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

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

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

Aloisio Dourado Neto

Supervisor
Prof. Dr. Teófilo Emidio de Campos

Commitee

Prof. Dr. Gabriela Csurka
Never Labs Europe

Prof. Dr. Anderson de Rezende Rocha
Unicamp

Prof. Dr. Bruno Luiggi Macchiavello Espinoza
CIC/UnB

Prof. Dr. Vinicius Ruela Pereira Borges (suplente)
CIC/UnB

Presentation Outline

- Introduction (Chapter 1)
  - Motivation
  - Problem statement
  - Objectives
  - Publications
- Background and related concepts (Chapter 2)
- Previous works (Chapter 3)

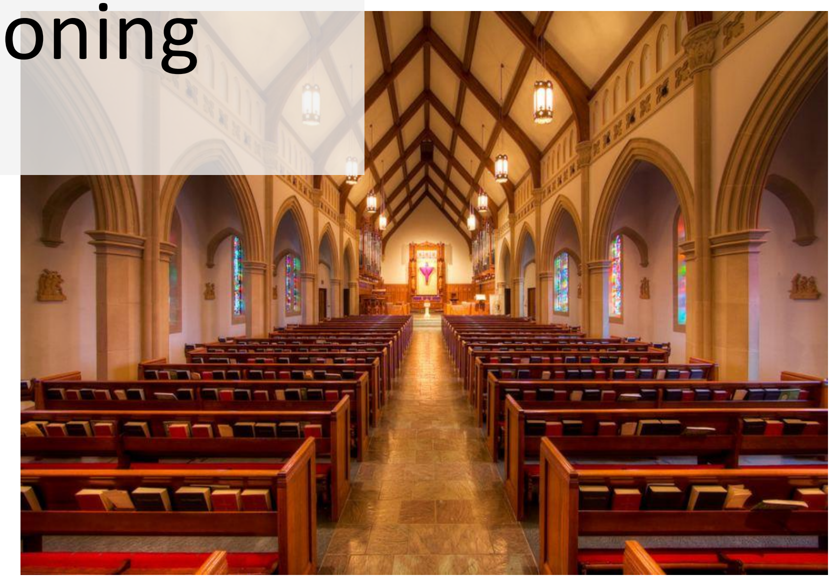
Presentation Outline

- Research steps (Chapters 4 to 8)
  - Semantic segmentation, FCN, domain adaptation, data augmentation and semi-supervision in 2D
    (Chapter 4)
  - First work in 3D: exploiting RGB input with EdgeNet
    (Chapter 5)
  - Going further in 3D: adding multiple input modes and data augmentation
    (Chapter 6)
  - Going even further in 3D: adding semi-supervision
    (Chapter 7)
  - Enhancing the field of view: 360 degree Semantic Scene Completion
    (Chapter 8)
- Conclusion (Chapter 9)

