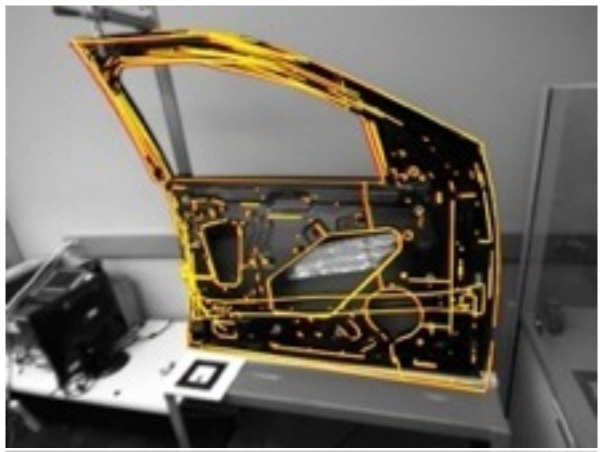
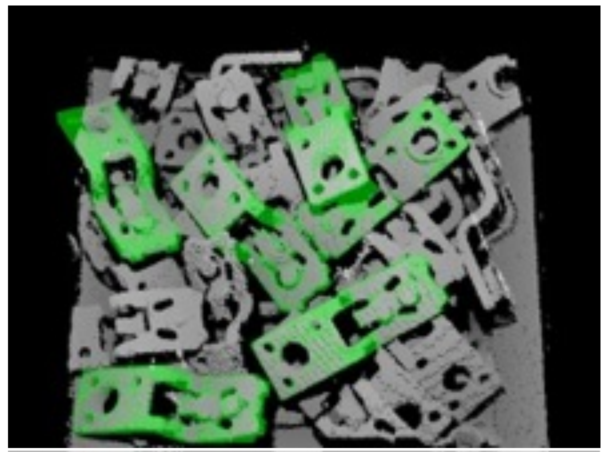
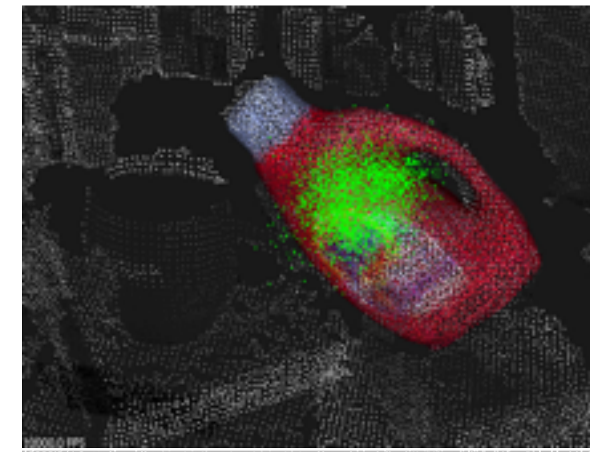
Visual Object Perception in Unstructured Environments







Changhyun Choi
Robotics Ph.D. Program
Interactive Computing, College of Computing
Georgia Institute of Technology

Prof. Henrik I. Christensen (Advisor)
Prof. James M. Rehg
Prof. Irfan Essa
Interactive Computing
College of Computing
Georgia Institute of Technology

Prof. Anthony Yezzi
Electrical and Computer Engineering
Georgia Institute of Technology

Prof. Dieter Fox
Computer Science and Engineering
University of Washington

Georgia Institute of Technology

Robotics@GT & Intelligent Machines

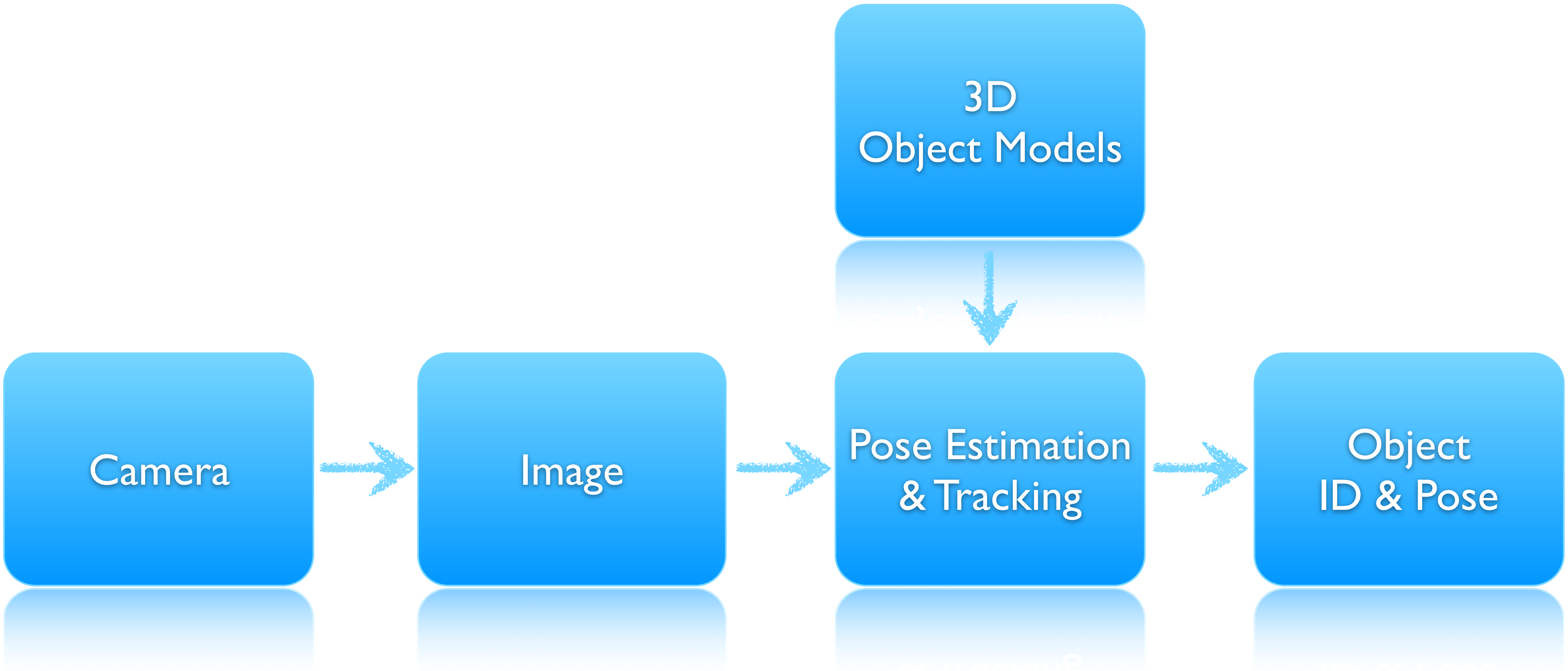
Outline

- Introduction
- State of the art
- Remaining challenges
- Motivations
- Thesis statement
- Approaches
- Conclusions & Future work

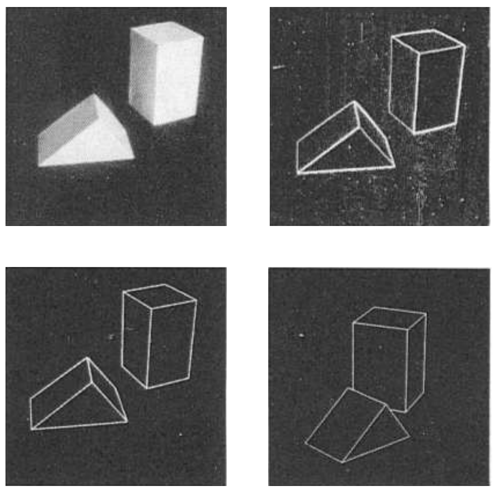
Introduction

- Pick-and-place task
- Robots moving from controlled settings to unstructured environments
- **Robust** object perception is crucial

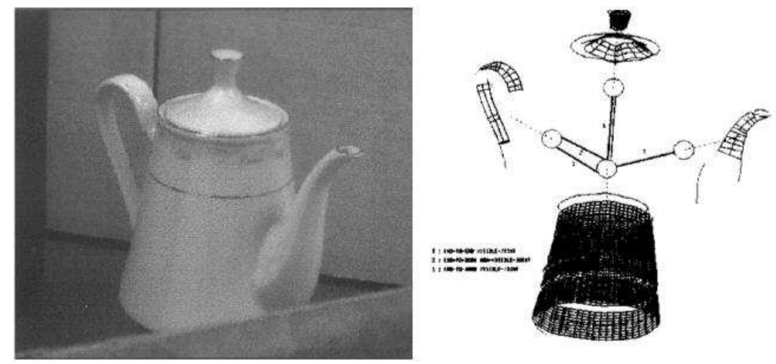
Problem Formulation



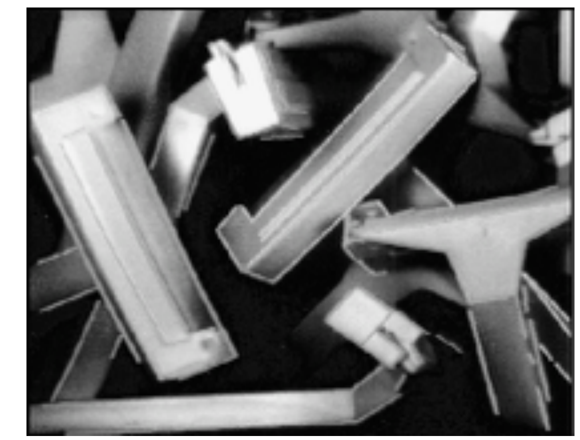
Early Object Perception



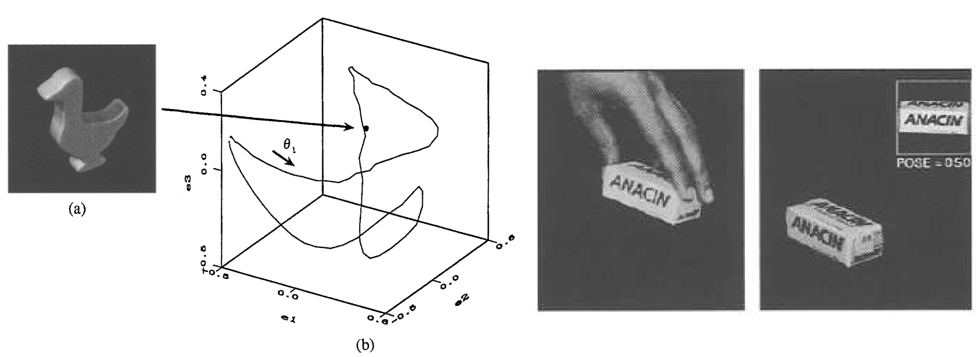
Convert 2D photo to 3D model
perspective projection
edge detection, line fitting
6-DOF transform
[Roberts 65]



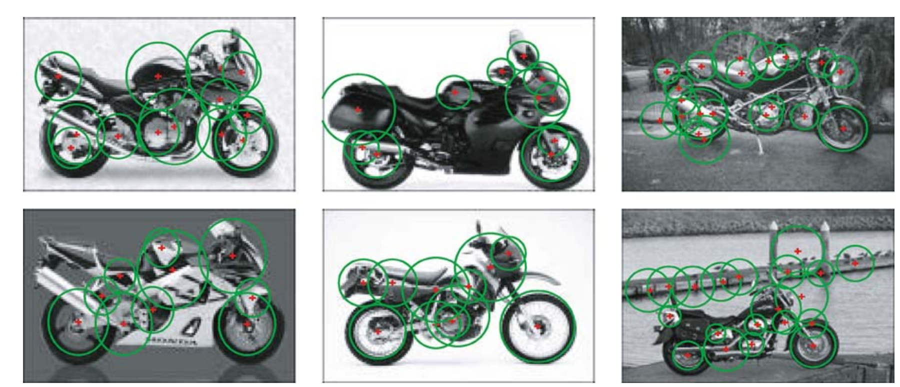
Categorical **3D** shape models: GC, geon
viewpoint invariant
textureless objects
[Binford 71, Brooks 83, Biederman 85,
Dickinson et al. 92, …]



Exact **3D** shape models: polyhedron or CAD
viewpoint invariant
textureless objects
[Lowe 87, Thompson and Mundy 87,
Huttenlocher and Ullman 90, , …]



