



Introduction to Embodied AI

Lecture 7 - Vision and Grasping III

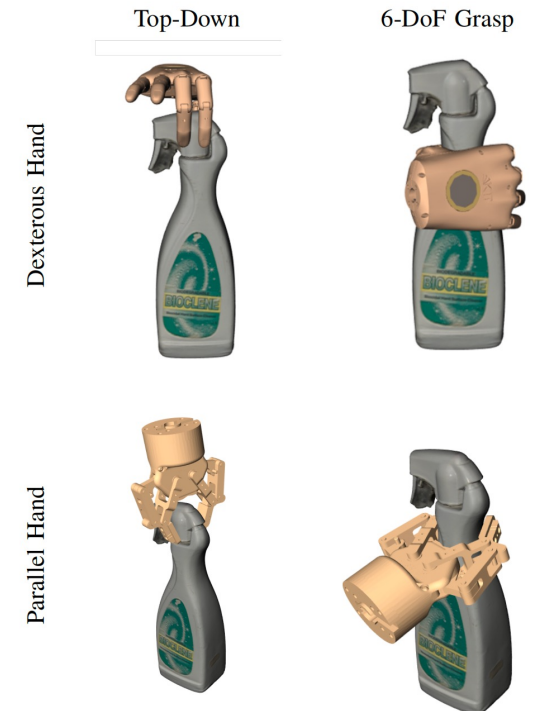
Prof. He Wang



Grasp Data Synthesis

Terminology

- **Grasp Pose** defines the position and orientation of a hand
 - **4-DoF grasp**: a 3D position and 1D hand orientation aligned with the direction of gravity, a.k.a. “top-down grasping”
 - **6-DoF grasp**: defined by a 3D position and 3D orientation
- **Force Closure** means if the forces that a grasp can be applied at the set of frictional contacts are sufficient to compensate for any external wrench applied to the object

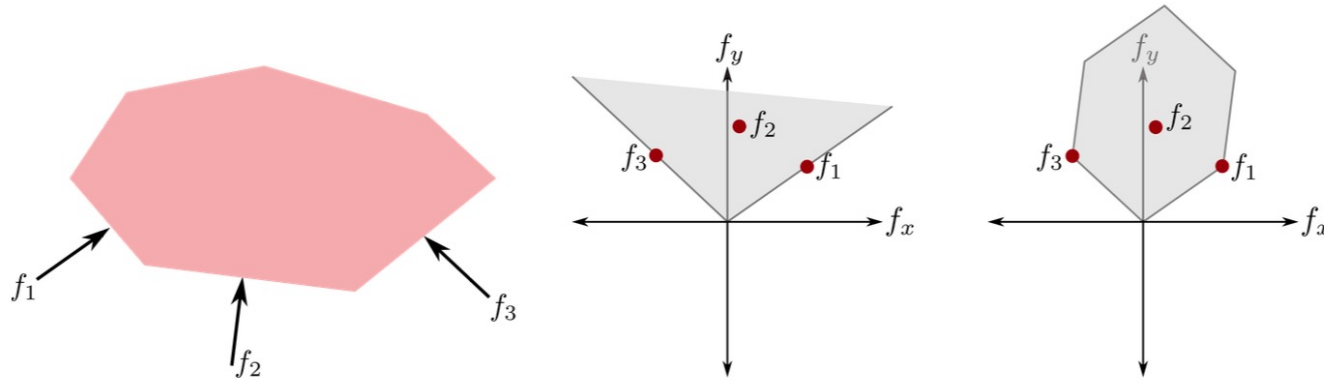


Force Closure

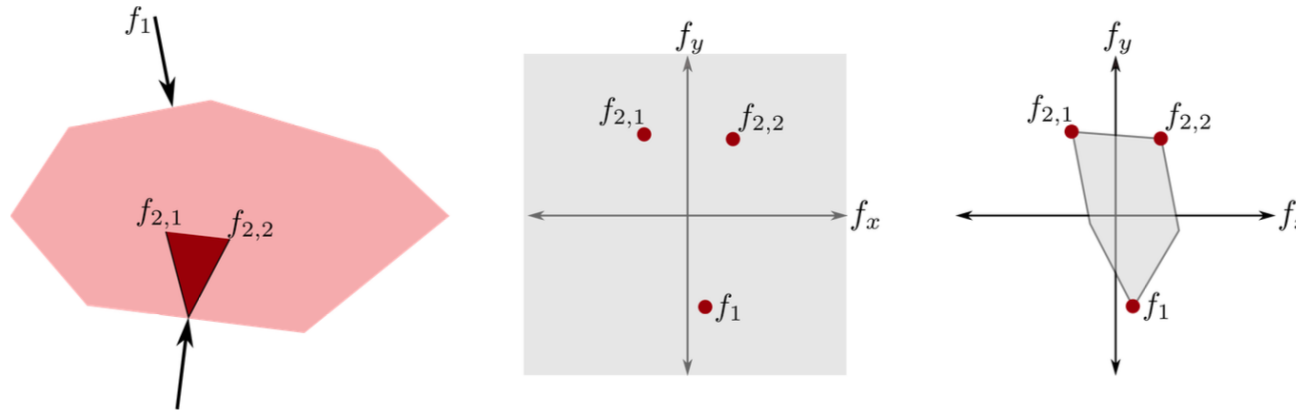
- In physics,
 - given a set of frictional contacts acting on a body, it is in **force closure** if the positive span of the wrench cones is the entire wrench space
 - if a rigid body is fully immobilized by a set of rigid stationary fixtures, we say it is in **form closure**
- The conditions for **force closure** are identical to the conditions for **first-order form closure** if all contacts are frictionless.
- When planning a grasp by a robot hand, **force closure** is a good minimum requirement. **Form closure** is usually too strict, requiring too many contacts.
- Simply, successful grasp $\leq$ force Closure $\leq$ form closure

Force Closure

- A simplified 2D problem



A 2D grasp with frictionless contacts



A 2D grasp consisting of a contact with friction and a contact without friction

- ✓ 3 possible force vectors with 2 linearly independent
- ✗ generate zero force with non-zero forces at each contact
- ✗ The origin is contained in the interior of the space of possible applied forces.

- ✓ 3 possible force vectors with 2 linearly independent
- ✓ generate zero force with non-zero forces at each contact
- ✓ The origin is contained in the interior of the space of possible applied forces.

- Watch: <https://www.youtube.com/watch?v=6RWFFMtD5k8>

Dataset

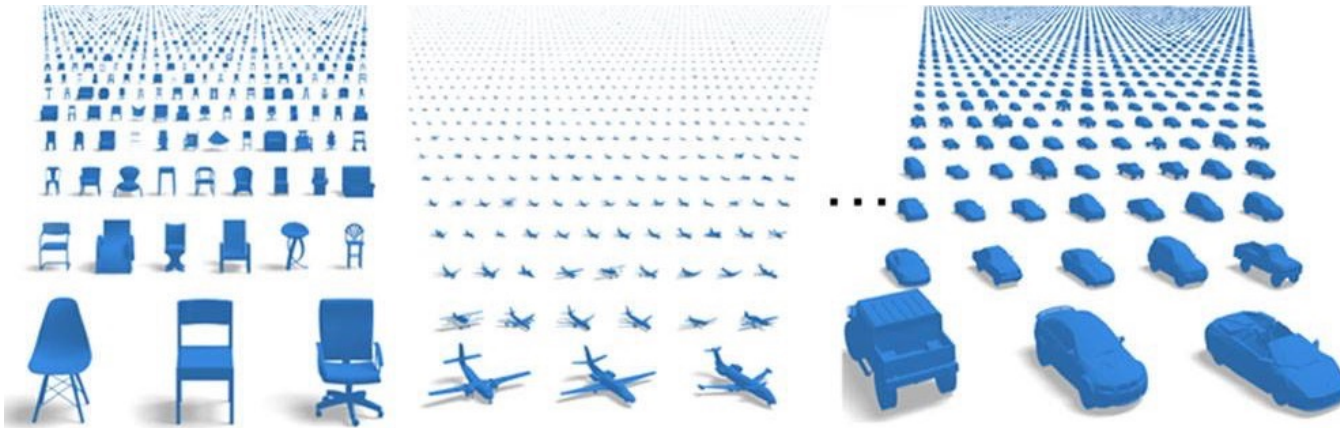
- Annotate real data
- Generate synthetic data with grasp label
 - sample grasp candidate with objects geometry
 - evaluate them with physical simulator
 - label grasps that form force closure as positive