Chapter 1

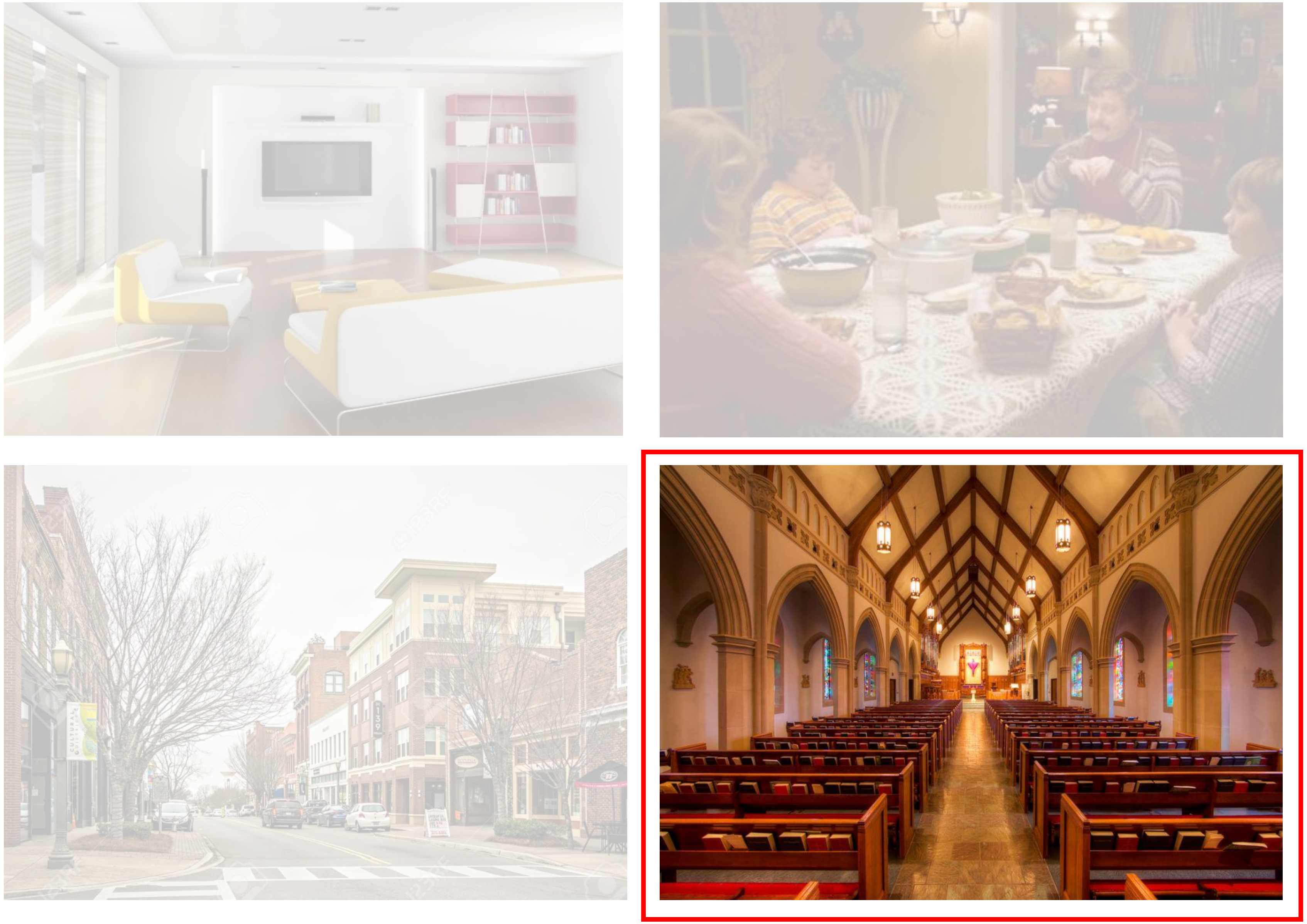
Introduction

3D Scene Reasoning



Chapter 1

Introduction



Chapter 1

Introduction

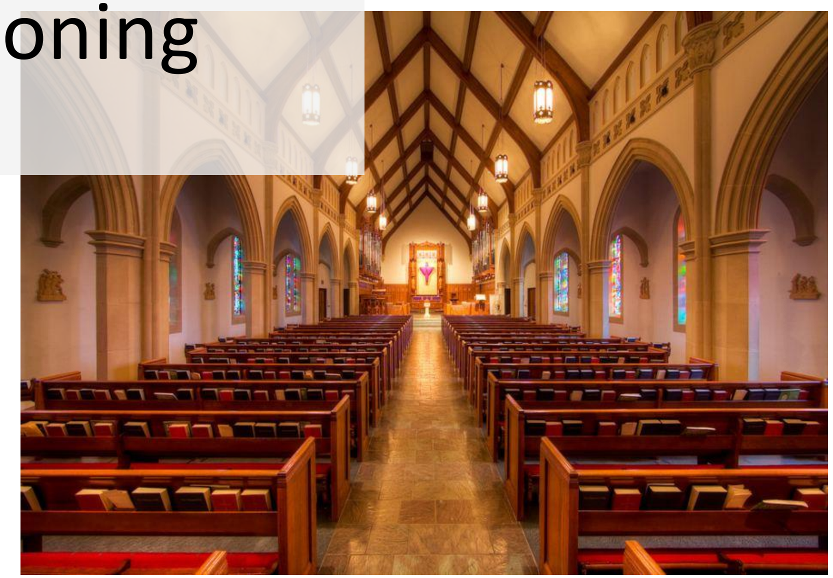
Chapter 1

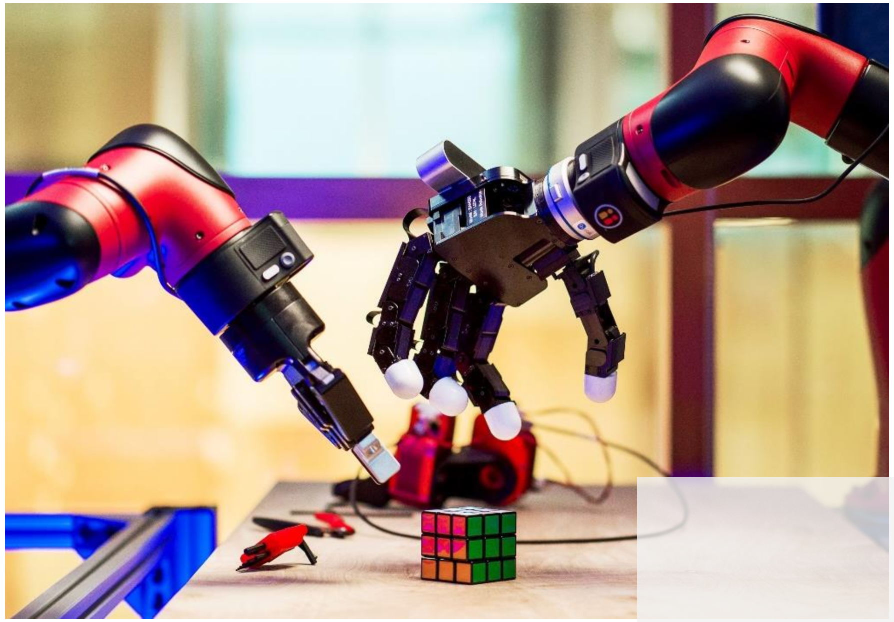
Introduction

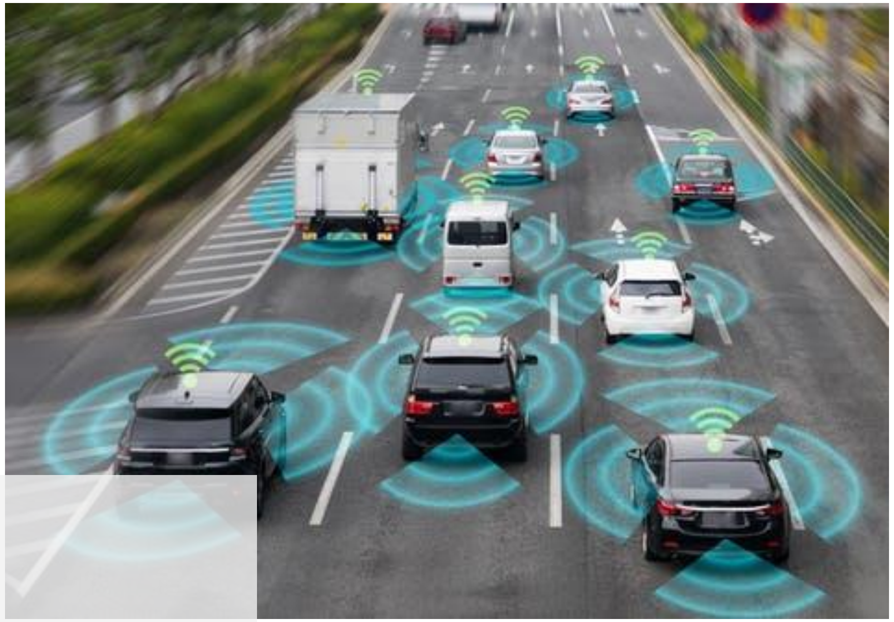
Chapter 1

Introduction

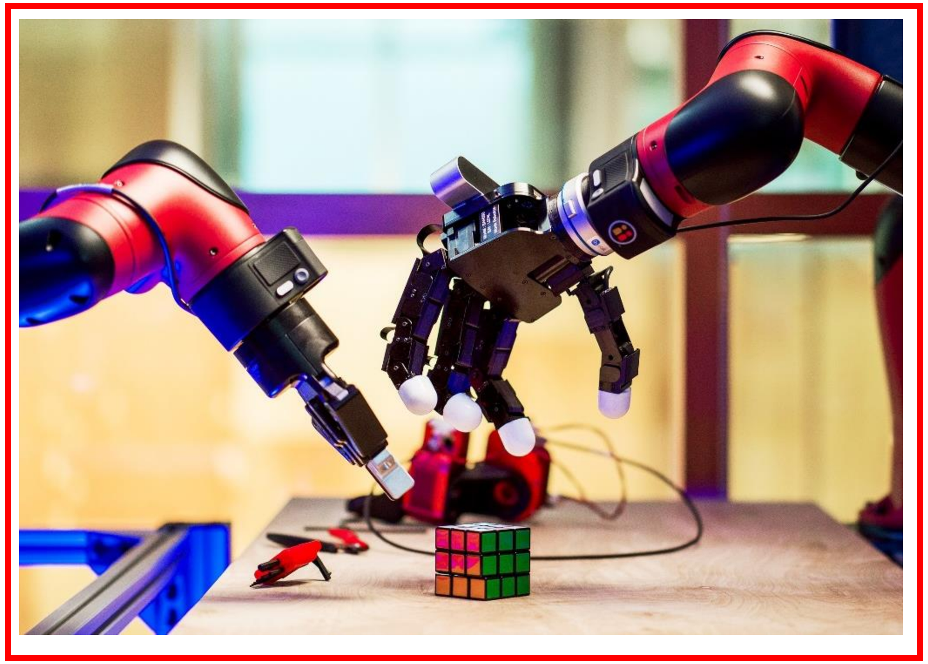

3D Scene Reasoning

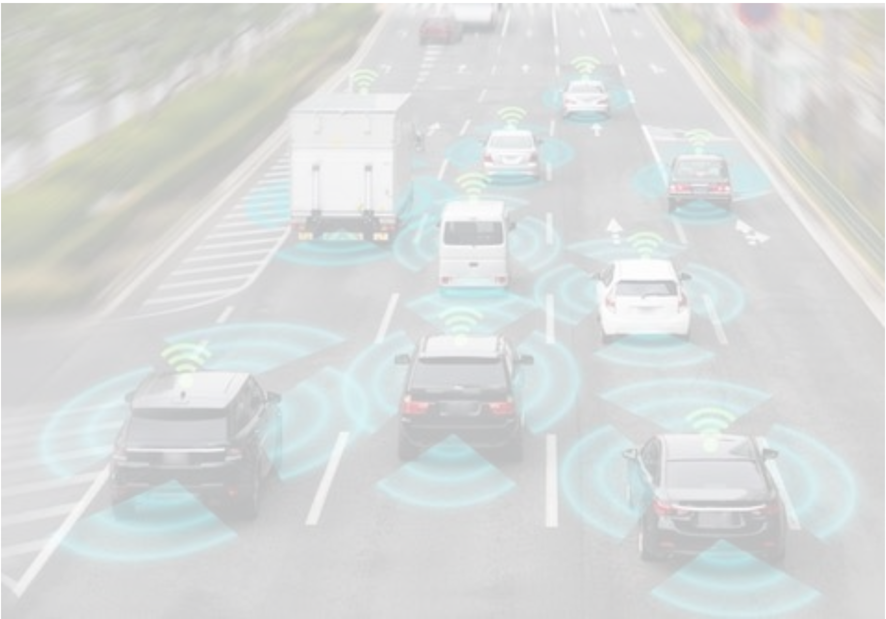



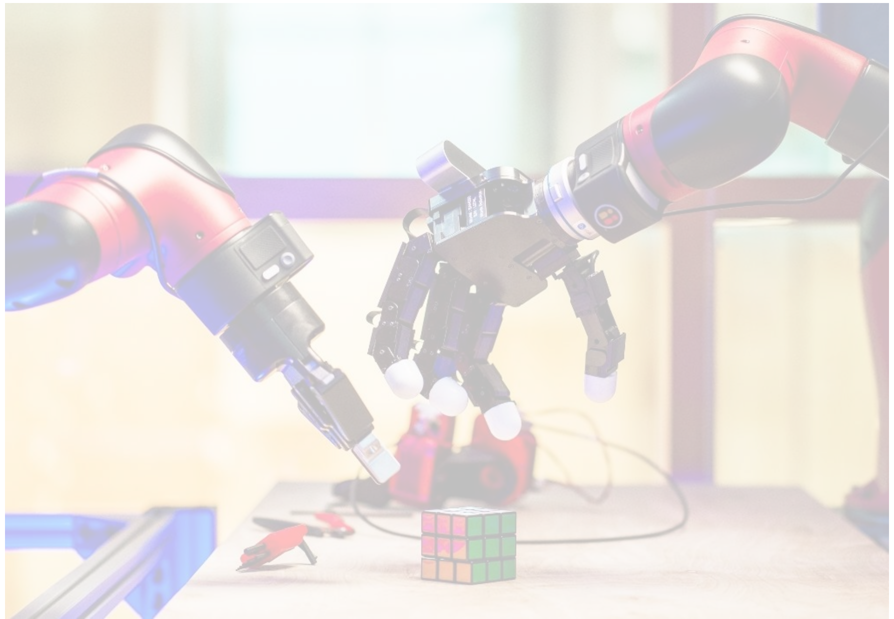


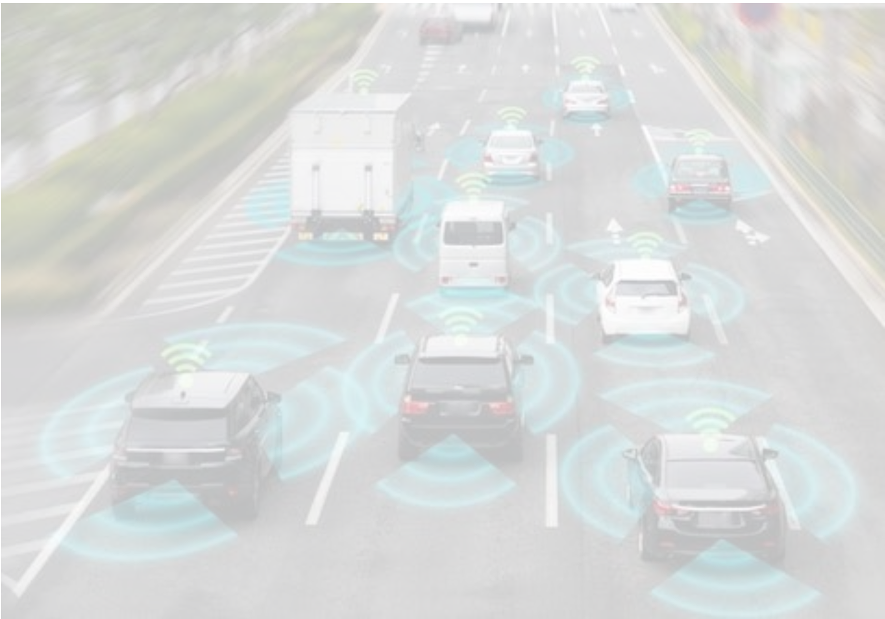


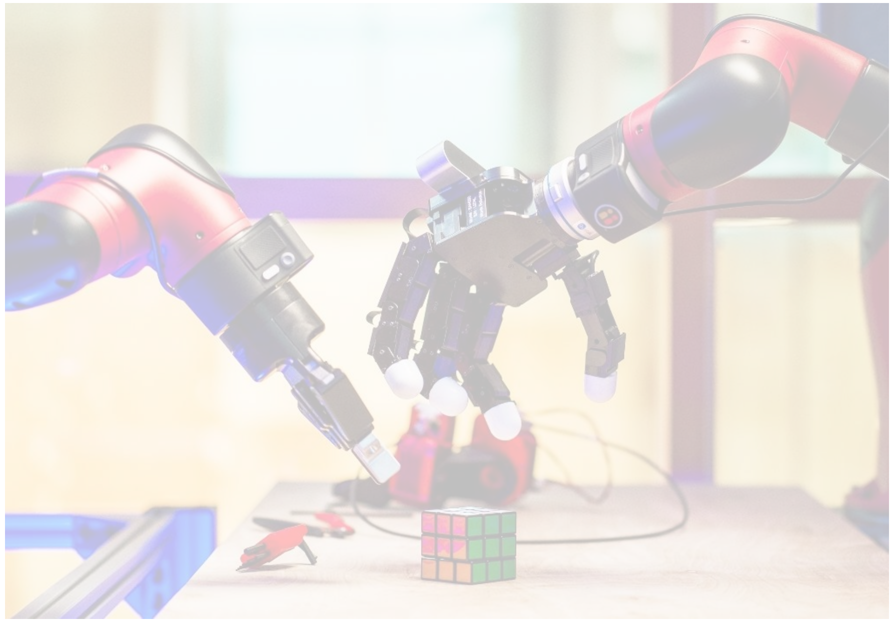
Applications

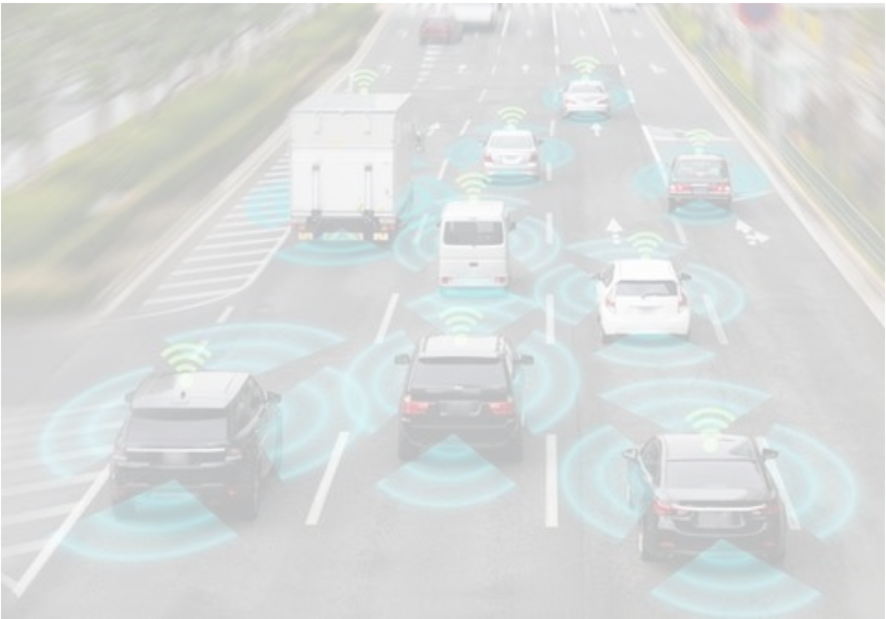


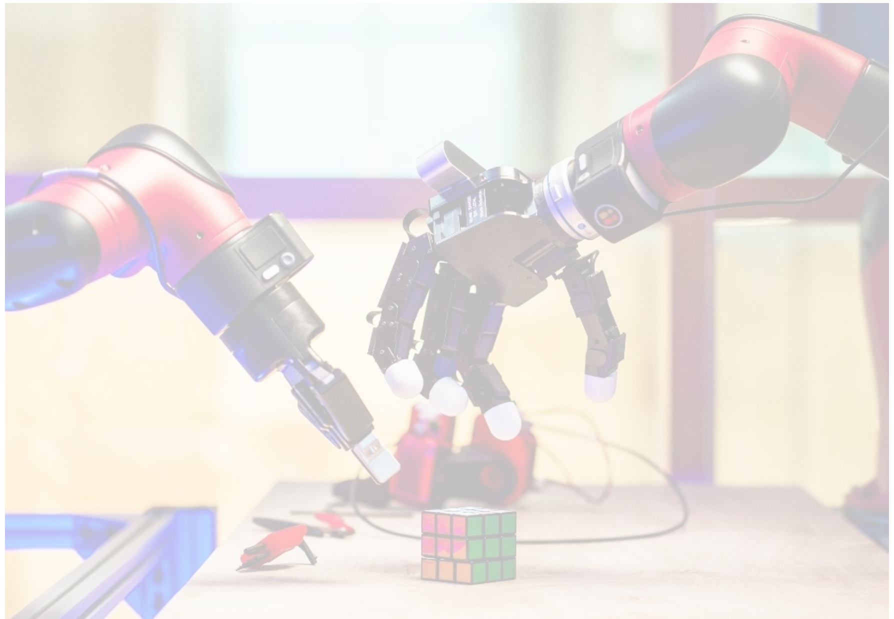


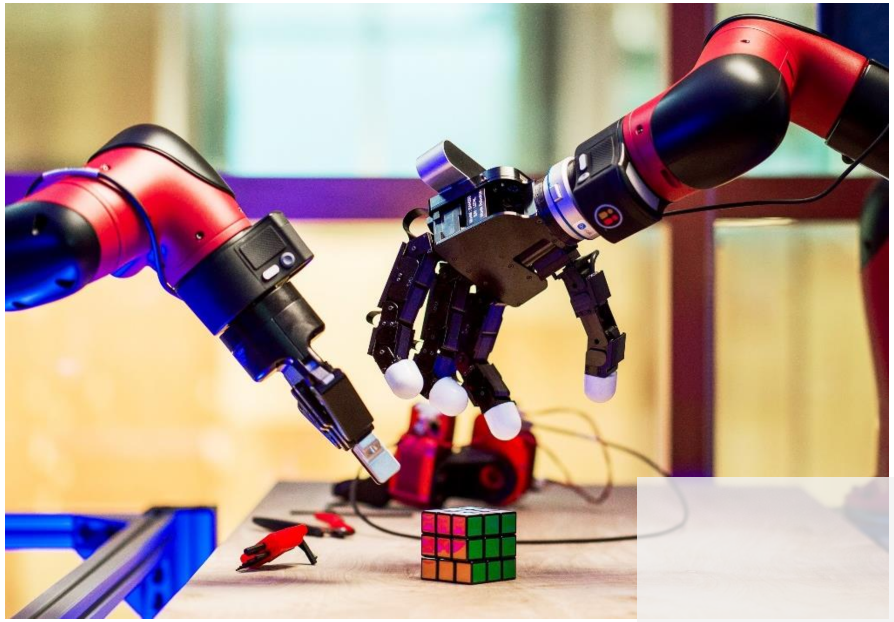


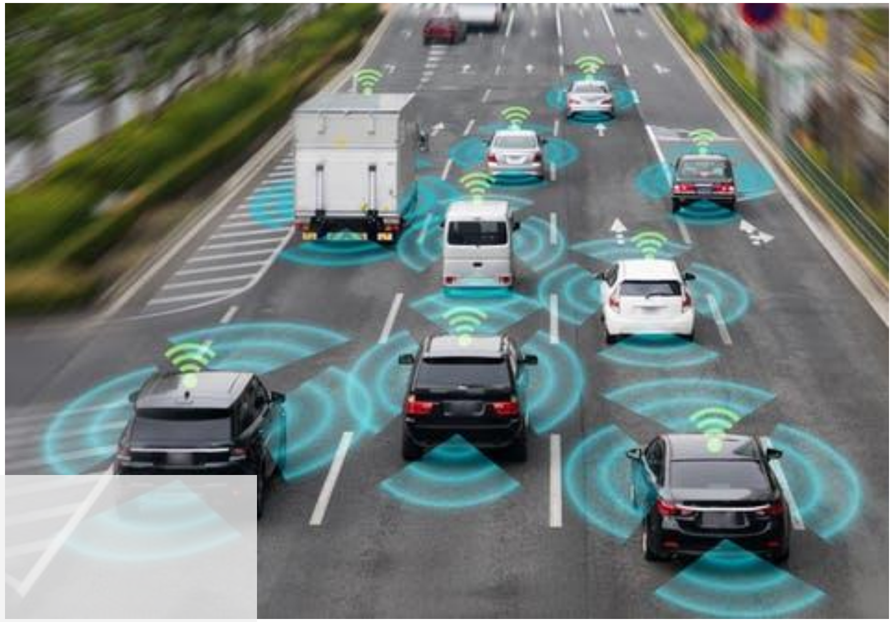






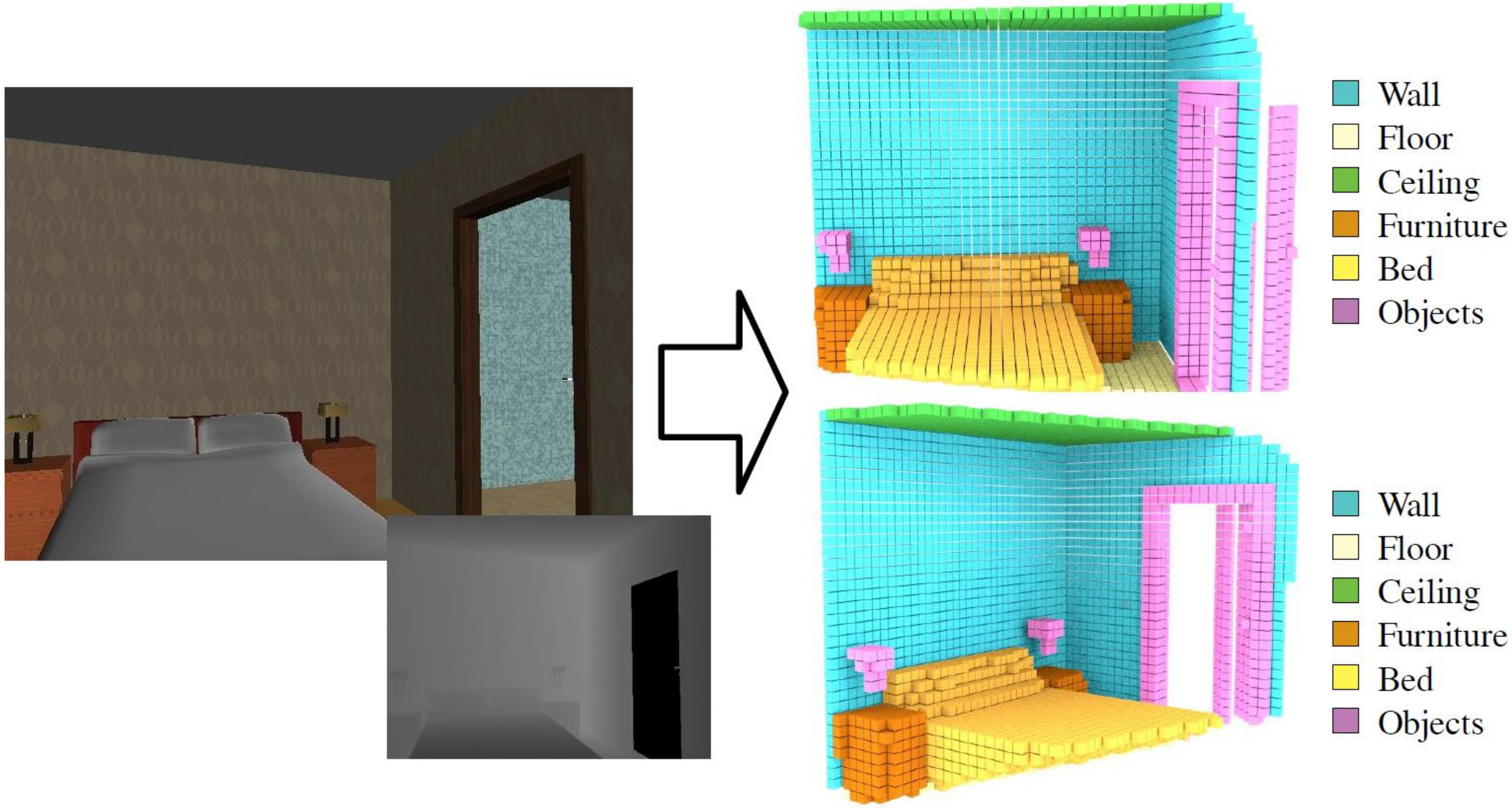






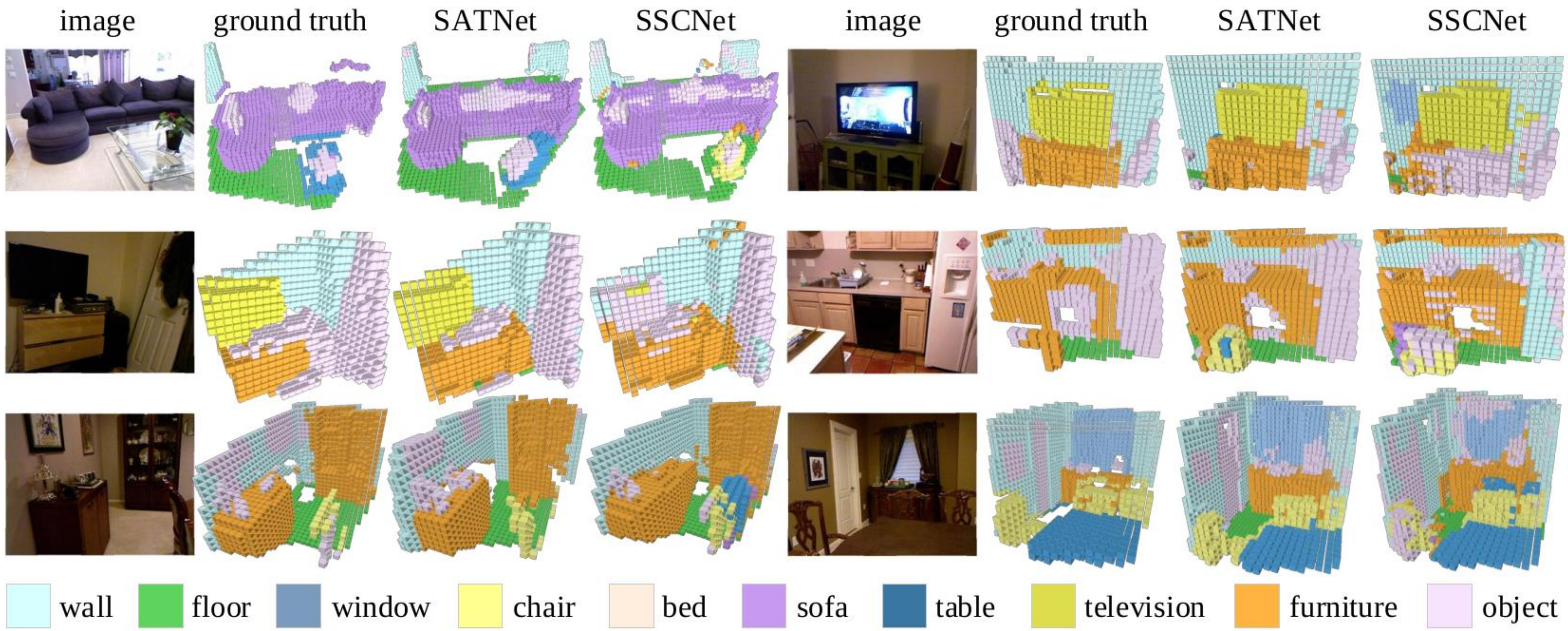
Applications

Semantic Scene Completion



[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

- Four main deficiencies of the approaches available by the time we started our research:

Problem Statement

- Four main deficiencies of the approaches available by the time we started our research:
  - **the RGB part and other modes of the RGB-D images are not completely explored;**

Problem Statement

- Four main deficiencies of the approaches available by the time we started our research:

    - the RGB part and other modes of the RGB-D images are not completely explored;

    - **techniques widely used in 2D deep CNN training are not used;**

Problem Statement

- Four main deficiencies of the approaches available by the time we started our research:
  - the RGB part and other modes of the RGB-D images are not completely explored;
  - techniques widely used in 2D deep CNN training are not used;
  - **available unlabelled data is not used;**

Problem Statement

- Four main deficiencies of the approaches available by the time we started our research:

  - the RGB part and other modes of the RGB-D images are not completely explored;

  - techniques widely used in 2D deep CNN training are not used;

  - available unlabelled data is not used;

  - **current solutions are limited to the restricted FOV of depth sensors**

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, semi-supervision and data augmentation in the 2D semantic segmenation context

3D Domain

- to apply current trends on 2D deep CNN training protocols to 3D SSC
- to propose and evaluate new SSC models that fully exploits the information in the RGB-D images
- to propose and evaluate the benefits of semi-supervised learning

360º 3D

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

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,** semi-supervision and data augmentation in **the** 2D semantic segmenation **context**

3D Domain

- to apply current trends on 2D deep CNN training protocols to 3D SSC
- to propose and evaluate new SSC models that fully exploits the information in the RGB-D images
- to propose and evaluate the benefits of semi-supervised learning

360º 3D

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

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, semi-supervision and data augmentation in the 2D semantic segmenation context

3D Domain

- **to apply current trends on 2D deep CNN training protocols to 3D SSC**
- **to propose and evaluate new SSC models that fully exploits the information in the RGB-D images**
- **to propose and evaluate the benefits of** semi-supervised **learning**

360^o 3D

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

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, semi-supervision and data augmentation in the 2D semantic segmenation context

3D Domain

- to apply current trends on 2D deep CNN training protocols to 3D SSC
- to propose and evaluate new SSC models that fully exploits the information in the RGB-D images
- to propose and evaluate the benefits of semi-supervised learning

360^o 3D

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

Publications

2D Domain

- Domain Adaptation for Holistic Skin Detection: 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI 2021)

3D Domain

- EdgeNet:Semantic Scene Completion from RGB-D images: International Conference on Pattern Recognition (ICPR 2020)
- Data Augmented 3D Semantic Scene Completion With 2D Segmentation Priors: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022)

360º 3D

- Semantic Scene Completion from a Single 360 degree Image and Depth Map: Conference on Computer Vision Theory and Applications (VISAPP 2020)
- Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras: Virtual Reality Journal

Publications

2D Domain

- Domain Adaptation for Holistic Skin Detection: **34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI 2021)**

3D Domain

- EdgeNet:Semantic Scene Completion from RGB-D images: **International Conference on Pattern Recognition (ICPR 2020)**
- Data Augmented 3D Semantic Scene Completion With 2D Segmentation Priors: **IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022)**

360^o 3D

- Semantic Scene Completion from a Single 360 degree Image and Depth Map: **Conference on Computer Vision Theory and Applications (VISAPP 2020)**
- Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras: **Virtual Reality Journal**

Publications

2D Domain

- Domain Adaptation for Holistic Skin Detection: **34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI 2021)**

3D Domain

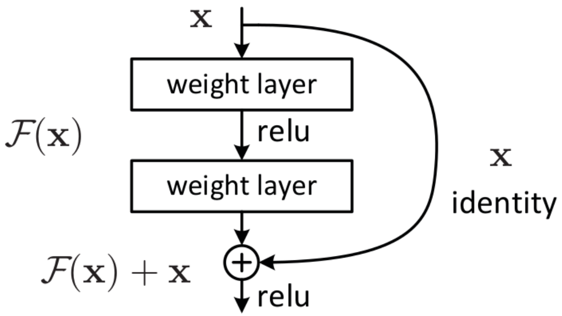
- EdgeNet:Semantic Scene Completion from RGB-D images: **International Conference on Pattern Recognition (ICPR 2020)**
- Data Augmented 3D Semantic Scene Completion With 2D Segmentation Priors: **IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022)**

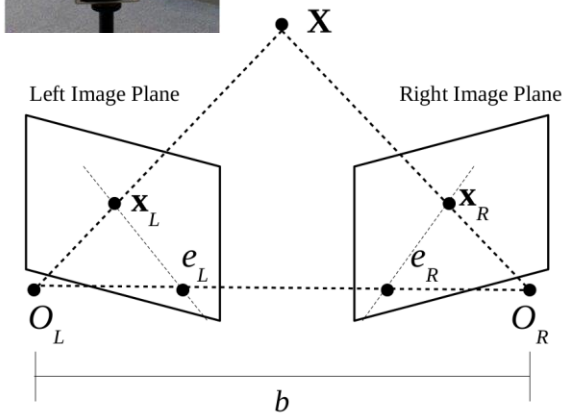
360^o 3D

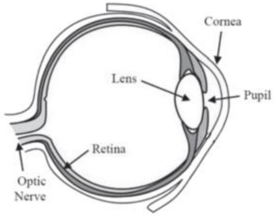
- Semantic Scene Completion from a Single 360 degree Image and Depth Map: **Conference on Computer Vision Theory and Applications (VISAPP 2020)**
- Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras: **Virtual Reality Journal**

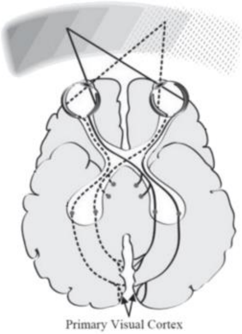
Chapter 2

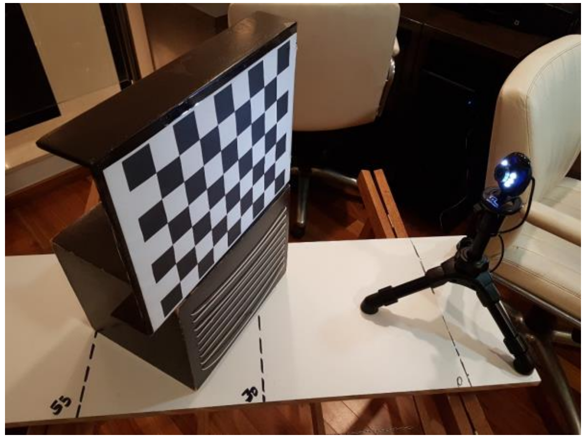
Background and Related Concepts

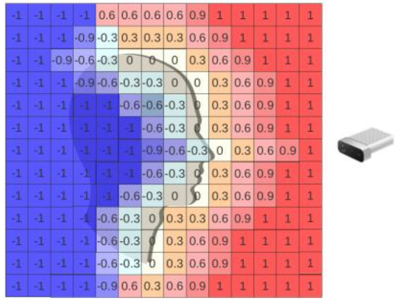


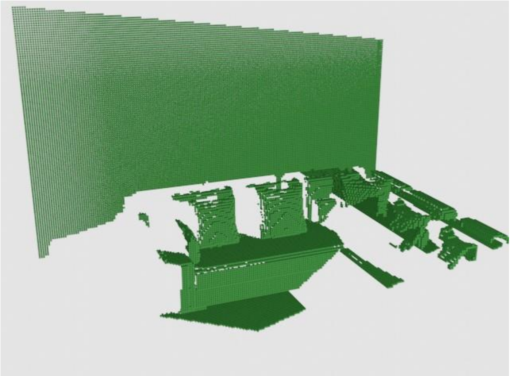








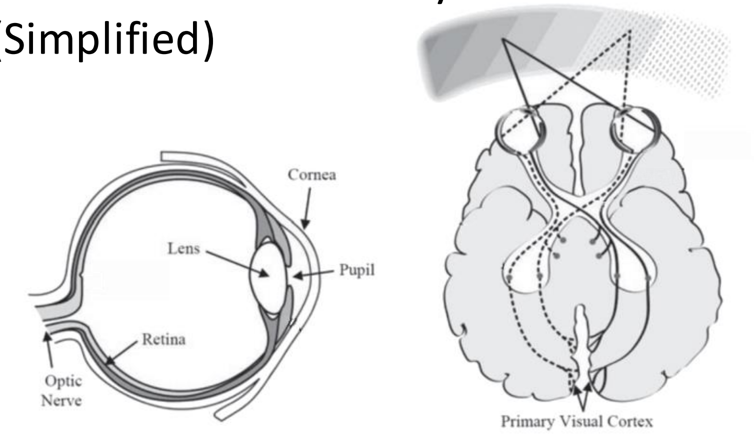




Stereo Vision

- Stereo Vision in Computer Vision relates to the stereo nature of human eyes

Human Stereo Vision System (Simplified)



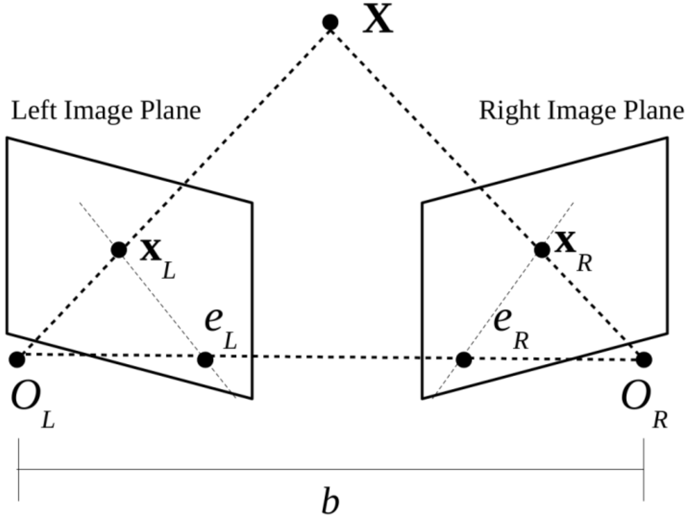
Example of a Digital Stereo Camera

Example of a Computer Vision Stereo image and corresponding dept map

(a) Left view (b) Right view (c) Depth map

Epipolar Geometry and Stereo Vision

Epipolar Geometry



Camera Calibration

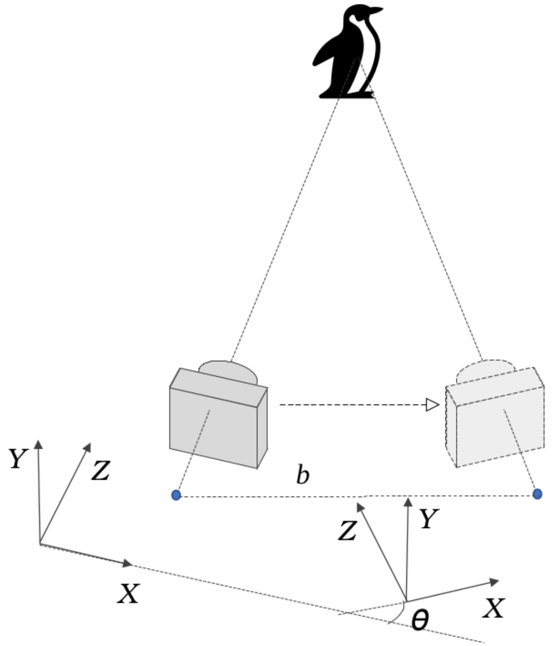


Image and Scene Reasoning Evolution

Image and Scene Reasoning Evolution

1956 Dartmouth Summer Research Project on Artificial Intelligence

Marvin Minsky, Claude Shannon, Ray Solomonoff and other scientists at the Dartmouth Summer Research Project on Artificial Intelligence. Photo by Margaret Minsky.

Image and Scene Reasoning Evolution

1958

Rosenblatt Perceptron

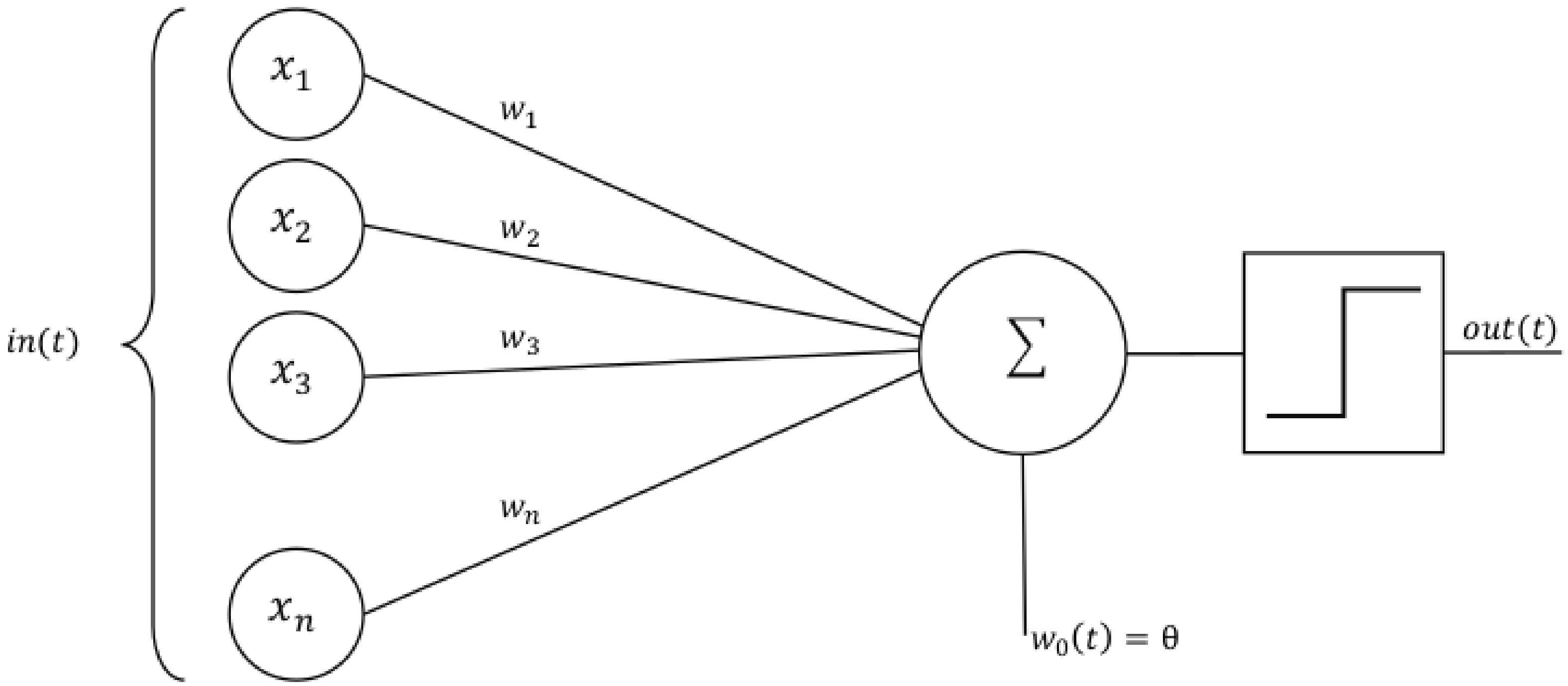


Image and Scene Reasoning Evolution

1959

MIT AI Lab – Marvin Minsky and Jonh McCarty

Image and Scene Reasoning Evolution

1966 The Summer Vision Project

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Image and Scene Reasoning Evolution

1966 The Summer Vision Project

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Logics-based approaches (Functionalism)

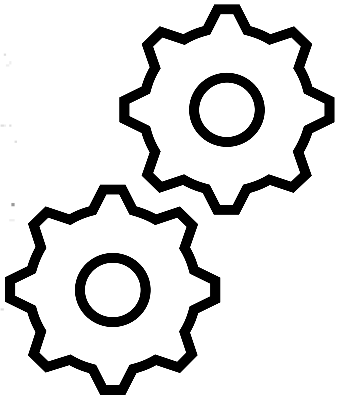
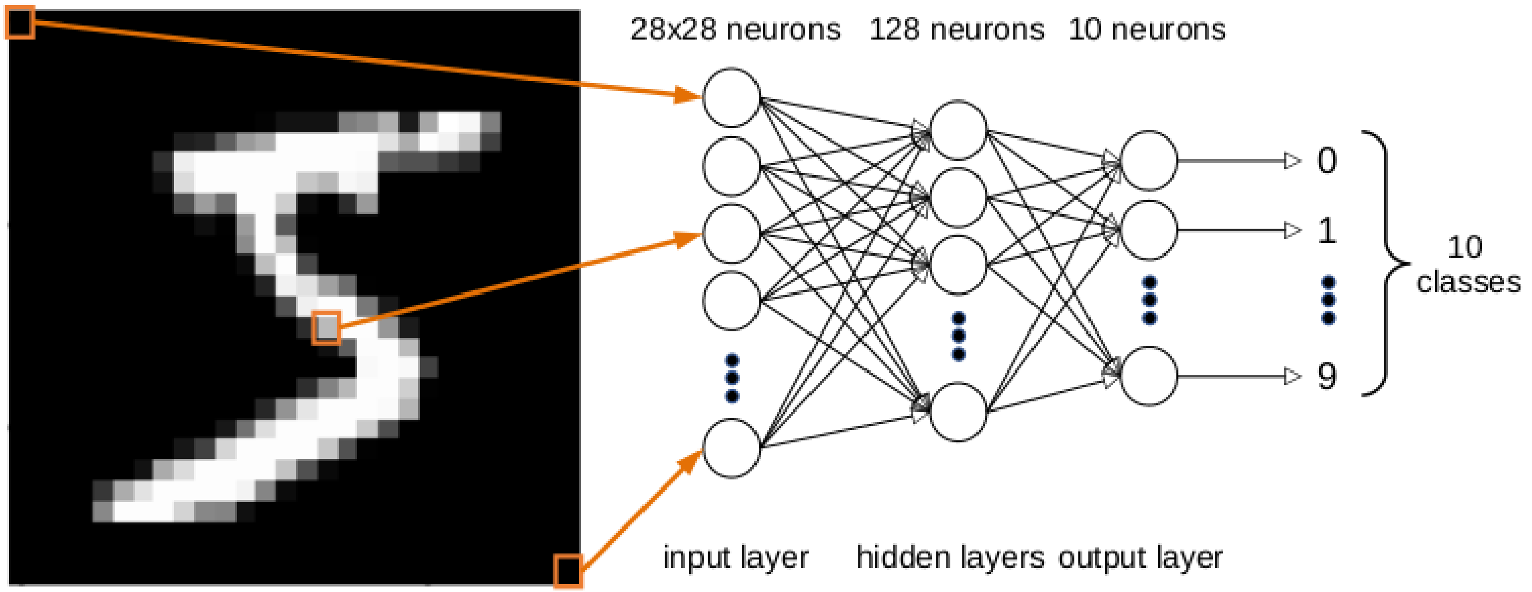


Image and Scene Reasoning Evolution

1986 Backpropagation Learning Algorithm and the Multilayer Perceptron



Multilayer perceptron for recognizing handwritten digits.

Image and Scene Reasoning Evolution

1986

AI Winter

Image and Scene Reasoning Evolution

1989

Yann LeCunn Convolutional Networks

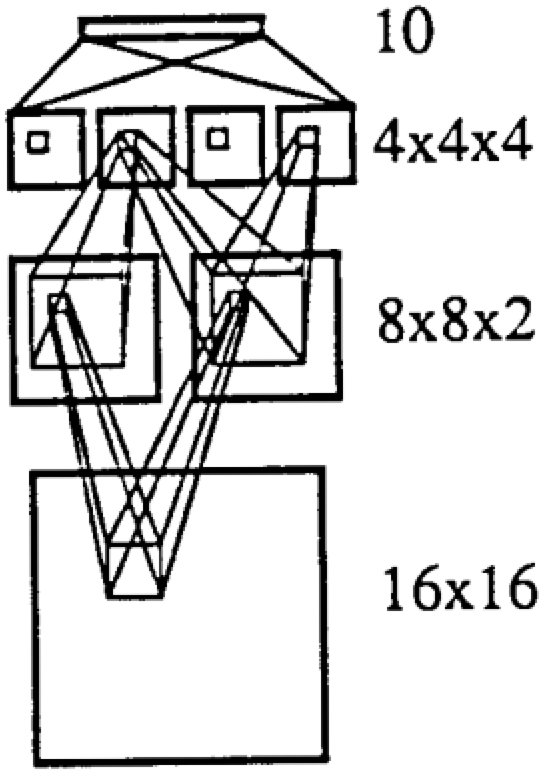
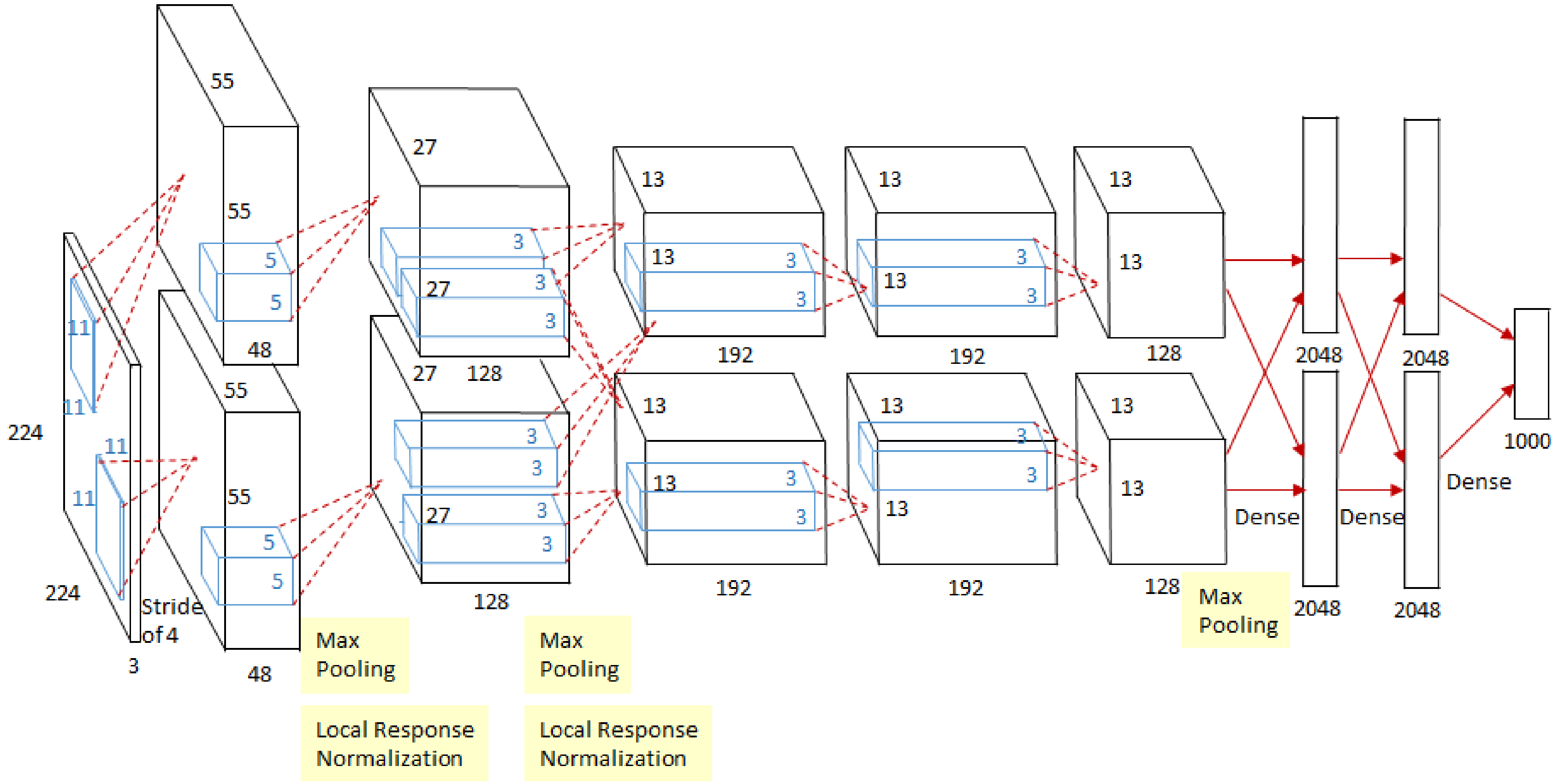


