CoReNet: Coherent 3D scene reconstruction from a single RGB image

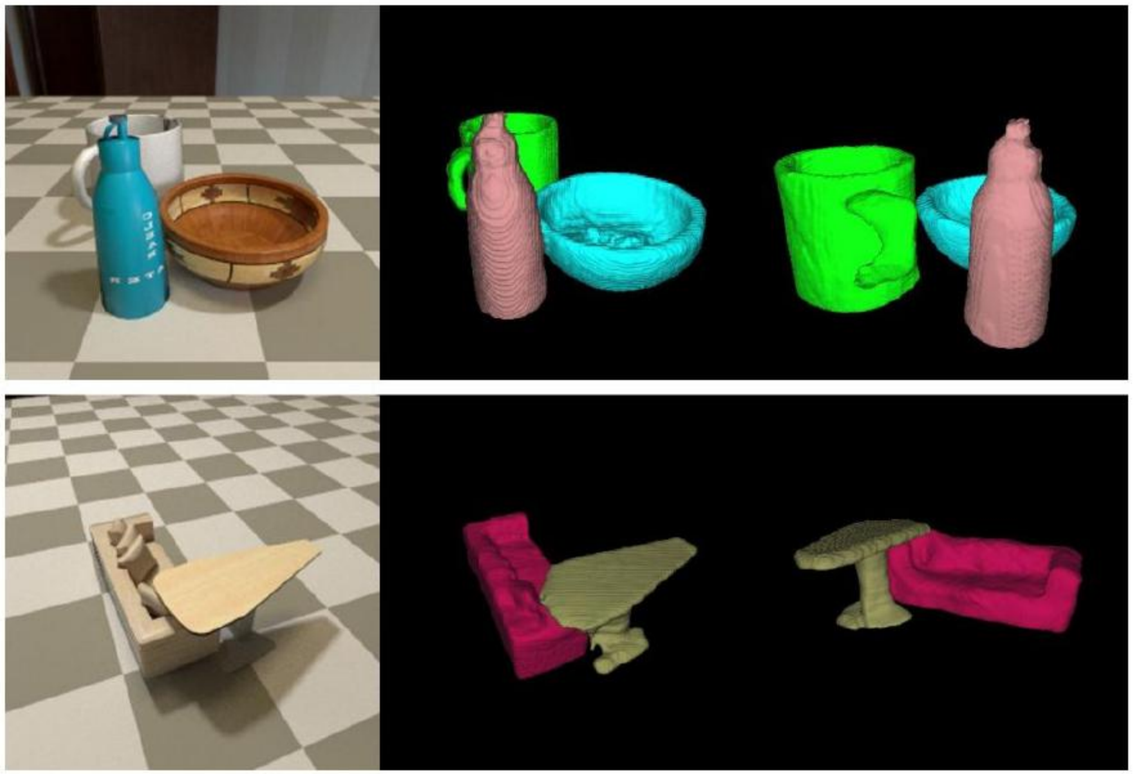
Stefan Popov, Pablo Bauszat, Vittorio Ferrari

ECCV 2020, oral presentation

Google Research

Contributions

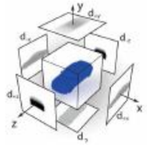
- Three improvements of common single object reconstruction architectures
- Coherent reconstruction of scenes with multiple objects
- Evaluation on a large-scale dataset with synthetic multi-object scenes



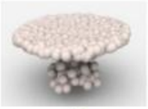
Related work: single object reconstruction

Voxel grids [Choy ECCV'16; Girdhar ECCV'16; Wu NIPS'16; Rezende NIPS'16; Gwak 3DV'17; Xie ICCV'19+IJCV'20]

Implicit functions [Park CVPR'19; Mescheder CVPR'19; Chen CVPR'19]



Fusion of fixed-viewpoint silhouettes [Soltani CVPR'17; Richter CVPR'18; Shin CVPR'18; Yao CVPR'20]



Point clouds [Fan CVPR'17; Mandikal BMVC'18]

Cuboidal primitives [Tulsiani CVPR'17; Niu CVPR'18]

Meshes [Liao CVPR'18; Wang ECCV'18; Henderson BMVC'18+IJCV'19; Kato CVPR '18; Chen CVPR '20]

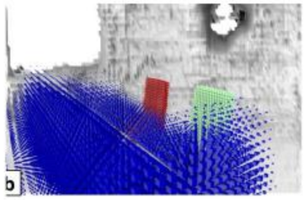
Typically on synthetic images with low realism (uniform background, local illumination):

Related work: multi object reconstruction

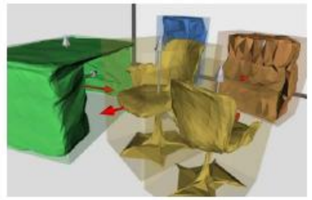
IM2CAD
[Izadinia CVPR'17]

3D-RCNN
[Kundu CVPR'18]

Mesh R-CNN
[Gkioxari ICCV'19]



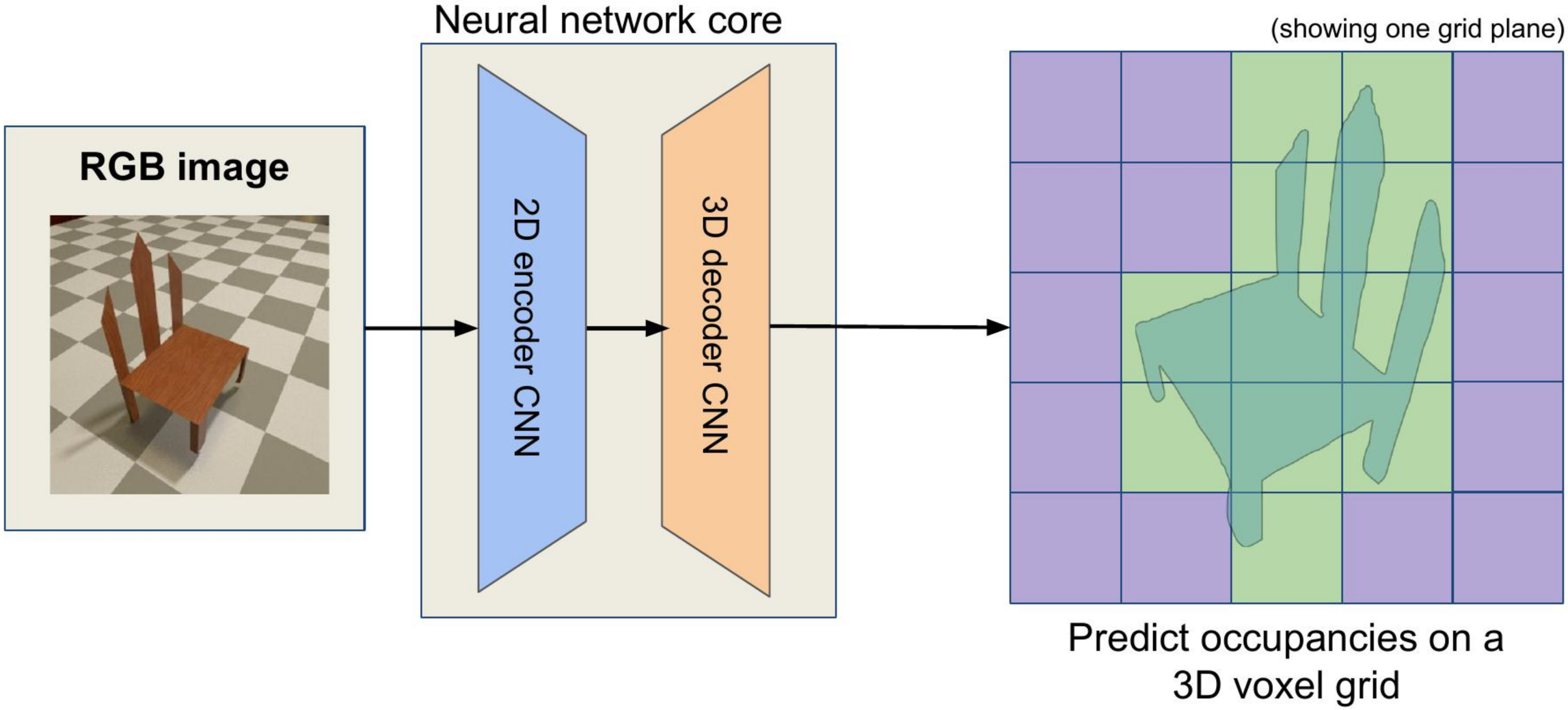
Factoring shape, pose, and layout
[Tulsiani CVPR'18]



Total3DUnderstanding
[Nie CVPR'20]

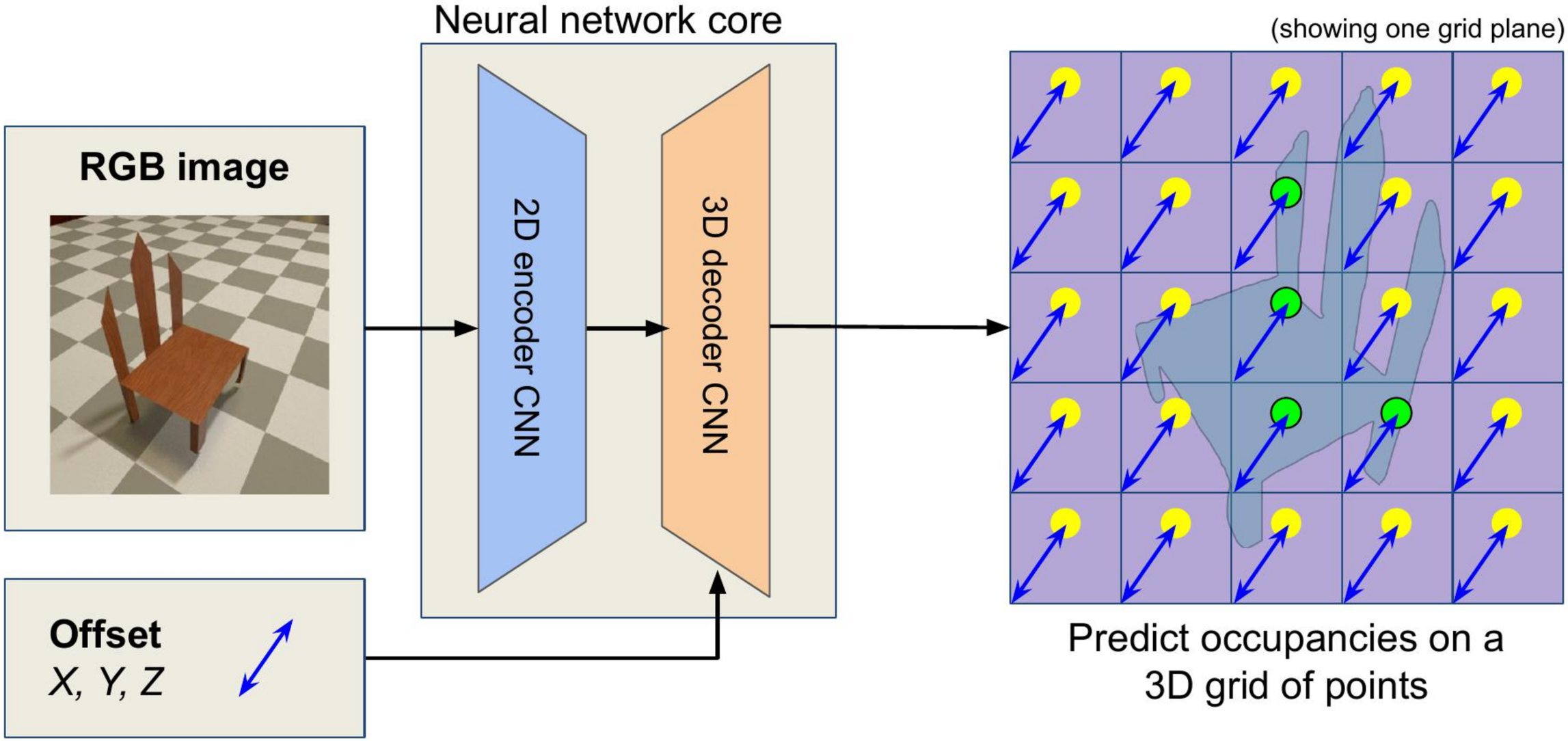
- Detect objects in 2D and then reconstruct their 3D shapes independently
- Reconstructed objects might intersect in 3D
- Most works don't evaluate 3D shape reconstruction for multiple object scenes

