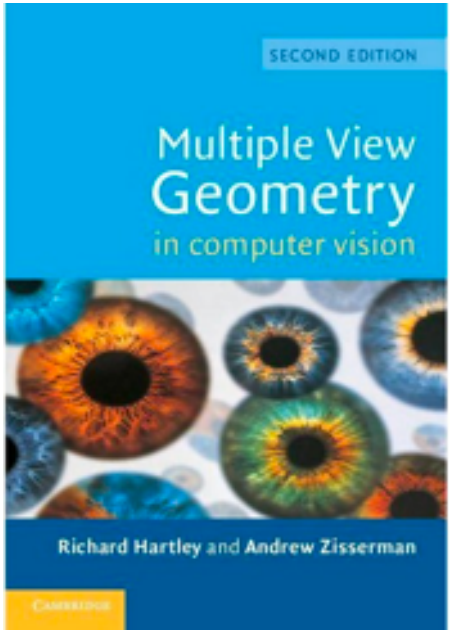
UC San Diego

3D Understanding Towards Object Manipulation

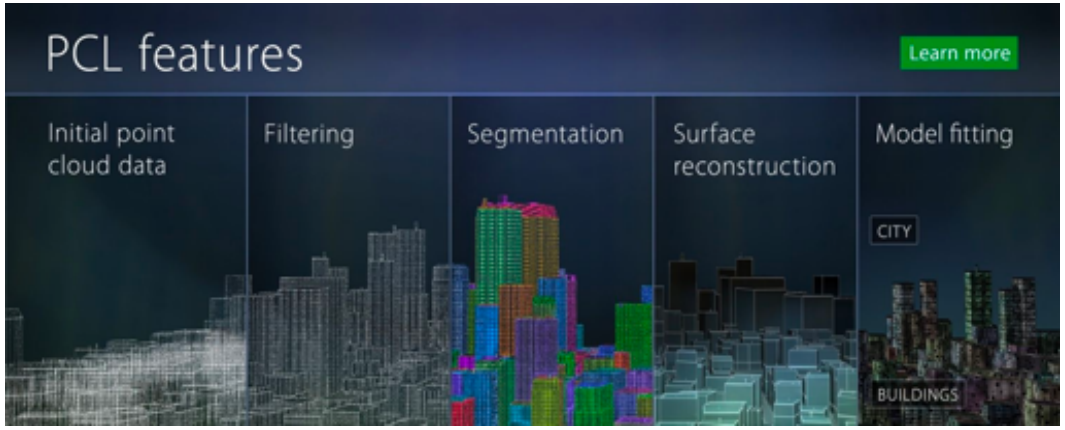
Some Thoughts and Progress

Hao Su

3D in CV/CG before DL Age



CGAL

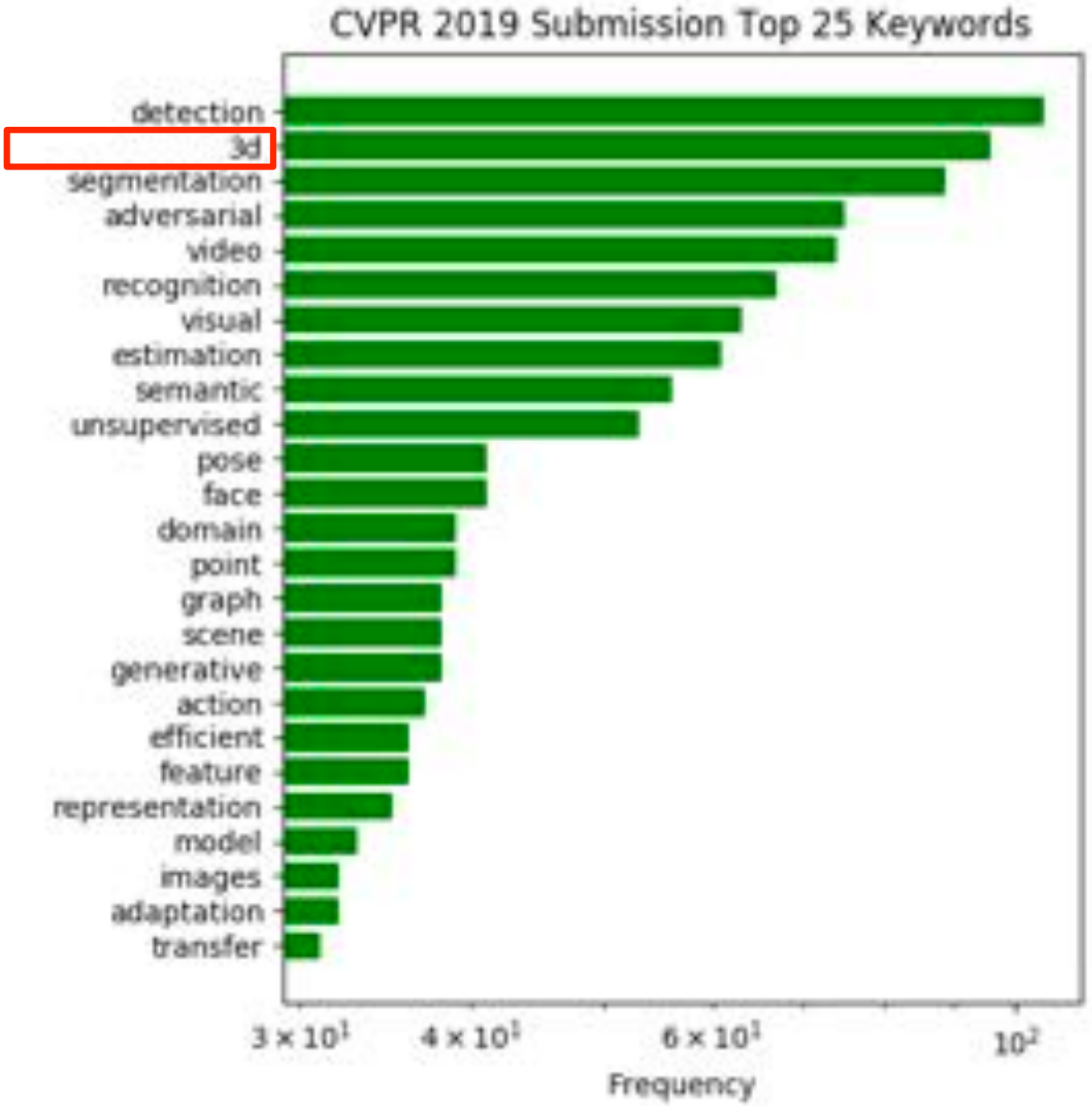


. . .

Recent Hype of 3D DL

Acquire Knowledge of 3D World by Learning

A priori knowledge of the 3D world


CVPR 2019 Submission Top 25 Keywords
detection
3d
segmentation
adversarial
video
recognition
visual
estimation
semantic
unsupervised
pose
face
domain
point
graph
scene
generative
action
efficient
feature
representation
model
images
adaptation
transfer
3 × 10^1
4 × 10^1
6 × 10^1
10^2
Frequency

Core Algorithms Invented

- Classification

Volumetric CNN, OctNet, O-CNN, SparseConvNet, PointNet, PointNet++, RS CNN, DGCNN, Point ConvNet, KPConv, Monte Carlo Point Convolution, PConv, Multi-View CNN, Spectral CNN, Synchronized Spectral CNN, Spherical CNN, …

- Detection/Segmentation

Sliding shape, 3D-SIS, Frustum PointNet, Point R-CNN, VoteNet, GSPN, SGPN, JSIS3D, ContFuse, PointPillar, Second, …

- Synthesize/Reconstruction

3D Autoencoder, PointSetGenNet, OctGenNet, AtlasNet, DeepSDF, Occupancy Networks, Implicit Fields, MarrNet, StructNet, 3DGAN, PointSetGAN, MVS, SurfaceNet, RMVS, PMVS, BA-Net, ……

Datasets Built

	Object	Part	Indoor Scene	Outdoor Scene
Synthetic	ShapeNet, ModelNet	ShapeNetPart, PartNet, Shape2Motion	SceneNet	vKITTI, Cala
Real	3DScan		ScanNet	KITTI, Semantic KITT, Waymo Open Dataset

My Tutorials on 3D Deep Learning

- 90min Summary (2020 March version):

https://youtu.be/vfL6uJYFrp4

- Can be found from my homepage: http://ai.ucsd.edu/~haosu

Timely to Think About Three Questions

- Many core algorithms developed.

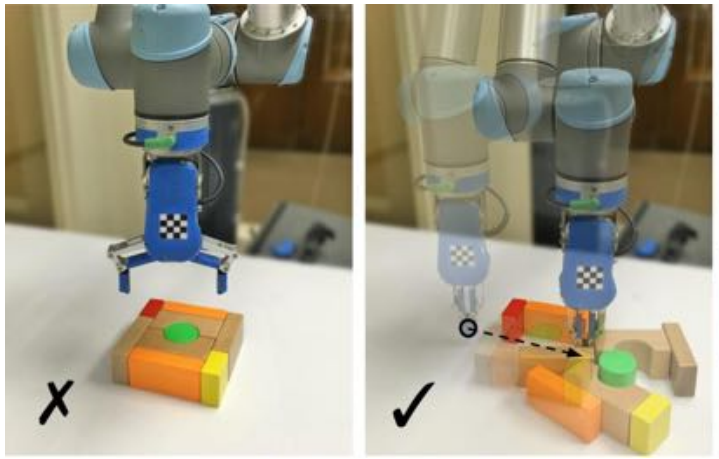
- But:

  1. How large is the performance gap for current algorithms to support **downstream applications**?
  2. What kind of new 3D deep learning problems have to be addressed?
  3. What efforts may be needed to build new benchmarks?

Exploratory Robots

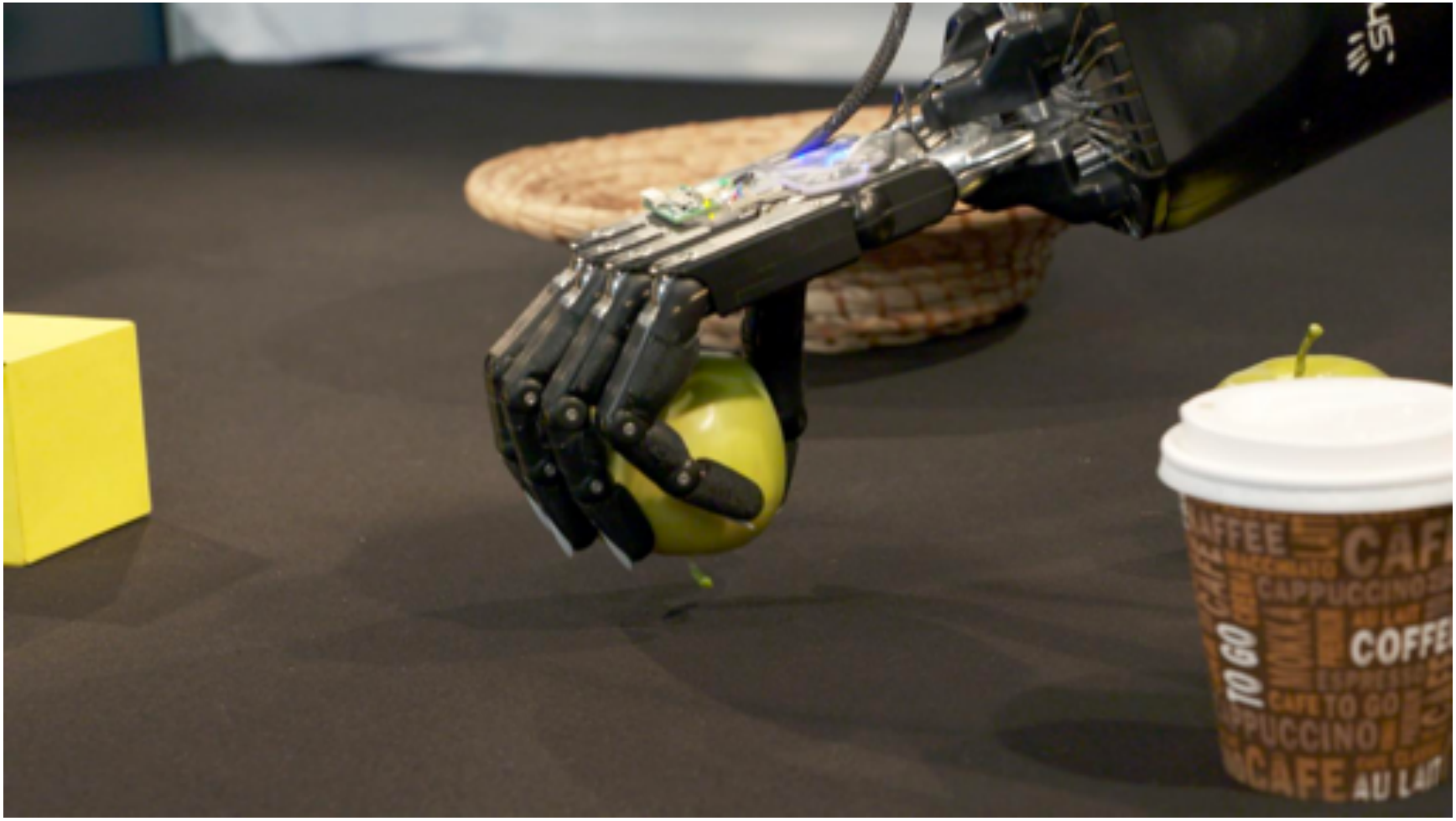
- Human-beings learn the unknowns via exploring the physical world

- An exploratory robot learn the environment dynamics via collecting interaction experience



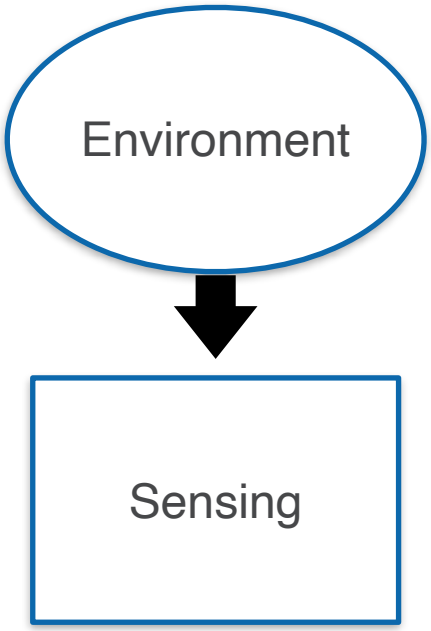
1. *Source: otteroo.com*
2. *Source: Andy Zeng*

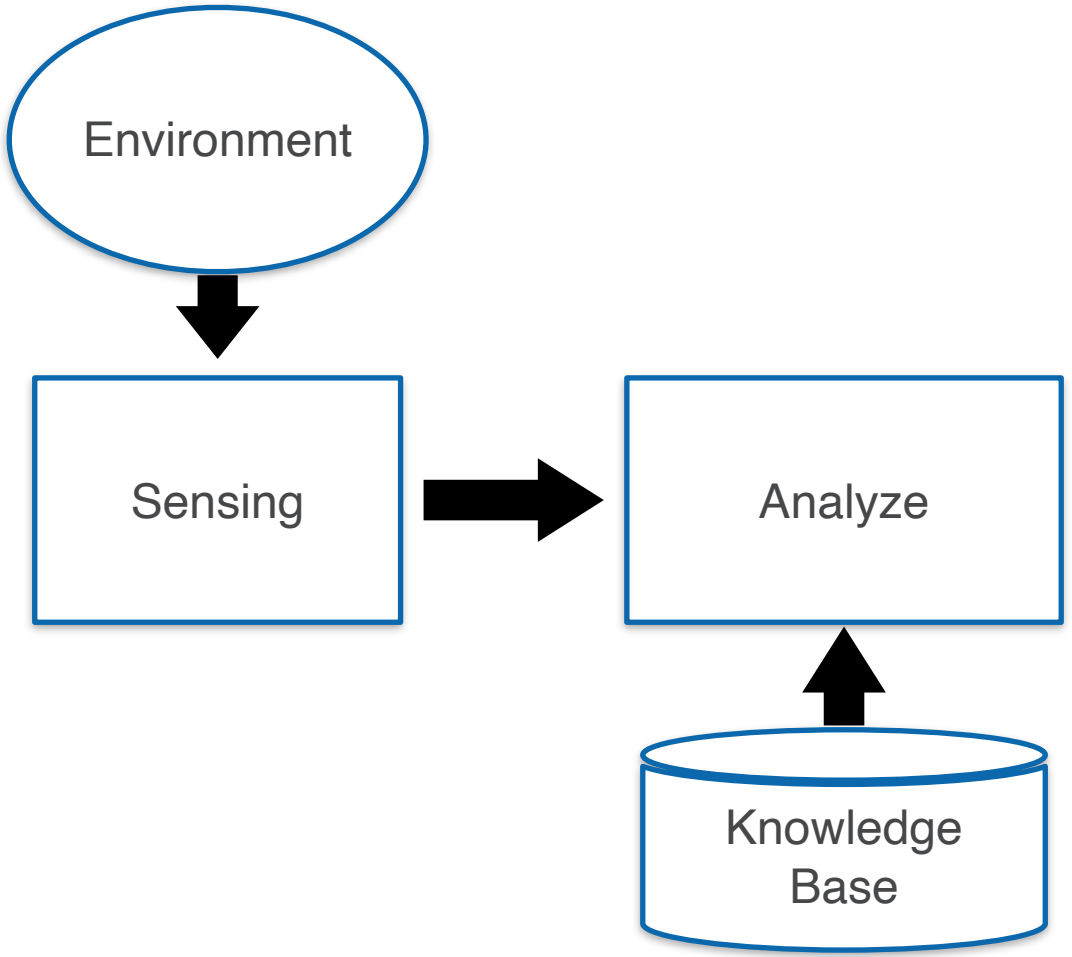
Object Manipulation



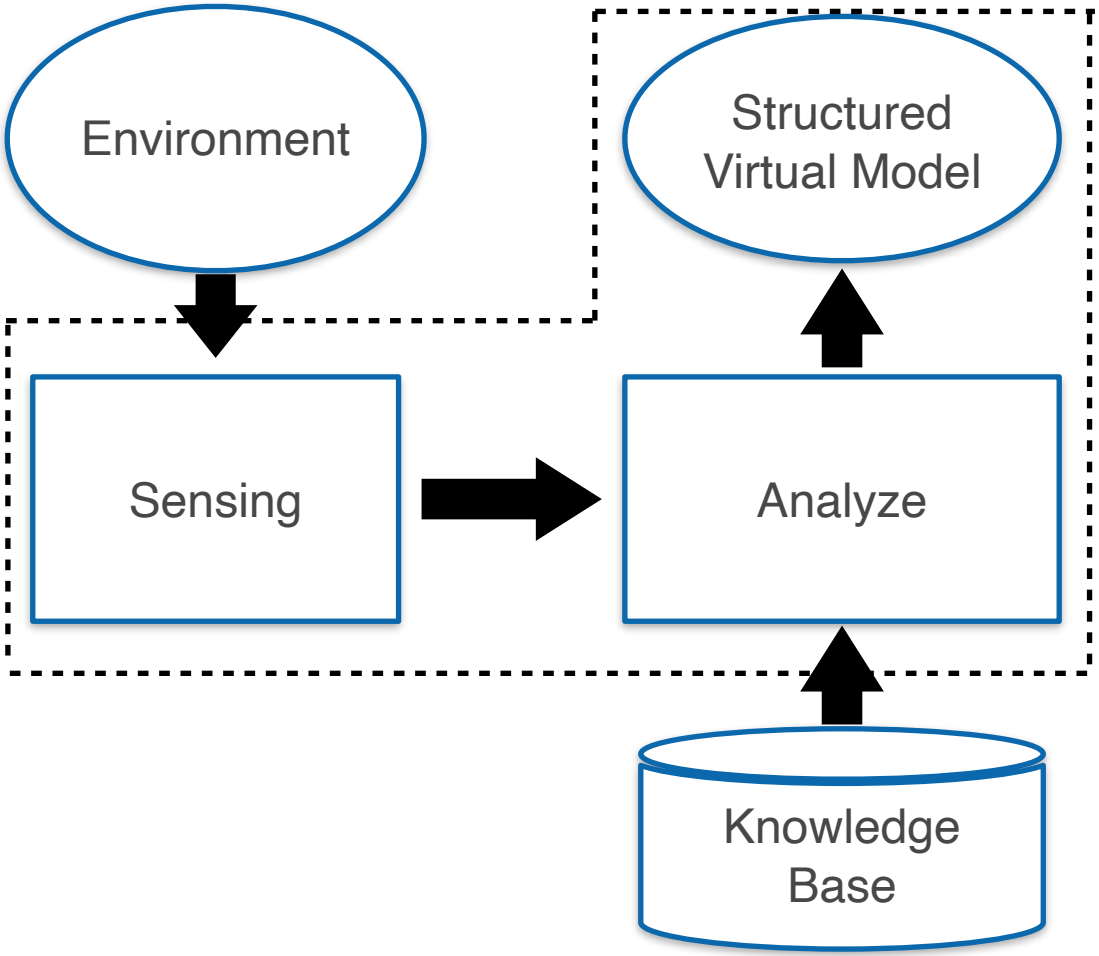
Credit: Bielefeld University
https://phys.org/news/2017-06-grasp.html