Exemplar **2D** appearance model: 2D templates
viewpoint dependent
textured objects
[Murase and Nayar 95, Ohba and Ikeuchi 97,
Black and Jepson 98, …]



Categorical **2D** appearance model:
spatial models with local appearance features
viewpoint dependent
textured objects with clutter and occlusion
[Lowe 99, Mikolajczyk and Schmid 04,
Fei Fei et al. 06, Fergus et al. 07, …]

State of the Art

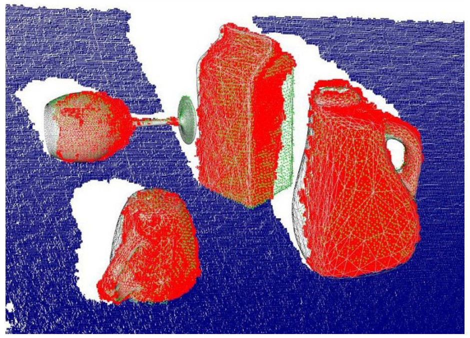
[Collet et al., IJRR'11]

- SiftGPU feature
- Sparse 3D keypoint models
- Iterative Clustering
- **Require** well textured objects
- **Cannot handle** textureless objects

[Hinterstoisser et al., PAMI'11]

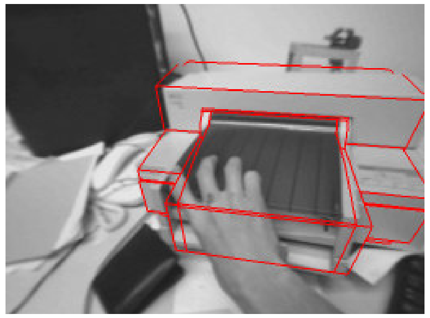
- Template Matching
- Combine *image gradients* and *surface normals*
- Can handle untextured objects
- **Require** large amount of templates (e.g. 2000)
- **Coarse** pose estimation
- Produce **jitter noises** in pose estimates

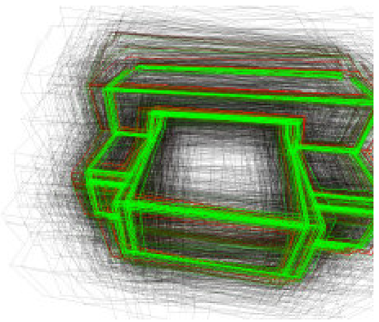
[Aldoma et al., ICCV workshop'11]
[Lai et al., AAAI'11]



- Table-top assumption
- Object segmentation + CVFH/kernel descriptors
- **Require** planar background
- **Hard** in cluttered environment

[Klein et al., BMVC'06]



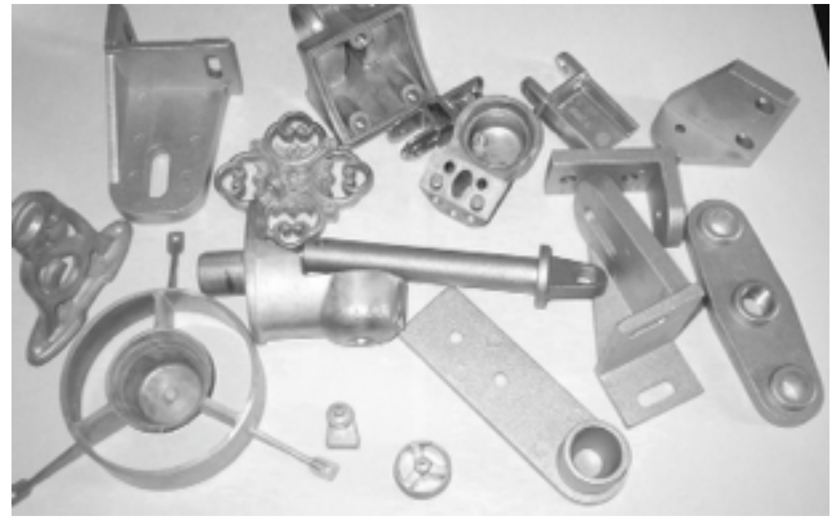


- Particle Filter
- Arbitrary shaped object
- **Require** a given starting pose
- **Do not** address challenging cases

Remained Challenges

1. Object with and without Textures
2. Background Clutter
3. Object Discontinuities
4. Real-time Constraints

Challenge 1: Texture


Handling both **textured** and **textureless** objects

- Textured objects
  - **Photometric**: *color*, *keypoints*, *edges* or textures from surfaces
- Textureless objects
  - **Geometric**: *point* coordinates, surface *normals*, depth discontinuities

Challenge 2: Clutter

Controlled environments

Unstructured environments

Difficulties = Degree of Clutter

- False measurements
- False pose estimates
- Stuck in local minima
- No table-top assumption

Challenge 3: Discontinuities

Occlusions

Out of FOV

Blur

- **Ideal** vs **Reality**
  - **Occluded** by other objects, human, or robots
  - Object goes **out** of the camera's field of view
  - **Blurred** in images
- Re-initialization problem

Challenge 4: Real-time

- Constrained by timing limitations
- Scarcely see real-time state-of-the-art

Definition and Scope 1

monocular
or RGB-D

object instance recognition

6-DOF p.e.
and tracking

Model-based Visual Object Perception

in Unstructured Environments

3D mesh models

cluttered &
obj. discontinuities

Definition and Scope 2

photometric image formation in 2D

intensity, color, edges from texture, keypoint descriptors, …

Visual features: Photometric & Geometric

3D geometric shapes

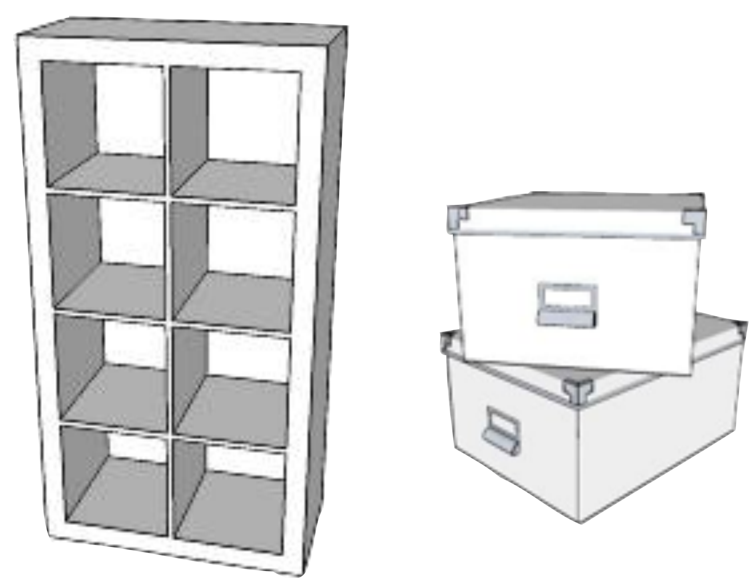
depth points, edges from geometric shapes, line segments, planes, normals, …

Motivations

Known 3D object model was strong assumption

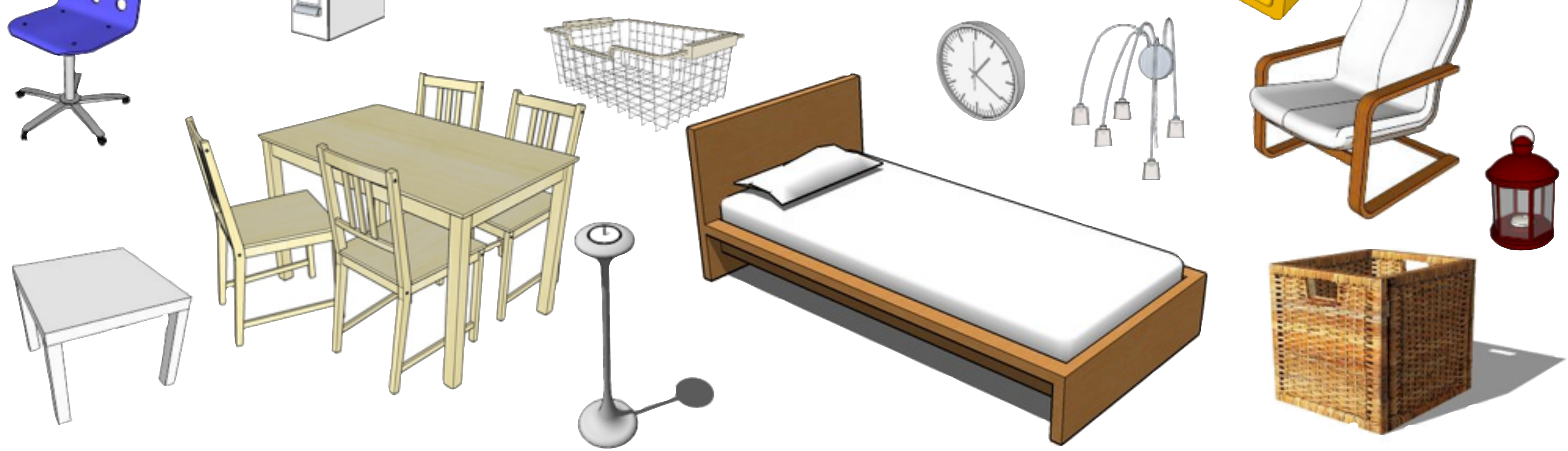
3D object models have been *accumulated* on the Internet!

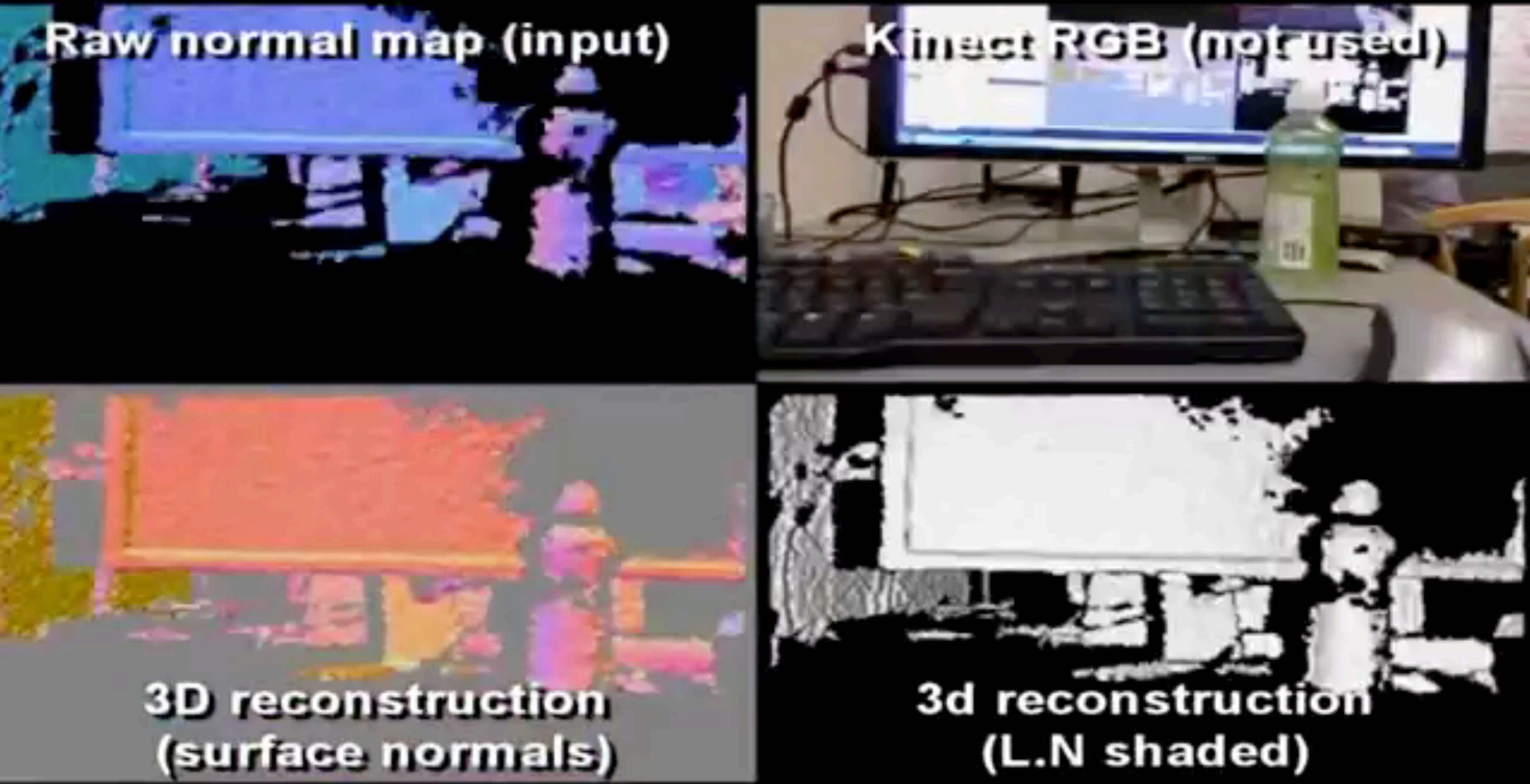
Motivations



Google 3D warehouse

(about 2.5 million models)