Starting point: Voxel grid encoder-decoder

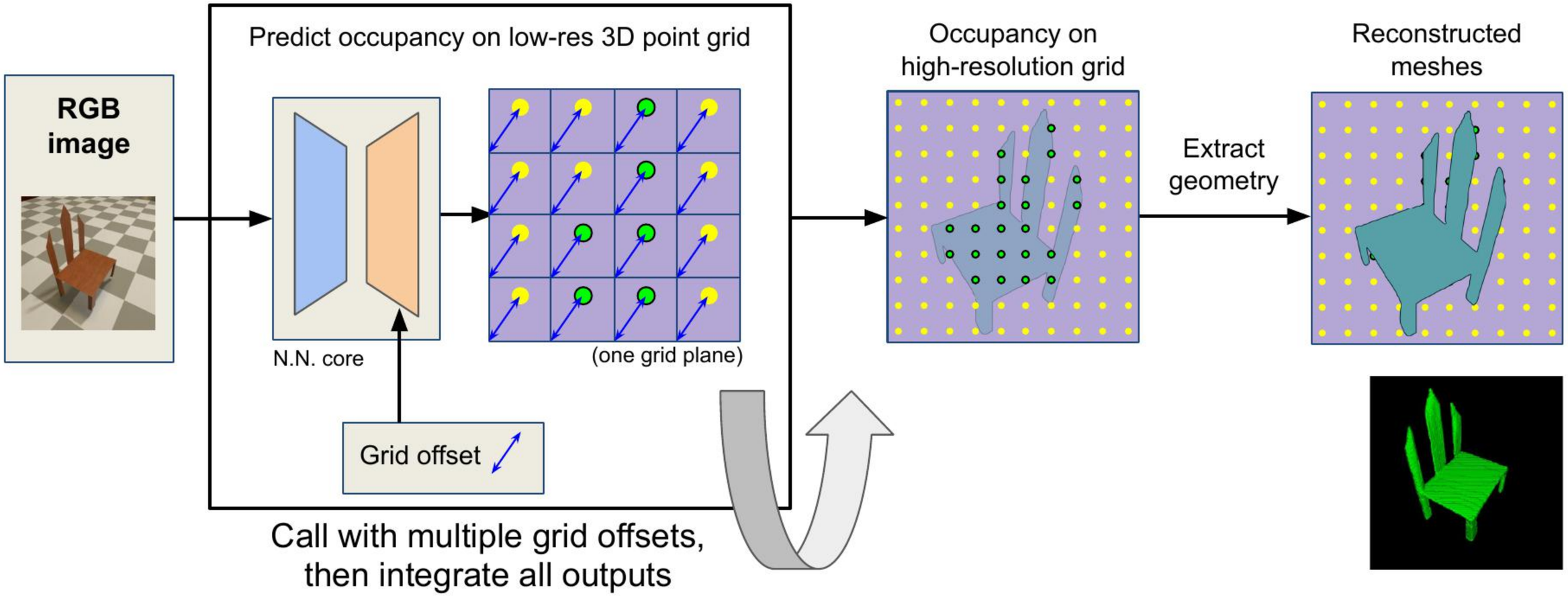


[Choy ECCV'16; Girdhar ECCV'16; Xie ICCV'19; Xie IJCV'20]

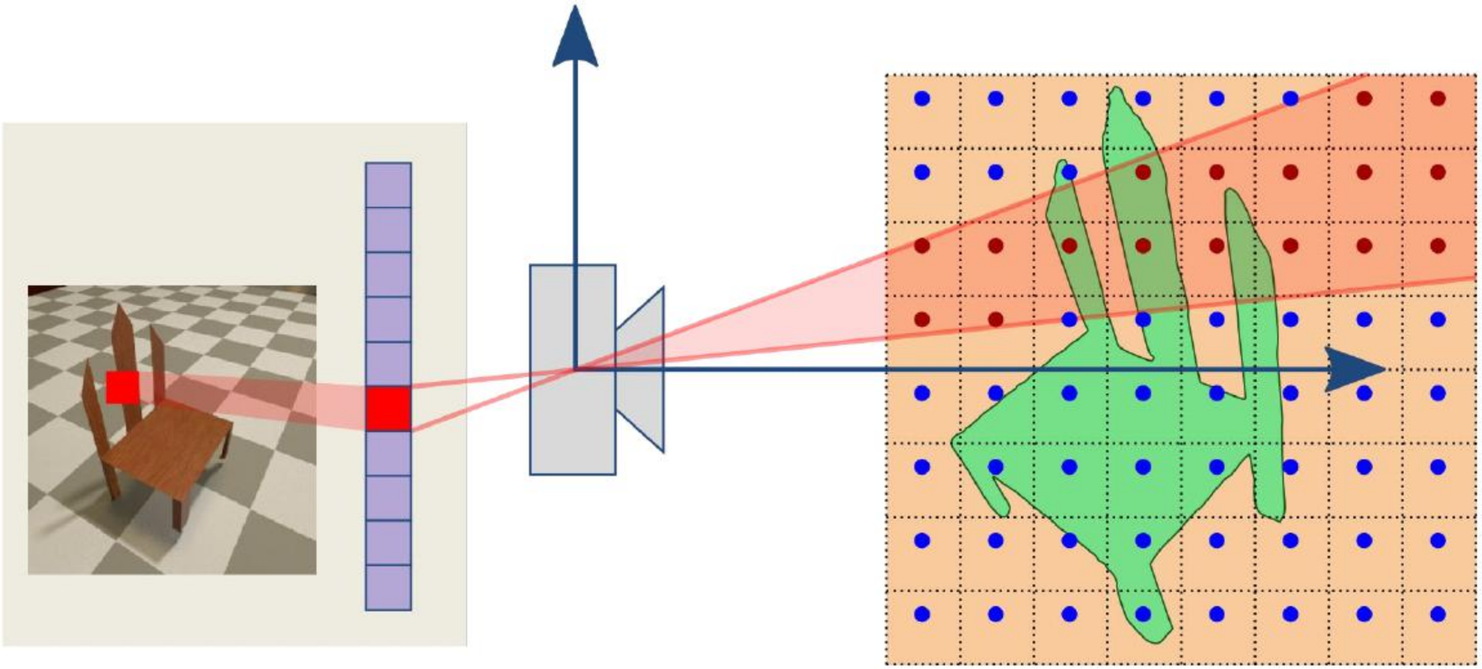
Improvement 1: Hybrid 3D volume representation



Fine object details without an excessive memory footprint

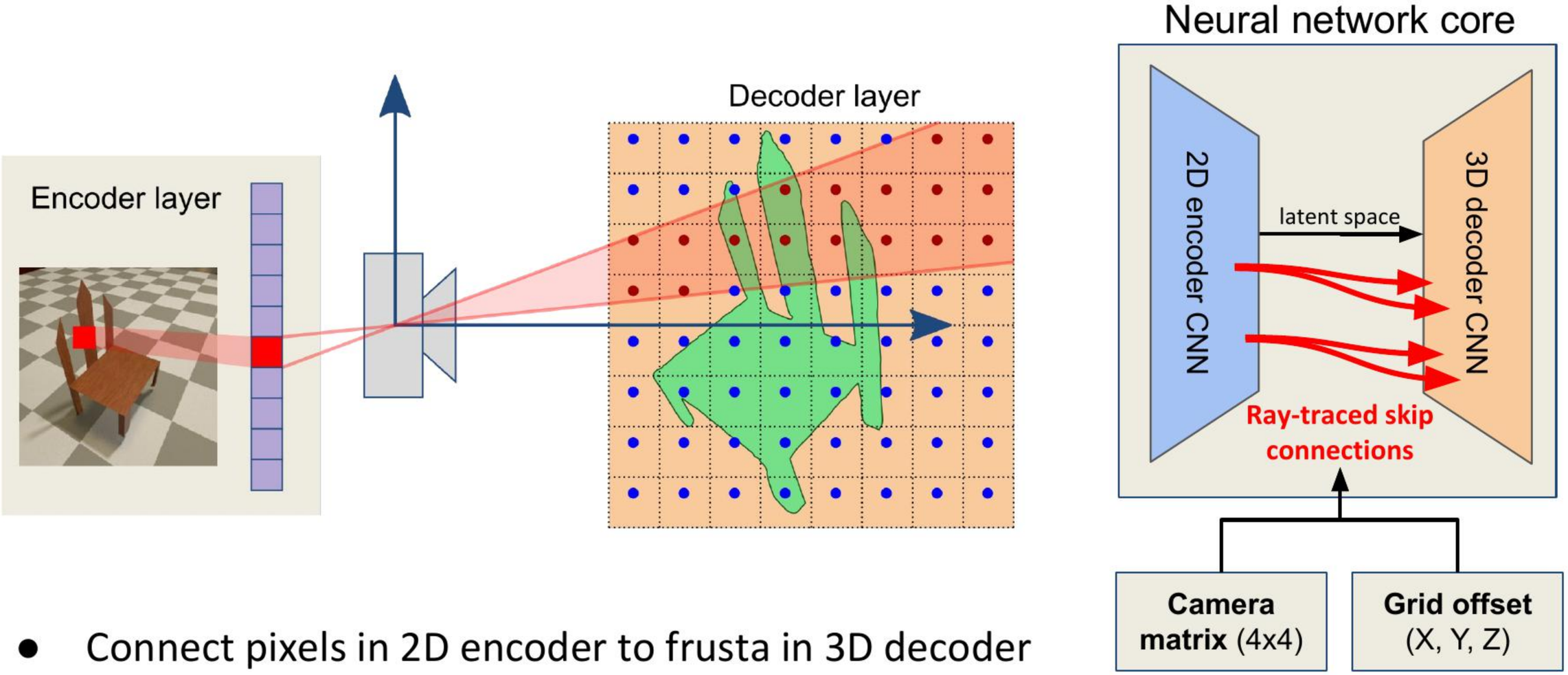


Improvement 2: Ray-traced skip connections



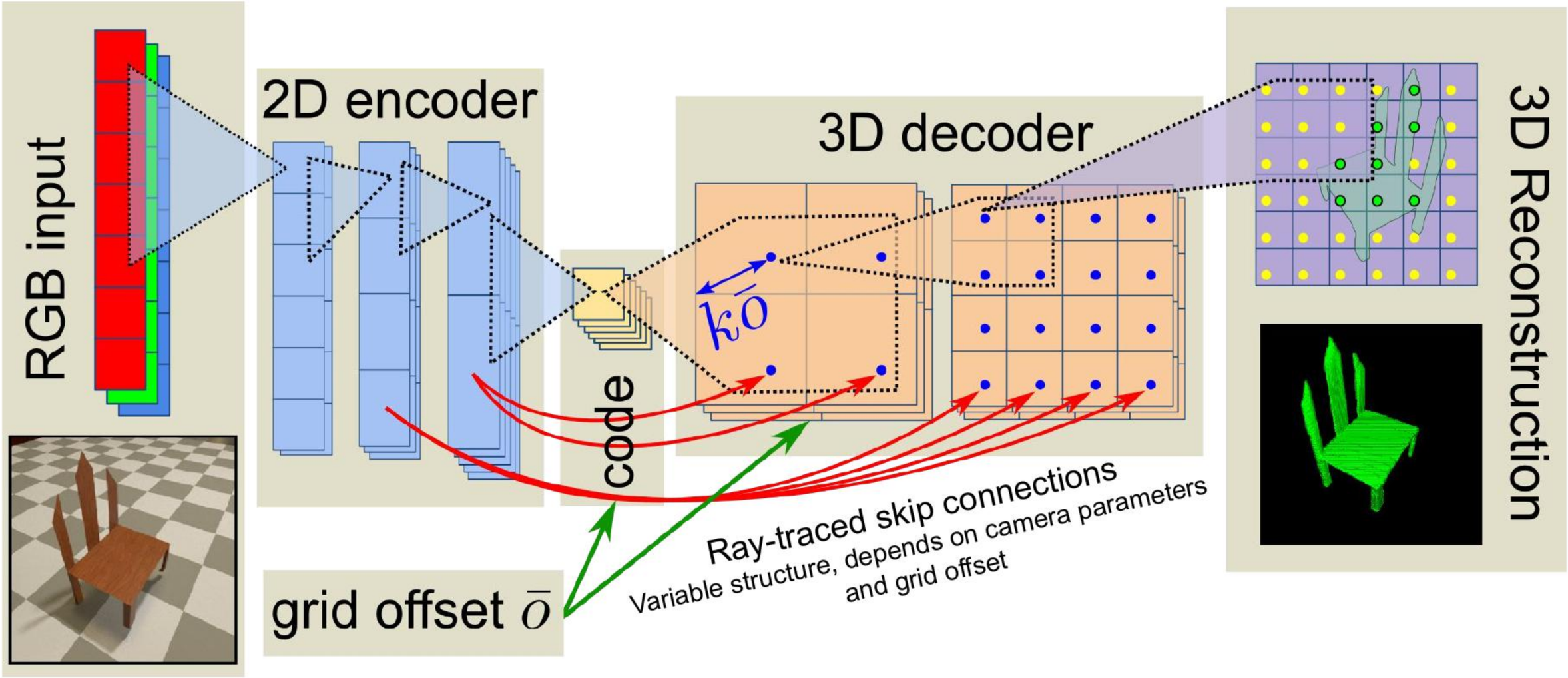
- Reconstruct in camera space
- Assume known camera intrinsics

Improvement 2: Ray-traced skip connections



- Connect pixels in 2D encoder to frusta in 3D decoder
- Propagates local image information into 3D, resulting in sharp reconstruction details.