Environment

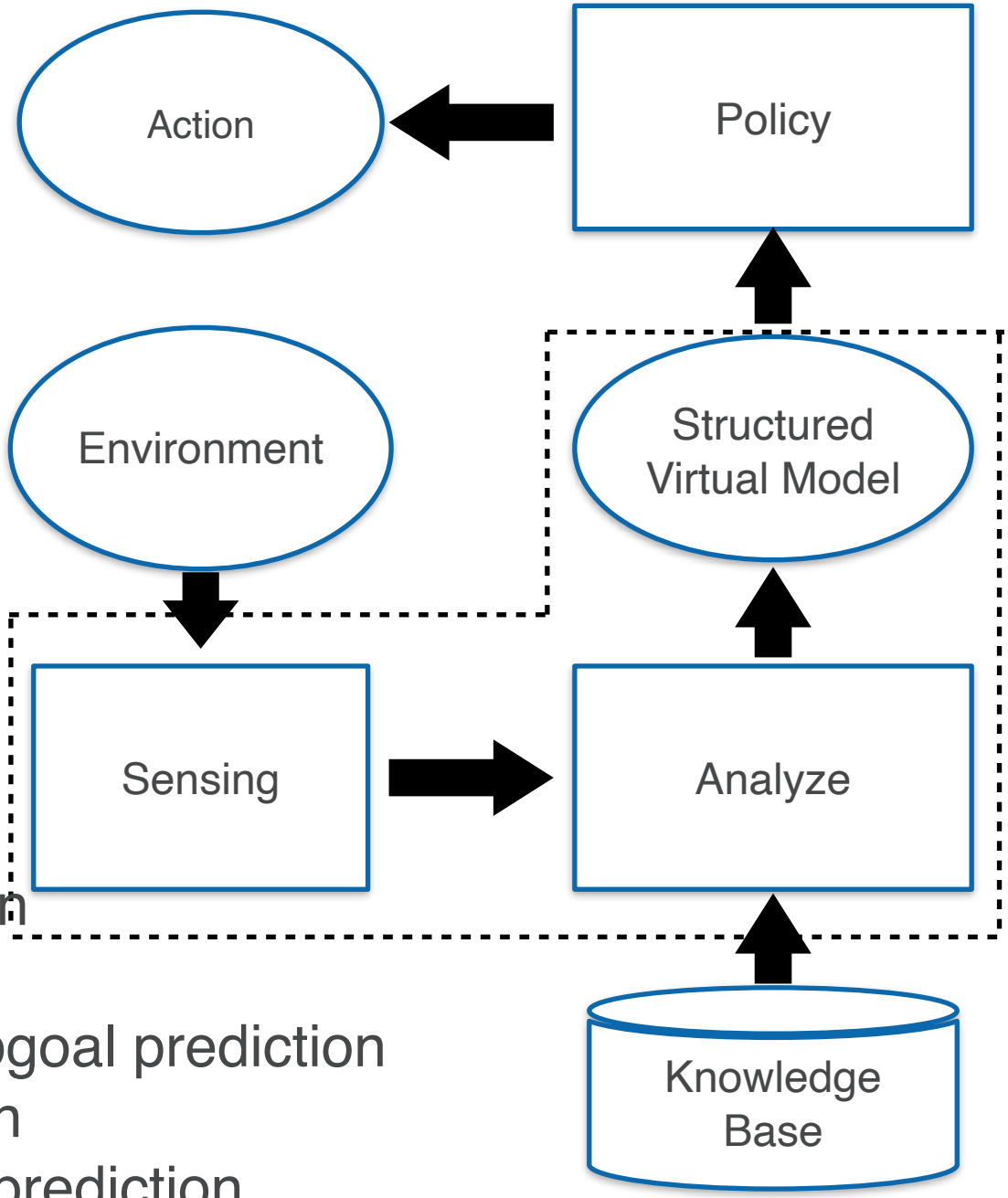

Environment
Sensing



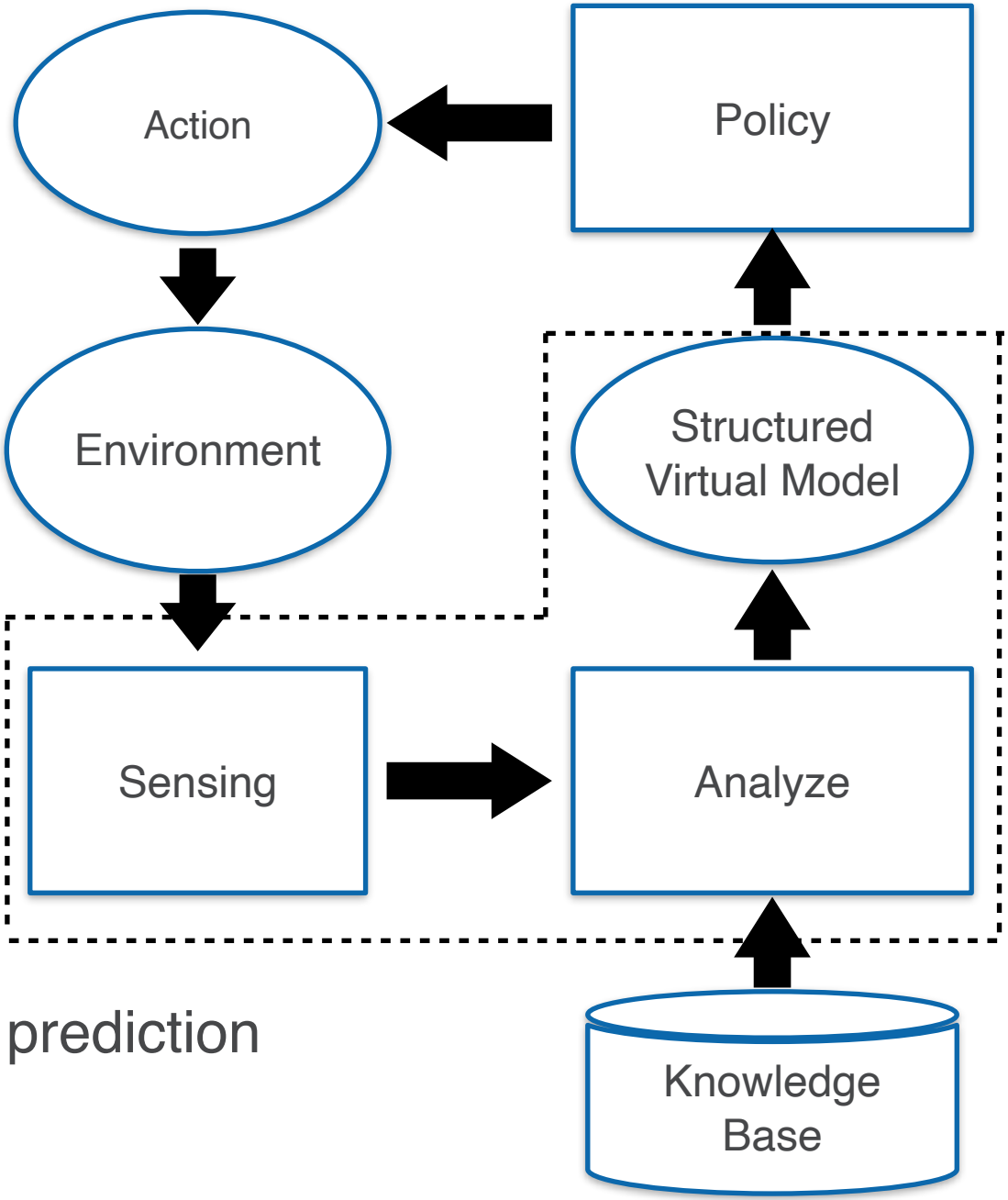
- Reconstruction
- Detection
- Segmentation



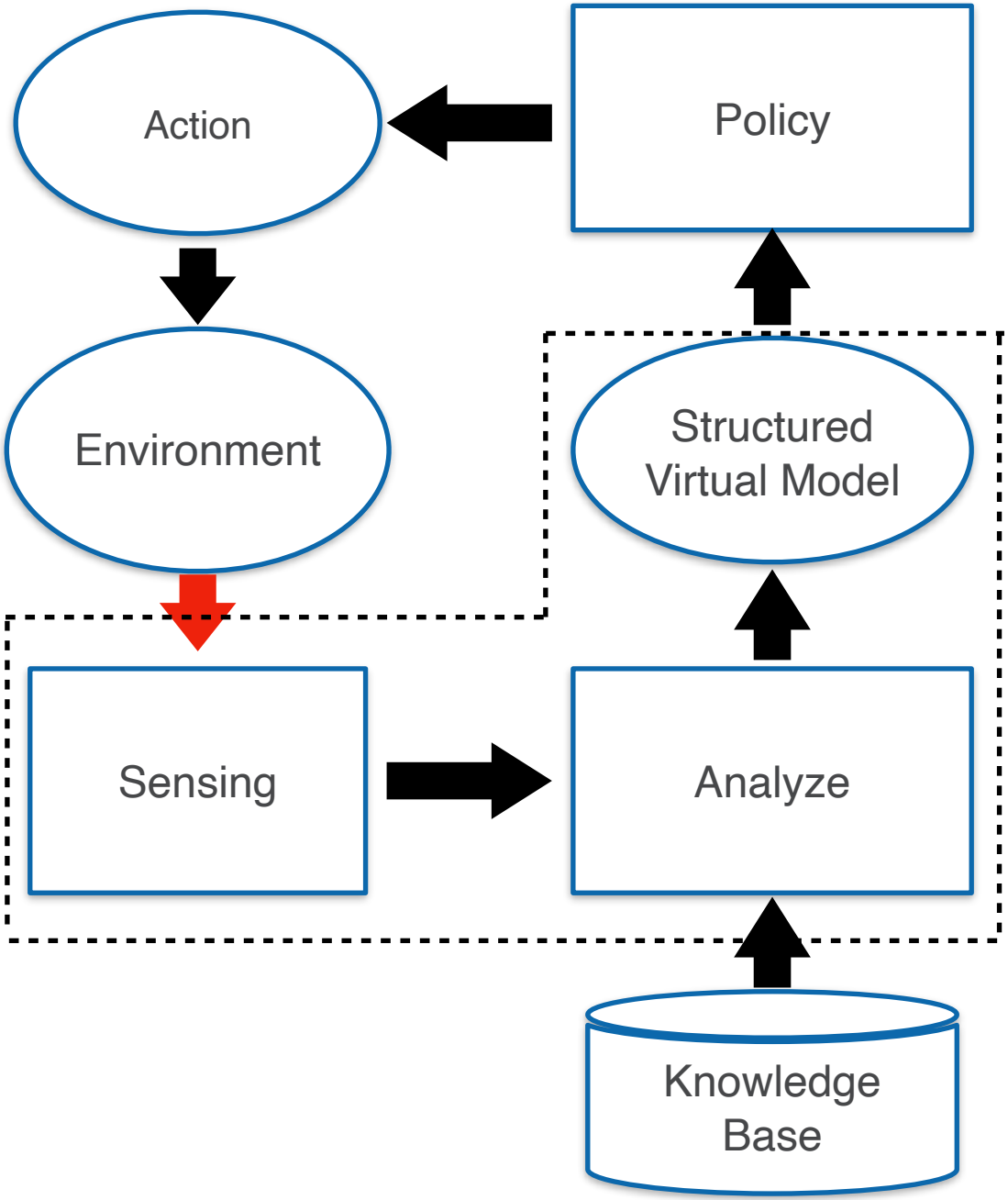
- Reconstruction
- Detection
- Segmentation

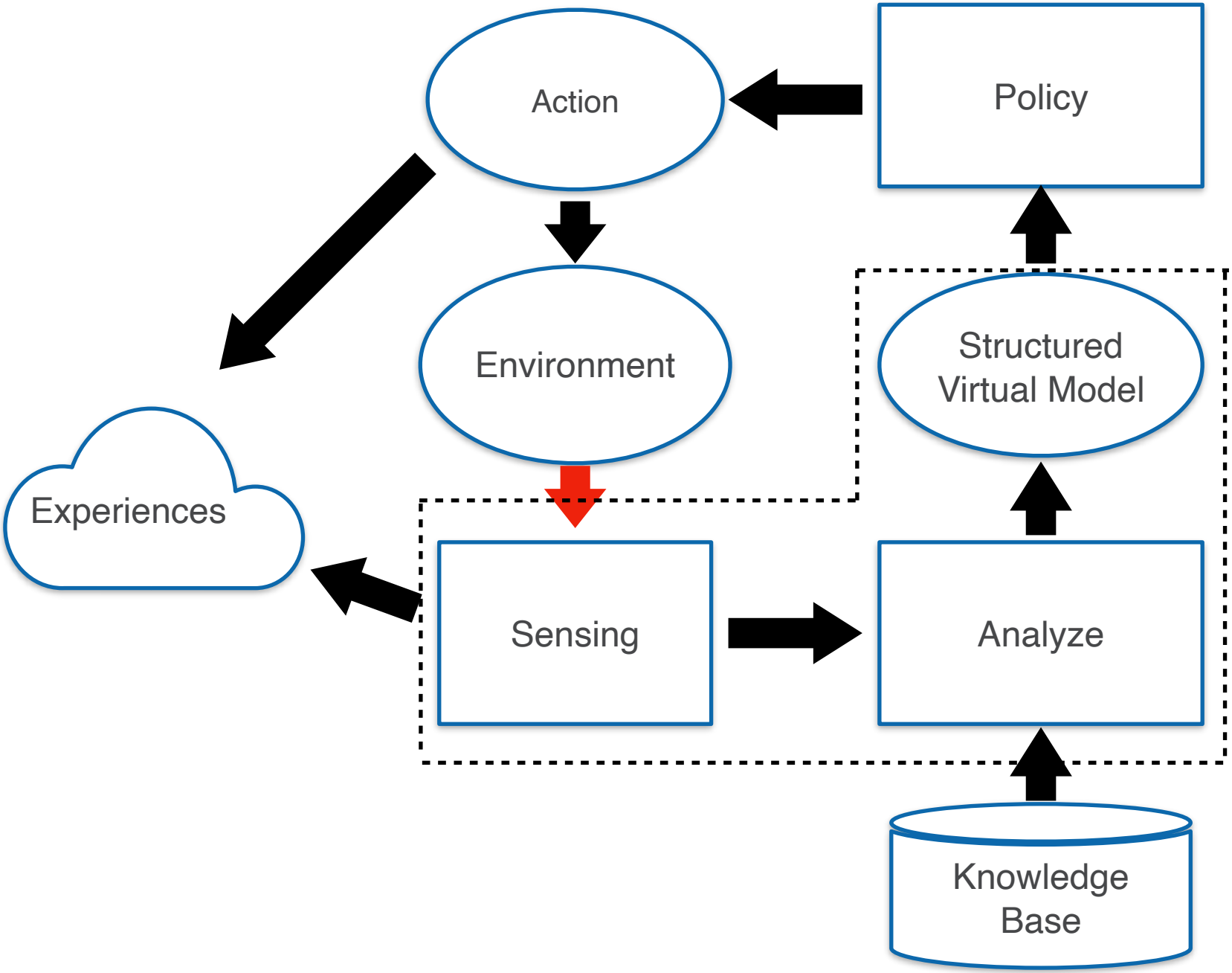


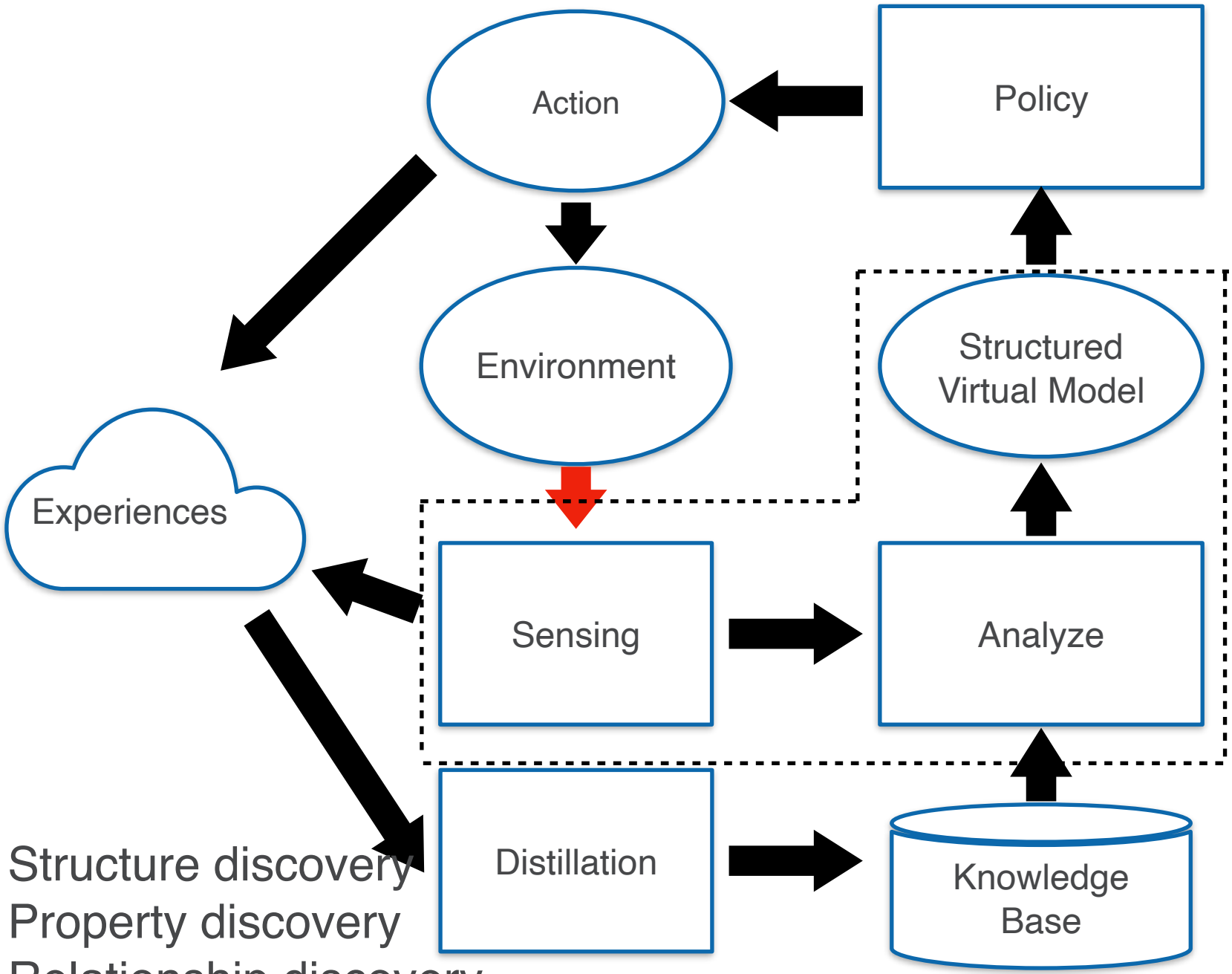
- Task representation
- Grasp proposal
- Plan synthesis/subgoal prediction
- Collision estimation
- Inverse dynamics prediction



- Forward dynamics prediction


Action
Policy
Environment
Structured
Virtual Model
Sensing
Analyze
Knowledge
Base


Action
Policy
Environment
Structured
Virtual Model
Experiences
Sensing
Analyze
Knowledge
Base



- Structure discovery
- Property discovery
- Relationship discovery

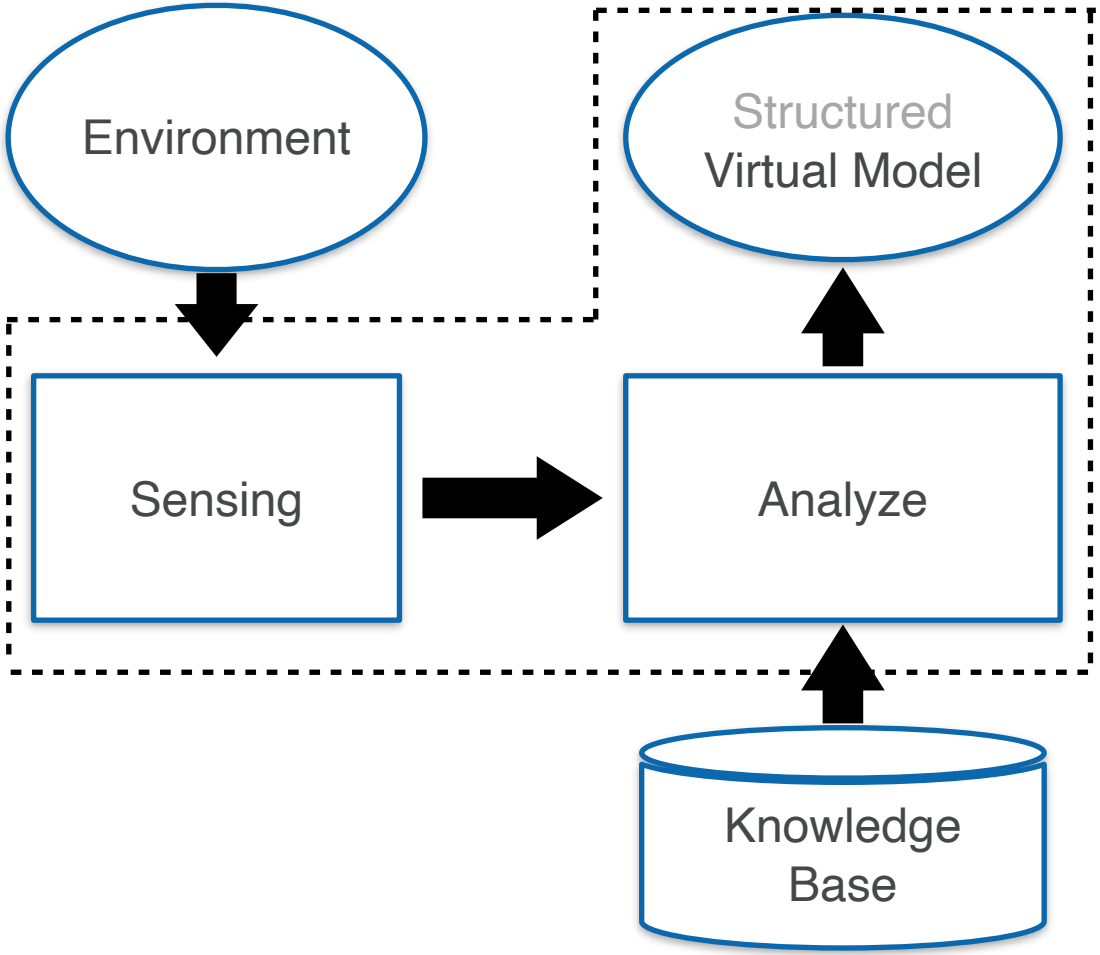
UC San Diego

Sampled Research Work (I)

Learning-based 3D Reconstruction

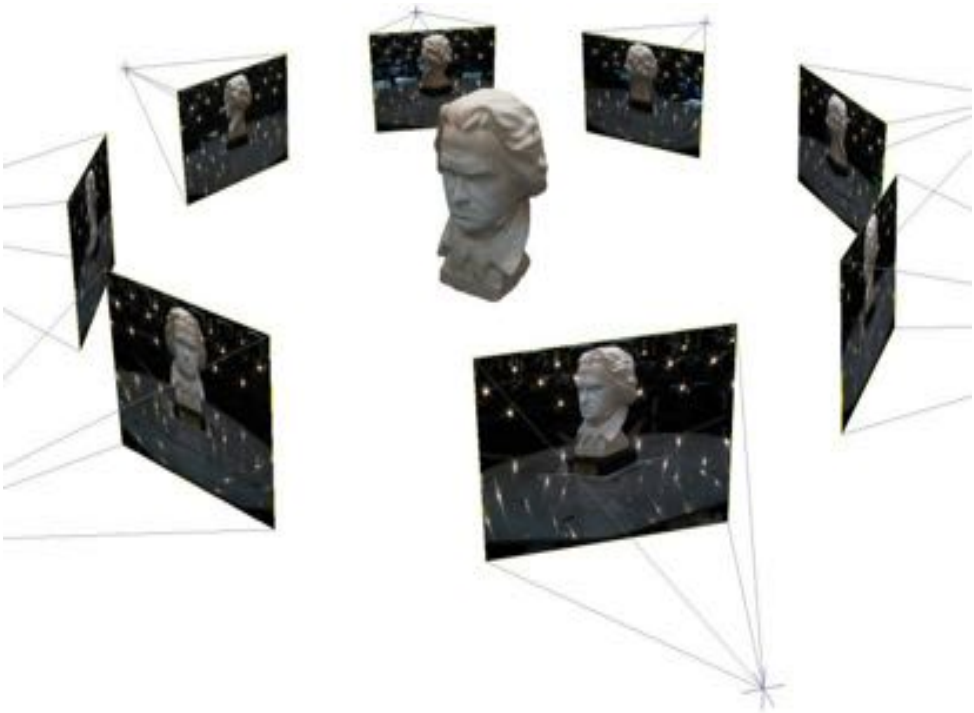
Deep Stereo using Adaptive Thin Volume Representation with Uncertainty Awareness, Shuo*, Xu*, et al. CVPR 2020 (oral)

Normal Assisted Stereo Depth Estimation, Kusupati, et al. CVPR 2020


Environment
Structured
Virtual Model
Sensing
Analyze
Knowledge
Base

Multi-View Stereo (MVS)

Reconstruct the dense 3D shape from a set of **images** and **camera parameters**



1. Goldlucke et al. “A Super-resolution Framework for High-Accuracy Multiview Reconstruction”

Requirements of MVS

Applications	Range	Accuracy	Time Efficiency	Computation Efficiency
Remote Sensing	★★★★★	★★	★	★★
Autonomous Driving	★★★★	★★	★★★★	★★★★
AR/VR	★★	★★★	★★★★★	★★★★★
Robot Manipulation	★	★★★★	★★★	★★★★
Inverse Engineering	★	★★★★★	★★	★★

