3D Scene Reconstruction with Multi-layer Depth and Epipolar Feature Transformers

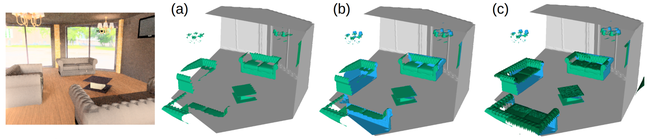
Daeyun Shin[1], Zhile Ren[2], Erik Sudderth[1], Charless Fowlkes[1]

[1]University of California, Irvine [2]Georgia Institute of Technology

Code, video, PDF:
https://research.dshin.org/iccv19/multi-layer-depth/

https://github.com/daeyun/single-view-3d-scene-recon

Goal: 3D **scene** reconstruction from a **single** RGB image

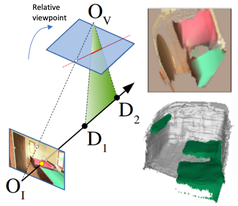


(a) Depth maps provide an efficient representation of the scene geometry but are incomplete.

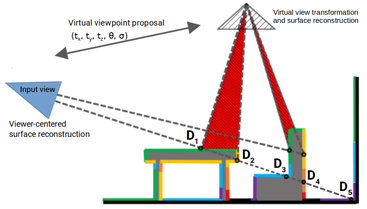
(b) We propose predicting occluded surfaces using viewer-centered, multi-layer depth maps.

(c) We introduce an epipolar transformer network that makes view-based predictions from virtual viewpoints.

Multi-layer depth and Epipolar Feature Transformers



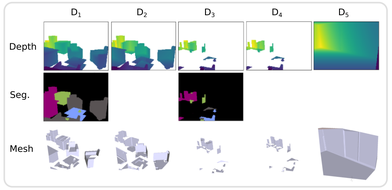
Virtual-view geometry prediction



Prediction of back surfaces allows viewer-centered **volumetric inference** in our **epipolar transformation**

$\bar{D}_1$	$\bar{D}_{1,2}$	$\bar{D}_{1,2,3}$	$\bar{D}_{1..4}$	$\bar{D}_{1..5}$	$\bar{D}_{1..5}$ + Overhead virtual view
0.237	0.427	0.450	0.480	0.924	0.932

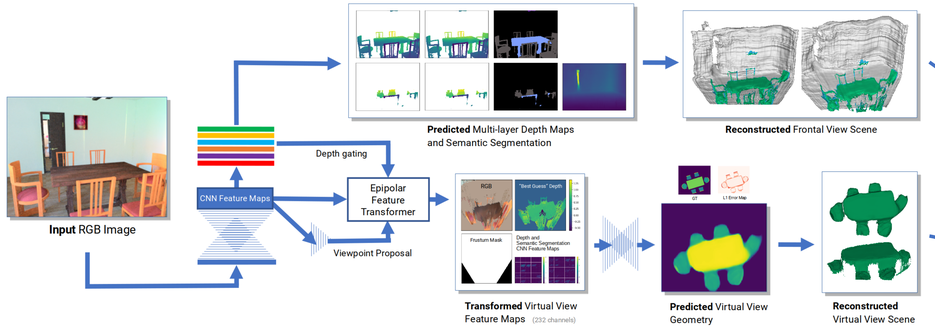
Scene surface coverage (recall) of ground truth depth layers with a 5cm threshold.
Our representation covers 93% of the scene geometry inside the viewing frustum.



To reconstruct 3D surfaces from predicted depth images, we triangulate point cloud vertices that correspond to a 2x2 neighborhood in image space.

While room envelopes (D_5) typically have a very simple shape, the prediction of occluded structure ($D_{2\text{-}4}$) behind visible object surfaces (D_1) is more challenging.

