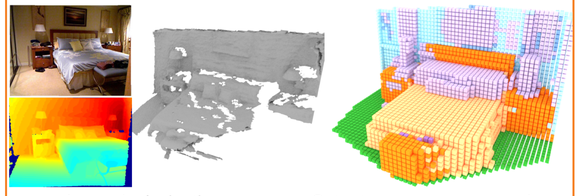
PRINCETON VISION & ROBOTICS

Semantic Scene Completion from a Single Depth Image

Shuran Song Fisher Yu Andy Zeng Angel X. Chang Manolis Savva Thomas Funkhouser

Code & Data
sscnet.cs.princeton.edu

New Task: Semantic Scene Completion



Input: single depthmap Output: occupancy + semantic

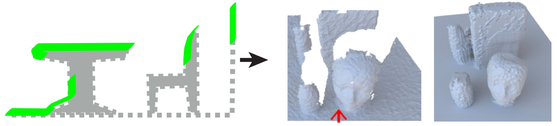
This paper focuses on semantic scene completion, a task for producing a complete 3D voxel representation of volumetric occupancy and semantic labels for a scene from a single-view depth map observation.

Prior Work

Surface Semantic labeling

[Silberman et al.]

Shape completion



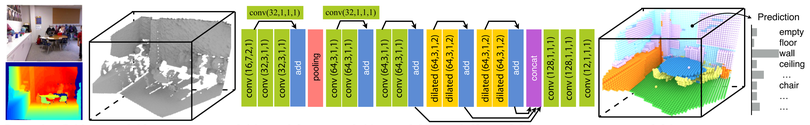
[Firman et al.]

Semantic + Completion

This Paper

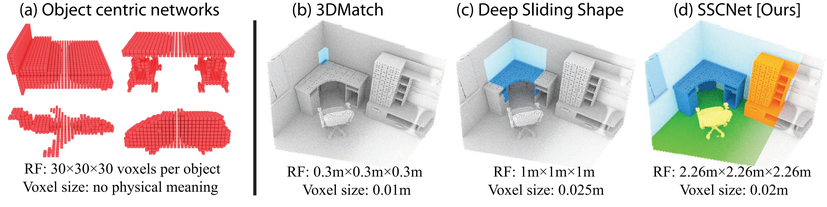
Previous work has considered scene completion and semantic labeling of depth maps separately. However, we observe that these two problems are tightly intertwined, and therefore should be addressed jointly.

Semantic Scene Completion Network



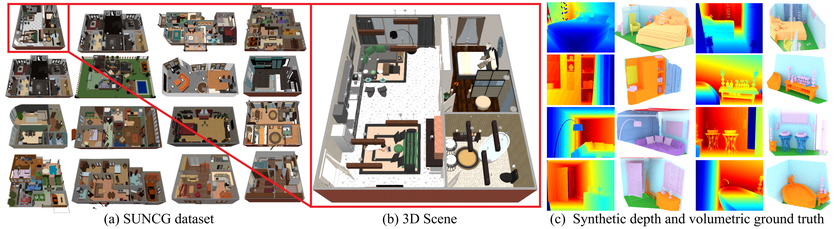
SSCNet is 3D convolutional network. Taking a single depth map as input, the network predicts occupancy and object labels for each voxel in the view. The parameters are shown as (#filters, kernel size, stride, dilation).

Receptive Field



(a) Object centric networks 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.

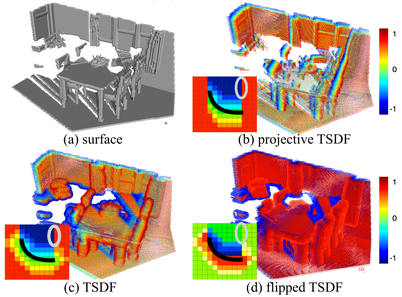
SUNCG Dataset: Over 40K Houses



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

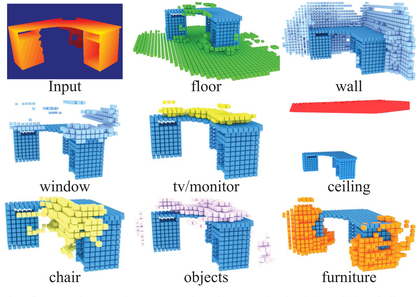
We collected a large-scale synthetic 3D scene dataset (SUNCG) to train our network. containing more than 40K manually created indoor environments. All these scenes are composed of individually labeled 3D objects, allowing us to compute full volumetric ground truth labels (suncg.cs.princeton.edu) .

Flipped TSDF



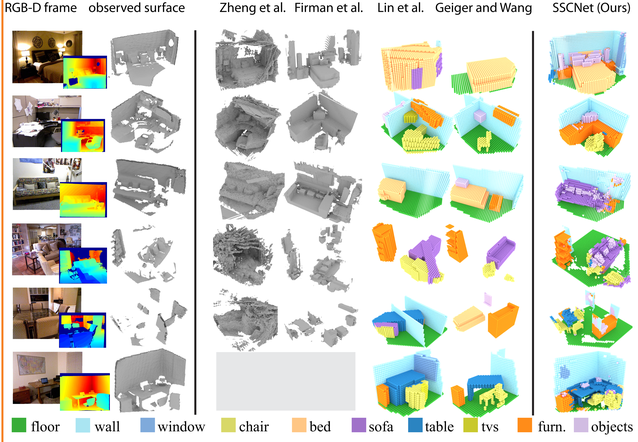
In our experiment we tested different surface encodings. The projective TSDF (b) is computed with respect to the camera and is therefore view-dependent. The accurate TSDF (c) has less view dependency but exhibits strong gradients in empty space along the occlusion boundary. In contrast, the flipped TSDF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

Results



Experiments

Is solving the tasks jointly better?

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

We evaluate two individual tasks: completion, surface labeling . And compare the model train on single task and joint task. The result shows even when evaluate the individual task the joint model outperforms the model trained on single task.

Does synthetic data help?

Test on NYU	NYU	SUNCG	SUNCG +NYU
semantic scene	24.7	20.2	30.5

Does a bigger receptive field help?

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

Comparisons

semantic scene completion	
method	mean IOU
Lin et al	12.0
Geiger and Wang	19.6
SSCNet (ours)	30.5

scene completion	
method	mean IOU
Zheng et al.	34.6
Firman et al	50.8
SSCNet (ours)	73.0