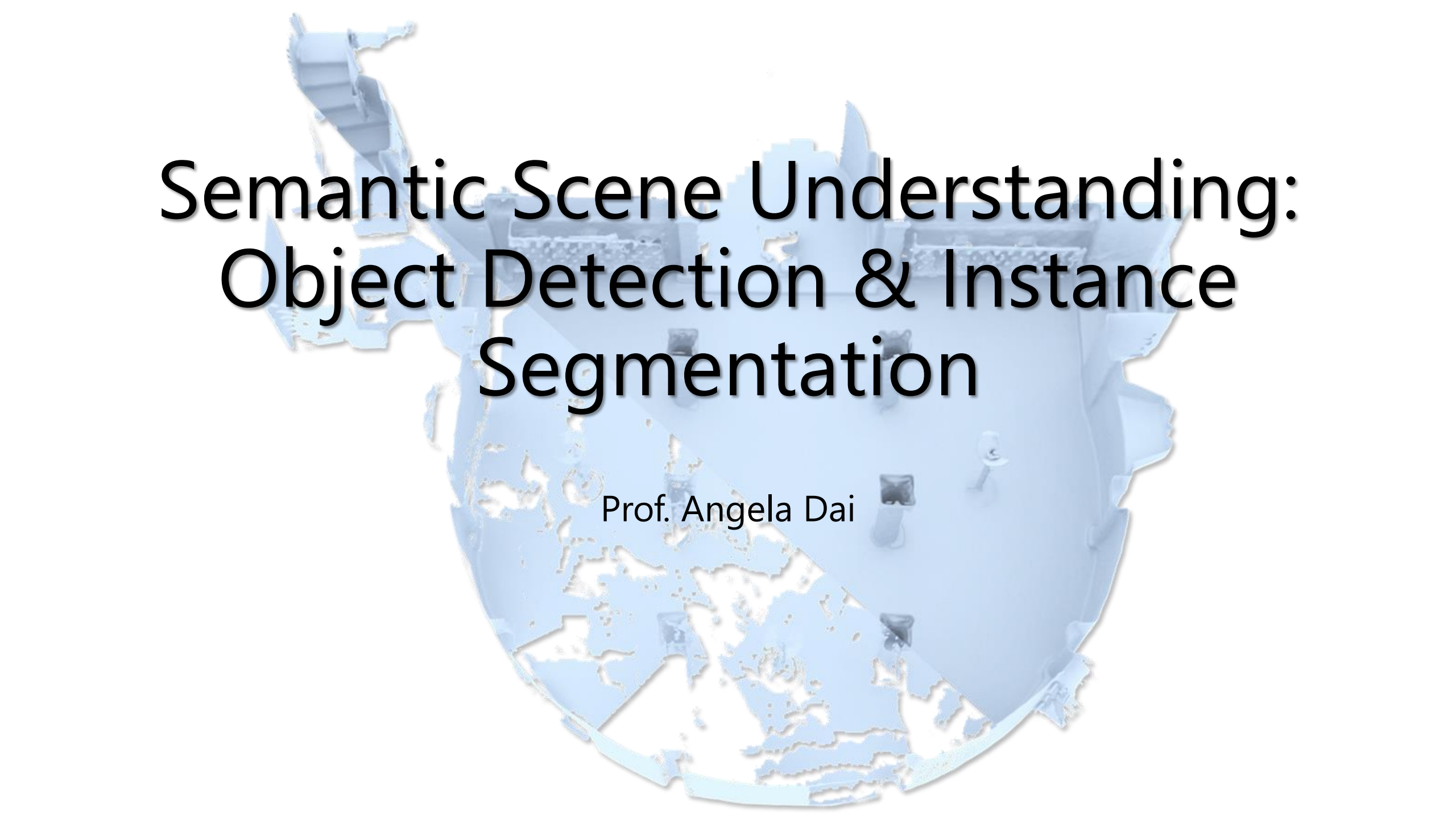
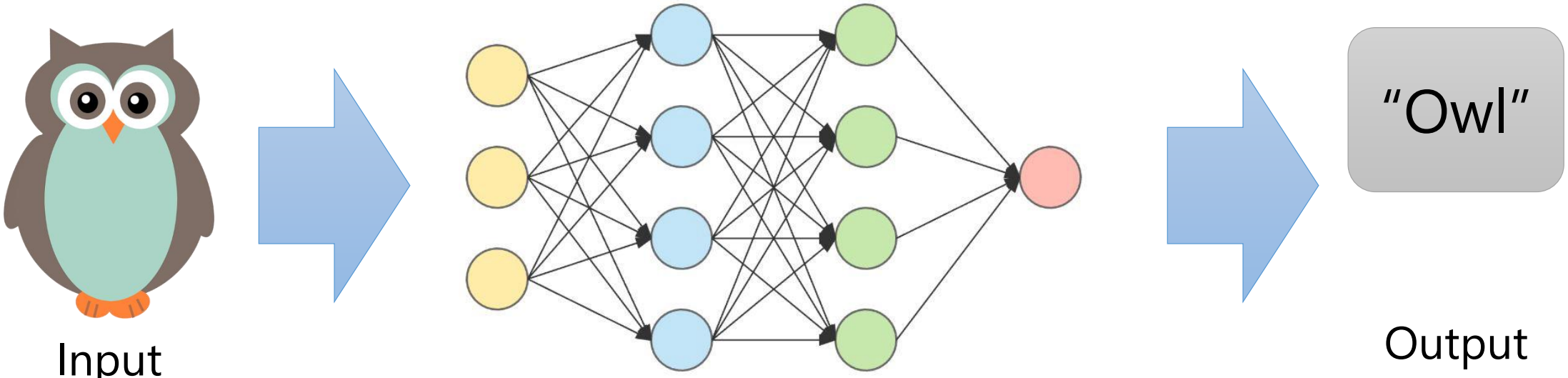


Semantic Scene Understanding: Object Detection & Instance Segmentation

Prof. Angela Dai

Brief Recap

Deep Learning



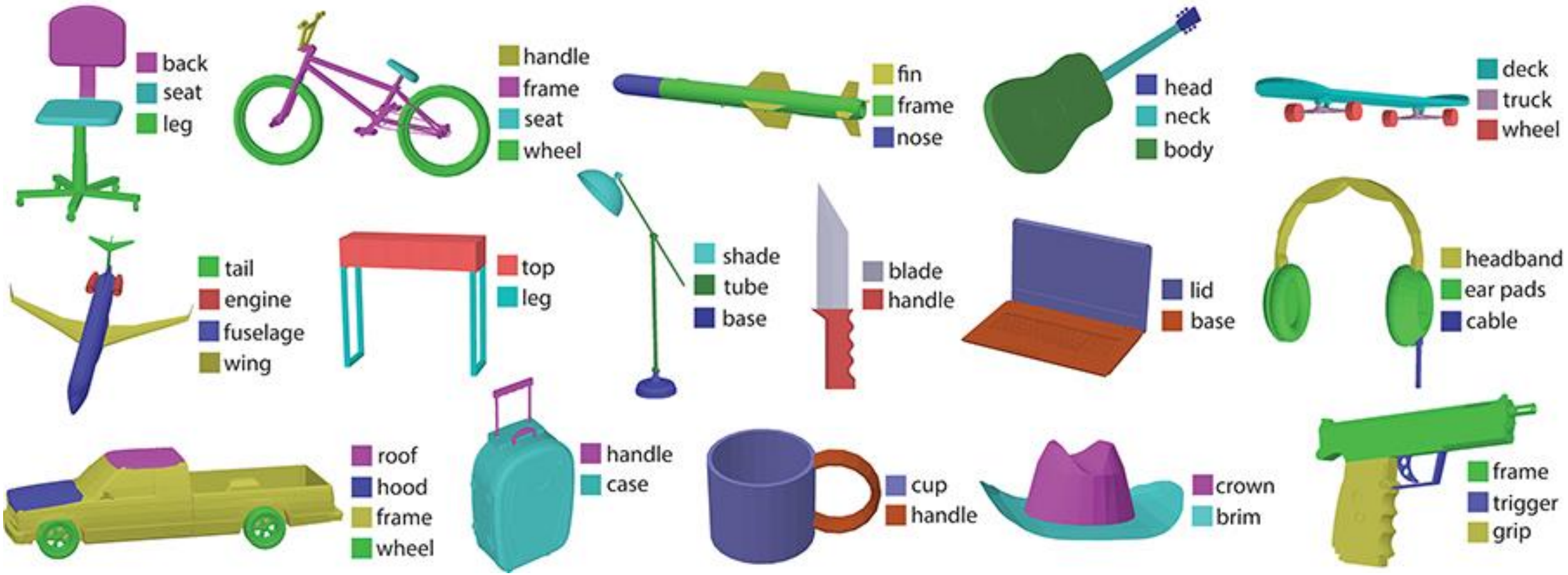
Input

Learned Feature Extraction
and Classification

Output

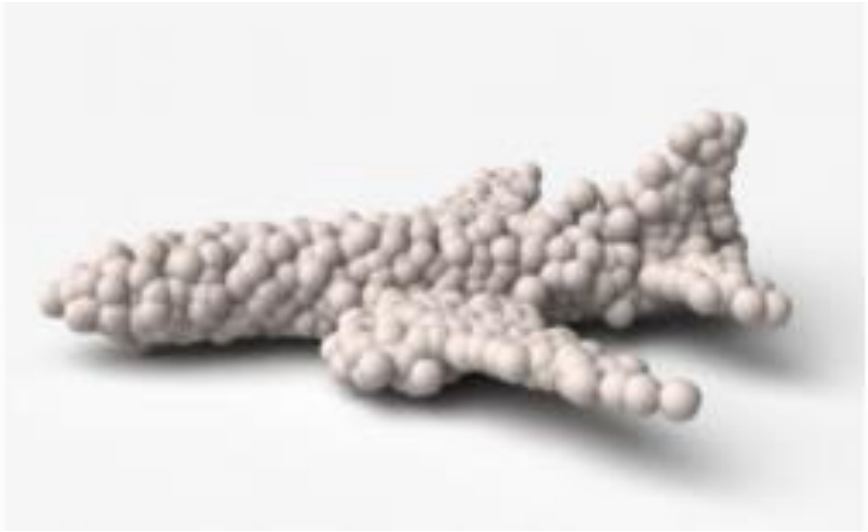
Want to automatically learn good feature representations for the task

Shape segmentation into parts

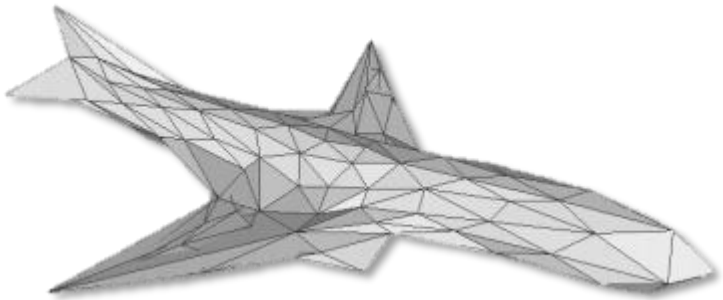


Generating Shapes

Signed Distance Fields

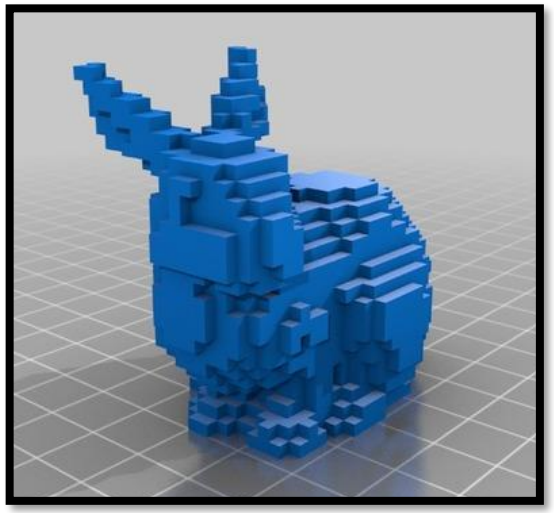


Point Clouds

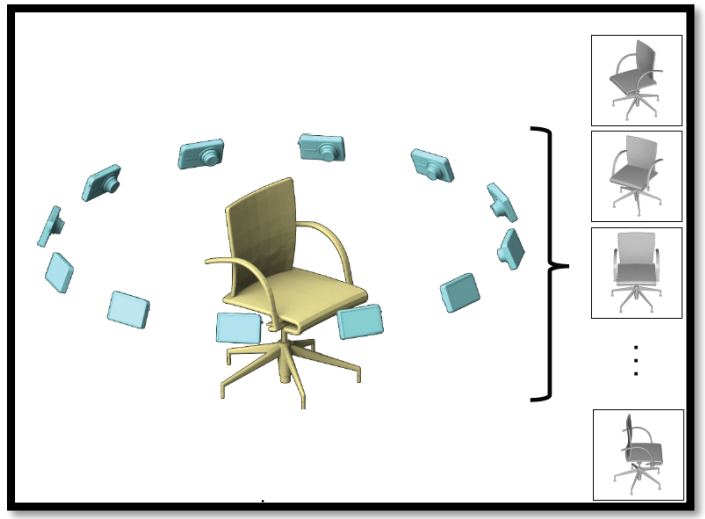


Meshes

3D Deep Learning by Representations

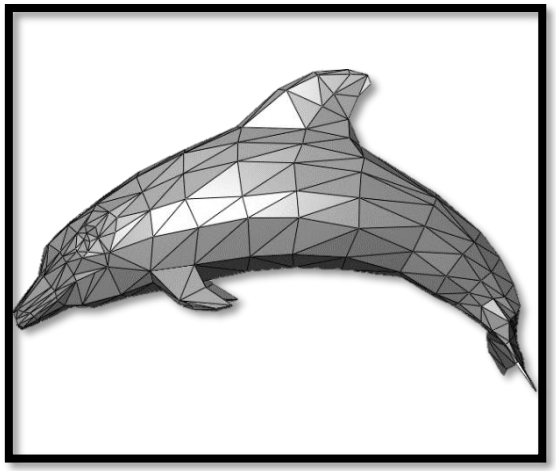


Volumetric
3D CNNs: Dense, Hierarchical, Sparse



Multi-View
(also: multi-view + volumetric/point/mesh)

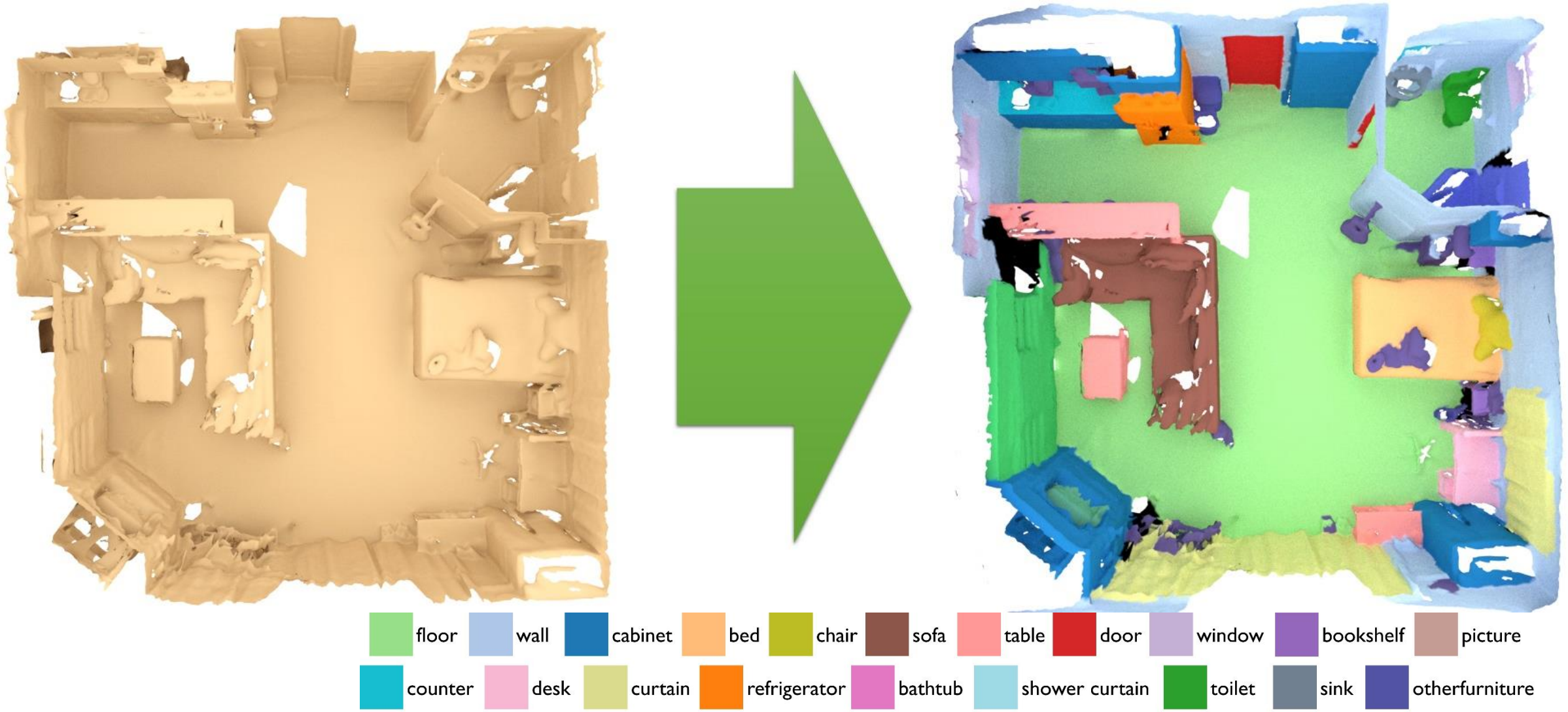
Point Cloud



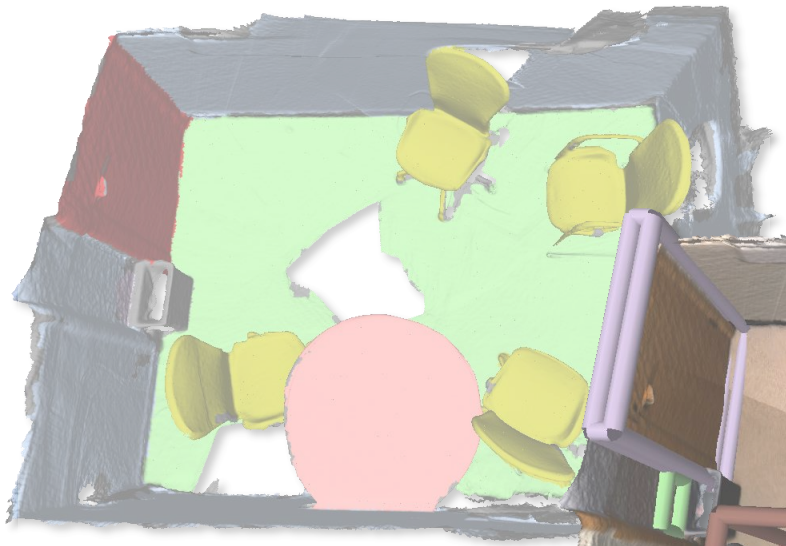
Mesh
Graph Neural Networks

and more!

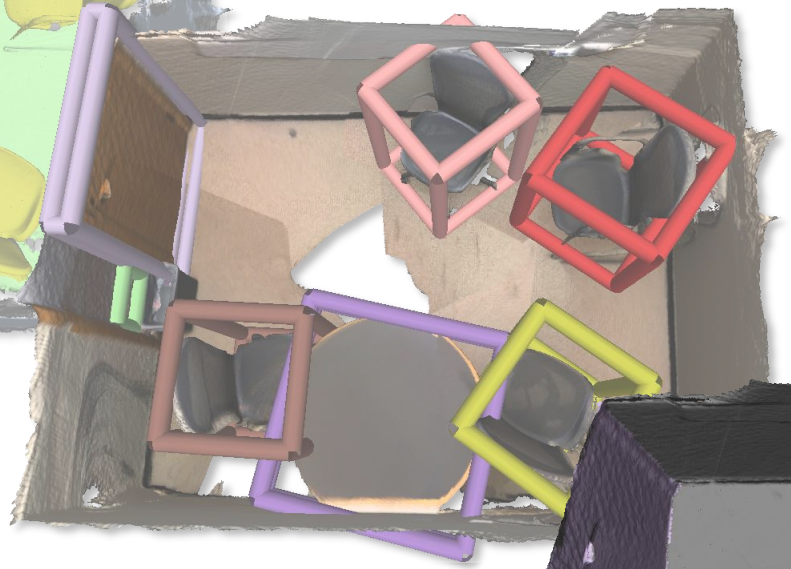
3D Semantic Segmentation



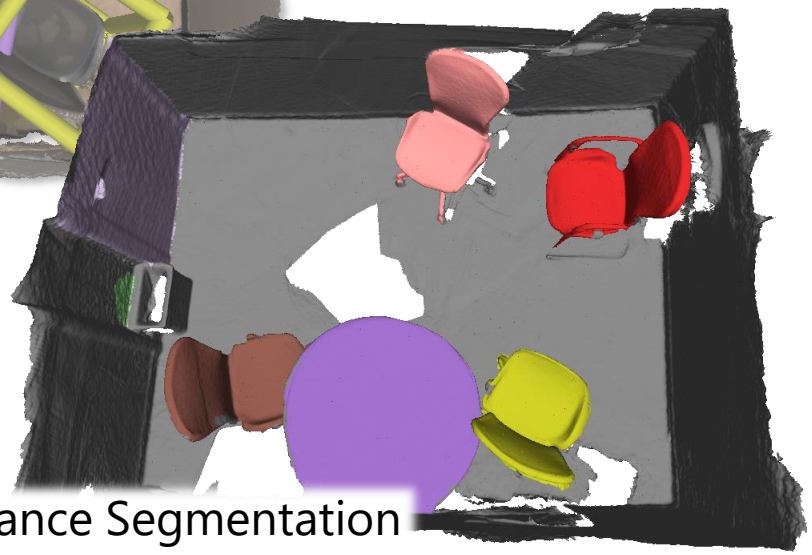
Understanding object-ness



Semantic Segmentation

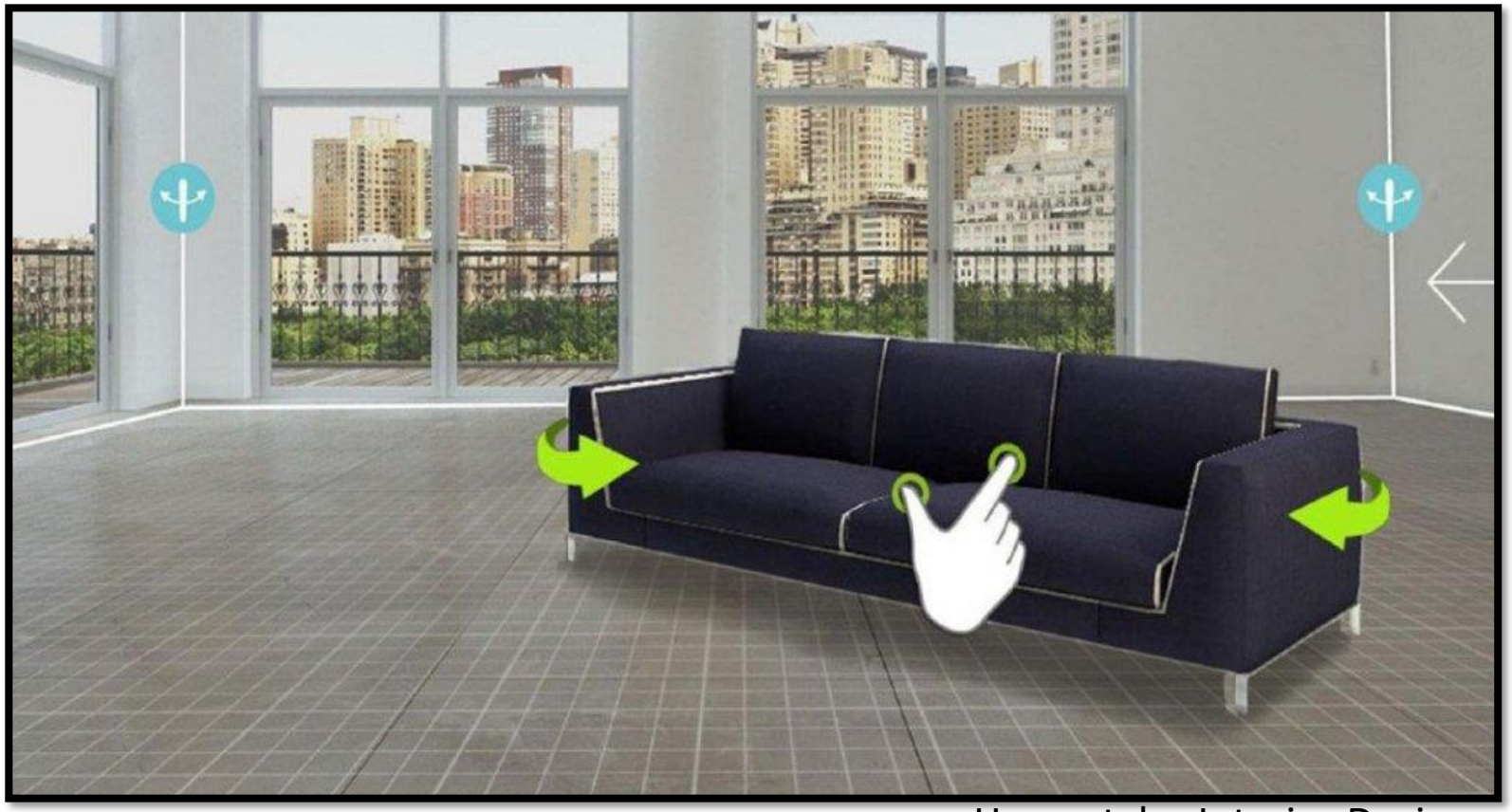


Object Detection



Instance Segmentation

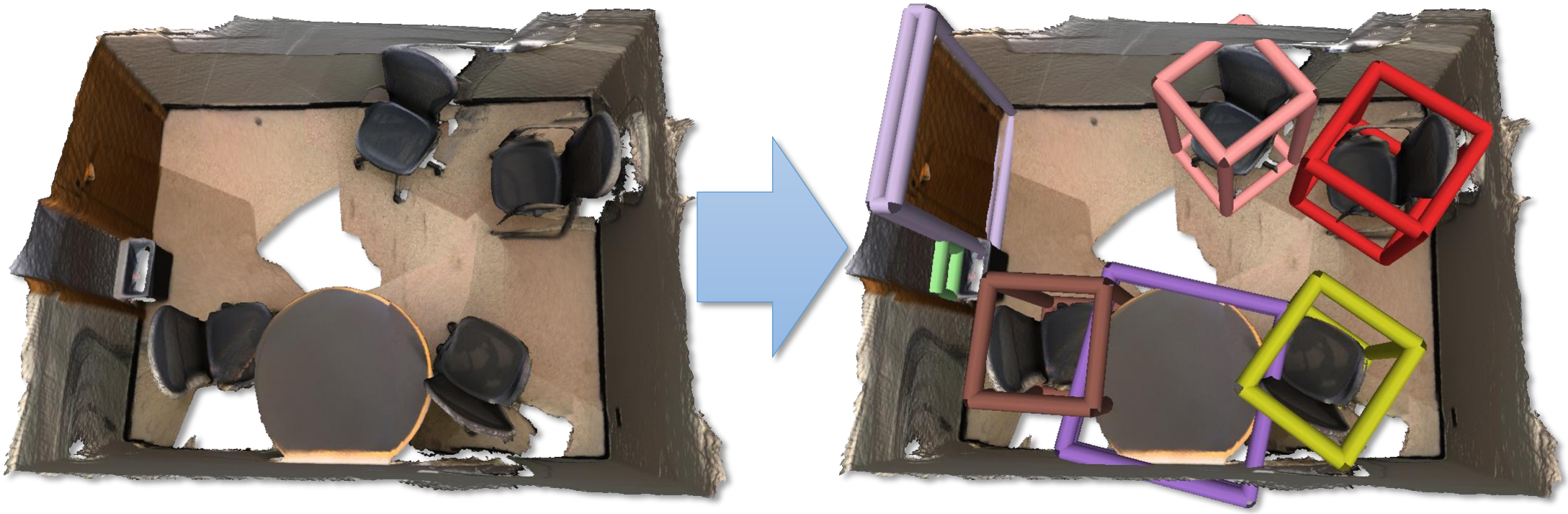
Understanding object-ness



Homestyler Interior Design

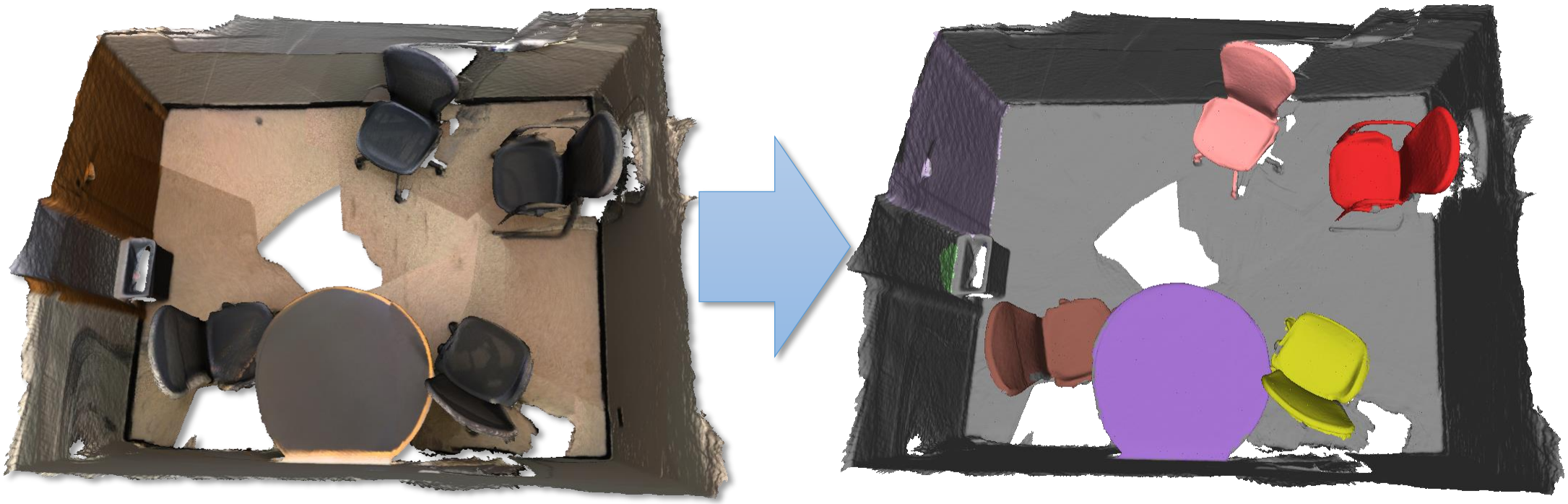
[Kent et al. ’17]

