UC San Diego

LEARNING TO MODEL THE ENVIRONMENT FOR INTERACTION

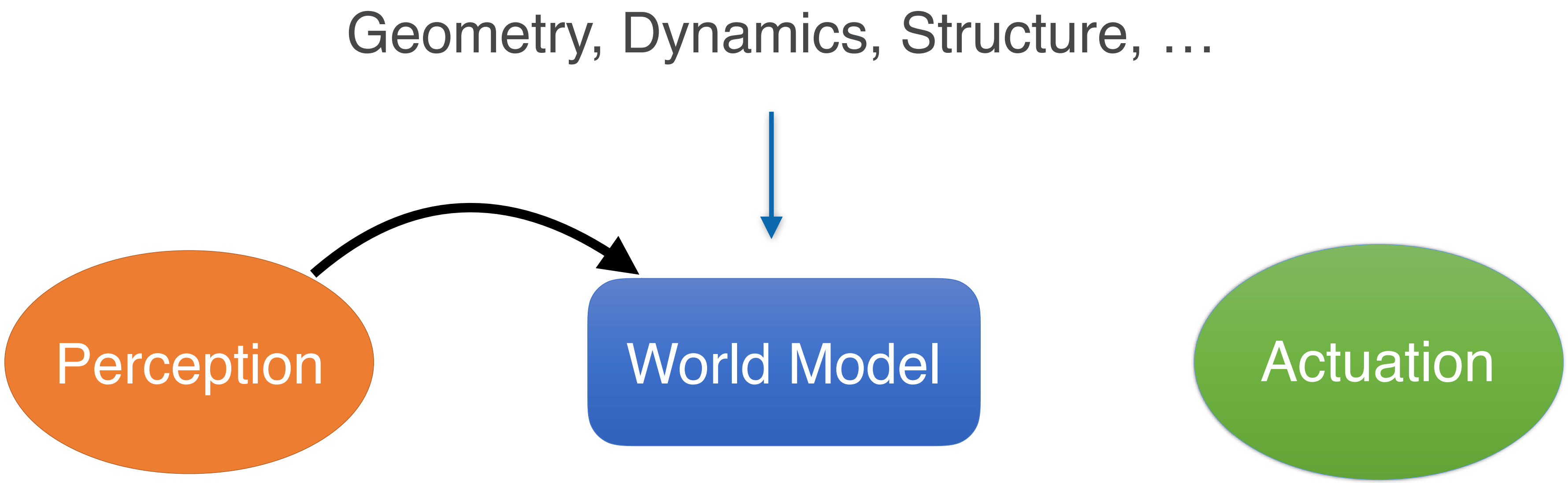
Hao Su

IROS Workshop on Perception and Grasping
Macau, China

Perception Models the Environment for Action

Perception Models the Environment for Action

Build Intelligent Agents to Live in Human Space

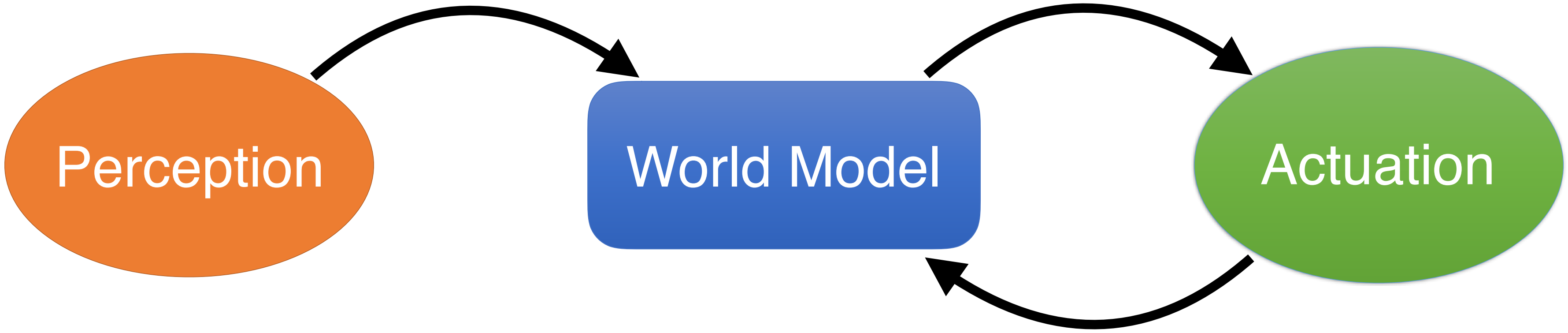


Build Intelligent Agents to Live in Human Space

Model-based Planning/Control

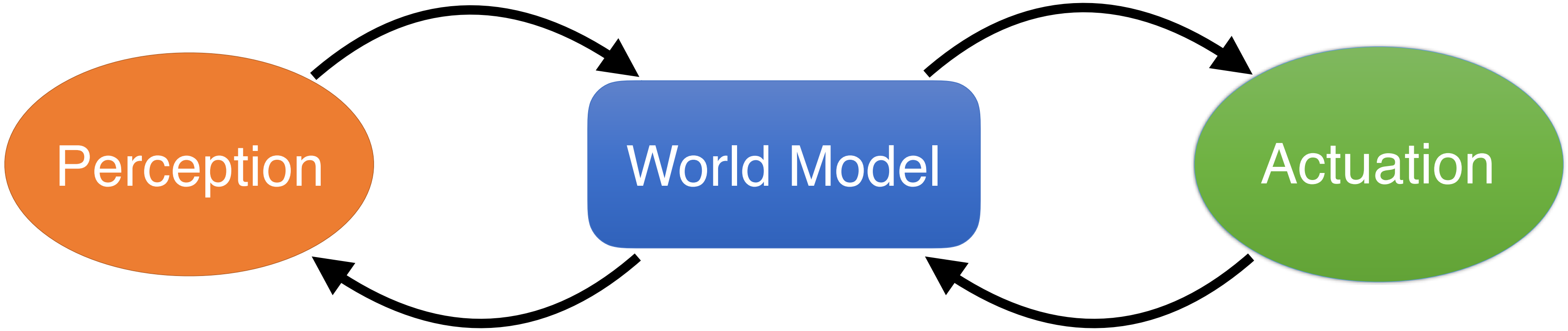
Perception → World Model → Actuation

Build Intelligent Agents to Live in Human Space



Emergence of Concepts

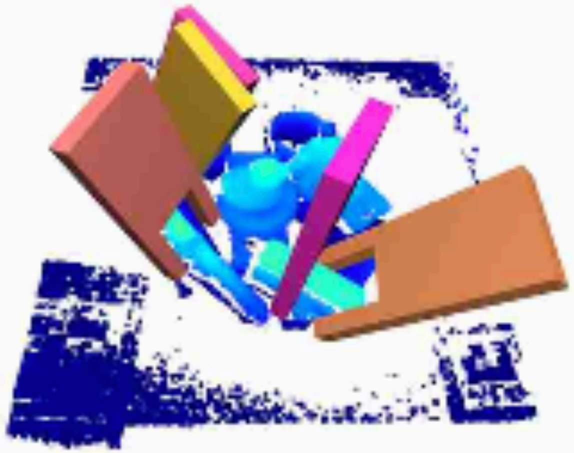
Build Intelligent Agents to Live in Human Space



Update of perceptron

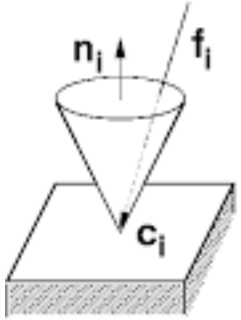
S^4G: Amodal Single-view Single-Shot $SE(3)$ Grasp Detection in Cluttered Scenes

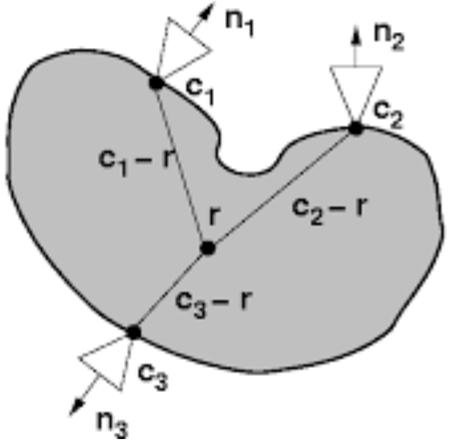
CoRL 2019

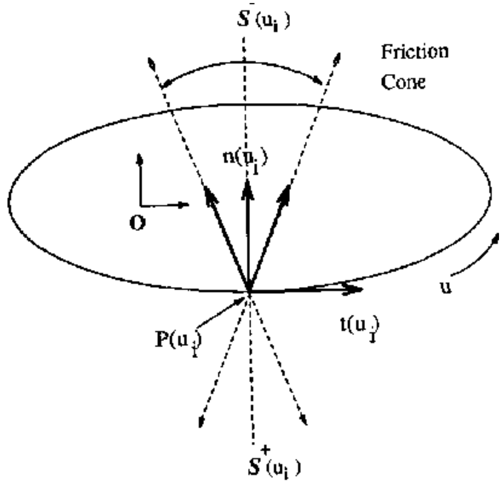


Robotics Grasping

- Most fundamental problem in robotics
- Serves as the initial step for other robot manipulation tasks
- E.g. open the door, use a hammer
- Analytical model of object grasping has already developed





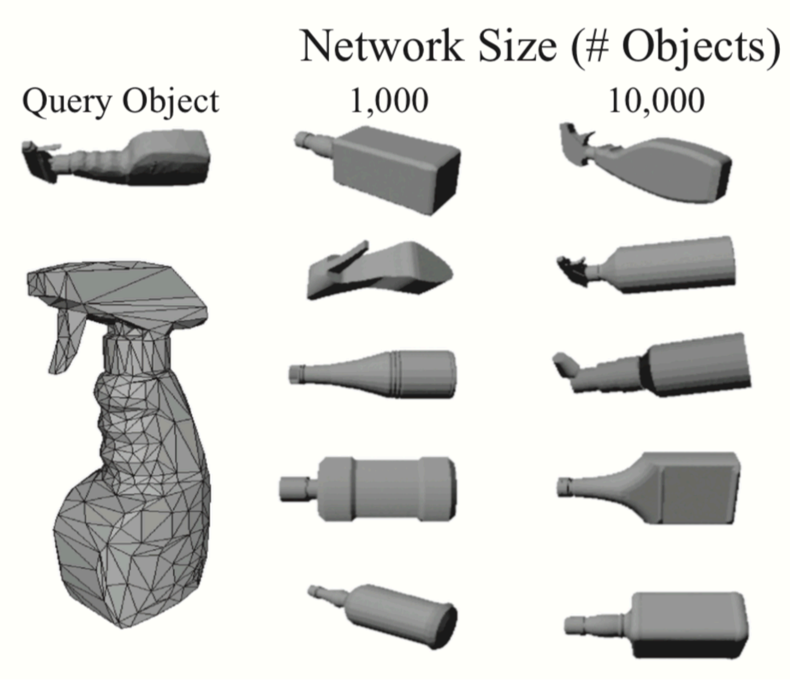


Nguyen, V. D. (1988). Constructing force-closure grasps. The International Journal of Robotics Research, 7(3), 3-16.

Classical Grasping

- However, classical method based on analytical model:
    - Needs detailed info about the object, e.g. complete geometry, friction, CAD

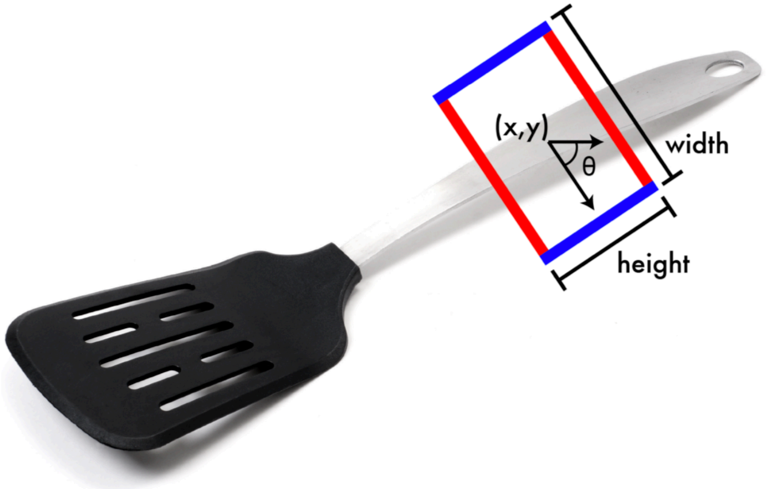
- Query based grasping
    - Built a database with pre-computed/labeled grasp
    - Match object with database, estimate the 6D pose

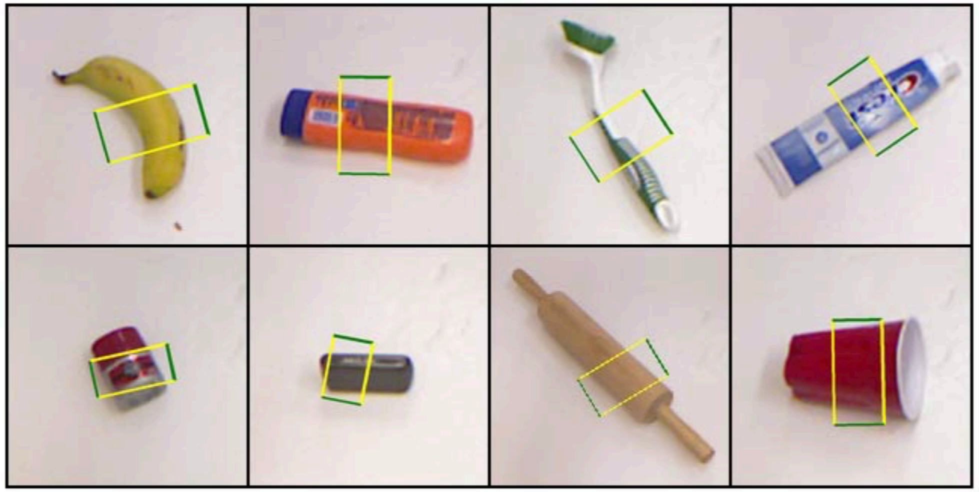


Mahler, J., Pokorny, F. T., Hou, B., Roderick, M., Laskey, M., Aubry, M., ... & Goldberg, K. (2016, May). Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In 2016 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1957-1964). IEEE.

Learning-based Grasping

- Can work for partially-observed geometry
- However, hard for human annotator to label full DOF ground-truth
- Limited to 3-4 DOF planar grasping for a long time

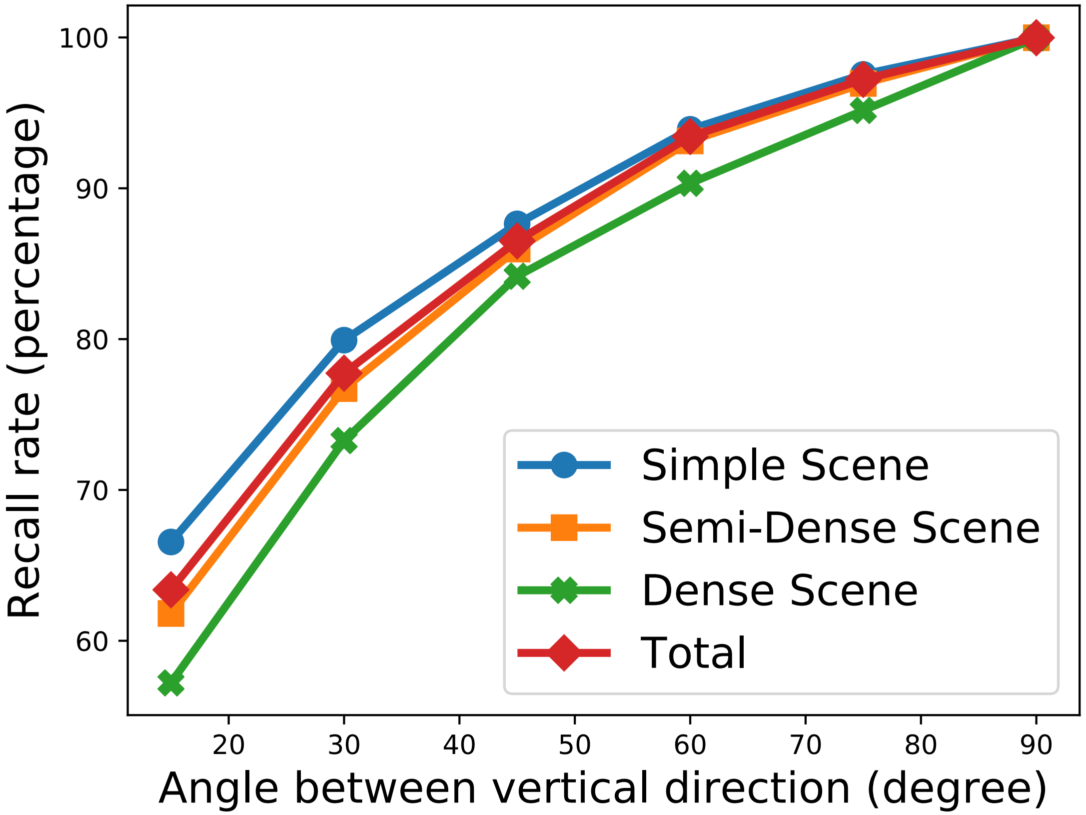




Industry assembly line, not domestic robot

SE(3) Grasp over 3/4 DoF Grasp

- Only 63.38% objects can be grasped by nearly vertical grasps (0°, 15°).



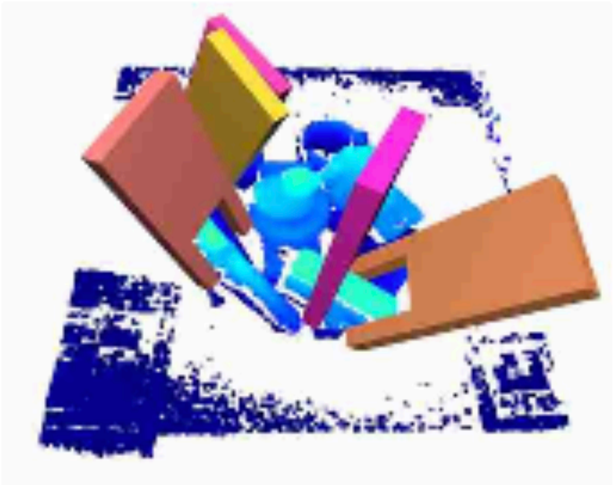
Classical Grasping Prediction: Sample-based

- Generate $SE(3)$ grasp from sampled point $c \in \mathscr{C}$
- Perform local search for antipodal grasp
- Using prior knowledge to remove naïve grasp
- Often use a Darboux frame to facilitate such search

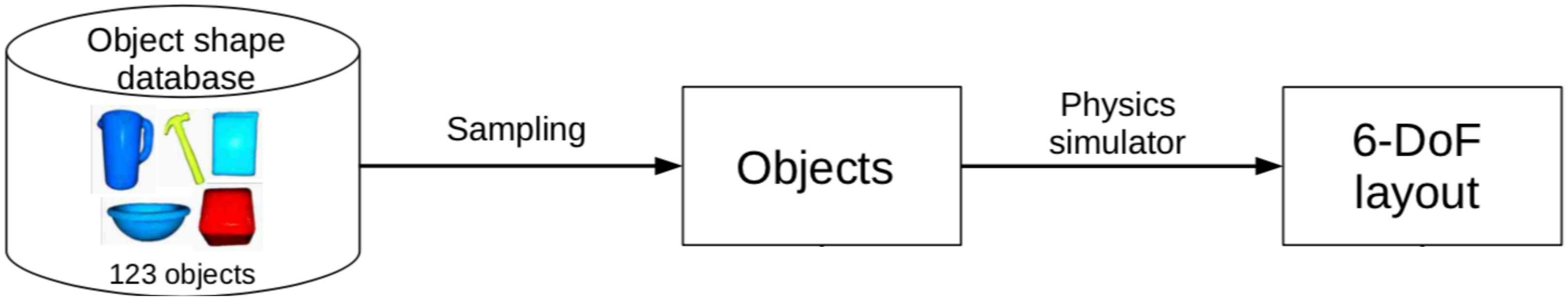
Costly, hard to sample from 6-D space!

