CNN) approaches on in-domain and cross-domain experiments and compare with state-of-the-art pixel-based approaches. We also propose a combination of inductive transfer learning and unsupervised domain adaptation methods, which are evaluated on different domains under several amounts of labelled data availability. We show a clear superiority of CNN over pixel-based approaches even without labelled training samples on the target domain. Furthermore, we provide experimental support for the counter-intuitive superiority of holistic over local approaches for human skin detection.

Keywords: Domain Adaptation, Skin segmentation, CNN.

1. Introduction

Human skin detection is the task of identifying which pixels of an image correspond to skin. The segmentation of skin regions in images has several applications: video surveillance, people tracking, human computer interaction, face detection and recognition and gesture detection, among many others.[28,38]

Before the boom of Convolutional Neural Networks (CNNs), most approaches

1

- The proposal of a new Domain Adaptation strategy that combines Pseudo-Labeling and Transfer Learning for cross-domain training
- A comparison between holistic and local approaches on in-domain and cross-domain
- comparison of CNN-based to pixel-based approaches
- an experimental assessment of the generalization power of different human skin datasets

*Submitted to International Journal of Pattern Recognition and Artificial Intelligence (Capes Qualis B1)

[30] Dourado, A., Guth, F., de Campos, T.E., and Weigang, L.: Domain adaptation for holistic skin detection. Tech. Rep. arXiv:1903.0969, Cornell University Library, 2019. http://arxiv.org/abs/1903.06969. 6, 26

Contributions

Use of RGB information present in RGB-D images

Paper – 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;

- A new end-to-end CNN architecture that fuses depth and 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 function
- A more efficient end-to-end training pipeline for semantic scene completion

*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

Contributions

360º Semantic Scene Completion

Paper – Sematic Scene Completion from a Single* 360º *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

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

- The extension of the SSC task to complete scene understanding using 360º imaging sensors or stereoscopic spherical cameras
- A novel approach to perform SSC for 360º images taking advantage of existing standard RGB-D datasets for network training
- A pre-processing method to enhance depth maps estimated from a stereo pair of low-cost 360º cameras

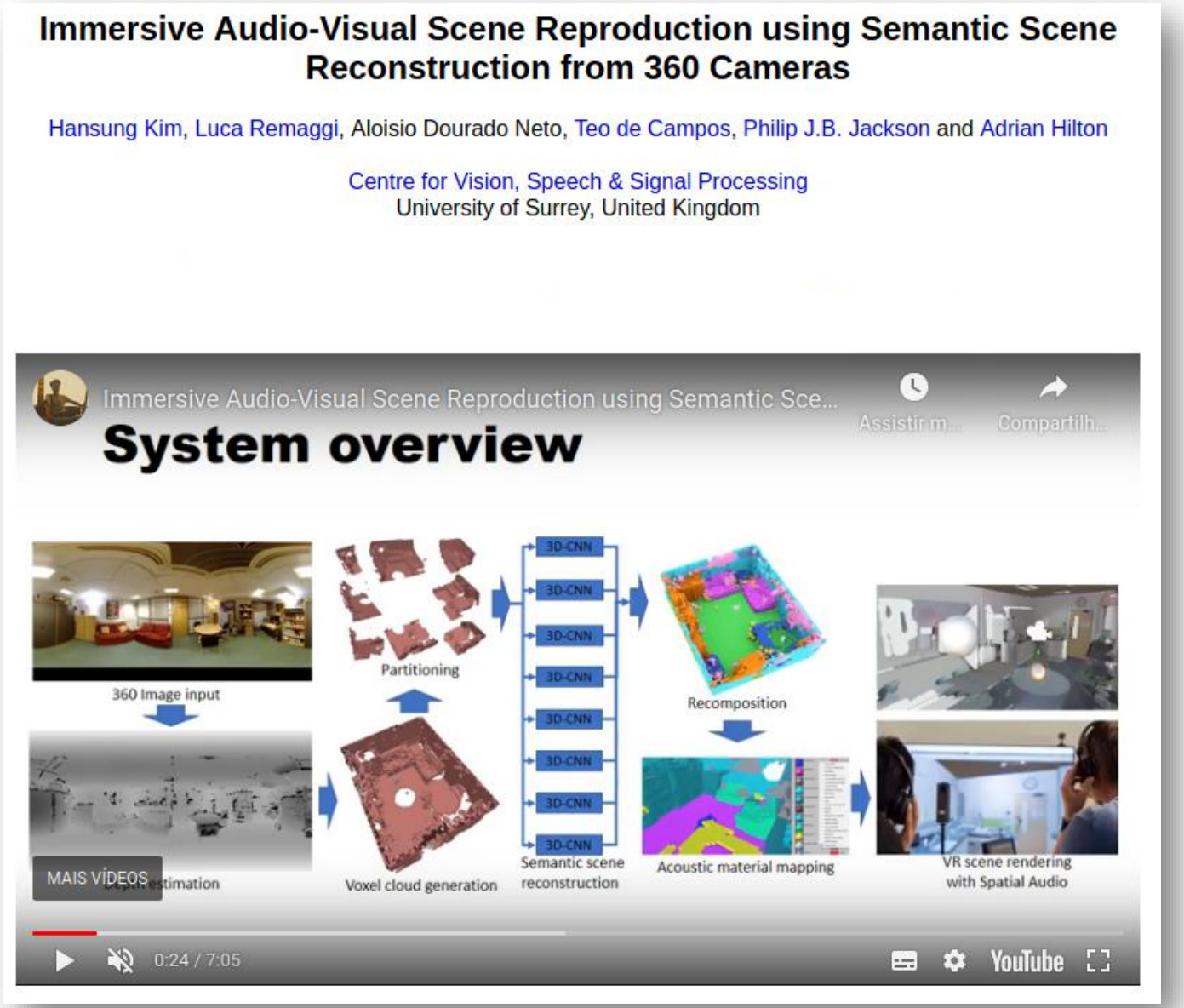
*Published in the proceedings of the 15th International Conference on Computer Vision Theory and Applications (VISAPP2020) (Qualis A1)

[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

Contributions

Application related to immersive audio-visual reproduction system:

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

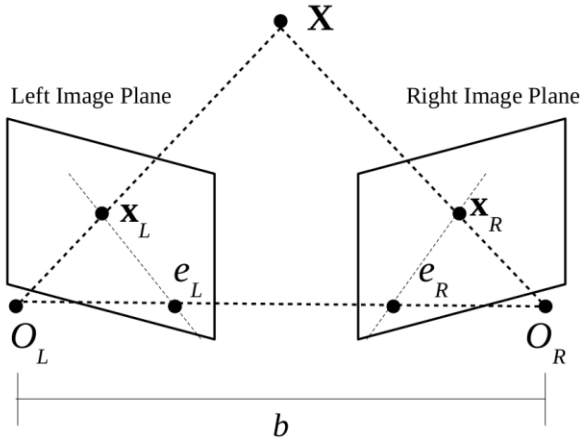


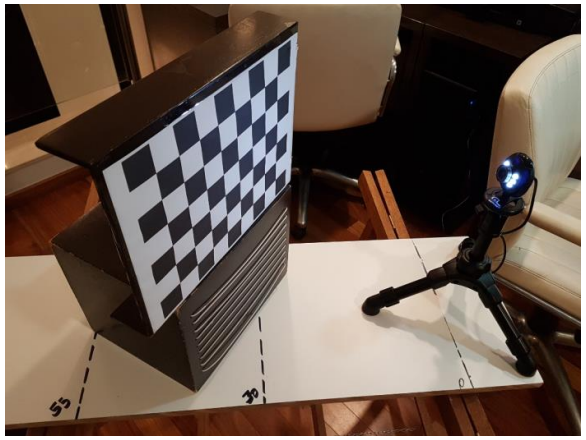
https://www.cvssp.org/hkim/paper/CVST2020/

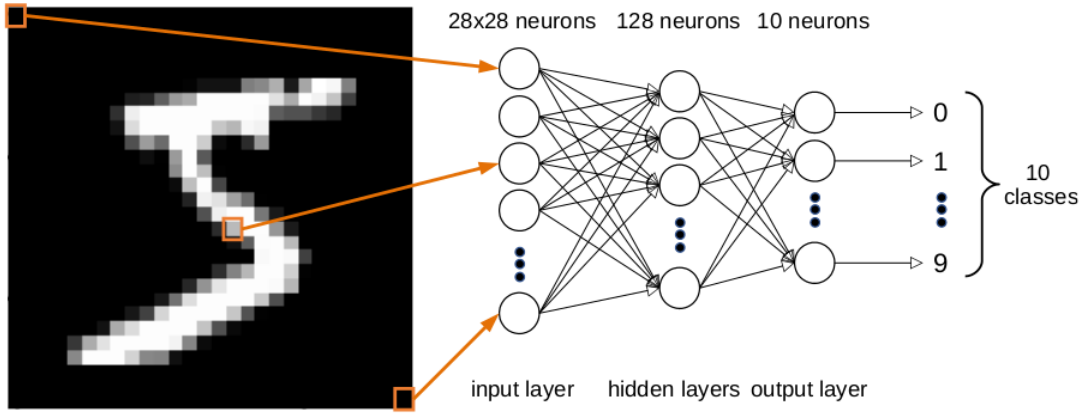
Chapter 2

Background

Relevant background knowledge for the remaining of this thesis





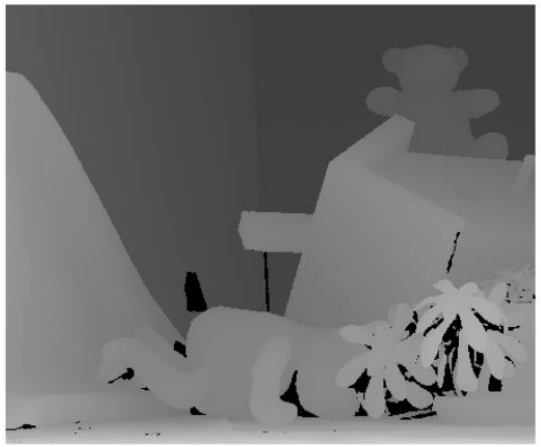


Background

Stereo Images and Depth Estimation

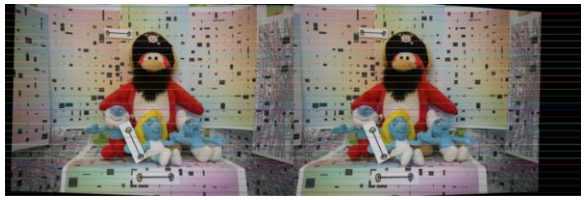
(a) Left view

(b) Right view



(c) Depth map

Epipolar Geometry



360^{O} Stereo

Background

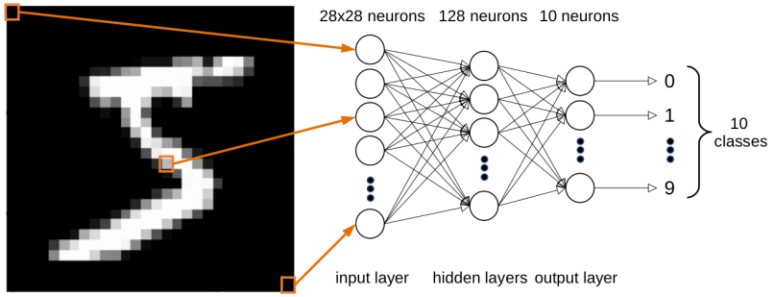
Depth Sensors

RealSense®. ©Intel Corporation.
Reproduced with permission.

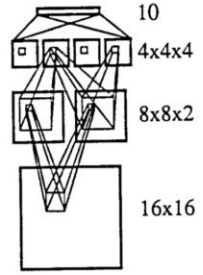
Matterport 360° Camera.
©Matterport Inc. Reproduced with permission.

Background

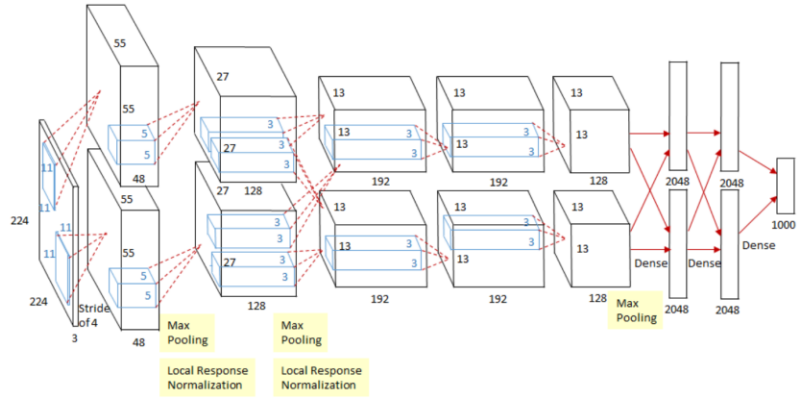
CNNs for Image Classification



Typical fully connected back-propagation neural network



Original image of the convolutional network proposed by LeCun in 1989 [62]. ©Elsevier/North Holland, 1989. Reproduced with permission.



AlexNet architecture (2012)[60]. Copyright held by the authors. Reproduced with permission.

[60] Krizhevsky, A., Sutskever, I., and Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J.C., Bottou, L., and Weinberger, K.Q. (eds.): Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc., 2012. 2, 20, 21
[62] Lecun, Y.: Generalization and network design strategies. Connectionism in perspective, 1989. https://ci.nii.ac.jp/naid/10008946620/en/. 19, 20

Background

Image Segmentation: from CNNs to FCNs

- Patch-based
- Fully Convolutional Network
- Encoder-decoder architectures
- U-Net

Background

Domain adaptation

- Transfer learning
- Inductive transfer learning and fine tunning
- Unsupervised domain adaptation
- Semi-supervised learning

Research steps

2D Domain

- Domain adaptation applied to the skin segmentation problem

3D Domain

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

360^O 3D

- Extending Semantic Scene Completion for 360^O Coverage

Chapter 3

Fully Convolutional Networks, Domain Adaptation and Semantic Segmentation

The Skin Detection Problem, Local vs Holistic, and the use of Domain Adaptation

Fully Convolutional Networks, Domain Adaptation and Semantic Segmentation

Why work on 2D?