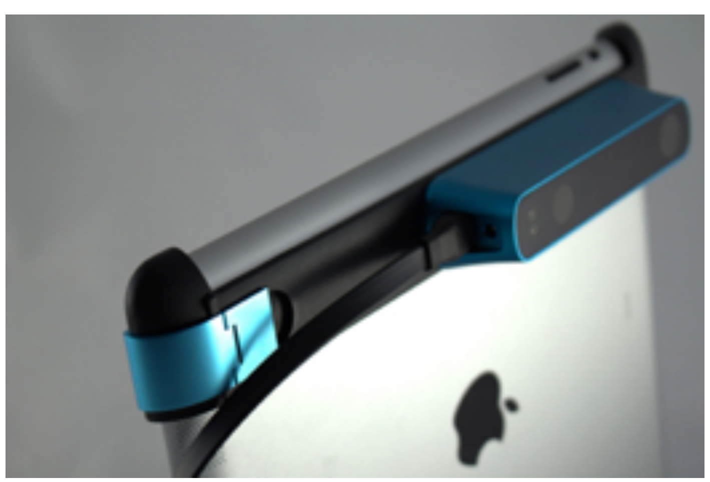
[Izadi et al., SIGGRAPH Talks 2011]

3D modeling will be a trivial task with **Kinect**!

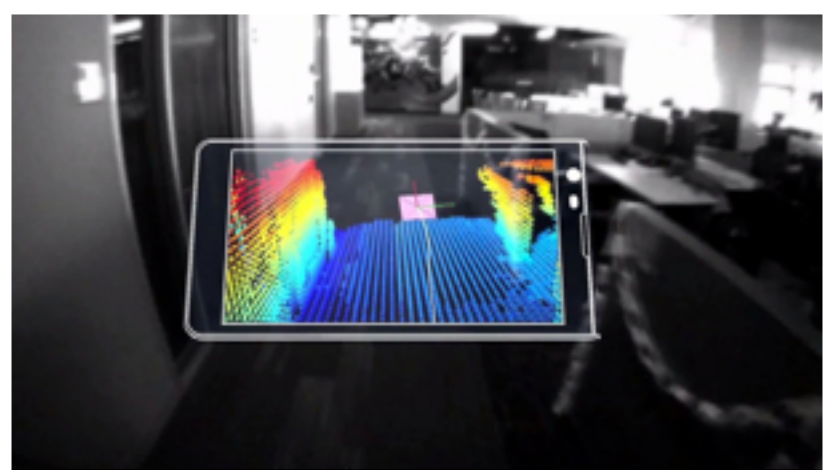
Motivations

24 million Kinects sold

Depth sensors are everywhere!



Occipital, Inc



Google Project Tango

Apple + PrimeSense

[AUTODESK 123D CATCH]

3D modeling will be a trivial task even with a **mobile phone**!

Motivations

- Promising in Robotics
  - exist in 3D space
  - interact with 3D world
  - 3D data is significant information for robots
- Advantages
  - Foreground object segmentation is trivial
  - Employ various geometric features from (3D models and 3D scene depth)

Thesis Statement

- To close the loop between the geometric era of early computer vision and the currently dominating appearance age, both photometric and geometric features need to be considered.
- The combination of these features enables object perception algorithms not only to be more effective but also to handle an increased spectrum of objects.
- Two theoretical frameworks using multiple pose hypotheses based on combined features are contributed in this thesis.
- These new frameworks are robust to significant clutter and occlusions, and are therefore efficacious solutions for visual object perception in unstructured environments.

Approaches

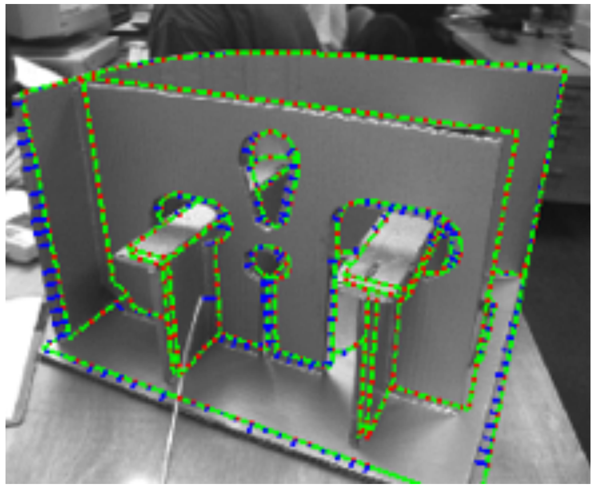
photometric

geometric

- 2D Visual Information (Monocular Camera)
  - **Combining Keypoint and Edge Features**
    [ICRA'10, ICRA'11, IJRR'12]
  - Extending to Textureless Objects
    [IROS'12]

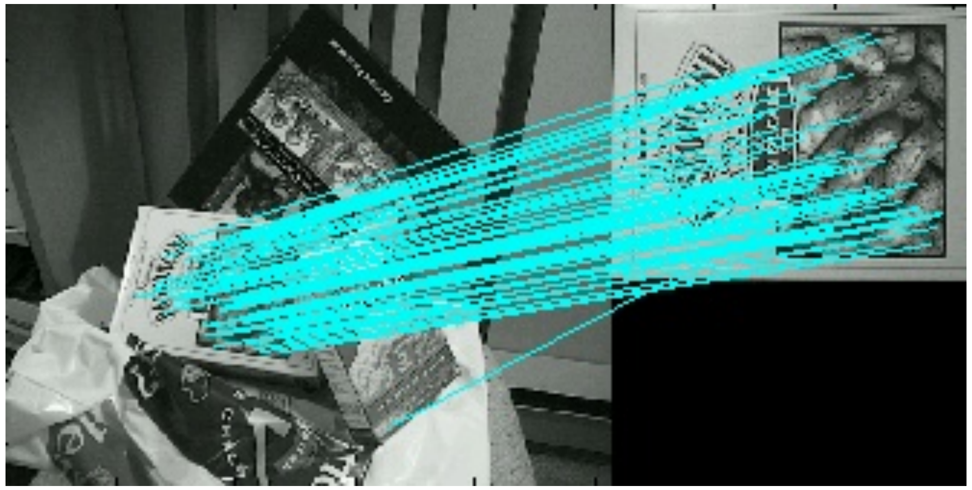
- 3D Visual Information (RGB-D Camera)
  - Voting-based Pose Estimation using Pair Features
    [ICRA'12, IROS'12]
  - Object Pose Tracking
    [IROS'13]

Related Work



[Harris, 92] [Drummond, PAMI'02]

- **Edge-based** approaches
  - Cheap to extract edges (real-time)
  - Applicable to textureless objects
  - Not distinctive enough
  - Might be stuck in local minima

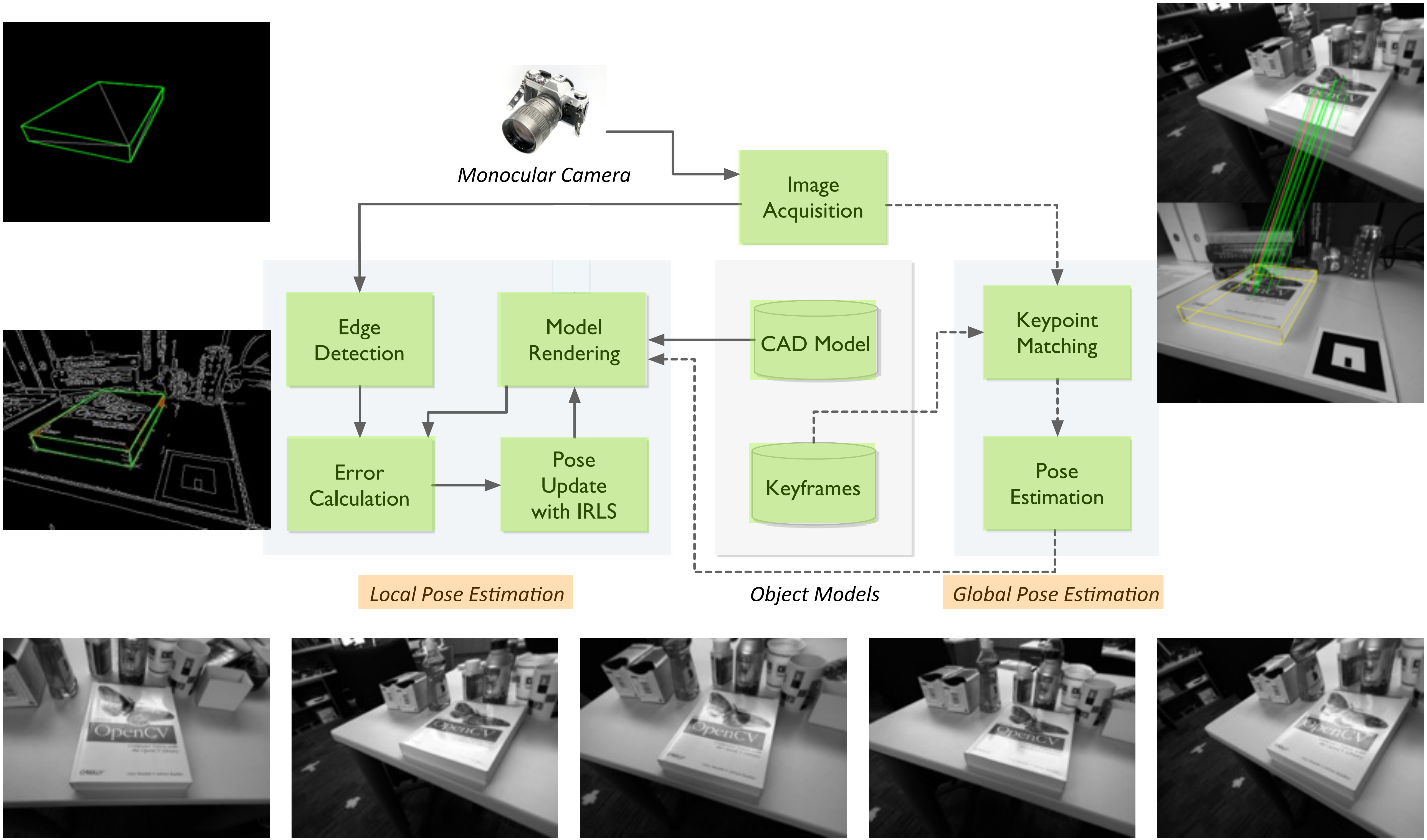


[Lowe, IJCV'04] [Gordon, 06] [Collet, IJRR'11]

