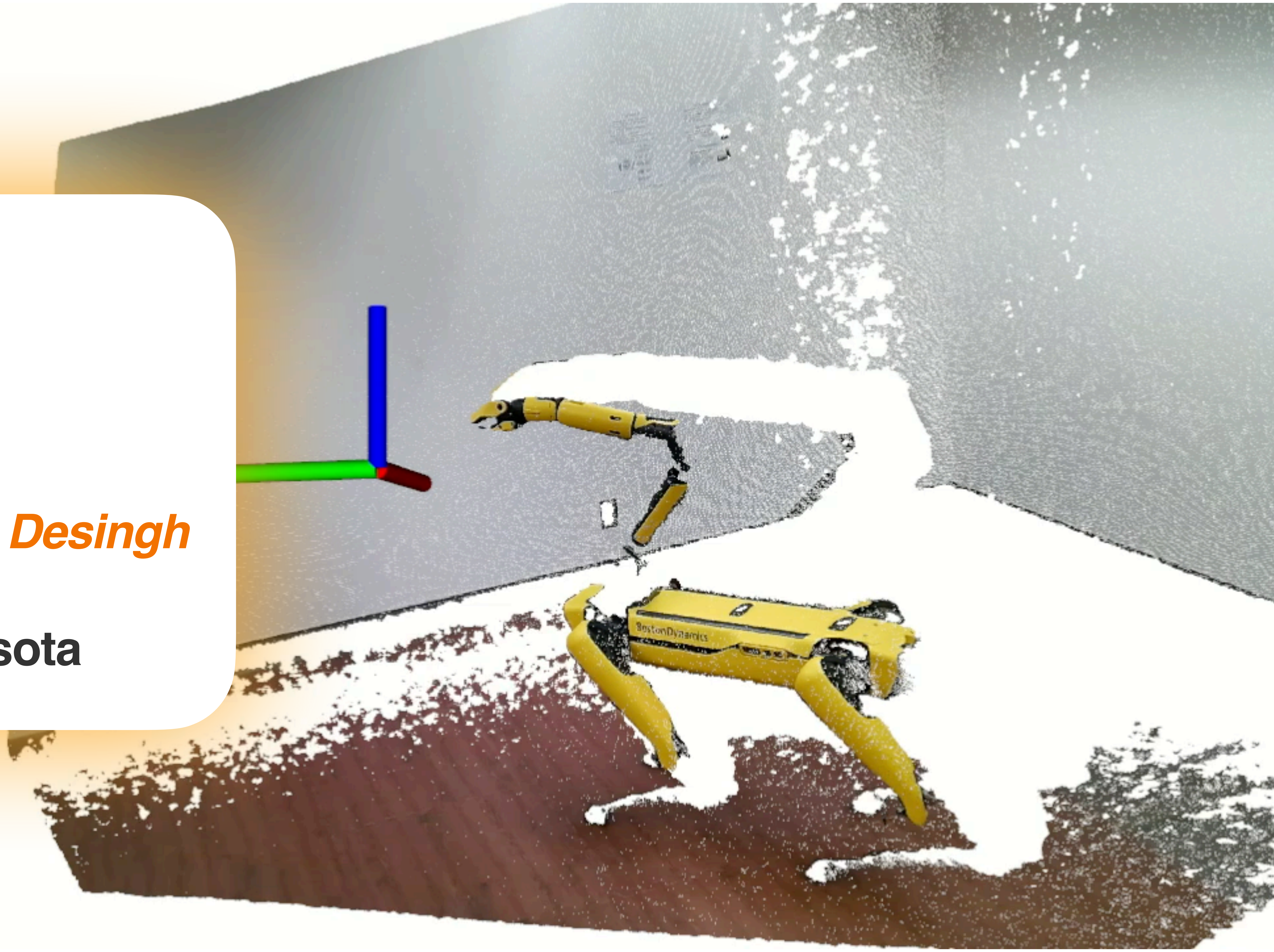
DeepRob

[Student] Lecture 14

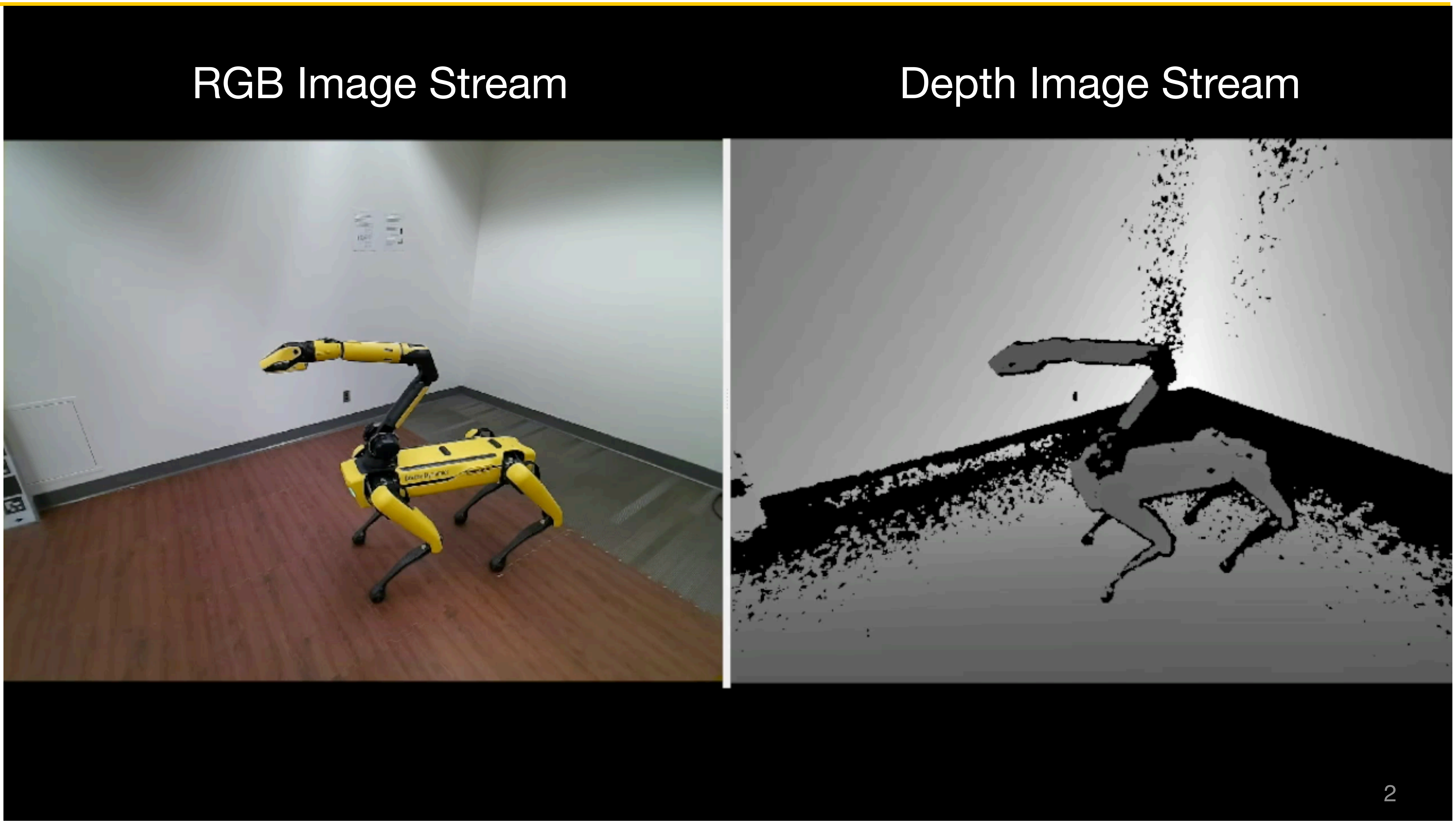
by Sidhanth Krishna, Shreyas Kallapur, Karthik Desingh

RGB-D Perception and Network Architectures

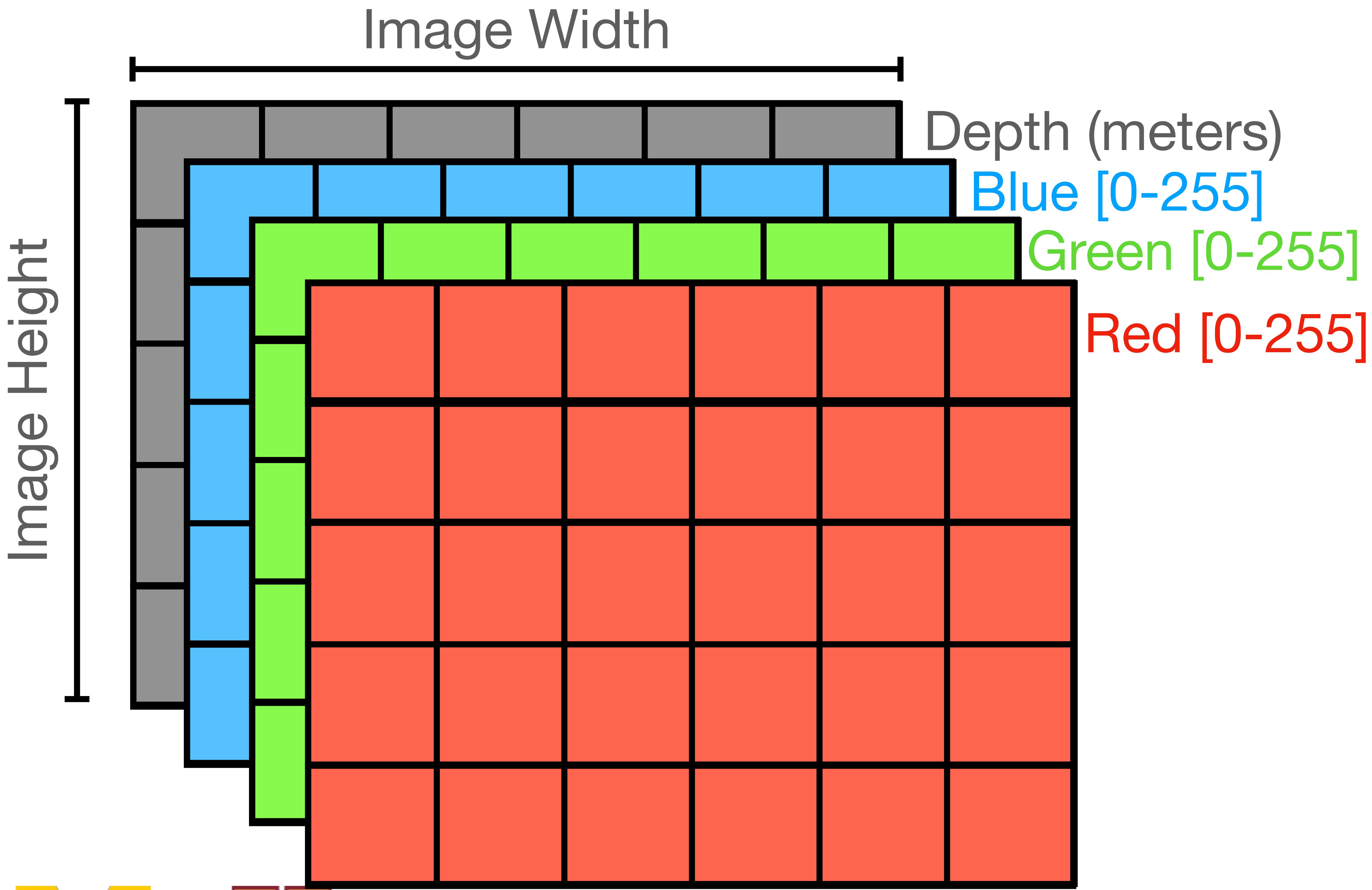
University of Michigan and University of Minnesota

