

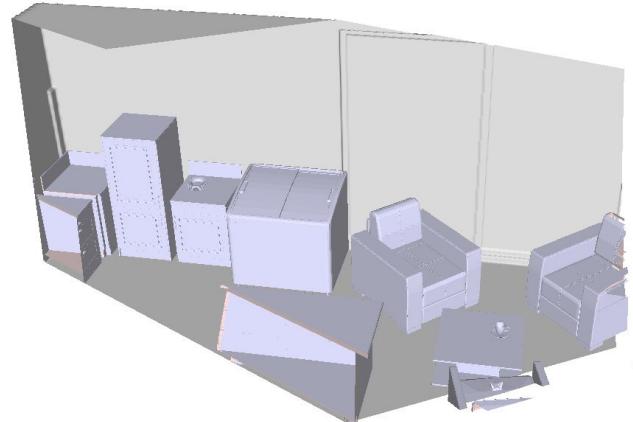
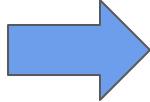
3D Scene Reconstruction with Multi-layer Depth and Epipolar Transformers

to appear, ICCV 2019

Daeyun Shin¹ Zhile Ren² Erik Sudderth¹ Charless Fowlkes¹



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

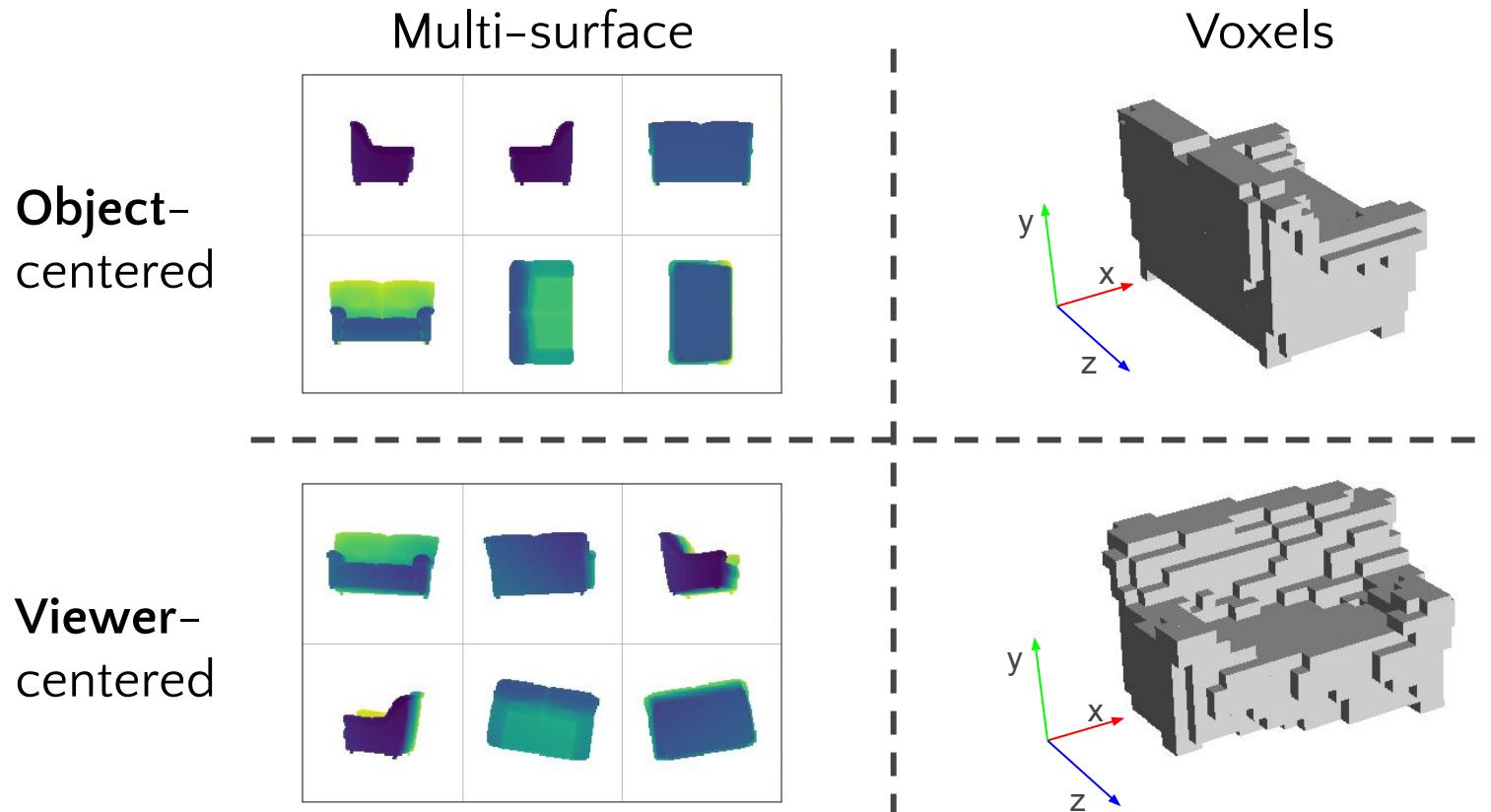


RGB Image

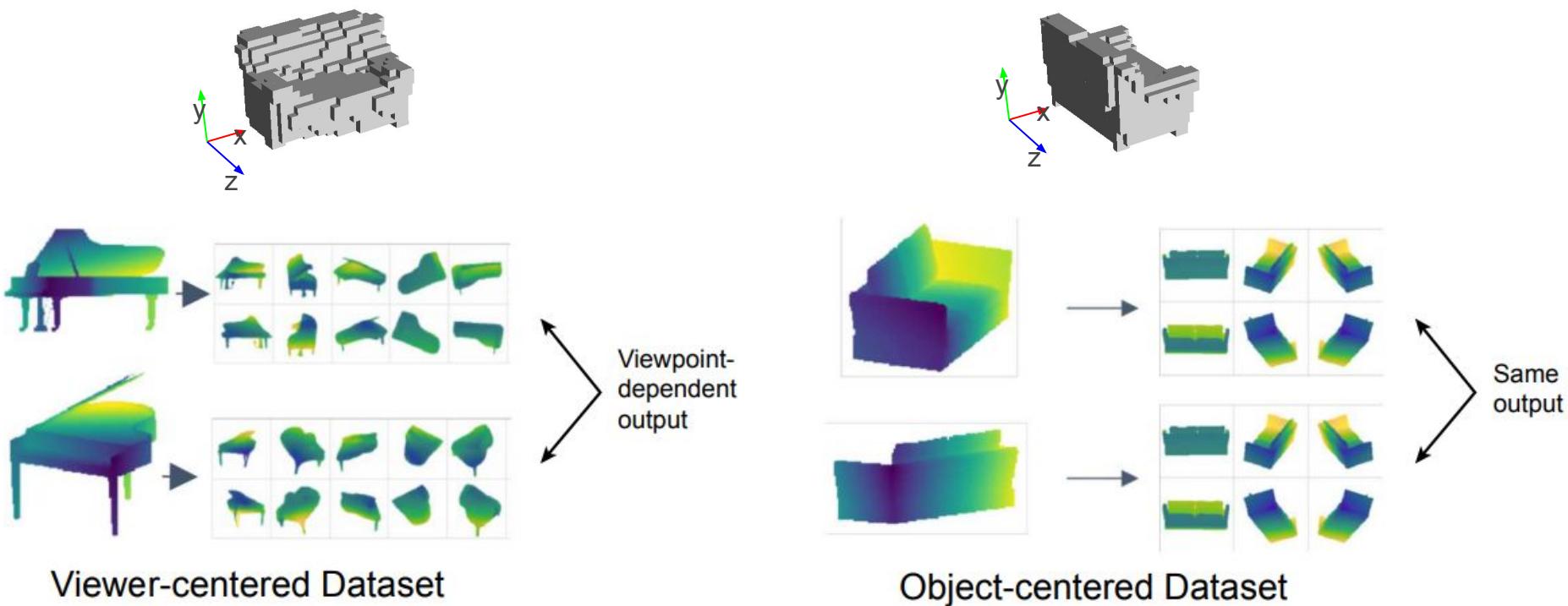
3D Scene Reconstruction
(SUNCG Ground Truth)

Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction (CVPR 18. Shin, Fowlkes, Hoiem)

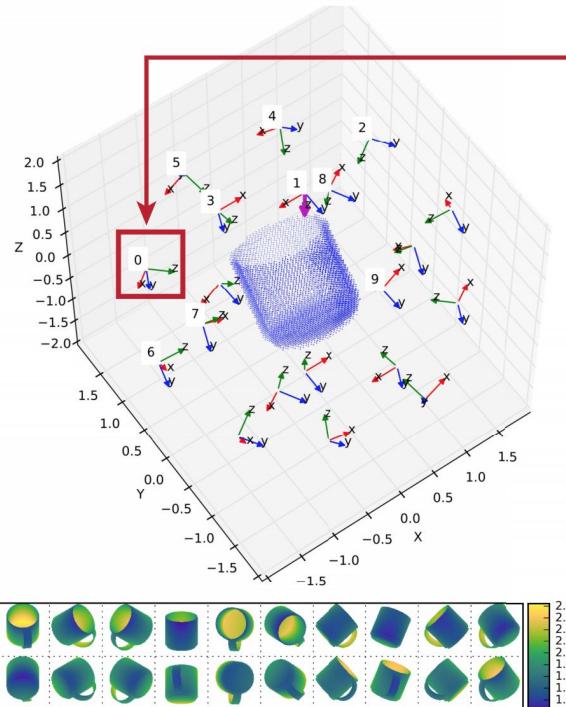
Question: What effect does shape representation have on prediction?



Coordinate system is an important part of shape representation



Synthetic training data



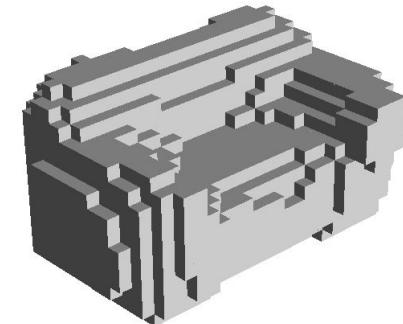
Top: RGB training images generated using RenderForCNN [3]. Our RGB dataset consists of 2.4M renderings of 34,000 3D CAD models from 12 object categories in ShapeNet.

Surfaces vs. voxels for 3D object shape prediction



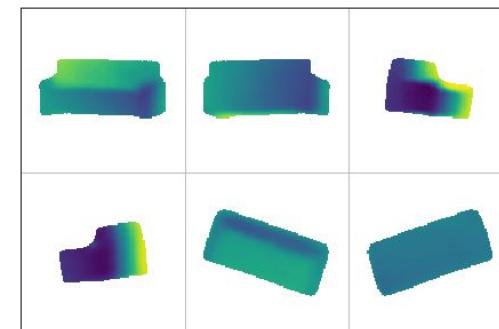
RGB Image

3D Convolution (most common approach) →



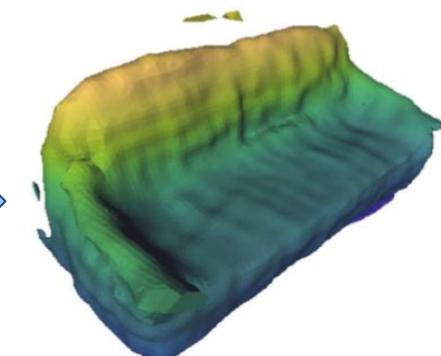
Predicted Voxels

2D Conv.



Multi-surface Prediction

3D Reconstruction →

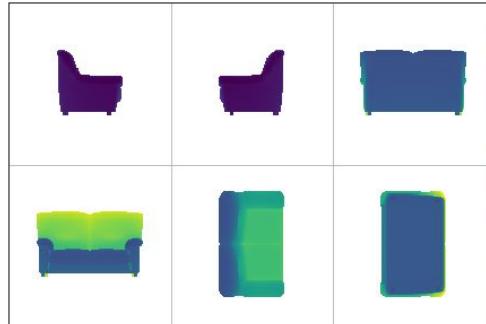


Predicted Mesh

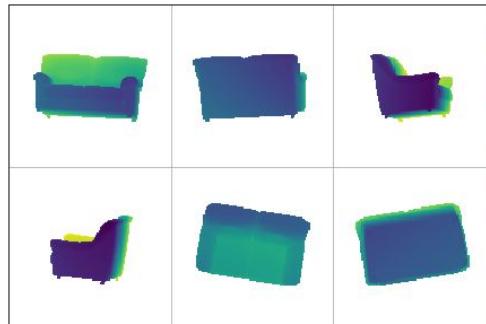
Question: What effect does shape representation have on prediction?

Object-centered

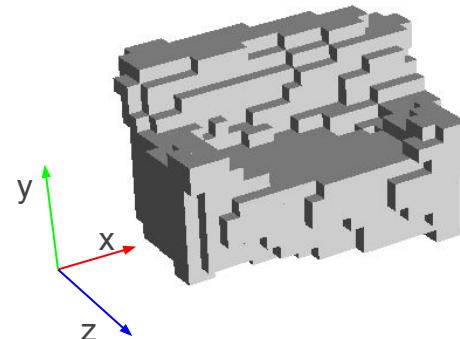
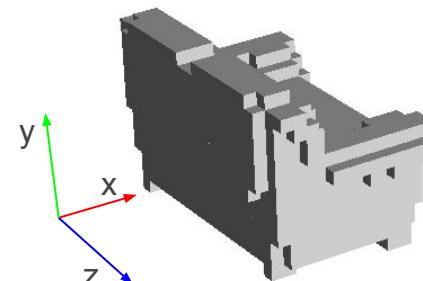
Multi-surface



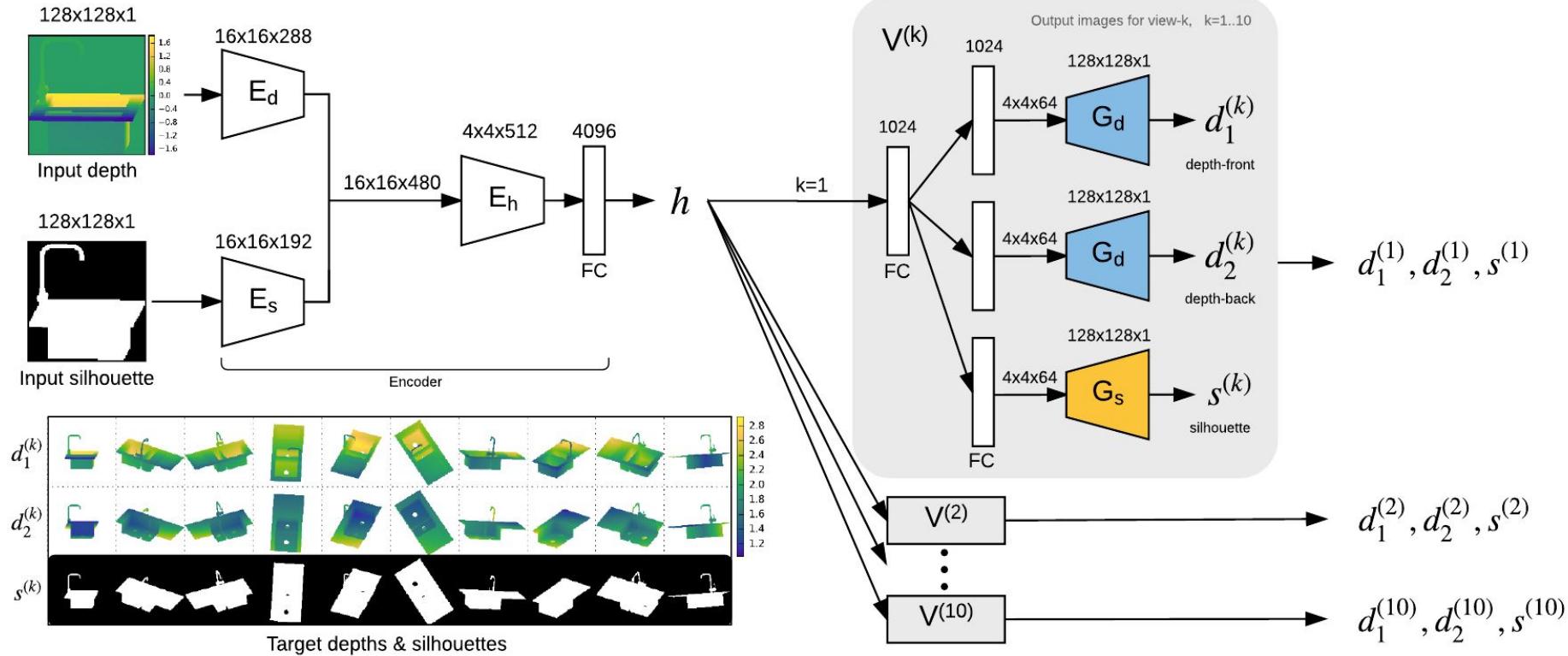
Viewer-centered



Voxels



Network architecture for surface prediction



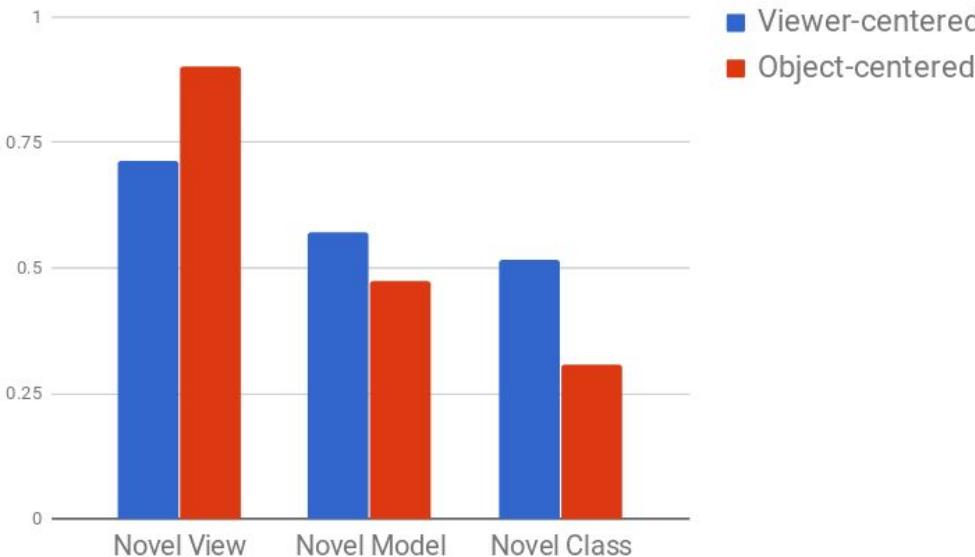
Experiments

- Three difficulty settings (how well does the prediction generalize?)
 - **Novel view**: new view of model that is in training set
 - **Novel model**: new model from a category that is in training set
 - **Novel category**: new model from a category that is not in the training set
- Evaluation metrics: Mesh surface distance, Voxel IoU, Depth L1 error
- Same procedure applied in all four cases.