- Work on 3D is hard, so it means less previous works to compare!
- We preferred to start to explore domain adaptation, FCNs and the segmentation problem in an easier domain

- Why the skin segmentation application?
  - Research field where some criticisms regardind the use of CNNs/FCNs are made:
    - the need for large training datasets [53]
    - the specificity or lack of generalization of neural nets
    - their prediction time [12]
  - One of our goals was to refute those criticisms

[53] Kakumanu, P., Makrogiannis, S., and Bourbakis, N.: A survey of skin-color modeling and detection methods. Pattern Recognition, 40(3):1106 – 1122, 2007, ISSN 0031-3203. 27
[12] Brancati, N., Pietro, G.D., Frucci, M., and Gallo, L.: Human skin detection through correlation rules between the YCb and YCr subspaces based on dynamic color clustering. Computer Vision and Image Understanding, 155:33 – 42, 2017, ISSN 1077-3142.
27, 28, 35, 36, 39, 42

The Skin Detection Problem

Human skin detection is the task of identifying which pixels of an image correspond to skin

Applications:

- video surveillance
- people tracking
- human-computer interaction

Previous Works

Historically, color-based or texture methods were preferred [49, 100]

Current state-of the-art works still rely on local approaches:

- Skin-color separation [12, 33]
- Patch-based CNN [74]

The use of domain adaptation methods for this problem is not common

[12] Brancati, N., Pietro, G.D., Frucci, M., and Gallo, L.: Human skin detection through correlation rules between the YCb andYCr subspaces based on dynamic color clustering. Computer Vision and Image Understanding, 155:33 – 42, 2017, ISSN 1077-3142.
27, 28, 35, 36, 39, 42
[33] Faria, R.A.D. and Hirata Jr., R.: Combined correlation rules to detect skin based on dynamic color clustering. In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), vol. 5, pp. 309–316. INSTICC, SciTePress, 2018, ISBN 978-989-758-290-5. 28, 35, 36
[49] Huynh-Thu, Q., Meguro, M., and Kaneko, M.: Skin-Color-Based Image Segmentation and Its Application in Face Detection. In MVA, pp. 48–51, 2002. 27, 39
[74] Lumini, A. and Nanni, L.: Fair comparison of skin detection approaches on publicly available datasets. Techn. rep., Cornell University Library, CoRR/cs.CV, August 2019. arXiv:1802.02531 (v3). 28, 43
[100] Shrivastava, V.K., Londhe, N.D., Sonawane, R.S., and Suri, J.S.: Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features. Comput. Methods Prog. Biomed., 126(C):98–109, Apr. 2016, ISSN 0169-2607. 27

Experiments

In-domain:

- Local CNN vs Holistic FCN
- Comparison to current color-based state-of-the-art

Cross-domain:

- Assessment of the gains of 3 simple methods

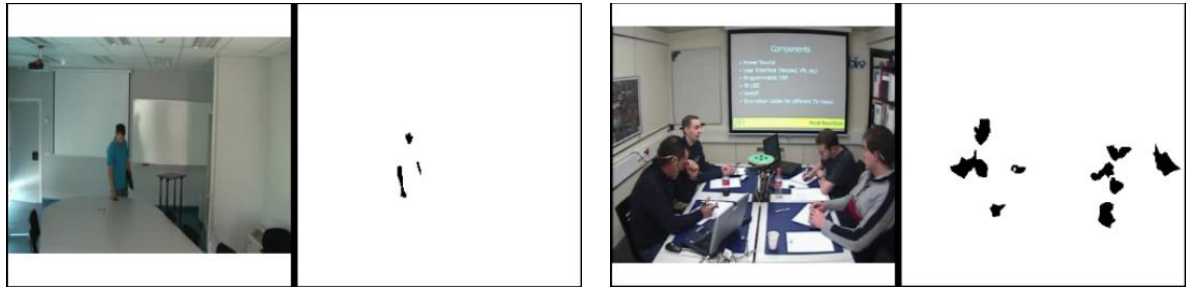
Datasets

SFA[15]
(1,118 images)

Pratheepan[117]
(78 images)

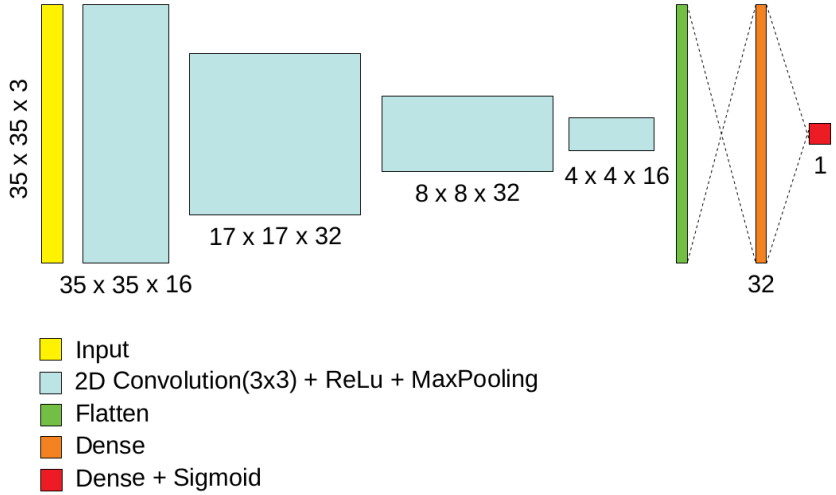
Compaq[51]
(4,670 images)

VPU[93]
(290 images)

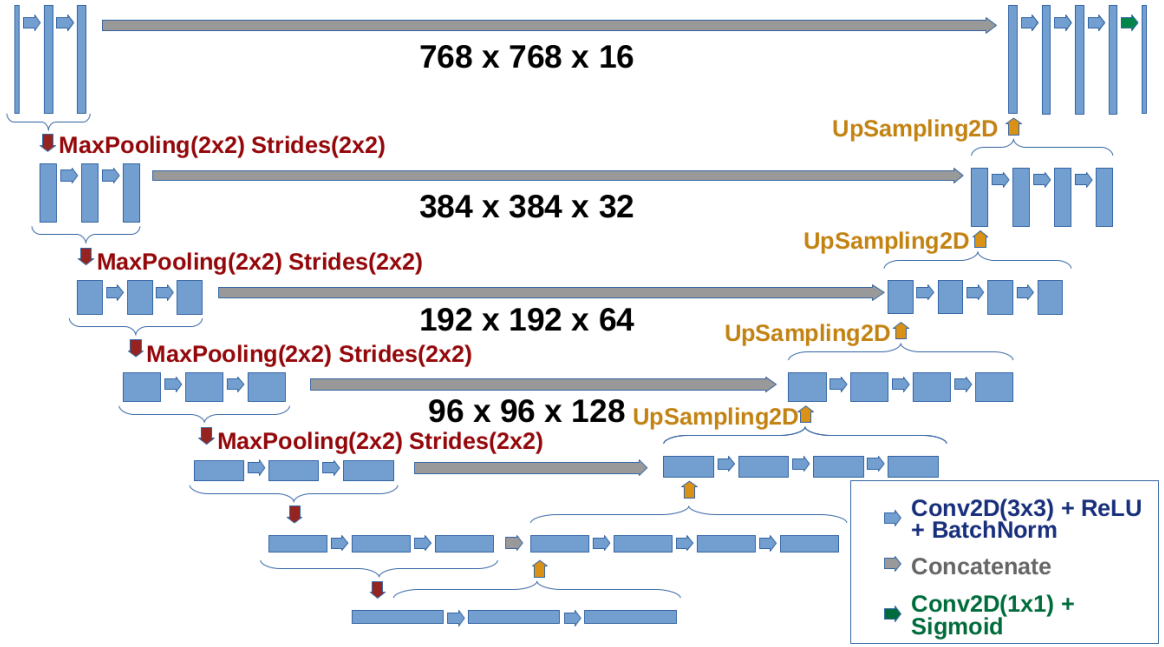


Models

Local, Patch-based CNN



Holistic, u-shaped FCN



In-domain evaluations and comparisons to the state-of-the-art

SFA dataset (in %)

Model	Acc	IoU	Prec	Recall	F_1
Faria and Hirata (2018) [33]	-	-	92.88	39.58	55.51
Our patch-based	91.14	82.17	89.71	91.00	90.35
Our U-Net	**97.94**	**92.80**	**96.65**	**95.89**	**96.27**

Compaq dataset (in %)

Model	Acc	IoU	Prec	Recall	F_1
Brancati *et al.* (2017) [12]	-	-	43.54	**80.46**	56.50
Our patch-based	90.18	46.00	58.92	73.59	65.45
Our U-Net	**92.62**	**54.47**	**68.49**	71.64	**70.03**

Pratheepan dataset (in %)

Model	Acc	IoU	Prec	Recall	F_1
Brancati *et al.* (2017) [12]	-	-	55.13	81.99	65.92
Faria and Hirata (2018) [33]	-	-	66.81	66.83	66.82
Our patch-based	87.12	55.57	59.83	**82.49**	69.36
Our U-Net	**91.75**	**60.43**	**72.91**	74.51	**73.70**

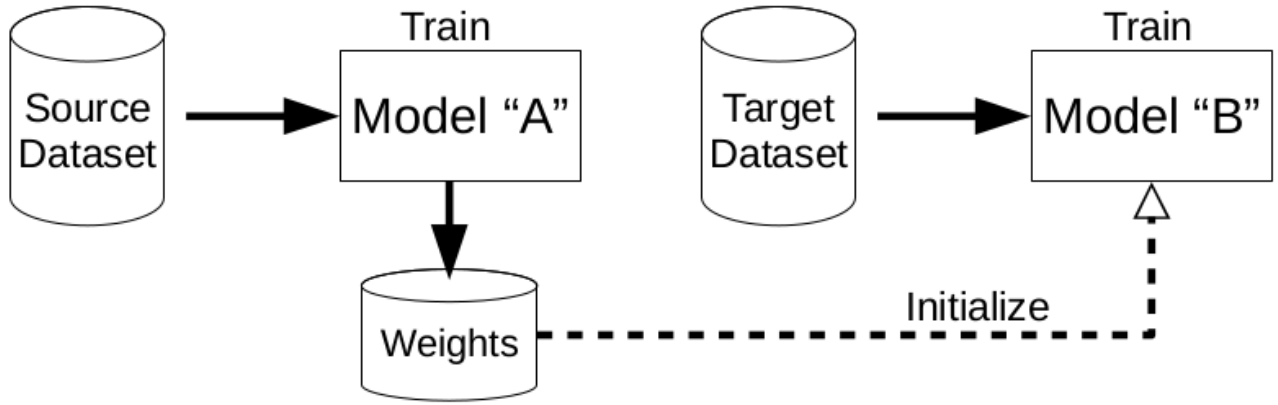
VPU dataset (in %)

Model	Acc	IoU	Prec	Recall	F_1
SanMiguel and Suja (2013) [93]	-	-	45.60	**73.90**	56.40
Our patch-based	93.48	14.14	46.34	42.82	44.51
Our U-Net	**99.04**	**45.29**	**57.86**	71.33	**63.90**

Cross-domain Baseline Results

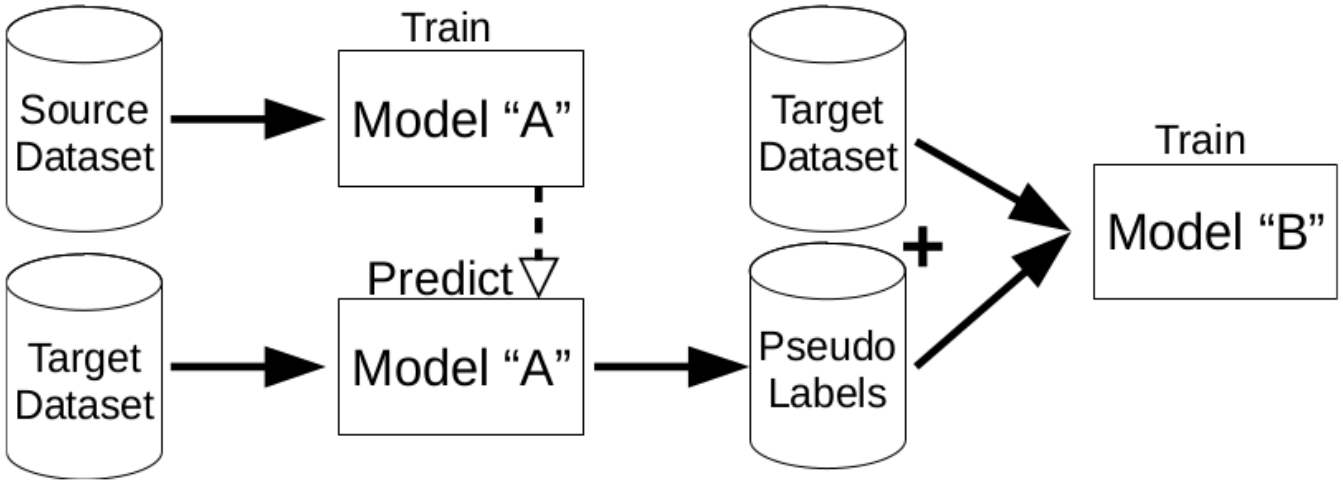
Model	Source Domain	Target Domain			
		SFA	Compaq	Prathee.	VPU
U-Net	SFA	-	18.92	44.98	11.52
	Compaq	**86.14**	-	**75.30**	23.67
	Prathee.	80.66	**63.49**	-	**36.68**
	VPU	14.83	44.71	48.02	-
Patch	SFA	-	54.80	62.92	21.60
	Compaq	71.28	-	72.59	19.94
	Prathee.	80.04	62.68	-	13.74
	VPU	82.63	51.48	58.34	-