Motivation

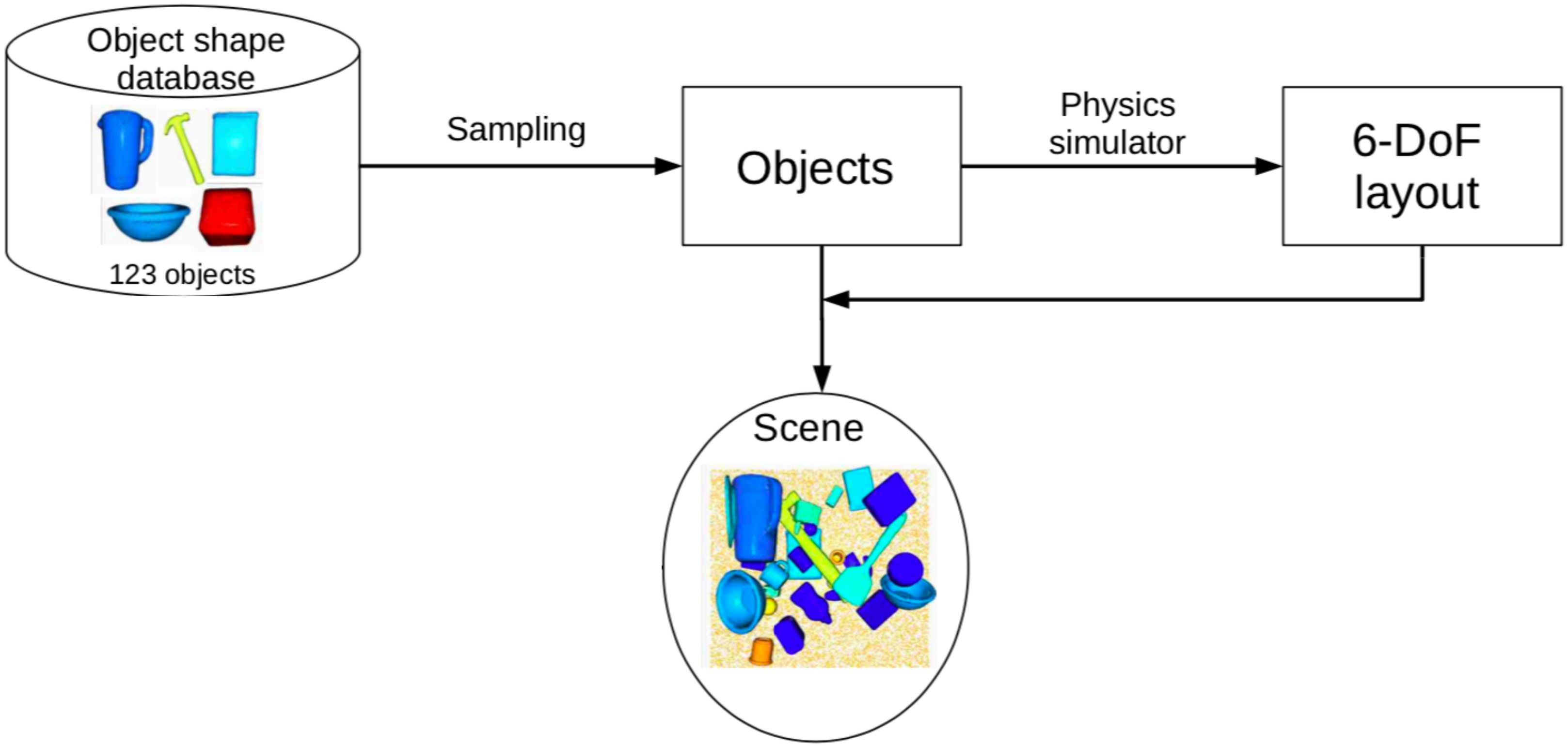
- Multi-view -> single-view
- Single object -> Whole scene
- Sampling -> Direct regression



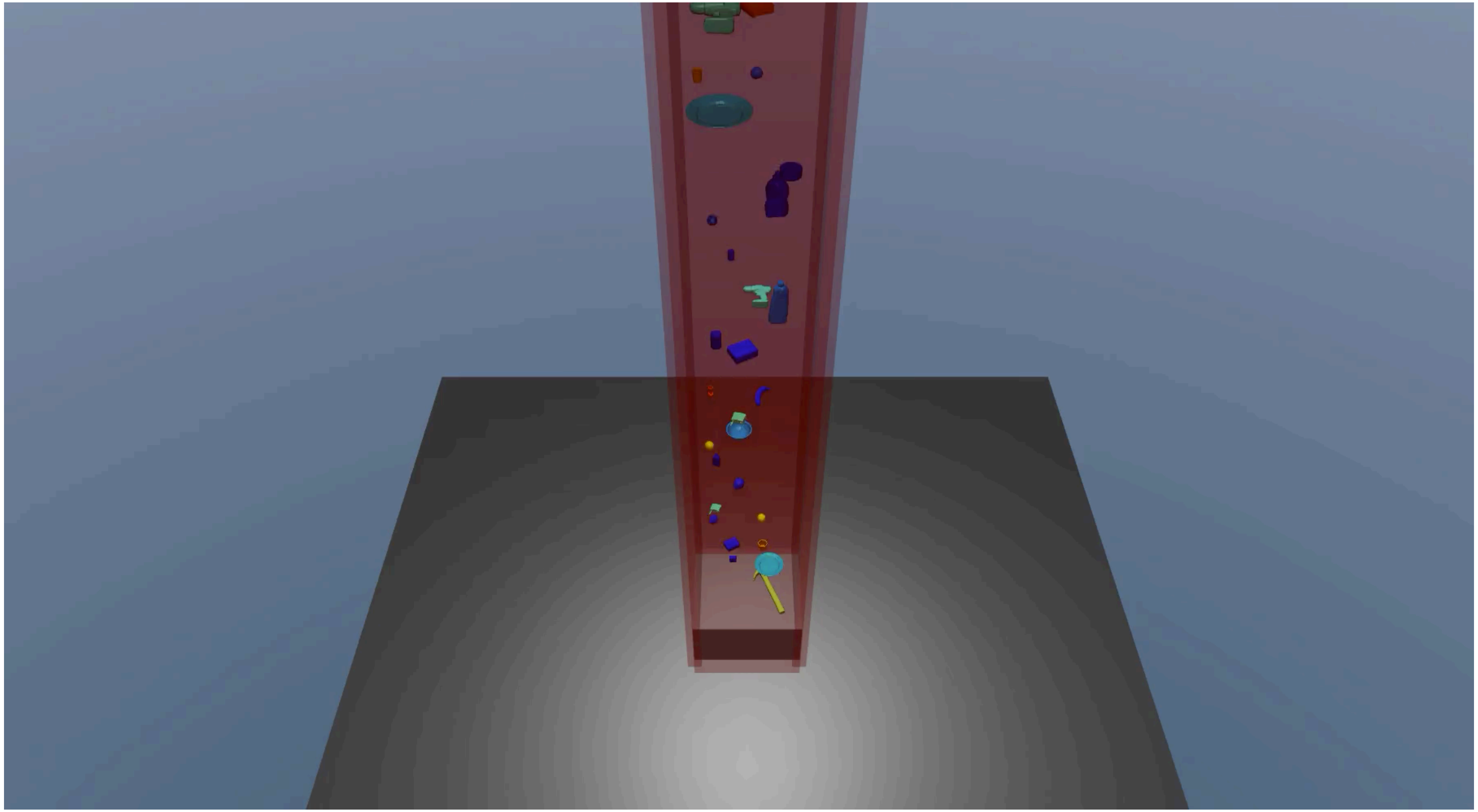
Generating Densely-labeled Training Data



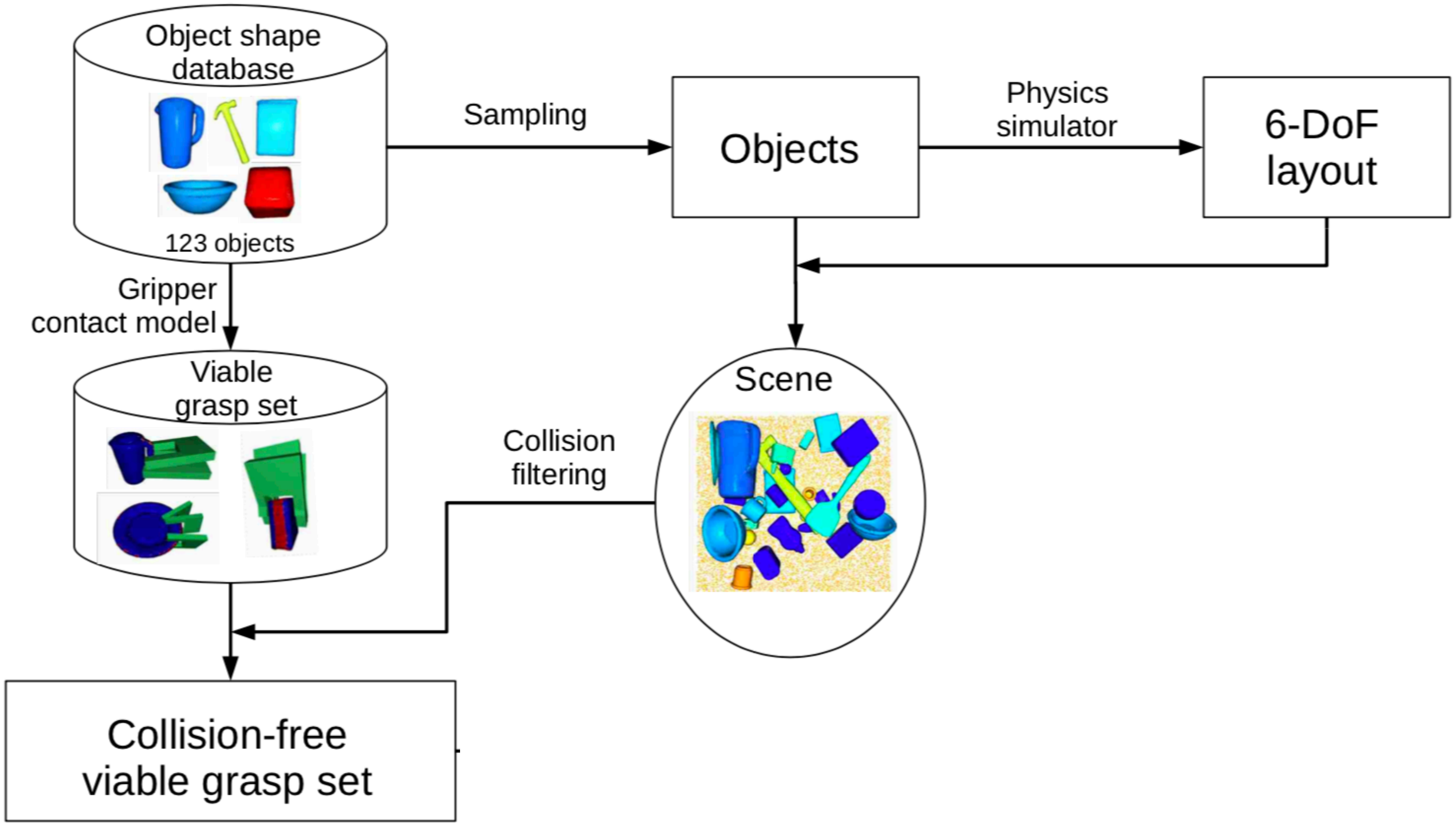
Generating Densely-labeled Training Data



Physically-plausible Scene Synthesis from Objects



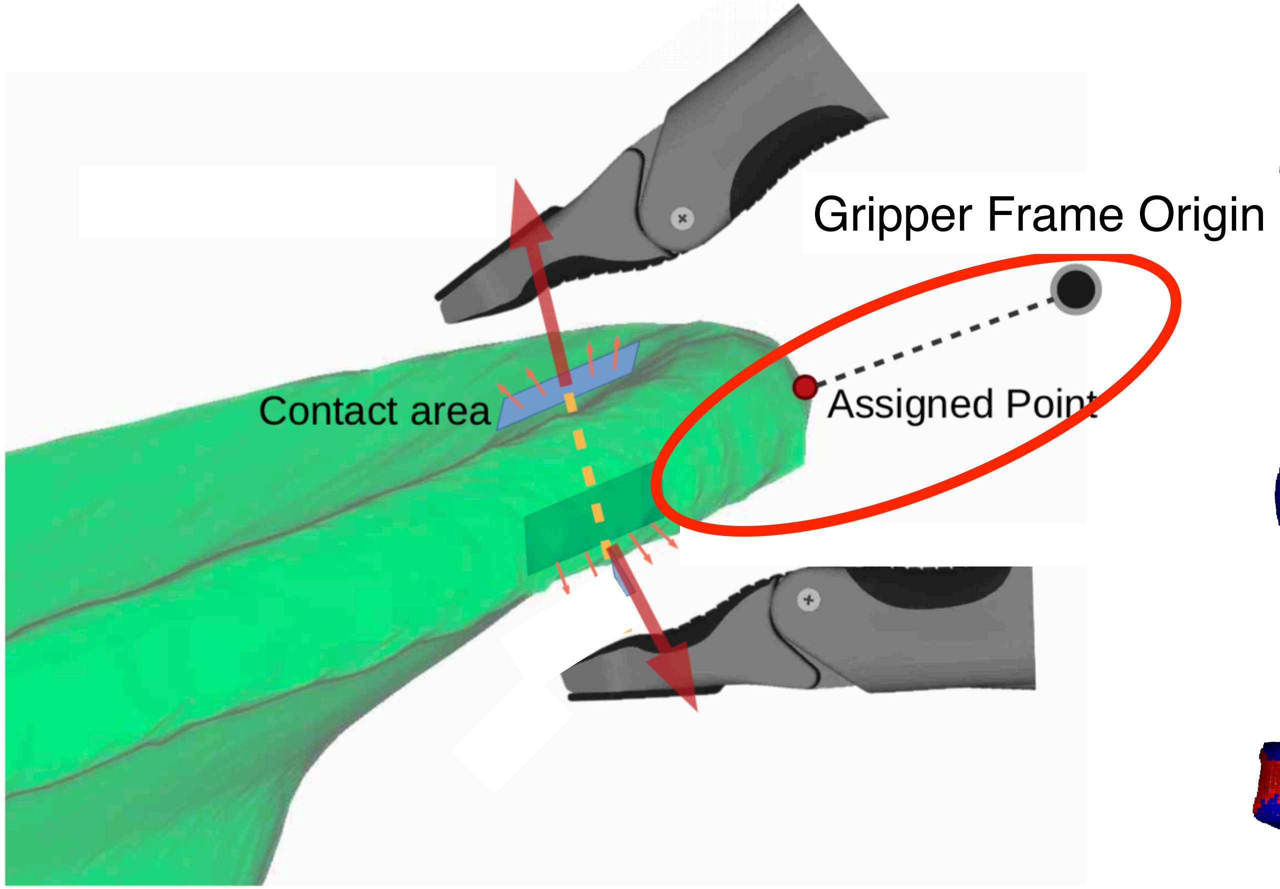
Generating Densely-labeled Training Data

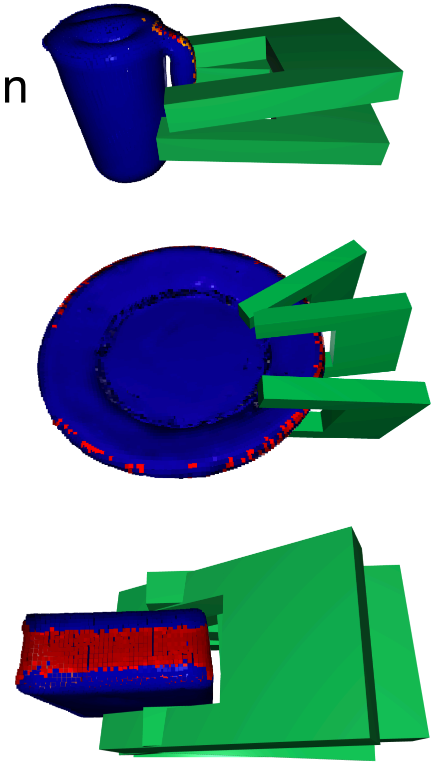


Store the Grasping Poses on the Surface

- Our frame: Sample contact points and check normal consistency

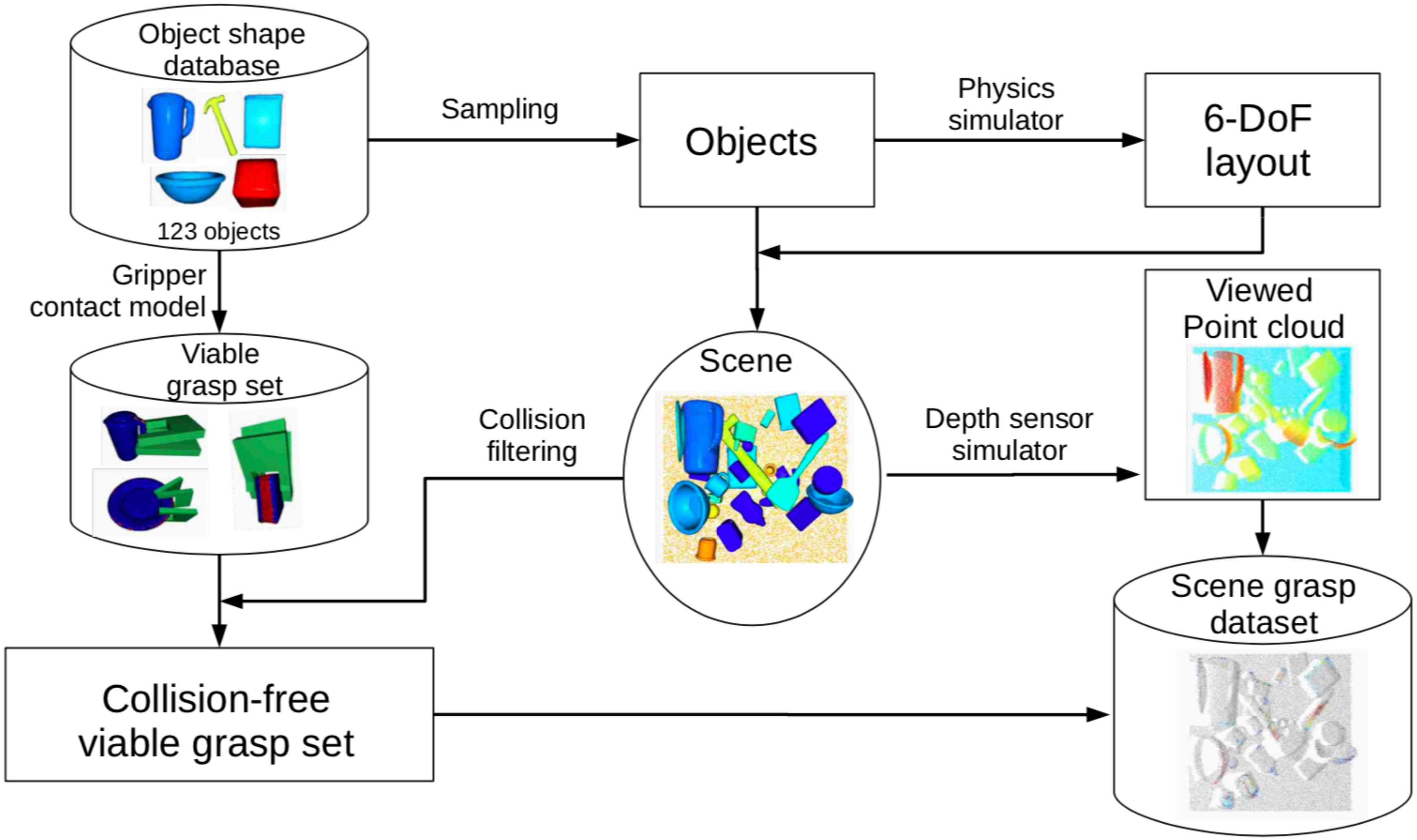
- Store on the points:
  - Rotation
  - Position
  - Score





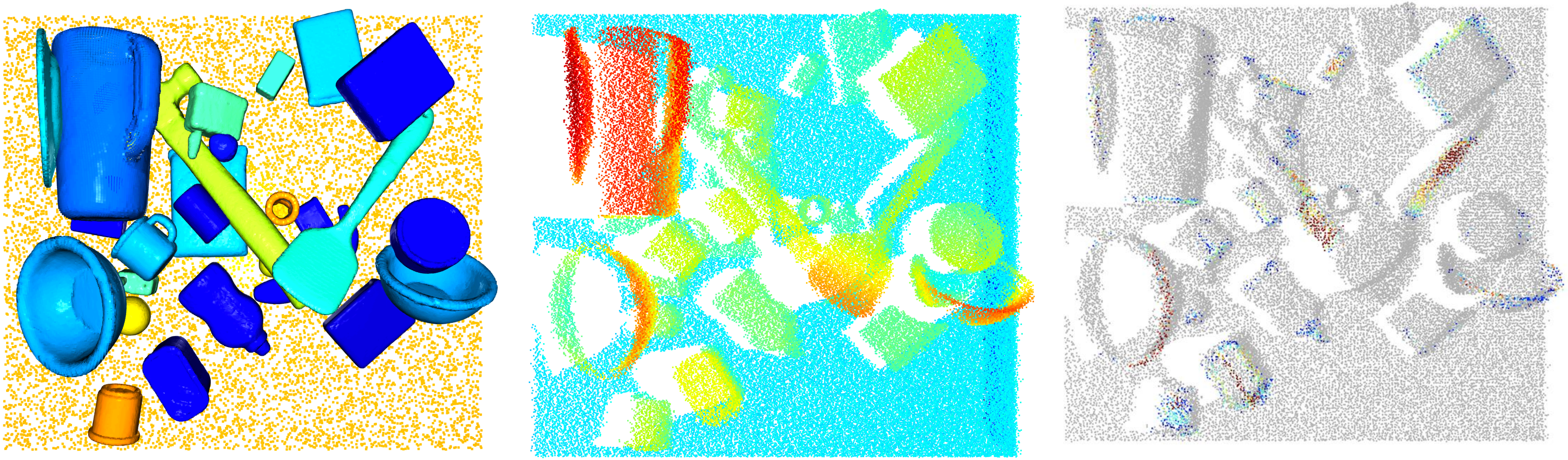
- Different from Darboux Frame, more suitable for **thin surfaces**

Generating Densely-labeled Training Data



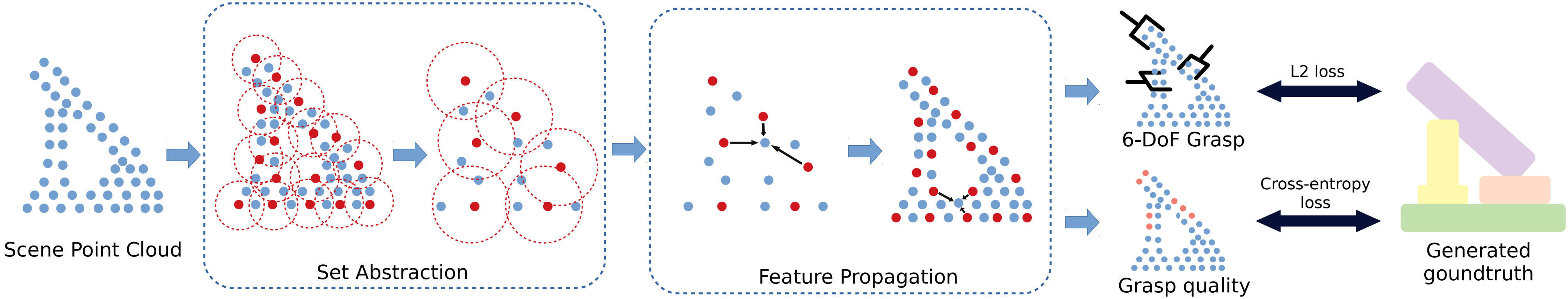
Scene-Level Collision and Robustness Evaluation

- From object-level grasp to scene level grasp
- Rendering noisy viewed point cloud as input for neural network
- Evaluate the quality metric under execution error



Grasp Proposal as Per-point Labeling

- Single-view
- Single-shot (v.s. sample-based)
- SE(3)



Qi, Charles Ruizhongtai, et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." *Advances in neural information processing systems*. 2017.