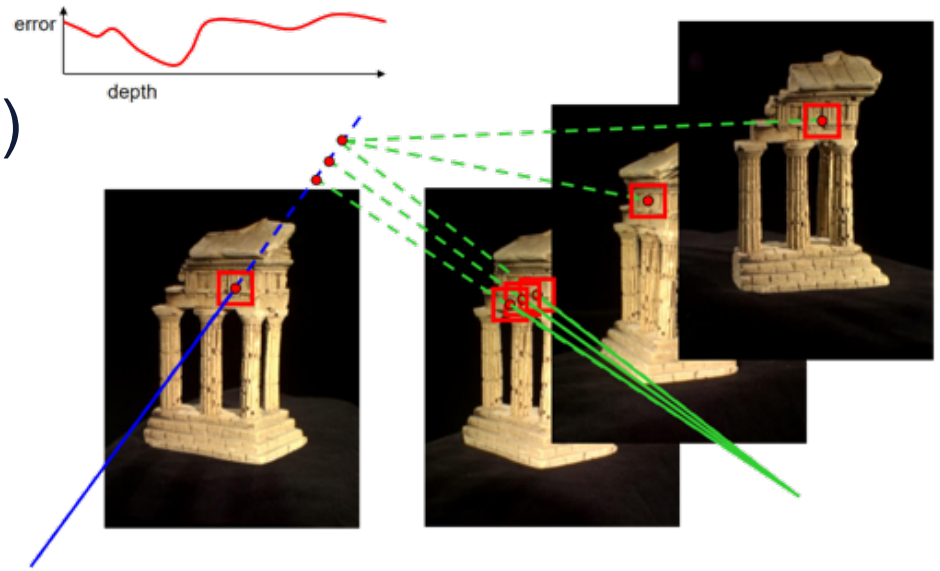
Reconstruction from Photo-Consistency

NCC (Normalized Cross Correlation)

$$\frac{\sum_{x,y}(W_1(x,y)-\overline{W_1})(W_2(x,y)-\overline{W_2})}{\sigma_{W_1}\sigma_{W_2}}$$

SSD (Sum Squared Distance)

$$\sum_{x,y}|W_1(x,y)-W_2(x,y)|^2$$

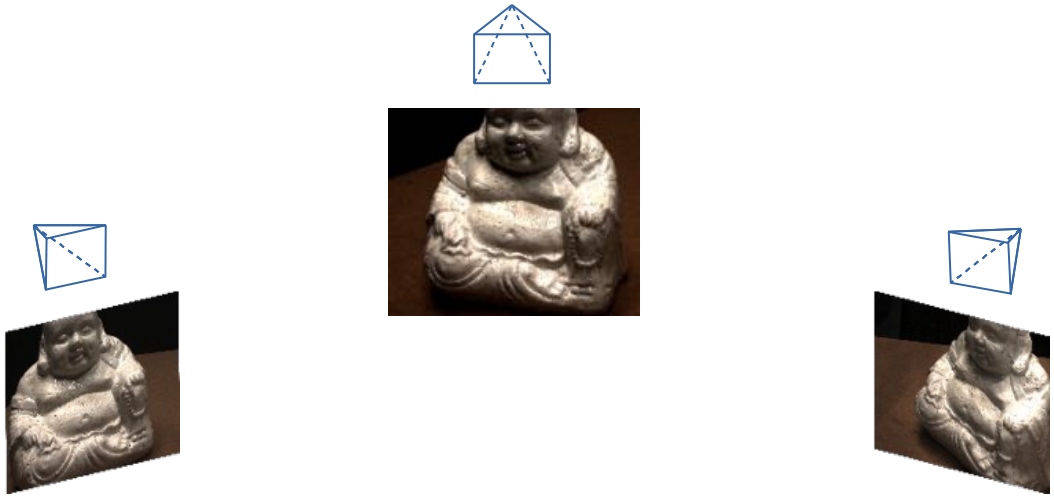


- Requires texture
- Sensitive to Non-lambertian area

Image source: UW CSE455

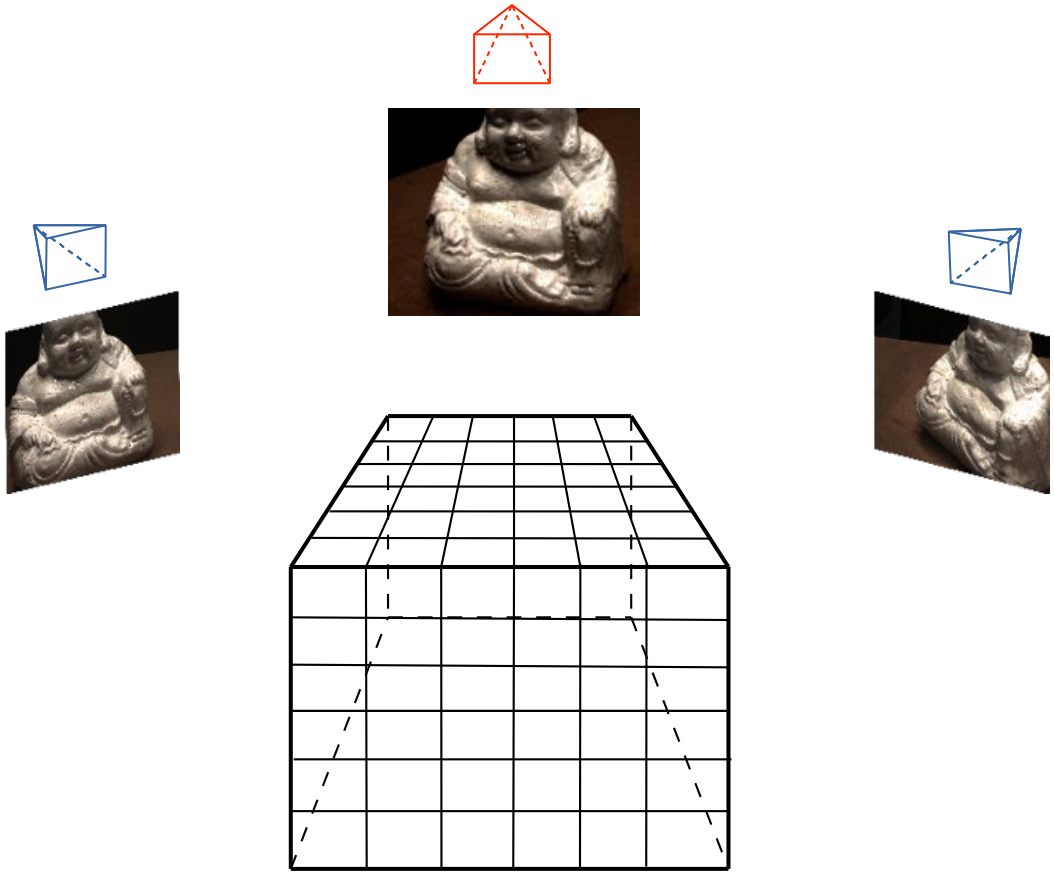
Cost-Volume-based MVS

Multi-view images and camera parameters

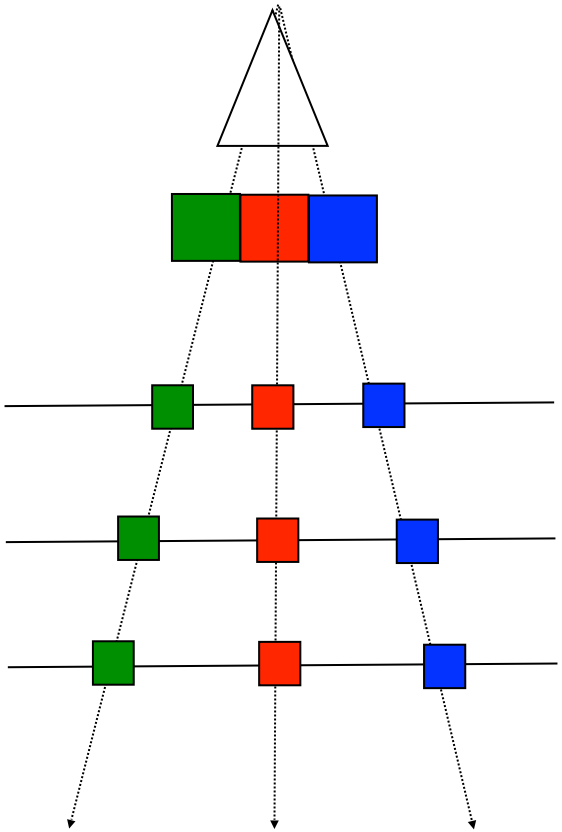


Cost-Volume-based MVS

Build 3D cost volume in reference view frustum



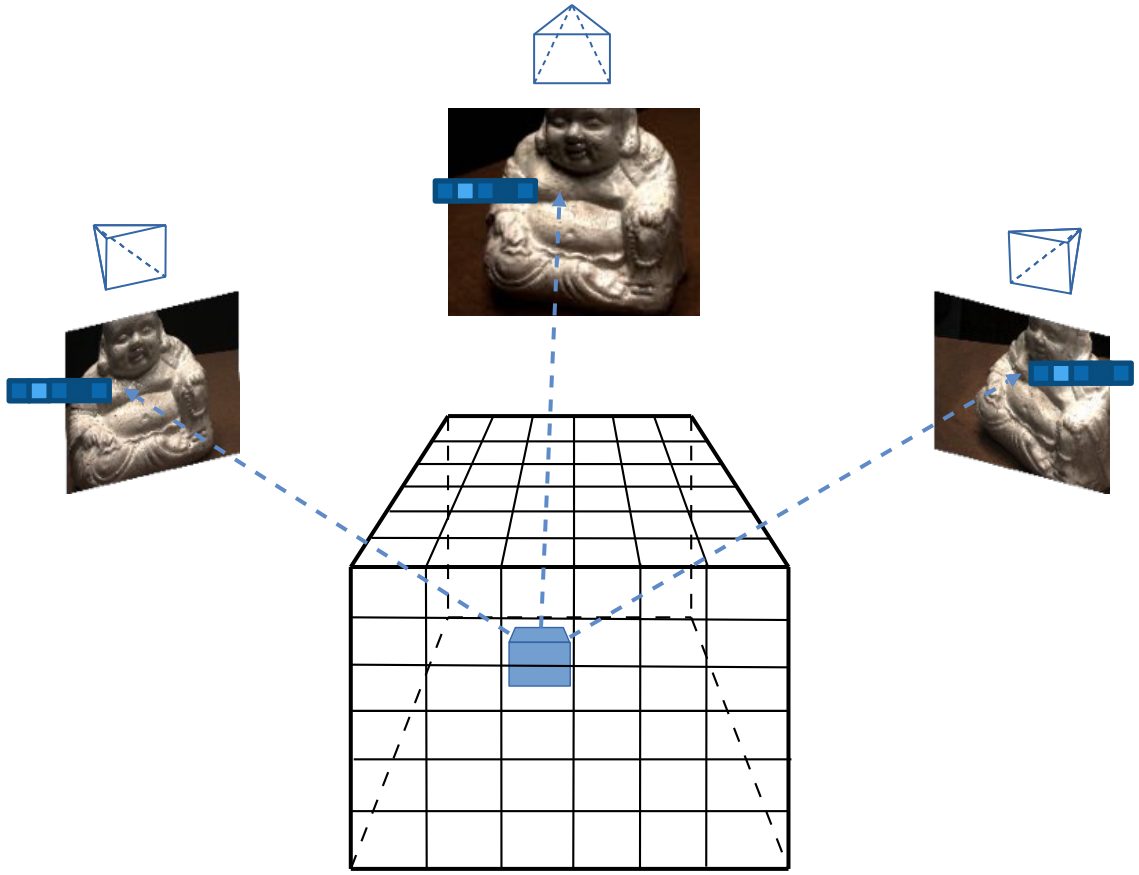
Topdown View of Cost Volume



Cost-Volume-based MVS

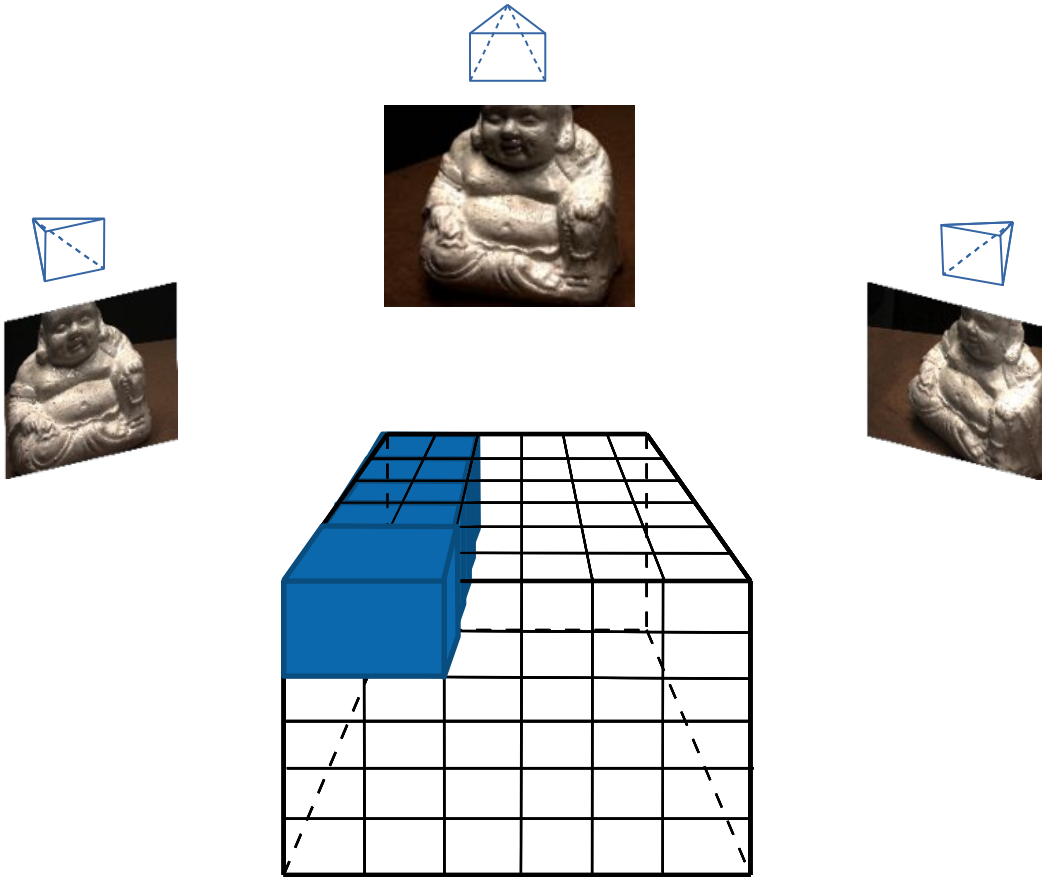
Fetch images features for each voxel

- Voxel in ground truth surface shows feature consistency



Cost-Volume-based MVS

Dense 3D CNNs

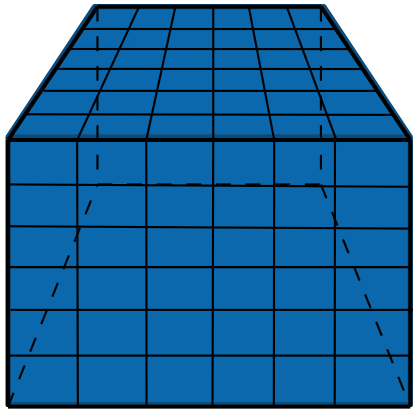


Cost-Volume-based MVS

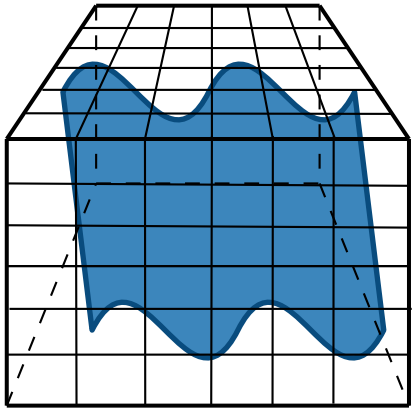
Are all these 3D CNNs necessary?

Cost-Volume-based MVS

- Convolution operations far from ground truth surface is wasting



Cost-volume



Target surface

High-level Idea

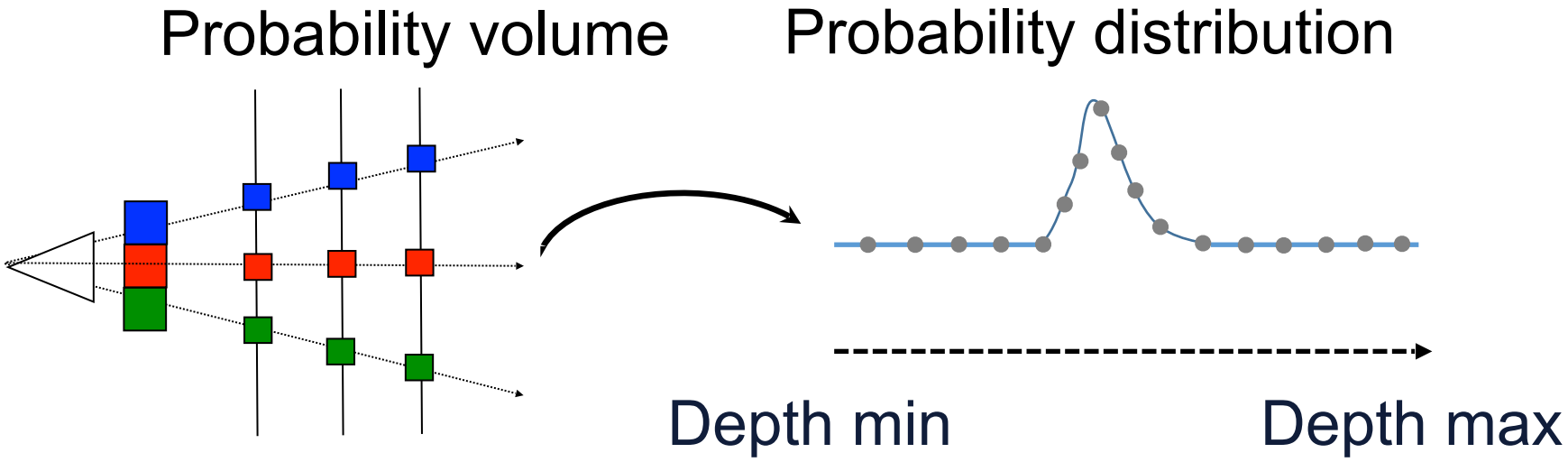
Previous:

Partition the space uniformly

This work:

Coarse-to-fine solution

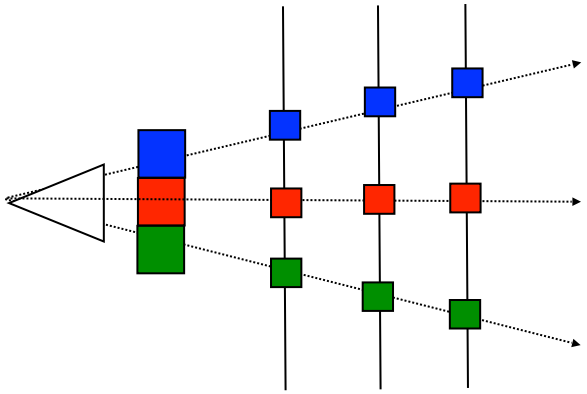
Adaptive sampling



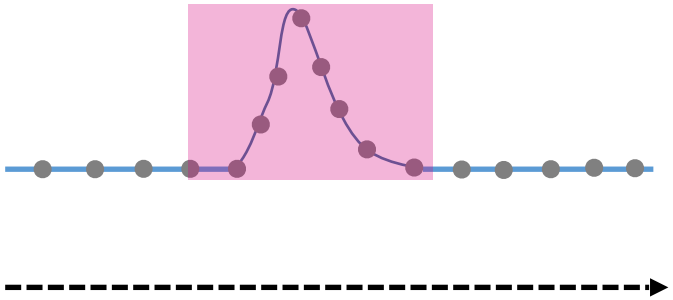
At the first stage, we uniformly sample the depth hypothesis and predict probability of depth