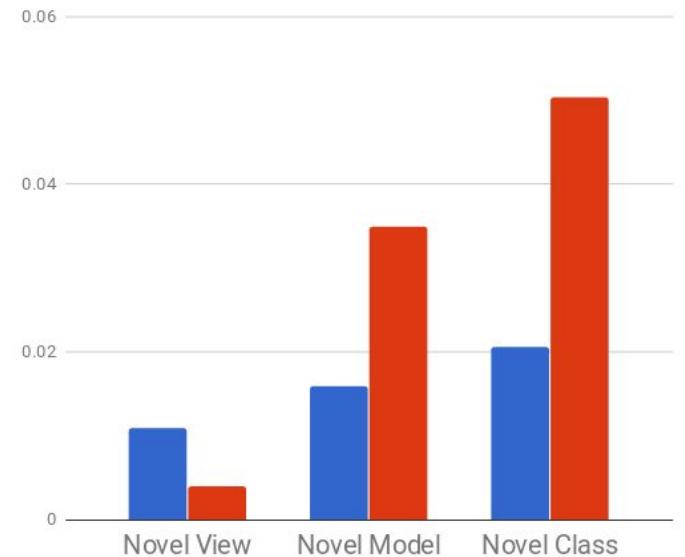
What effect does **coordinate system** have on prediction?

Viewer-centered vs. Object-centered

Voxel IoU (mean, higher is better)



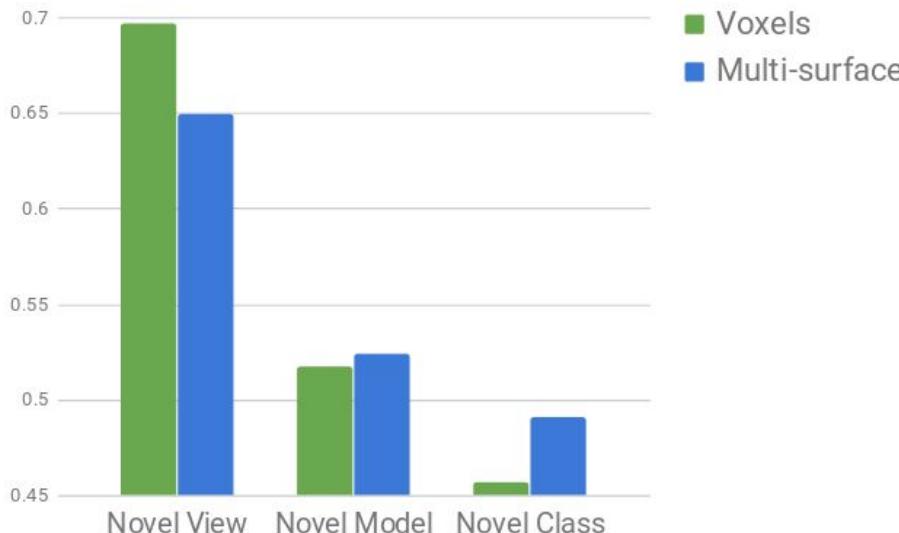
Depth error (mean, lower is better)



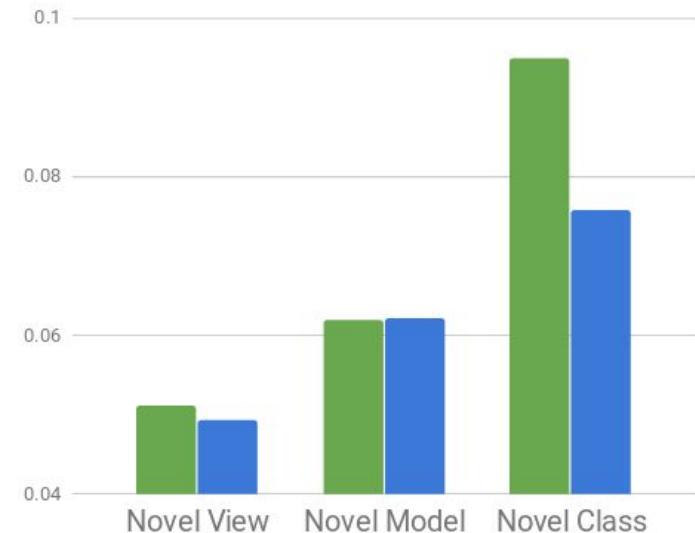
What effect does **shape representation** have on prediction?

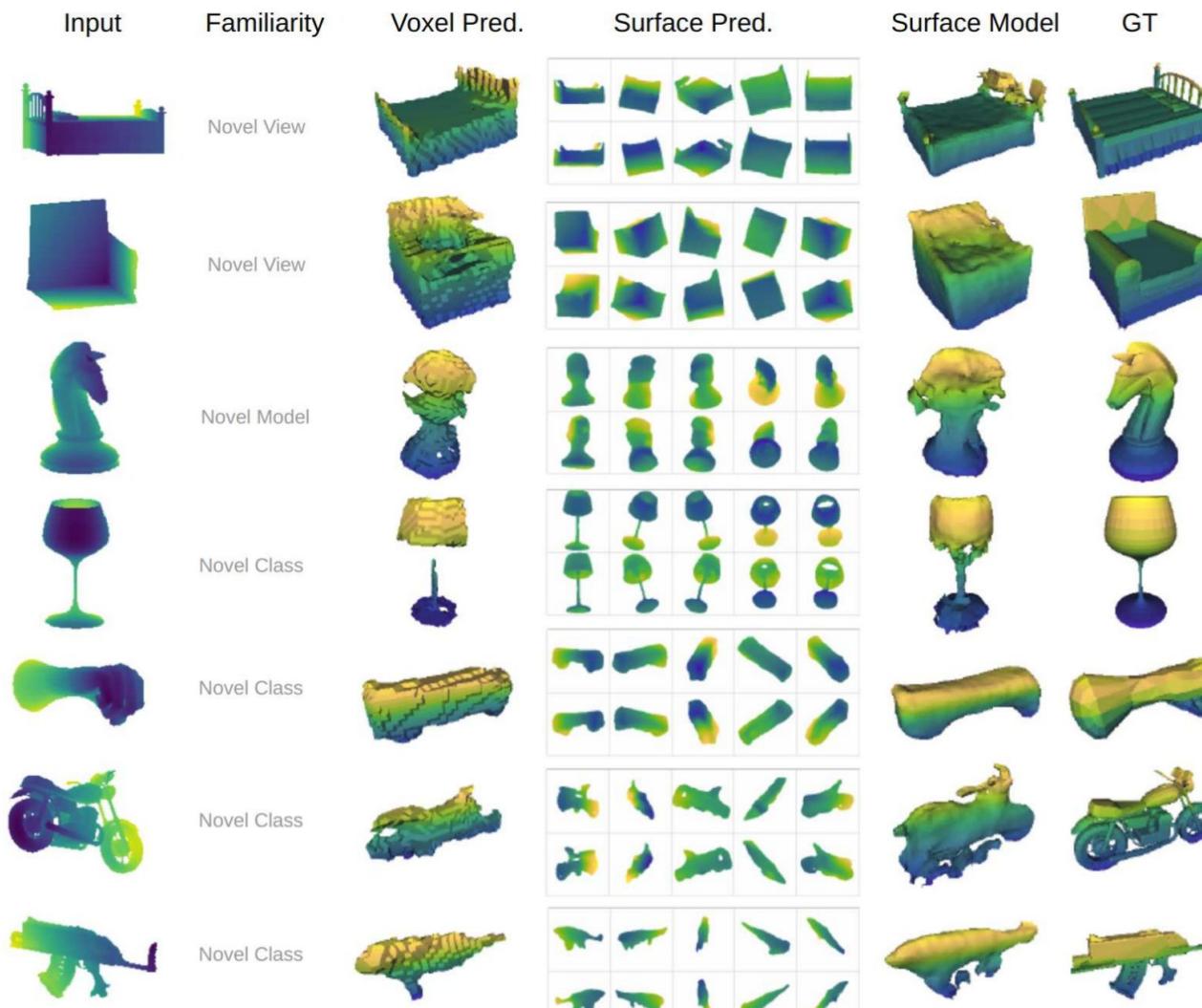
Voxels vs. multi-surface

Voxel IoU (mean, higher is better)



Surface distance (mean, lower is better)

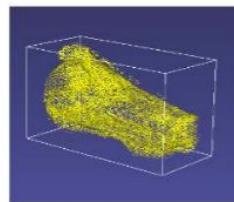
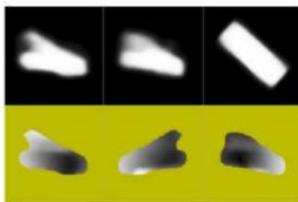




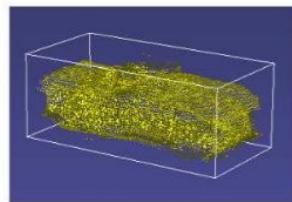
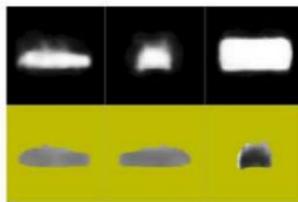
Input



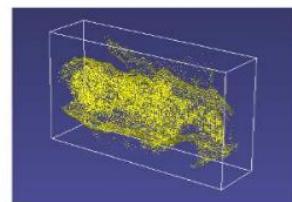
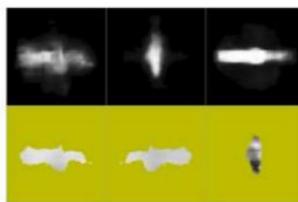
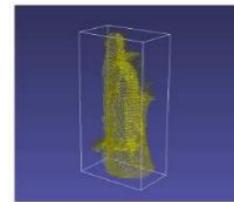
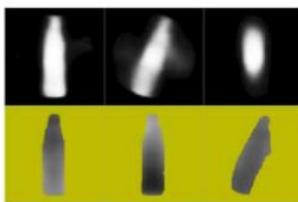
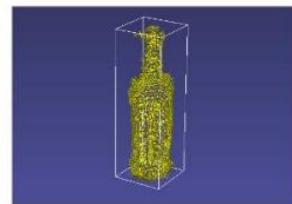
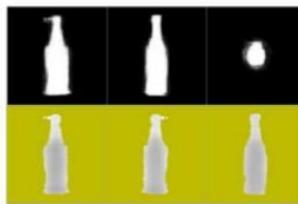
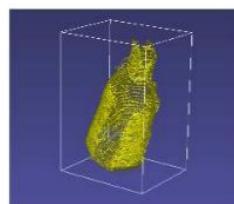
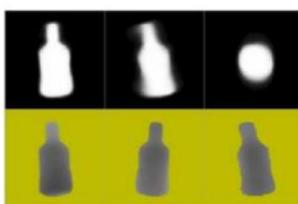
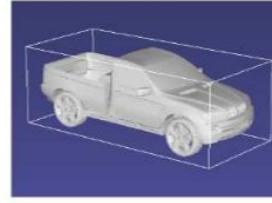
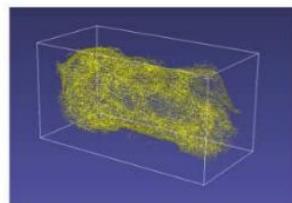
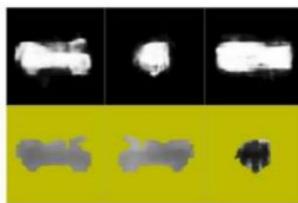
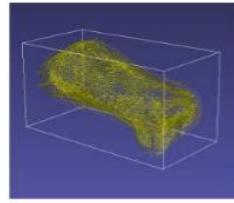
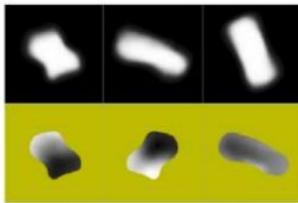
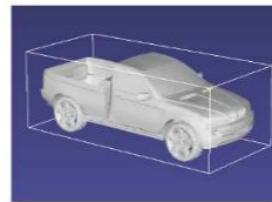
Viewer-centered



Object-centered



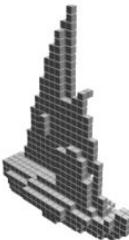
GT



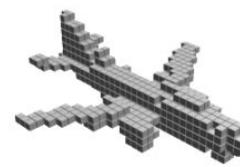
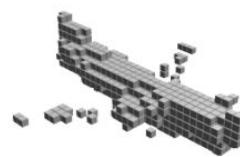
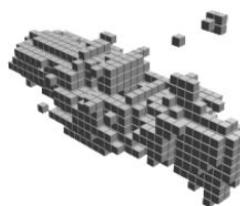
Input



GT



Object-centered prediction (3D-R2N2)



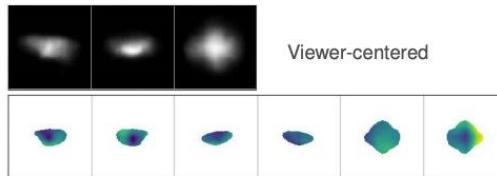
Inspiring examples from 3D-R2N2's Supplementary Material

Input

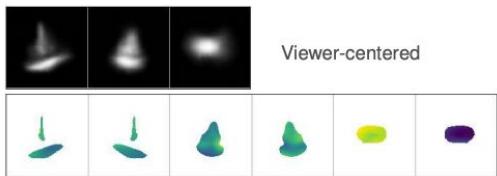


Multi-surface Pred.

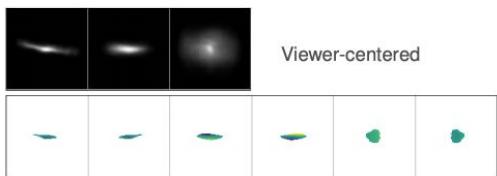
Viewer-centered



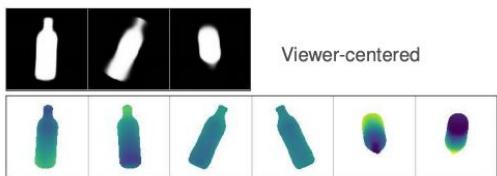
Viewer-centered



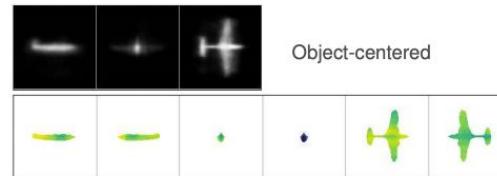
Viewer-centered



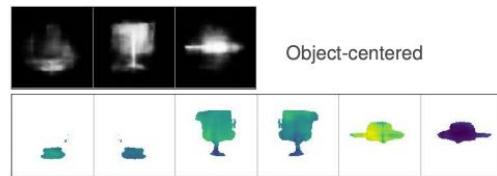
Viewer-centered



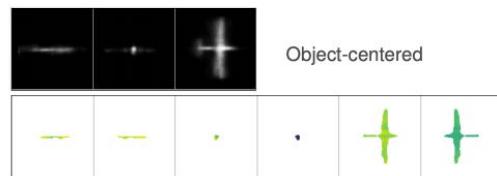
Object-centered



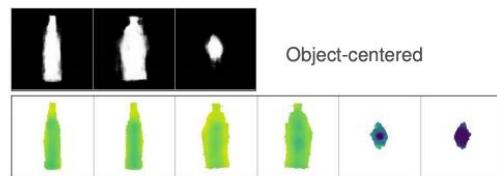
Object-centered



Object-centered



Object-centered



Shape representation is important in learning and prediction.

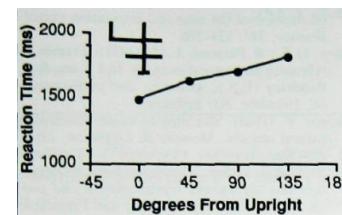
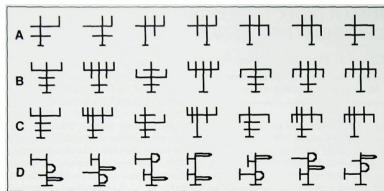
- Viewer-centered representation generalizes better to difficult input, such as, novel object categories.
- 2.5D surfaces (depth and segmentation) tend to generalize better than voxels and predicts higher fidelity shapes (thin structures)



2.5D segmentation, depth

Viewer-centered vs. Object-centered: Human vision

- Tarr and Pinker ¹: Found that human perception is largely tied to viewer-centered coordinate, in experiments on 2D symbols
- McMullen and Farah ²: Object-centered coordinates seem to play more of a role for familiar exemplars, in line drawing experiments.
- We do not claim our computational approach has any similarity to human visual processing.