Ray-traced skip connections enhance our hybrid representation



Improvement 3: IoU Loss (intersection-over-union)

- Volume is mostly empty, leading to class imbalance
- Focus on overall object geometry rather than independent points
- Generalizes the standard IoU metric

$$IoU_g(g,p) = \frac{\sum_{i \in G} \min(g_i \ , p_i \ )}{\sum_{i \in G} \max(g_i \ , p_i \ )}$$

Improvement 3: IoU Loss (intersection-over-union)

- Support for multiple classes

$$IoU_g(g,p) = \frac{\sum_{i \in G} \sum_{c=1}^{C-1} \min(g_{ic}, p_{ic})}{\sum_{i \in G} \sum_{c=1}^{C-1} \max(g_{ic}, p_{ic})}$$

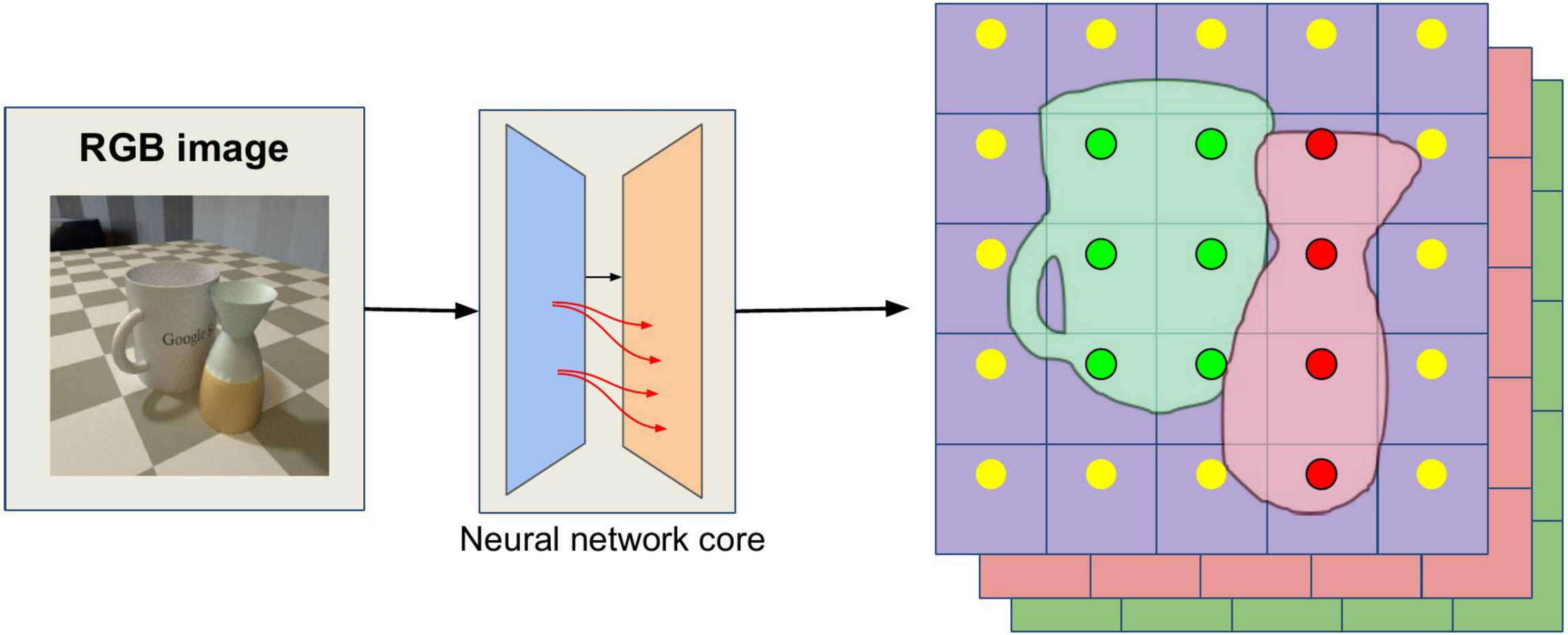
Improvement 3: IoU Loss (intersection-over-union)

- Support for multiple classes
- Balance for additional sparsity of multiple classes

$$IoU_g(g,p) = \frac{\sum_{i \in G} \sum_{c=1}^{C-1} \min(g_{ic}, p_{ic}) \cdot \mu(g_{ic})}{\sum_{i \in G} \sum_{c=1}^{C-1} \max(g_{ic}, p_{ic}) \cdot \mu(g_{ic})}$$

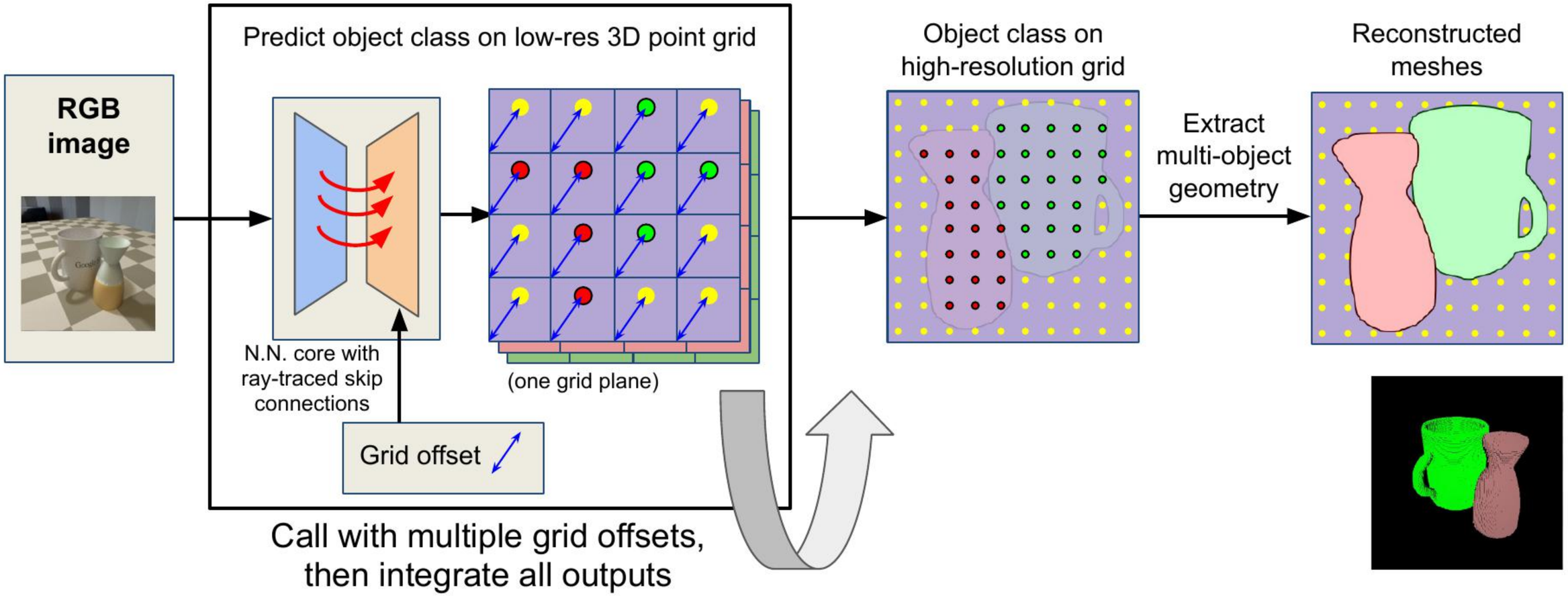
$$\mu(g_{ic}) = \begin{cases} 1 & \text{, if } g_{ic} = 1 \\ \frac{1}{C-1} & \text{, if } g_{ic} = 0 \end{cases}$$

Reconstructing multiple objects in the same image



Predict multinomial distribution over classes, for each point of the grid

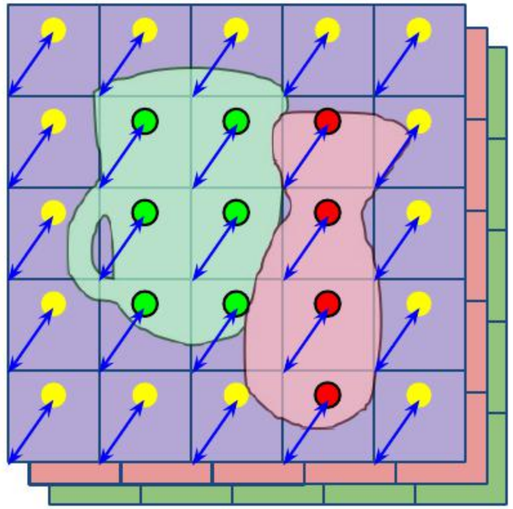
Multi-object reconstruction pipeline



Our improvements help even more with multiple objects

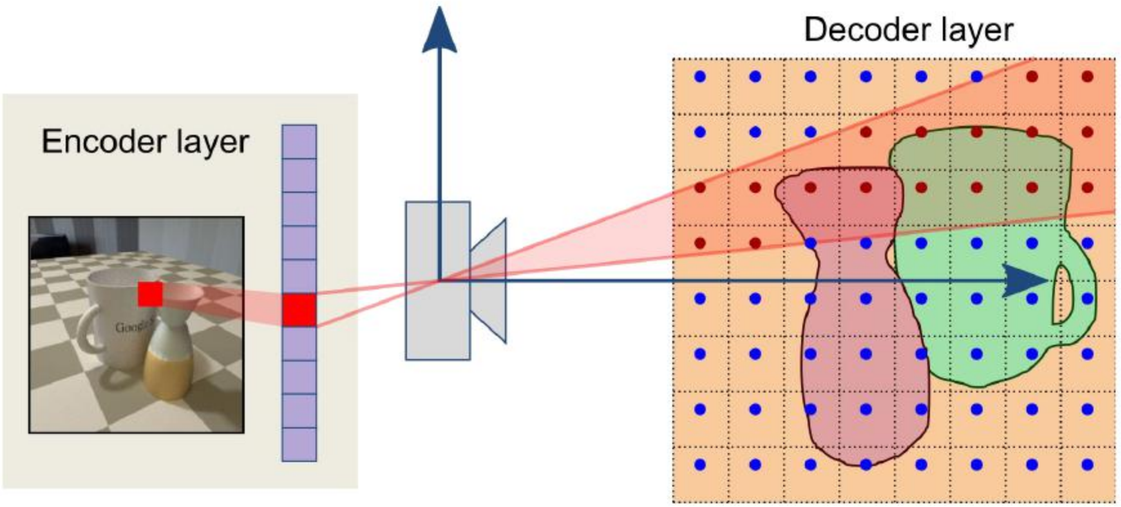
Hybrid volume representation:

- objects can be anywhere in space
- occupy a smaller fraction of the volume



Ray-traced skip connections:

- propagate occlusion boundaries and contact points from 2D to 3D
- help understand depth relations among objects locally



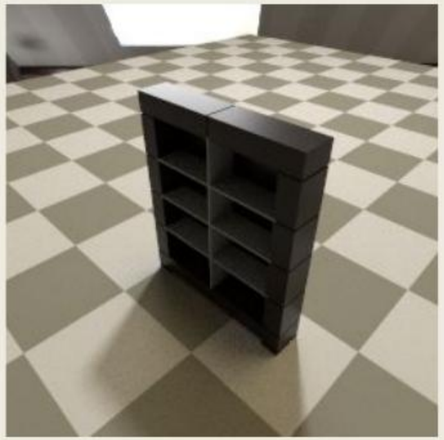
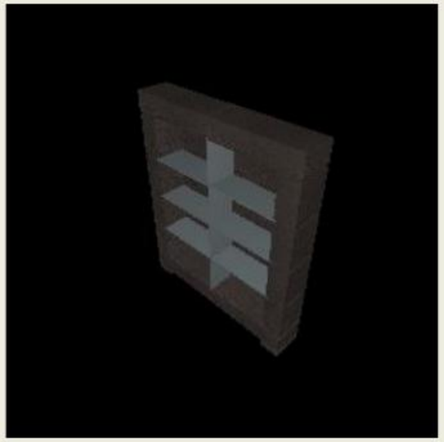
Our improvements help even more with multiple objects

- Focus on overall object geometry rather than independent points
- Generalizes the standard IoU metric