Image and Scene Reasoning Evolution

2012 The boom of convolutional networks

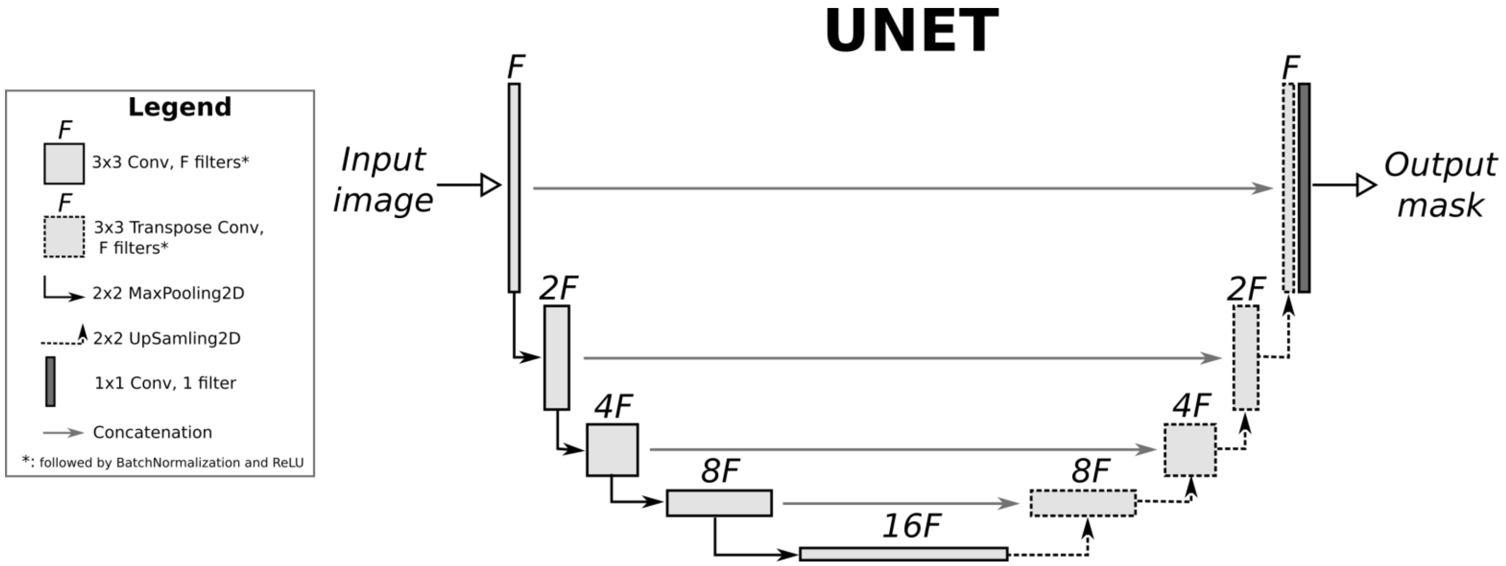


AlexNet: ImageNet LSVRC-2010

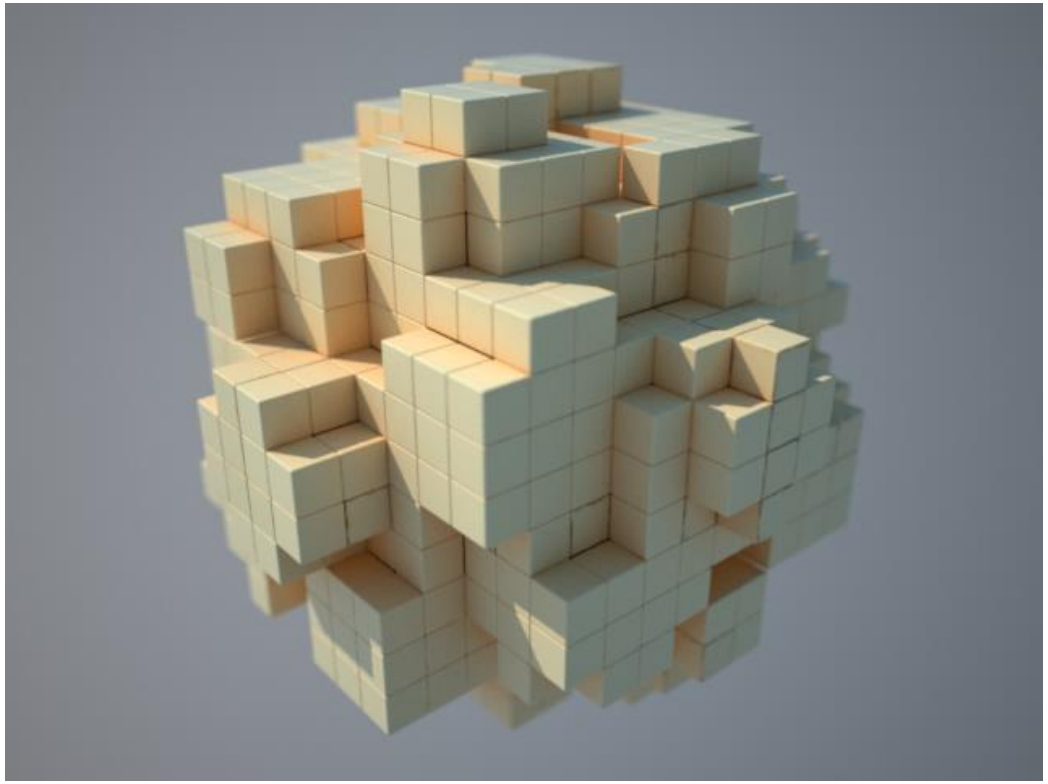
Image and Scene Reasoning Evolution

2014

Fully Connected for Image Segmentation



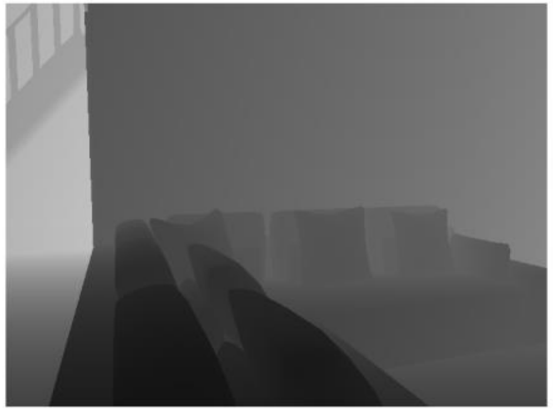
3D Representation: Voxel Volume Encoding



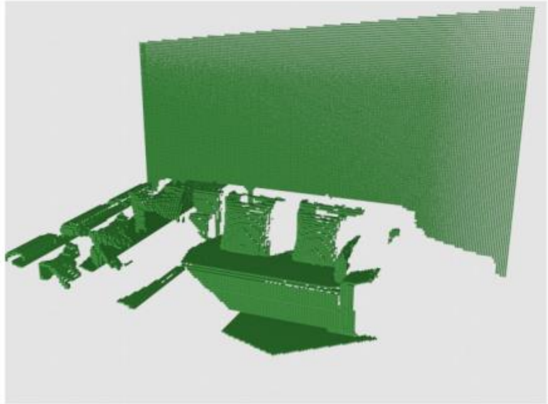
3D Representation: Voxel Volume Encoding

Lifting from Depth Maps to Voxels

RGB



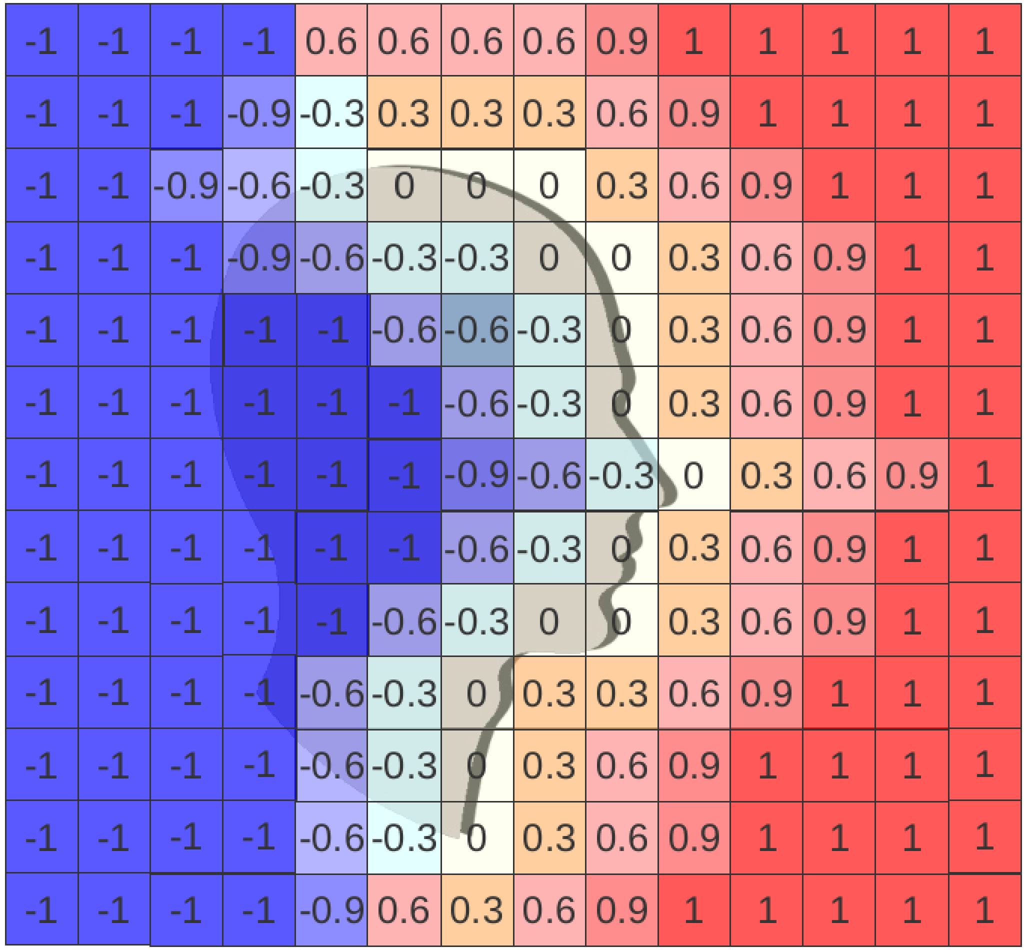
Depth Map



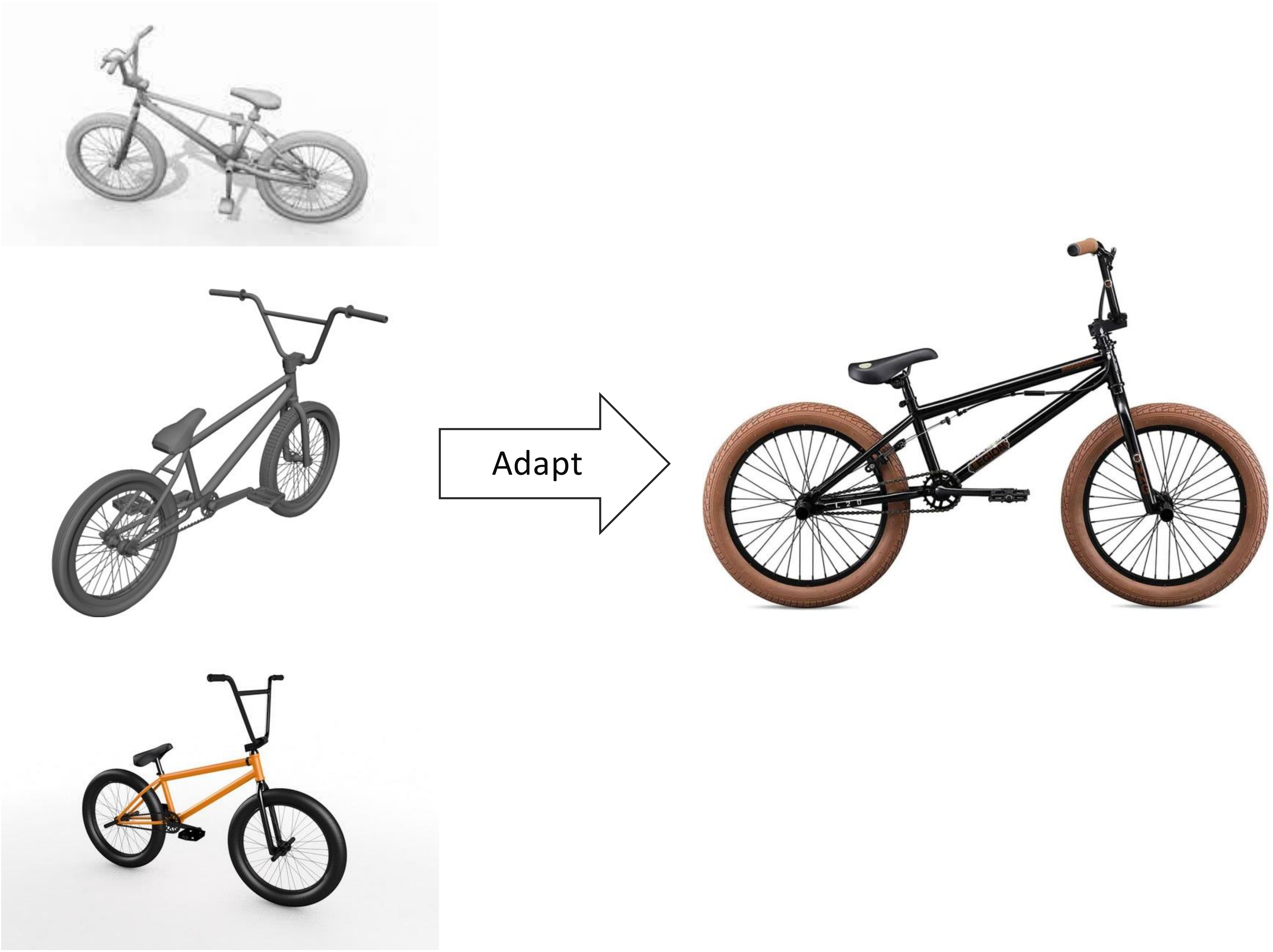
Voxel Representation

3D Representation: Voxel Volume Encoding

Truncated Signed Distance Function (TSDF)

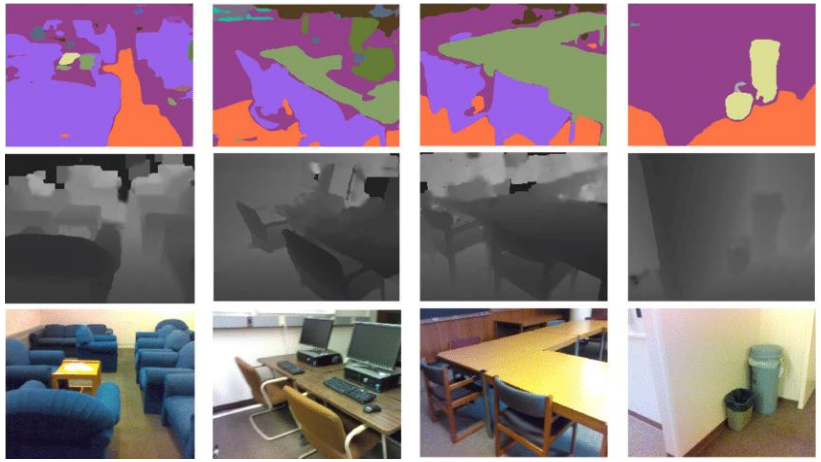


Domain Adaptation



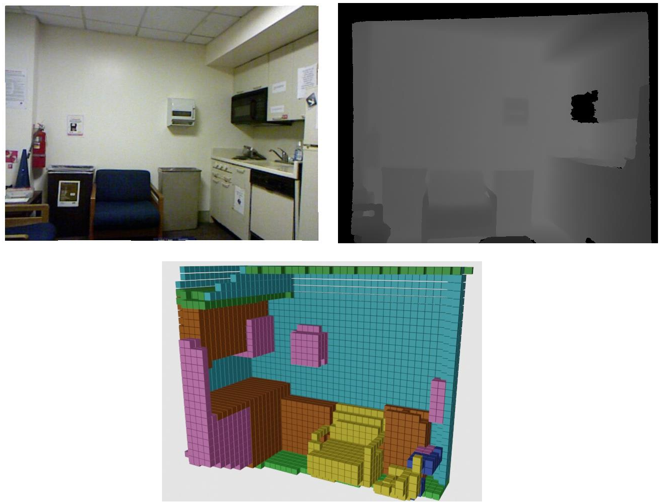
Chapter 3

Related Works



2D RGB-D Semantic Segmentation

Partial 3D Scene Reasoning from RGB-D

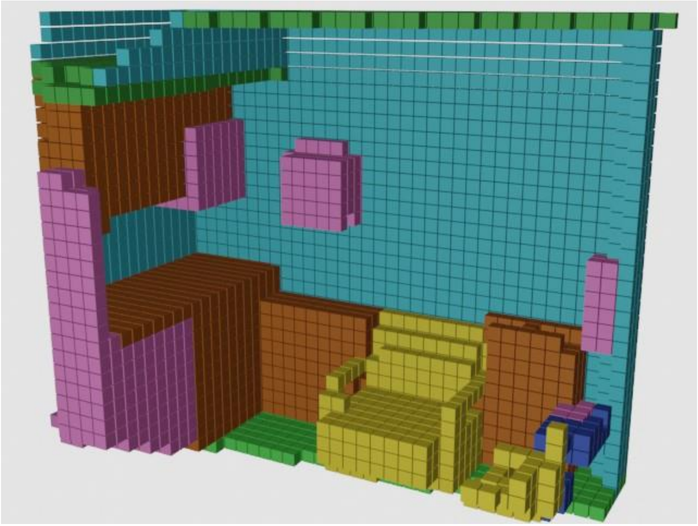


3D Semantic Scene Completion

360 degree Scene Understanding

Semantic Scene Completion

RGB-D Input



Output

Semantic Scene Completion

The Seminal Work

Semantic Scene Completion from a Single Depth Image

Shuran Song Fisher Yu Andy Zeng Angel X. Chang Manolis Savva Thomas Funkhouser
Princeton University
http://sscnet.cs.princeton.edu

Abstract

*This paper focuses on semantic scene completion, a task for producing a complete 3D voxel representation of volumetric occupancy and semantic labels for a scene from a single-view depth map observation. Previous work has considered scene completion and semantic labeling of depth maps separately. However, we observe that these two problems are tightly intertwined. To leverage the coupled nature of these two tasks, we introduce the semantic scene completion network (*__SSCNet__*), an end-to-end 3D convolutional network that takes a single depth image as input and simultaneously outputs occupancy and semantic labels for all voxels in the camera view frustum. Our network uses a dilation-based 3D context module to efficiently expand the receptive field and enable 3D context learning. To train our network, we construct SUNCG - a manually created large-scale dataset of synthetic 3D scenes with dense volumetric annotations. Our experiments demonstrate that the joint model outperforms methods addressing each task in isolation and outperforms alternative approaches on the semantic scene completion task. The dataset and code is available at http://sscnet.cs.princeton.edu.*

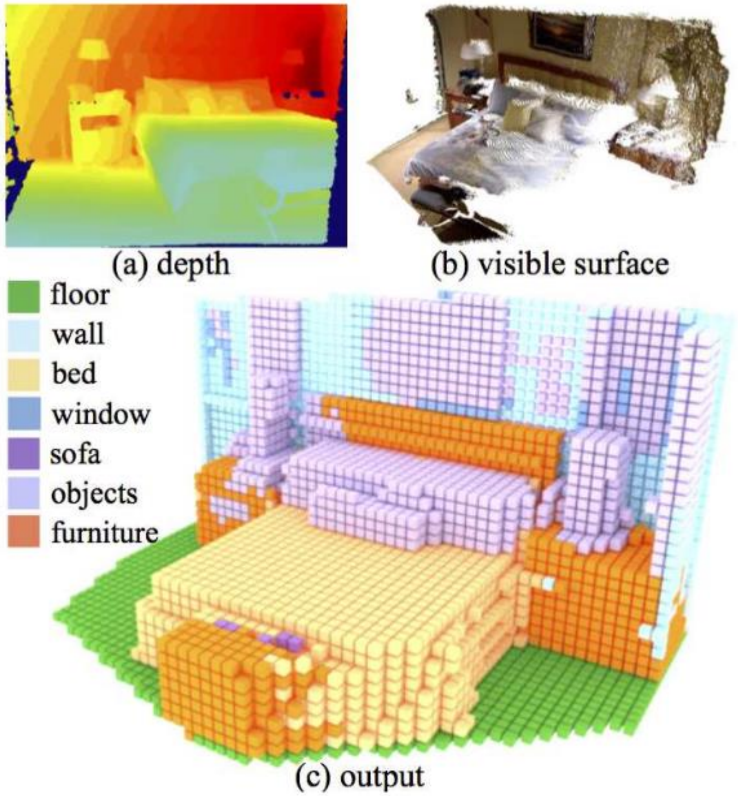


Figure 1. **Semantic scene completion.** (a) Input single-view depth map (b) Visible surface from the depth map; color is for visualization only. (c) Semantic scene completion result: our model jointly predicts volumetric occupancy and object categories for each of

Semantic Scene Completion

The Seminal Work

Dilated Convolutions to Enhance Receptive Field

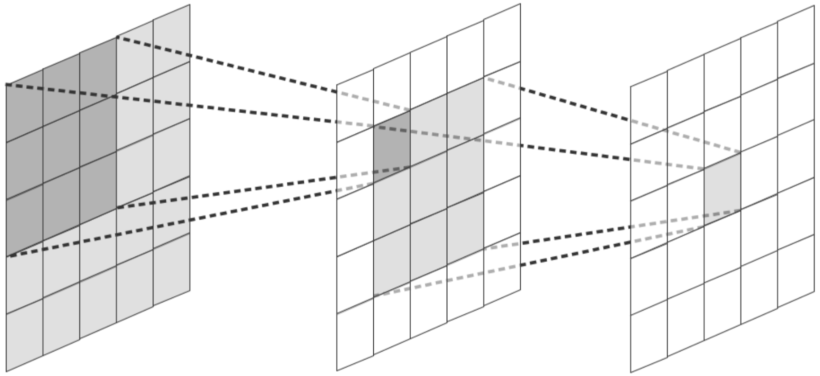
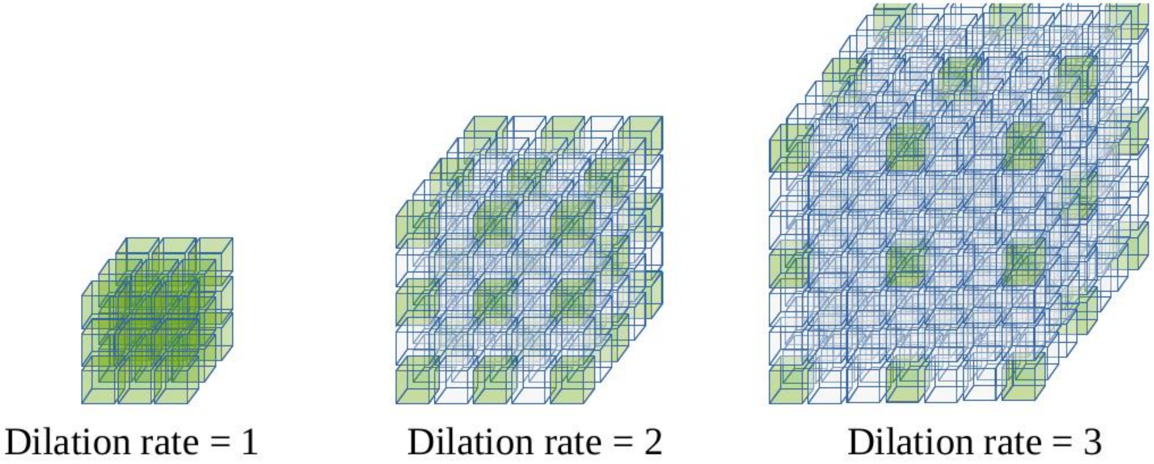


Illustration of a 2D CNN’s Receptive Field

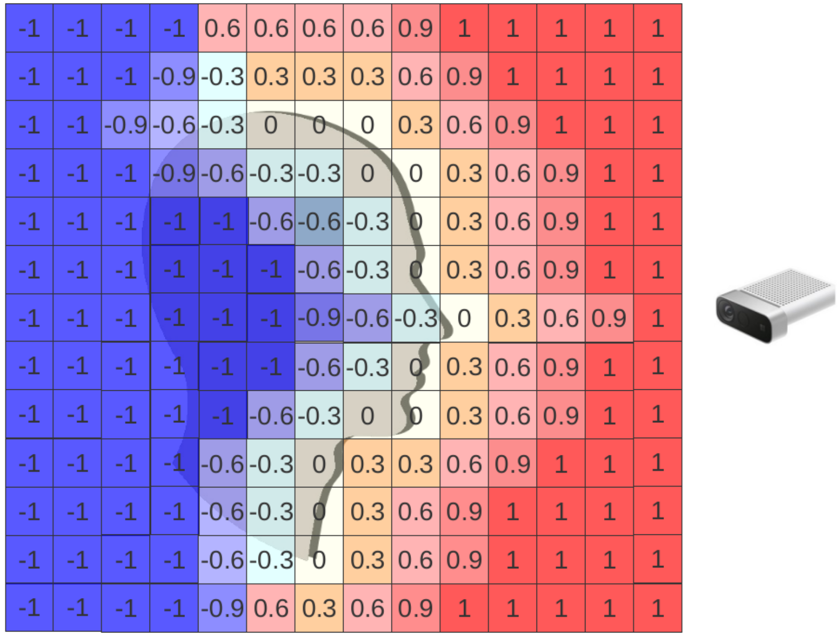


Dilated 3D Convolution Kernels

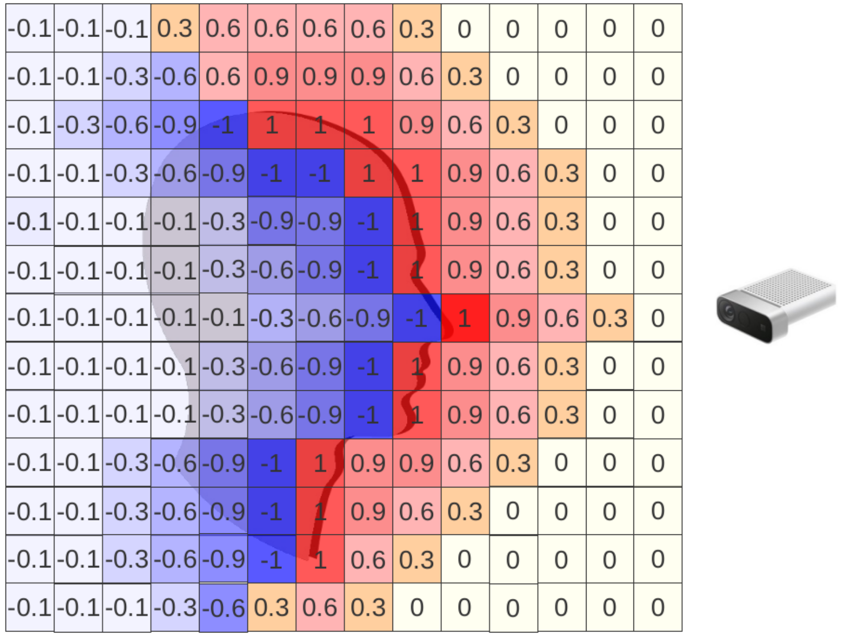
Semantic Scene Completion

The Seminal Work

Better 3D volume encoding



Original TSDF



Proposed F-TSDF

$$F-TSDF = sign(TSDF) \times (1 - |TSDF|)$$

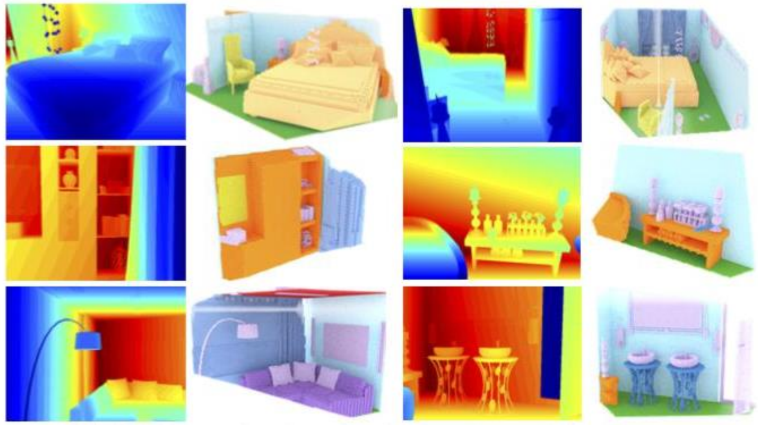
Semantic Scene Completion

The Seminal Work

Training on synthetic data

SUNCG Synthetic Scenes

Generated Views



Semantic Scene Completion

SSC Prior Works

Semantic Scene Completion

SSC Prior Works

- Depth maps only

Semantic Scene Completion

SSC Prior Works

- Depth maps only
- Depth maps plus RGB

Semantic Scene Completion

SSC Prior Works

- Depth maps only
- Depth maps plus RGB
- Depth maps plus 2D Segmentation

360° Scene Understanding

Datasets

- NYUD v2
- NYUCAD
- SUNCG

Chapter 4

Fully Convolutional Networks, Domain Adaptation and Semantic Segmentation

Fully Convolutional Networks, Domain Adaptation and Semantic Segmentation

Why work on 2D?