[1]: M. J. Tarr and S. Pinker. *When does human object recognition use a viewer-centered reference frame?* Psychological Science, 1(4):253–256, 1990

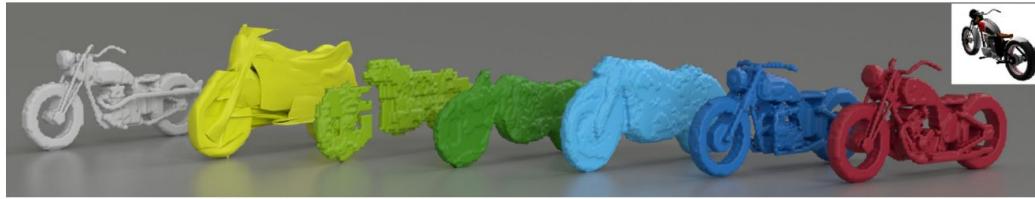
[2]: P. A. McMullen and M. J. Farah. *Viewer-centered and object-centered representations in the recognition of naturalistic line drawings.* Psychological Science, 2(4):275–278, 1991.

Follow-up work (Tatarchenko et al., CVPR 19):

What Do Single-view 3D Reconstruction Networks Learn?

Maxim Tatarchenko^{*1}, Stephan R. Richter^{*2}, René Ranftl², Zhuwen Li²,
Vladlen Koltun², and Thomas Brox¹

¹University of Freiburg ²Intel Labs



- They observe that SoA single-view 3D object reconstruction methods actually perform image classification, and retrieval performance is just as good.
- Following our CVPR 18 work, they recommend the use of viewer-centered coordinate frames.

Follow-up work (Zhang et al., NIPS 18 oral):

Learning to Reconstruct Shapes from Unseen Classes

Xiuming Zhang*
MIT CSAIL

Zhoutong Zhang*
MIT CSAIL

Chengkai Zhang
MIT CSAIL

Joshua B. Tenenbaum
MIT CSAIL

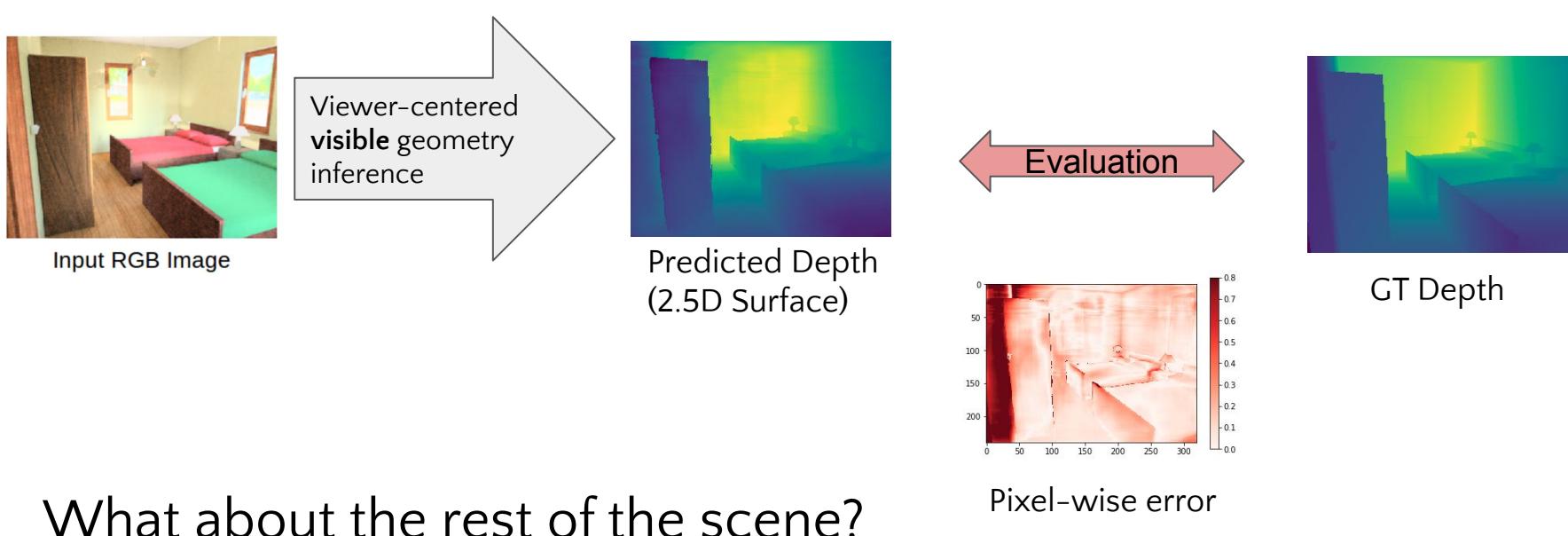
William T. Freeman
MIT CSAIL, Google Research

Jiajun Wu
MIT CSAIL

- Zhang et al. performs single-view reconstruction of objects in novel categories.
- Their viewer-centered approach achieves SoA results.
- Following our CVPR 18 work, they experiment with both object-centered and viewer-centered models and validate our findings.

How can we extend viewer-centered,
surface-based object representations
to **whole scenes**?

Background: Typical monocular depth estimation pipeline

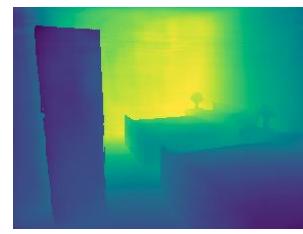


What about the rest of the scene?

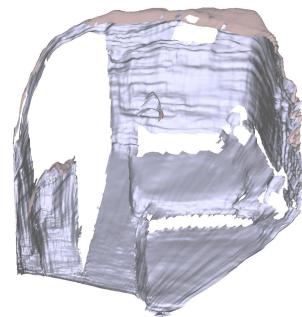
2.5D in relation to 3D



Input RGB Image

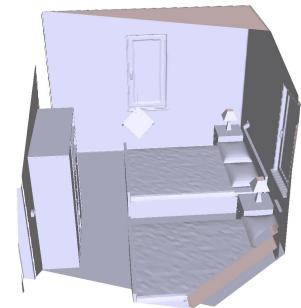


Predicted Depth



Predicted Depth as 3D mesh

Evaluation

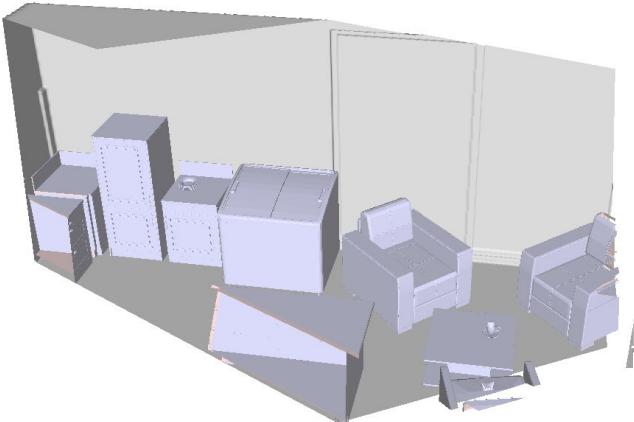


Ground Truth 3D Mesh

- 3D requires predicting **both** visible and occluded surfaces!

Multi-layer Depth

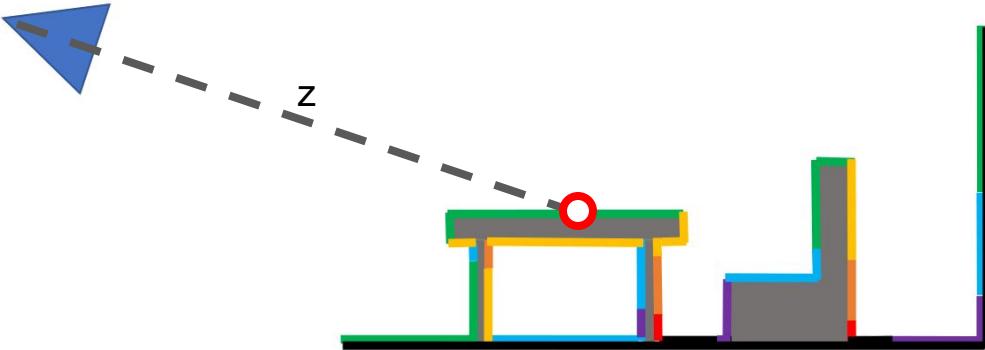
Synthetic dataset



CAD model of 3D Scene
(SUNCG Ground Truth, CVPR 17)

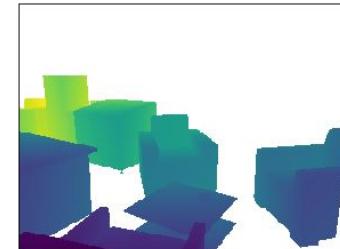
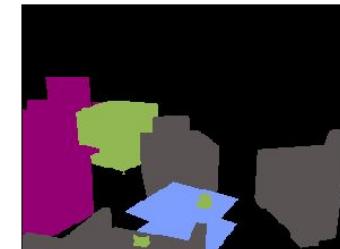


RGB Rendering
Physically-based rendering
(PBRS, CVPR 17)

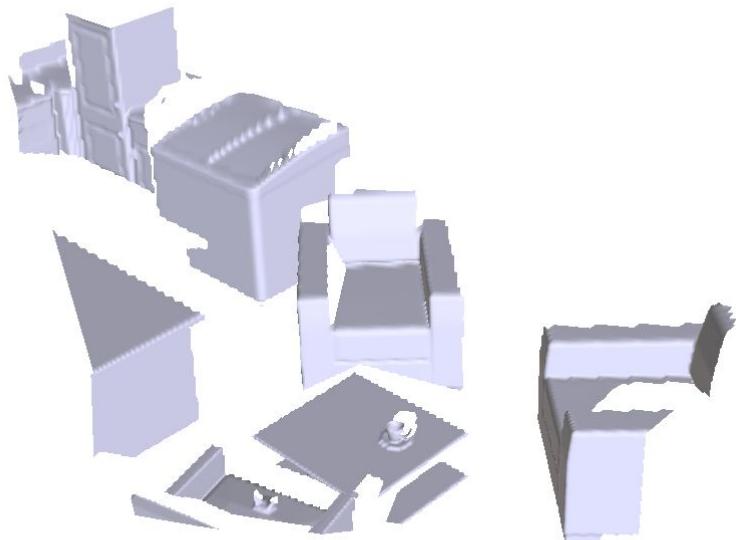


D_1

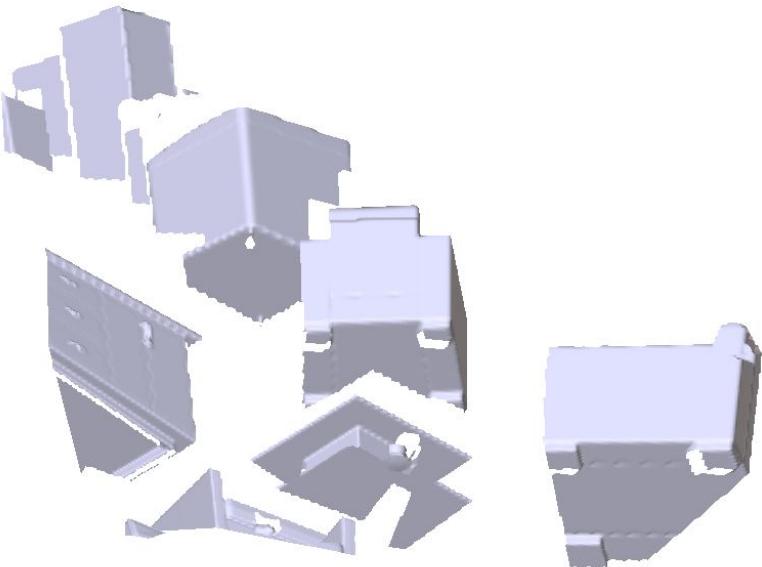
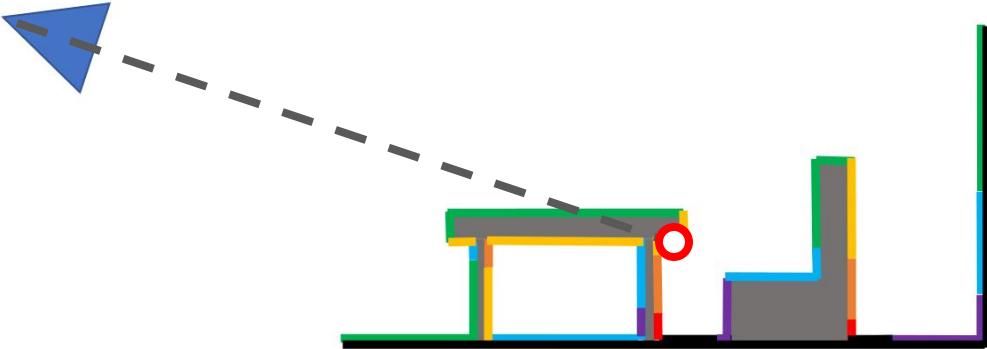
Learning Target:



Object First-hit Depth Layer

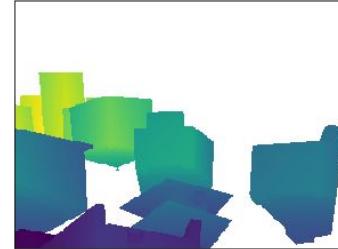


"Traditional depth image with segmentation"



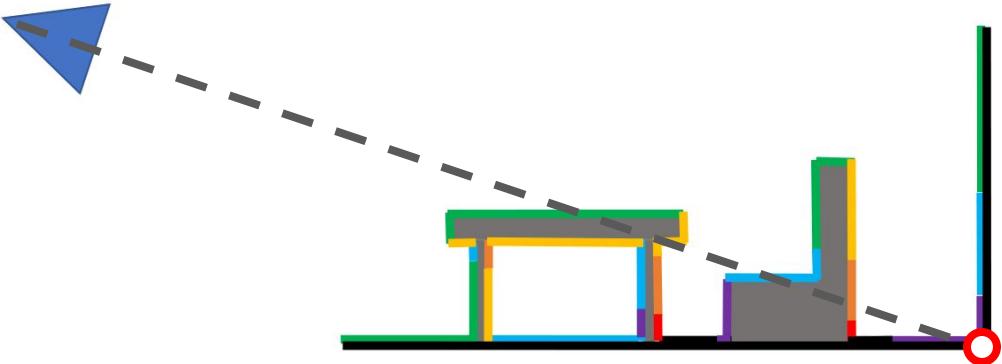
D_2

Learning Target:

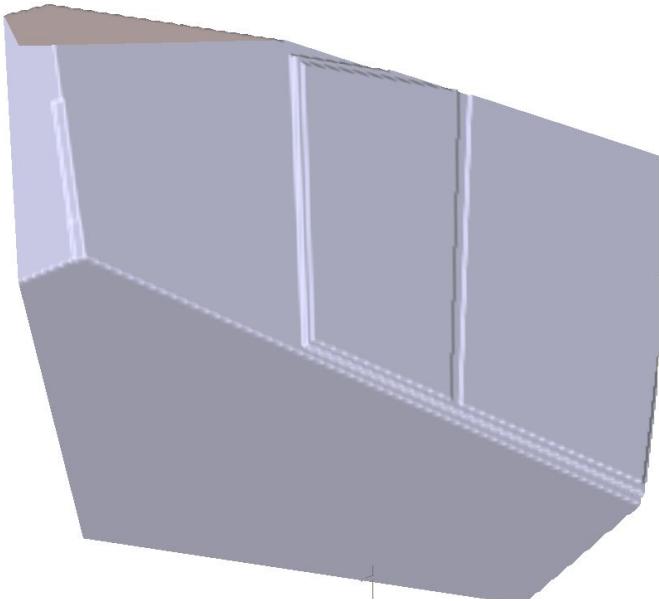


Object Instance-exit Depth Layer

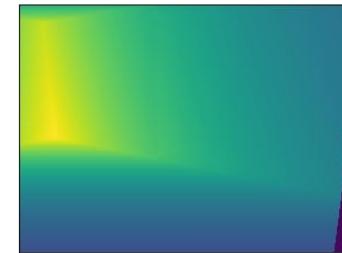
“Back of the first object instance”



D_5



Learning Target:



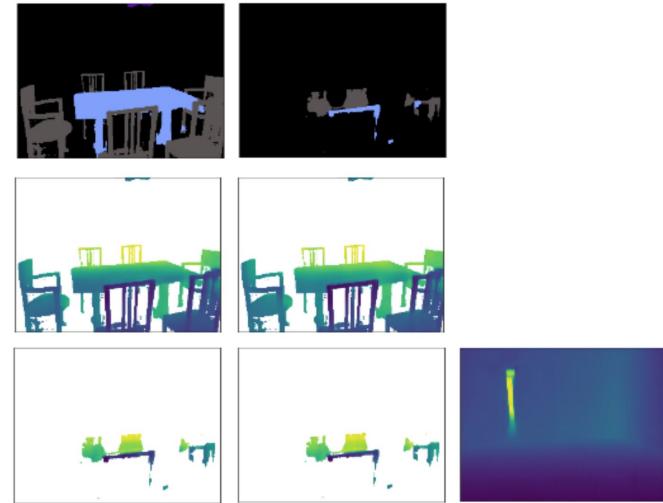
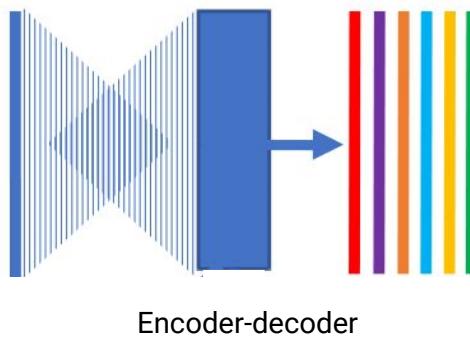
Room Envelope Depth Layer



Multi-layer Surface Prediction



Input RGB Image

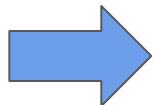


**Predicted Multi-layer Depth
and Semantic Segmentation**

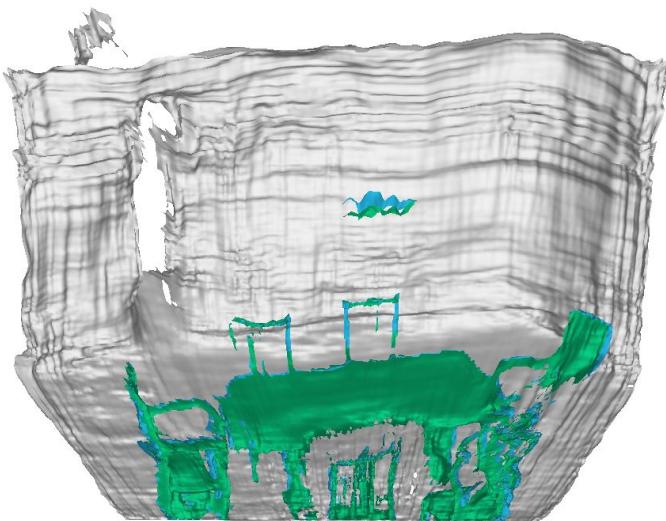
Multi-layer Surface Prediction



Input RGB Image



Multi-layer
Depth Prediction
and Segmentation



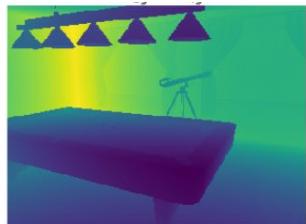
Surface Reconstruction from multi-layer depth

3D scene geometry from depth (2.5D)

- How much geometric information is present in a depth image?



