From Pixels to Scene:

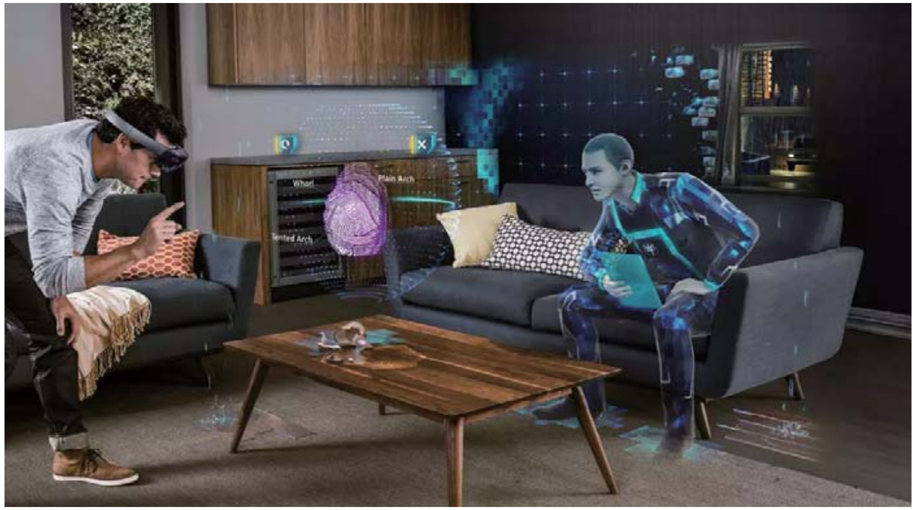
Recovering 3D Geometry and Semantics for Indoor Environments

Yinda Zhang

Google

Email: yindaz@gmail.com
Homepage: www.zhangyinda.com

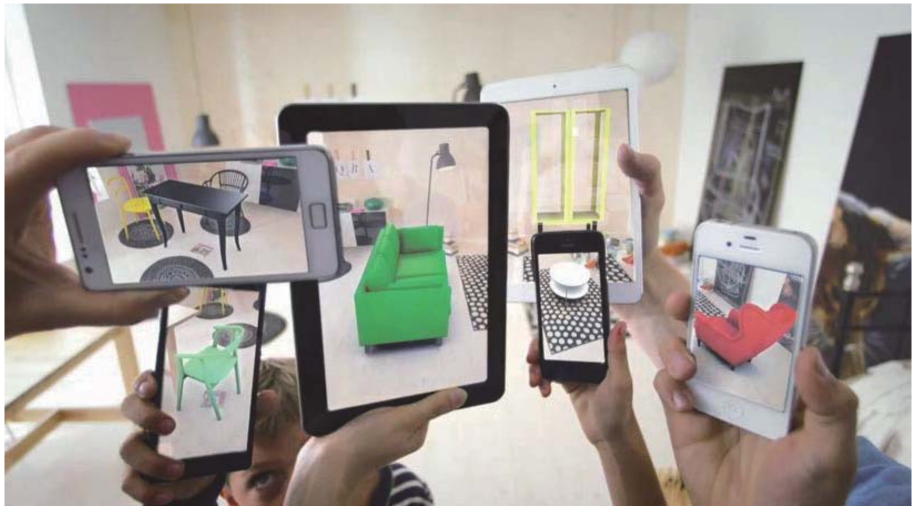
Scene Understanding



Augmented Reality Game

Autonomous driving

Indoor Robotics



Ecommerce

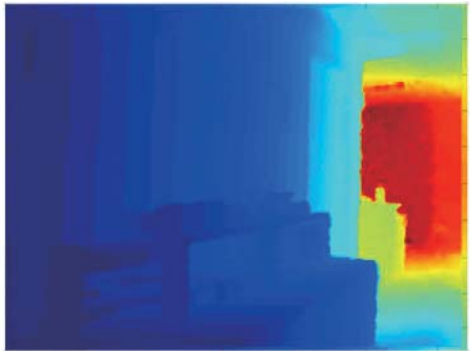
Challenging

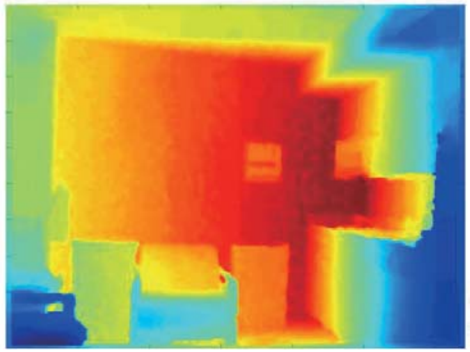
- Data
- Task
- Formulation

Challenging

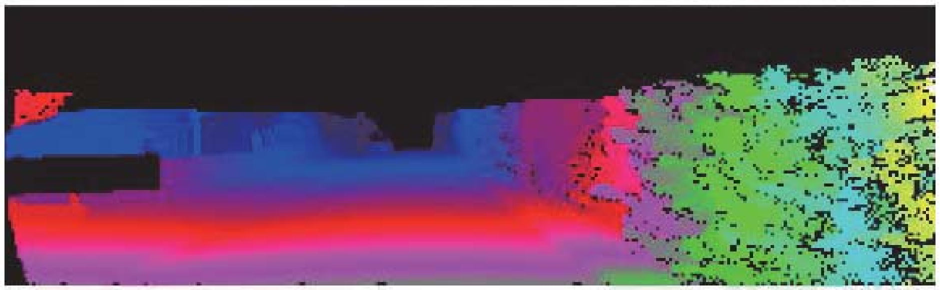
- Data
- Task
- Formulation

- Indoor, Depth ~ 3m



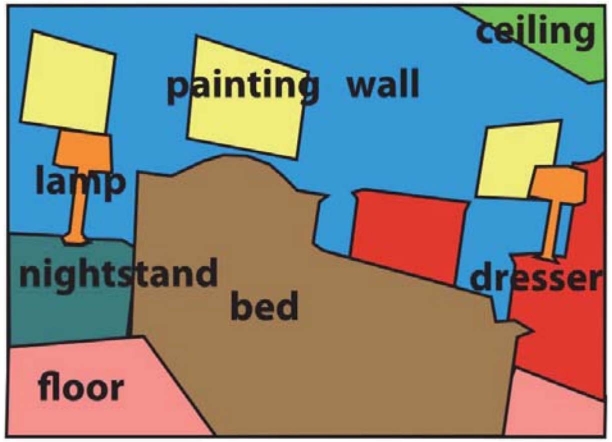
- Outdoor, Depth ~ 50m

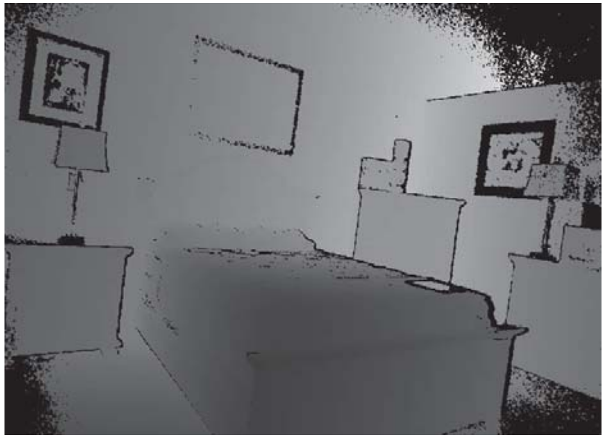
Challenging

- Data
- Task
- Formulation

Image



Semantic



Geometry

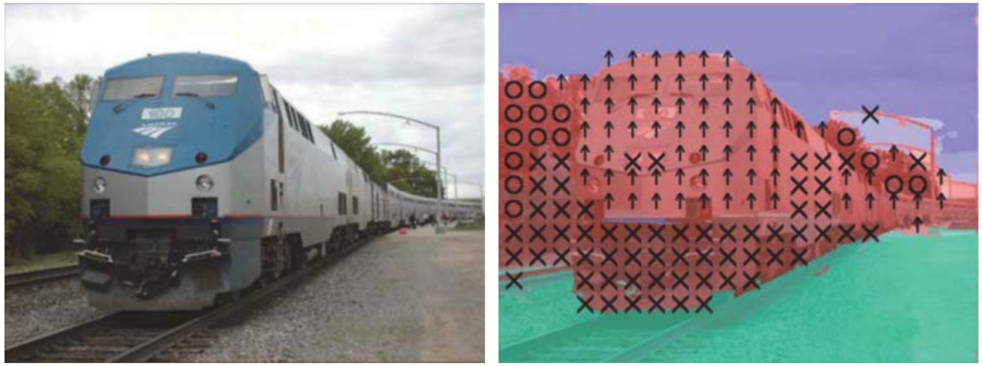
Challenging

- Data
- Task
- Formulation

Surface Label Estimation

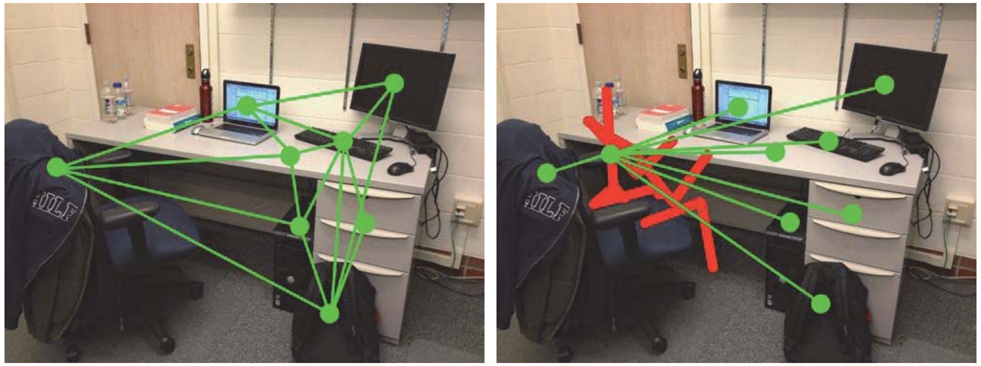
Hedau et al. ICCV2009

Geometric Context



Hoiem et al. ICCV2005

Predict Human Interaction

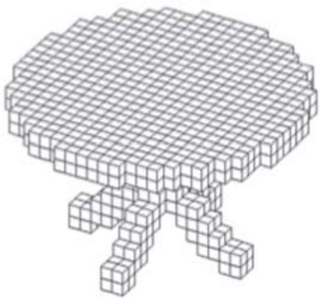


Jiang et al. ICCV2009

Challenging

- Data
- Task
- Formulation

- Geometry Representation



Volumetric

Point Cloud

Mesh

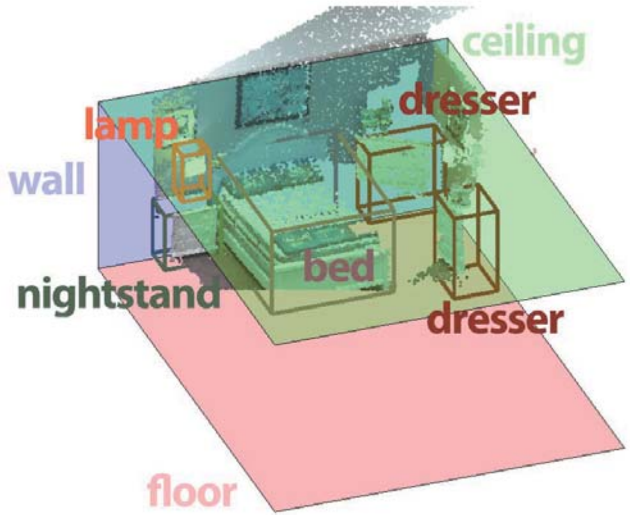
- Semantic Representation

Classification

Detection

Segmentation

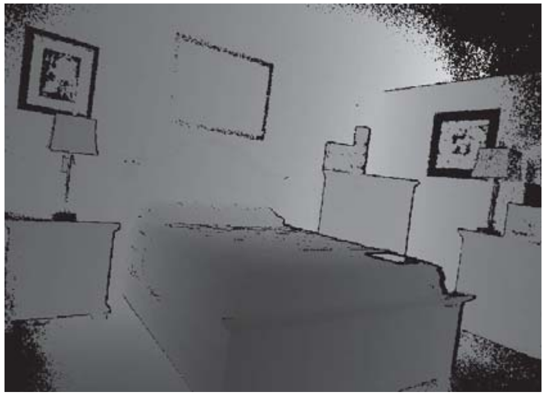
- Holistic



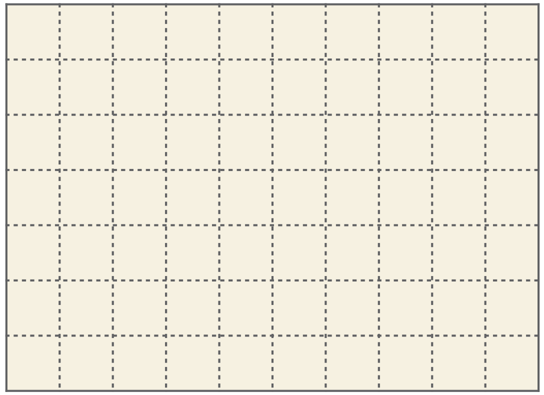
We’ll discuss…

- Data
- Task
- Formulation

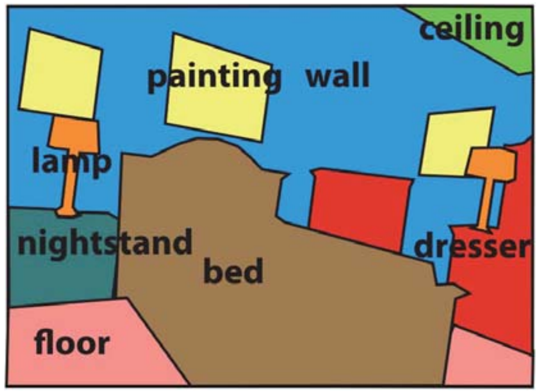
Indoor Scene



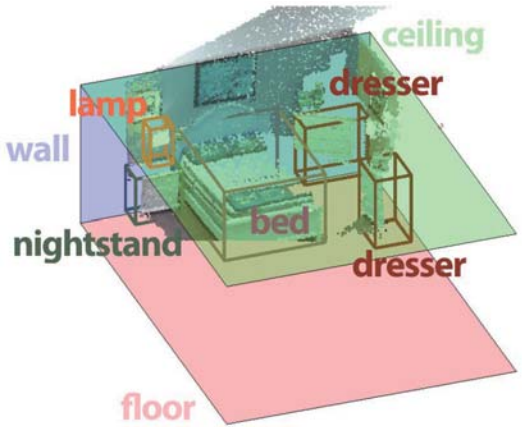
Geometry



Pixelwise



Semantic

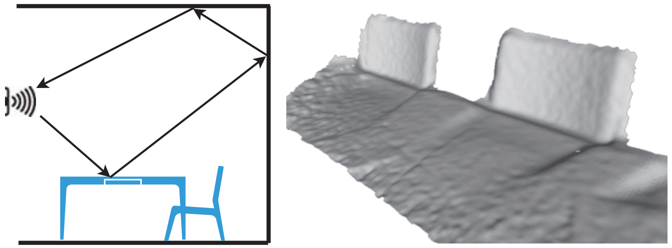


Holistic

Active Depth Sensing

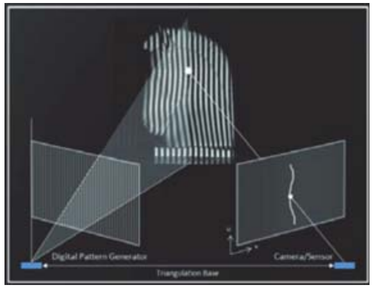
- Time of Flight
  - ✗ Fast motion
  - ✗ Multi-Path

Time of Flight

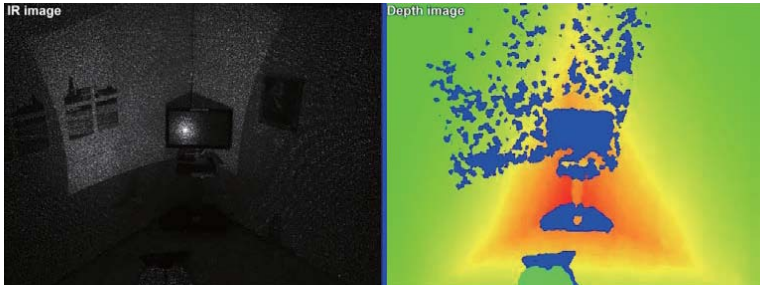


Multi-Path Interference

- Structured Light
  - ✗ Calibration
  - ✗ Multi-Device



Structured Light



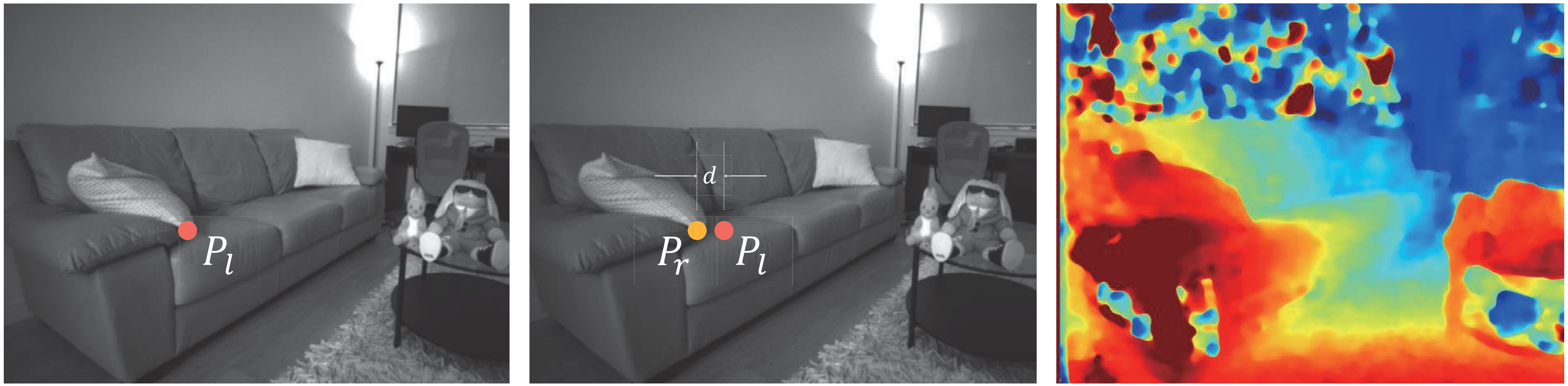
Multi-Device Interference

Passive Depth Sensing

- Stereo Matching
  - ✗ Texture-less Region

$$Z = \frac{bf}{d}$$

b: distance between camera centers
f: camera focal length
d: disparity
Z: depth



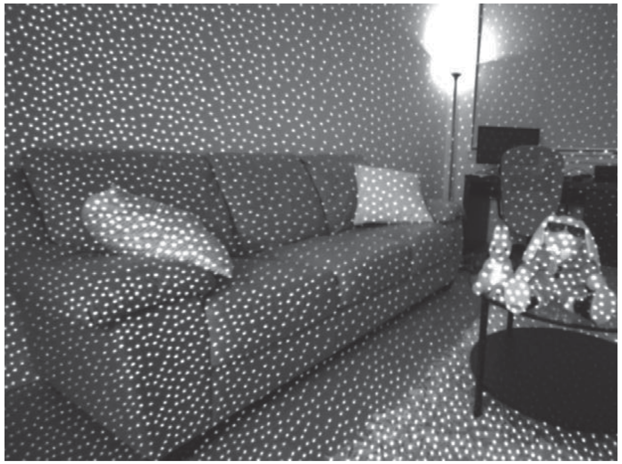
Left View   Right View   Disparity

Active Stereo System

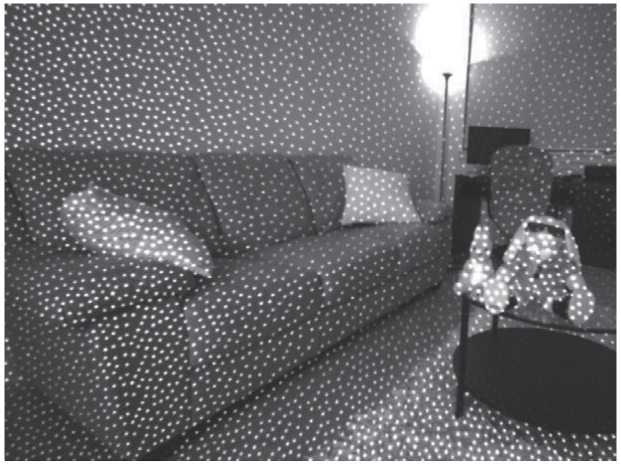
- Stereo Matching
  - ✗ ~~Texture-less Region~~

$$Z = \frac{bf}{d}$$

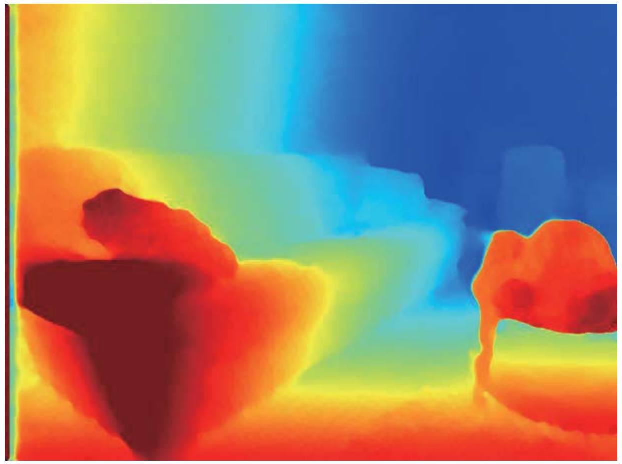
b: *distance between camera centers*
f: *camera focal length*
d: *disparity*
Z: *depth*



Left View

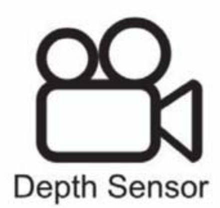


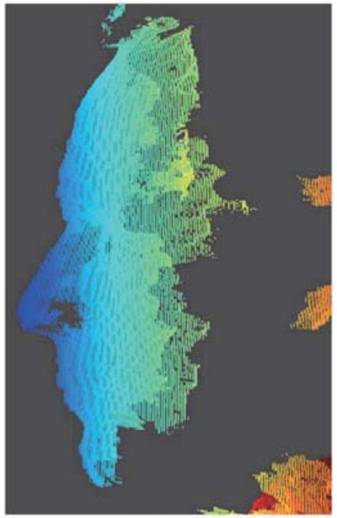
Right View



Disparity

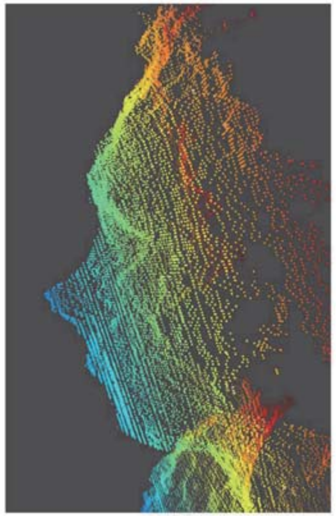
Active Stereo System





300mm

500mm



700mm

2500mm

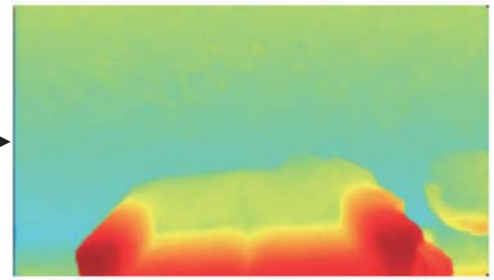
Use deep learning!