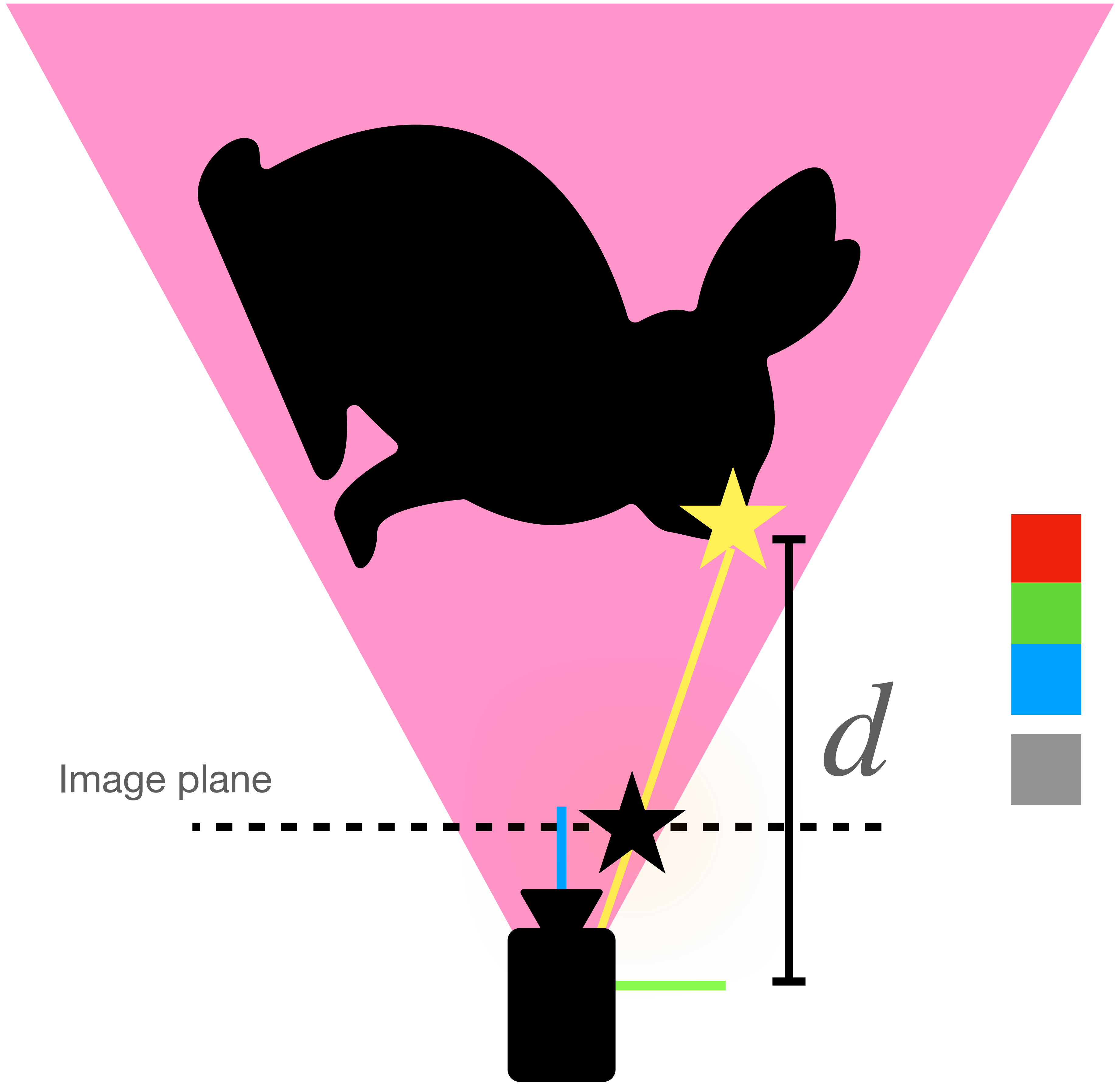


What is RGB-D data?



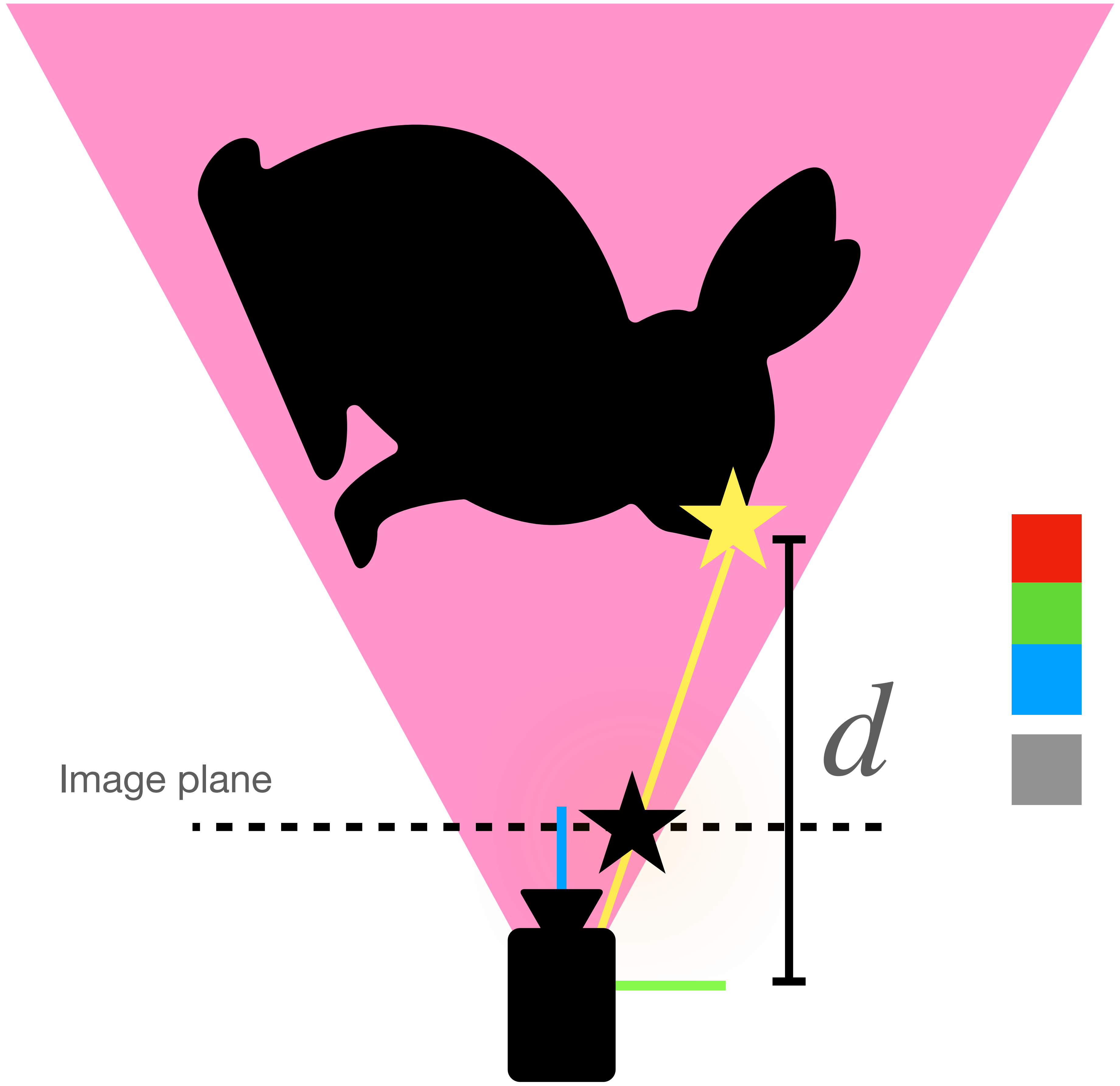
What is RGB-D data?



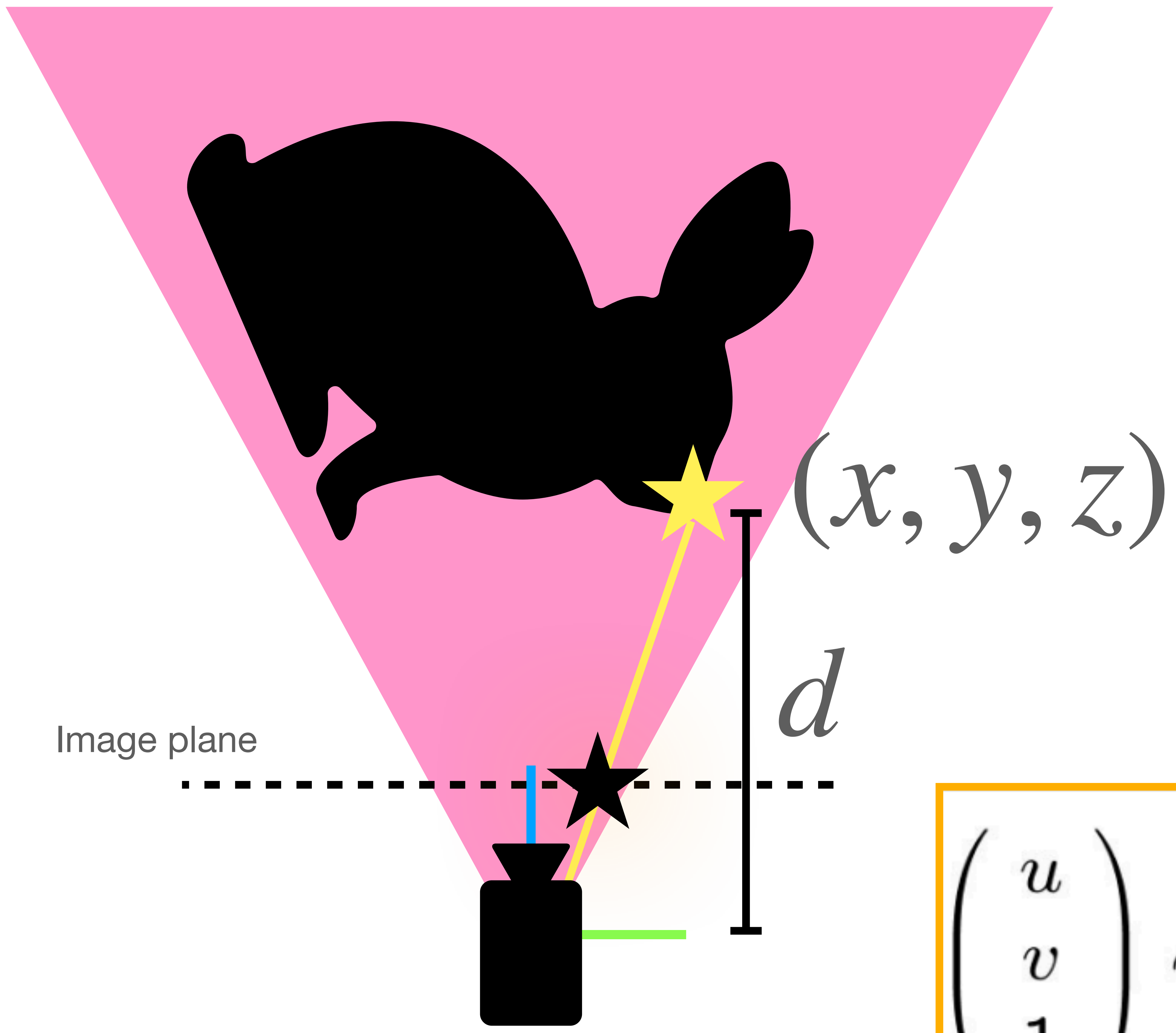


RGB-D to RGB XYZ

RGB-D to RGB XYZ



RGB-D to RGB XYZ



$$z = \frac{d}{d_{scale}}$$

$$x = \frac{(u - c_x) \times z}{f_x}$$

$$y = \frac{(v - c_y) \times z}{f_y}$$

d_{scale} is mm to meters (or something like that)

$$\underbrace{\begin{pmatrix} u \\ v \\ 1 \end{pmatrix}}_{\text{Image Coordinates}} \sim \underbrace{\begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix}}_{\text{Intrinsic Matrix}} \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{pmatrix}}_{\text{Extrinsic Matrix}} \underbrace{\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}}_{\text{3D point}}$$