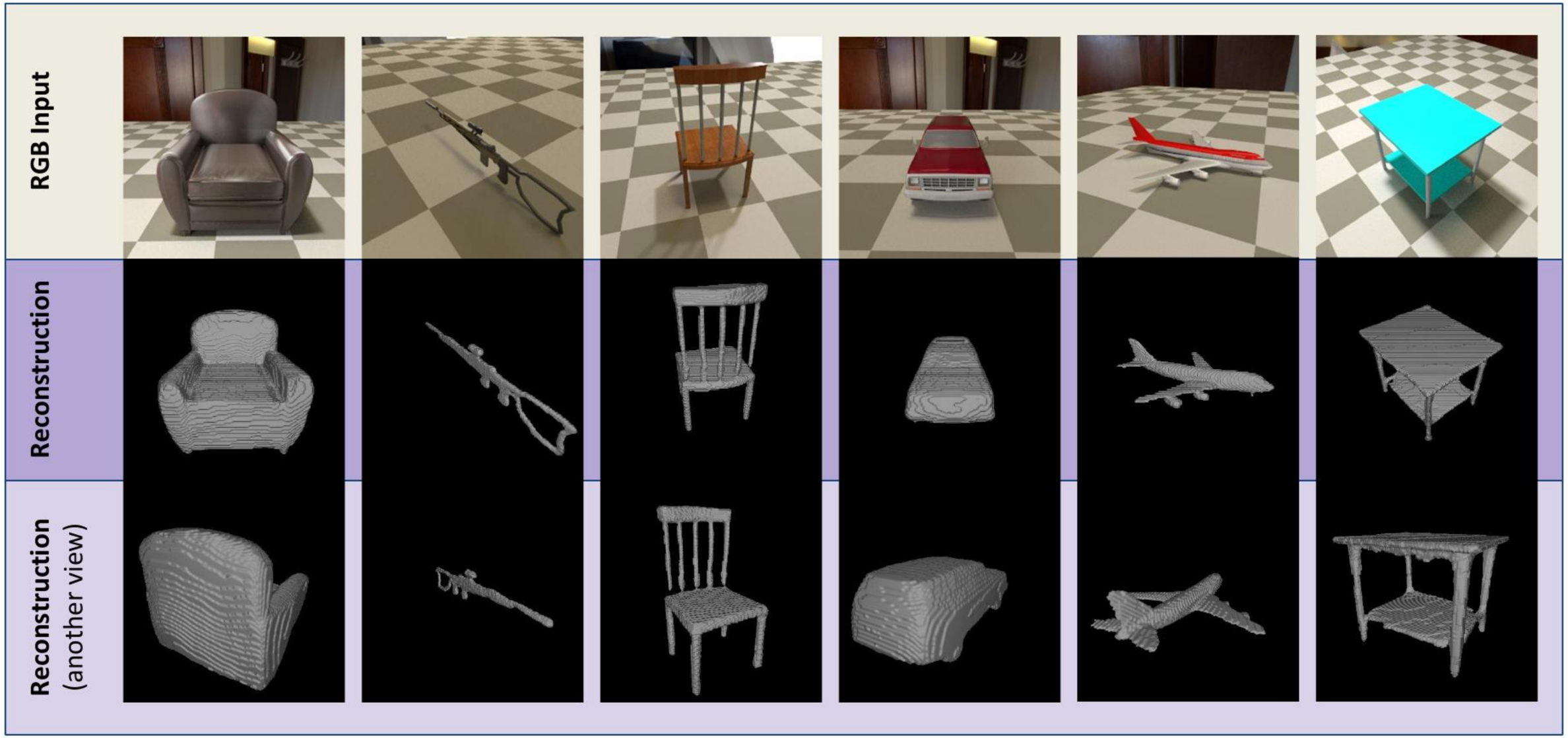
$$IoU_g(g,p) = \frac{\sum_{i \in G} \sum_{c=1}^{C-1} \min(g_{ic}, p_{ic}) \cdot \mu(g_{ic})}{\sum_{i \in G} \sum_{c=1}^{C-1} \max(g_{ic}, p_{ic}) \cdot \mu(g_{ic})}$$

$$\mu(g_{ic}) = \begin{cases} 1 & \text{, if } g_{ic} = 1 \\ \frac{1}{C-1} & \text{, if } g_{ic} = 0 \end{cases}$$

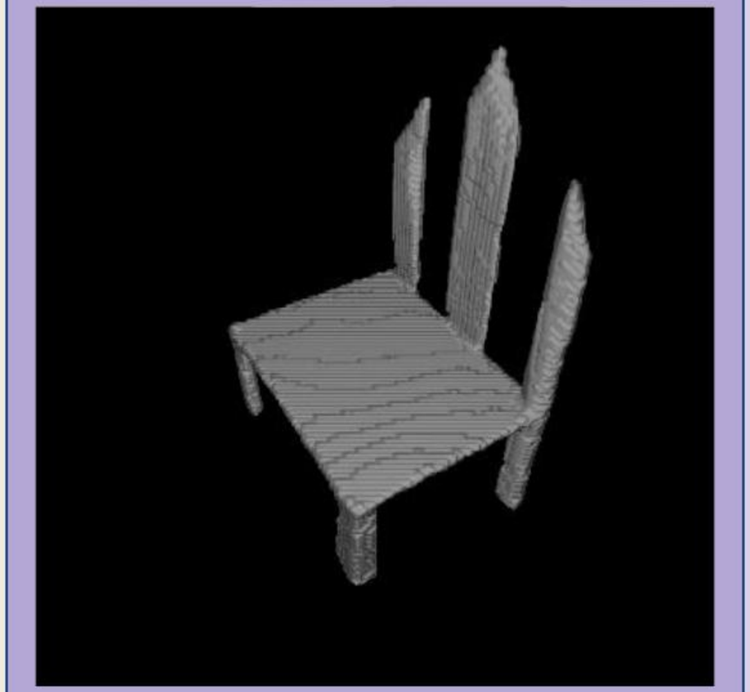
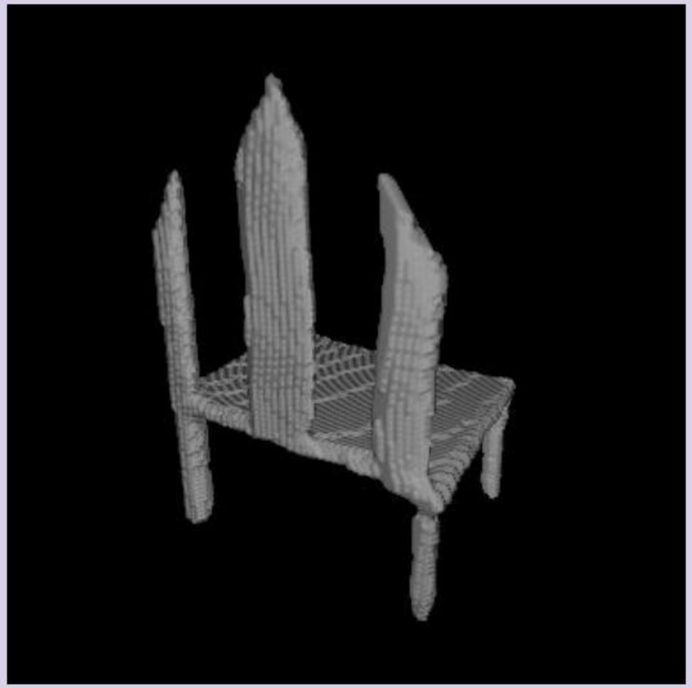
Experiment datasets

ShapeNet	Pix3D	ShapeNet - Pairs	ShapeNet - Triplets
  Setting of [Choy ECCV'16]	Setting of [Gkioxari ICCV'19], [Sun CVPR'18]	7 pairs out of 13 classes 366K train, 91K test	2 triplets out of 6 classes 91K train, 23K test

Qualitative results for single object reconstruction

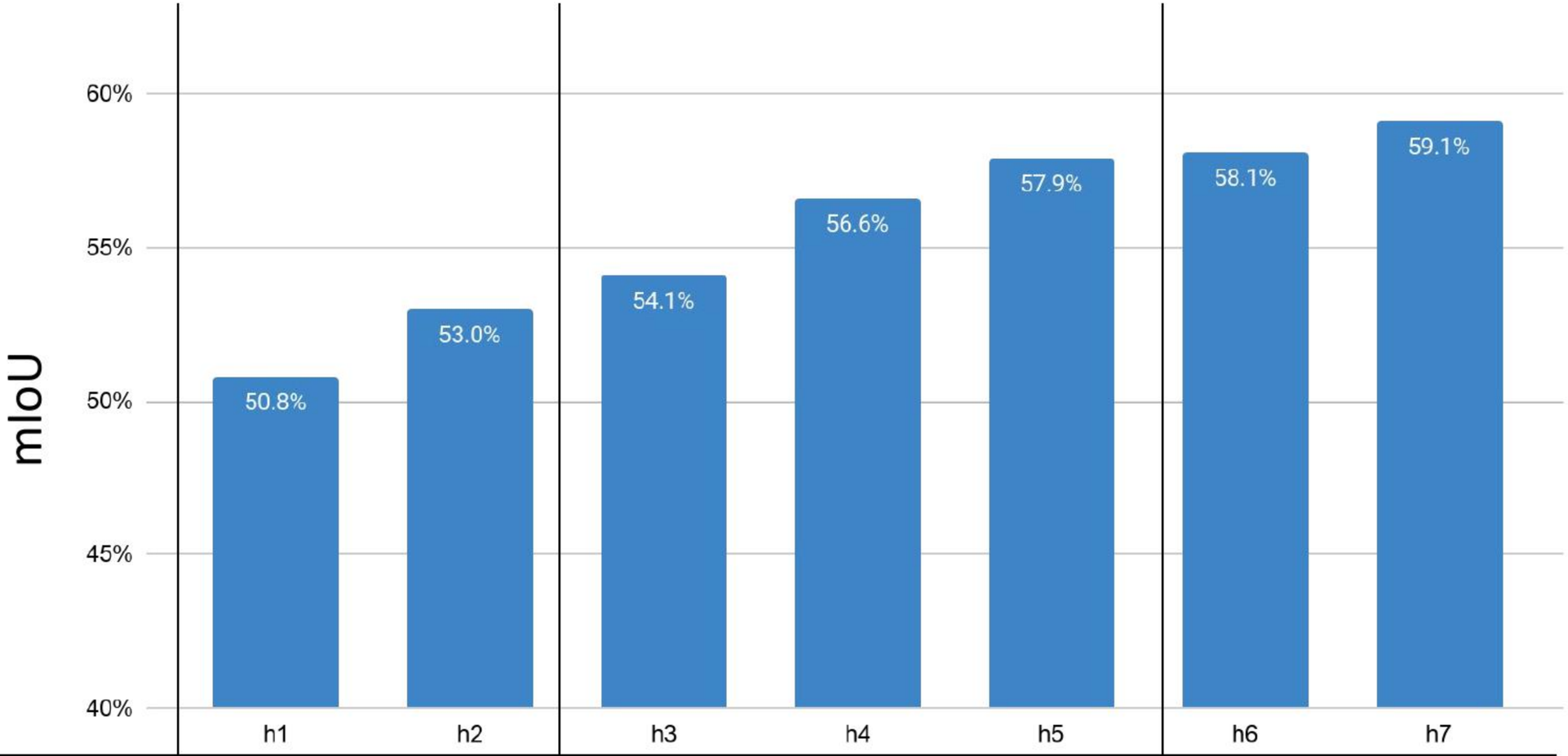


Qualitative results for single object reconstruction

RGB Input	Reconstruction	Reconstruction (another view)
		

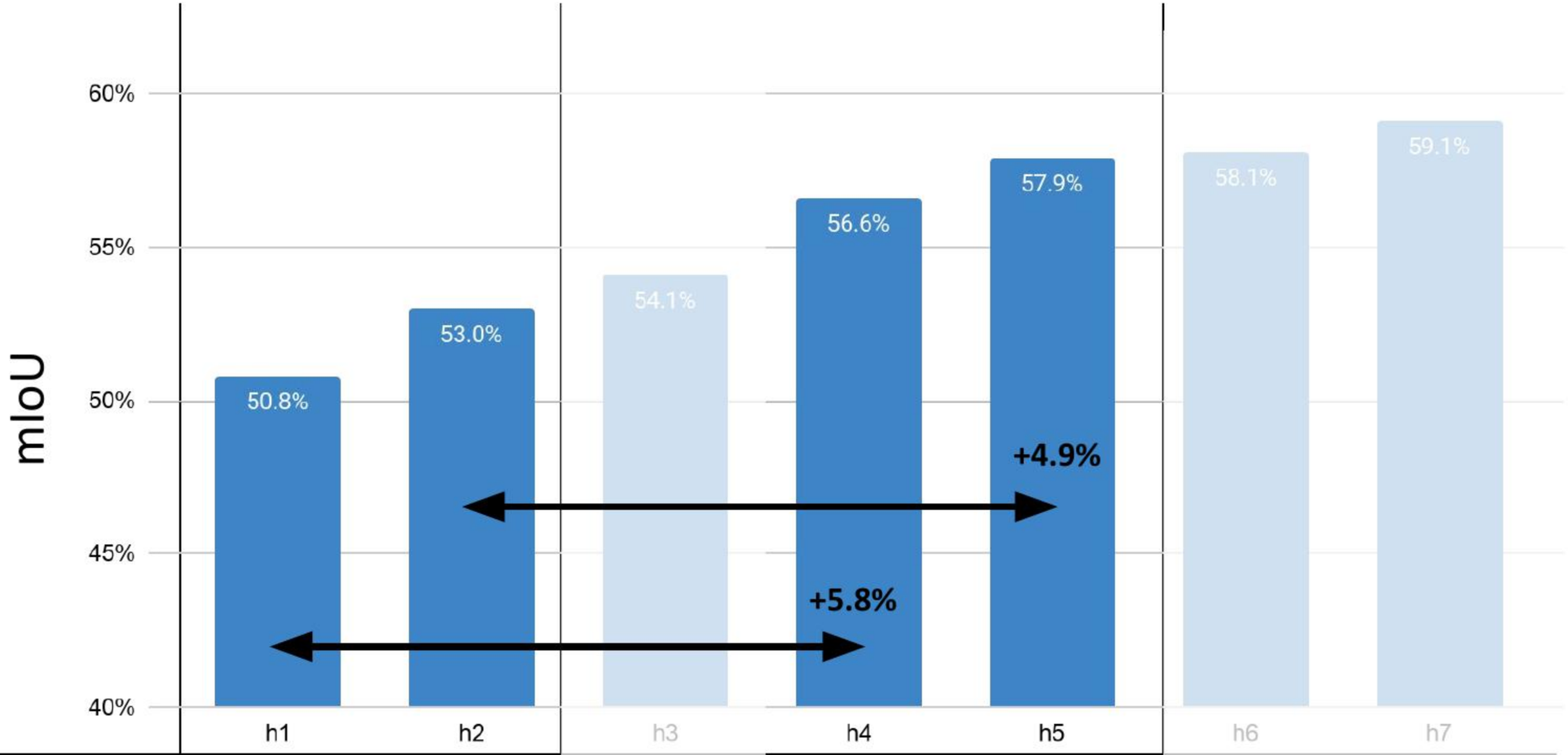
Properly reconstructs hidden parts and handles thin structures