3D Object Detection



Localize individual objects by estimating their bounding boxes

3D Instance Segmentation



Recognize individual objects by estimating the geometry belonging to each object

Successes in 2D Perception

[He et al. ’17] Mask R-CNN

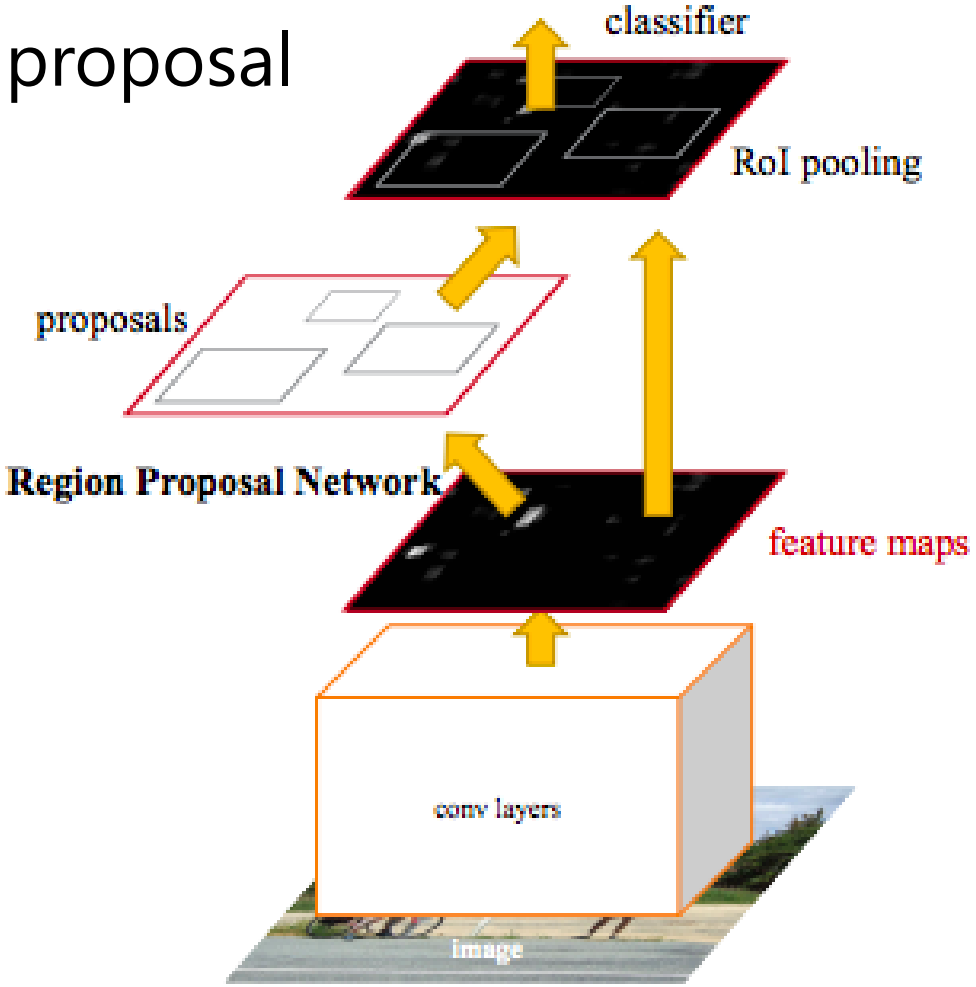
In 2D: Mask R-CNN

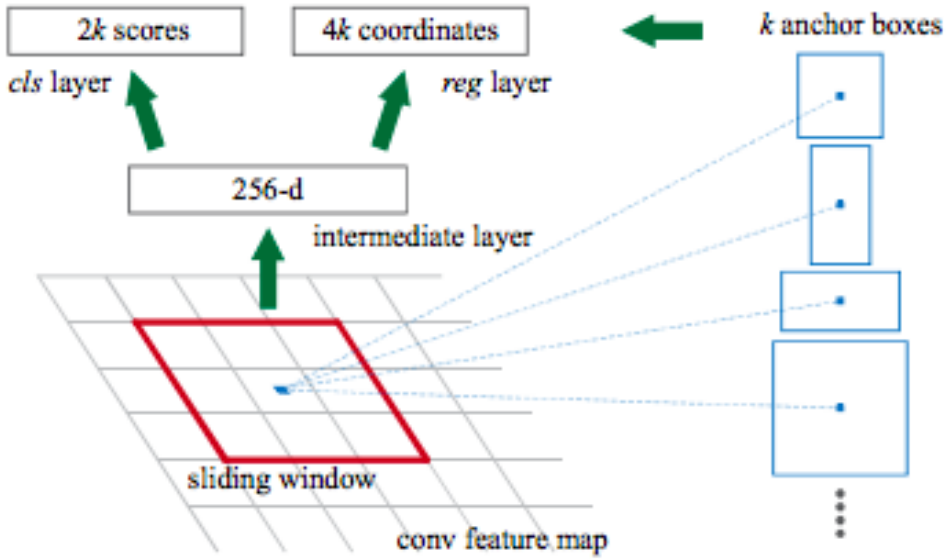
- Region proposal

[He et al. ’17]

In 2D: Mask R-CNN

- Region proposal





[Ren et al. '15]

In 2D: Mask R-CNN

- Bounding box refinement

Proposed bounding box $P = (p_x, p_y, p_w, p_h)$

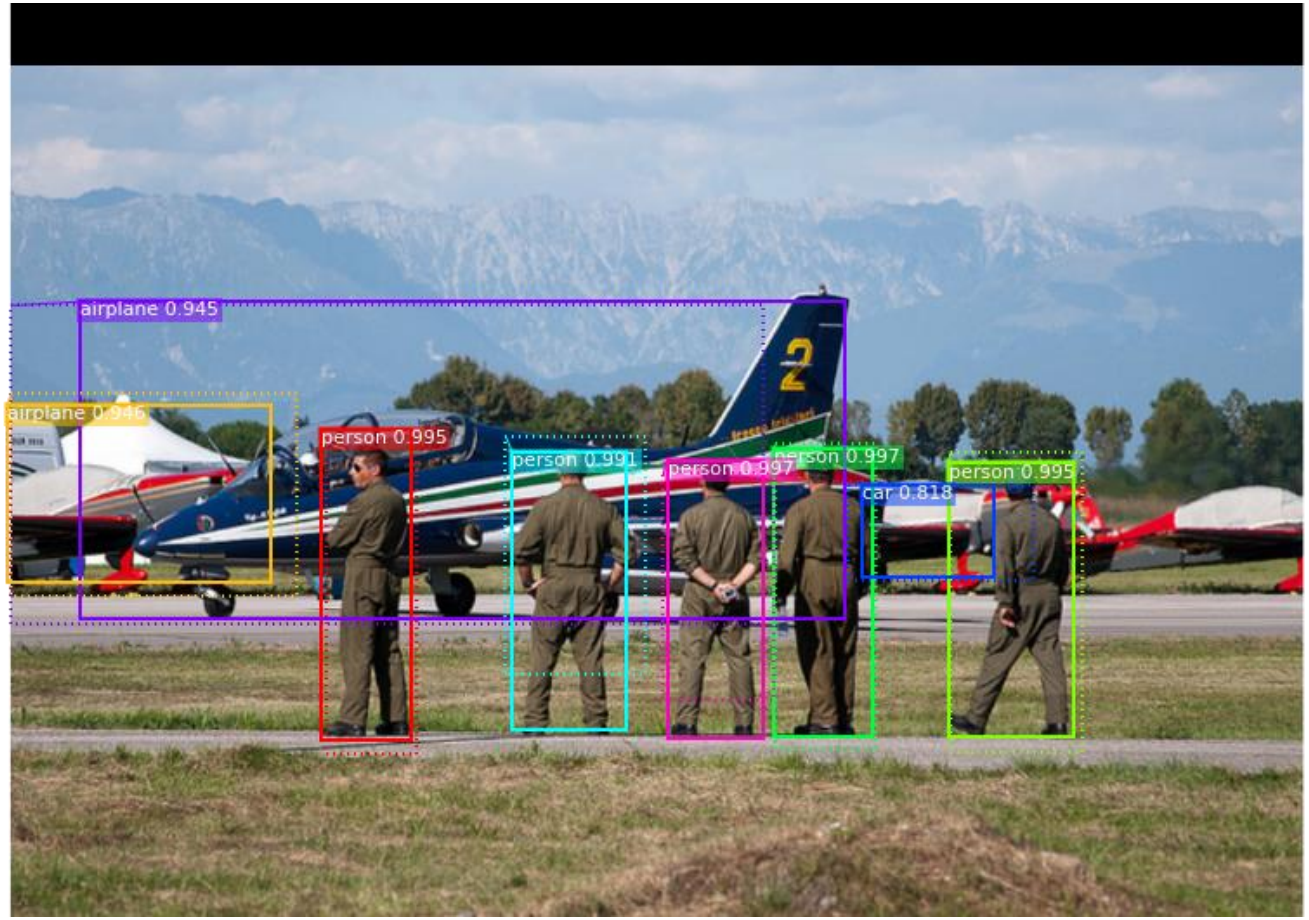
Ground truth bounding box $G = (g_x, g_y, g_w, g_h)$

Predicted refinement as offsets:

$$d(\mathrm{P}) = (d_x, d_y, d_w, d_h)$$

Transform proposal P:

$$\hat{g}_x = p_w d_x + p_x$$
$$\hat{g}_y = p_h d_y + p_y$$
$$\hat{g}_w = p_w \exp(d_w)$$
$$\hat{g}_h = p_h \exp(d_h)$$



[He et al. '17]

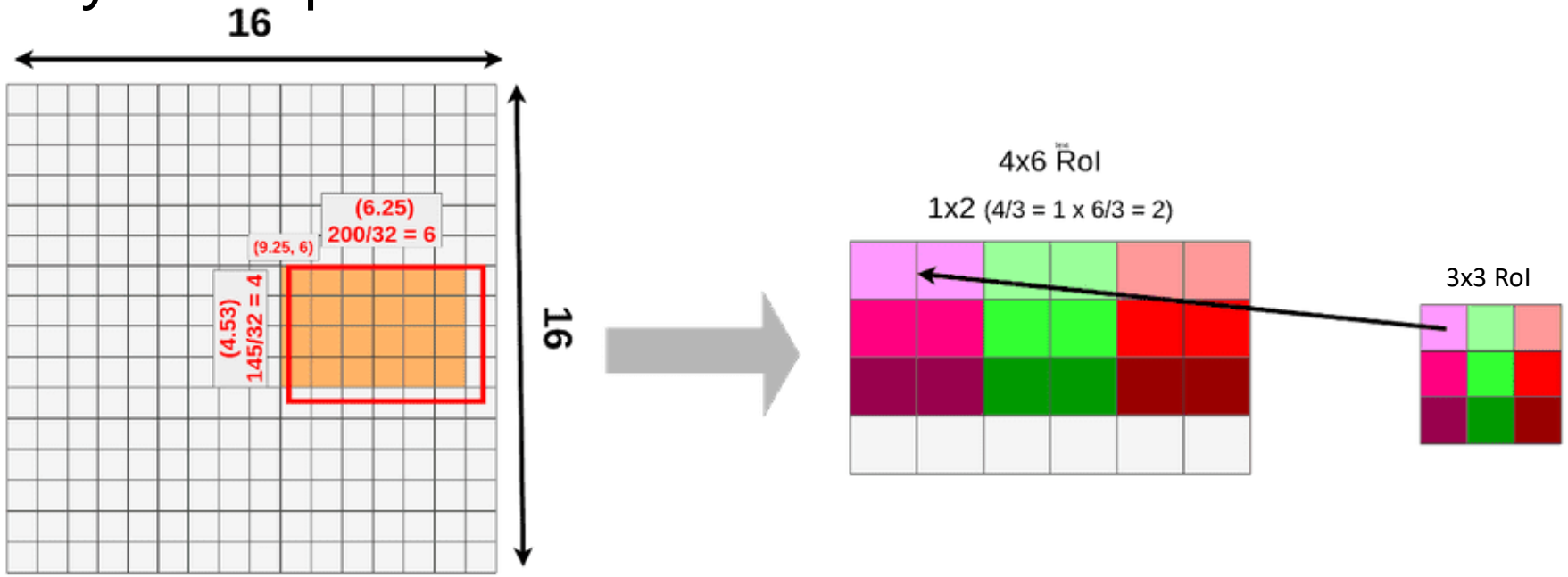
In 2D: Mask R-CNN

- Mask generation

[He et al. ’17]

In 2D: Mask R-CNN

- Mask generation
- Extract feature maps from each region of interest
- Binary mask prediction

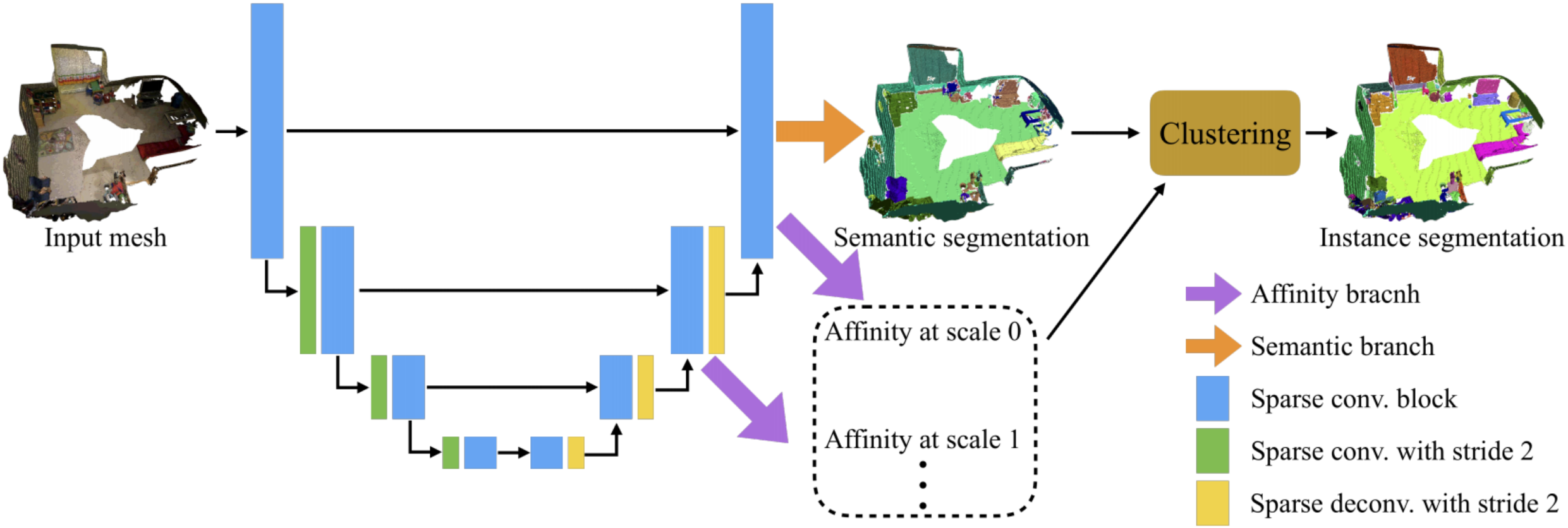


[He et al. '17]

Object Detection and Instance Segmentation in 3D?

- 3rd dimension -> more anchors?
- Can we leverage the scale information from 3D?
- In 3D: more spatial separation between objects
- In 3D: more difficult to capture high resolution

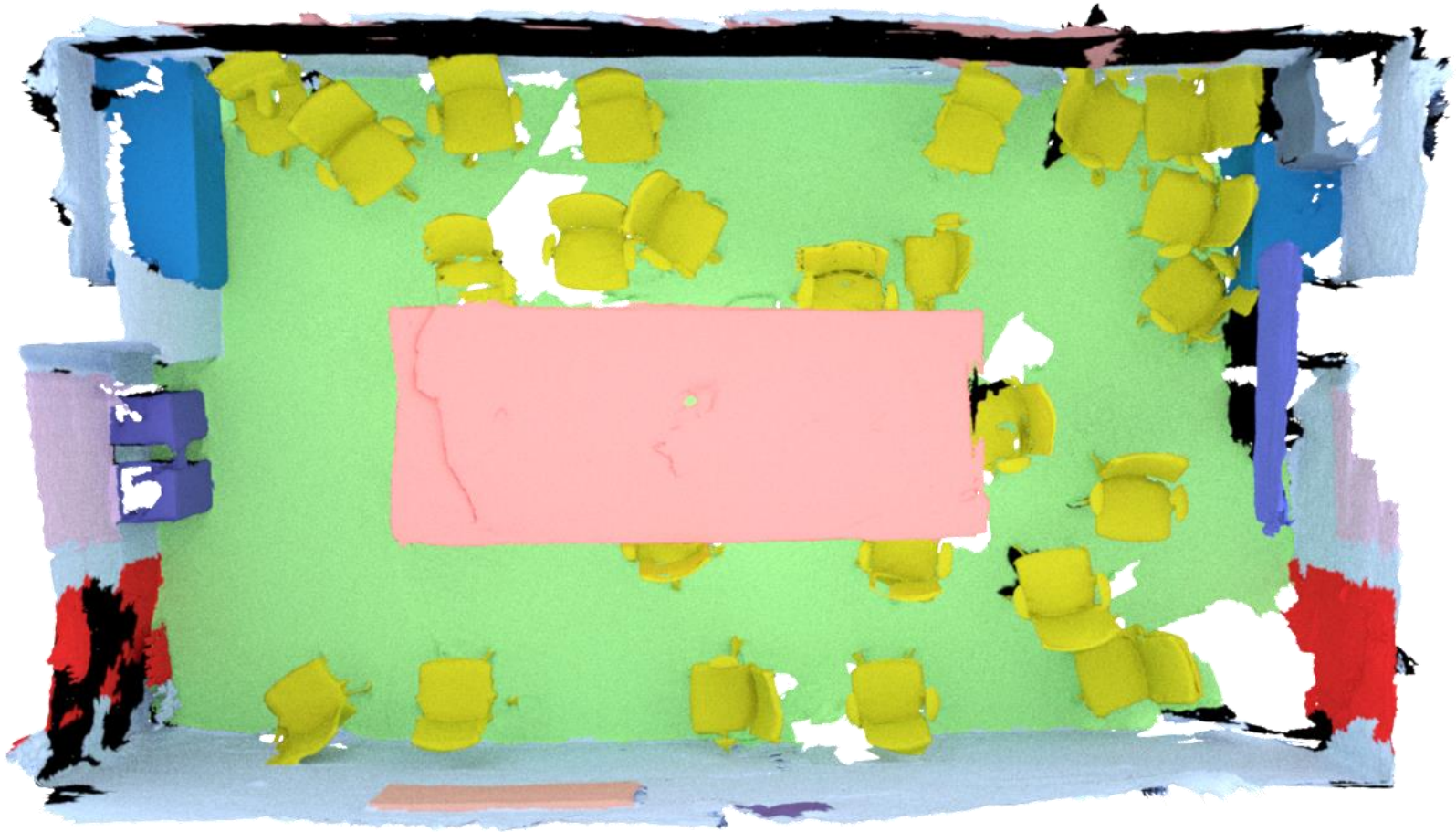
In 3D: Group Semantic Segmentations?



[Liu et al. '19]

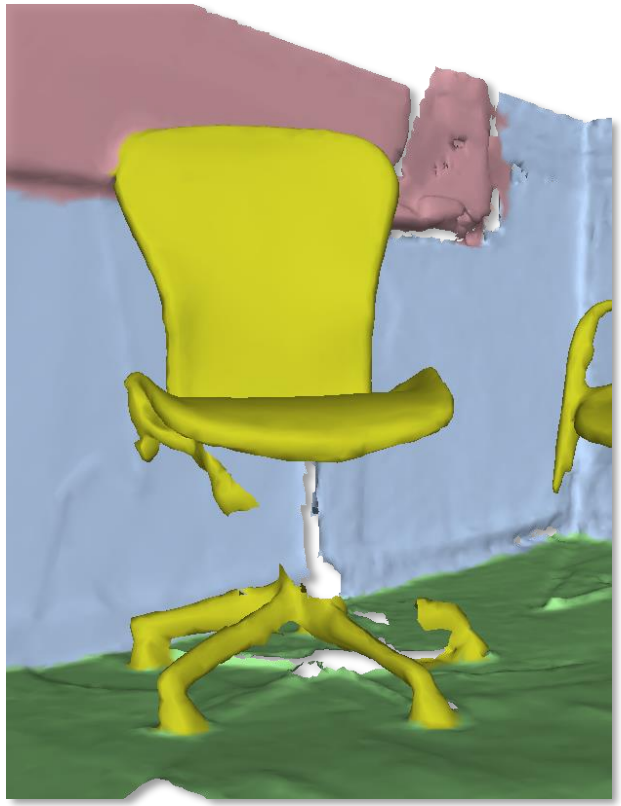
In 3D: Group Semantic Segmentations?

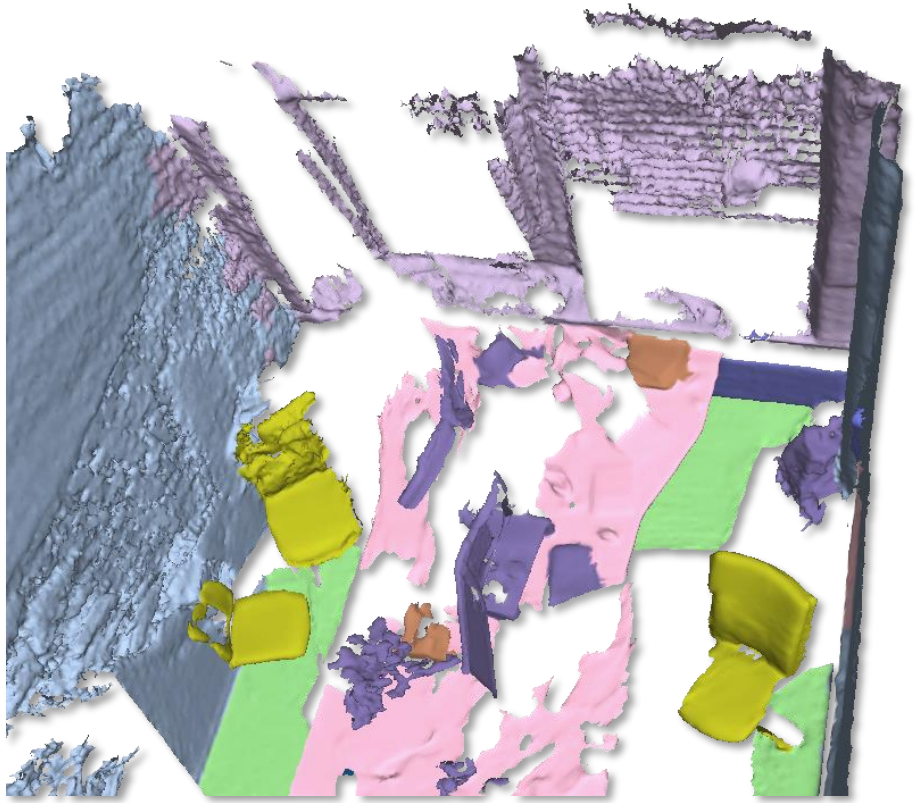
- Challenge: distinguishing objects that lie spatially near each other



In 3D: Group Semantic Segmentations?