- Work on 3D is hard
- Start to explore domain adaptation and segmentation in an easier domain

[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

Fully Convolutional Networks, Domain Adaptation and Semantic Segmentation

- Why the skin segmentation application?
  - Research field where some criticisms regarding the use of CNNs/FCNs are made:
    - the need for large training datasets [53]
    - the specificity or lack of generalization of neural nets
    - long prediction time [12]
  - We wanted to try 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

Previous Works

Historically, color-based or texture methods were Preferred

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

- Skin-color separation
- Patch-based CNN

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

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

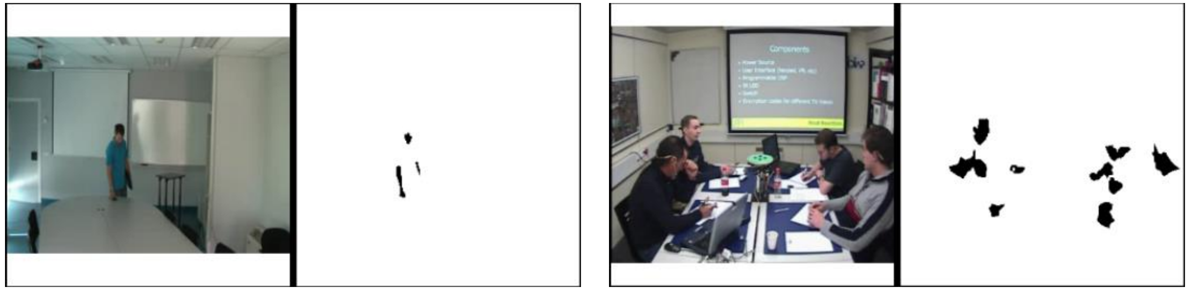
Datasets

SFA[15]
(1,118 images)

Pratheepan[117]
(78 images)

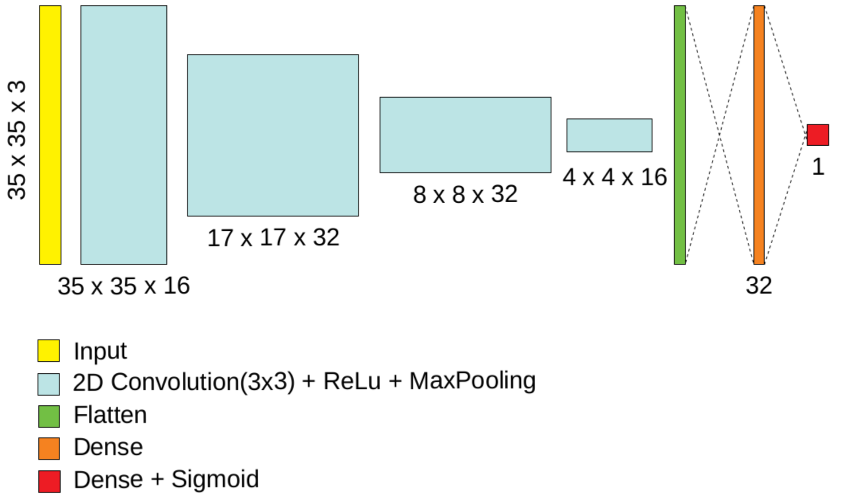
Compaq[51]
(4,670 images)

VPU[93]
(290 images)

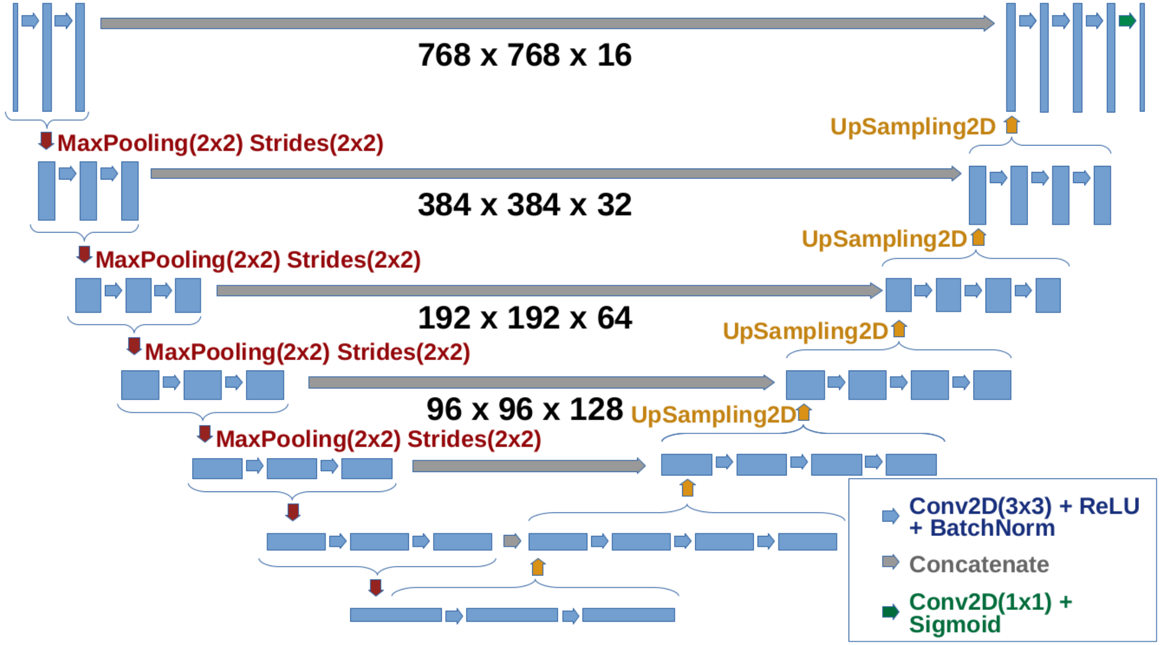


Models

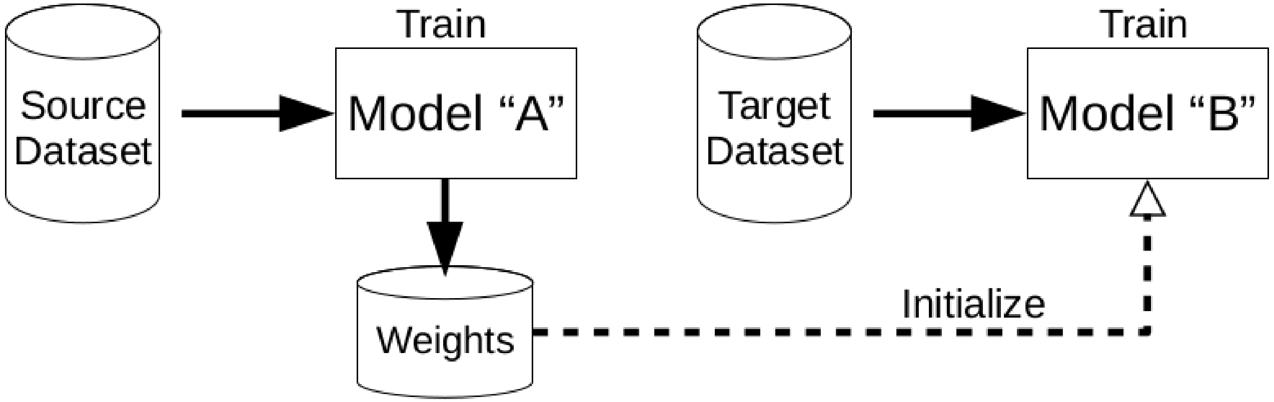
Local, Patch-based CNN



Holistic, u-shaped FCN

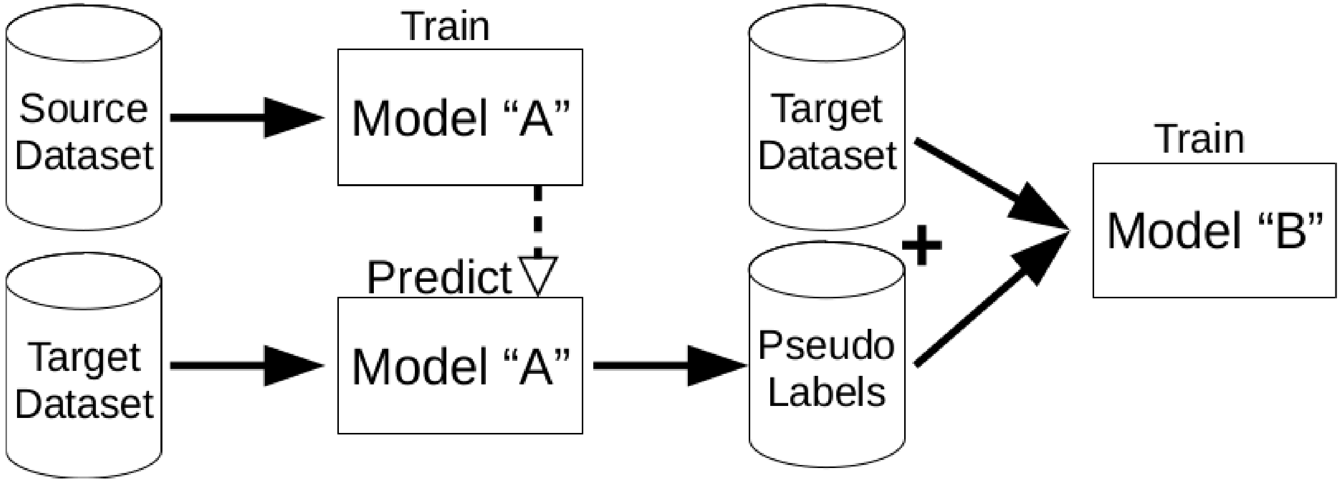


Domain adaptation approaches



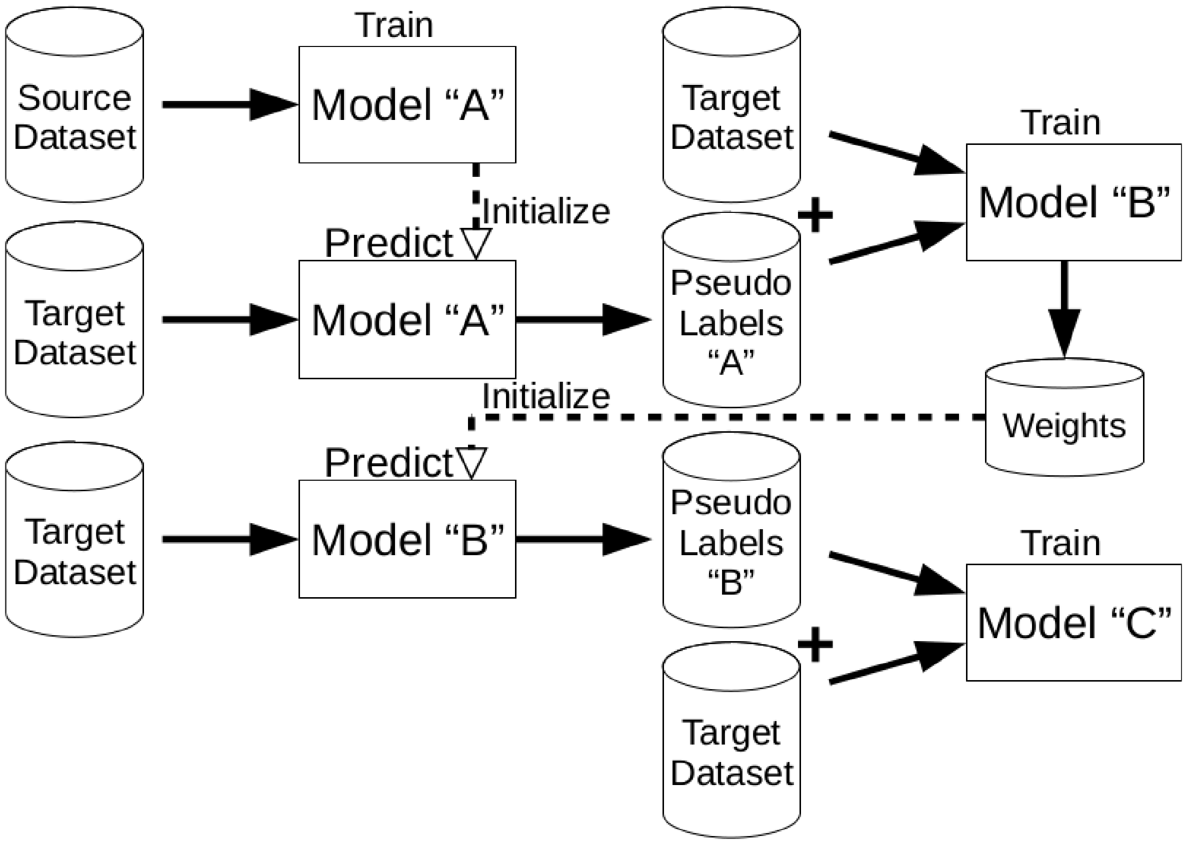
Inductive Transfer Learning by fine-tuning parameters of a model to a new domain

Domain adaptation approaches



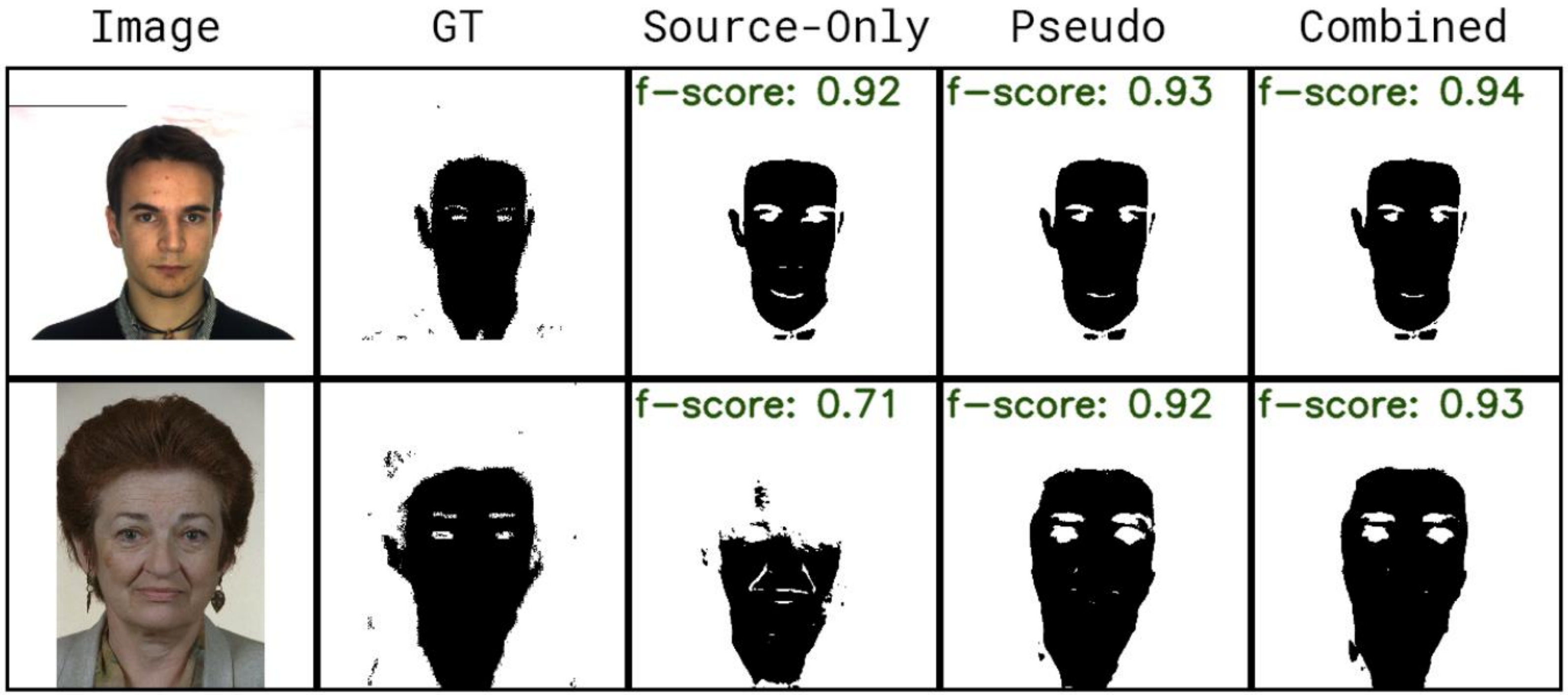
Semi-supervised and unsupervised Domain Adaptation by cross-domain pseudo-labeling

Domain adaptation approaches



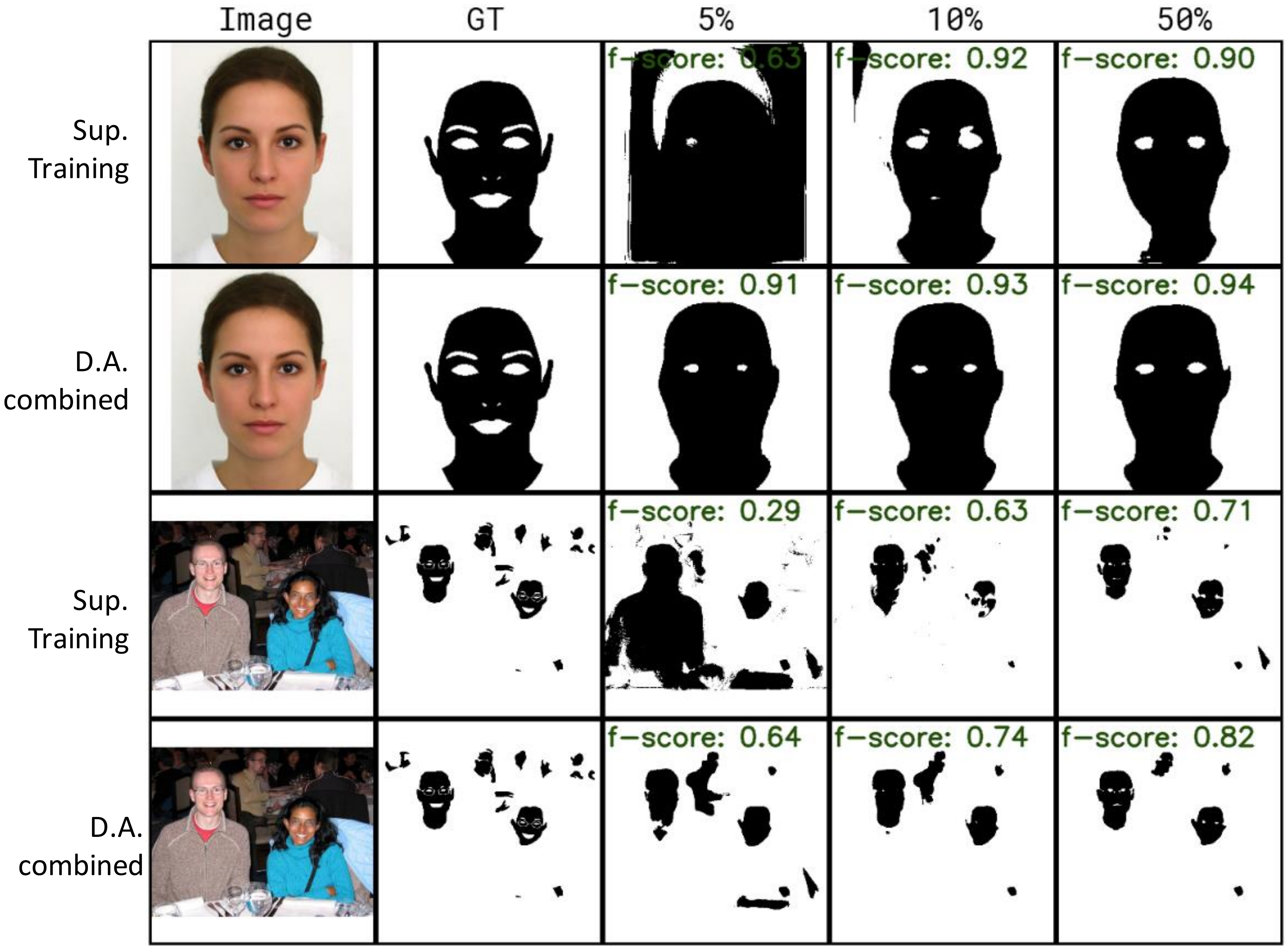
Combined transfer learning and domain adaptation approach

Domain adaptation qualitative results



Domain adaptation from Compaq to SFA 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

Chapter 4 Summary

Refuted criticisms regarding the use of Deep Convolutional Networks for skin segmentation

- Color or texture separation may suffice:
  - Our two CNN approaches performed much better than the color-based state-of-the-art
- CNNs are slow:
  - Our U-Net inference time was enough for real-time applications
- CNNs need too much data to generalize:
  - With no labeled data -> 60% improvement

Publication

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]

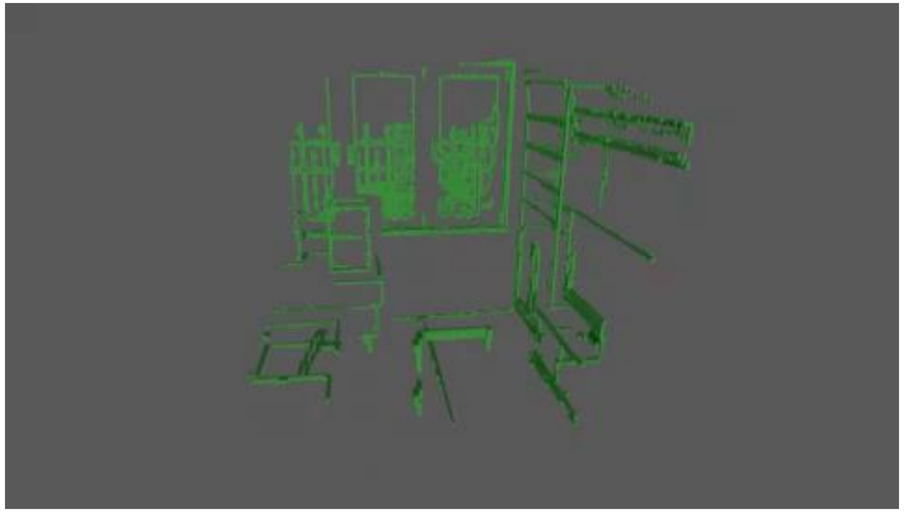
Before the boom of Convolutional Neural Networks (CNNs), most approaches

1

Published in the 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI 2021)

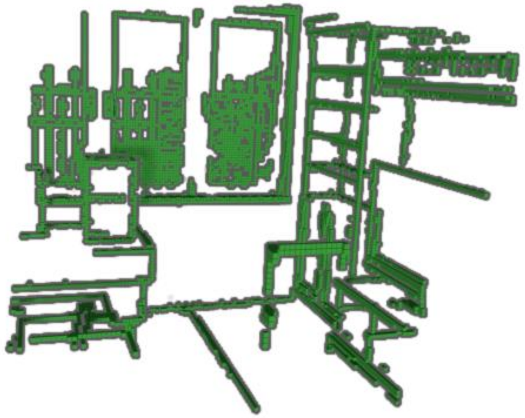
Chapter 5

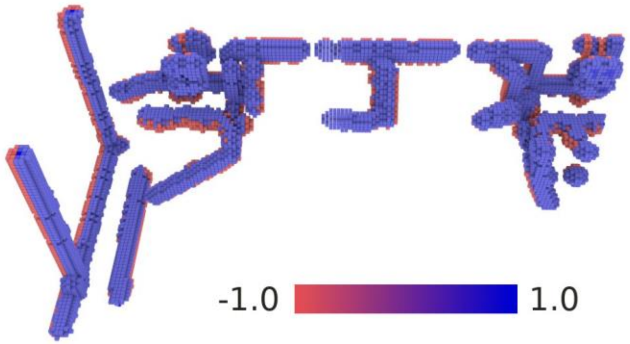
Using RGB Edges to improve Semantic Scene Completion from RGB-D Images



Our Approach: EdgeNet

- We extract information from RGB data using image Canny Edge detector

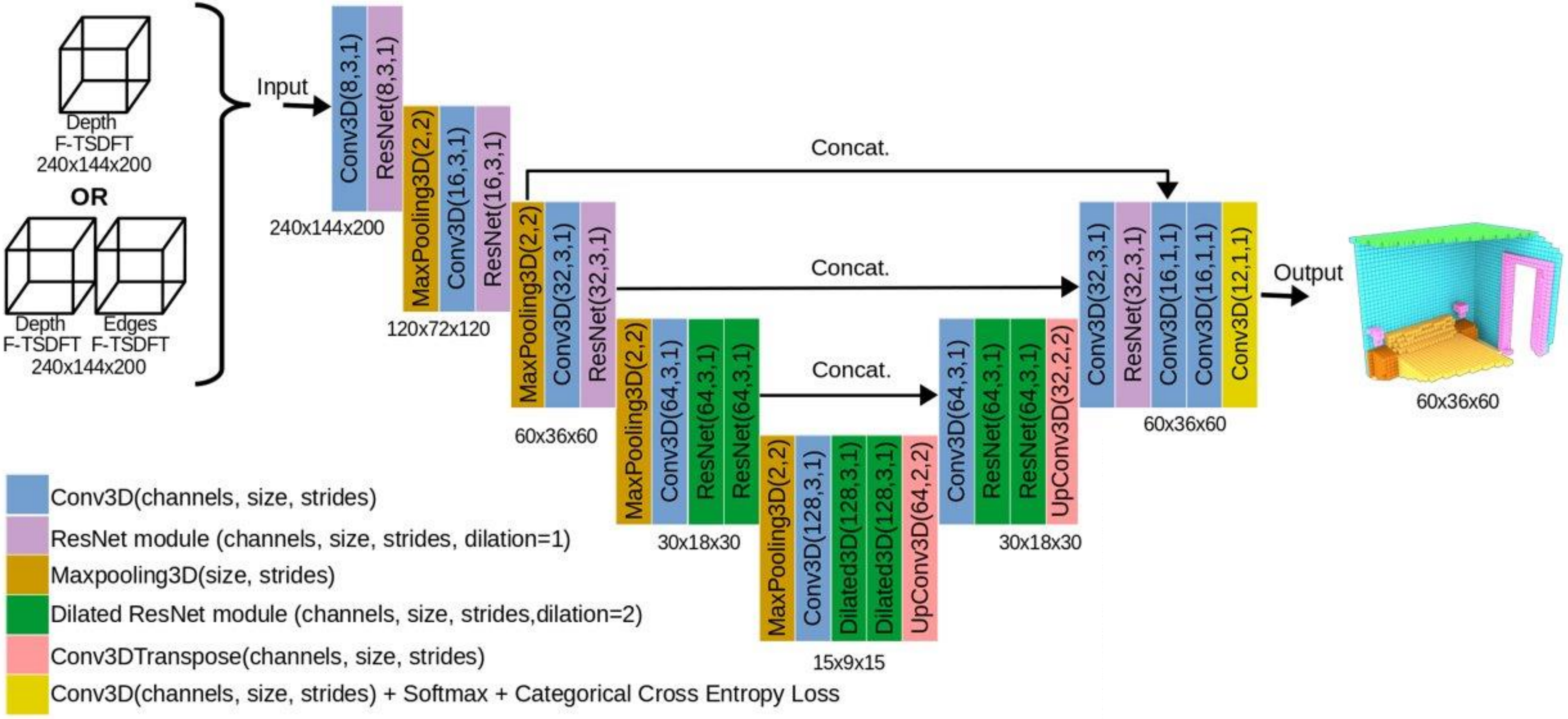




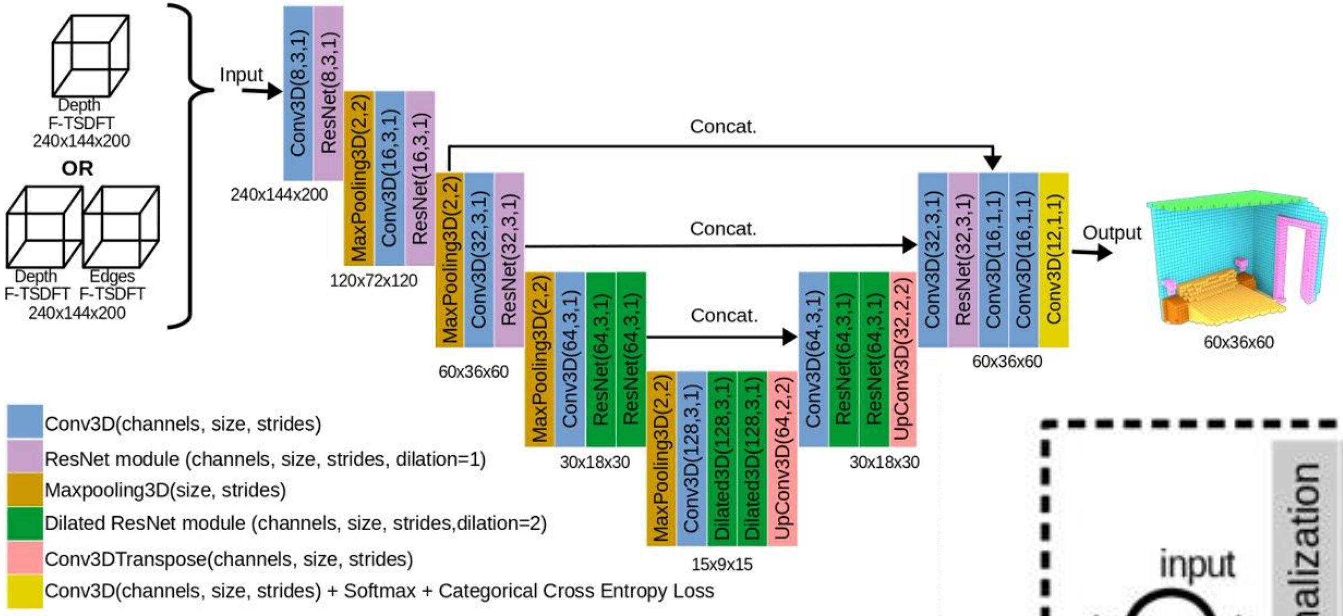
Our implementation

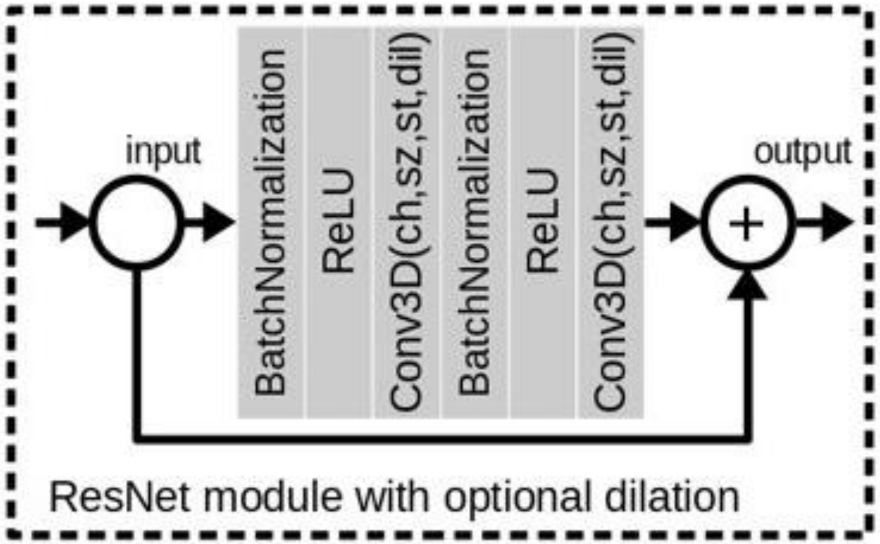
- Offline F-TSDF calculation using portable C++ CUDA code
- We provide a software interface between CUDA and Python
- Preprocessing code is independent from the deep learning framework