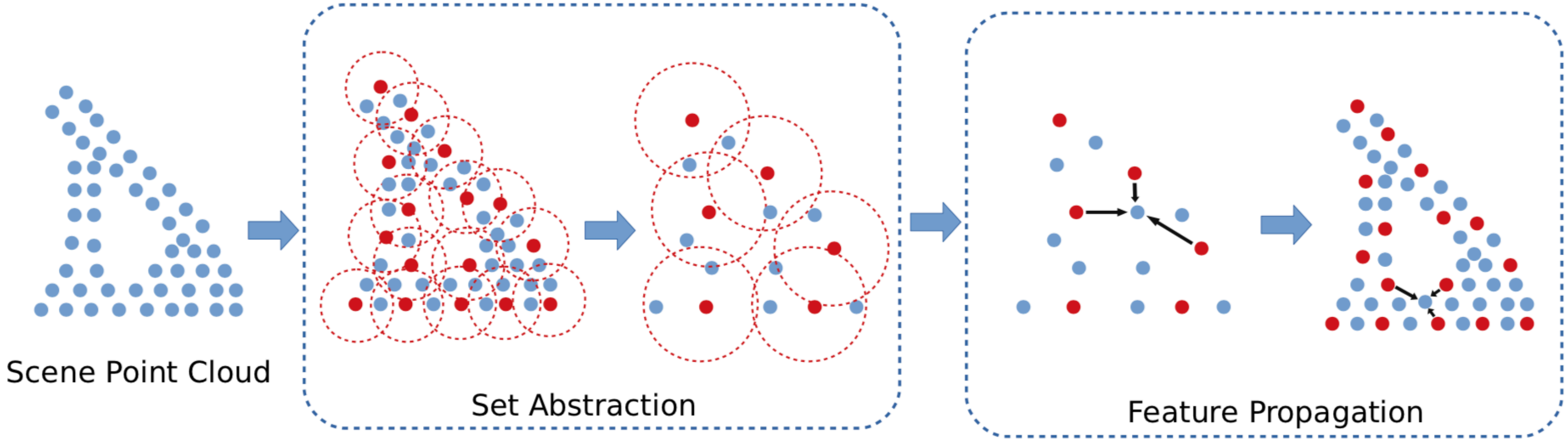
Rotation Representation for Regression

- Quaternion and Euler angle are discontinuous at certain point
- We regress 6D representation of rotation matrix with redundancy
- L2 loss for regression

$$\begin{matrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{matrix}$$

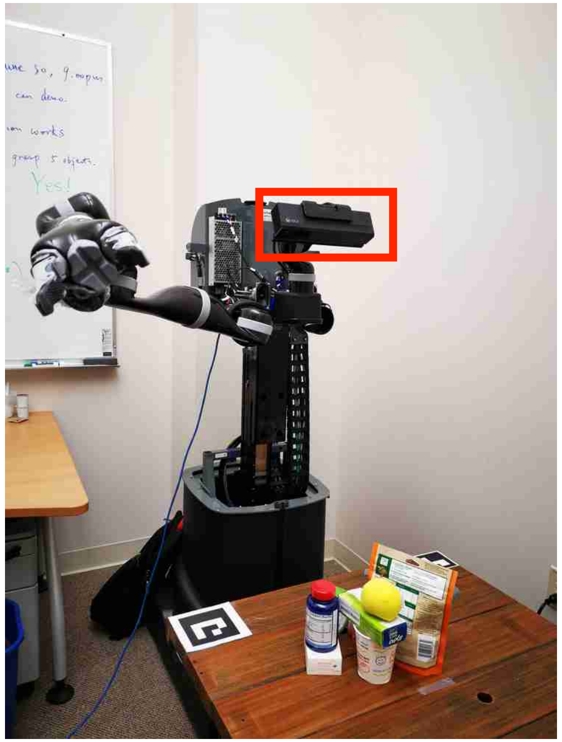
PointNet++ based Architecture

- Extracts hierarchical point set features
- Robust to partial and noisy observation
- Infer geometry relationship between objects in the scene.



Experiments

- Single-view depth from Kinect
- Zero-shot learning setting (novel test objects)



Experiments

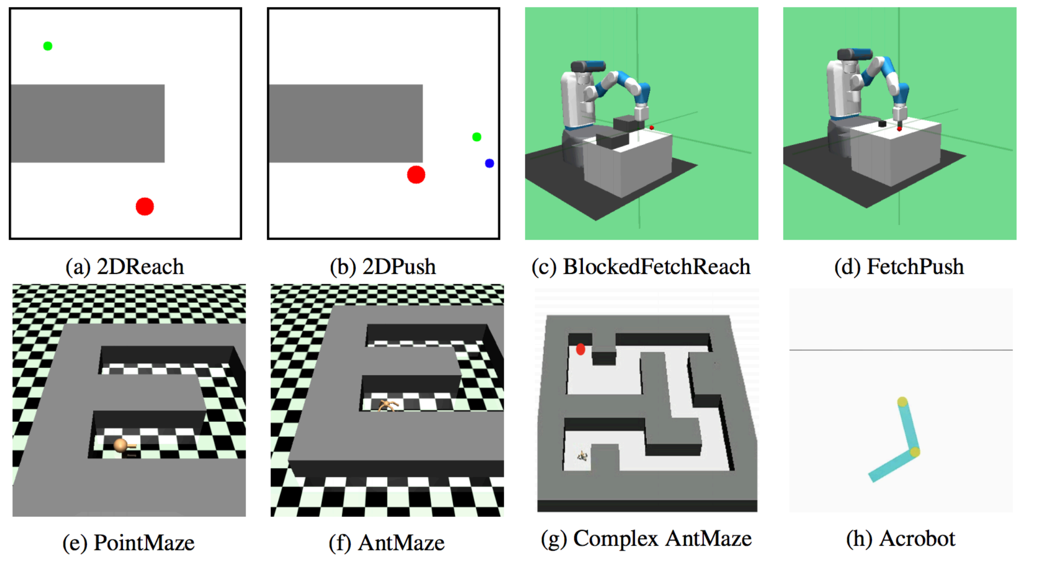
- Robotics experiments with cluttered scene
- 30 objects not present in the training data

Experimental

	Grasp quality		**Time-efficiency**		
	Success rate	Completion rate	Processing	Inference	**Total**
GPD (3 channels)	40.0%	60.0%	24106 ms	**1.50 ms**	24108 ms
GPD (12 channels)	33.3%	50.0%	27195ms	1.70ms	27197ms
PointNetGPD	40.0%	60.0%	17694ms	2.86ms	17697ms
Ours	**77.1%**	**92.5%**	**5804ms**	12.60 ms	**5817 ms**

Mapping State Space using Landmarks

NeurIPS 2019



Background: Universal Goal Reaching

- Learn a policy to reach given goals

- The agent will receive -1 penalty (reward) every time step, unless it reaches the given goal.
  - It's a sparse reward setting

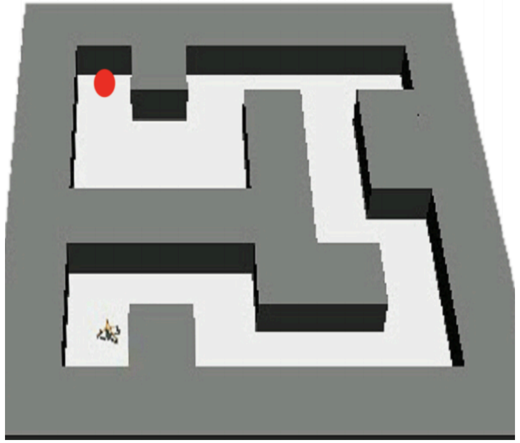
- Finding a shortest path on a graph?
  - Reward becomes negative shortest path distance (if no discount)

Universal Function Value Approximator

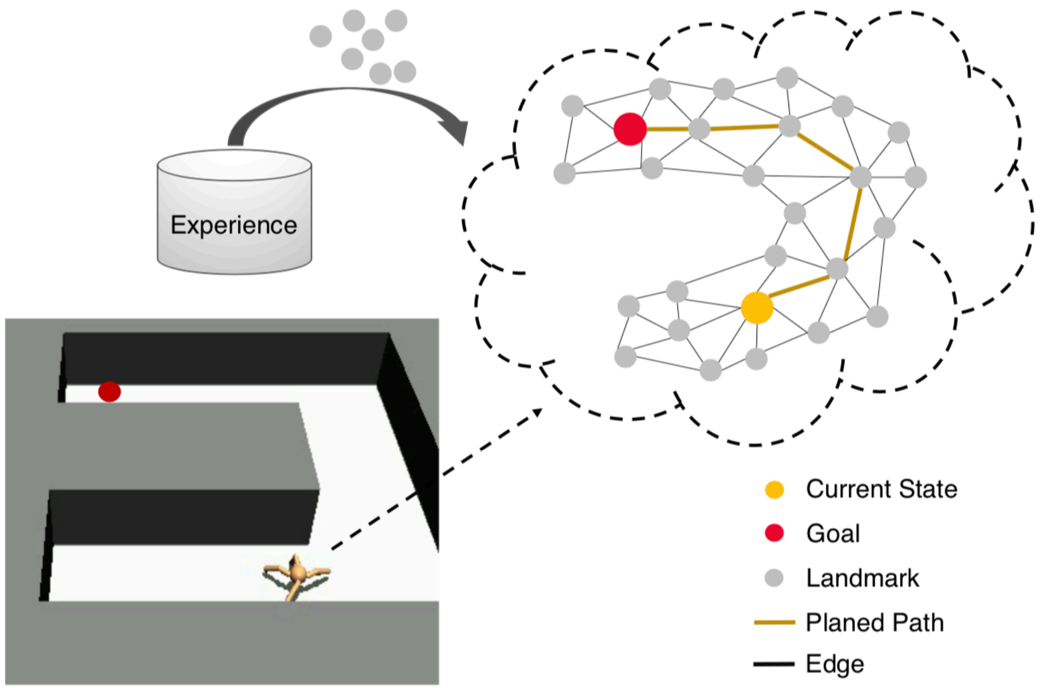
- $Q(s, a, g)$: goal-conditioned Q value
- Hindsight Experience Replay (HER) is the SOTA (baseline) for this problem

Long-horizon RL is Difficult

- Q(s, a, g) is not accurate if g is faraway from s
    - Reason 1: The number of state-goal pairs increases quadratically while the network capacity is limited.
    - Reason 2: if $(s,\ g)$ is some unseen pairs, long-term extrapolation is not reliable (take the maze as an example).
    - …

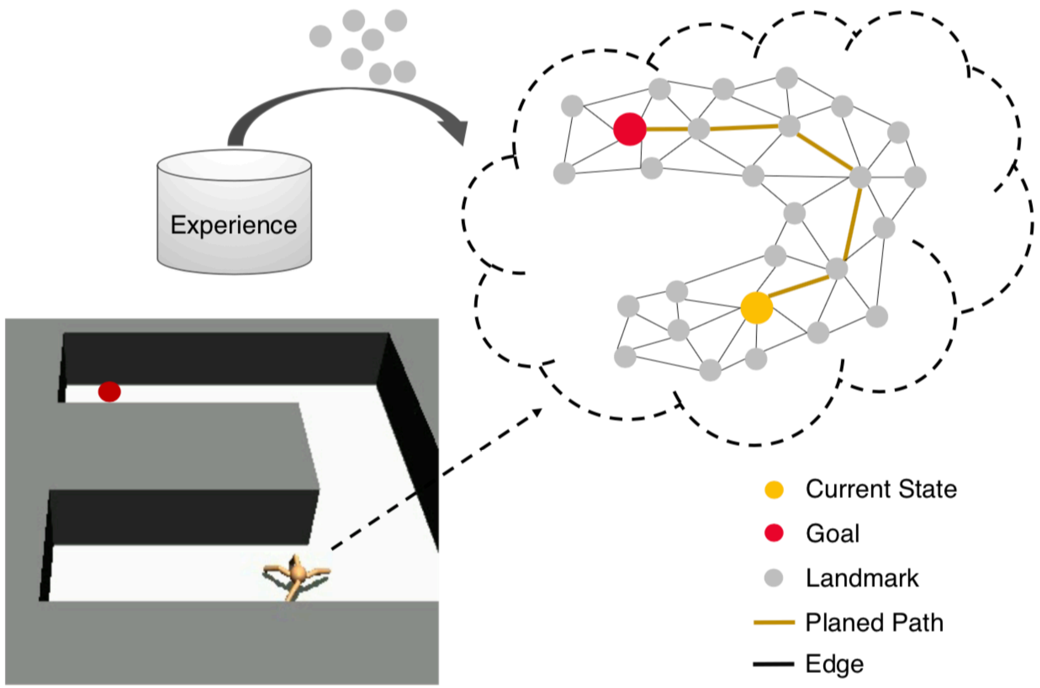


Our Approach: Planning with a Landmark-based Map



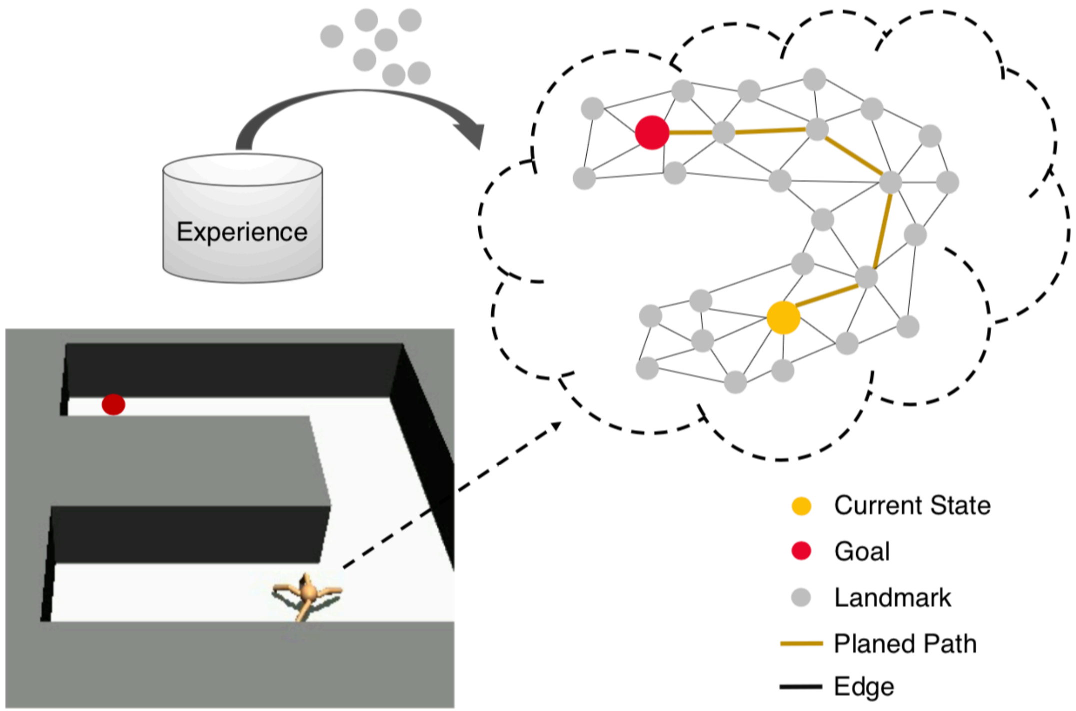
- View the MDP as a graph
- Sample Landmarks
- Build the Graph
- Planning

Landmark Sampling



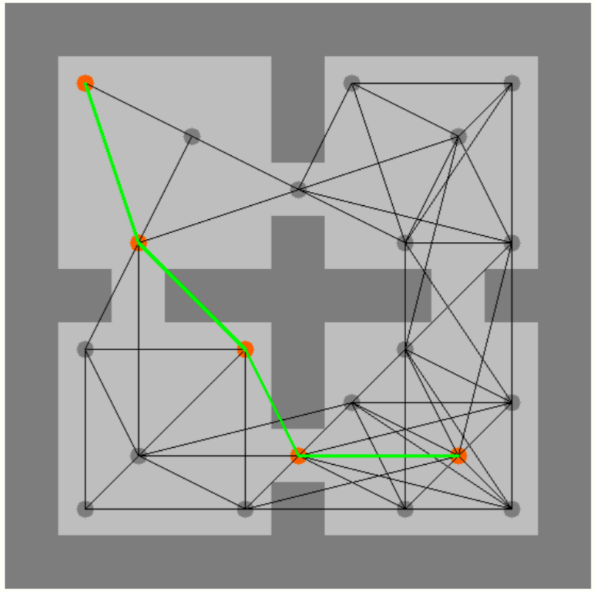
- **Replay buffer**: stores the transition collected by HER
- **Landmarks Sampling**: Using **farthest point sampling (based on Q)** from the replay buffer
- A compressed state space representation
- Encourages exploration to boundary states

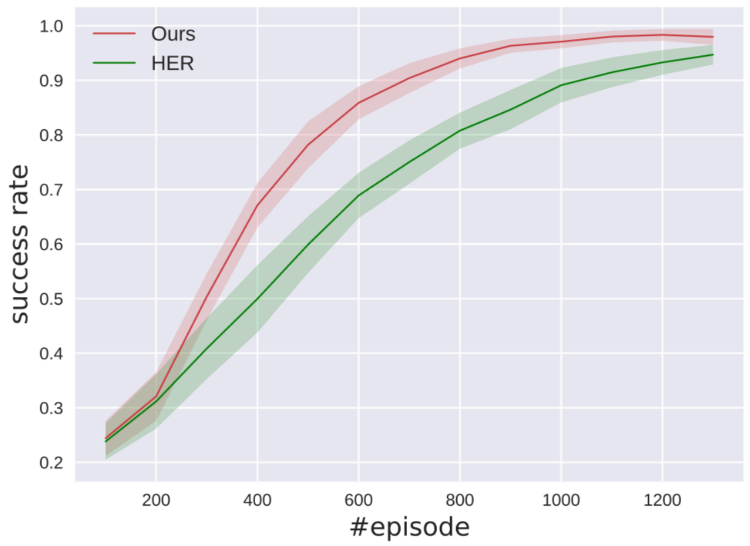
Planning



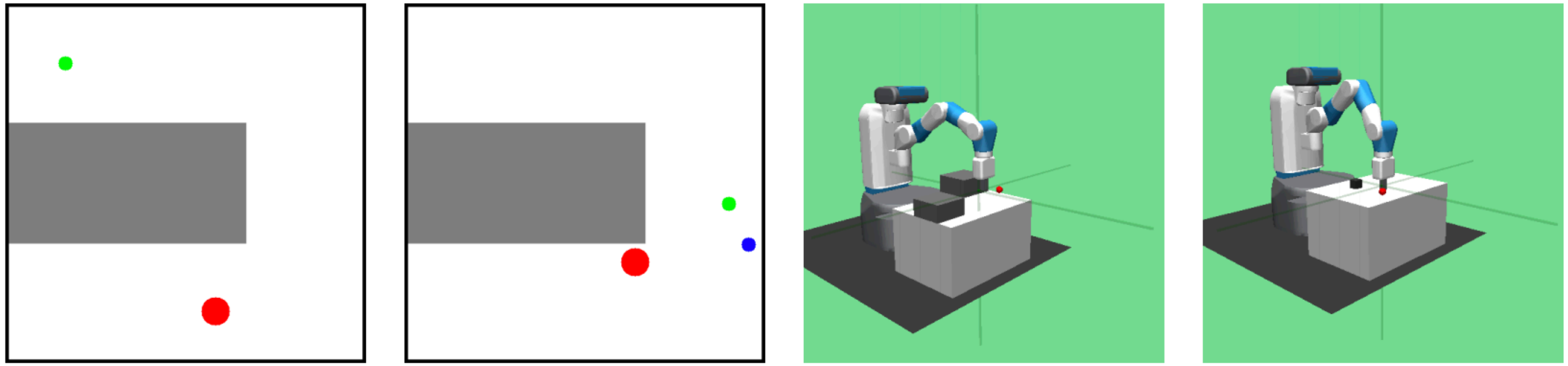
- Add the current state s and goal into the graph
- Find shortest path using Bellman-ford algorithm
- Find the next landmark l_i in the path
- Use HER to generate a local policy $\pi(s, l_i)$. (DDPG or DQN)