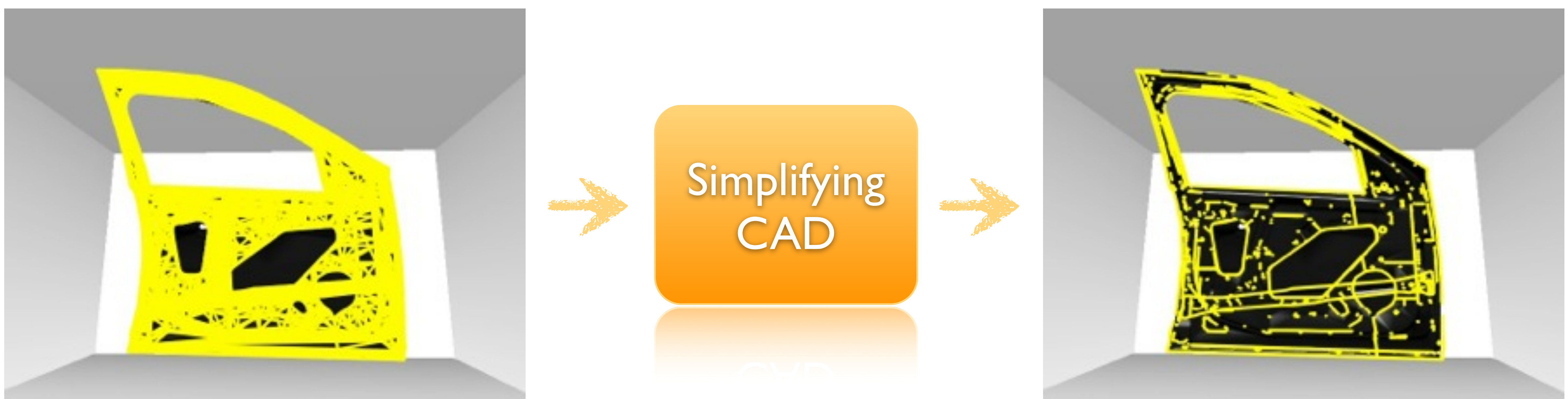
- **Keypoint-based** approaches
  - Good for initialization
  - Invariant to scale and rotation
  - Only applicable to textured objects
  - Computationally expensive

Complementary → Combining

Overview

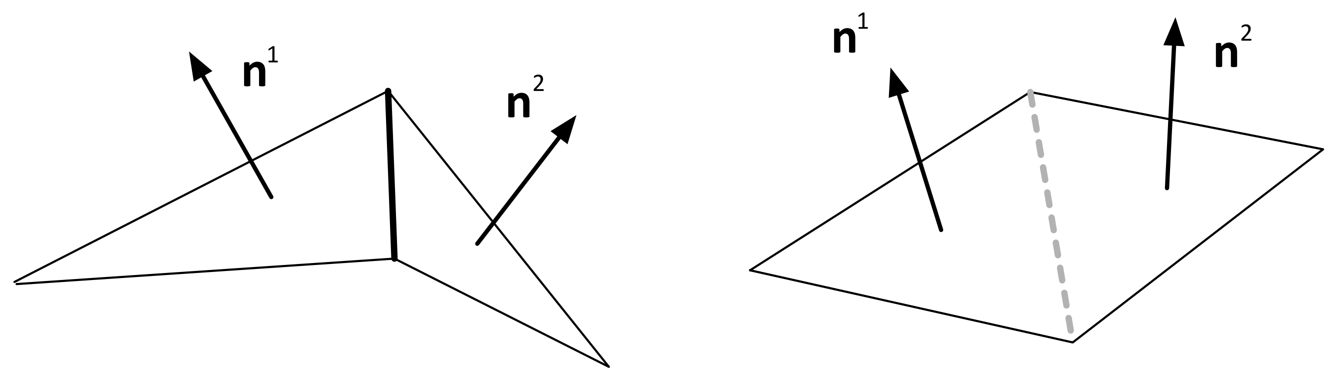

Monocular Camera
Image Acquisition
Edge Detection
Model Rendering
Error Calculation
Pose Update with IRLS
CAD Model
Keyframes
Keypoint Matching
Pose Estimation
Local Pose Estimation
Object Models
Global Pose Estimation

Simplifying CAD Model



- Original CAD models are too **complex**.
- Most edges in CAD do not appear in the real edge image.
- We should **simplify** in some way.

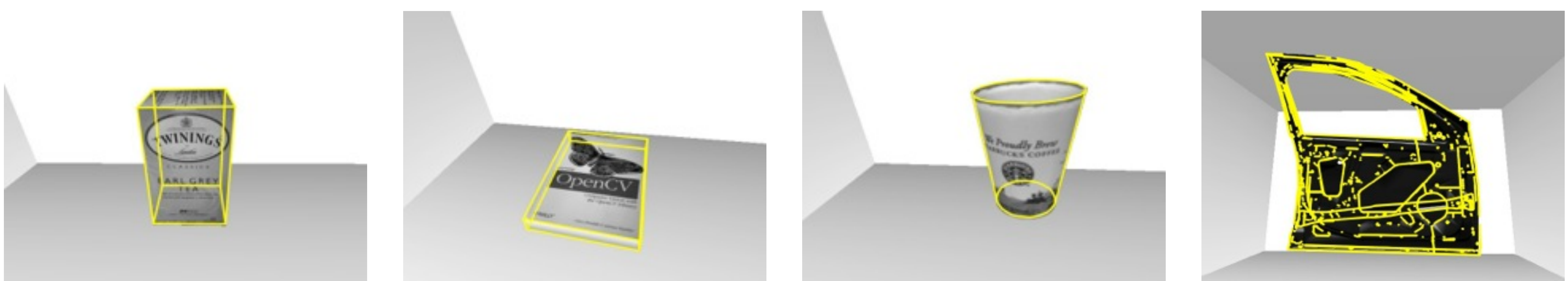
Salient Edges

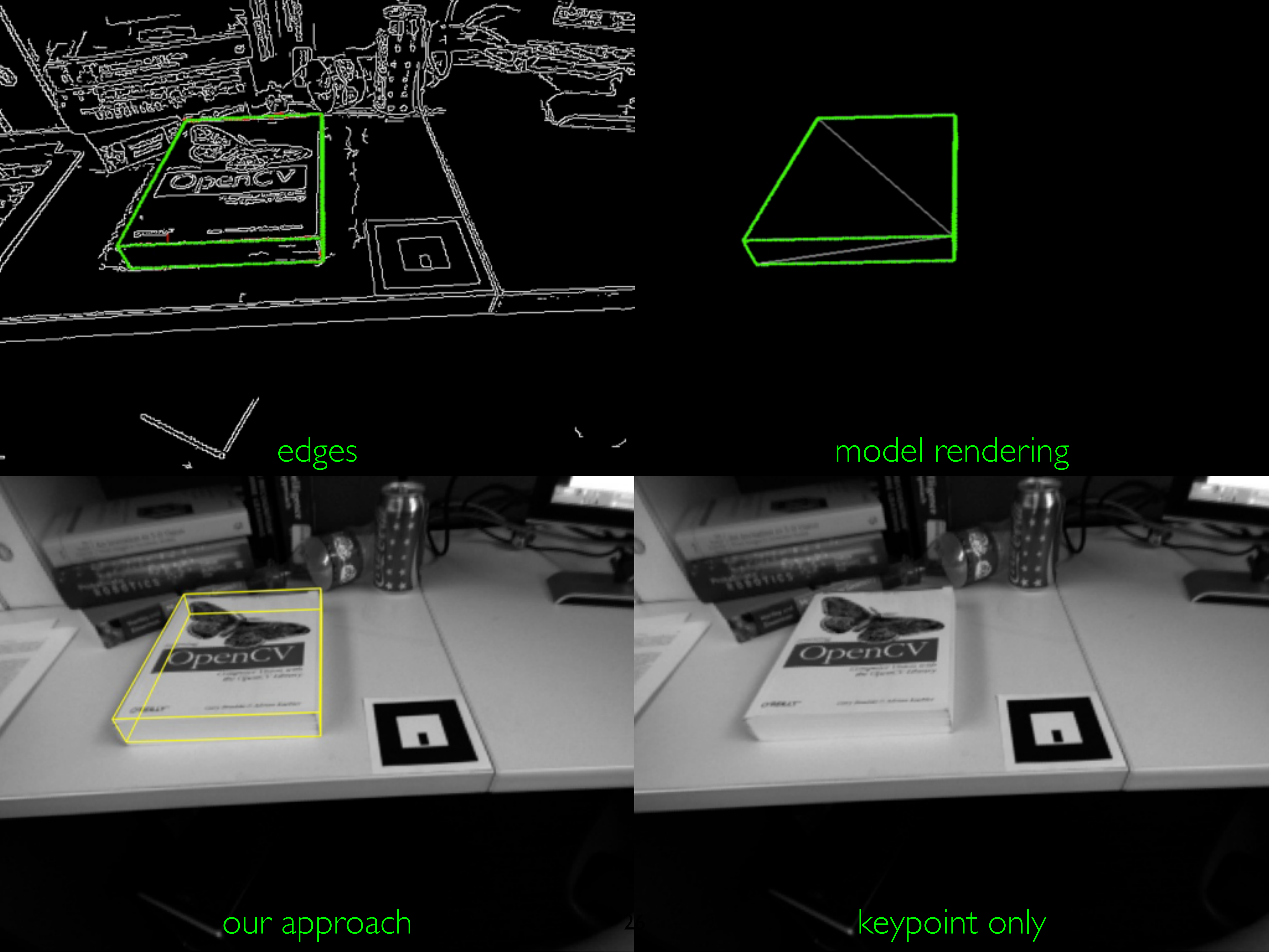


Sharp Edge **Dull Edge**

$$I(edge_i) = \begin{cases} 1 & \text{if } |\mathbf{n_i^1} \cdot \mathbf{n_i^2}| \leq \tau_s \\ 0 & \text{otherwise} \end{cases}$$

- Use face normal vectors
- Automatically determine salient edges which are more likely to be **visible** in images




edges
model rendering
our approach
keypoint only
OpenCV

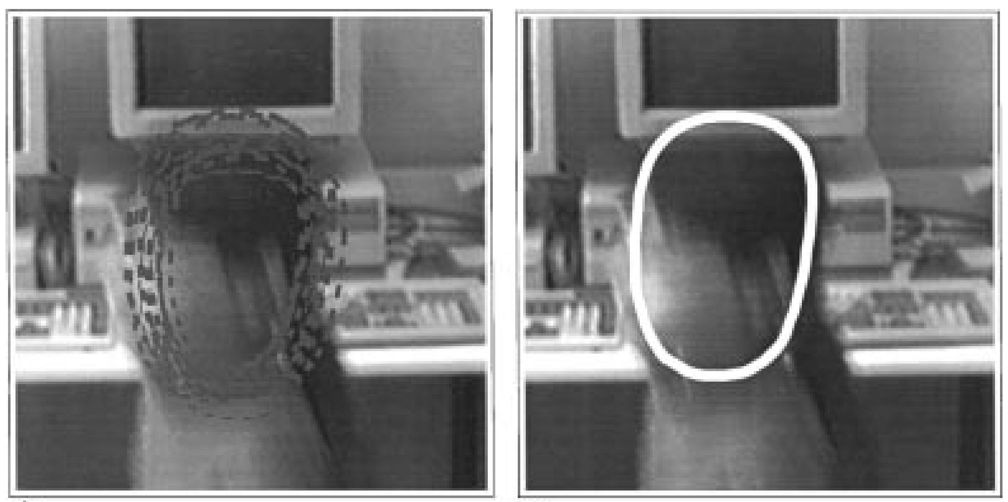
Limitation

- **Single** pose hypothesis
  - **Wrong** prior pose → not converging to global optimum
- **Ambiguous** edges
- Stuck in **local minima**
  - Highly cluttered environment
  - Occlusions

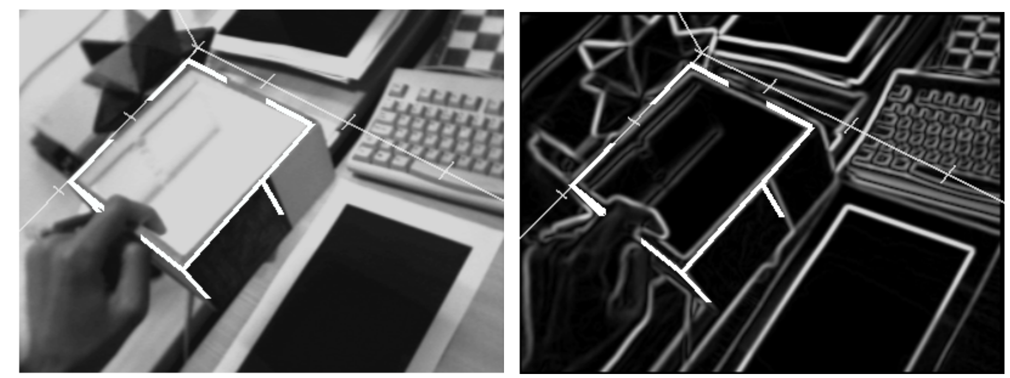
- **Multiple** pose hypotheses
- Particle Filtering