No ground truth…

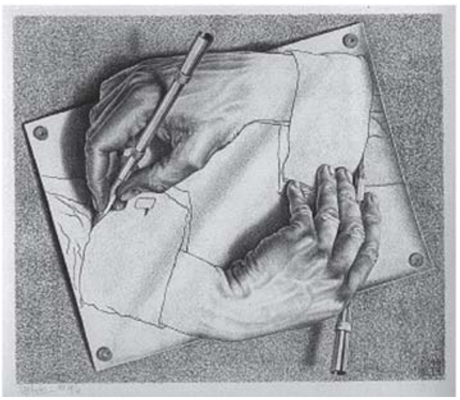
ActiveStereoNet

Input: Left/Right View

End-to-End System



Output: Disparity

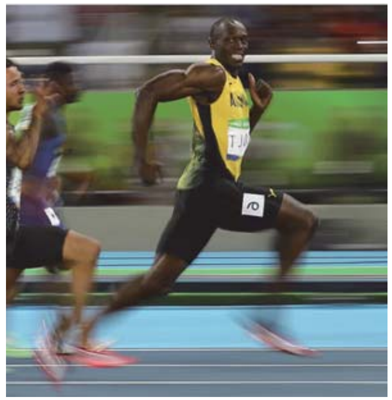


Self-supervised Learning

=

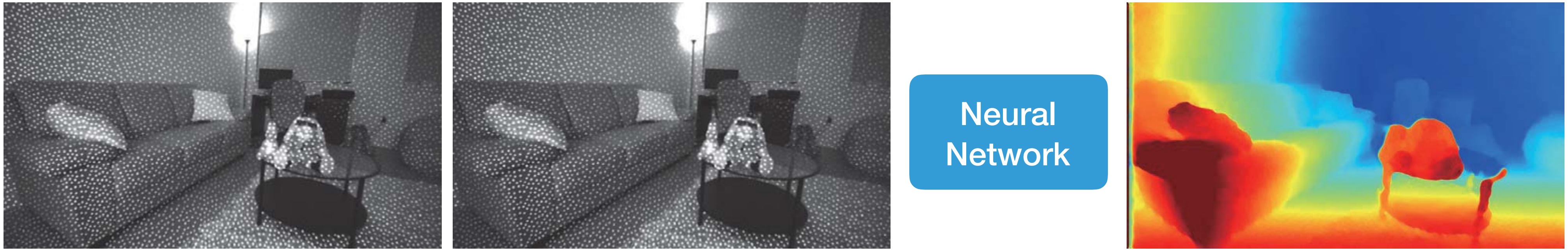
Annotation

Supervision



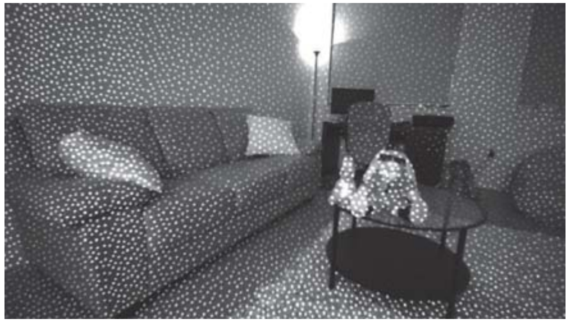
Just keep running…

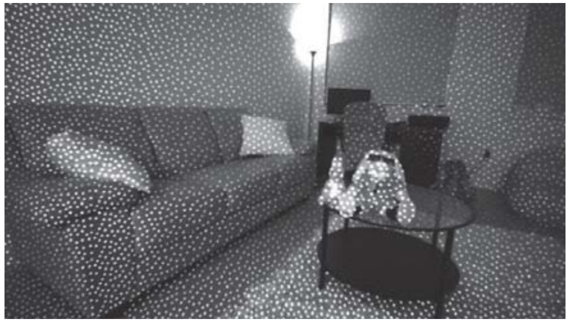
Self-Supervised Learning



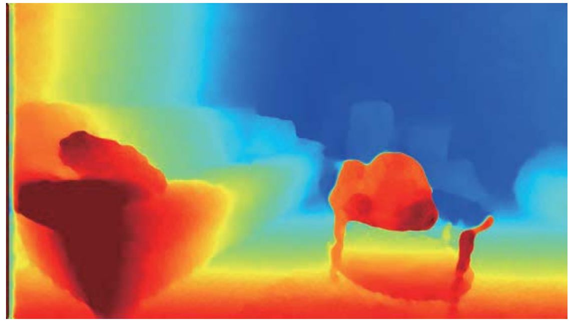
Left View Right View Estimated Disparity