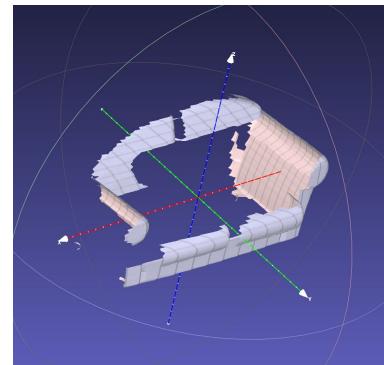
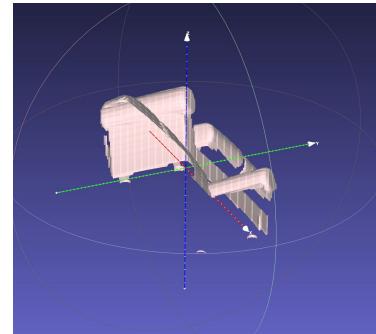
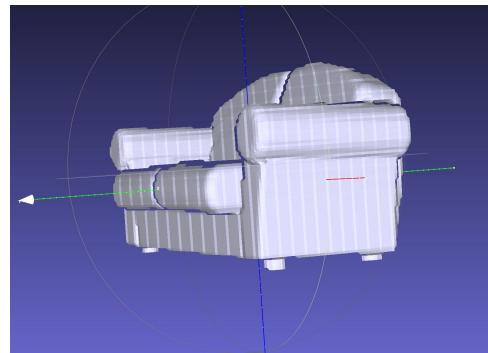
RGB image (2D)



2.5D depth

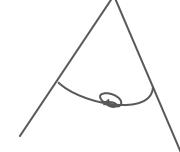


Mesh representation of a synthetically generated depth image (SUNCG).

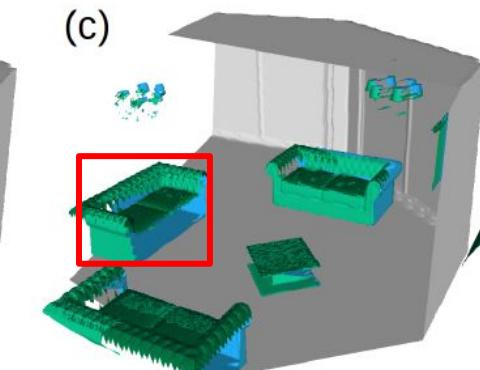
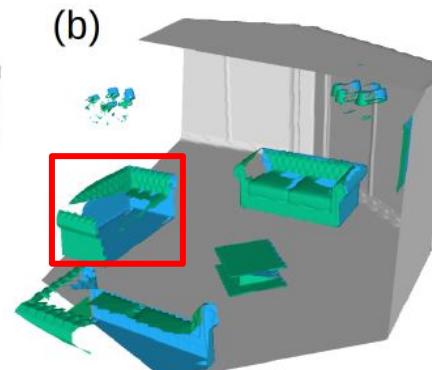
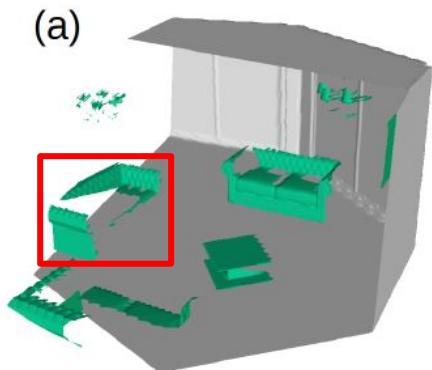


Epipolar Feature Transformers

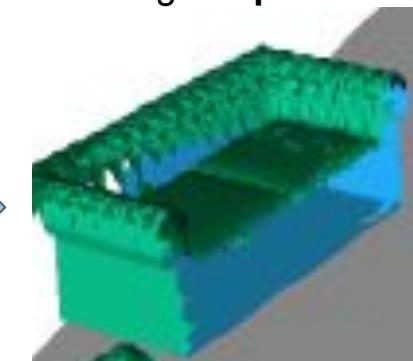
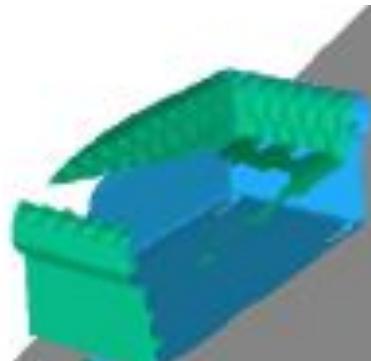
Multi-layer is not enough. Motivation for multi-view prediction



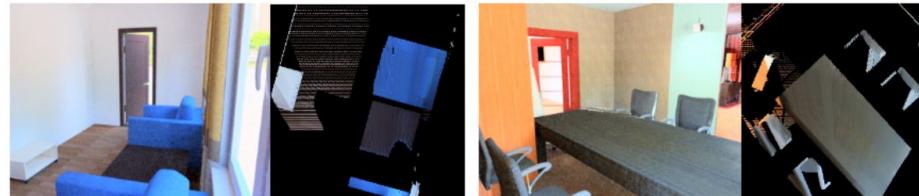
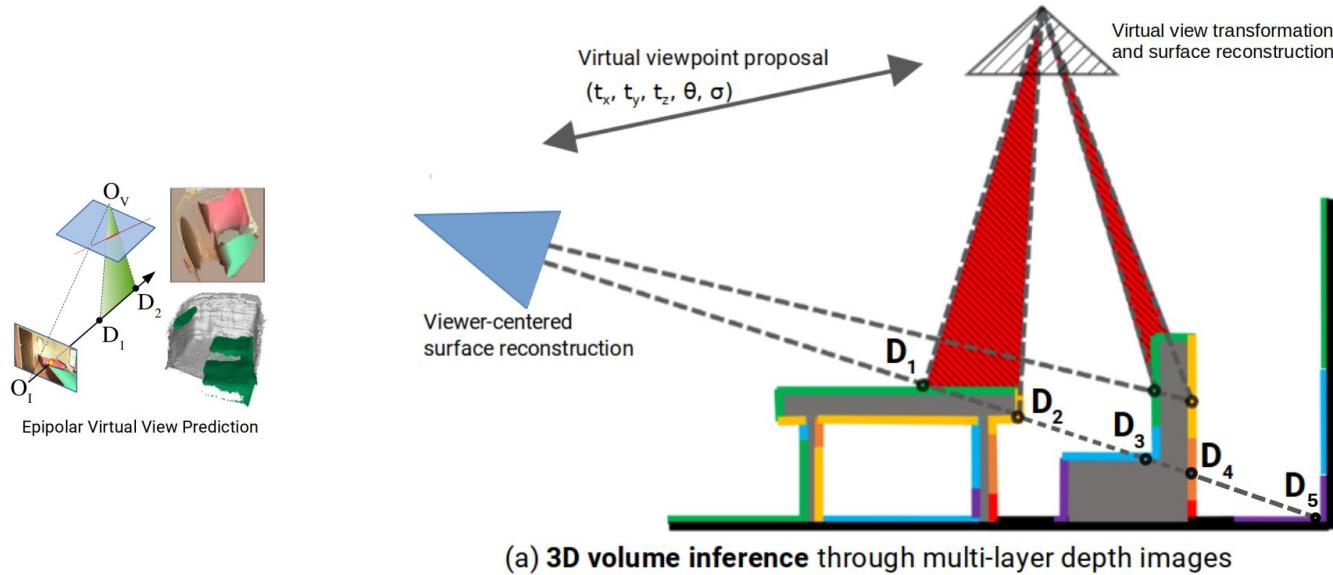
RGB Image



Ground truth depth
visualization

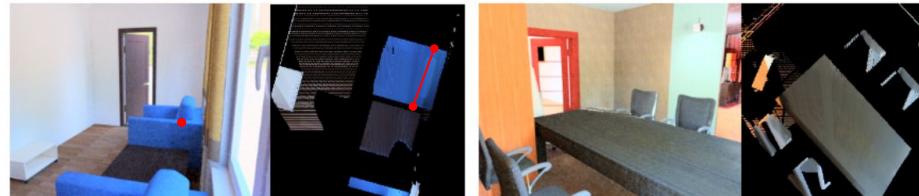
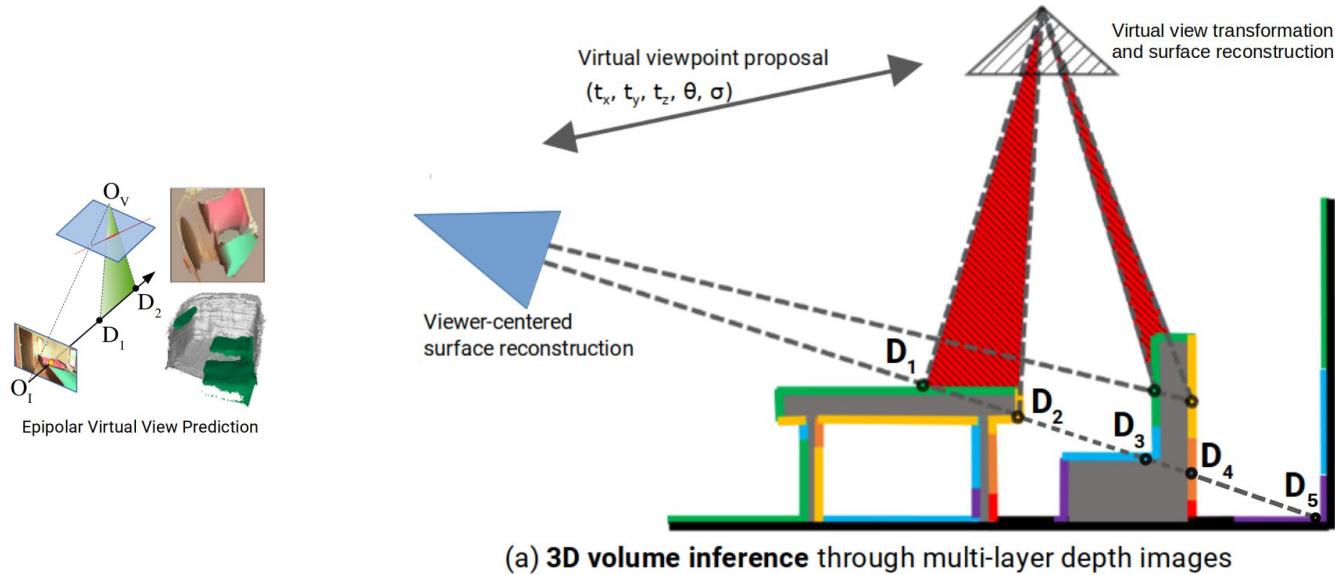


Multi-view prediction from a single image: **Epipolar Feature Transformer Networks**



(b) **Input** image and **transformed** color features using \bar{D}_1 and \bar{D}_2 .

Multi-view prediction from a single image: **Epipolar Feature Transformer Networks**



(b) **Input** image and **transformed** color features using \bar{D}_1 and \bar{D}_2 .

Virtual View Surface Prediction



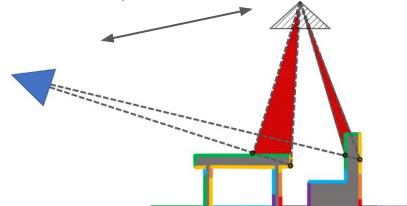
Input Image

Depth gating

Epipolar Feature Transformer

Virtual Viewpoint Proposal

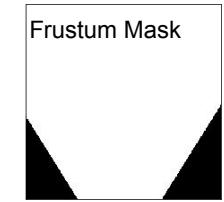
$$(t_x, t_y, t_z, \theta, \sigma)$$



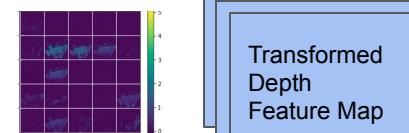
3 channels



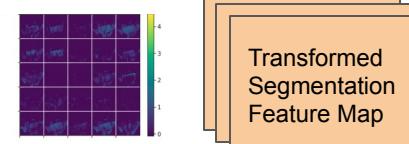
1 channel



1 channel

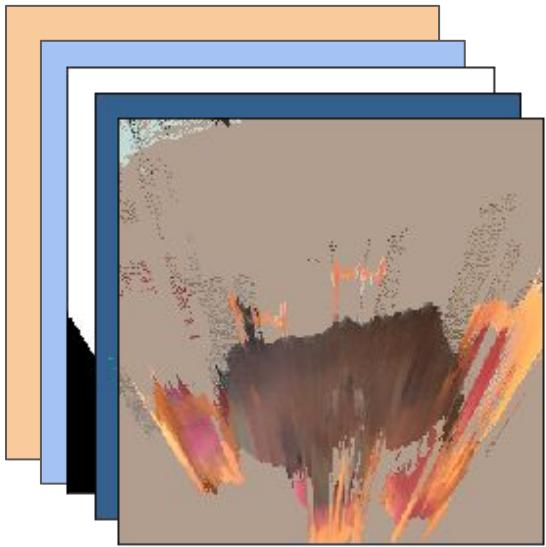


48 channels

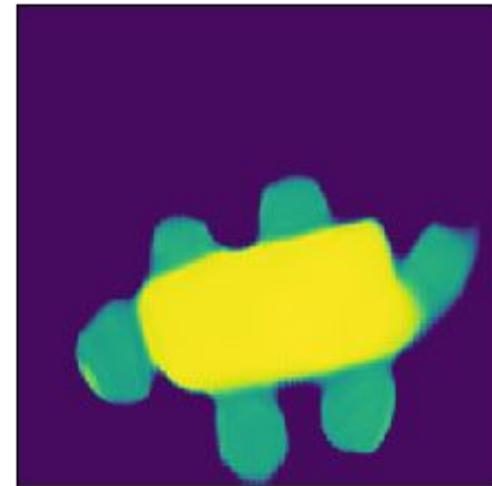
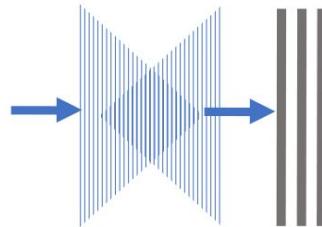


64 channels

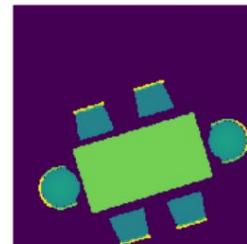
Transformed Virtual View Features
117 channels total