Related Work

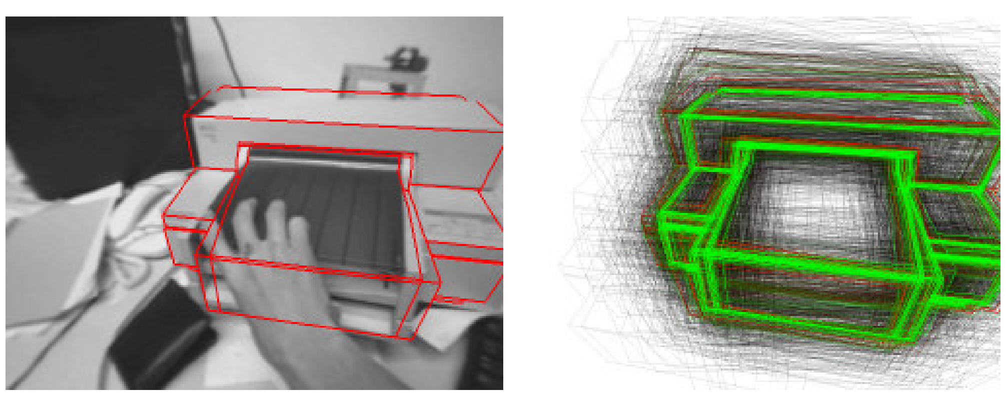
Particle Filtering using Edges



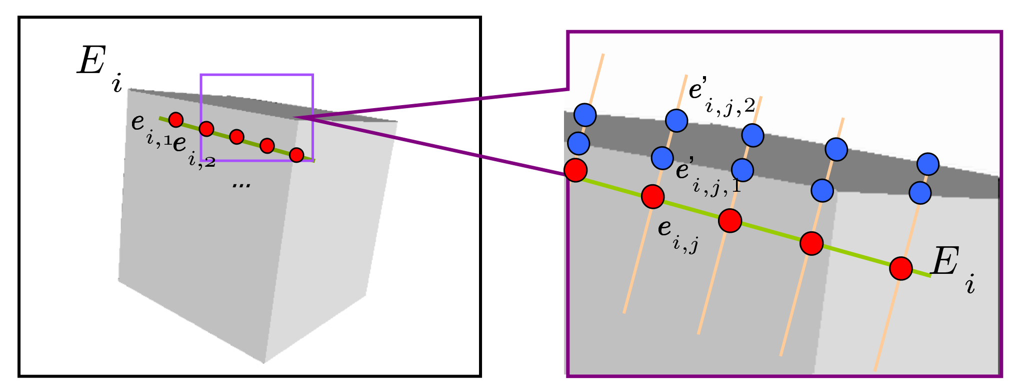
[Isard, IJCV'98] Condensation in 2D



[Pupilli, ICPR'06] PF for 3D edge-based tracking



[Klein, BMVC'06] PF for complex object tracking



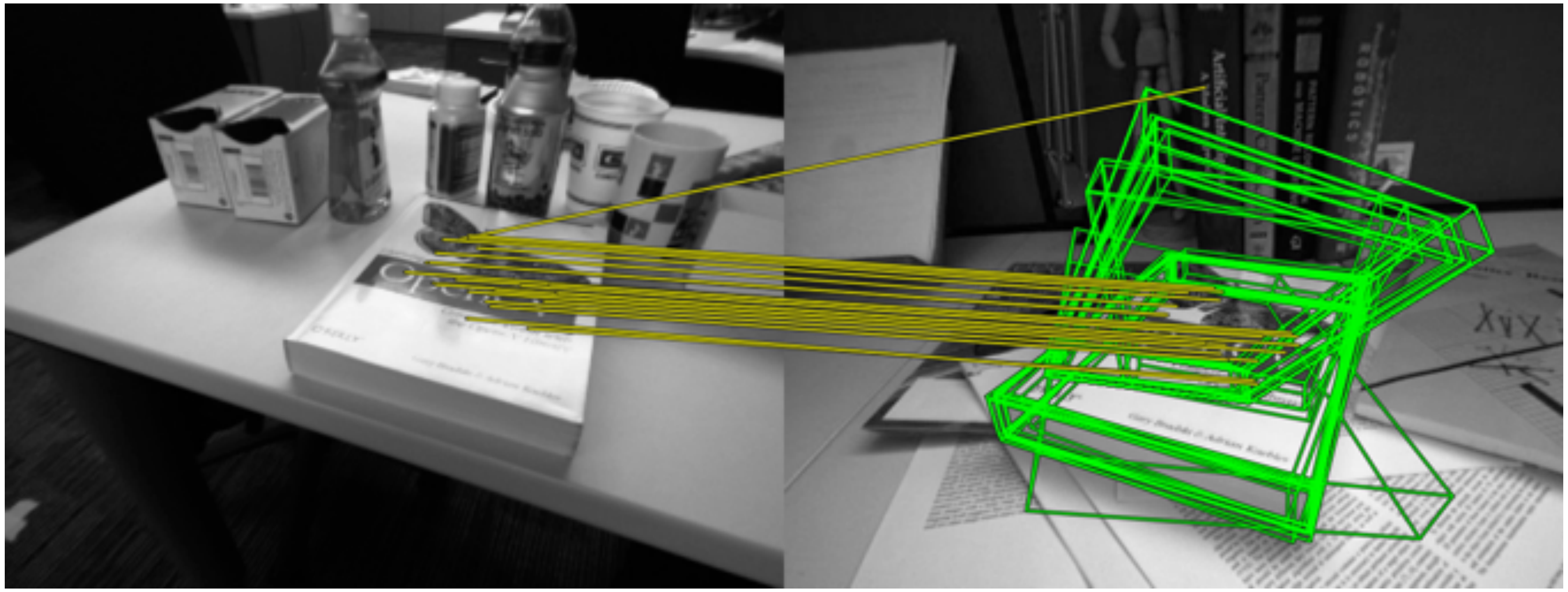
[Teuliere, ICRA'10] Multiple edge correspondences

Contributions

- Given starting pose
- Gaussian random walk
- No re-initialization

- Initialization
- AR(1) state dynamics
- Auto re-initialization

Initialization



Keyframe (image, 2D & 3D keypoints) **Input image**

Initialize the particle filter using keypoints

- Given 2D-3D keypoints correspondences
- Randomly choose a set of minimum correspondences
- Solve PnP problem to estimate candidate poses
- Weights proportional to inlier ratio of remaining correspondences
- Importance sampling

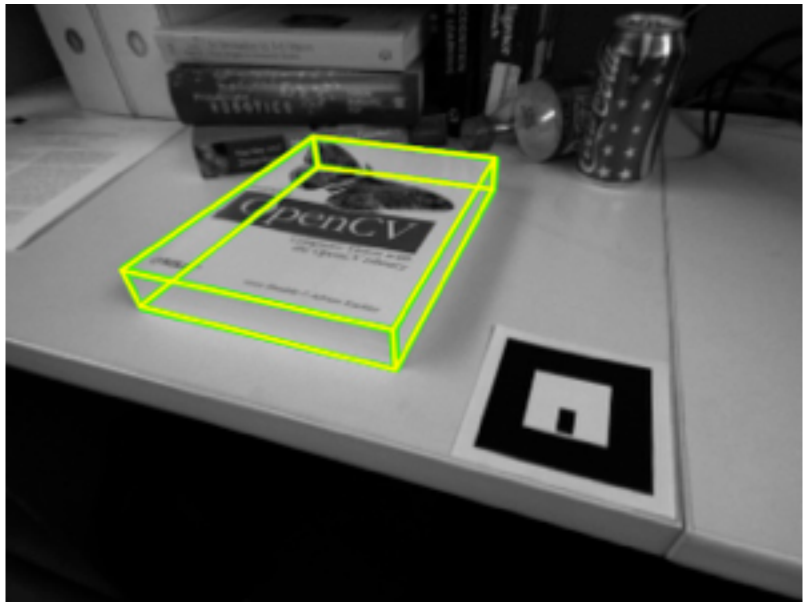
AR Dynamics

$$X_t = X_{t-1} \cdot \exp(A_{t-1} + dW_t \sqrt{\Delta t}),$$

$$A_{t-1} = a \log(X_{t-2}^{-1} X_{t-1})$$

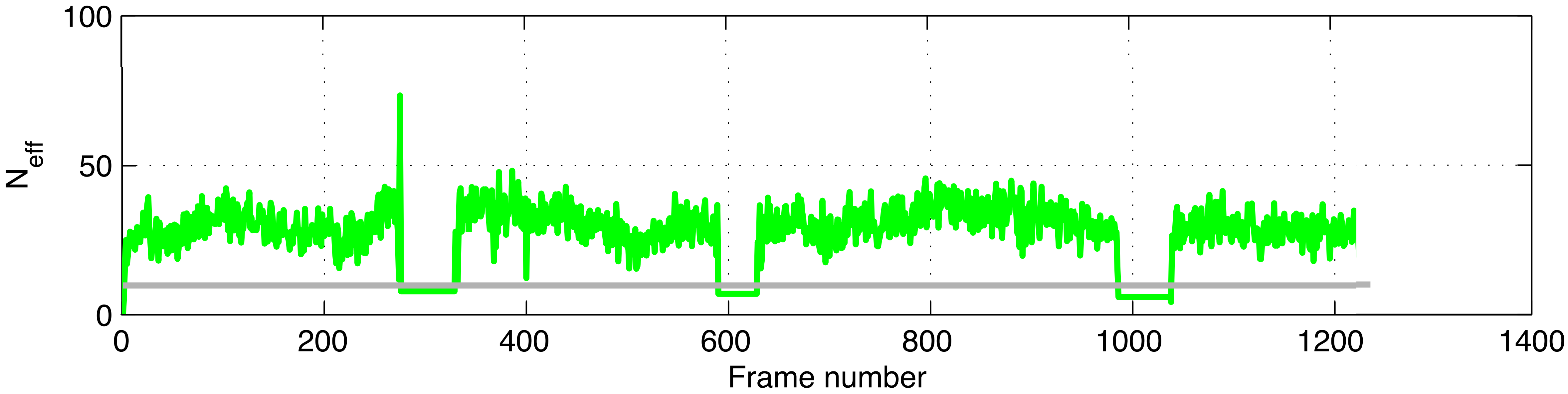
- Instead of Gaussian random walk models
- Linear prediction based on previous states
- Propagate particles more effectively