What are RGB-D sensors?

RGB-D sensors

Microsoft Kinect V1

ENSENSO

Azure Kinect

ZED

ASUS Xtion

Pico Flexx2

Intel Realsense D405

Photoneo

Intel RealSense L515

What are their imaging mechanisms?

1. Stereoscopic

ENSENSO

ZED

Intel Realsense D405

2. Structured Light

Microsoft Kinect V1

ASUS Xtion

Photoneo

3. Time-of-Flight

Azure Kinect

Pico Flexx2

Intel RealSense L515

What are their imaging mechanisms?

1. Stereoscopic

ENSENSO

Intel Realsense D405

2. Structured Light

Microsoft Kinect V1

Photoneo

3. Time-of-Flight

Azure Kinect

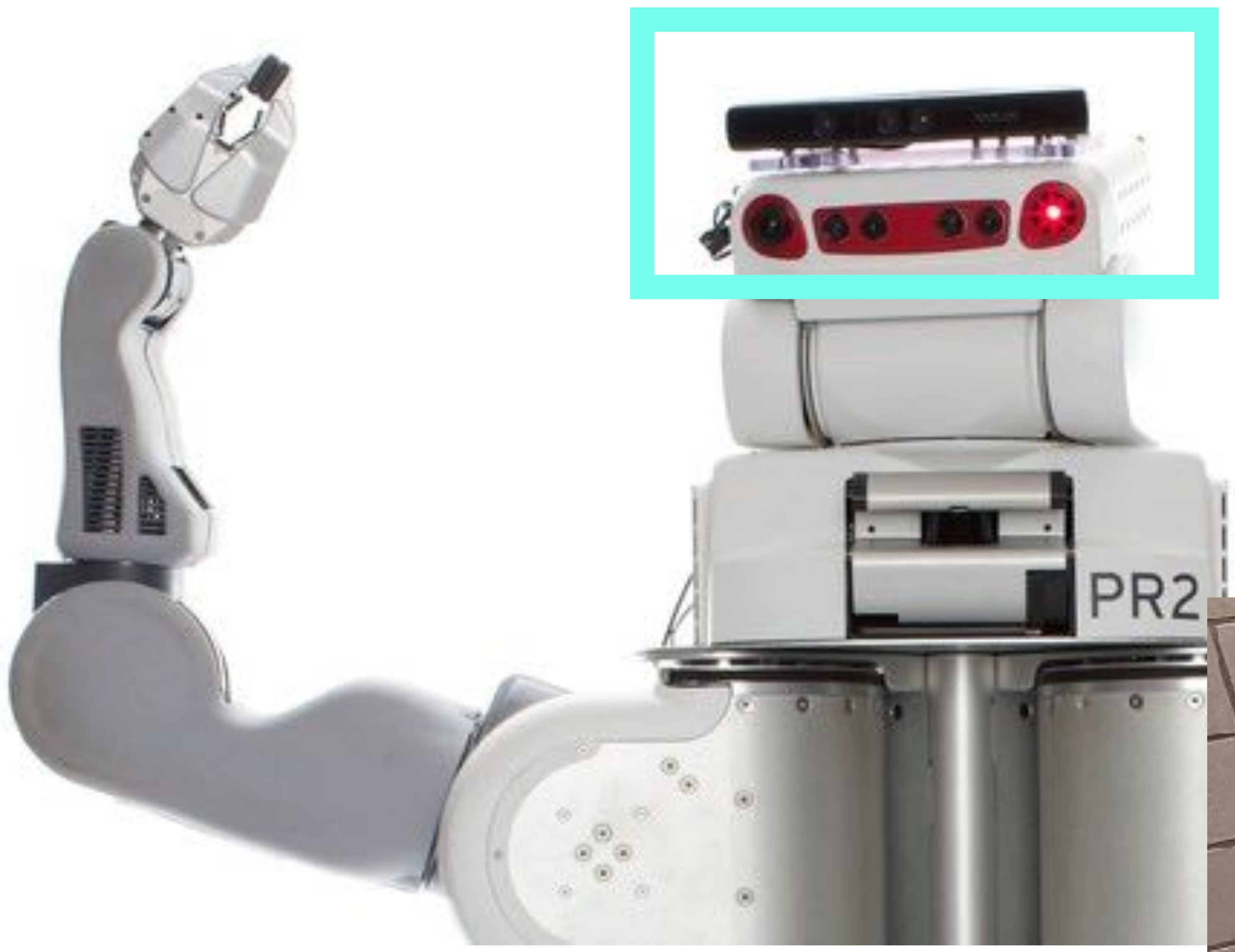
Pico Flexx2

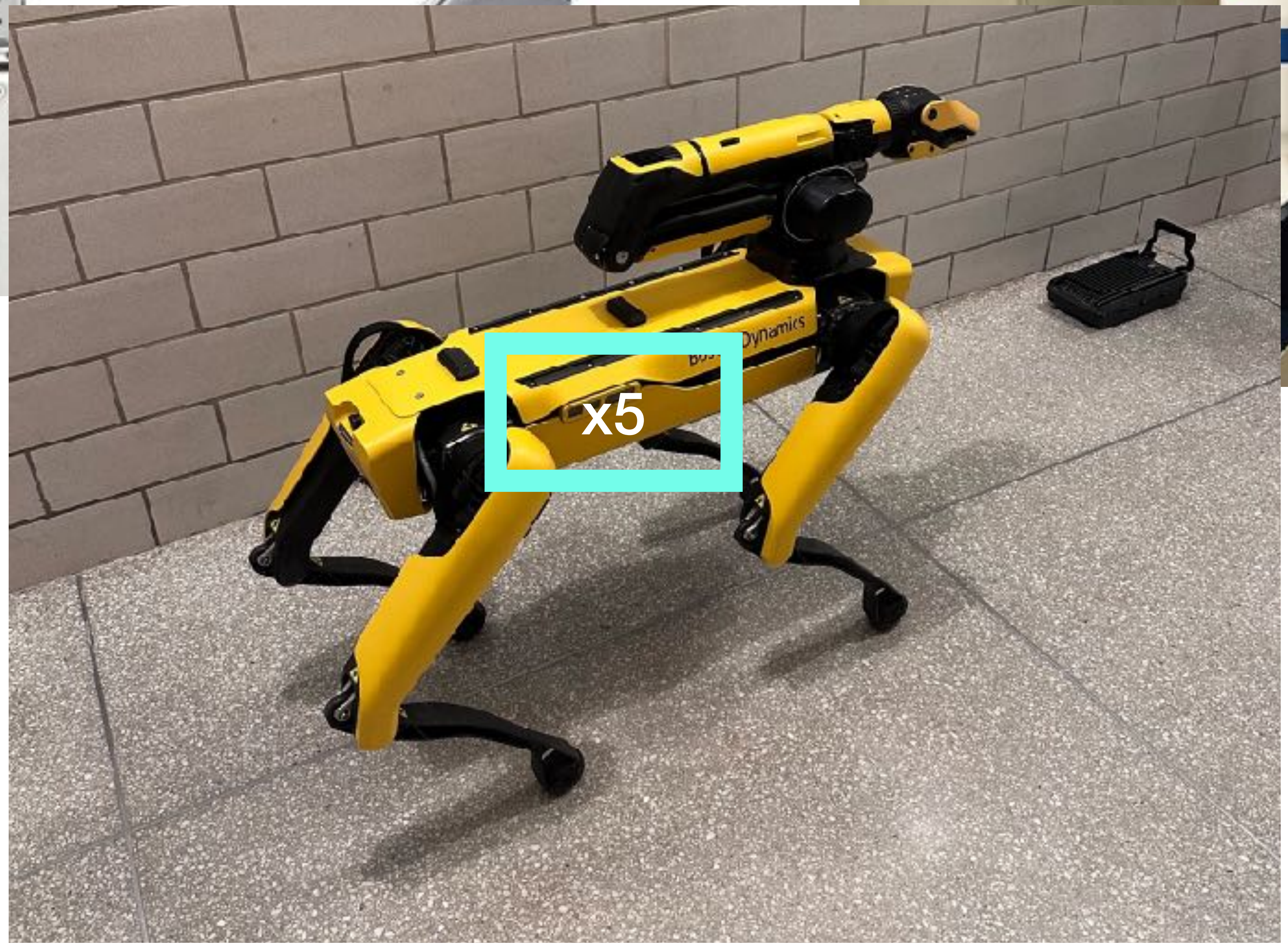
Intel RealSense L515

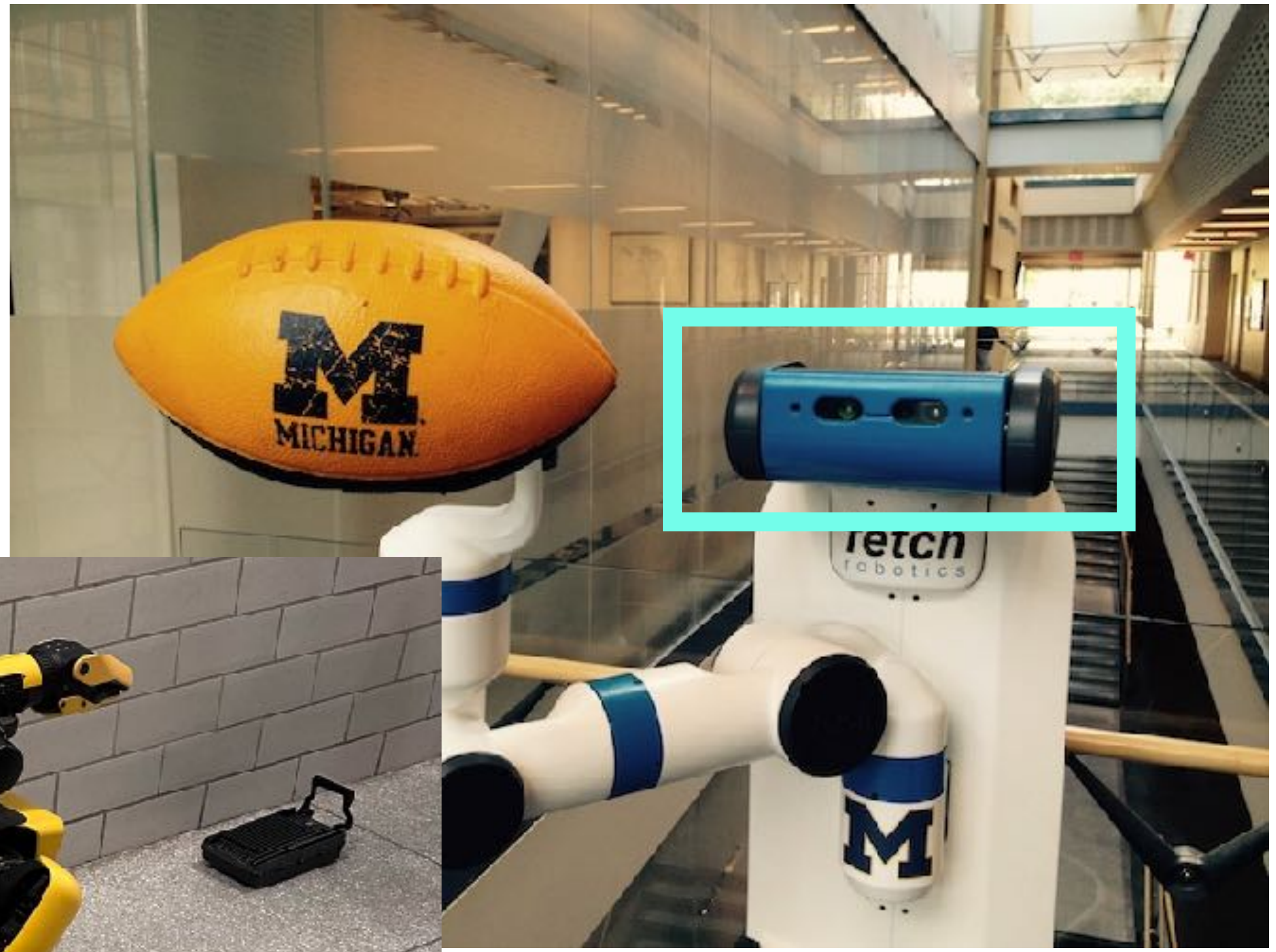
Choosing a sensor

- **Range**
- **Resolution**
- **Reliability**
  - **Data acquisition**
  - **Mechanical**

RGB-D sensors on Robots





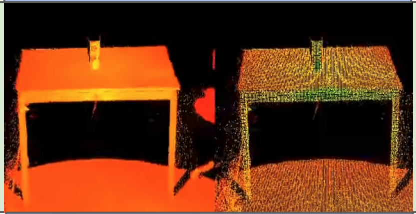
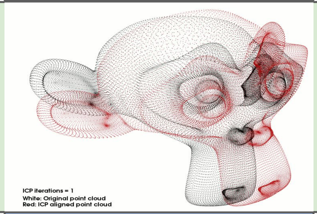
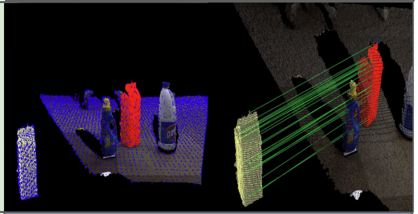
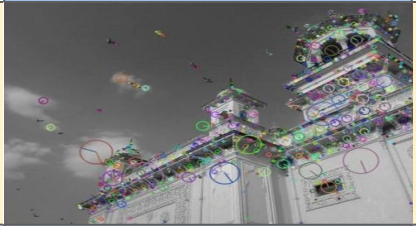
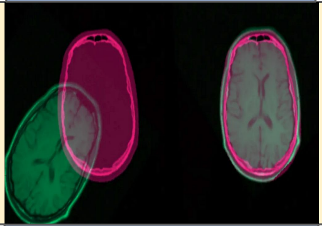
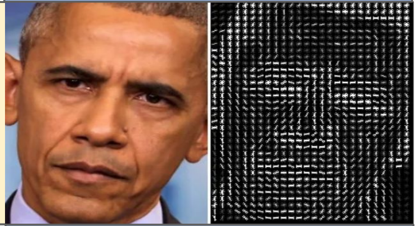


Fusion methods to get RGB-D

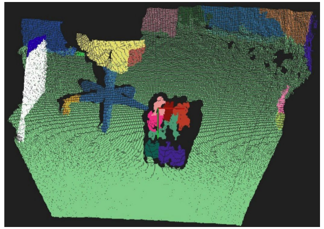
Kumar, G Ajay, Jin Hee Lee, Jongrak Hwang, Jaehyeong Park, Sung Hoon Youn, and Soon Kwon. 2020. "LiDAR and Camera Fusion Approach for Object Distance Estimation in Self-Driving Vehicles" *Symmetry* 12, no. 2: 324. https://doi.org/10.3390/sym12020324