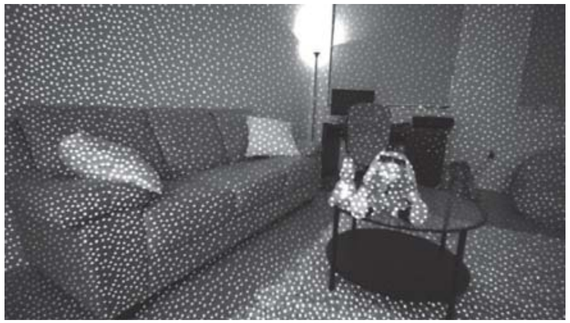
Self-Supervised Learning

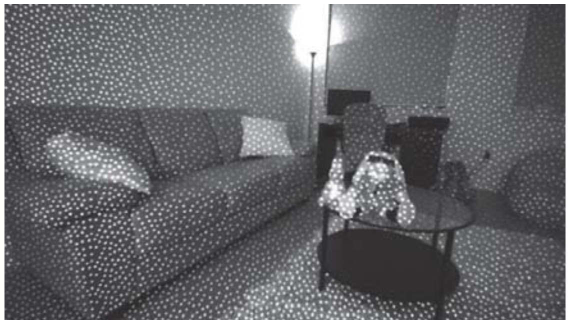

Left View


Right View

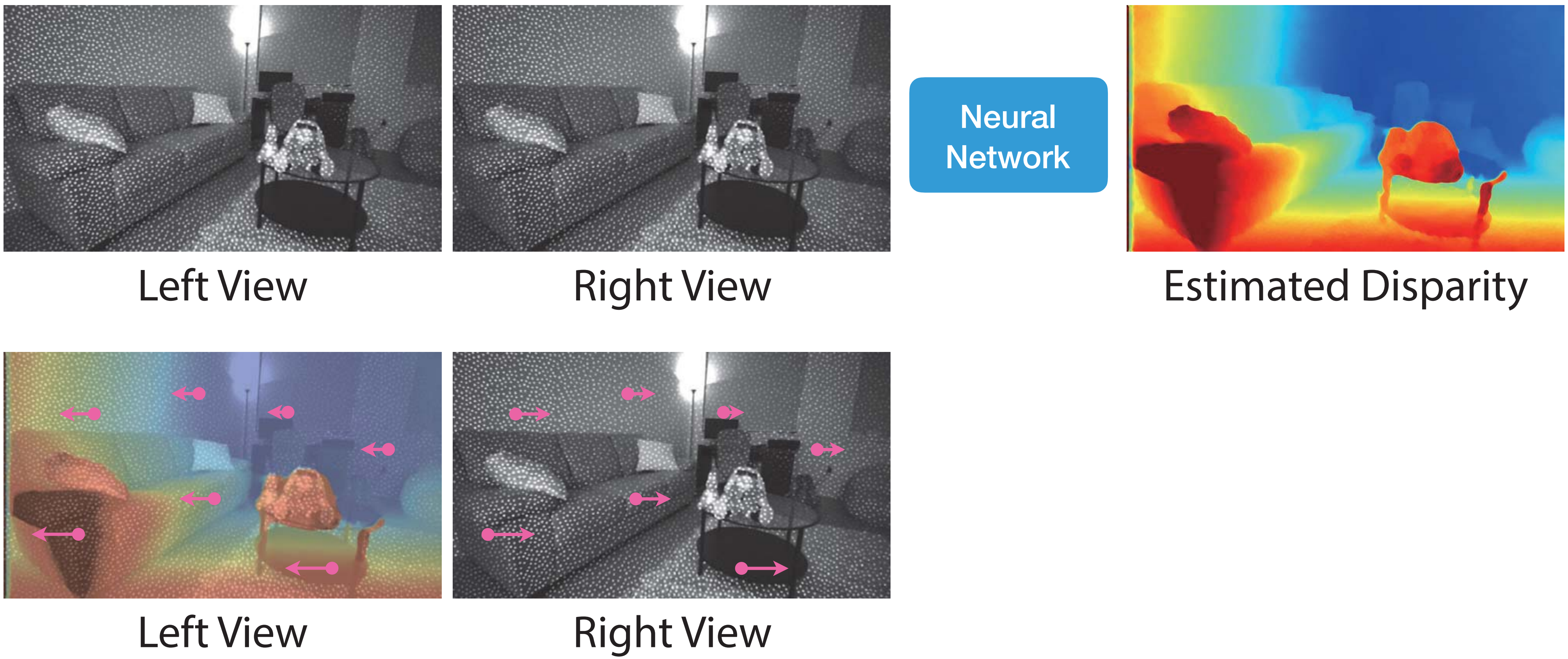
Neural Network


Estimated Disparity

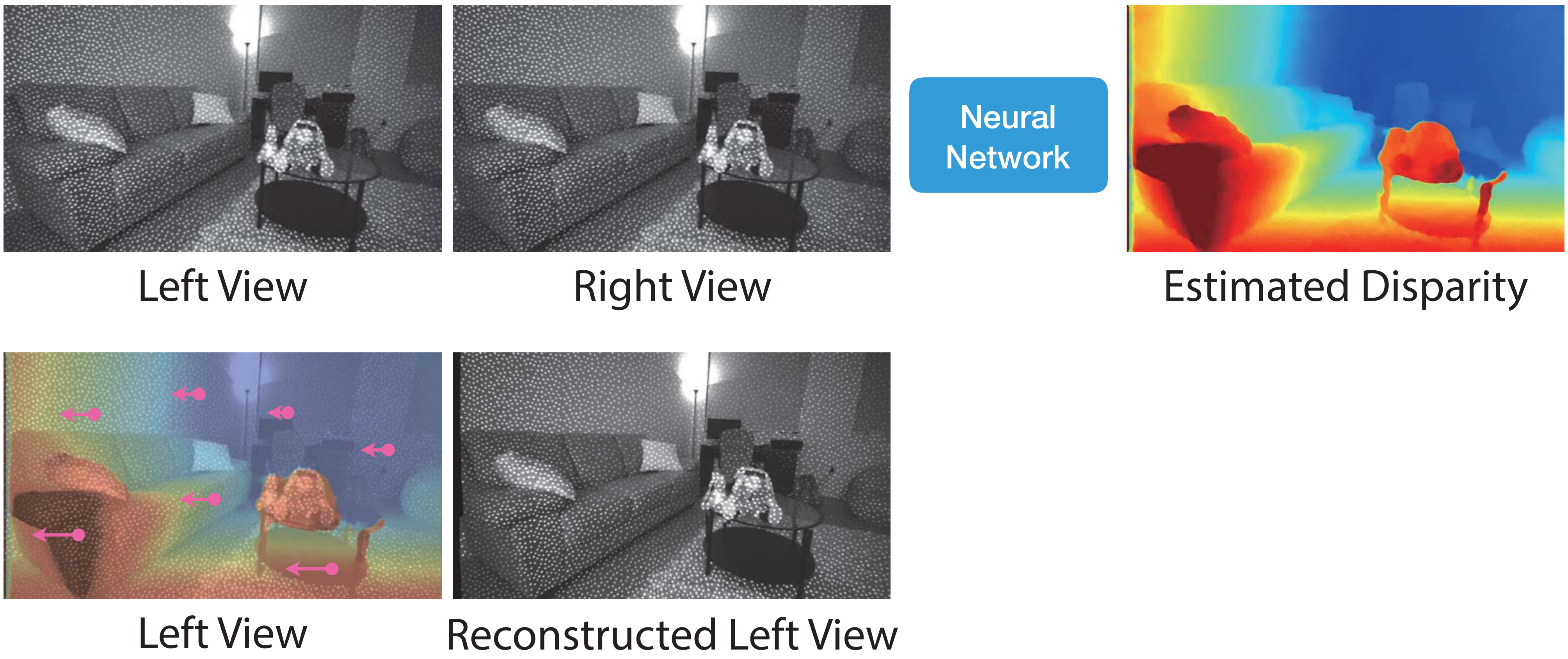

Left View


Right View

Self-Supervised Learning

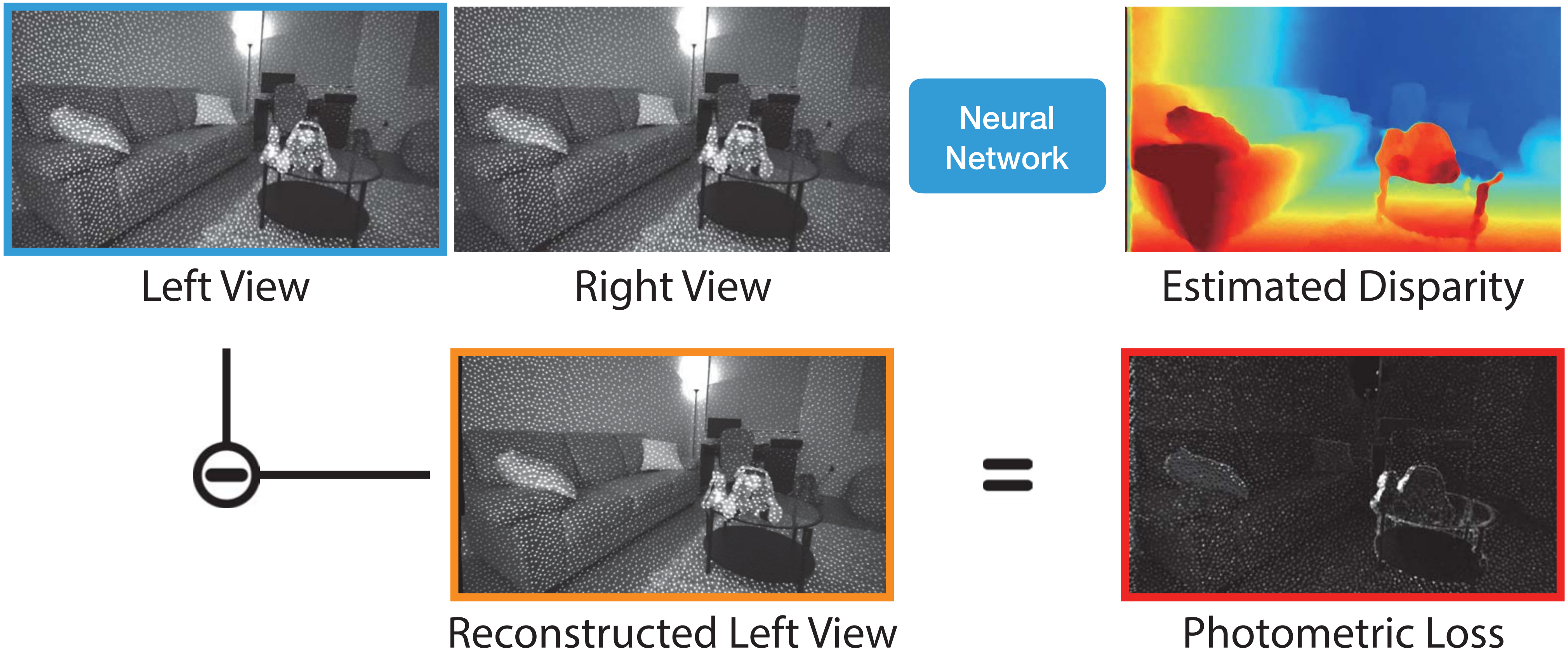


Self-Supervised Learning



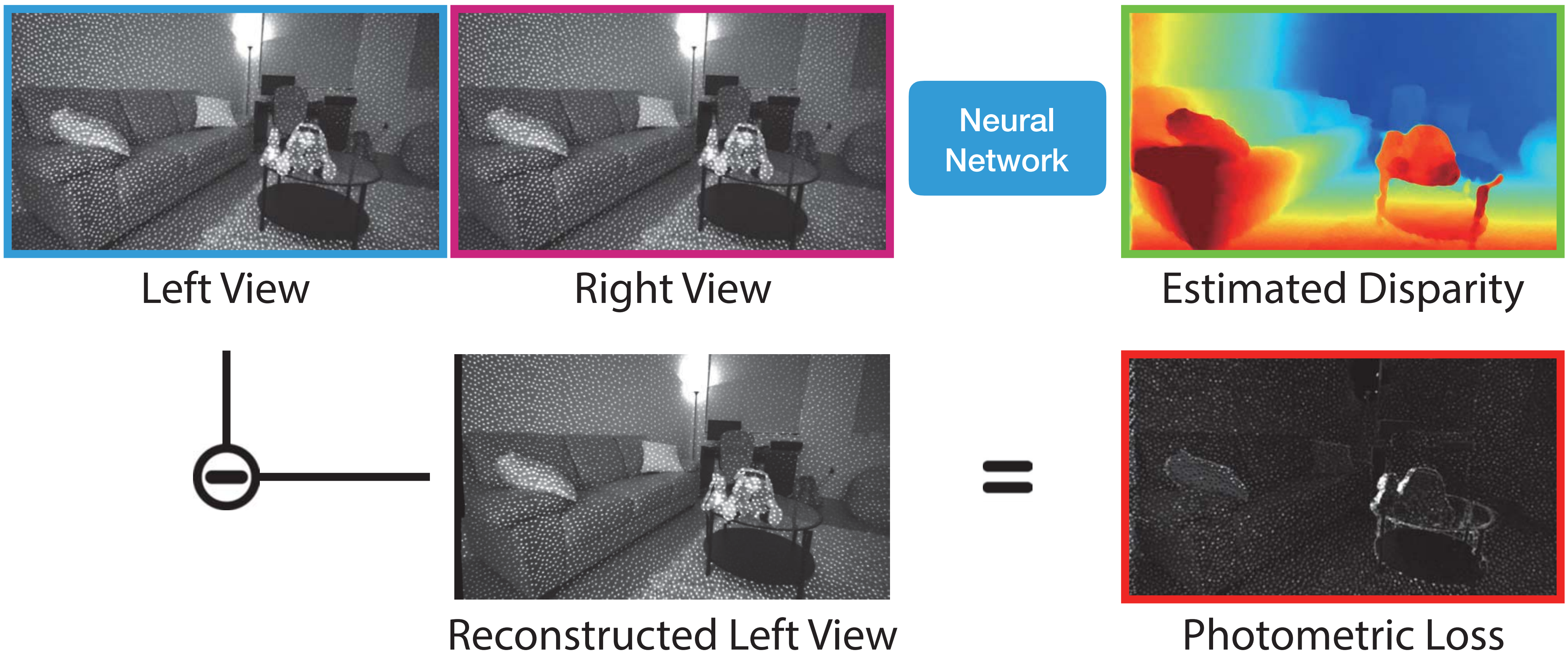
Self-Supervised Learning

Photometric Loss = | Left View - Reconstructed Left View |



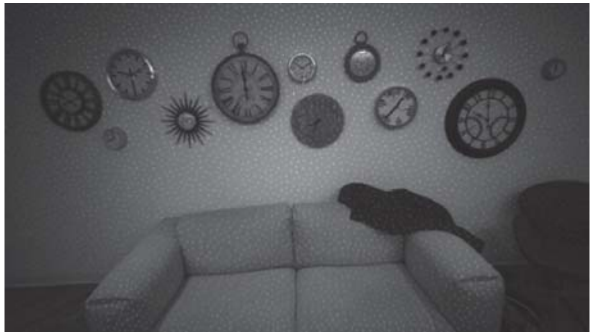
Self-Supervised Learning

Photometric Loss = | Left View - Warping(Right View, Left Disparity) |



Experiments

Experiments



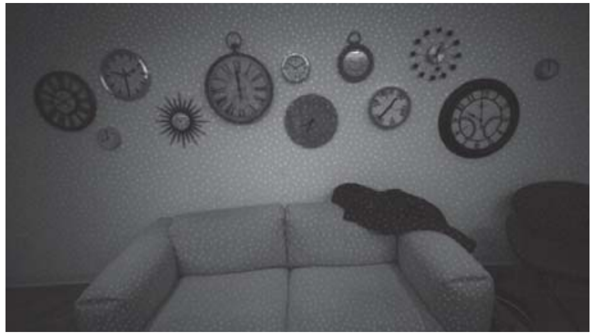
Left IR Image

Intel RealSense D435

IR Stereo Camera

IR Projector

Color Camera

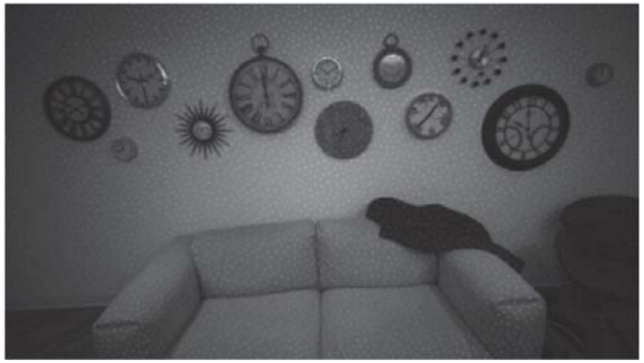


Right IR Image

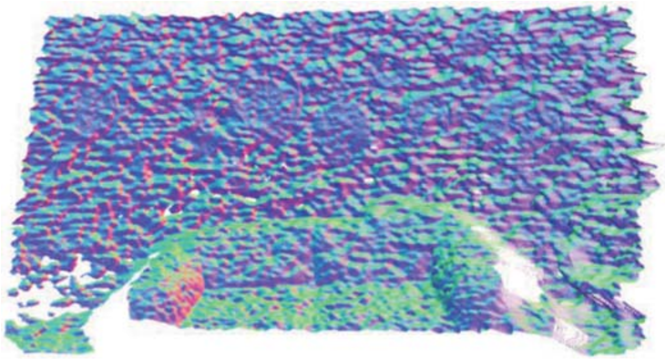
Color Image

Experiments

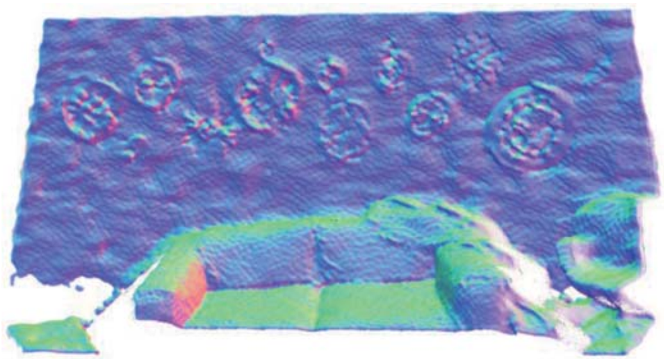
Intel RealSense D435



Left View



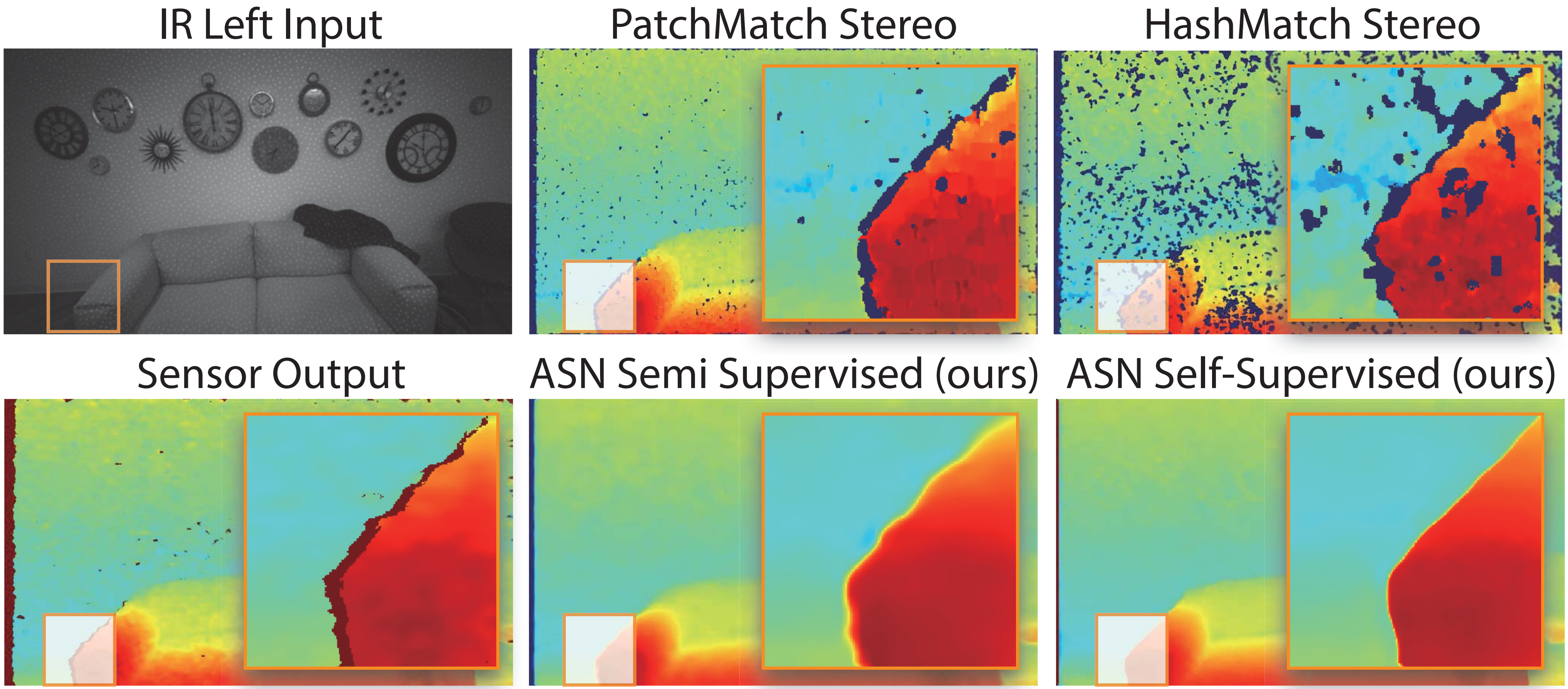
Intel RealSense D435



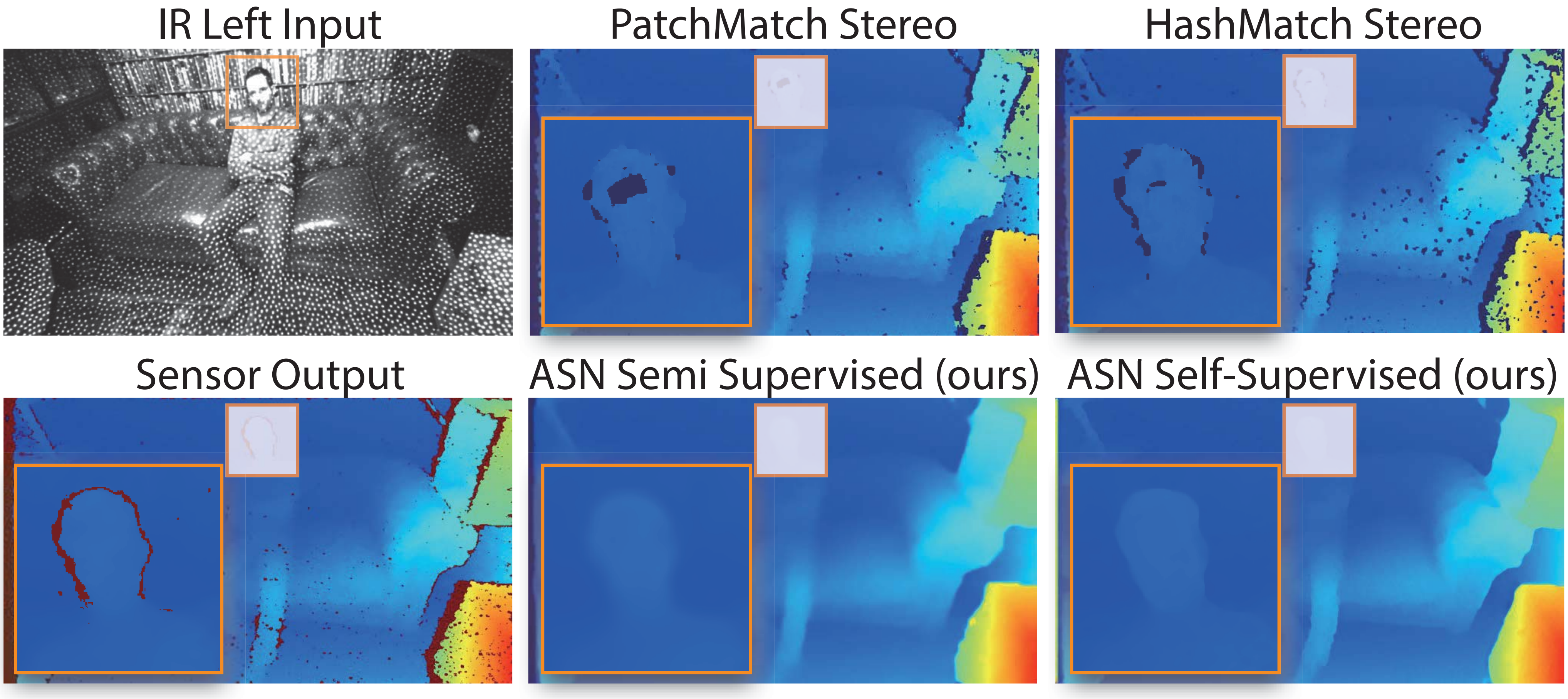
ActiveStereoNet

Color Image

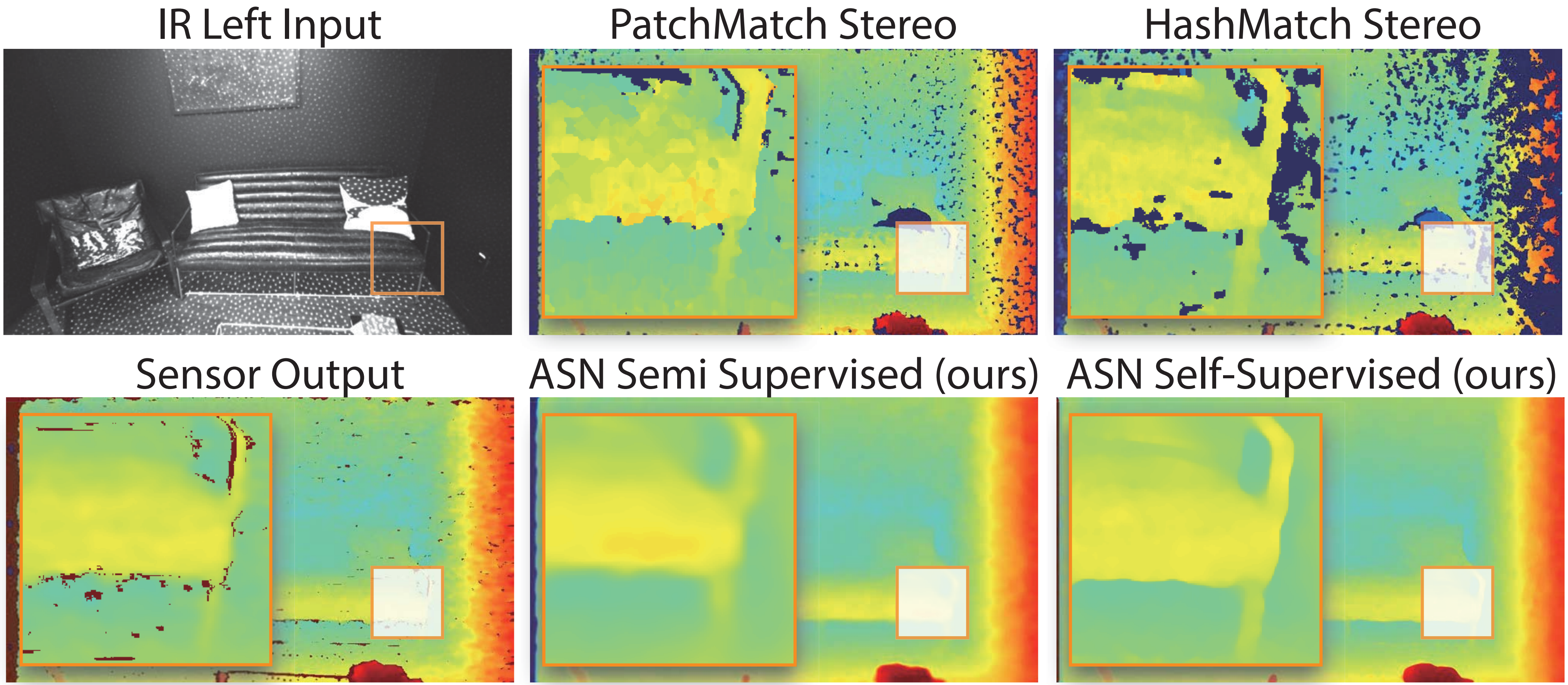
Disparity Qualitative Result



Disparity Qualitative Result



Disparity Qualitative Result



Disparity Quantitative Result

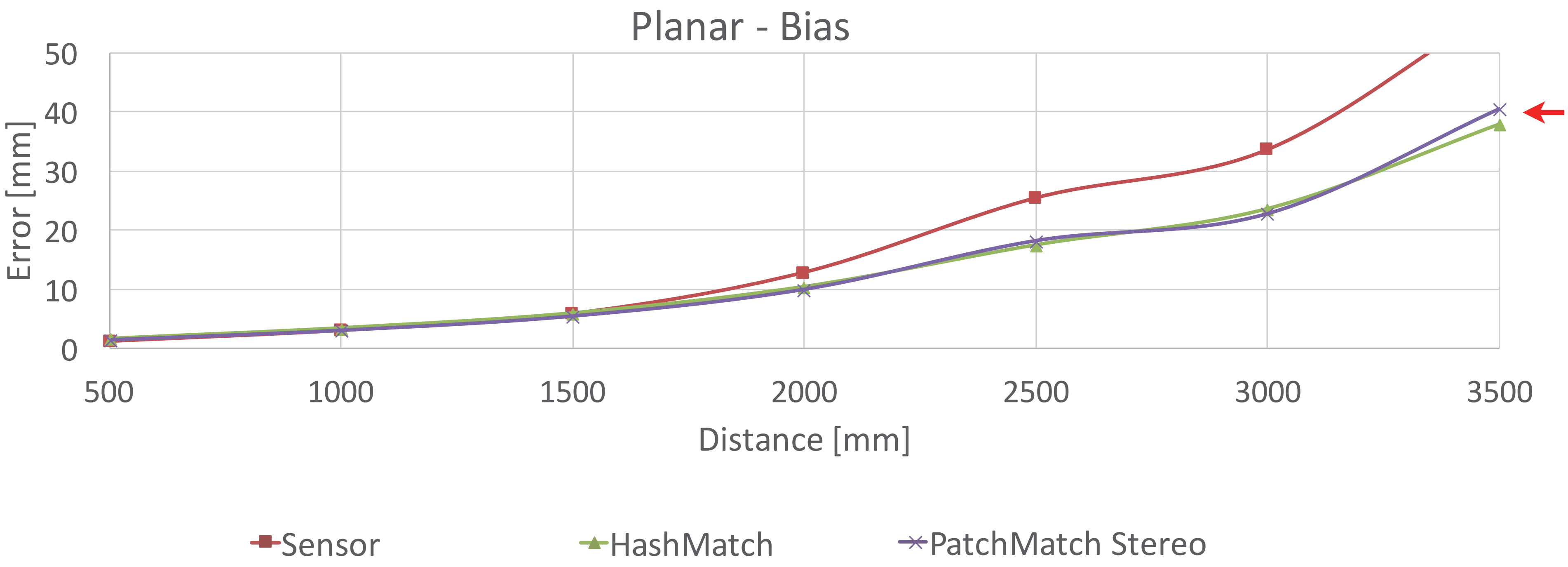
Fit plane on planar-wise scene as ground truth.

Disparity Quantitative Result

Sensor — Semi-Global Matching

Disparity Quantitative Result

Traditional Methods

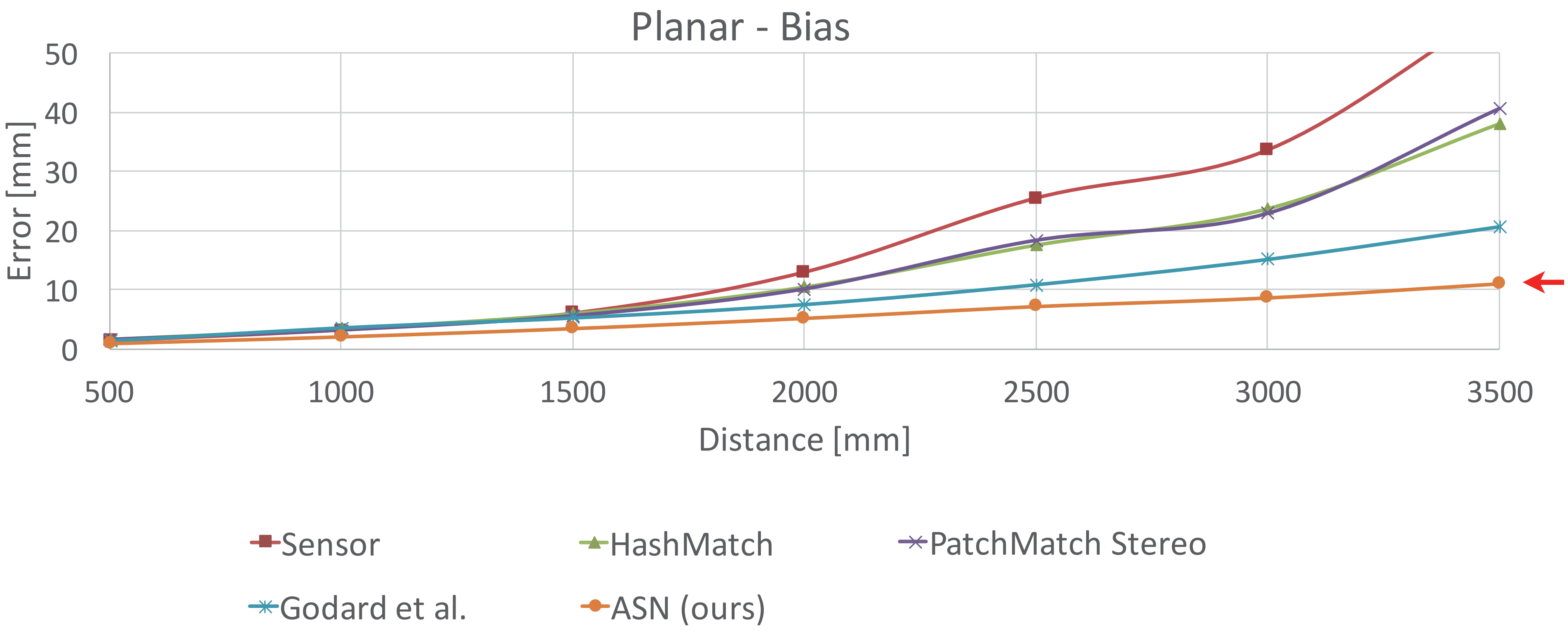


Disparity Quantitative Result

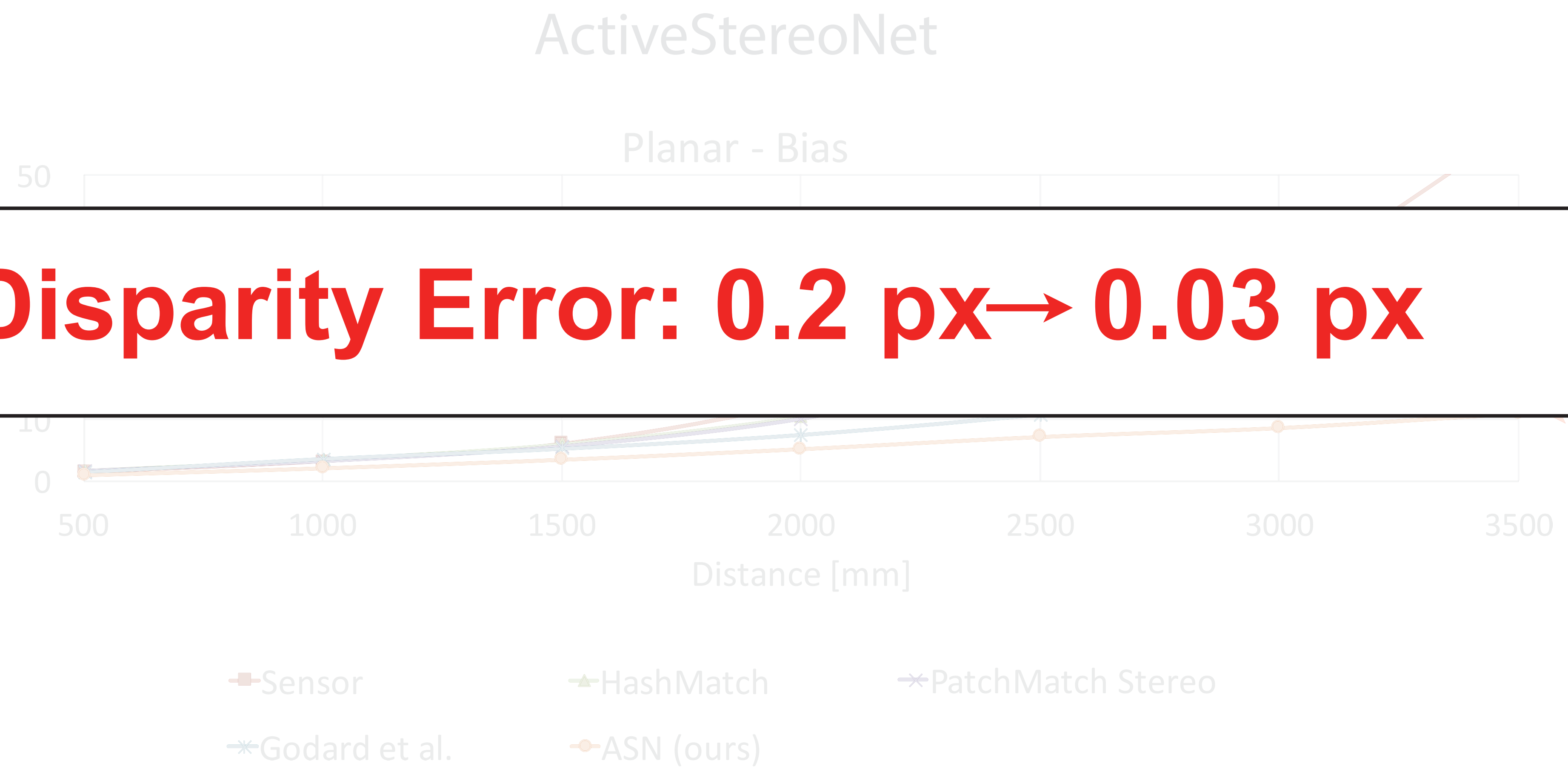
Previous Self-Supervised Method

Disparity Quantitative Result

ActiveStereoNet



Disparity Quantitative Result



Disparity Error: 0.2 px→ 0.03 px

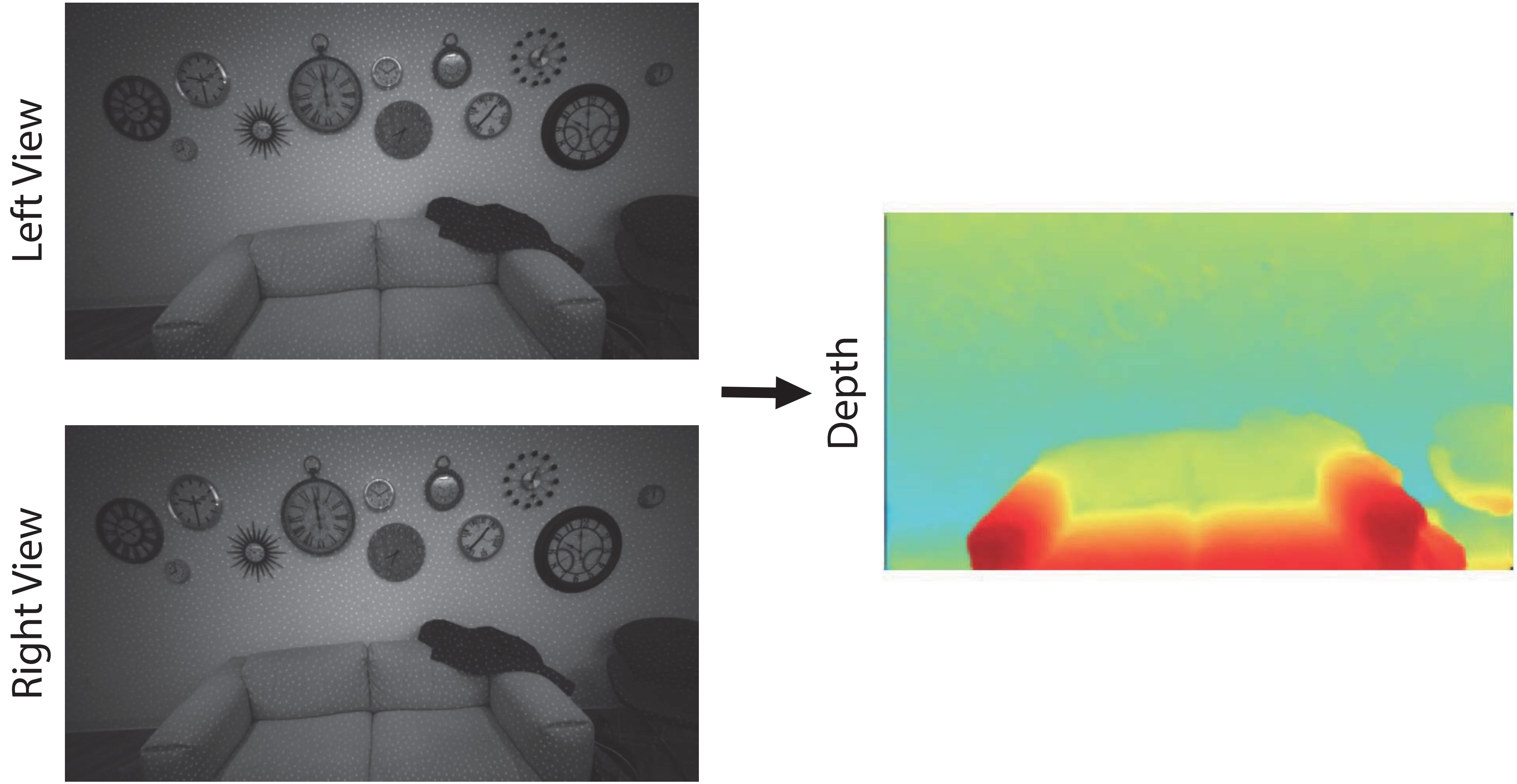
Active Stereo Net: End-to-End Self-Supervised Learning for Active Stereo Systems

Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, Sean Fanello

ECCV 2018

Project Webpage:
http://asn.cs.princeton.edu/

Pixel-wise Depth Estimation



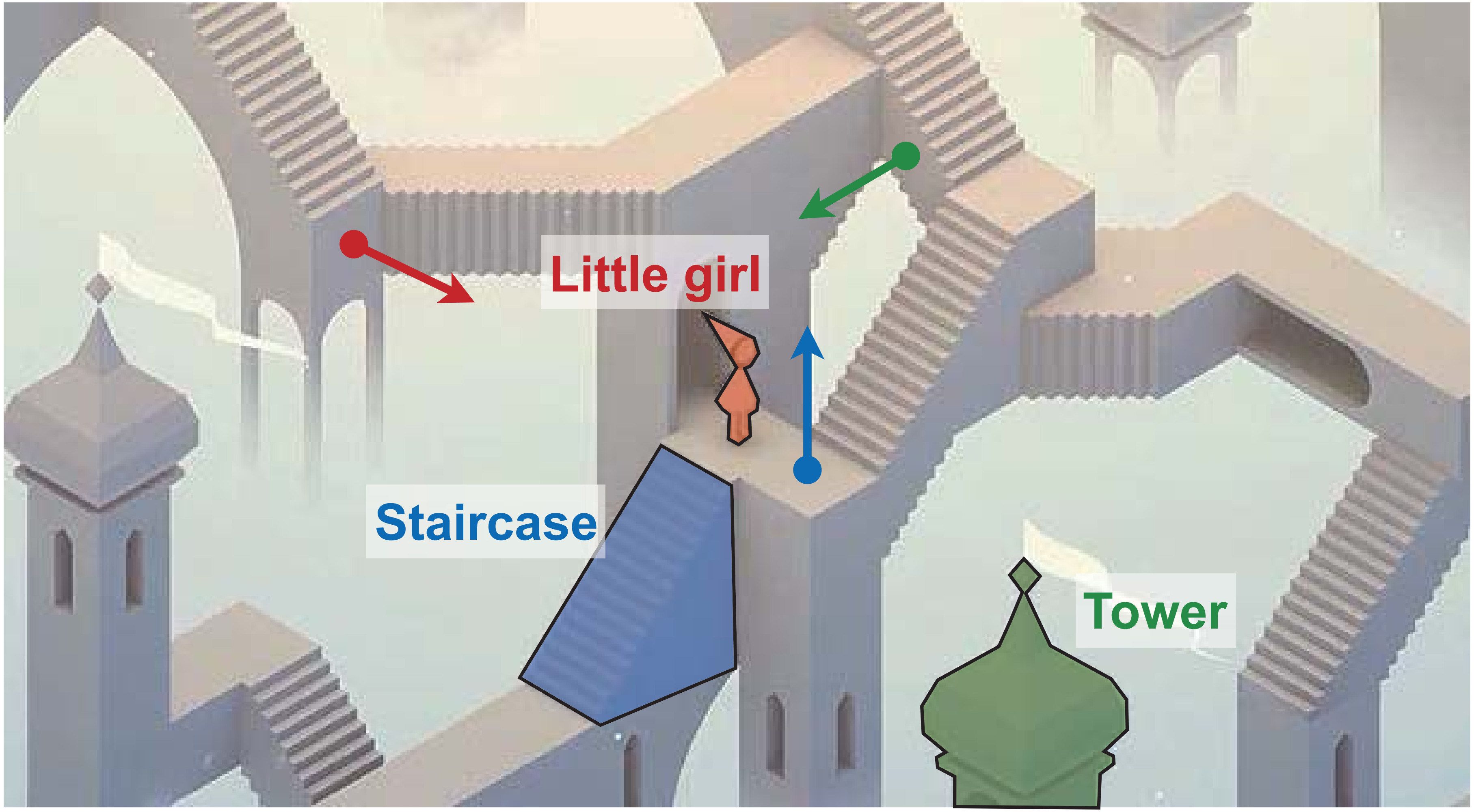
What if only one eye?

Left View

Right View

What if only one eye?

…in Monument Valley



Understand from Single Image

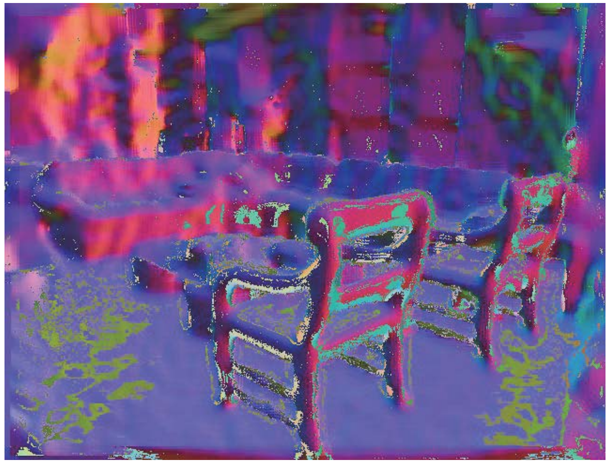
...in Reality

Color

Depth

Instance Boundary

Semantic Segmentation



Surface Normal

Data for Scene Understanding

	NYU Dataset
❖ Aligned color and gnd	1449
❖ Diversity	Limited
❖ Ground Truth	Noisy

Image | Raw Depth | Refined Depth | Normal Map

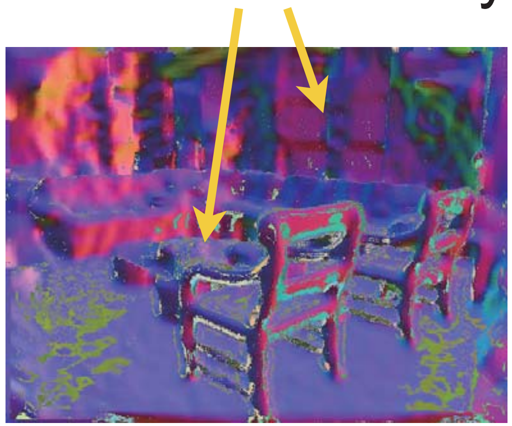
Data for Scene Understanding

	NYU Dataset	**Synthetic**
❖ Aligned color and gnd	1449	∞
❖ Diversity	Limited	∞
❖ Ground Truth	Noisy	Accurate