Uncertainty Estimation

Probability volume



Probability distribution

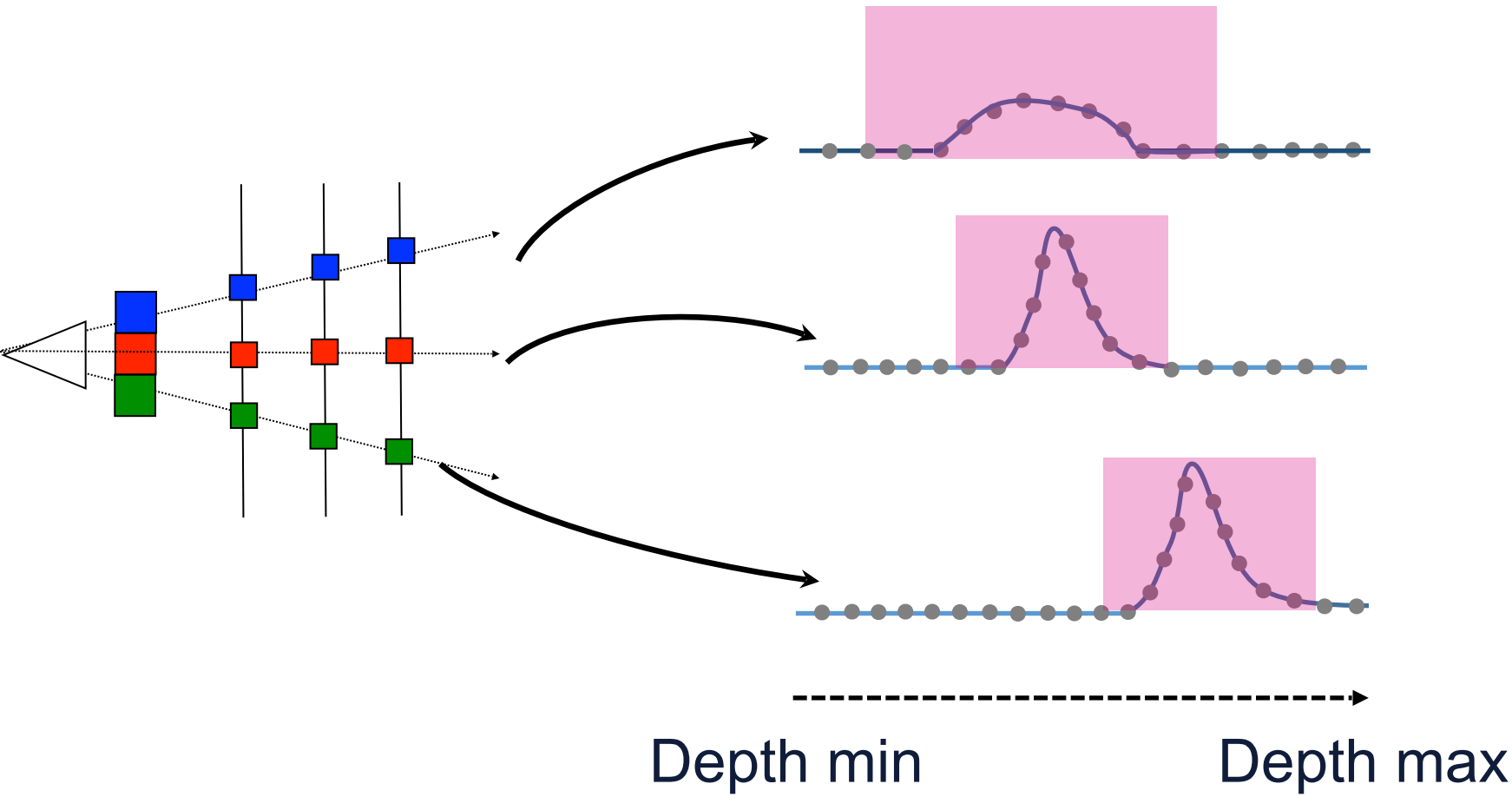


Depth min

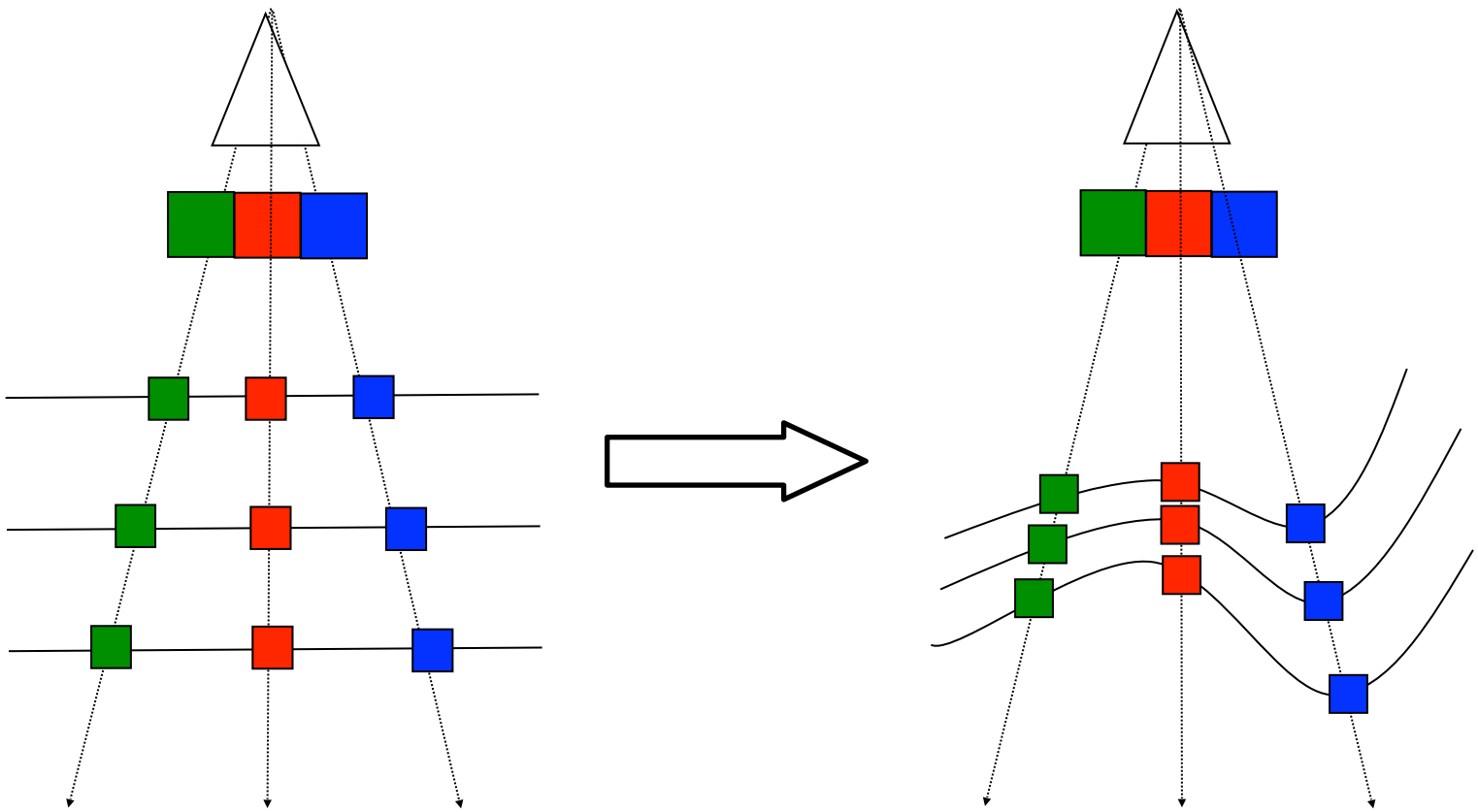
Depth max

Variance: $\hat{\mathbf{V}}_k(x) = \sum_{j=1}^{D_k} \mathbf{P}_{k,j}(x) \cdot (\mathbf{L}_{k,j}(x) - \hat{\mathbf{L}}_k(x))^2,$

Uncertainty Aware Warping



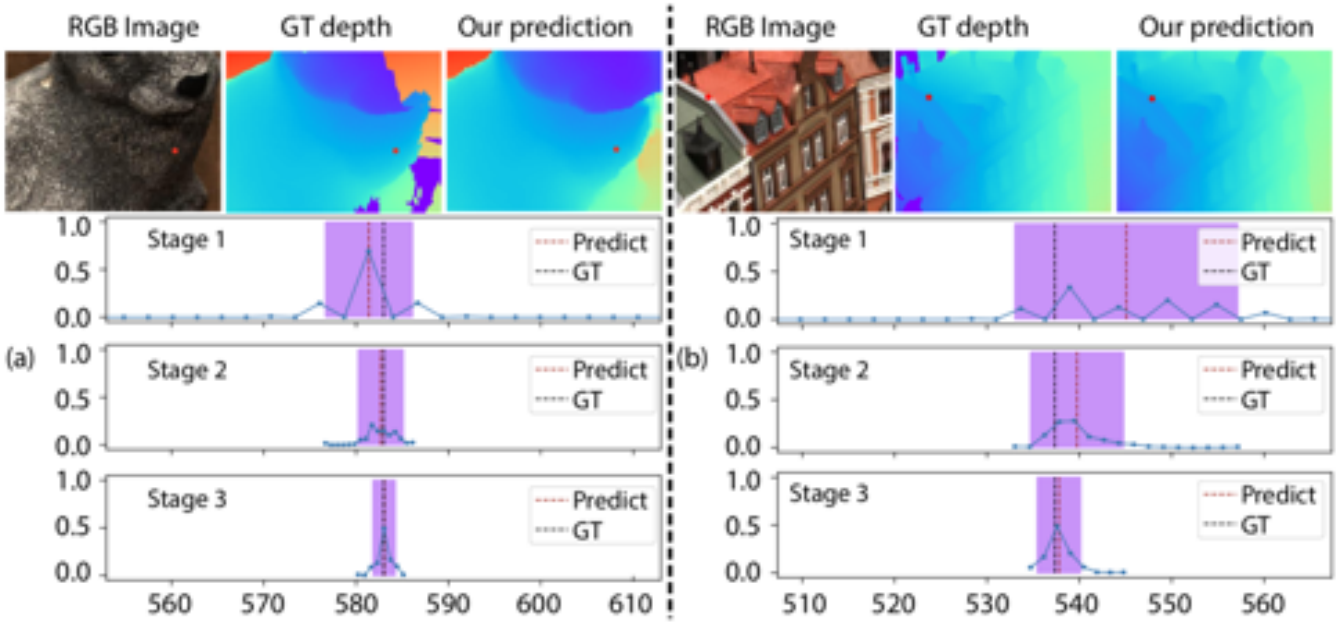
Form a New Cost Volume



Uniform
depth hypotheses

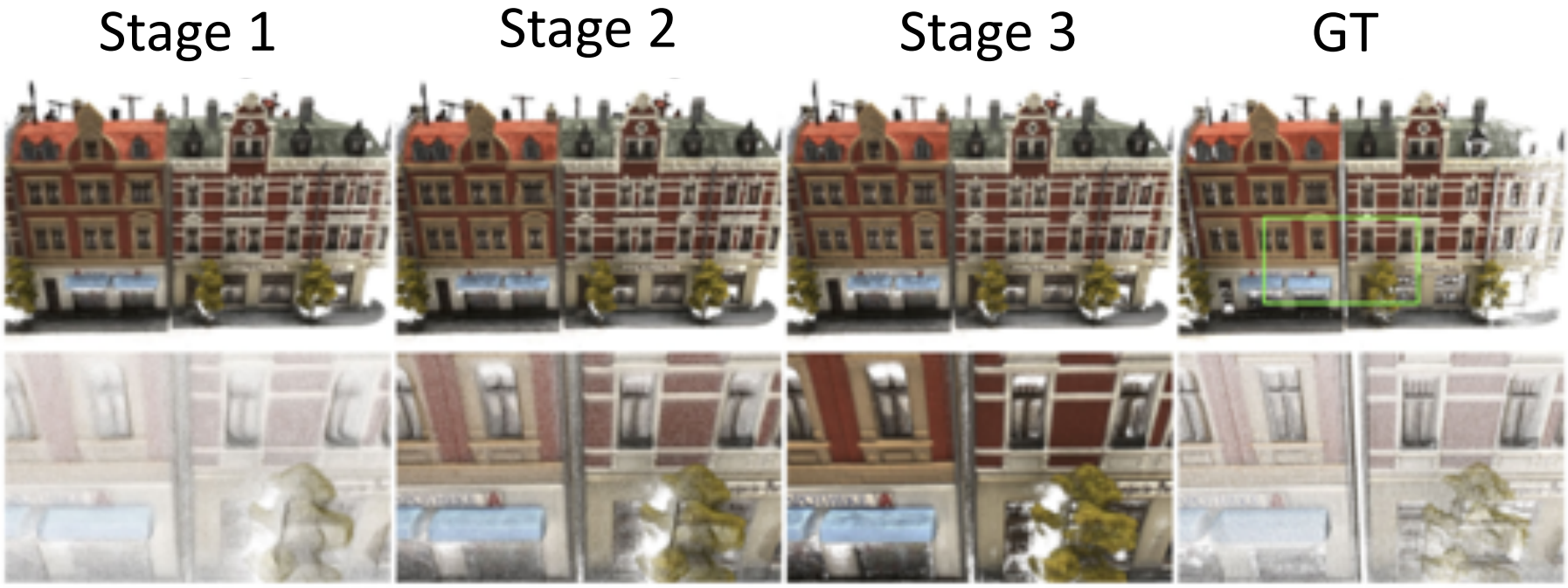
Spatially-varying
depth hypotheses

Narrowing Process Visualization



Y axis: probability
X axis: depth values
Purple region: estimated uncertainty

Point Cloud Comparison



gradually densify the local geometry

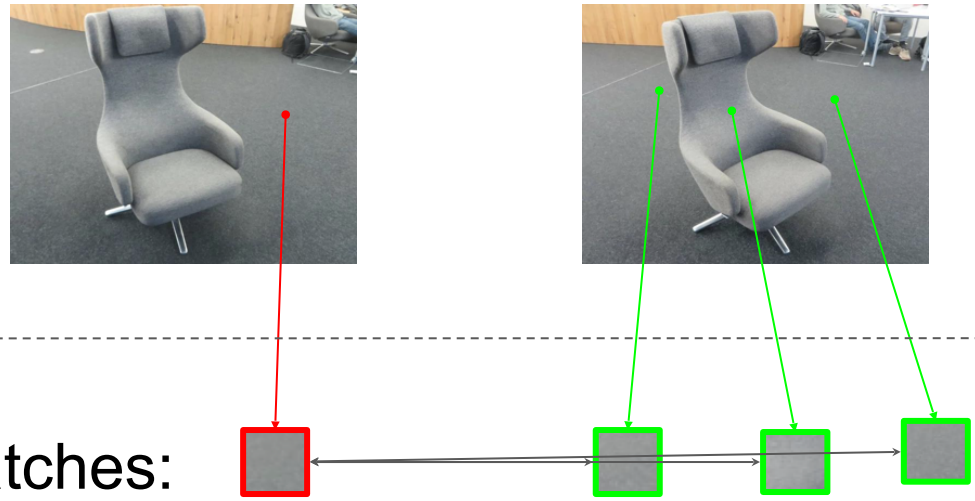
Speed & Memory Comparison

Method	Running time (s)	Memory (MB)	Input size	Prediction size
One stage	0.065	1309		160x120
Two stages	0.114	1607	640x480	320x240
Our full model	0.257	1647		640x480
MVSNet [58]	1.049	4511	640x480	160x120
R-MVSNet [59]	1.421	4261	640x480	160x120

Table 5: Performance comparisons. We show the running time and memory of our method by running the first stage, the first two stages and our full model.

Resolution (Speed) is OK. But Difficulty Still Exists

Weak texture or repetitive patterns



Observed patches:

???

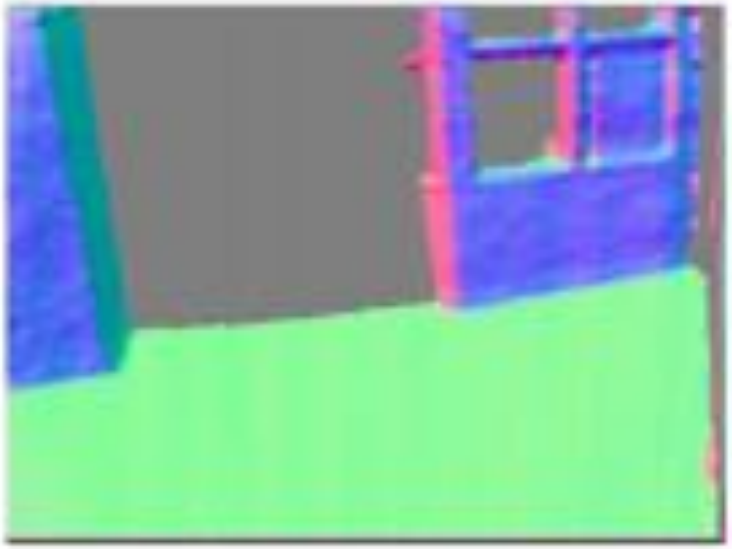
Resolution (Speed) is OK.
But Difficulty Still Exists

GT point cloud

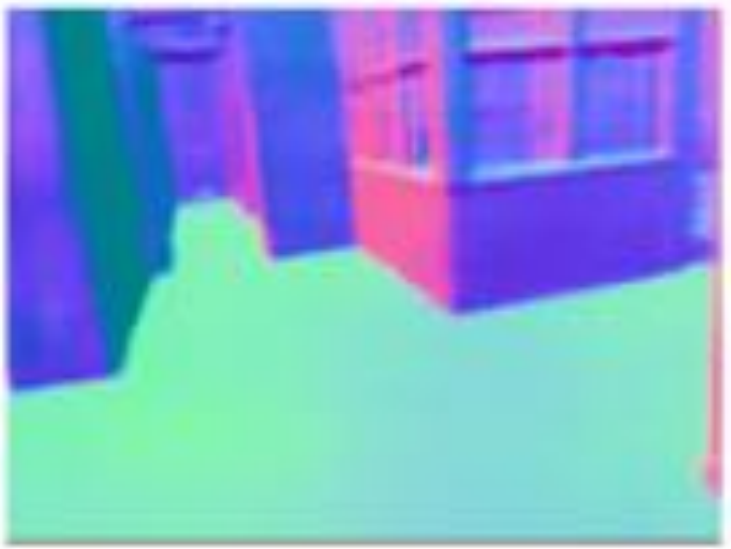
Predicted point cloud

High-order Differential Quantity is Easier to Estimate

GT normal



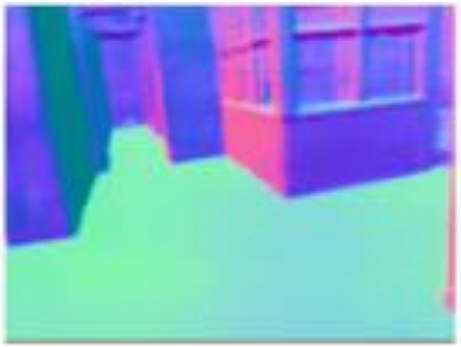
Predicted normal



Normal Prediction is Easier
(from single view)