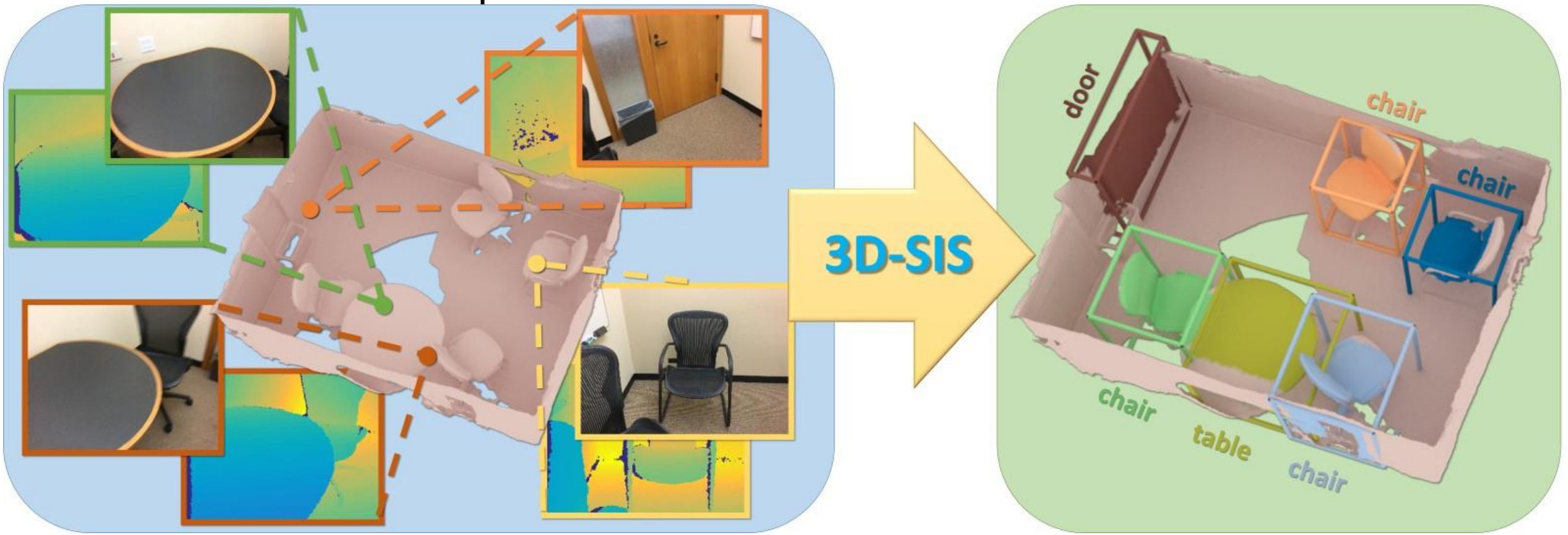
- Challenge: merging objects with thin pieces





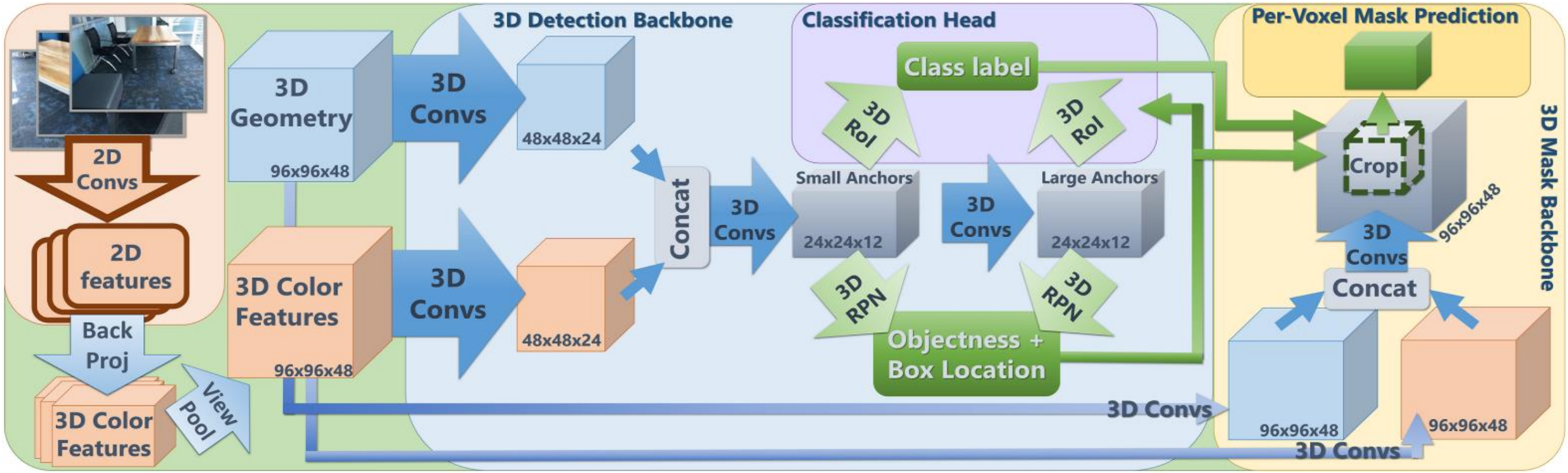
3D-SIS: 3D Semantic Instance Segmentation

- Mask R-CNN inspiration



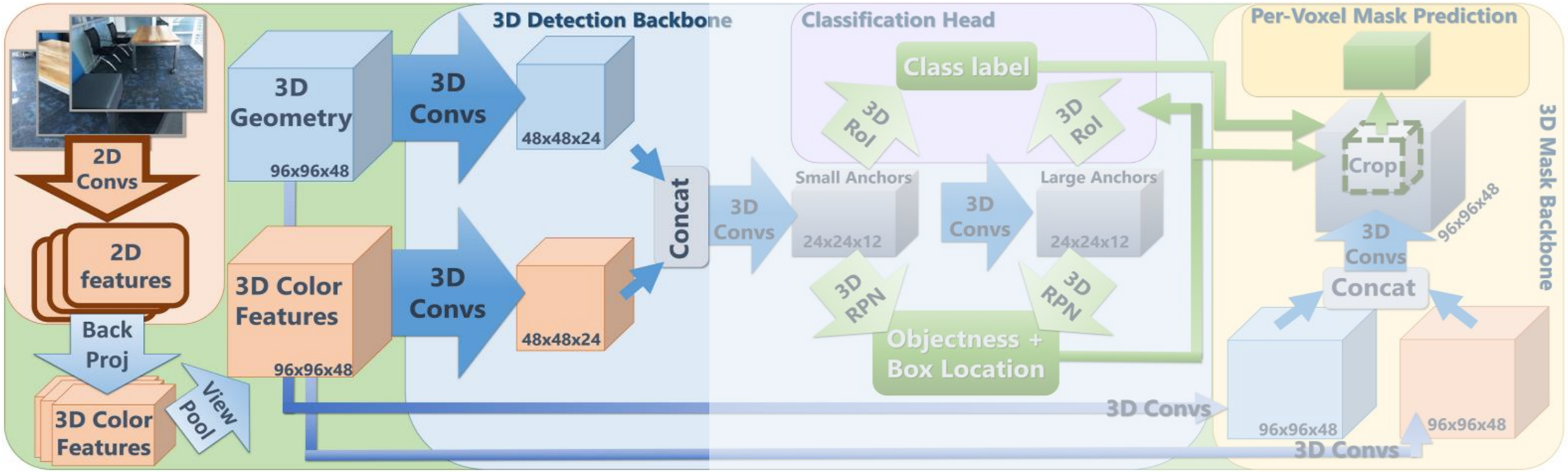
[Hou et al. '19]

3D-SIS: 3D Semantic Instance Segmentation



[Hou et al. '19]

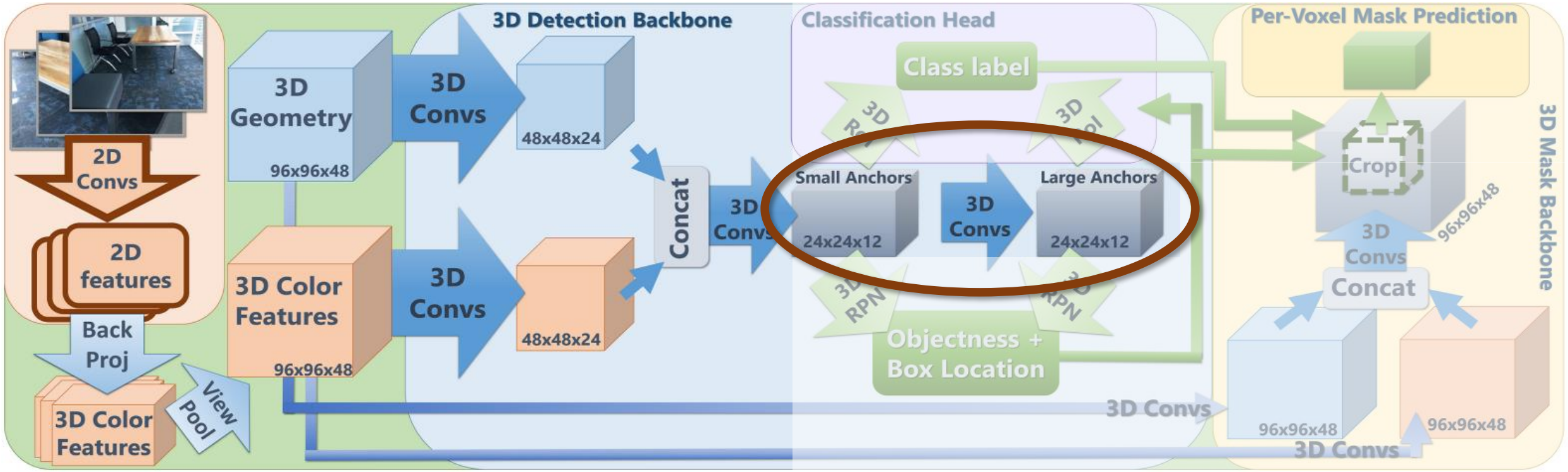
3D-SIS: 3D Semantic Instance Segmentation



Joint color-geometry feature extraction

[Hou et al. '19]

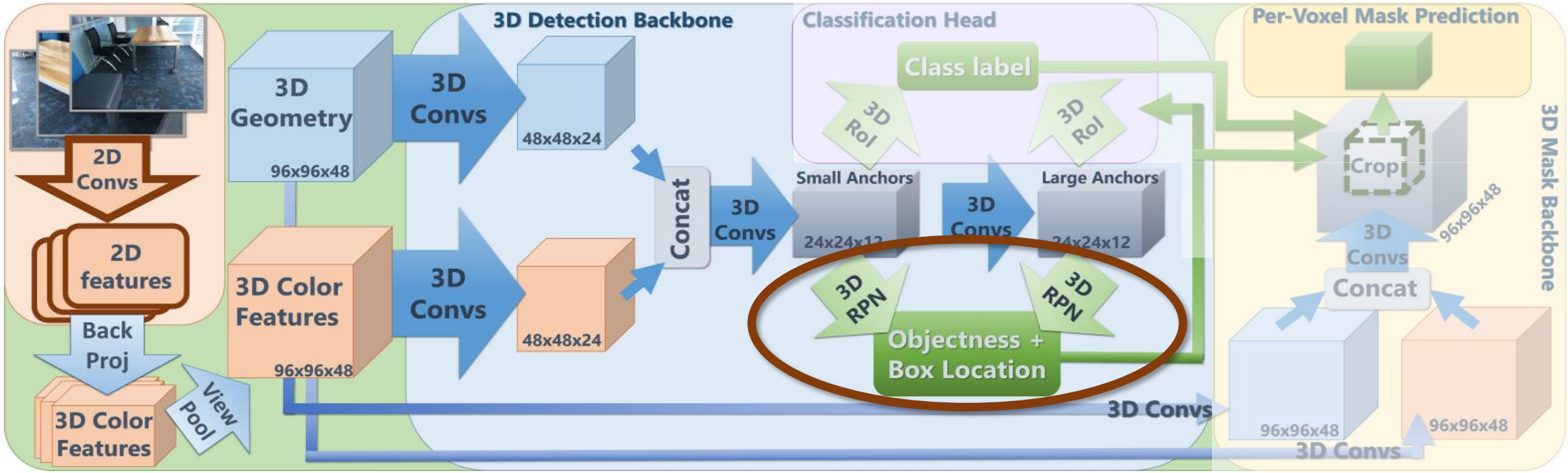
3D-SIS: 3D Semantic Instance Segmentation



3D anchors: from clustering train object boxes.
Large-sized anchors have larger receptive field size.

[Hou et al. '19]

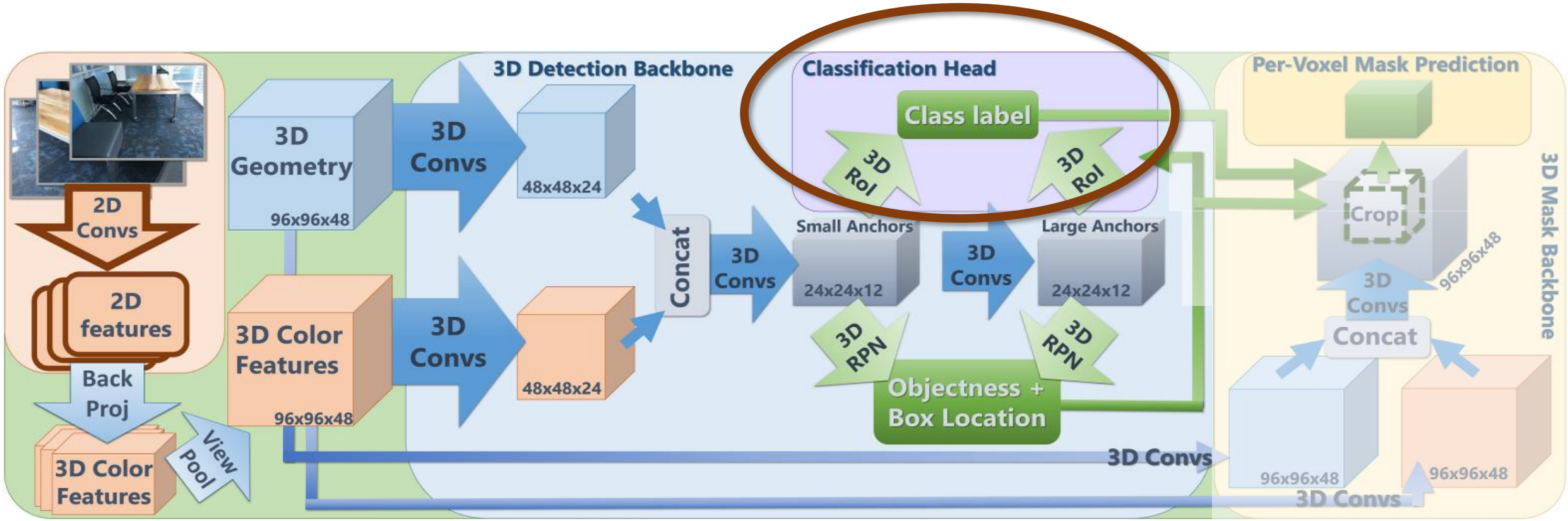
3D-SIS: 3D Semantic Instance Segmentation



Predict if anchor box corresponds to an object
+ 3D box refinement if object

[Hou et al. '19]

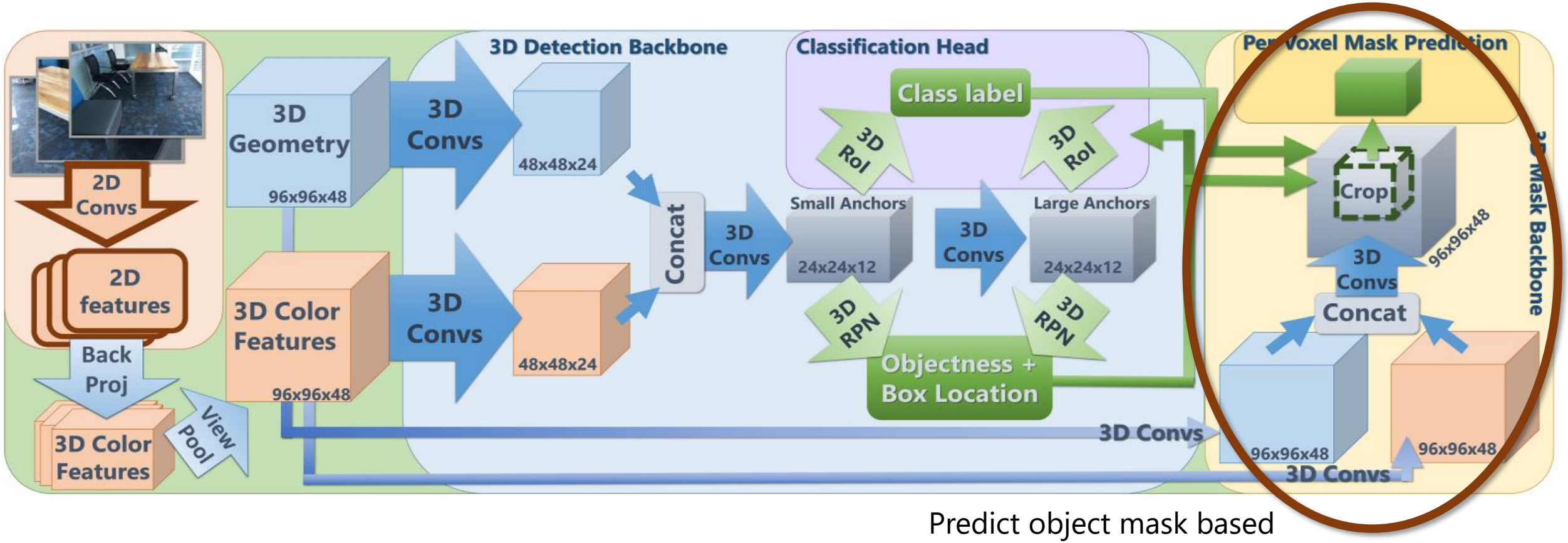
3D-SIS: 3D Semantic Instance Segmentation



3D region of interest pooling.
Predict object class category.

[Hou et al. '19]

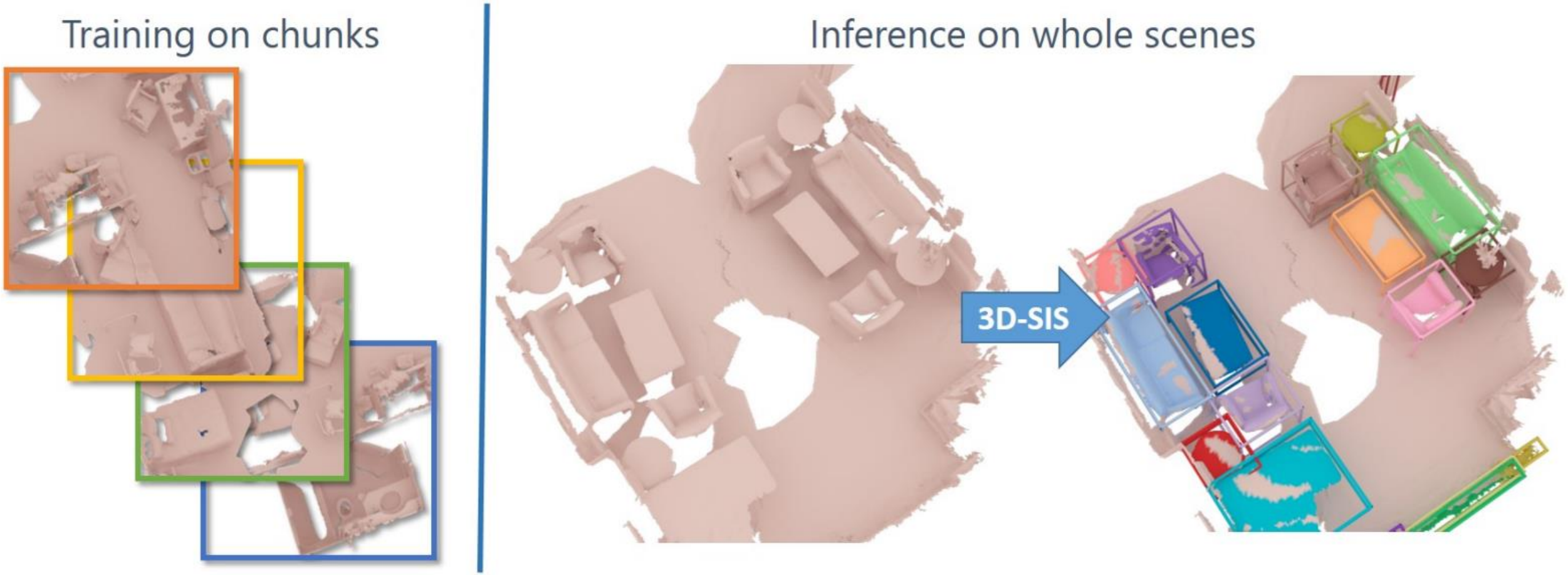
3D-SIS: 3D Semantic Instance Segmentation



Predict object mask based on full spatial feature grid

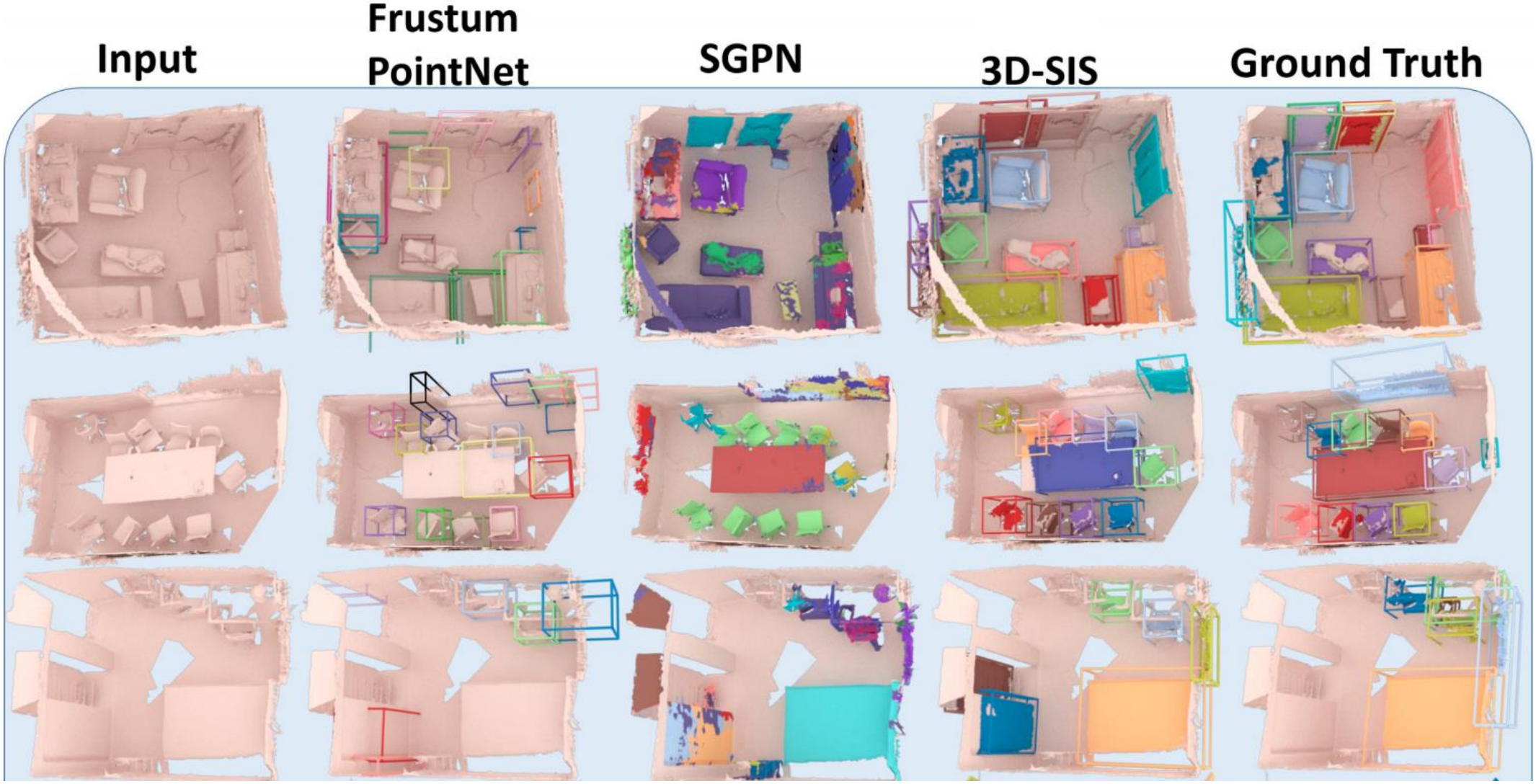
[Hou et al. '19]

3D-SIS: 3D Semantic Instance Segmentation



[Hou et al. '19]

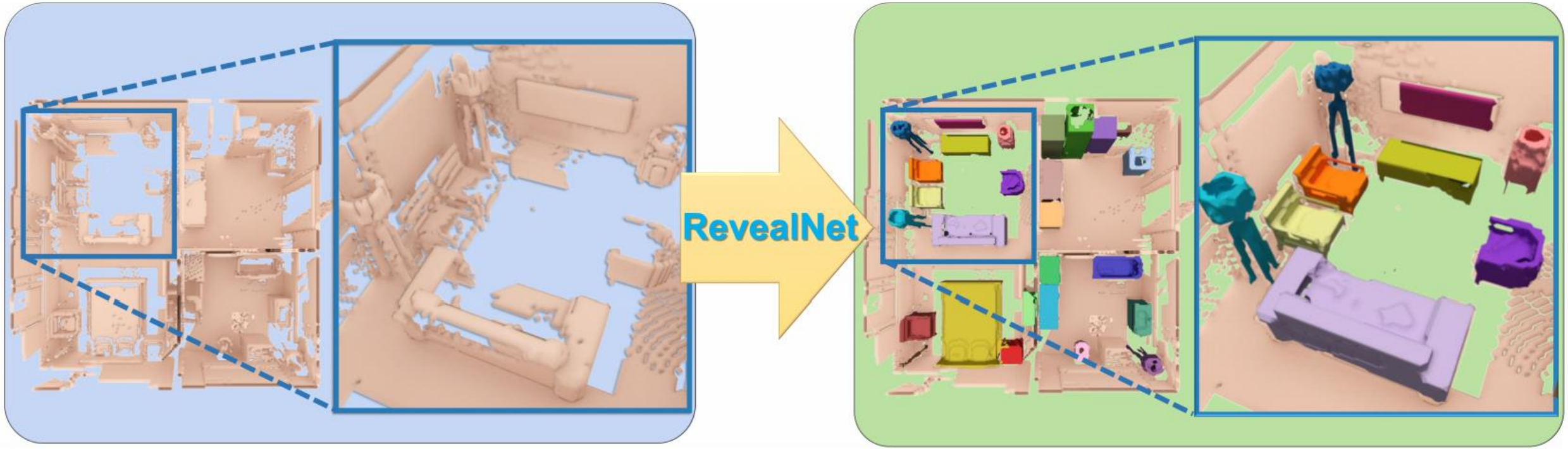
3D-SIS: 3D Semantic Instance Segmentation



[Hou et al. '19]

RevealNet

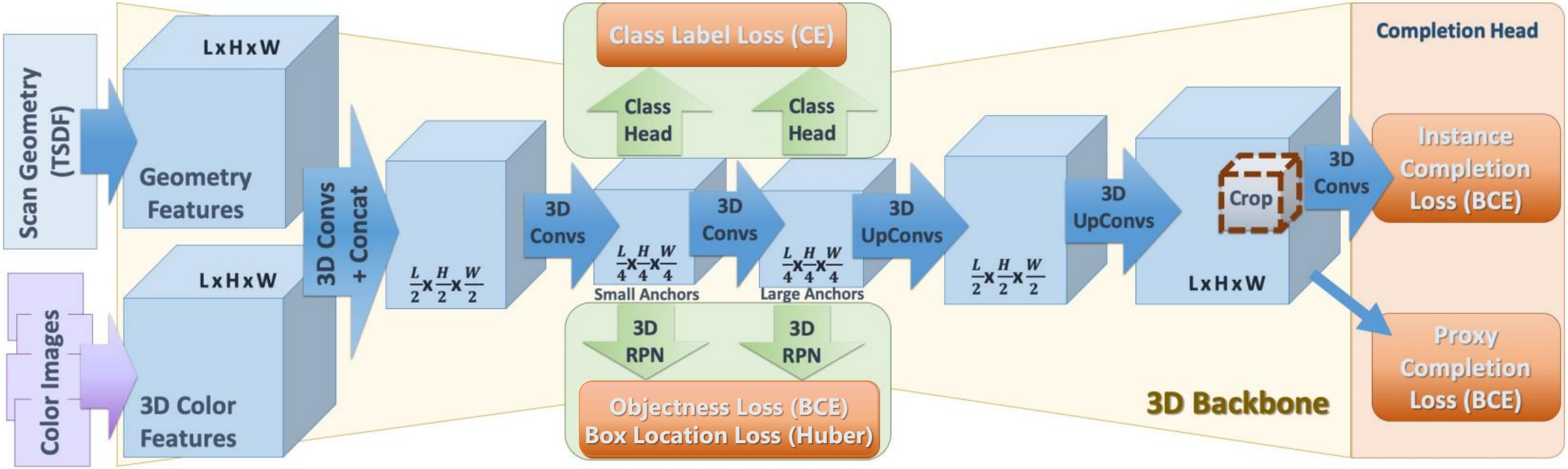
- Can we get better priors by hallucinating missing geometry?