Heuristics before Deep Learning

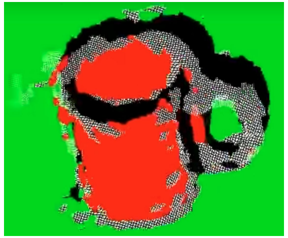
Traditional Methods

	Representation	Registration	Recognition
Depth			
	Voxelization	ICP Algorithm	Correspondence Grouping
RGB			
	SIFT	Homography	HoG + SVM

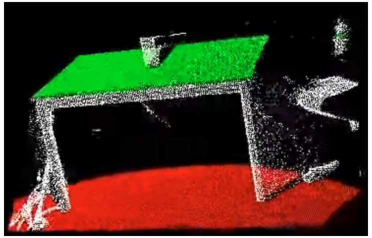
Traditional Methods - Segmentation



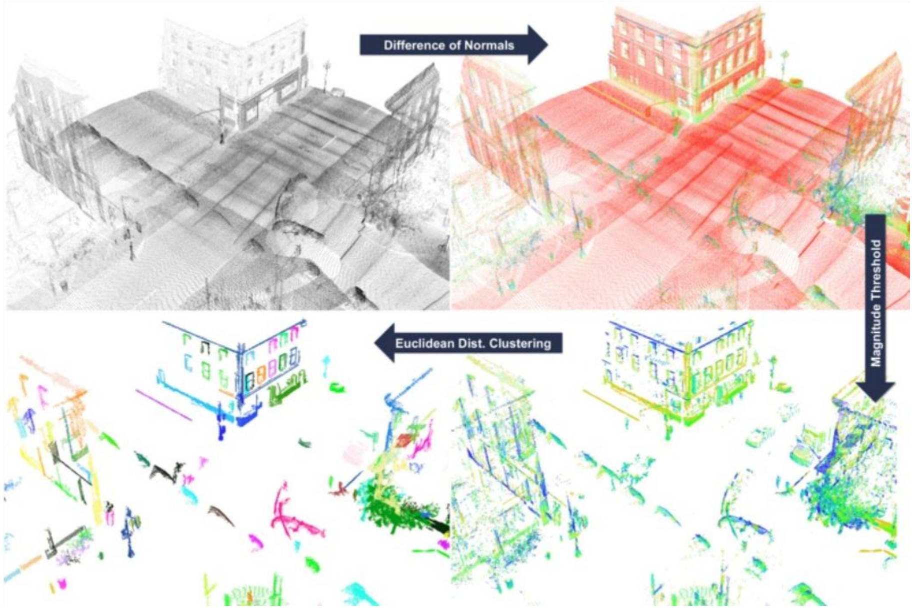
Color-based region growing segmentation



Cylinder Model



Plane Model

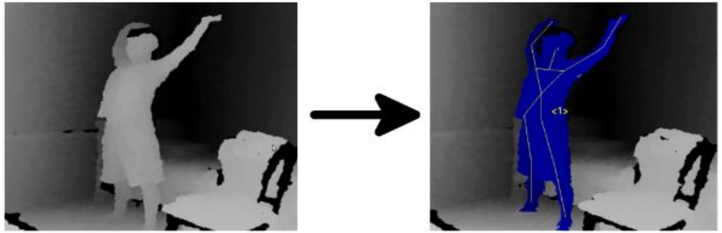


Difference of Normals Segmentation

Traditional Methods - Video Recognition

Figure 3: Samples from our dataset. Row-wise, from left: brushing teeth, cooking (stirring), writing on whiteboard, working on computer, talking on phone, wearing contact lenses, relaxing on a chair, opening a pill container, drinking water, cooking (chopping), talking on a chair, and rinsing mouth with water.

- ➢ Task: Human Activity Detection
- ➢ Data: RGBD frames

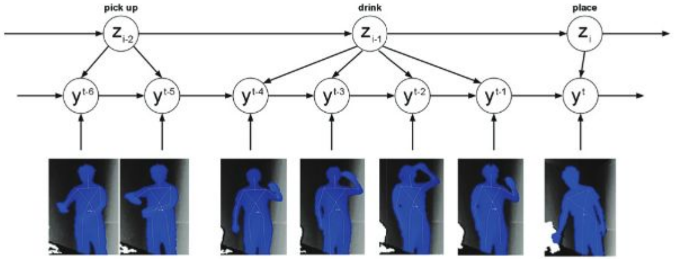


- ➢ Kinect depth map

Sung, Jaeyong, et al. "Human Activity Detection from RGBD Images." *plan, activity, and intent recognition* 64 (2011)

Modeling Pipeline

RGBD image → RGBD Features (Body Pose features, Hand Position, Motion Information) → ME-Markov Model → Activity Recognition