Object Datasets

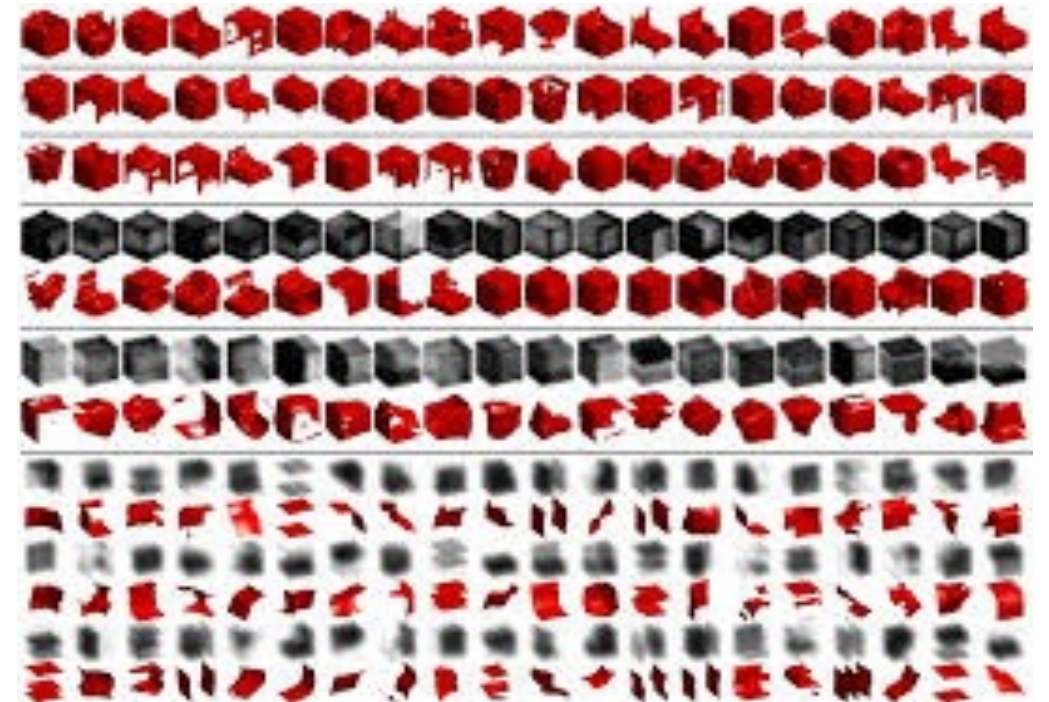
Synthetic dataset

- Object set: without grasping annotation

ShapeNet



ModelNet



Object Datasets

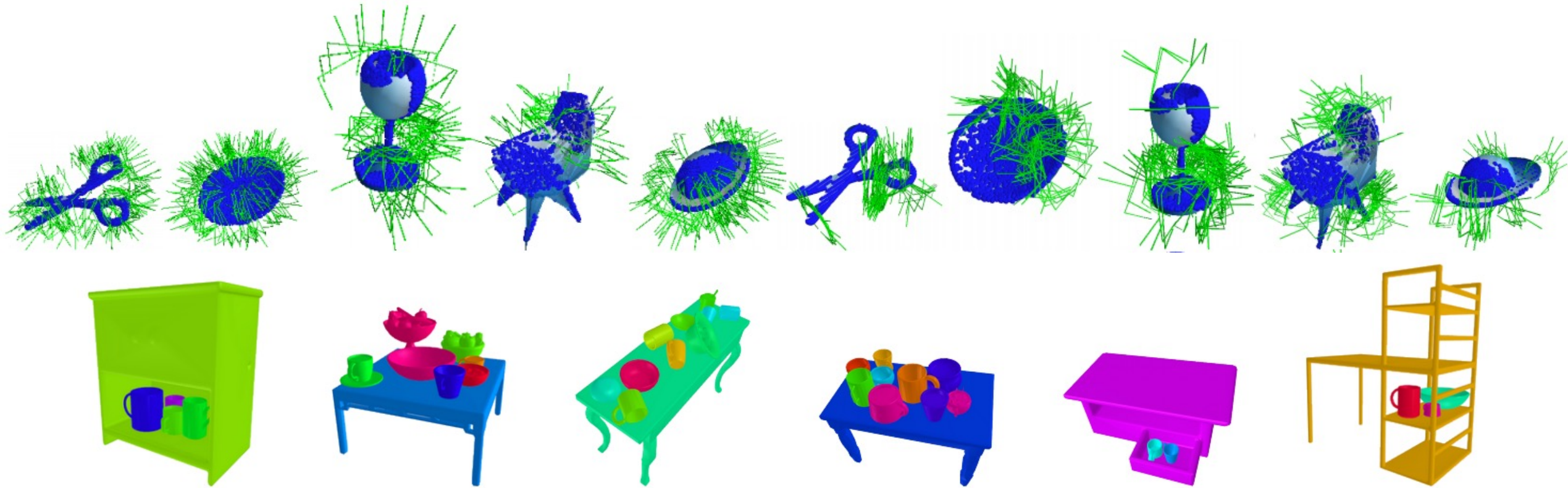
- **Objaverse-XL**: 10M+ 3D Objects



Grasp Dataset

Synthetic dataset

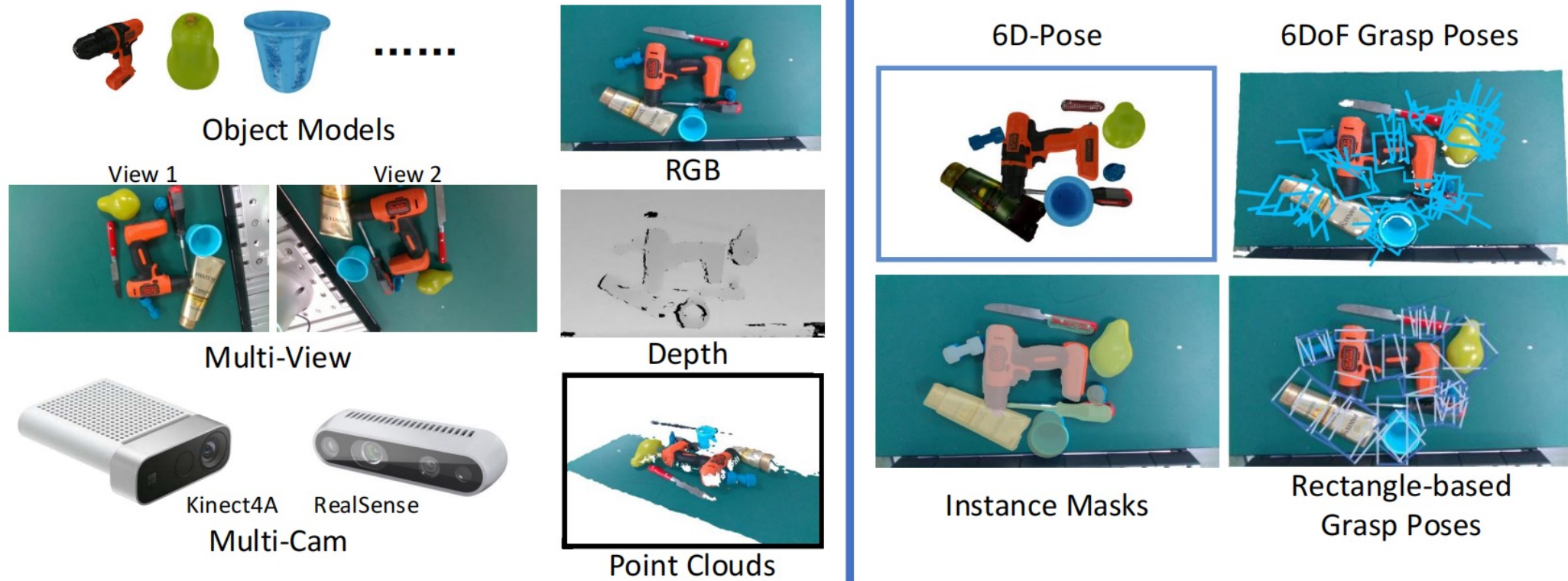
- **ACRONYM**: with grasping annotation



Grasp Dataset

Real dataset

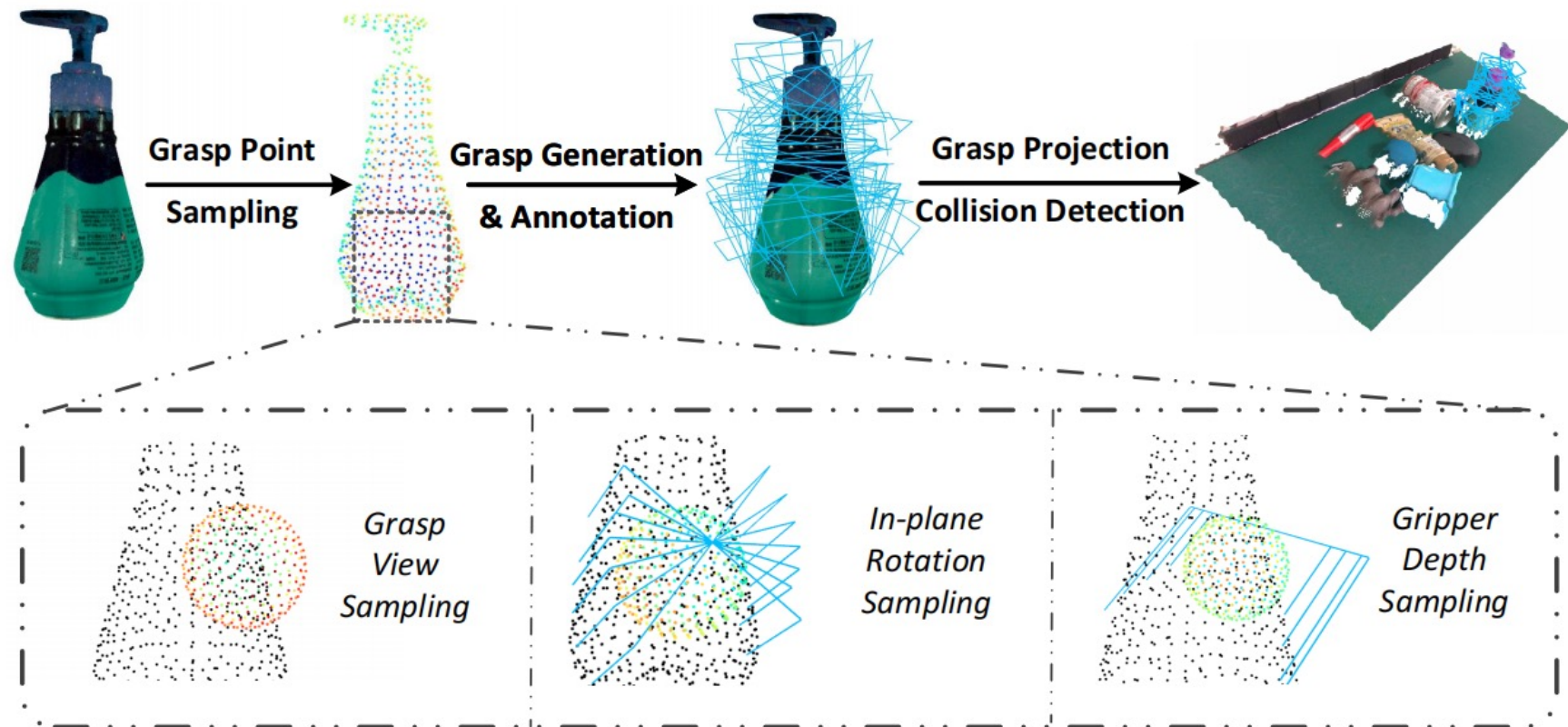
- **GraspNet-1Billion**: with grasping annotation



GraspNet-1Billion: How Grasp Annotations Generated?

Grasp pose annotation pipeline

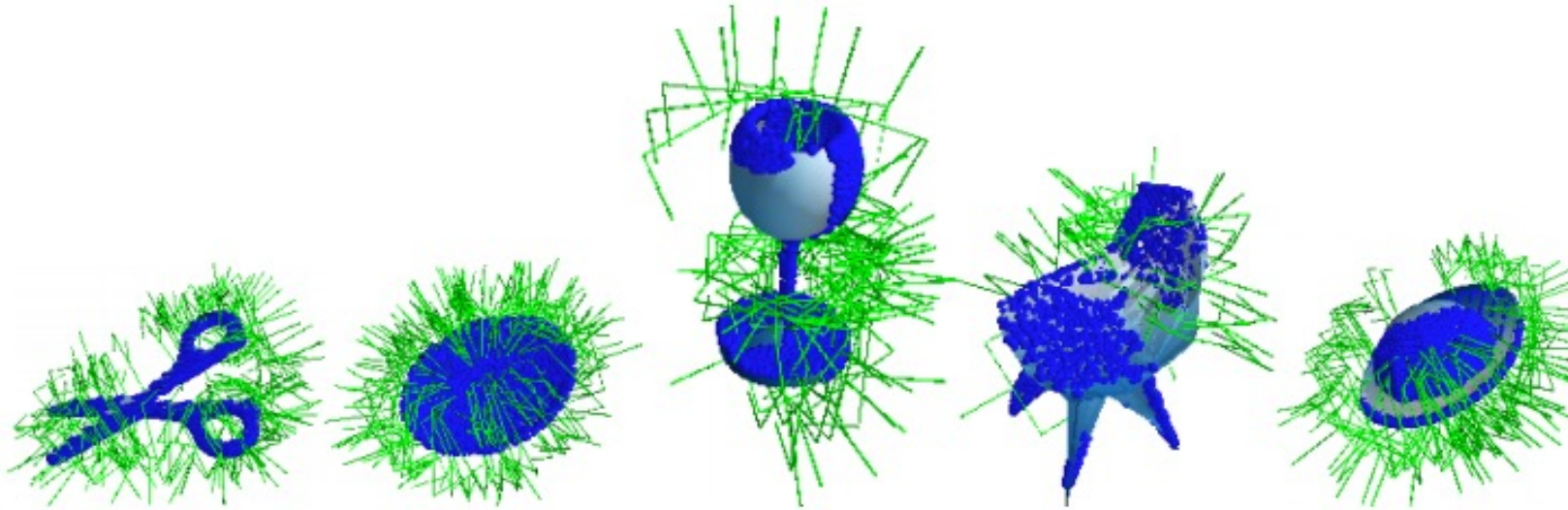
- The grasp point is firstly sampled from point cloud.
- Then the grasp view, the in-plane rotation and the gripper depth are sampled and evaluated. Finally, the grasps are projected on the scene using the 6D pose of each object.
- Collision detection is also conducted to avoid the collision between grasps and background or other object.



Grasp Detection

Grasp Detection

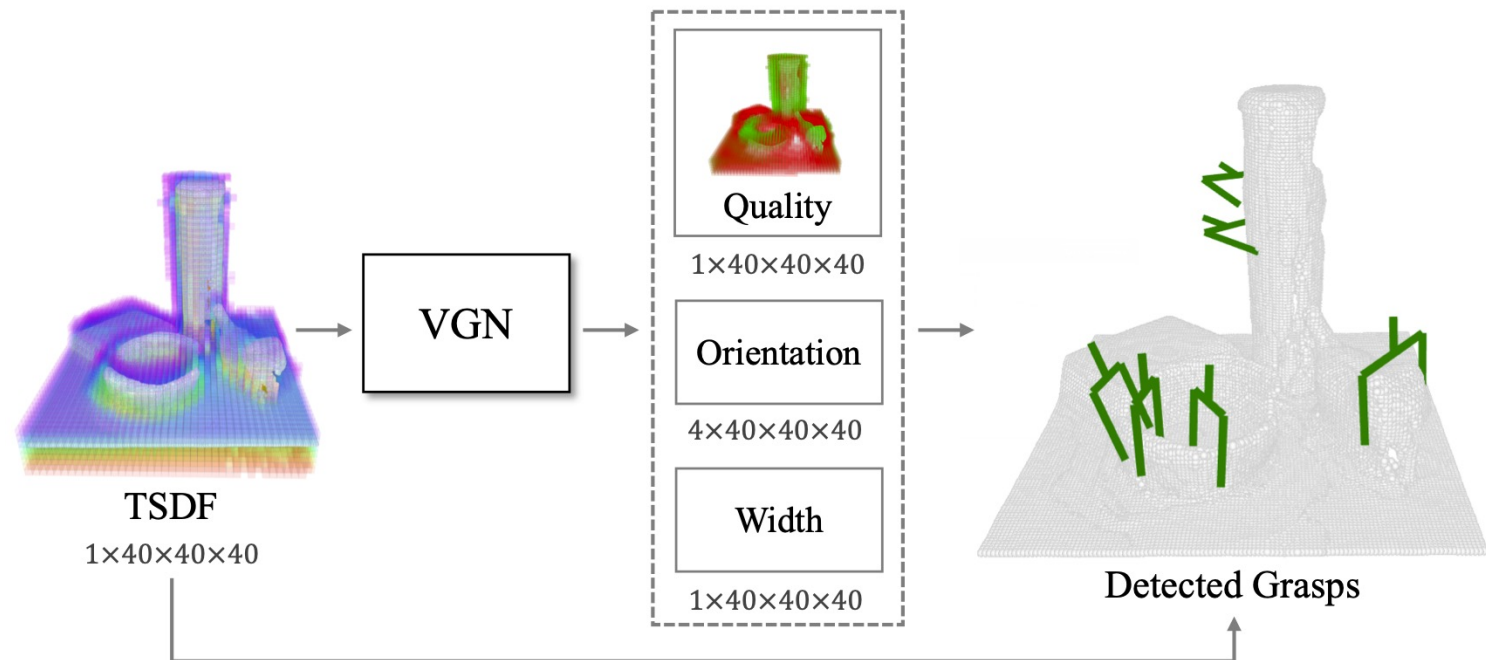
- Due to multi-modal grasp distribution, formulate grasping as a detection problem.



Visual Input Representation

Voxel Grids

- A voxel represents a value in a regular 3D grid. Voxel Grids are analogous to images in 3D space, where a voxel is similar to a pixel over a 3D grid instead of a 2D image
- Network: VGN
- Explicit geometry; limited by the volume resolution

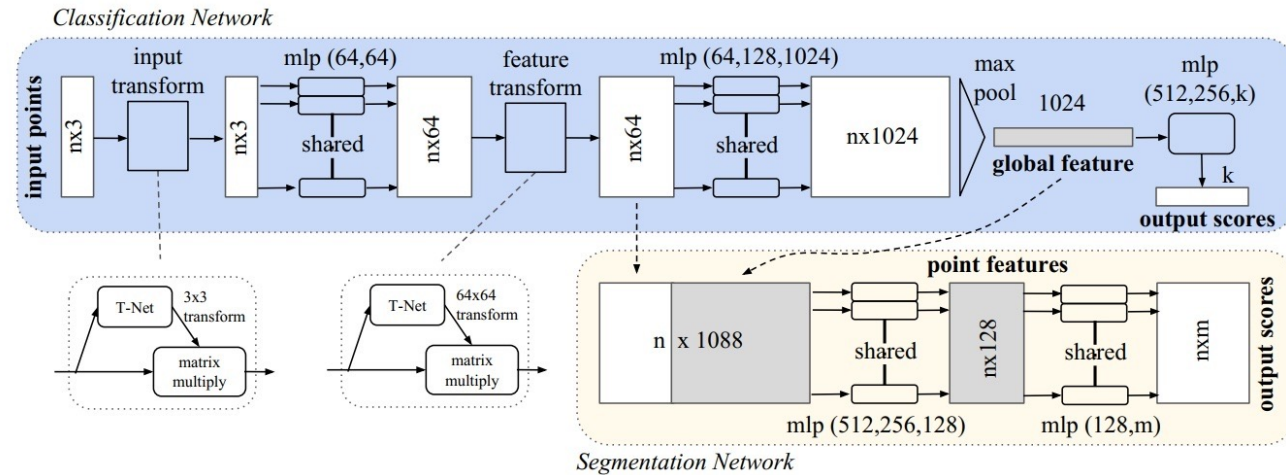


M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto. Volumetric Grasping Network: Real-time 6 DOF Grasp Detection in Clutter. CoRL 2020.

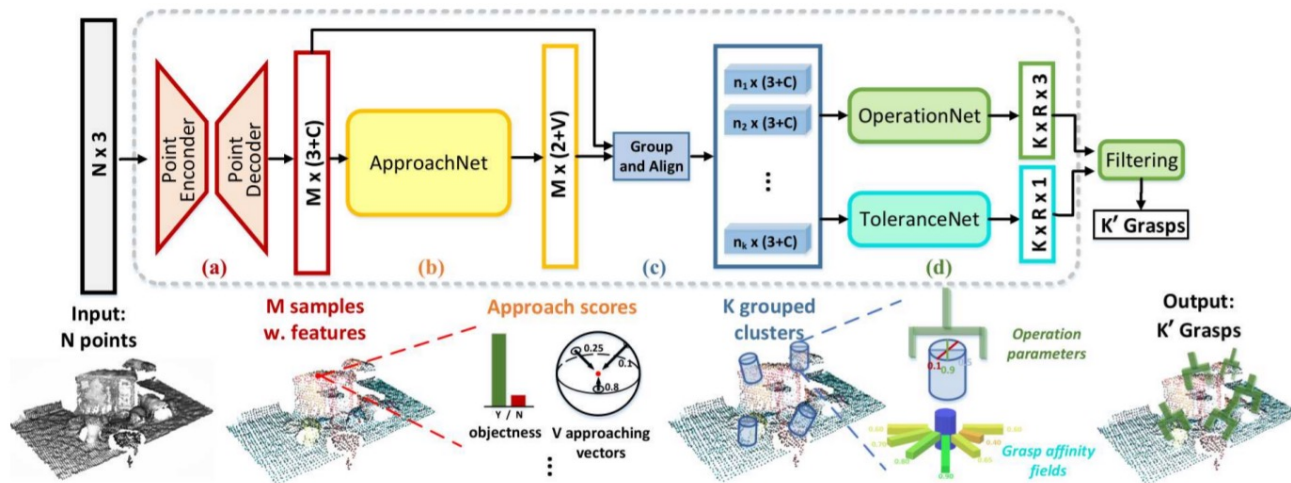
Visual Input Representation

Point Cloud

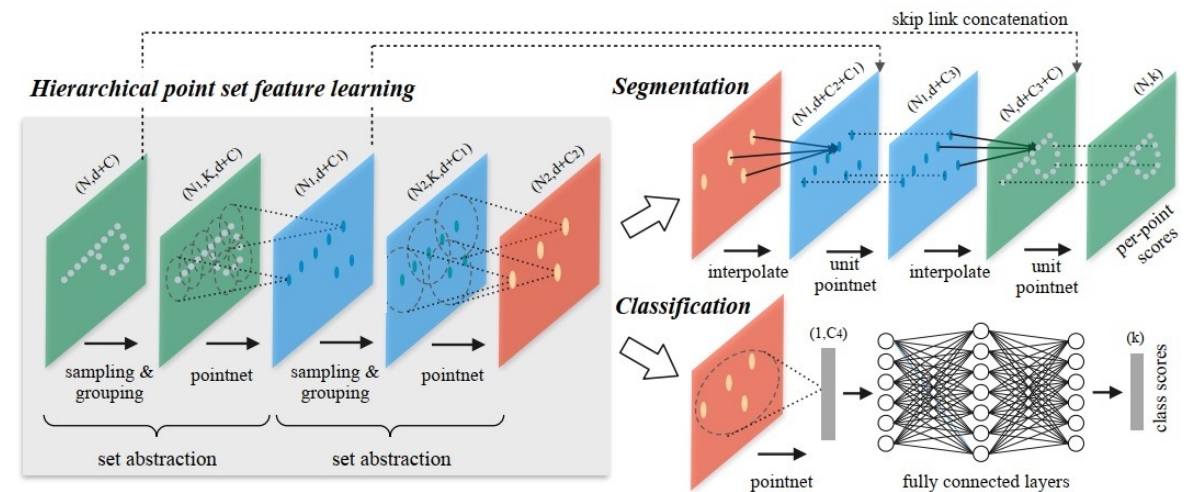
- Backbone: PointNet, PointNet++
- Network: GraspNet-baseline
- Explicit geometry; sensor noise, depth error (specular/transparent)