Impact of our contributions



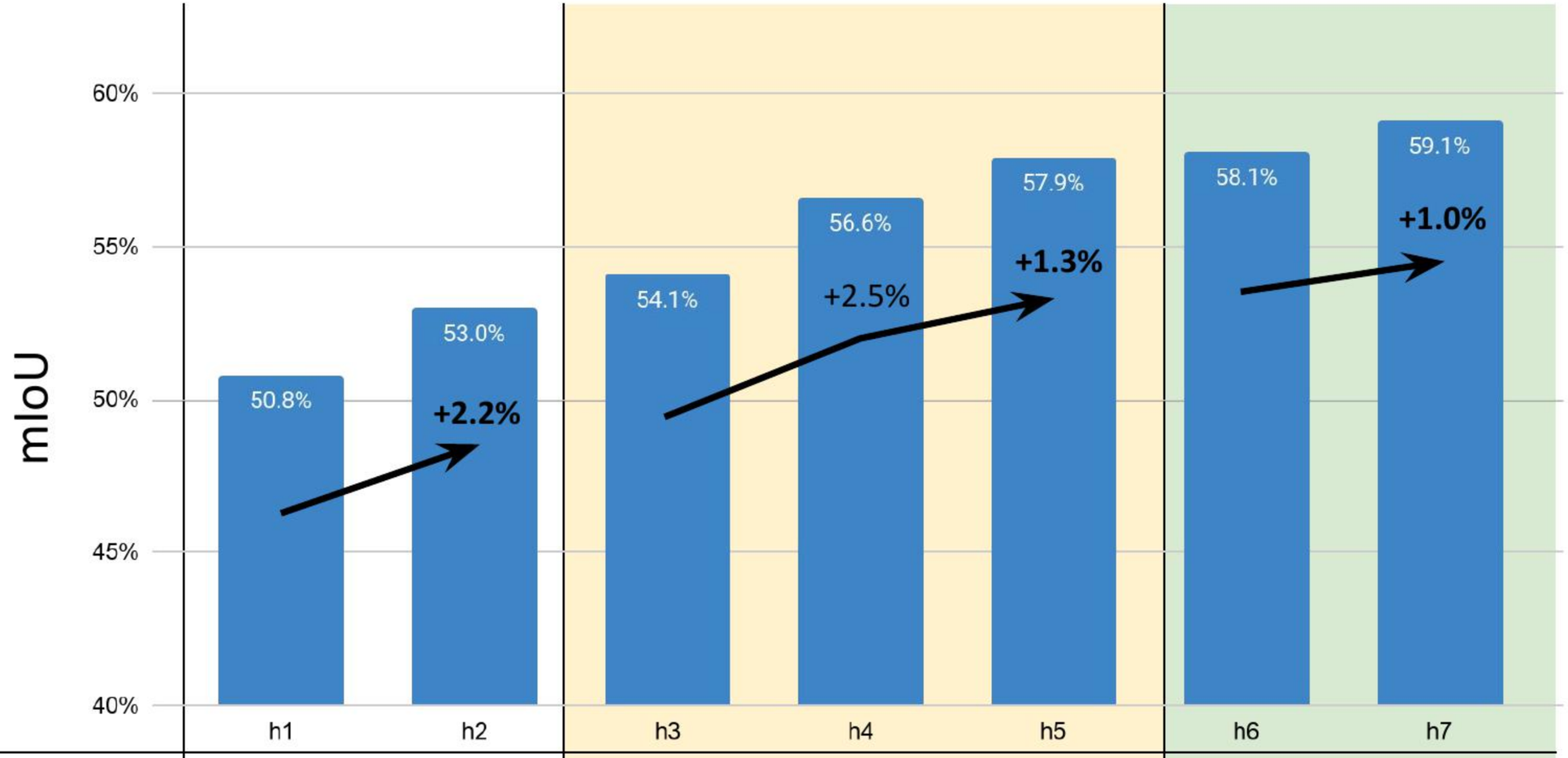
loss	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes			yes	
realism	low		low			high	

Ray-traced skip connections improve mIoU substantially



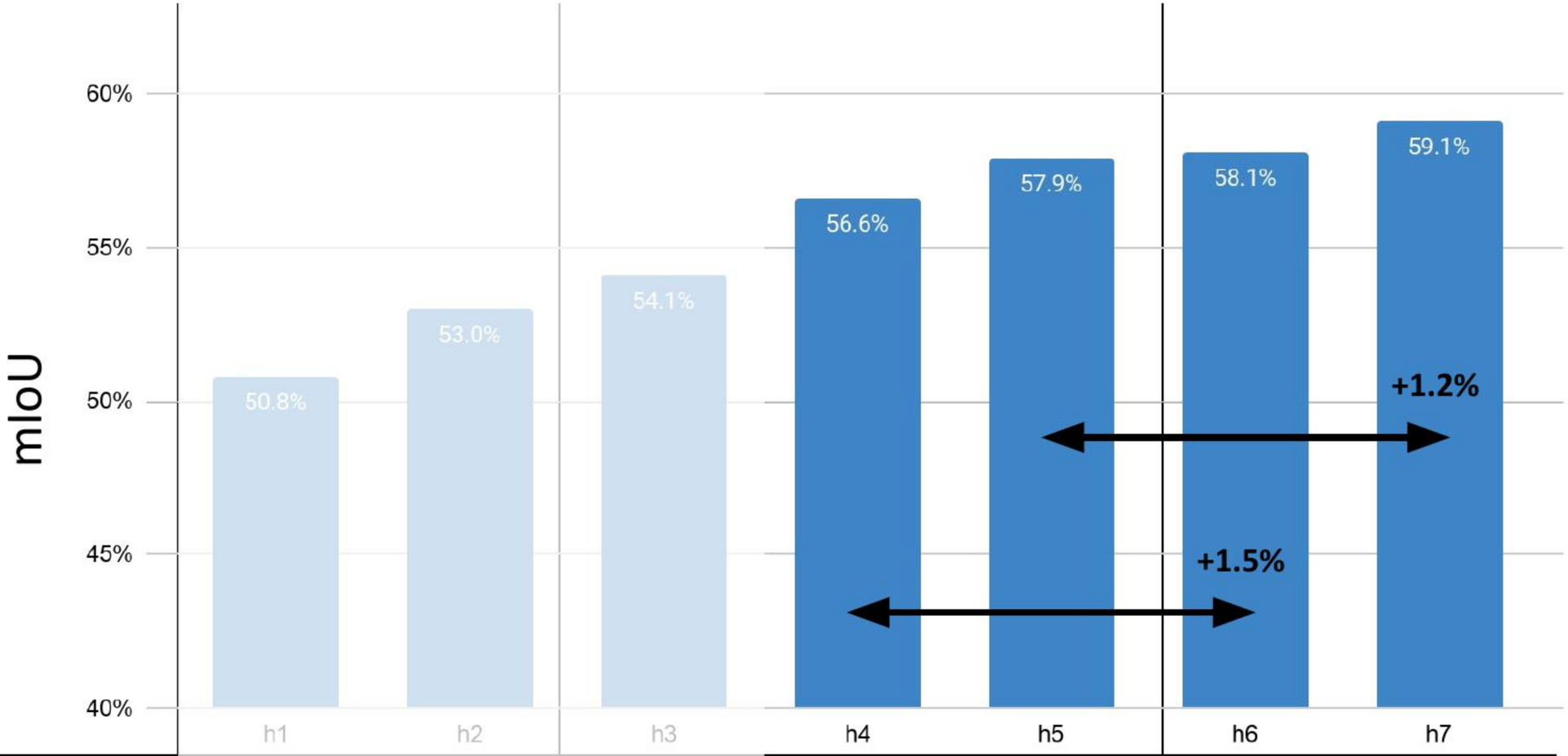
loss	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes			yes	
realism	low		low			high	

Our IoU loss performs best



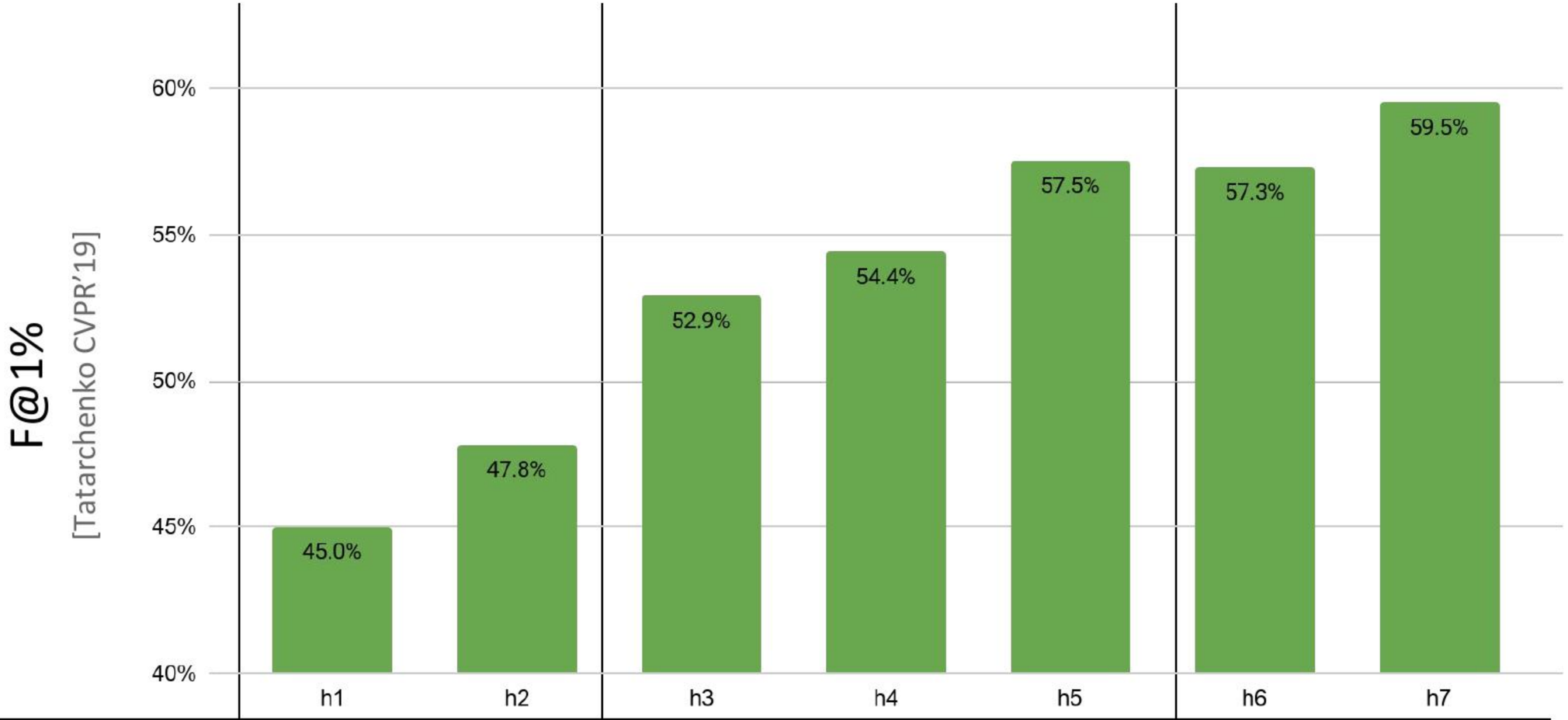
loss	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes			yes	
realism	low		low			high	

Better performance on high compared to low realism



loss	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes			yes	
realism	low		low			high	

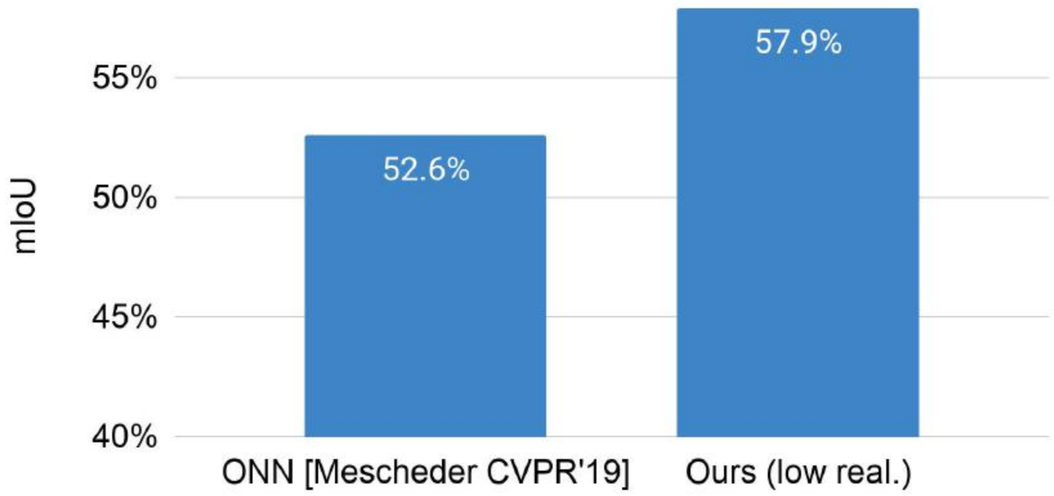
Same trends using a surface metric (F-score)