[Hou et al. ’20]

RevealNet

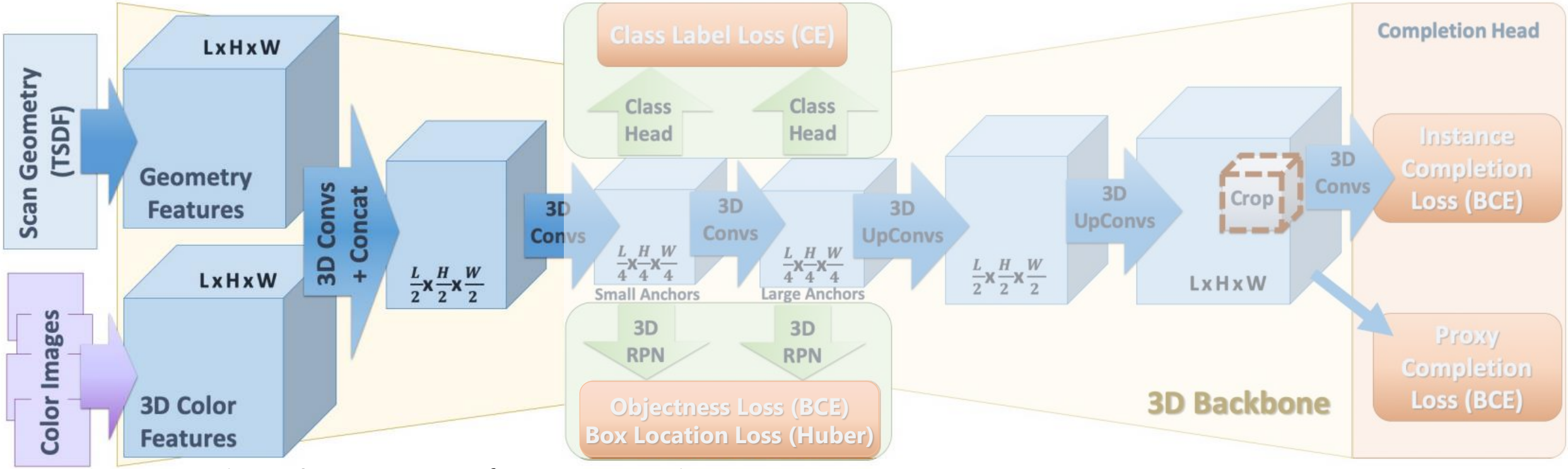
- Can we get better priors by hallucinating missing geometry?



[Hou et al. '20]

RevealNet

- Can we get better priors by hallucinating missing geometry?

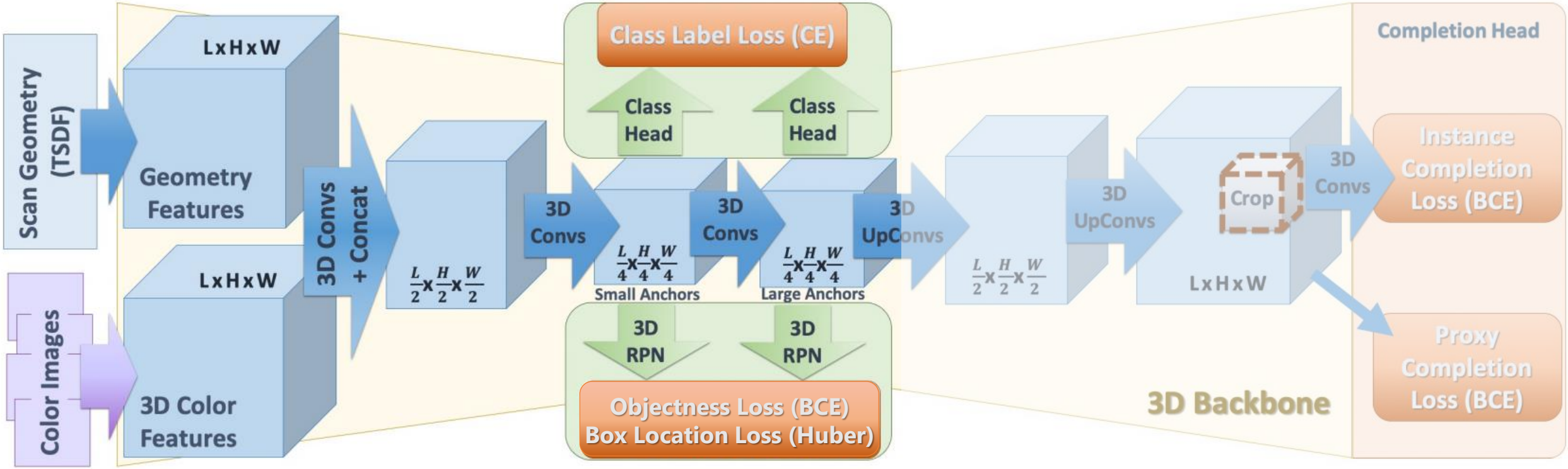


Joint color-geometry feature extraction

[Hou et al. '20]

RevealNet

- Can we get better priors by hallucinating missing geometry?

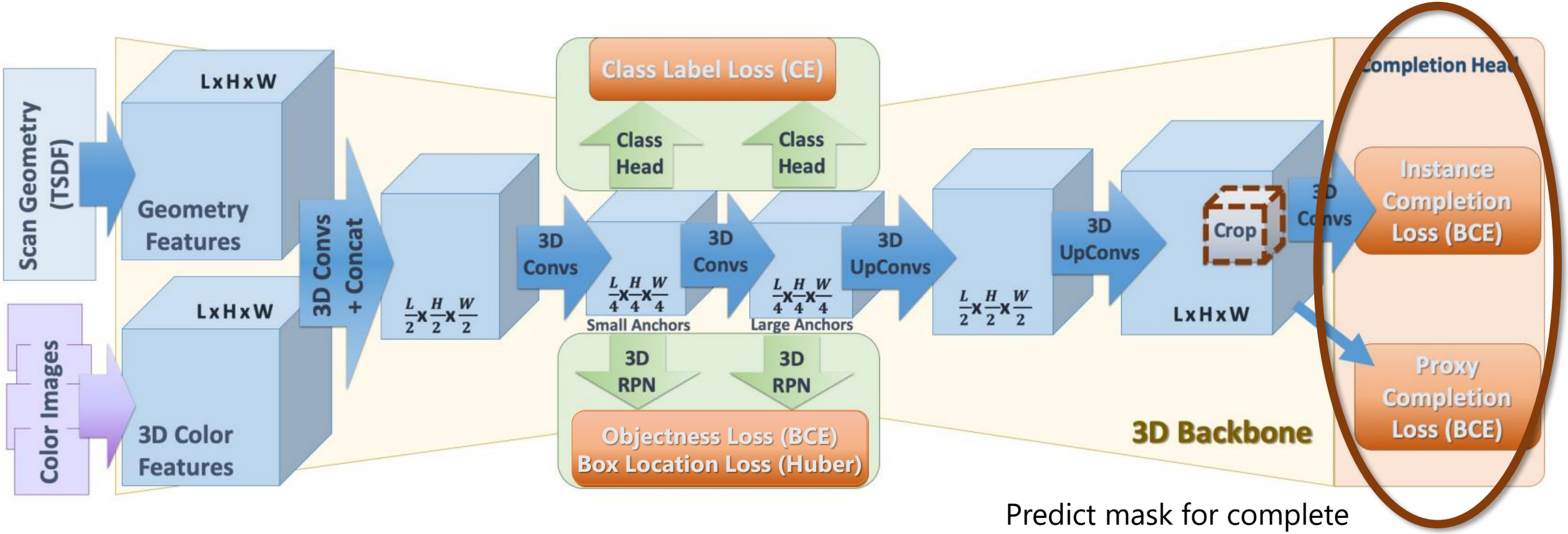


Object detection: bounding box regression and classification

[Hou et al. '20]

RevealNet

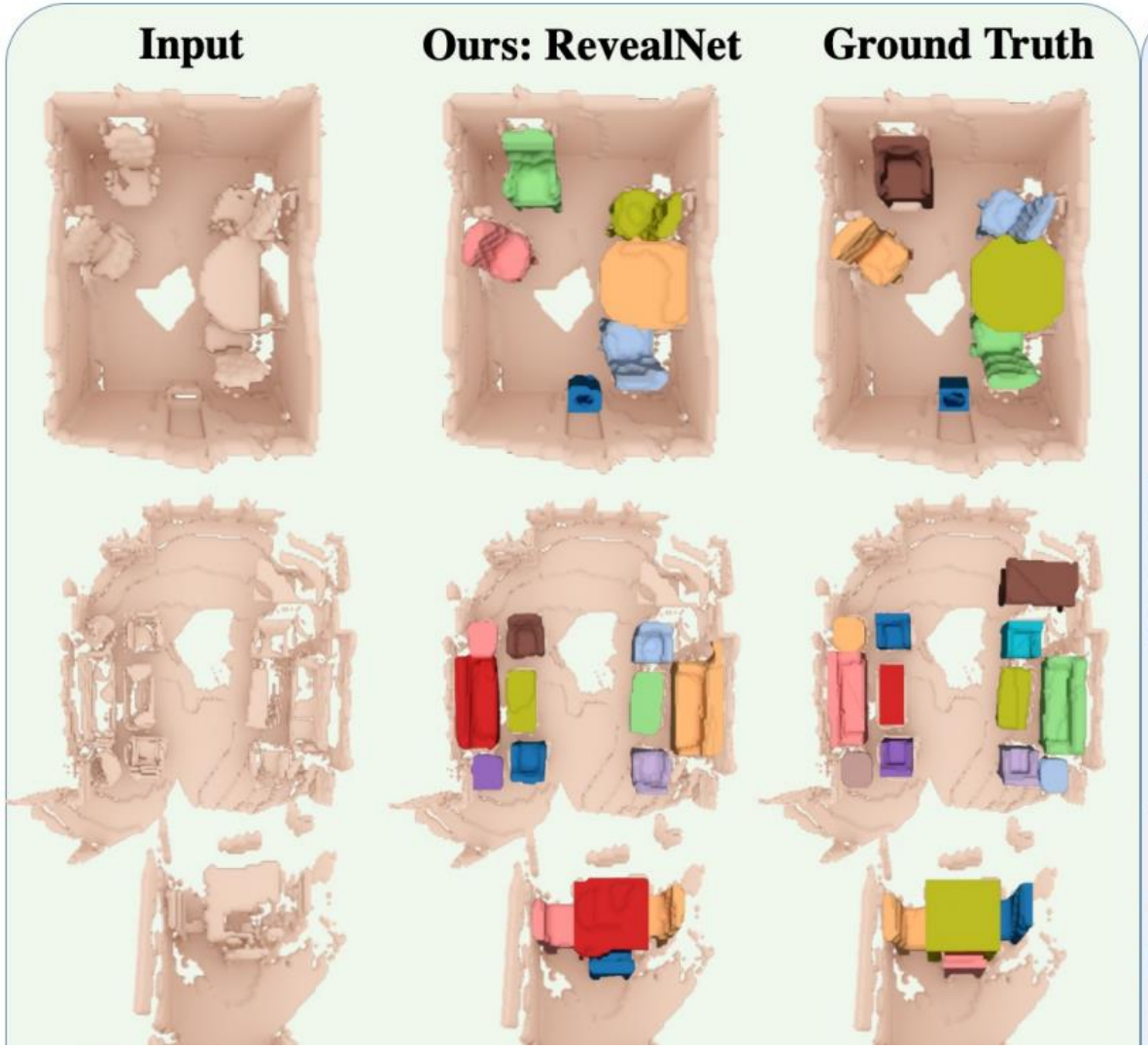
- Can we get better priors by hallucinating missing geometry?

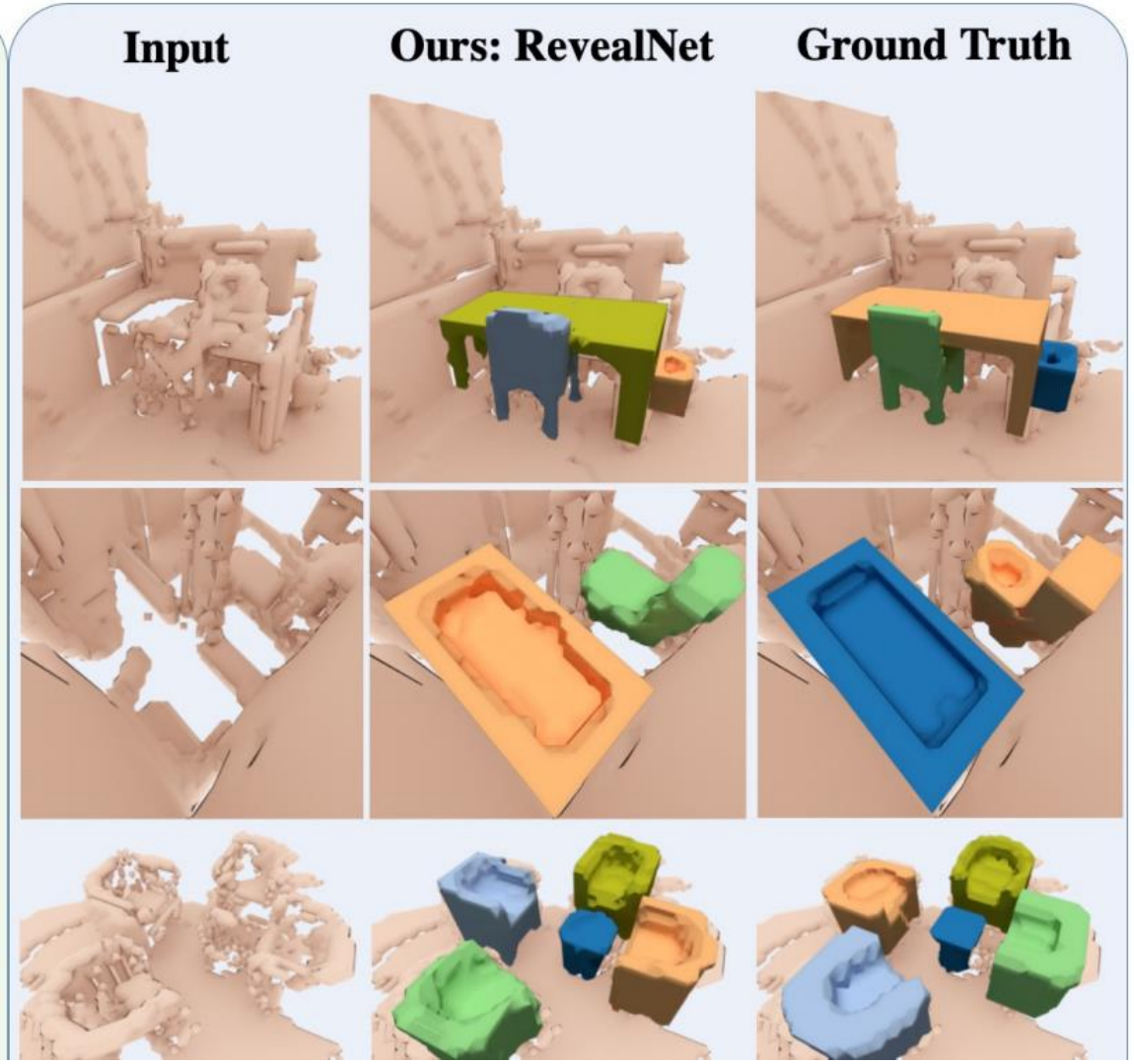


Predict mask for complete object geometry

[Hou et al. '20]

RevealNet





[Hou et al. ’20]

RevealNet

- End-to-end training for instance completion task.

	mAP@0.5
Completion + Instance	5.24
Instance + Completion	5.49
RevealNet (end-to-end)	**21.77**

[Hou et al. '20]

RevealNet

- Geometric completion helps instance segmentation

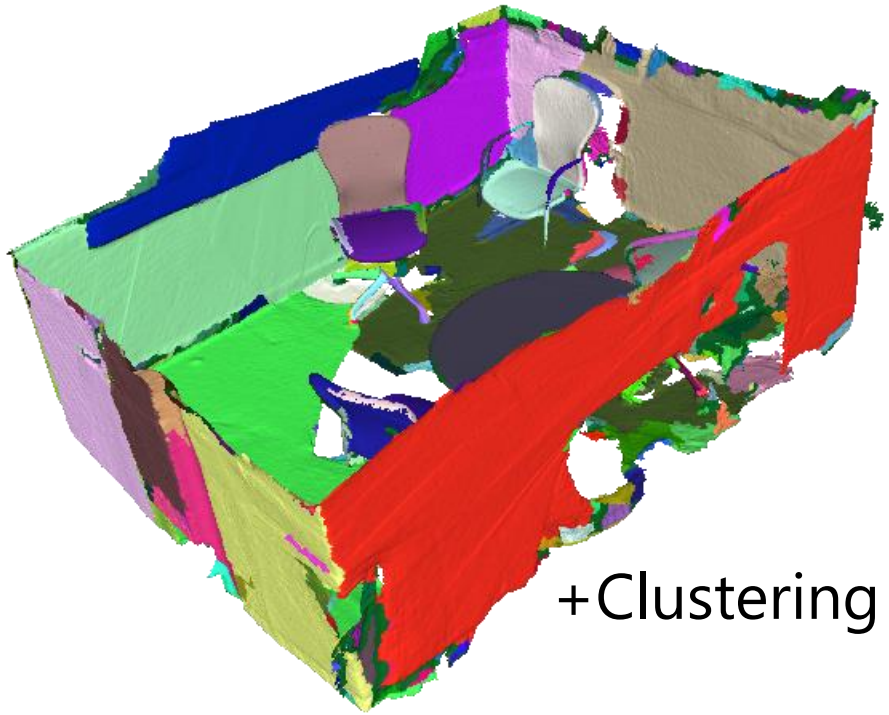
	mAP@0.5
3D-SIS	20.78
RevealNet (no completion)	24.49
RevealNet	**30.52**

[Hou et al. '20]

Top-down, anchor-based challenges

- More diverse objects -> more anchors to cover possibilities
- Easy for thin/very anisotropic objects to miss overlap with target
- Predicting objectness for many empty space locations -> inefficient

Top-down vs bottom-up

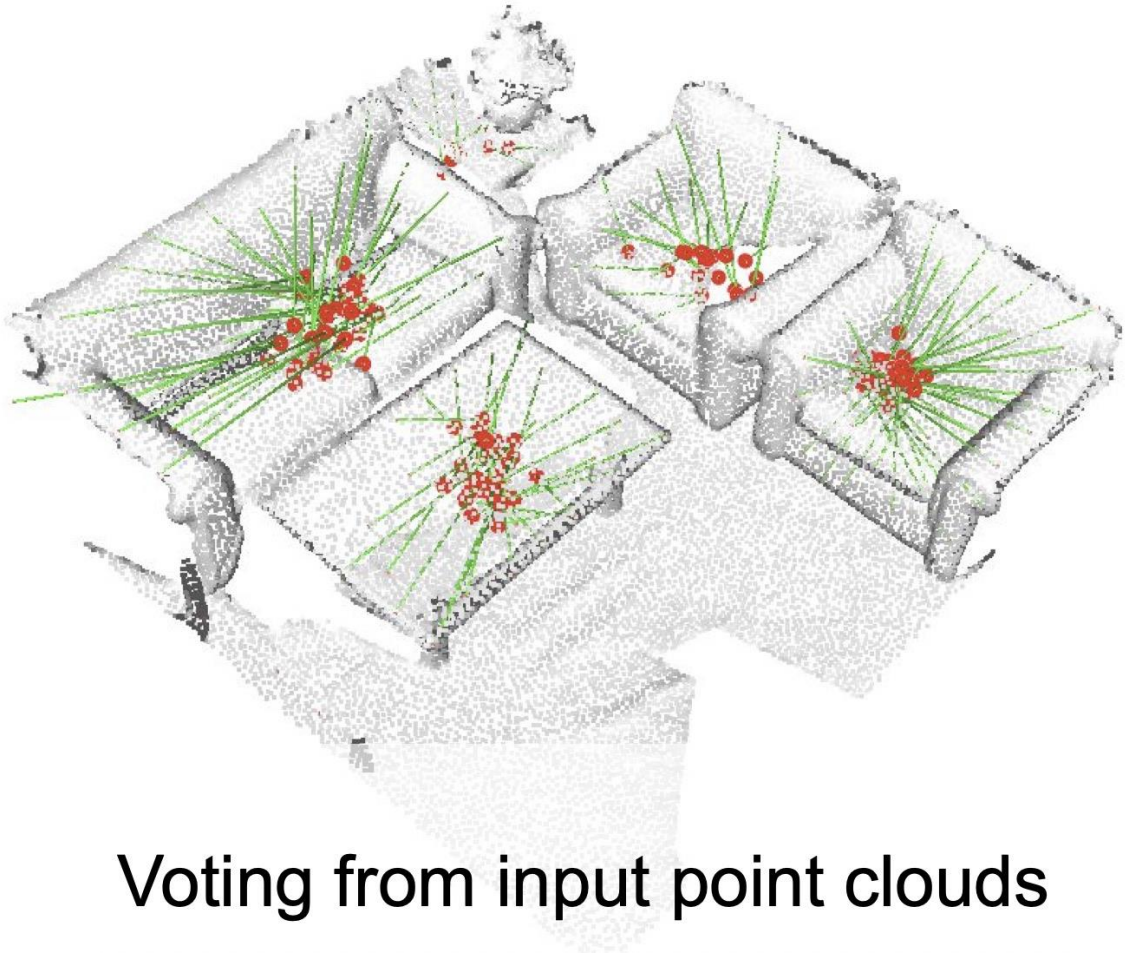


+Clustering

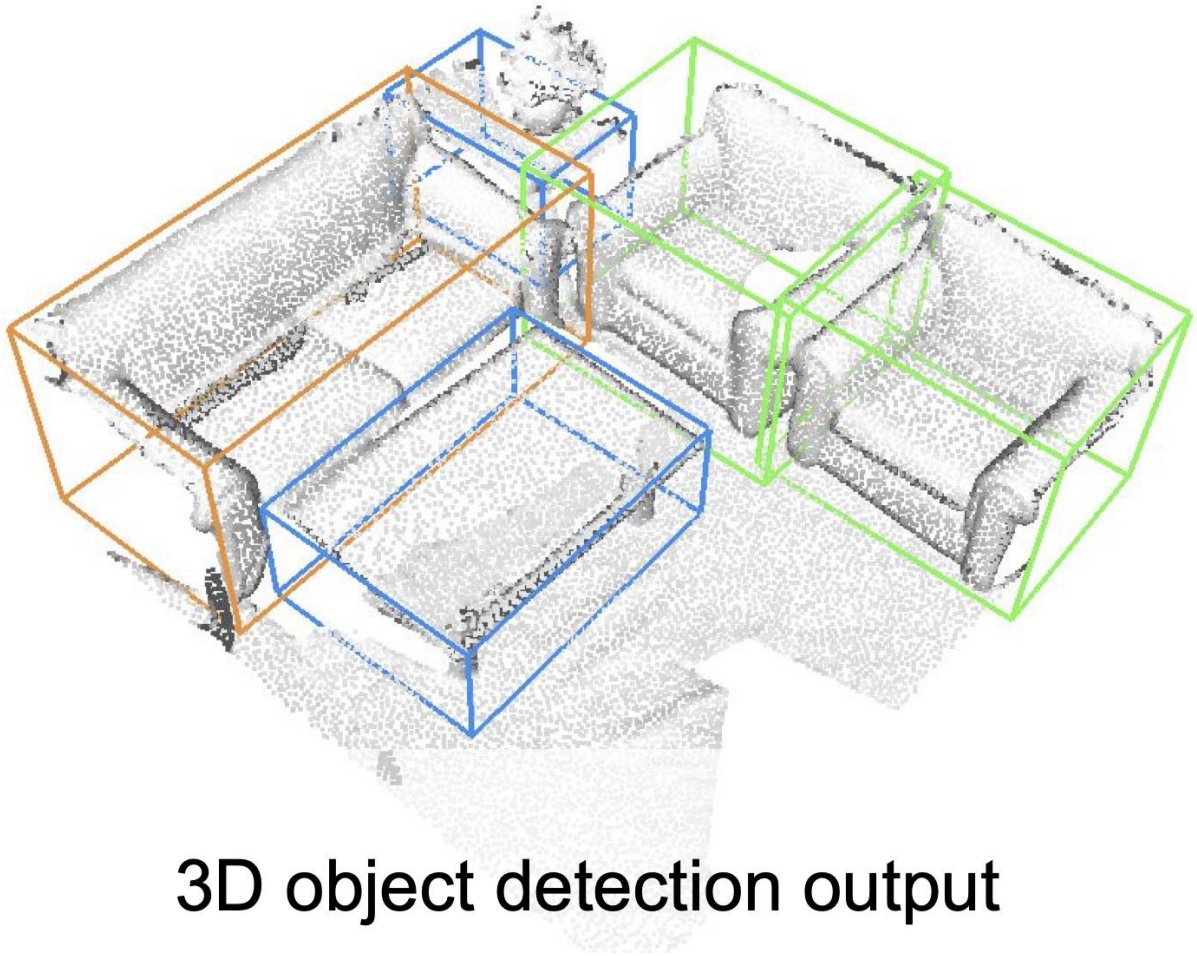
Bottom-Up

Top-Down

VoteNet



Voting from input point clouds



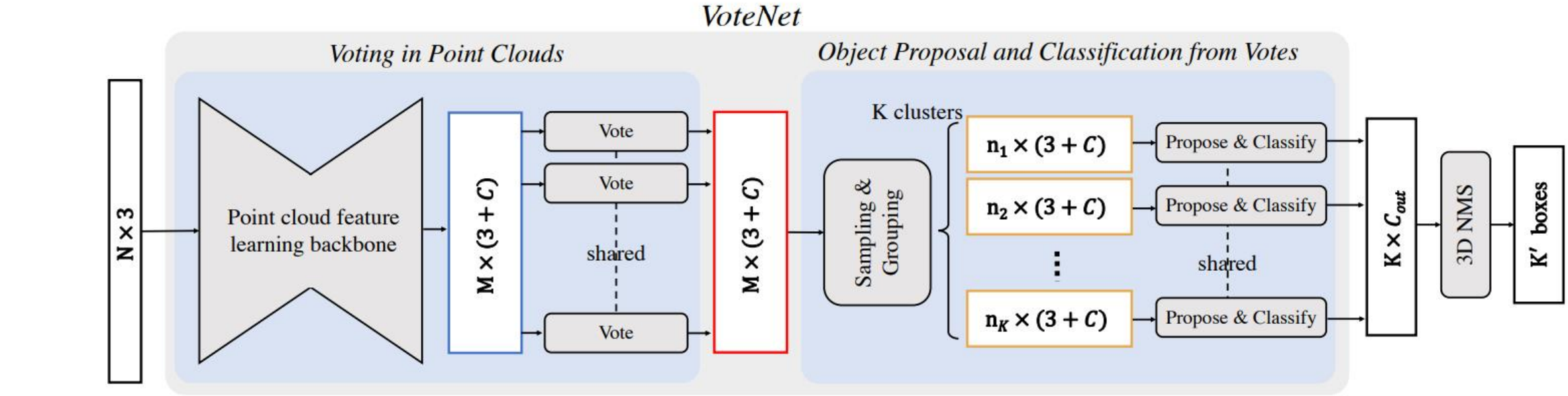
3D object detection output

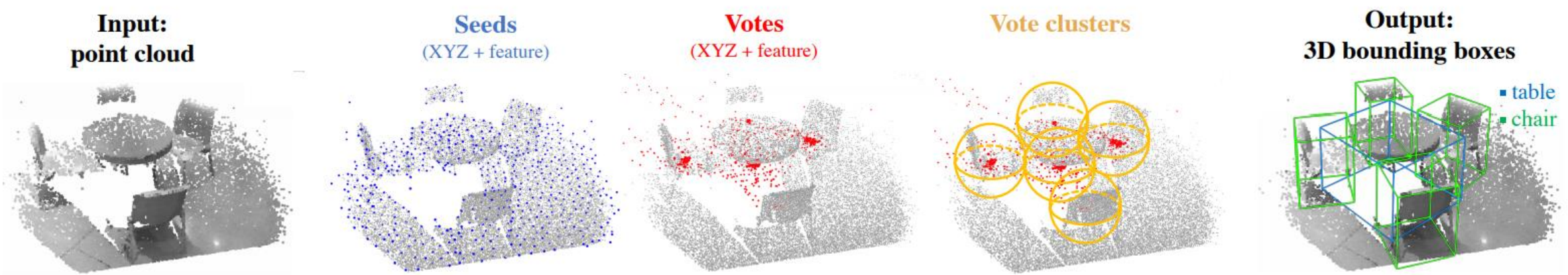
[Qi et al. '19]