PointNet



GraspNet-baseline

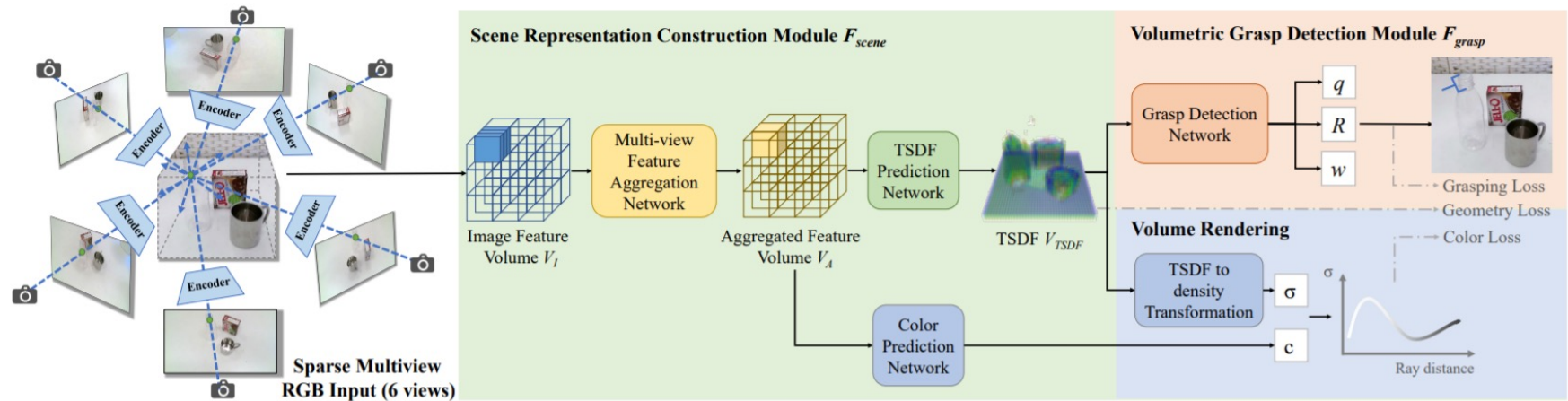
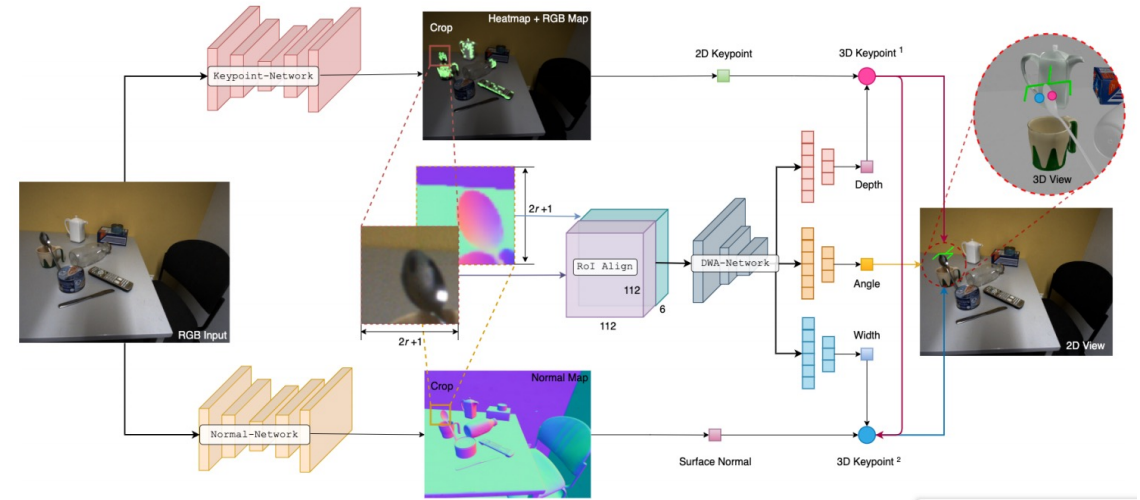


PointNet++

Visual Input Representation

Images

- RGB only: MonoGraspNet
- Multi-view RGB: GraspNeRF
- Implicit geometry



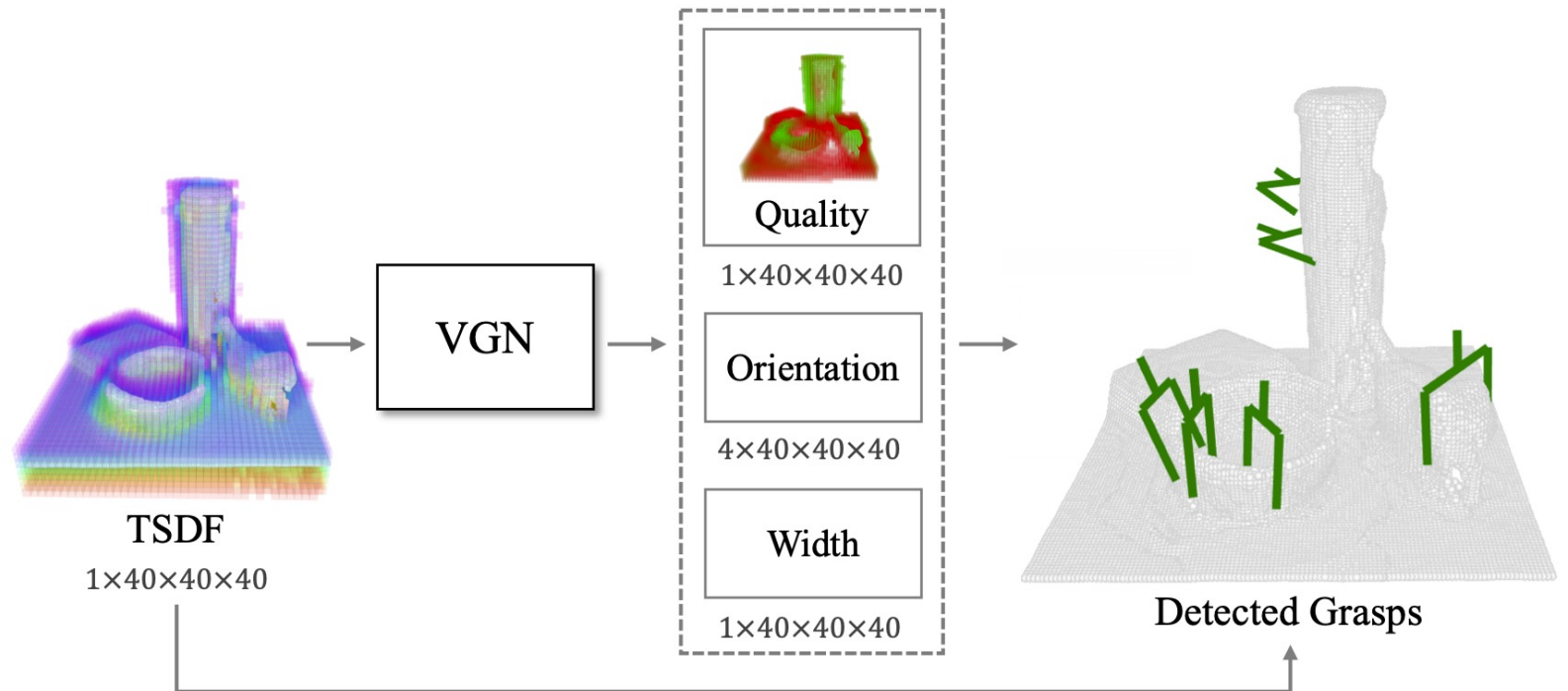
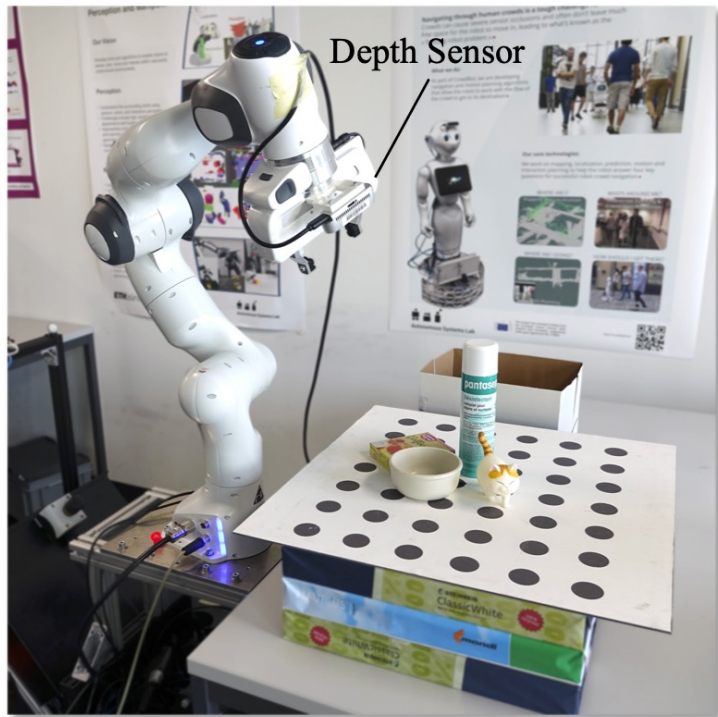
Zhai G, Huang D, Wu S C, et al. MonoGraspNet: 6-DoF Grasping with a Single RGB Image. arXiv:2209.13036, 2022.

Dai, Q., Zhu, Y., Geng, Y., Ruan, C., Zhang, J., and Wang, H. GraspNeRF: Multiview-based 6-DoF Grasp Detection for Transparent and Specular Objects Using Generalizable NeRF. arXiv:2210.06575.

Metrics

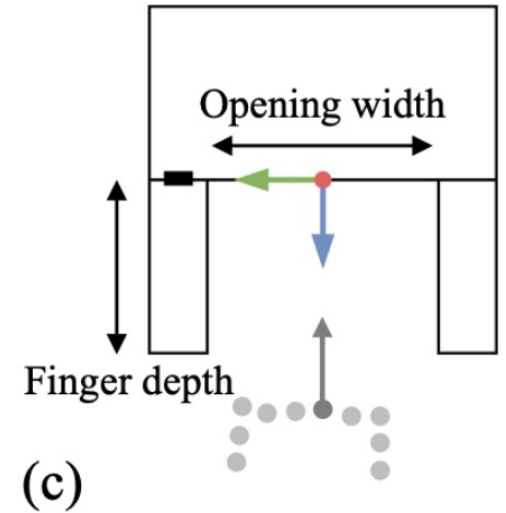
- **Success Rate**
 - the ratio of successful grasp executions
- **Percent cleared**
 - the percentage of objects removed during each round
- **Planning time**
 - the time between receiving input and returning grasps

Volumetric Grasping Network: Real-time 6 DOF Grasp Detection in Clutter (VGN)



How to define grasps?

- Define a grasp g by:
 - The position t (index in the volume)
 - Orientation r (quaternion)
 - The opening width w of the gripper
- Each pose is associated with a scalar quantity $q \in [0, 1]$ capturing the **probability of grasp success**
- The edges of the volume correspond to the boundaries of the grasping workspace



Detection Network

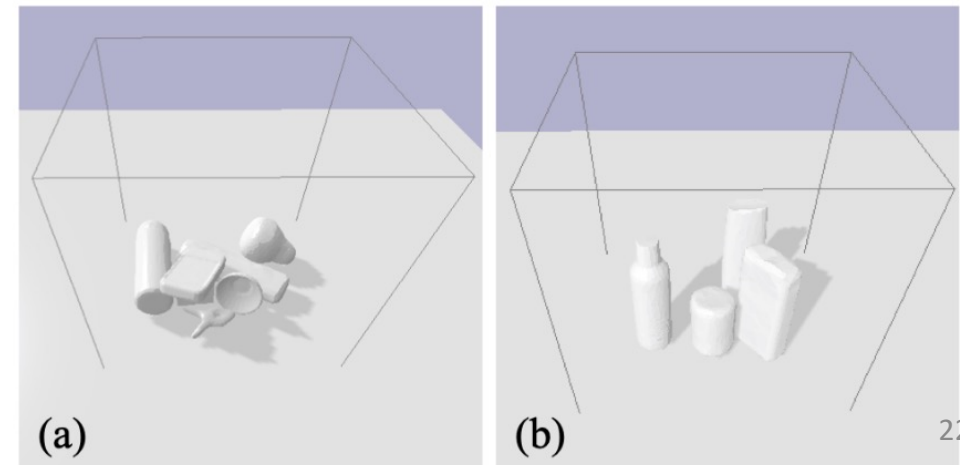
- Input is represented as a TSDF volume fused from multi-view depth maps
- The output is three volume, with each voxel contains quality, orientation and width respectively
- Follows a 3D FCN encoder-decoder architecture
- Training loss is similar to S4G

Post processing

- The grasp quality tensor is smoothed with a 3D Gaussian kernel which favors grasps in regions of high grasp quality
- Mask out voxels whose distance to the nearest surface is smaller than the finger depth
- Apply non-maxima suppression

Synthetic data generation

- Create virtual scenes for simulated grasping trials using two different strategies, “pile” and “packed” using PyBullet Simulator
- Sample a point and associated normal from the ground truth point cloud
- Test 6 different grasp orientations about the normal vector by spawning the gripper at the given pose
- Pose that lead to a collision with the environment are labeled as negative

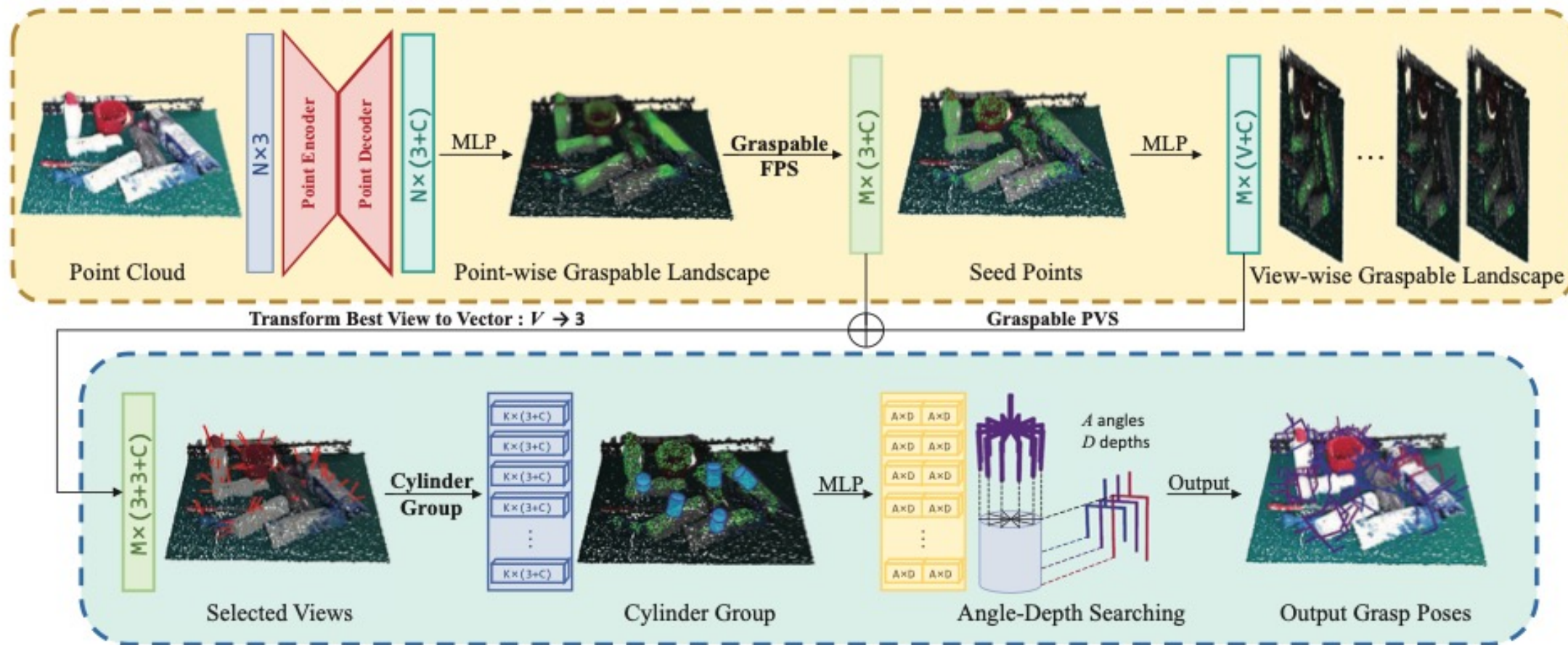


- Real demo



- More robust to occlusion thanks to multiview observation
- Limited to predefined workspace
- Limited to volume resolution