Network Architecture

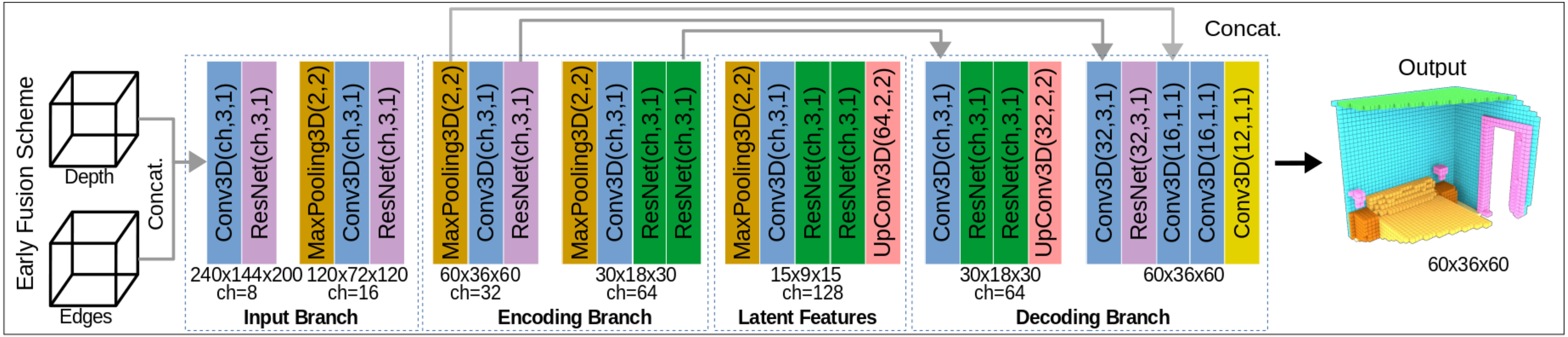


Network Architecture

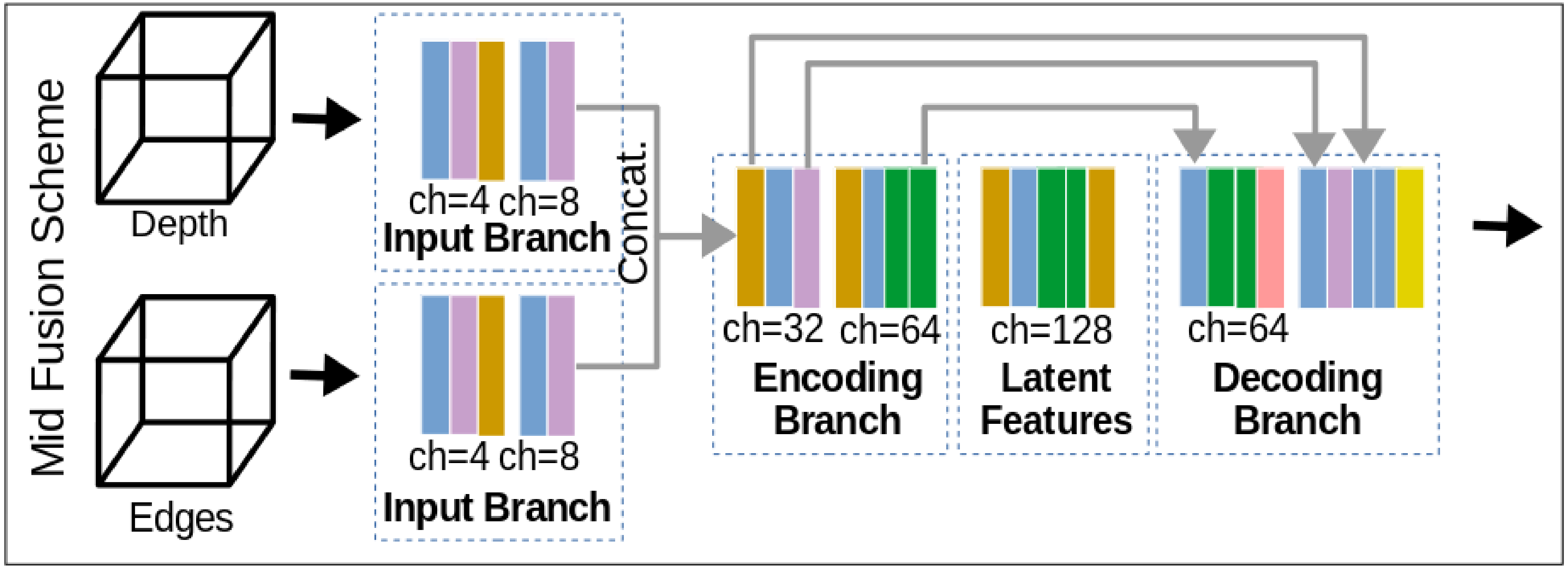




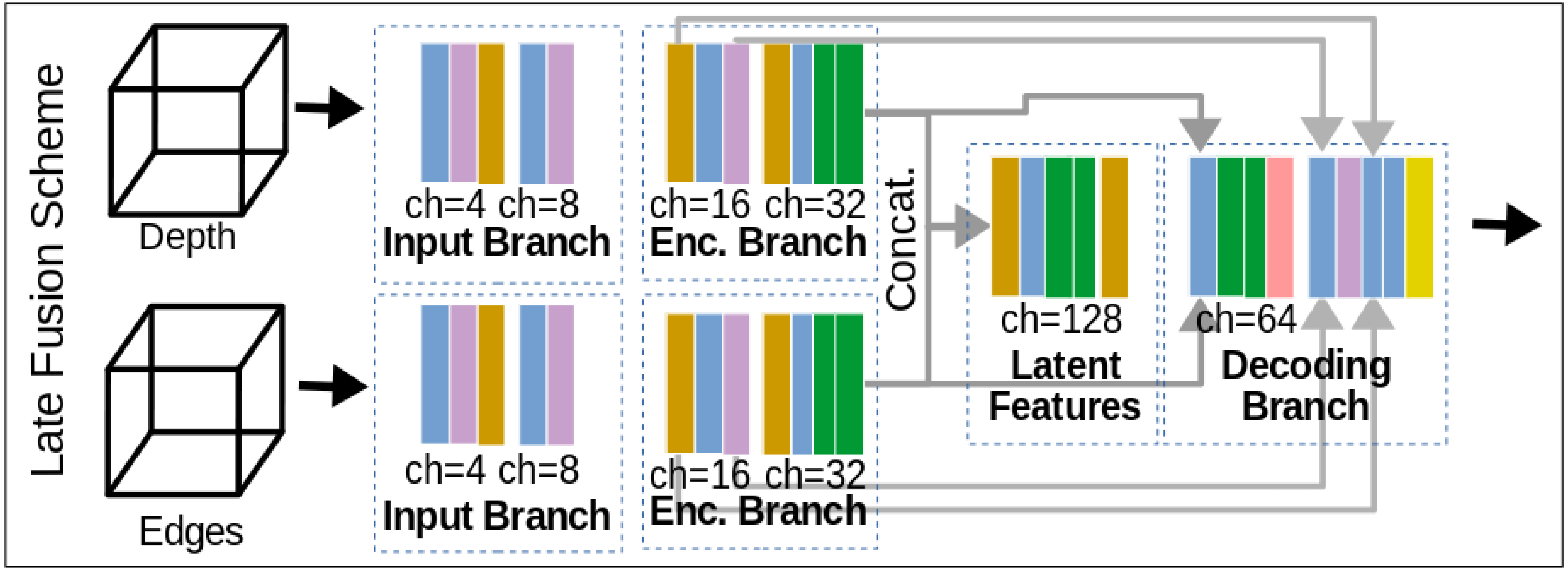
Network Architecture - Fusion Schemes



Network Architecture - Fusion Schemes



Network Architecture - Fusion Schemes



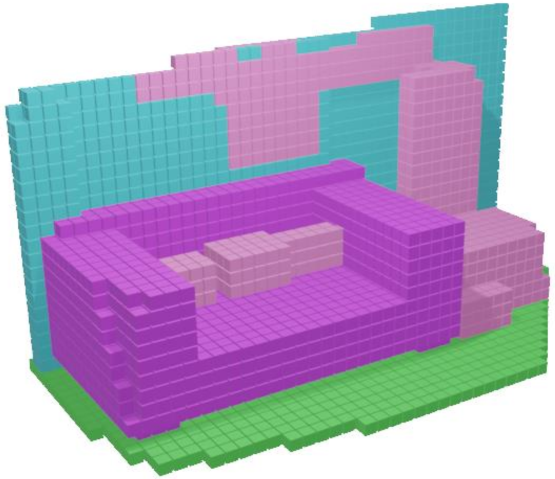
Training Time

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

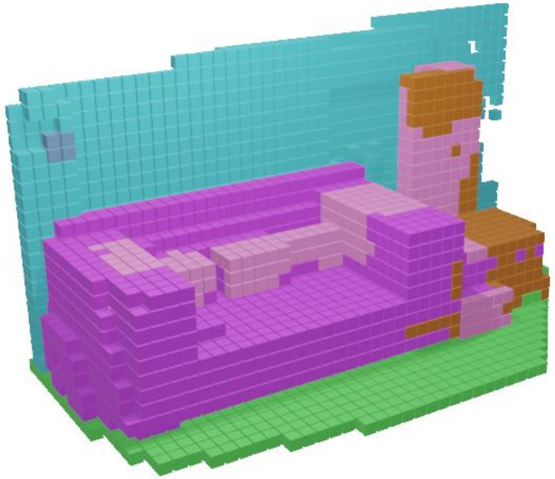
Quantitative Results

- 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

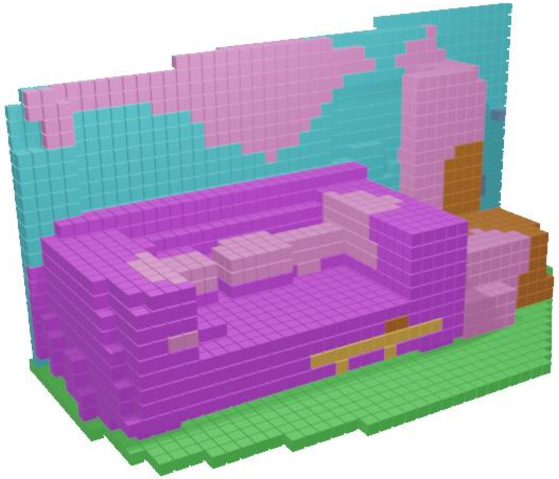
Qualitative Results



Ground Truth

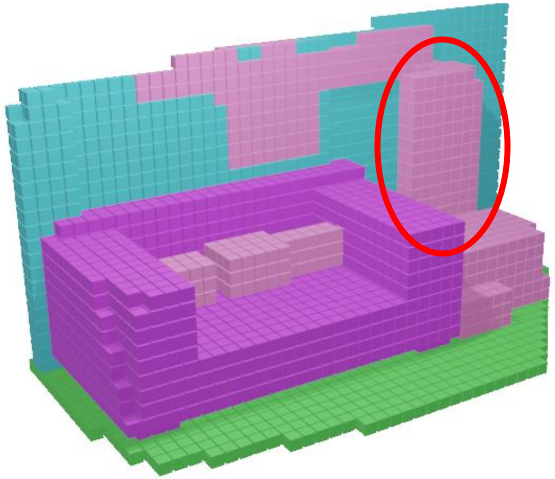


SSCNet

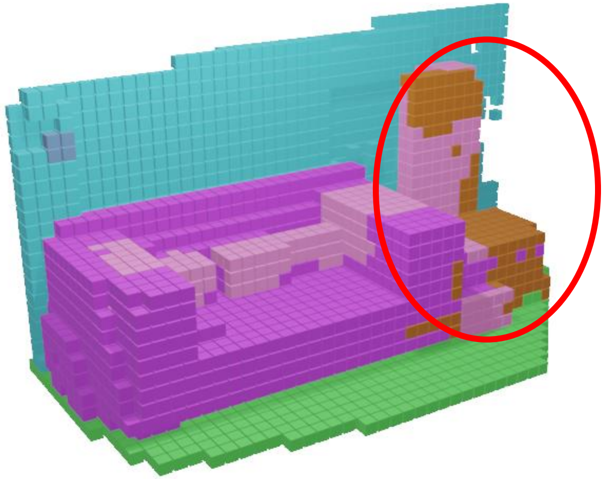


EdgeNet-MF

Qualitative Results



Ground Truth

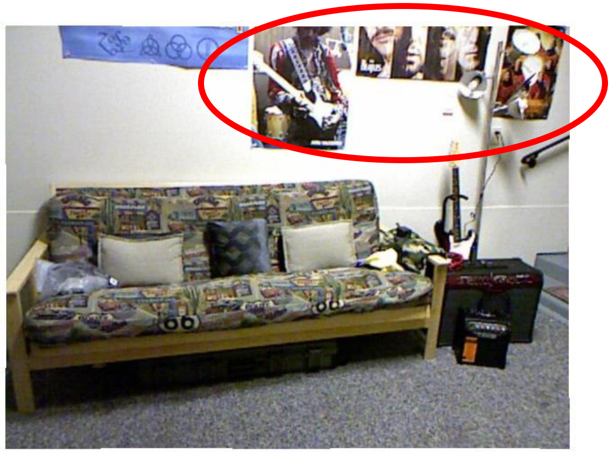


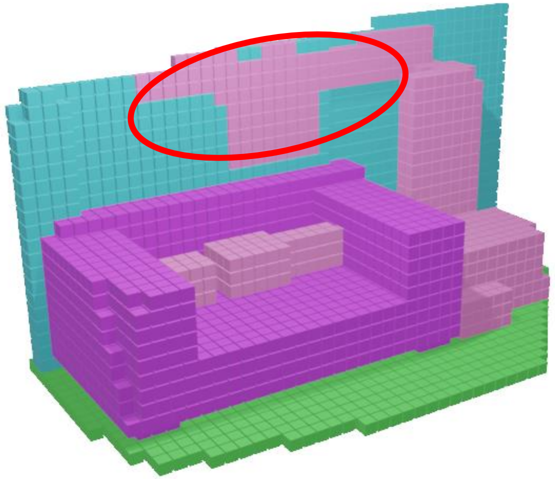
SSCNet

EdgeNet-MF

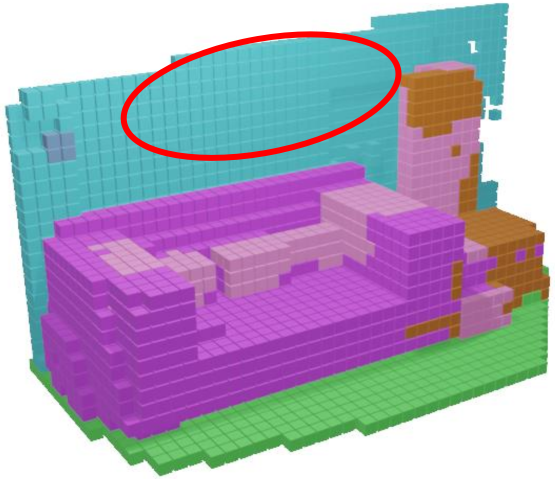
Higher overall accuracy

Qualitative Results

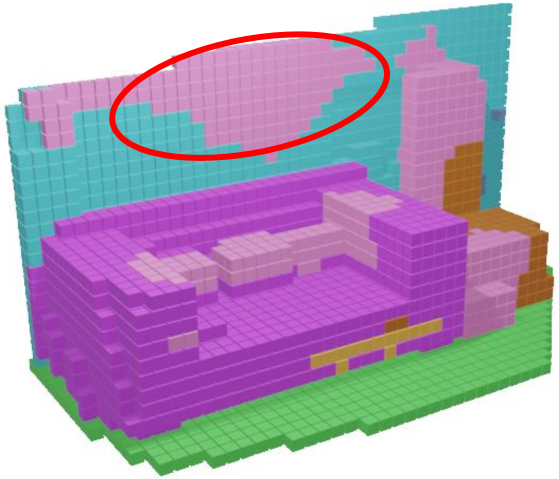




Ground Truth



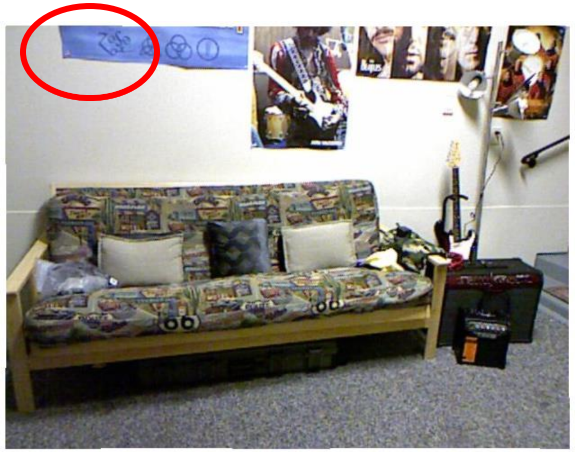
SSCNet

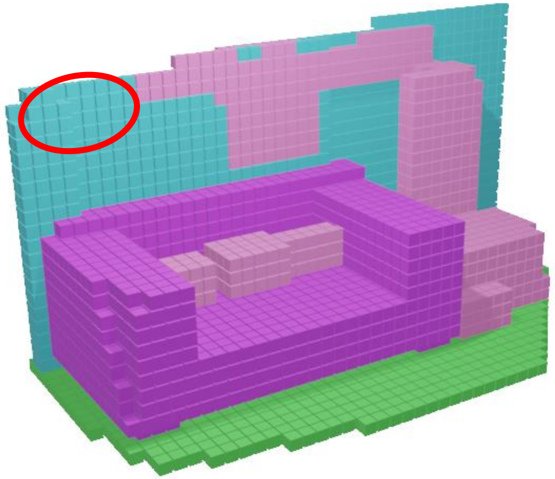


EdgeNet-MF

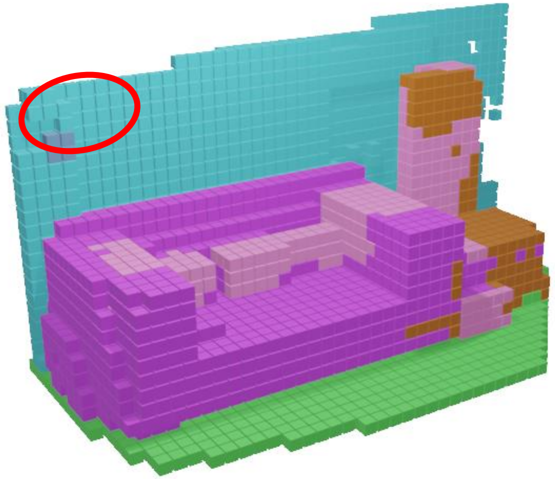
Hard-to-detect classes

Qualitative Results

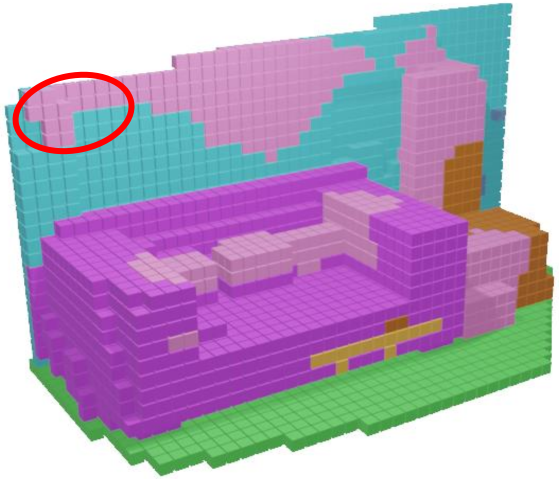




Ground Truth



SSCNet



EdgeNet-MF

NYU Ground Truth errors

Chapter 5 Summary

Contributions

- A new end-to-end network architecture
- A new RGB encoding strategy
- Visually perceptible improvements in 3D
- 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: Semantic 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.

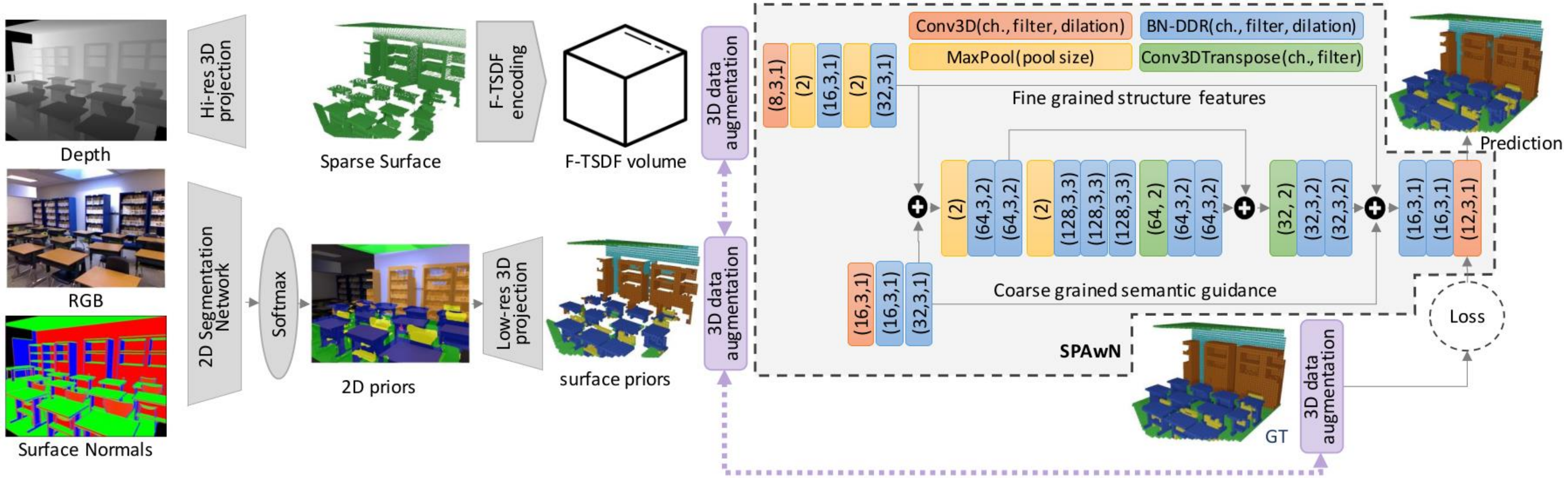
To summarise, our main contributions are:

- 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;

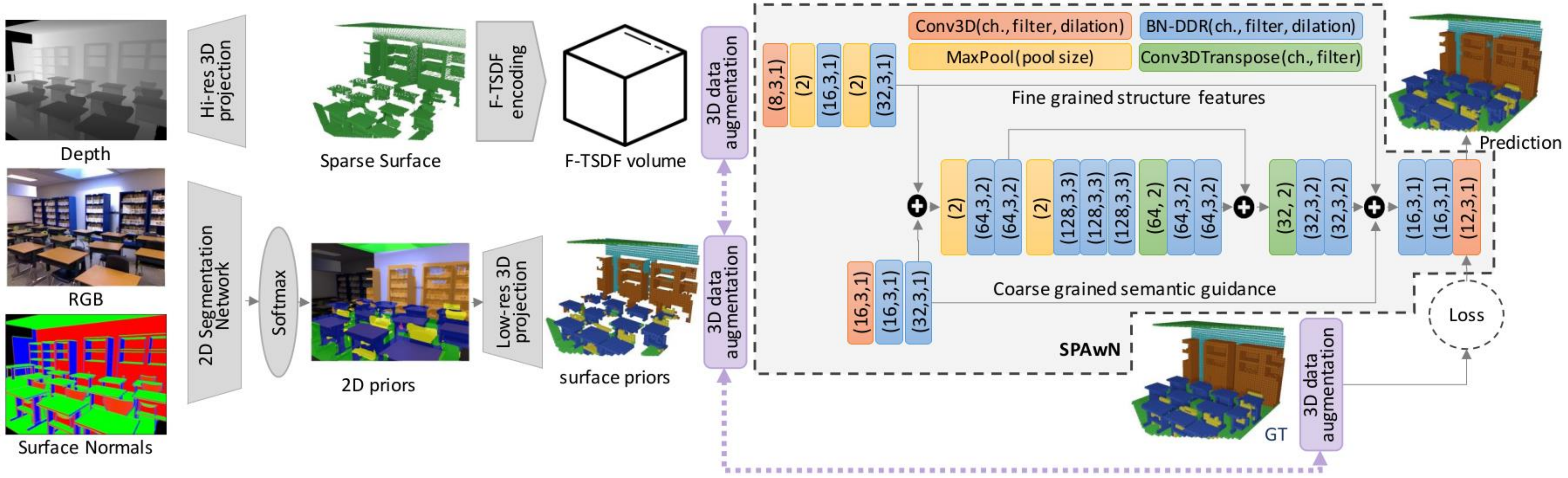
Published in the proceedings of the 25th International Conference on Pattern Recognition (ICPR2020)