Results

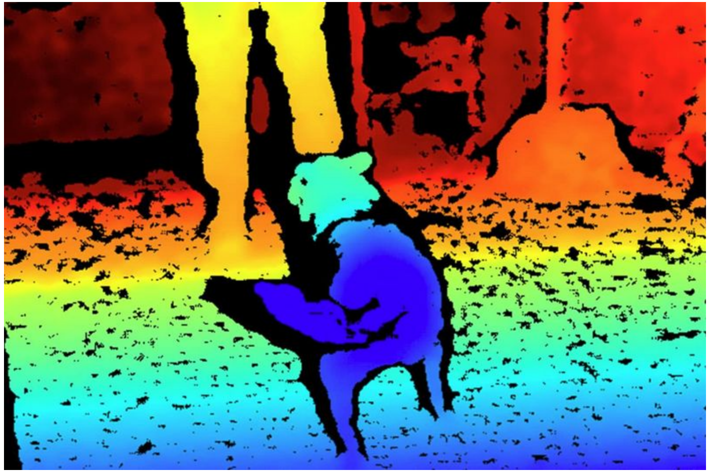
Location	Activity	Person seen before			New Person		
		Prec	Rec	$F_{0.5}$	Prec	Rec	$F_{0.5}$
bathroom	rinsing mouth	69.8	59.5	67.4	41.3	60.1	44.0
	brushing teeth	96.8	74.2	91.2	97.1	28.6	65.6
	wearing contact lens	80.3	91.2	82.3	74.1	91.6	77.0
	Average	82.3	75.0	80.3	70.8	60.1	62.2
bedroom	talking on the phone	88.2	80.2	86.5	74.7	54.6	69.6
	drinking water	88.5	78.2	86.2	65.8	67.3	66.1
	opening pill container	91.2	81.8	89.2	92.1	58.5	82.6
	Average	89.3	80.1	87.3	77.5	60.1	72.8
kitchen	cooking (chopping)	80.2	88.1	81.6	73.4	78.3	74.4
	cooking (stirring)	88.1	46.8	74.8	65.5	43.9	59.7
	drinking water	93.2	82.8	90.9	87.9	80.8	86.4
	opening pill container	86.6	82.2	85.7	86.4	58.0	78.7
	Average	87.0	75.0	83.3	78.3	65.2	74.8
living room	talking on the phone	75.7	82.1	76.9	61.2	54.9	59.8
	drinking water	84.5	80.3	83.6	64.1	68.7	64.9
	talking on couch	91.7	74.0	87.5	45.1	37.4	43.3
	relaxing on couch	85.7	84.6	85.4	24.4	8.3	17.5
	Average	84.4	80.3	83.4	48.7	42.3	46.4
office	talking on the phone	87.3	81.3	86.0	74.3	55.0	69.4
	writing on whiteboard	91.6	84.9	90.2	74.8	89.4	77.3
	drinking water	84.6	78.5	83.3	67.3	69.1	67.7
	working on computer	93.7	76.7	89.7	61.5	21.1	44.5
	Average	89.3	80.3	87.3	69.5	58.6	64.7
Overall Average		**86.5**	**78.1**	**84.3**	**69.0**	**57.3**	**64.2**

Generalizability

Overall results

Datasets

Organised and Unorganised point clouds



- ➢ Organized point clouds are arranged in a regular grid pattern
- ➢ Typically generated by sensors such as 3D lidar or RGB-D cam

- ➢ Unorganized point clouds don’t have any inherent spatial arrangement
- ➢ Relies on multi-view geometry
- ➢ Typically generated by sensors such as Lidar or photogrammetry

Computer Vision Datasets

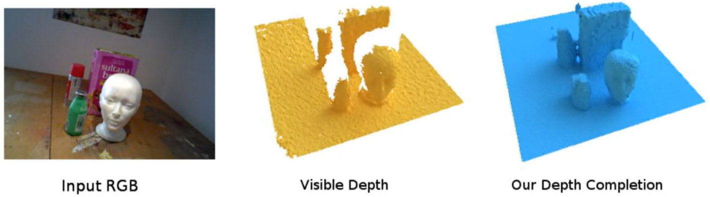
Datasets	Target Application	Device	Year
RGBD Object	Single objects in isolation	Kinect-v1	2011
TUM	Camera pose and scene recognition	Kinect-v1	2012
LINEMOD RGBD	Pose estimation	Kinect-v1	2012
SUN RGBD	Semantic reasoning and segmentation	Kinect-v1	2013
NTU RGBD	Activity and gestures	Kinect-v2	2016
VT-KFER	Face pose and recognition	Kinect-v1	2015
BIWI RGBD-ID	Human recognition	Kinect-v2	2012

J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In Intelligent Robots and Systems (IROS), 2012

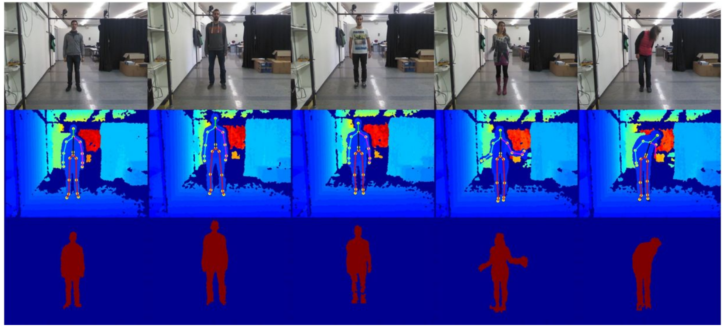
. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In International Conference on Robotics and Automation (ICRA), 2014

M. Firman, O. Mac Aodha, S. Julier, and G. Brostow. Structured prediction of unobserved voxels from a single depth image. In Computer Vision and Pattern Recognition (CVPR), 2016

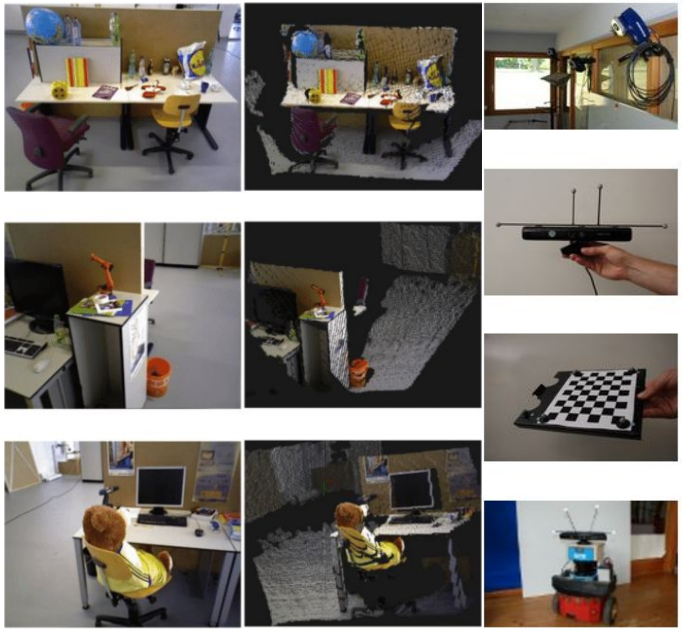
Datasets



TUM Dataset



BIWI Dataset



RGB RGBD Sensors

J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In Intelligent Robots and Systems (IROS), 2012

. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In International Conference on Robotics and Automation (ICRA), 2014

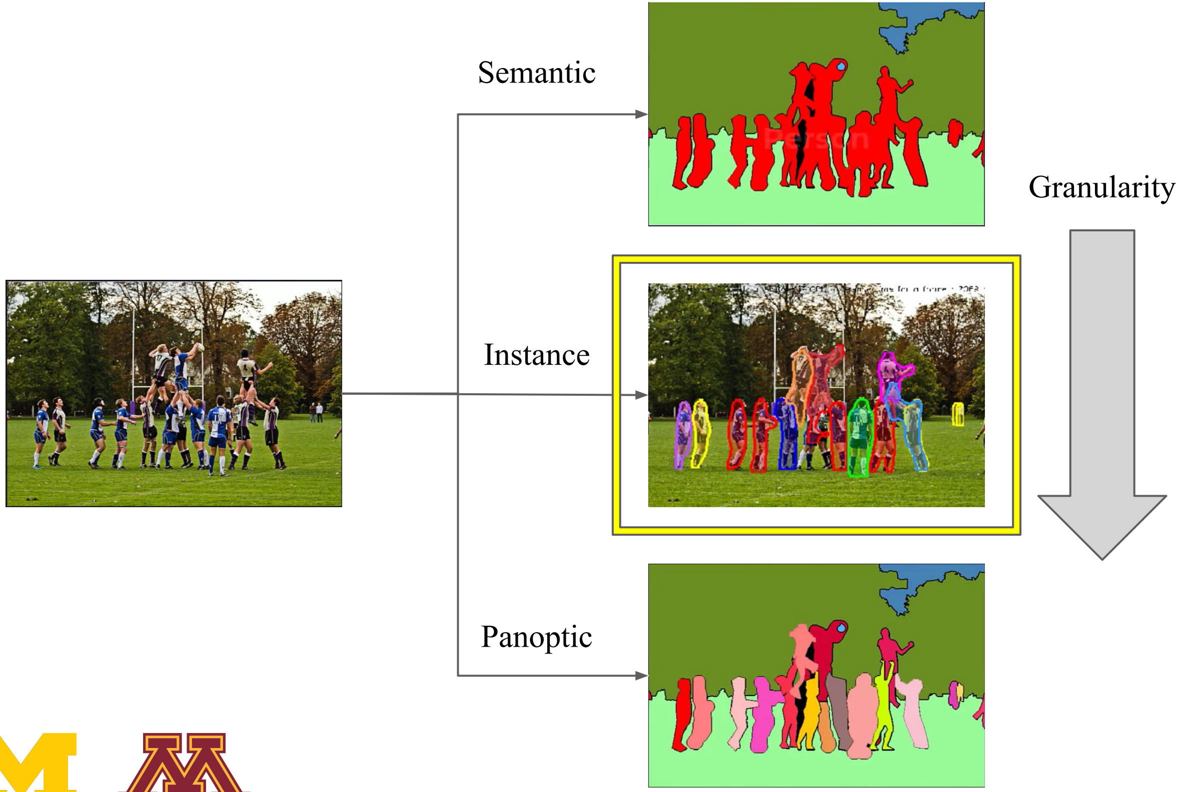
M. Firman, O. Mac Aodha, S. Julier, and G. Brostow. Structured prediction of unobserved voxels from a single depth image. In Computer Vision and Pattern Recognition (CVPR), 2016

Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation

Yu Xiang, Christopher Xie, Arsalan Mousavian, Dieter Fox

Conference on Robot Learning (CoRL), 2020

Types of Segmentation

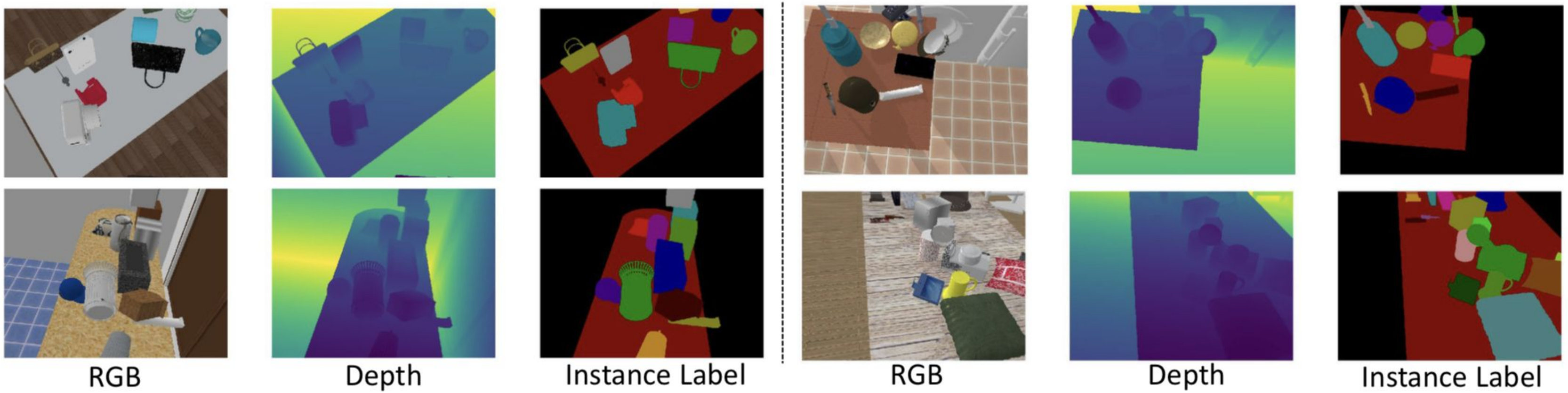


Instance Segmentation

Panoptic Segmentation

https://analyticsindiamag.com/semantic-vs-instance-vs-panoptic-which-image-segmentation-technique-to-choose