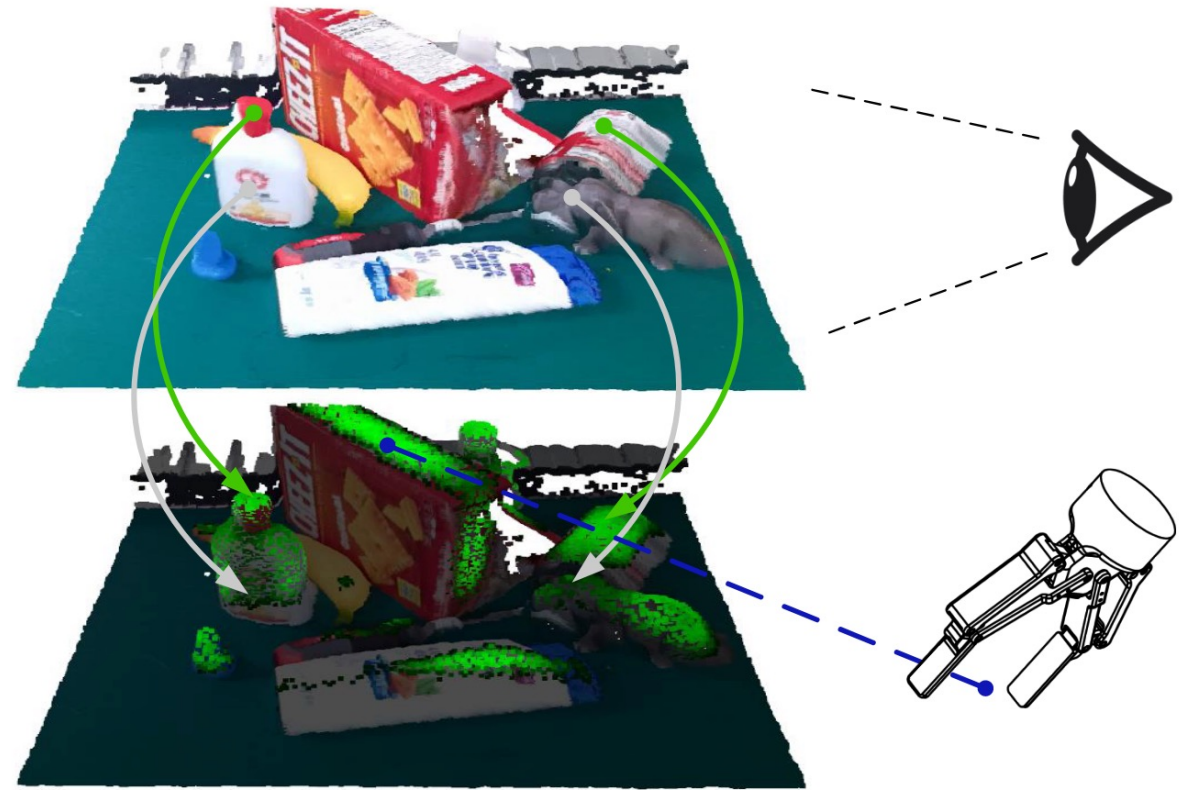
GS-Net



Graspness Discovery in Clutters for Fast and Accurate Grasp Detection (ICCV 2021)

GS-Net

- For grasp pose detection in the wild, it can be regarded as a two-stage problem.
- ✓ **Where stage** (locations with high graspability)
- ✓ **How stage** (decide grasp parameters like in-plane rotation, approaching depth, grasp score and gripper width)



Graspness Discovery

- A novel geometrically based quality, *graspness*, for distinguishing graspable area in cluttered scenes.
- For point sets P, V approach directions uniformly distributed in a sphere space.

- Point-wise graspness scores

$$\mathcal{S}^p = \{s_i^p | s_i^p \in [0, 1], i = 1, \dots, N\}$$

- View-wise graspness scores

$$\mathcal{S}^v = \{s_i^v | s_i^v \in [0, 1]^V, i = 1, \dots, N\}$$

Graspness Measure

- **Single Object Graspness.**

$$\tilde{s}_i^p = \frac{\sum_{j=1}^V \sum_{g \in \mathbf{G}_{i,j}} \mathbf{1}(g)}{\sum_{j=1}^V |\mathbf{G}_{i,j}|}, i = 1, \dots, N,$$
$$\tilde{s}_i^v = \left\{ \frac{\sum_{g \in \mathbf{G}_{i,j}} \mathbf{1}(g)}{|\mathbf{G}_{i,j}|} \mid 1 \leq j \leq V \right\}, i = 1, \dots, N.$$

- For each grasp, calculate a grasp quality score using a force analytic model

$$\tilde{s}_i^p = \frac{\sum_{j=1}^V \sum_{k=1}^L \mathbf{1}(q_k^{i,j} > c)}{\sum_{j=1}^V |\mathcal{G}_{i,j}|}, i = 1, \dots, N,$$
$$\tilde{s}_i^v = \left\{ \frac{\sum_{k=1}^L \mathbf{1}(q_k^{i,j} > c)}{|\mathcal{G}_{i,j}|} \mid 1 \leq j \leq V \right\}, i = 1, \dots, N.$$

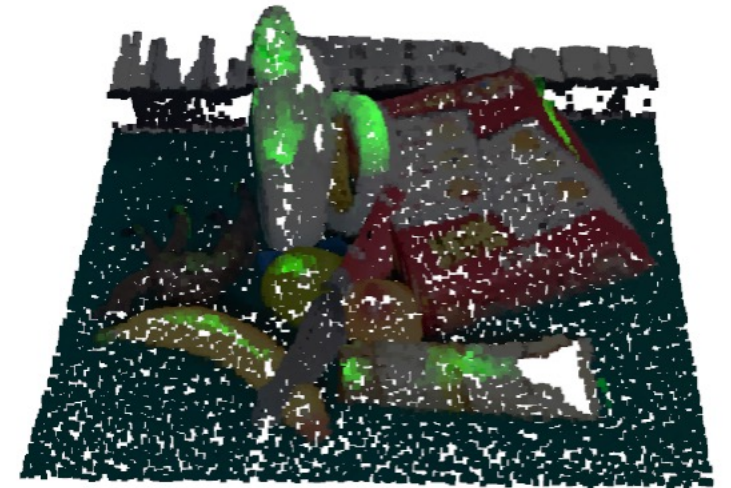
Graspness Measure

- **Scene-Level Graspness**

- ✓ Each grasp is evaluated by a collision checking process and assigned a collision label

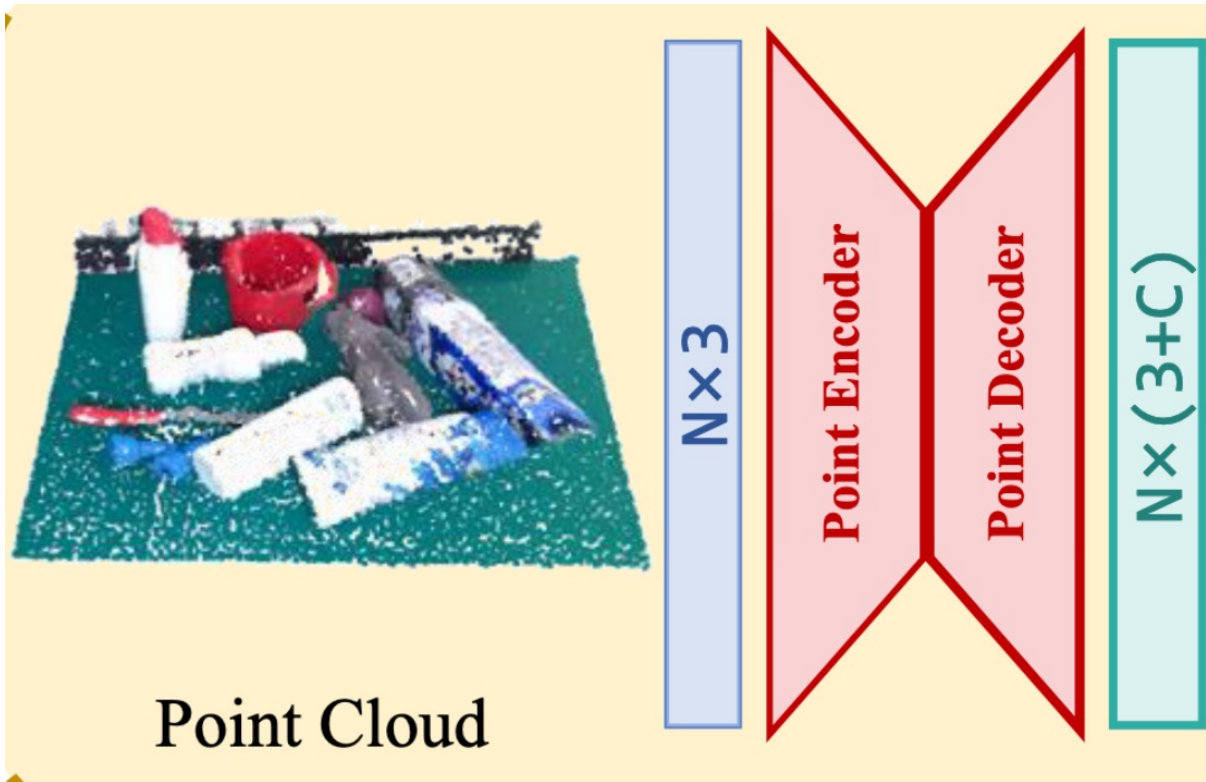
$$\tilde{s}_i^p = \frac{\sum_{j=1}^V \sum_{k=1}^L \mathbf{1}(q_k^{i,j} > c) \cdot \mathbf{1}(c_k^{i,j})}{\sum_{j=1}^V |\mathcal{G}_{i,j}|}, i = 1, \dots, N,$$

$$\tilde{s}_i^v = \left\{ \frac{\sum_{k=1}^L \mathbf{1}(q_k^{i,j} > c) \cdot \mathbf{1}(c_k^{i,j})}{|\mathcal{G}_{i,j}|} \mid 1 \leq j \leq V \right\}, i = 1, \dots, N.$$



- Project the object points to scene by object 6D poses.
- For each point in the scene, associate it with the nearest projected object point.

GSNet Architecture

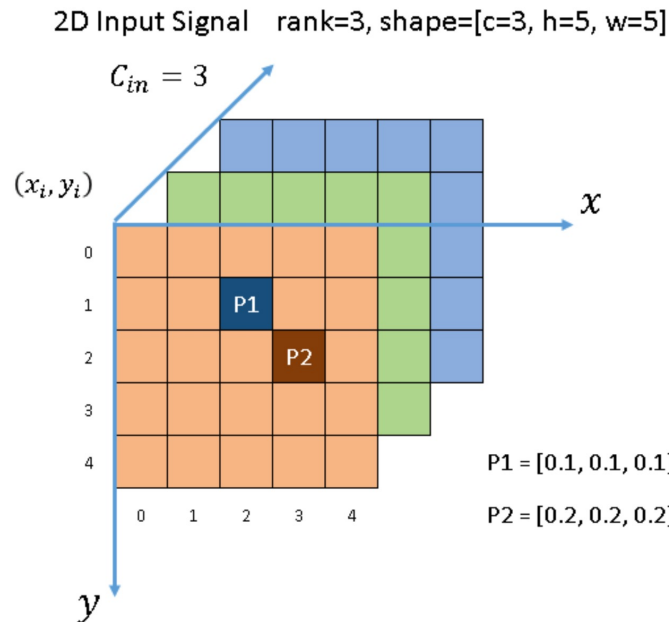
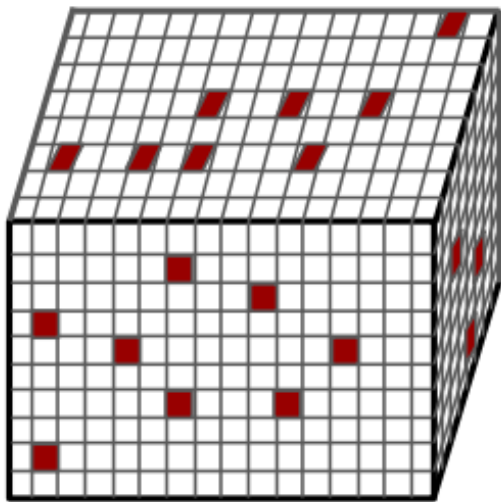


- **Backbone Network**

- ✓ ResUNet14 build upon MinkowskiEngine
- ✓ U-shape architecture with residual blocks
- ✓ For point cloud of size $N \times 3$, extract a C-channel feature, output a point set of size $N \times (3+C)$

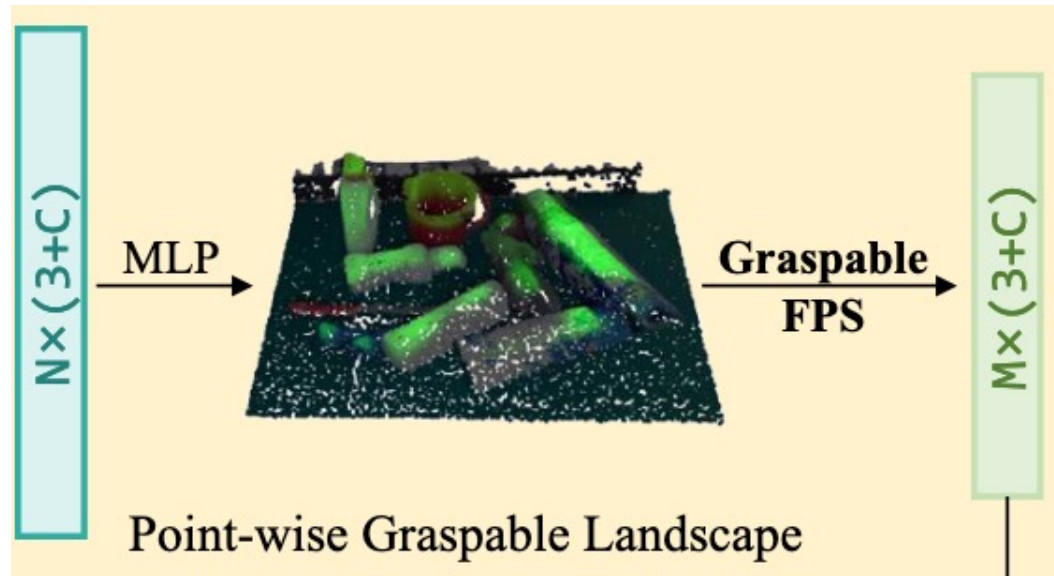
SparseConv

- Most voxels in a 3D point cloud are empty, which makes the point cloud data in 3D voxels typically a sparse signal.
- Compute convolutions on sparse data efficiently, instead of scanning all image pixels or spatial voxels.
- The convolution output is only computed when the kernel covers an active input site.