Experiments: FourRoom

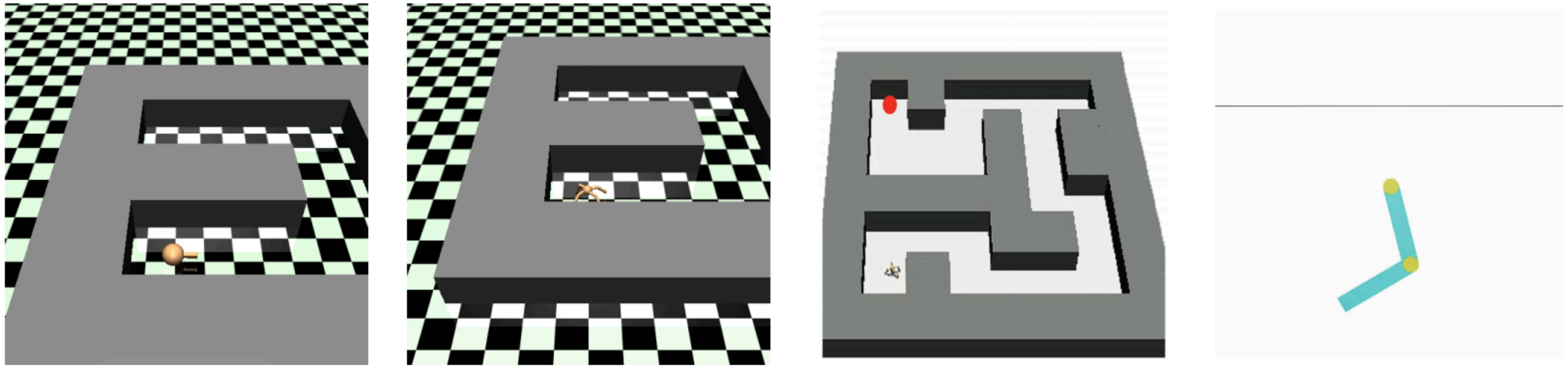




Experiments: Continuous Control task

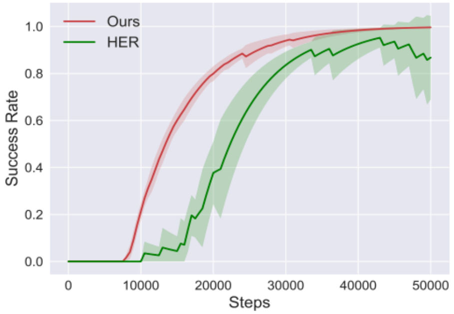


(a) 2DReach (b) 2DPush (c) BlockedFetchReach (d) FetchPush

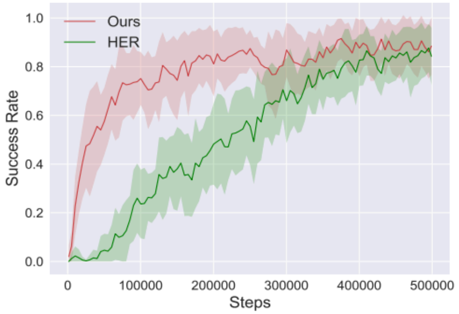


(e) PointMaze (f) AntMaze (g) Complex AntMaze (h) Acrobot

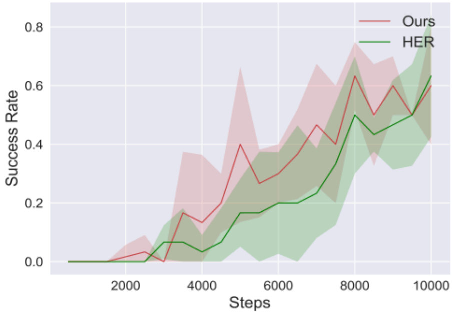
Experiments: Continuous Control task



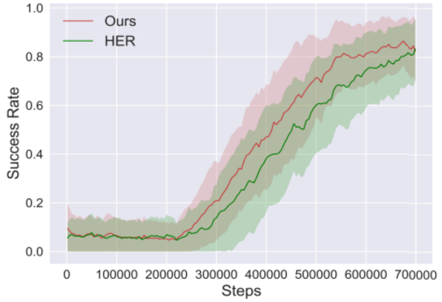
(a) 2DReach



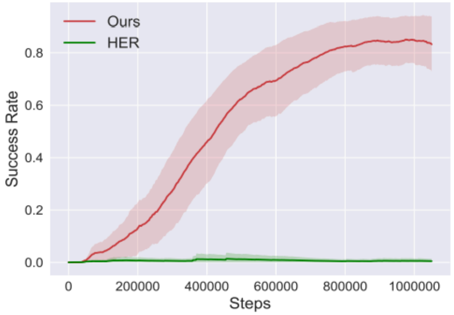
(b) 2DPush



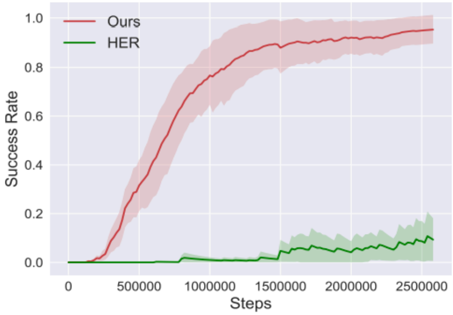
(c) BlockedFetchReach



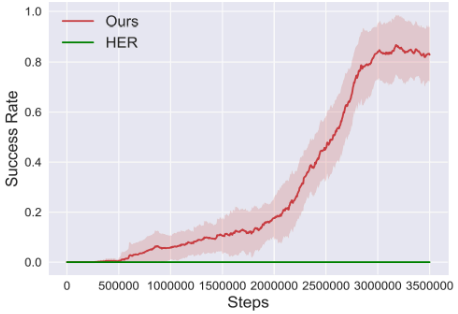
(d) FetchPush



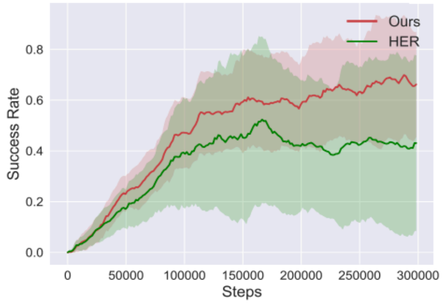
(e) PointMaze



(f) AntMaze



(g) Complex AntMaze



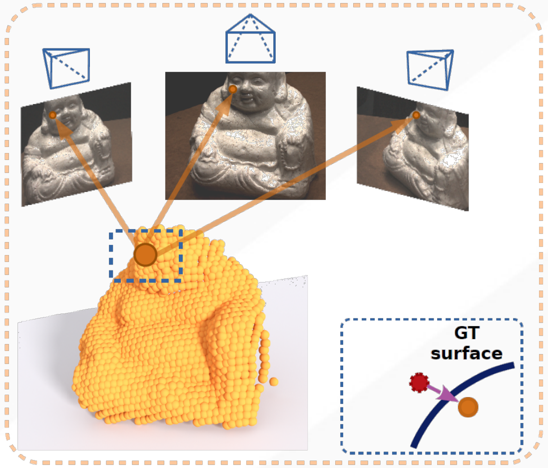
(h) Acrobot

Sum up

- Pure model-free RL is not good for learning long-horizon actions
- Decouple planning and local polices (network-based policy)
- Landmark-based map helps planning and exploration

Point-based Multi-View Stereo Network

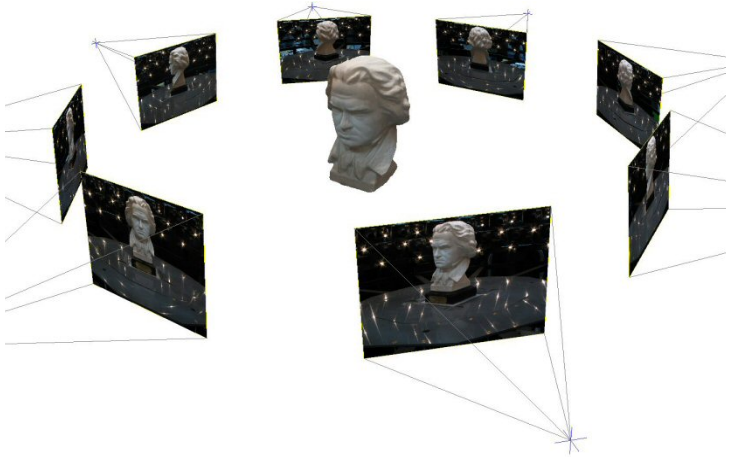
ICCV 2019 (oral)



Multi-view Stereo (MVS)

Target:

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



Learning-based MVS

Learning-based MVS

Learned feature ➡ more robust matching

Learning-based MVS

Learned feature ➡ more robust matching

Shape prior ➡ more complete reconstruction

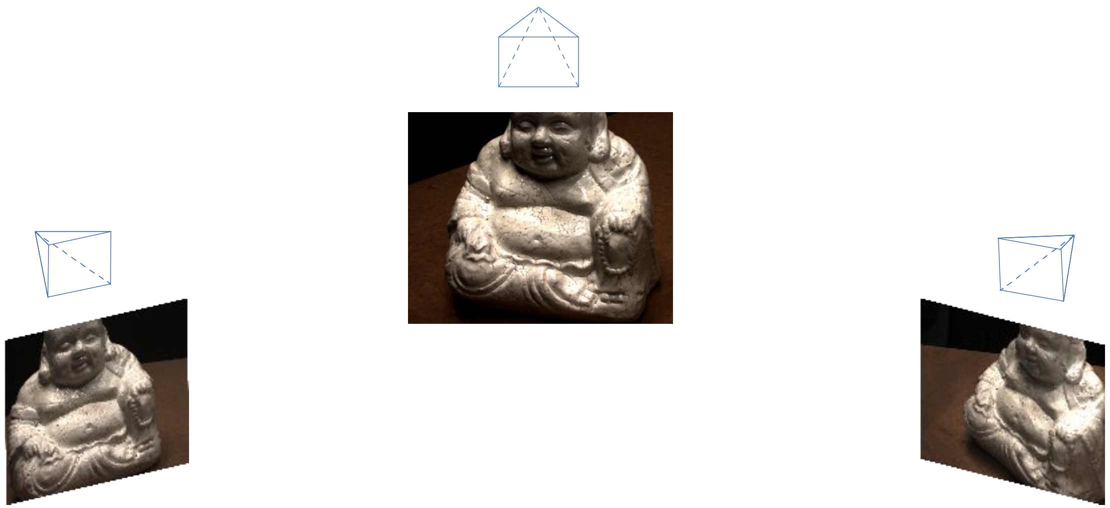
Learning-based MVS

Learned feature ➡ more robust matching

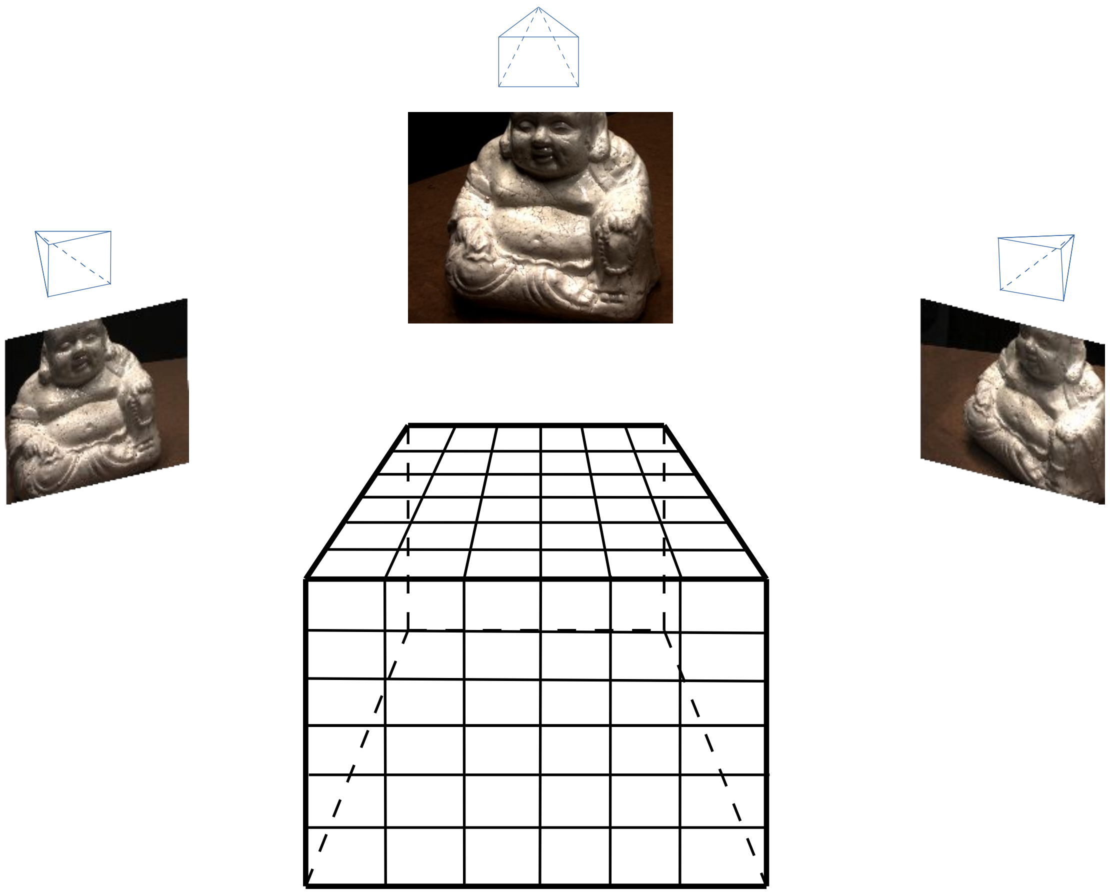
Shape prior ➡ more complete reconstruction

Key Component: 3D Cost Volume

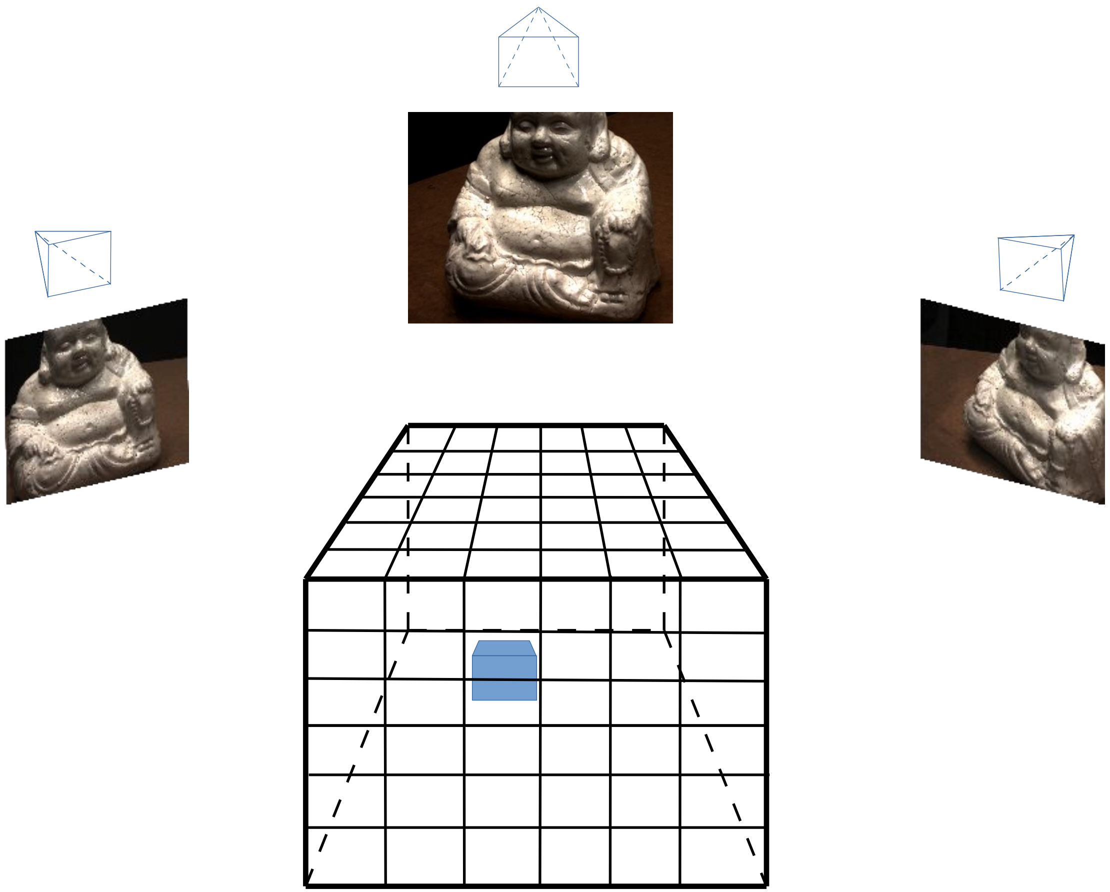
Cost-volume based methods



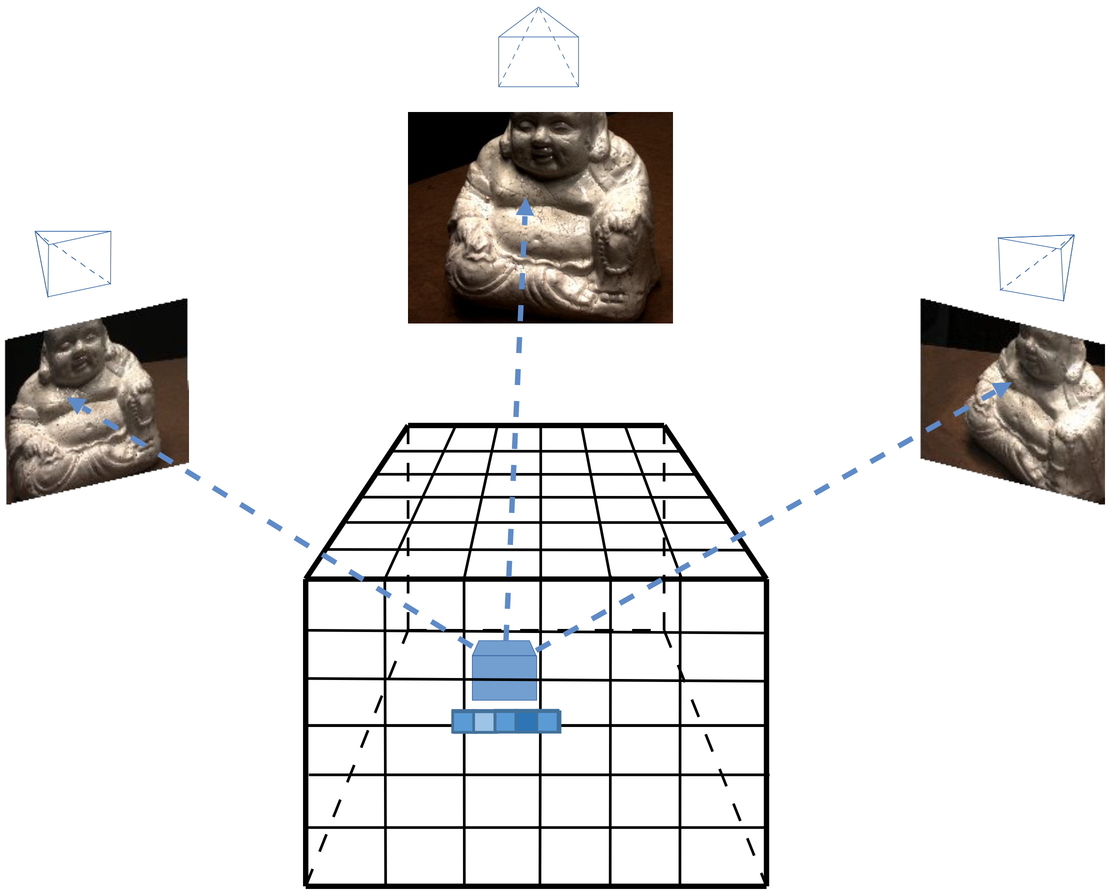
Cost-volume based methods



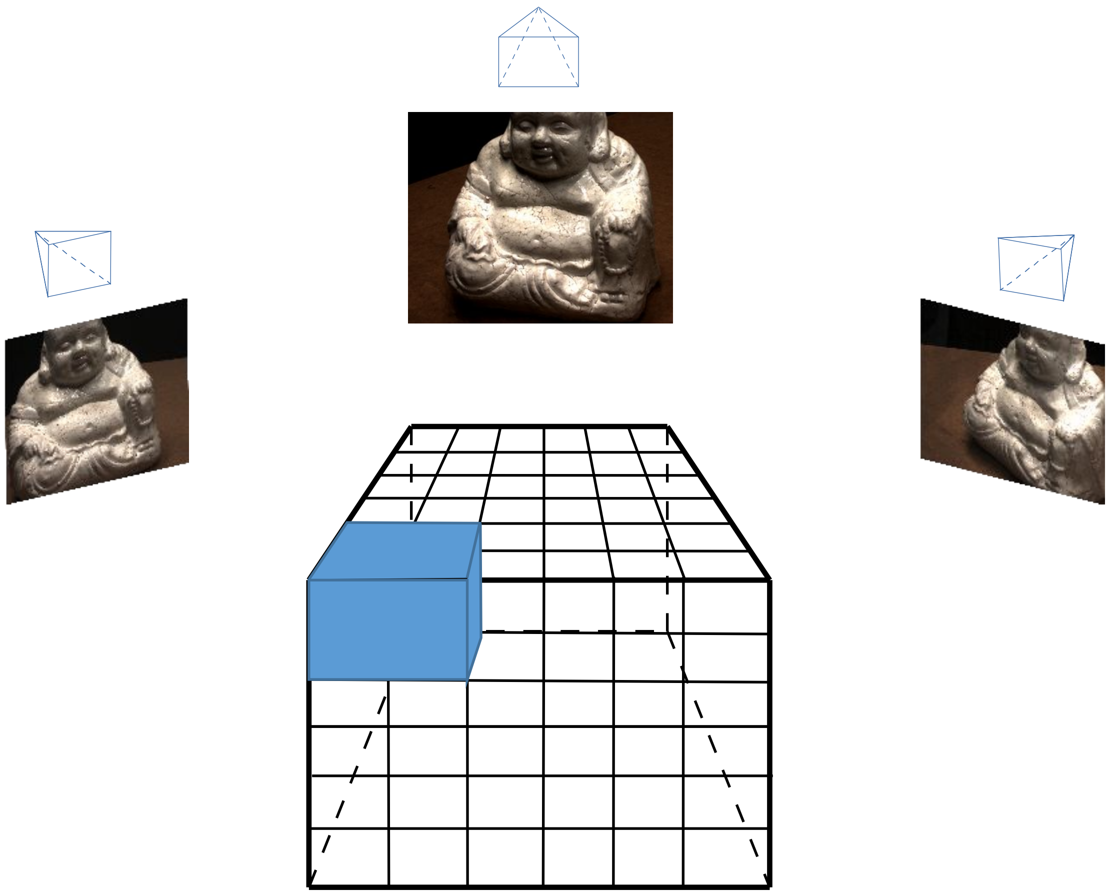
Cost-volume based methods



Cost-volume based methods

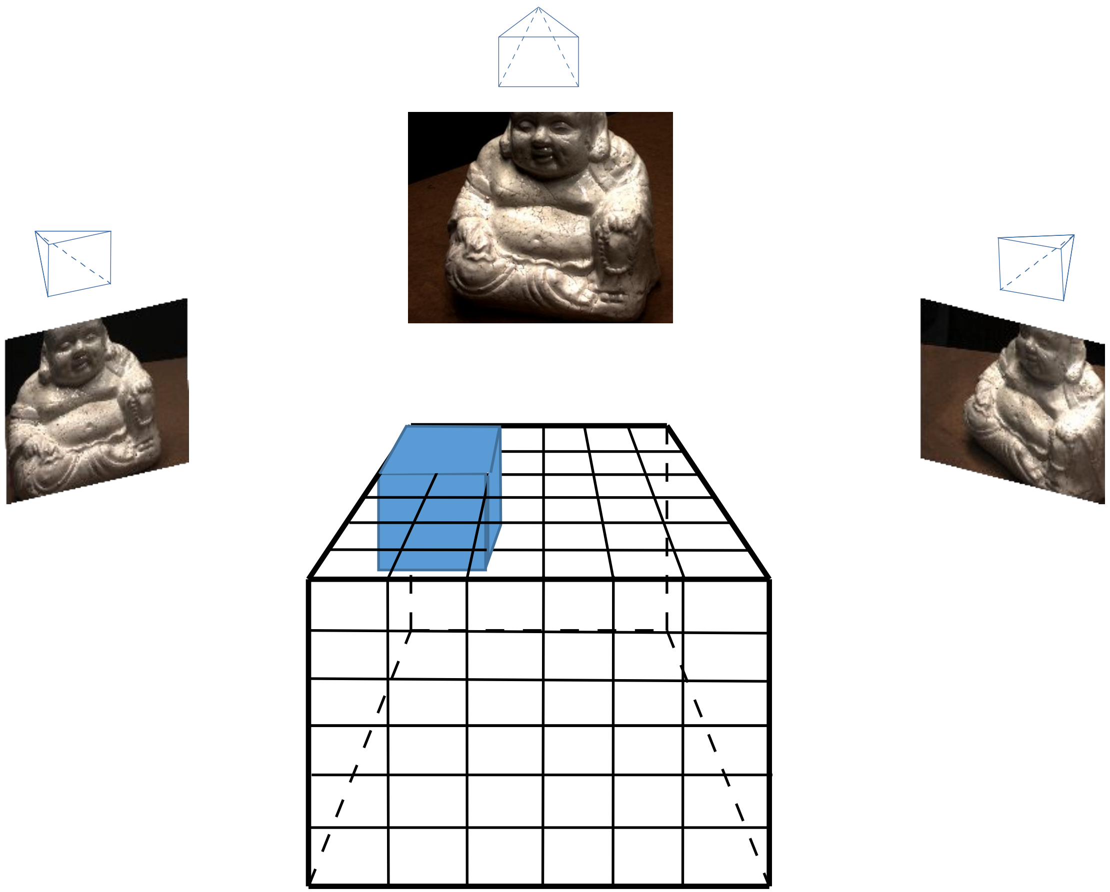


Cost-volume based methods



Dense 3D CNNs

Cost-volume based methods



Dense 3D CNNs

Cost-volume based methods

Limitation:

Memory consumption **cubic** to resolution

Cost-volume based methods

Limitation:

Memory consumption **cubic** to resolution

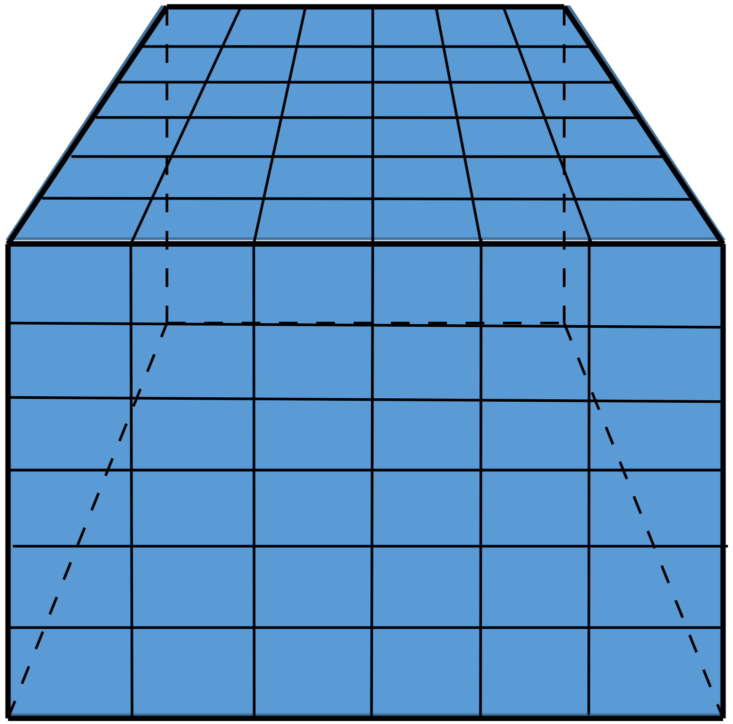
not feasible for
high-resolution accurate
reconstruction

Cost-volume based methods

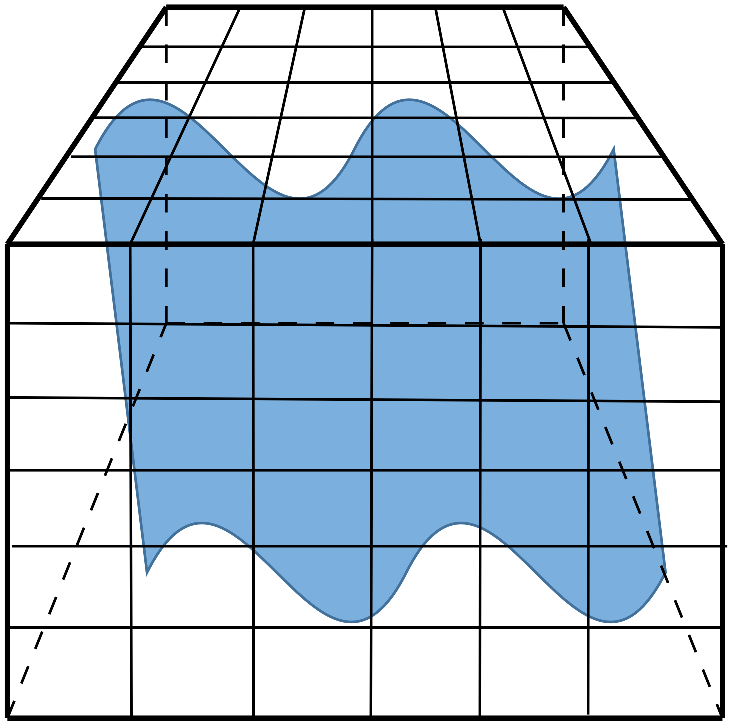
Are all these 3D CNNs necessary?

Cost-volume based methods

Are all these 3D CNNs necessary?



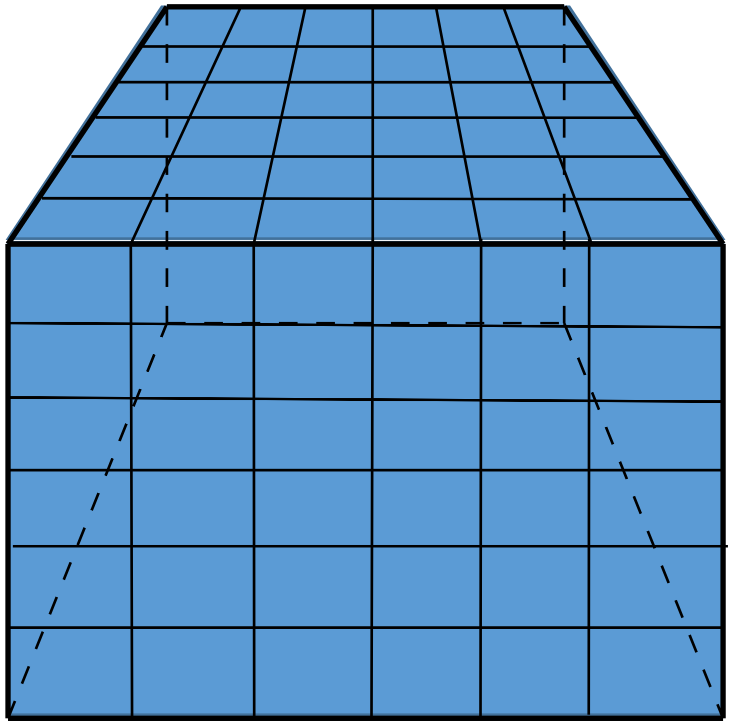
Cost-volume



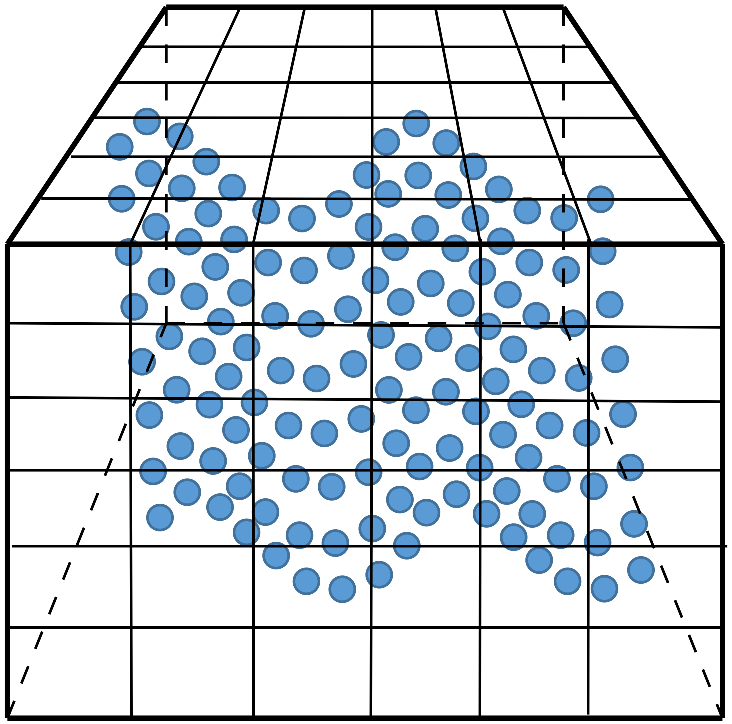
Target surface

Cost-volume based methods

Are all these 3D CNNs necessary?



Cost-volume

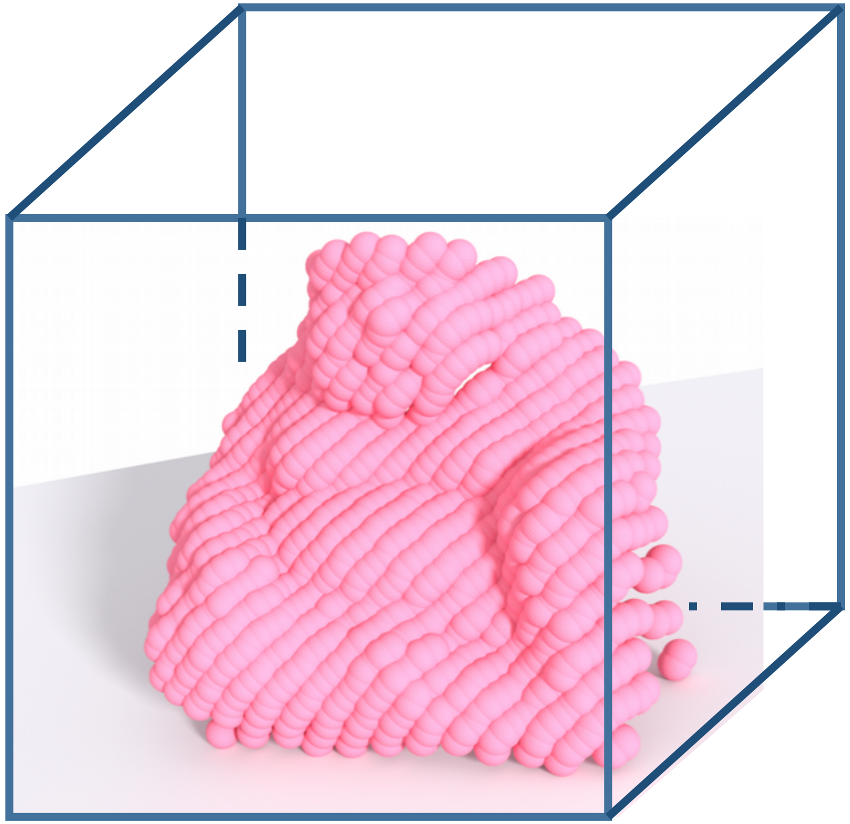


Surface points

Point MVSNet

Point Cloud Representation

Suitable for sparse occupancy memory-efficient



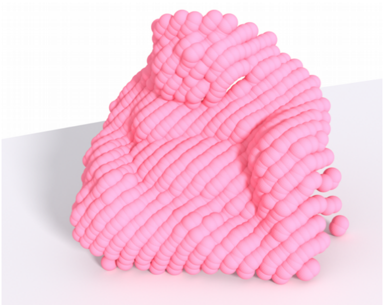
Architecture

Viewed images

Architecture

Viewed images

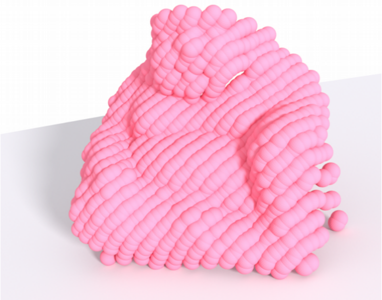
Coarse depth map



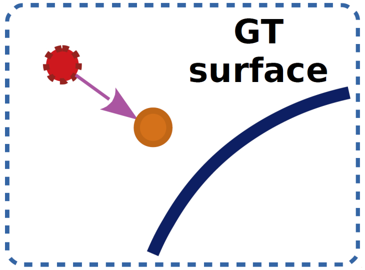
Architecture

Viewed images

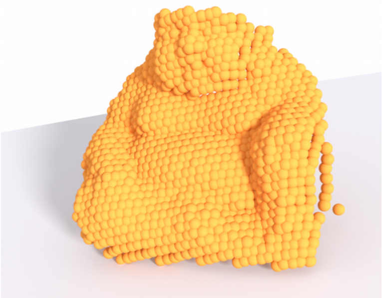
Coarse depth map



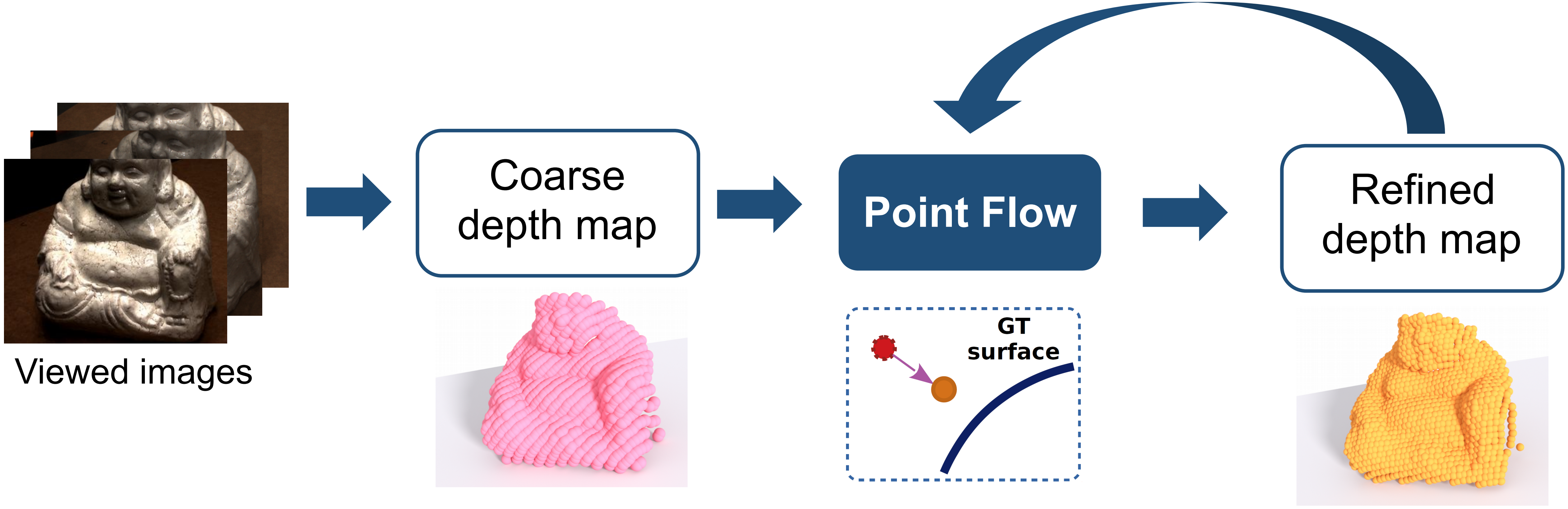
Point Flow



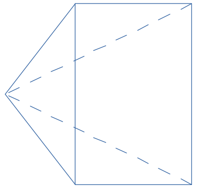
Refined depth map



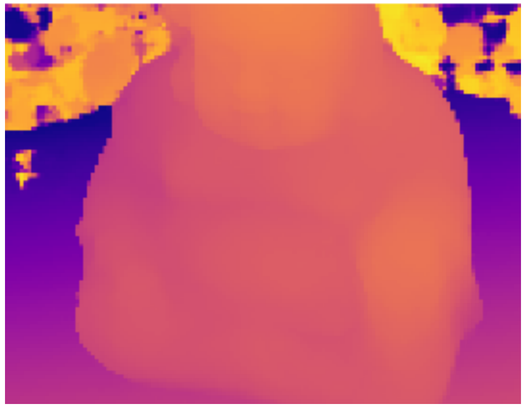
Architecture



Initial Point Cloud

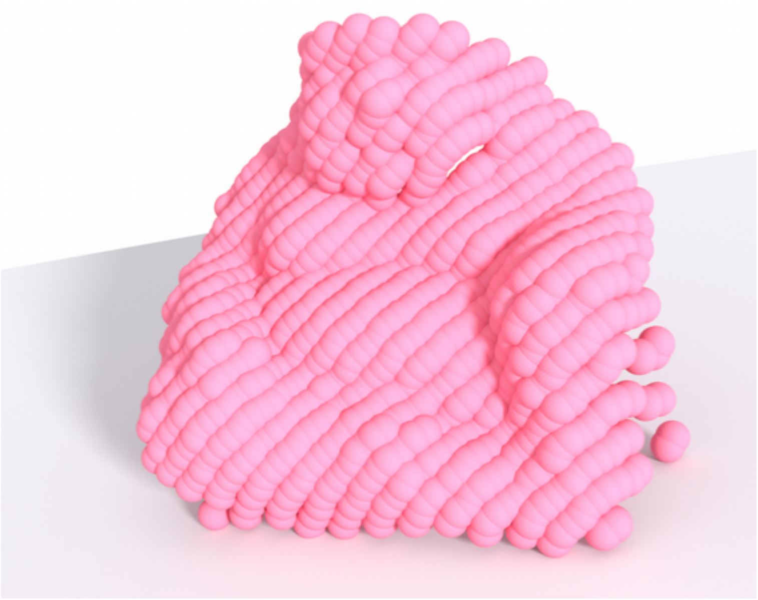


Reference camera



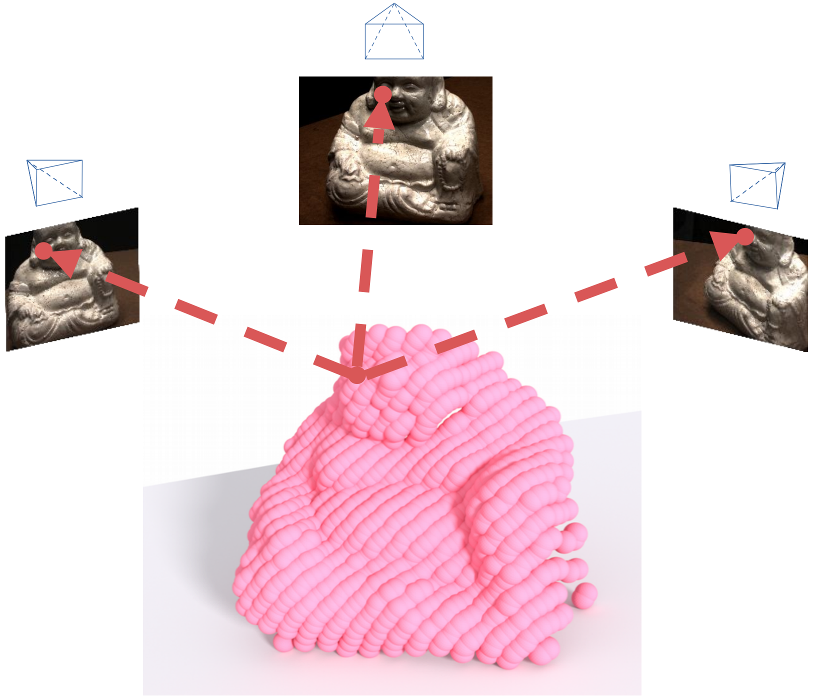
Coarse Depth map

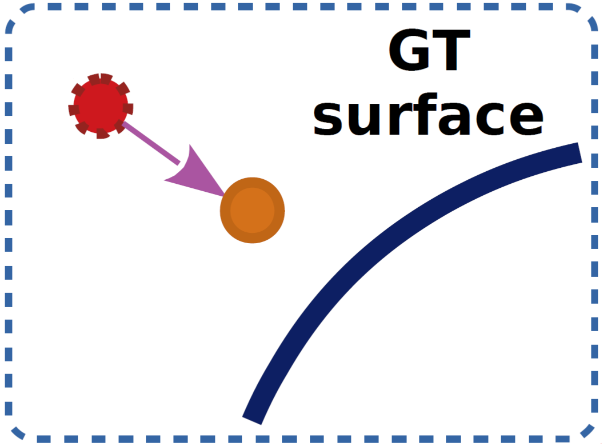
Unprojection



Initial point cloud

Point Flow

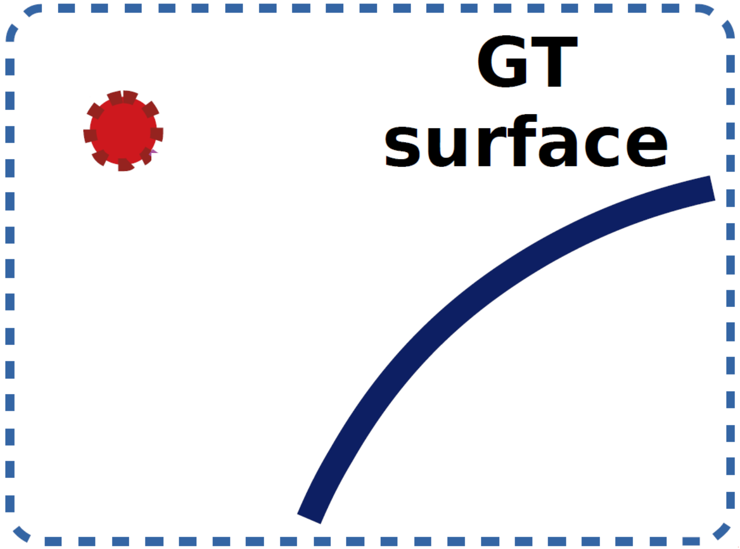




Flow prediction

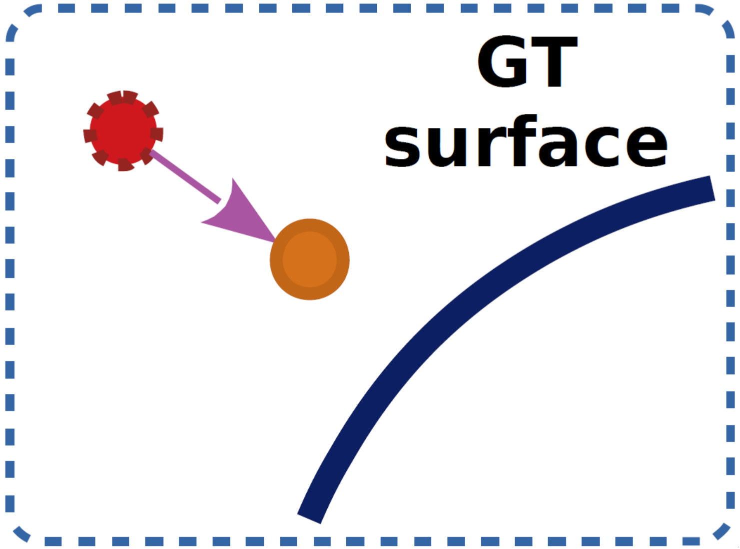
Point Flow

How to predict the flow to the GT surface?