Depth-Normal Joint Learning

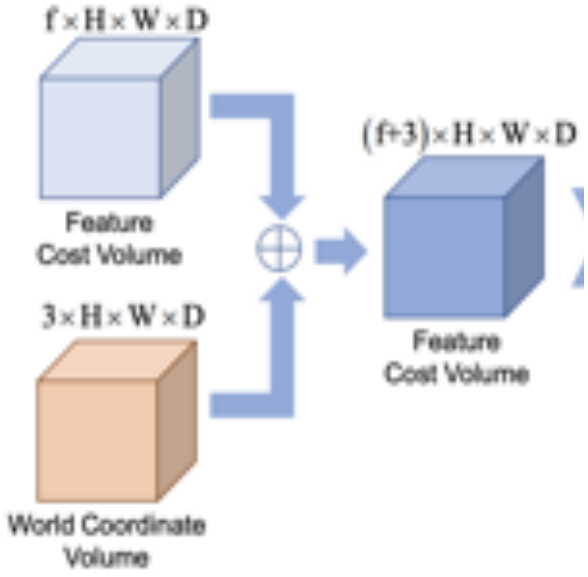
Normal



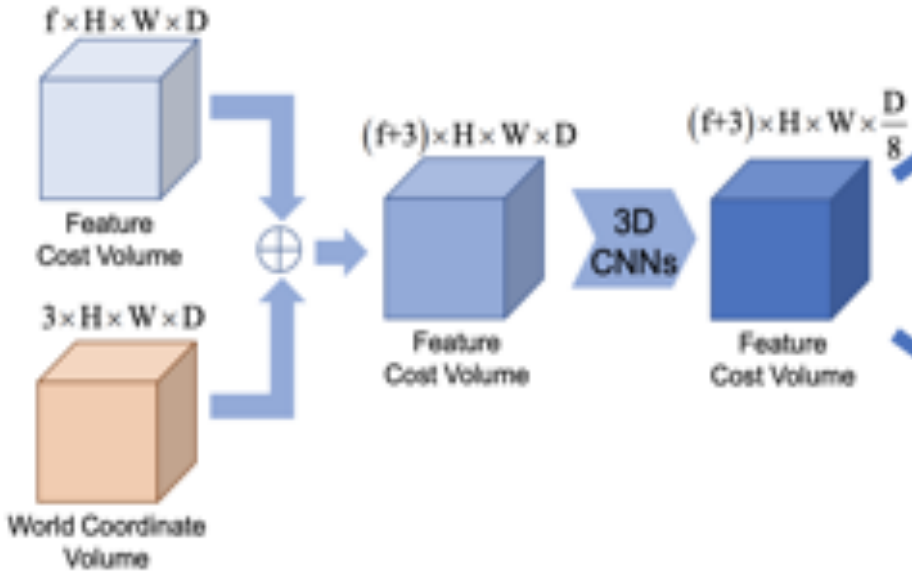
Depth

Depth-Normal Consistency

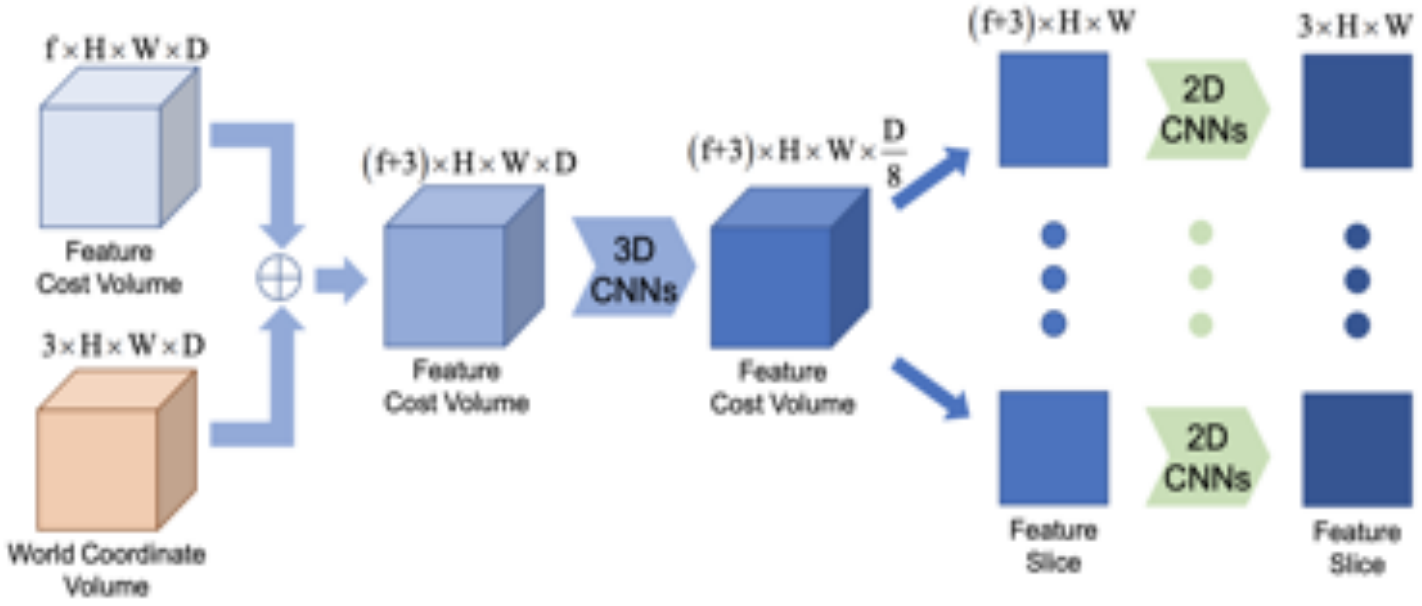
Multi-View Normal Estimation



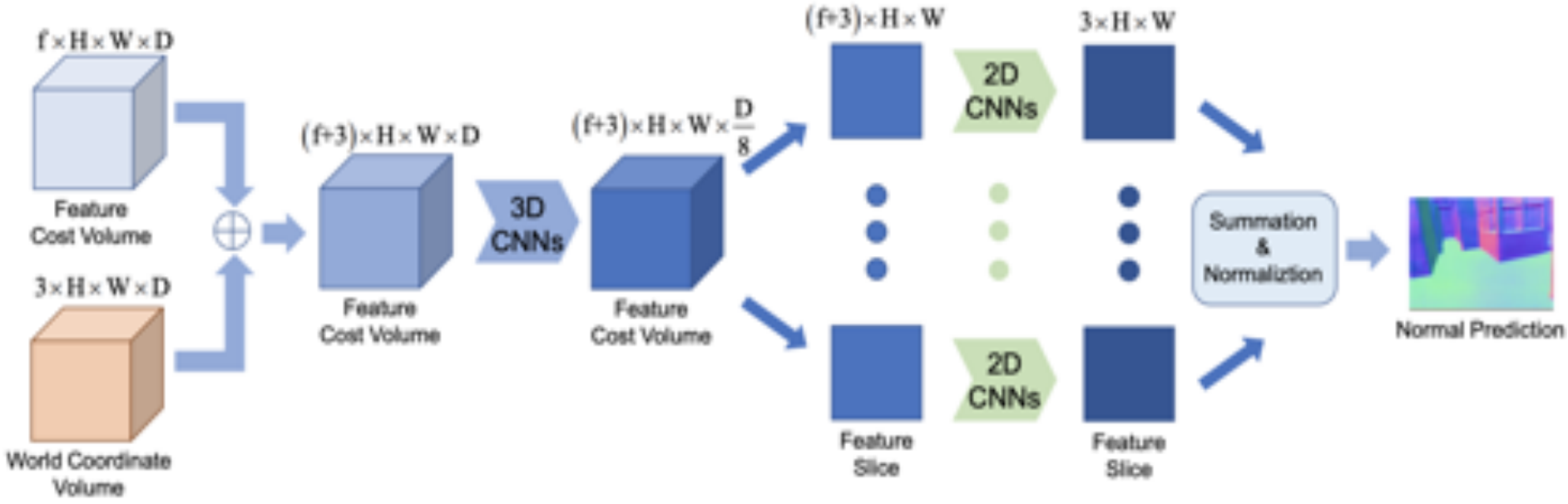
Multi-View Normal Estimation



Multi-View Normal Estimation



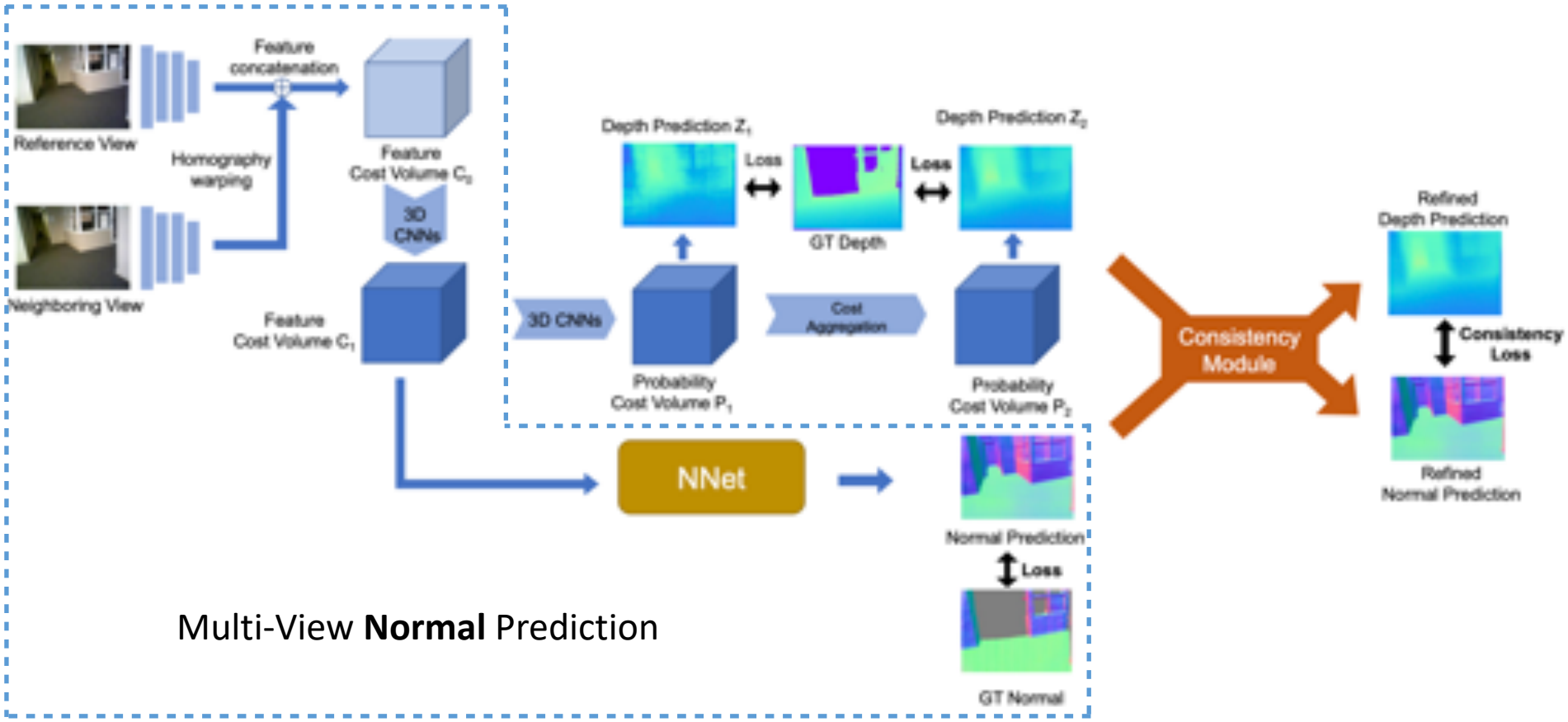
Multi-View Normal Estimation



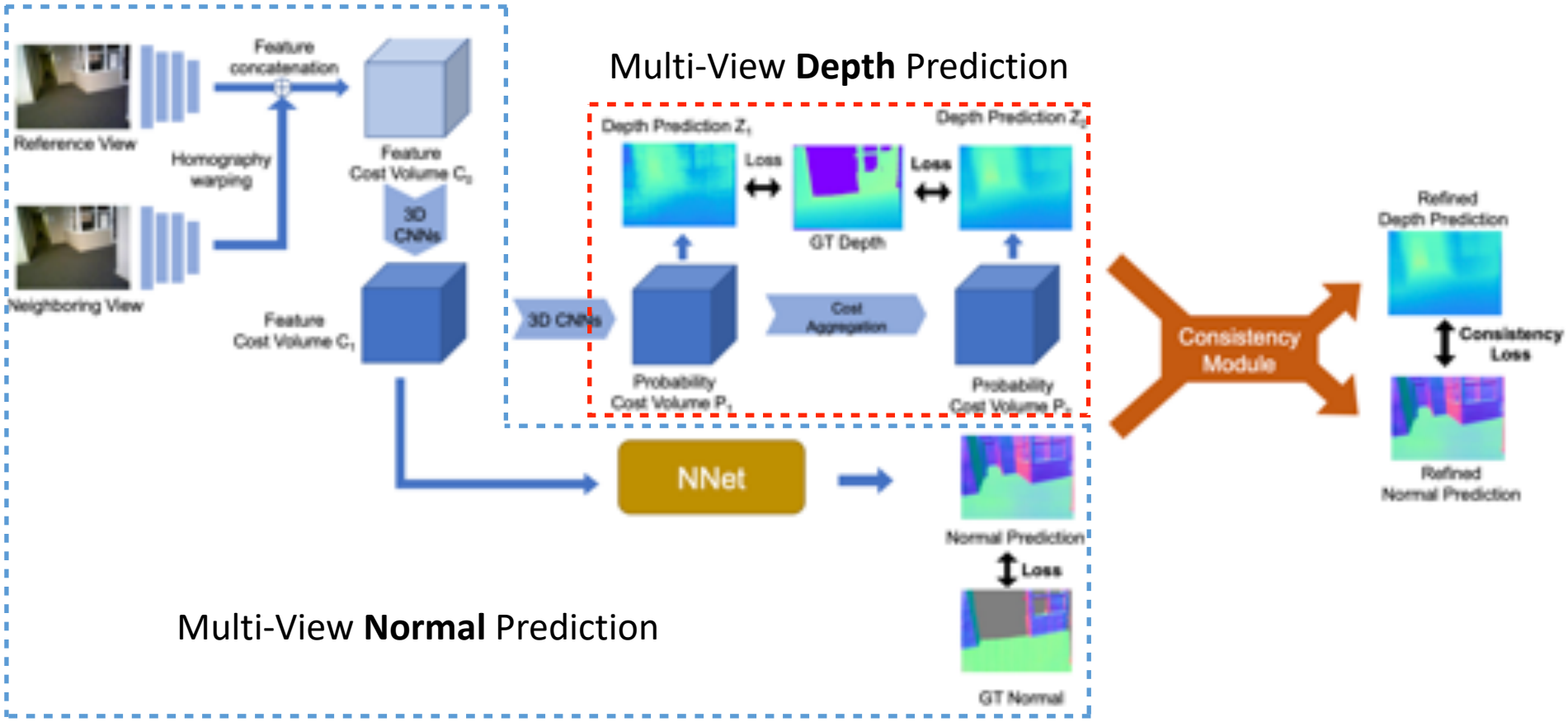
Multi-View Normal Estimation Result

Image RGB based Ours

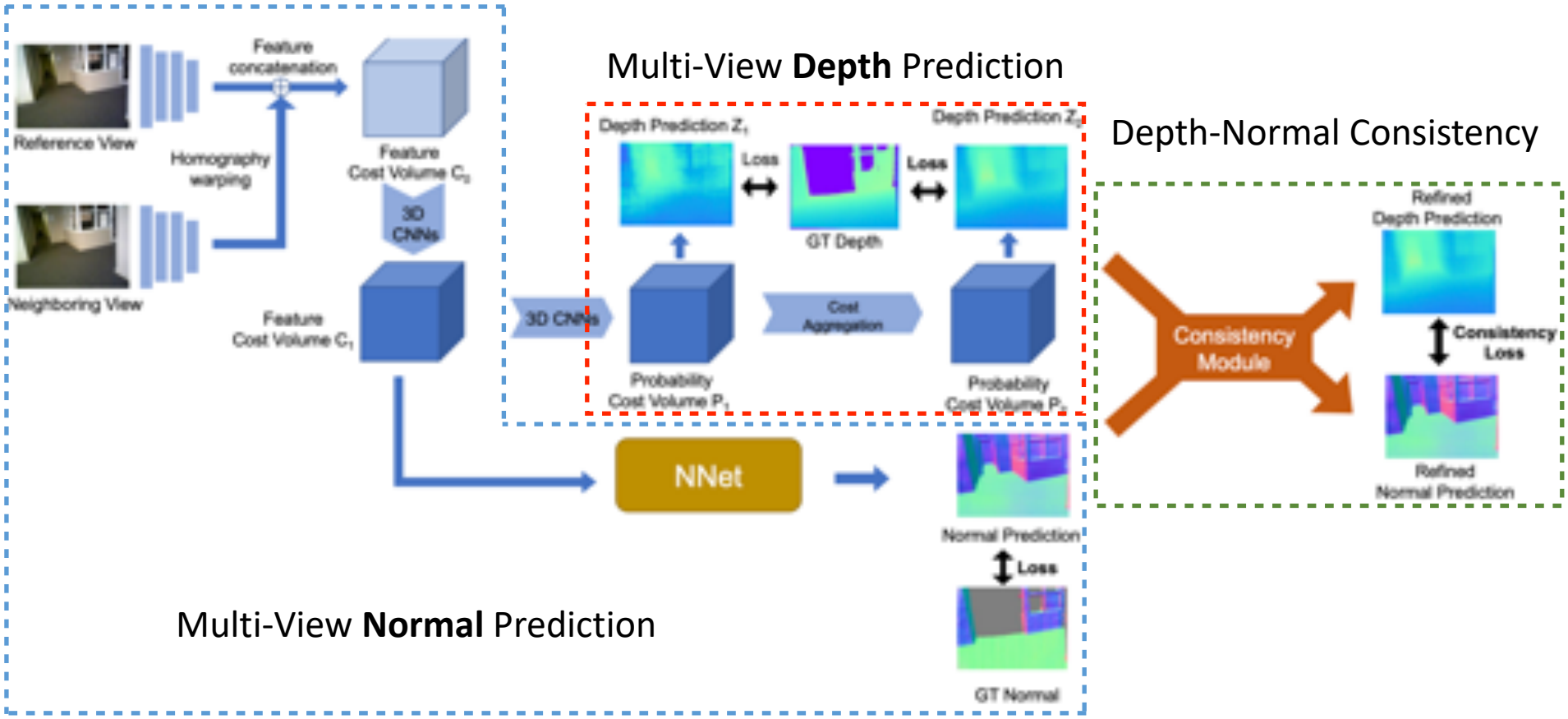
Overall Architecture



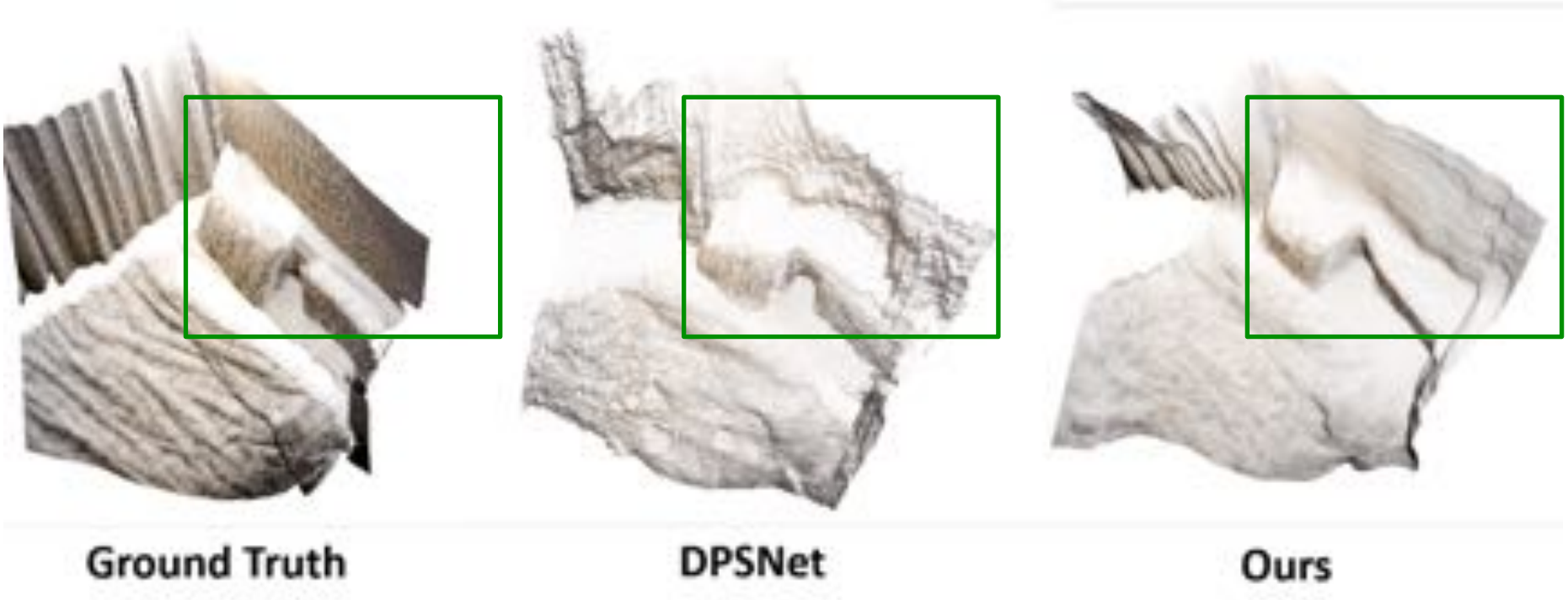
Overall Architecture



Overall Architecture



Qualitative results

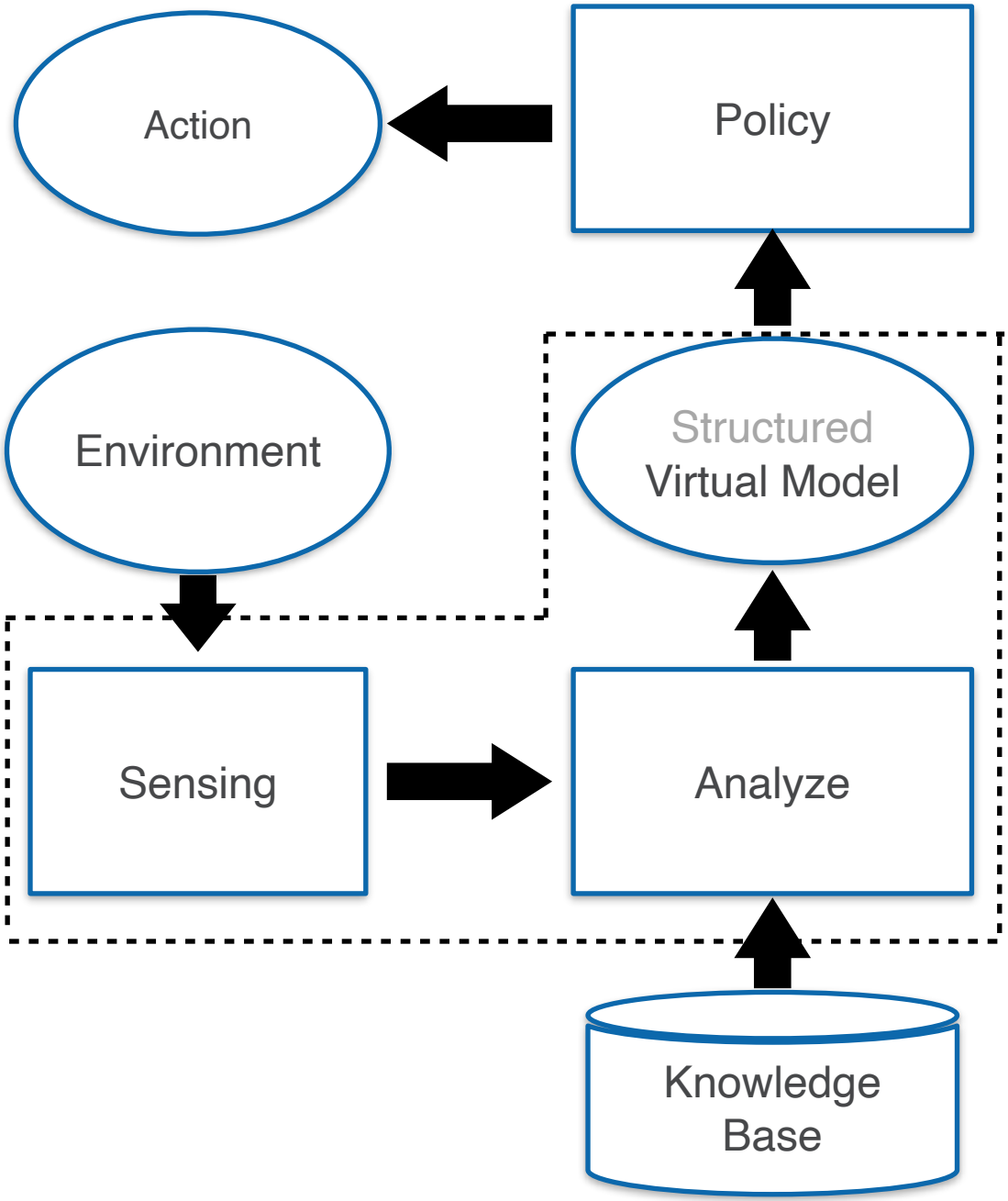


UC San Diego

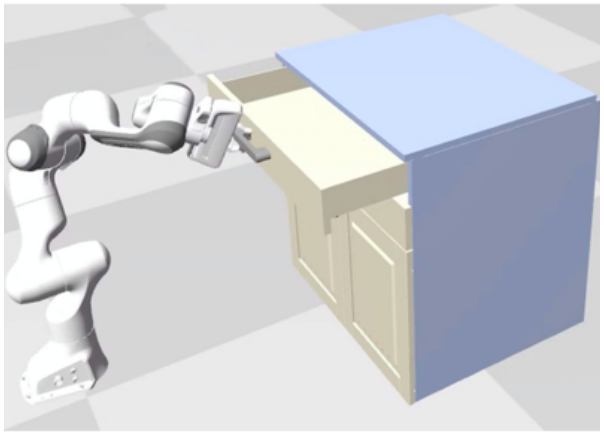
Sampled Research Work (II)

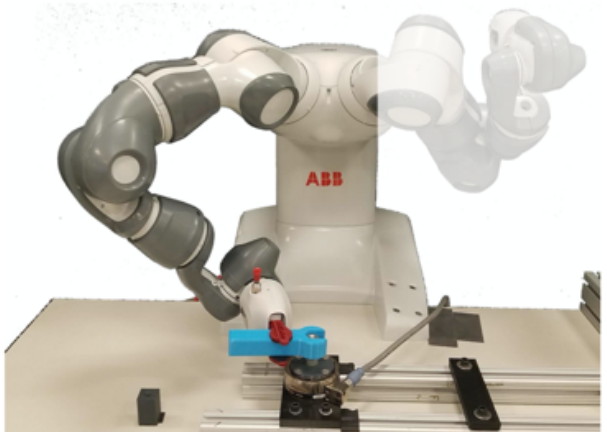
Grasp Proposal Prediction

S4G: Amodal Single-view Single-Shot SE(3) Grasp Detection in Cluttered Scenes, Qin*, Chen*, et al. CoRL 2019


Action
Policy
Environment
Structured
Virtual Model
Sensing
Analyze
Knowledge
Base

- Exploratory robot needs to infer the structure, hence functionality, of the environment
- Reconstruction only does not permit interaction!





1. *Source: Boston Dynamics*
2. *Source: Nvidia Robotics Research*
3. *Source: Eckovation*
4. *Source: MCube Lab*

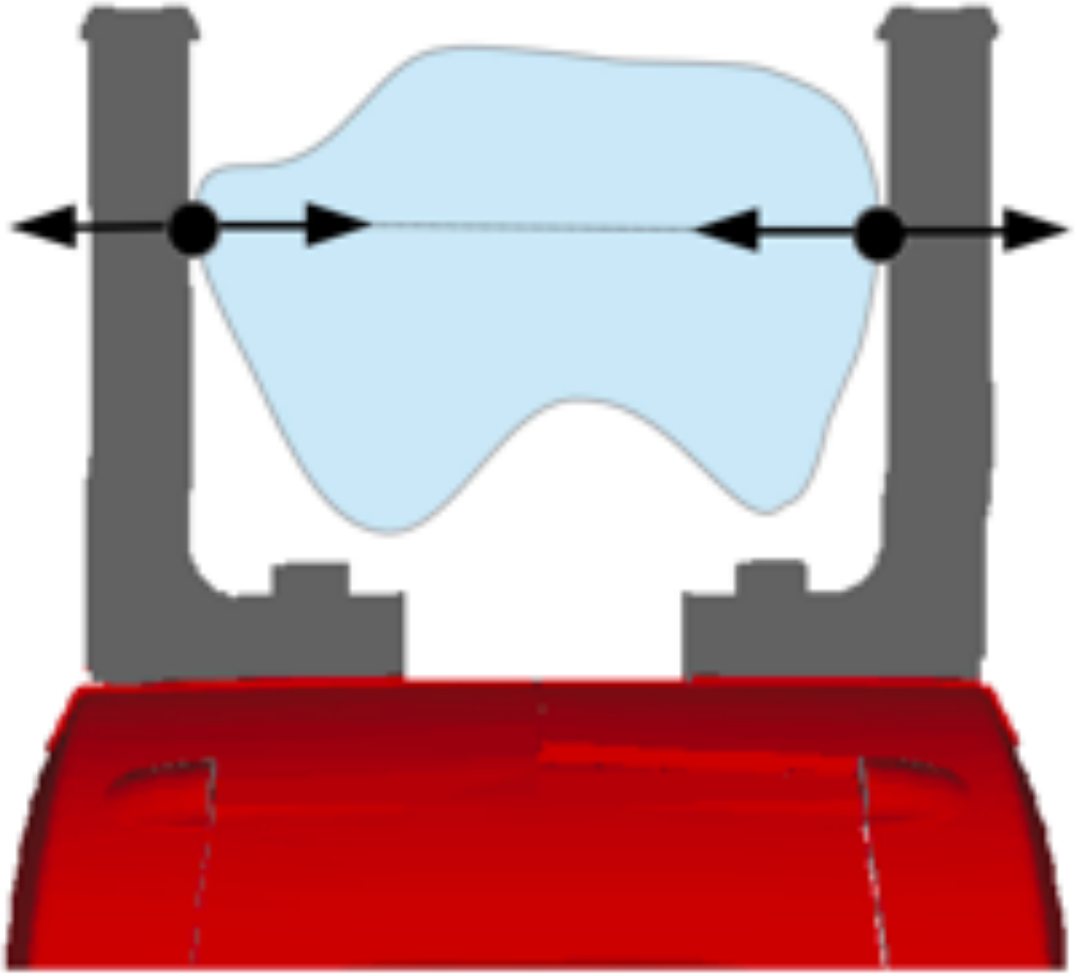
Primary Action: Grasping

Most structure action requires first to grasp the object before any specific action

