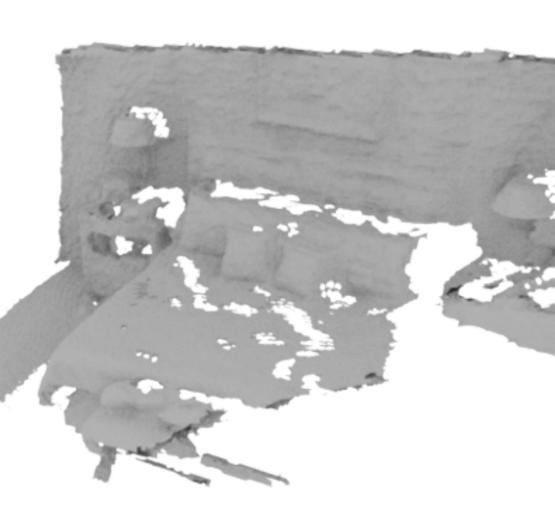


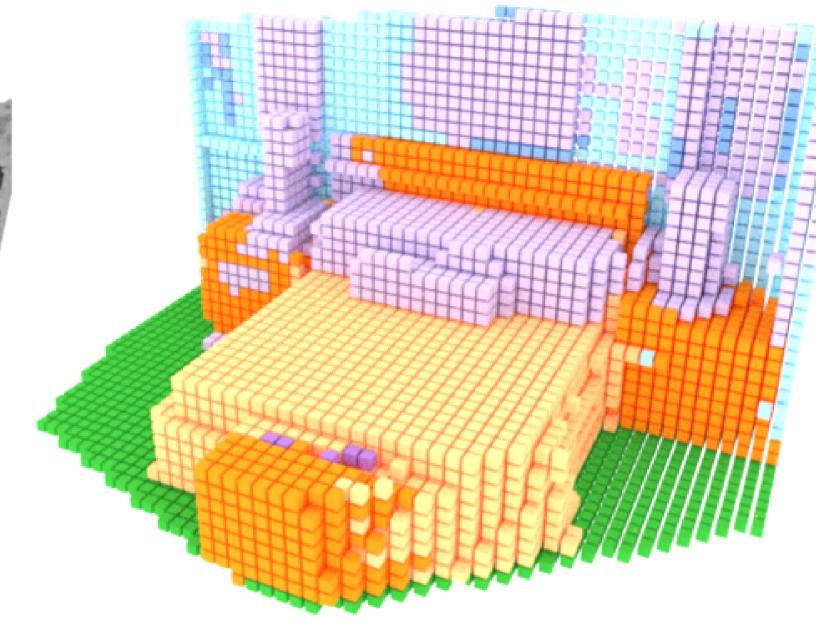
# VET NOV TAM TYM PRINCETON VISI N & ROBOTICS



## **New Task: Semantic Scene Completion**







Input: single depthmap

Output: occupancy + semantic

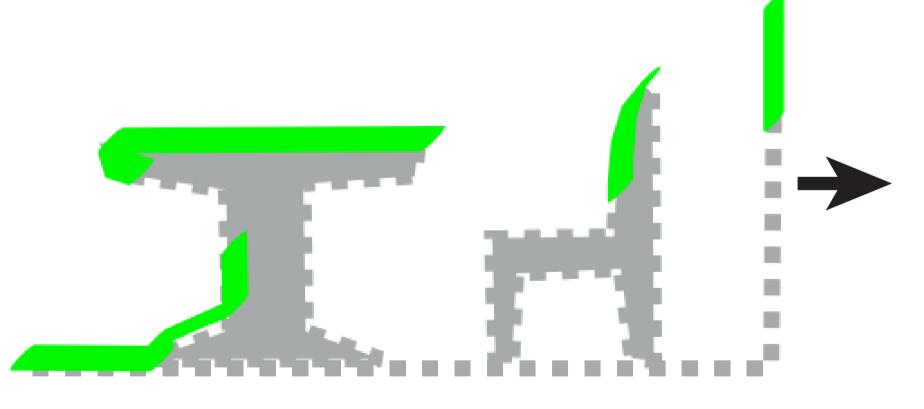
This paper focuses on semantic scene completion, a task for producing a complete 3D voxel representation of volumetric occupancy and semantic labels for a scene from a single-view depth map observation.