Multimodal 3D SSC with 2D Segmentation Priors and Data Augmentation

Chapter 6

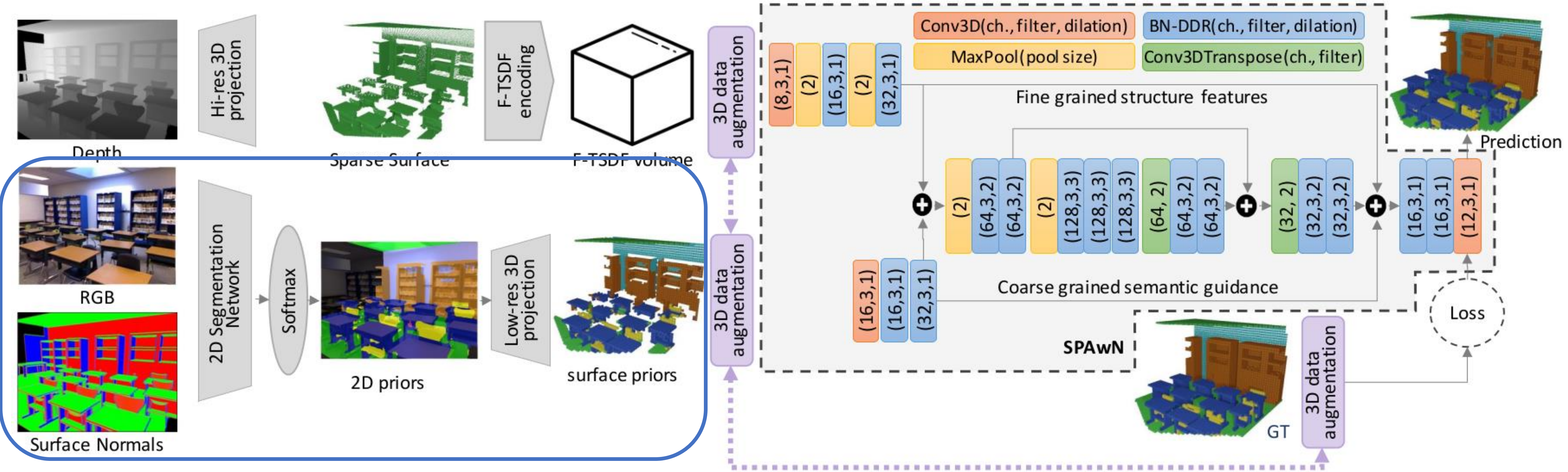


Proposed Solution



SPAwN: Segmentation Priors Aware Network

Proposed Solution

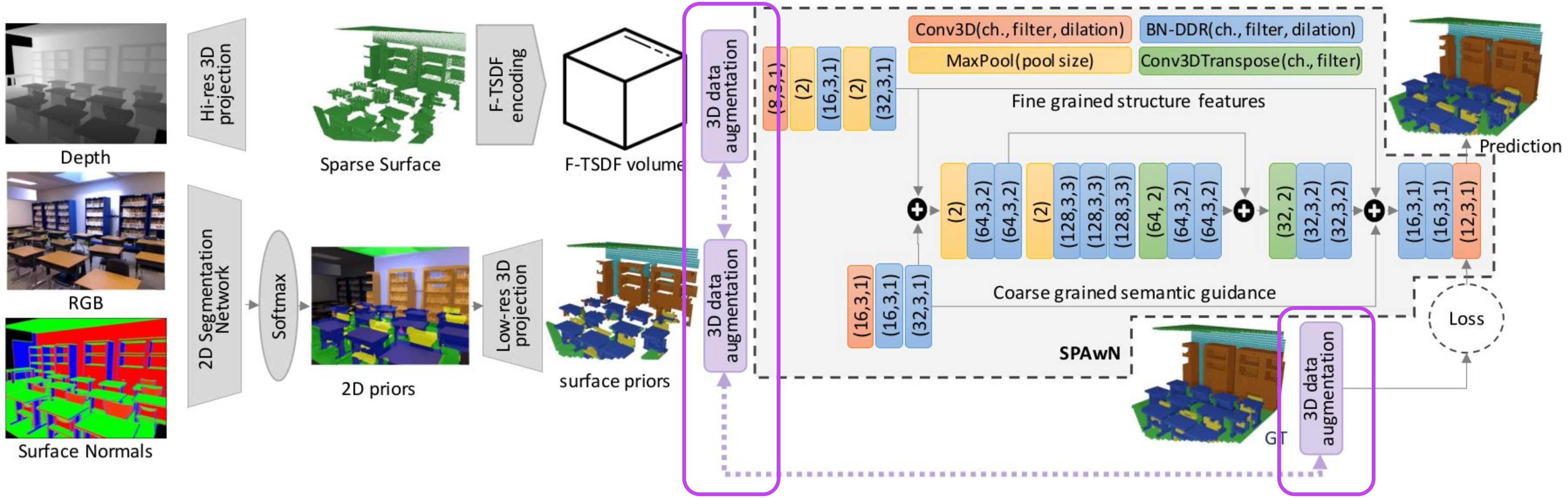


Two Hypothesis

The use of 2D predicted probabilities instead of inner segmentation features

SPAwN: Segmentation Priors Aware Network

Proposed Solution

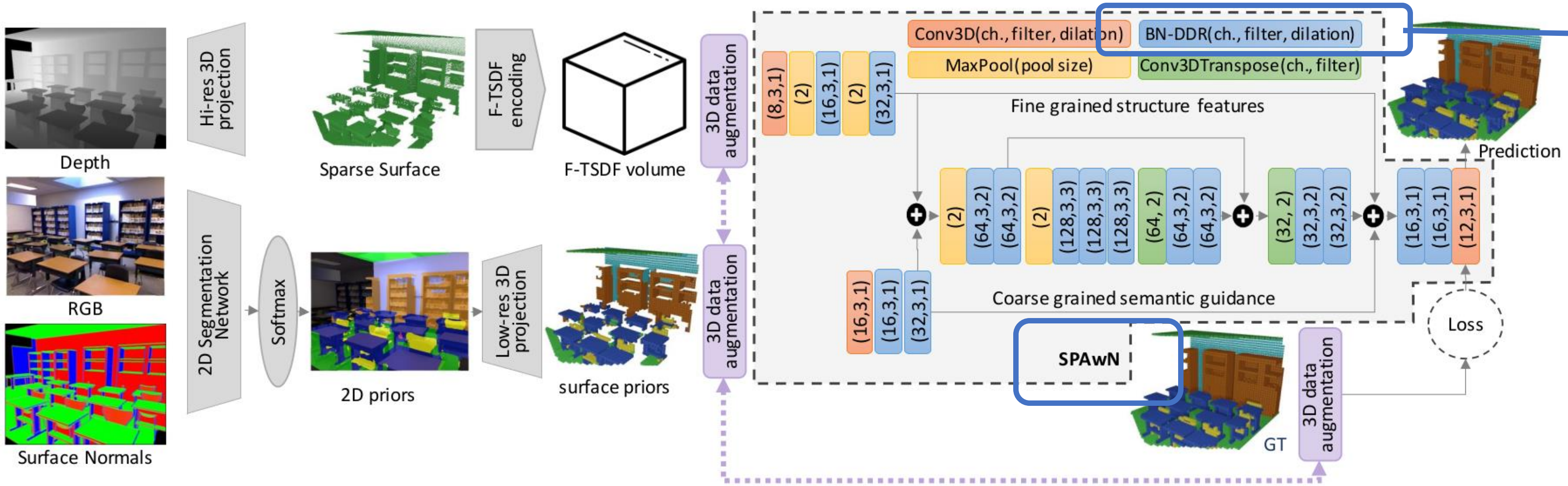


Two Hypothesis

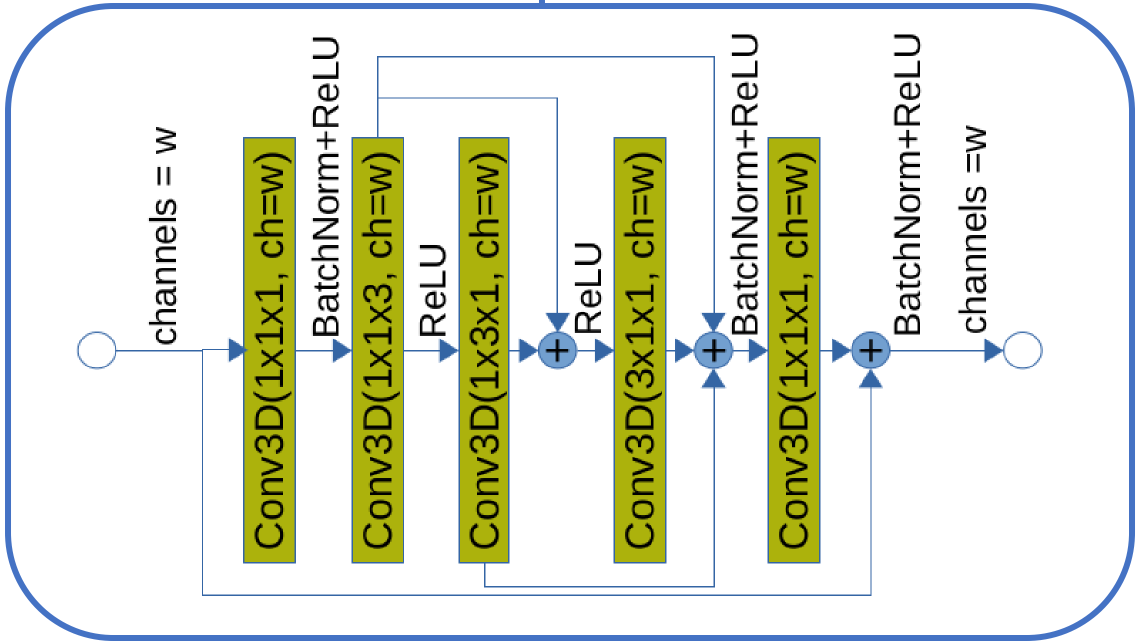
- The use of 2D predicted probabilities instead of inner segmentation features
- The use 3D Data Augmentation

SPAwN: Segmentation Priors Aware Network

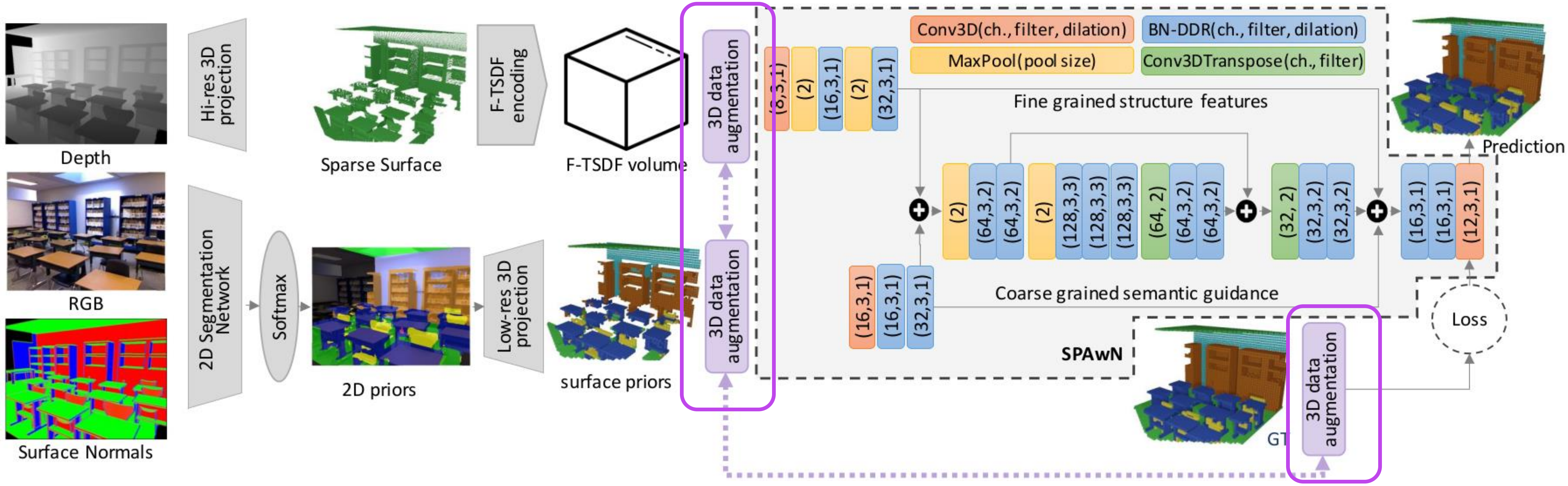
Proposed Solution



BN-DDR: Batch-normalized Dimensional Decomposition Residual Block



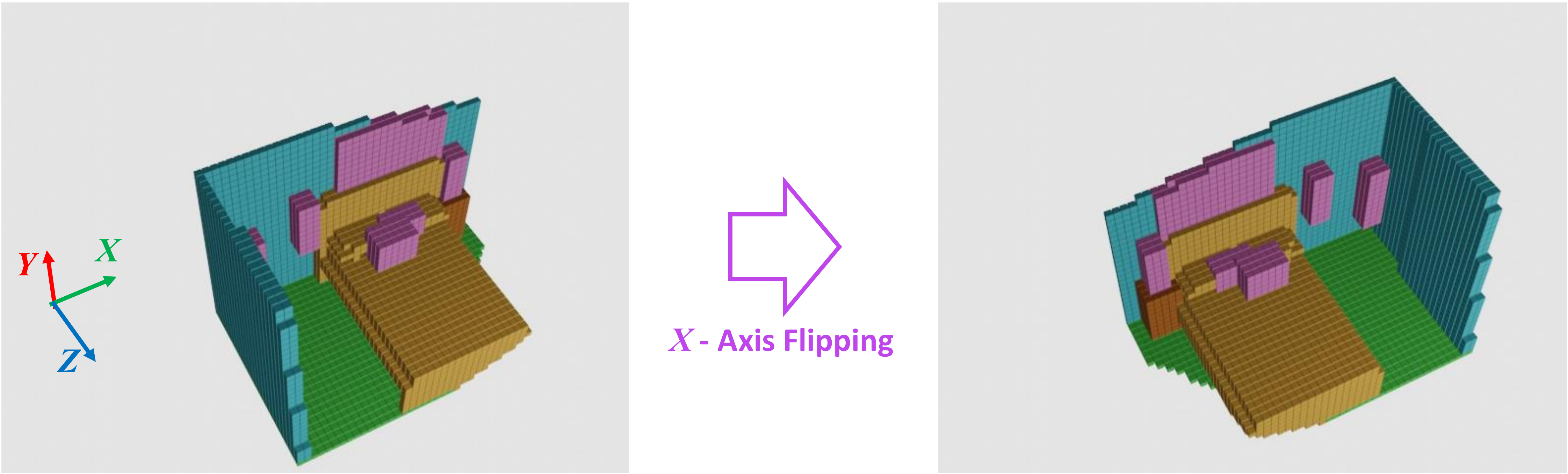
Proposed Solution: Second Hypotesis



3D Data Augmentation

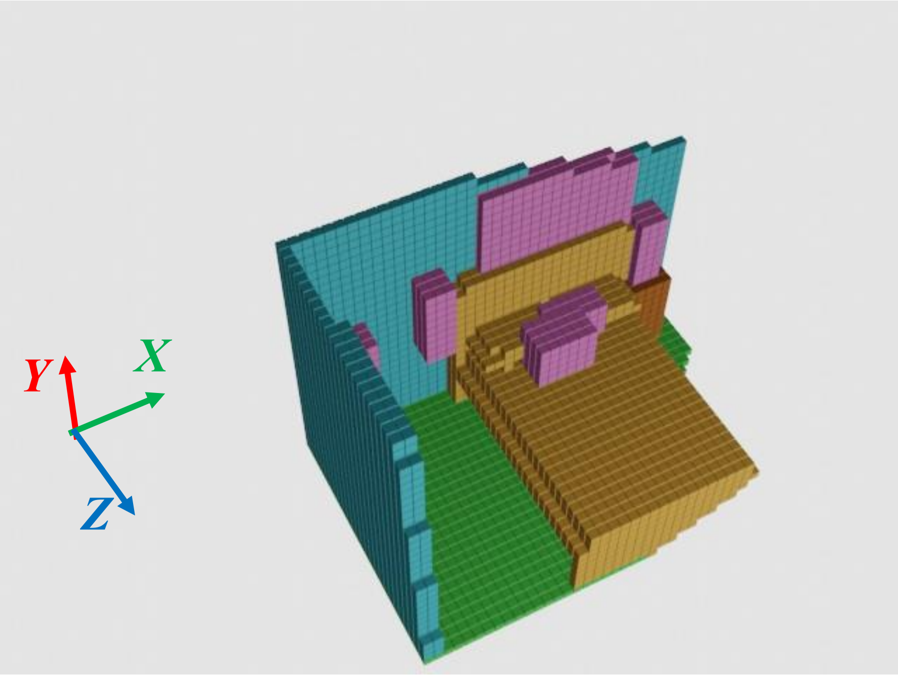
Proposed Solution: Second Hypotesis

3D Data Augmentation – Base Transformations

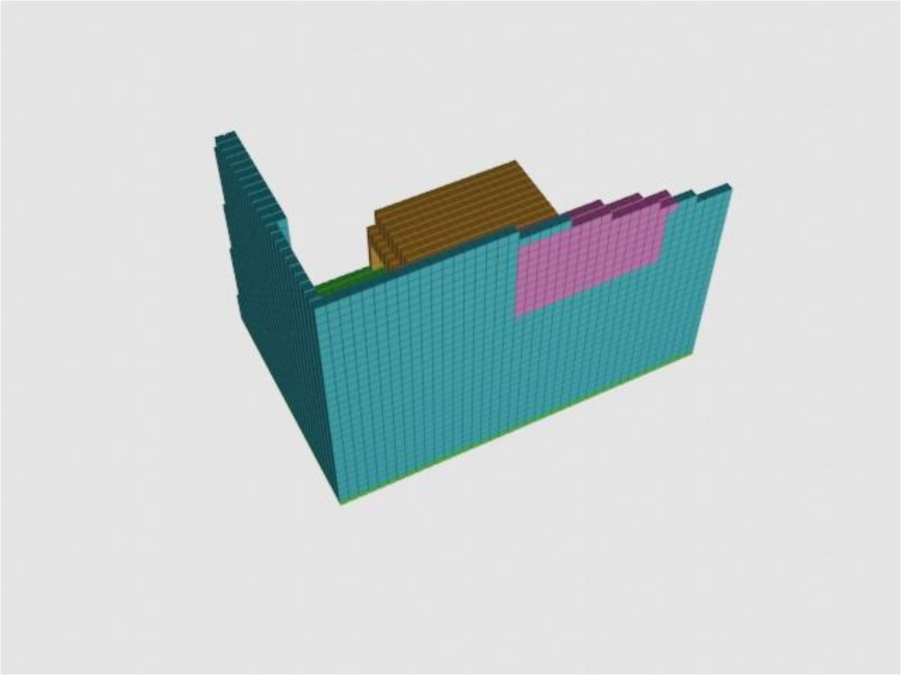


Proposed Solution: Second Hypotesis

3D Data Augmentation – Base Transformations

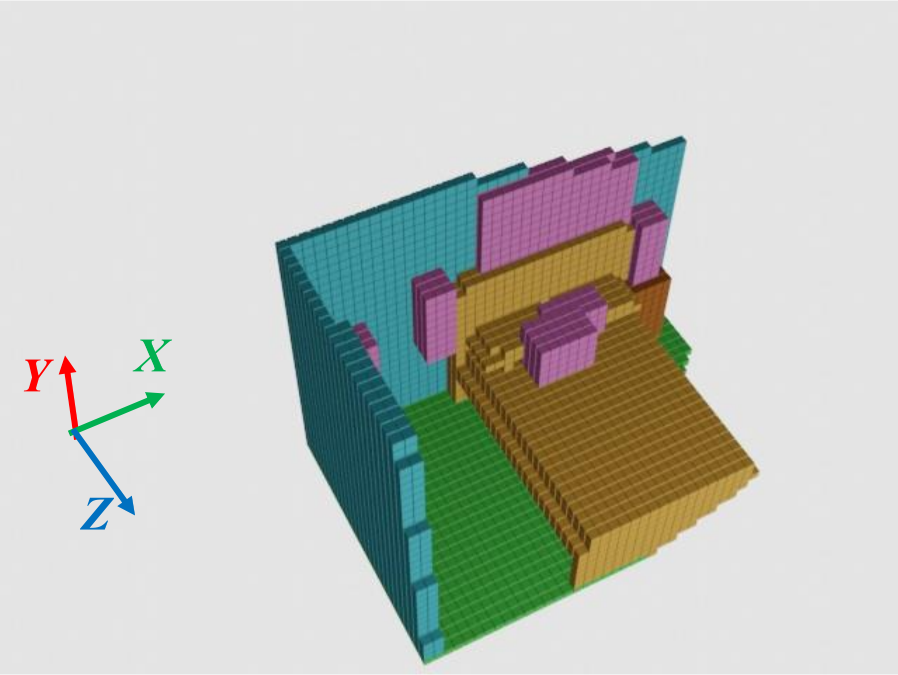


Z - Axis Flipping

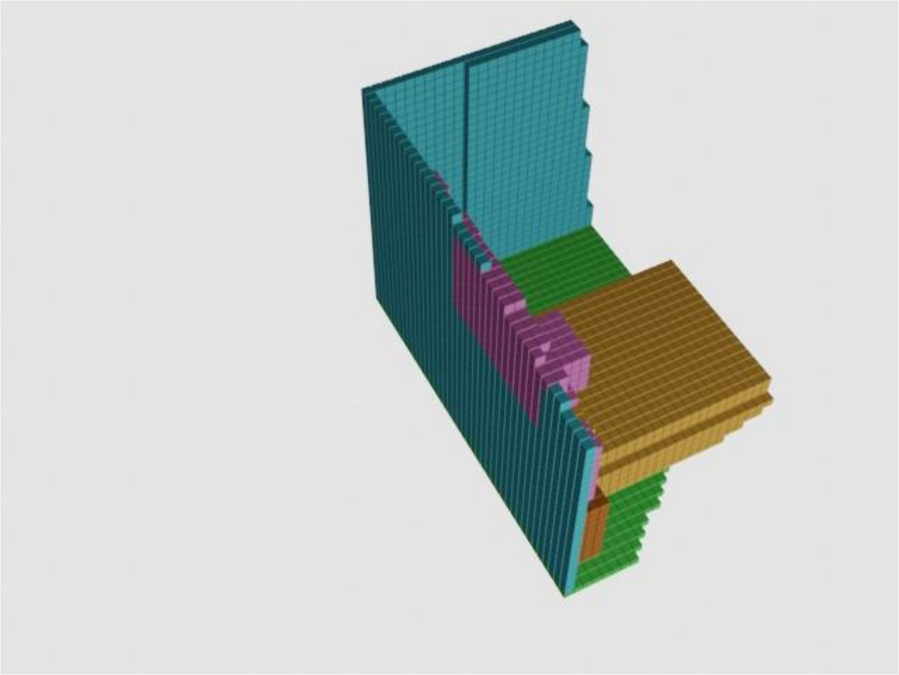


Proposed Solution: Second Hypotesis

3D Data Augmentation – Base Transformations

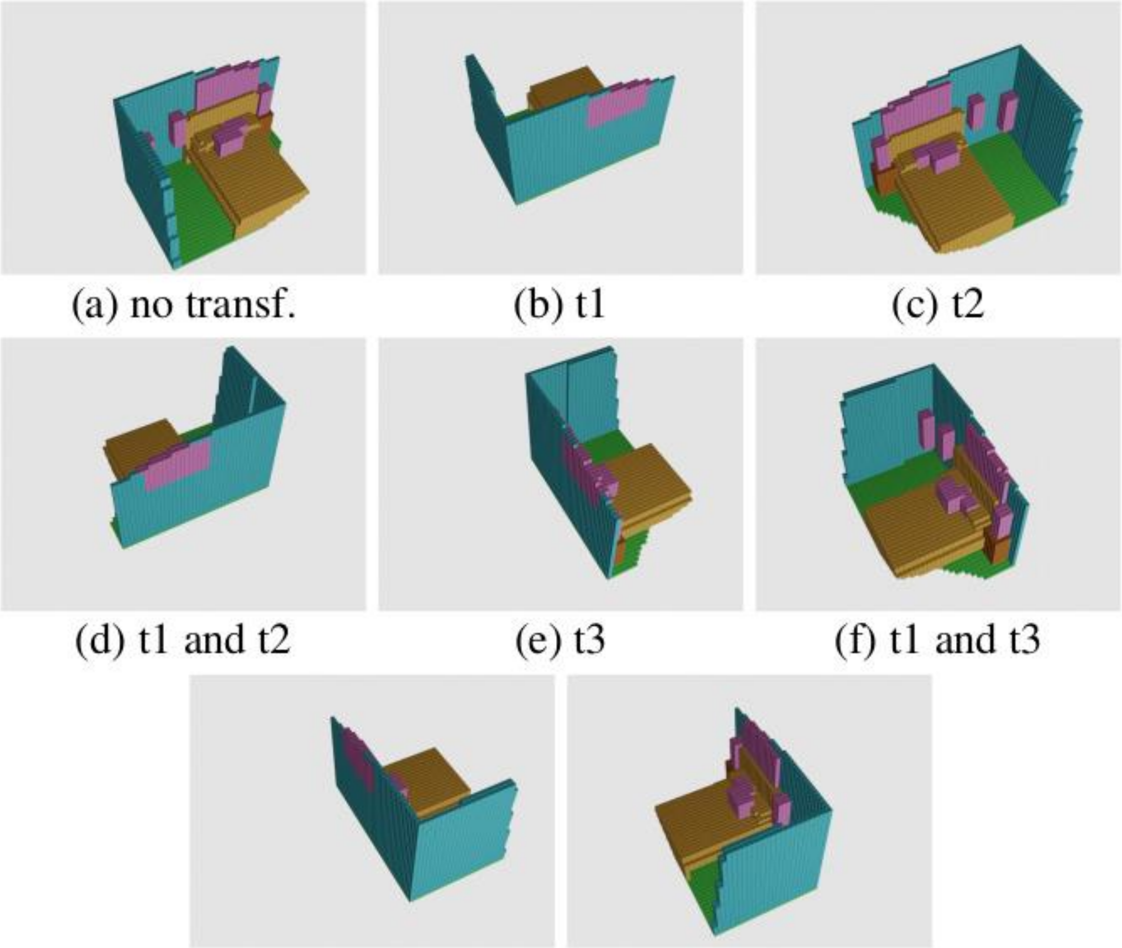


$X \leftrightarrow Z$ axes swapping



Proposed Solution: Second Hypotesis

3D Data Augmentation – All augmented volumes generated from a single scene



(a) no transf. (b) t1 (c) t2

(d) t1 and t2 (e) t3 (f) t1 and t3

(g) t2 and t3 (h) t1, t2 and t3

Ablation Study

input modes	DDR type	class bal.	DA	TTDA	comp. IoU	SSC mIoU
depth	Regular	no	no	no	55.5	24.5
	BN-DDR	no	no	no	60.8	31.8
	BN-DDR	yes	no	no	60.8	32.2
depth rgb	Regular	no	no	no	60.9	38.6
	BN-DDR	no	no	no	63.0	41.0
	BN-DDR	yes	no	no	64.4	42.2
depth rgb sn	Regular	no	no	no	61.3	39.2
	BN-DDR	no	no	no	63.4	41.4
	BN-DDR	yes	no	no	63.8	43.4
	BN-DDR	yes	yes	no	65.7	47.7
	BN-DDR	yes	yes	yes	66.2	48.0
oracle test	*BN-DDR*	yes	no	no	76.7	67.9

Table 1: **Progressive impact of SPAwN components on NYUDv2.** No pretraining was performed. “sn” means surface normals, DA means data augmentation and TTDA means test-time data augmentation.

Comparison to the State-of-the-Art

model	pipeline type	scene completion prec.	scene completion rec.	scene completion IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SISNet-BiSeNet [1]	iterative	93.3	96.1	89.9	85.2	90.0	83.7	80.8	60.0	83.5	80.8	68.6	77.3	86.7	70.1	78.8
SISNet-DeepLabv3 [1]		92.6	96.3	89.3	85.4	90.6	82.6	80.9	62.9	84.5	82.6	71.6	72.6	85.6	69.7	79.0
EdgeNet[3]	straight-forward	93.3	90.6	85.1	97.2	95.3	78.2	57.5,	51.4	80.7	74.1	54.5	52.6	70.3	60.1	70.2
ESSC[34]		92.6	90.4	84.5	96.6	83.7	74.9	59.0	55.1	83.3	78.0	61.5	47.4	73.5	62.9	70.5
CCPNet[36]		**98.2**	**96.8**	**91.4**	99.2	89.3	76.2	63.3	58.2	**86.1**	**82.6**	65.6	53.2	**76.8**	65.2	74.2
SPAwN (ours)		91.9	88.7	82.3	**99.3**	**96.1**	**84.4**	**75.1**	**59.2**	81.5	78.1	**67.3**	**80.1**	76.3	**70.4**	**78.9**

Table 2: **Results on SUNCG test set**. "Straight-forward" means that training and inference are done in a direct pipeline, and iterative means that the pipeline has an iterative loop. Our SPAwN semantic scene completion overall results surpass by far all known previous straight-forward solutions on SUNCG synthetic images, and are comparable to both SISNet models, even though they have a much higher parameter count and operate with a complext iterative pipeline for both training and inference. We highlight the best (bold) and second best (underline) results for the straight-forward models.

Comparison to the State-of-the-Art

model	pipeline type	train	scene compl.			semantic scene completion (IoU, in percentages)											
			prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SISNet-BiSeNet[1]	iterative	NYU	90.7	84.6	77.8	53.9	93.2	51.3	38.0	38.7	65.0	56.3	37.8	25.9	51.3	36.0	49.8
SISNet-DLabv3[1]	iterative	NYU	92.1	83.8	78.2	54.7	93.8	53.2	41.9	43.6	66.2	61.4	38.1	29.8	53.9	40.3	52.4
TS3D[7]	straight-forward	NYU	-	-	60.0	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
SketchAware[2]	straight-forward	NYU	**85.0**	**81.6**	**71.3**	**43.1**	93.6	**40.5**	24.3	30.0	57.1	49.3	29.2	14.3	42.5	28.6	41.1
SPAwN (ours)	straight-forward	NYU	82.3	77.2	66.2	41.5	**94.3**	38.2	**30.3**	**41.0**	**70.6**	**57.7**	**29.7**	**40.9**	**49.2**	**34.6**	**48.0**
TNetFuse[22]	straight-forward	NYU + SUNCG	67.3	85.8	60.6	17.3	92.1	28.0	16.6	19.3	57.5	53.8	17.7	18.5	38.4	18.9	34.4
ForkNet[33]	straight-forward	NYU + SUNCG	-	-	63.4	36.2	93.8	29.2	18.9	17.7	61.6	52.9	23.3	19.5	45.4	20.0	37.1
CCPNet[36]	straight-forward	NYU + SUNCG	**91.3**	**92.6**	**82.4**	25.5	**98.5**	**38.8**	27.1	27.3	64.8	**58.4**	21.5	30.1	38.4	23.8	41.3
SPAwN (ours)	straight-forward	NYU + SUNCG	81.2	80.4	67.8	**44.2**	94.2	**40.9**	**33.5**	**42.5**	**69.3**	58.4	**32.4**	**44.3**	**53.4**	**36.3**	**49.9**

Table 3: **Results on NYUDv2 test set**. SUNCG + NYU means trained on SUNCG and fine-tuned on NYUDv2. Our SPAwN models hold the best and second-best overall semantic scene completion results for real-world images, on both training scenarios, when compared to previous straight-forward solutions.

Comparison to the State-of-the-Art

model	pipeline type	train	scene compl. prec.	rec.	IoU	semantic scene completion (IoU, in percentages) ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SISNet-BiSeNet[1]	iterative	NYUCAD	94.2	91.3	86.5	65.6	94.4	67.1	45.2	57.2	75.5	66.4	50.9	31.1	62.5	42.9	59.9
SISNet-DLabv3[1]	iterative	NYUCAD	94.1	91.2	86.3	63.4	94.4	67.2	52.4	59.2	77.9	71.1	58.1	46.2	65.8	48.8	63.5
CCPNet[36]	straight-forward	NYUCAD + SUNCG	**91.3**	**92.6**	82.4	56.2	**96.6**	58.7	35.1	44.8	68.6	65.3	37.6	35.5	53.1	35.2	53.2
SketchAware[2]	straight-forward	NYUCAD + SUNCG	90.6	92.2	**84.2**	59.7	94.3	**64.3**	32.6	51.7	72.0	68.7	45.9	19.0	60.5	38.5	55.2
SPAwN (ours)	straight-forward	NYUCAD + SUNCG	84.5	87.8	75.6	**65.3**	94.7	61.9	**36.9**	**69.6**	**82.2**	**72.8**	**49.1**	**43.6**	**63.4**	**44.4**	**62.2**
SSCNet[31]	straight-forward	NYUCAD + SUNCG	75.4	**96.3**	73.2	32.5	92.6	40.2	8.9	40.0	60.0	62.5	34.0	9.4	49.2	26.5	40.0
CCPNet[36]	straight-forward	NYUCAD + SUNCG	**93.4**	91.2	**85.1**	58.1	**95.1**	60.5	36.8	47.2	69.3	67.7	39.8	37.6	55.4	37.6	55.0
SPAwN (ours)	straight-forward	NYUCAD + SUNCG	86.3	90.1	78.9	**77.6**	95.0	**68.0**	**38.1**	**67.9**	**82.2**	**77.1**	**56.8**	**50.0**	**65.7**	**46.5**	**65.9**

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

Qualitative Results

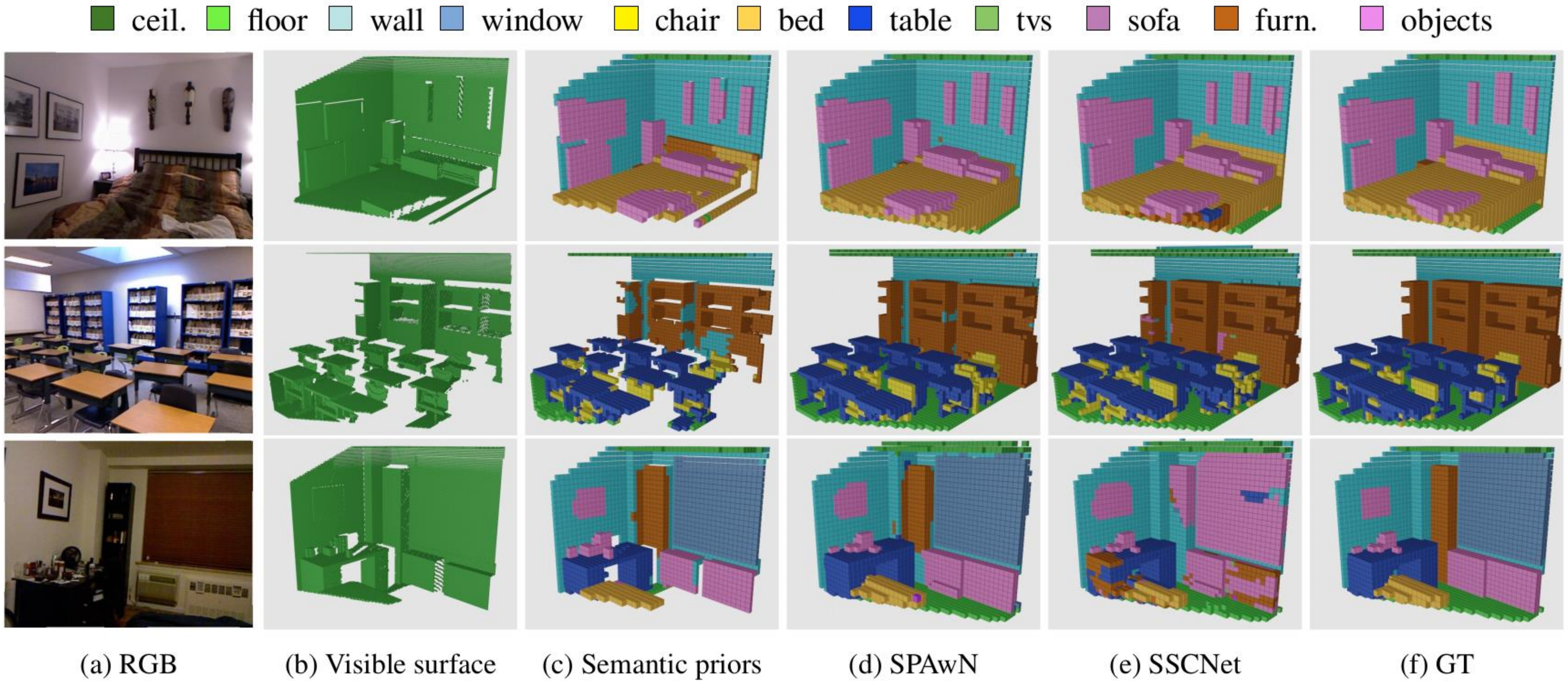


Figure 5: **SPAwN qualitative results on NYUCAD.** 2D segmentation priors projected to 3D provide good semantic guidance while SPAwN complete and refine the predictions, achieving results visually close to perfection. Compared to baseline SSCNet [31], results are much more accurate. (Best viewed in color).