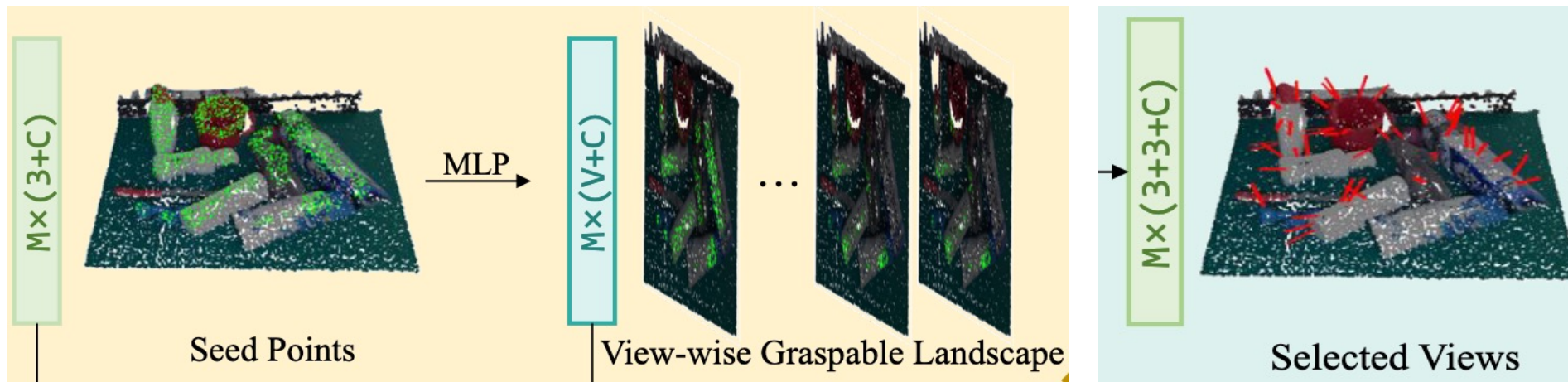
A1	A1A2	A1A2
A1	A1A2	A1A2
	A2	A2

GSNet Architecture



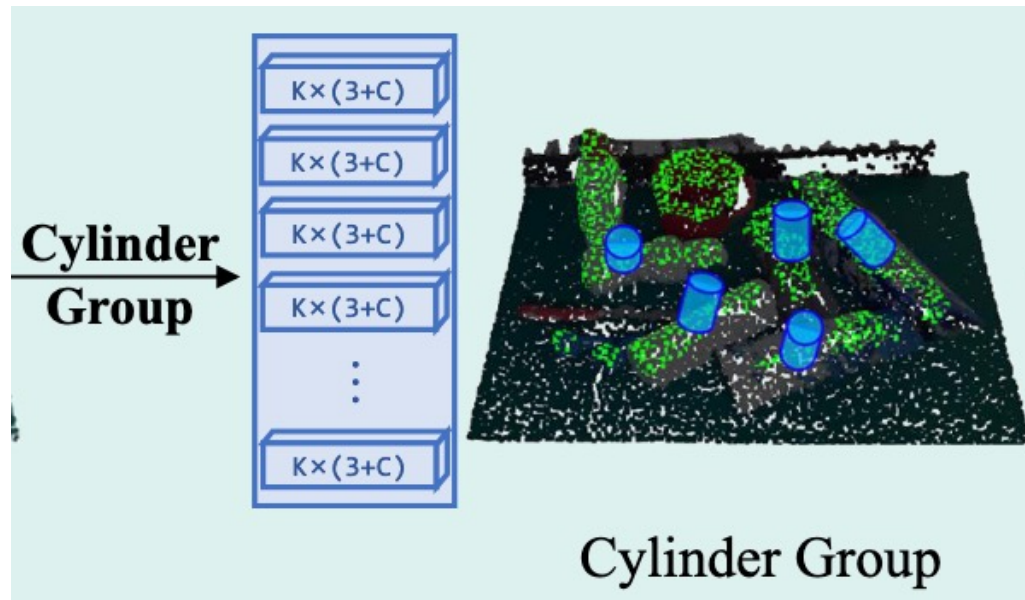
- **Graspable Farthest Point Sampling**
- an MLP generate point-wise graspable landscape, and output
 - ✓ graspable landscape ($N \times 1$)
 - ✓ binary objectness classification score ($N \times 2$)
- Select points with graspness score larger than threshold and adopt FPS to maximize distances among sampled points.
- **M seed points** are sampled with $(3+C)$ -dim features

GSNet Architecture



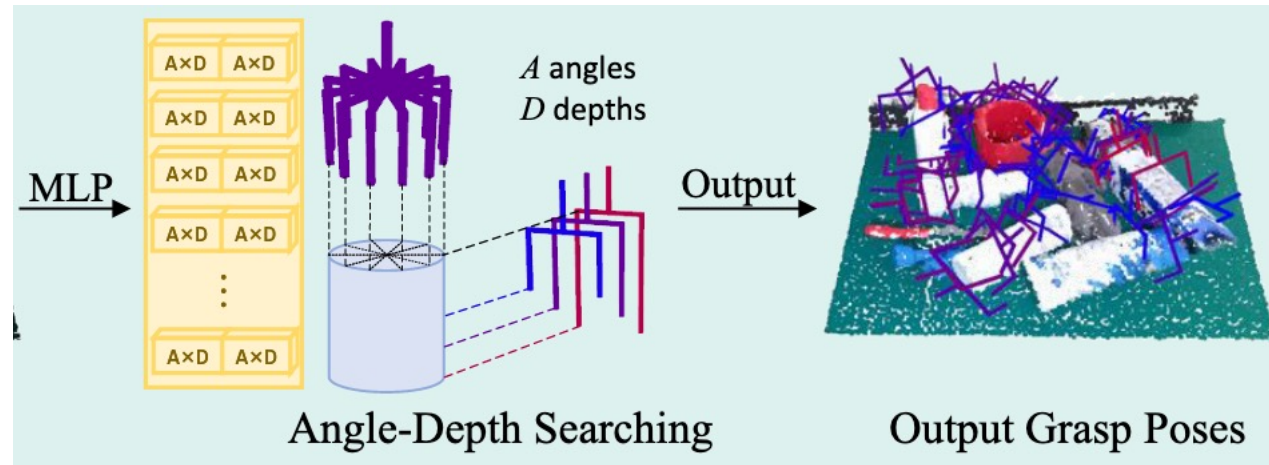
- **Graspable Probabilistic View Selection**
- V views are sampled from a unit sphere using Fibonacci lattices.
- An MLP output $M \times V$ vectors for view-wise graspable landscapes, and $M \times C$ residual features for grasp generation.
- Normalize the graspness scores of all views to $(0,1)$, probability score, and use probabilistic view selection(PVS) to get the best view.

GSNet Architecture



- **Cylinder-Grouping from Seed Points**
- For each of the M point-view pairs, group and sample K points using the cylinder with fixed height and radius.
- Grasp candidates of size $M \times K \times (3+C)$
 - ✓ point coordinates normalized by cylinder radius
 - ✓ feature vector

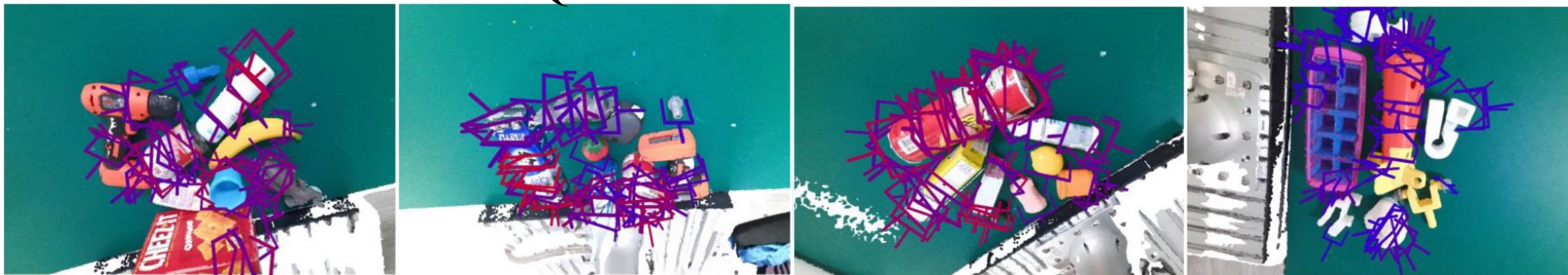
GSNet Architecture



- **Grasp Generation from Candidates**
- Grasp candidates are processed by an MLP and max-pooling layer.
- Output scores and widths for different (in-plane rotation)-(approach depth) combinations by an MLP.
- Output size is $M \times (A \times D \times 2)$ where A is the number of in-plane rotation angles, D is the number of gripper depths, and 2 is score and width.

GSNet

Qualitative results of GSNet



Real grasping experiments

IDs	#Objects	#Attempts	Success Rate
4, 10, 22, 32, 36, 57	6	6	100%
2, 38, 58, 59, 61, 69	6	7	85.7%
34, 37, 64, 66, 68, 72, 77	7	9	77.8%
0, 2, 23, 29, 39, 56, 62	7	7	100%
1, 10, 40, 41, 44, 48, 65, 69	8	8	100%
3, 9, 10, 23, 33, 42, 63, 68	8	9	88.9%
Total	42	46	91.3%

Table 7. Results of cluttered scene grasping. #Objects denotes the number of objects, and so does #Attempts.

Landscape types

Landscape	AP
object-level	55.33
scene-level	59.70

View graspness types

View Graspness	AP
mean score	50.62
max score	56.95
feasible ratio	59.70

Inference speed on GraspNet-1Billion. “CGM” is cascaded graspness model and “GOM” is grasp operation model.

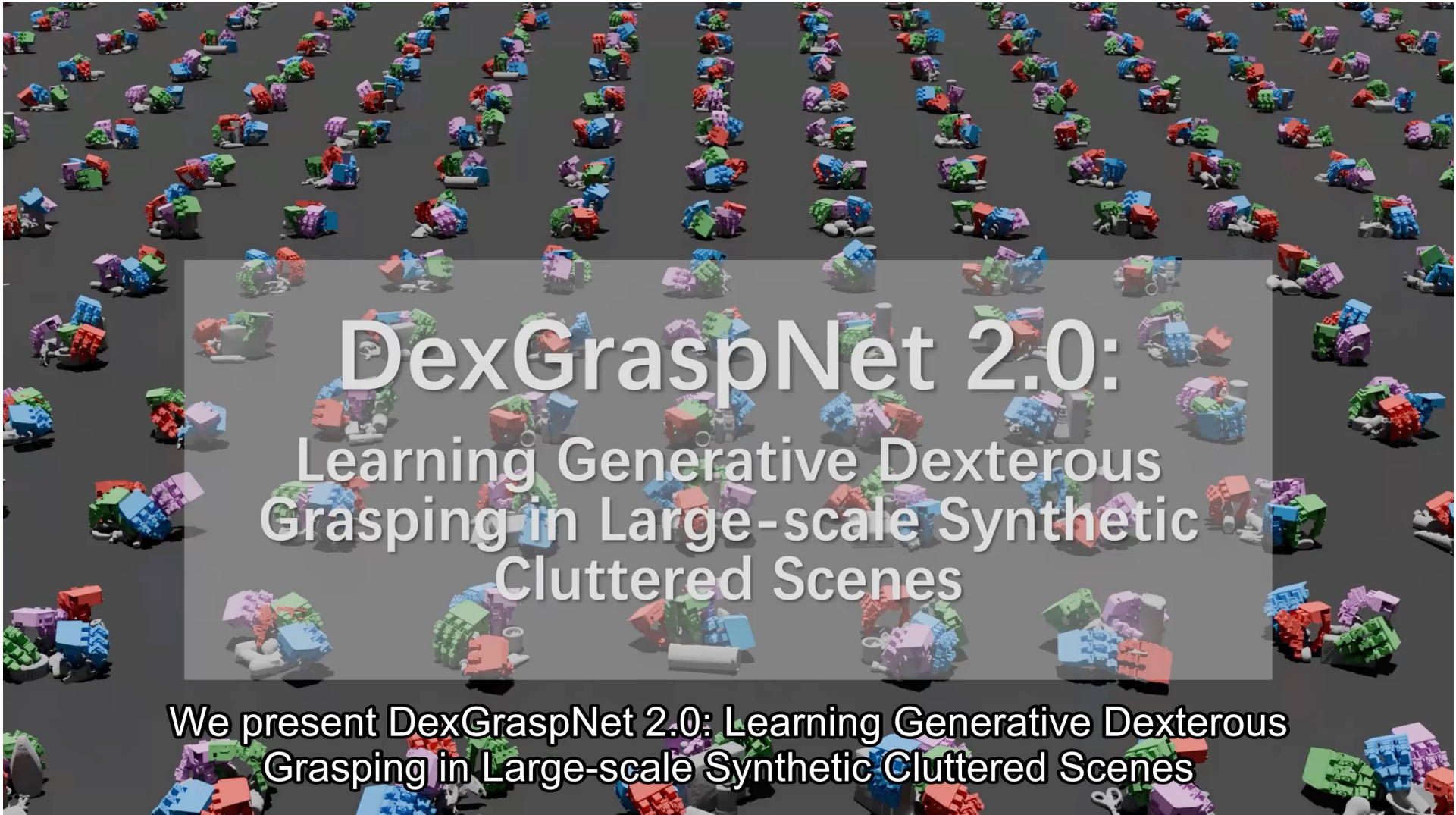
Camera	CGM	GOM	Total
RealSense	0.08s	0.02s	0.10s
Kinect	0.10s	0.02s	0.12s

➤ Conclusion

- A novel geometrically based quality named graspnes.
- A look-ahead searching method is adopted as the graspness measure.
- An end-to-end network is developed to incorporate graspness into grasp pose detection problem.
- An independent model learns the graspable landscapes.

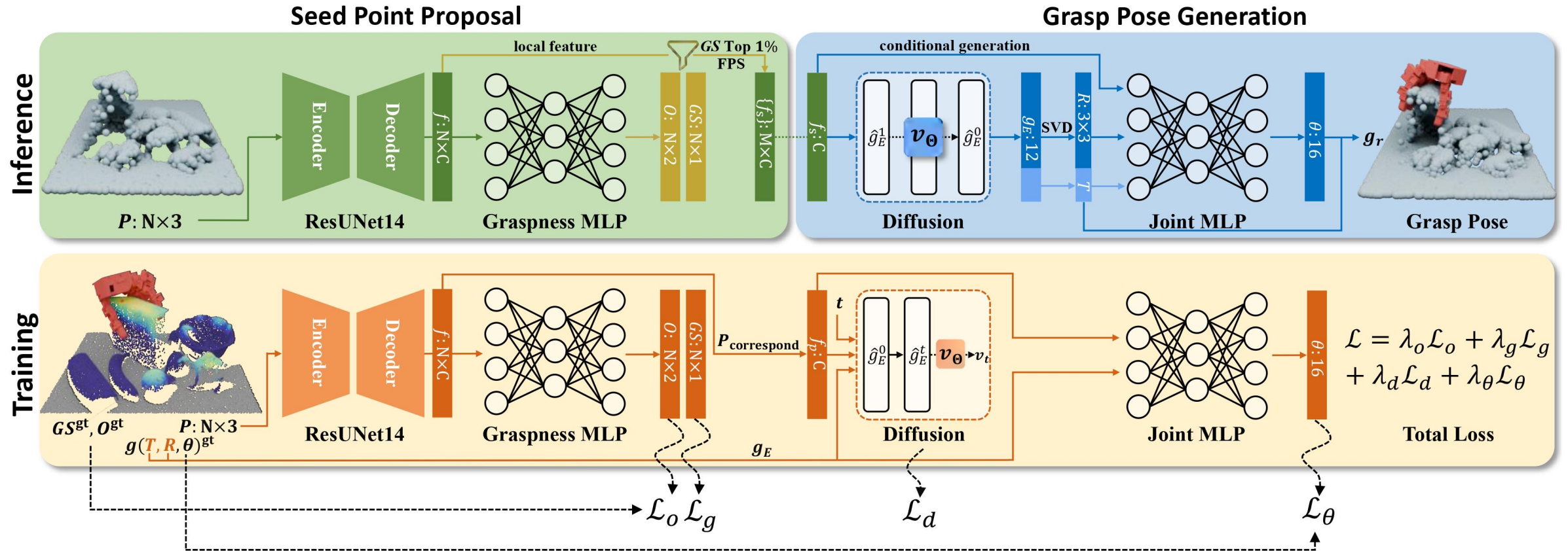
Conditional Grasp Generative Model

Data: DexGraspNet 2.0



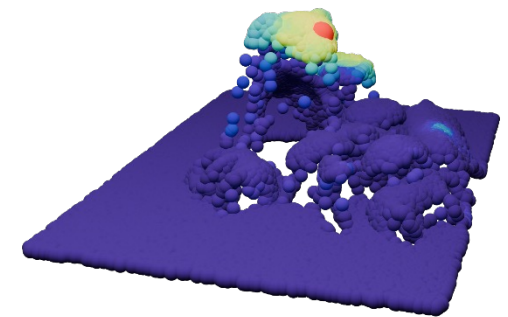
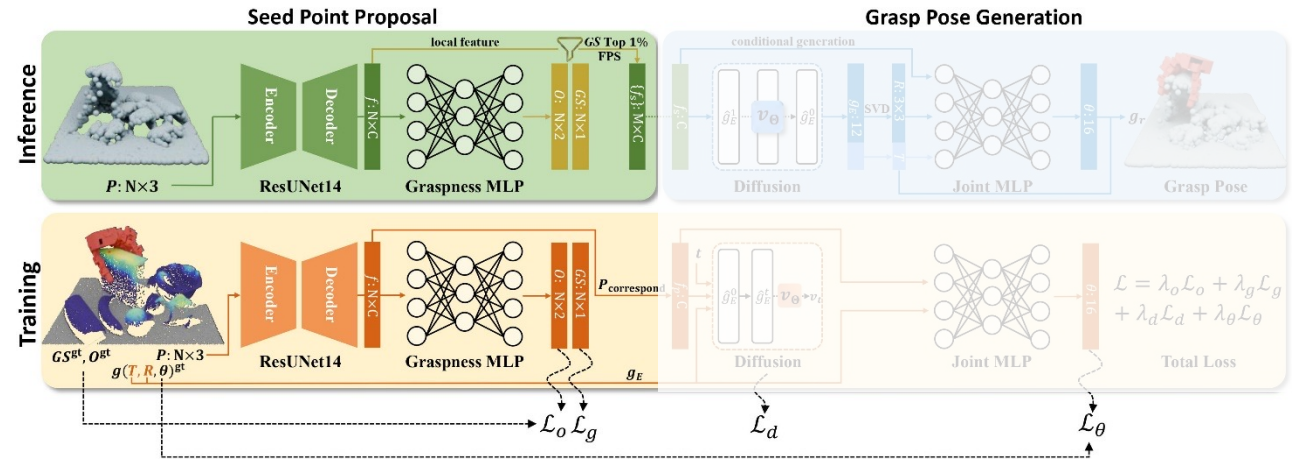
In submission.