#### **Prior work**

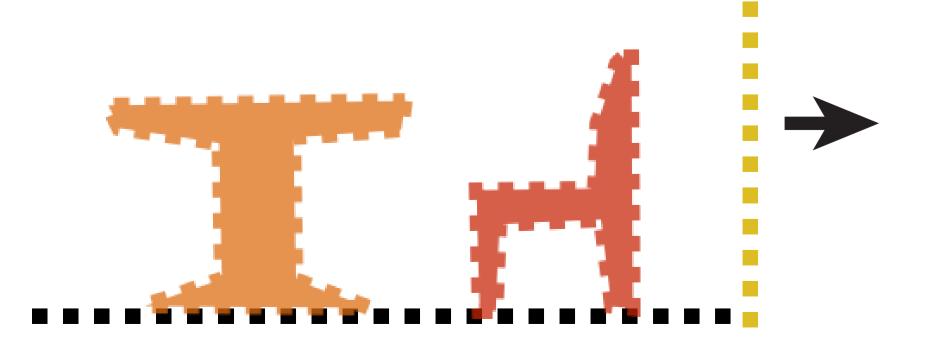
#### Surface Semantic labeling



#### Shape completion

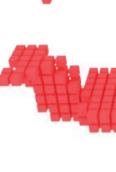


#### Semantic + Completion

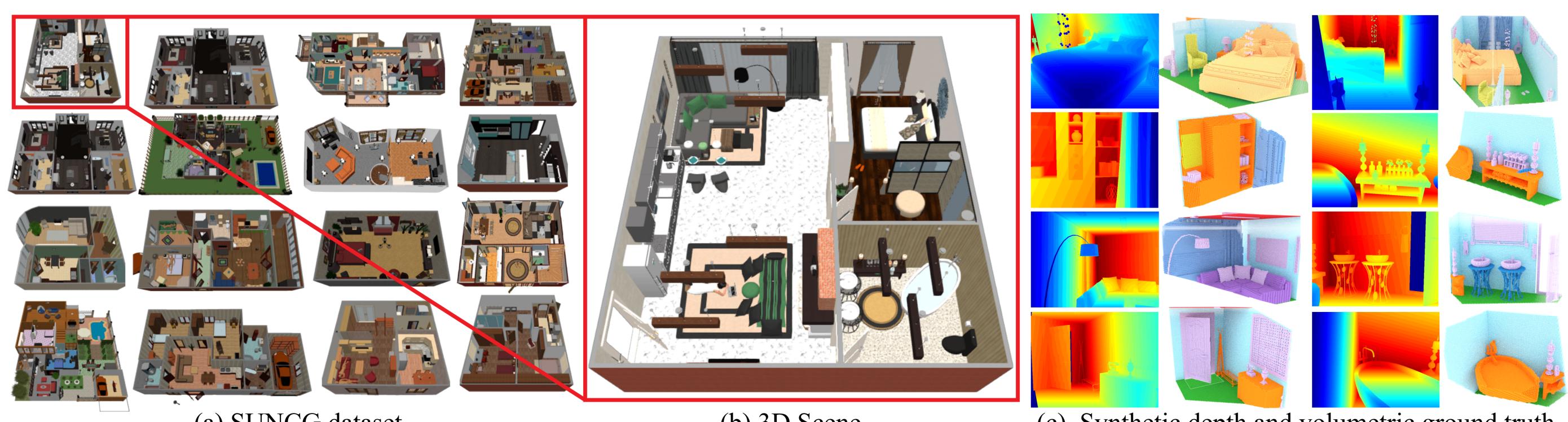


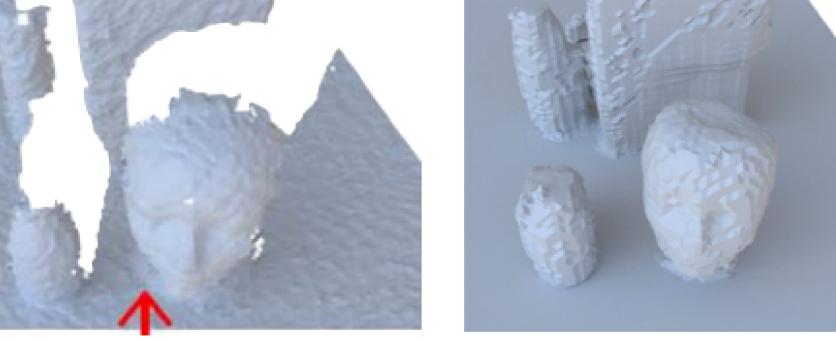
[Silberman et al.]