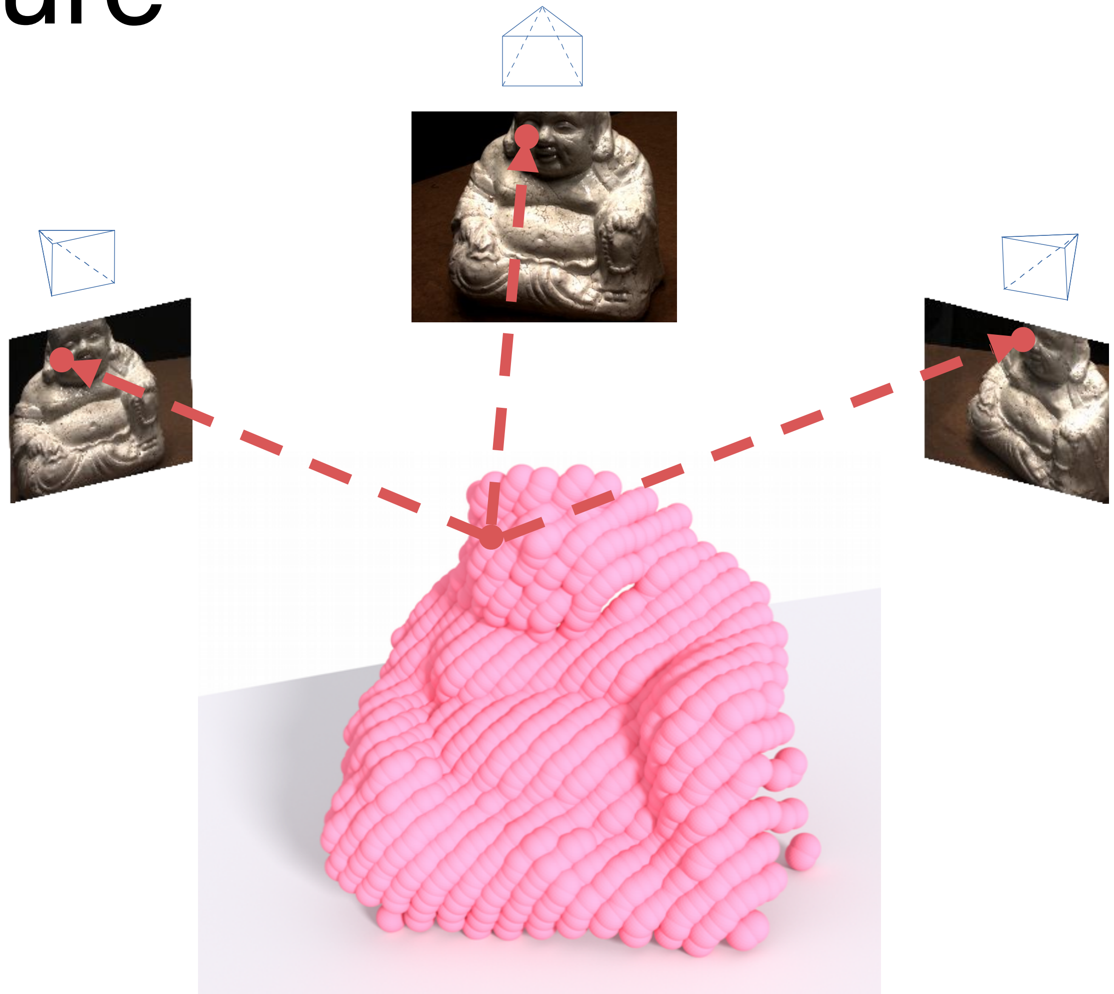


Point Flow

How to predict the flow to the GT surface?

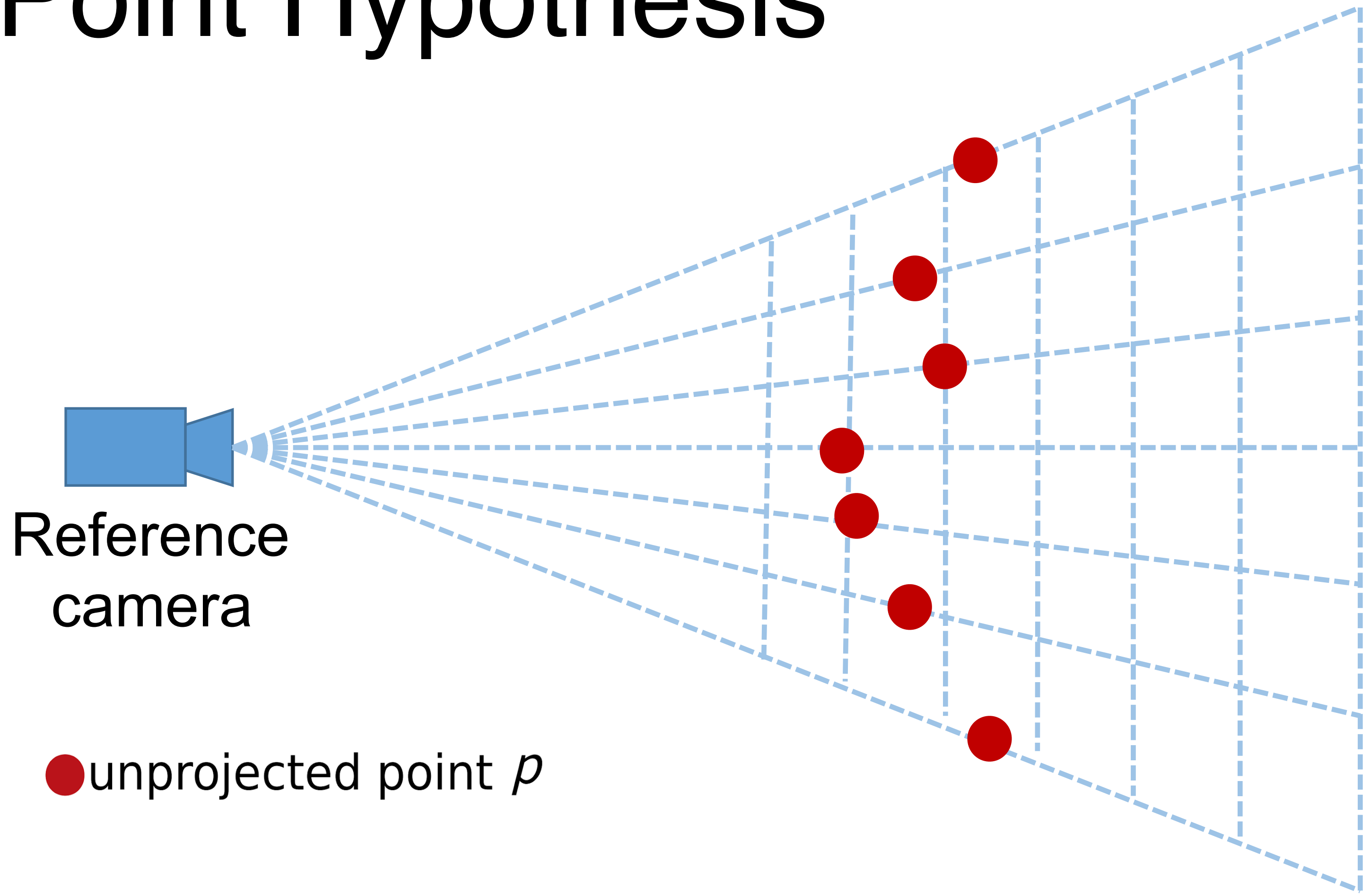


Point Feature

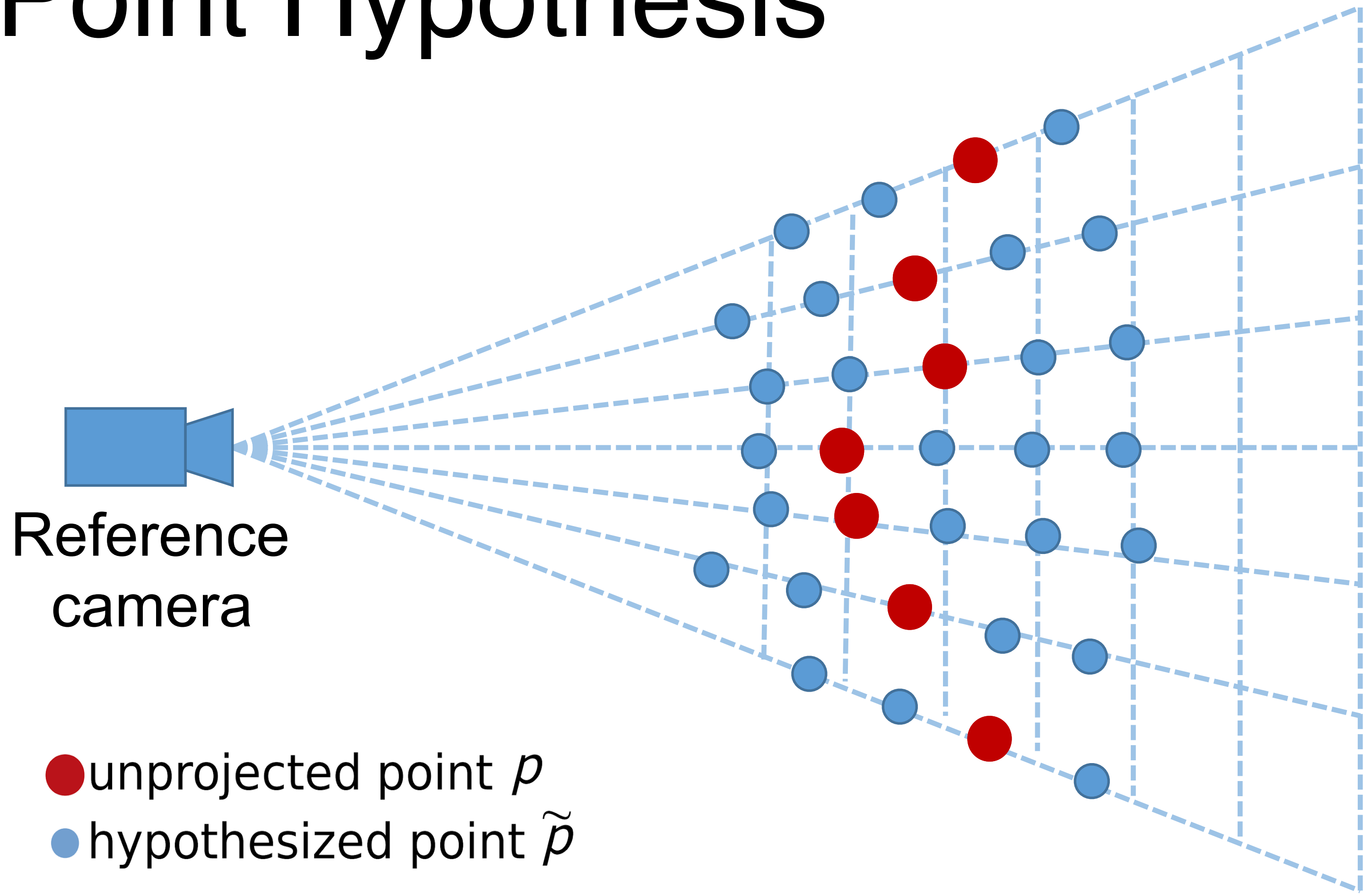


Dynamic feature fetching

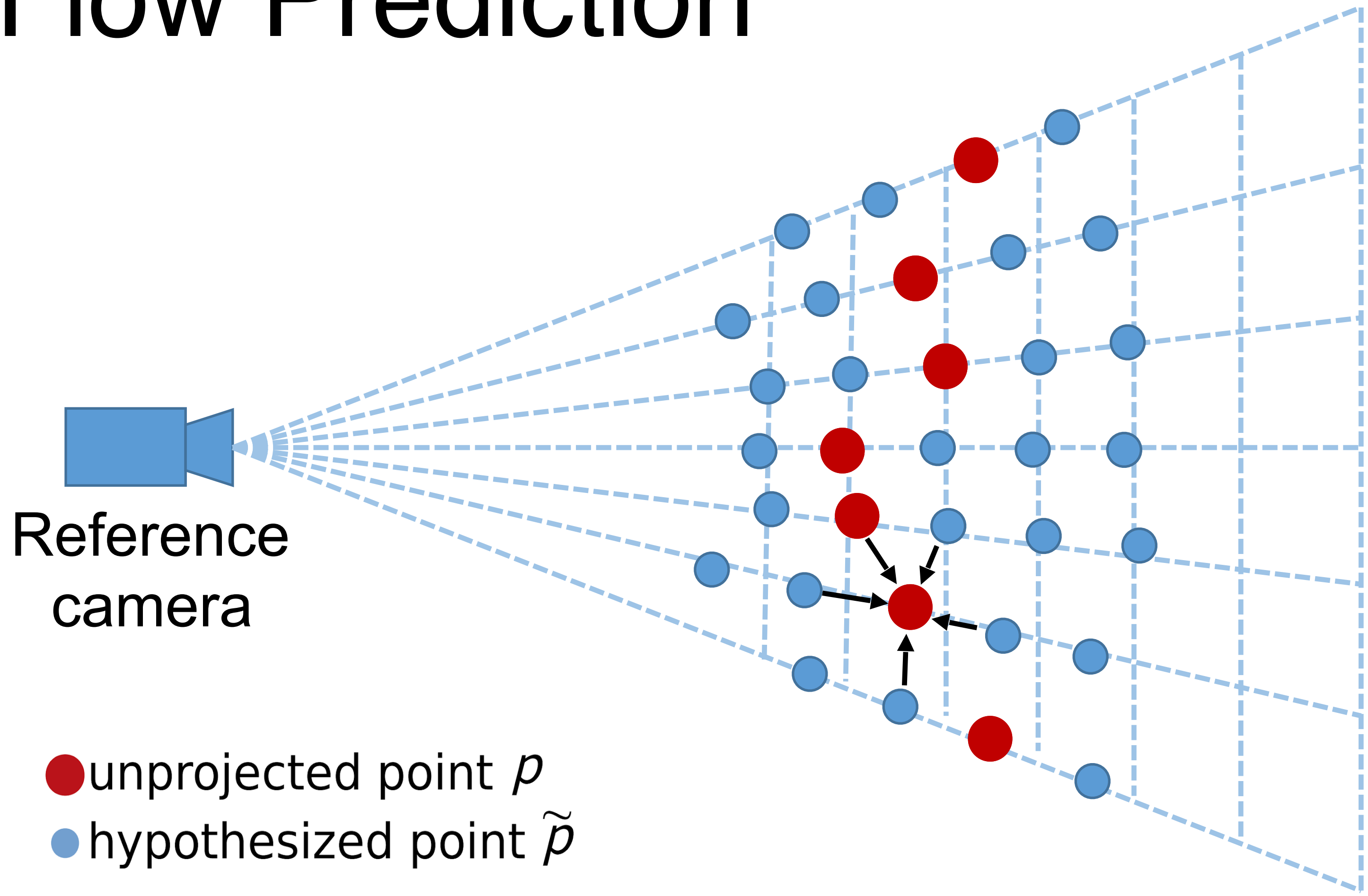
Point Hypothesis



Point Hypothesis



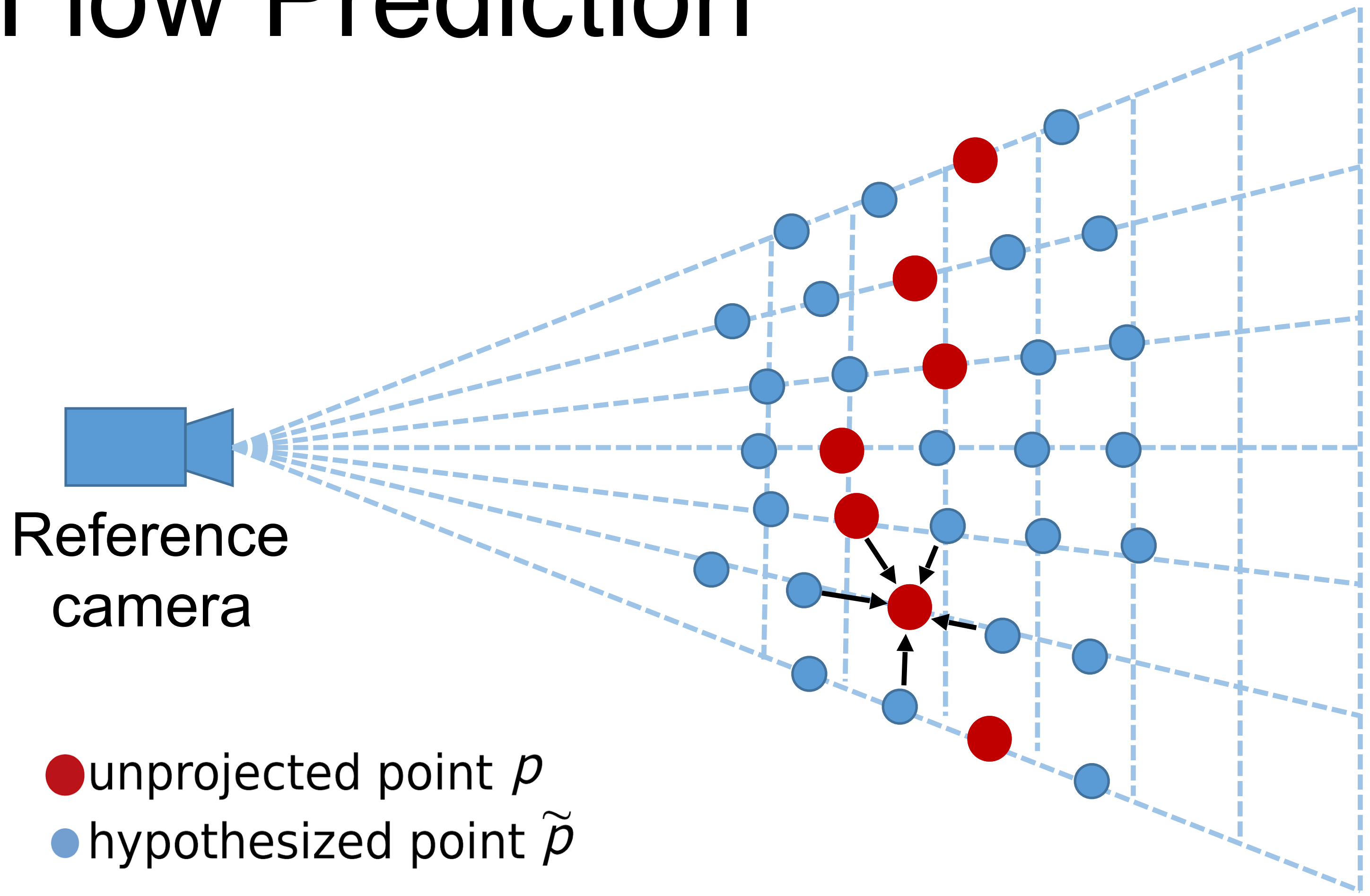
Flow Prediction



expected offset

$$\Delta d_p = \mathbf{E}(ks) = \sum_{k=-m}^{m} ks \times \mathrm{Prob}(\tilde{\mathbf{p}}_k)$$

Flow Prediction



expected offset

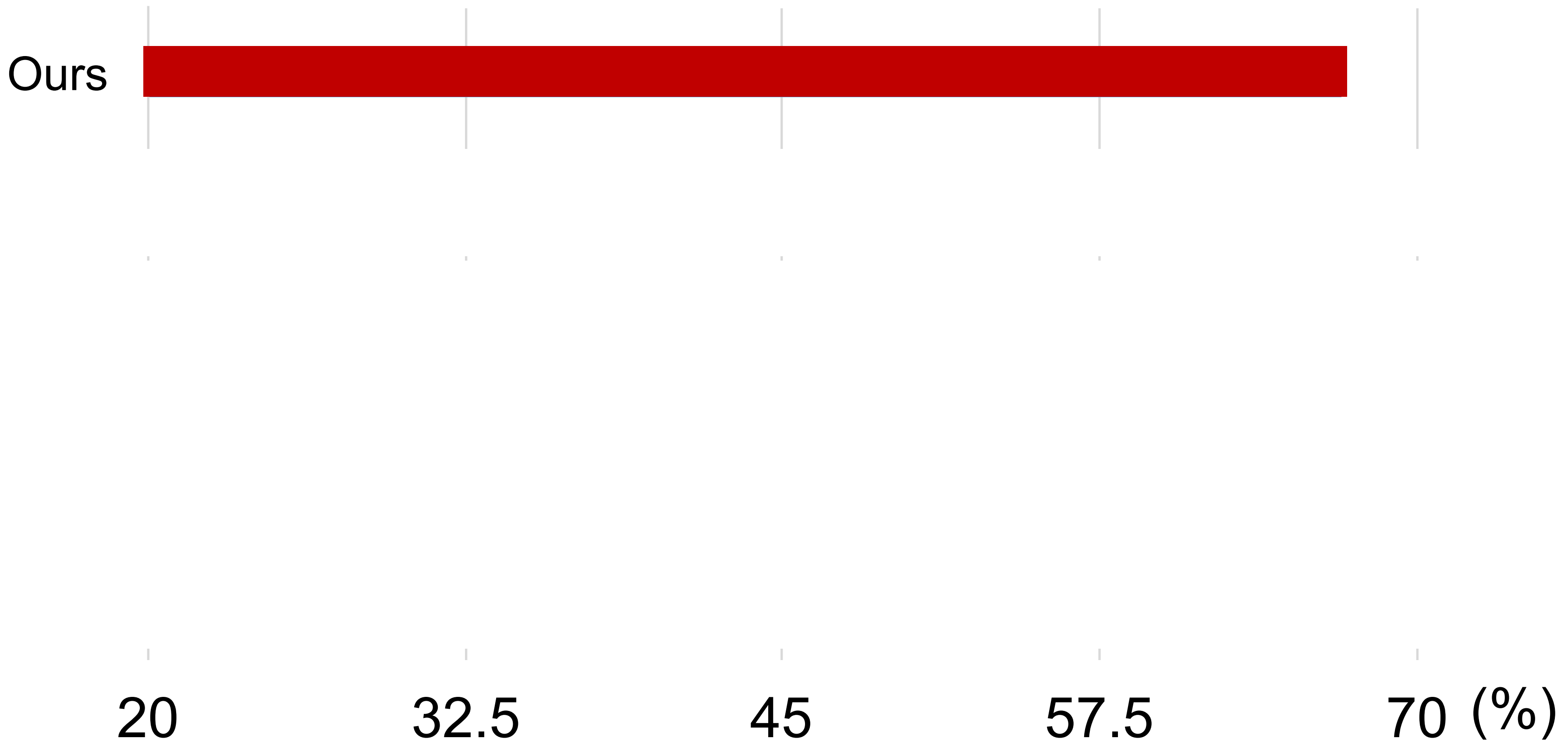
$$\Delta d_p = \mathbf{E}(ks) = \sum_{k=-m}^{m} ks \times \mathrm{Prob}(\tilde{\mathbf{p}}_k)$$