Image

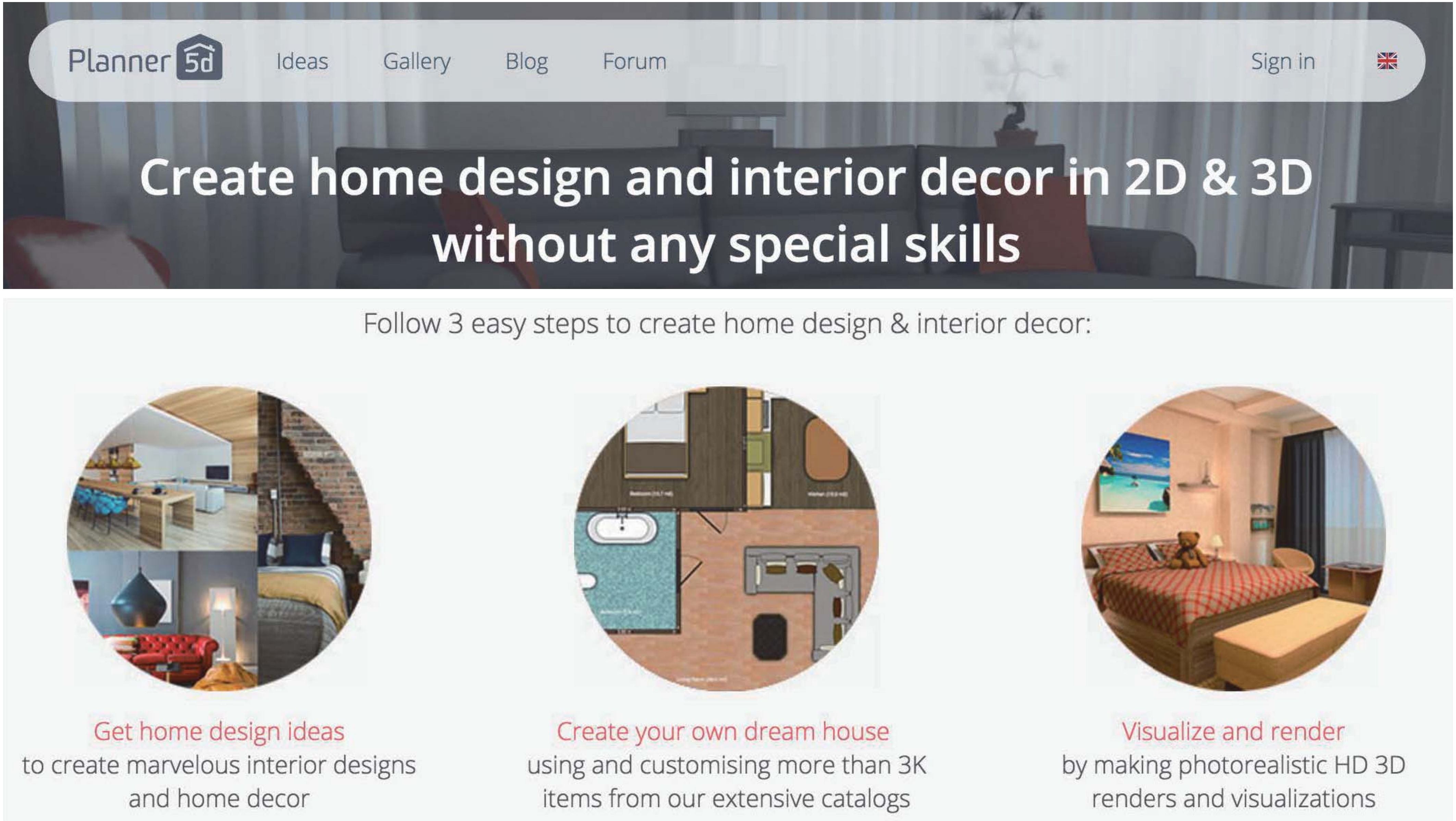
Raw Depth

Refined Depth



Normal Map

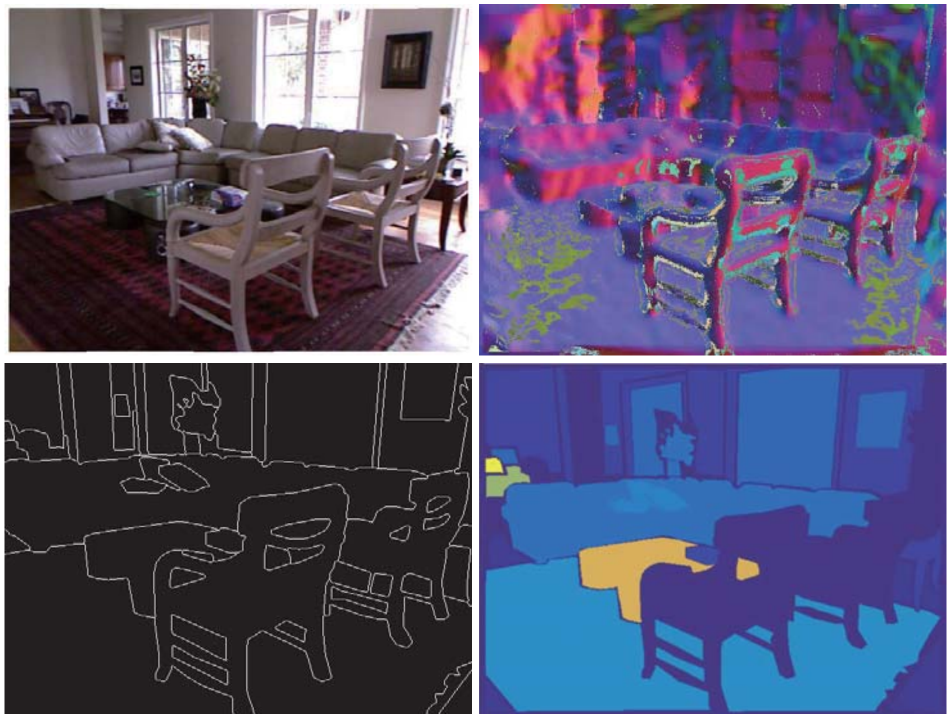
Synthetic Virtual Scene: SUNCG



Synthetic Data Generation

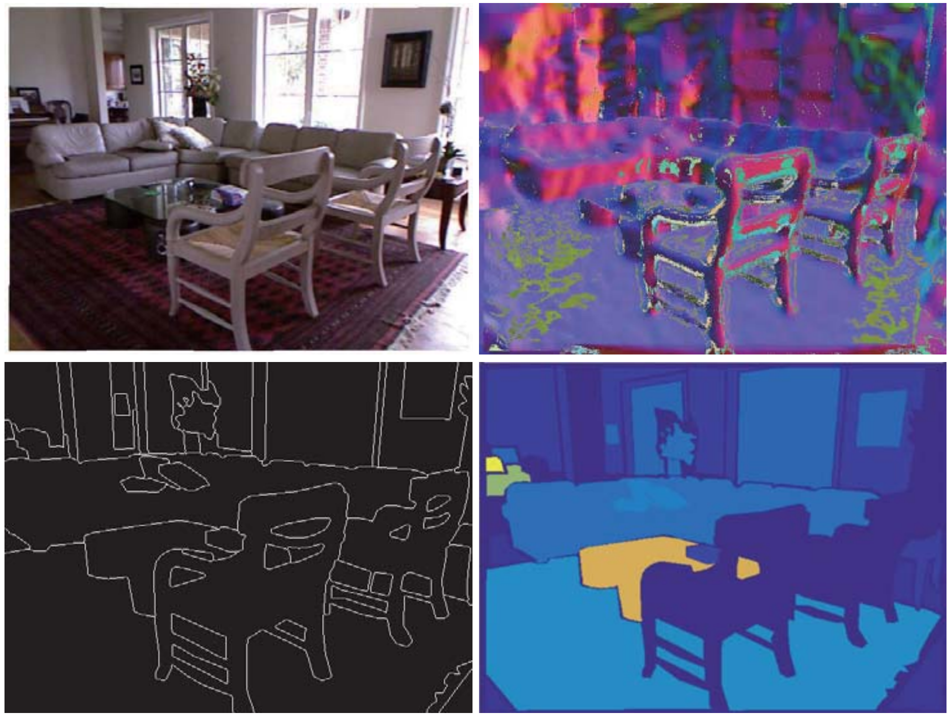
- **Source Domain**

- **Target Domain**



Synthetic Data Generation

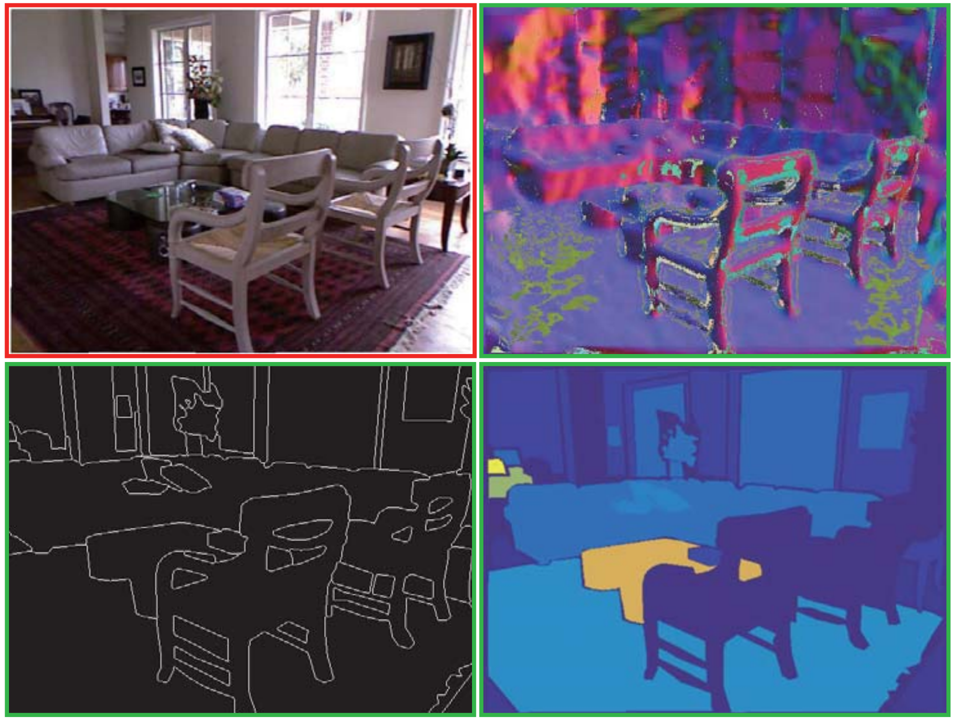
- **Real Data**



- **Synthetic Data**

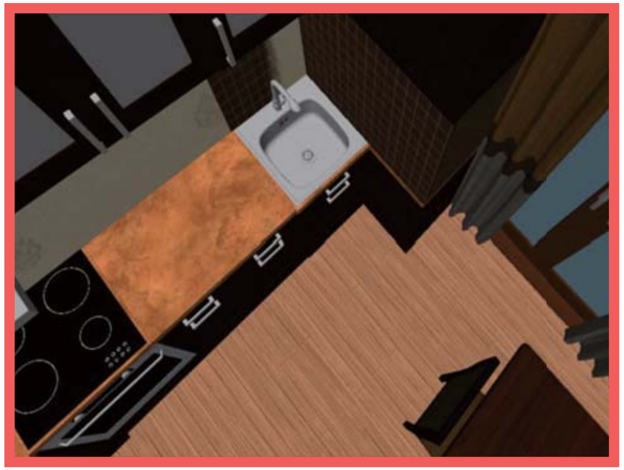
Synthetic Data Generation

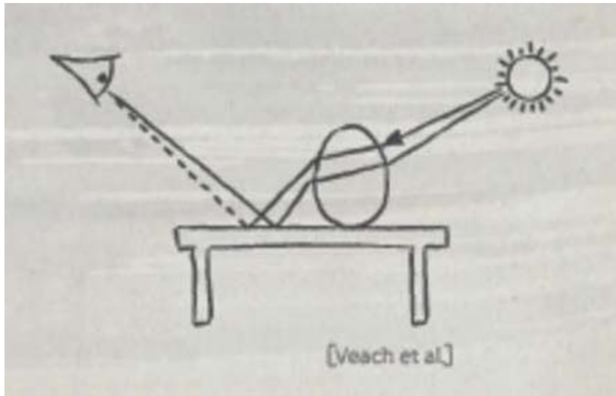
- **Real Data**
- **Synthetic Data**



Realistic Synthetic Data?

- **Viewpoint**
- **Physically based Rendering**
- **Illumination**





Mitsuba
PHYSICALLY BASED RENDERER

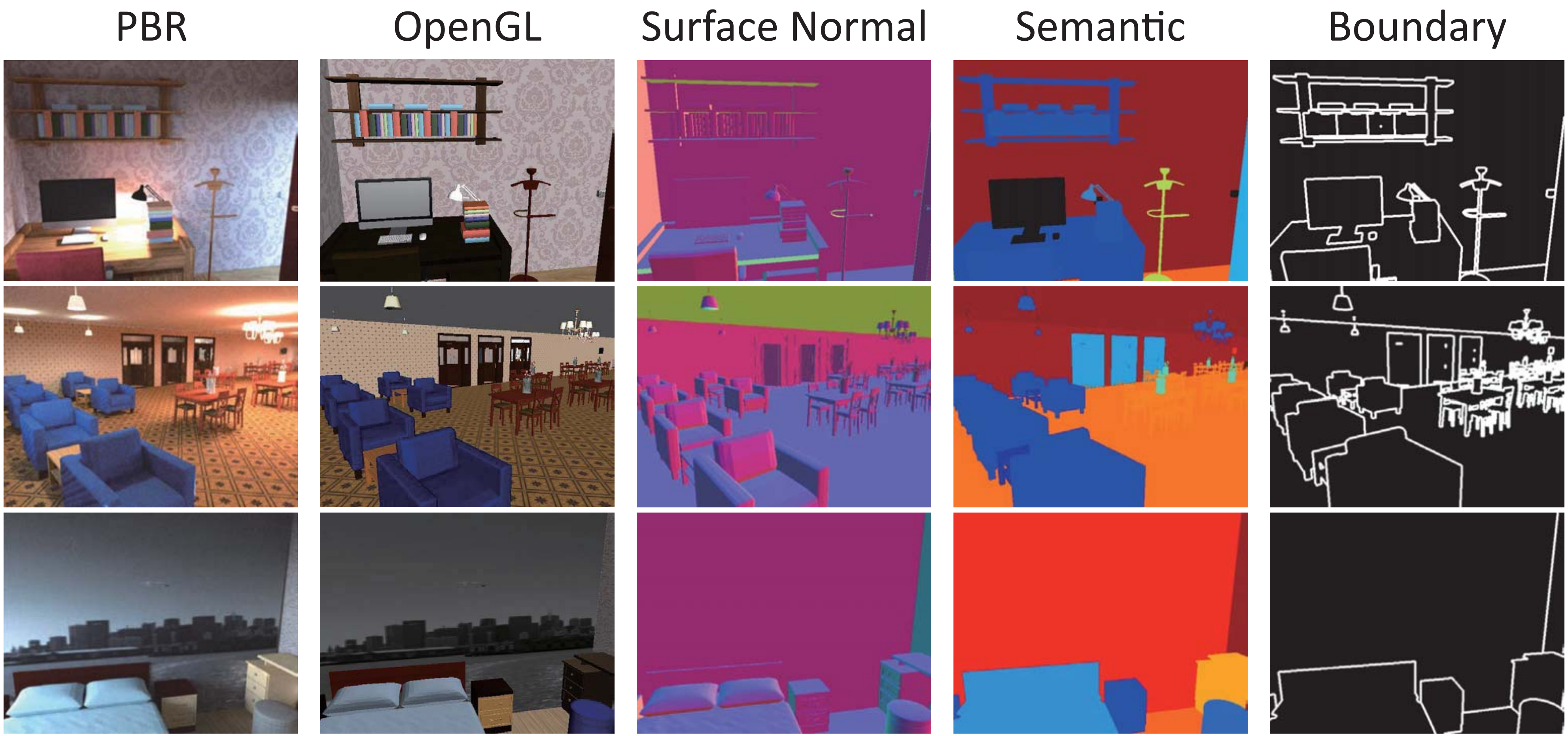
Realistic Synthetic Data?

OpenGL Rasterization

Physically Based Rendering

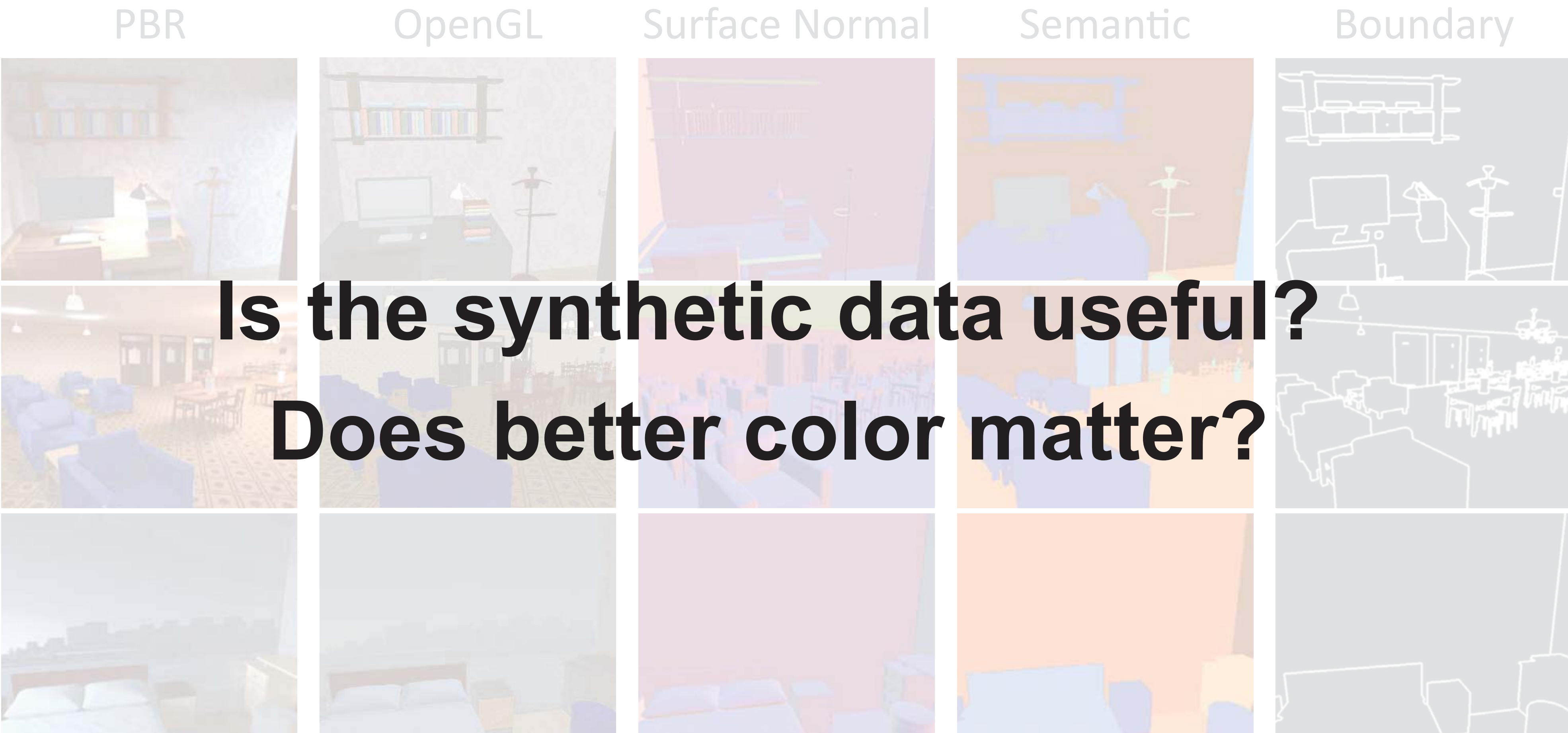
Realistic Synthetic Data!

- Speed: 30s per camera
- # Image: ~500,000



Realistic Synthetic Data!

- Speed: 30s per camera
- # Image: ~500,000



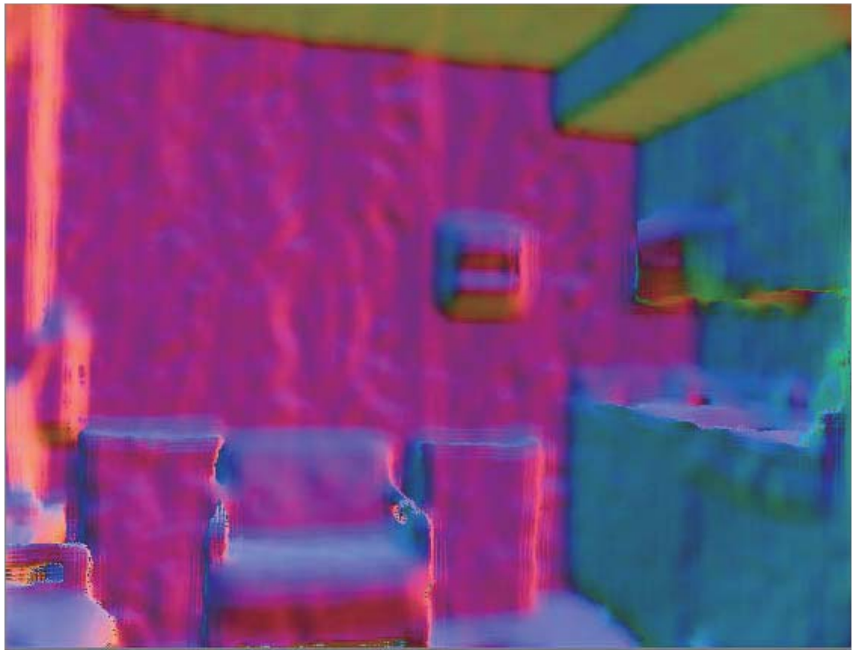
Is the synthetic data useful?

Does better color matter?

Normal Estimation

- Evaluation Metric:
  - Mean angle error

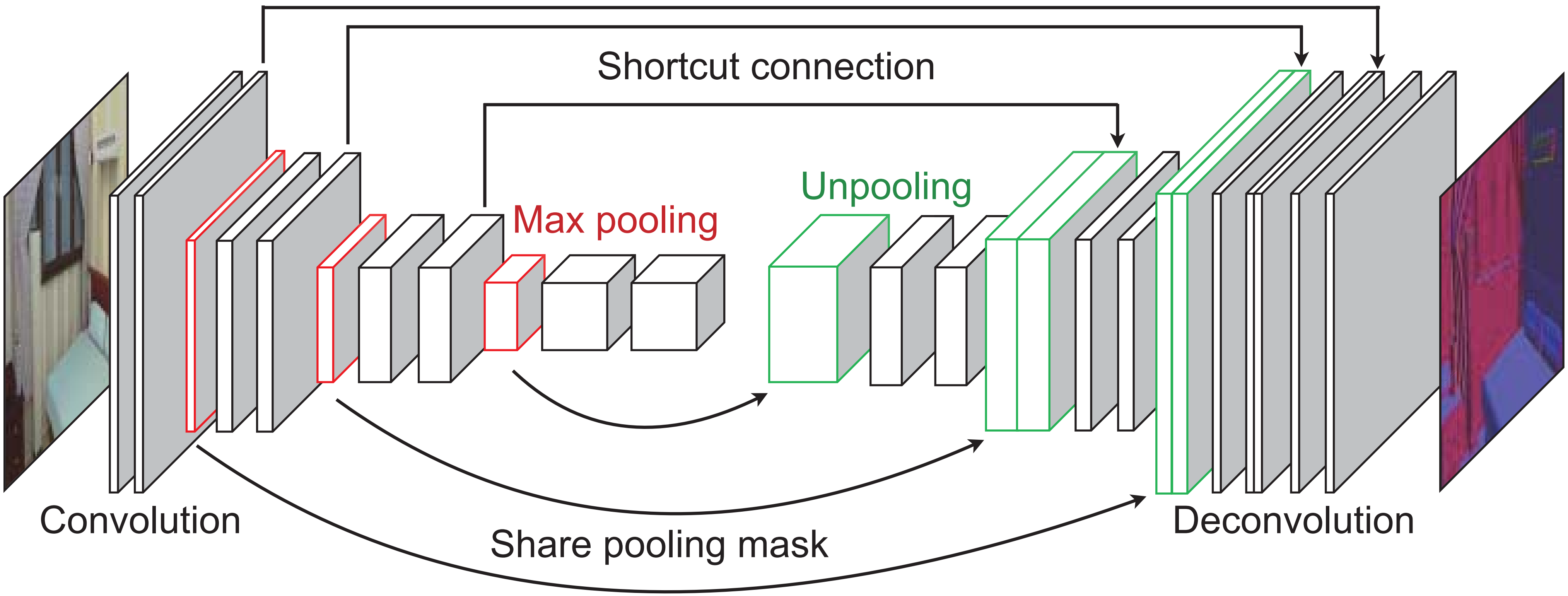
Color



Surface Normal

Normal Architecture

- Fully Convolutional Neural Network

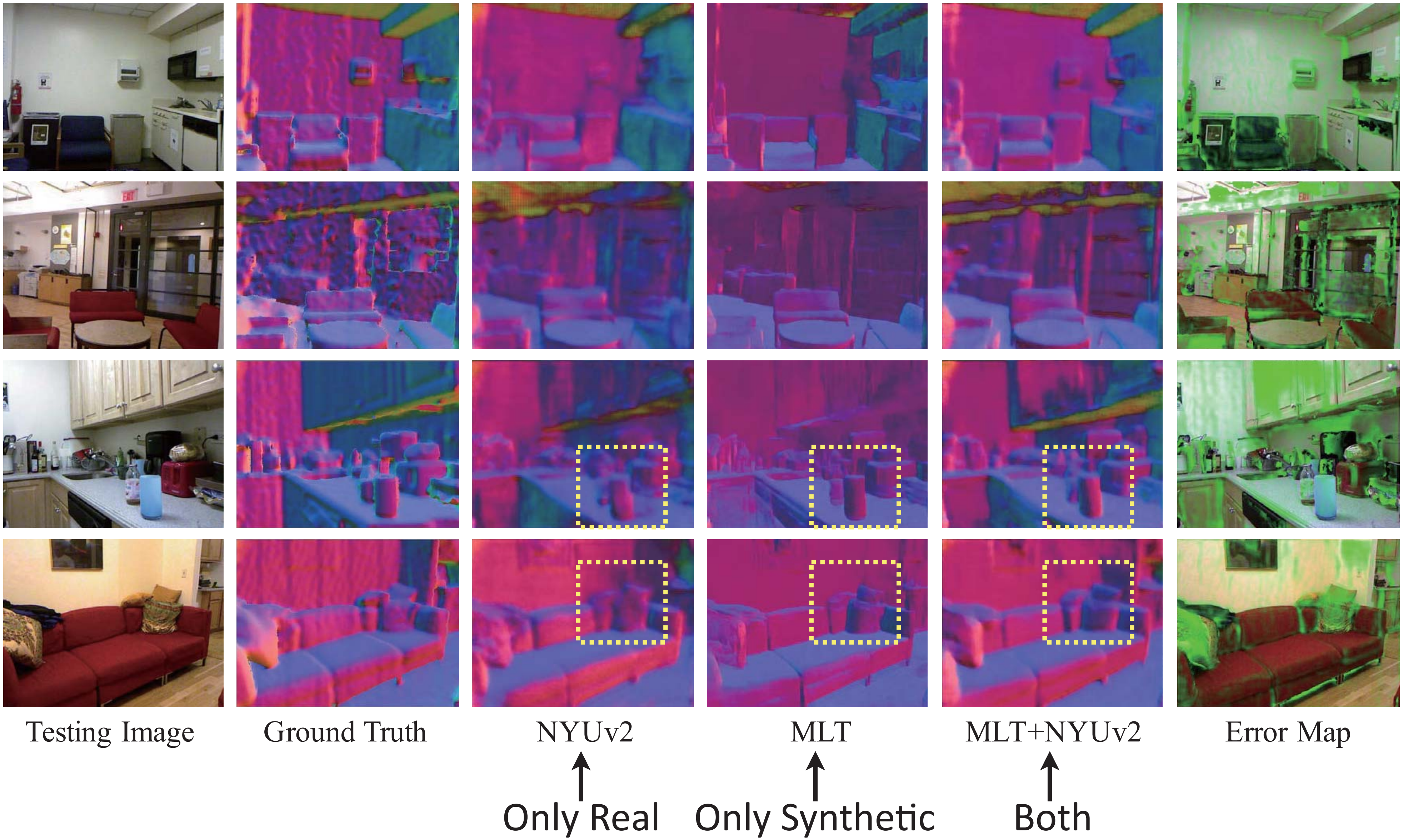


Quantitative Evaluation

- Evaluation Metric: mean angular error

❖ Pre-Train	❖ Fine-tune	❖ Test	❖ Evaluation
N/A	NYUv2	NYUv2	27.30
OpenGL	N/A	NYUv2	33.06
Mitsuba	N/A	NYUv2	27.90
OpenGL	NYUv2	NYUv2	23.38
Mitsuba	NYUv2	NYUv2	21.74