Diffusion-based Dexterous Hand Pose Generation



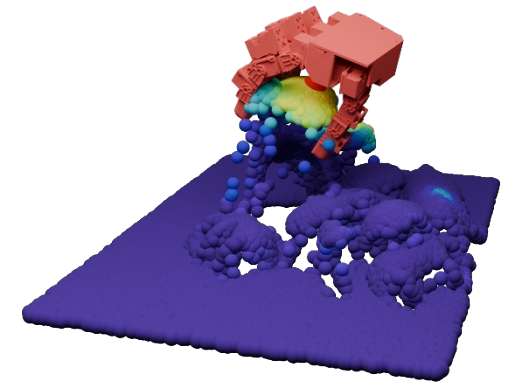
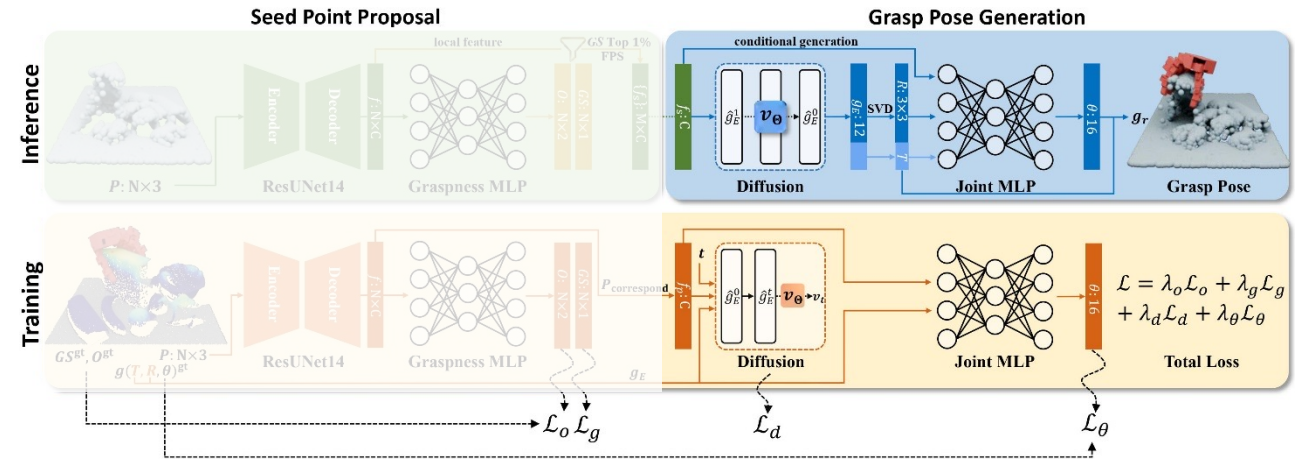
Method

- Stage 1: Where to grasp
 - Use sparse convolution to get the **local feature**
 - Predict graspness and objectness scores
 - Accumulated with nearby gt seed points

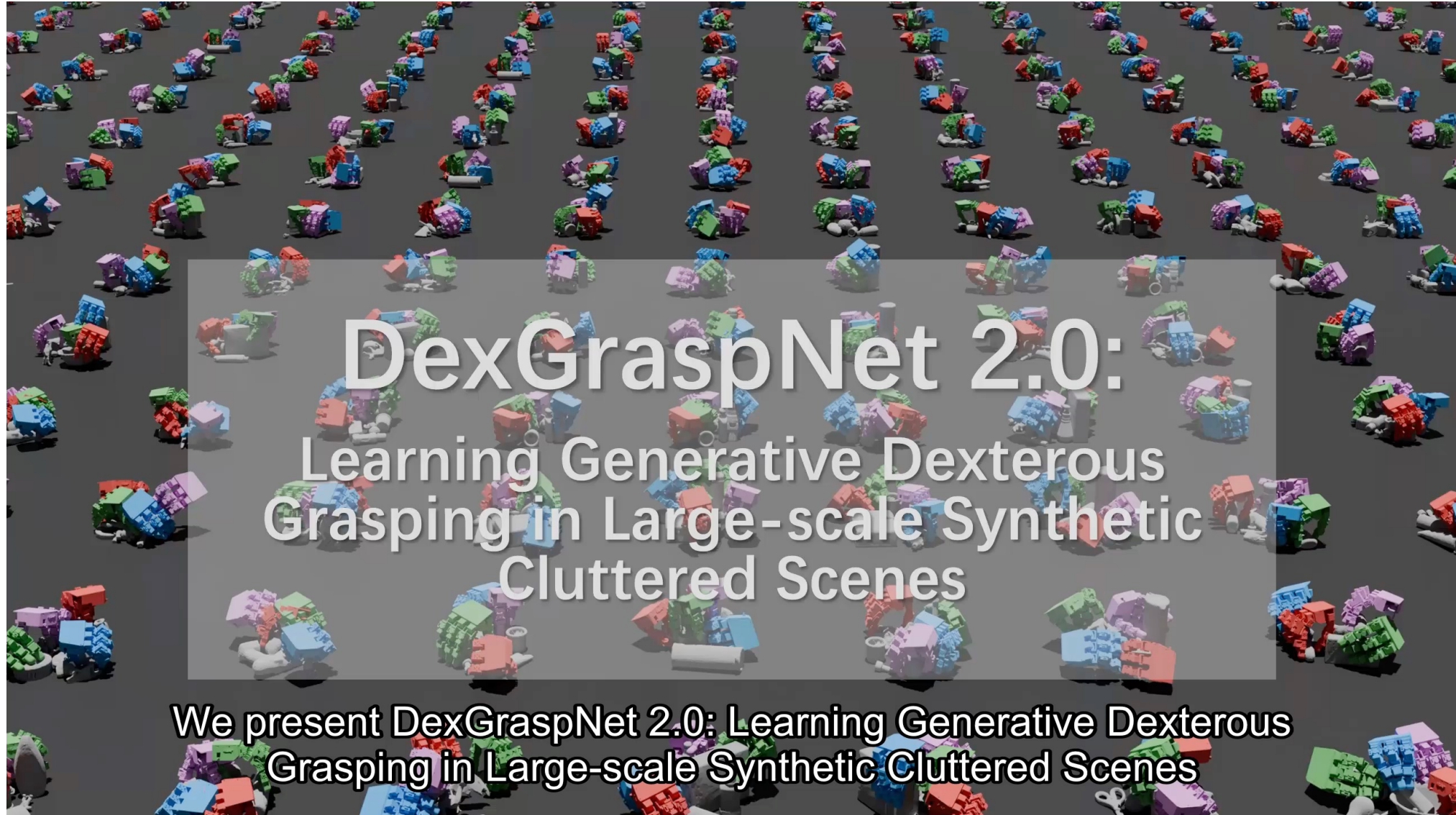


Method

- Stage 2: How to grasp
 - Use only the **local feature** in stage 1
 - Predict residual position T , rotation R , finger joint angles θ
 - To handle the multi-modal grasping pose distribution, we use diffusion model for T and R
 - Then predict θ directly



Dexterous Grasping Pose Diffusion Model



Experiments

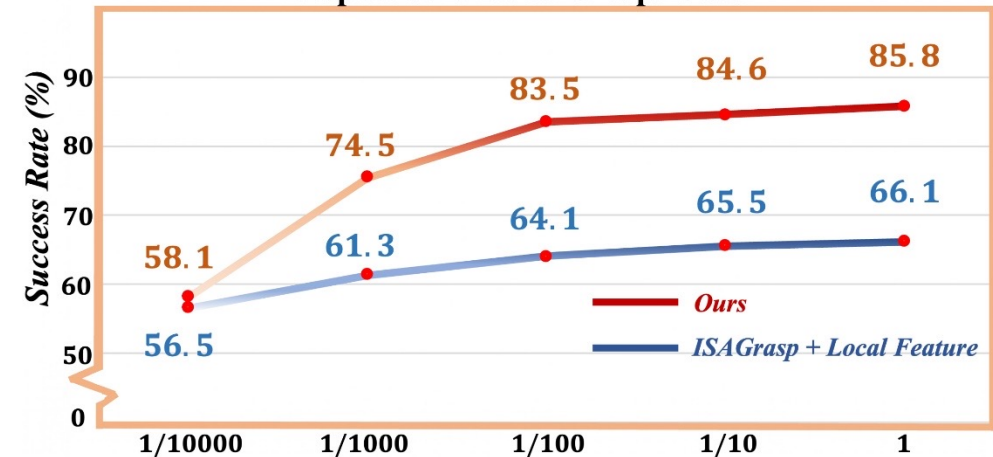
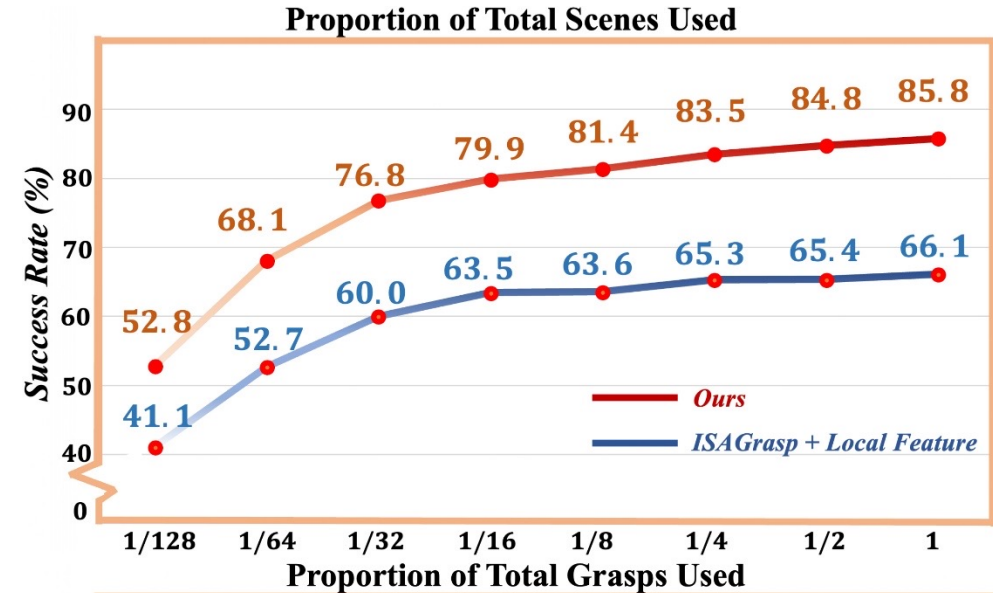
- Achieve SOTA on DexGraspNet 2.0 benchmark

	Method	GraspNet-1Billion			ShapeNet		
		Dense	Random	Loose	Dense	Random	Loose
Baseline	HGC-Net [8]	46.0	37.8	26.7	46.4	44.8	30.4
	GraspTTA [†] [11]	62.5	54.1	42.8	56.6	57.8	46.4
	ISAGrasp [†] [5]	63.4	60.7	51.4	64.0	56.3	52.7
Ablation	local feature	16.8	10.9	4.8	21.3	17.6	10.9
	decomposed model	84.2	80.7	71.3	74.9	72.5	66.4
	random scene	90.0	84.1	68.2	78.9	78.8	71.3
	Ours	90.6	83.7	73.2	81.0	85.4	74.2

Table 1: **Baseline and ablation studies for dexterous grasping.** Modified baseline methods are indicated with [†]. Ablation studies are conducted on three aspects as shown in the lower half of the table. Each **Dense** scene contains 8-11 objects, and each **Random** scene contains 1-10 objects, obtained by deleting objects from Dense scenes, and each **Loose** scene contains 1-2 objects.

Scaling Law of Grasp Data

- We explore the impact of scaling the number of grasping poses and scenes
- Significant performance degradation with less poses/scenes



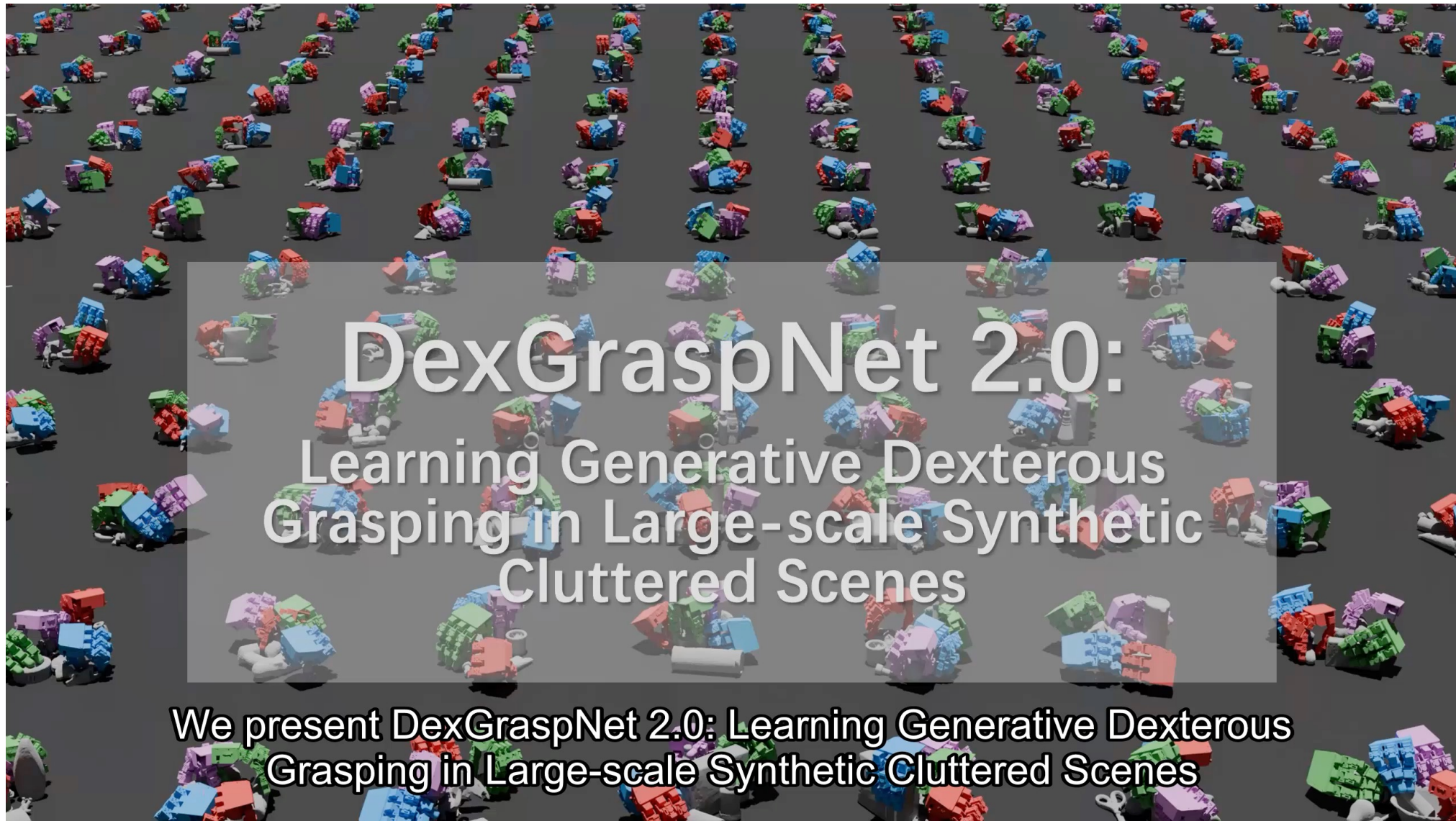
Experiments

- Our method can be used for robots with varying morphologies
- For parallel grippers, our method can achieve comparable performance with GSNet and outperform others

Method	AP Seen	AP Similar	AP Novel
GG-CNN [31]	15.5/16.9	13.3/15.1	5.5/7.4
Chu <i>et al.</i> [32]	16.0/17.6	15.4/17.4	7.6/8.0
GPD [33]	22.9/24.4	21.3/23.2	8.2/9.6
PointGPD [34]	26.0/27.6	22.7/24.4	9.2/10.7
GraspNet1B[15]	27.6/29.9	26.1/27.8	10.6/11.5
GSNet [13]	67.1/63.5	54.8/49.2	24.3/19.8
GSNet [13]*	68.7	59.8	24.6
Ours*	56.0	53.2	23.2
Ours # *	66.6/61.5	61.7/53.3	27.4/23.2
GSNet (render)*	78.4	69.7	36.7
Ours (render) # *	77.4	72.5	39.1

Table 2: **Experiment results on GraspNet-1Billion.** * means using grasps that scores over ≥ 0.9 . # means using refined poses, and render means using rendered depth images. Statistics in the table follows a **RealSense/Kinect** format, where results with a single number uses Realsense setting.