Chapter 6 Summary

Contributions:

- **SPAwN**: novel 3D SSC network that explicitly fuses semantic priors with high-resolution structural information from depth maps.
- **BN-DDR:** batch normalized DDR module with higher discrimination power than its predecessors
- **3D Data Augmentation:** mode and resolution agnostic strategy that may be applied to other SSC solutions to reduce overfitting

Chapter 6 Summary

Results

- **SPAwN alone consistently suparssed all previous straightforward solutions:**
  - **All evaluated datasets**
  - **Multiple training scenarios**

- **SPAwN** when combined with our Data Augmentation strategy presented unprecedent levels of SCC scores achieving a boost of 19.8% (10.9 p.p.) on **NYUCAD**

Publication

Data Augmented 3D Semantic Scene Completion with 2D Segmentation Priors

Data Augmented 3D Semantic Scene Completion with 2D Segmentation Priors

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

Abstract

Semantic scene completion (SSC) is a challenging Computer Vision task with many practical applications, from robotics to assistive computing. Its goal is to infer the 3D geometry in a field of view of a scene and the semantic labels of voxels, including occluded regions. In this work, we present SPAwN, a novel lightweight multimodal 3D deep CNN that seamlessly fuses structural data from the depth component of RGB-D images with semantic priors from a bimodal 2D segmentation network. A crucial difficulty in this field is the lack of fully labeled real-world 3D datasets which are large enough to train the current data-hungry deep 3D CNNs. In 2D computer vision tasks, many data augmentation strategies have been proposed to improve the generalization ability of CNNs. However those approaches cannot be directly applied to the RGB-D input and output volume of SSC solutions. In this paper, we introduce the use of a 3D data augmentation strategy that can be applied to multimodal SSC networks. We validate our contributions with a comprehensive and reproducible ablation study. Our solution consistently surpasses previous works with a similar level of complexity.

1. Introduction

Reasoning about scenes in 3D is a natural human ability that remains a challenge for Computer Vision. In the past, the two most common scene understanding tasks were scene completion [6] and semantic labeling of visible surfaces [11, 26, 27]. Noticing that these are intertwined tasks, in 2017, Song et al. [31] introduced the Semantic Scene Completion (SSC) task for simultaneously completing occluded voxels and inferring their semantic labels and proposed SSCNet, achieving better results than dealing with these tasks separately. Early approaches only used depth information, ignoring the RGB channels [31, 10]. The use of color channels was introduced later [8].

We present a new approach for exploring information from the RGB-D input (explained in section 3), as shown in Figure 1. The solution uses 2D prior probabilities from a bimodal 2D segmentation network as semantic guidance to the depth map's structural data. The proposed multimodal 3D network, *SPAwN*, uses a new memory-saving batch-normalized dimensional decomposition residual building block (BN-DDR) and can be trained on a single 10Gb GPU with a 4 scene mini-batch.

To overcome the limitations imposed by the lack of sizeable real-world datasets, we are the first to apply 3D data augmentation for the SSC task. Data augmentation is widely used in the training of 2D deep CNNs [17, 14] and its goal is to reduce overfitting by artificially increasing the variety of samples in the training dataset using transformations like flipping, cropping, rotation and color transforms. However, those transformations can not naïvely be used in 3D applications like semantic Scene completion because of the difference in the number of dimensions of the input (2D) and output (3D). In this paper, we propose to apply data augmentation to inner 3D volumes of the solution with three fast 3D transformations in voxel space that preserve the main characteristics of the scene. Our proposed data augmentation approach reduces overfitting and achieves unprecedented levels of semantic completion when compared to previous works of similar memory footprint and complexity.

We evaluated our contributions with and without pre-training on synthetic data and observed that our method surpasses, by far, all previous state-of-the-art results on both scenarios. We demonstrate the benefits of the proposed architecture and the data augmentation approach separately, with several experiments in a comprehensive and reproducible ablation study. Regarding the proposed augmentation scheme, we evaluate it for training (regular data augmentation) and test (test-time data augmentation).

Supplementary material provides additional graphs and data regarding all experiments. All models and training code necessary to reproduce our results and the ablation experiments are publicly available[1] .

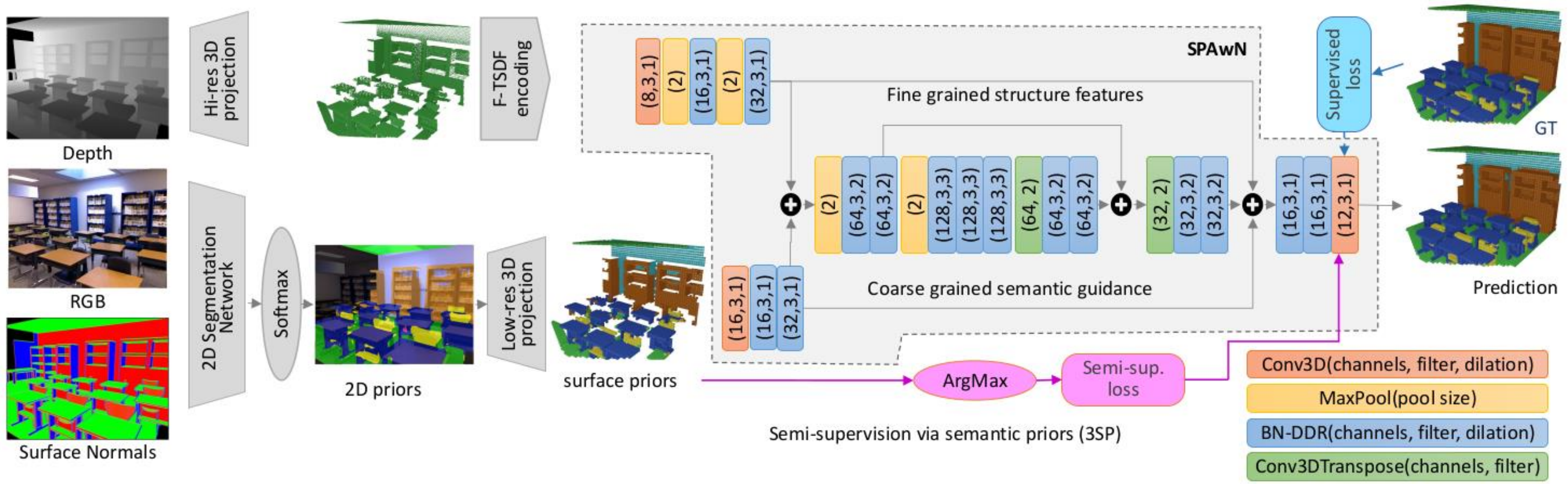
Our contributions are listed below.

[1] Source code: https://cic.unb.br/~teodecampos/aloisio/

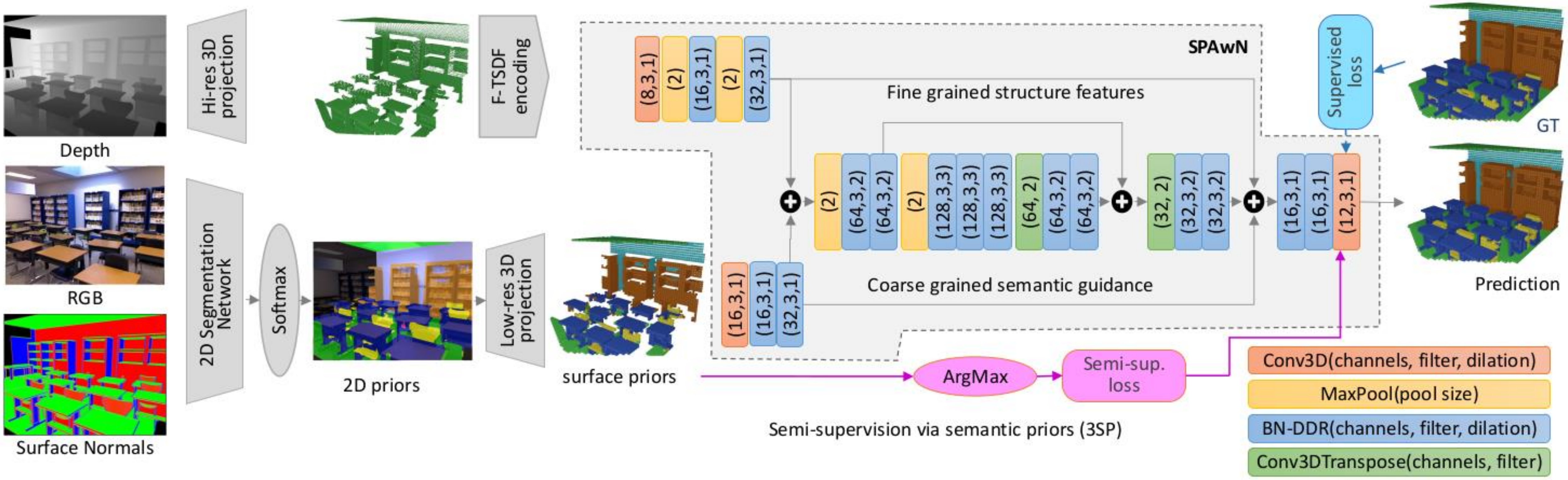
Published in the proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022)

Chapter 7

Exploying unlabeled data to enhance SSC scores

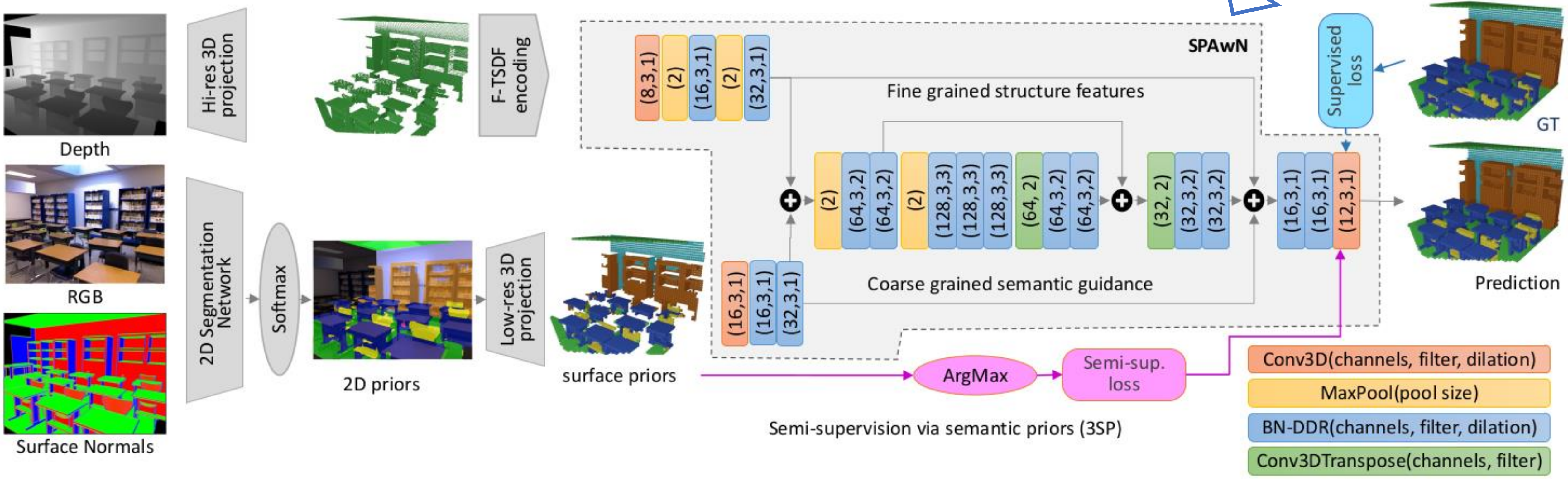


Proposed Solution:
Semi-Supervision via Segmentation Priors (S3P)

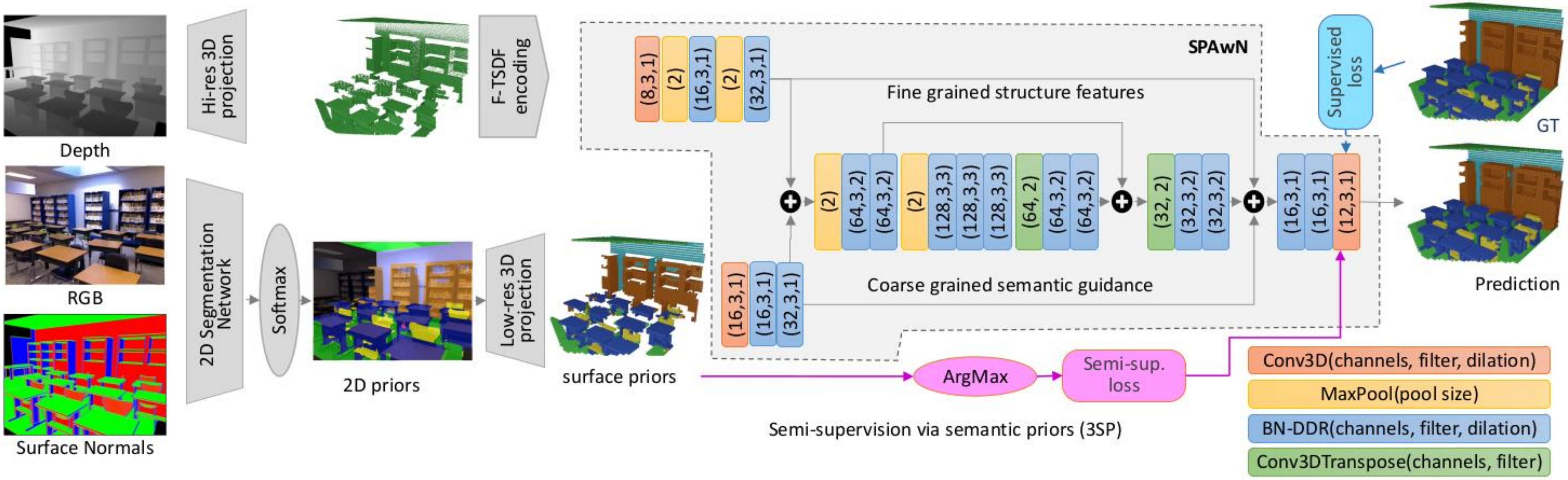


Proposed Solution:
Semi-Supervision via Segmentation Priors (S3P)

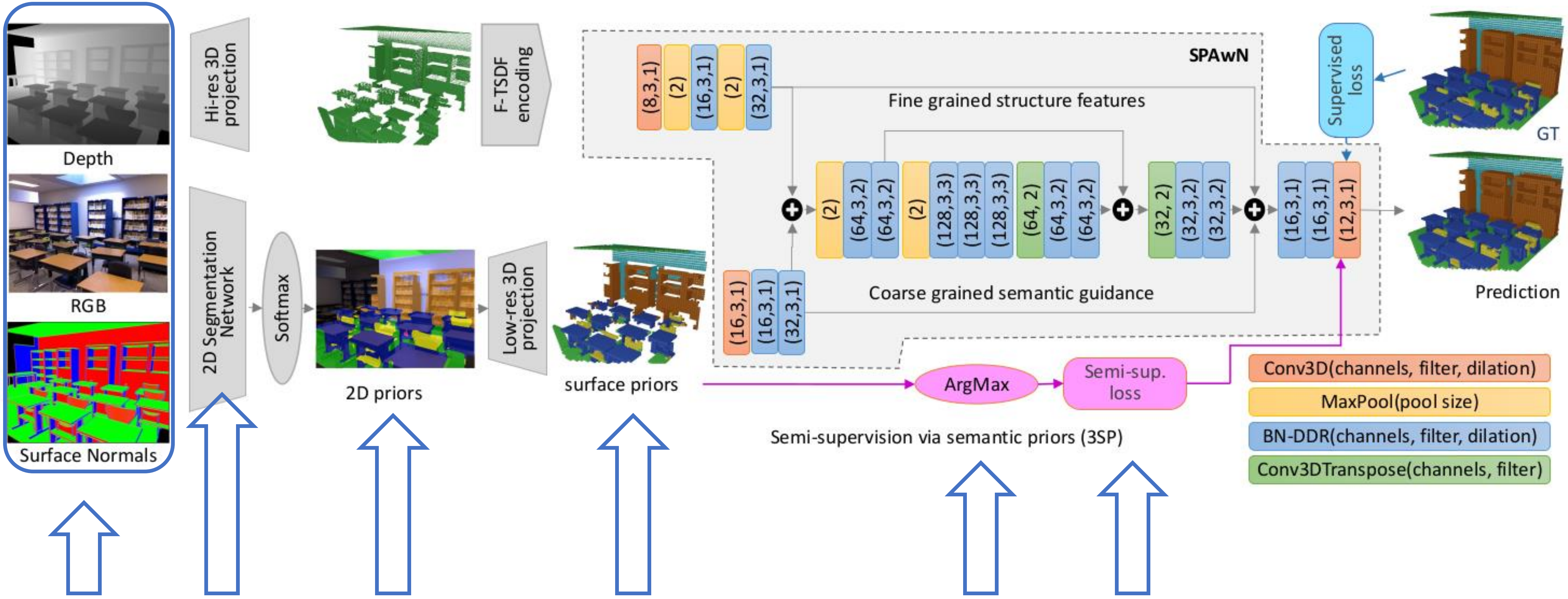
SPAwN



Proposed Solution:
Semi-Supervision via Segmentation Priors (S3P)



Proposed Solution:
Semi-Supervision via Segmentation Priors (S3P)



Unlabeled data

Pre-trained 2D FCN

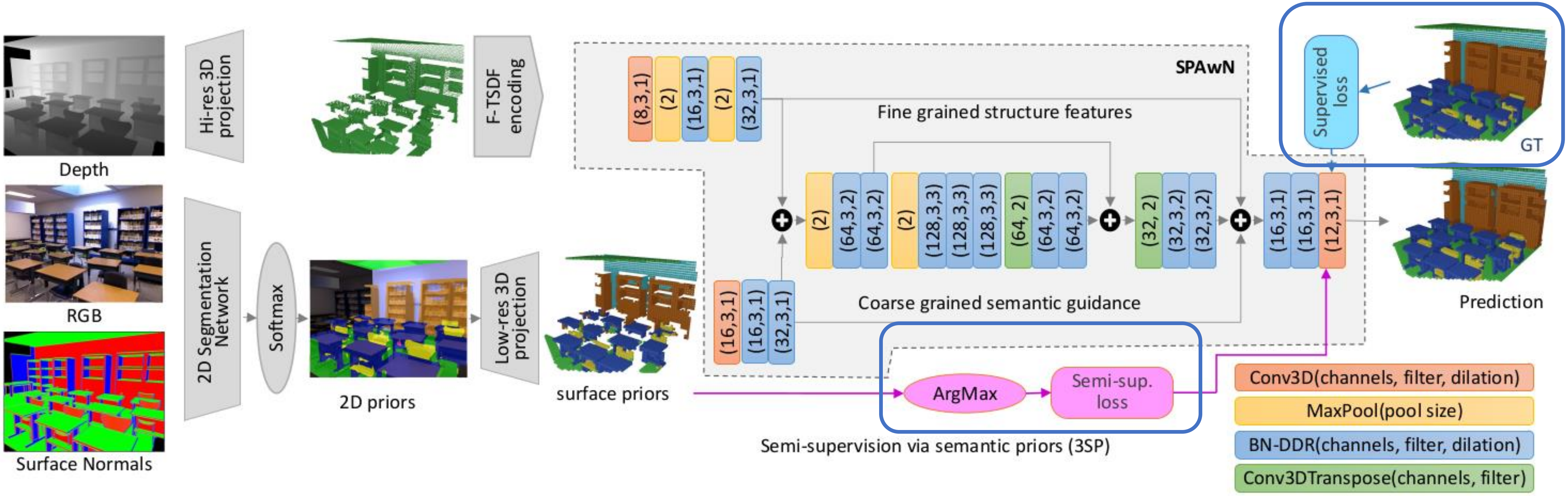
High quality 2D priors

2D-3D projection

Semi-sup GT generation

Semi-sup loss

Proposed Solution:
Semi-Supervision via Segmentation Priors (S3P)



Ablation Study

input modes	DDR type	class balancing	train type	SSC IoU
depth	Regular	no	Sup.	21.6
	BN-DDR	no	Sup.	28.4
	BN-DDR	yes	Sup.	30.1
	BN-DDR	yes	S-Sup.	39.1
depth+rgb	Regular	no	Sup.	34.9
	BN-DDR	no	Sup.	38.4
	BN-DDR	yes	Sup.	39.4
	BN-DDR	yes	S-Sup.	<u>43.5</u>
depth+rgb+sn	Regular	no	Sup.	35.2
	BN-DDR	no	Sup.	39.2
	BN-DDR	yes	Sup.	41.4
	BN-DDR	yes	S-Sup.	**45.1**
oracle test	*BN-DDR*	yes	Sup.	67.9
	BN-DDR	yes	S-Sup.	67.9

Table 1: **Progressive impact of our contributions on NYUDv2.** No pretraining was performed. "sn" means surface normals. "Sup." and "S-Sup." mean supervised and semi-supervised training respectively.

Ablation Study

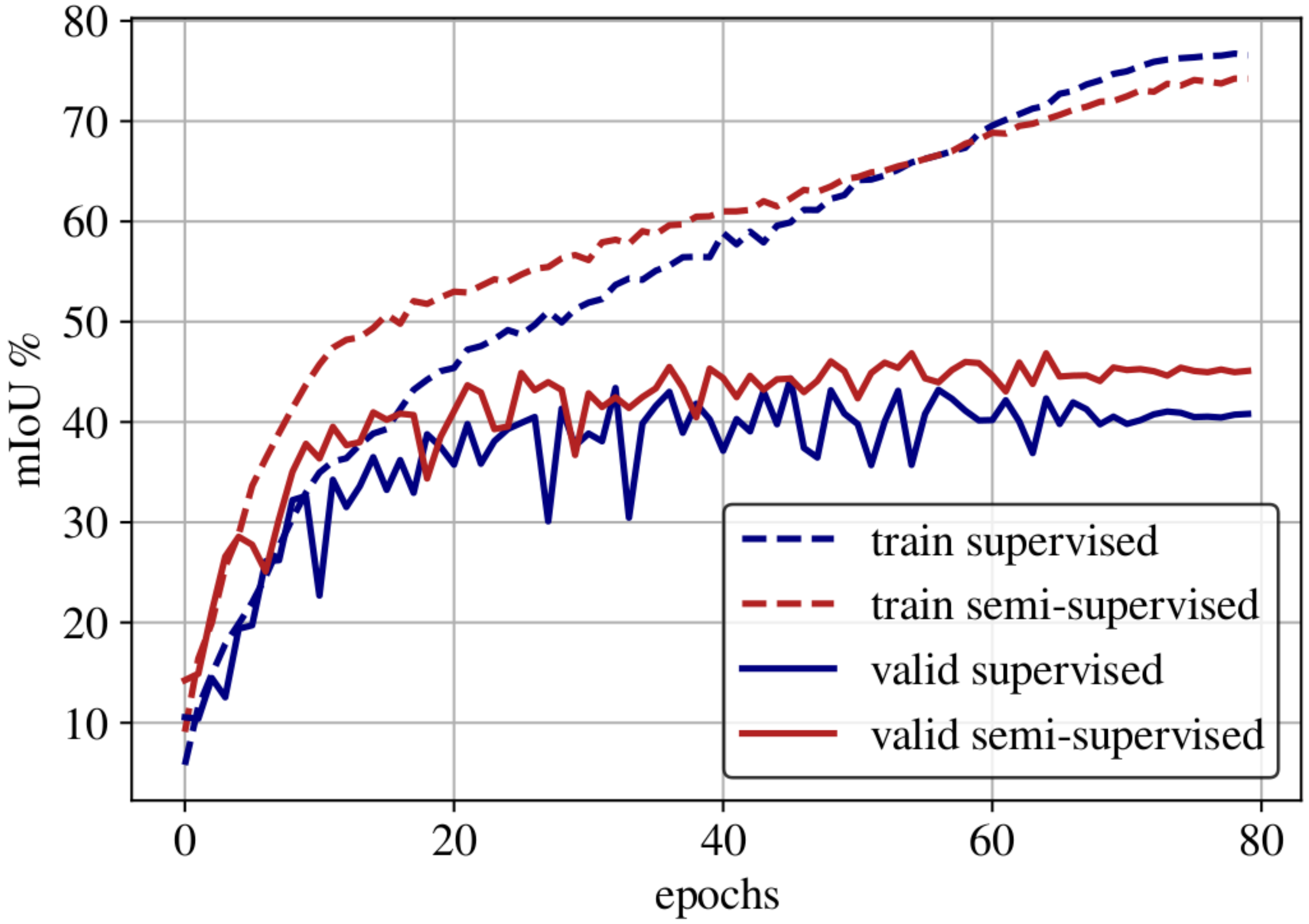


Figure 5: **Effect of the semi-supervised training** over model overfitting and regularization on NYDv2.

Comparison to the State-of-the-Art

train	model	semantic scene completion (IoU, in percentages)											
		ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
NYUDv2	TS3D[6]	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
	CCPNet[40]	23.5	**96.3**	35.7	20.2	25.8	61.4	<u>56.1</u>	18.1	28.1	37.8	20.1	38.5
	SketchAware[1]	**43.1**	93.6	**40.5**	24.3	30.0	57.1	49.3	**29.2**	14.3	42.5	<u>28.6</u>	41.1
	SPAwN (sup.)	22.9	<u>94.8</u>	35.8	<u>25.4</u>	<u>33.2</u>	<u>65.6</u>	54.4	20.0	<u>33.5</u>	<u>44.2</u>	25.7	<u>41.4</u>
	SPAwN+S3P (s-sup.)	<u>35.6</u>	94.4	<u>37.0</u>	**30.4**	**36.8**	**68.5**	**58.9**	<u>23.4</u>	**32.3**	**47.9**	**30.6**	**45.1**
SUNCG + NYUDv2	TNetFuse[23]	17.3	92.1	28.0	16.6	19.3	57.5	53.8	17.7	18.5	38.4	18.9	34.4
	ForkNet[36]	36.2	93.8	29.2	18.9	17.7	61.6	52.9	<u>23.3</u>	19.5	45.4	20.0	37.1
	CCPNet[40]	25.5	**98.5**	**38.8**	<u>27.1</u>	27.3	64.8	**58.4**	21.5	<u>30.1</u>	38.4	23.8	41.3
	SPAwN (sup.)	<u>31.5</u>	<u>94.5</u>	<u>38.7</u>	27.0	<u>32.8</u>	<u>67.6</u>	57.2	20.9	**30.7**	<u>47.5</u>	<u>27.2</u>	<u>43.2</u>
	SPAwN+S3P (s-sup.)	**37.5**	93.6	37.8	**35.0**	**39.4**	**71.9**	<u>58.2</u>	**23.4**	29.7	**50.7**	**34.2**	**46.5**

Table 3: **Results on NYUDv2 test set**. The column "train" indicates datasets used for training the models. SUNCG + NYU means trained on SUNCG and fine-tuned on NYUDv2. Our SPAwN semi-supervised and supervised models hold the best and second-best overall semantic scene completion results for real-world images, on both training scenarios.