Re-initialization

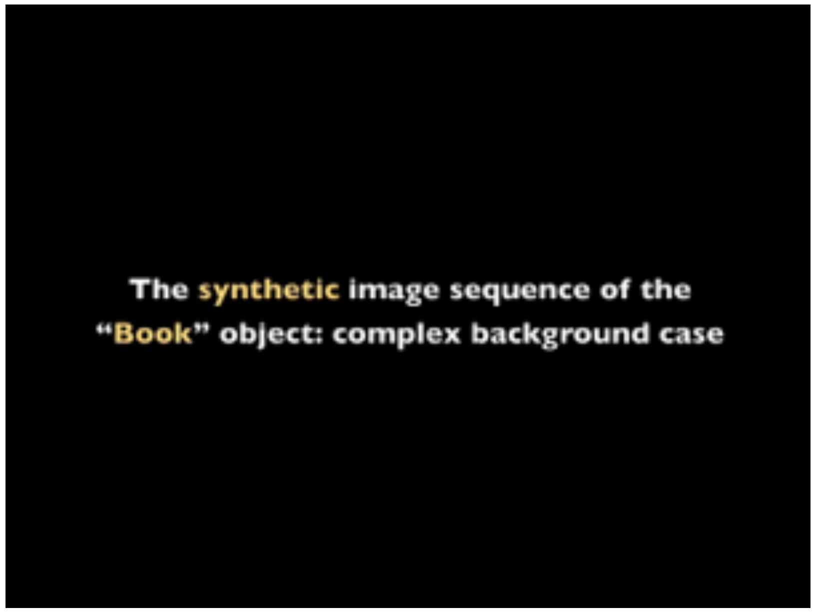


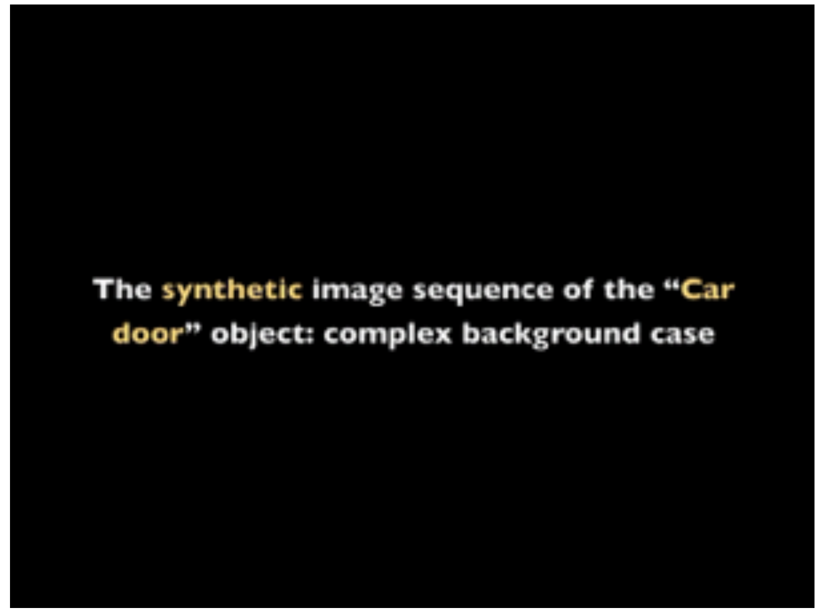
$$\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N} (\tilde{\pi}^{(i)})^2}$$

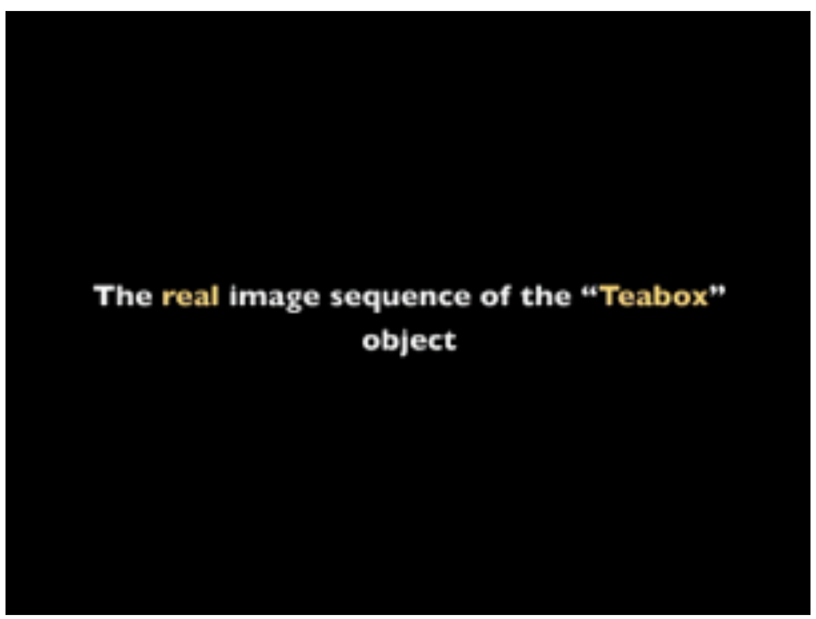
Effective number of particle size



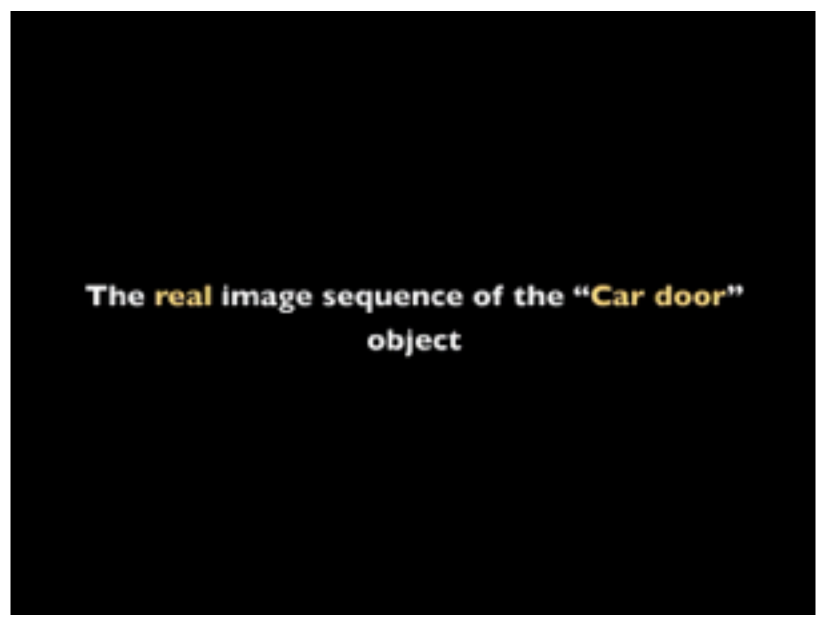
Experiments







Single vs. Multiple pose hypotheses

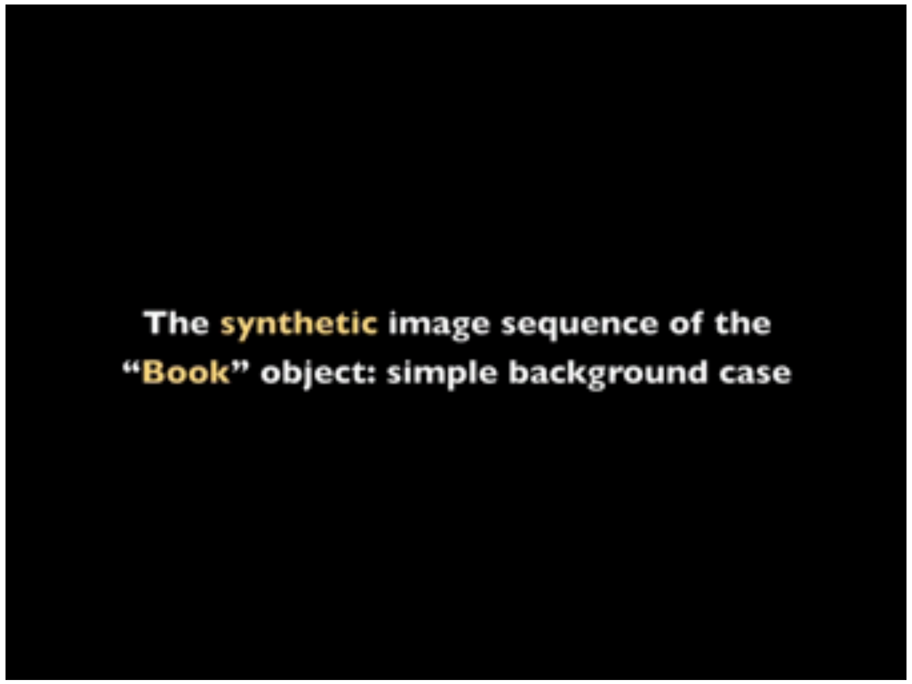


with vs. without AR state dynamics

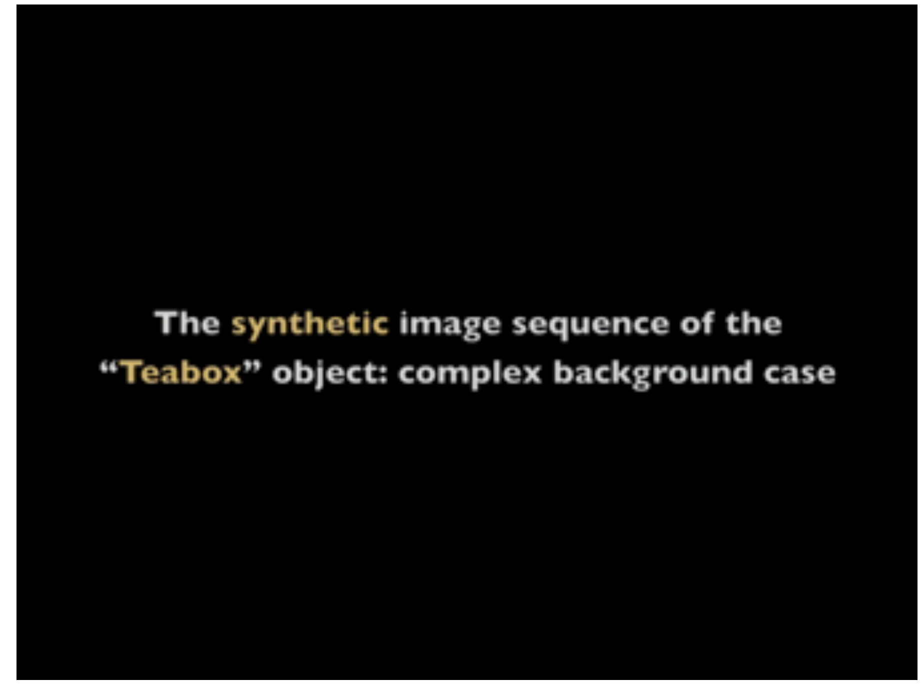
Reinitialization exp.