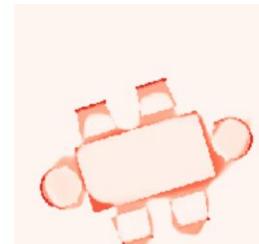
Transformed Virtual View Features



Height Map Prediction



Ground Truth

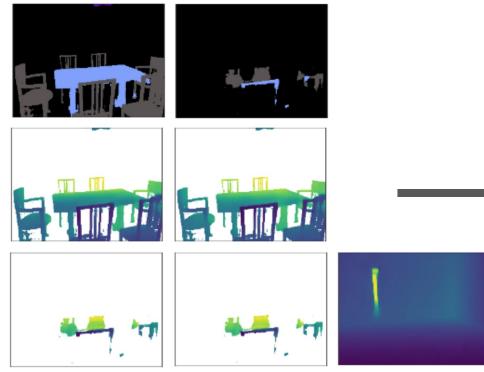


L1 Error Map

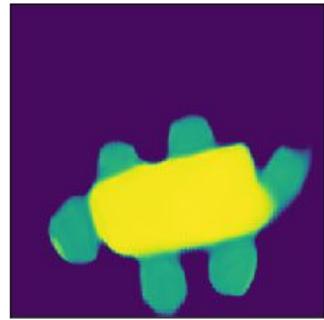
Multi-layer Multi-view Inference



Input Image



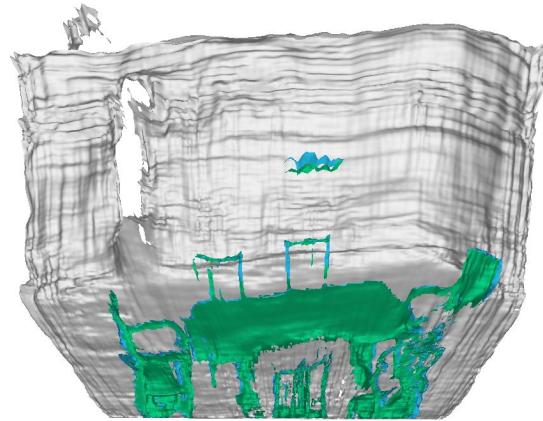
Frontal Multi-layer Prediction



Height Map Prediction



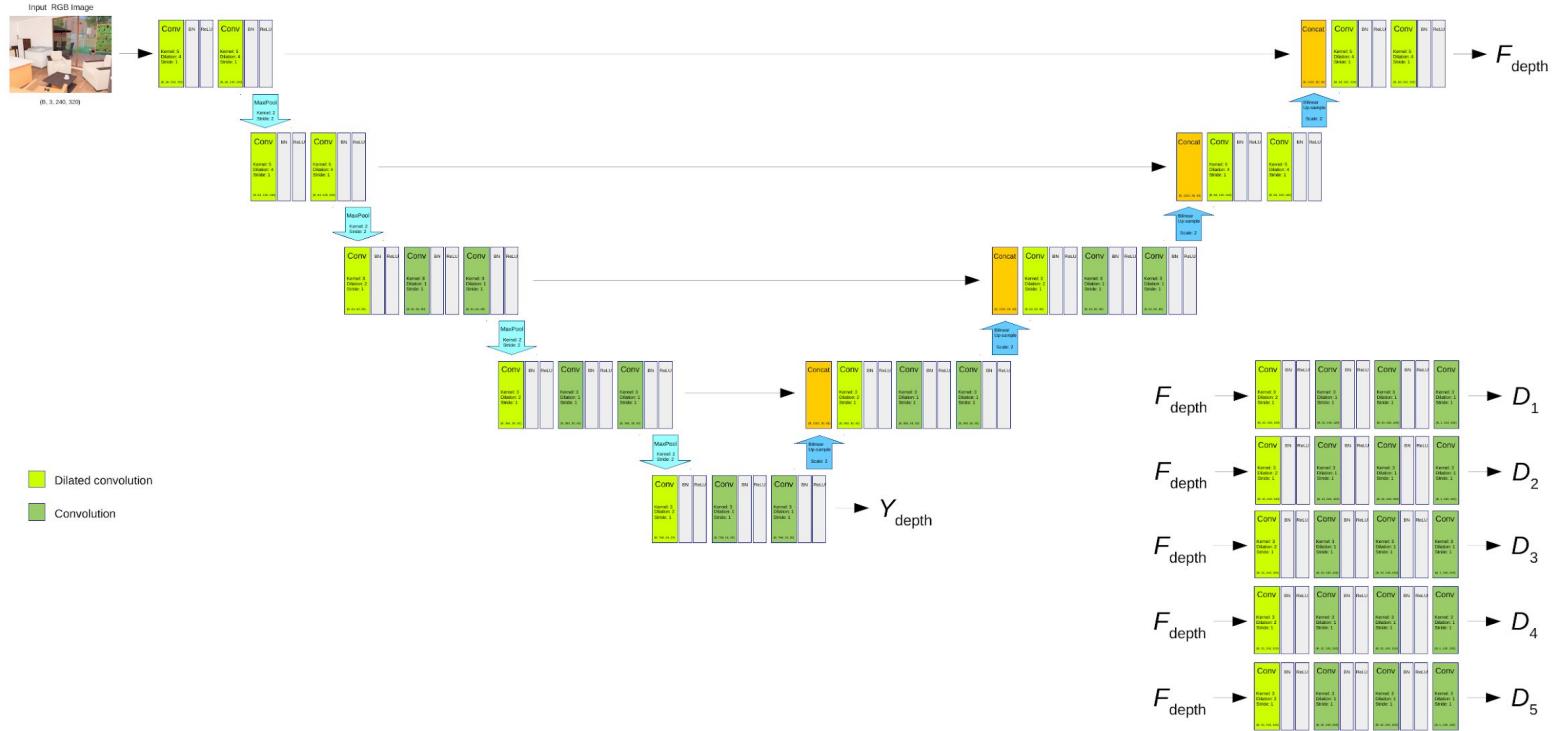
Virtual View Surface Reconstruction



Frontal View Surface Reconstruction



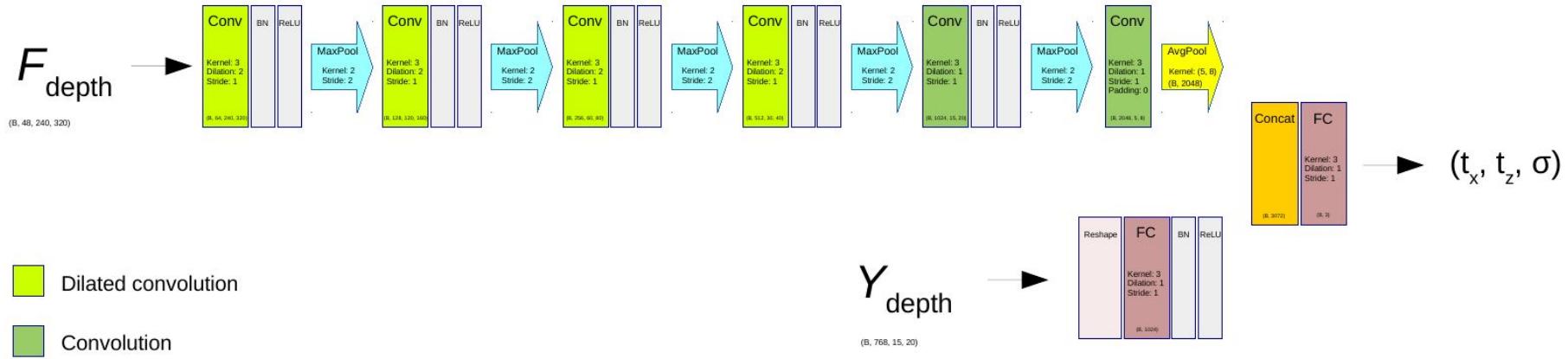
Network architecture for multi-layer depth prediction



Network architecture for multi-layer semantic segmentation



Network architecture for virtual camera pose proposal

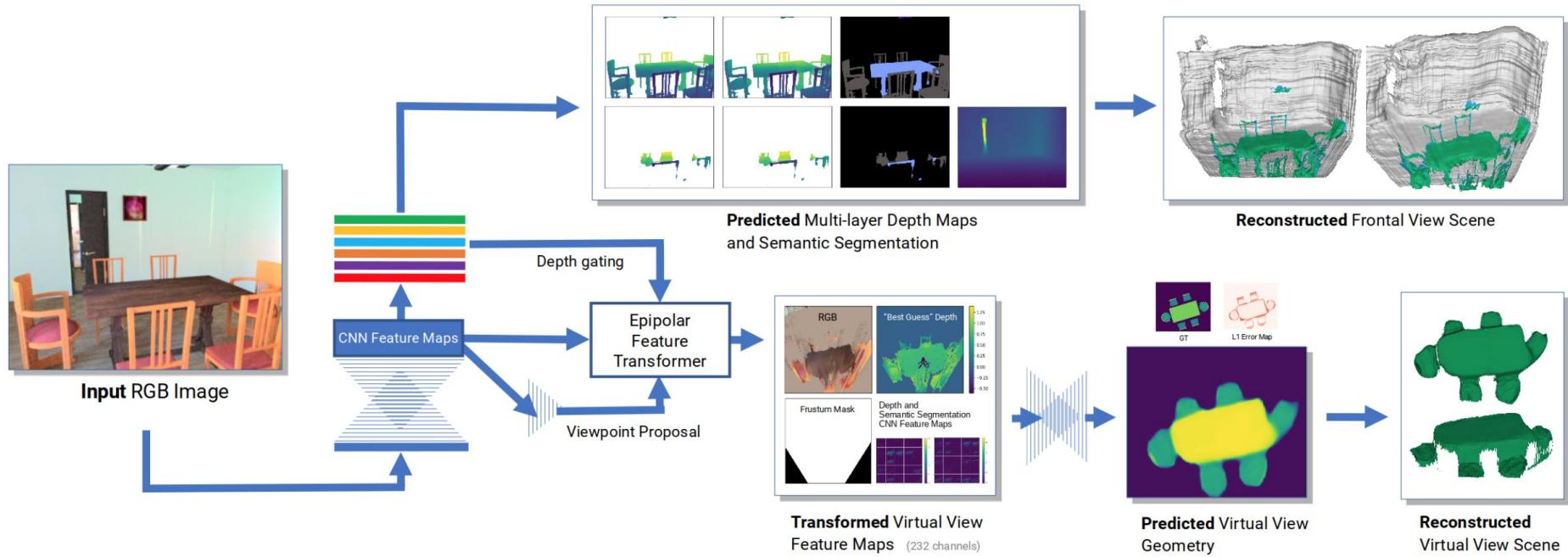


Network architecture for virtual view surface prediction



Network architecture for virtual view semantic segmentation





Layer-wise cumulative surface coverage

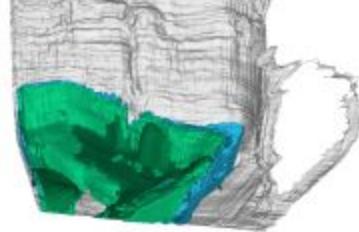
\bar{D}_1	$\bar{D}_{1,2}$	$\bar{D}_{1,2,3}$	$\bar{D}_{1..4}$	$\bar{D}_{1..5}$	$\bar{D}_{1..5} + \text{Ovh.}$
0.237	0.427	0.450	0.480	0.924	0.932

Table 1: Scene surface coverage (recall) of ground truth depth layers with a 5cm threshold. Our predictions cover 93% of the scene geometry inside the viewing frustum.

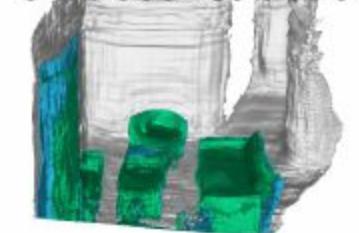
Results



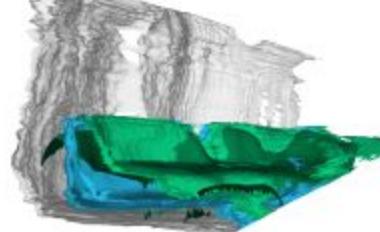
3D Reconstruction



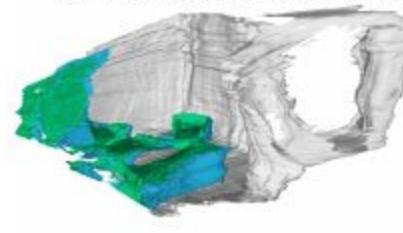
3D Reconstruction



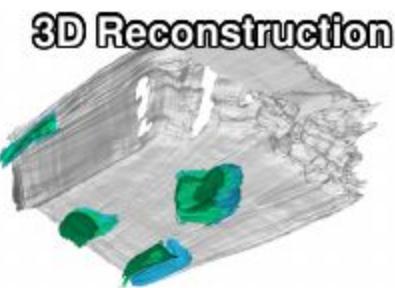
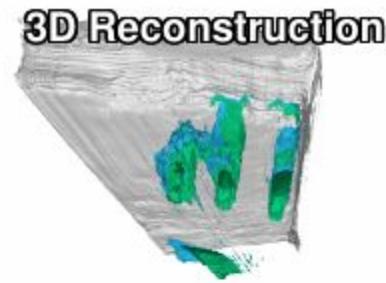
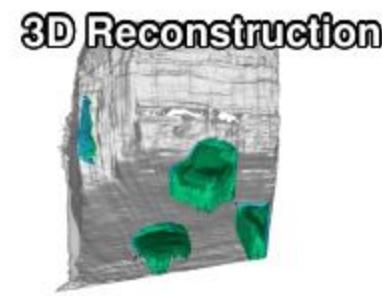
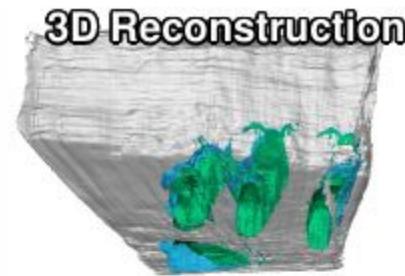
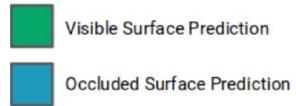
3D Reconstruction



3D Reconstruction



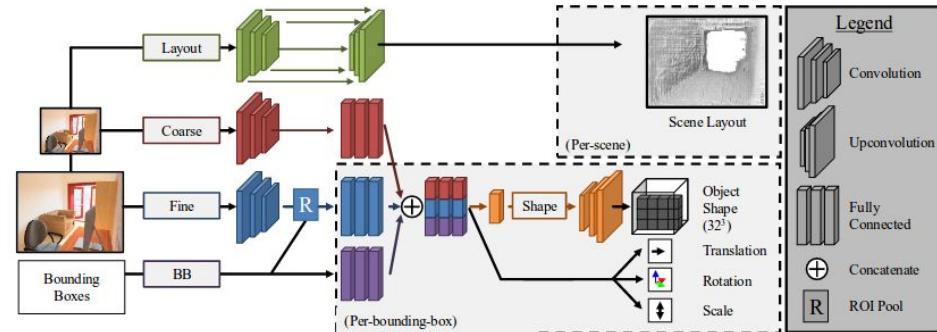
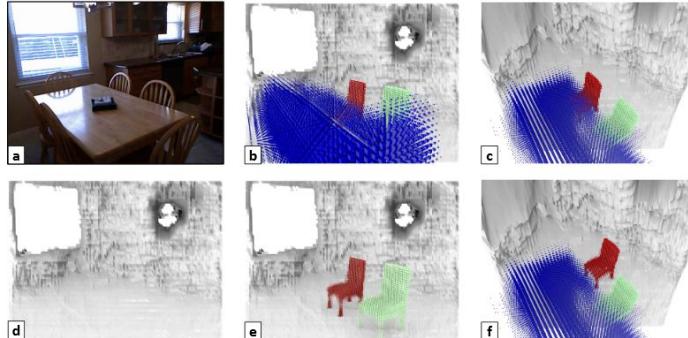
Input View / Alternate viewpoint



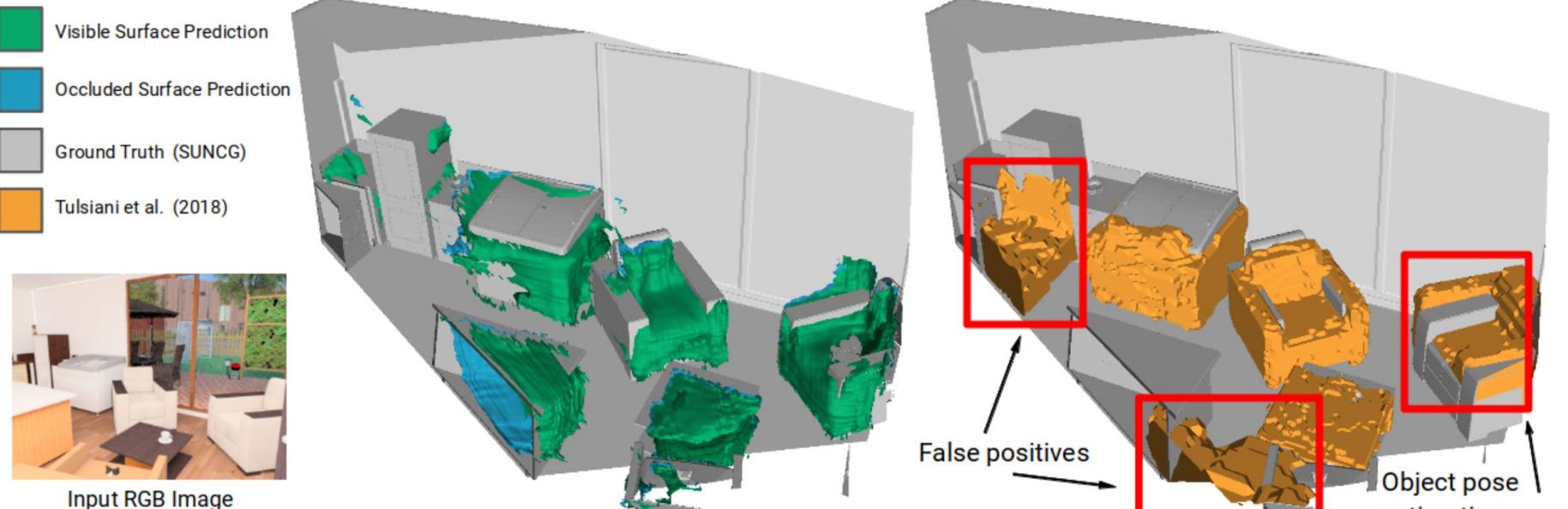
Input View / Alternate viewpoint

Previous state-of-the-art based on **object detection** and **volumetric** object shape prediction

- CVPR 2018
 - "Factoring Shape, Pose, and Layout from the 2D Image of a 3D Scene" by Tulsiani et al.
 - 3D scene geometry prediction from a single RGB image