Point Cloud Object Detection

- Consider: anchors on point cloud?
  - If anchors on points in point cloud -> often don't coincide with object center
  - If anchors not restricted to point cloud -> how to distribute them?
- Different way to generate object proposals?
- Hough transform (typically line-finding): vote for candidates

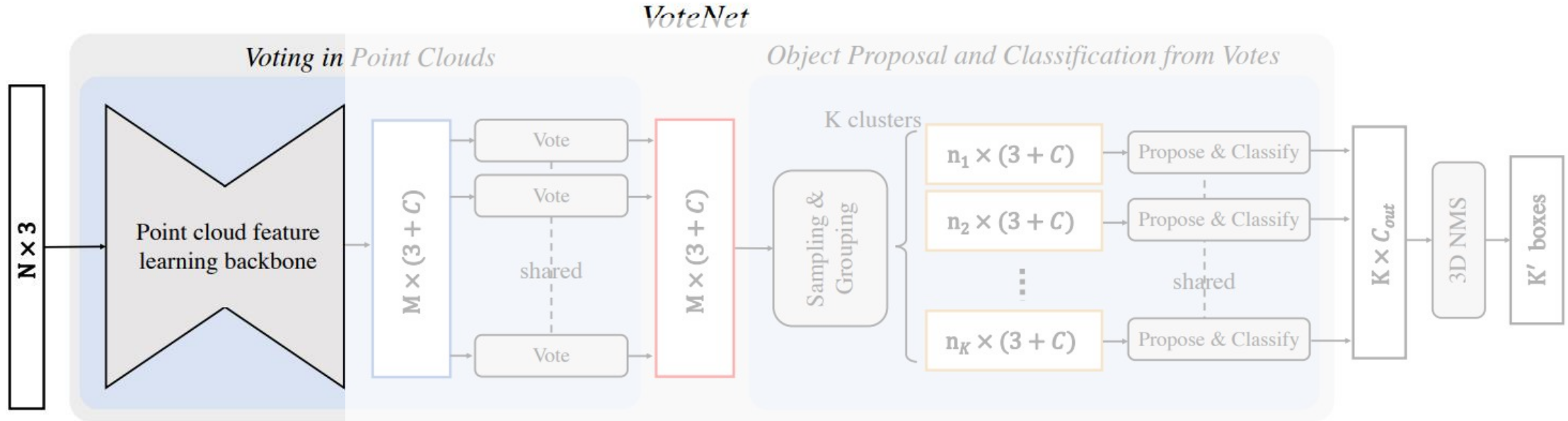
VoteNet





[Qi et al. '19]

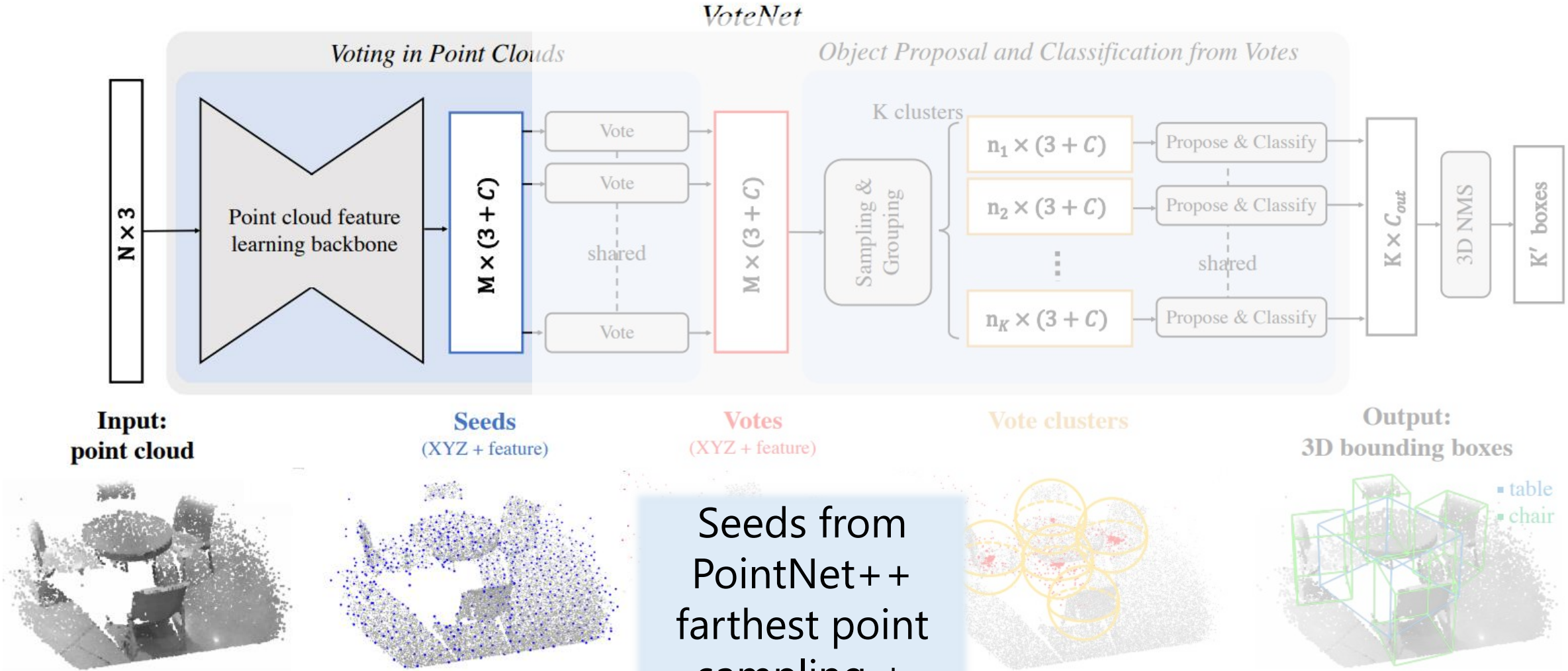
VoteNet



PointNet++

[Qi et al. '19]

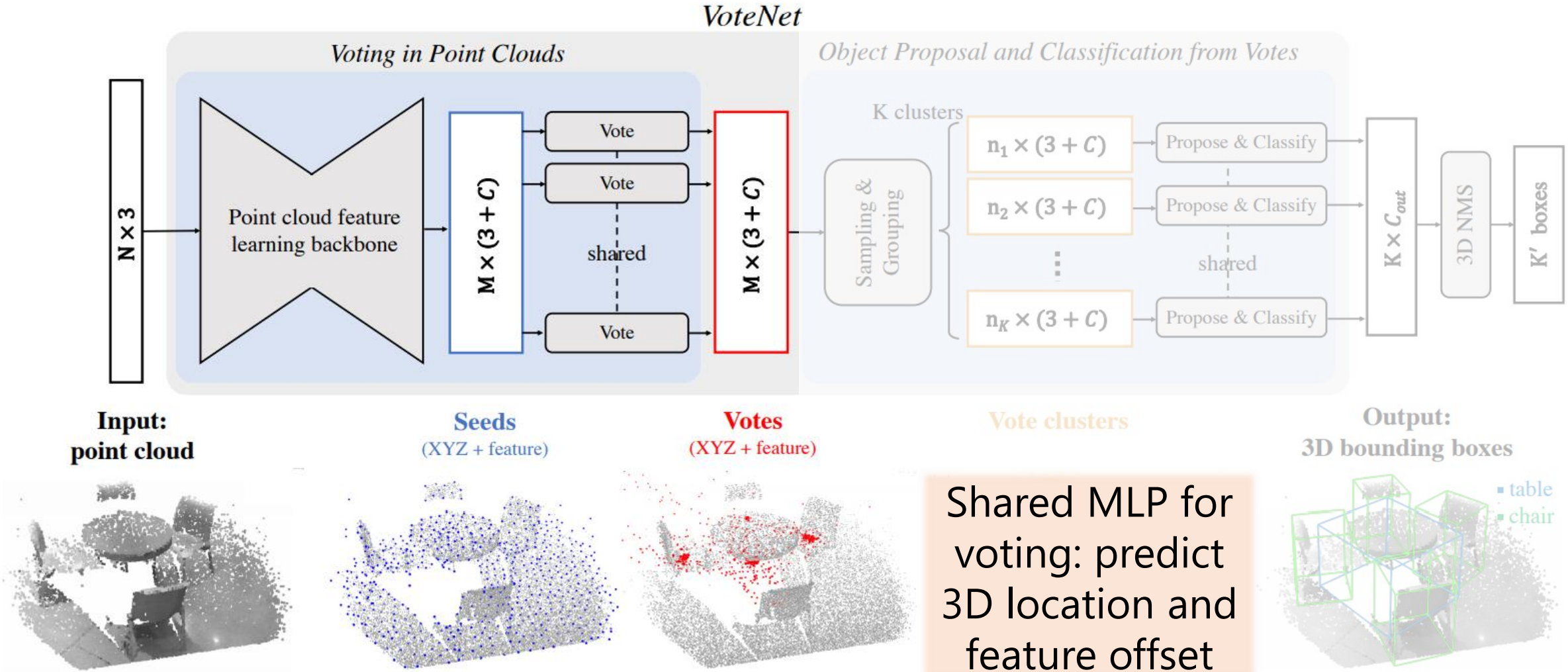
VoteNet



Seeds from PointNet++ farthest point sampling + PointNet

[Qi et al. '19]

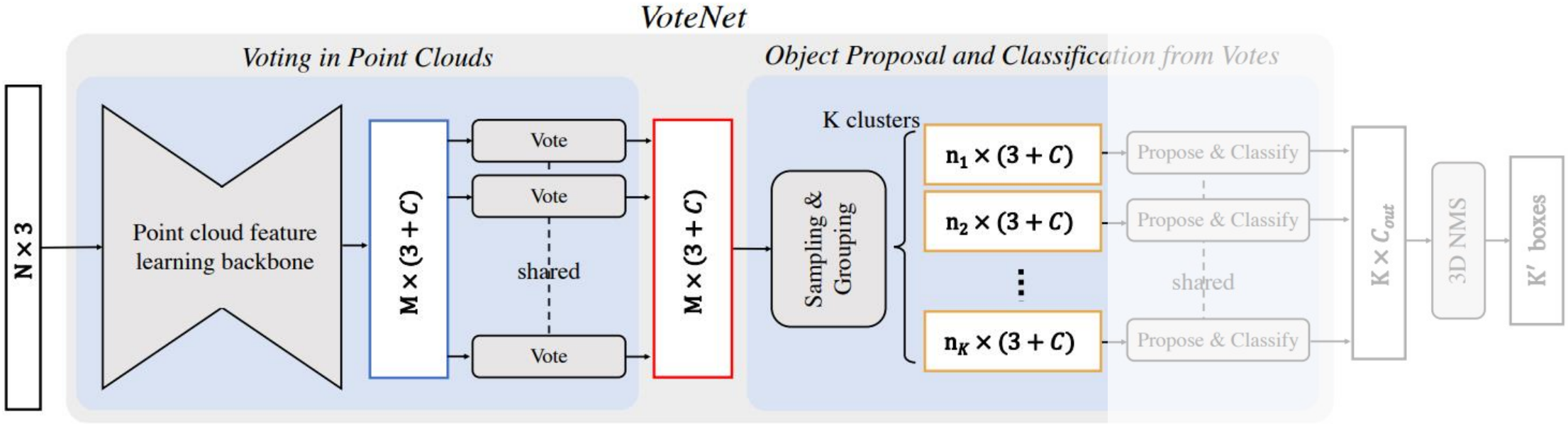
VoteNet

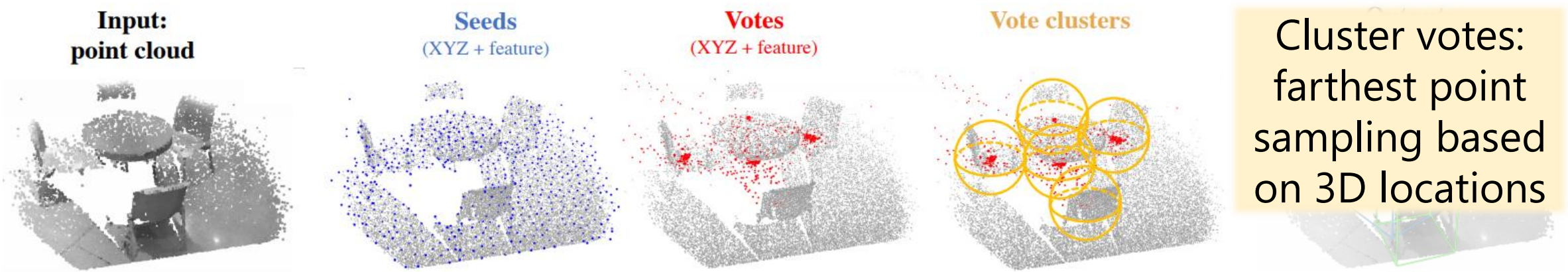


Shared MLP for voting: predict 3D location and feature offset

[Qi et al. '19]

VoteNet

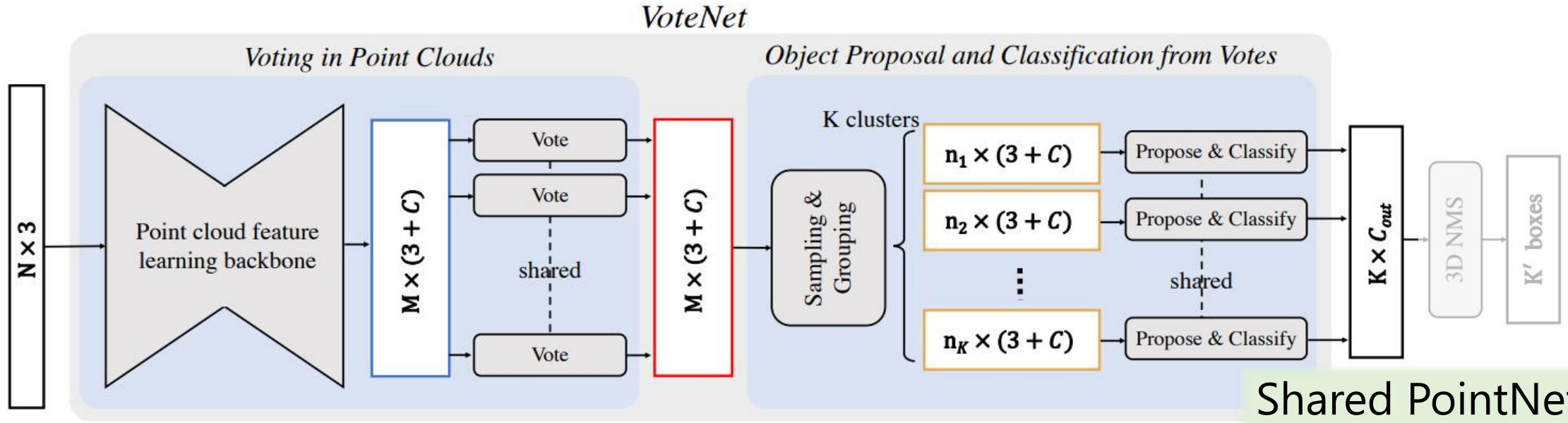


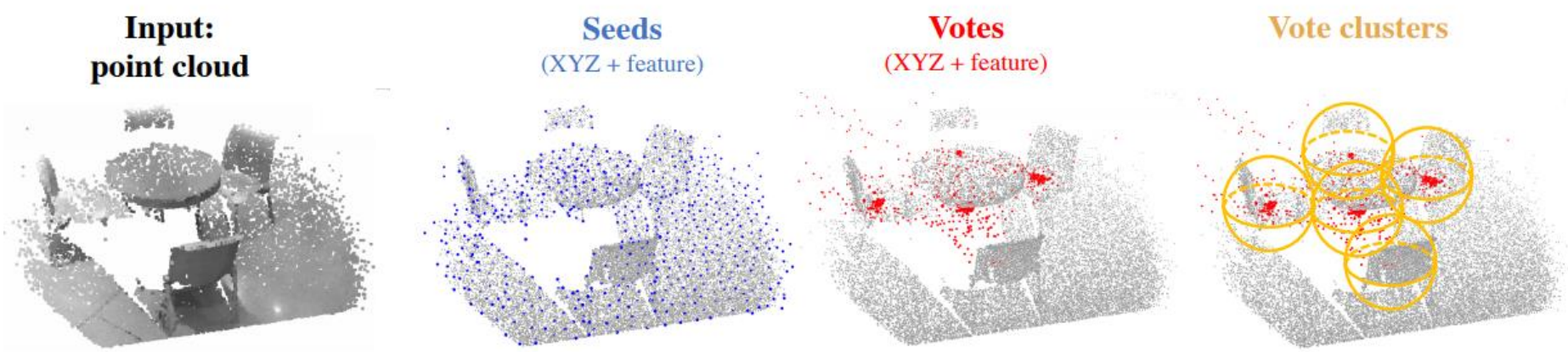


Cluster votes: farthest point sampling based on 3D locations

[Qi et al. '19]

VoteNet



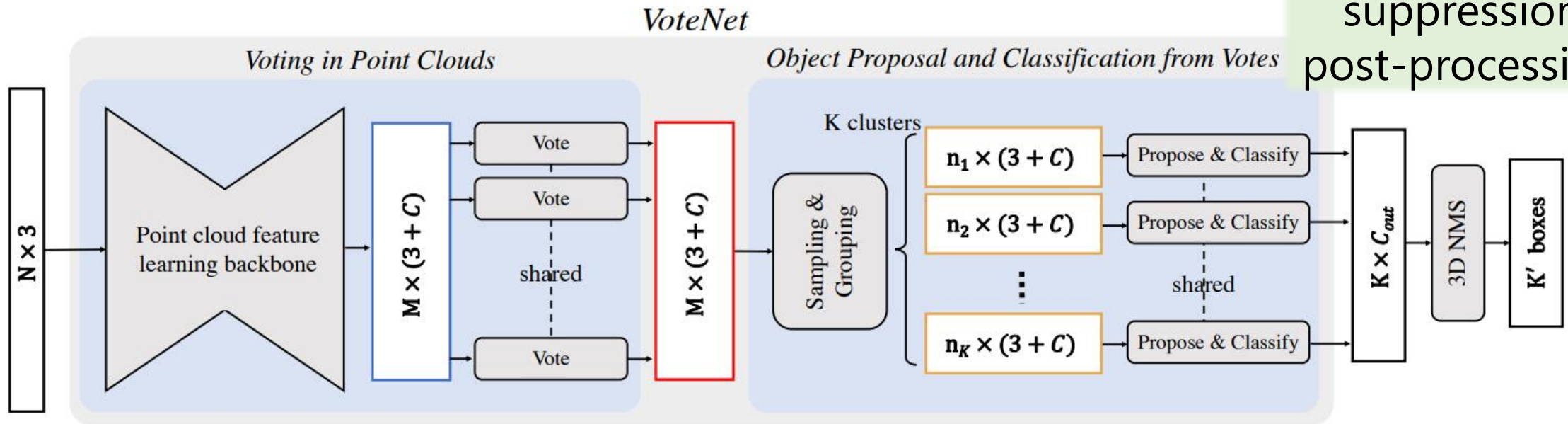


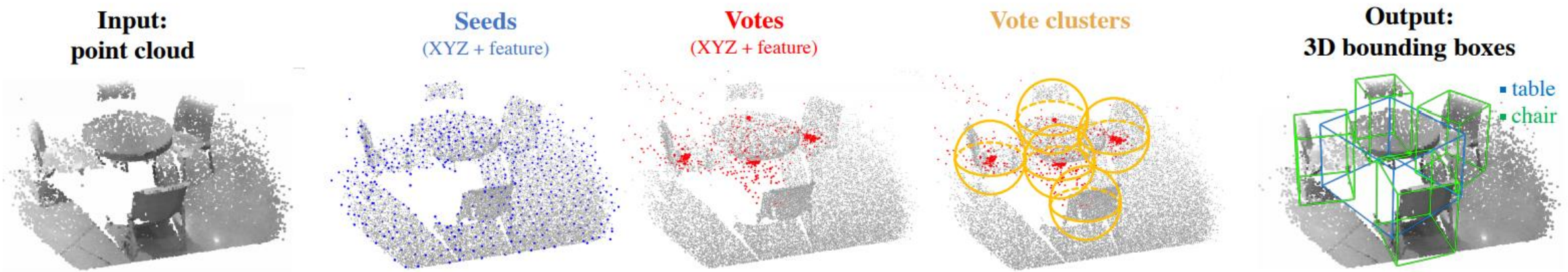
Shared PointNet to process vote cluster: predict objectness, bounding box parameters, class category

[Qi et al. '19]

VoteNet

3D non-max suppression post-processing





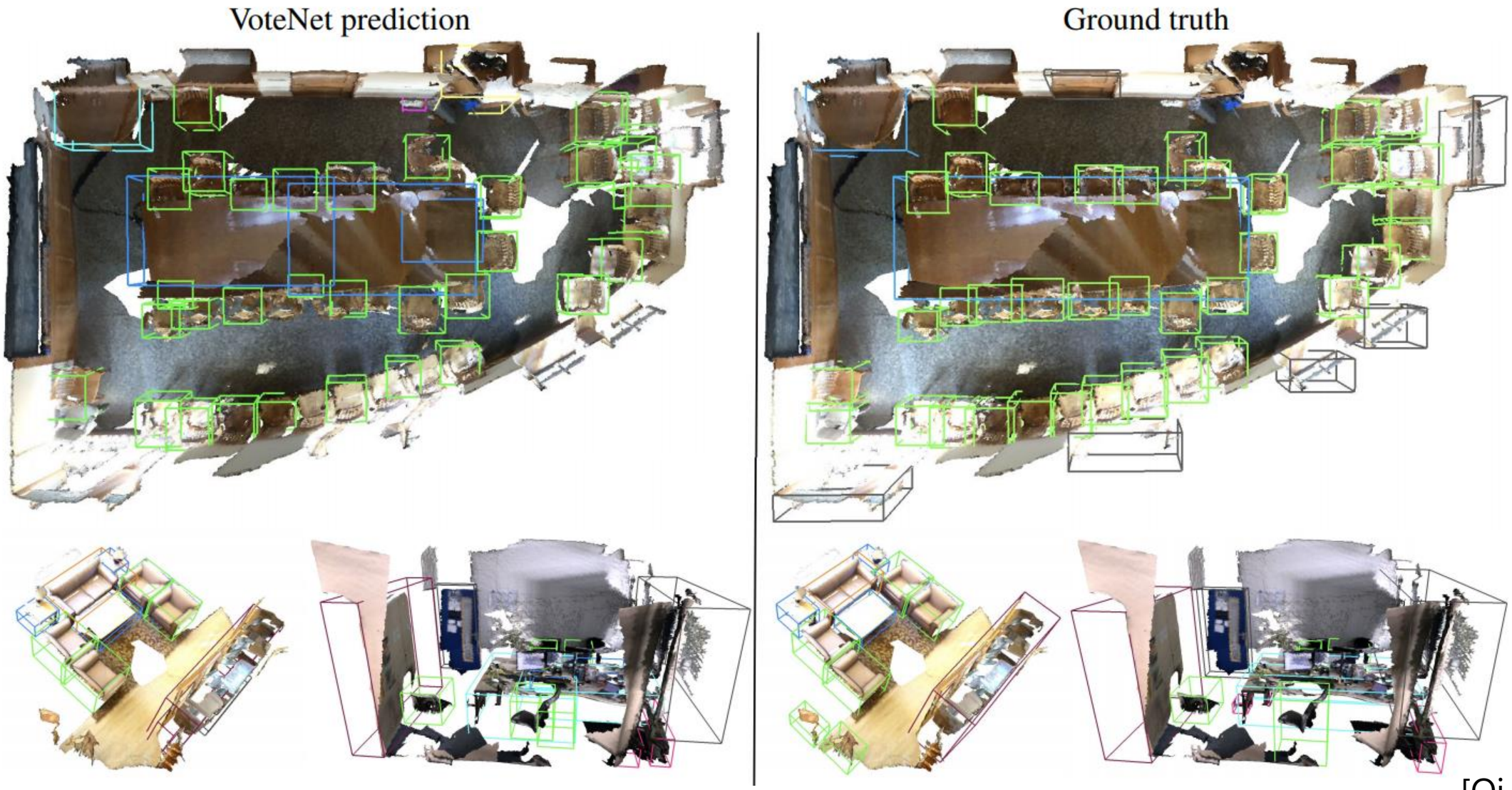
[Qi et al. '19]

VoteNet

	Input	mAP@0.25	mAP@0.5
DSS [42, 12]	Geo + RGB	15.2	6.8
MRCNN 2D-3D [11, 12]	Geo + RGB	17.3	10.5
F-PointNet [34, 12]	Geo + RGB	19.8	10.8
GSPN [54]	Geo + RGB	30.6	17.7
3D-SIS [12]	Geo + 1 view	35.1	18.7
3D-SIS [12]	Geo + 3 views	36.6	19.0
3D-SIS [12]	Geo + 5 views	40.2	22.5
3D-SIS [12]	Geo only	25.4	14.6
VoteNet (ours)	Geo only	**58.6**	**33.5**

[Qi et al. '19]

VoteNet



[Qi et al. '19]

3D-MPA

- Multi Proposal Aggregation for 3D Semantic Instance Segmentation

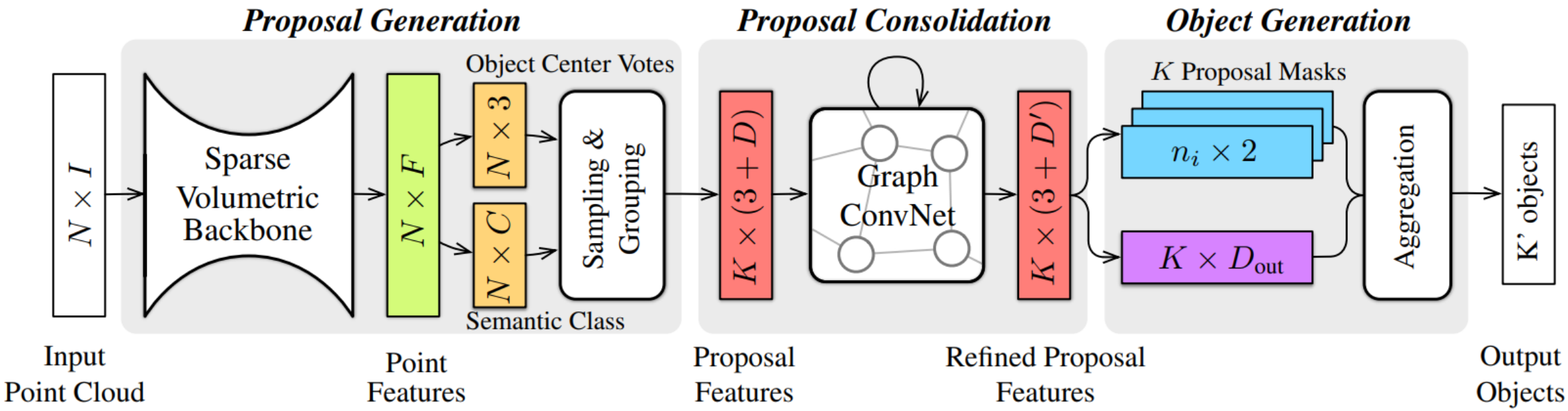
Input: 3D Point Cloud

Object Center Votes & Aggregated Proposals

Output: 3D Semantic Instances

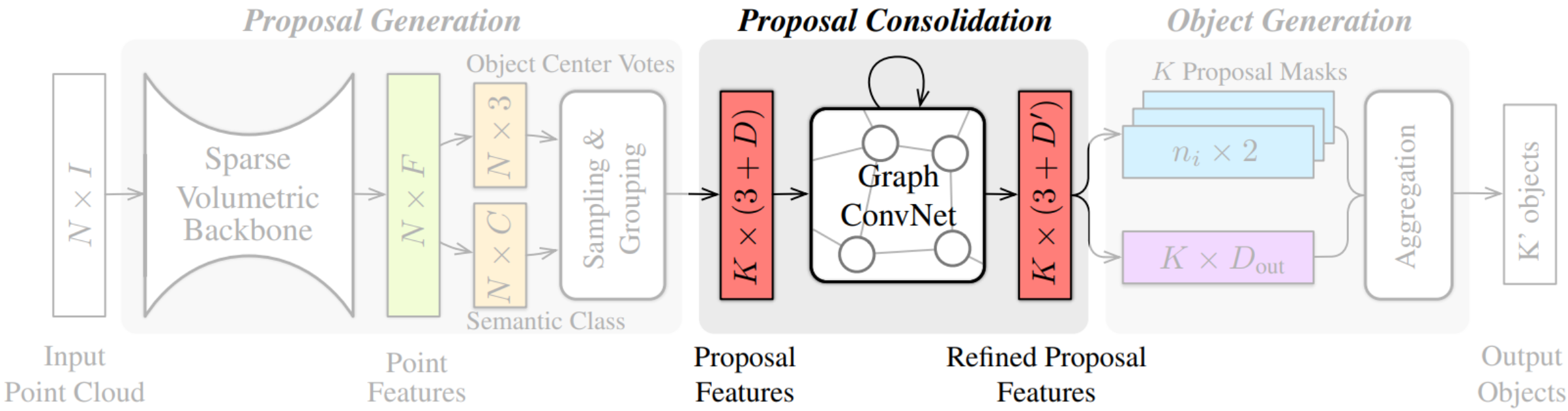
[Engelmann et al. '20]