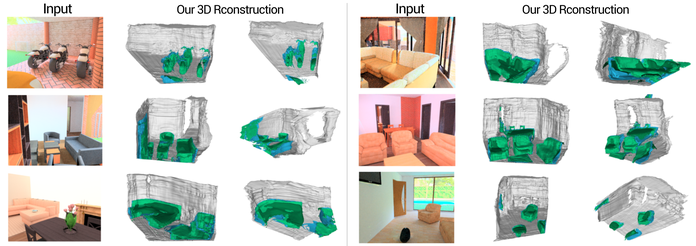
Fully convolutional depth-based 3D scene geometry prediction and completion



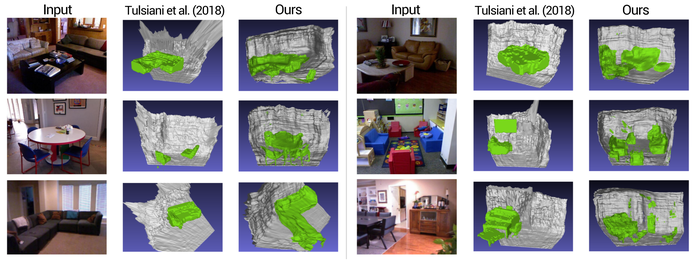
Overview of our system for reconstructing a complete 3D scene from a single RGB image. We first predict a multi-layer depth map that encodes the depths of front and back object surfaces as seen from the input camera. The epipolar feature transformer network then transfers CNN features from the input view to a virtual overhead view, where the orthographic heights of observed surfaces are predicted.

Reconstruction from synthetic and real-world images

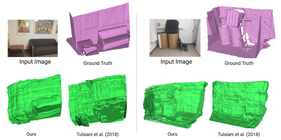
Results on **SUNCG** (PBRS renderings):



Results on **NYUv2**:



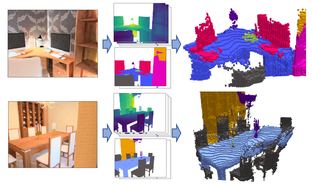
Results on **ScanNet**:



We **quantitatively** evaluate the **synthetic-to-real transfer** of 3D scene geometry on the ScanNet dataset. Our network model is trained entirely on synthetically generated images.

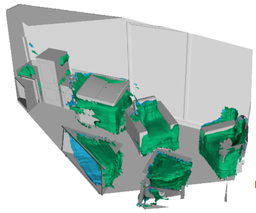
	Precision	Recall
$D_{1,2,3,4,5}$ & Overhead	**0.221**	**0.358**
Tulsiani *et al.*	0.132	0.191

Voxelization on **SUNCG**:

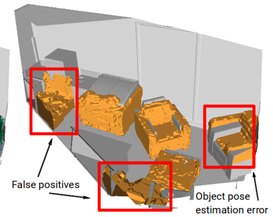


Comparison to object detection-based approach

Object-based compositional scene reconstruction is sensitive to detection and pose estimation errors.

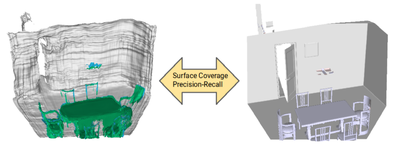


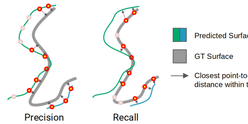
Ours: fully convolutional viewer-centered scene surface prediction



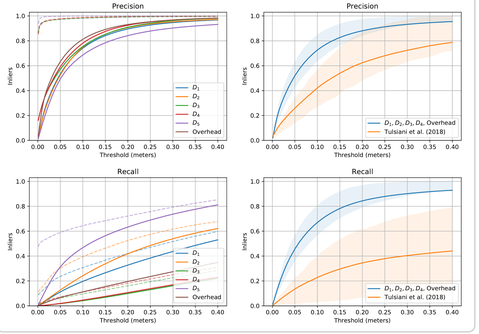
SotA: Tulsiani et al. (2018) - Object-based detection and voxel shape prediction

Surface precision-recall evaluation





Precision and recall of scene geometry as a function of match distance **threshold** on SUNCG



Left: Accuracy of different model layers evaluated on the whole scene. Dashed lines are the performance bounds provided by ground-truth depth layers.
Right: Accuracy of our model relative to the state-of-the-art, evaluated on objects only.

We uniformly sample points on surface of the ground truth mesh then compute the distance to the closest point on the predicted mesh (recall), and vice versa (precision). Sampling density is $\rho = 10000/m^2$.

We measure the percentage of inlier distances for given threshold.

	Precision	Recall
D_1	0.525	0.212
D_1 & Overhead	**0.553**	**0.275**
$D_{1,2,3,4}$	0.499	0.417
$D_{1,2,3,4}$ & Overhead	**0.519**	**0.457**

Augmenting the frontal depth prediction with the predicted virtual view height map improves both precision and recall of objects in the scene (match threshold of 5cm).