Domain adaptation approaches



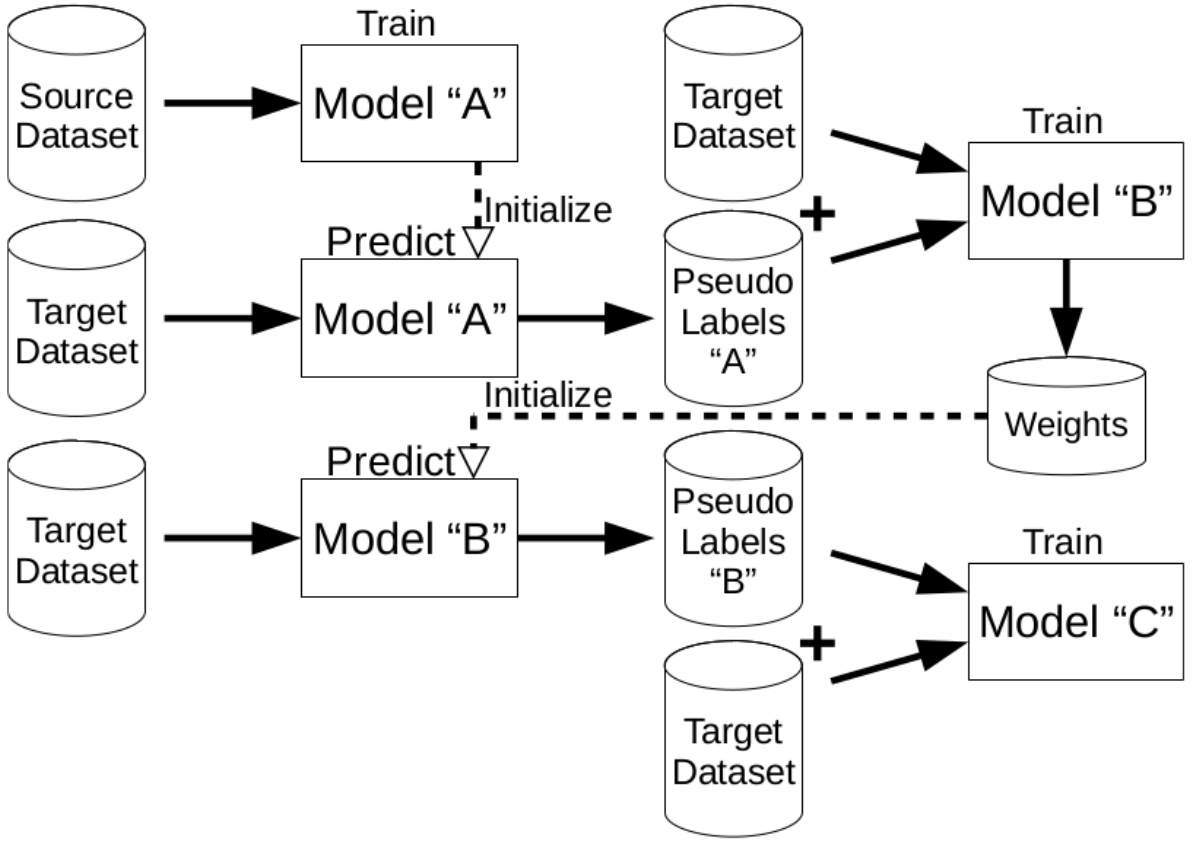
Inductive Transfer Learning by fine-tuning parameters of a model to a new domain

Domain adaptation approaches



Semi-supervised and unsupervised Domain Adaptation by cross-domain pseudo-labeling

Domain adaptation approaches

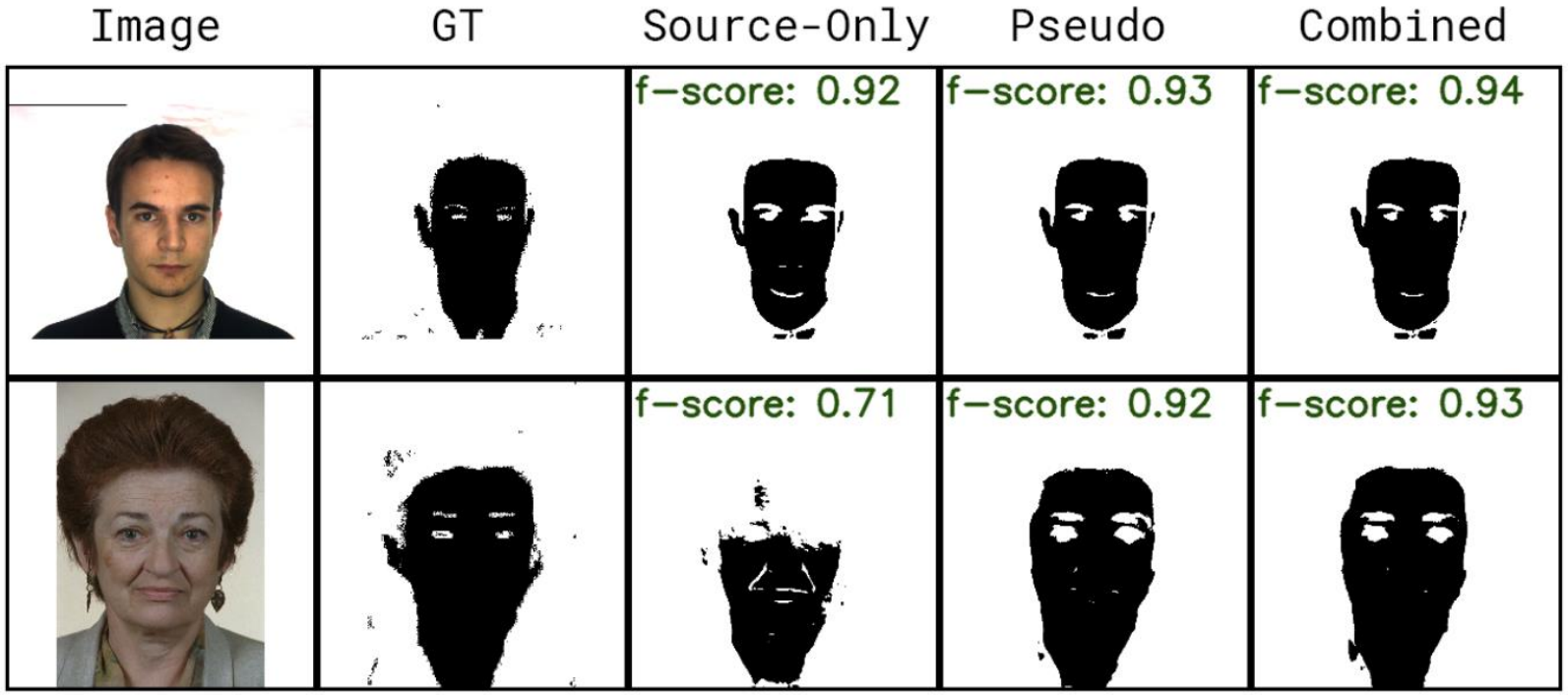


Combined transfer learning and domain adaptation approach

Domain adptation quantitative results

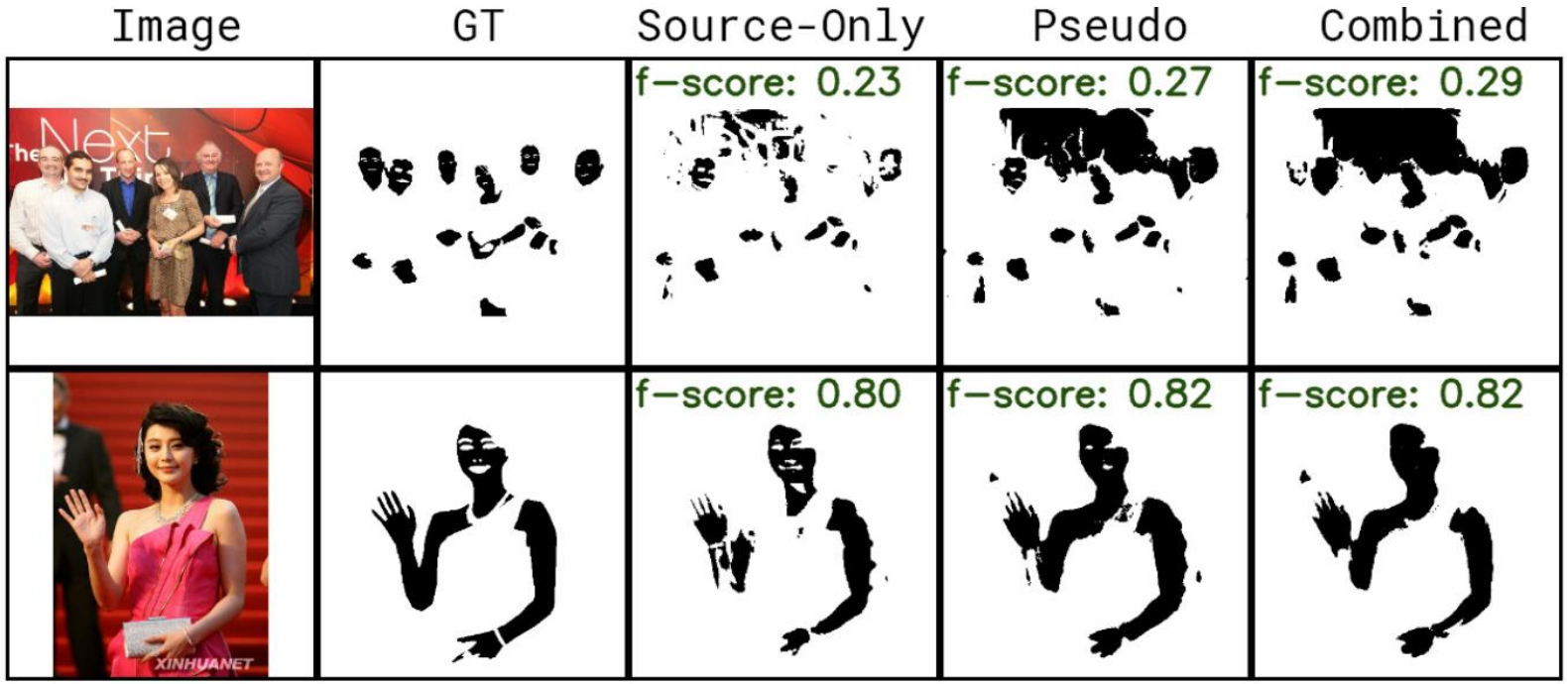
Source	Target	Approach	Target Training Label Usage				
			0%	5%	10%	50%	100%
Target only	SFA	Target only	-	93.49	94.50	95.72	**96.27**
	Compaq		-	**66.84**	**67.78**	**69.37**	70.03
	Pratheepan		-	46.36	59.86	69.04	73.70
	VPU		-	41.27	53.44	63.18	63.90
Compaq	SFA	Source only	86.14	-	-	-	-
		Fine-tuning only	-	92.89	94.04	**95.86**	95.98
		Cross-domain pseudo-label only	88.80	88.90	89.69	93.22	-
		Combined approach	**89.24**	90.05	90.36	94.57	-
	Pratheepan	Source only	75.30	-	-	-	-
		Fine-tuning only	-	72.52	74.69	76.47	**77.16**
		Cross-domain pseudo-label only	75.58	75.52	77.18	**80.08**	-
		Combined approach	**76.80**	**75.67**	**77.84**	79.87	-
	VPU	Source only	**23.67**	-	-	-	-
		Fine-tuning only	-	**51.51**	46.50	67.47	**69.62**
		Cross-domain pseudo-label only	02.67	02.86	02.68	02.77	-
		Combined approach	02.66	02.68	02.67	02.66	-
Pratheepan	SFA	Source only	80.66	-	-	-	-
		Fine-tuning only	-	**93.68**	**94.70**	**95.69**	95.99
		Cross-domain pseudo-label only	82.50	83.36	83.63	90.60	-
		Combined approach	**82.96**	84.12	84.47	92.93	-
	Compaq	Source only	**63.49**	-	-	-	-
		Fine-tuning only	-	64.88	66.10	68.97	**70.52**
		Cross-domain pseudo-label only	39.50	41.26	44.69	62.39	-
		Combined approach	34.72	36.22	39.05	57.06	-
	VPU	Source only	**36.68**	-	-	-	-
		Fine-tuning only	-	**51.61**	**60.19**	**68.15**	**69.44**
		Cross-domain pseudo-label only	02.66	02.66	02.67	02.77	-
		Combined approach	02.65	02.66	02.67	02.74	-

Domain adptation qualitative results



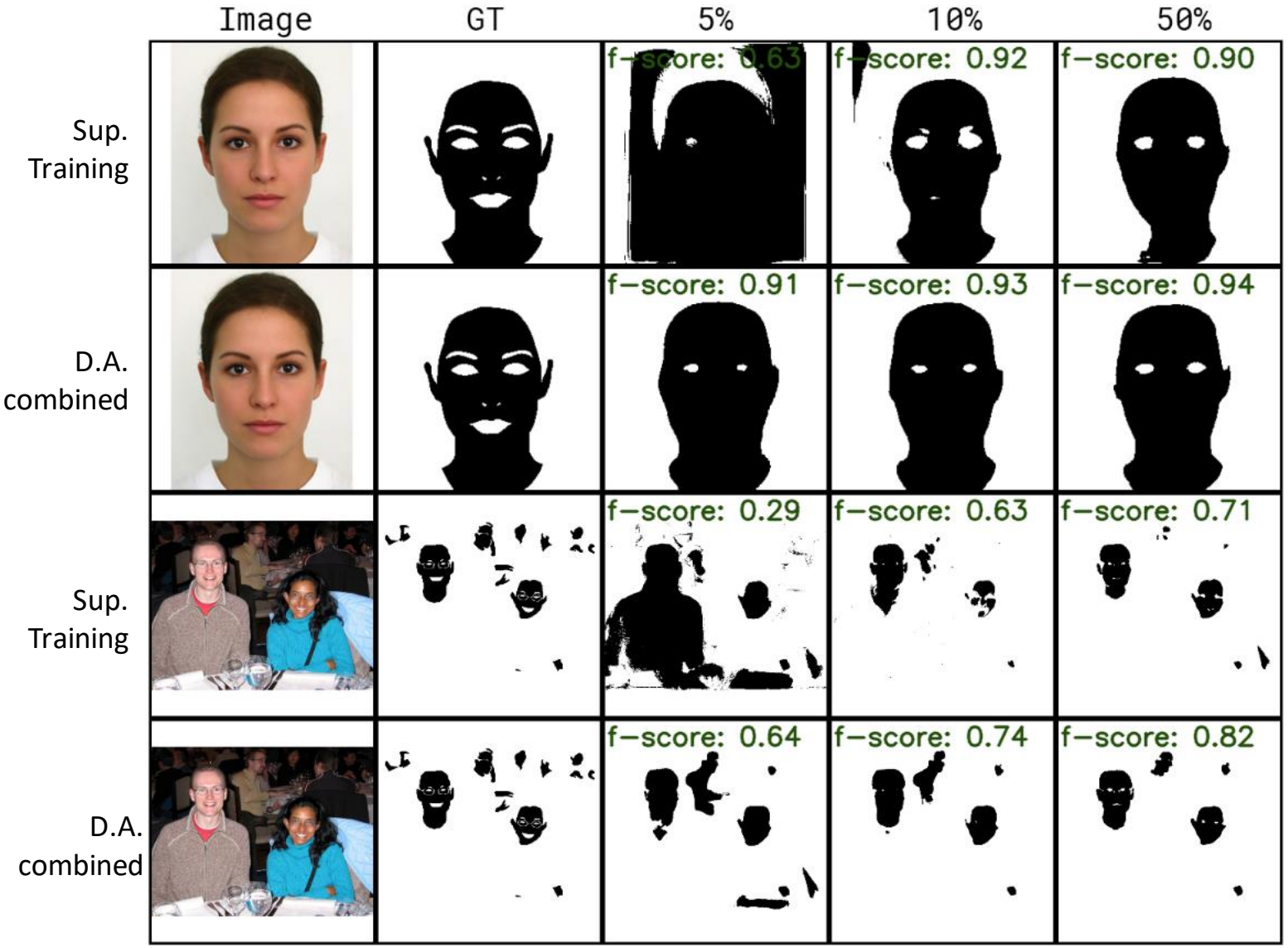
Domain adaptation from Compaq to SFA using no real labels from target