Qualitative Results

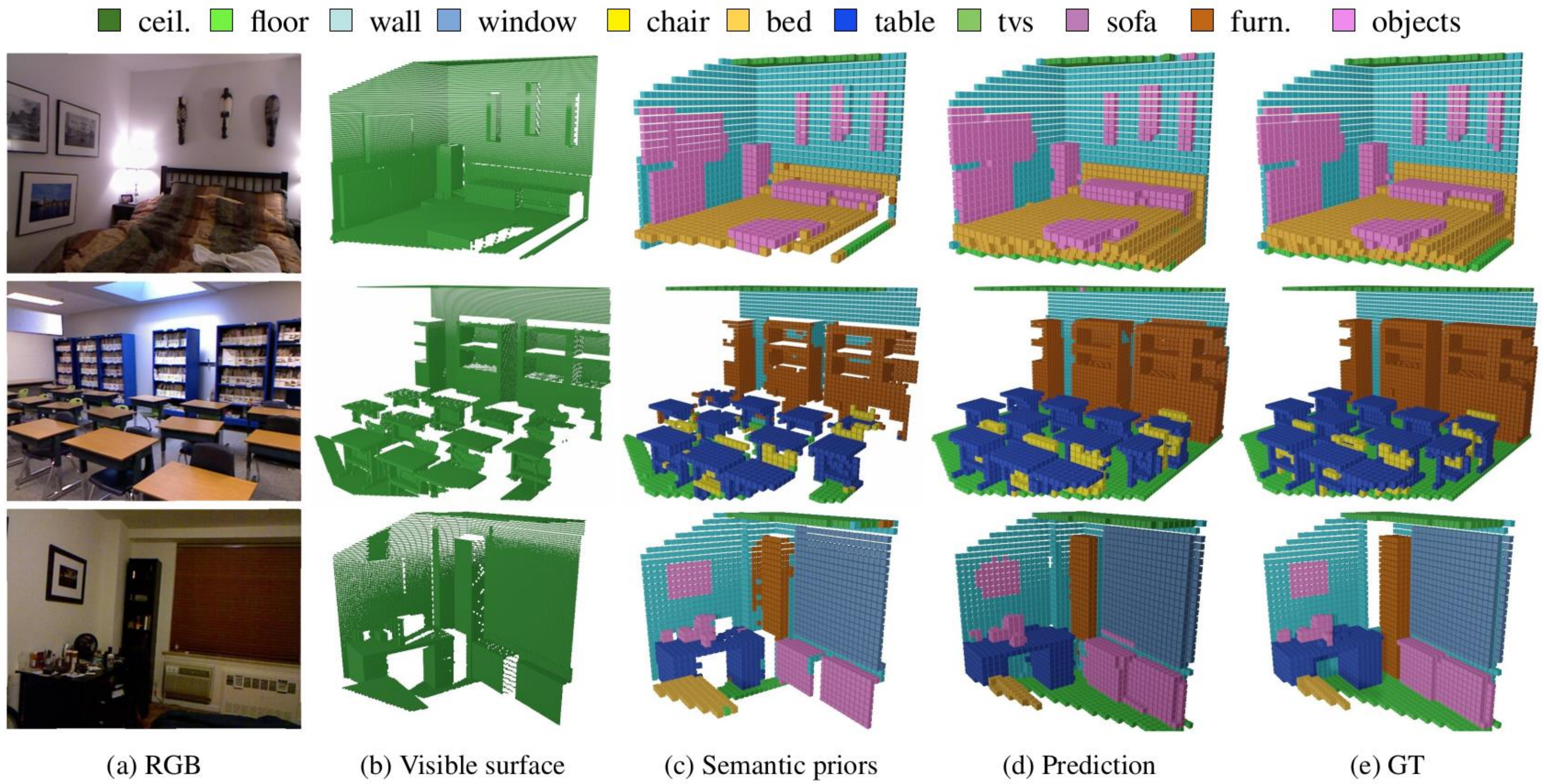


Figure 7: **SPAwN & S3P qualitative results on NYUCAD.** 2D segmentation priors projected to 3D provide good semantic guidance. However, the resulting volume is incomplete and still presents some errors. SPAwN & S3P together complete and refine the predictions, and final results are visually close to perfection. (Best viewed in color).

Chapter 7 Summary

- Remarkable Results

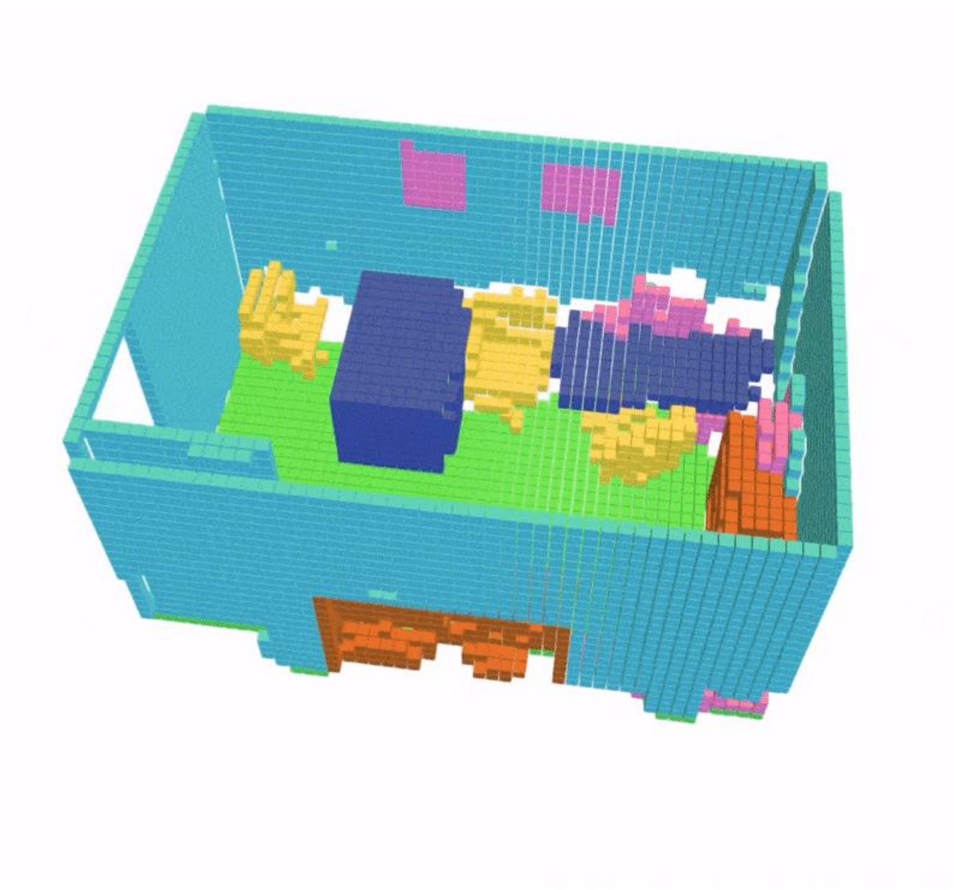
  - **SPAwN alone had consistently suparssed previous state-of-the-art:**
    - **All evaluated datasets**
    - **Multiple training scenarios**

  However,

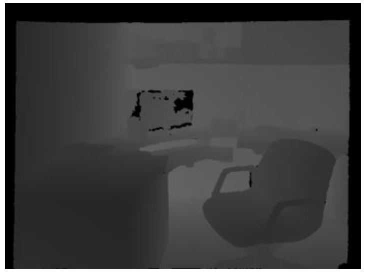
  - **SPAwN** when combinedwith **S3P** presented unprecedent levels of SCC scores achieving a boost of 12.6% (5.2 p.p.) on **NYUdV2**

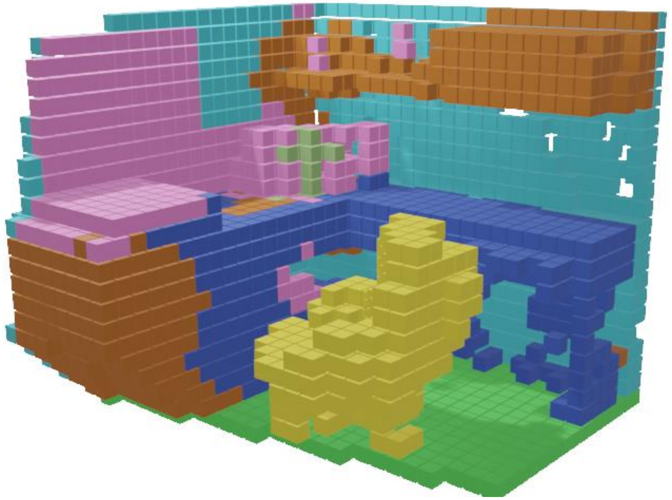
Chapter 8

Extending Semantic Scene Completion for 360^{O} Coverage

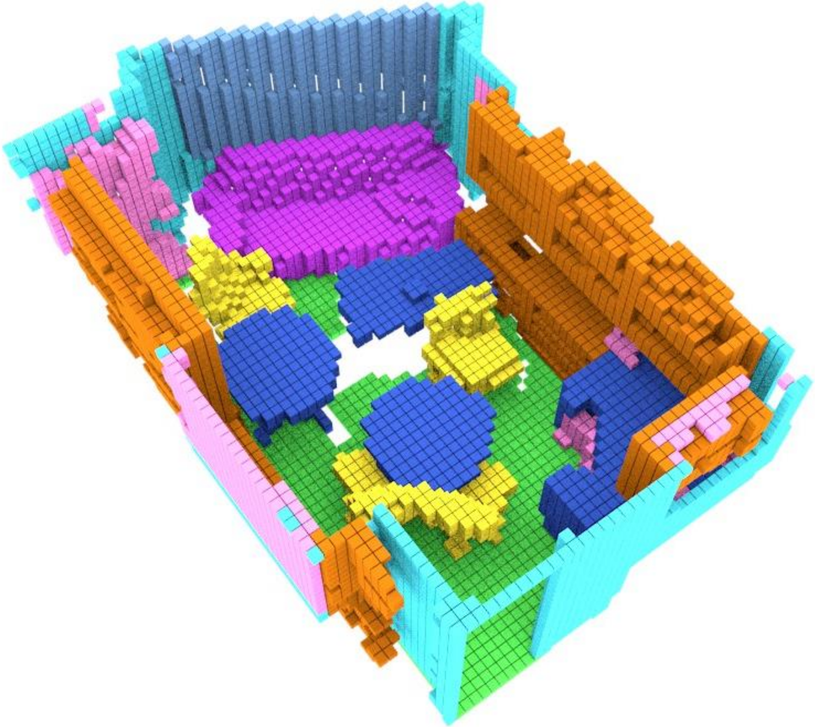


Current Semantic Scene Completion Limitations



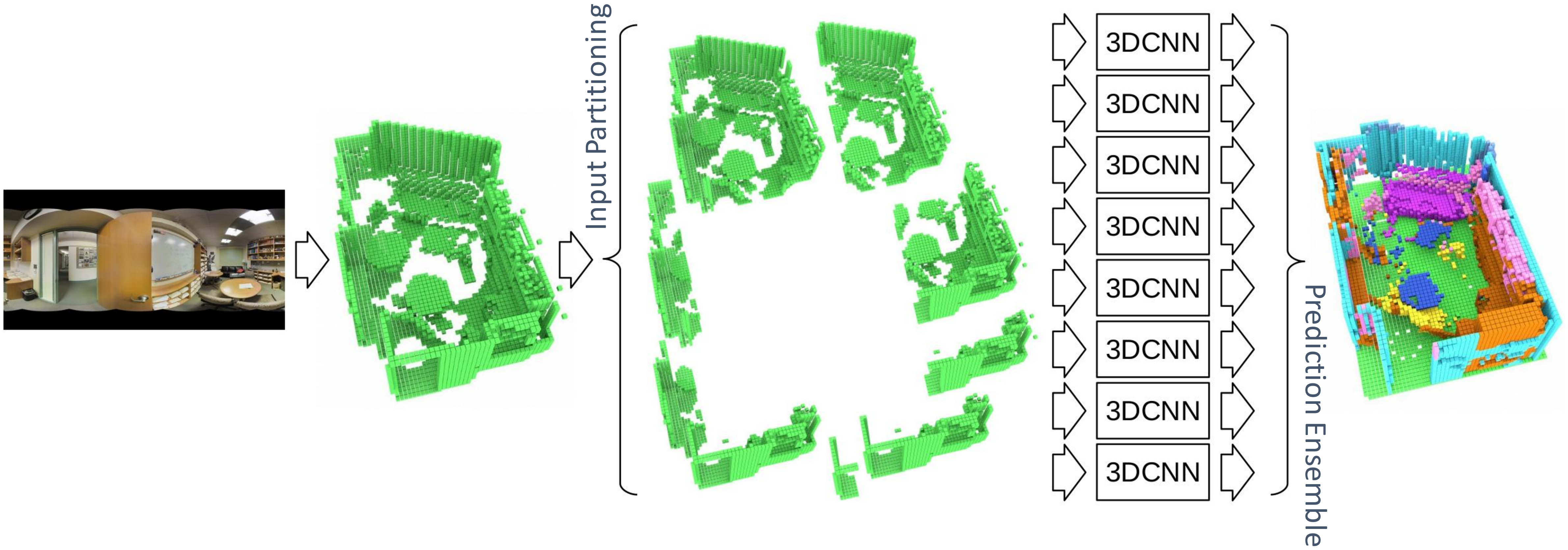


Regular RGB-D Sensor



Panoramic Image from
Matterport Camera

Our approach



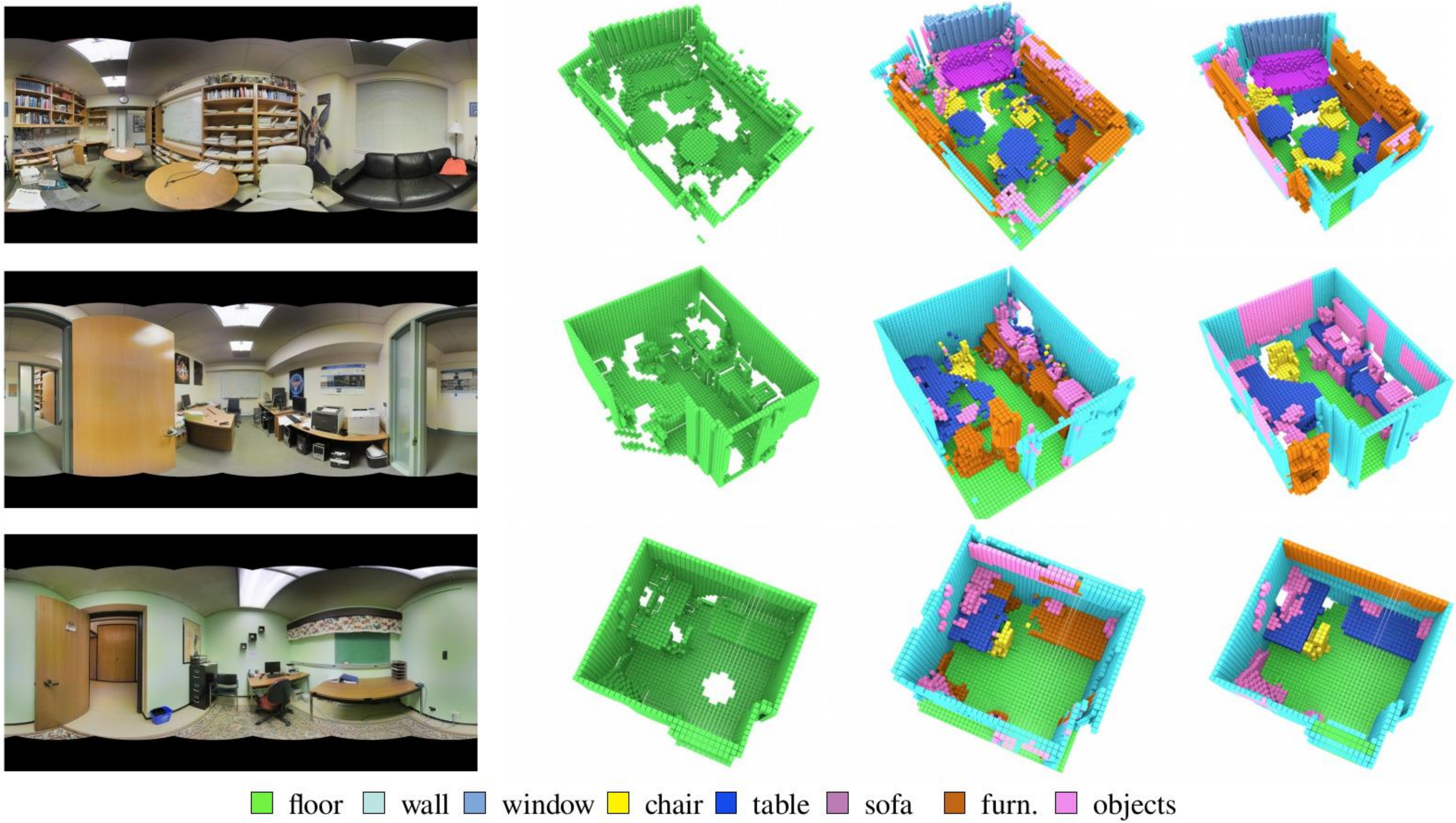
The 3DCNN is trained using SUNCG and fine-tuned in NYUDV2

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

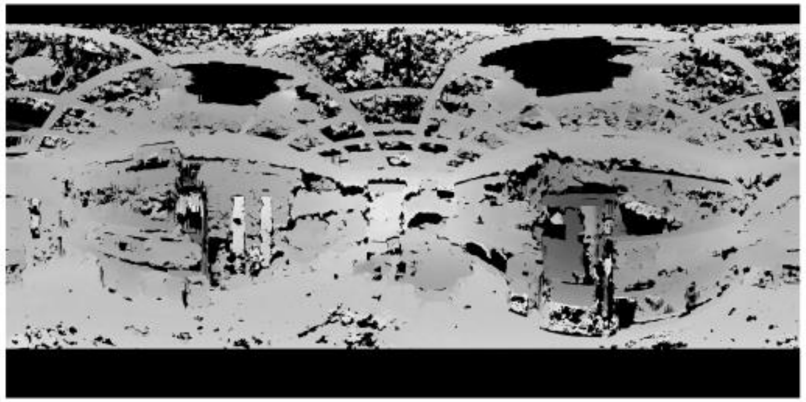
GT Pred.

Results on Stanford 2D-3DS Dataset

RGB Image | Input Volume | Predicted Volume | GT

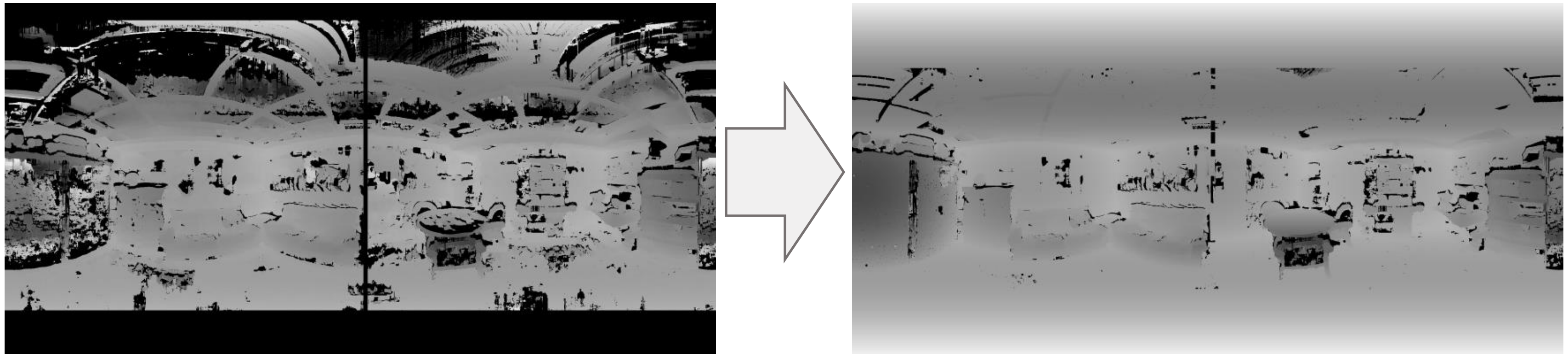


Experiments on Spherical Stereo Images

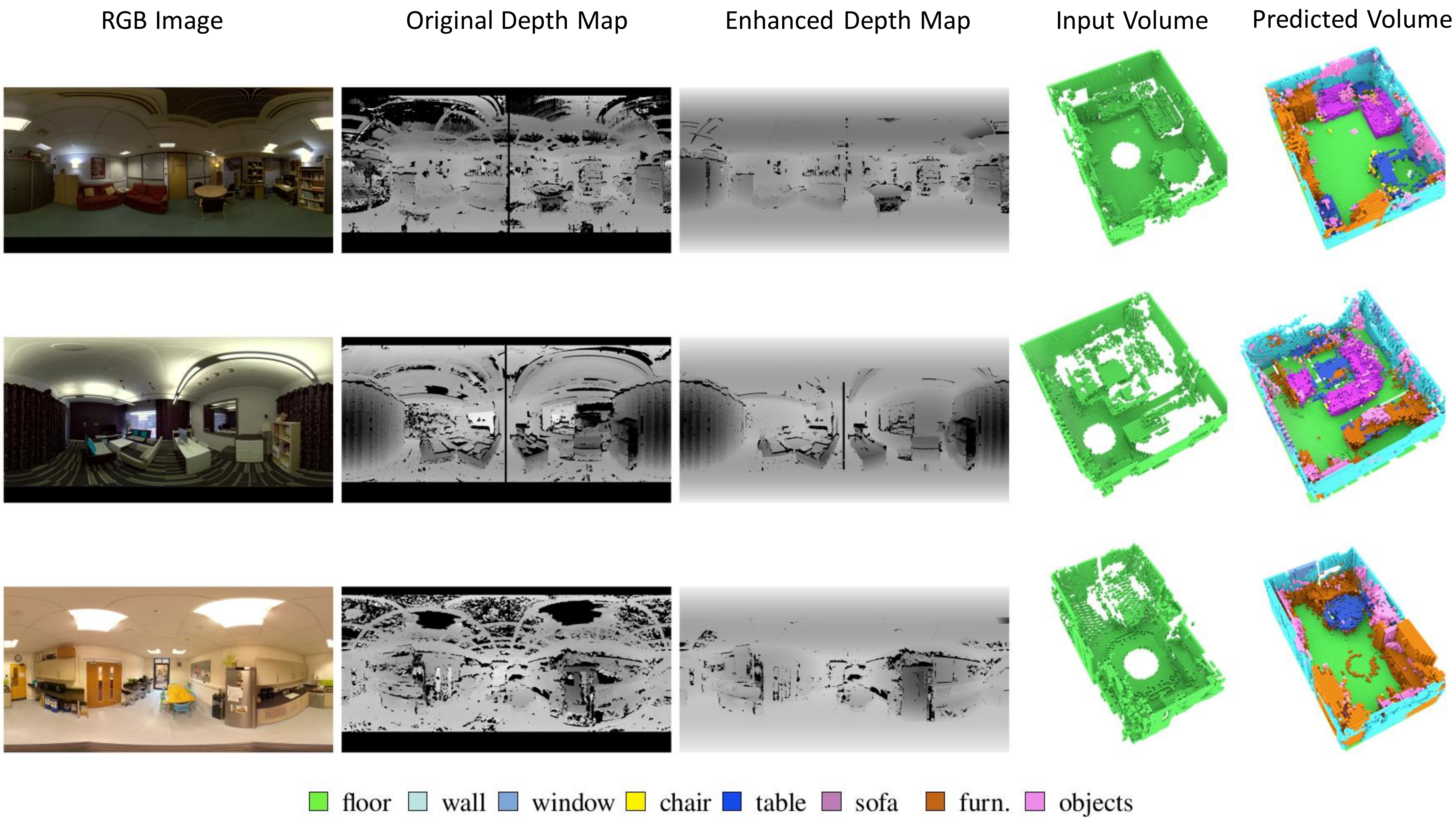


Our approach

- Depth map enhancement procedure:
  - Align the scene (Manhattan principle)
  - Apply Canny Edge Detector
  - RANSAC to fit a plane over coherent regions with similar colours



Results on Spherical Images



Chapter 8 Summary

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

Publication 1

Sematic Scene Completion from a Single 360° 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-

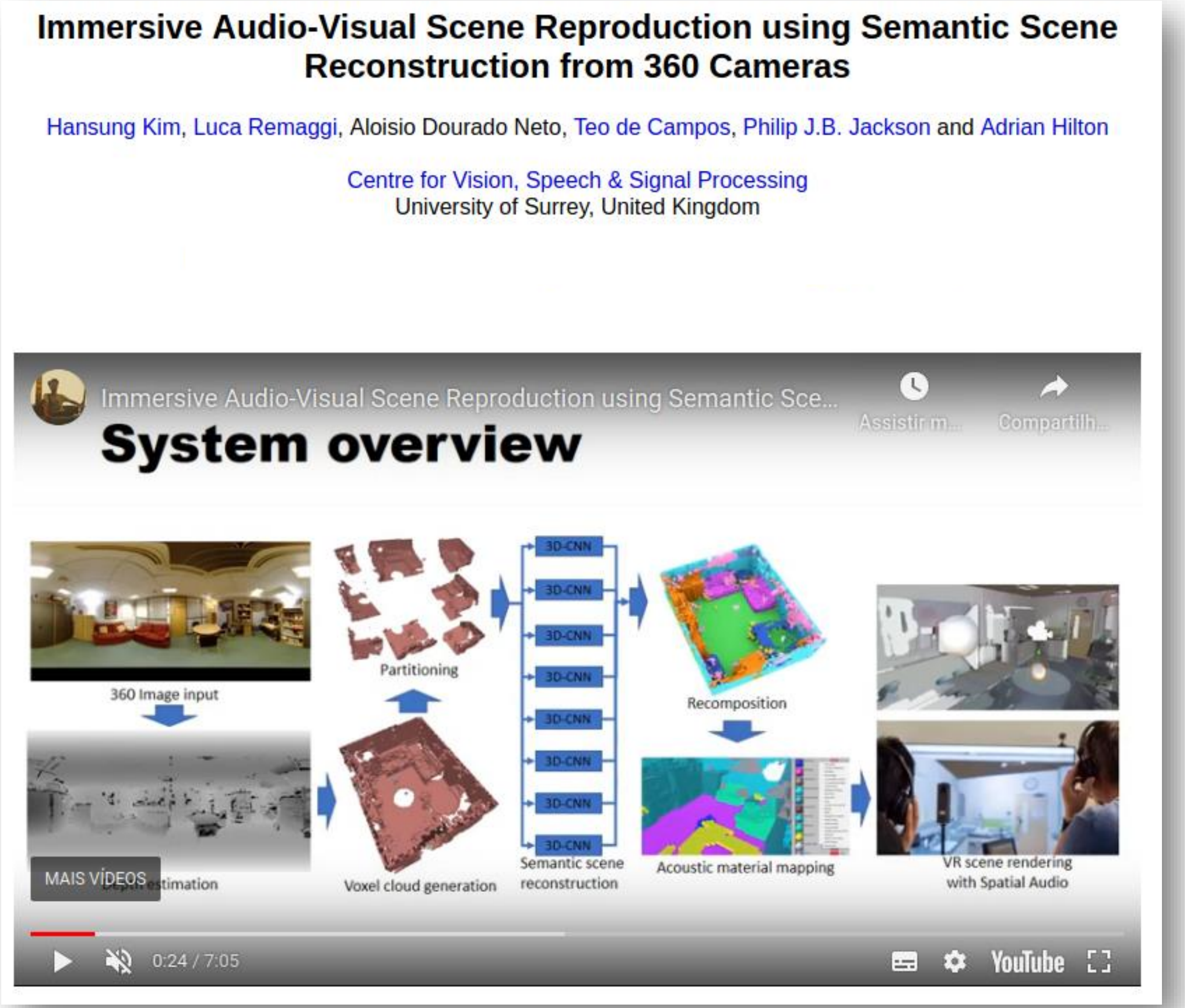
[a] https://orcid.org/0000-0002-5037-7178
[b] https://orcid.org/0000-0003-4907-0491
[c] https://orcid.org/0000-0001-6172-0229
[d] https://orcid.org/0000-0003-4223-238X

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

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

Publication 2: *Application Paper*

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



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

Chapter 9

Conclusion

Research Objectives Achievement

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

2D Domain

- to assess the benefits of domain adaptation, semi-supervision and data augmentation in the 2D semantic segmenation context

3D Domain

- to apply current trends on 2D deep CNN training protocols to 3D SSC
- to propose and evaluate new SSC models that fully exploits the information in the RGB-D images
- to propose and evaluate the benefits of semi-supervised learning

360º 3D

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

Contributions

1. A new Domain Adaptation strategy for skin detection;
2. EdgeNet, a new end-to-end CNN architecture that fuses depth and RGB edges;
3. a new 3D volumetric edge representation using F-TSDF;
4. a more efficient end-to-end training pipeline for SSC;
5. SPAwN, a novel lightweight multimodal 3D SSC CNN;
6. BN-DDR, a memory-saving batch-normalized building block for 3D CNNs;
7. a novel strategy to apply data augmentation technique for 3D SSC;
8. S3P, a novel 2D-prior-based semi-supervised training approach to the SSC task.

Publications

4 high level conferences
1 Journal

1. **Domain Adaptation for Holistic Skin Detection:** proceedings of the 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI 2021);
2. **EdgeNet: Semantic Scene Completion from RGB-D images:** proceedings of the International Conference on Pattern Recognition (ICPR 2020);
3. **Data Augmented 3D Semantic Scene Completion With 2D Segmentation:** proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022)
4. **Semantic Scene Completion from a Single 360◦ Image and Depth Map:** proceedings of the Conference on Computer Vision Theory and Applications (VISAPP 2020);
5. **Immersive audio-visual scene reproduction using semantic scene reconstruction from 360 cameras:** Virtual Reality Journal (VIRE).

Future Work

1. Combining chapter 6 and 7: data augmentation and semi-supervision combined into a single model;

2. extending S3P to explore large-scale real 3D datasets without dense 3D labels, but with 2D labels;

3. the resulting model could be used to replace EdgeNet as base model for the 360 degree SSC approach.

Thank you!