Ours vs. BLORT

SYNTHETIC

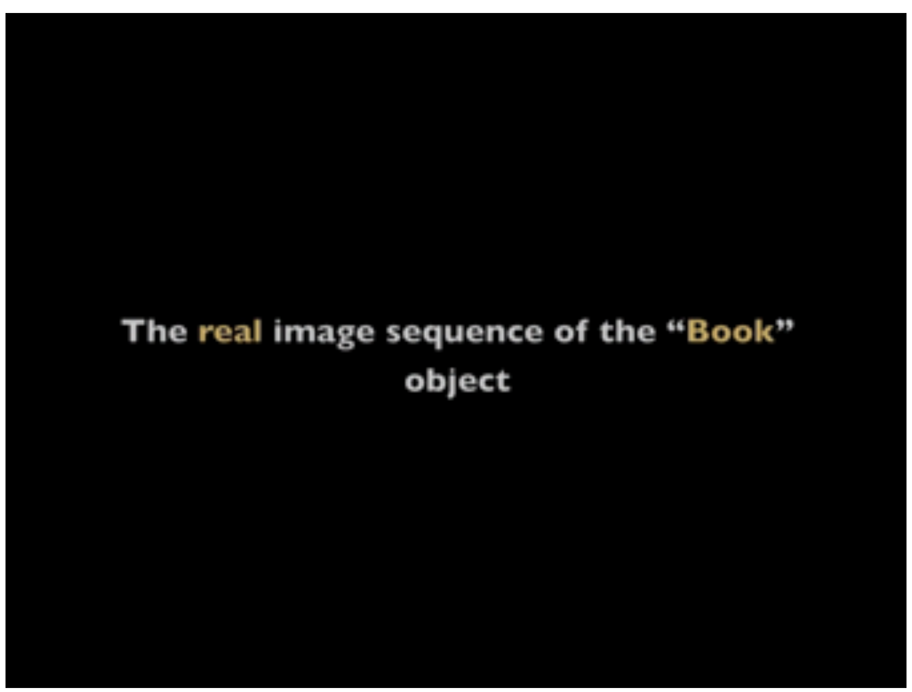


Book



Teabox

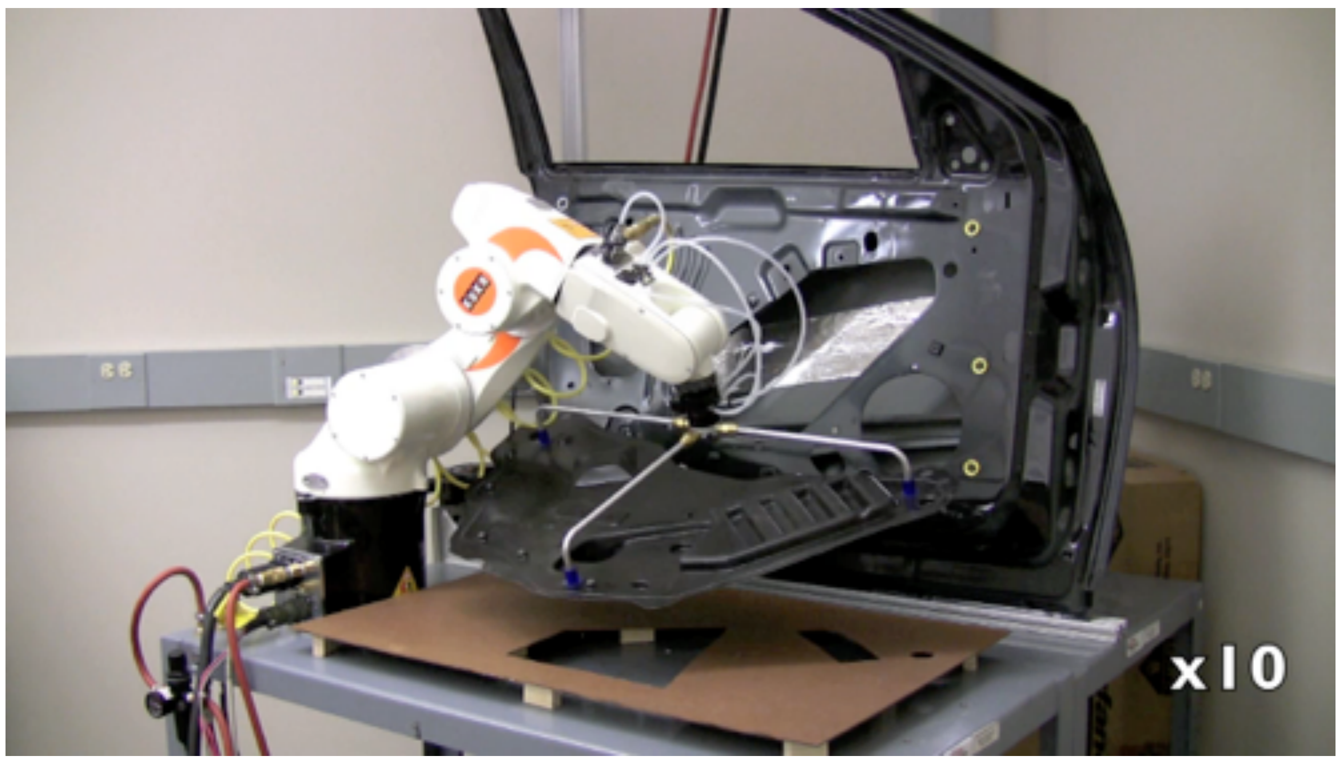
REAL

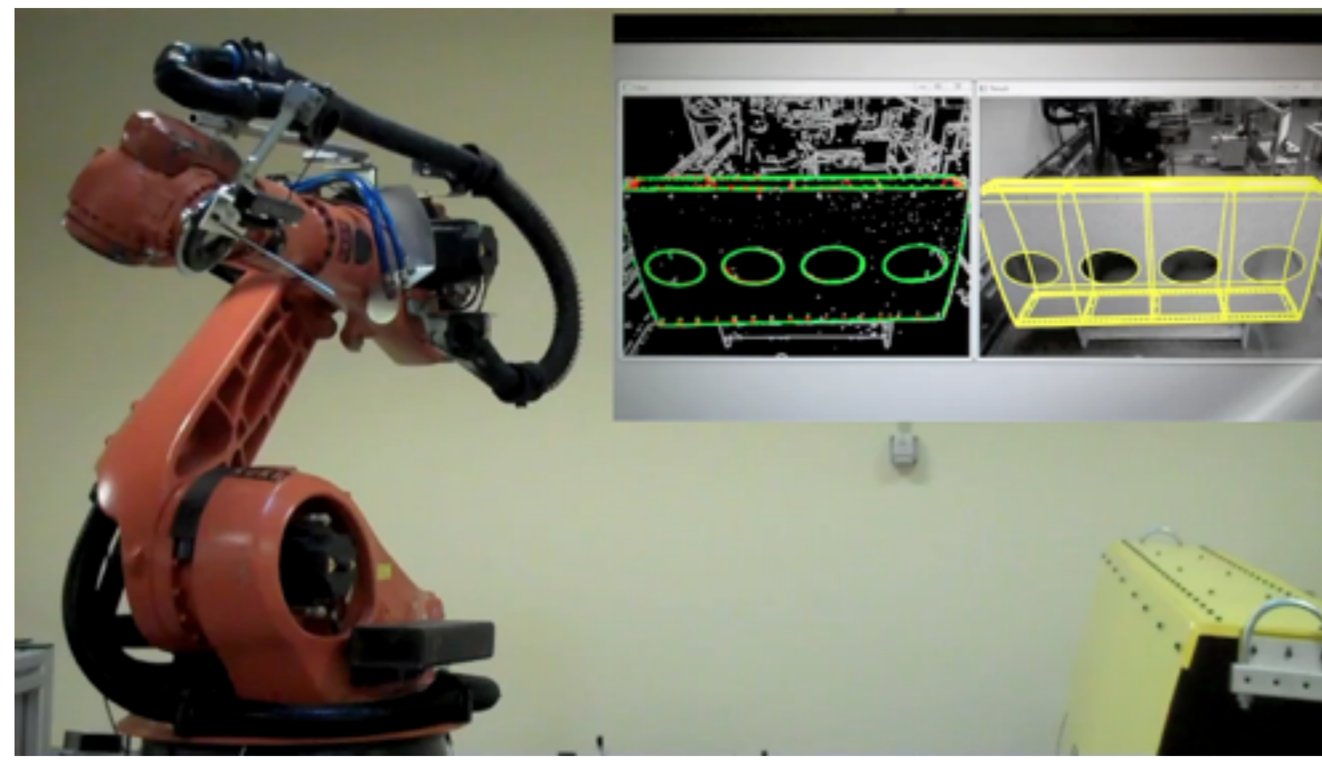


Book

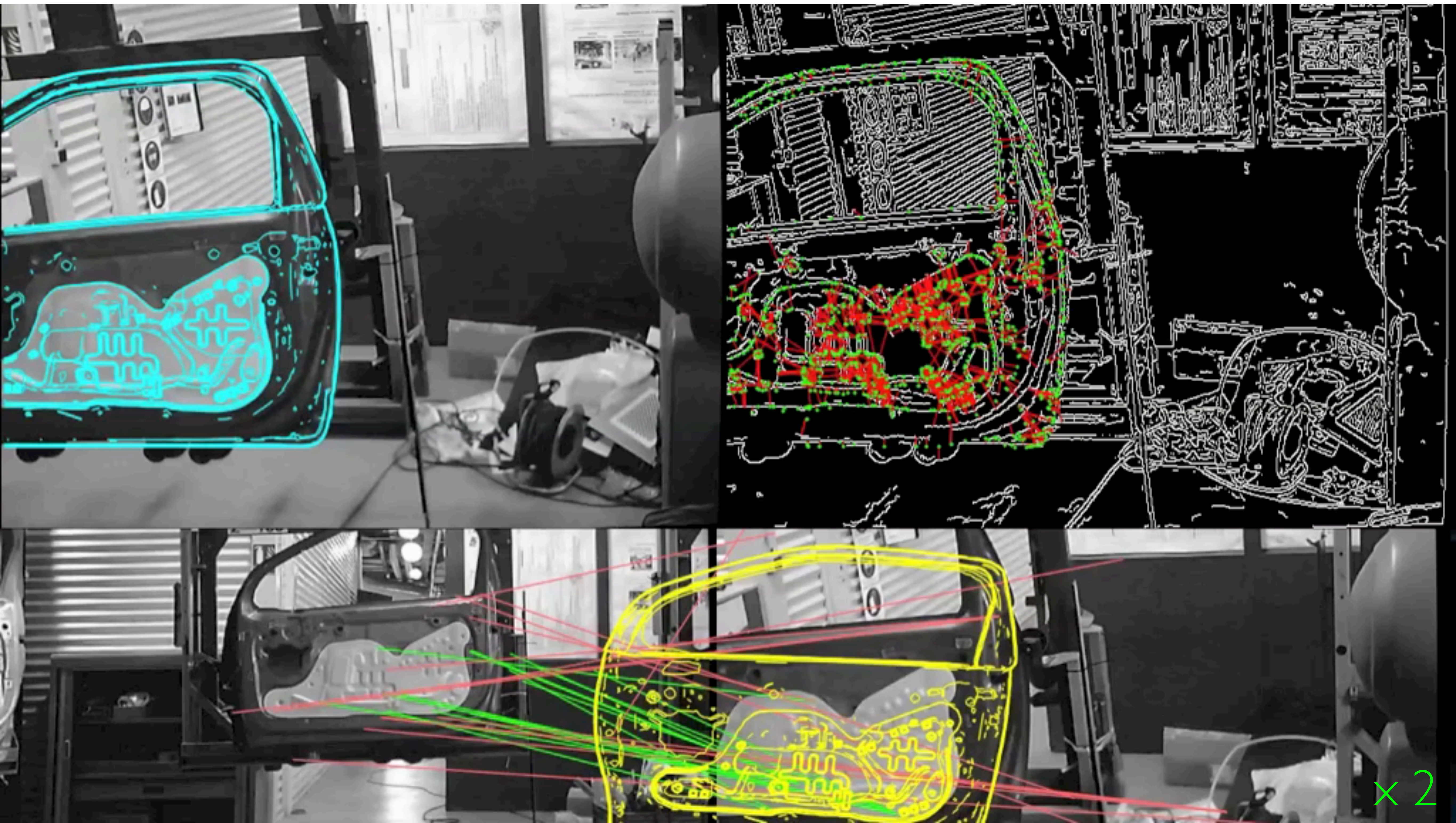
Cup

Robotic Assembly





Robotic Assembly



Approaches

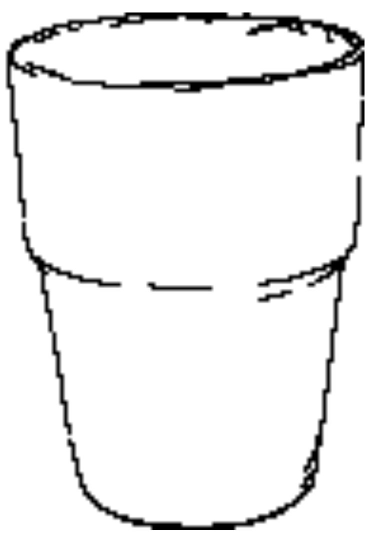
- 2D Visual Information (Monocular Camera)
  - Combining Keypoint and Edge Features [ICRA'10, ICRA'11, IJRR'12]
  - **Extending to Textureless Objects** [IROS'12]

geometric

- 3D Visual Information (RGB-D Camera)
  - Voting-based Pose Estimation using Pair Features [ICRA'12, IROS'12]
  - Object Pose Tracking [IROS'13]