1. Approach the object with appropriate direction
2. Grasp and hold the object
3. Execute object-specific manipulation

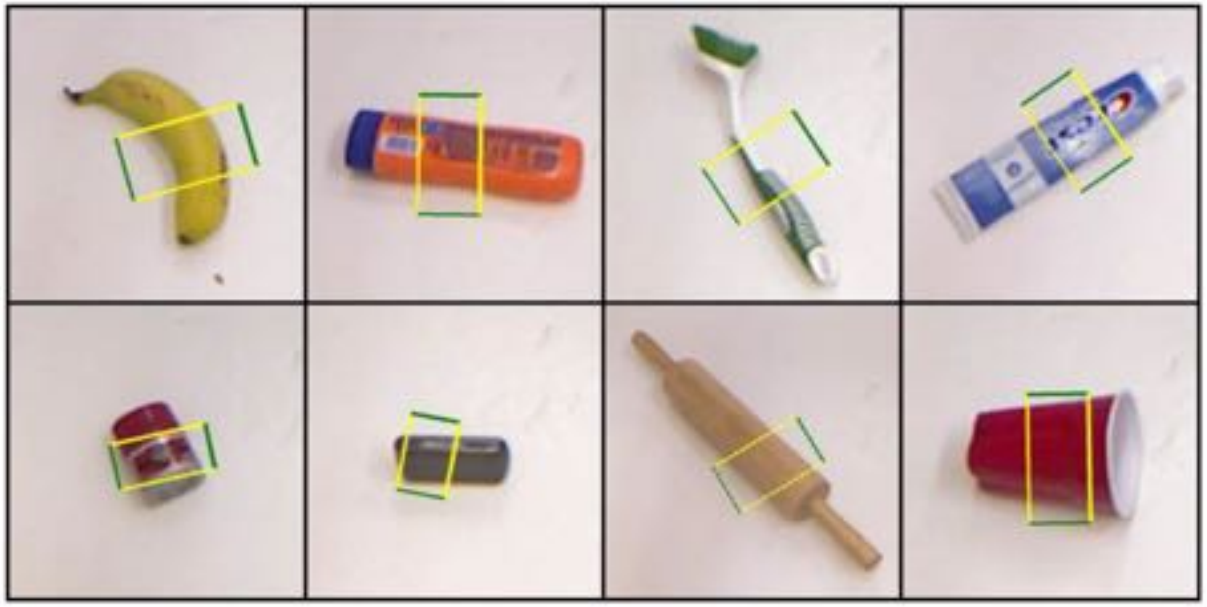
The ability to grasp any object is the preliminary for efficient robot exploration

Antipodal Grasp



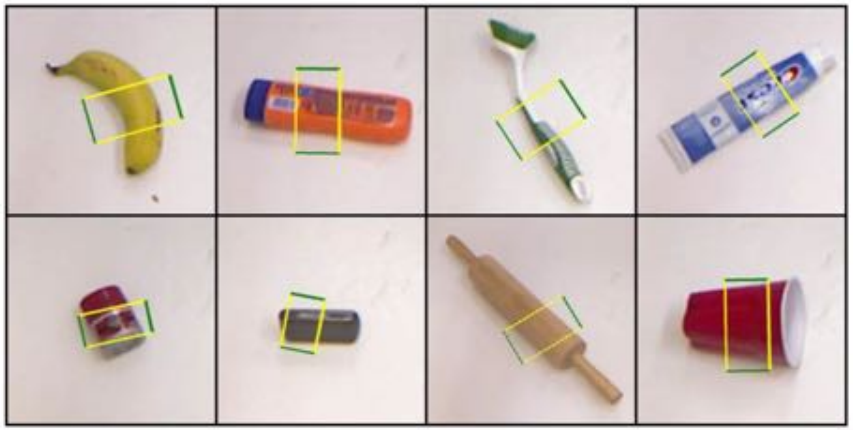
Current Fashion: Data-driven

- Formulate grasping as a object-detection problem
- Represent grasp pose as bounding box



Redmon et al., "**Real-Time Grasp Detection Using Convolutional Neural Networks**", *ICRA 2015*

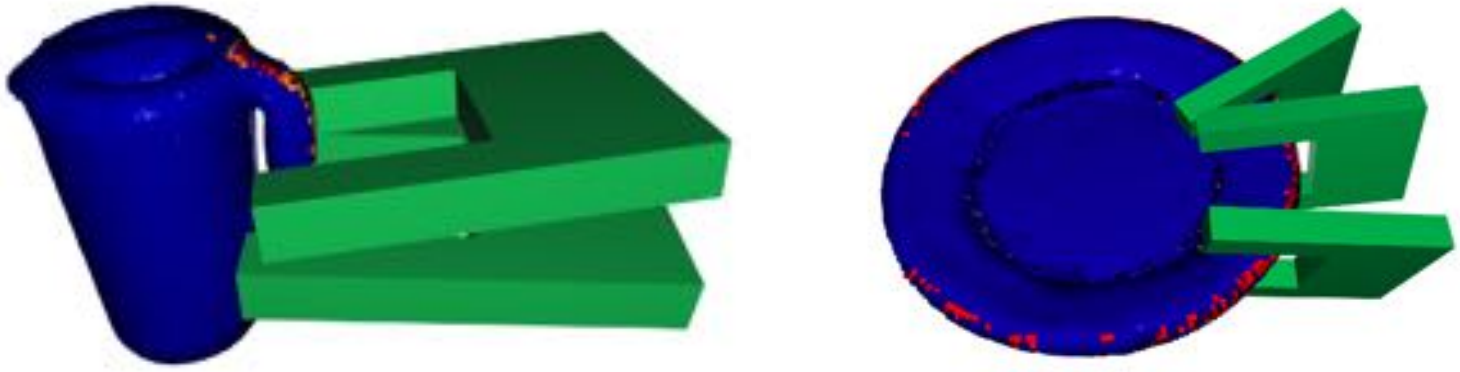
2D Detection-based Grasping



Limit approach direction to top-down

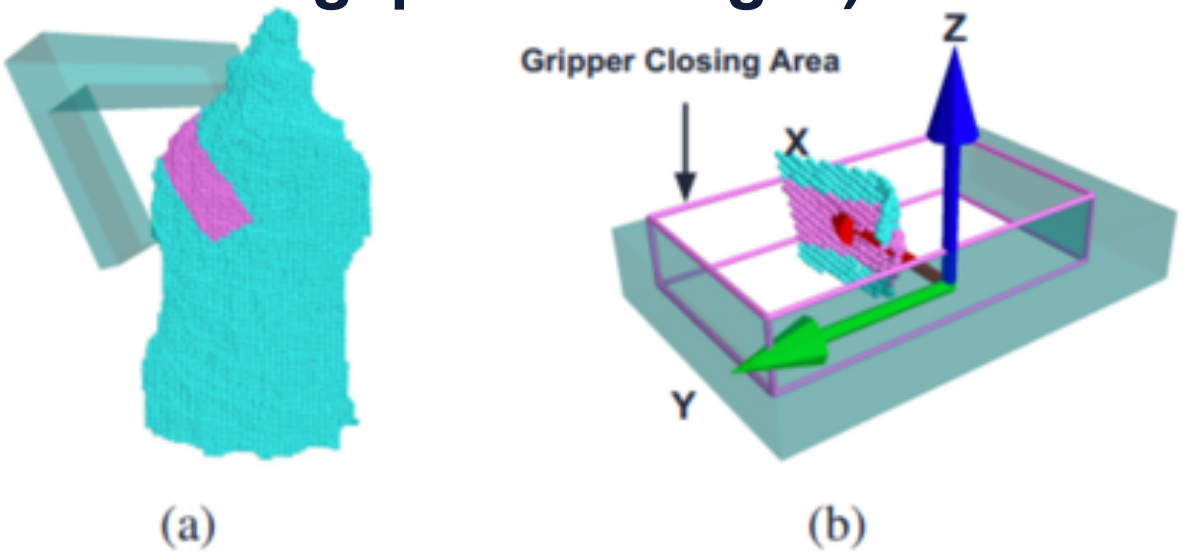
Not applicable for highly-free exploration

Grasp in SE(3)



3D Geometry-based Grasping

- Utilize 3D representation for grasp evaluation
- Detect grasp poses based on geometric structure but not object semantics
- **Better generalizability to unknown objects (PC has smaller domain gap than images)**



Liang et al., "**PointNetGPD: Detecting Grasp Configuration from Point Sets**", *ICRA 2015*

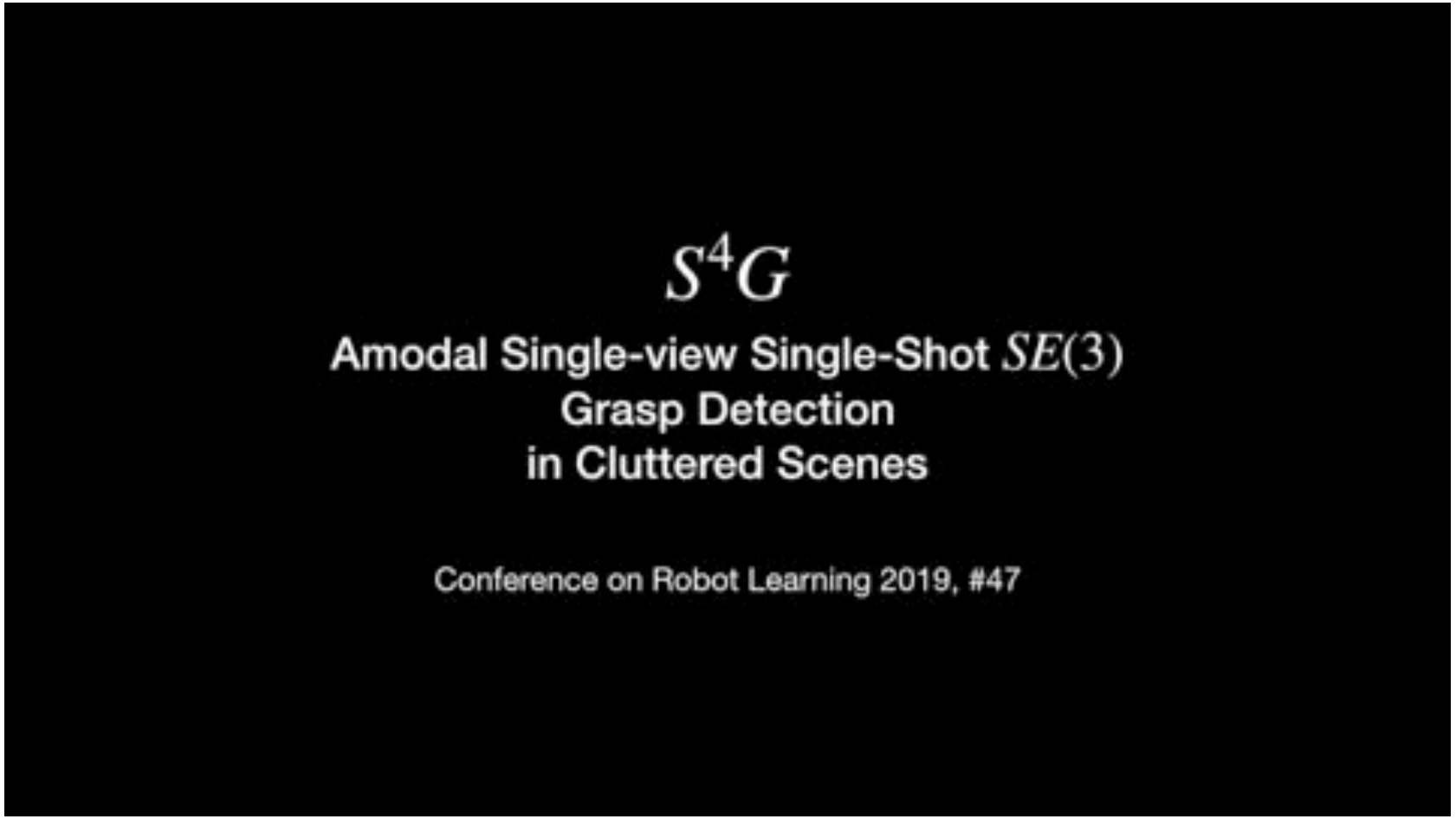
Challenge for Geometry-based Grasping

- High-quality grasp is hard to annotate
  - Human do not know either
  - Infinite answers for an single object
- Grasp pose in 3D is hard to regress
  - Representation of $SE(3)$
- Low quality of current commodity 3D sensor

Problem Setting of S4G

- Single-view: only see partial point cloud
- Commercial Kinect2: noisy sensing
- 6 Degree of Freedom: No direction limitation
- Clutter scene: stacked objects with occlusion

S4G: SE(3) Grasp Generation from 3D Point Cloud



Qin et al., "**S4G:Amodal Single-view Single-Shot SE(3) Grasp Detection in Cluttered Scenes**", *CoRL 2019*

High-level Idea

- Sim2Real+Imitation Learning

High-level Idea

- Sim2Real+Imitation Learning

- For objects in training data
  - sample grasps (gripper pose)
  - verify by force closure (using full geometry)
  - **record good ones on the shape surface (grasp pose function defined on the surface)**

High-level Idea

- Sim2Real+Imitation Learning
- For objects in training data
  - sample grasps (gripper pose)
  - verify by force closure (using full geometry)
  - **record good ones on the shape surface (grasp pose function defined on the surface)**
- Simulate partial scan of objects in the training data

High-level Idea

- Sim2Real+Imitation Learning (Search+NN)

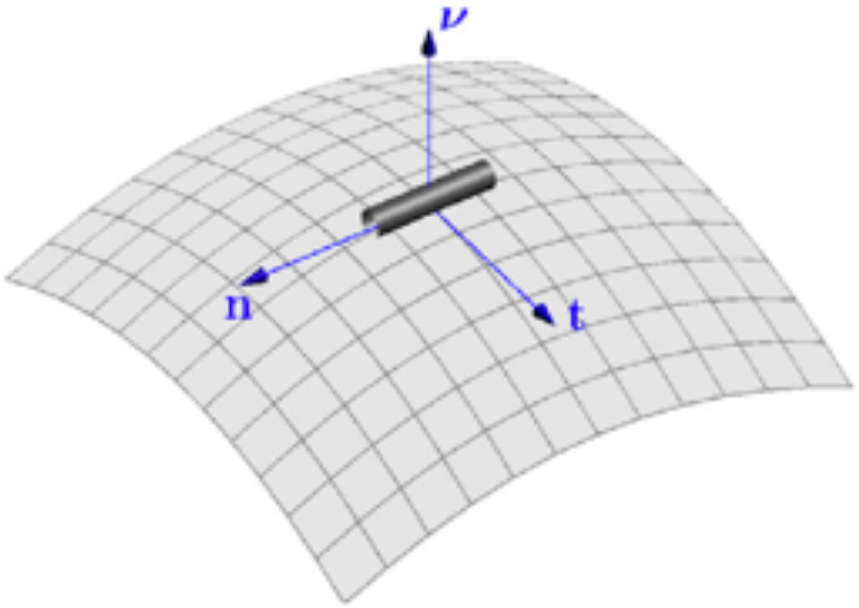
- For objects in training data
  - sample grasps (gripper pose)
  - verify by force closure (using full geometry)
  - **record good ones on the shape surface (grasp pose function defined on the surface)**

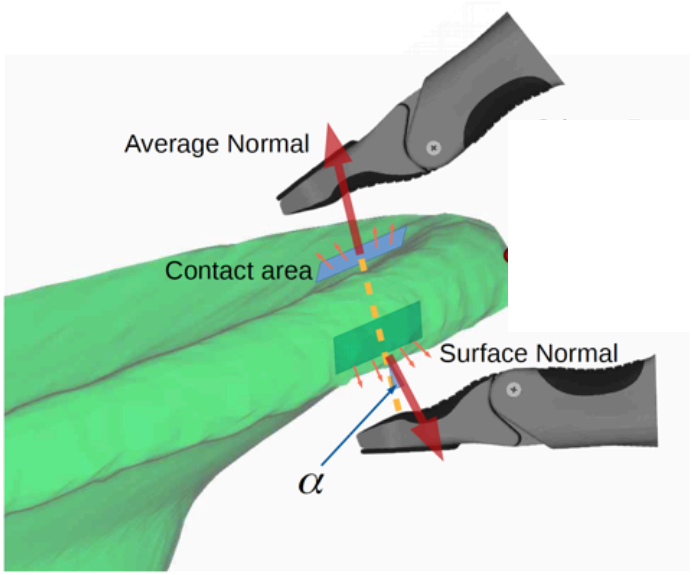
- Simulate partial scan of objects in the training data

- Use neural network to learn the grasp pose function from partial scans

Search For Object-Centric Grasps

- Enumerate possible grasps based on local geometry around contact points



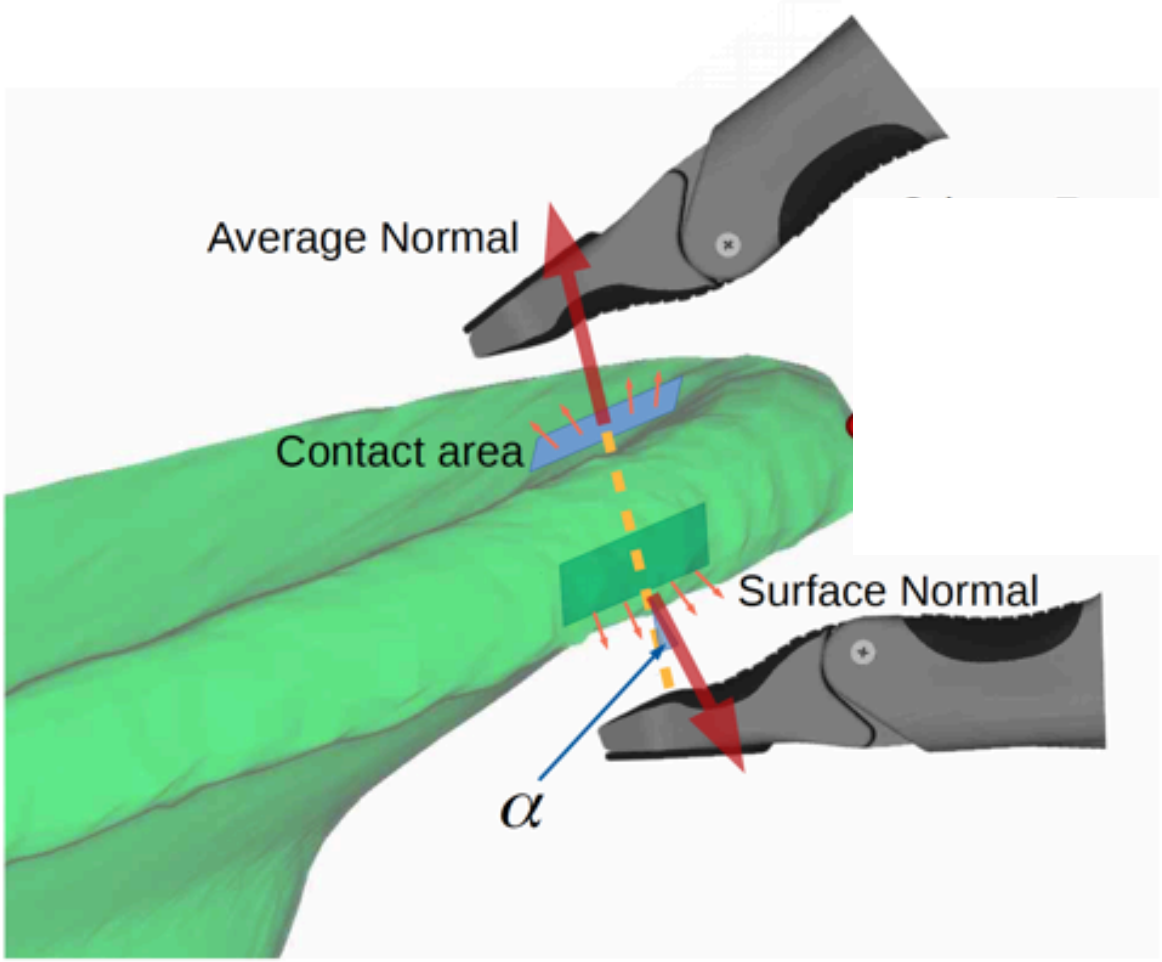


Classical: Daboux Frame

Ours

Search For Object-Centric Grasps

- Verify by **force-closure** (can resist external forces)

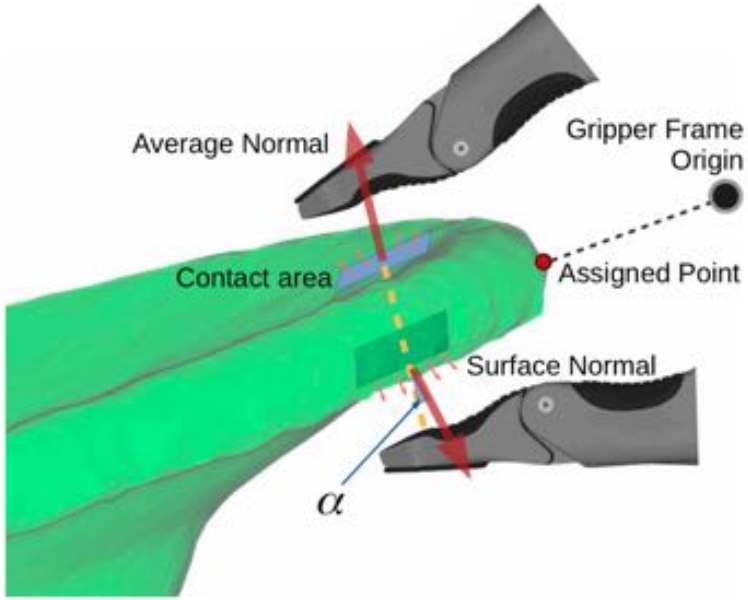


Good Grasps as Surface Function

Regression grasp pose precisely is hard globally

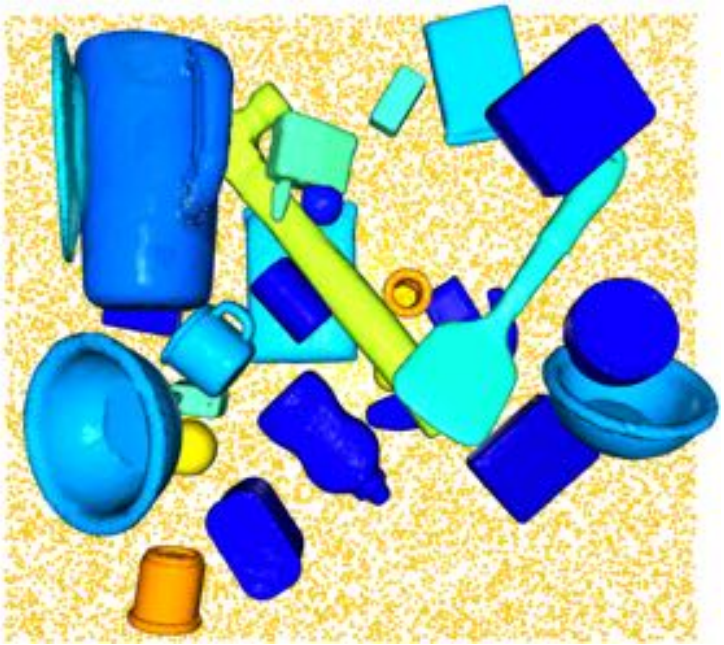
- The size of arena: 1.5m
- However, 1.5 cm (1%) error is large enough for failure

- Solution: regress local poses
  - In dataset, register each grasp with nearest point
  - Predict local offset with respect to this point