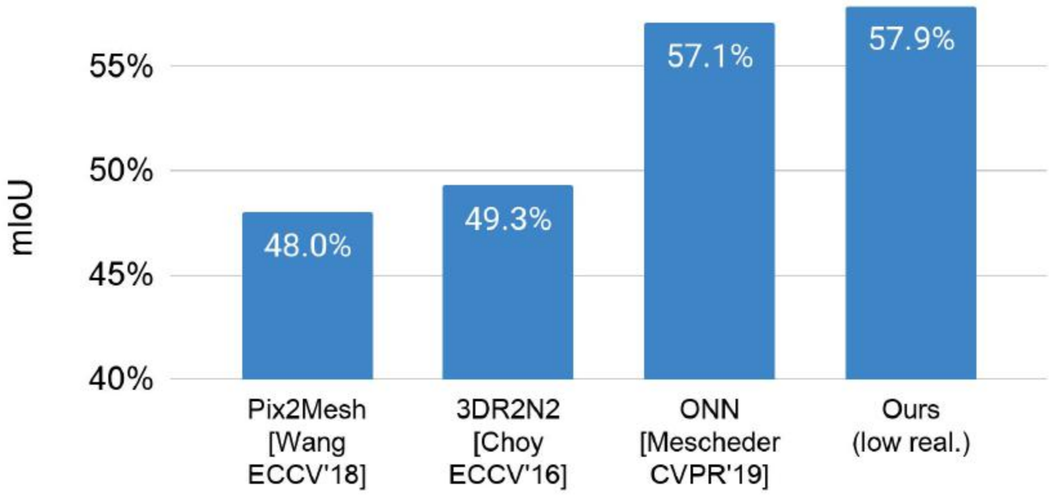
loss	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes			yes	
realism	low		low			high	

Comparison to state-of-the-art (see paper)

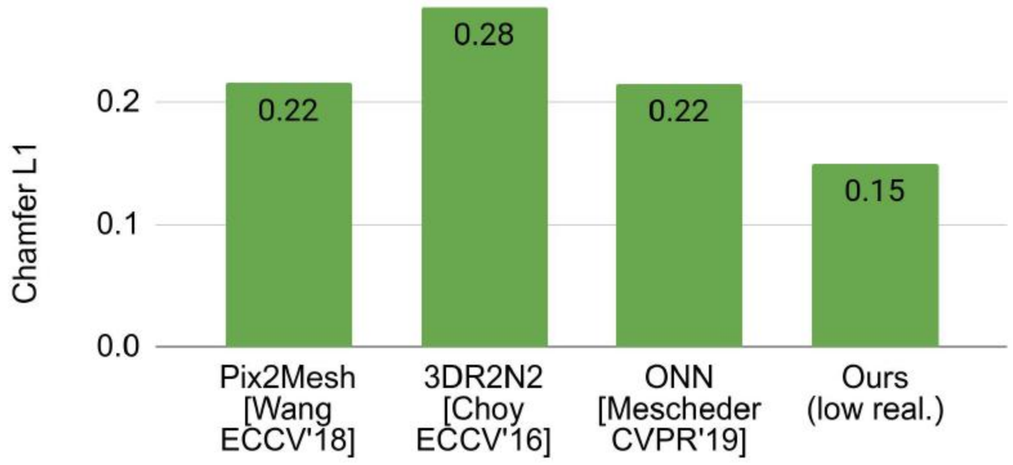
mIoU, in our setting



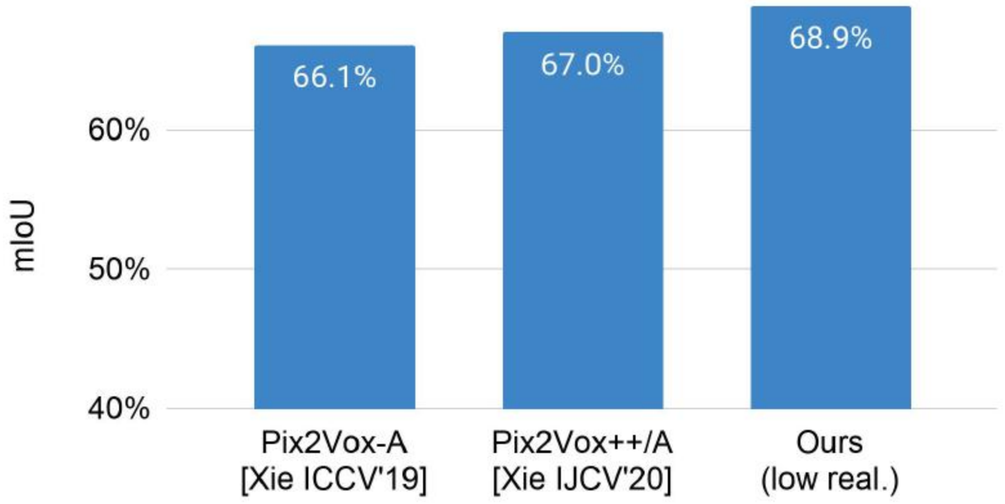
mIoU, as reported in [Mescheder CVPR'19]

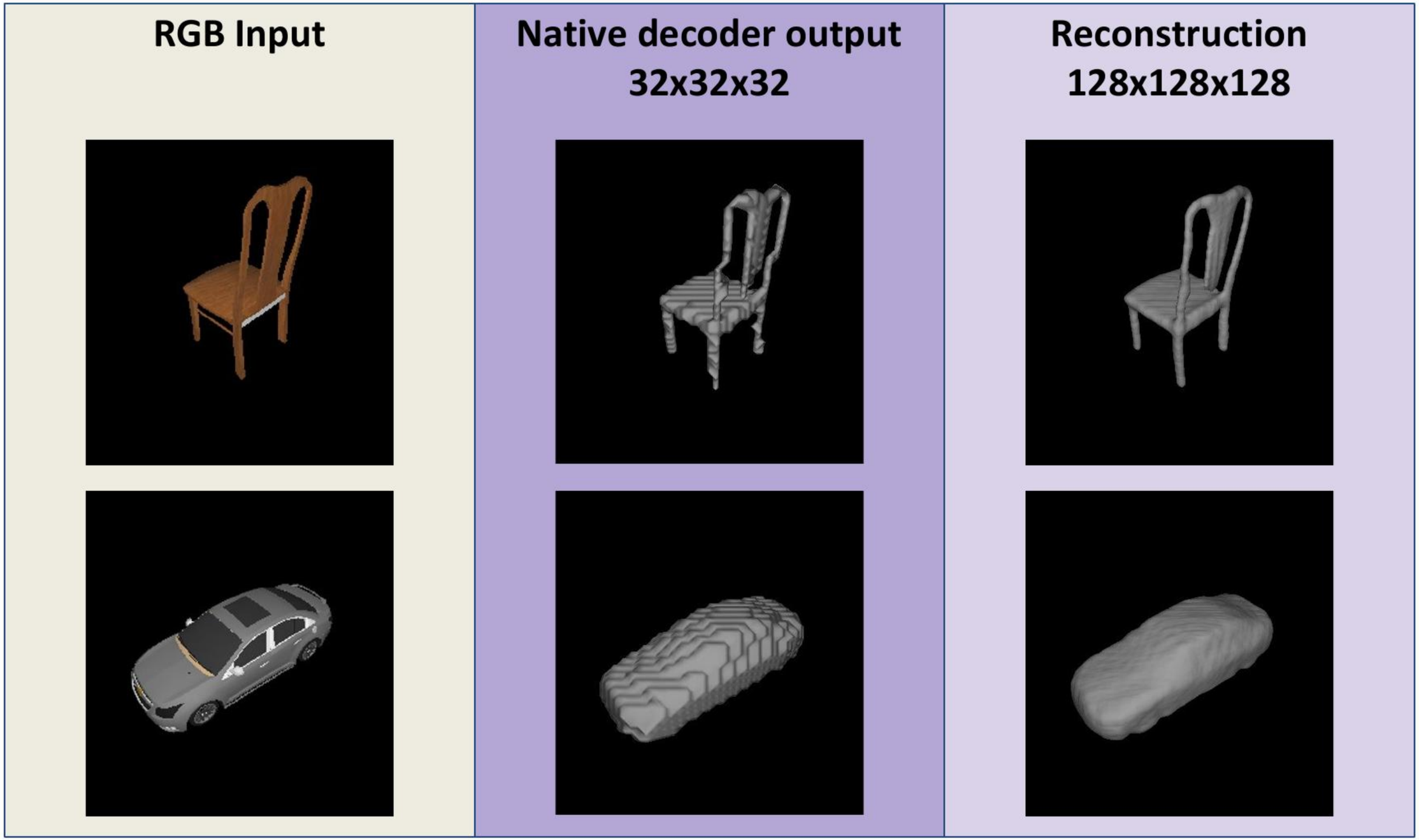


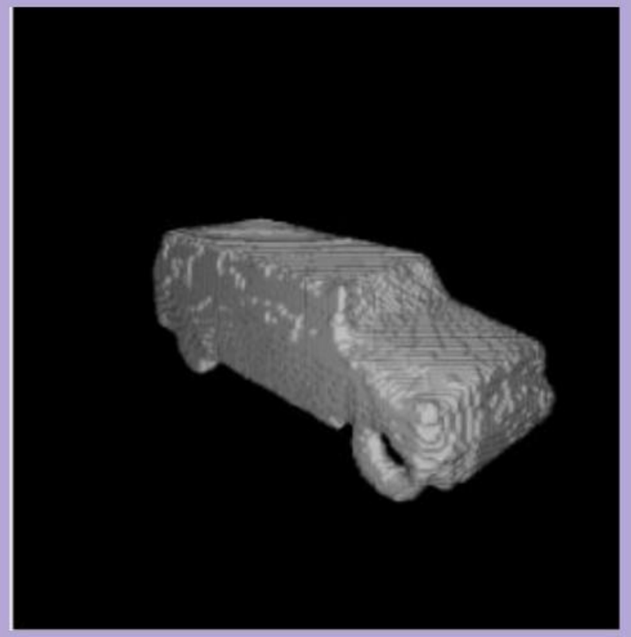
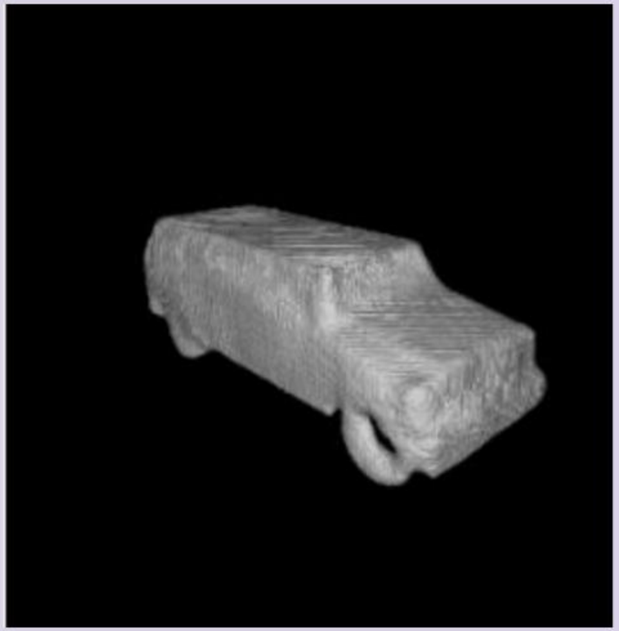
Chamfer distance, as reported in [Mescheder CVPR'19]



mIoU, measured on 32^3 voxelization (Pix2Vox)

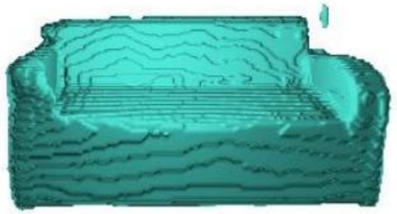



RGB Input
Native decoder output
32x32x32
Reconstruction
128x128x128

RGB Input	Native decoder output 128x128x128	Reconstruction 256x256x256
		

Single object reconstruction from real images (Pix3D)