Results

DTU Benchmark

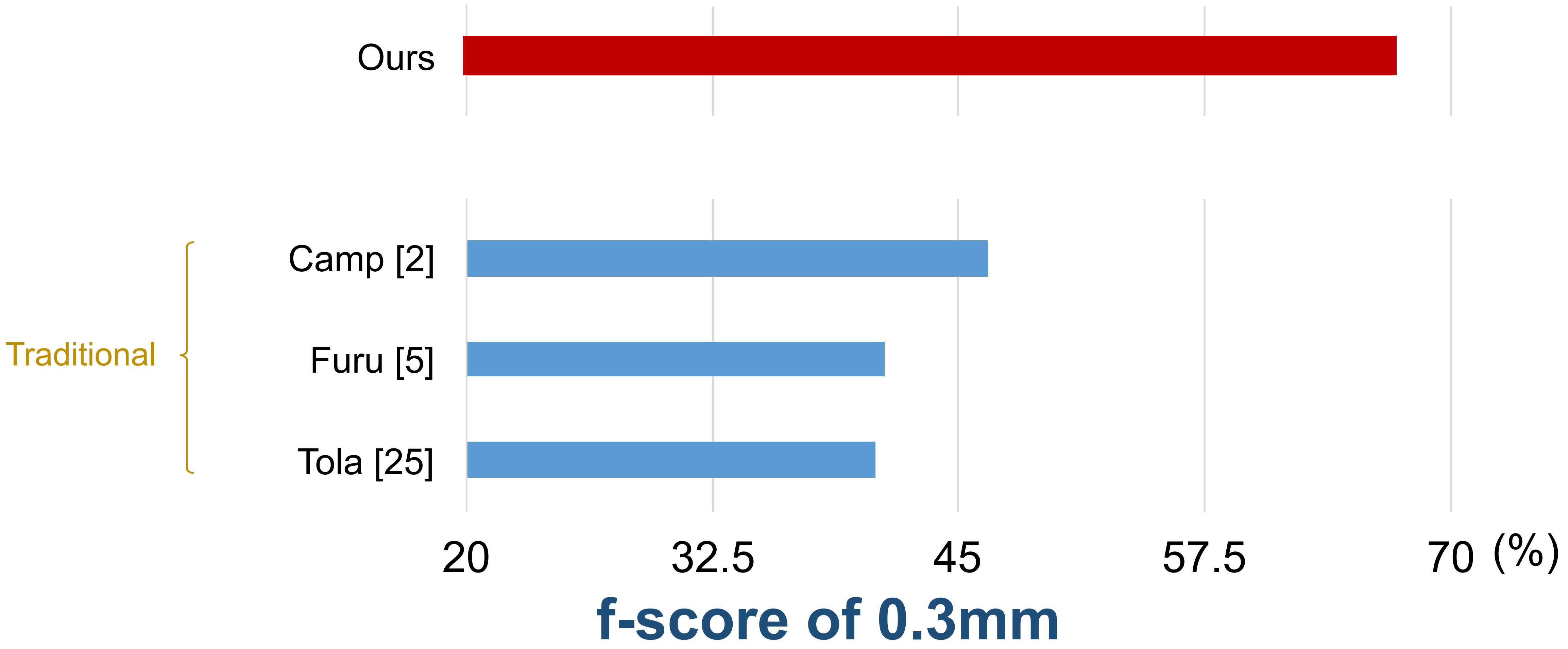
49 views / scene

DTU Benchmark
49 views



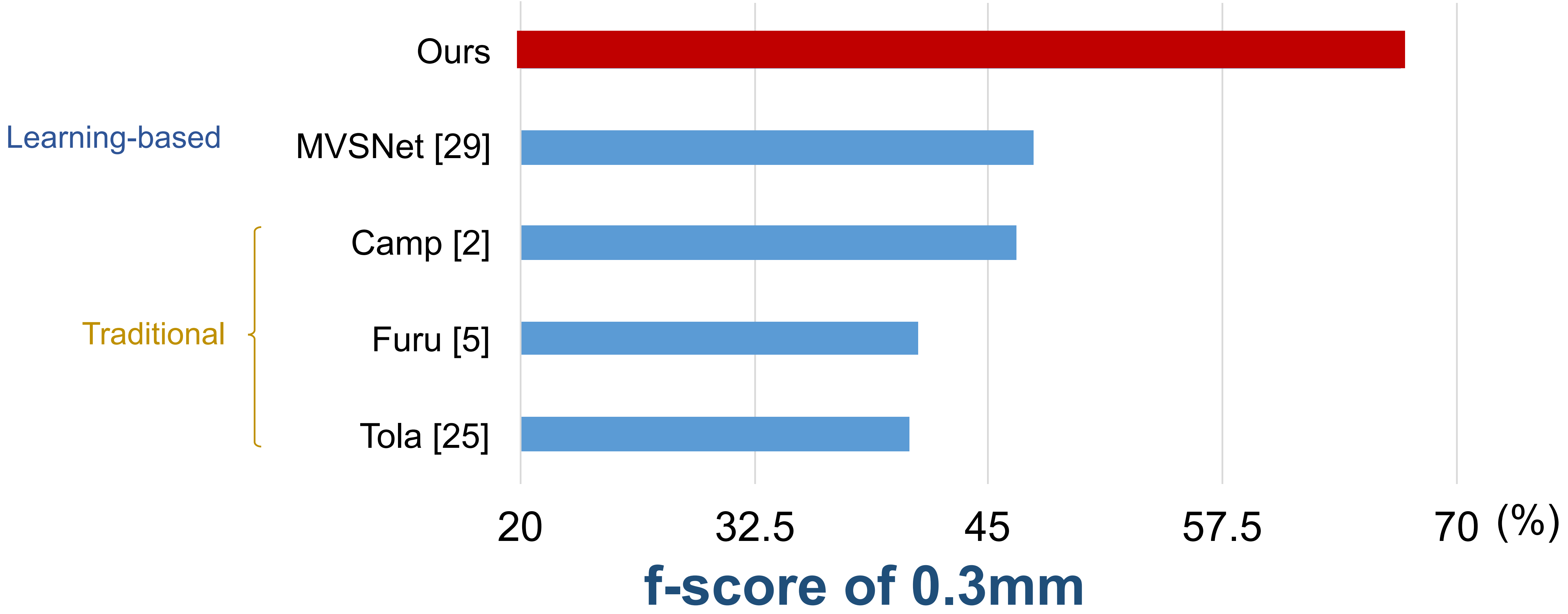
f-score of 0.3mm

DTU Benchmark
49 views



DTU Benchmark

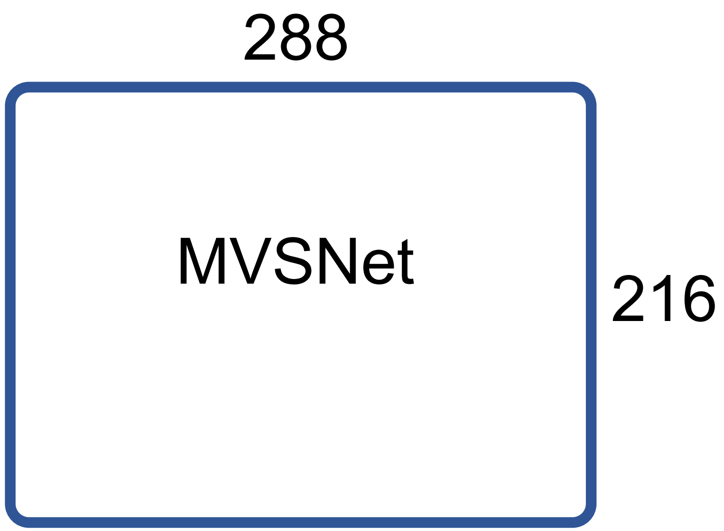
49 views



Memory Efficiency

Memory Efficiency

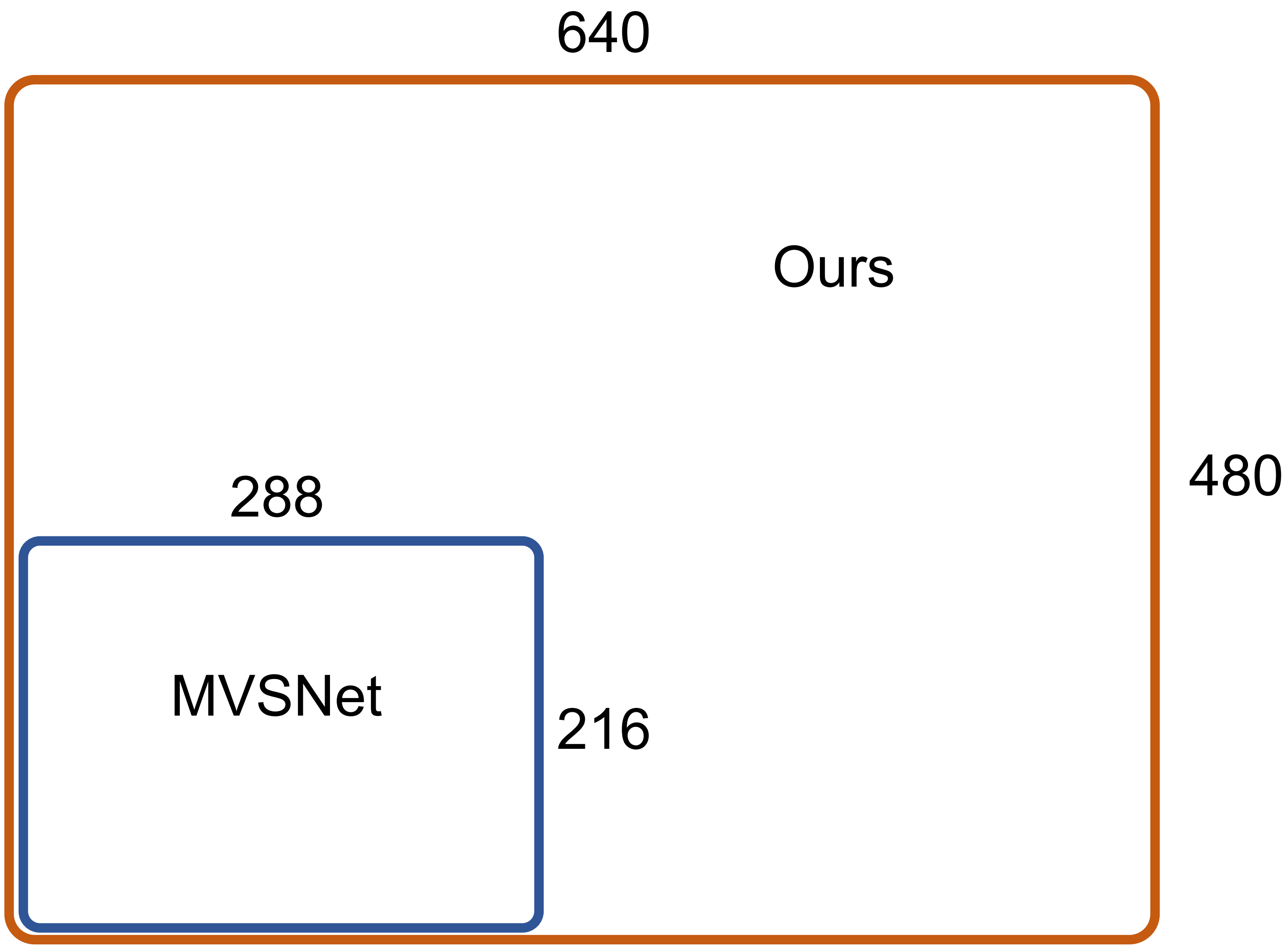
Depth map resolution



Memory Efficiency

Depth map resolution

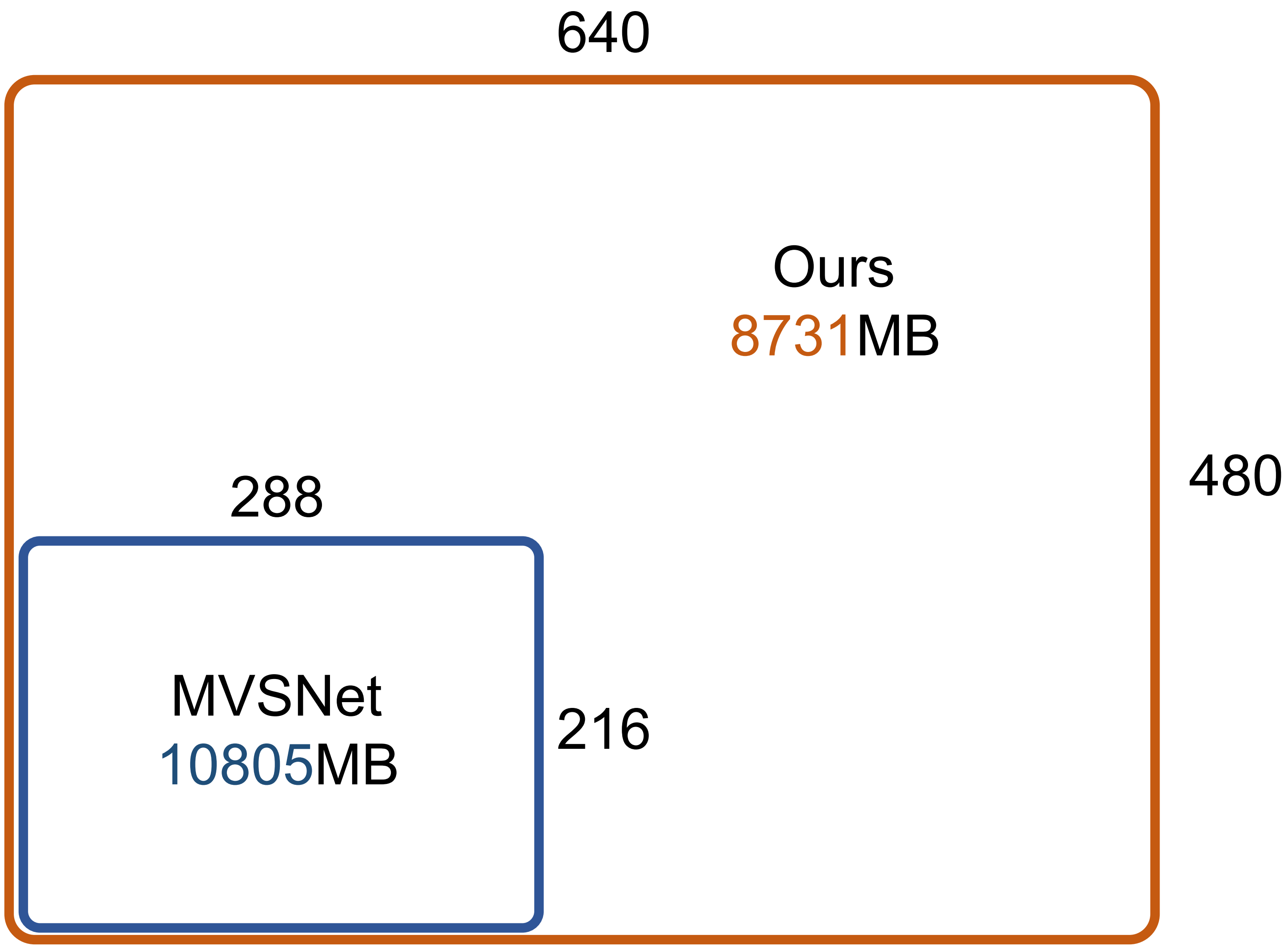
5X



Memory Efficiency

Depth map resolution

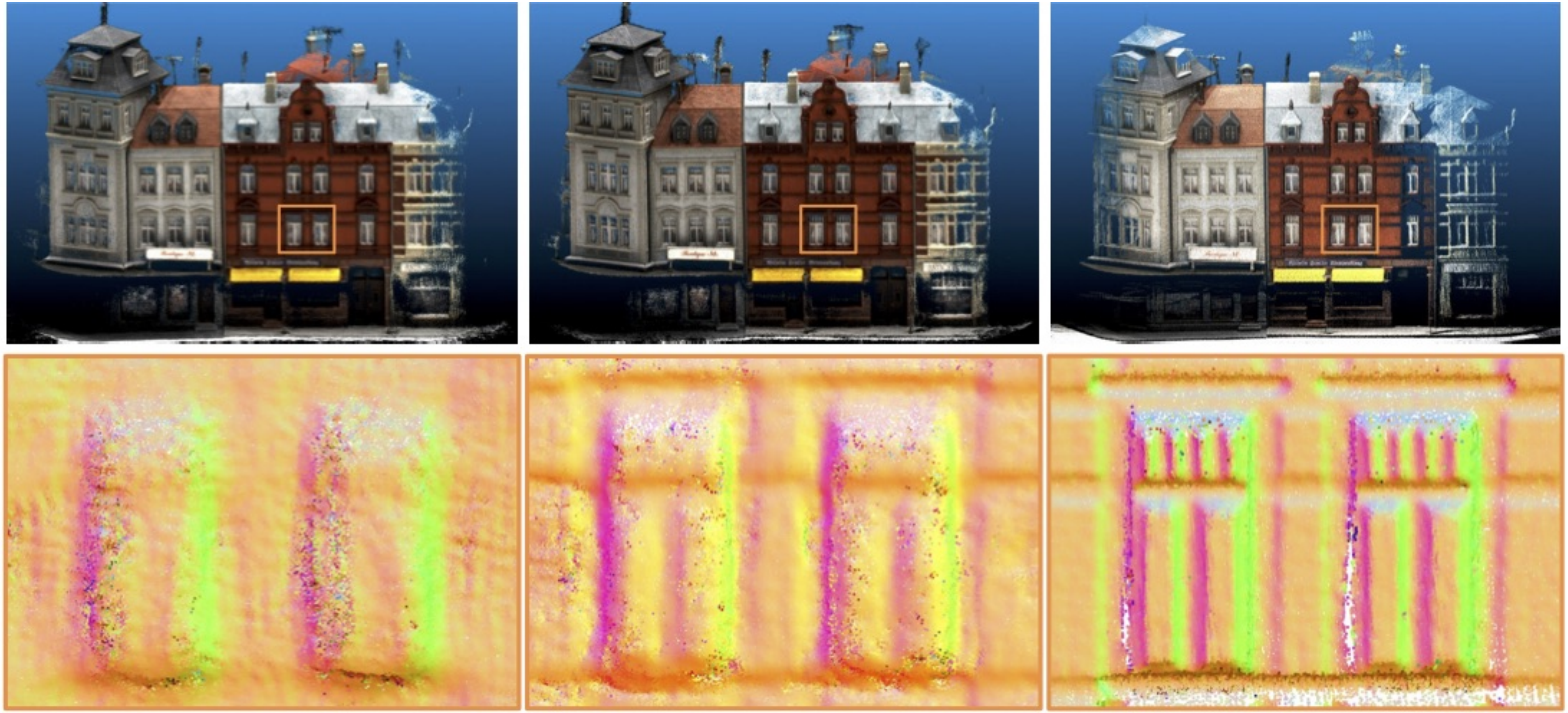
5X



Accurate Reconstruction on DTU

Reconstructed Point Cloud

Computed Normals



MVSNet | Ours | Ground truth

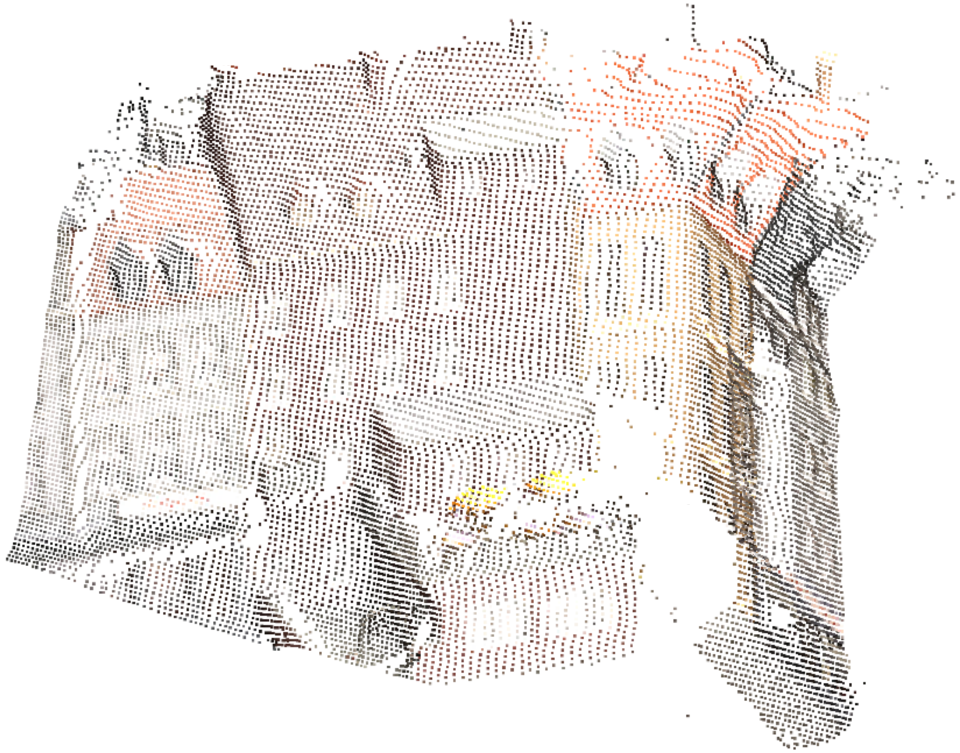
Reconstruction is More Complete

Camp [2]