3D-MPA



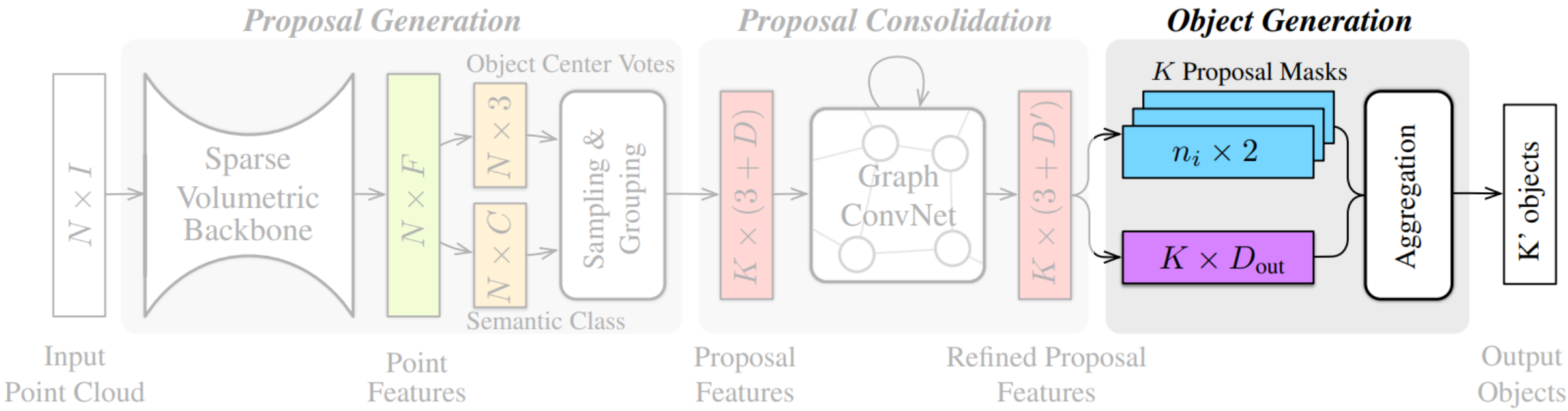
[Engelmann et al. '20]

3D-MPA



[Engelmann et al. '20]

3D-MPA



Avoid non-max suppression

- Aggregate proposals together

[Engelmann et al. '20]

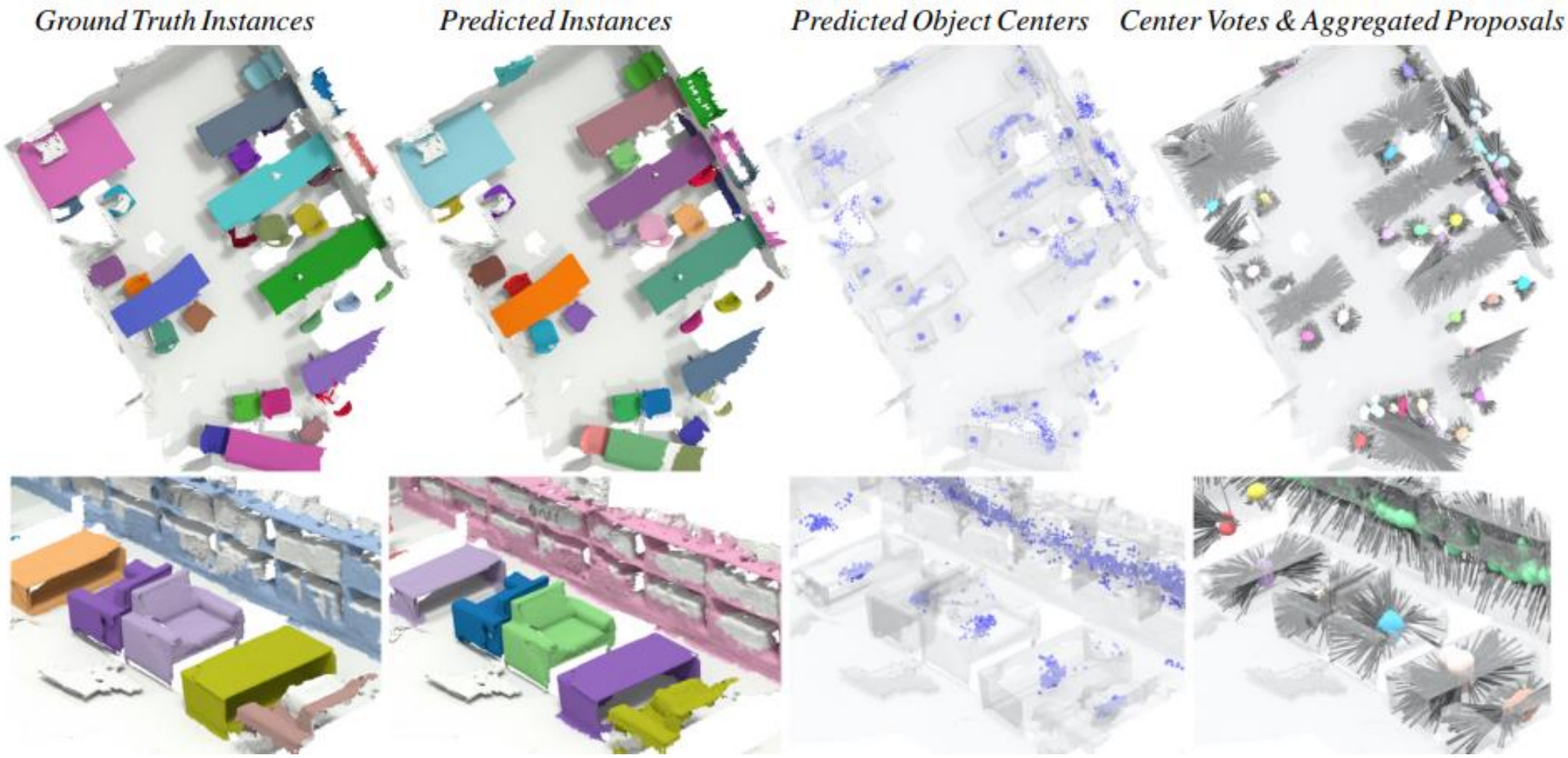
3D-MPA

3D Object Detection		
ScanNetV2	mAP@25%	mAP@50%
DSS [37]	15.2	6.8
MRCNN 2D-3D [17]	17.3	10.5
F-PointNet [30]	19.8	10.8
GSPN [50]	30.6	17.7
3D-SIS [18]	40.2	22.5
VoteNet [29]	58.6	33.5
3D-MPA (Ours)	**64.2**	**49.2**

3D Instance Segmentation						
ScanNetV2	Validation Set			Hidden Test Set		
	mAP	@50%	@25%	mAP	@50%	@25%
SGPN [44]	-	11.3	22.2	4.9	14.3	39.0
3D-BEVIS [10]	-	-	-	11.7	24.8	40.1
3D-SIS [18]	-	18.7	35.7	16.1	38.2	55.8
GSPN [50]	19.3	37.8	53.4	15.8	30.6	54.4
3D-BoNet [49]	-	-	-	25.3	48.8	68.7
MTML [19]	20.3	40.2	55.4	28.2	54.9	73.1
3D-MPA (Ours)	**35.3**	**59.1**	**72.4**	**35.5**	**61.1**	**73.7**

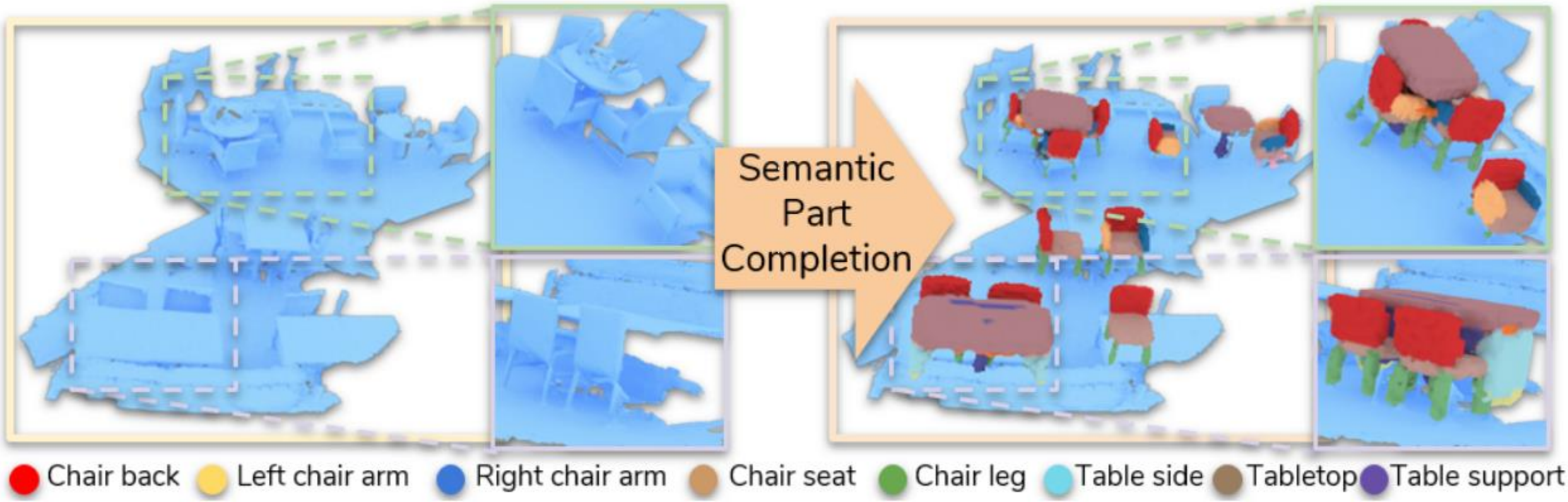
[Engelmann et al. '20]

3D-MPA



[Engelmann et al. ’20]

Towards Part-Based Scene Understanding



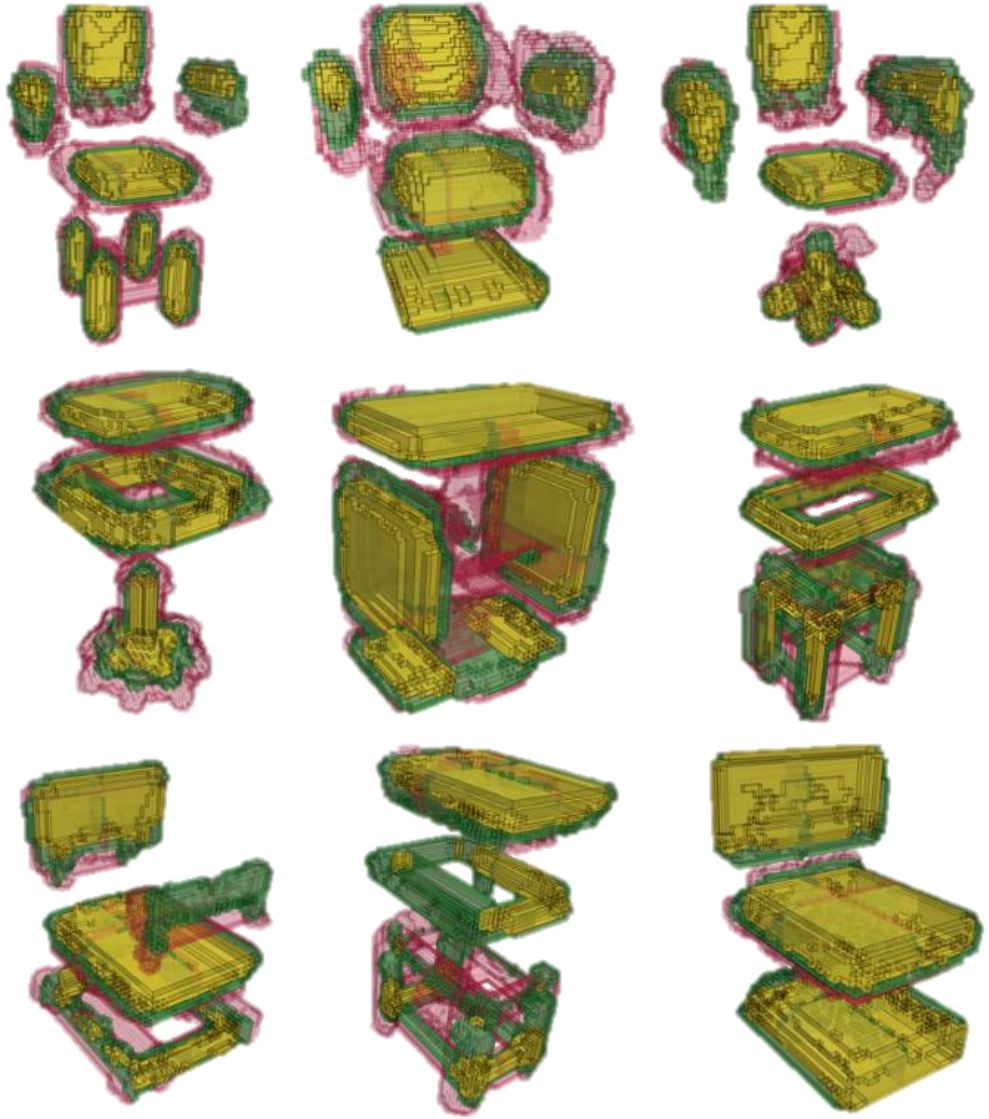
[Bokhovkin et al. '21]

Towards Part-Based Scene Understanding

Chairs

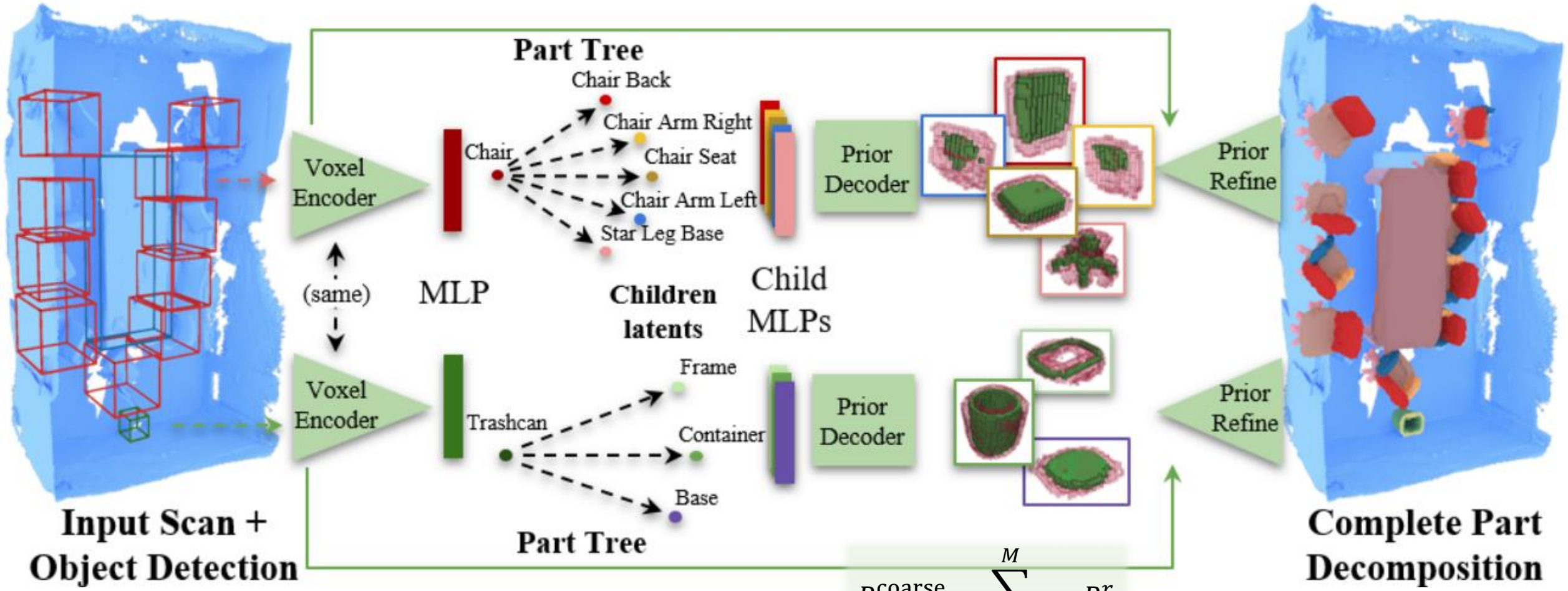
Tables

Beds



[Bokhovkin et al. ’21]

Towards Part-Based Scene Understanding



$$P_k^{\text{coarse}} = \sum_{m=1}^{M} w_m P_m^r$$

[Bokhovkin et al. '21]

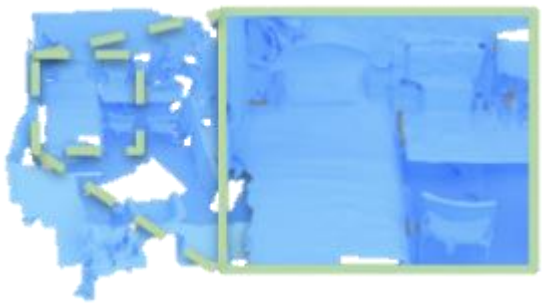
Towards Part-Based Scene Understanding

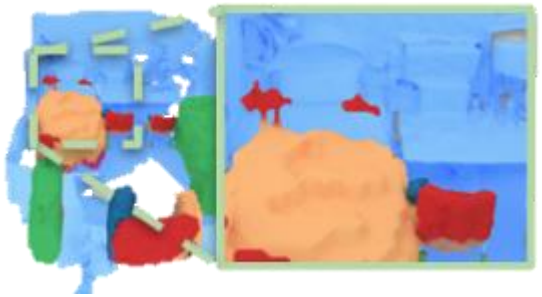
Input

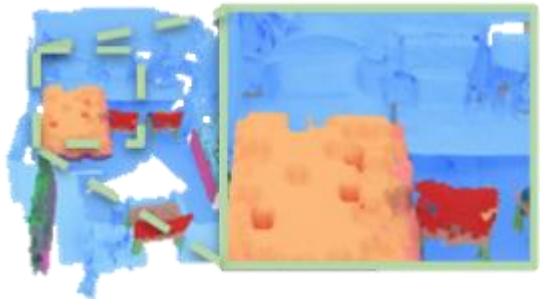
SG-NN + MLCVNet + PointGroup

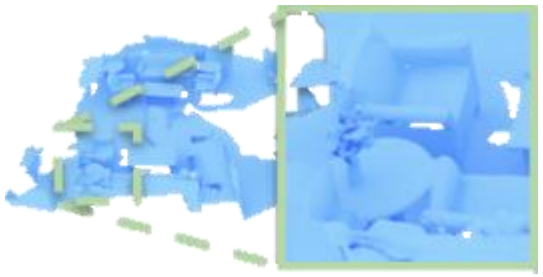
SG-NN + MLCVNet + StructureNet

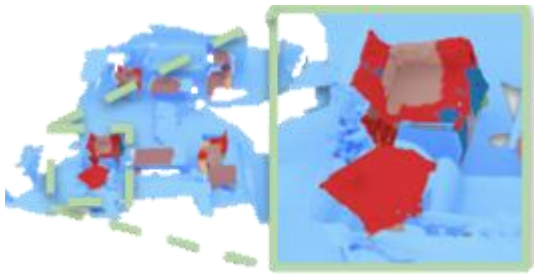
Ours

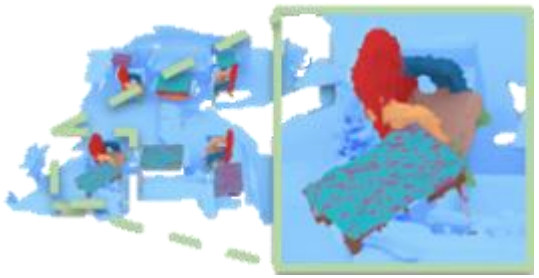


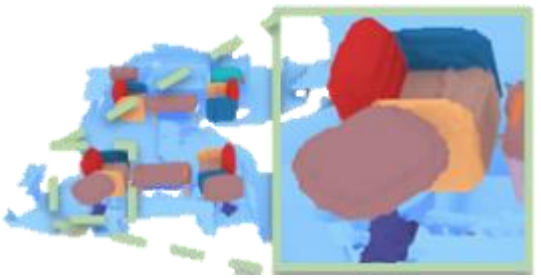








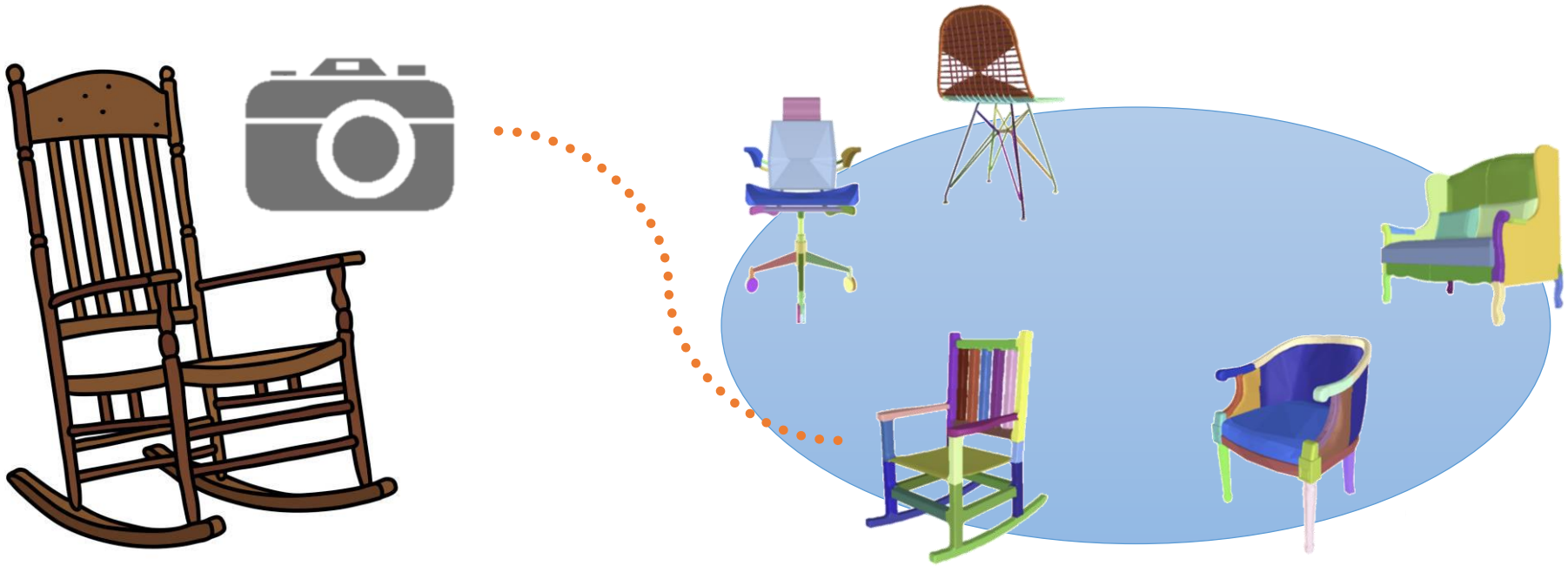




[Bokhovkin et al. ’21]

Learned Synthetic Object Priors

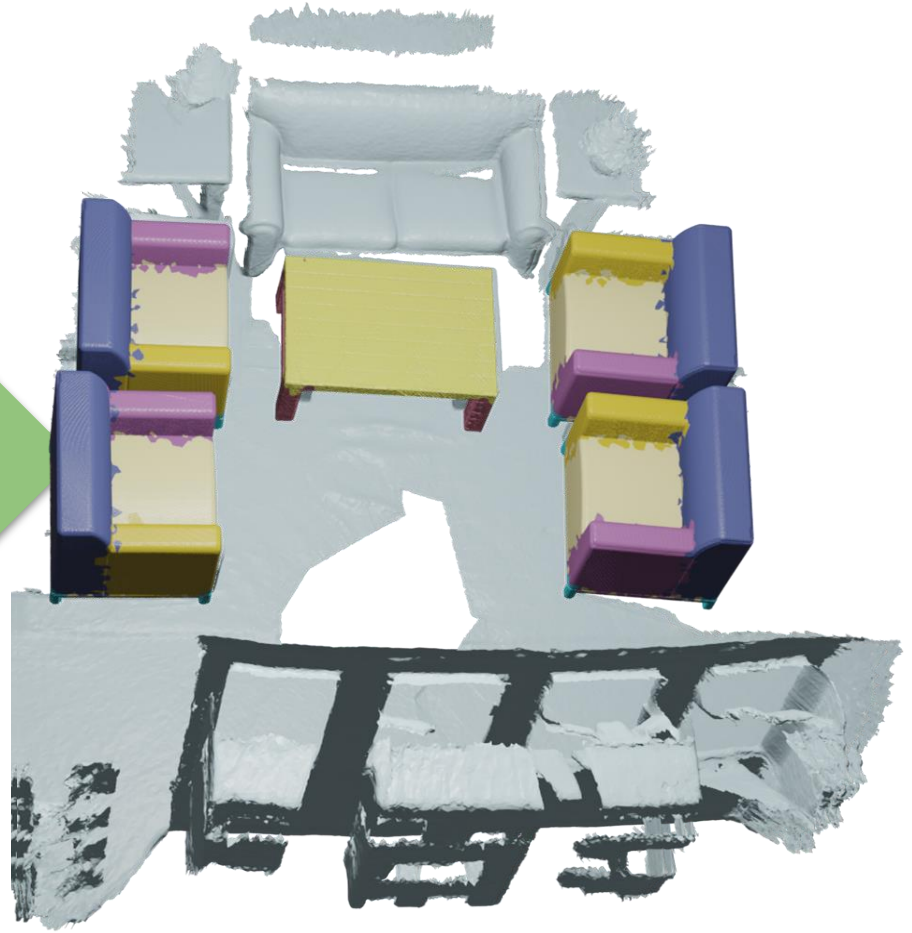
- Learn manifold of synthetic shapes and parts
- Optimize over manifold at test time to fit to input observations