Domain adptation qualitative results



Domain adaptation from Compaq to Pratheepan using no real labels from target

Supervised training vs domain adaptation



Comparison of source only vs. domain adaptation combined approach in the Compaq→Pratheepan scenario

Conclusions

We refuted some common criticisms regarding the use of Deep Convolutional Networks for skin segmentation

In-domain:

- We compared two CNN approaches (patch-based and holistic) to the state-of-the-art pixel-based solutions for skin detection
- Our U-Net model obtained F1 scores which were on average 30% better than state-of-the-art recent published color based results
- In more homogeneous and clean datasets, like SFA, our F1 score was 73% better
- We experimentaly showed that an holistic approach like U-Net, besides being much faster, gives better results than a patch-based local approach
- Our experiments also showed that a FCN based solution is fast enough for real-time applications

Conclusions

We also concluded that the common critique of lack of generalization of CNNs does not hold true against our experimental data.

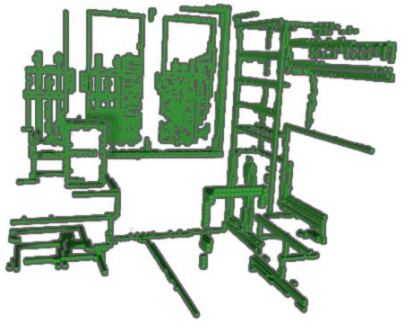
Cross-domain:

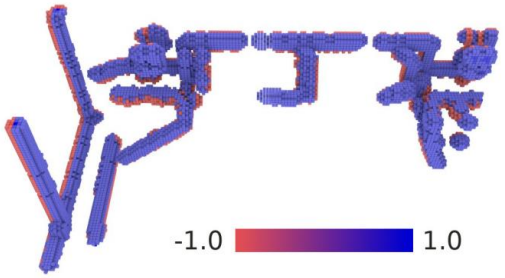
- We proposed novel approaches for semi-supervised and unsupervised domain adaptation applied to skin segmentation using CNNs
- With no labeled data on the target domain, our domain adaptation method's F1 score is an improvement of 60% over color-based results for homogeneous target datasets like SFA and 13% in heterogeneous datasets like Pratheepan.
- Despite the simplicity of the chosen methods, they greatly contribute to the improvement in the performance on skin segmentation across different datasets

Chapter 4

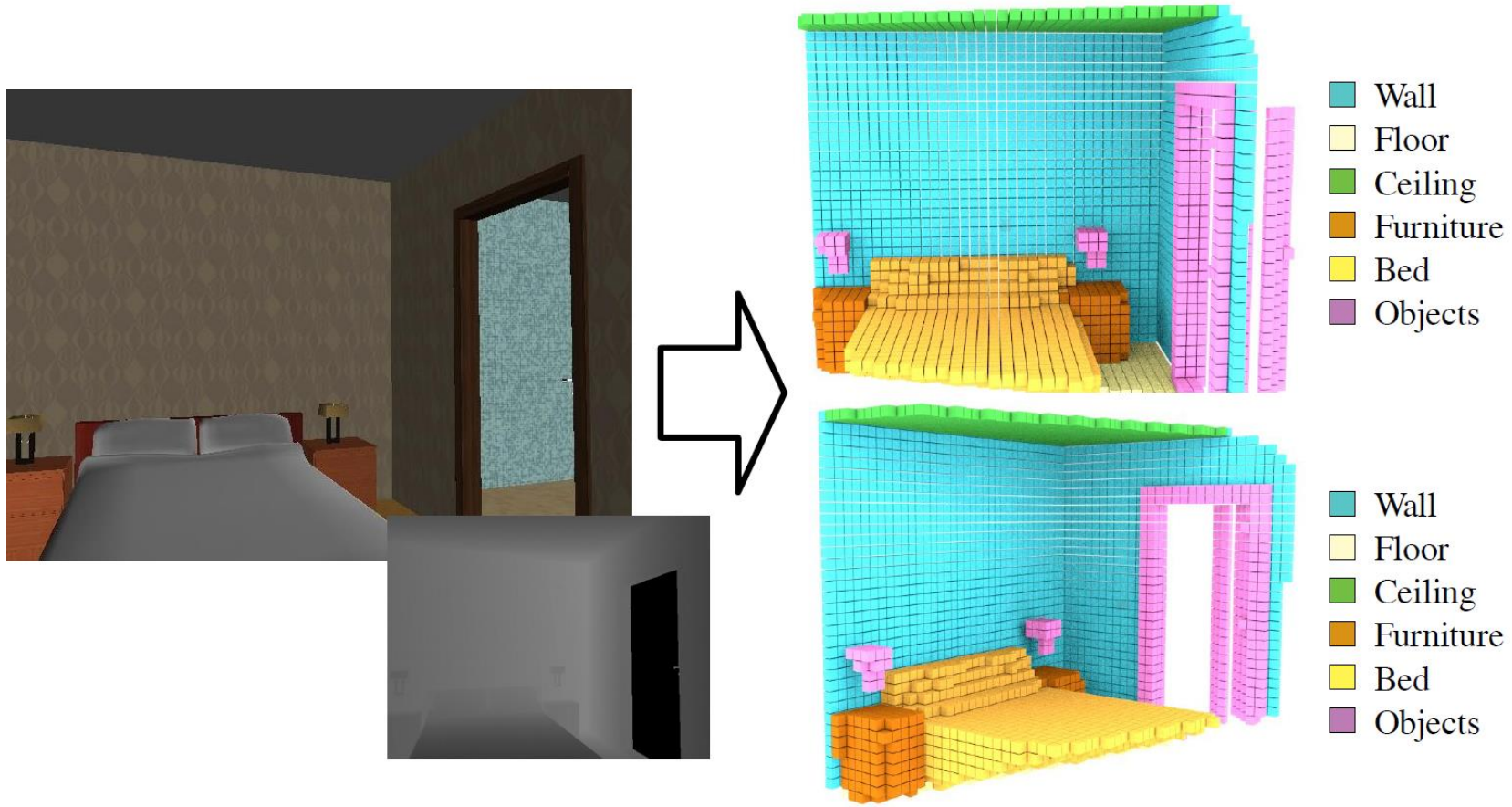
Using RGB Edges to improve Semantic Scene Completion from RGB-D Images

EdgeNet: Semantic Scene Completion from RGB-D images





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

Previous Works

Depth maps only

- SSCNET: Song et al. [107]
  - Typical contracting only 3D CNN with dilated convolutions
  - Depth map encoded with F-TSDF
  - Weighted softmax loss
  - Train on SUNCG, Fine tune on NYU

[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

Previous Works

Depth maps only

- SSCNET: Song et al. [107]
  - Typical contracting only 3D CNN with dilated convolutions
  - Depth map encoded with F-TSDF
  - Weighted softmax loss
  - Train on SUNCG, Fine tune on NYU
- Zhang et al. [119]:
  - dense conditional random field

[119] Zhang, L., Wang, L., Zhang, X., Shen, P., Bennamoun, M., Zhu, G., Shah, S.A.A., and Song, J.: Semantic scene completion with dense CRF from a single depth image. Neurocomputing, 318:182–195, Nov. 2018, ISSN 09252312. https://doi.org/10.1016/j.neucom.2018.08.052. 2, 4, 18, 46, 52, 53

Previous Works

Depth maps only

- SSCNET: Song et al. [107]
  - Typical contracting only 3D CNN with dilated convolutions
  - Depth map encoded with F-TSDF
  - Weighted softmax loss
  - Train on SUNCG, Fine tune on NYU
- Zhang et al. [119]:
  - dense conditional random field
- Guo and Tong [40]:
  - applied a sequence of 2D convolutions to the depth maps
  - used a projection layer to projected the features to 3D and

[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

- SSCNET: Song et al. [107]
  - Typical contracting only 3D CNN with dilated convolutions
  - Depth map encoded with F-TSDF
  - Weighted softmax loss
  - Train on SUNCG, Fine tune on NYU
- Zhang et al. [119]:
  - dense conditional random field
- Guo and Tong [40]:
  - applied a sequence of 2D convolutions to the depth maps
  - used a projection layer to projected the features to 3D and

Neglects the RGB channels from the input data

Previous Works

Depth maps plus RGB

- Guedes *et al*.[38]
  - 3 channels of RGB data projected to 3D
  - Same architecture as SSCNET
  - no significant improvement

[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]
  - 3 channels of RGB data projected to 3D
  - Same architecture as SSCNET
  - no significant improvement

Suffers from RGB data sparsity after projection to 3D

Previous Works

Depth map plus 2D segmentation

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

[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

- Two stream 3D semantic scene completion: Garbade *et al*.[36]
  - 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.*[70]
  - 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

[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

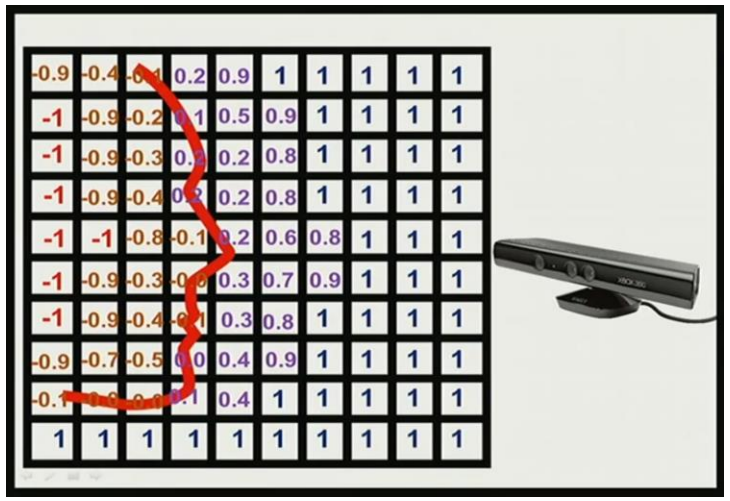
Depth map plus 2D segmentation

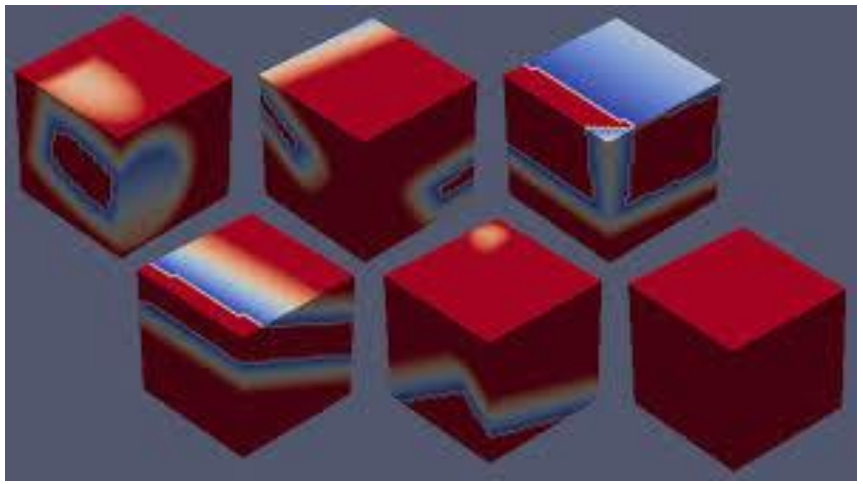
- Two stream 3D semantic scene completion: Garbade *et al.*[36]
  - 2D pretrained segmentation CNN and a Fully Connected CRF to generate a 2D segmentation map from RGB
  - Predicted 2D labels are projected to 3D and fused to the 3D branch (no one-hot-encoding)
- TNetFusion: Liu *et al.*[70]
  - depth 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