Real Experiments

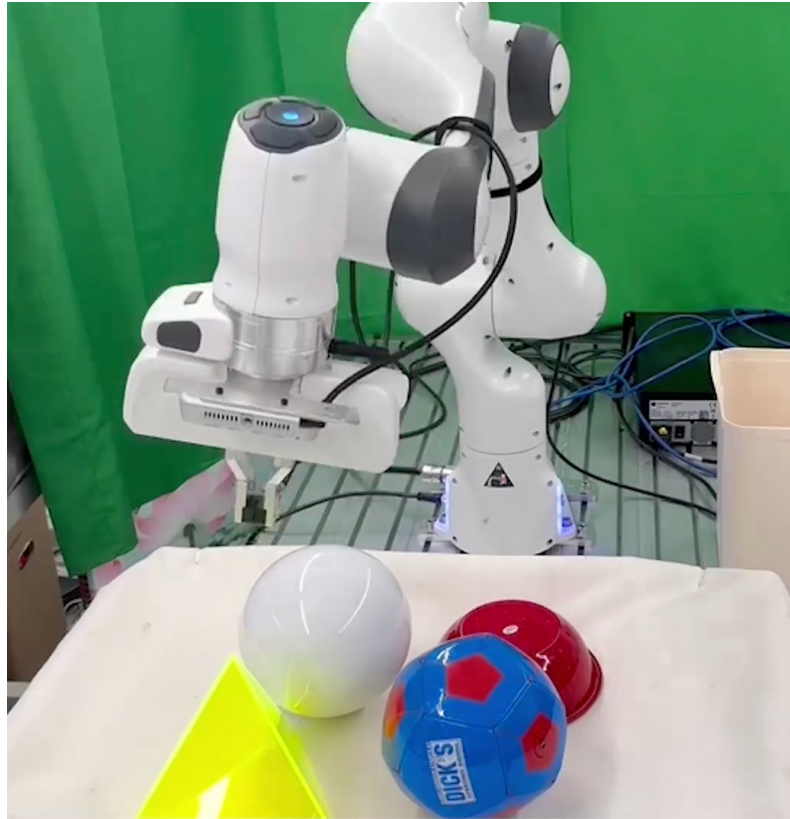
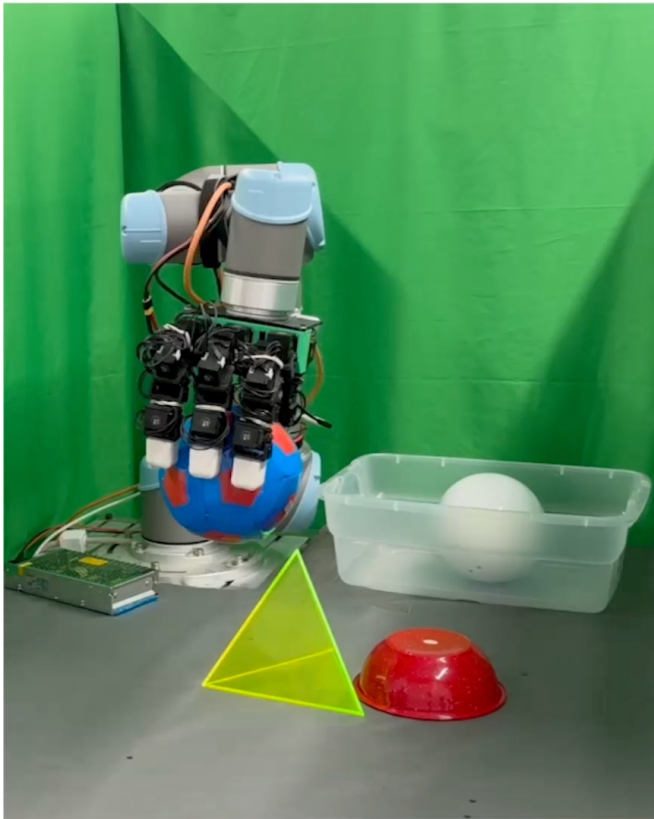


Real Experiments



Compare to gripper

The dexterous hand has stronger wrapping ability



End Effector	Normal	Large
Parallel Gripper	92.4	0.0
Dexterous Hand	81.5	100.0

Summary

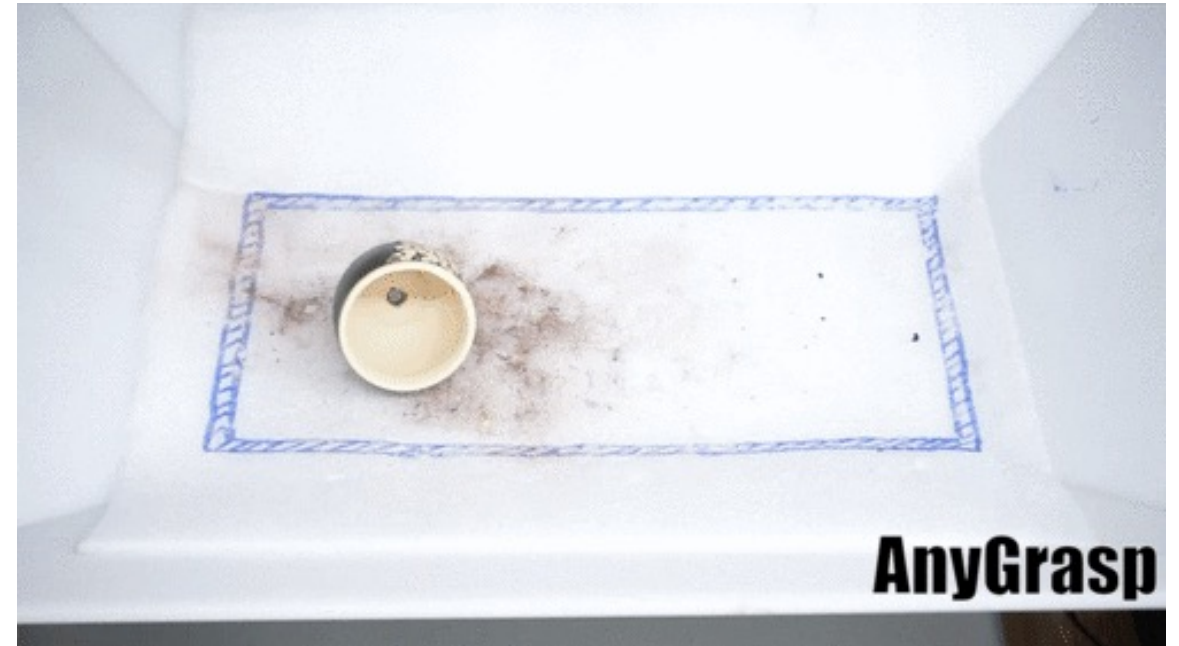
➤ Main contribution:

- **Large-scale Benchmark** for dexterous grasping in cluttered scenes
7500 synthetic training scenes and 670 test scenes.
- Propose an **end-to-end generative grasp prediction pipeline**
generates wrist 6D pose via diffusion model.
regress hand joint poses conditioned on wrist pose.

➤ Limitations:

- **Limited to power grasping**
- **Data only covers force closure cases.**

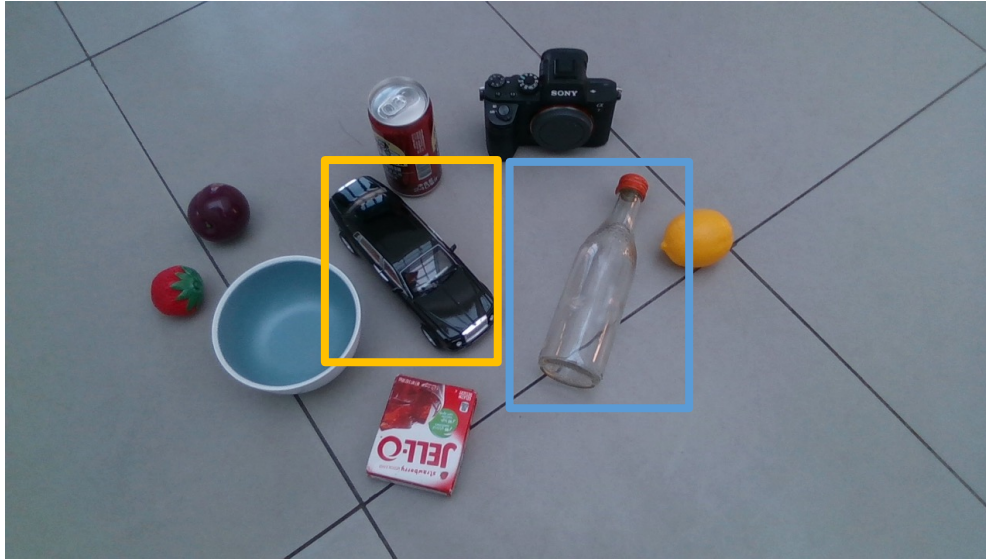
Generalizable Grasping for Diffuse Objects



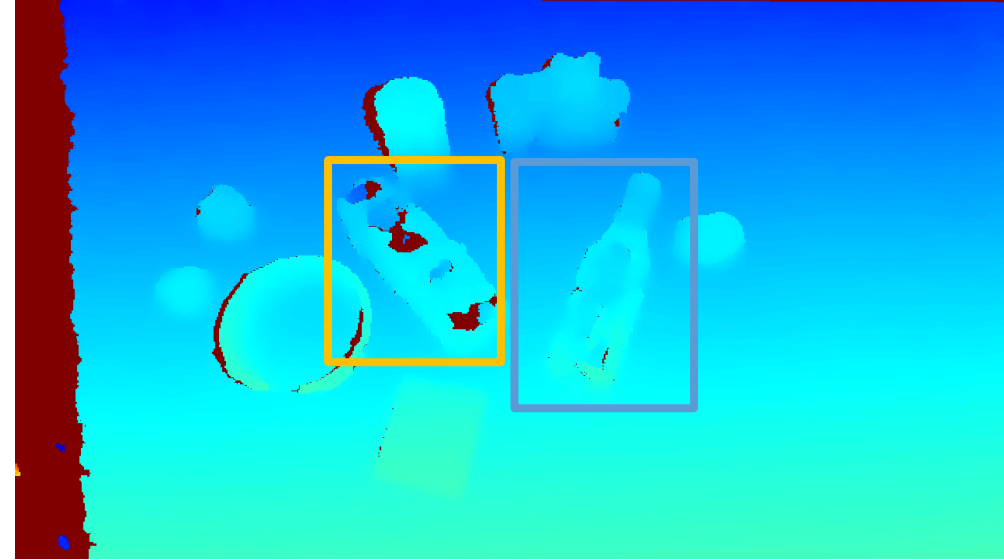
Fang, Hao-Shu, et al. "Graspnet-1billion: A large-scale benchmark for general object grasping." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

Depth Sensing Problem

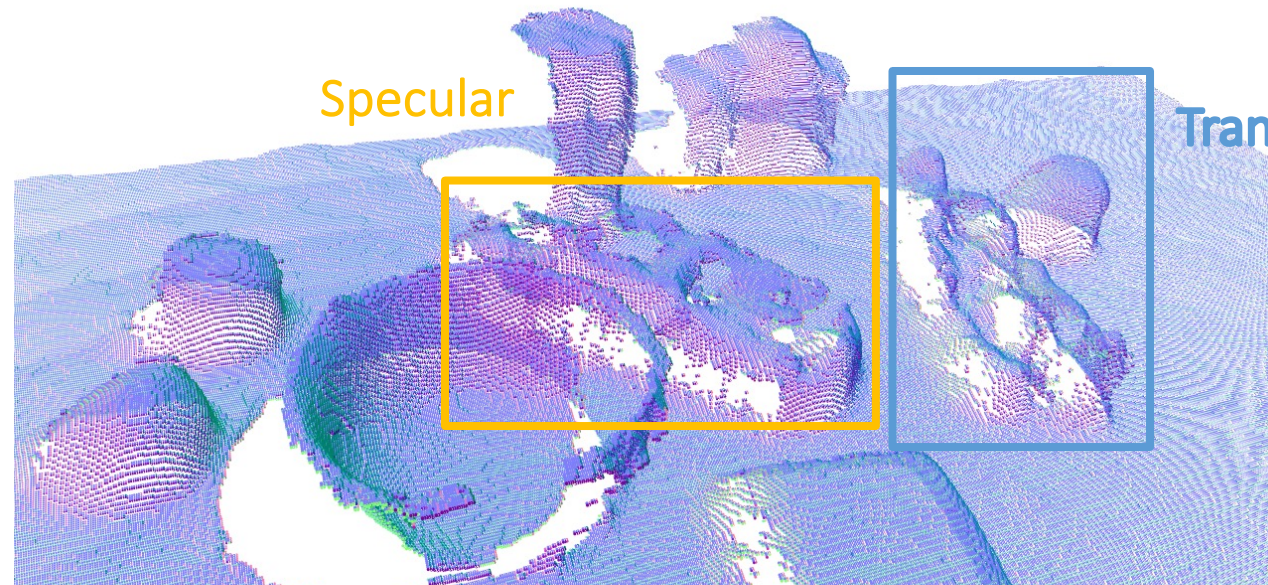
RGB



Depth



Point cloud



Specular

Transparent

Demo: Robotic Grasping

GraspNet-baseline
only

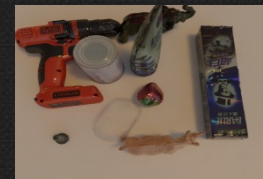
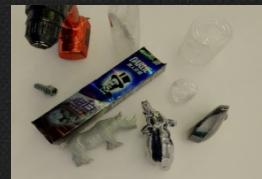
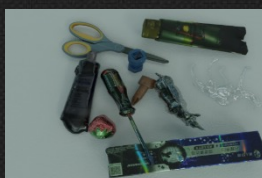
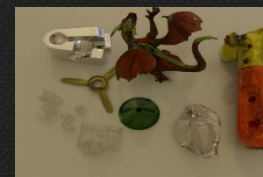


Demo: Robotic Grasping

Refined point cloud
+
GraspNet-baseline



Data Generation (STD-GraspNet)