Understanding Object Parts in Scenes

Part-based Shape Representation

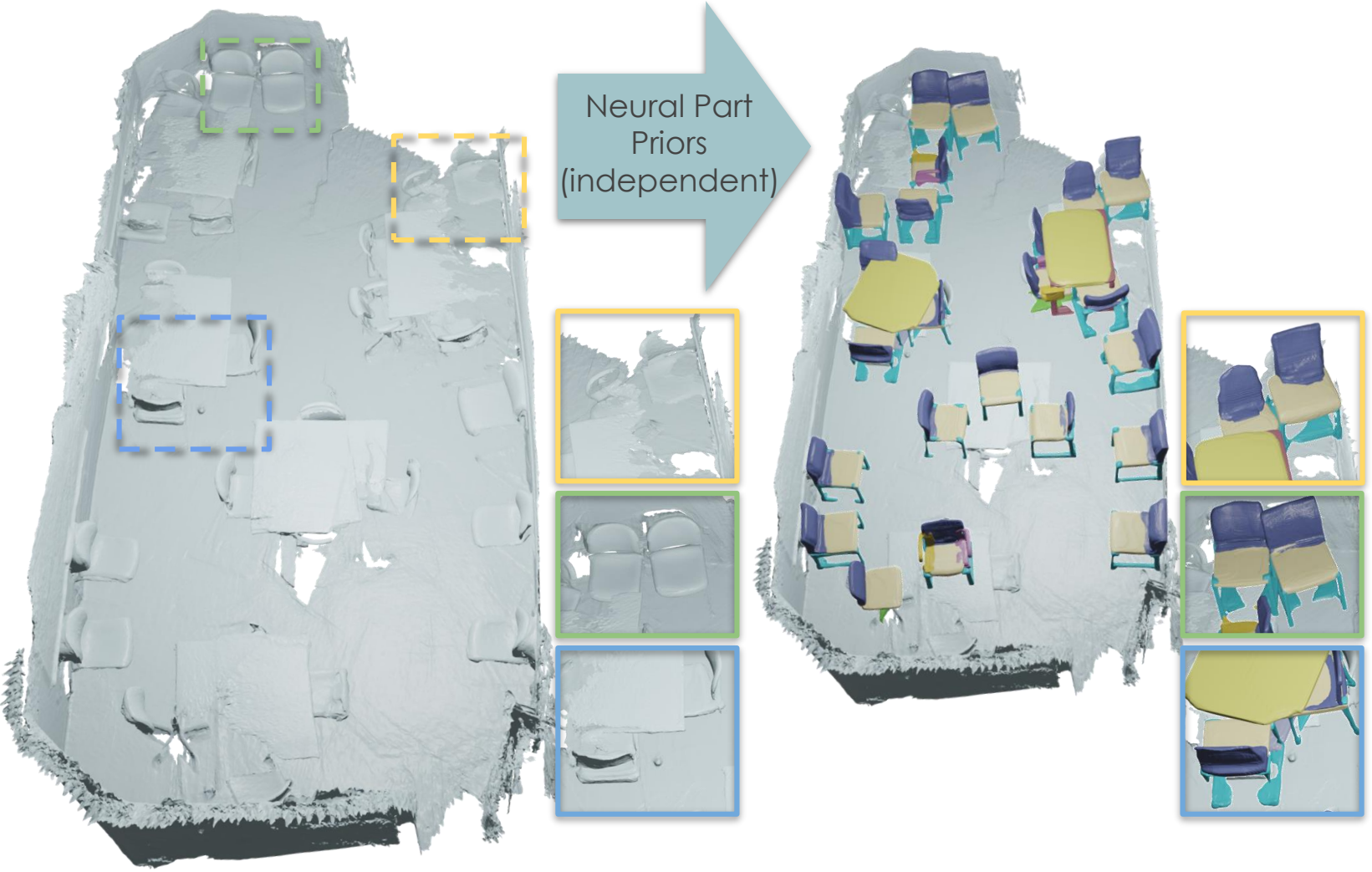
Part Decomposition of Scenes



Neural Part Priors

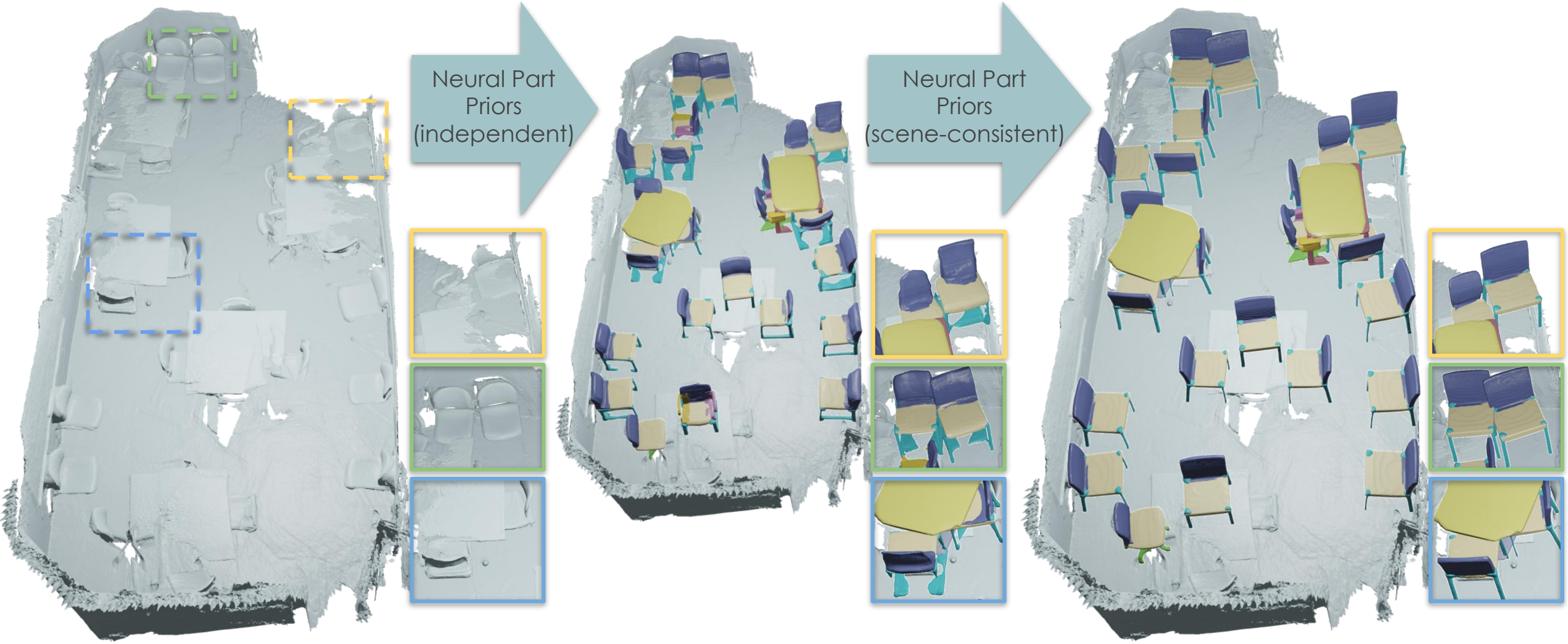
[Bokhovkin et al. ‘23]

Neural Part Priors



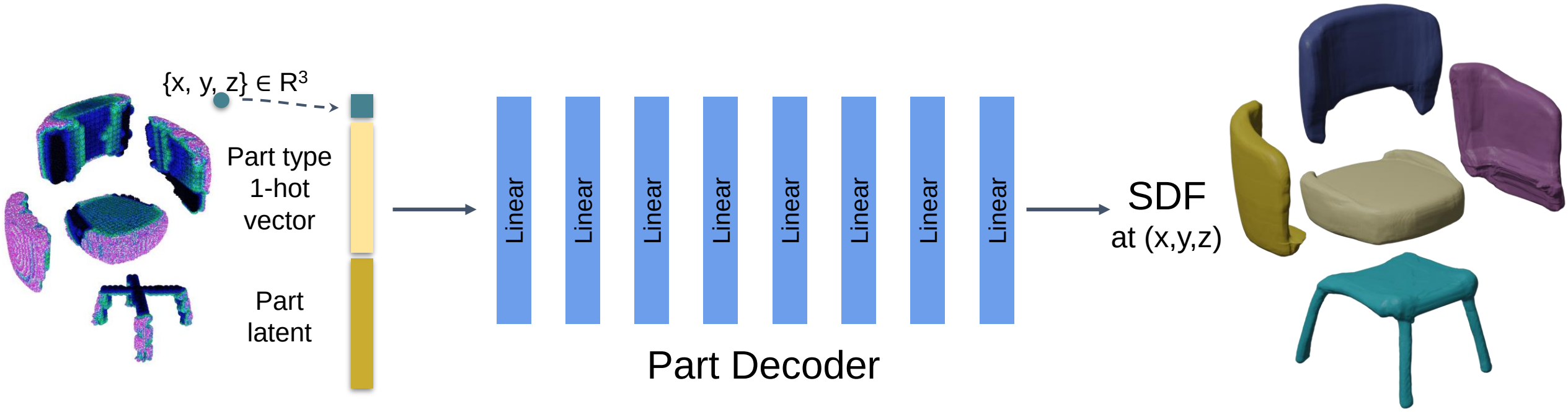
[Bokhovkin et al. ‘23]

Neural Part Priors



[Bokhovkin et al. '23]

Learned Part and Shape Priors

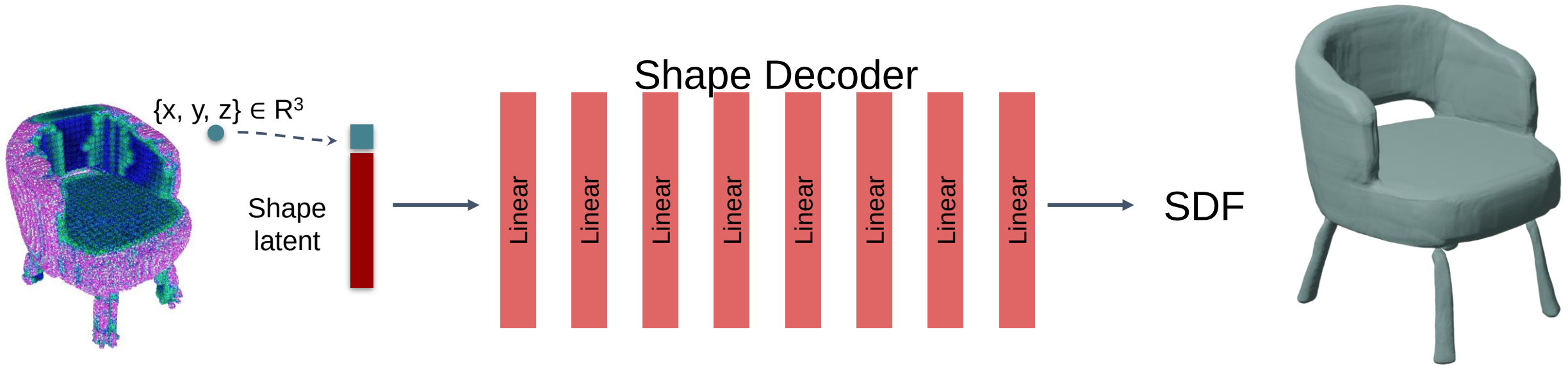


$$f_p : \mathbb{R}^3 \times \mathbb{R}^{256} \times \mathbb{Z}_2^{N_c} \rightarrow \mathbb{R}, \qquad f_p(\mathbf{x}, \mathbf{z}_k^{\mathrm{p}}, \mathbb{1}_{\mathrm{part}}) = d$$

$$L = \sum_{j=1}^{N_p} |f_p(\mathbf{x}_j, \mathbf{z}_k^{\mathrm{p}}, \mathbb{1}_{\mathrm{part}}) - D^{\mathrm{gt}}(\mathbf{x}_j)|_1 + ||\mathbf{z}_k^{\mathrm{p}}||_2^2$$

[Bokhovkin et al. '23]

Learned Part and Shape Priors



$$f_s : \mathbb{R}^3 \times \mathbb{R}^{256} \rightarrow \mathbb{R}, \qquad f_p(\mathbf{x}, \mathbf{z}_i^{\mathrm{s}}) = d$$

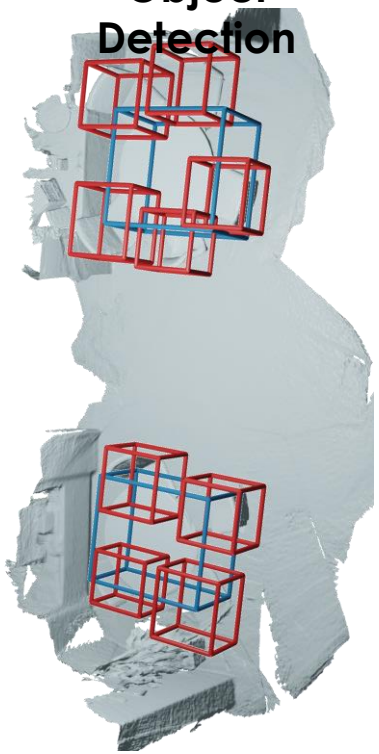
[Bokhovkin et al. ’23]

Interpolations in Part Space

[Bokhovkin et al. ’23]

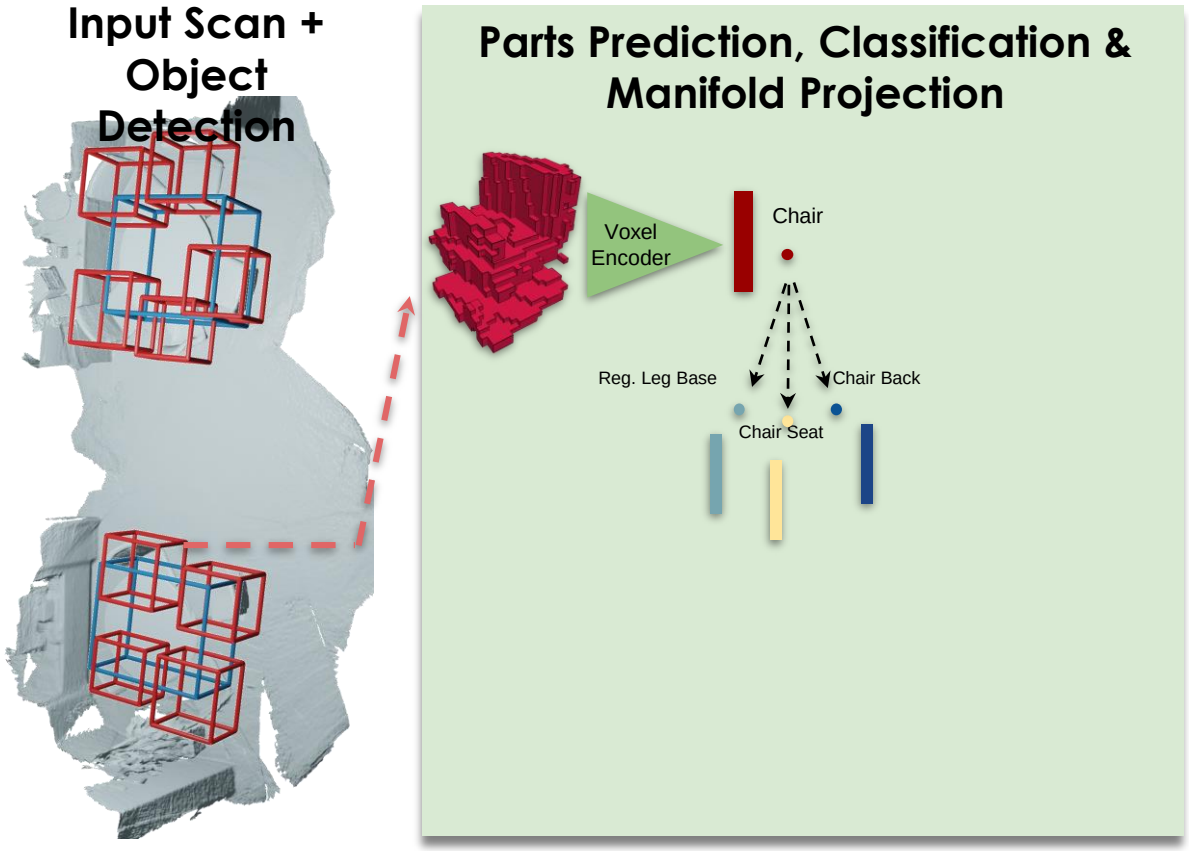
Part-Based Scene Understanding

Input Scan + Object Detection



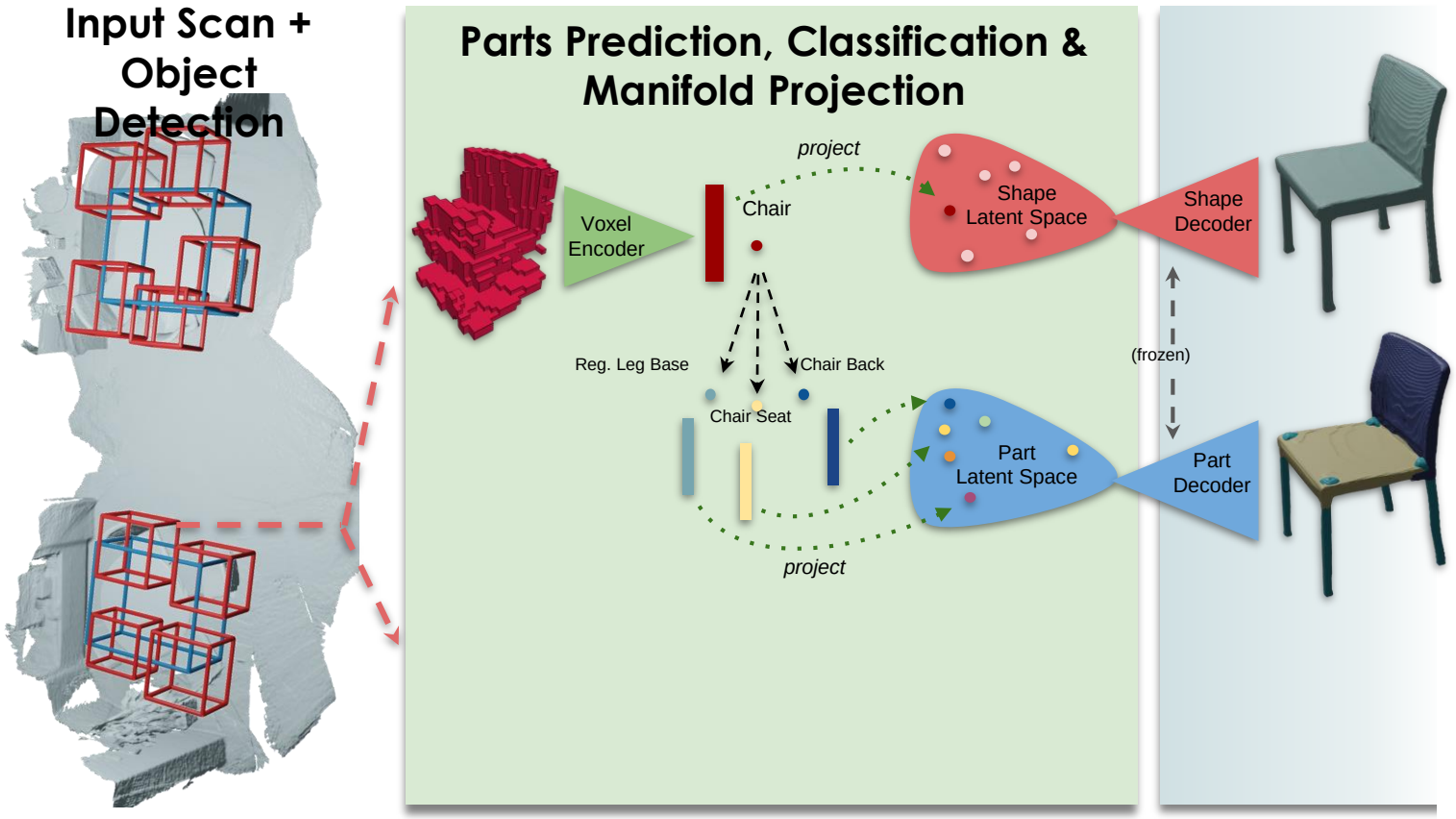
[Bokhovkin et al. ’23]

Predicting Part Semantic Labels



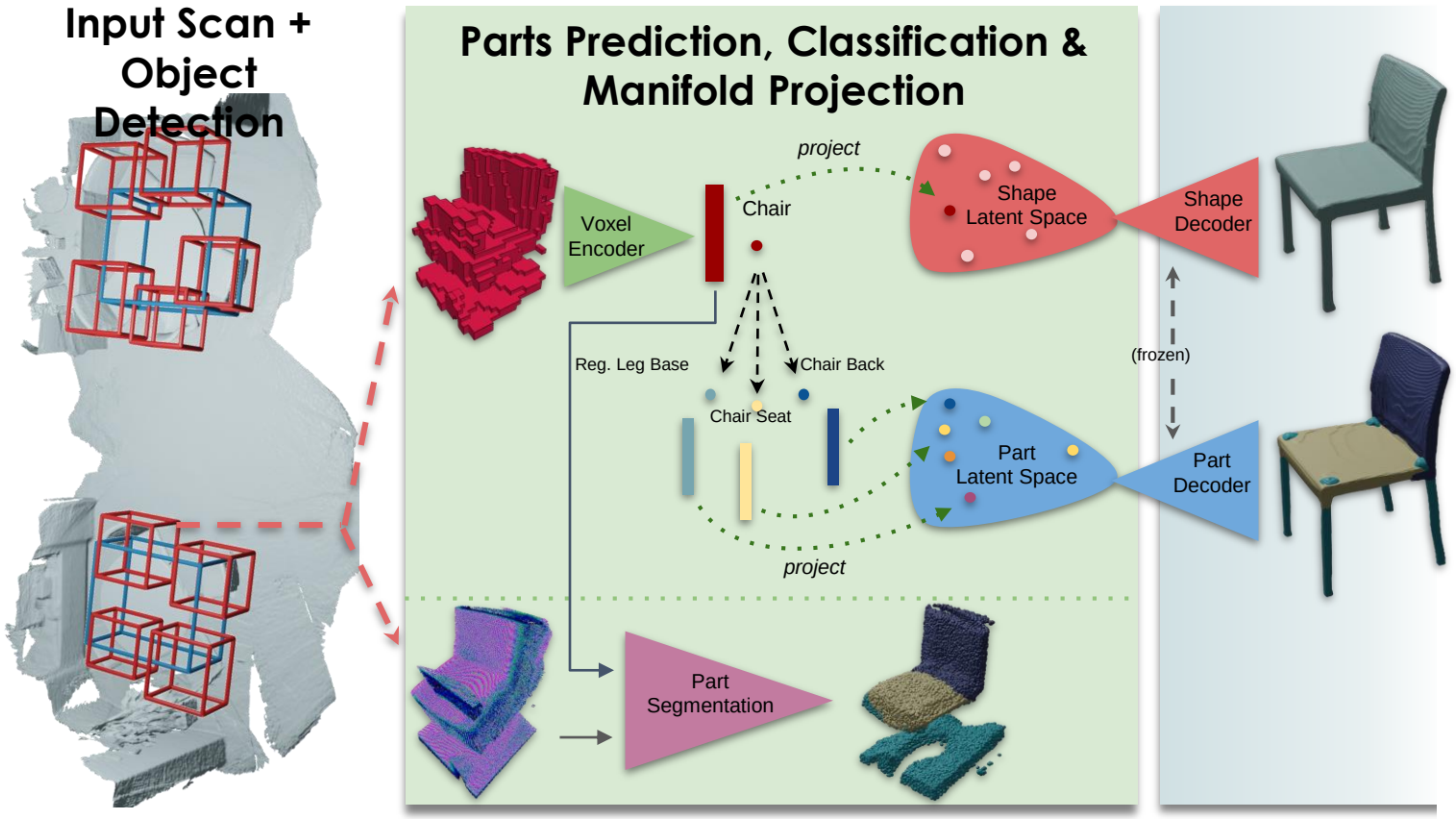
[Bokhovkin et al. '23]

Projection into Learned Part Spaces



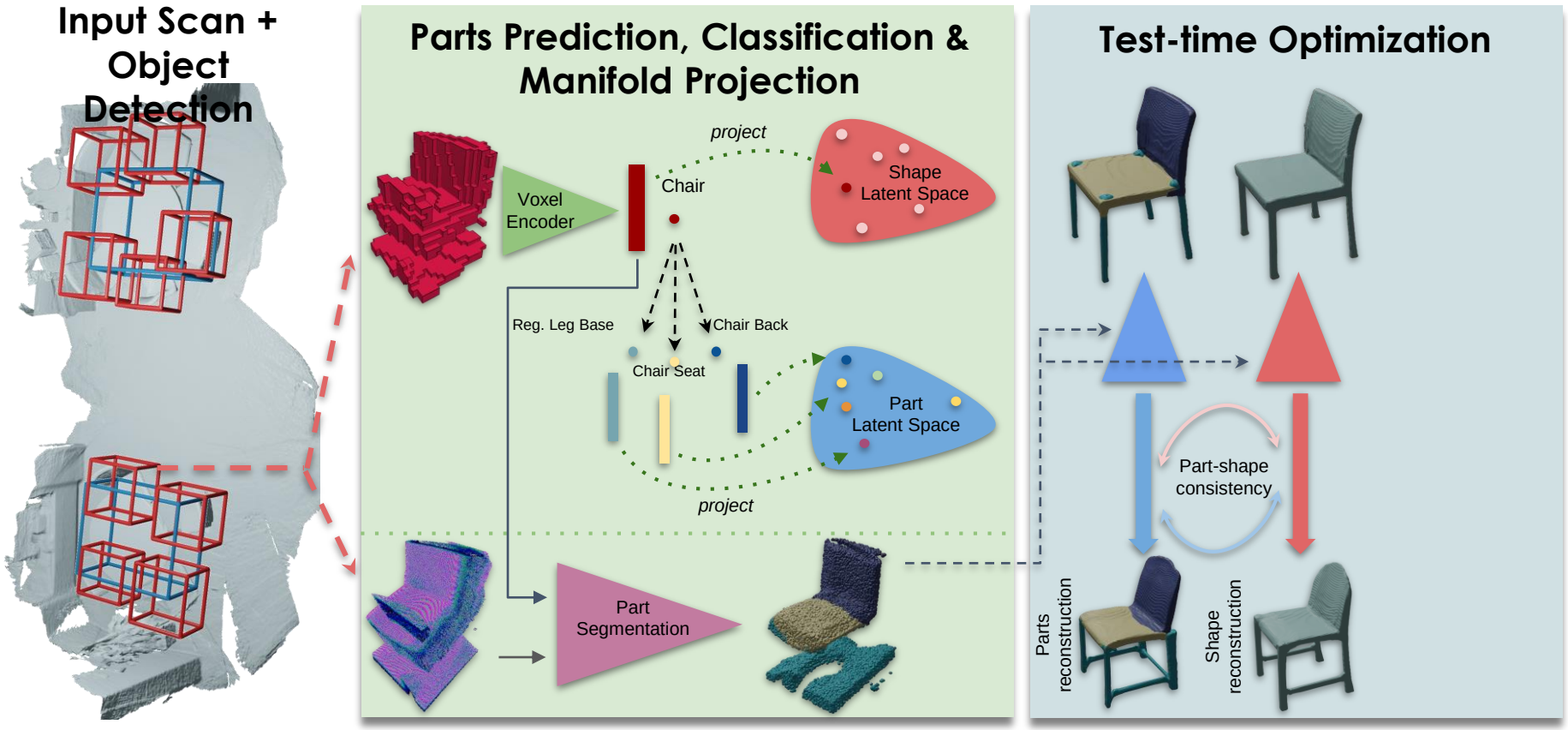
[Bokhovkin et al. '23]