Ours

Foveated depth inference

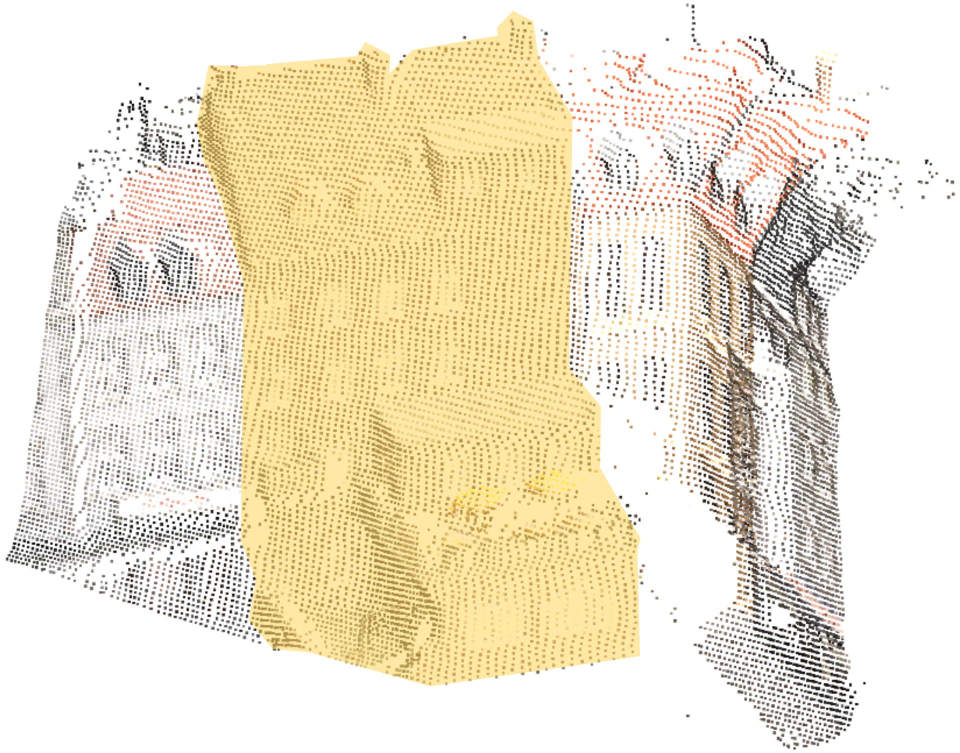
only refine the ROI depth



sparse

Foveated depth inference

only refine the ROI depth

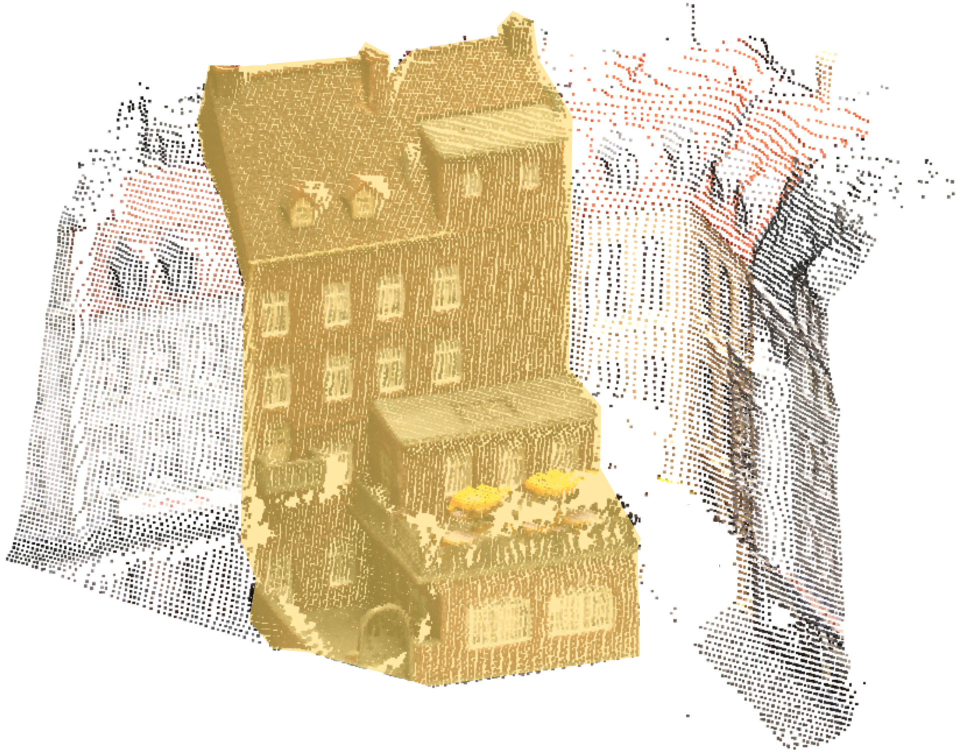


sparse

denser

Foveated depth inference

only refine the ROI depth

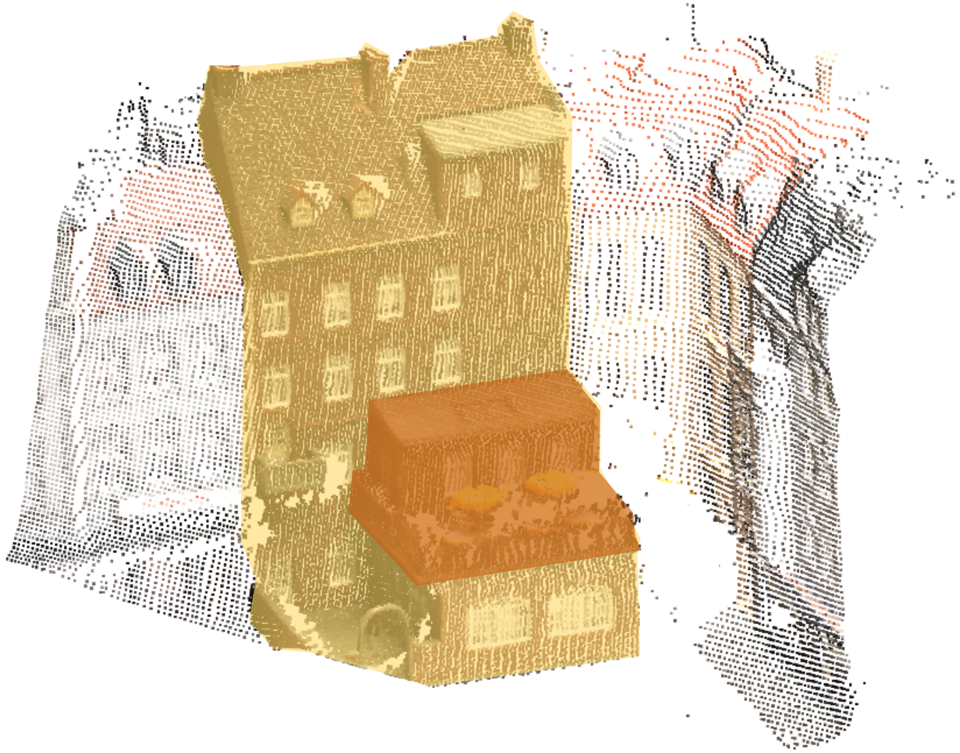


sparse

denser

Foveated depth inference

only refine the ROI depth



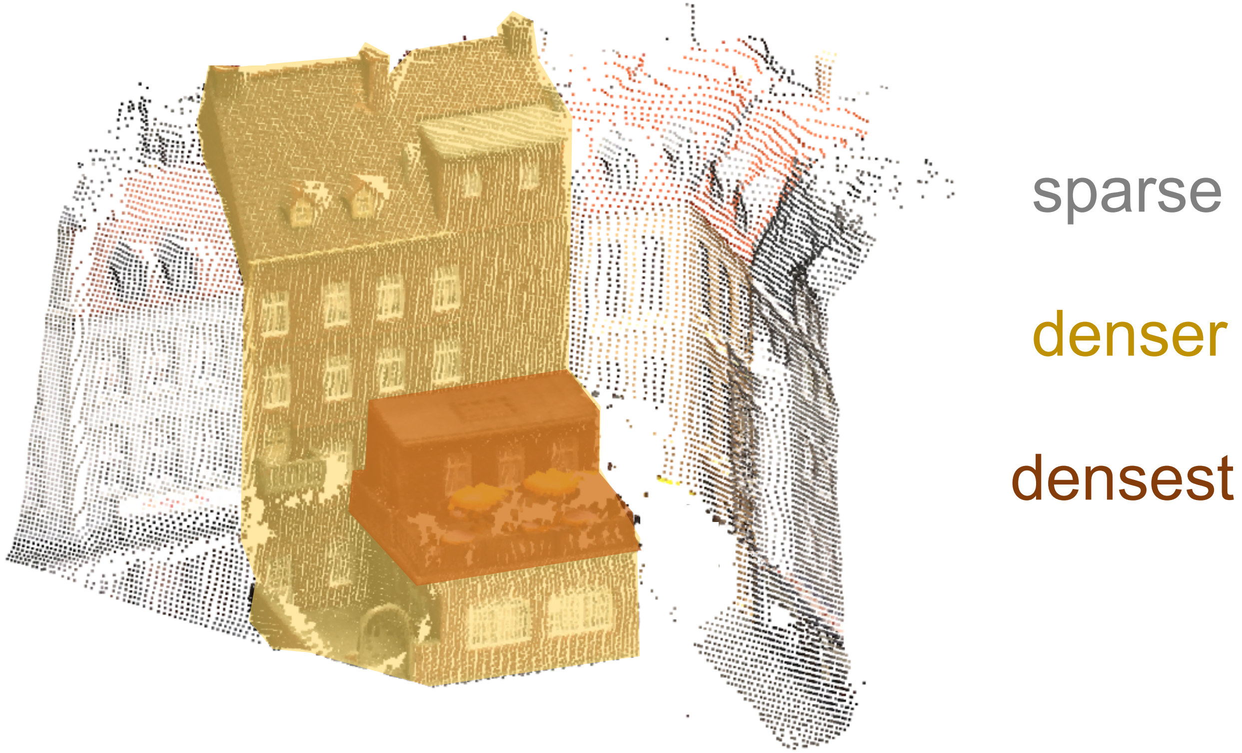
sparse

denser

densest

Foveated depth inference

only refine the ROI depth



Conclusion

Conclusion

- MVS target surface is **sparse** in 3D space

Conclusion

- MVS target surface is **sparse** in 3D space
- **Point MVSNet** process the surface **points directly**

Conclusion

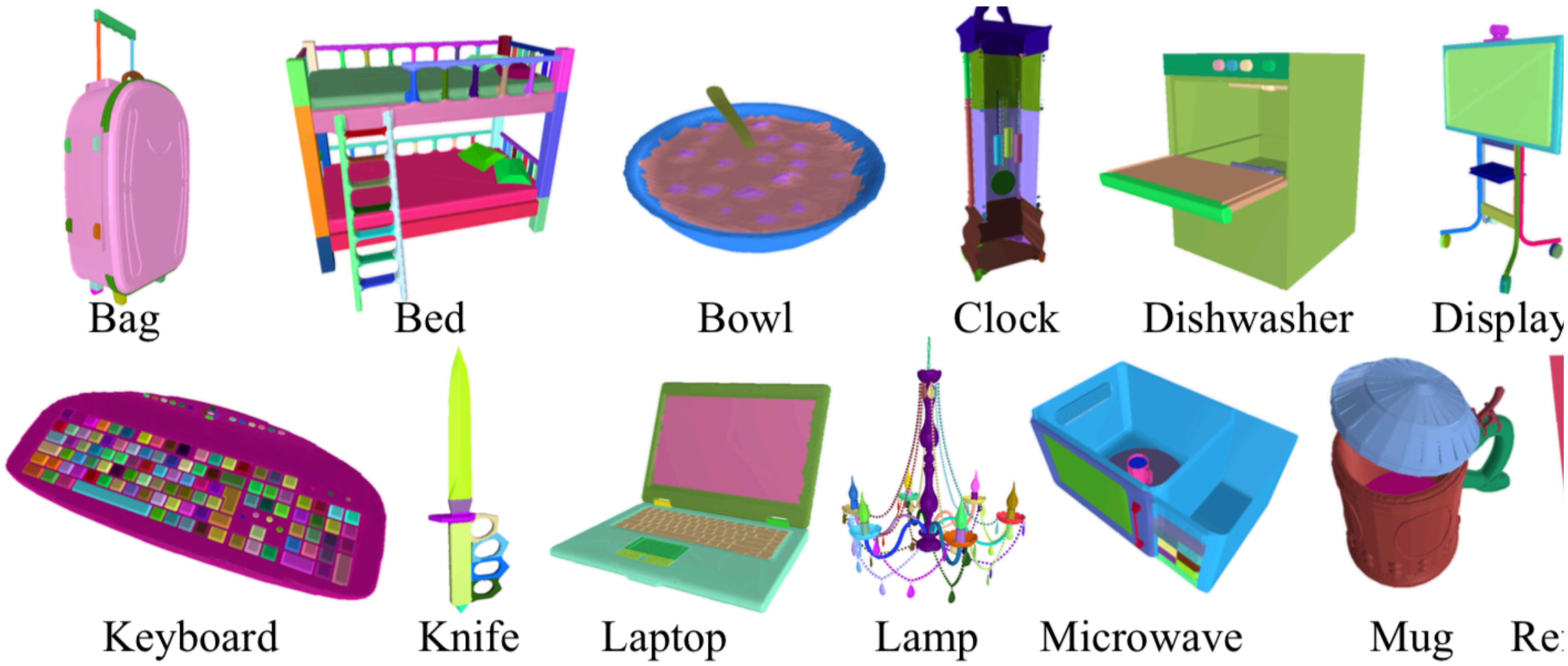
- MVS target surface is **sparse** in 3D space
- **Point MVSNet** process the surface **points directly**
- Better **time and memory efficiency**

Conclusion

- MVS target surface is **sparse** in 3D space
- **Point MVSNet** process the surface **points directly**
- Better **time and memory efficiency**
- **Iterative** refinement

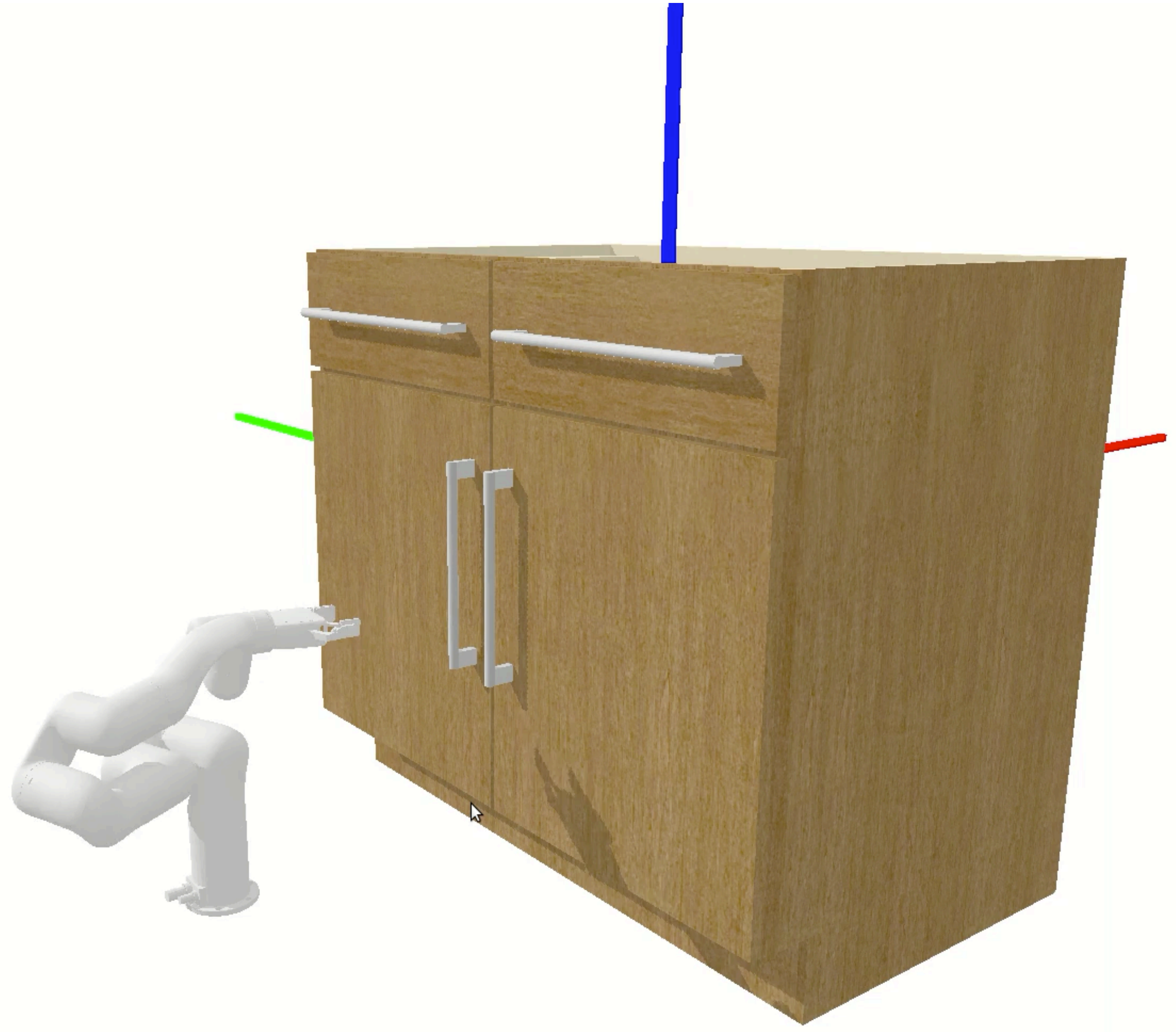
PartNet: A Database for Actionable Information

573,585 part instances over **26,671** 3D models covering **24** object categories

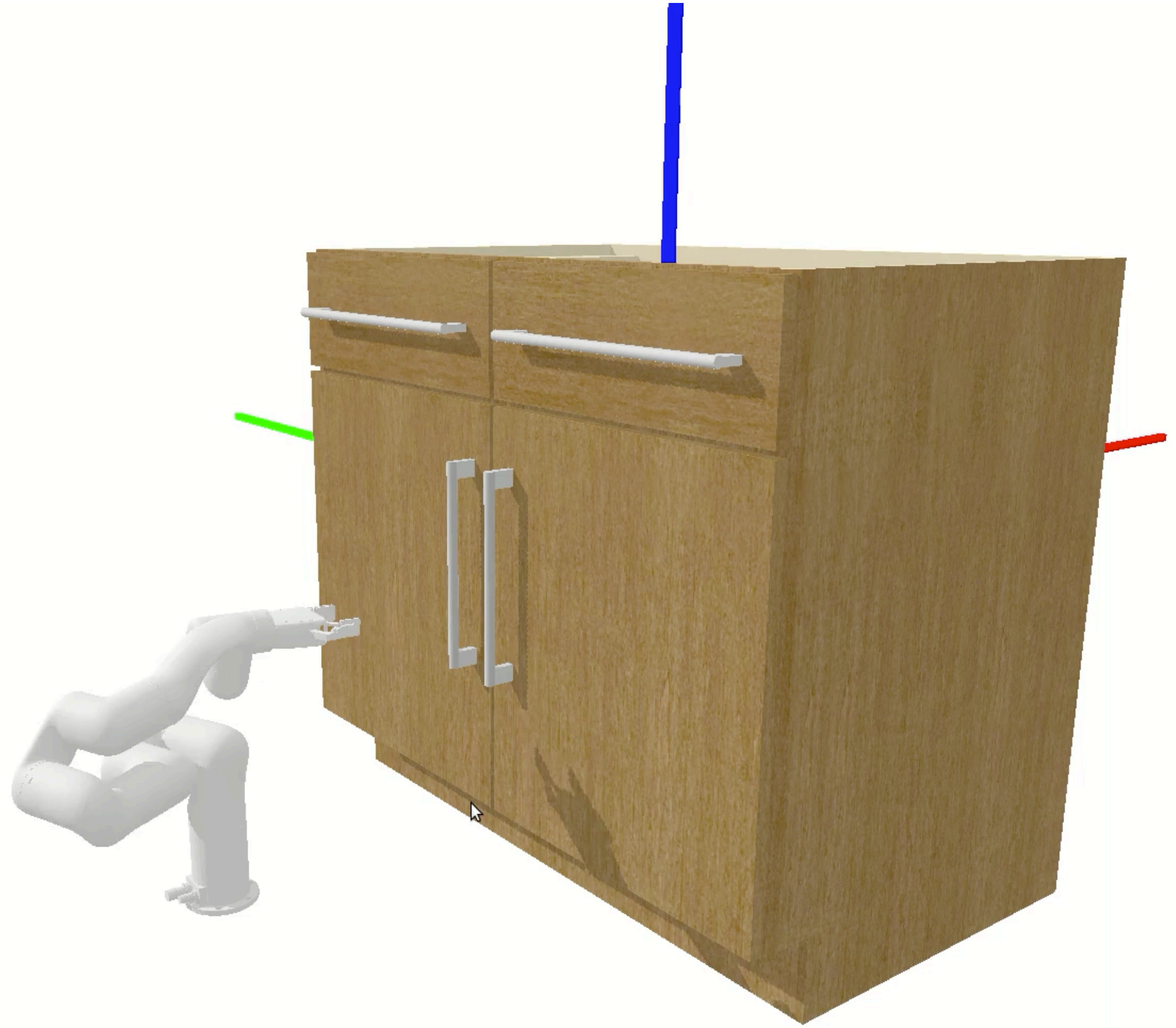


CVPR2019

Interactive Simulated Environment Modeling



Interactive Simulated Environment Modeling



Thank you!