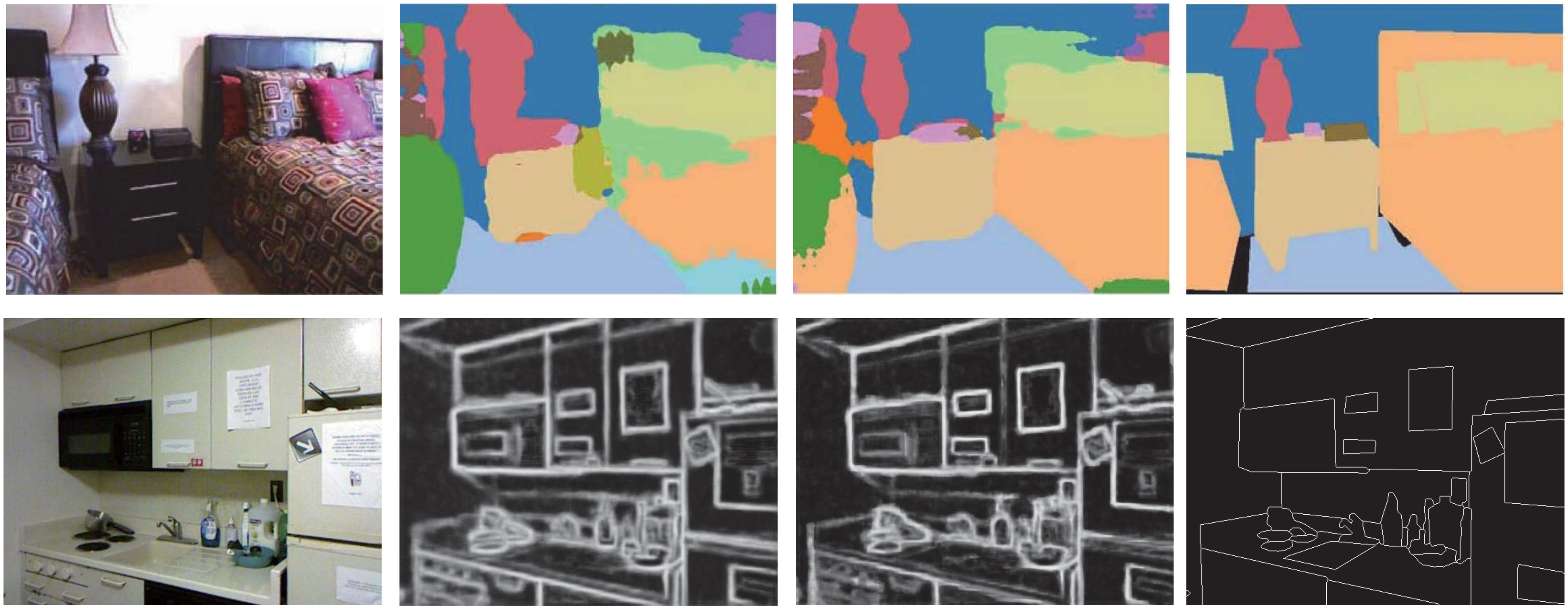
Qualitative Results



Other Tasks

- Semantic Segmentation — Accuracy: 31.7 → 33.2
- Instance Boundary Detection — Accuracy: 71.3 → 72.5

Color Input	w/o Syn. Data	w/ Syn. Data	Ground Truth



Physically-Based Rendering for Indoor Scene Understanding Using Convolutional Neural Networks

Y. Zhang, S. Song, E. Yumer, M. Savva,
J. Lee, H. Jin, T. Funkhouser.

CVPR 2017

Project Webpage:
http://pbrs.cs.princeton.edu

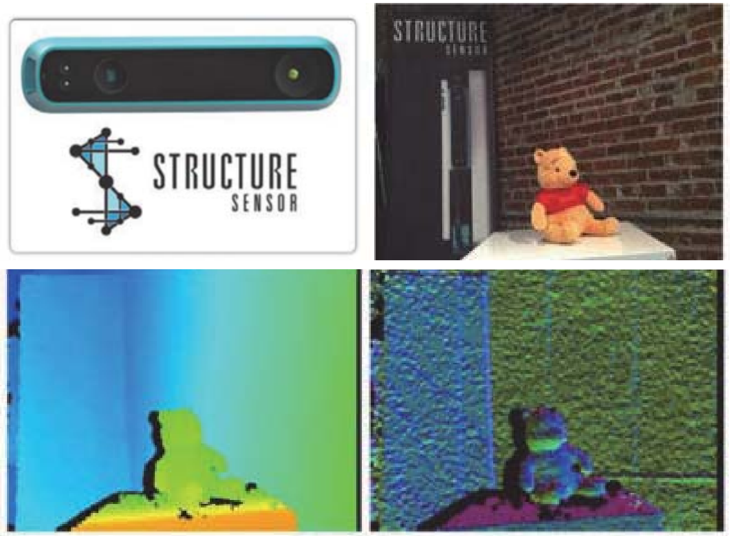
Dual v.s. Single

iPhone XS:
Powerful but Expensive

iPhone XR:
Compact and Cheap

Dual v.s. Single

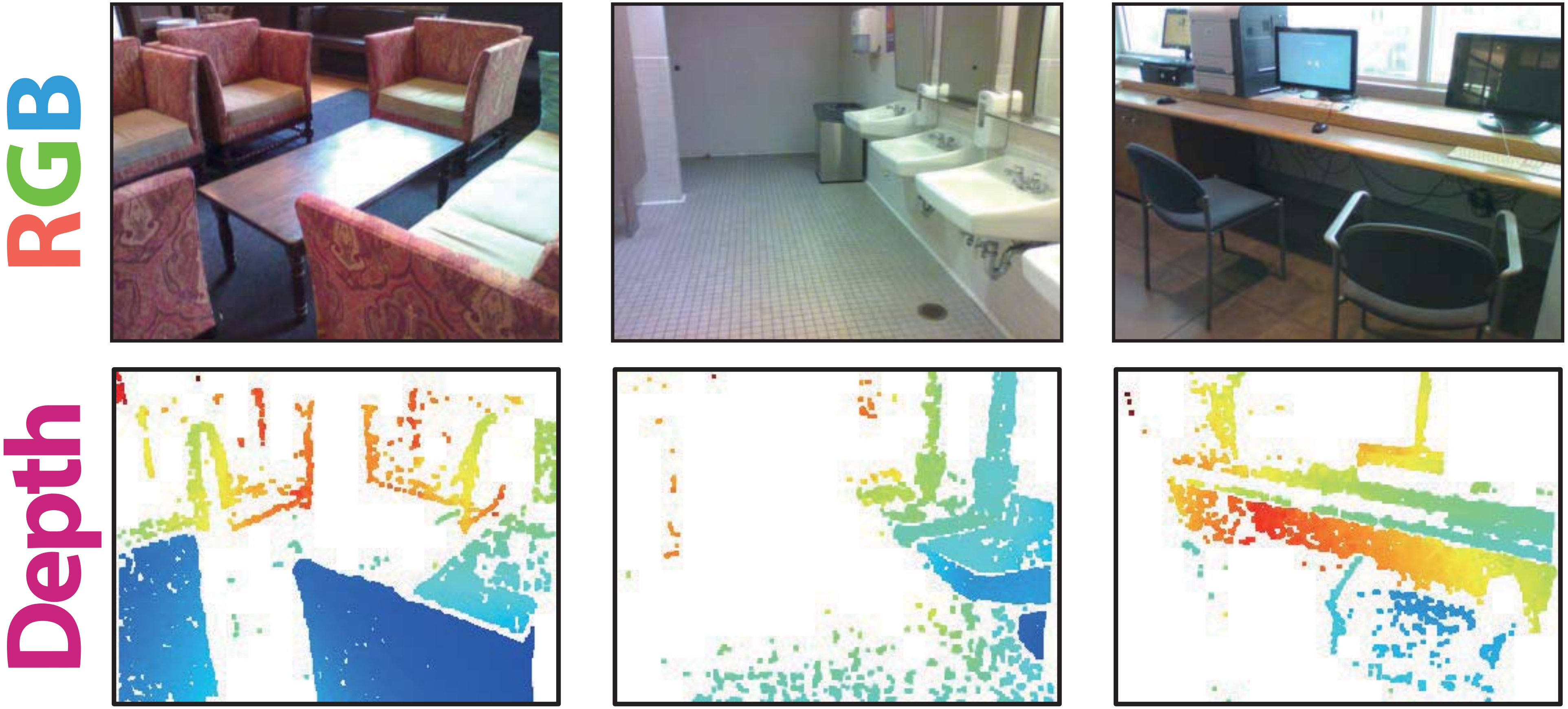
- Multiple Cameras
  - More information
  - Analytical solution
  - Reliable result
  - Expensive setting
- Single Cameras
  - Limited information
  - Data-driven, learn prior
  - Plausible result
  - Cheap and available



Color is beautiful!

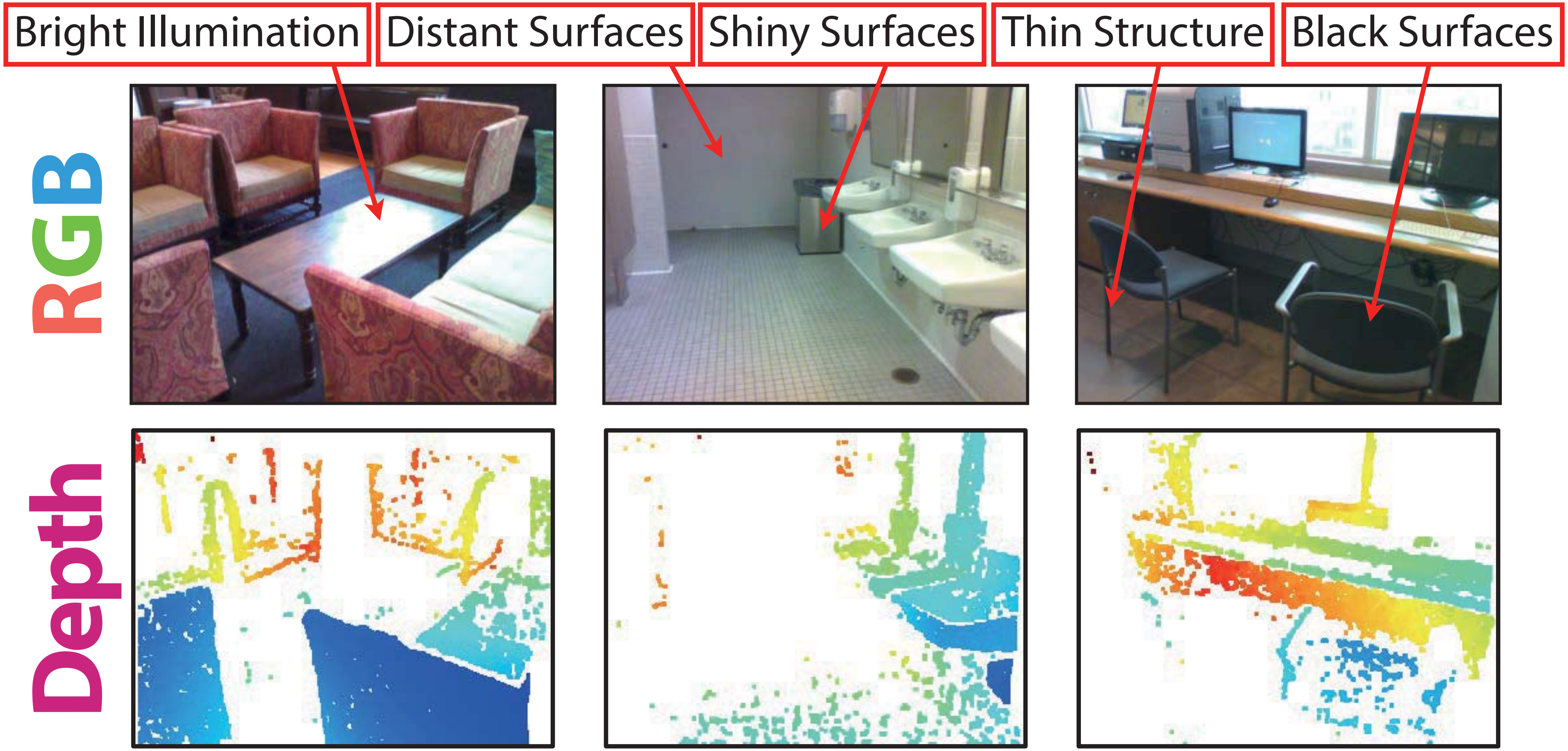
RGB

But… Depth is not so great!



From Intel Realsense R200

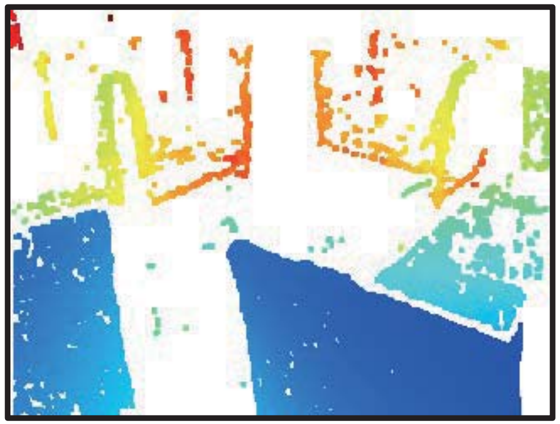
Why?



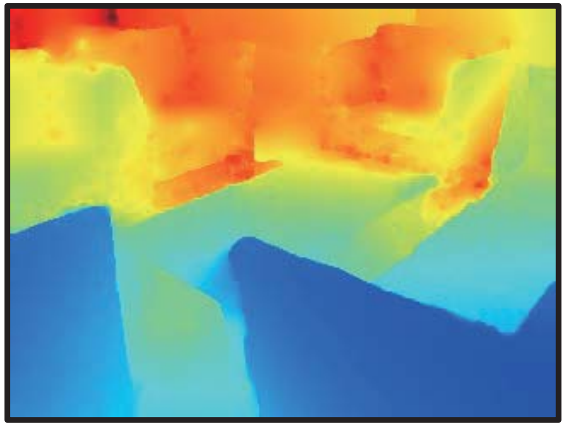
From Intel Realsense R200

Goal: Depth Completion

Input: Color Image



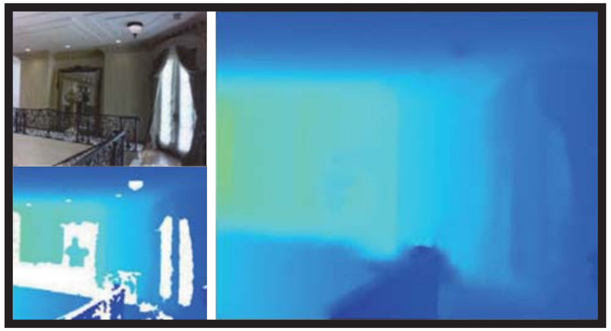
Input: Sensor Depth



Output: Completed Depth

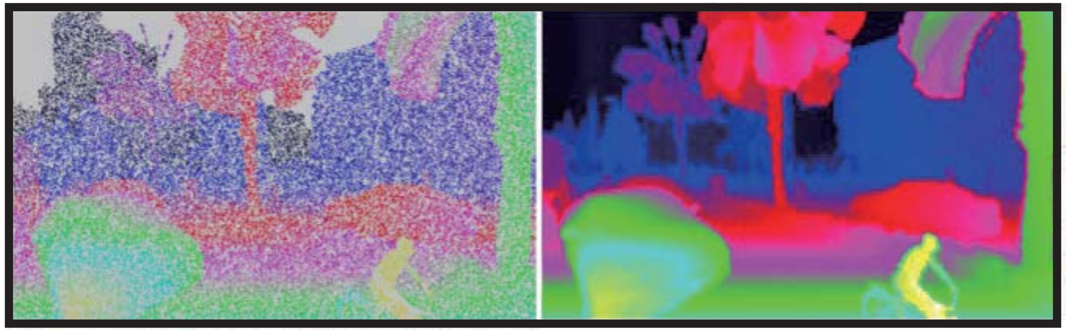
Related Work

- Depth Completion from RGB-D



Joint Bilateral Filter [Silberman, 2012]

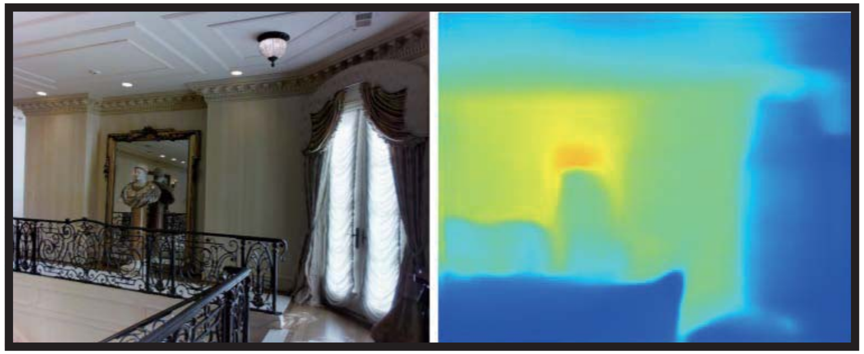
- Depth Upsampling from Sparse



Sparsity Invariant CNNs [Uhrig, 2017]

- Depth Estimation from RGB

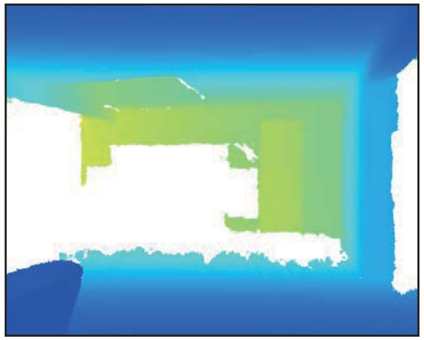
Deeper Depth Prediction [Laina, 2016]



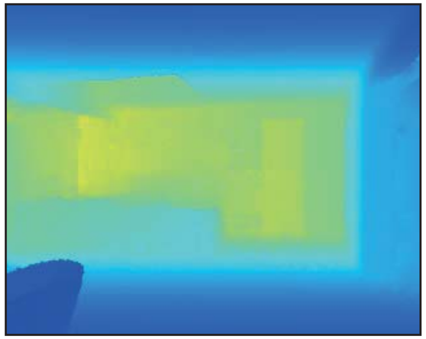
Harmonizing Overcomplete Predictions [Chakrabarti, 2016]

Goal: Depth Completion

Color Image

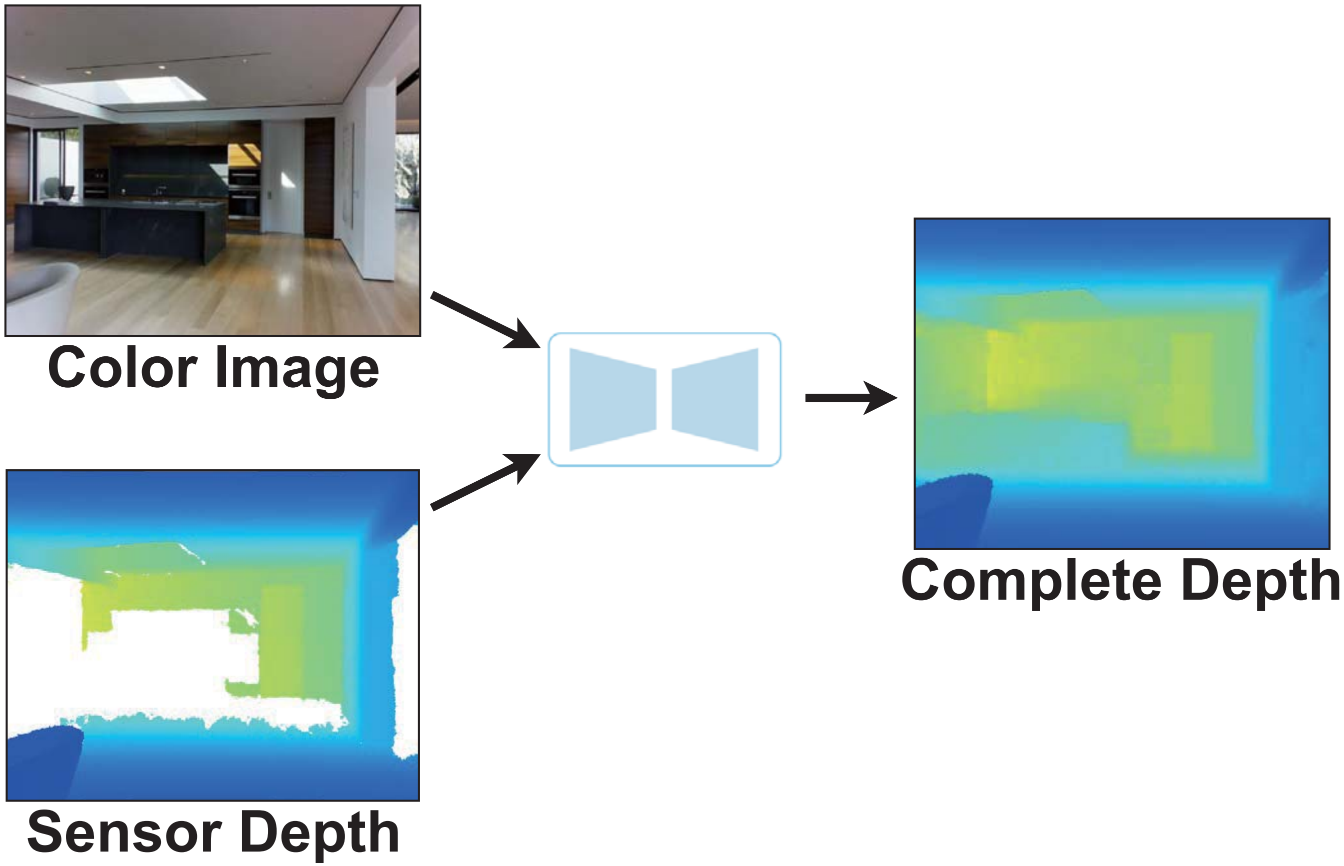


Sensor Depth

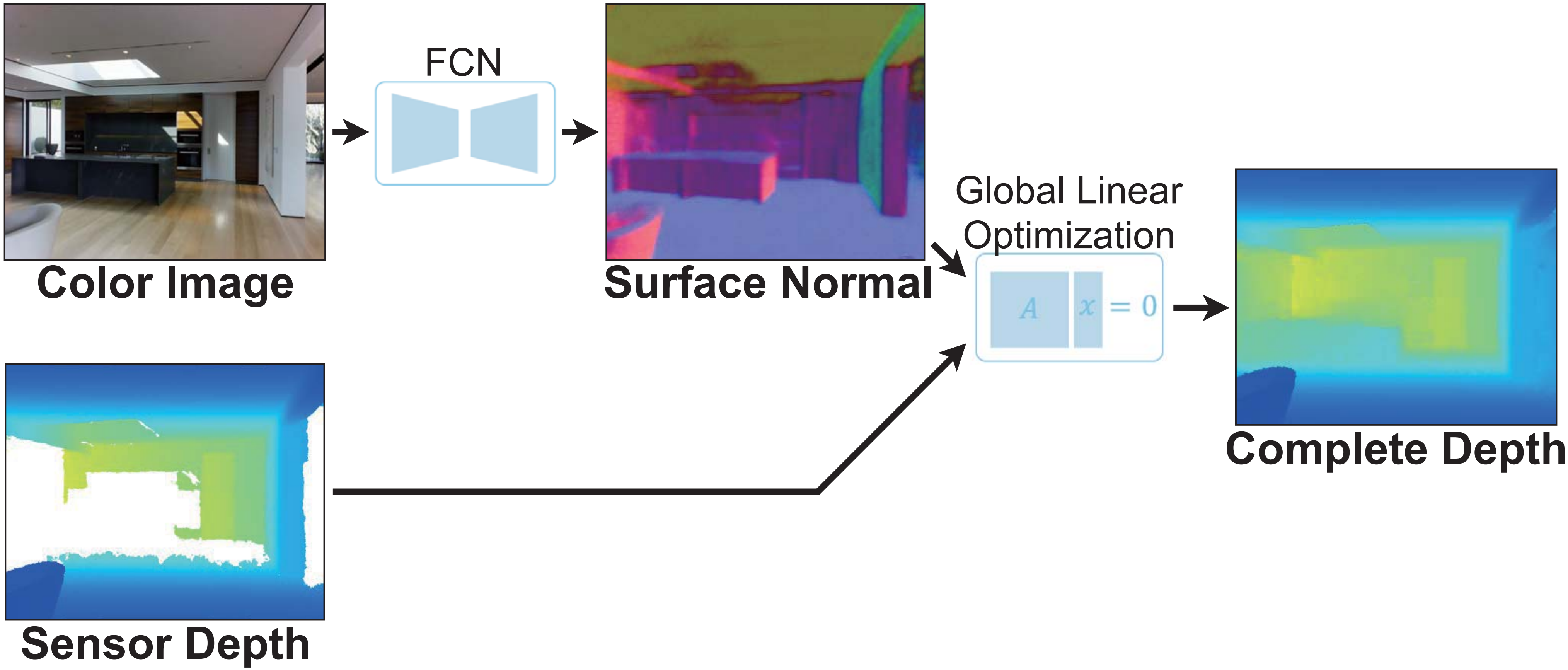


Complete Depth

Obvious Approach



Our Approach

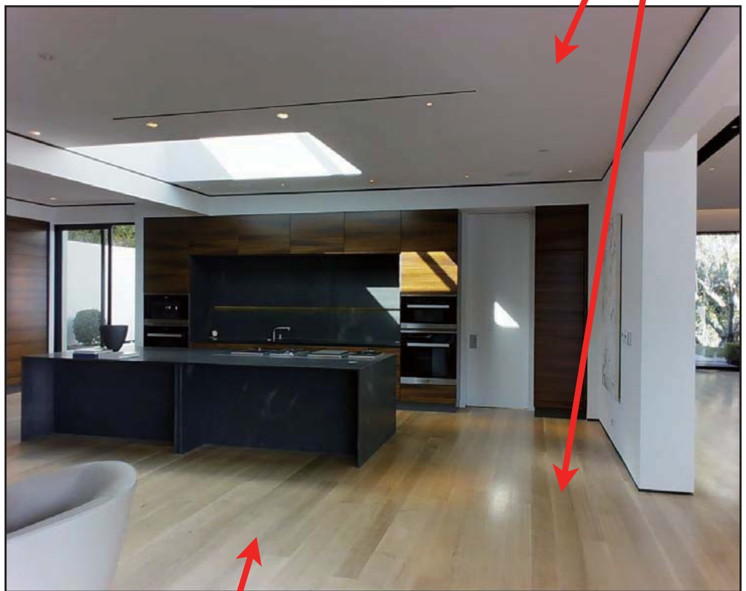


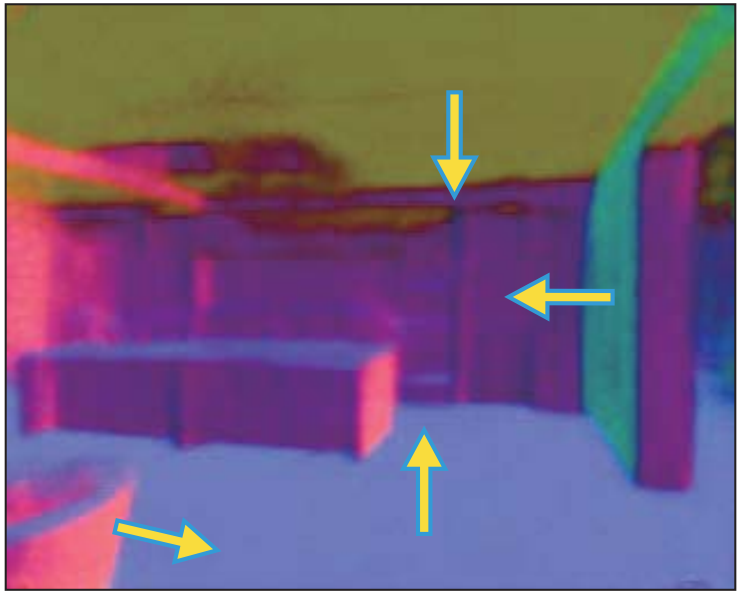
Why Surface Normals?