[Firman et al.]

# This Paper

Previous work has considered scene completion and semantic labeling of depth maps separately. However, we observe that these two problems are tightly intertwined, and therefore should be addressed jointly.

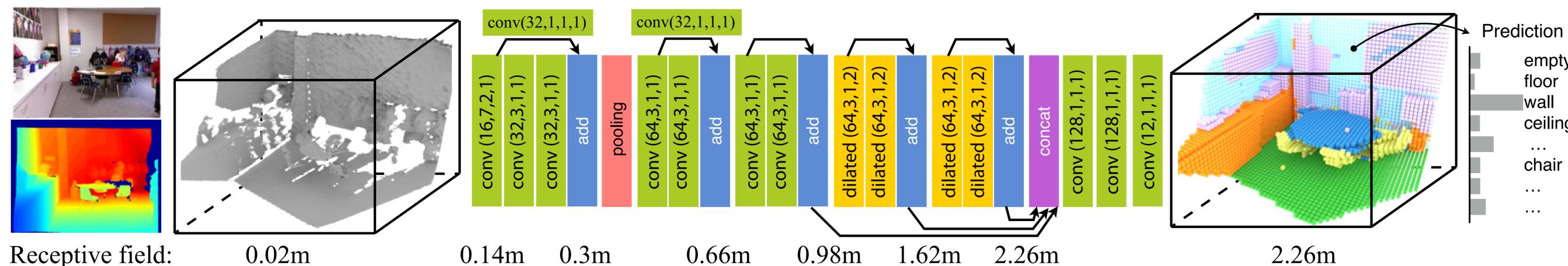
# **Semantic Scene Completion from a Single Depth Image**

Shuran Song

Fisher Yu

Andy Zeng

### **Semantic Scene Completion Network**



SSCNet is 3D convolutional network. Taking a single depth map as input, the network predicts occupancy and object labels for each voxel in the view. The parameters are shown as (#filters, kernel size, stride, dilation).

#### **Receptive field**

(a) Object centric networks (b) 3DMatch (c) Deep Sliding Shape RF:  $30 \times 30 \times 30$  voxels per object RF: 0.3m×0.3m×0.3m RF: 1m×1m×1m Voxel size: no physical meaning Voxel size: 0.01m Voxel size: 0.025m

(a) Object centric networks scale objects into the same 3D voxel grid thus discarding physical size information. In (b)-(d), colored regions indicate the effective receptive field of a single neuron in the last layer of each 3D ConvNet. With the help of 3D dilated convolution SSCNet drastically increases its receptive field compared to other 3D ConvNet architectures thus capturing richer 3D contextual information.

#### **SUNCG Dataset: over 40K houses**

(a) SUNCG dataset

(b) 3D Scene

(c) Synthetic depth and volumetric ground truth

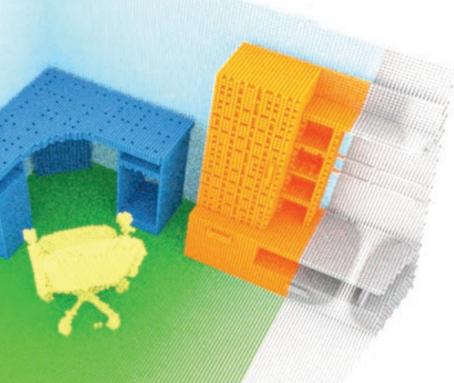
We collected a large-scale synthetic 3D scene dataset (SUNCG) to train our network. containing more than 40K manually created indoor environments. All these scenes are composed of individually labeled 3D objects, allowing us to compute full volumetric ground truth labels (suncg.cs.princeton.edu).