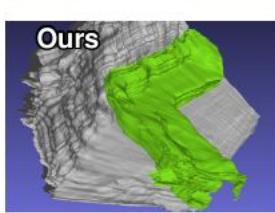
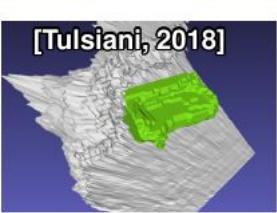
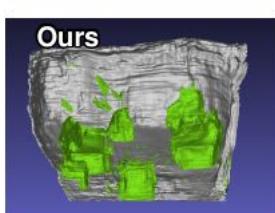
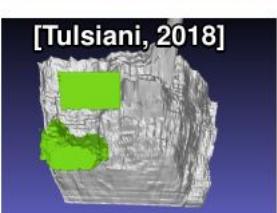
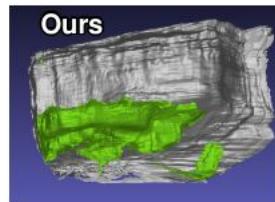
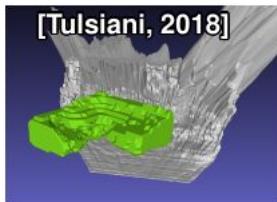
Object-based reconstruction is sensitive to detection and pose estimation errors



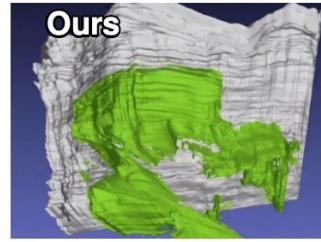
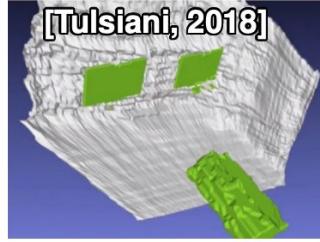
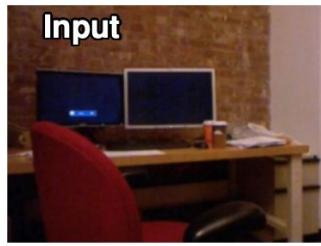
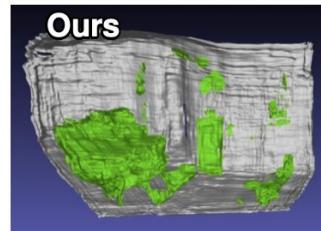
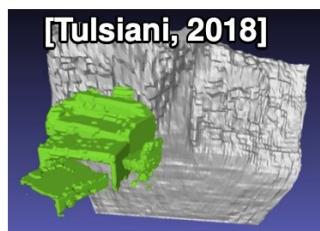
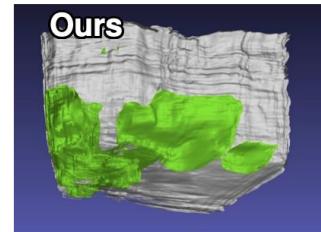
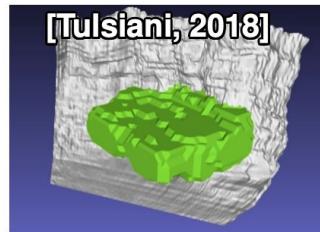
Our viewer-centered, end-to-end
scene surface prediction

Object-detection-based state of the art
(Tulsiani et al., CVPR 18)

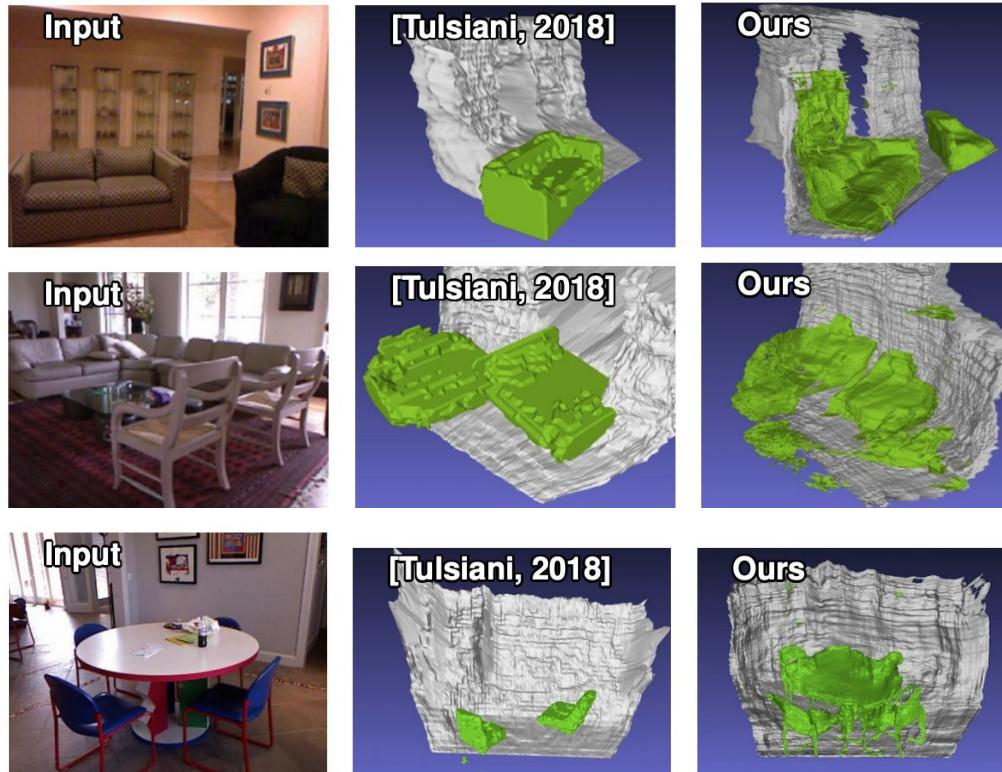
Results on real-world images: object detection error and geometry



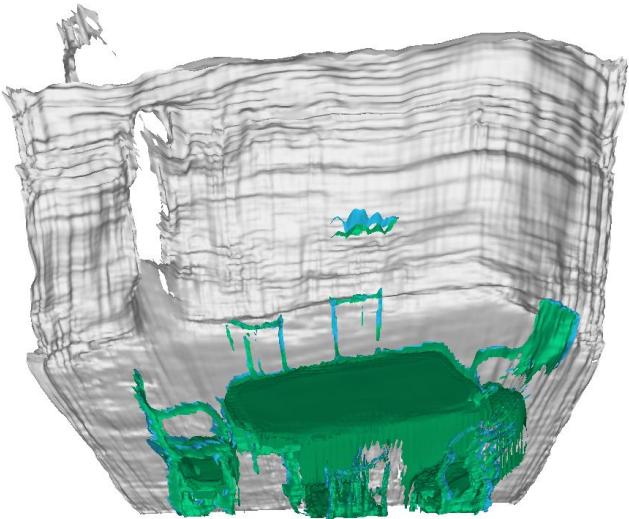
Results on real-world images



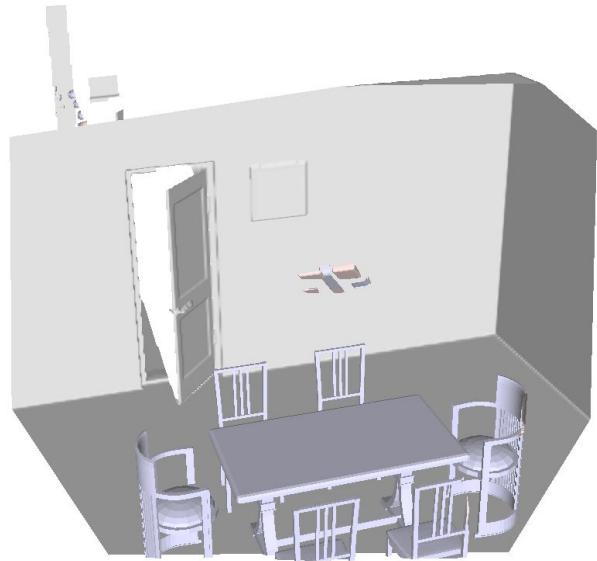
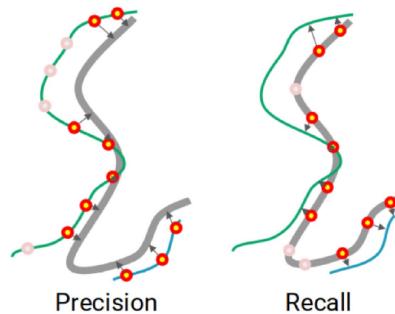
Results on real-world images



Quantitative Evaluation Metric



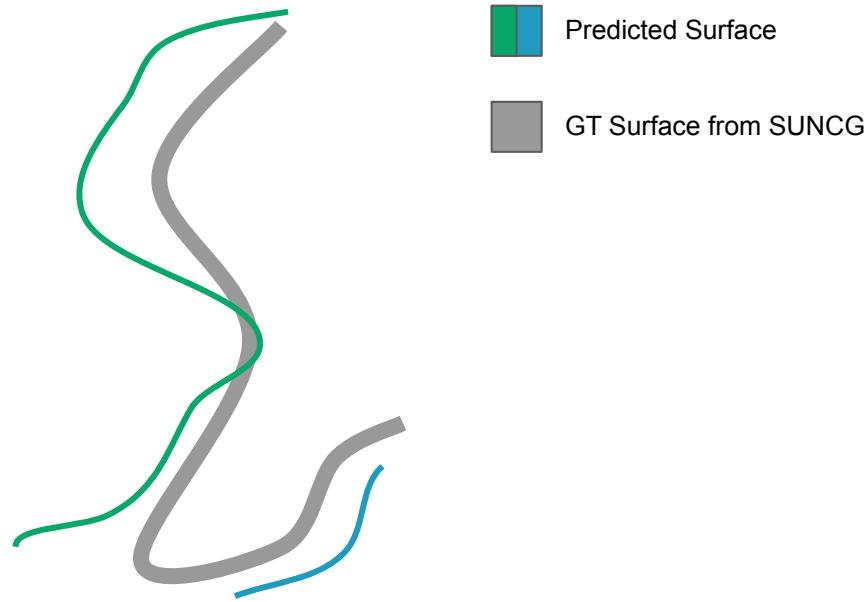
Predicted 3D Mesh



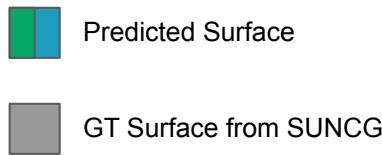
Ground Truth 3D Mesh

"Inlier" Threshold: ← →

Surface Coverage
Precision-Recall Metrics



Surface Coverage
Precision-Recall Metrics



i.i.d. point sampling on predicted mesh

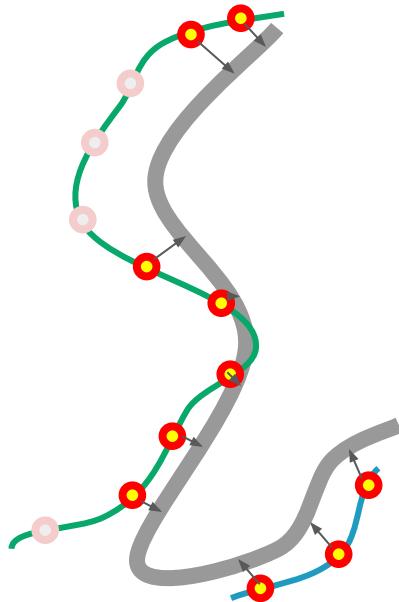
(with constant density $\rho = 10000$ points per
unit area, m^2 in real world scale)



Surface Coverage **Precision**-Recall Metrics

Precision =

$$\frac{\text{Number of points within threshold } (\textcolor{red}{\bullet})}{\text{Total number of sampled points } (\textcolor{red}{\bullet} + \textcolor{pink}{\circ})}$$



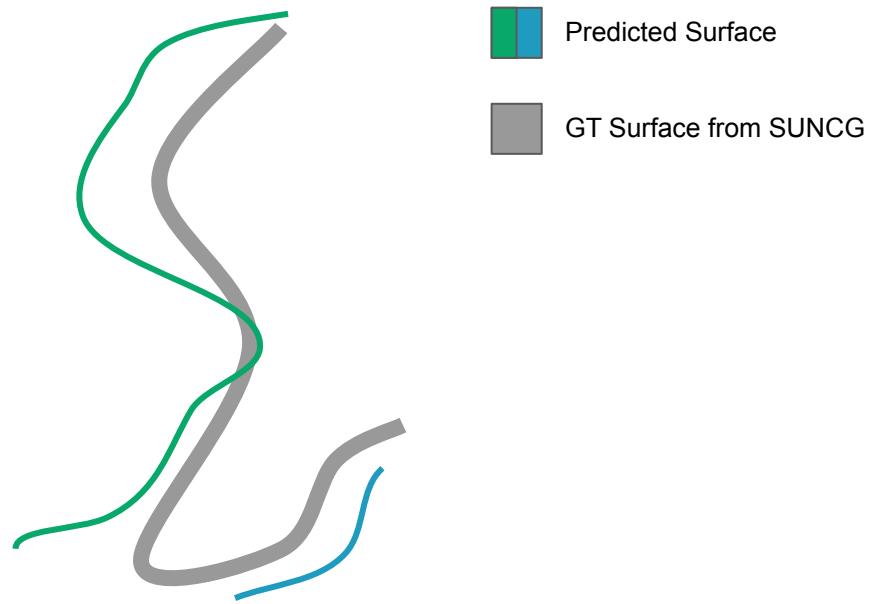
 Predicted Surface

 GT Surface from SUNCG

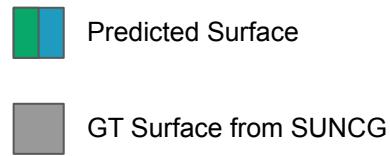
→ Closest distance from point to surface, within threshold

"Inlier" Threshold: 

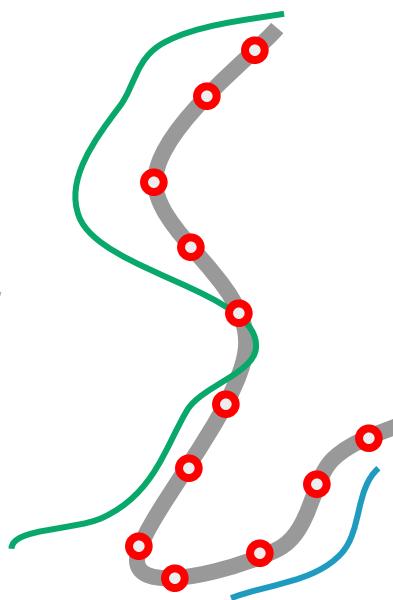
Surface Coverage
Precision-**Recall** Metrics



Surface Coverage
Precision-**Recall** Metrics



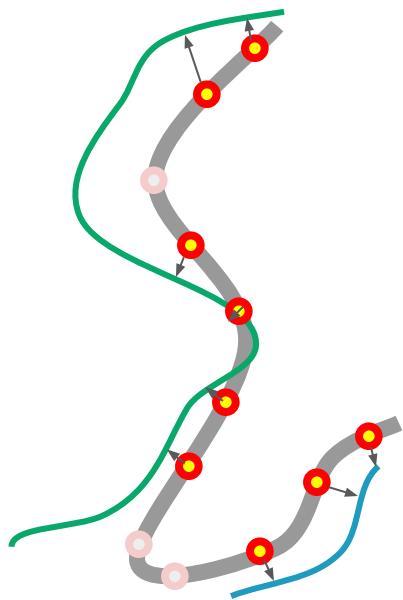
i.i.d. point sampling on GT mesh
(with constant density $\rho = 10000$ points per
unit area, m^2 in real world scale)



Surface Coverage Precision-**Recall** Metrics

Recall =

$$\frac{\text{Number of points within threshold } (\textcolor{red}{\bullet})}{\text{Total number of sampled points } (\textcolor{red}{\bullet} + \textcolor{pink}{\circ})}$$



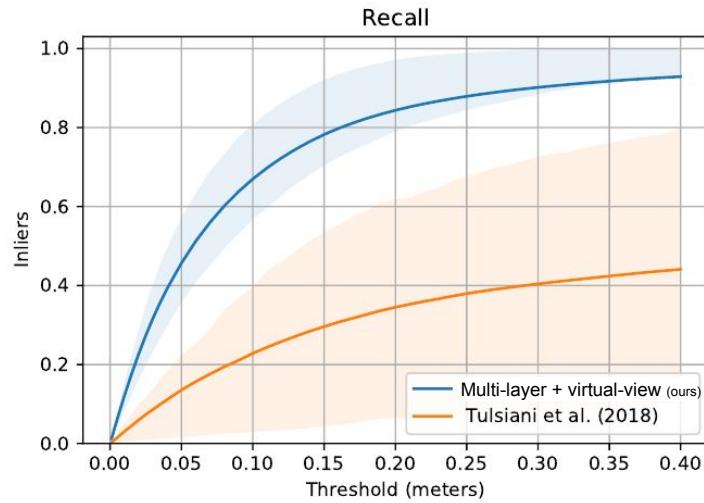
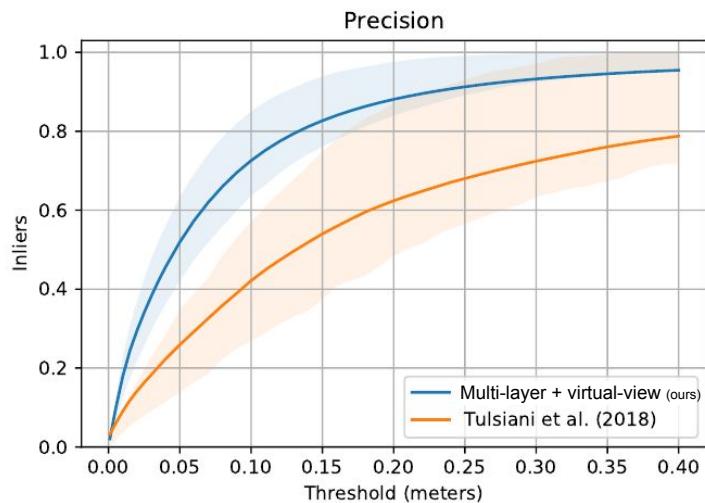
Predicted Surface