- Estimating surface normals is easier than estimating depths.

Determined by local shading or texture

Unit vector (-1,1)

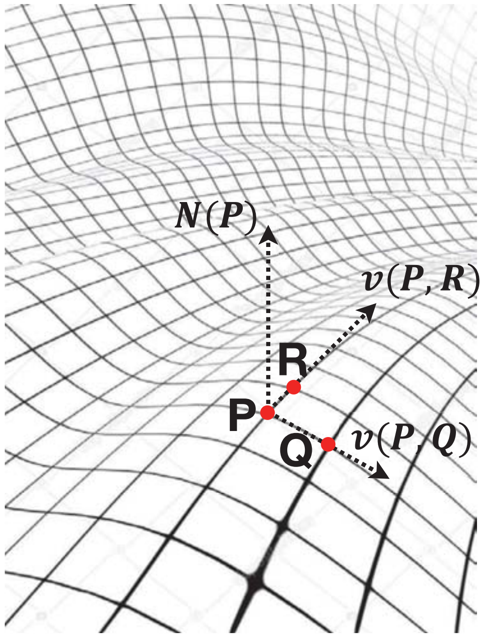




Constant within planar regions

Why Surface Normals?

- Estimating surface normals is easier than estimating depths.
- Depths can be estimated robustly from normals.



$N(P)$: Estimated surface normal at pixel P

Non-linear Constraints:

$$N_d(P) = v(P,Q) \times v(P,R)$$

$$\left\langle \frac{N_d(P)}{\|N_d(P)\|} \cdot N(P) \right\rangle = 1$$

$\rightarrow$

Linearized Constraints:

$$\langle N(P) \cdot v(P,Q) \rangle = 0$$

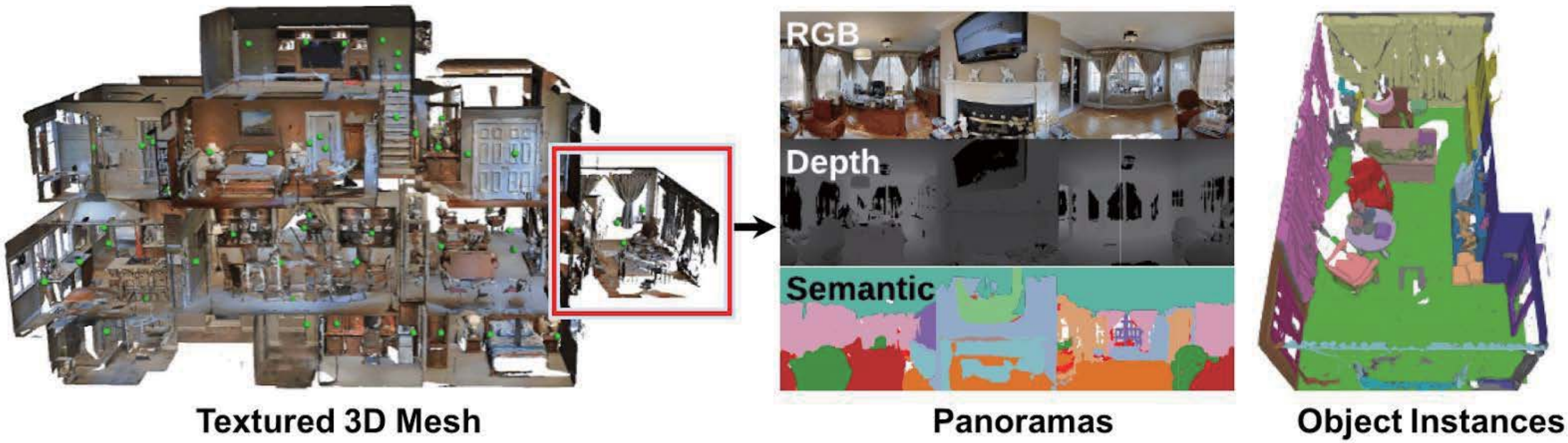
$$\langle N(P) \cdot v(P,R) \rangle = 0$$

Global Solution!

Experiment

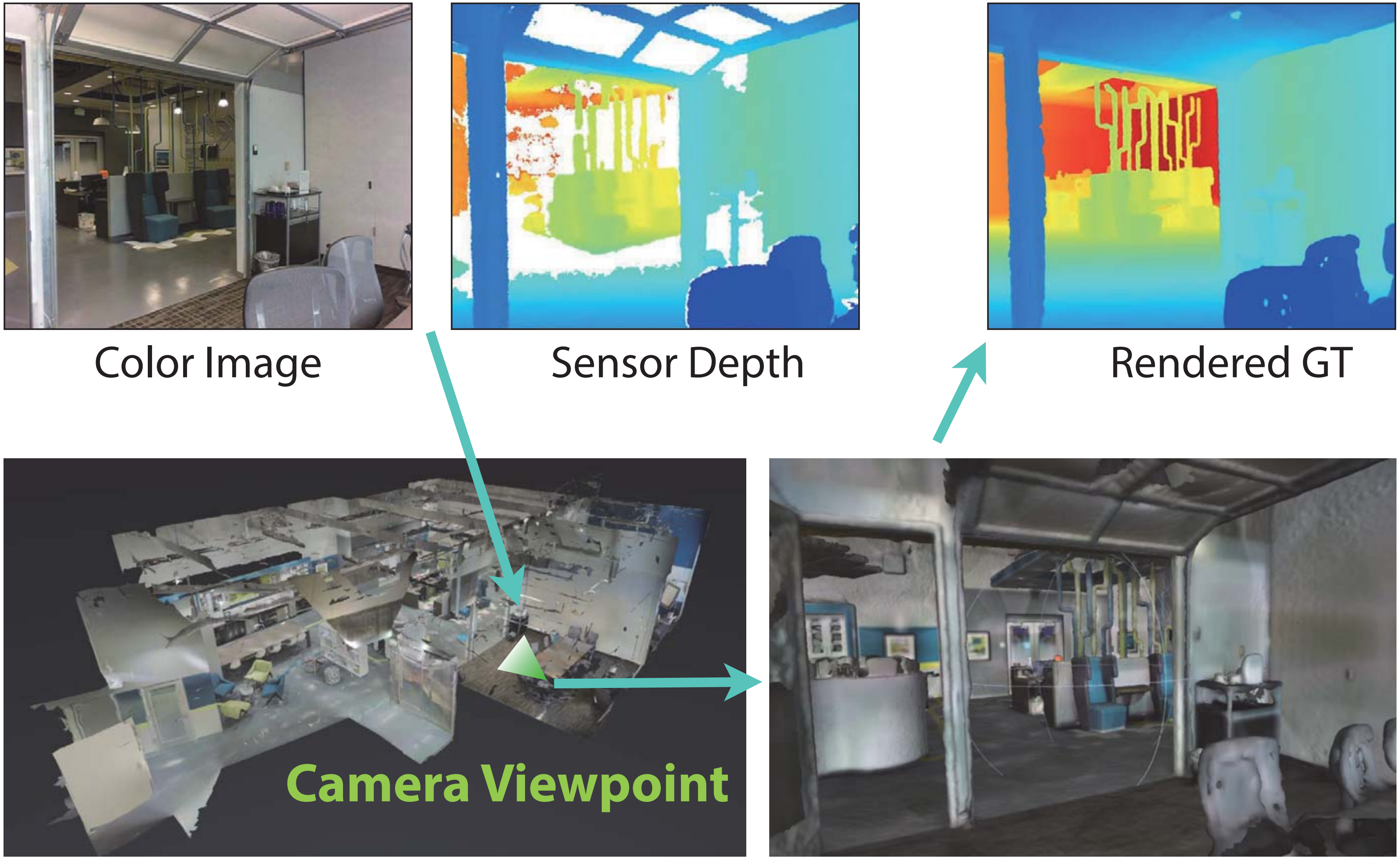
Ground Truth for Missing Area

- Matterport3D/ScanNet Dataset:
  - Environment with dense RGBD scans
  - High quality mesh reconstruction



[1] Dai et.al, ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes, CVPR 2017
[2] Chang et.al, Matterport3D: Learning from RGB-D Data in Indoor Environments, 3DV 2017

Ground Truth for Missing Area



Results on Matterport3D

- Evaluation metrics:
  - Relative Error: $median\left(\frac{\left|D_e(p) - D_g(p)\right|}{D_g(p)}\right)$
  - Squared Error: $median\left(\left(D_e(p) - D_g(p)\right)^2\right)$
  - Relative Depth: $max\left(\frac{D_e(p)}{D_g(p)}, \frac{D_g(p)}{D_e(p)}\right)$

- Comparison to depth completion methods:

[5] J. T. Barron and B. Poole. The fast bilateral solver. ECCV 2016.
[20] D. Ferstl et al. Image guided depth upsampling using anisotropic total generalized variation. ICCV 2013.
[23] D. Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. Comp. stat. & data anal., 2010.
[64] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. ECCV 2012.
[80] Y. Zhang et al. Physically-based rendering for indoor scene understanding using convolutional neural networks. CVPR 2017.

Results on Matterport3D

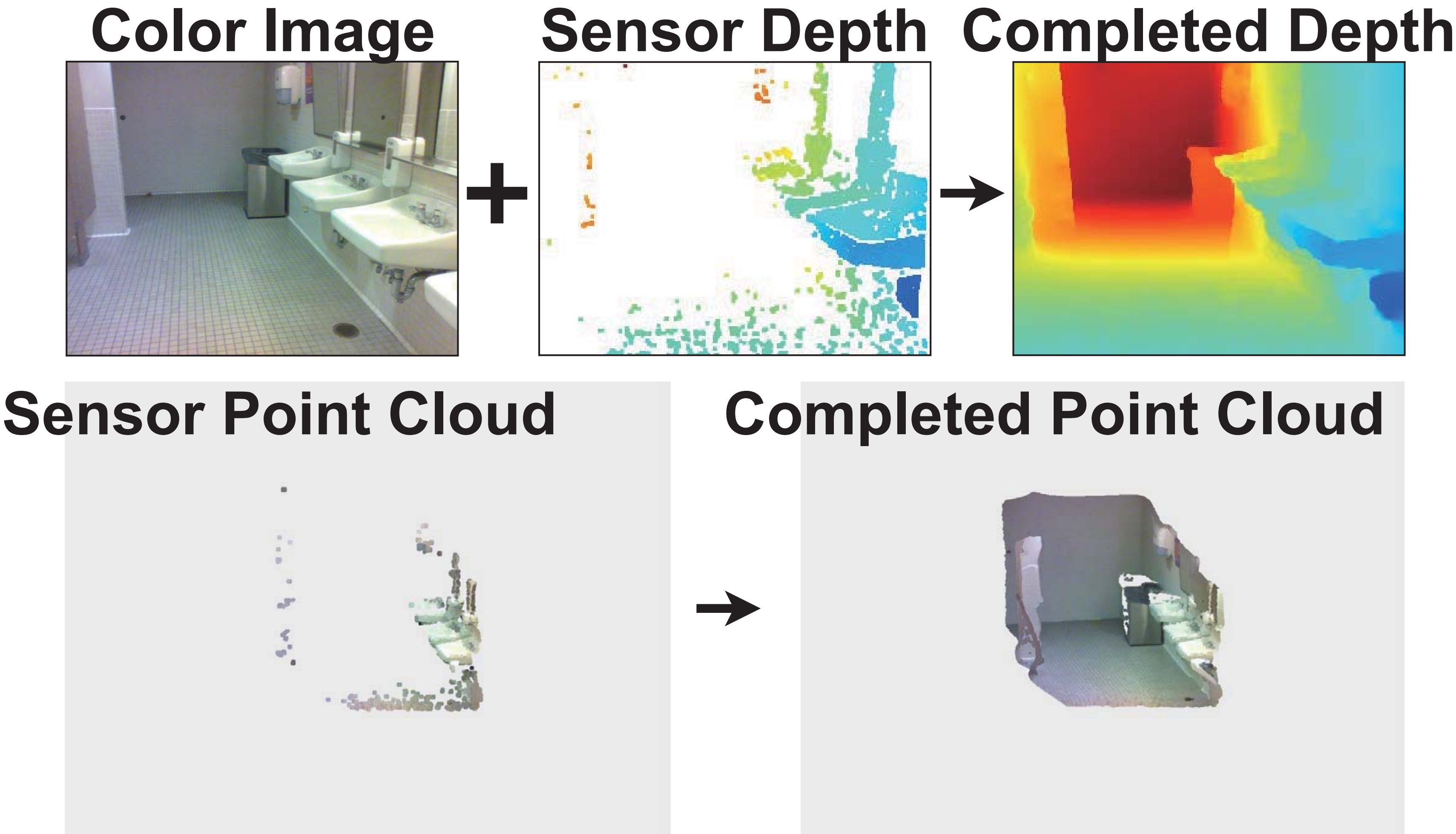
- Comparison to depth completion methods:

Method	Rel↓	RMSE↓	1.05↑	1.10↑	1.25↑	1.25^2↑	1.25^3↑
Smooth	0.151	0.187	32.80	42.71	57.61	72.29	80.15
Bilateral [64]	0.118	0.152	34.39	46.50	61.92	75.26	81.84
Fast [5]	0.127	0.154	33.65	45.08	60.36	74.52	81.79
TGV [20]	0.103	0.146	37.40	48.75	62.97	75.00	81.71
Garcia et.al [23]	0.115	0.144	36.78	47.13	61.48	74.89	81.67
FCN [80]	0.167	0.241	16.43	31.13	57.62	75.63	84.01
Ours	**0.089**	**0.116**	**40.63**	**51.21**	**65.35**	**76.74**	**82.98**

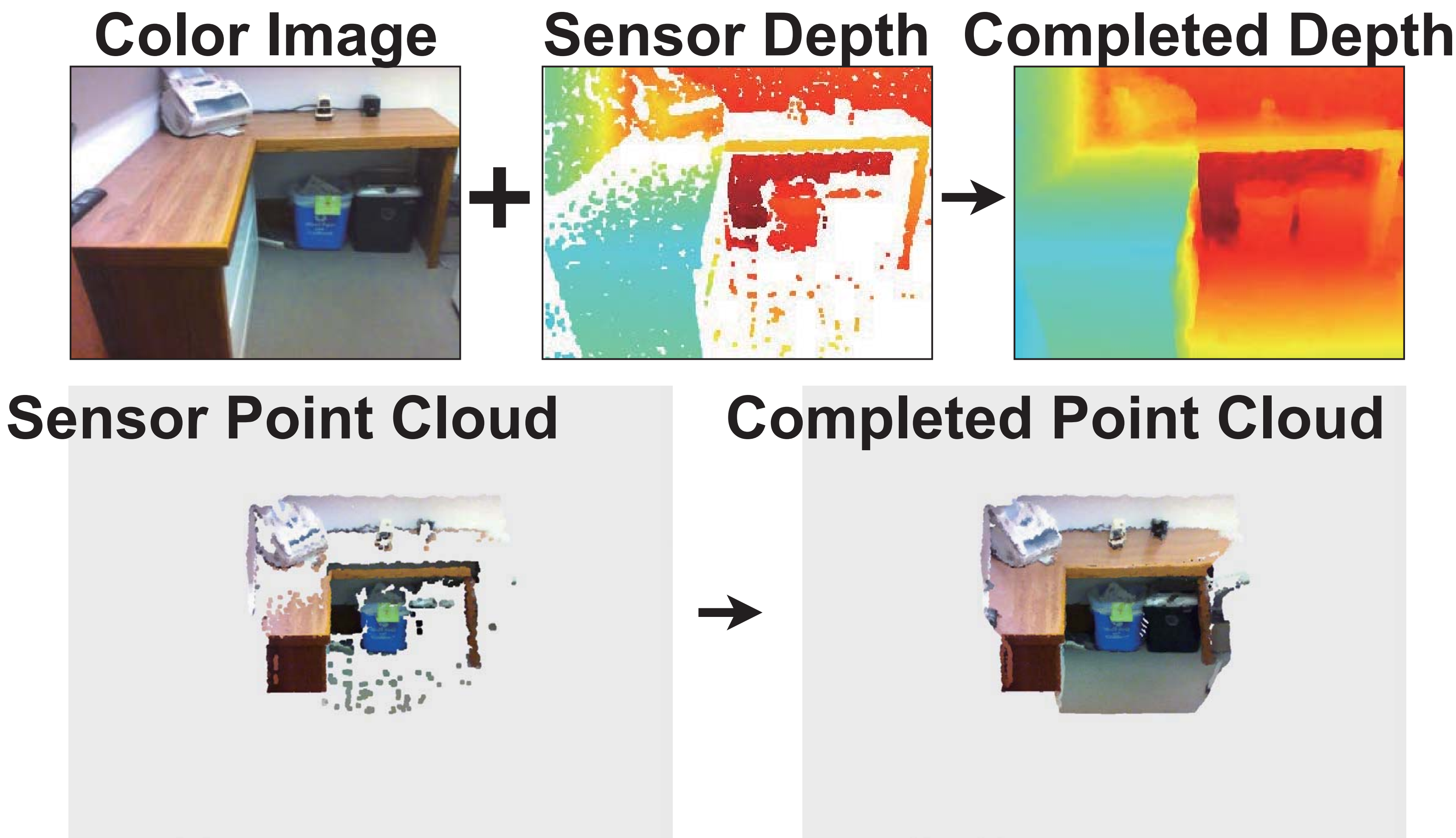
[5] J. T. Barron and B. Poole. The fast bilateral solver. ECCV 2016.
[20] D. Ferstl et al. Image guided depth upsampling using anisotropic total generalized variation. ICCV 2013.
[23] D. Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. Comp. stat. & data anal., 2010.
[64] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. ECCV 2012.
[80] Y. Zhang et al. Physically-based rendering for indoor scene understanding using convolutional neural networks. CVPR 2017.

More Challenging Case…

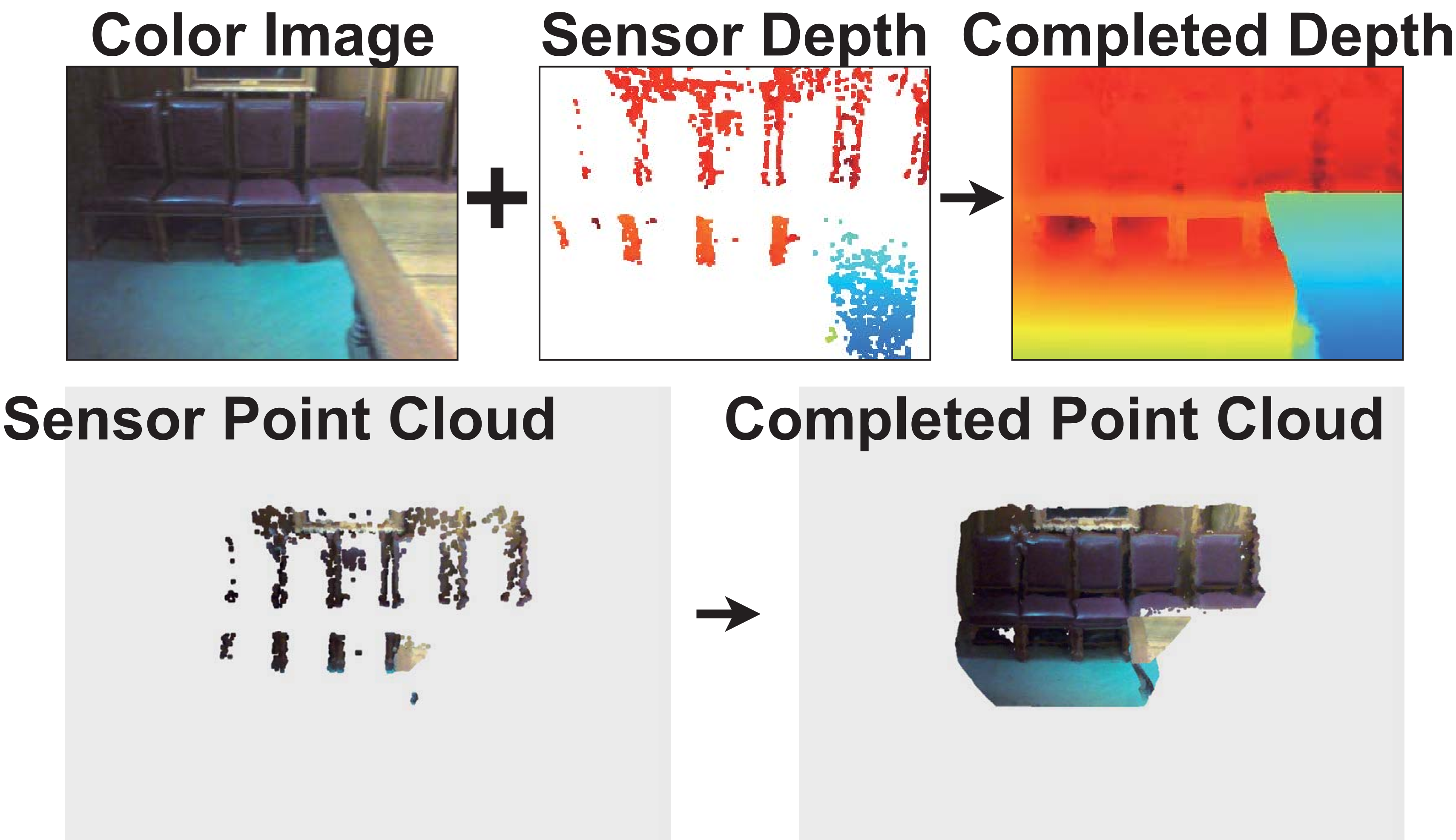
Results on Realsense R200



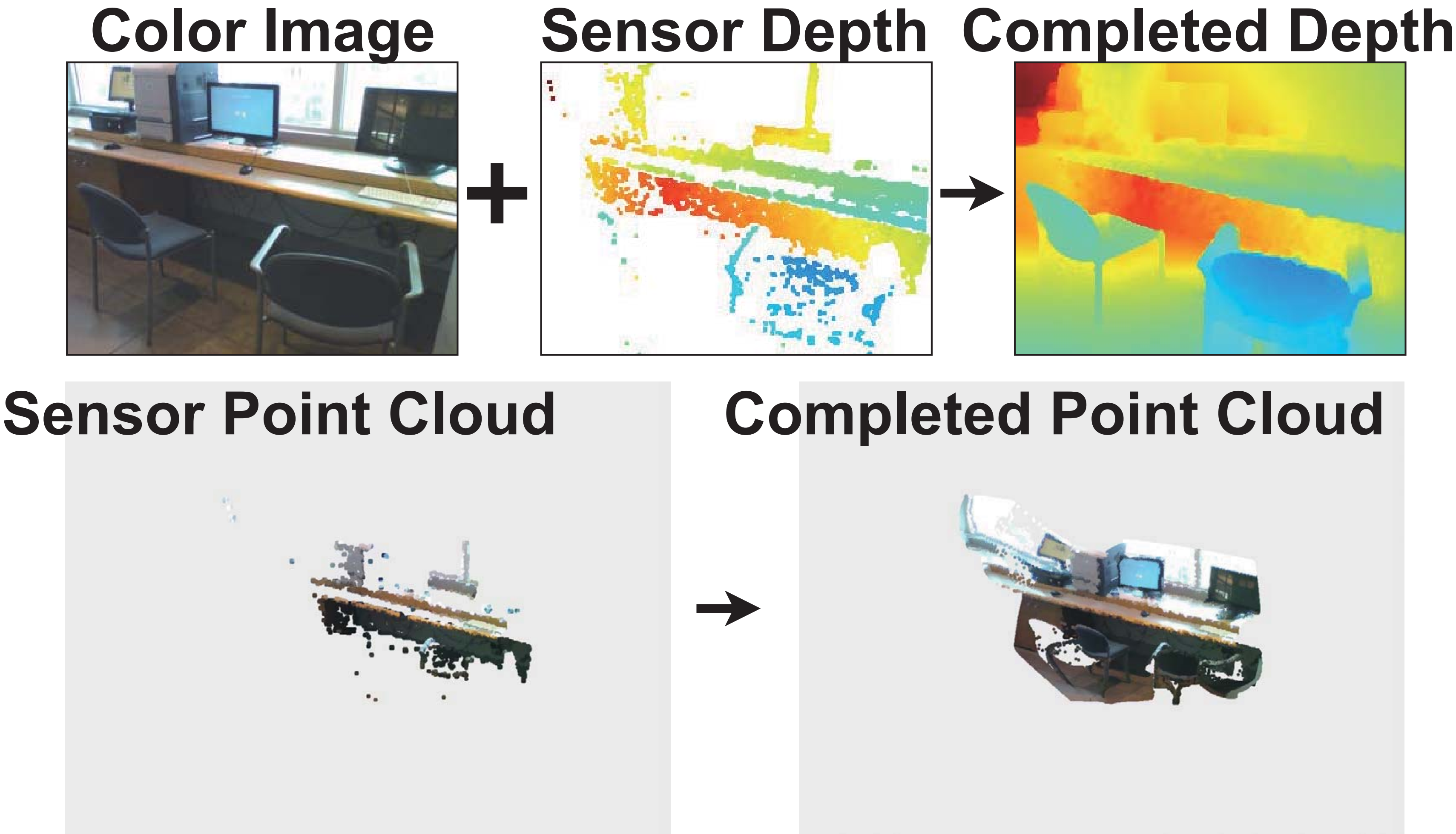
Results on Realsense R200



Results on Realsense R200

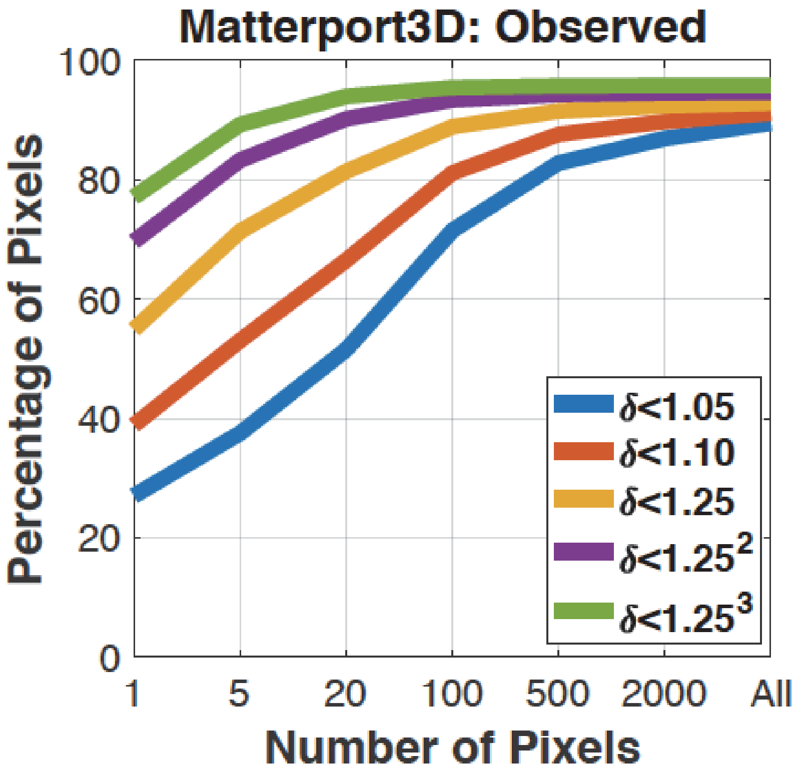


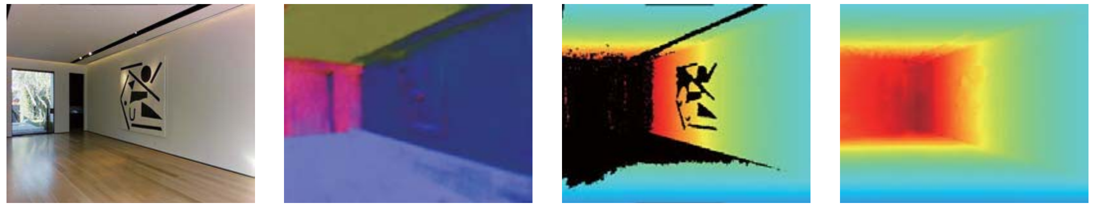
Results on Realsense R200



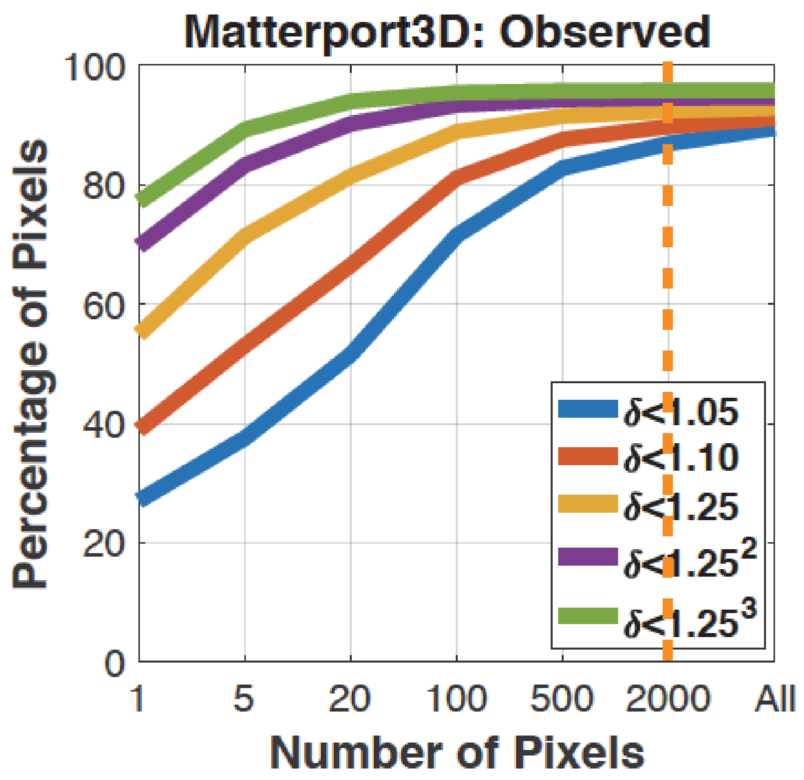
Go crazy!