Reconstruction

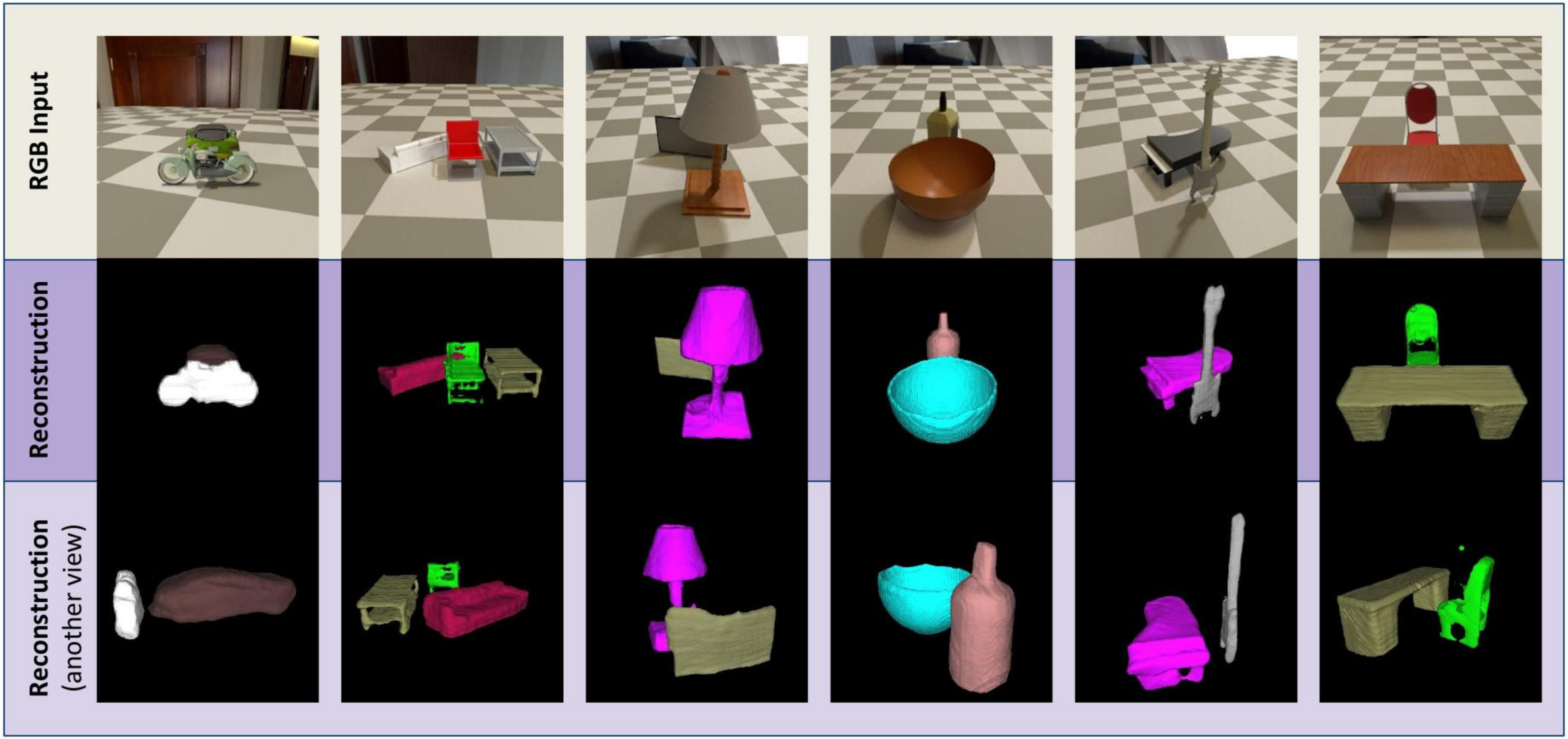


Reconstruction
(other views)

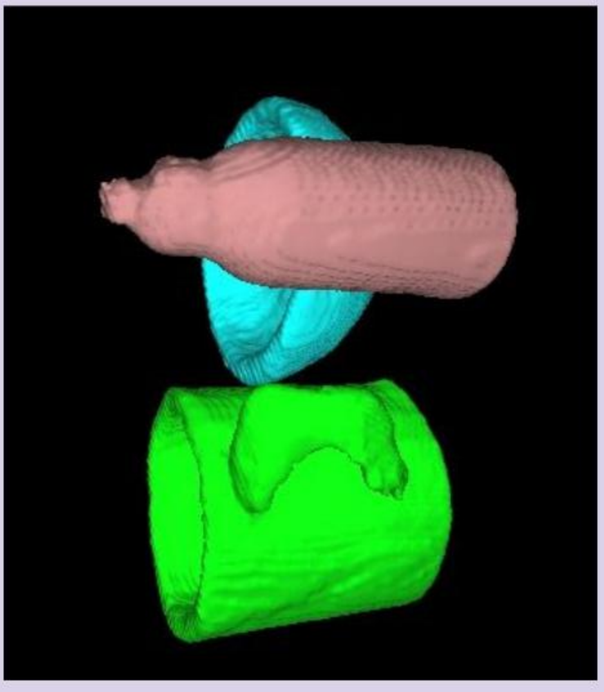
Single object reconstruction from real images (Pix3D)

See paper for quantitative comparison to 3D-R2N2 [Choy ECCV'16],
the Pix3D method [Sun CVPR'18], Pix2Vox [Xie ICCV'19], and Pix2Vox++ [Xie IJCV'20]

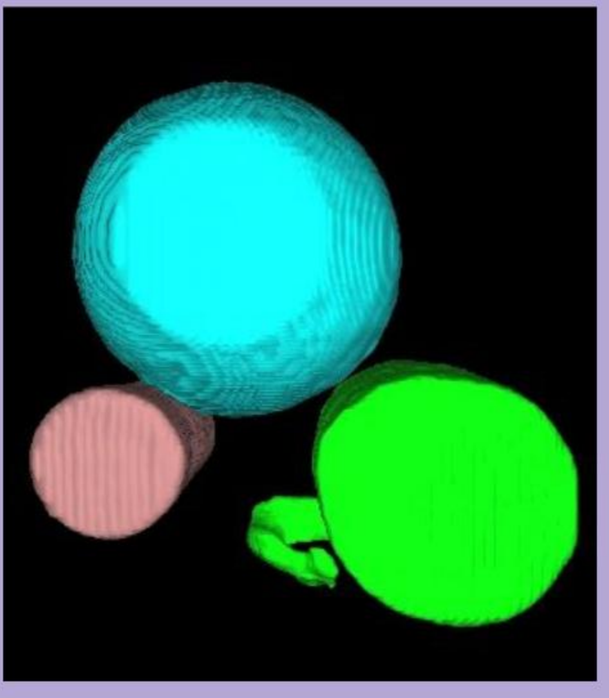
Qualitative results for multi-object reconstruction



Reconstruction
(another view)

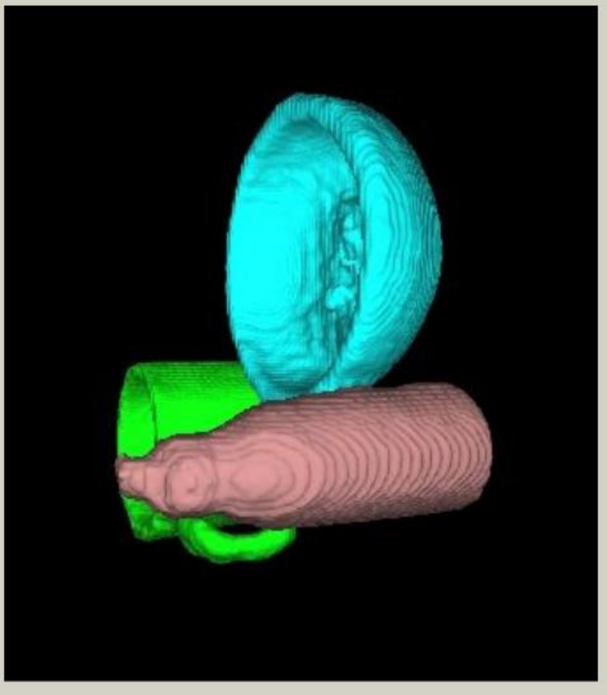


Reconstruction
(from below)

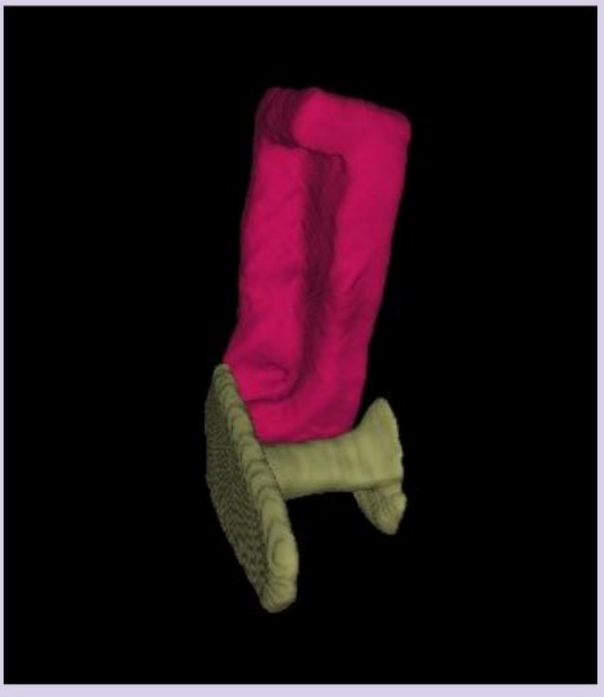


RGB Input

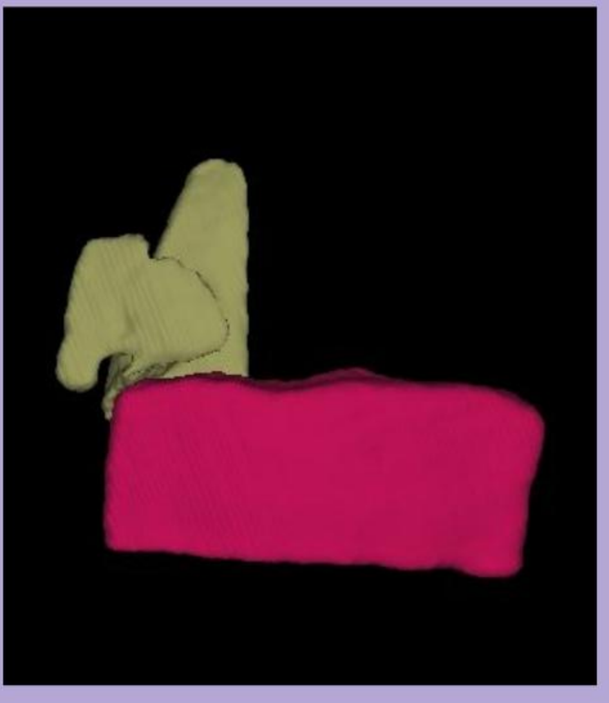
Reconstruction



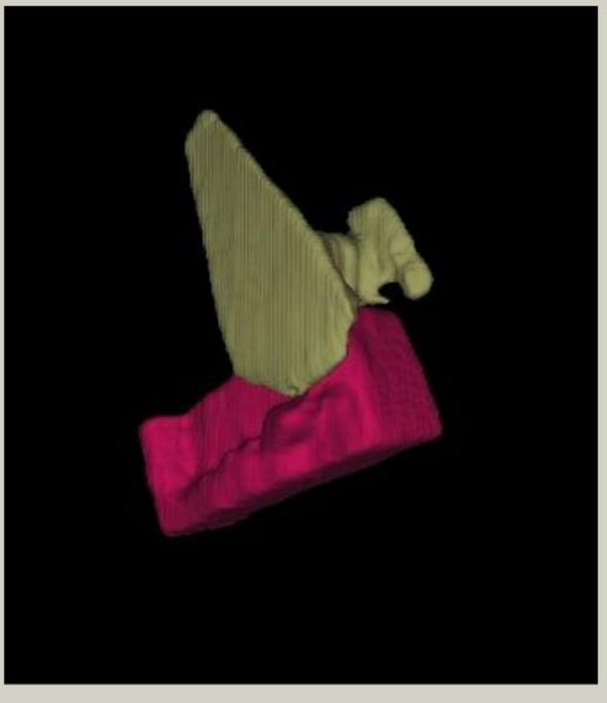
Reconstruction
(another view)



Reconstruction
(from below)