Requires a complex two step training procedure

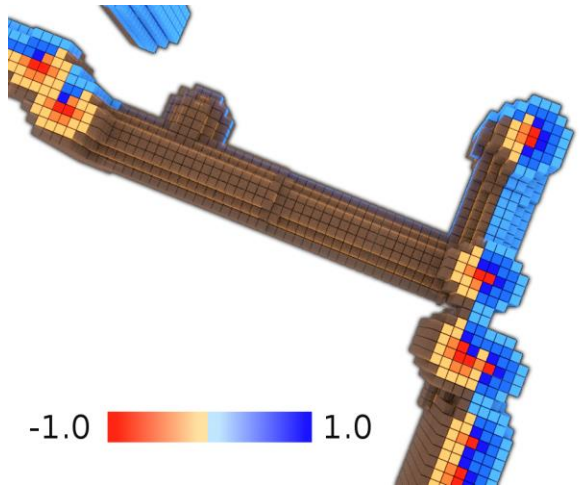
One note about previous works

- 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

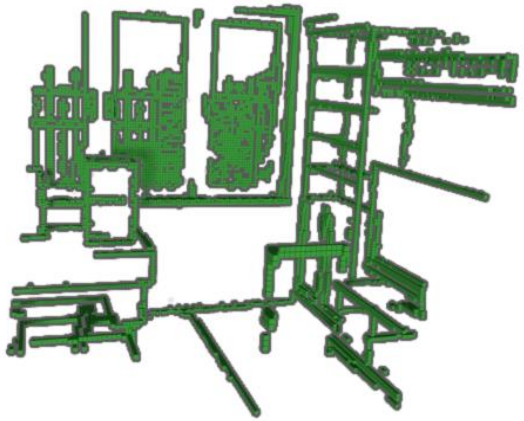


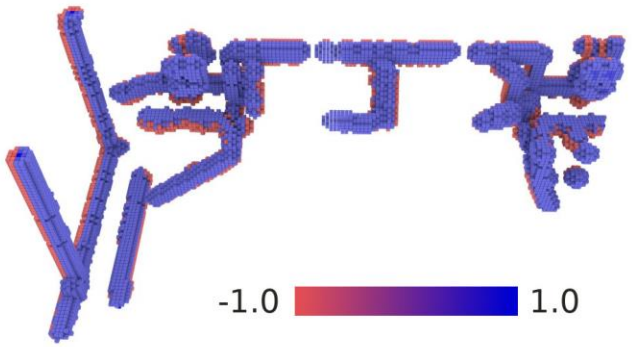
F-TSDF

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

Our approach: EdgeNet

- We extract information from RGB data using image edges:
  - Easy to get: Canny Edge Detector[14]
  - It is possible to apply F-TSDF to image edges (binary data)



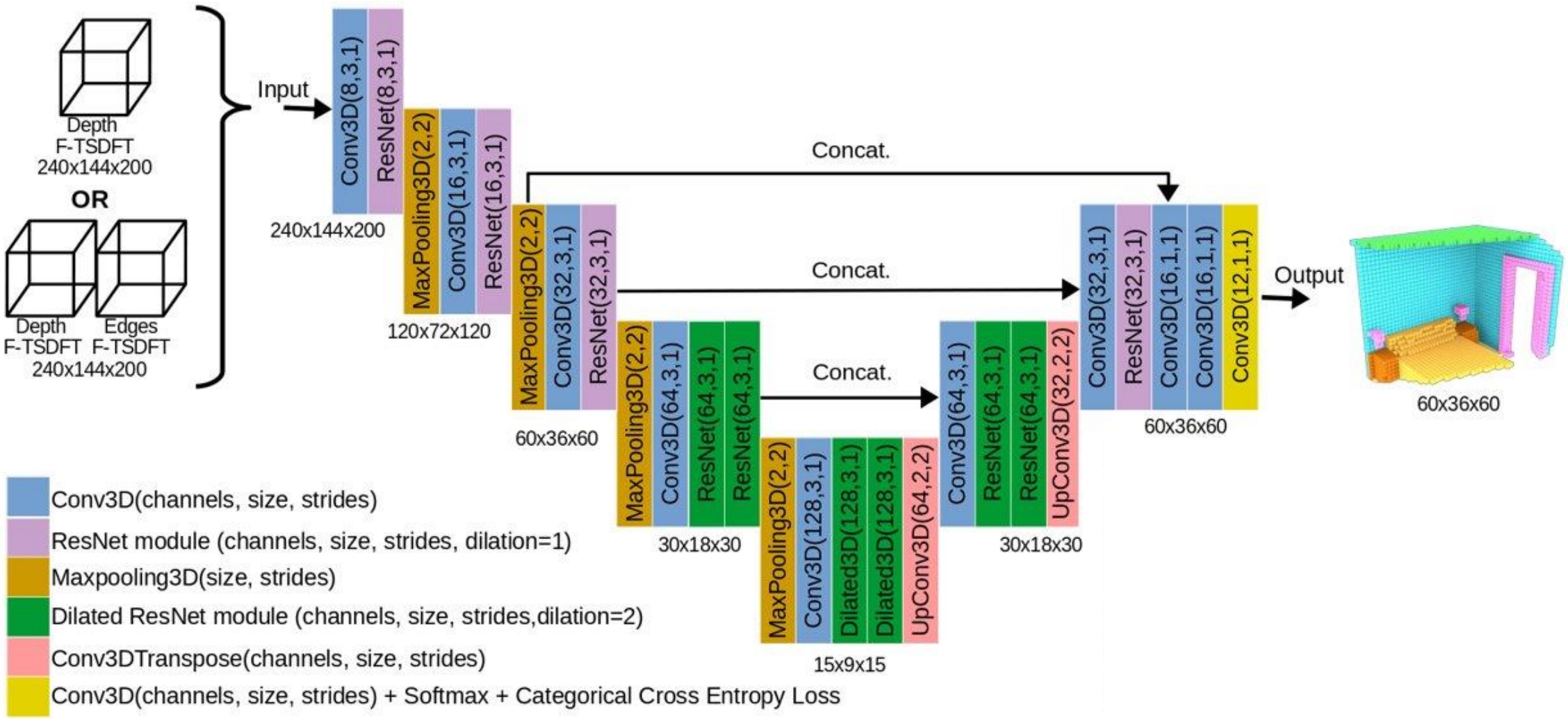


[14] Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 8(6):679–698, Nov 1986, ISSN 0162-8828.
48, 67

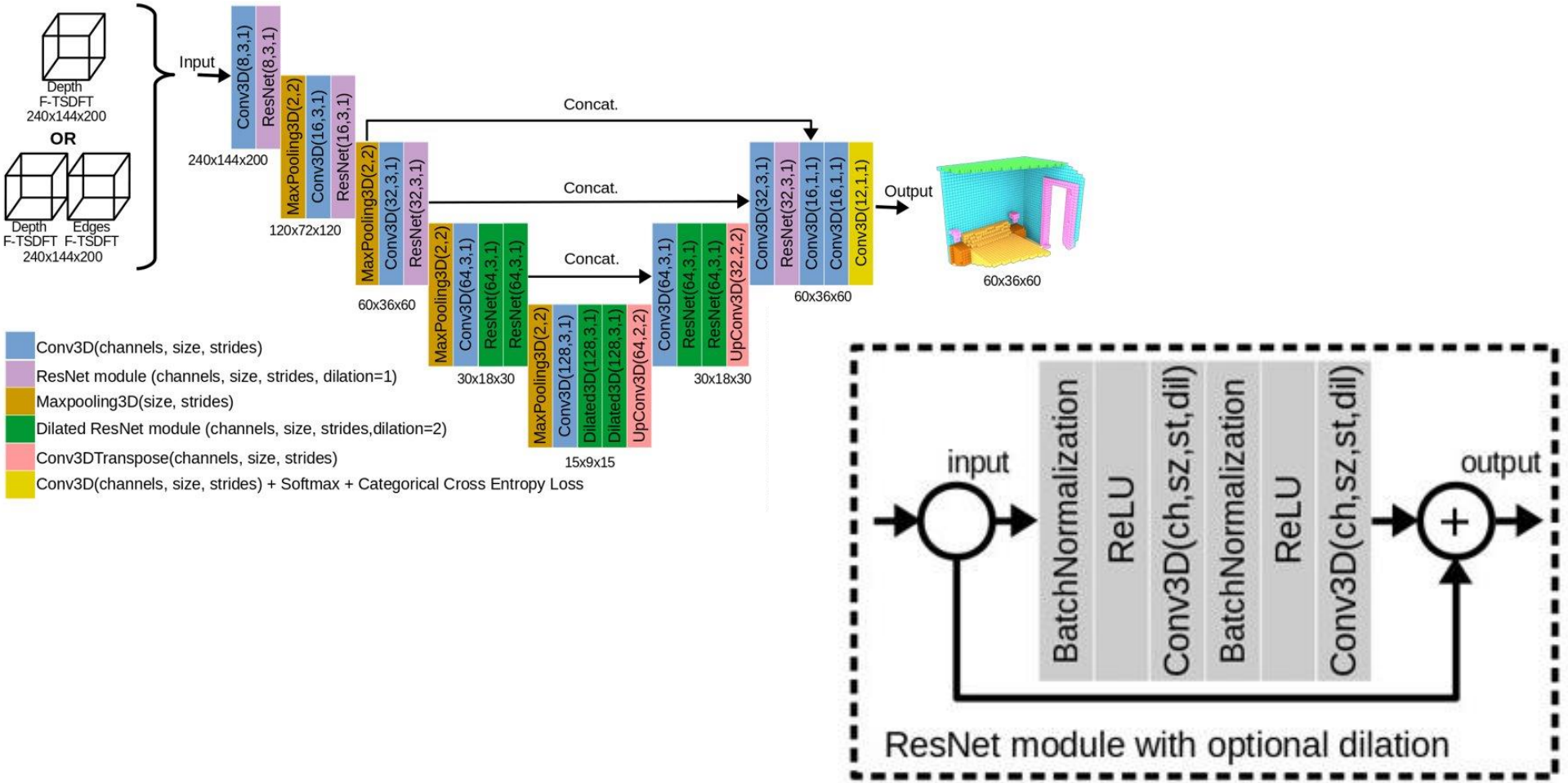
Our implementation

- Offline FTSD-F calculation using portable c++ cuda code
- We provide a software interface between cuda and python
- Preprocessing code is independent from the deep learning framework

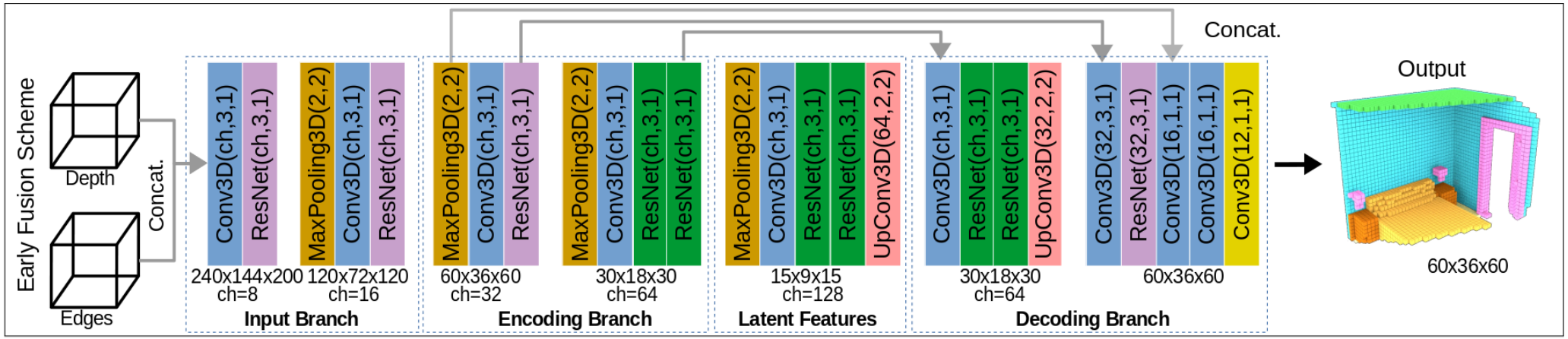
Network Architecture

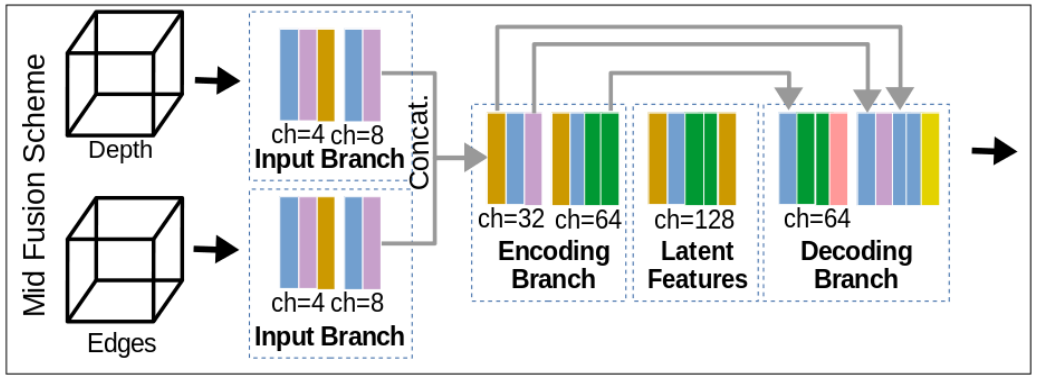


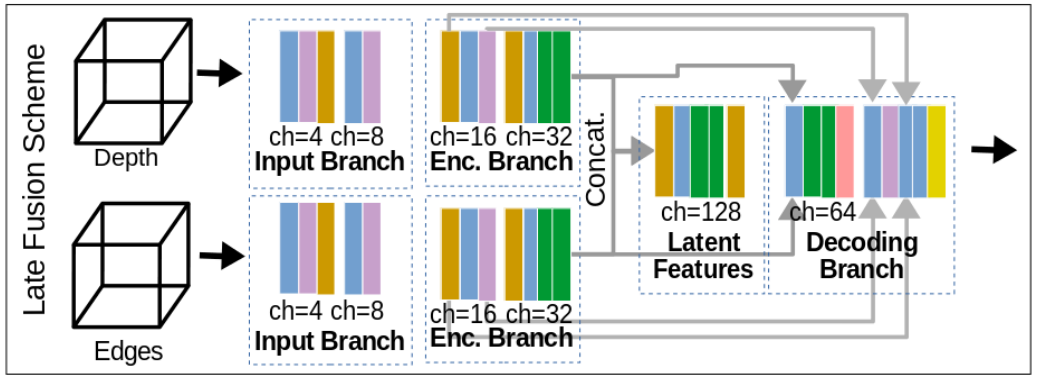
Network Architecture

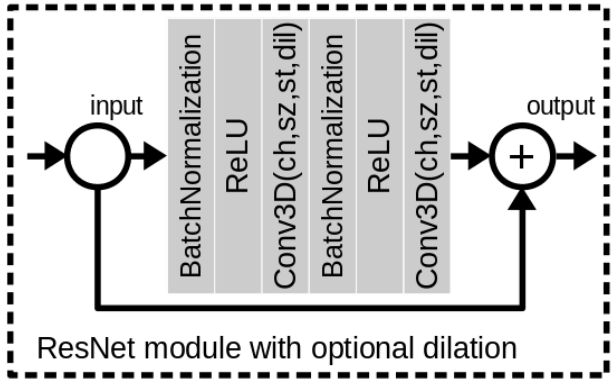


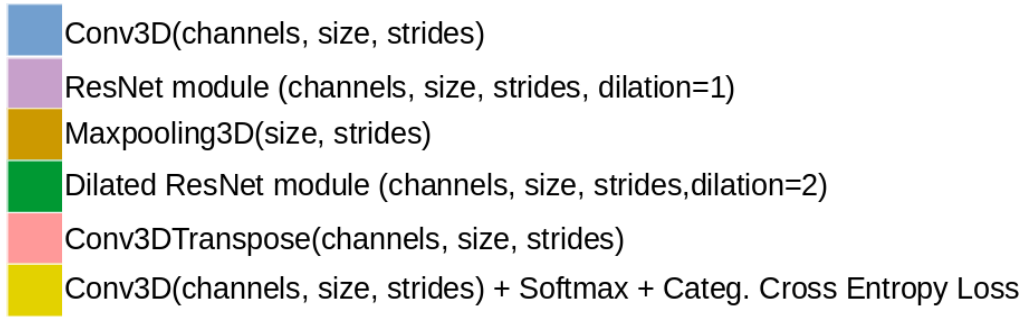
Network Architecture - Fusion Schemes











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
- NYUDv2
  - Real 3D scenes captured with Kinect
  - 795 scenes for training and 654 for testing