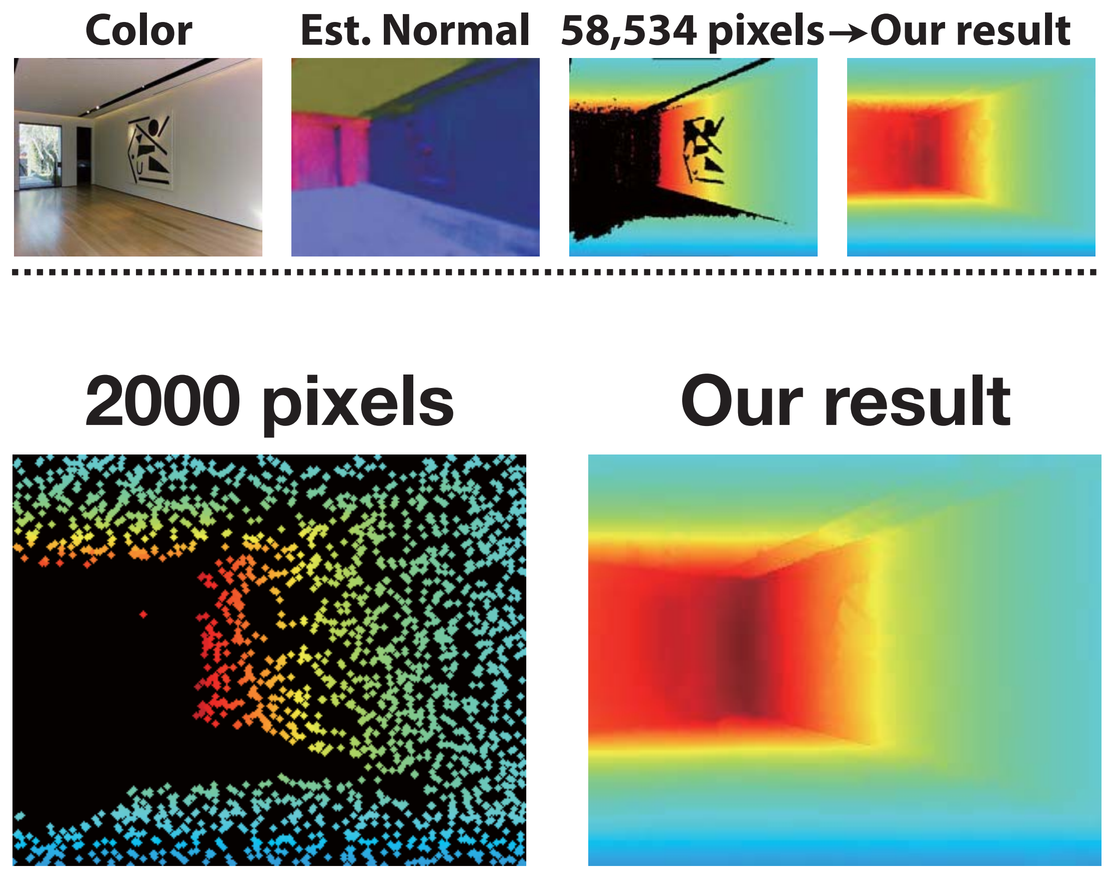
How Much Depth Do We Need?



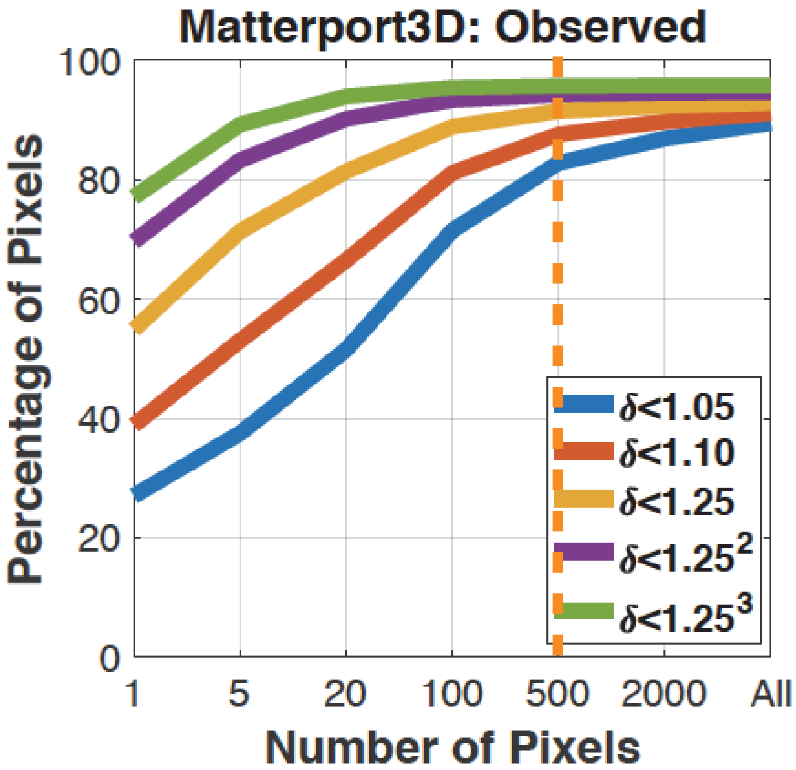


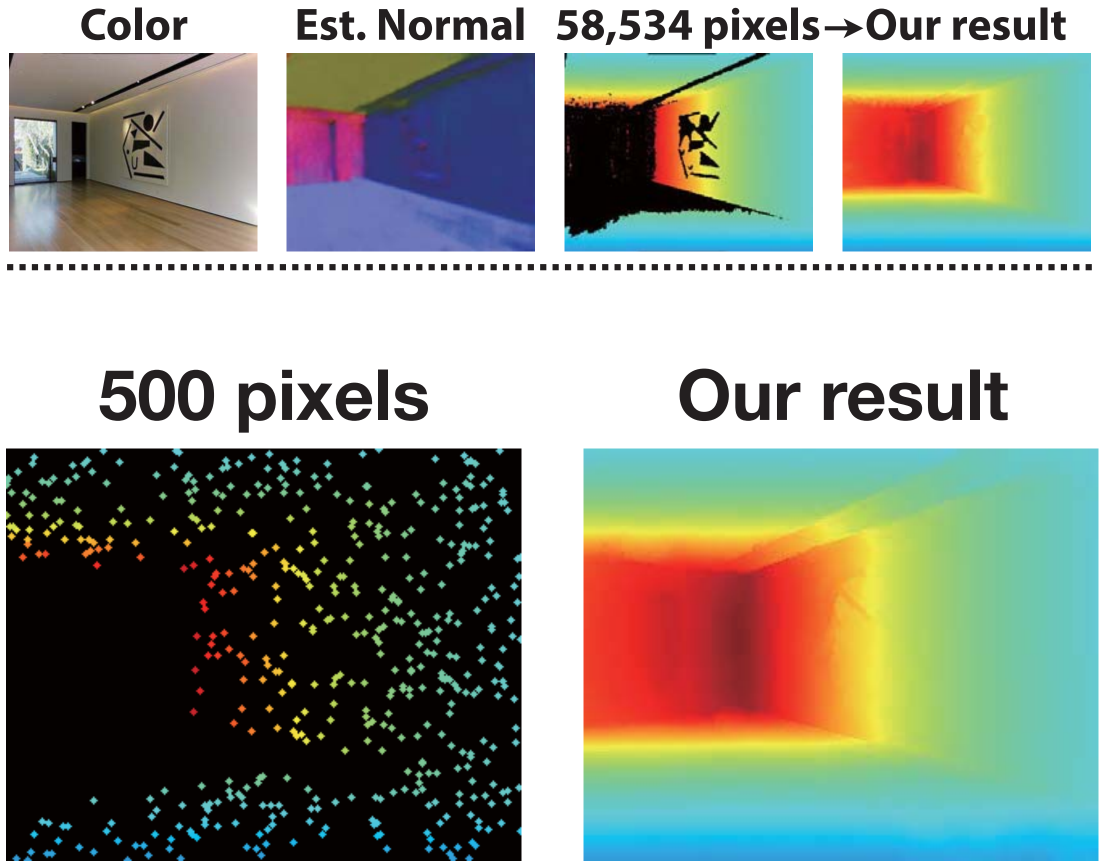
How Much Depth Do We Need?



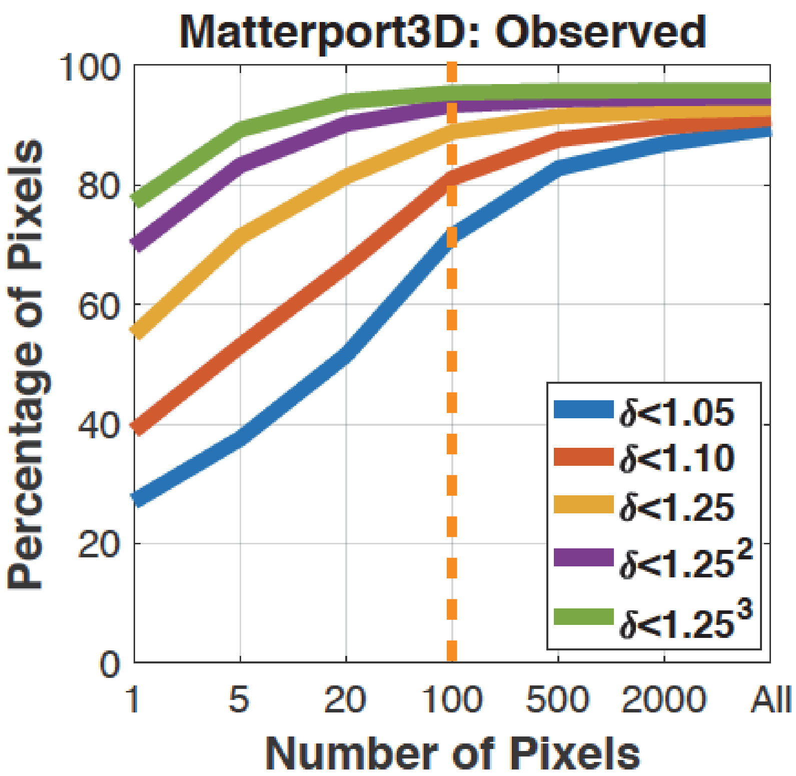


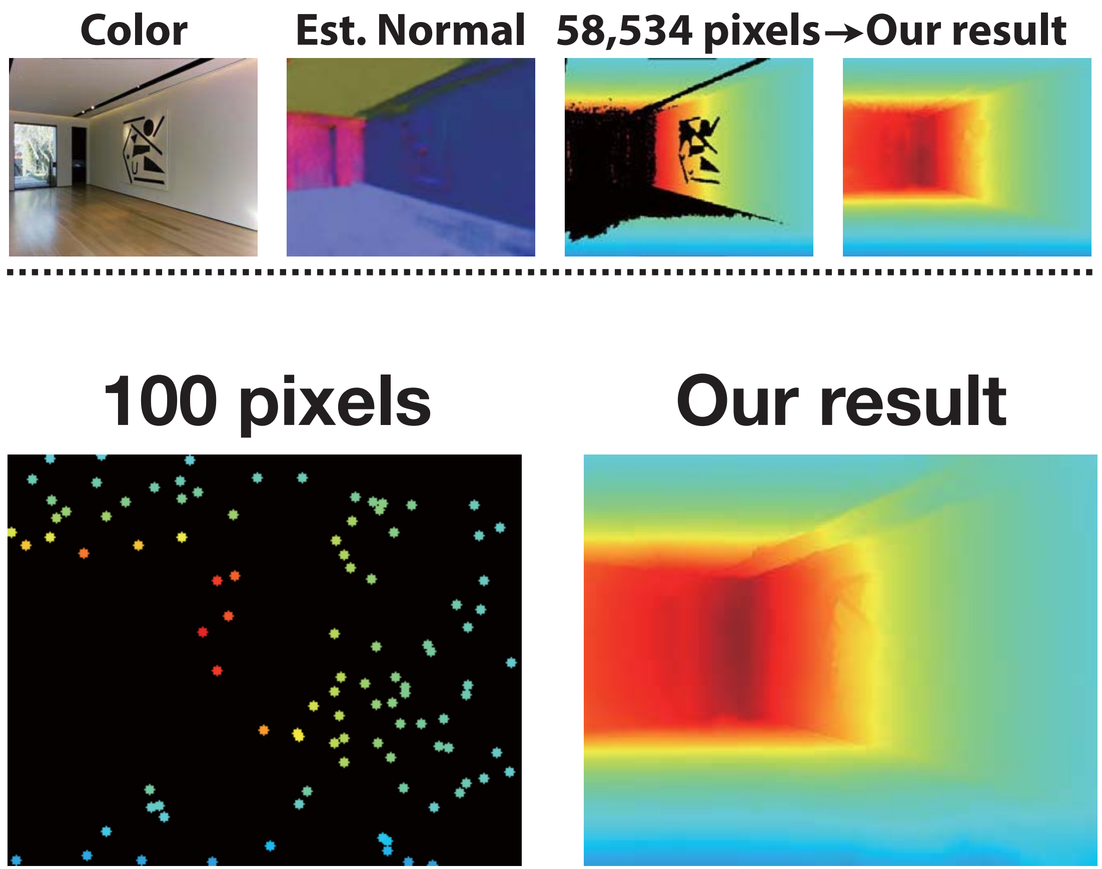
How Much Depth Do We Need?



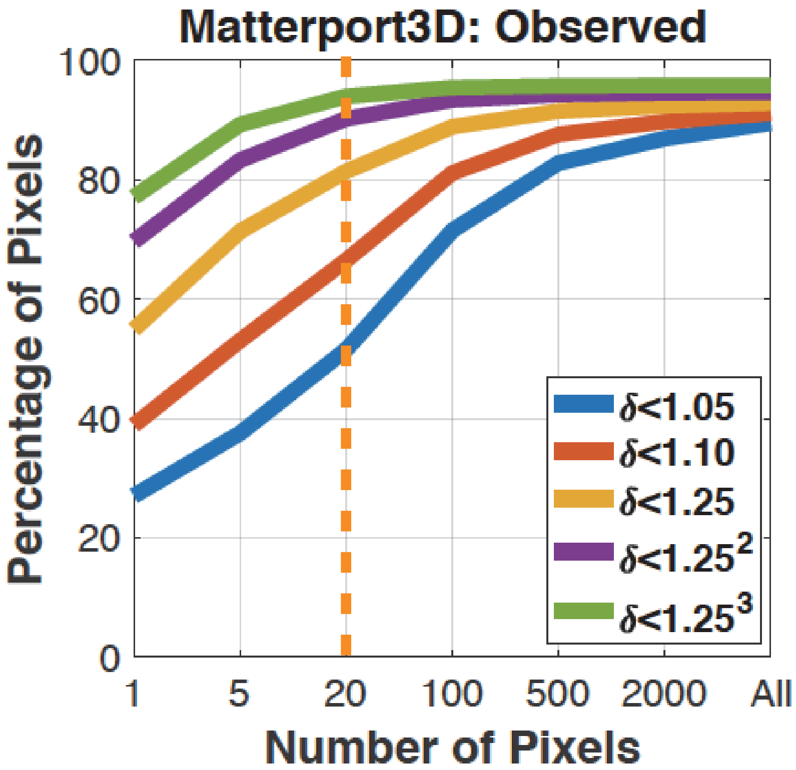


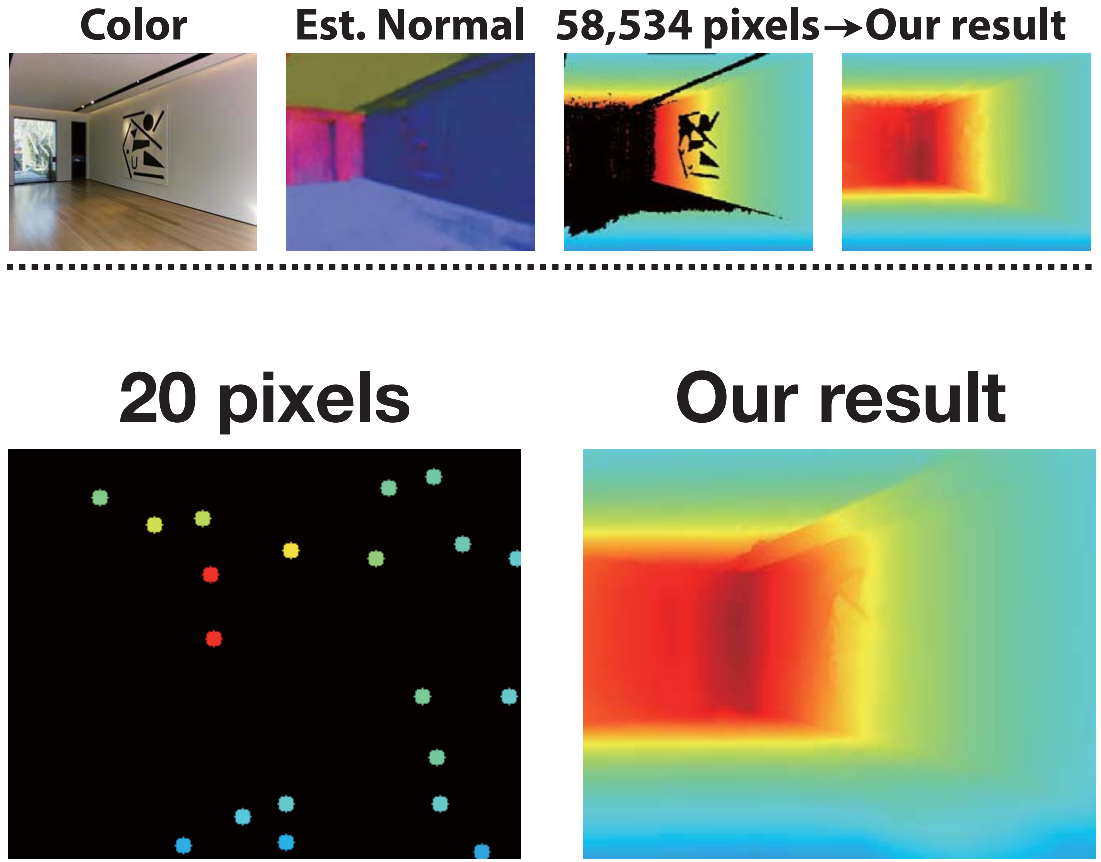
How Much Depth Do We Need?





How Much Depth Do We Need?





Couldn’t be harder… 1px

Depth Completion with 1px

- Comparison to depth estimation methods:
  - Run our method assuming center pixel is at 3 meter, and globally scale with the 1 known depth.
  - Run single image based depth estimation, and globally scale with the 1 known depth.

Obs	Meth	Rel↓	RMSE↓	1.05↑	1.10↑	1.25↑	1.25^2↑	1.25^3↑
Y	[37]	0.190	0.374	17.90	31.03	54.80	75.97	85.69
	[7]	0.161	0.320	21.52	35.5	58.75	77.48	85.65
	Ours	**0.130**	**0.274**	**30.60**	**43.65**	**61.14**	**75.69**	**82.65**
N	[37]	0.384	0.572	8.86	16.67	34.64	55.60	69.21
	[7]	0.352	0.610	11.16	20.50	37.73	57.77	70.10
	Ours	**0.283**	**0.537**	**17.27**	**27.42**	**44.19**	**61.80**	**70.90**

[7] Chakrabarti, A. et al., Depth from a single image by harmonizing overcomplete local network predictions. NIPS 2016.
[37] Laina, C. et al., Deeper depth prediction with fully convolutional residual networks. 3DV 2016.

Deep Depth Completion of a Single RGB-D Image

Yinda Zhang, Thomas Funkhouser

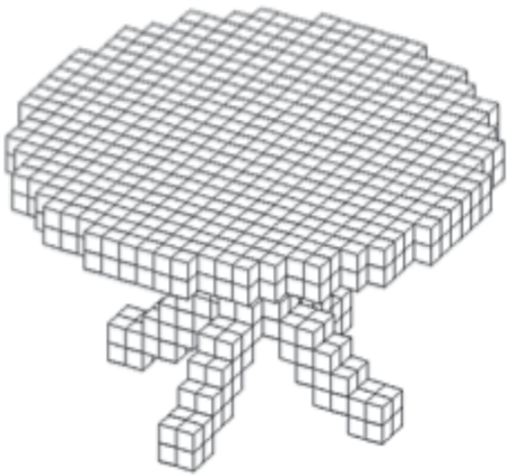
CVPR 2018

Project Webpage:
http://deepcompletion.cs.princeton.edu/

3D Geometry

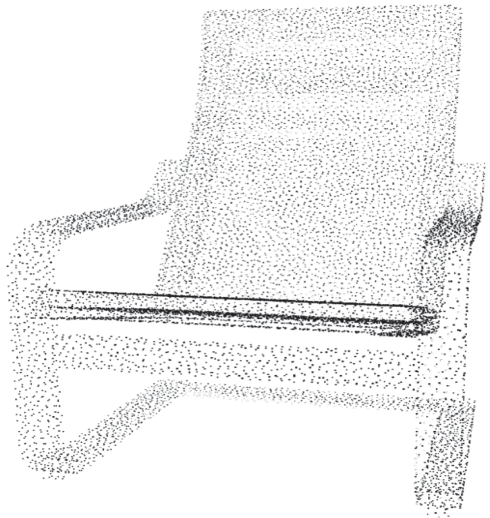
- Representation

  - Volumetric



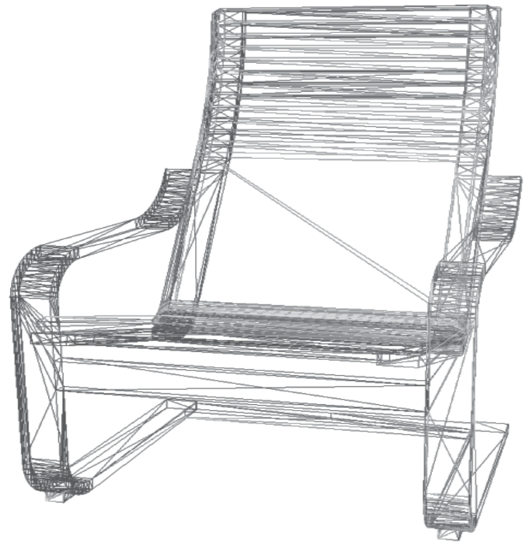
✓ Easy for deep learning
▸ High memory cost
▸ Slow computation
▸ Low resolution

  - Point cloud



✓ Can work with deep learning
▸ Point cloud are un-ordered
▸ No high order surface detail
▸ Non-trivial to form closed shape

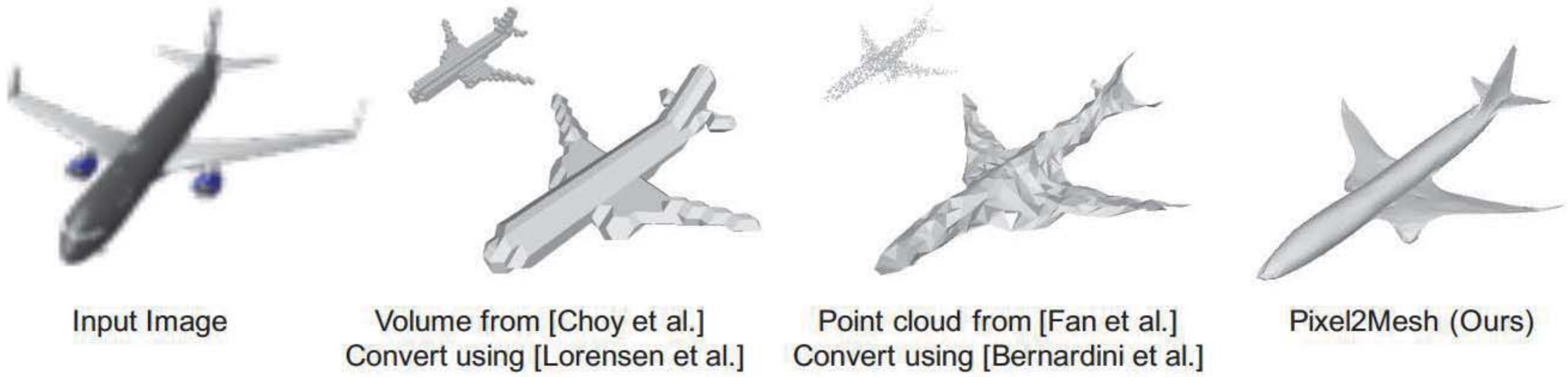
  - Mesh



✓ Connectivity
✓ Surface details
▸ Tricky for deep learning

Pixel2Mesh

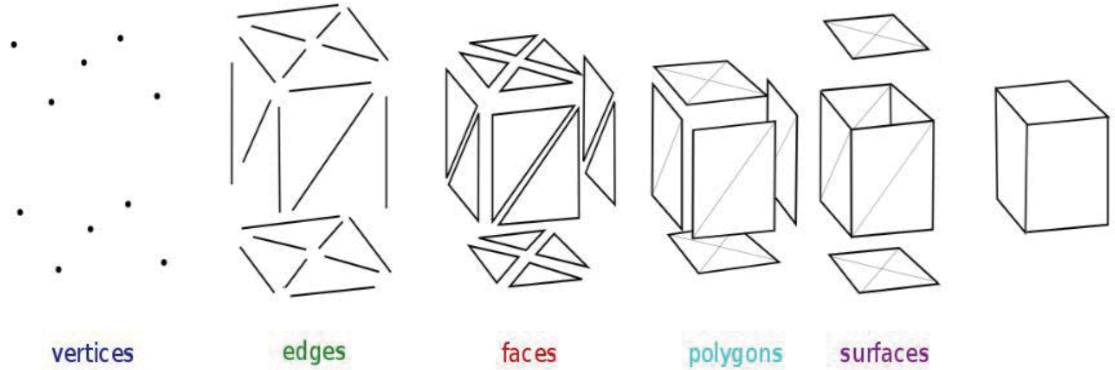
- End-to-end system produces mesh from a single color image.