GT Surface from SUNCG

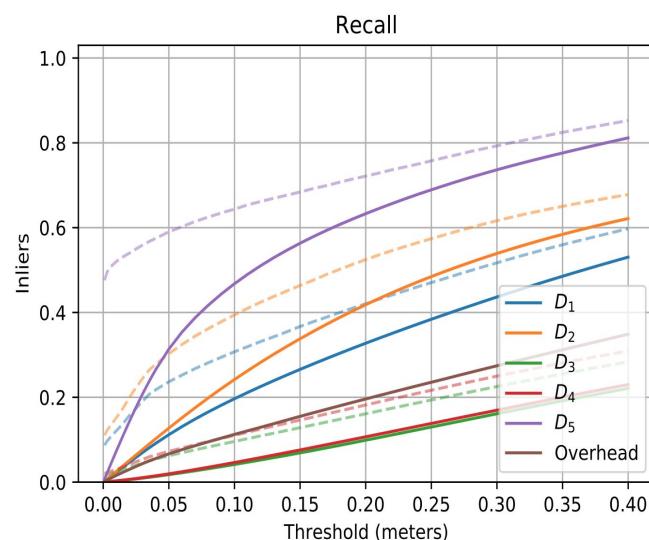
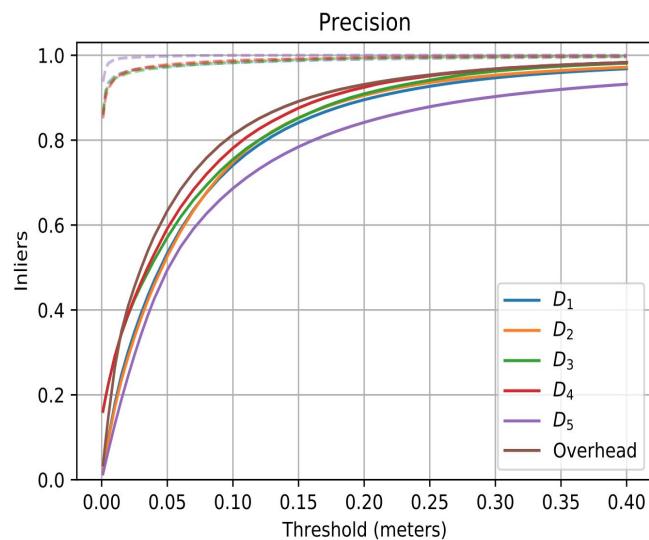
→ Closest distance from point to surface, within threshold

"Inlier" Threshold:

Our multi-layer, virtual-view depths vs. Object detection based state-of-the-art, 2018



Layer-wise evaluation



Top-down virtual-view prediction improves both precision and recall

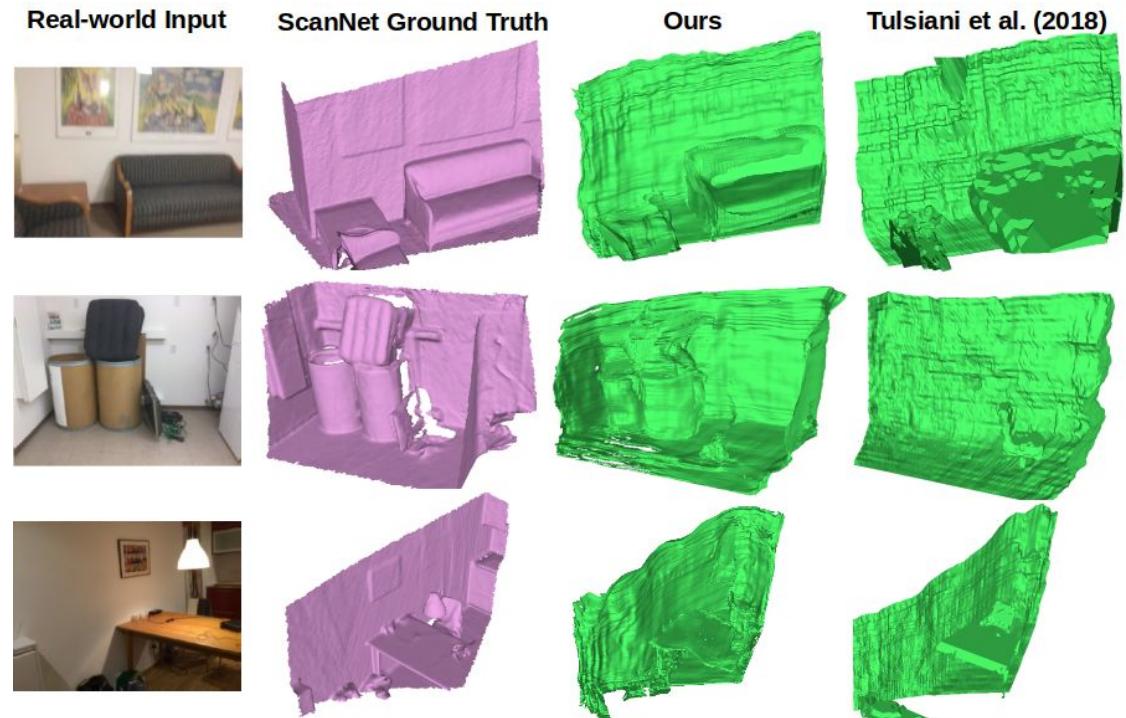
	Precision	Recall
$D_{1,2,3,4}$	0.499	0.417
$D_{1,2,3,4}$ & Overhead	0.519	0.457

(Match threshold of 5cm)

Synthetic-to-real transfer of 3D scene geometry on ScanNet

$D_{1,2,3,4,5}$ & Overhead
Tulsiani et al. [43]

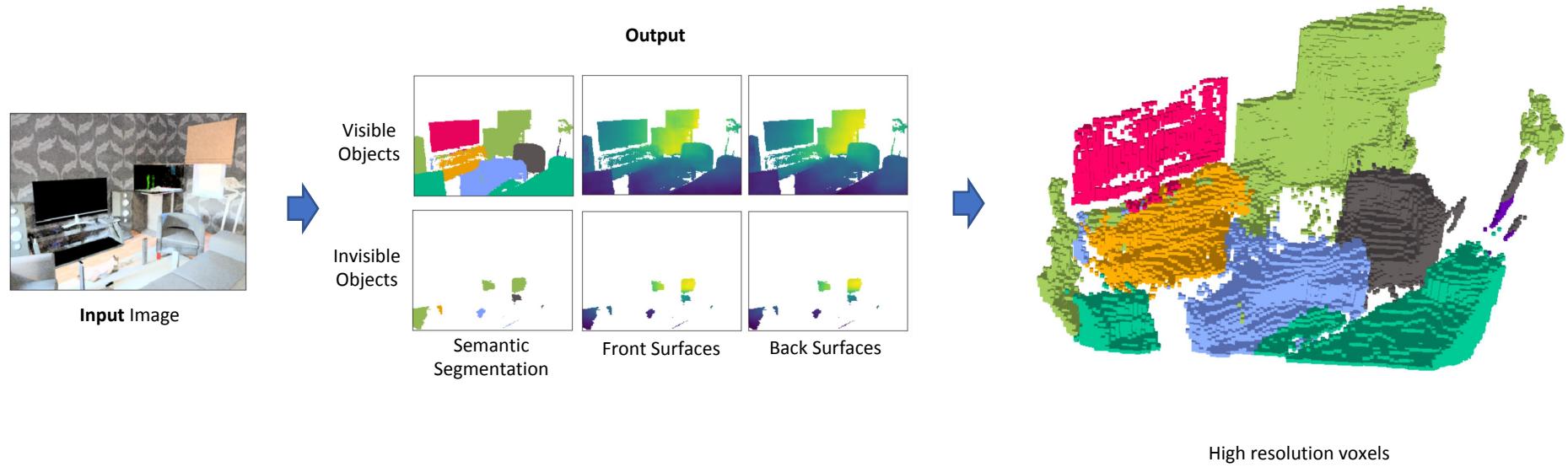
Precision	Recall
0.221	0.358
0.132	0.191



We measure recovery of true object surfaces and room layouts within the viewing frustum (threshold of 10cm).

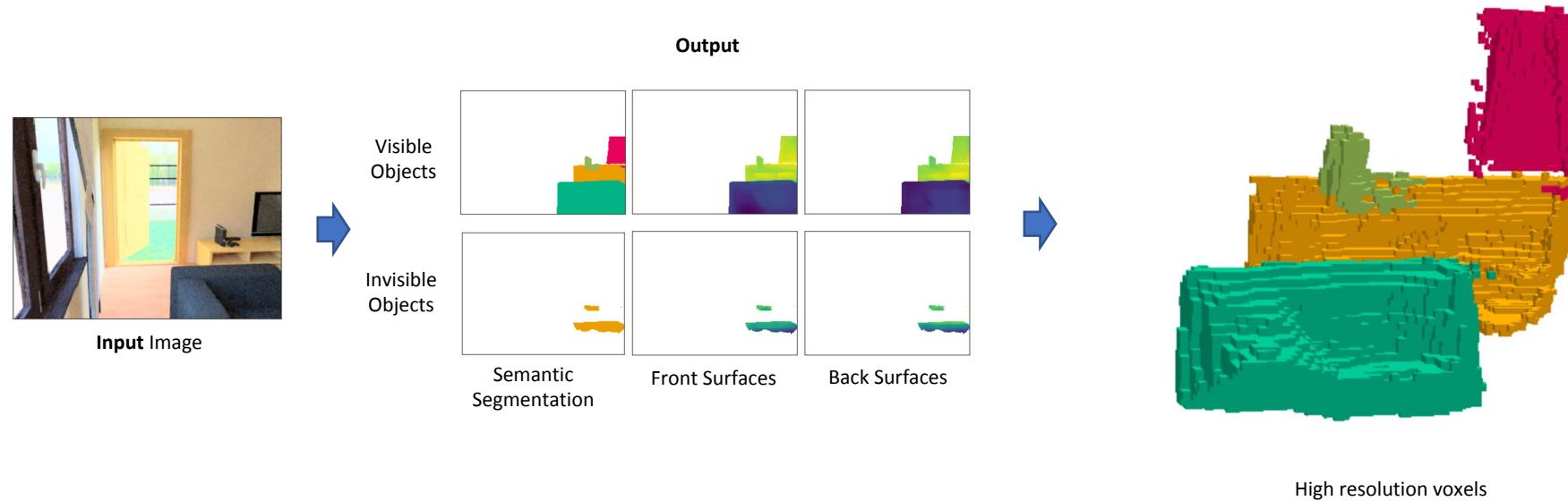
Voxelization of multi-layer depth maps

We project the center of each voxel into the input camera, and the voxel is marked occupied if the depth value falls in the first object interval (D_1, D_2) or the occluded object interval (D_3, D_4).