Training Protocol

- Train on SUNCG, fine tune on NYUD v2
- One Cycle Learning [103]
- SGD optimizer
- Batch size = 3
- training time (comparison to Song et al.[107]:
  - SUNCG: from 7 days to 4 days
  - NYU: from 30 hours to 6 hours

[103] Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. Tech. Rep. arXiv:1803.09820, Cornell University Library, 2018. http://arxiv.org/abs/1803.09820. 52

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

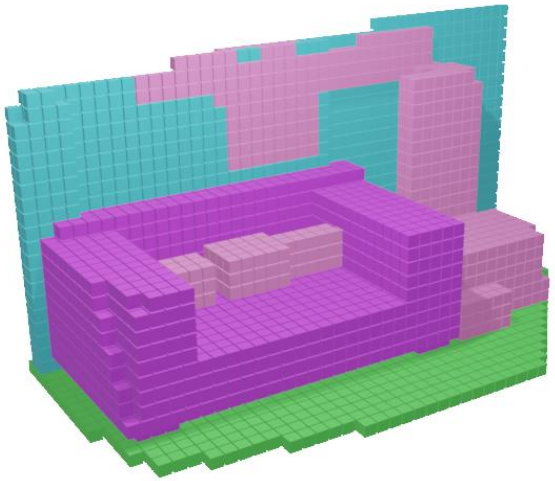
input	model	scene completion			semantic scene completion (IoU, in percentages)											
		prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
d	SSCNet[24]	76.3	**95.2**	73.5	96.3	84.9	56.8	28.2	21.3	56.0	52.7	33.7	10.9	44.3	25.4	46.4
	SSCNet*	92.7	89.7	83.8	97.0	94.6	74.3	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

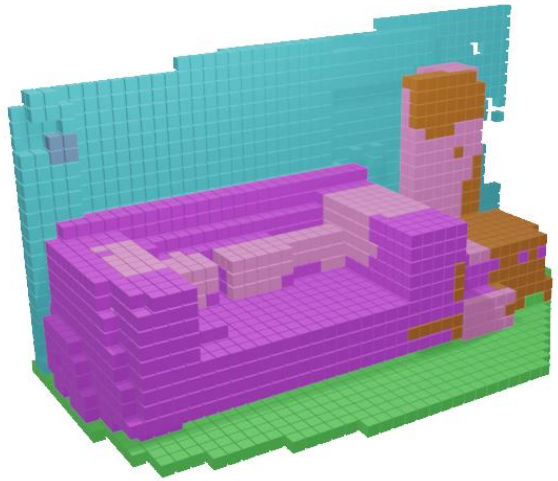
Results on NYU-DV2

train	input	model	scene completion			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