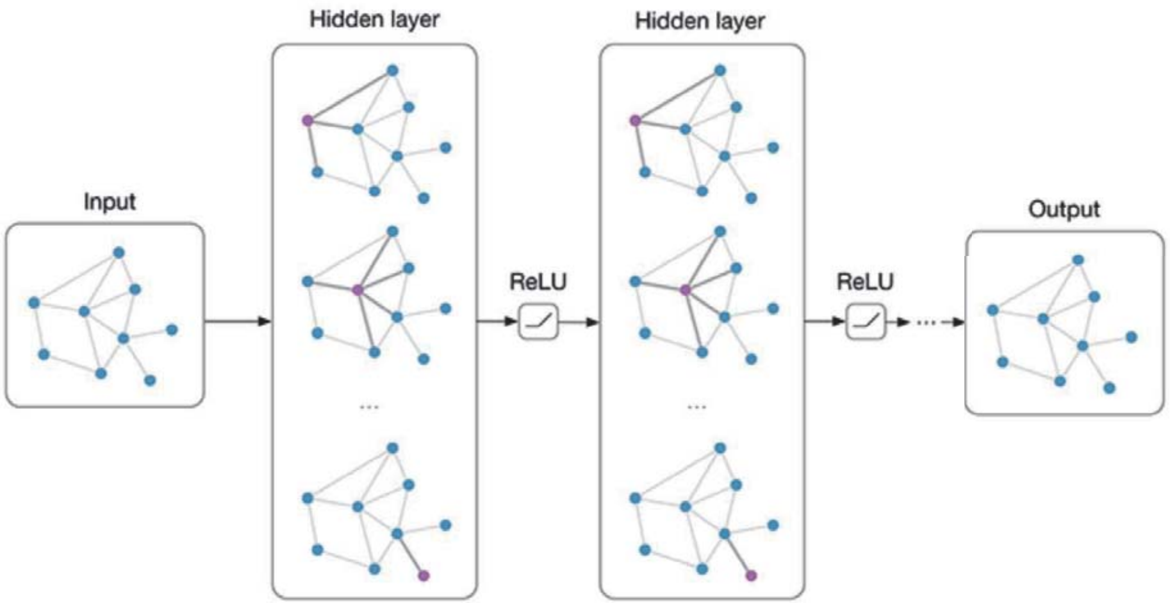


[1] Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV (2016)
[2] Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: CVPR (2017)
[3] Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: SIGGRAPH (1987)
[4] Bernardini, F., Mittleman, J., Rushmeier, H.E., Silva, C.T., Taubin, G.: The ball-pivoting algorithm for surface reconstruction. IEEE Trans. Vis. Comput. Graph. 5(4), 349–359 (1999)

Graph-based CNN

- Mesh can be represented as a graph.
- Run CNN on graph.
- Belief propagation.

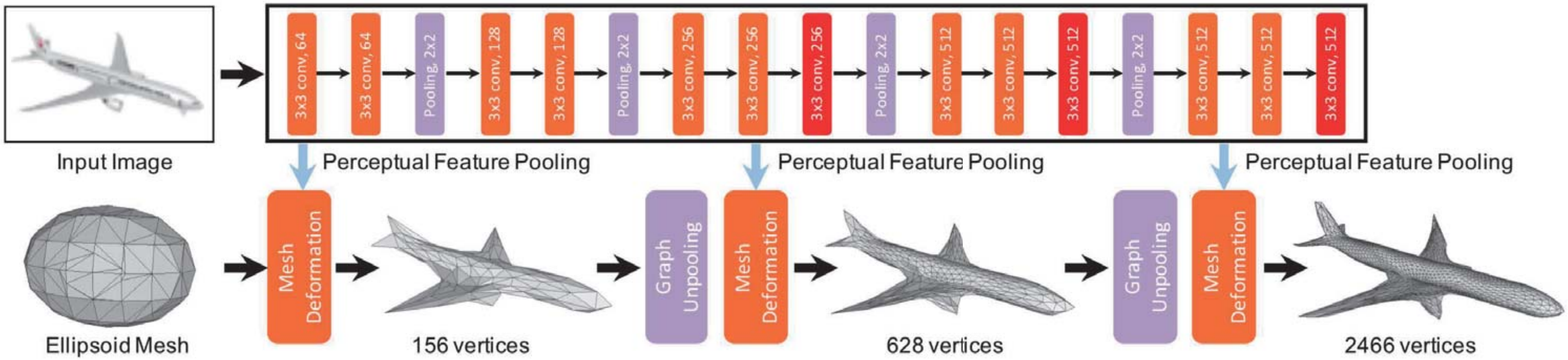




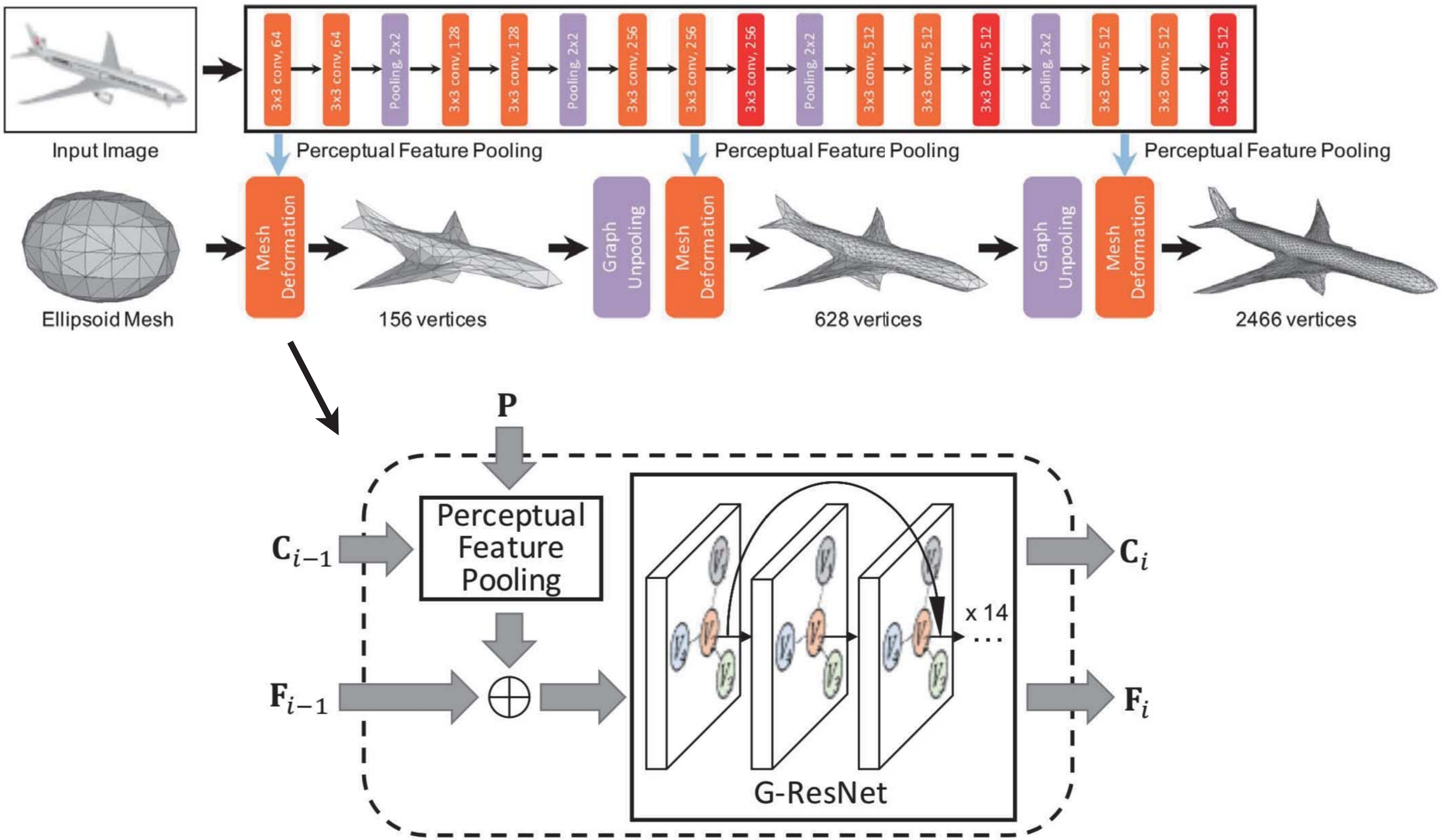
$$f_p^{l+1} = w_0 f_p^l + \sum_{q \in \mathcal{N}(p)} w_1 f_q^l$$

Pixel2Mesh

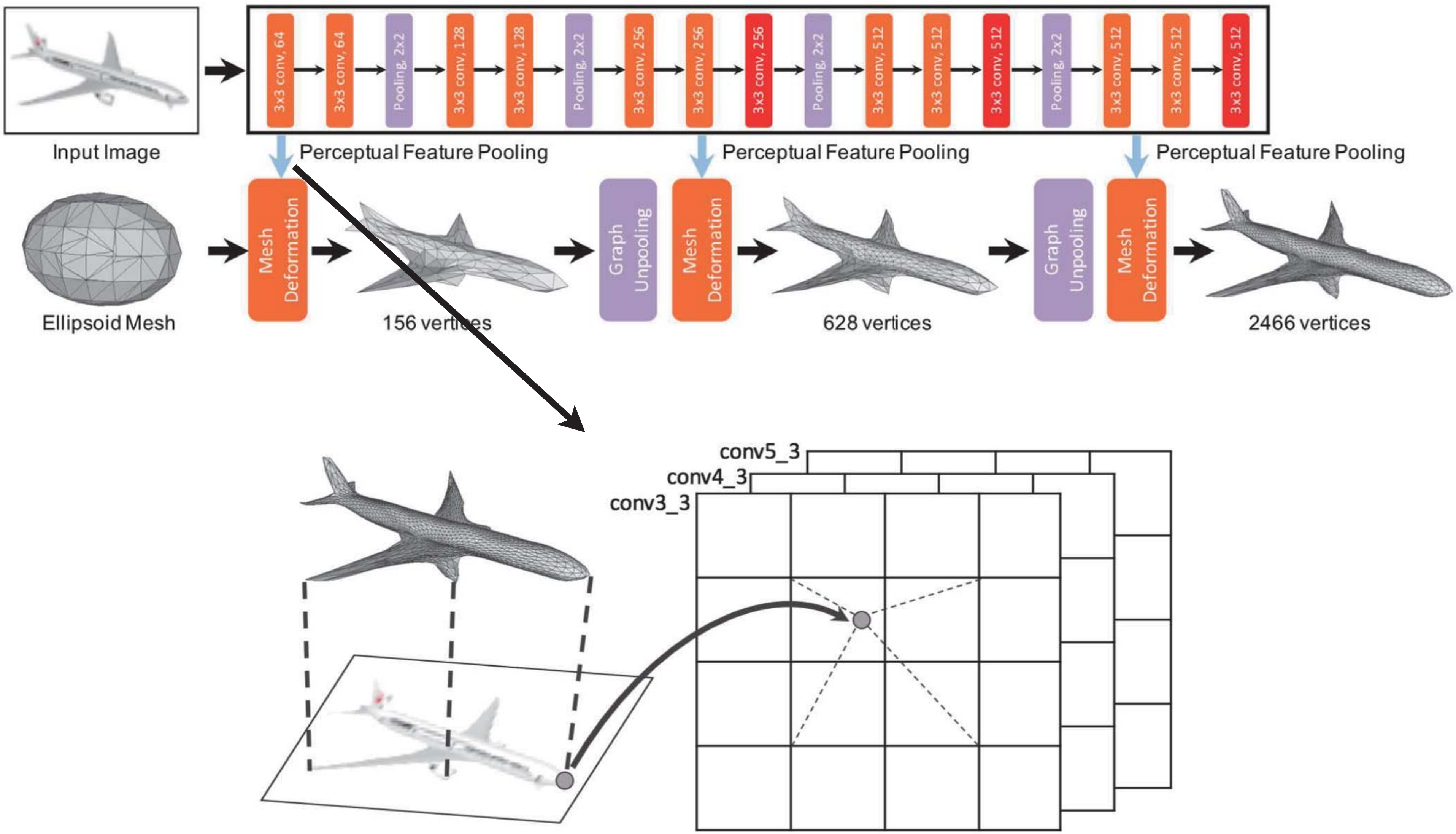
- Deform from a ellipsoid to target mesh.
- From coarse to fine.
- Explainable model.



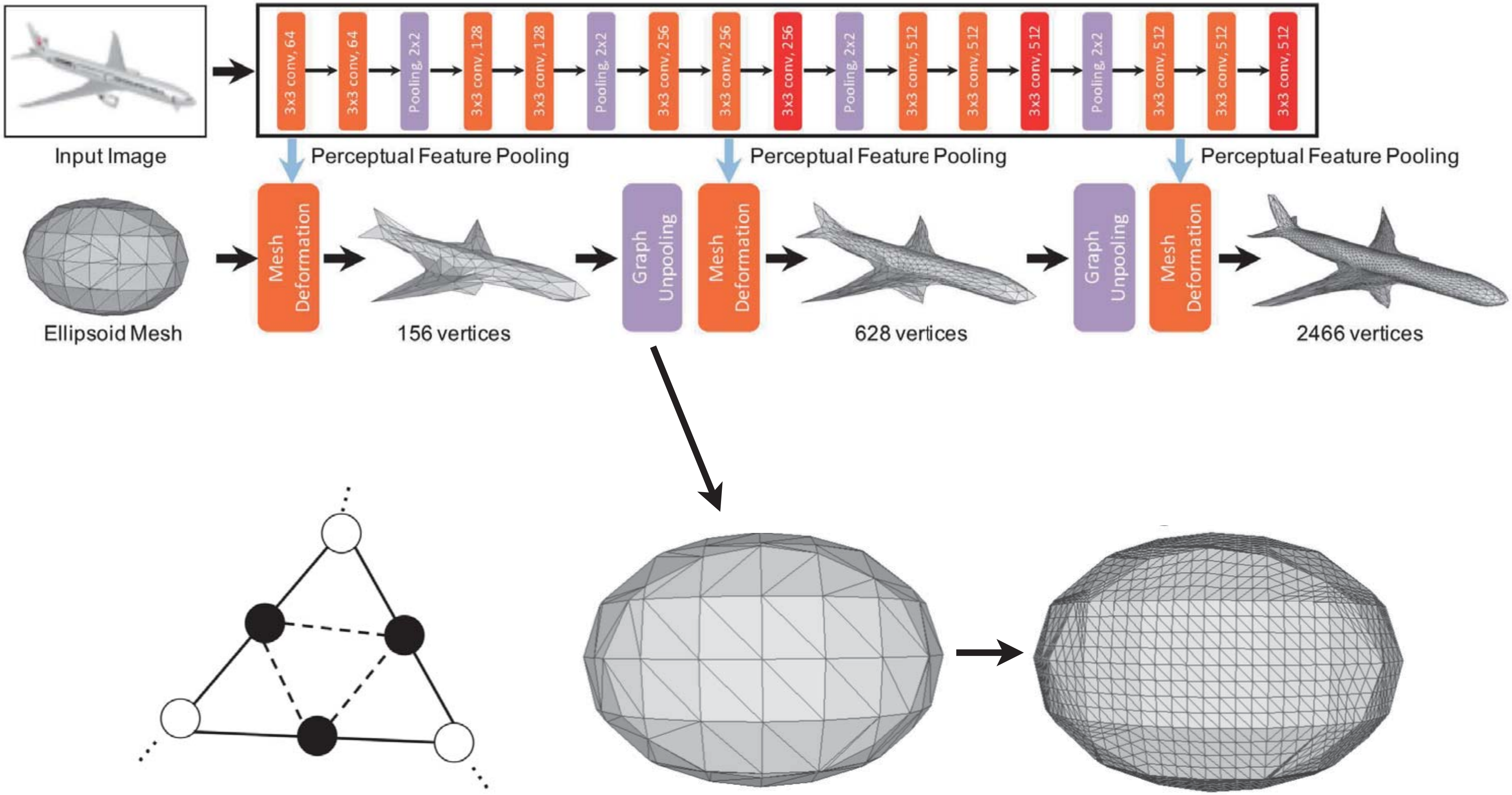
Mesh Deformation Block



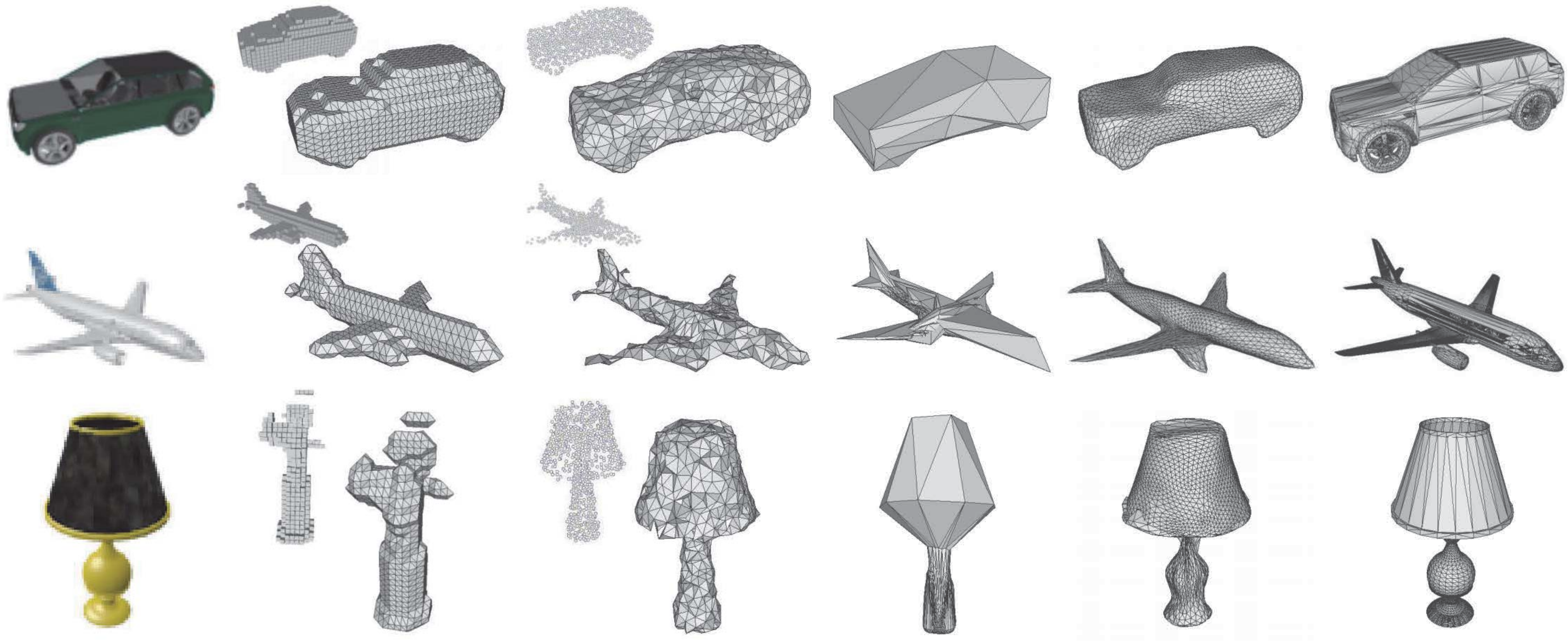
Perceptual Feature Pooling



Graph Unpooling



Experiment Results



(a) (b) (c) (d) (e) (f)

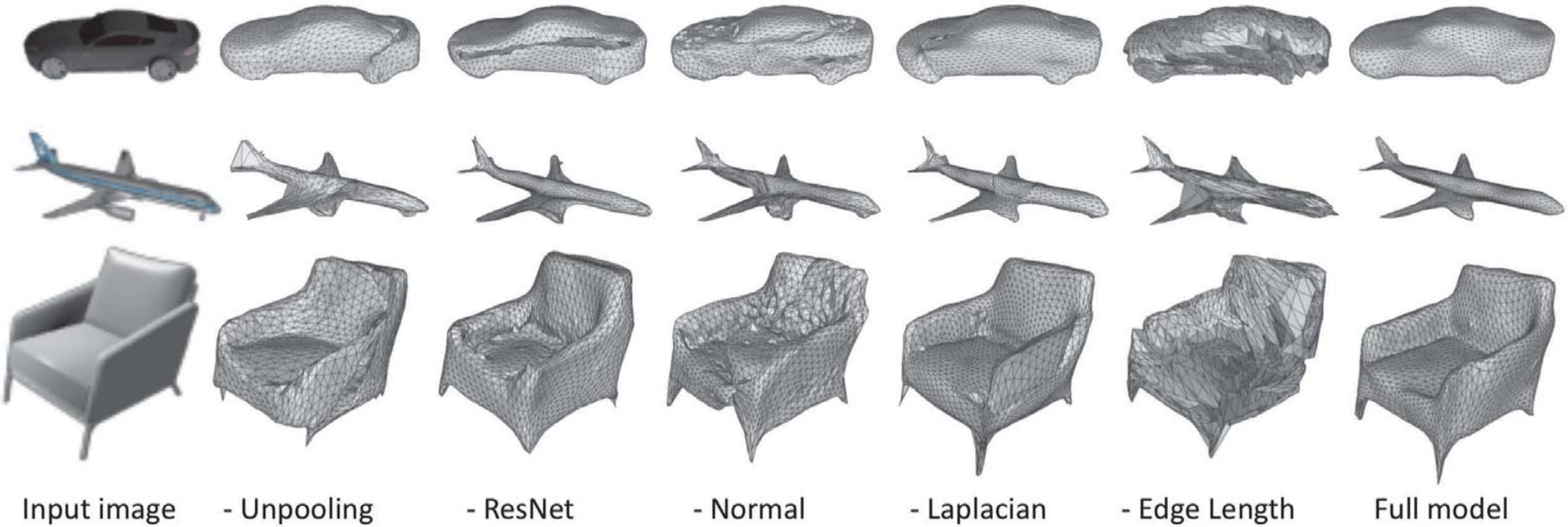
On ShapeNet rendering. (a) Input image; (b) Volume from 3D-R2N2 [1], converted using Marching Cube [4]; (c) Point cloud from PSG [2], converted using ball pivoting [5]; (d) N3MR[3]; (e) Ours; (f) Ground truth.

Experiment Results

Category	CD				EMD			
	3D-R2N2	PSG	N3MR	Ours	3D-R2N2	PSG	N3MR	Ours
plane	0.895	**0.430**	0.450	0.477	0.606	**0.396**	7.498	0.579
bench	1.891	0.629	2.268	**0.624**	1.136	1.113	11.766	**0.965**
cabinet	0.735	0.439	2.555	**0.381**	**2.520**	2.986	17.062	2.563
car	0.845	0.333	2.298	**0.268**	1.670	1.747	11.641	**1.297**
chair	1.432	0.645	2.084	**0.610**	1.466	1.946	11.809	**1.399**
monitor	1.707	**0.722**	3.111	0.755	1.667	1.891	14.097	**1.536**
lamp	4.009	**1.193**	3.013	1.295	1.424	1.222	14.741	**1.314**
speaker	1.507	0.756	3.343	**0.739**	**2.732**	3.490	16.720	2.951
firearm	0.993	**0.423**	2.641	0.453	0.688	**0.397**	11.889	0.667
couch	1.135	0.549	3.512	**0.490**	2.114	2.207	14.876	**1.642**
table	1.116	0.517	2.383	**0.498**	1.641	2.121	12.842	**1.480**
cellphone	1.137	0.438	4.366	**0.421**	0.912	1.019	17.649	**0.724**
watercraft	1.215	**0.633**	2.154	0.670	0.935	0.945	11.425	**0.814**
mean	1.445	0.593	2.629	**0.591**	1.501	1.653	13.386	**1.380**

CD and EMD on the ShapeNet test set. Smaller is better.

Ablation Study



Pixel2Mesh: Generating 3D Mesh Models from Single RGB Image

Nanyang Wang, Yinda Zhang, Zhuwen Li,
Yanwei Fu, Wei Liu, Yu-Gang Jiang

ECCV 2018

Project Webpage:
http://bigvid.fudan.edu.cn/pixel2mesh/

What's next?

- Better geometry
  - Thin structure
  - Distant area
  - Dynamic scene
- Availability
  - Quality v.s. Computation
  - Multi-view, Temporal sequence
  - IR, Projector, Color, Event image
- Integration
  - Semantics, Motion Planing
  - Rendering
  - VR/AR