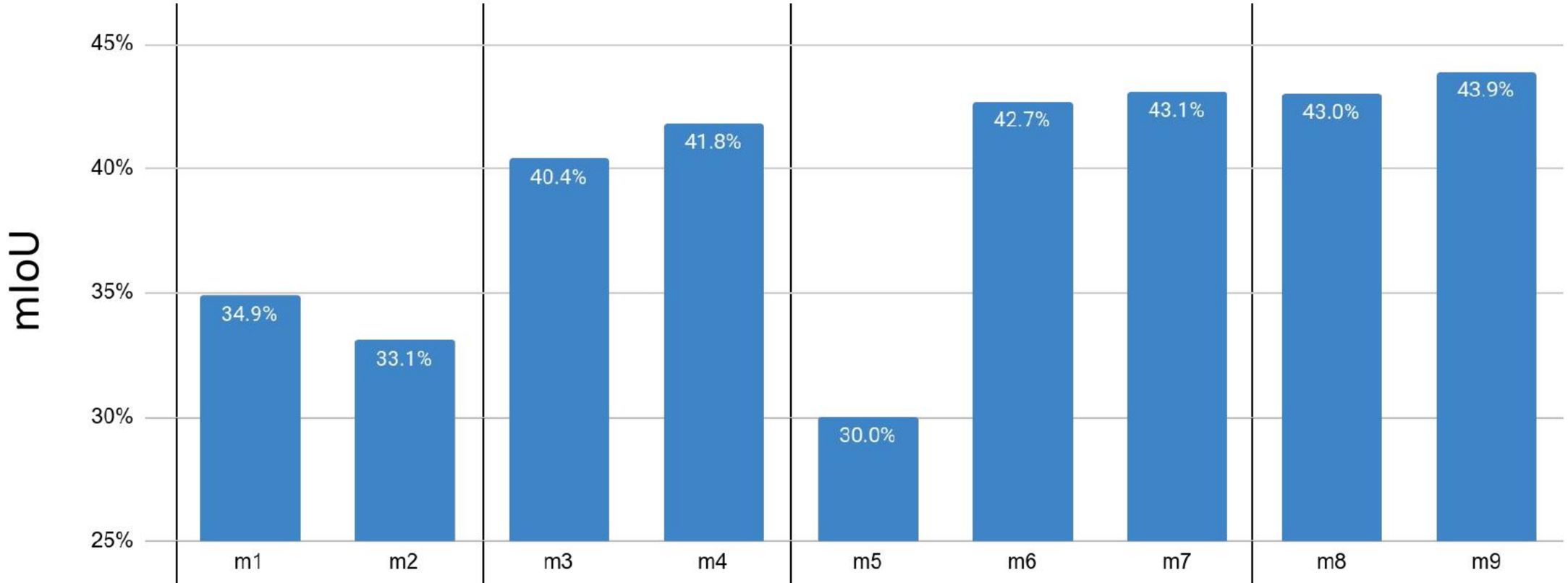


RGB Input



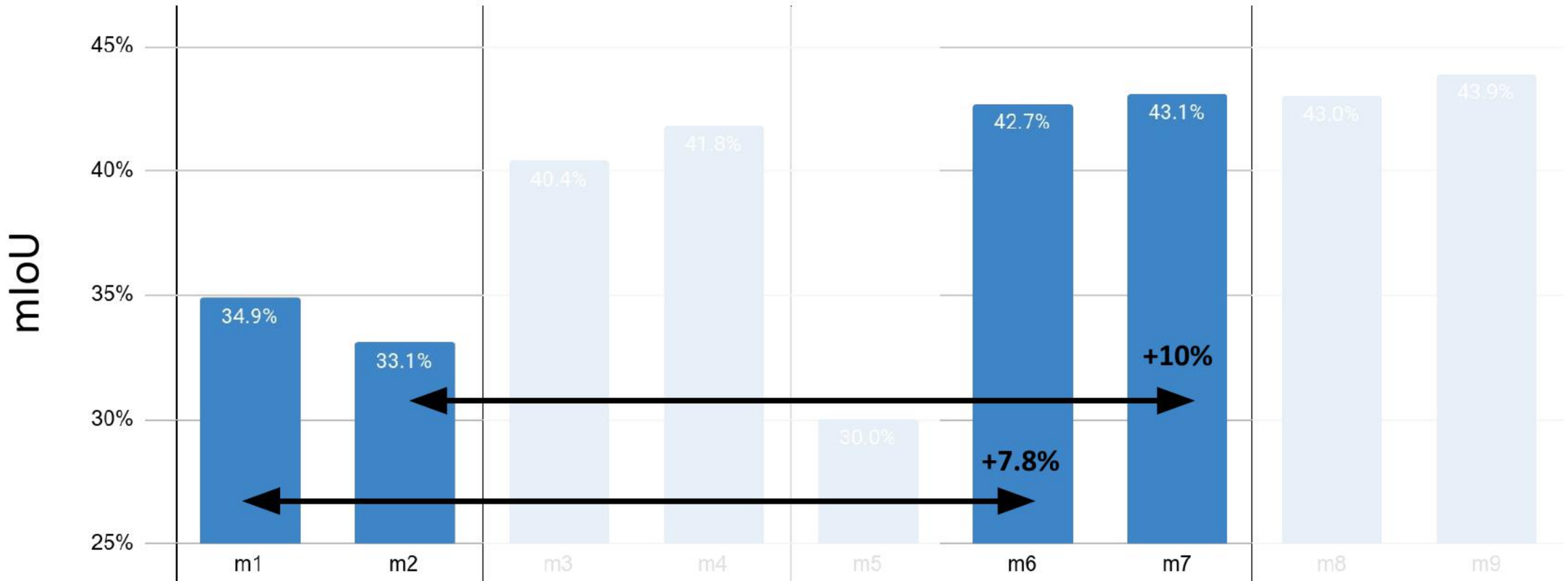
Reconstruction

Impact of our contributions on multi-object scenes



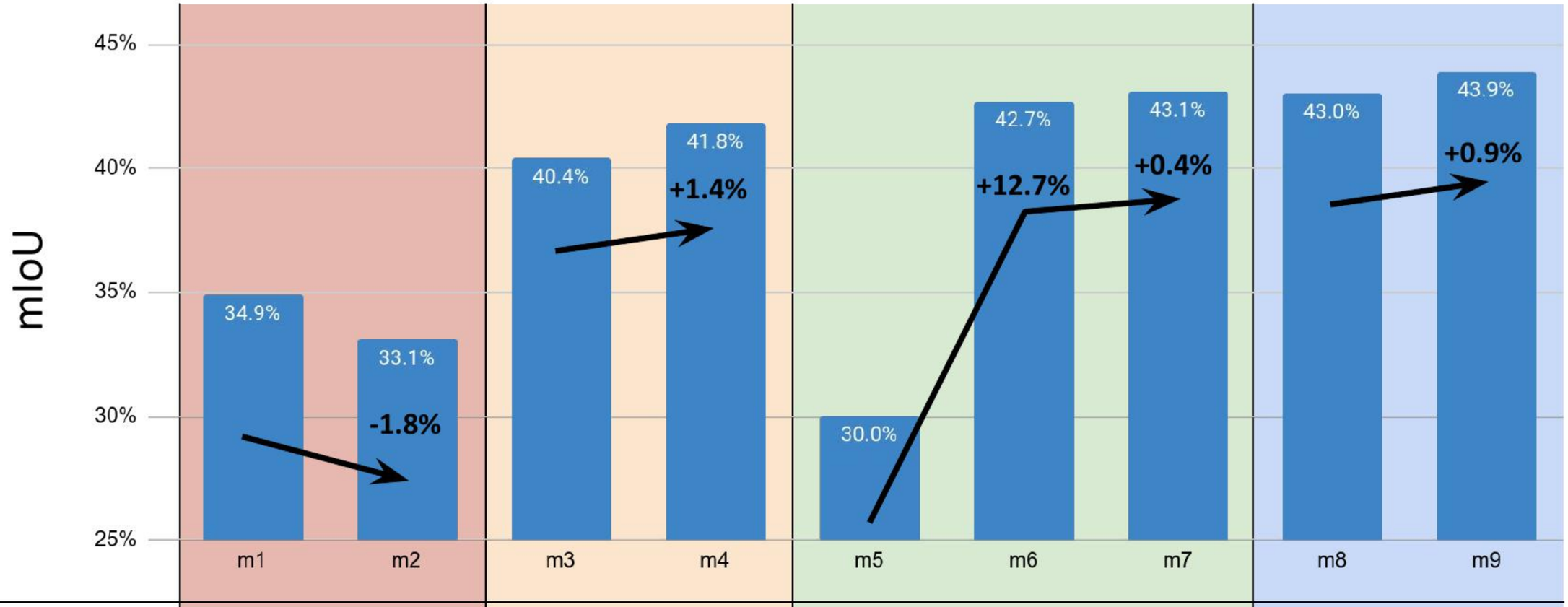
loss	focal	IoU	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes		yes			yes	
realism	high		low		high			high	
dataset	pairs		pairs		pairs			triplets	

Ray-traced skip connections improve mIoU substantially



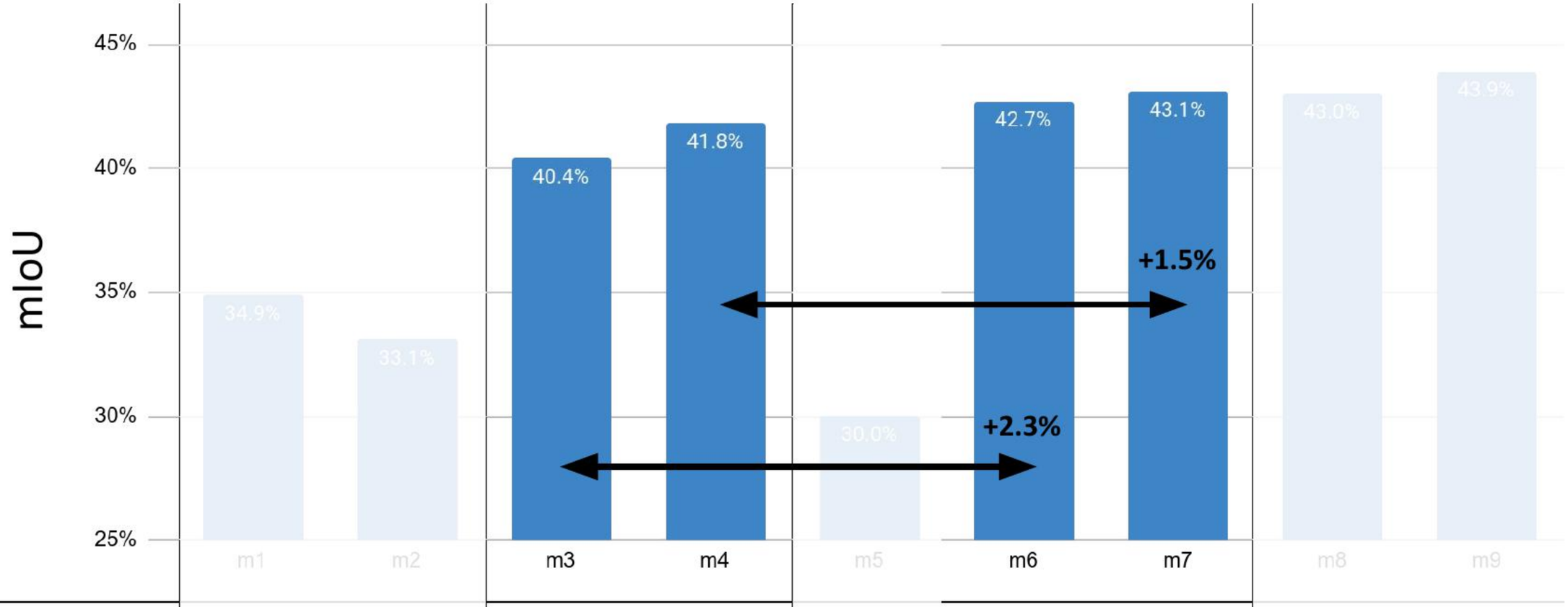
loss	focal	IoU	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes		yes			yes	
realism	high		low		high			high	
dataset	pairs		pairs		pairs			triplets	

IoU loss performs best with ray-traced skip connections



loss	focal	IoU	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes		yes			yes	
realism	high		low		high			high	
dataset	pairs		pairs		pairs			triplets	

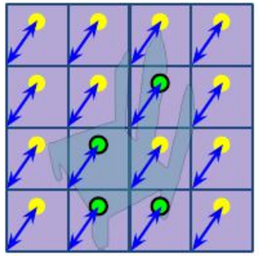
Better performance on high compared to low realism

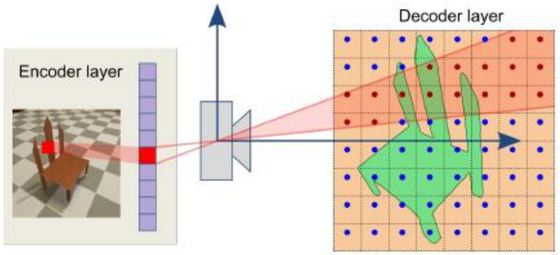


	m1	m2	m3	m4	m5	m6	m7	m8	m9
loss	focal	IoU	focal	IoU	xent	focal	IoU	focal	IoU
ray-traced skip conn.	no		yes		yes			yes	
realism	high		low		high			high	
dataset	pairs		pairs		pairs			triplets	

Conclusion

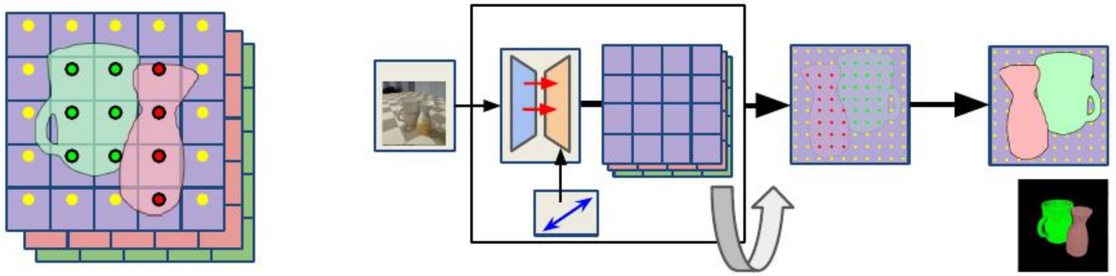
- Three improvements of common single object reconstruction architectures





$$IoU_g(g,p) = \frac{\sum_{i \in G} \sum_{c=1}^{C-1} \min(g_{ic}, p_{ic}) \cdot \mu(g_{ic})}{\sum_{i \in G} \sum_{c=1}^{C-1} \max(g_{ic}, p_{ic}) \cdot \mu(g_{ic})}$$

- Coherent reconstruction of scenes with multiple objects



- Evaluation on large-scale datasets with synthetic multi-object scenes and Pix3D