# Angel X. Chang

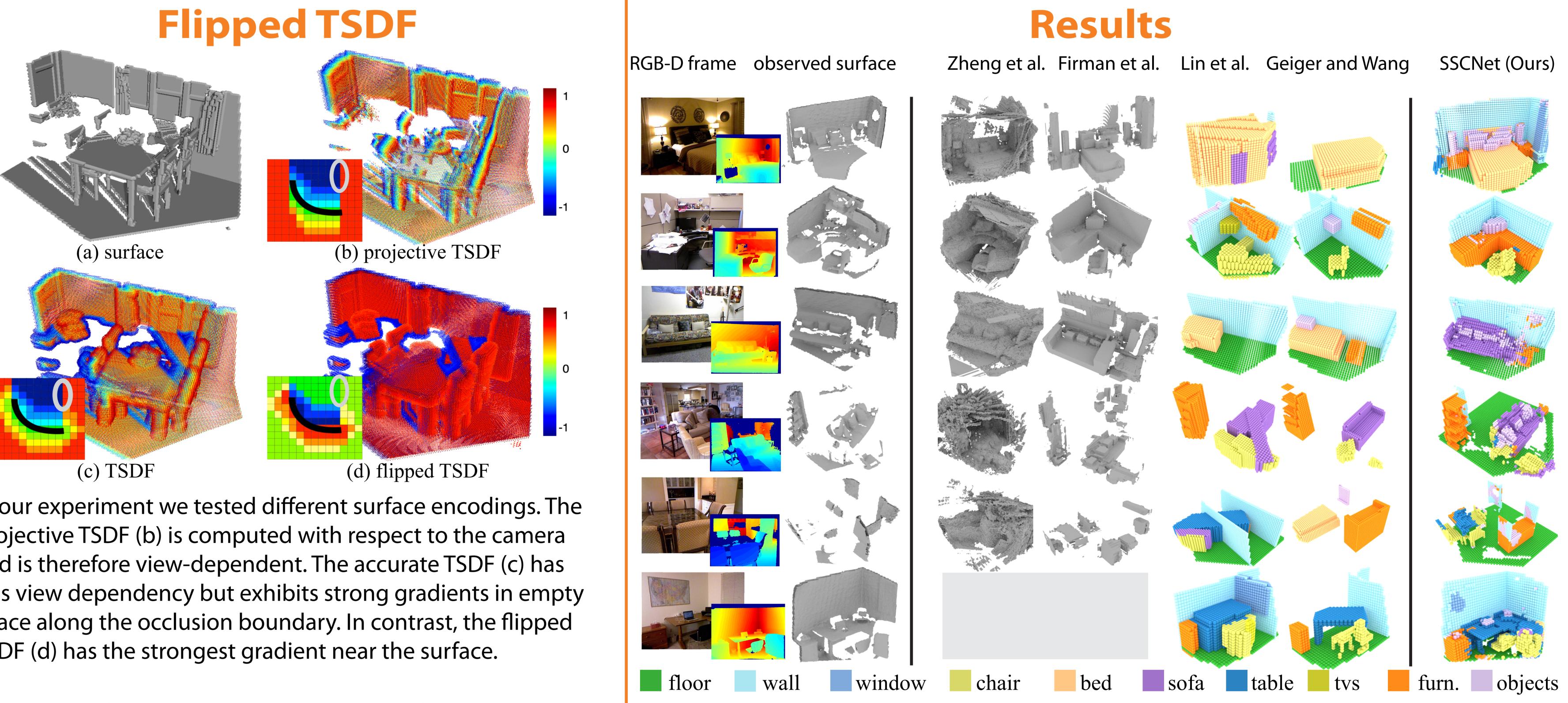
#### Manolis Savva

## Thomas Funkhouser

#### (d) SSCNet [Ours]

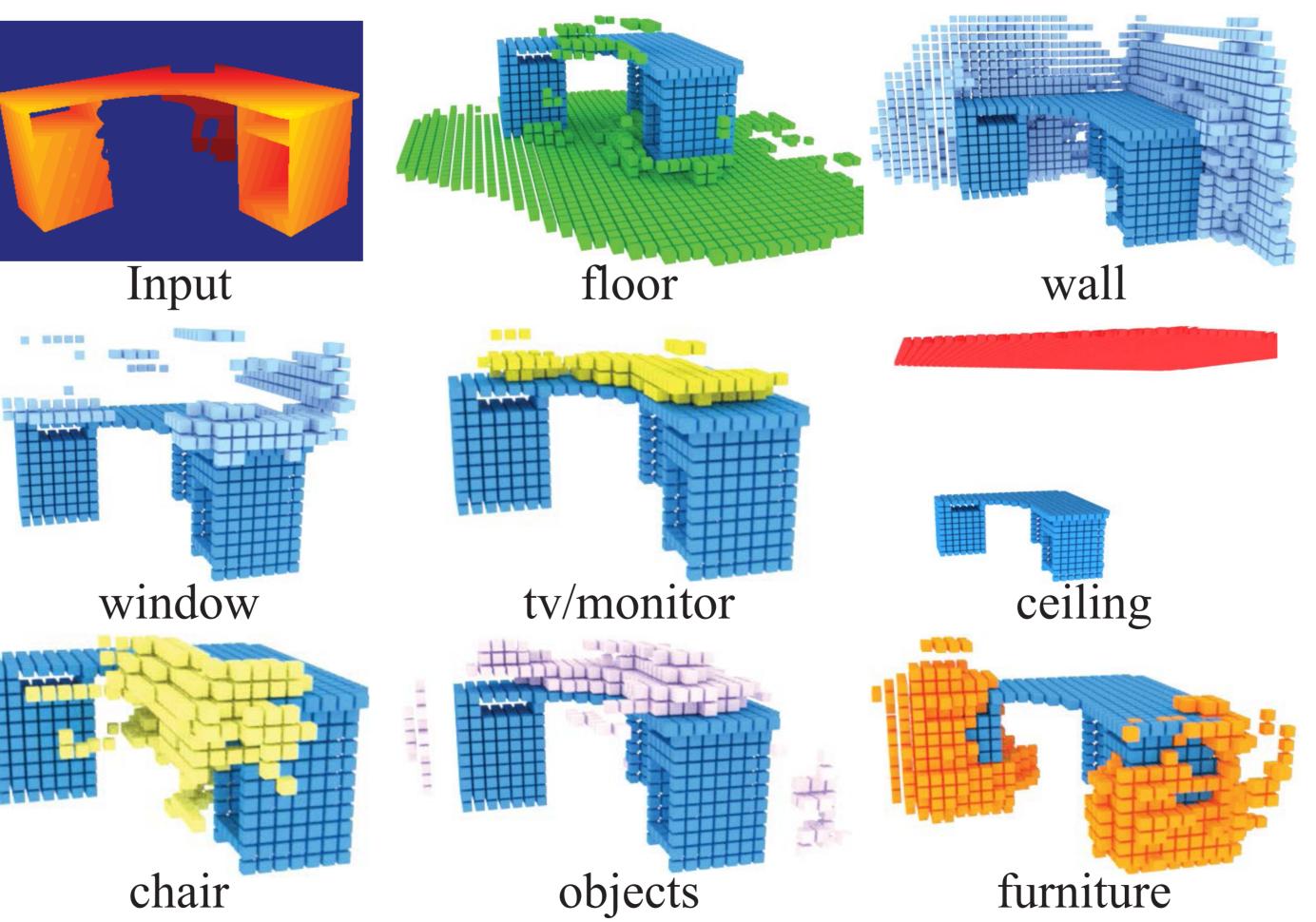


RF: 2.26m×2.26m×2.26m Voxel size: 0.02m



In our experiment we tested different surface encodings. The projective TSDF (b) is computed with respect to the camera and is therefore view-dependent. The accurate TSDF (c) has less view dependency but exhibits strong gradients in empty space along the occlusion boundary. In contrast, the flipped TSDF (d) has the strongest gradient near the surface.

## What 3D context does the network learn?



The figure shows the input depth map (a desk) and the following figures show the predictions for other objects. Without observing any information for other objects the SSCNet is able to hallucinate their locations based on the observed object and the learned 3D context.

#### Code & Data sscnet.cs.princeton.edu

#### Experiments

#### Is solving the tasks jointly better?

task/molel	scene completion w/o semantics	semantic labeling w/o completion
completion only	64.8	-
semantic only		51.2
joint	73.0	54.2

We evaluate two individual tasks, which is completion (without semantic) and surface labeling (without completion). And we compare the model train on single task and joint task. The result shows even when we evaluate the individual task the joint model outperforms the model trained on single task.

#### **Does a bigger receptive field help?**

#### **Does synthetic data help?**

model/task	Basic	Basic + Dilation
semantic scene completion	38.0	44.3

Test on NYU	NYU	SUNCG+NYU
semantic scene completion	24.7	30.5