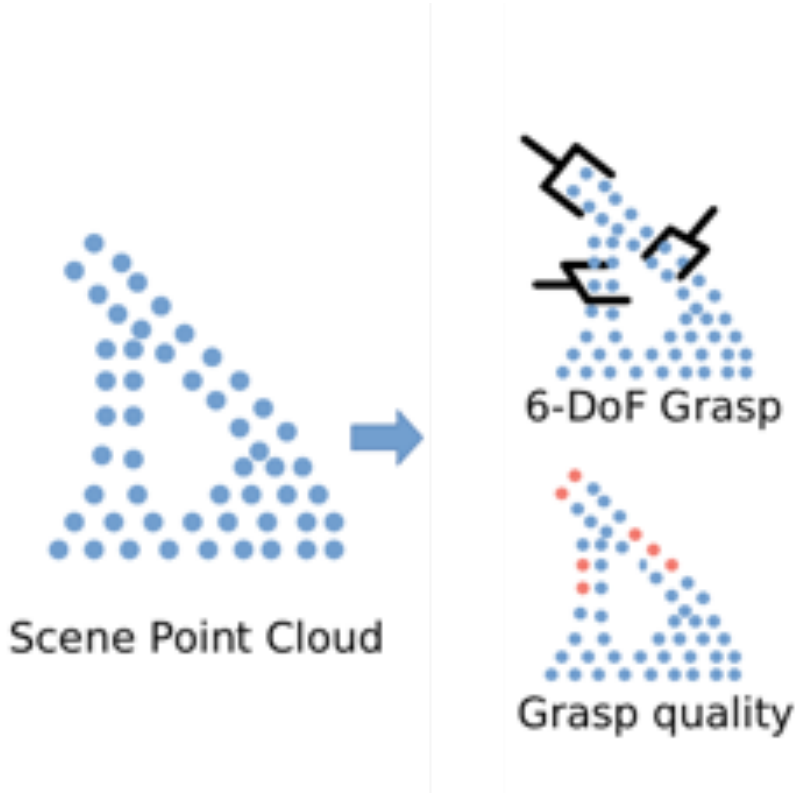
Scene-level Considerations

- Collision checking with the whole scene
- Render depth from different views as input for network



Single-shot Grasp Proposal

- Input: single-view observation
- Output: grasp poses and corresponding quality scores



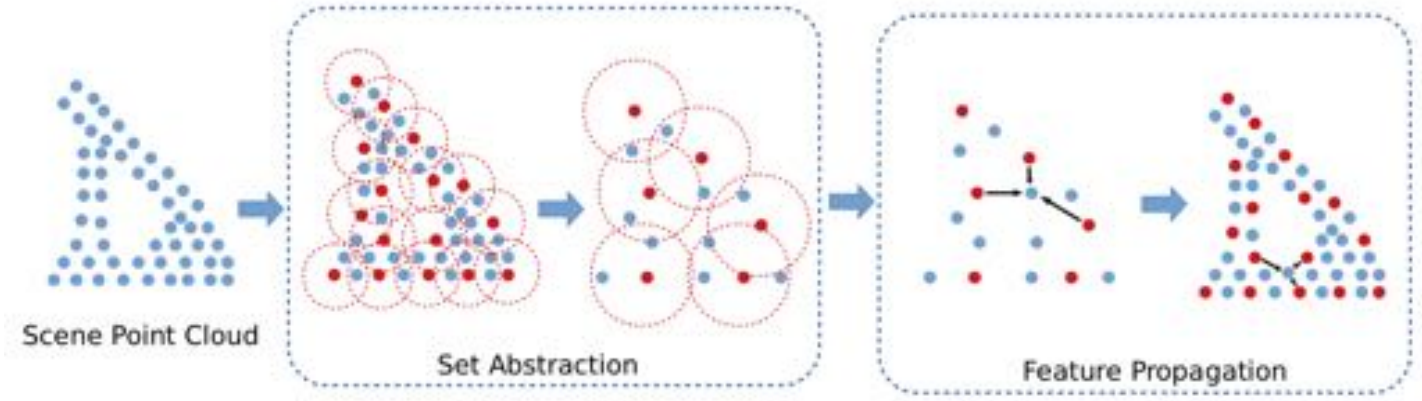
PointNet++: Extract Hierarchical Features

Local features:

- How to grasp the object

Global features:

- Avoid collision with other object



Quantitative Result

Outperform other SOTA methods with large margin in accuracy and efficiency

	Grasp quality		Time-efficiency		
	Success rate	Completion rate	Preprocess	Inference	Postprocess
GPD (3 channels)	40.0%	60.0%	24103 ms	**1.50 ms**	**3.27 ms**
GPD (12 channels)	33.3%	50.0%	27190ms	1.70ms	4.66ms
PointNetGPD	40.0%	60.0%	17687ms	2.86ms	6.73ms
Ours	**77.1%**	**92.5%**	**5617ms**	12.60 ms	186.58 ms

Discussion

- Main error source
  - Low depth map quality (precision+completeness)

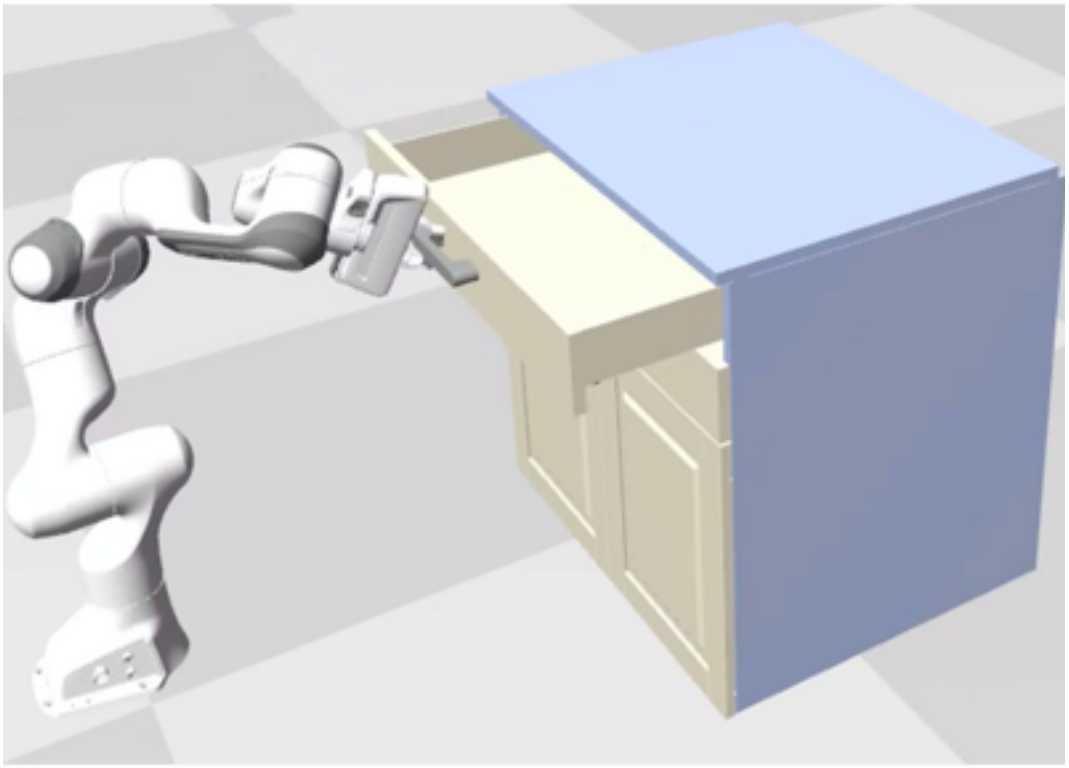
Cheap and high-quality 3D sensor is vital

- Sim2Real:
  - Model trained on sim directly applied on real:
  - RGB information is not used in this work

Point Cloud representation: lower domain gap

So far, Purely Mechanics-based

- Exploratory robot should use manipulation as a mean to **verify structure hypothesis** of objects



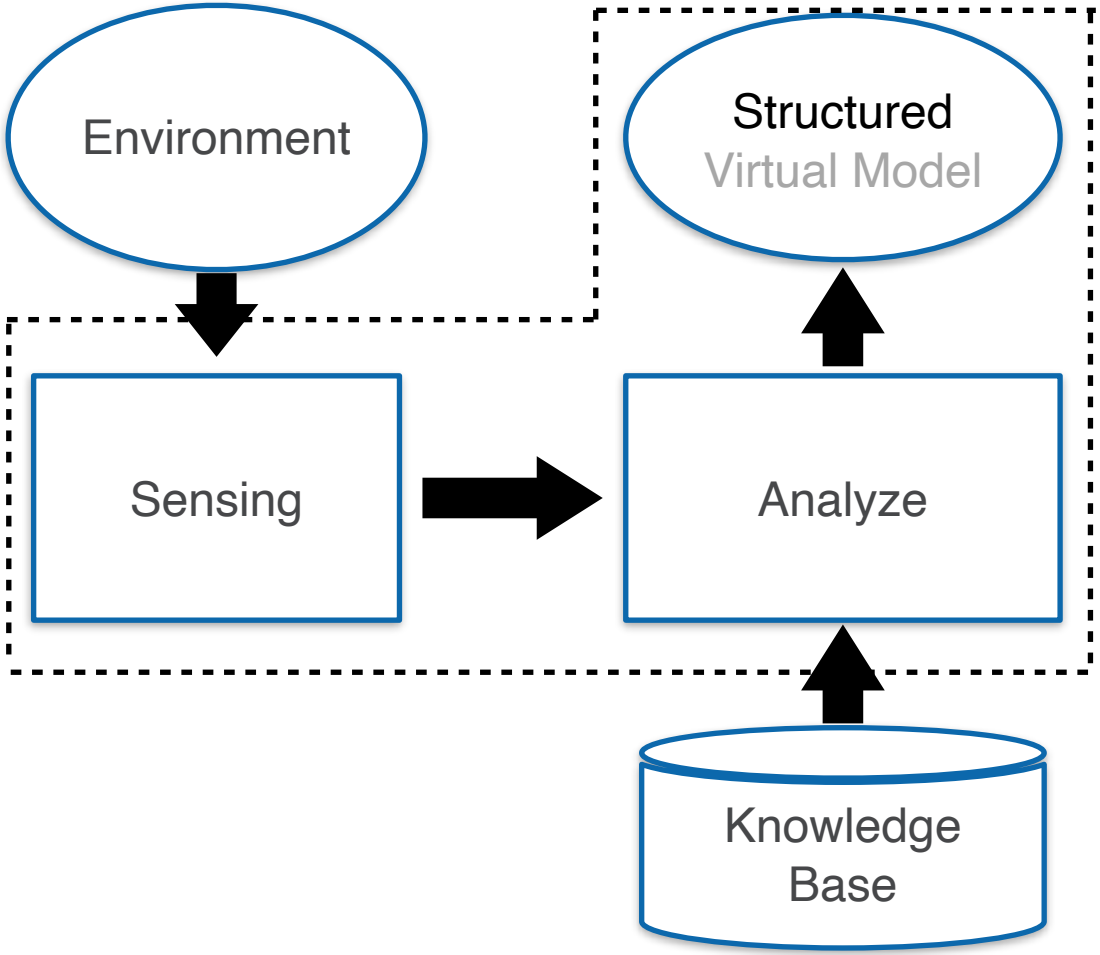
Source: Eckovation

UC San Diego

Sampled Research Work (III)

Structure Hypothesis Generation: Zero-shot 3D Part Proposal

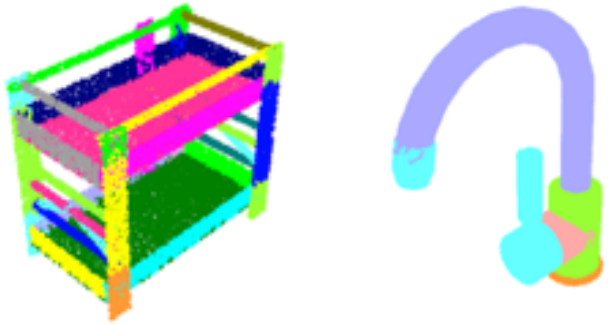
Learning to Group: A Bottom-Up Framework for 3D Part Discovery in Unseen Categories, Luo et al. ICLR 2020


Environment
Structured
Virtual Model
Sensing
Analyze
Knowledge
Base

Task

Data in Knowledge Base

New Data

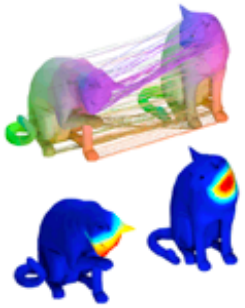


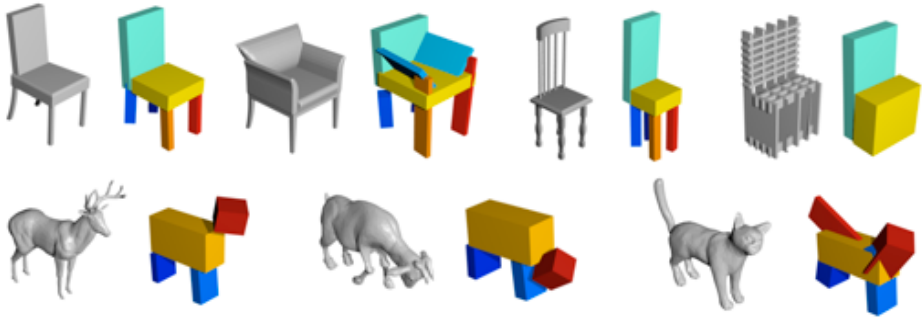
Training set and test set are of different categories, but reuse local structures

Why Few-shot/Zero-shot Learning in 3D?

Algorithmically, 3D shapes are:

- easier to be **related (correspondence)**
- easier to be **compared**
- easier to **abstracted**





Revisit 3D Part Segmentation

- Learning-Based Methods
  - Fully Convolutional [PartNet-InsSeg, Mo et al.]
  - Clustering Based [SGPN, Wang et al.]
  - Segmentation by Synthesis [GSPN, Yi et al.]

Train on chair, storage furniture and lamp, Test on faucet

Fully Conv **Clustering** **Seg by Synth** **Reference**

Revisit 3D Part Segmentation

- Traditional Methods
  - Use part geometry heuristics
  - convexity, flatness, etc [WCSeg, Kaick et al.]

Train on chair, storage furniture and lamp, Test on faucet

Fully Conv

Clustering

Seg by Synth

Reference

Traditional

Key Idea

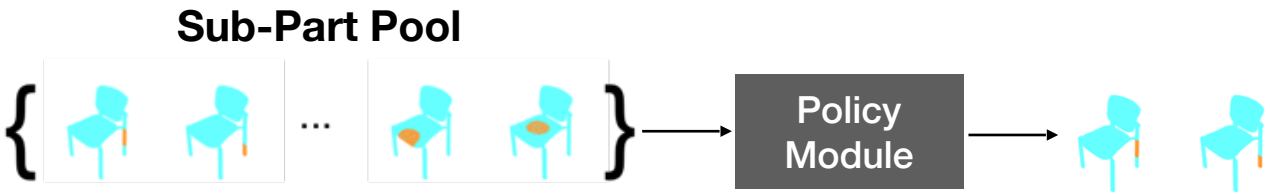
Incorporating **global context** is likely to hurt zero-shot generalization.

Should be **parsimonious** in using context information.