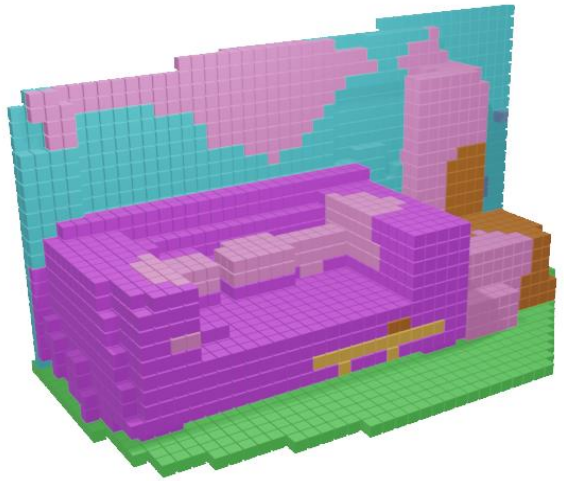
Qualitative Results



Ground Truth

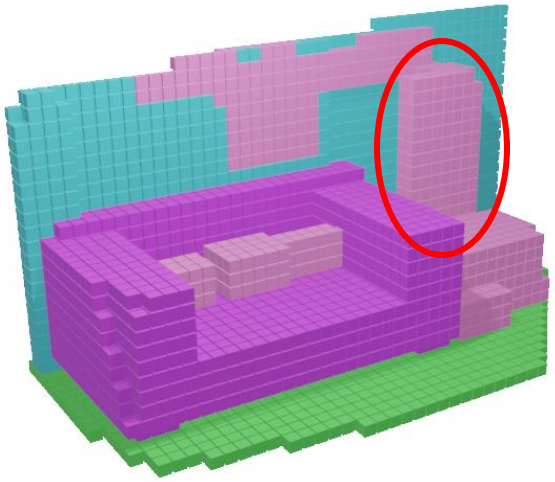


SSCNet

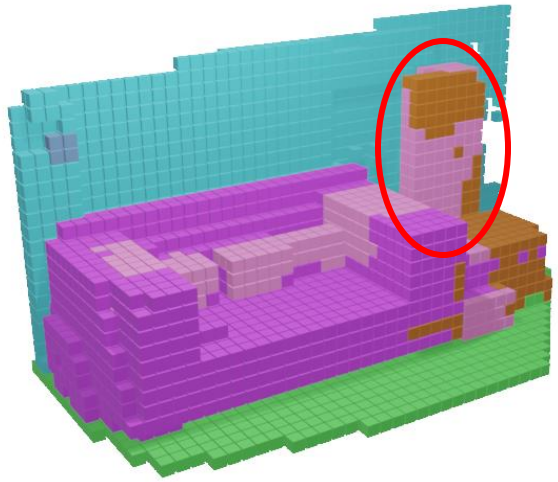


EdgeNet-MF

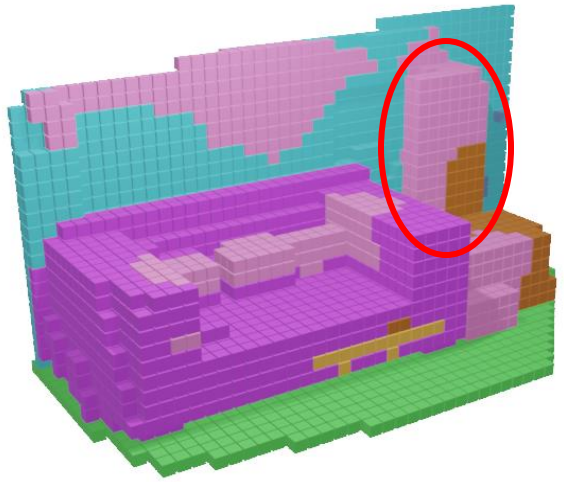
Qualitative Results



Ground Truth



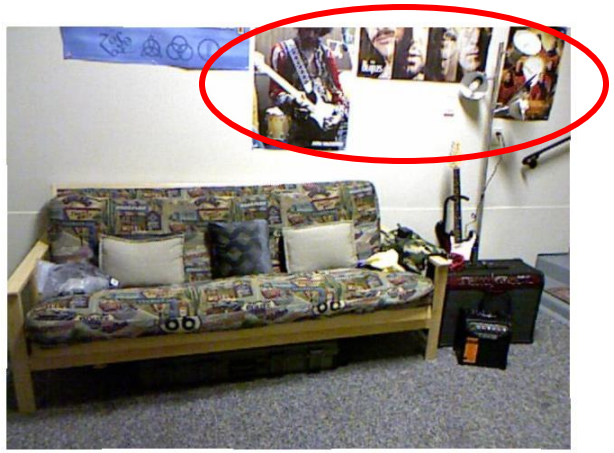
SSCNet

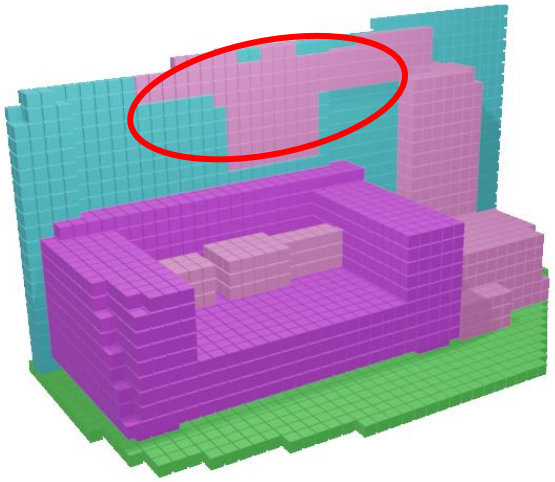


EdgeNet-MF

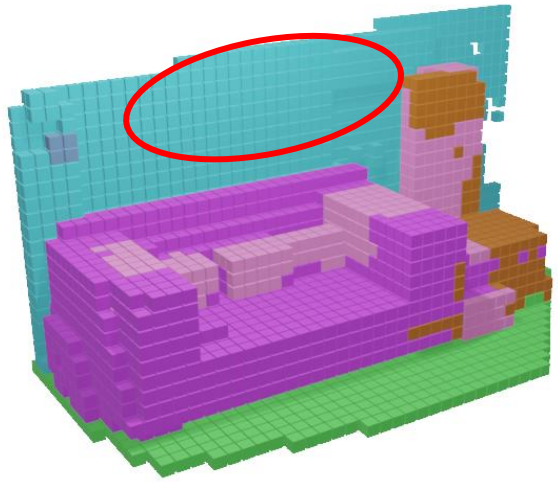
Higher overall accuracy

Qualitative Results

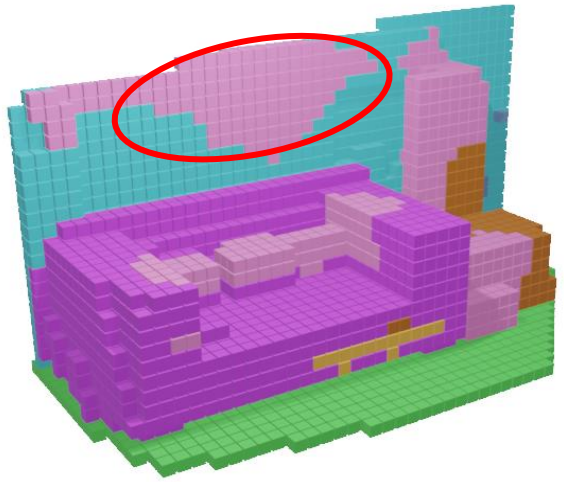




Ground Truth



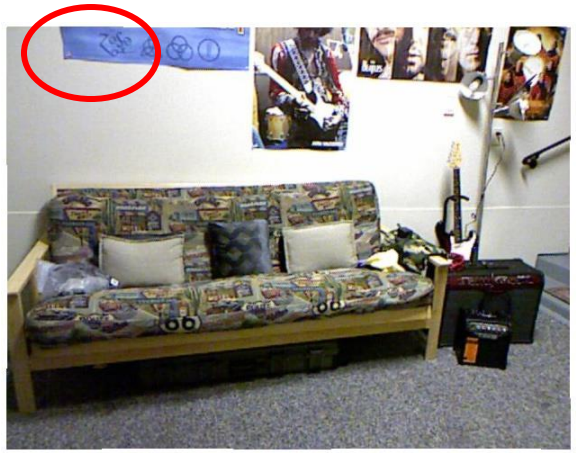
SSCNet

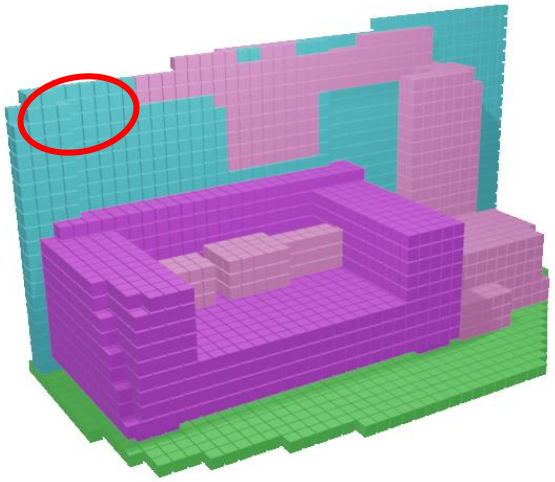


EdgeNet-MF

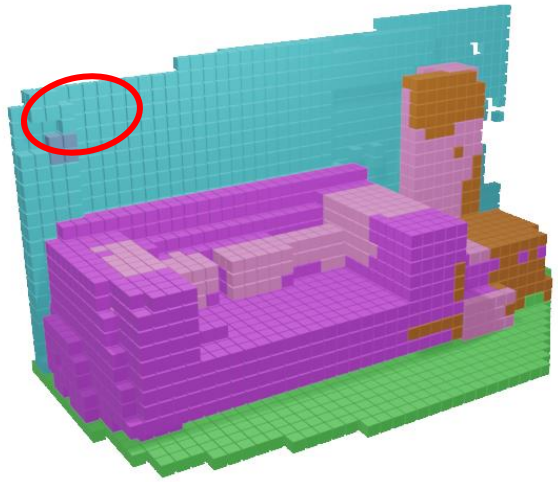
Hard-to-detect classes

Qualitative Results

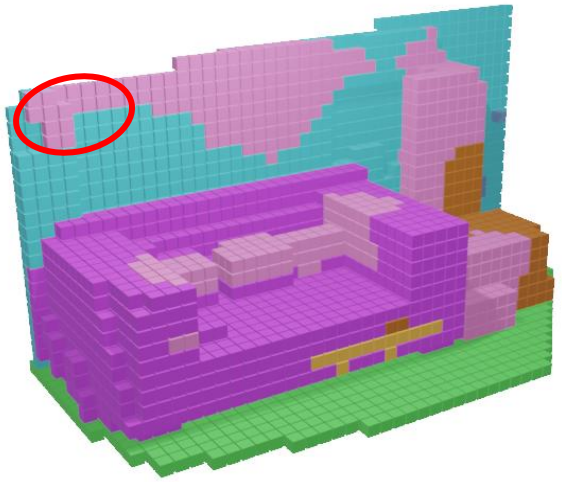




Ground Truth



SSCNet



EdgeNet-MF

NYU Ground Truth errors

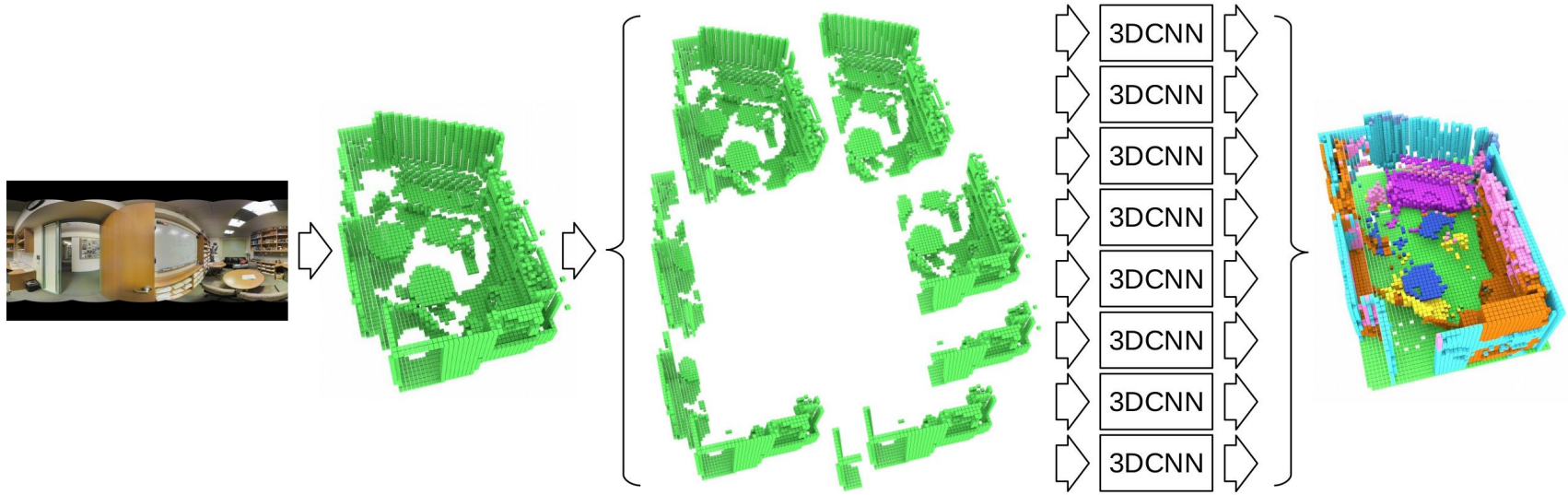
Conclusions

- We presented
  - a new approach to fuse depth and colour into a CNN for semantic scene completion
  - a new end-to-end network architecture capable of properly aggregating edges and depth
- 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 SUNCG dataset
- We surpased other end-to-end approaches on NYUv2 dataset
- We developed a faster lightweight training pipeline for the task, which showed to be efficient and effective

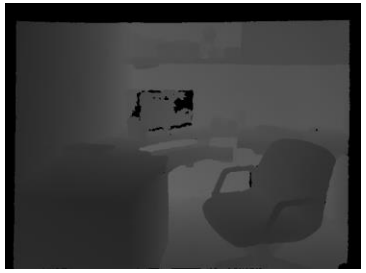
Chapter 5

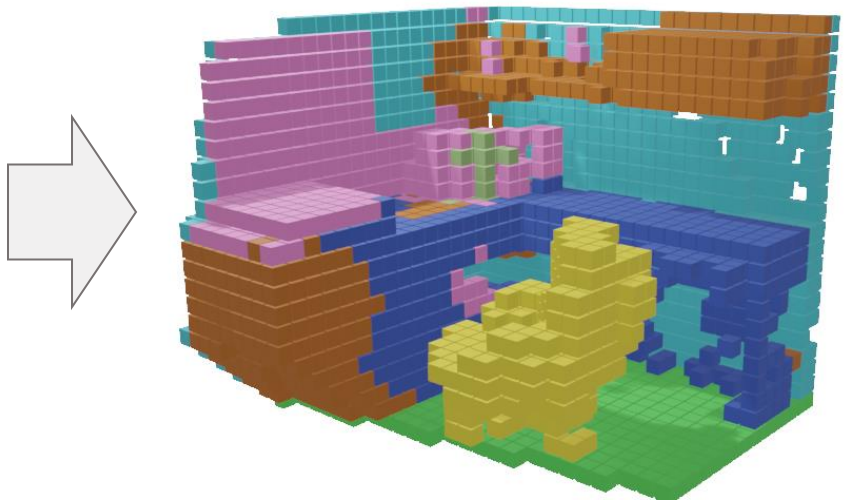
Extending Semantic Scene Completion for 360^{O} Coverage

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

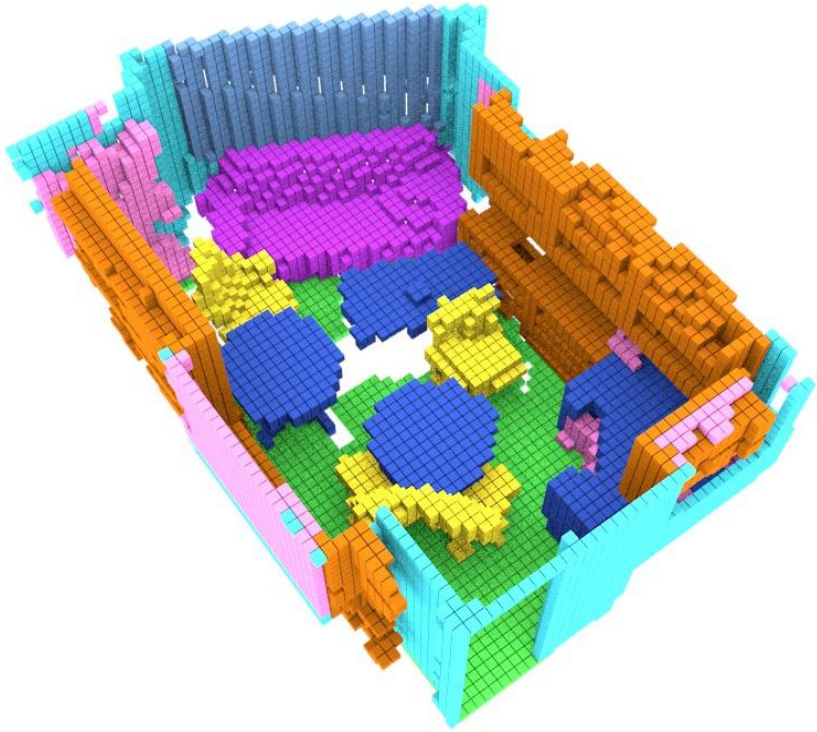


Current Semantic Scene Completion Limitations





Regular RGB-D Sensor

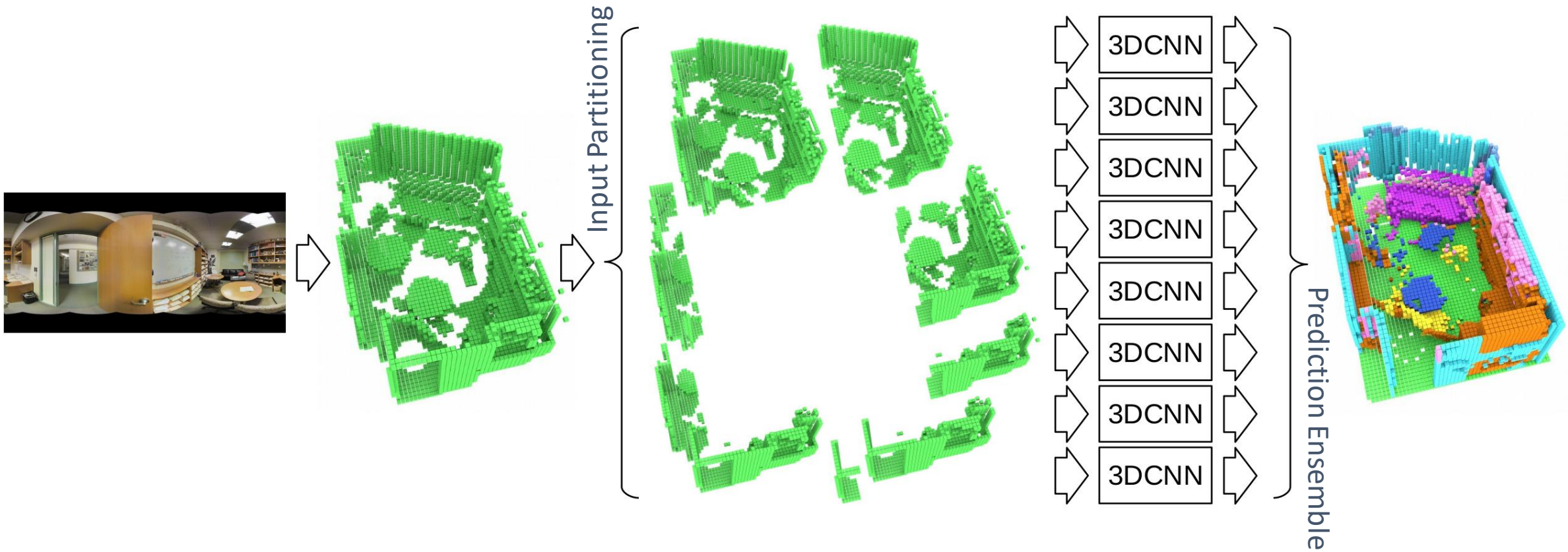


Panoramic Image from Matterport Camera

Obstacles to 360^{O} Semantic Scene Completion

- The task is highly memory consuming – a naive 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



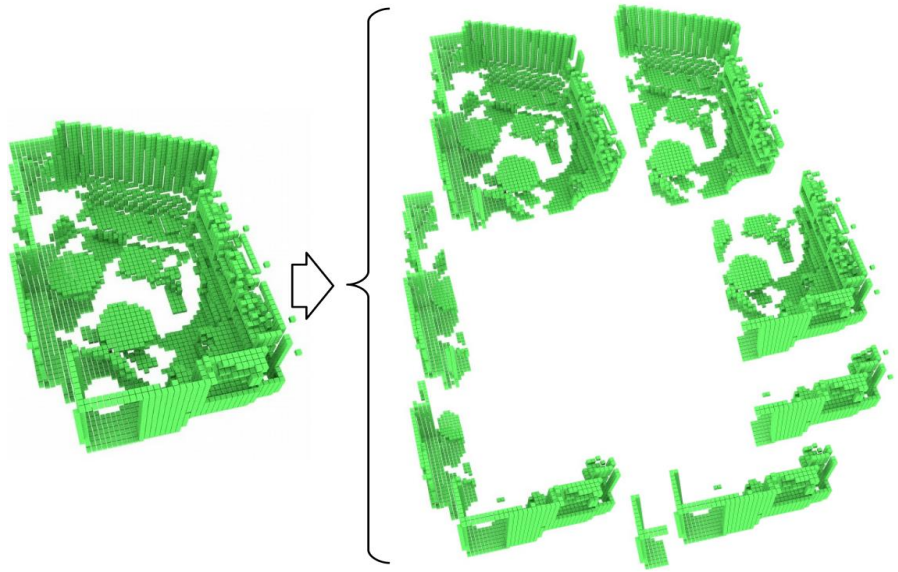
The 3DCNN is trained using SUNCG and fine-tuned in NYUDV2