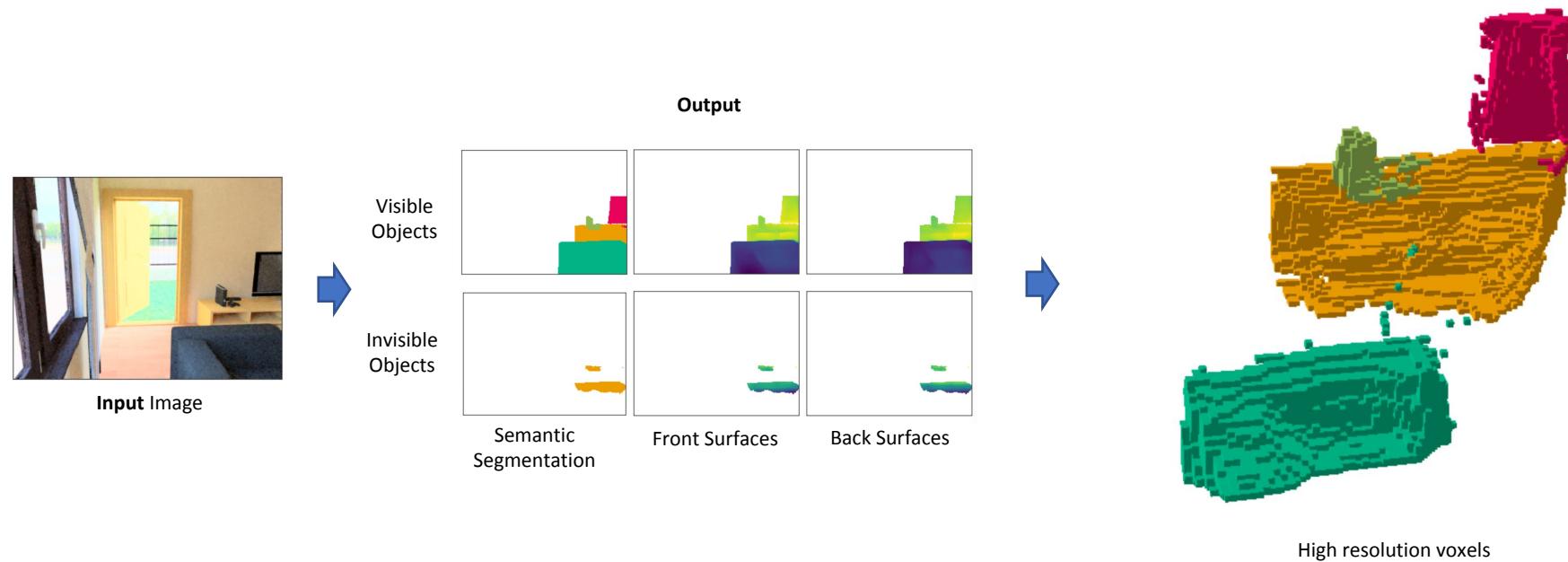


Our fully convolutional, viewer-centered inference of 3D scene geometry

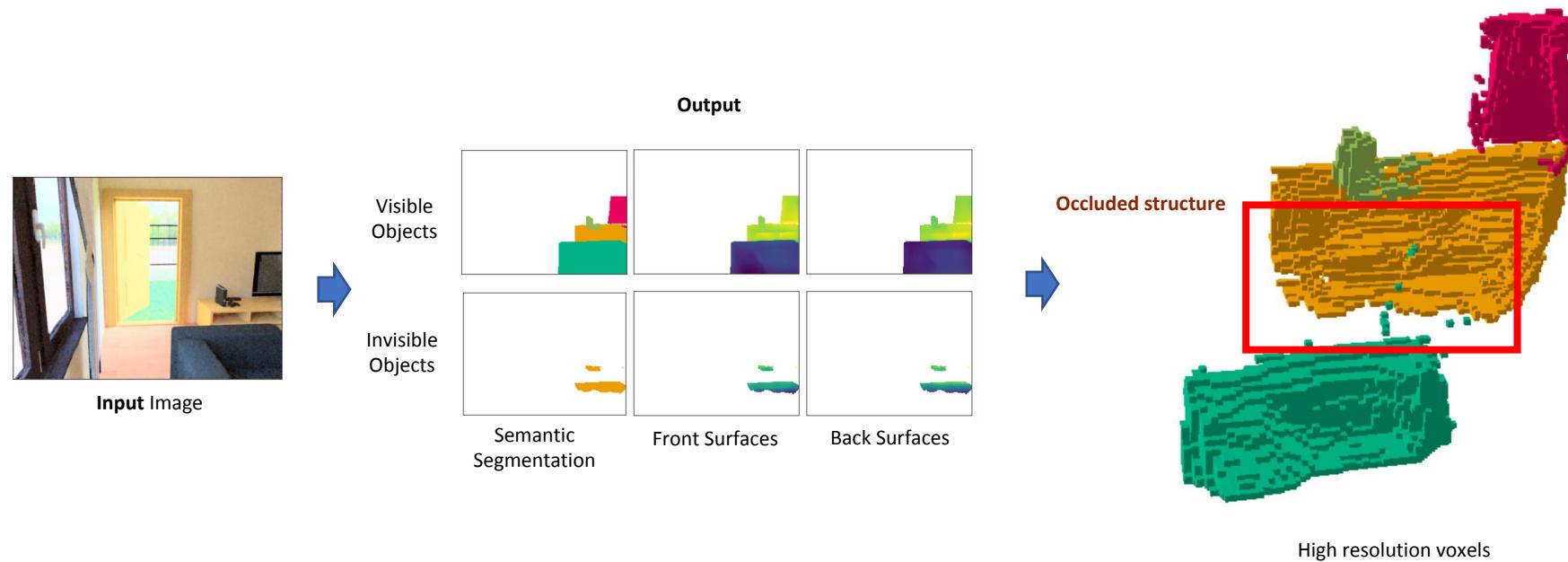
Voxelization of multi-layer depth maps



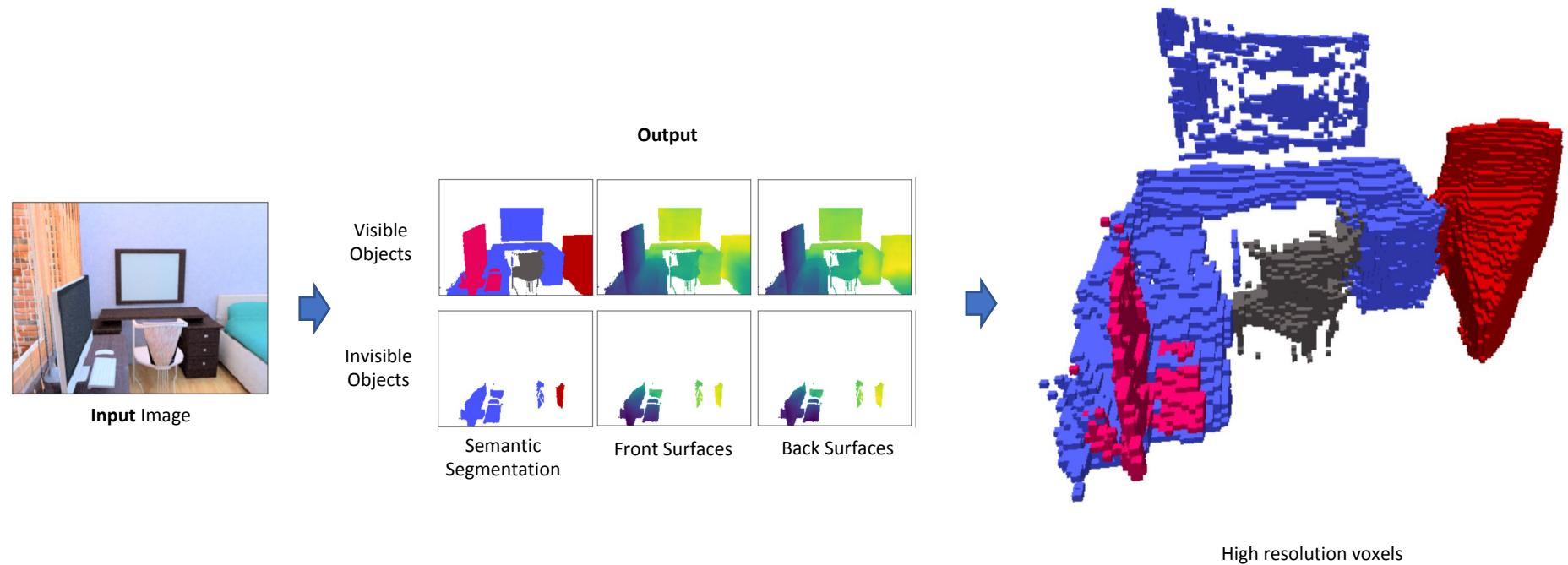
Voxelization of multi-layer depth maps



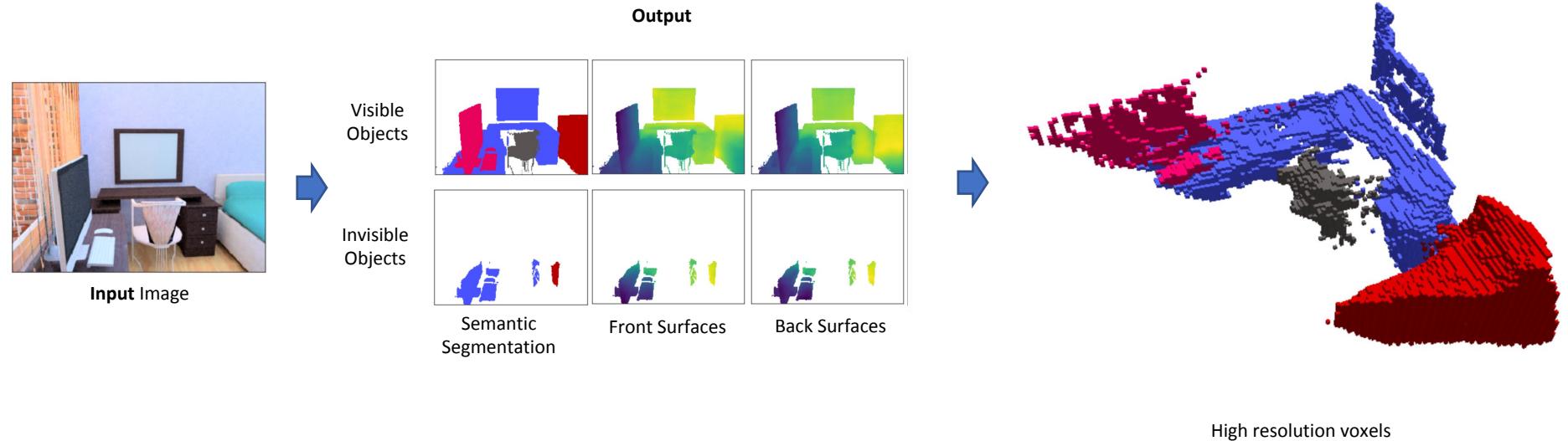
Voxelization of multi-layer depth maps



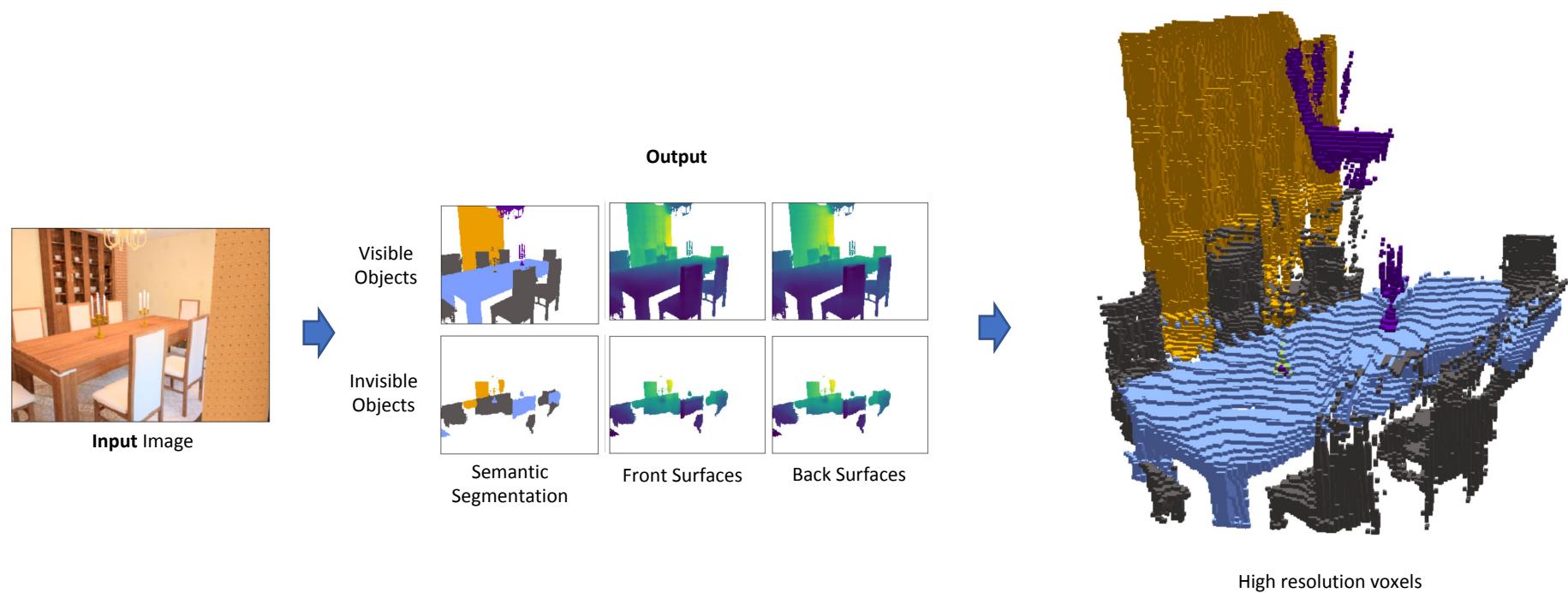
Voxelization of multi-layer depth maps



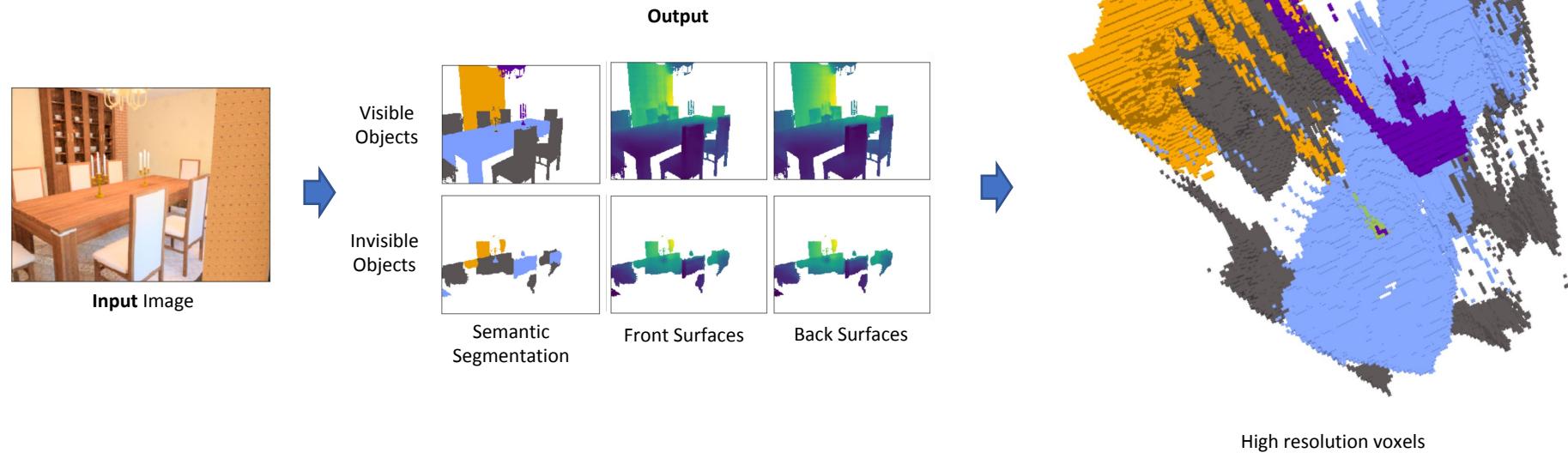
Voxelization of multi-layer depth maps



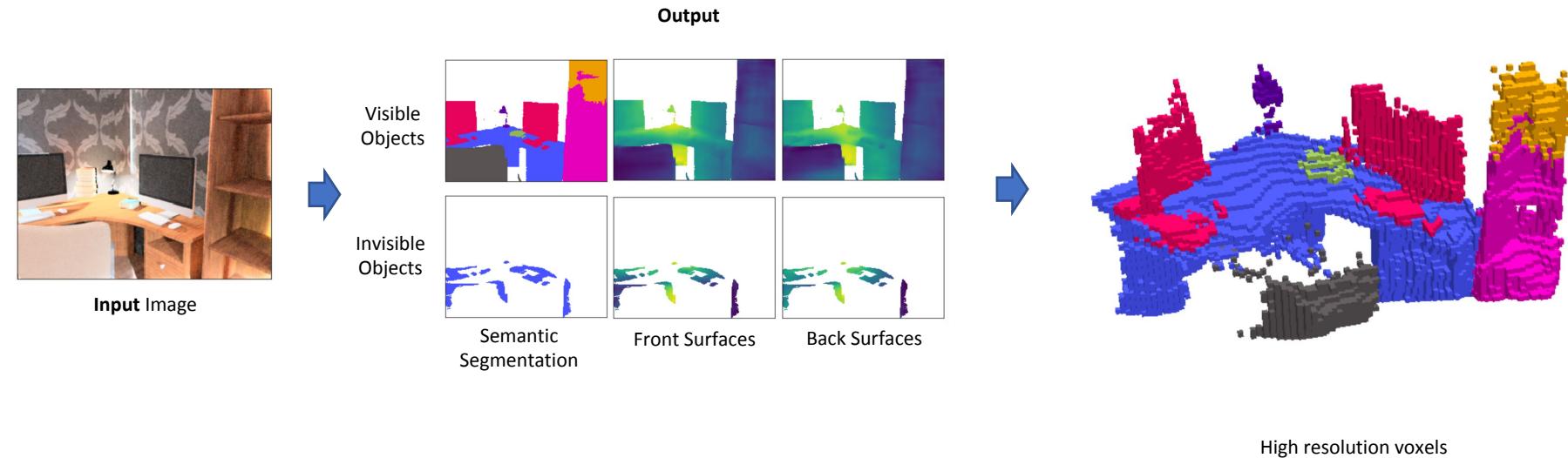
Voxelization of multi-layer depth maps



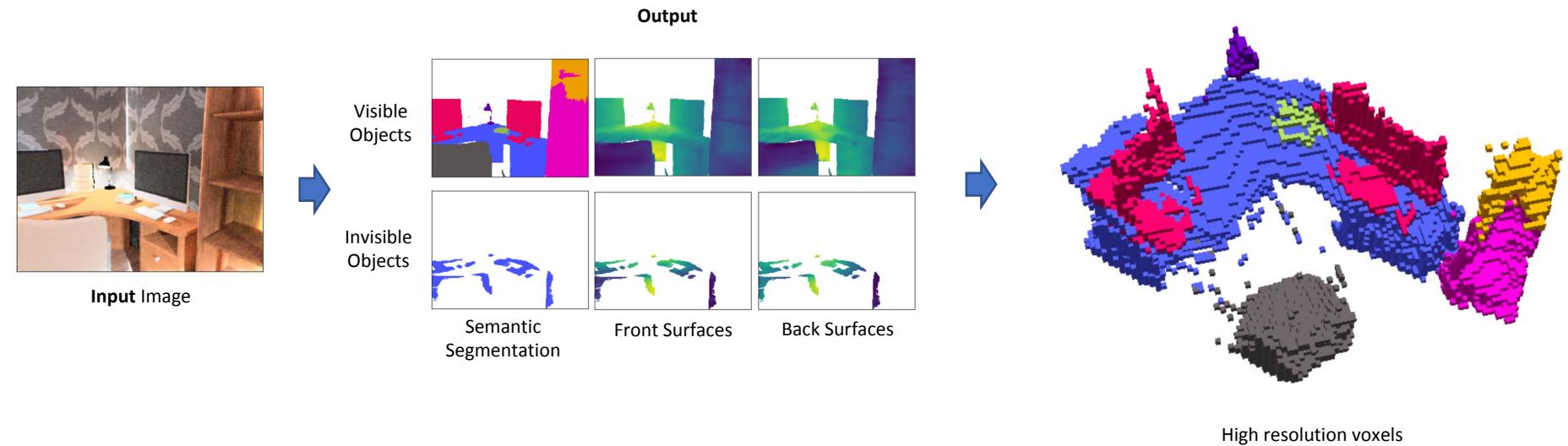
Voxelization of multi-layer depth maps



Voxelization of multi-layer depth maps



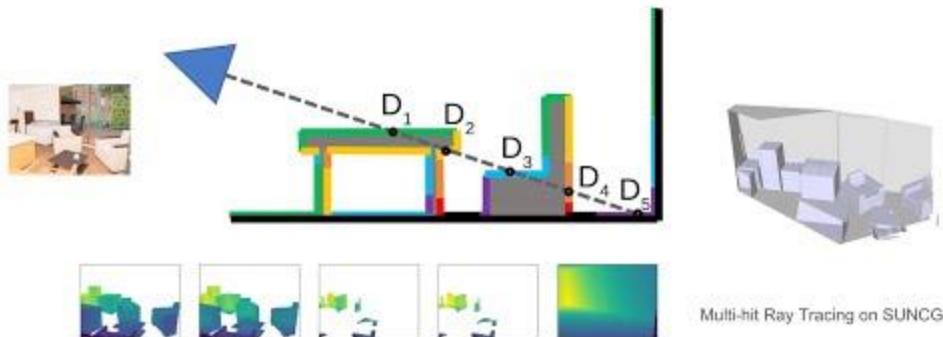
Voxelization of multi-layer depth maps



Supplemental Video

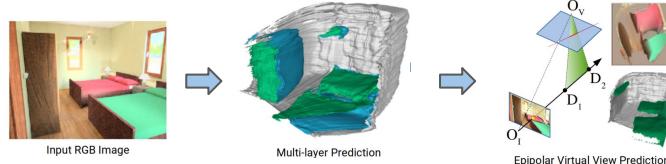


Our approach: Multi-layer Depth Representation



Conclusion

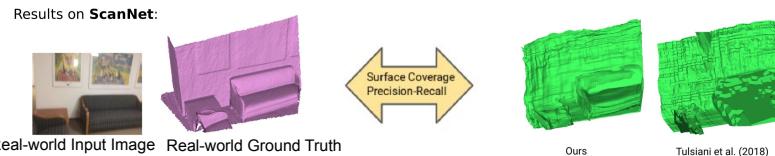
- Multi-layer and virtual-view prediction from a **single** image



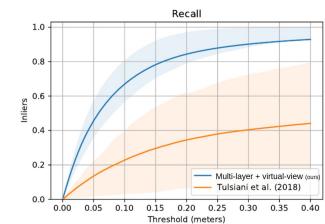
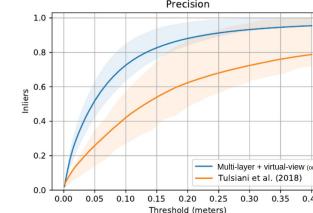
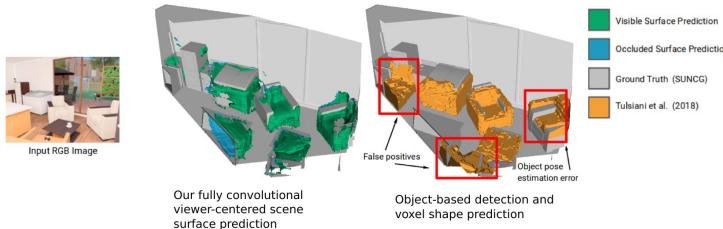
- Surface-based accuracy evaluation



- Synthetic-to-real transfer of 3D scene geometry prediction, evaluated quantitatively



- Geometric comparison with detection-based voxel prediction methods



Code and dataset coming soon.
Follow on Twitter for updates!



@DaeyunShin



UCIRVINE