Our Approach

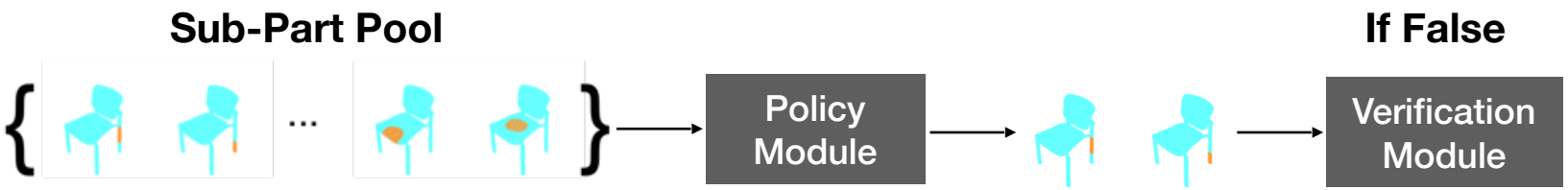
Sub-Part Pool

Our Approach

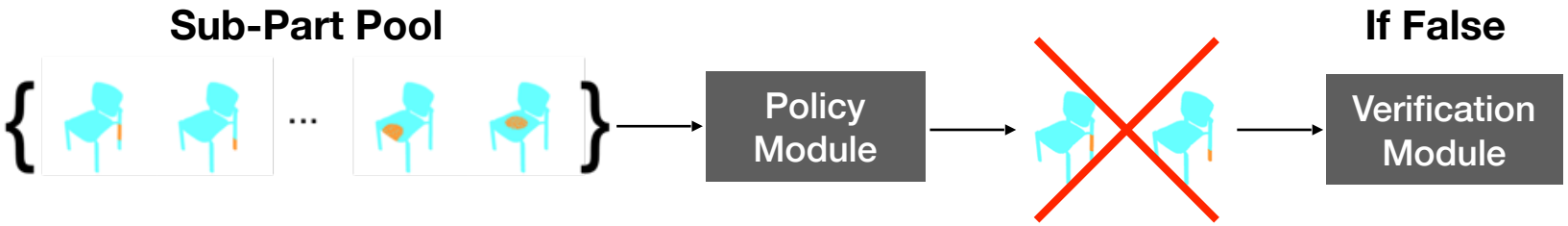


Our Approach

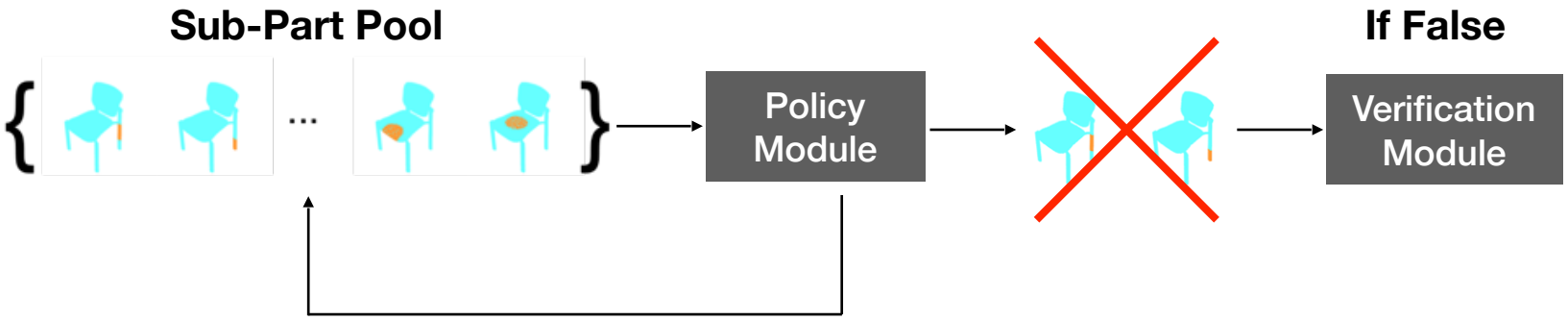
Our Approach



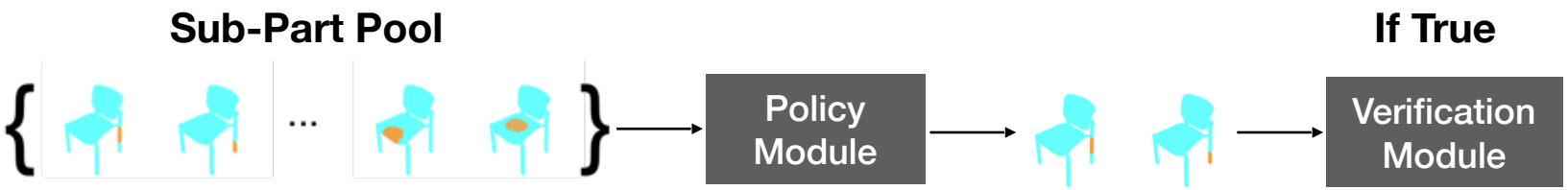
Our Approach



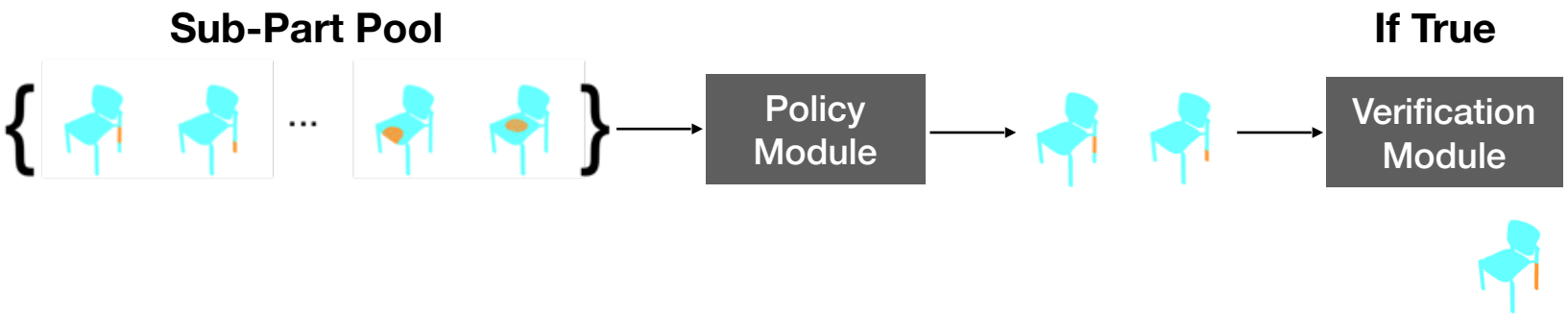
Our Approach



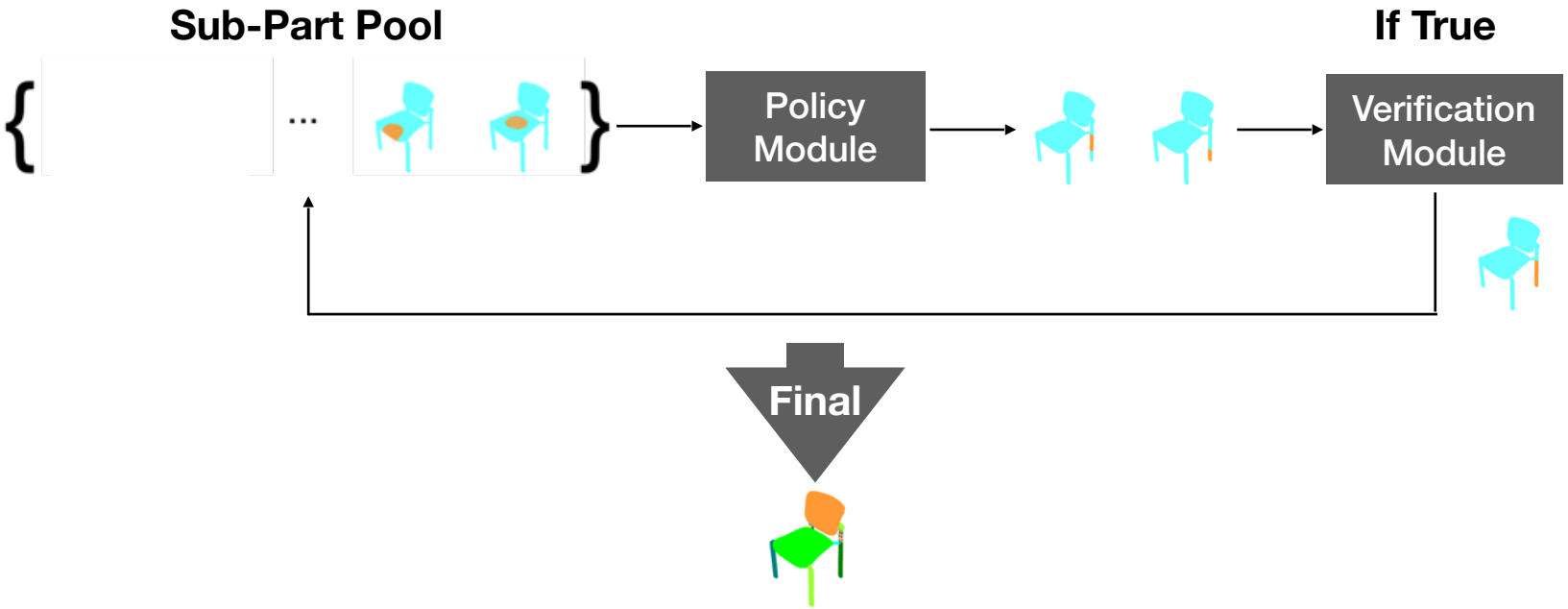
Our Approach



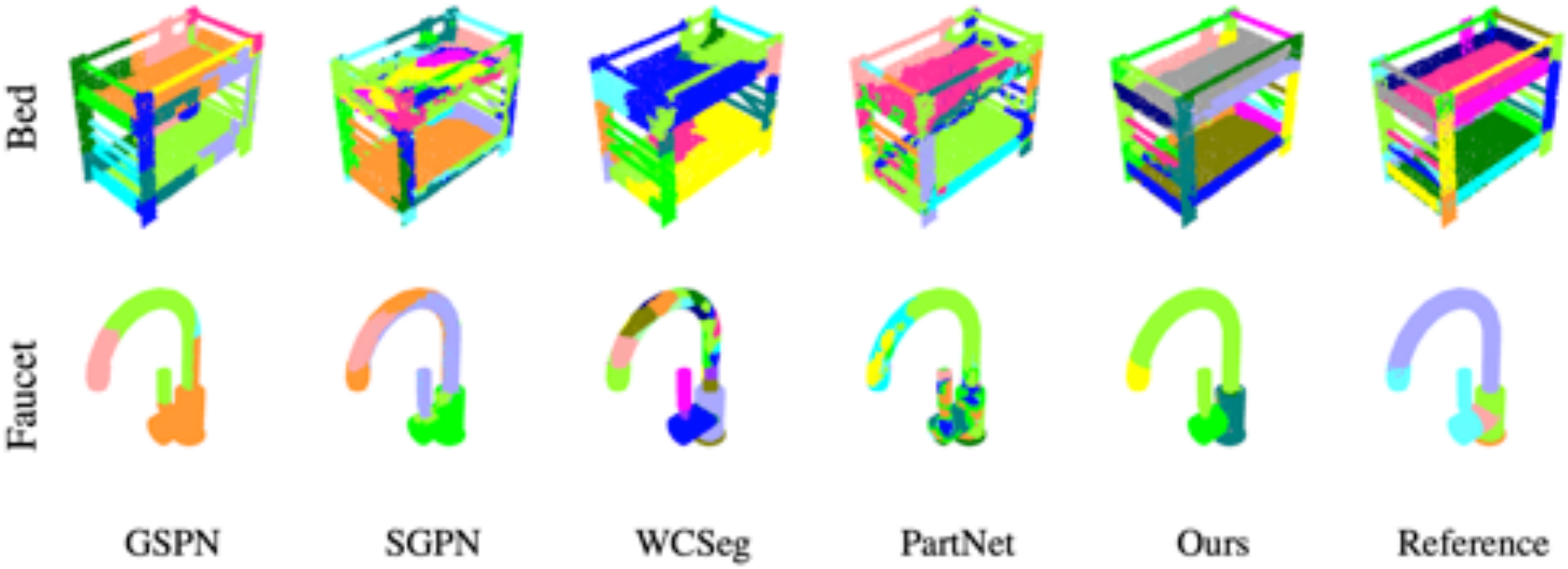
Our Approach



Our Approach



Qualitative Results



Train on chair, storage furniture, and lamp.
Test on bed and faucet, respectively.

Quantitative Results

	Seen Category					Unseen Category								
				Avg	WAvg									
PartNet	**55.3**	50.3	**23.4**	43	**47.4**	18.2	9.7	40.7	**73.5**	30.3	29.3	43.6	32.1	16.5
SGPN	42.2	44.2	11.5	32.6	36	21.4	7	46.7	53.3	27.7	8.7	34.8	28.9	25.5
GSPN	39.7	43.7	14.4	32.6	35	34.4	8.4	46.9	72.8	40.6	**40.6**	57.8	36.7	28.4
WCSeg	33.1	56.8	3.2	31	31.4	**41.9**	8.6	**56.3**	69.3	34.2	27.6	59.7	30.2	**37.3**
Our	50.6	**57**	21.7	**43.1**	45.6	41.6	**10.4**	49.2	72.2	**42.4**	31.2	**67**	**37.2**	33.1

	Unseen Category													
													Avg	WAvg
PartNet	16.6	**52.5**	0.4	33.6	82.1	29.6	33	25	0.8	38.9	12.2	36.8	31.2	35.7
SGPN	20	37	0.4	31	67.3	7.2	13.3	5.9	6.4	34.8	7.8	27.5	24.4	30.8
GSPN	25.3	31.7	0.4	18.9	92.9	**39.2**	40.6	26.4	3.7	34.6	12.7	41.4	35.1	34.7
WCSeg	**48.2**	48.7	0.3	**60.1**	64.8	30.8	46	19.5	**39**	31.4	12.3	29	37.9	33.5
Ours	30.9	34.1	0.4	44.1	**96.6**	34.3	**48.2**	**26.6**	16.7	**44.1**	**13**	**43.1**	**38.9**	**42.1**

Train on chair, storage furniture, and lamp.
Test on both seen categories and unseen categories.
Number is the average recall.

UC San Diego

Sampled Research Work (IV)

Environment For End-to-End Learning & Evaluation of Interaction Tasks

SAPIEN: A SimulAted Part-based Interactive ENvironment,
Xiang et al. CVPR 2020 (oral)

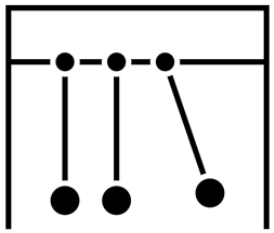
An Accessible Platform to Explore Object Manipulation Problems

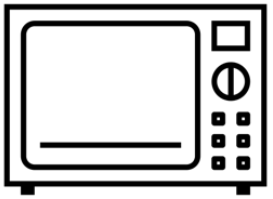
- Real robot/experiments are costly
- When it comes robotics planning/execution
  - Time: cannot speed up real-world physics
  - Cost: costly to maintain hardware
  - Hardware stability: hard to reproduce experiments
  - Safety
- Alternative: Simulation

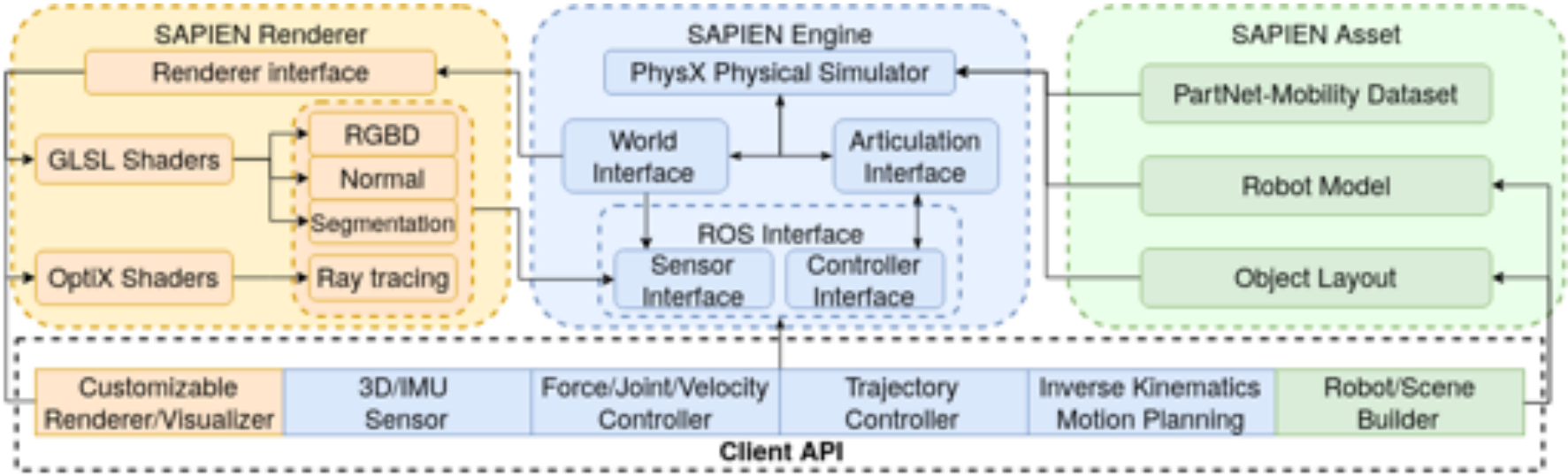
SAPIEN

Xiang et al., "**SAPIEN: A SimulAted Part-based Interactive ENvironment**", CVPR 2020

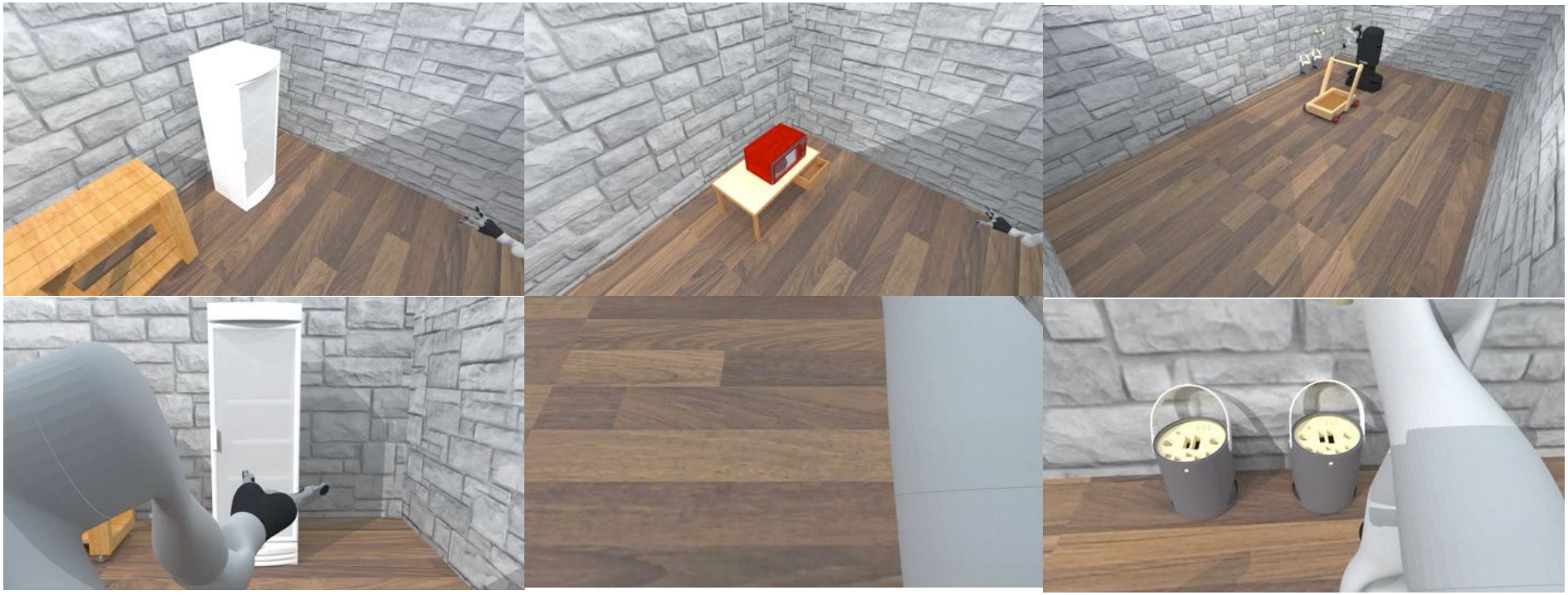
SAPIEN System







SAPIEN

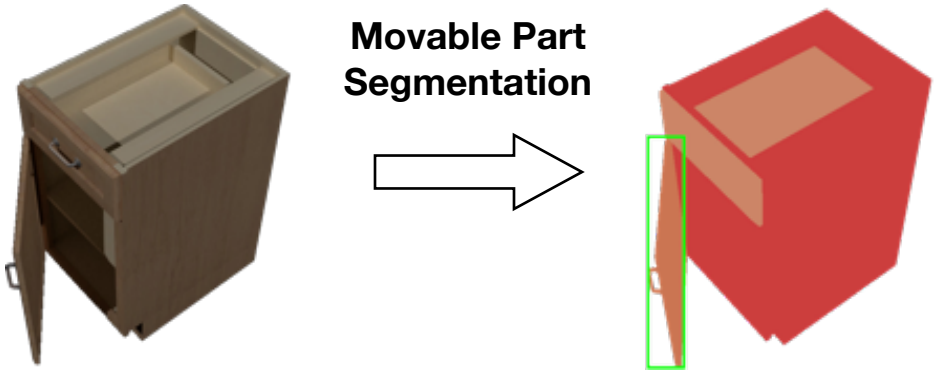


Xiang et al., "**SAPIEN: A SimulAted Part-based Interactive ENvironment**", CVPR 2020

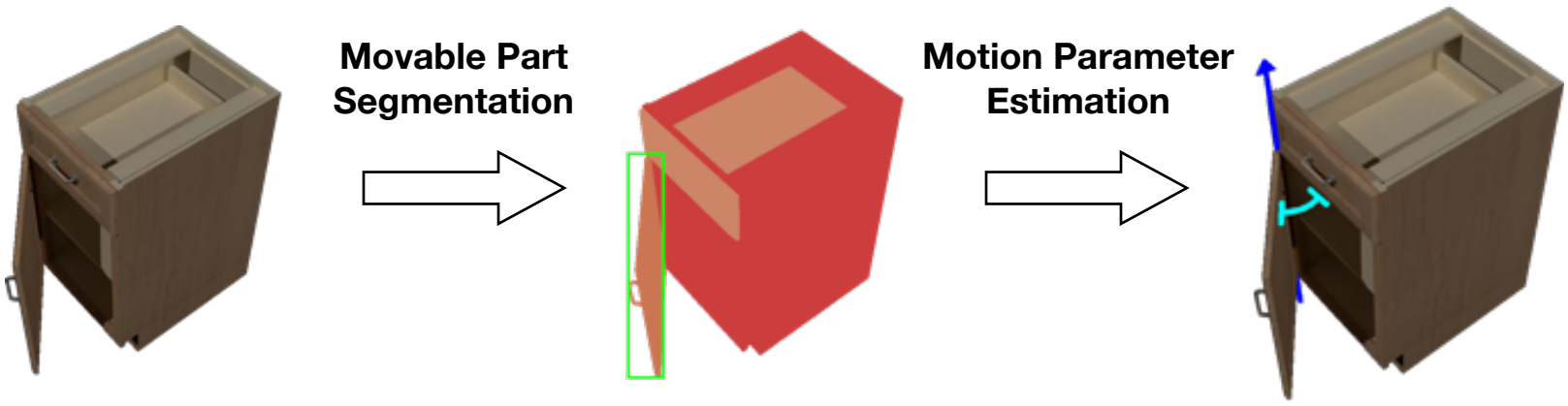
SAPIEN Asset
PartNet-Mobility Dataset

Task Demonstrations

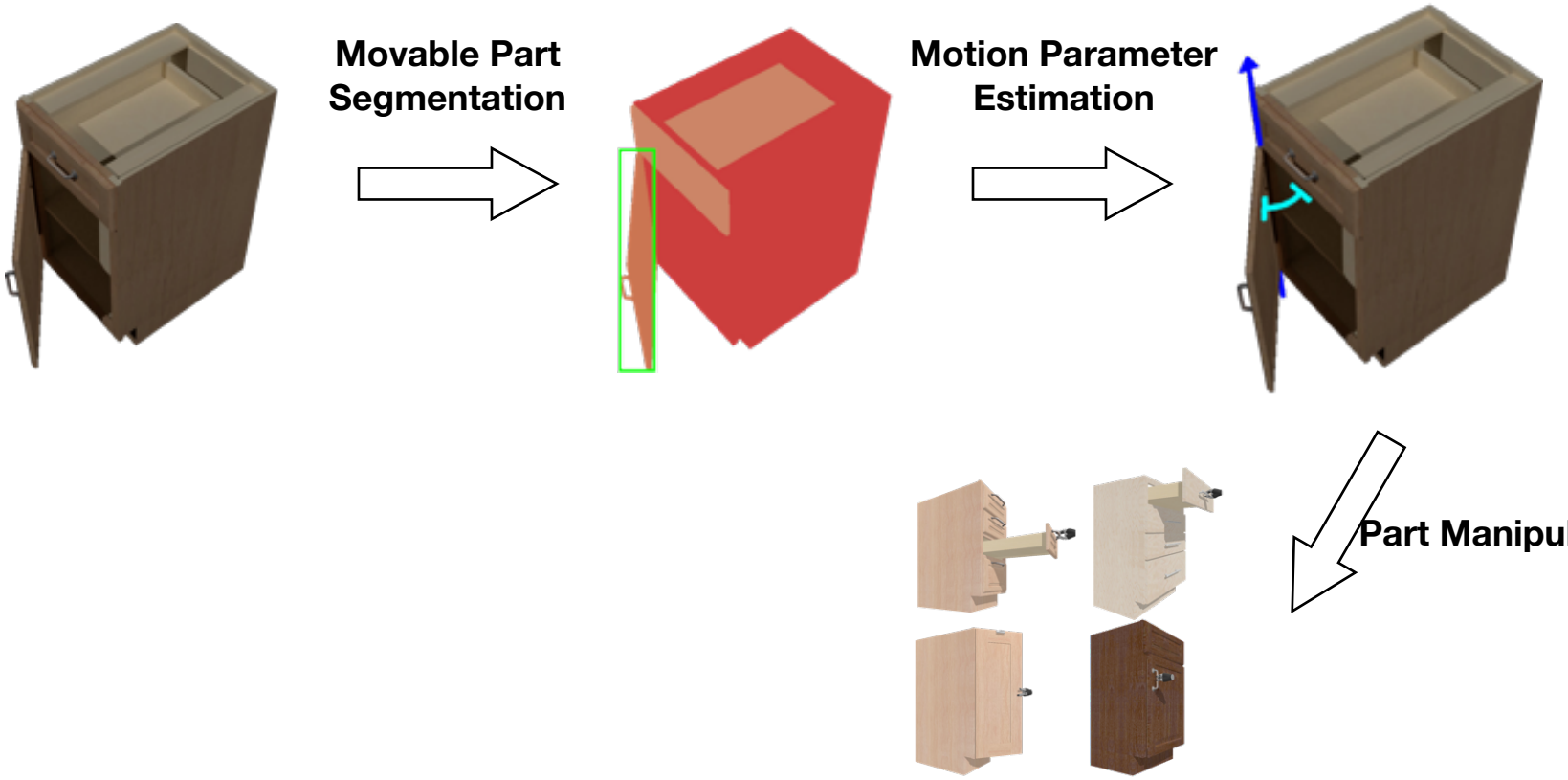
Task Demonstrations



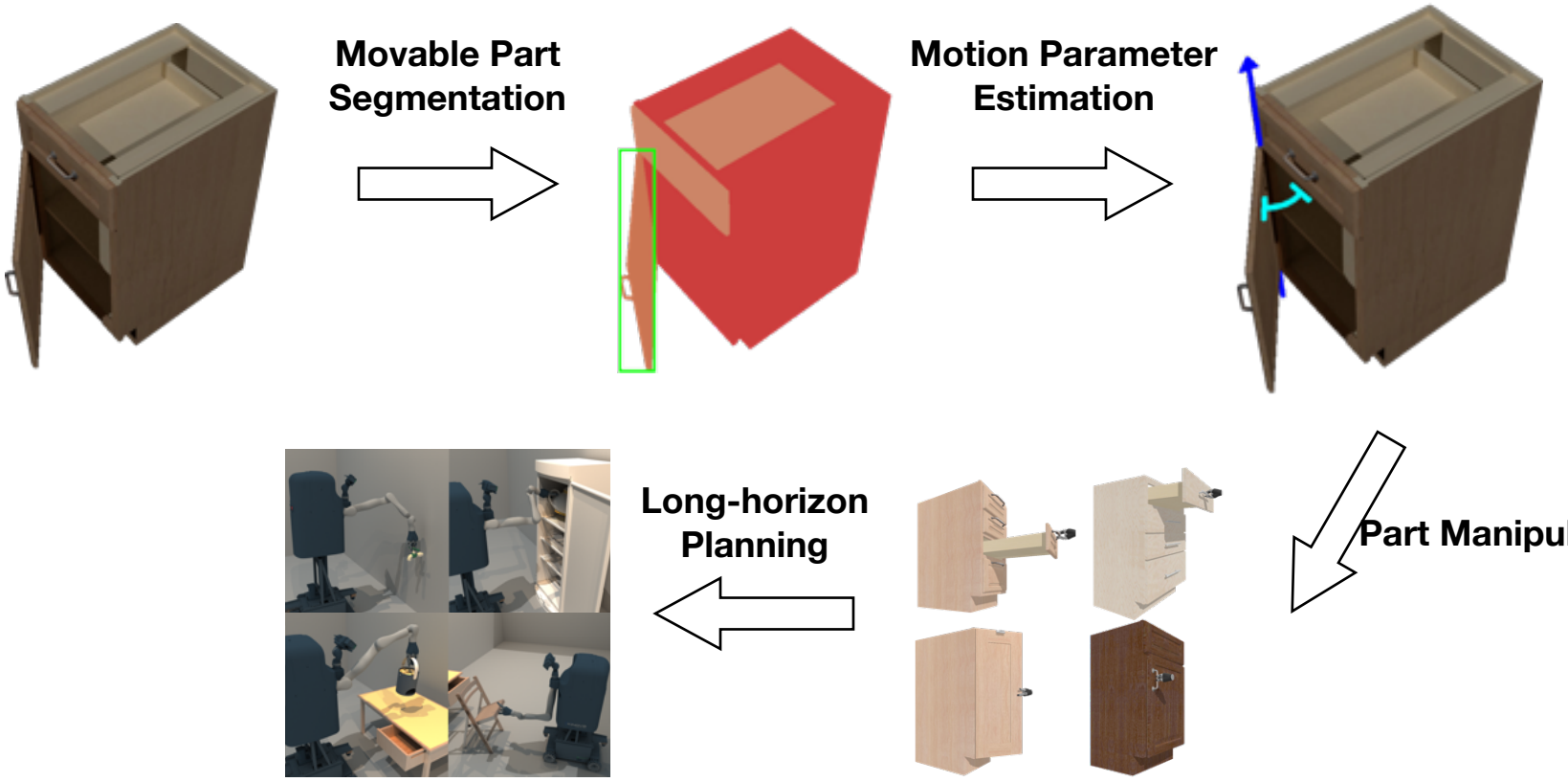
Task Demonstrations



Task Demonstrations



Task Demonstrations



Usage Information

- pip install sapien
- http://sapien.ucsd.edu
- **SAPIEN Challenge** to come later in the year

Conclusion

We still have a long way to go to develop really useful learning algorithms for building exploratory robots!

- Sensing, Representation, Composable Unit Discovery, …