Part Segmentation



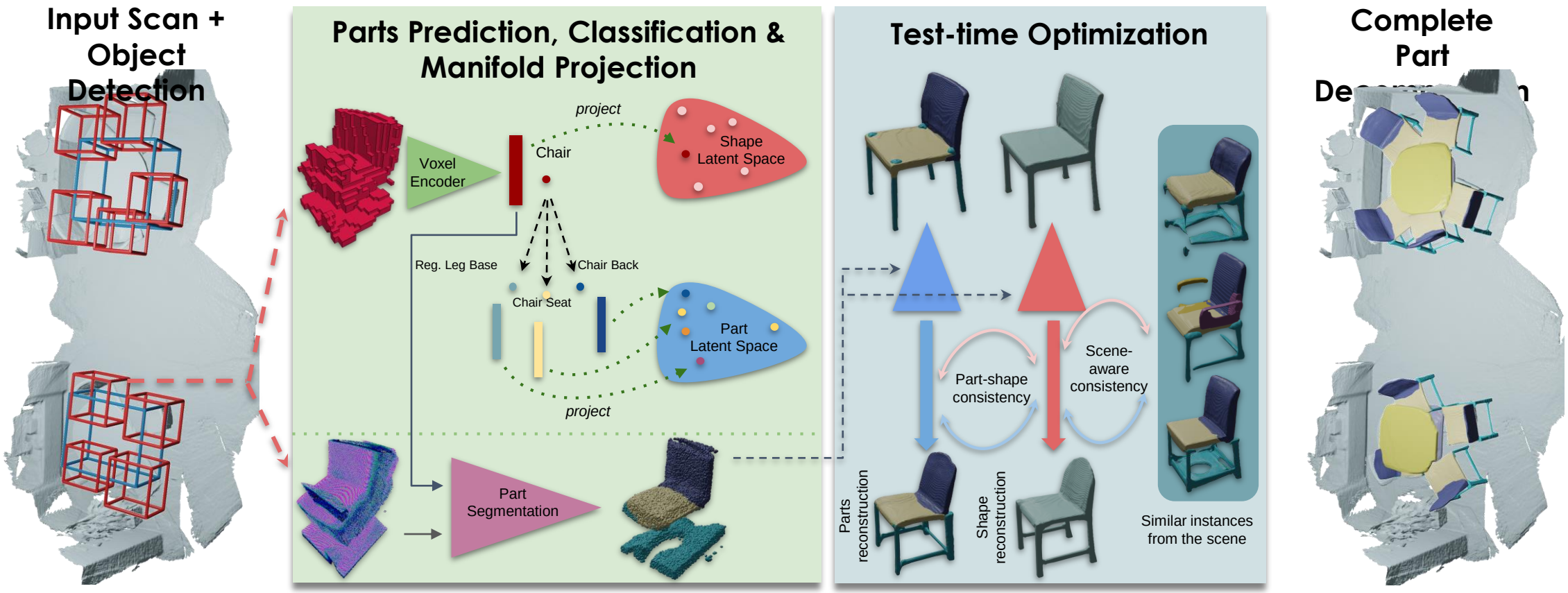
[Bokhovkin et al. '23]

Optimize to Fit Input Observations



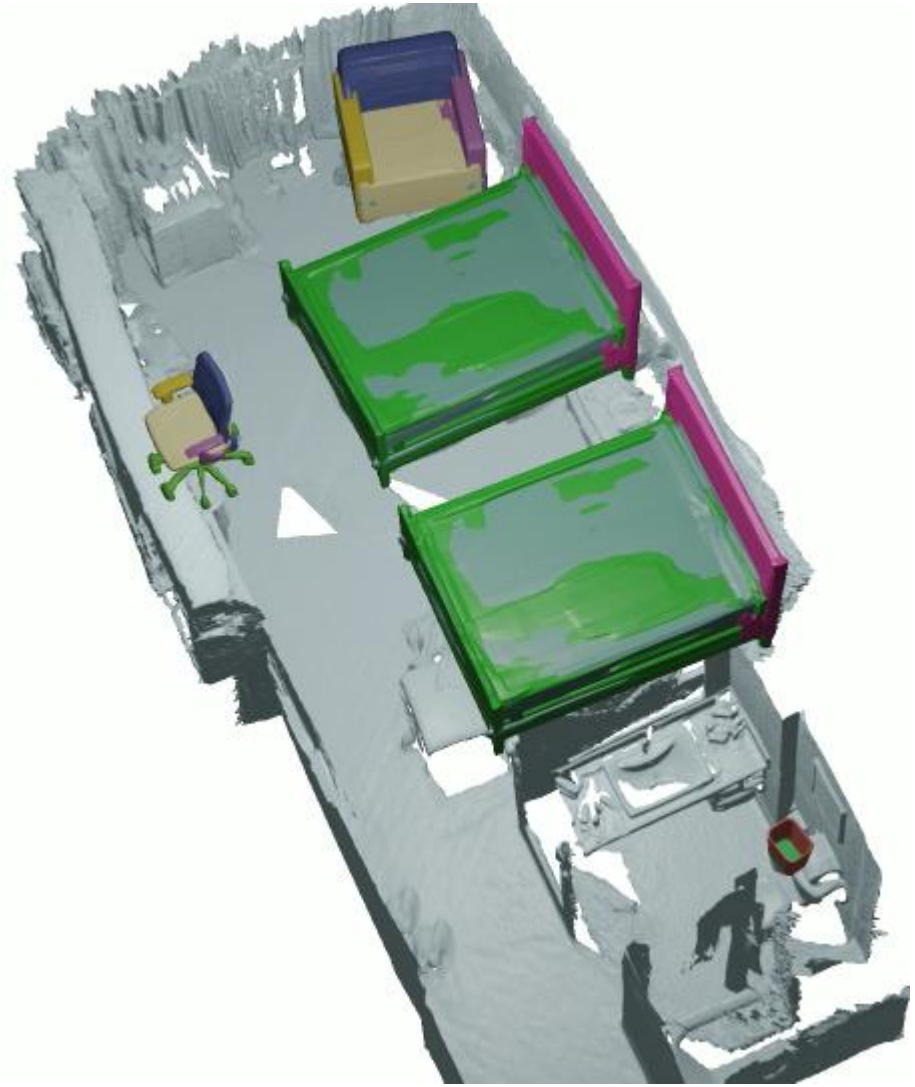
[Bokhovkin et al. '23]

Scene-Consistent Optimization



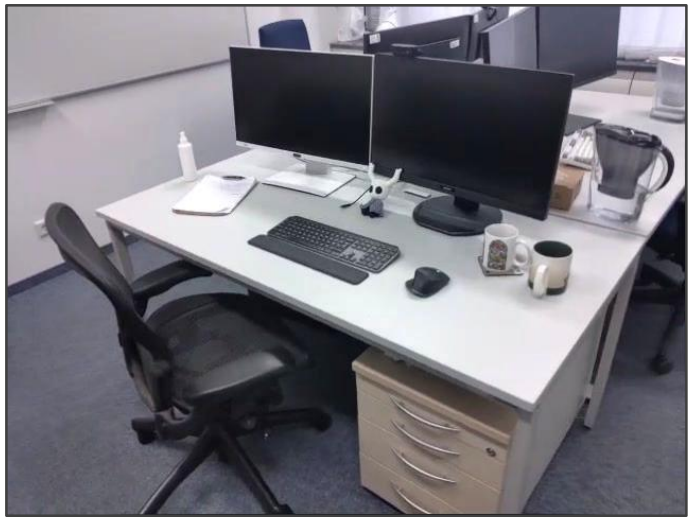
[Bokhovkin and Dai]: NPPs (arXiv'22)

Neural Part Prior Optimization



[Bokhovkin et al. ’23]

What about 2D Recognition?



RGB Image

Semantic Segmentation

Instance Segmentation

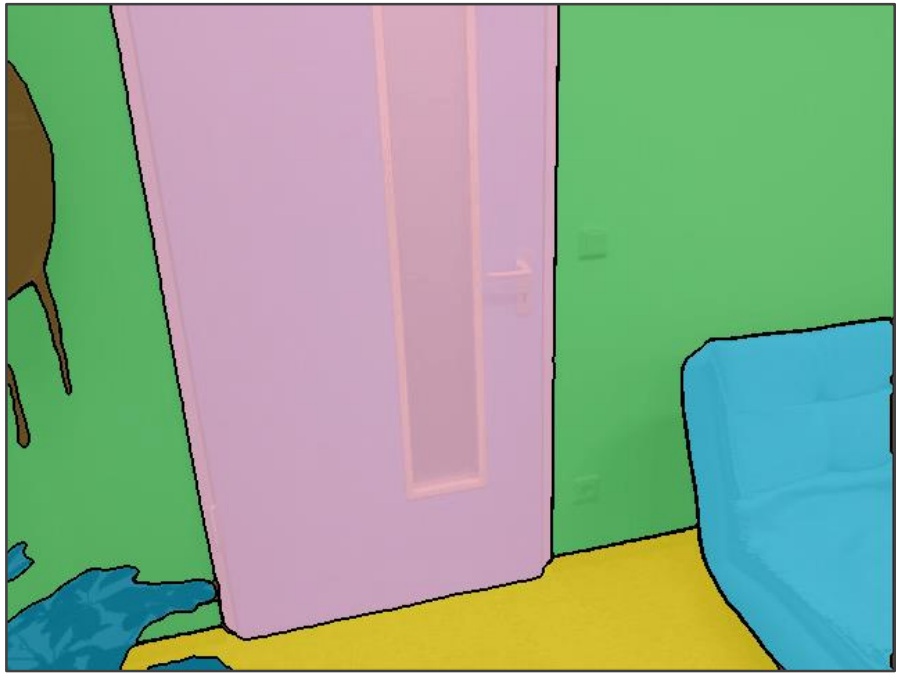
stuff: void, wall, table, floor, cabinet, other

things: chair, screen, cup, keyboard, bottle, mouse, sofa, bed

Mask2Former [Cheng et al. '22]

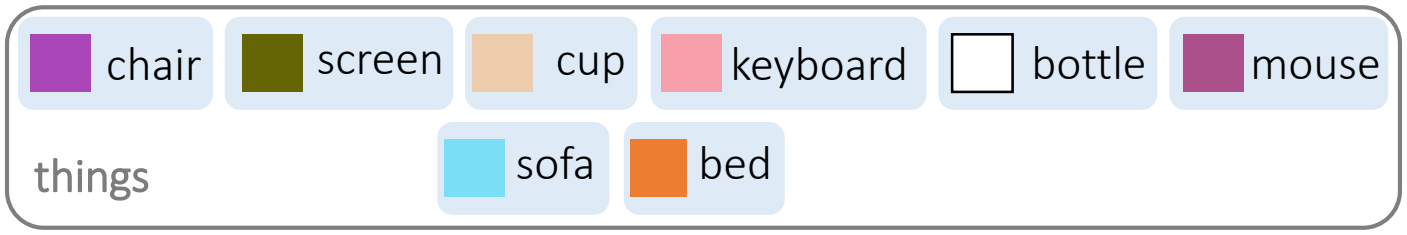
2D Recognition: Multi-View (In)Consistency

RGB Images from a Scene



Semantic Segmentation

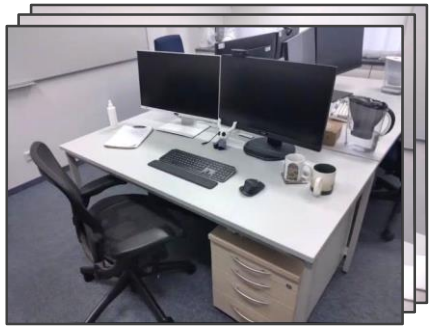
Instance Segmentation



Mask2Former [Cheng et al. '22]

Optimizing for View Consistency

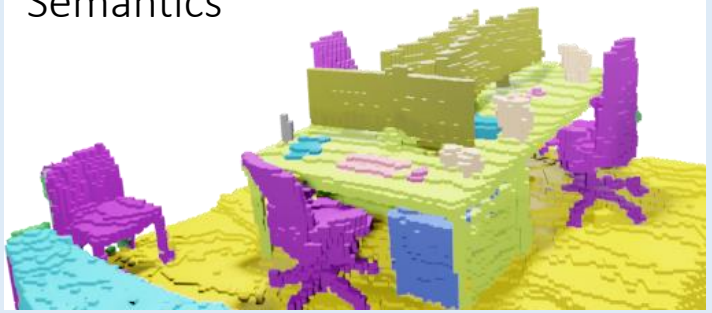
RGB Images from a Scene



Predicted 2D Panoptic Seg.

Density + Color

Semantics



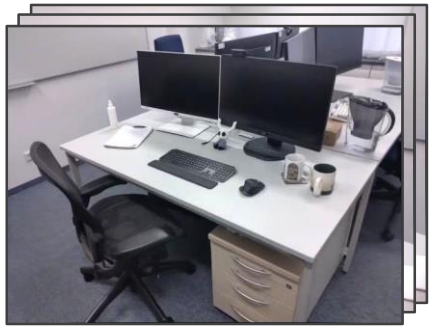
Instances

Volumetric Scene Representation

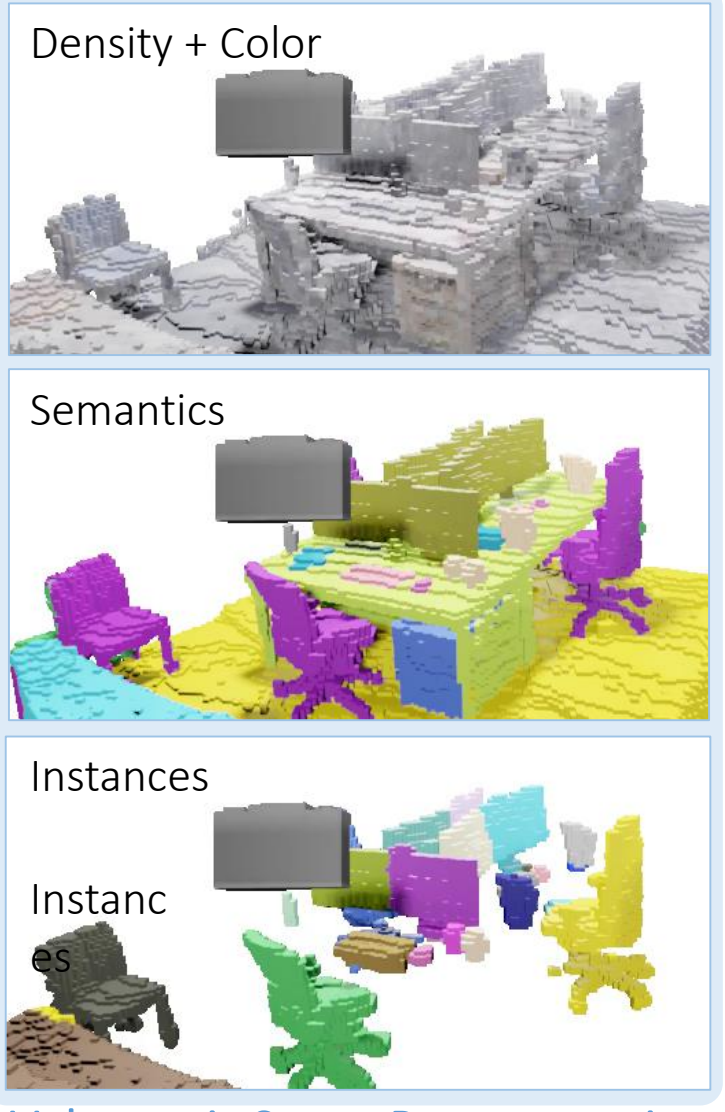
[Siddiqui et al. '23]

Optimizing for View Consistency

RGB Images from a Scene

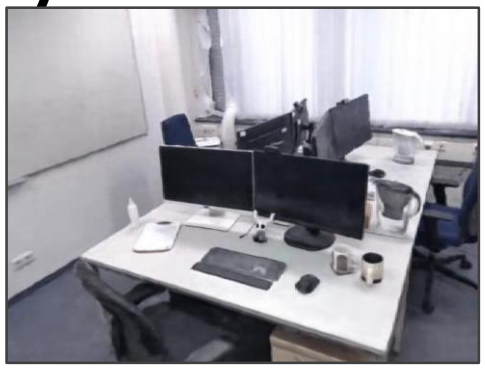


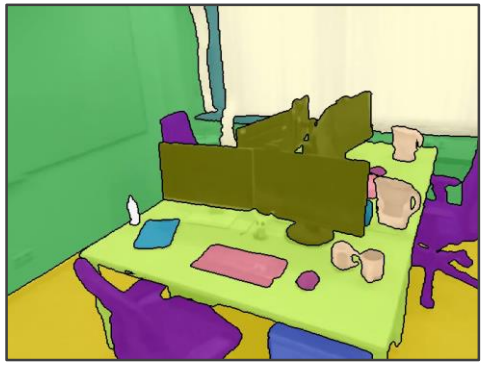
Predicted 2D Panoptic Seg.



Volumetric Scene Representation

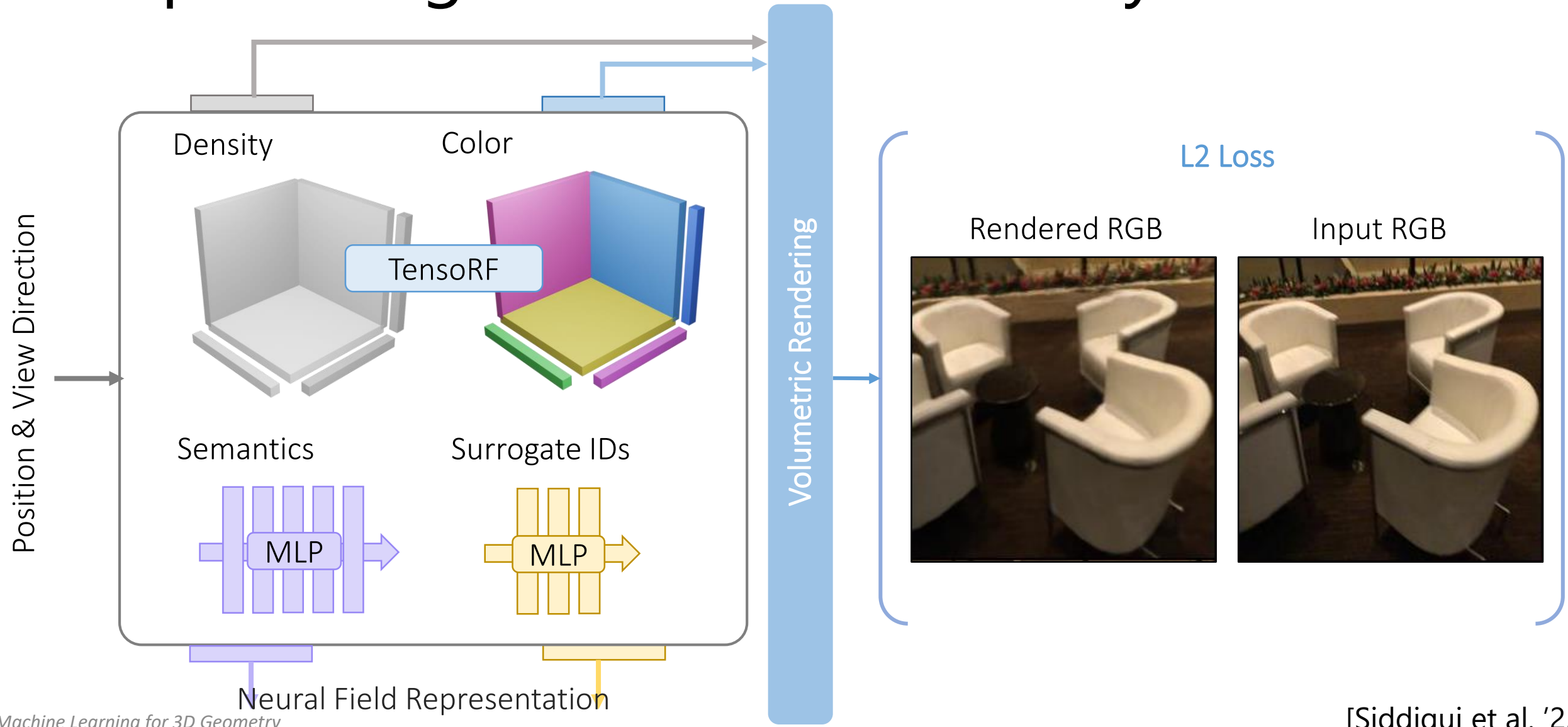
Volumetric Rendering





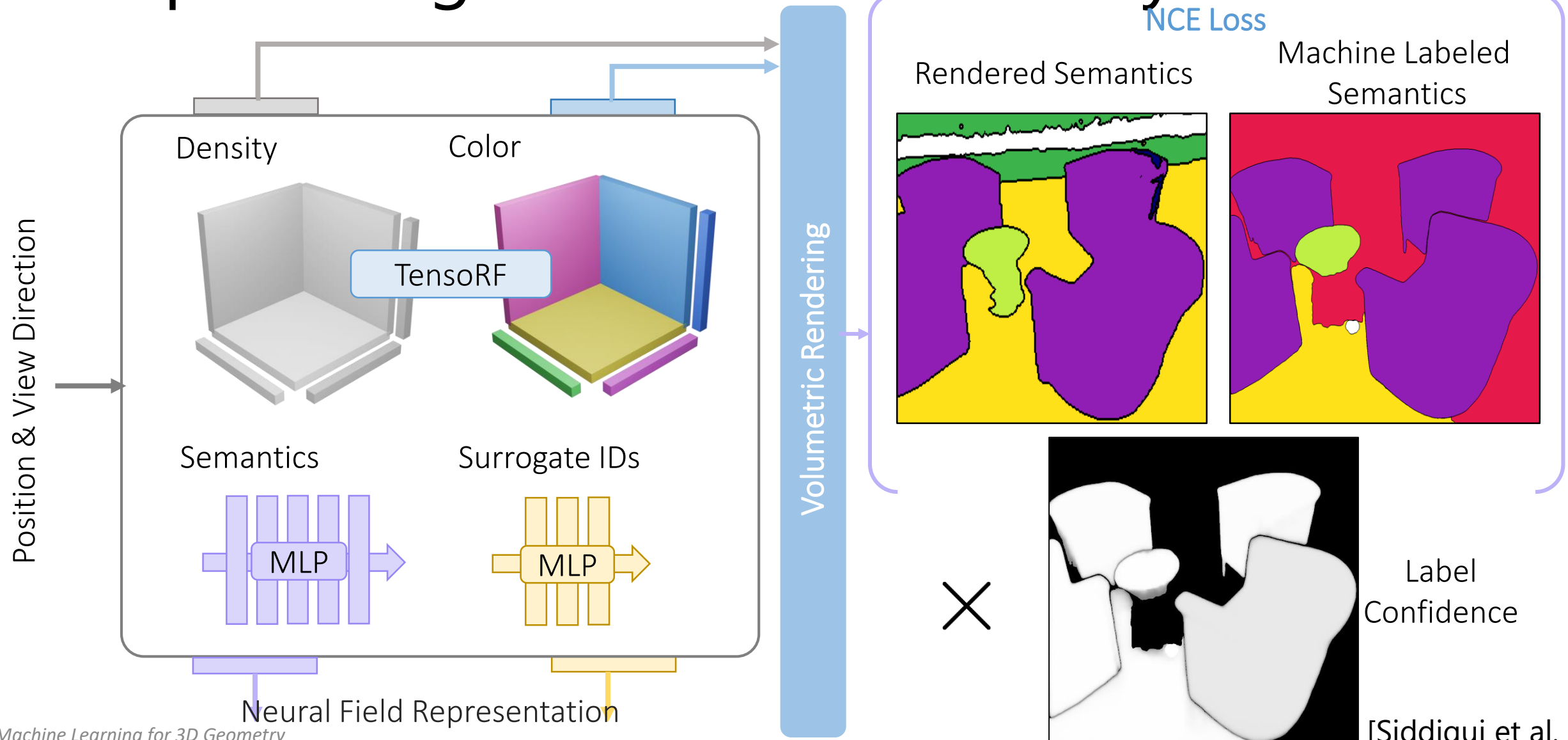
[Siddiqui et al. '23]

Optimizing for View Consistency



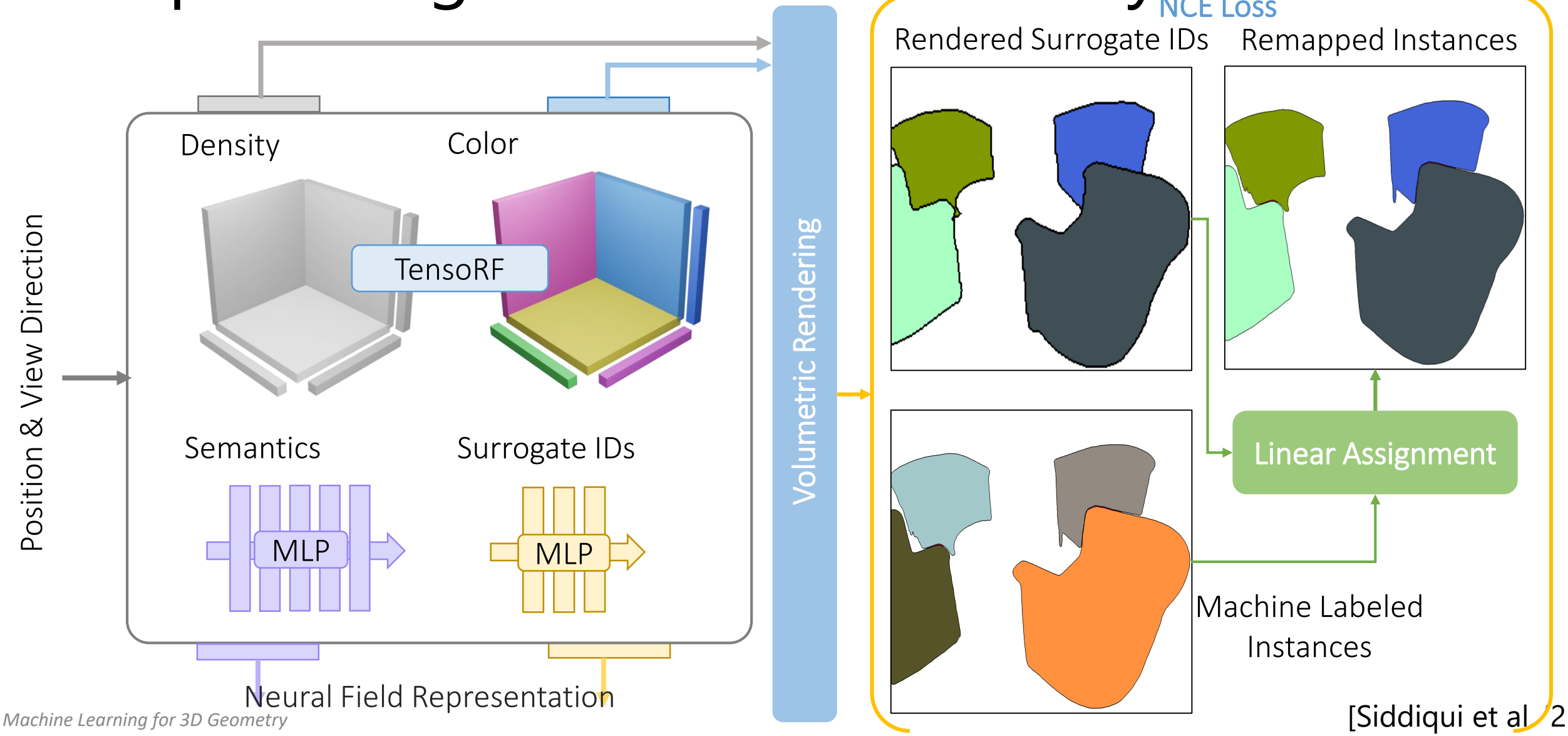
[Siddiqui et al. '23]

Optimizing for View Consistency



[Siddiqui et al. '23]

Optimizing for View Consistency



[Siddiqui et al '23]