Datasets

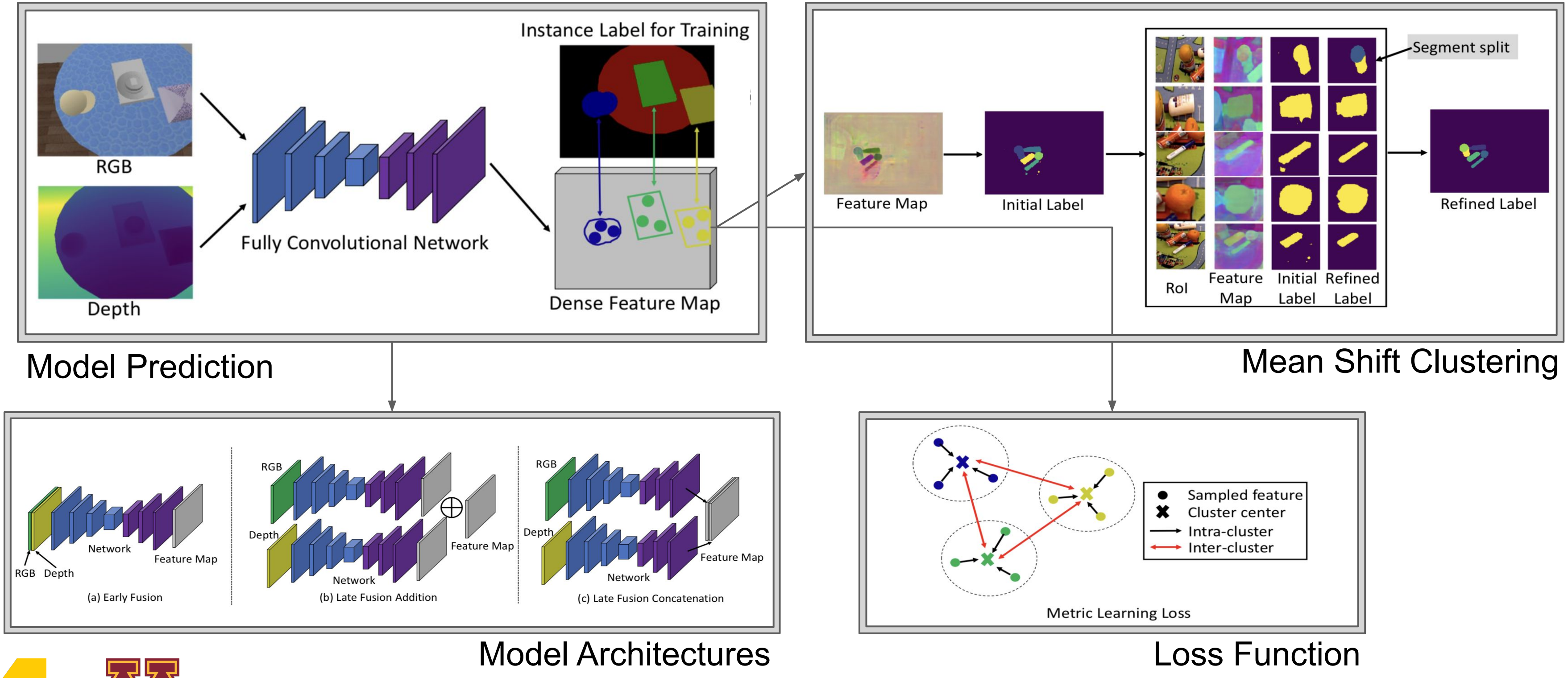


Tabletop Dataset (40,000 synthetic scenes)

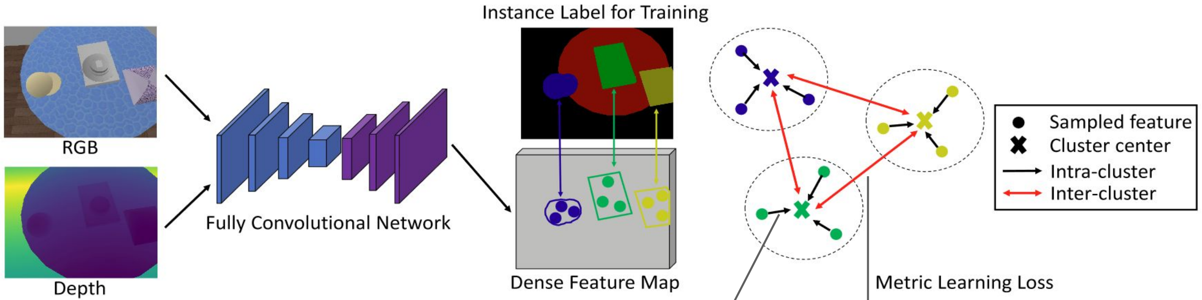
SUNCG house dataset - sample home environment
ShapeNet dataset - sample table and arbitrary objects (5-25)
PyBullet - physics simulator to place objects
7 RGB-D images captured for every scene

Training	TableTop	Synthetic Dataset - RGBD
Evaluation	OCID, OSD	Real world Data - RGBD

Unseen Object Segmentation



Loss - inter cluster and intra cluster

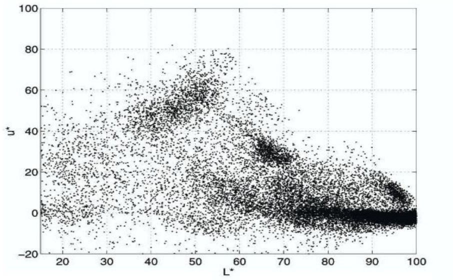


$$\ell_{\text{intra}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\mathbb{1}\left\{d(\mu^k, \mathbf{x}_i^k) - \alpha \geq 0\right\} \; d^2(\mu^k, \mathbf{x}_i^k)}{\sum_{i=1}^{N} \mathbb{1}\left\{d(\mu^k, \mathbf{x}_i^k) - \alpha \geq 0\right\}}$$

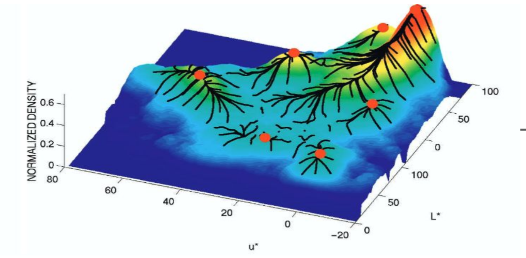
$$\ell_{\text{inter}} = \frac{2}{K(K-1)} \sum_{k<k'} \left[\delta - d(\mu^k, \mu^{k'})\right]_+^2$$

Mean Shift Clustering

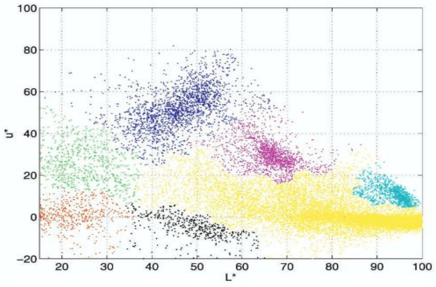
- ➢ MeanShift clustering aims to discover blobs in a smooth density of samples



2D Feature space

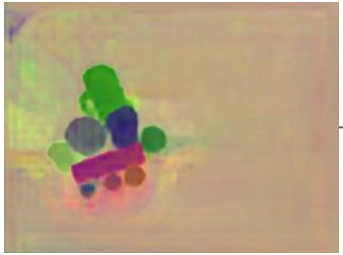


Density Estimate



Classification

- ➢ Two stage clustering



Feature Map

Stage-1

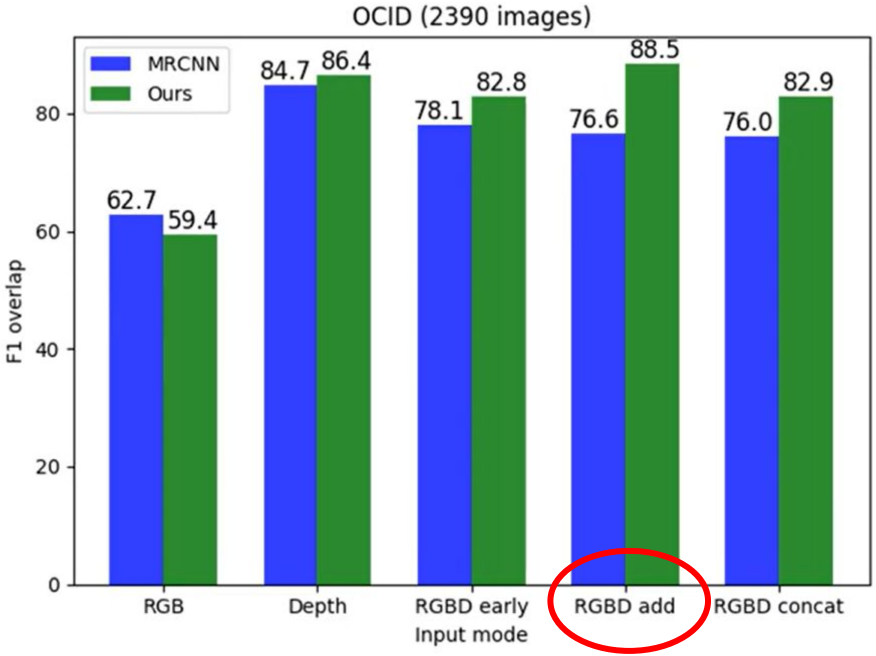
Stage-2

- ➢ Generates sharper object boundaries
- ➢ Separates objects that are under-segmented from stage-1