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

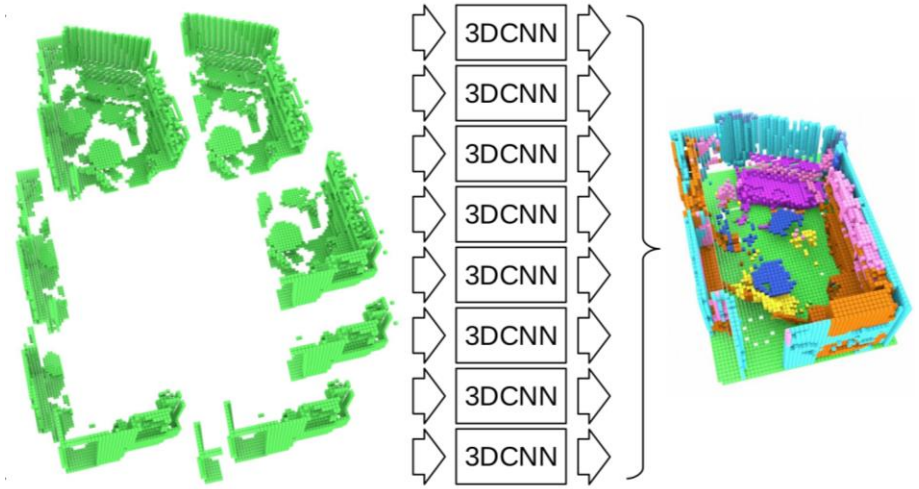
Input Partitioning



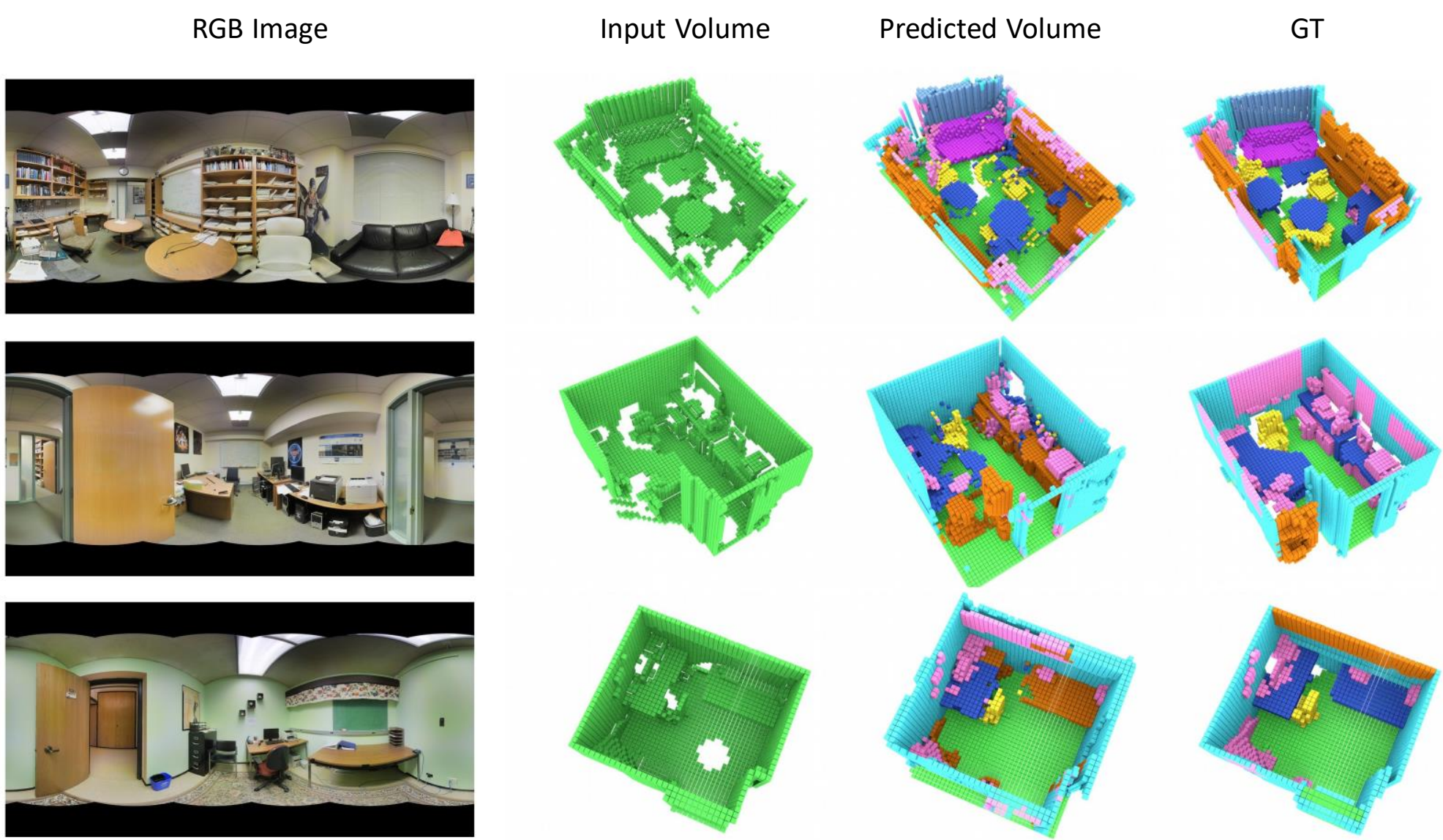
Our approach

- 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

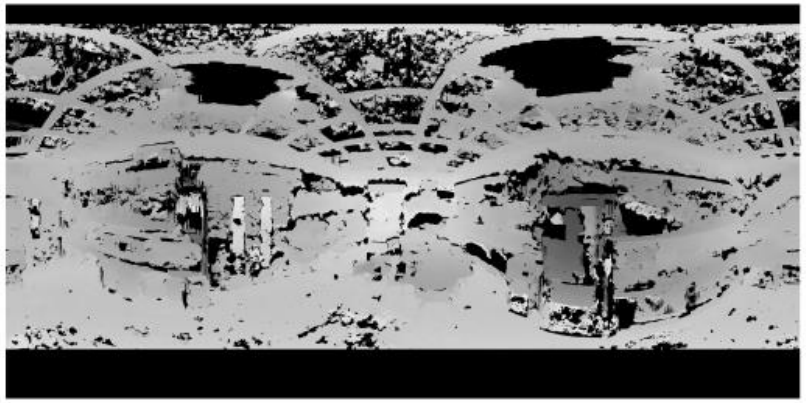


Results on Stanford 2D-3DS Dataset

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 scanning
- However, depth estimation is subject to errors due to occlusions between two camera views and correspondence matching errors



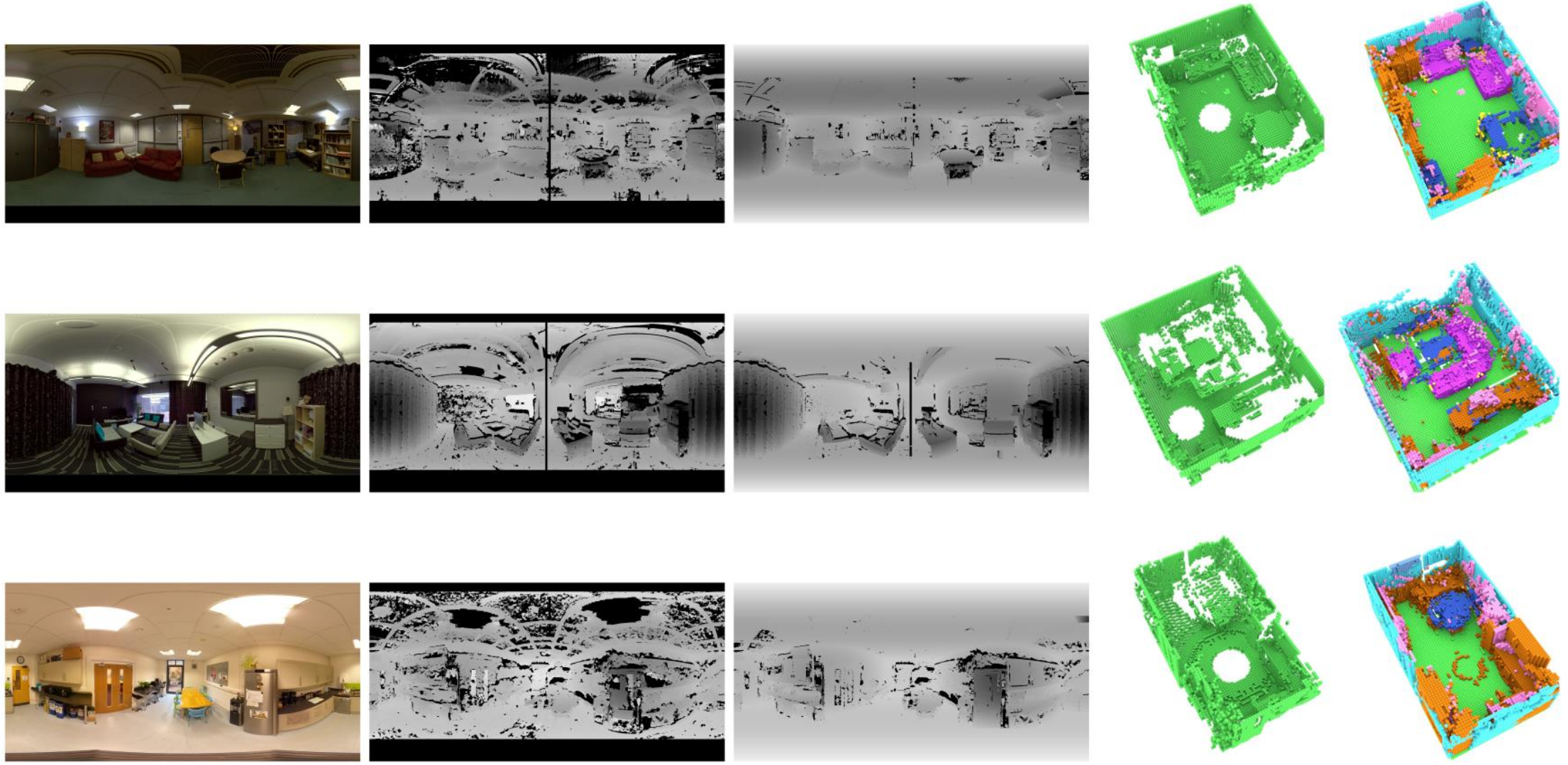
Our approach

- The scenes are captured as a vertical stereo image pair
- Dense stereo matching with spherical stereo geometry [56] is used to recover depth information
- 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

[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

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



floor wall window chair table sofa furn. objects

Conclusions

- We introduced the task of Semantic Scene Completion from a pair of 360^{o} 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 360^{o} cameras
- We evaluated on the publicly available Stanford 2D-3D-Semantics dataset and on a collection of 360^{o} 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.

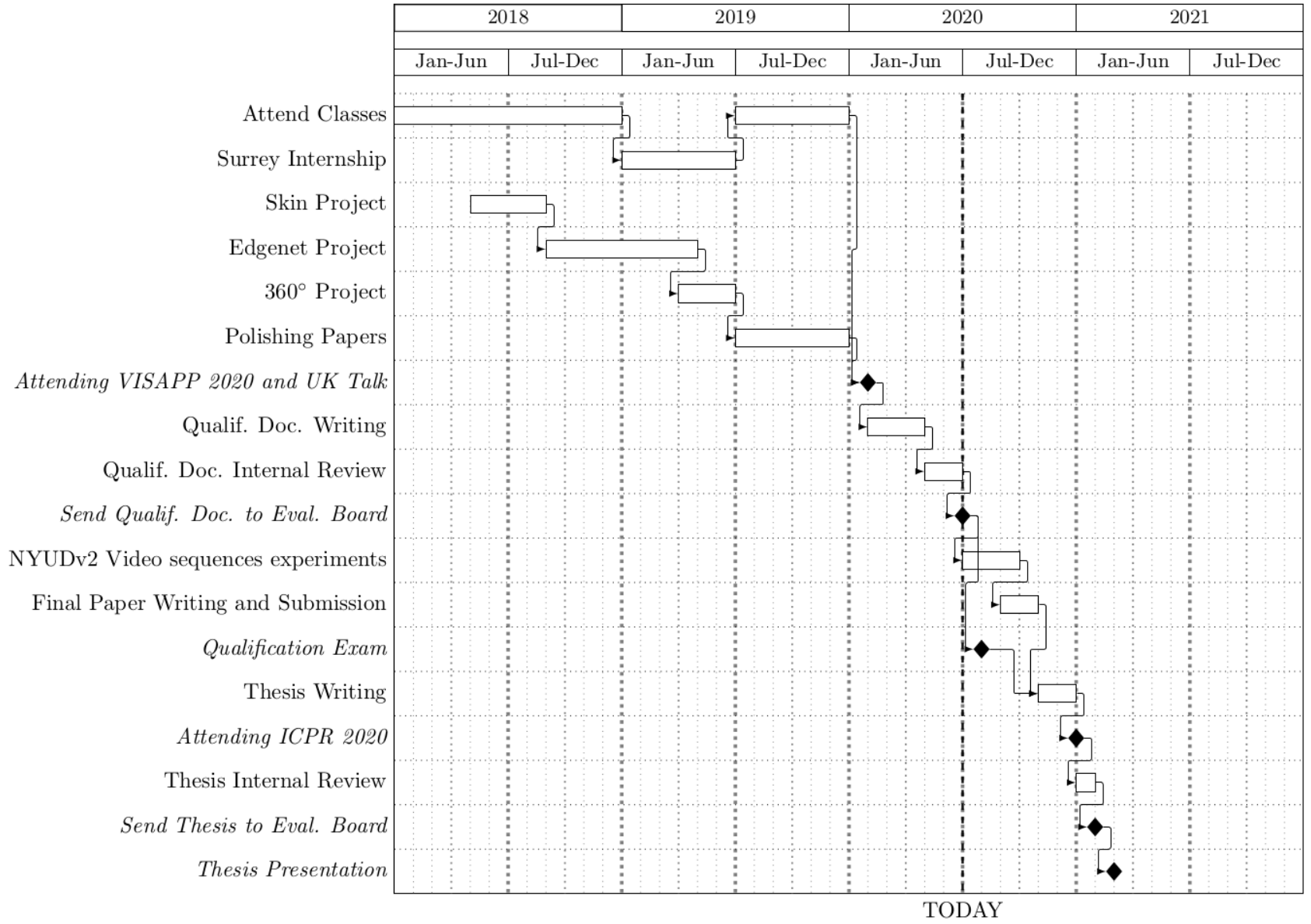
Chapter 6

Work Plan

Remaining Activities

- Review of the qualification document following feedback from the qualification exam review board
- Consolidating Chapters 4 to 5 into a single journal paper
- Missing experiments: review the 360^{O} solution, fine tune the network to the 360^{O} domain
- thesis writing
- attending ICPR 2020
- thesis presentation

Timeline


2018
2019
2020
2021
Jan-Jun
Jul-Dec
Jan-Jun
Jul-Dec
Jan-Jun
Jul-Dec
Jan-Jun
Jul-Dec
Attend Classes
Surrey Internship
Skin Project
Edgenet Project
360° Project
Polishing Papers
Attending VISAPP 2020 and UK Talk
Qualif. Doc. Writing
Qualif. Doc. Internal Review
Send Qualif. Doc. to Eval. Board
NYUDv2 Video sequences experiments
Final Paper Writing and Submission
Qualification Exam
Thesis Writing
Attending ICPR 2020
Thesis Internal Review
Send Thesis to Eval. Board
Thesis Presentation
TODAY

Thank you!