DREDS: Domain Randomization

- Randomization of object layout: object instances, arrangement



DREDS: Domain Randomization

- Randomization of object material: BSDF



DREDS: Domain Randomization

- Randomization of background



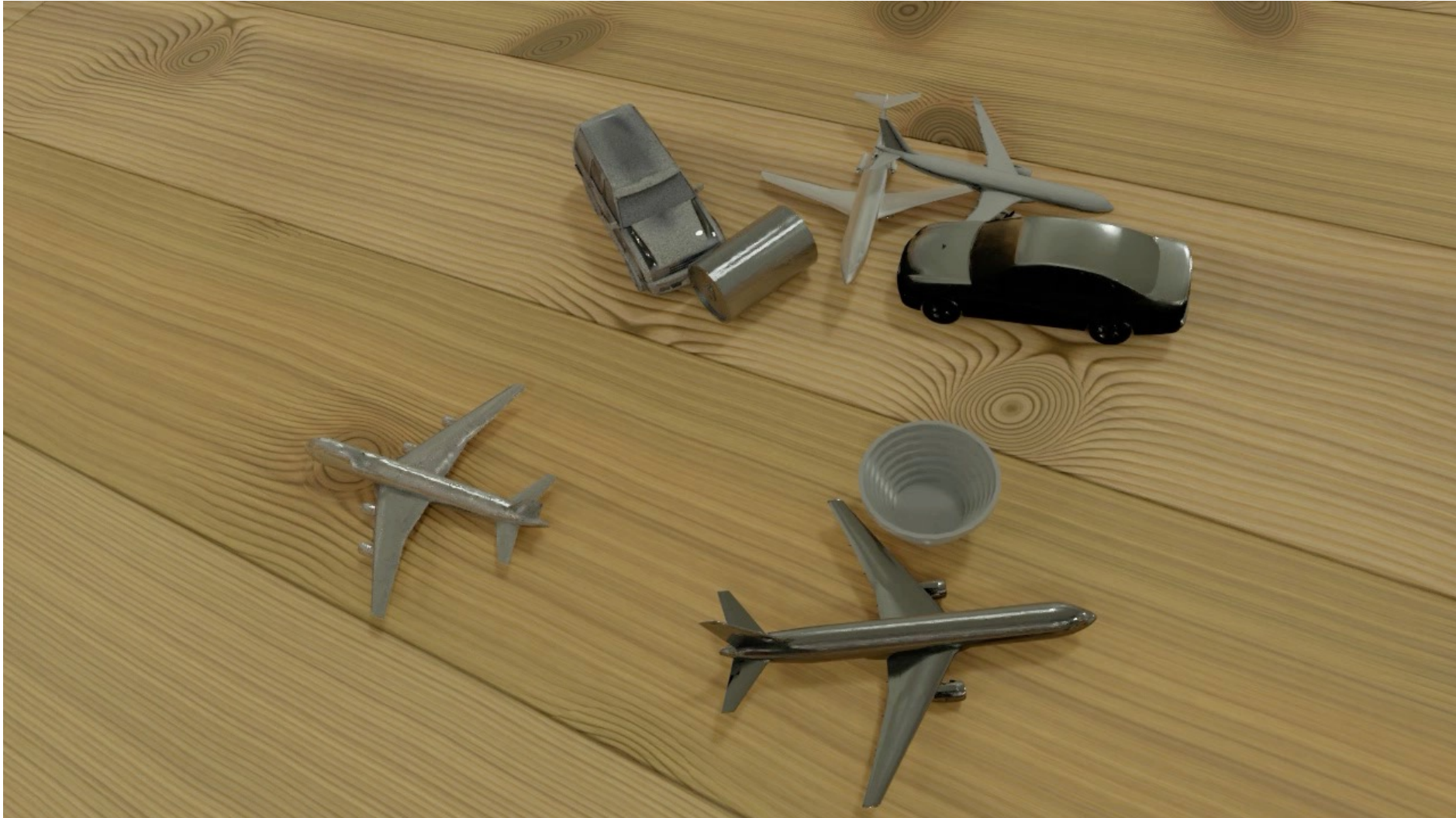
DREDS: Domain Randomization

- Randomization of illumination



DREDS: Domain Randomization

- Randomization of camera viewpoint





ICRA2024
YOKOHAMA | JAPAN

ASGrasp: Generalizable Transparent Object Reconstruction and 6-DoF Grasp Detection from RGB-D Active Stereo Camera

Jun Shi¹, Yong A¹, Yixiang Jin¹, Dingzhe Li¹, Haoyu Niu¹, Zhezhu Jin¹, He Wang^{†2,3,4}

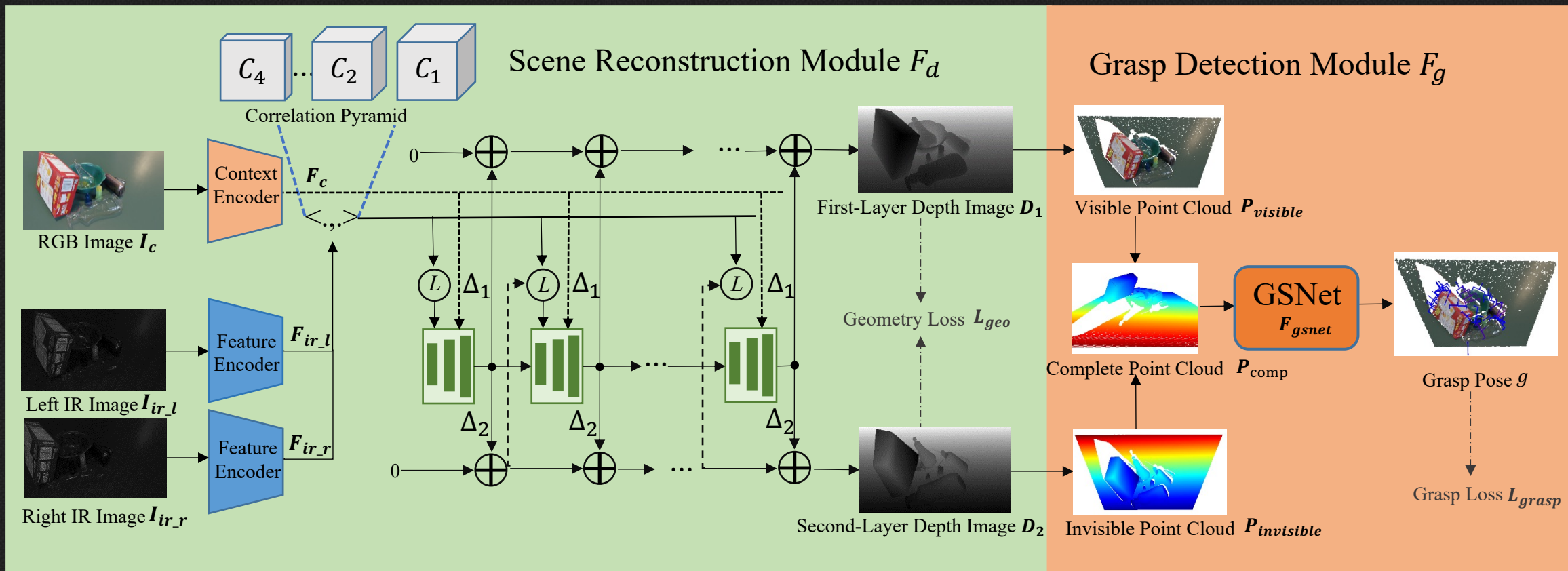
¹ Samsung R&D Institute China, Beijing (SRC-B)

² CFCS, School of Computer Science, Peking University ³ Galbot

⁴ Beijing Academy of Artificial Intelligence (BAAI)

†: corresponding author

ASGrasp Pipeline



World-wide First Material-Agnostic Grasping Method



With a success rate $> 95\%$

Speed X10

Grasp Poses

RViz Visualization

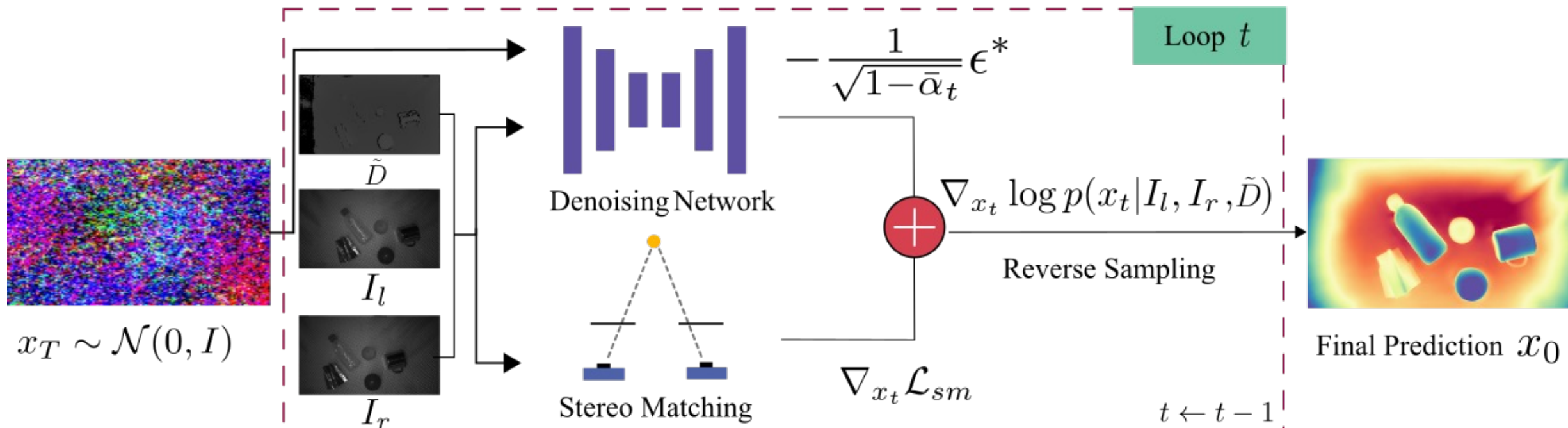




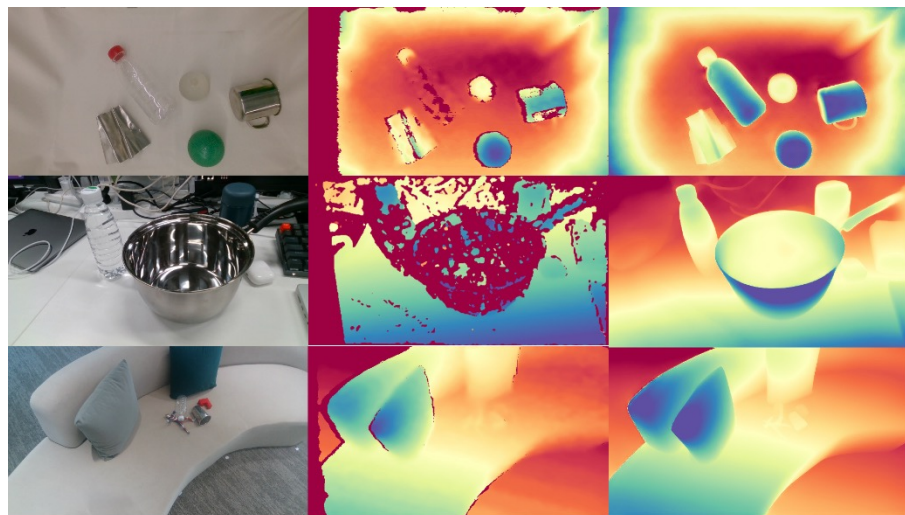
D³RoMa: Disparity Diffusion-based Depth Sensing for Material-Agnostic Robotic Manipulation

CoRL 2024

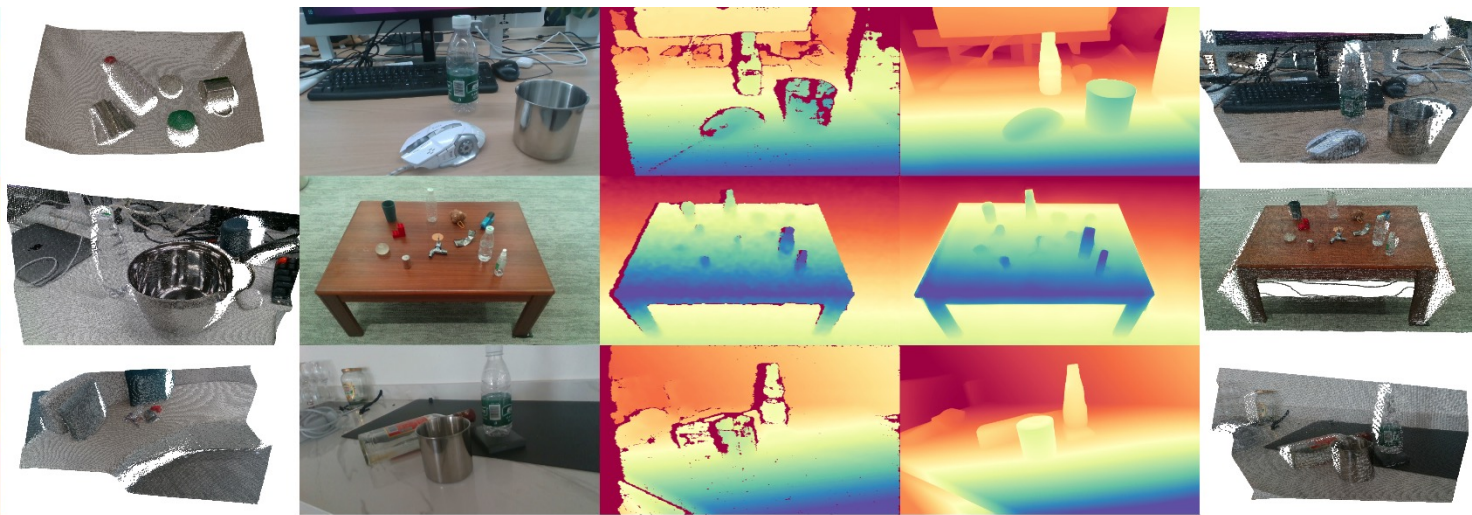
Method: Geometry guided disparity diffusion



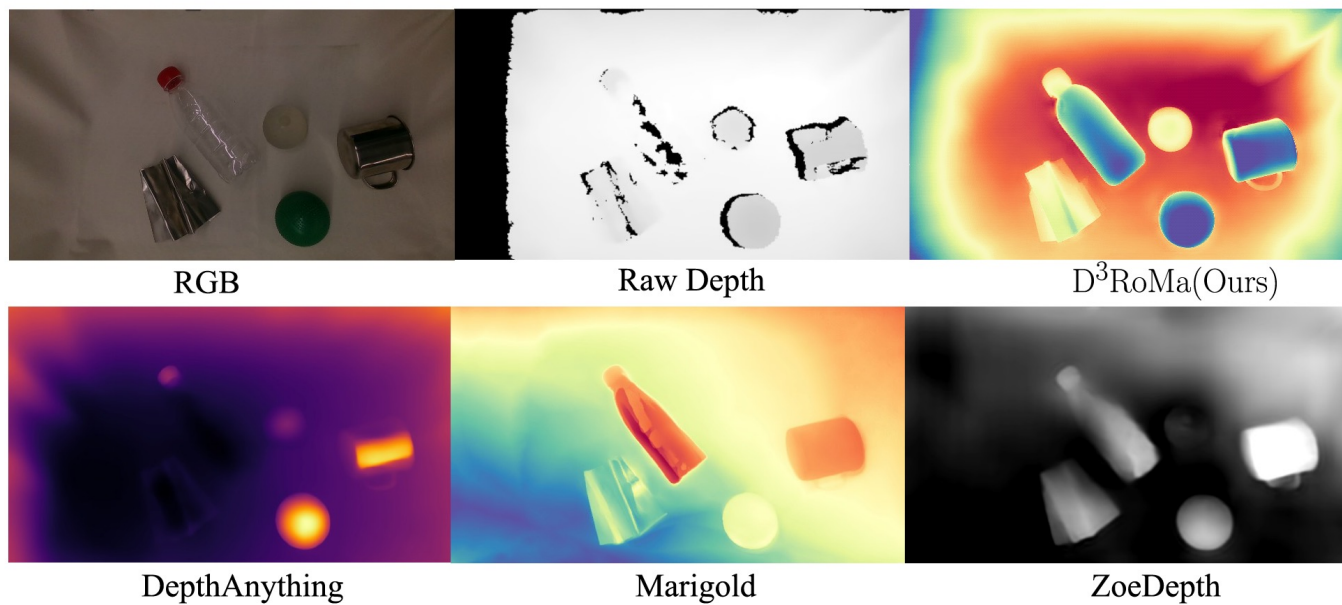
Experiments on Sim2Real



(a) RealSense D415

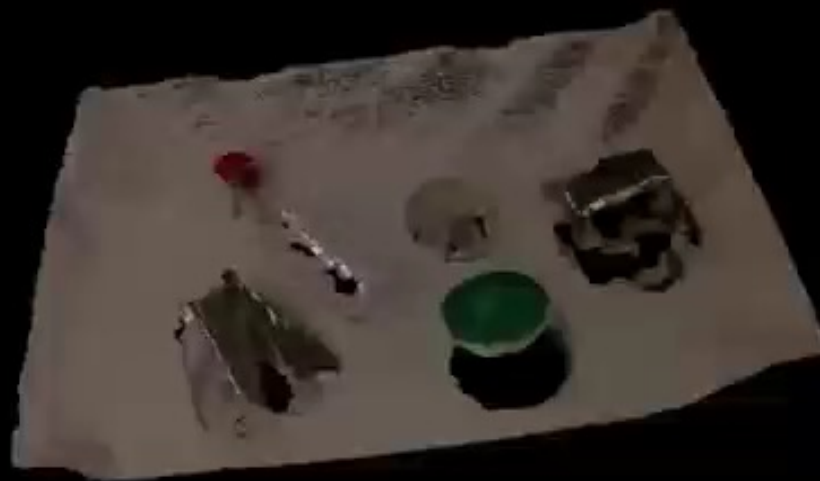


(b) RealSense D435



Result Visualization

Raw Depth



D3RoMA



* Captured by RealSense D415



Introduction to Embodied AI

Next week: Lecture 7, Vision and Grasping III