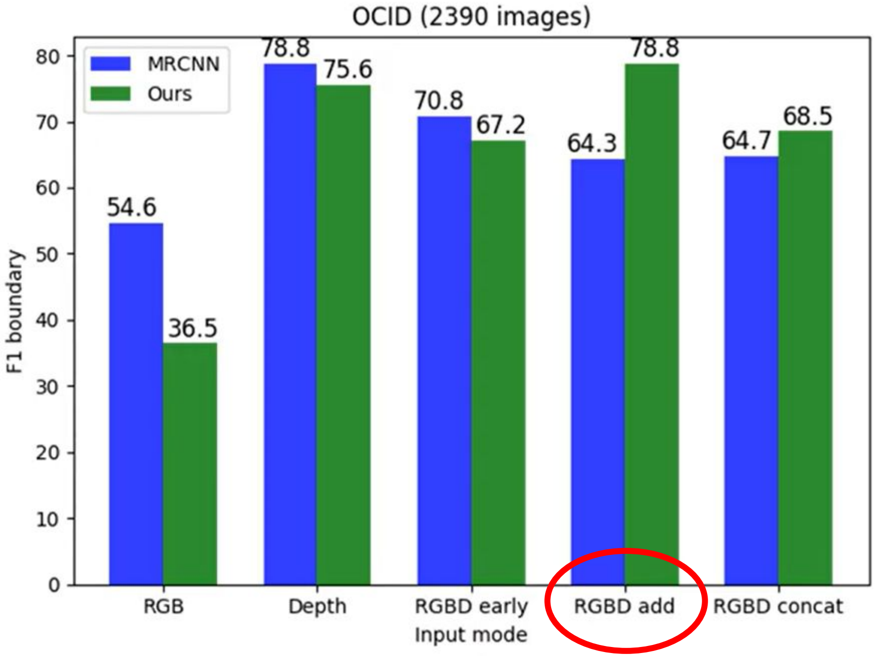
Evaluation Metrics

- Overlap
  - Precision - $P = \frac{\sum_i |c_i \cap g(c_i)|}{\sum_i |c_i|}$
  - Recall - $R = \frac{\sum_i |c_i \cap g(c_i)|}{\sum_j |g_j|}$
  - F-score - $F = \frac{2PR}{P+R}$
- c_i denotes the set of pixels belonging to predicted object *i*, $g(c_i)$ is the set of pixels of the matched ground truth object of c_i, and g_j is the set of pixels for ground truth object *j*
- Boundary
  - Precision - $P = \frac{\sum_i |c_i \cap D[g(c_i)]|}{\sum_i |c_i|}$
  - Recall - $R = \frac{\sum_i |D[c_i] \cap g(c_i)|}{\sum_j |g_j|}$
  - F-score - $F = \frac{2PR}{P+R}$
- Overlap measures don't take object boundaries into account
- To remedy this, we only consider pixels belonging to the boundaries of objects using the *D[.]* (Dilation) operation

Results



F-score overlap



F-score boundary

Results

Method	Input	OCID [11] (2390 images)							OSD [10] (111 images)						
		Overlap			Boundary				Overlap			Boundary			
		P	R	F	P	R	F	%75	P	R	F	P	R	F	%75
MRCNN	RGB	77.6	67.0	67.2	65.5	53.9	54.6	55.8	64.2	61.3	62.5	50.2	40.2	44.0	31.9
UCN (Ours)	RGB	54.8	76.0	59.4	34.5	45.0	36.5	48.0	57.2	73.8	63.3	34.7	50.0	39.1	52.5
MRCNN	Depth	85.3	85.6	84.7	**83.2**	76.6	**78.8**	72.7	77.8	85.1	80.6	52.5	57.9	54.6	77.6
UCN (Ours)	Depth	83.1	90.7	86.4	77.7	74.3	75.6	75.4	78.7	83.8	81.0	52.6	50.0	50.9	72.1
MRCNN	RGBD early	78.7	79.0	78.1	73.4	70.3	70.8	62.2	78.3	78.4	78.3	65.2	62.2	63.2	61.2
UCN (Ours)	RGBD early	78.8	89.2	82.8	66.9	69.7	67.2	73.5	77.4	81.8	79.2	53.9	53.0	53.0	69.0
MRCNN	RGBD add	79.6	76.7	76.6	68.7	63.7	64.3	62.9	66.4	64.8	65.5	53.7	43.8	47.5	37.1
UCN (Ours)	RGBD add	**86.0**	**92.3**	**88.5**	80.4	**78.3**	**78.8**	**82.2**	**84.3**	**88.3**	**86.2**	**67.5**	**67.5**	**67.1**	**79.3**
MRCNN	RGBD concat	79.6	76.2	76.0	68.2	63.5	63.7	63.0	67.0	63.8	65.3	53.1	42.7	46.5	37.1
UCN (Ours)	RGBD concat	79.2	87.8	82.9	70.6	67.5	68.5	68.3	76.4	83.3	79.7	50.5	48.5	48.8	67.5

Evaluation of proposed method and Mask R-CNN [32] trained on different input modes

Qualitative Results

Input Image

Feature Map

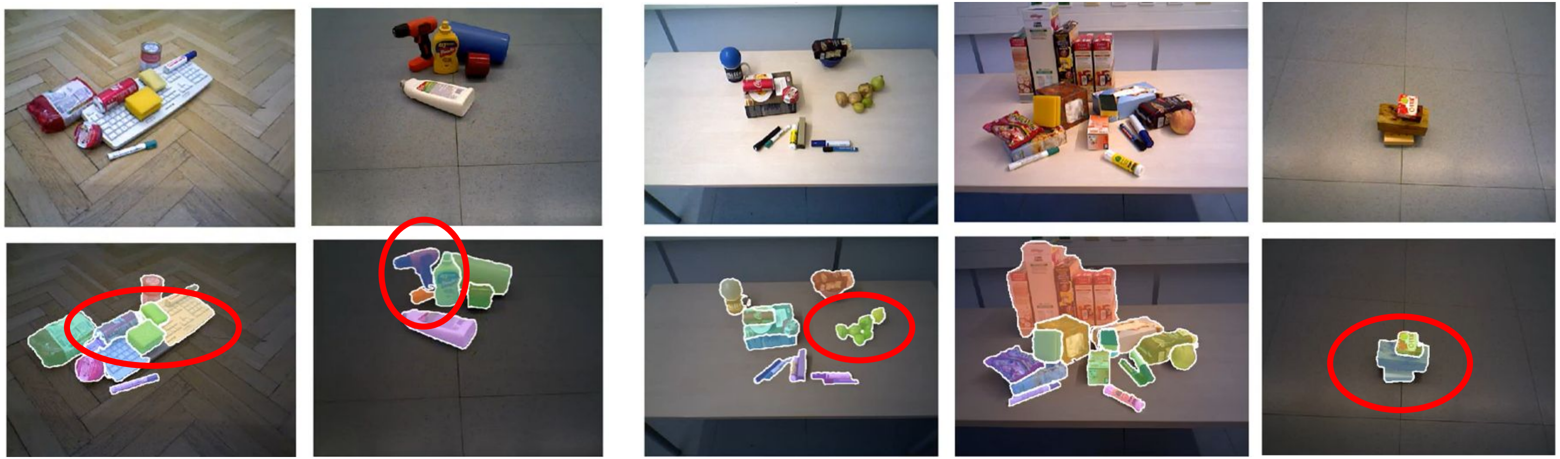
Initial Label

Refined Label

Failure Cases

Input Image

Final Label



Over-Segmentation

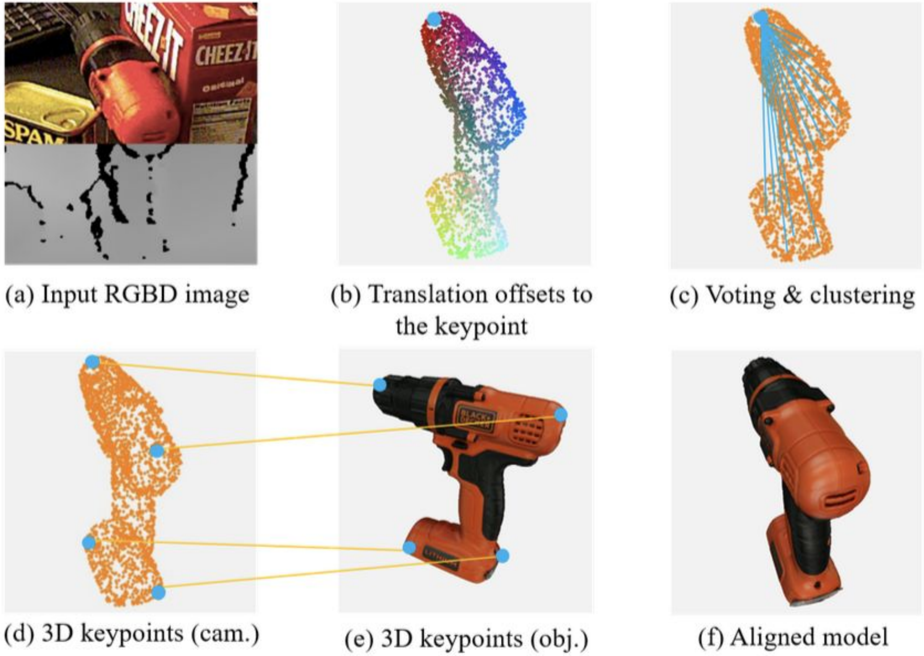
Under-Segmentation

Segmentation of Transparent Objects

ClearGrasp
Sajjan et al. ICRA'20

Relevant Works

PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DoF Pose Estimation



- ➢ Deep learning-based approach proposed for estimating the 6 degrees of freedom (6DoF) pose of an object in 3D space
- ➢ Deep Hough voting network to predict the per-point translation offset to the selected keypoint (b)
- ➢ Least Square fitting is applied to estimate 6D pose parameters (d)-(e)

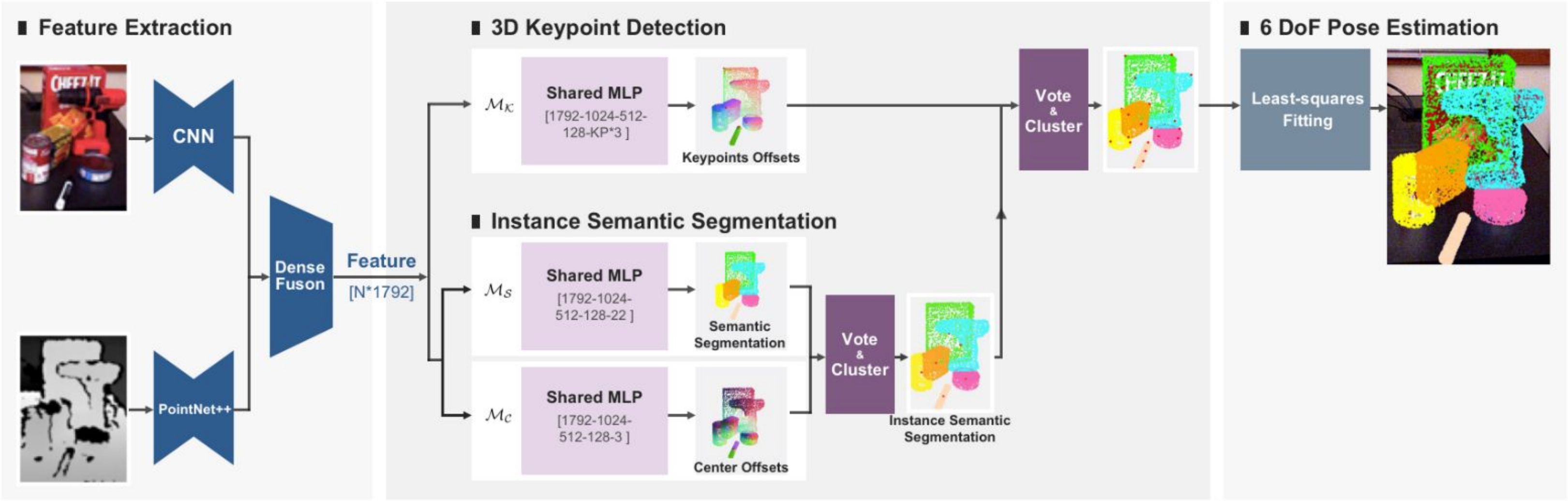
He, Yisheng, et al. "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020

PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DoF Pose Estimation

Dataset	Sensor	Number of Objects	Annotation	Data-split
YCB Video Dataset	Kinect V2 RGB-D camera	21 objects	ground truth poses - 3D translations and 4D quaternions.	Training
LINEMOD Dataset	Kinect RGB-D camera	13 objects		Evaluation
OccludedLINEMOD Dataset	Kinect RGB-D camera	13 objects with added occlusions		

He, Yisheng, et al. "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020

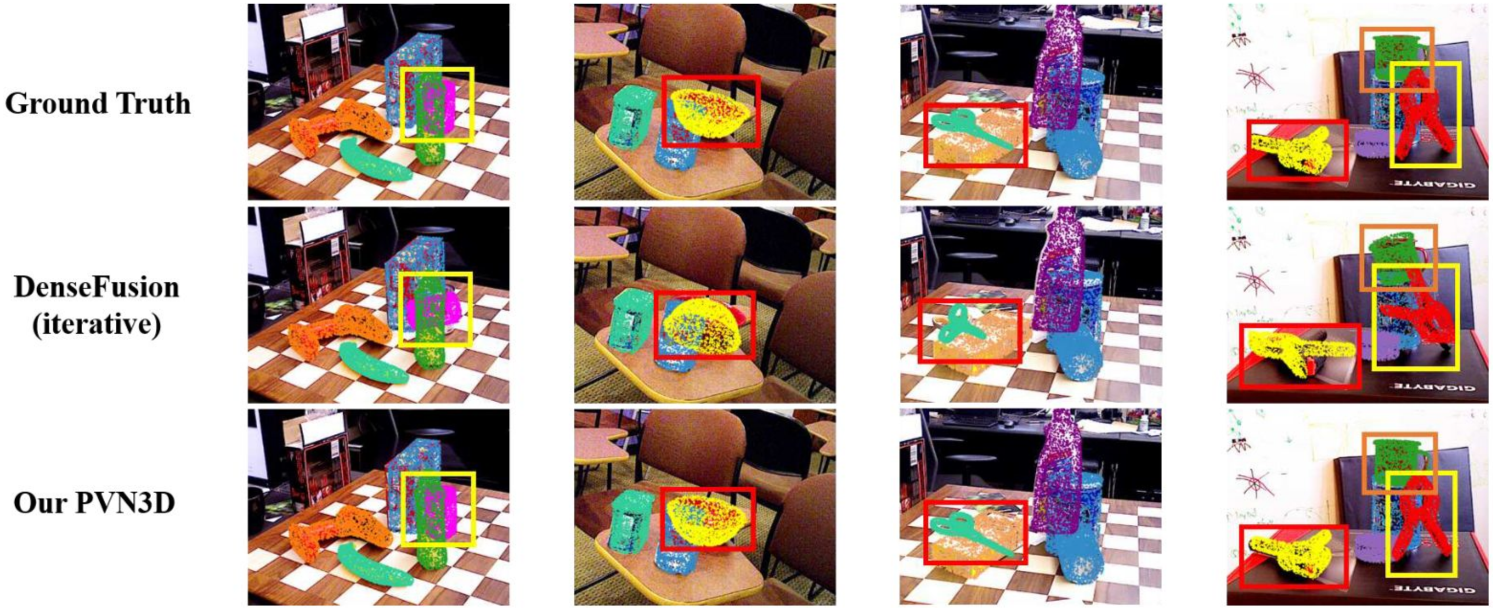
PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DoF Pose Estimation



Overview of PVN3D

He, Yisheng, et al. "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020

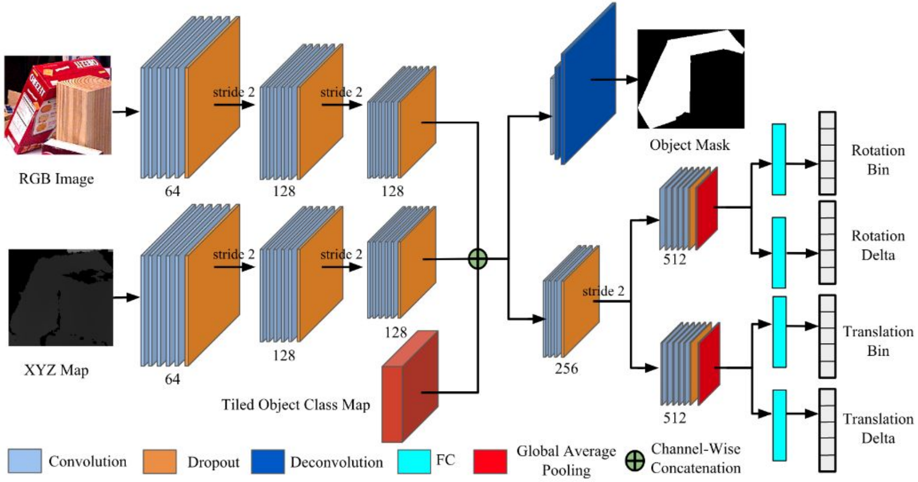
PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DoF Pose Estimation



Qualitative results on the YCB-Video dataset.

He, Yisheng, et al. "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020

A Unified Framework for Multi-View Multi-Class Object Pose Estimation



Class network architecture
XYZ map - normalized 3D coordinates of each image pixel

ROI, MCN estimates on RGB, MCN estimates on RGB-D and MV5-MCN estimates on RGB-D

Li, Chi, Jin Bai, and Gregory D. Hager. "A unified framework for multi-view multi-class object pose estimation." Proceedings of the european conference on computer vision (eccv). 2018.

A Unified Framework for Multi-View Multi-Class Object Pose Estimation

ShapeNet Dataset - 55,000 3D models in 55 categories

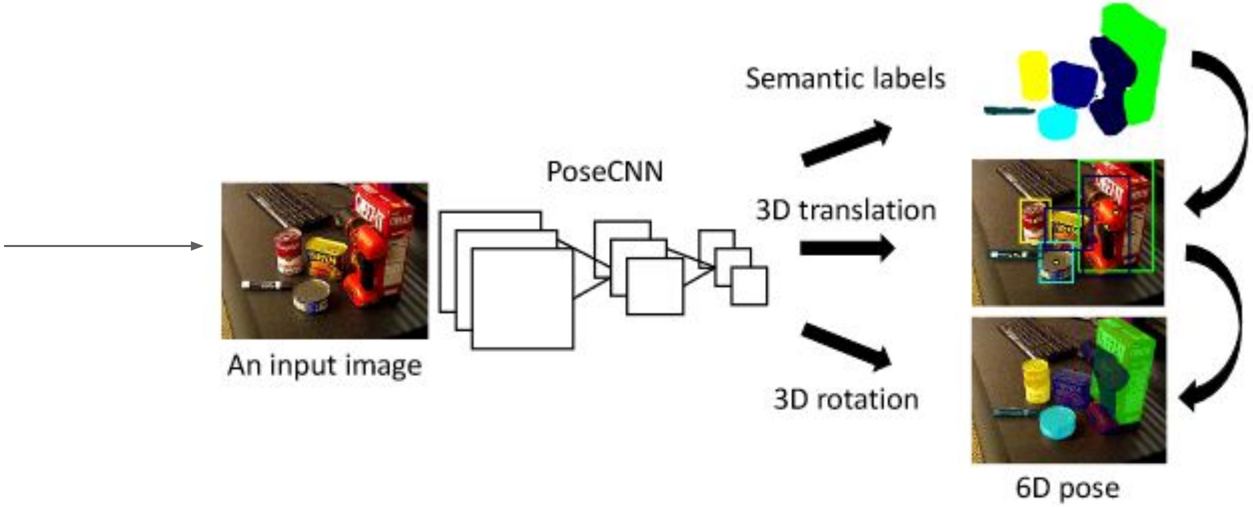
Training	rendered the ShapeNet models from multiple viewpoints with varying lighting and camera positions
Evaluation	Occluded LINEMOD dataset and the YCB-Video dataset

Li, Chi, Jin Bai, and Gregory D. Hager. "A unified framework for multi-view multi-class object pose estimation." *Proceedings of the european conference on computer vision (eccv)*. 2018.

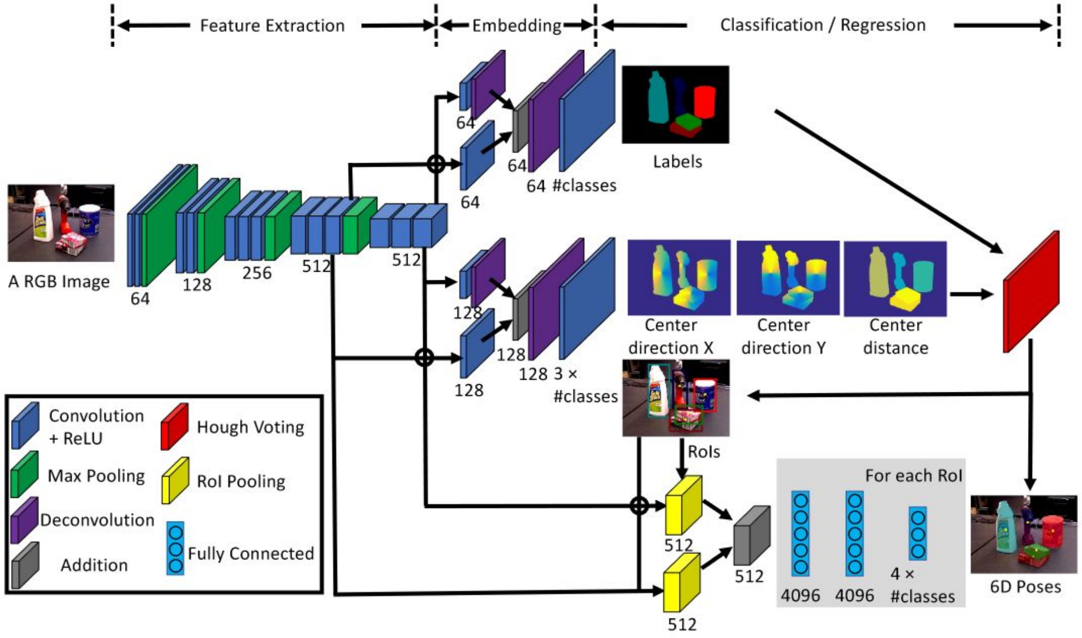
Chang, Angel X., et al. "Shapenet: An information-rich 3d model repository." arXiv preprint arXiv:1512.03012 (2015).

Calli, Berk, et al. "Yale-CMU-Berkeley dataset for robotic manipulation research." The International Journal of Robotics Research 36.3 (2017): 261-268.

PoseCNN:A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes



Xiang, Yu, et al. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes." *arXiv preprint arXiv:1711.00199* (2017)



Xiang, Yu, et al. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes." *arXiv preprint arXiv:1711.00199* (2017)

DeepRob

[Student] Lecture 14

by Sidhanth Krishna, Shreyas Kallapur, Karthik Desingh

RGB-D Perception and Network Architectures

University of Michigan and University of Minnesota