Textureless Objects

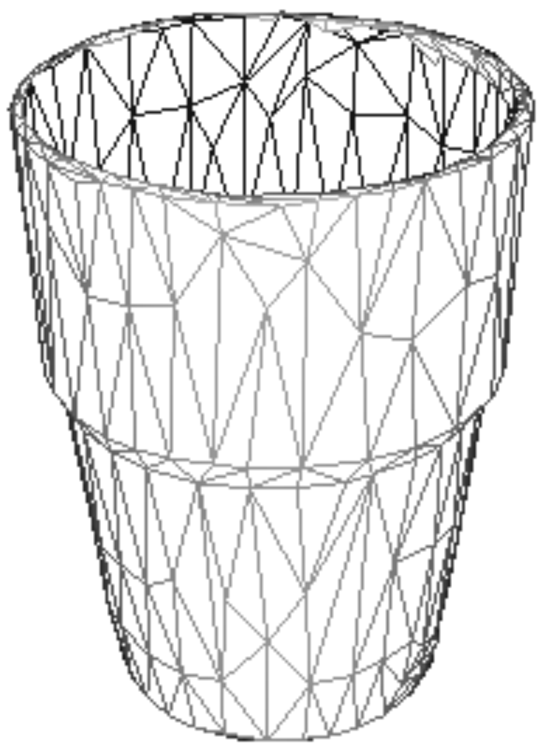
Textureless object



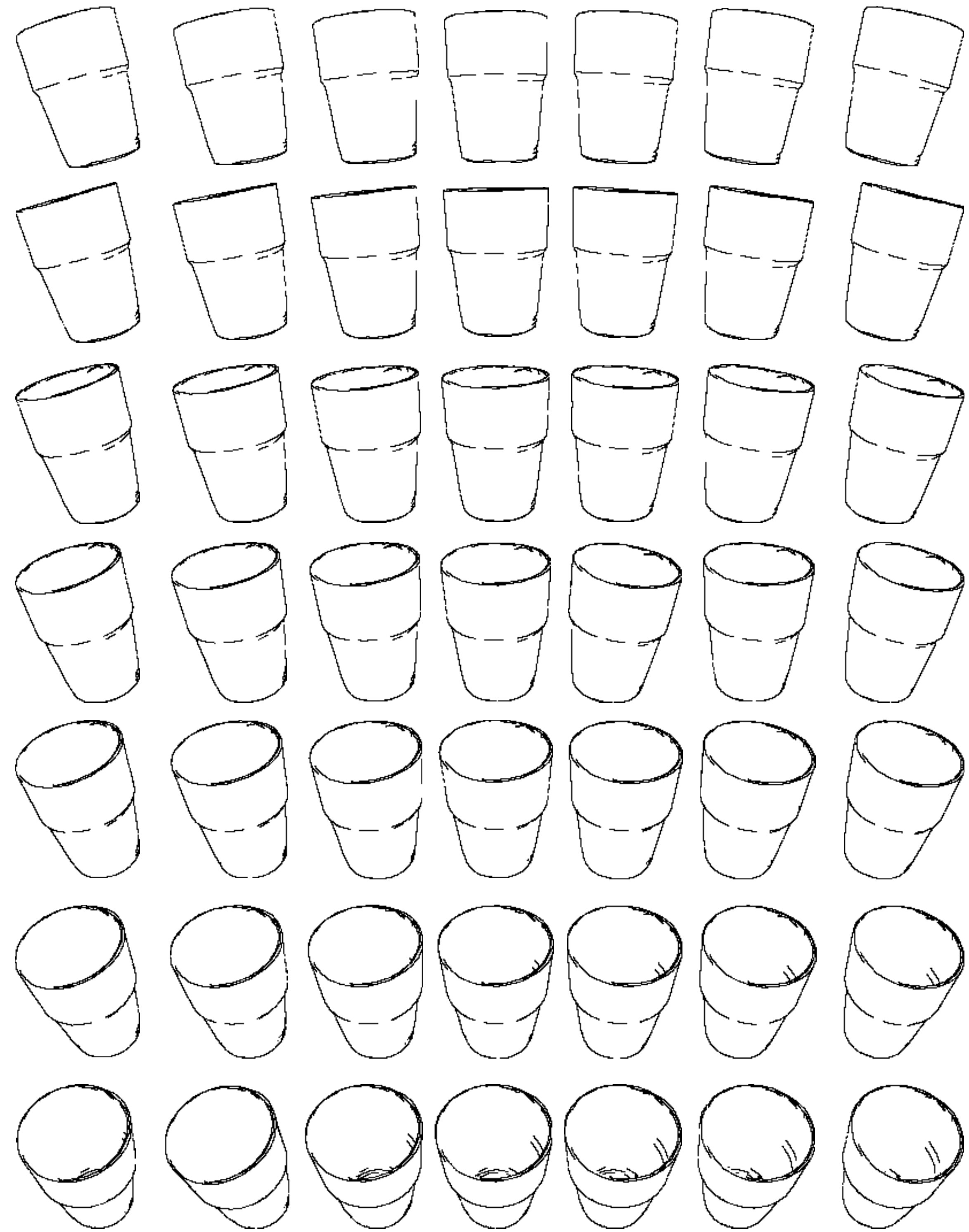
Edge template

- Edges (or boundaries) are preferred for textureless objects.
- From CAD model to Edge templates
- Efficient chamfer matching [Liu, CVPR'10]
- Coarse 3D pose estimation from 2D chamfer matching results
- Annealing Process after Initialization

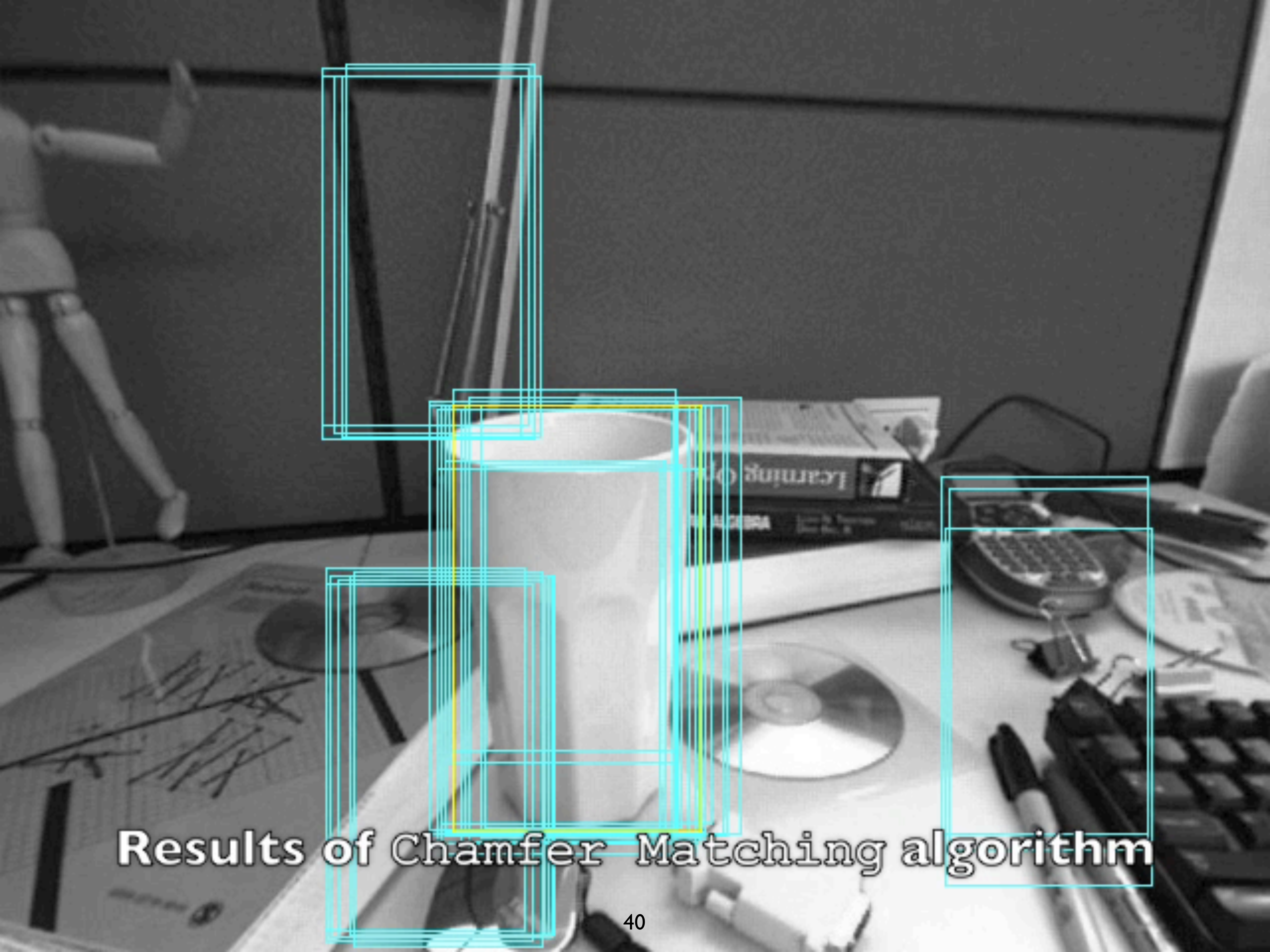
Edge Templates

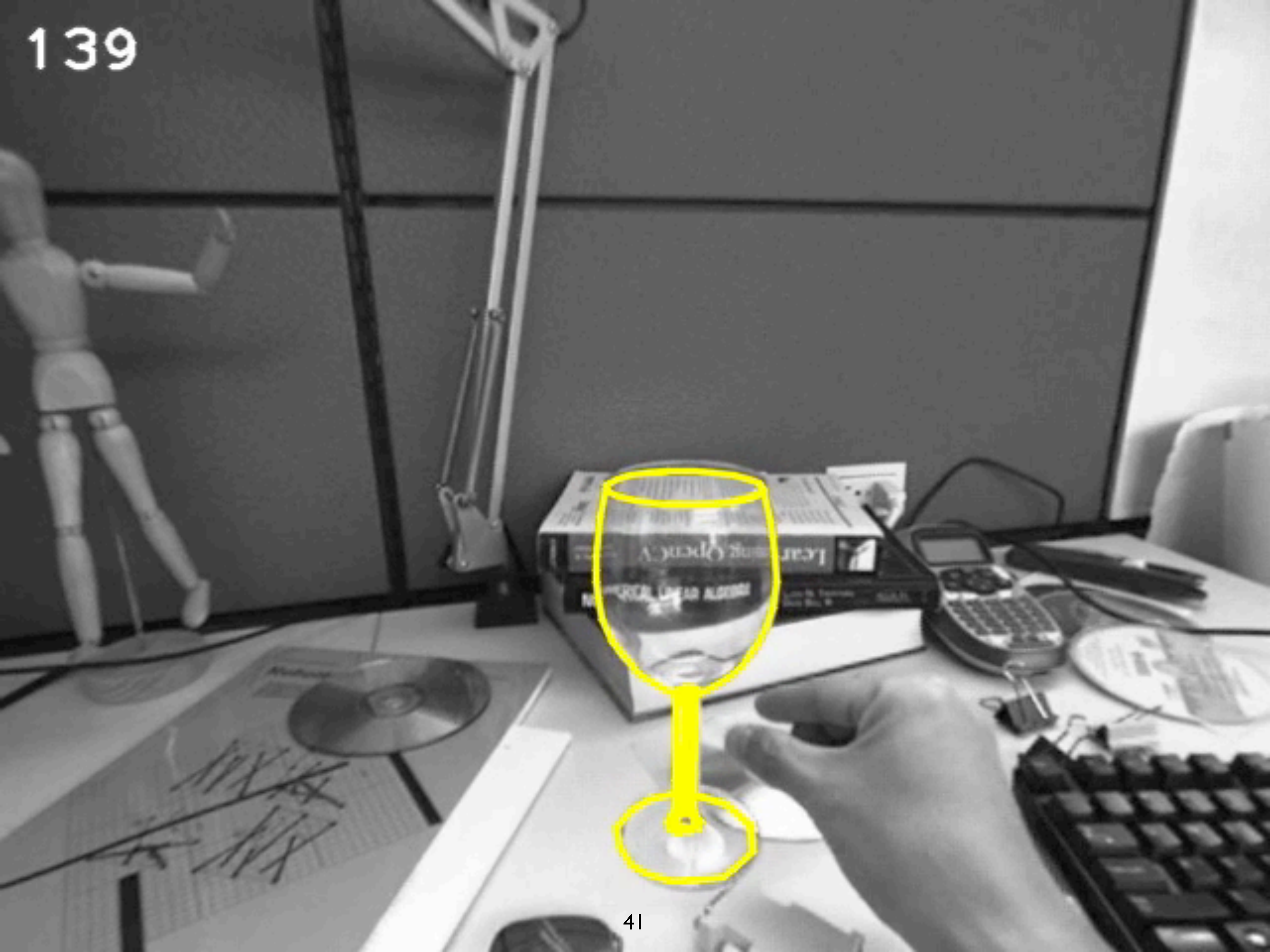


CAD model



Edge Templates


Results of Chamfer Matching algorithm


139

Approaches

- 2D Visual Information (Monocular Camera)
  - Combining Keypoint and Edge Features [ICRA'10, ICRA'11, IJRR'12]
  - Extending to Textureless Objects [IROS'12]

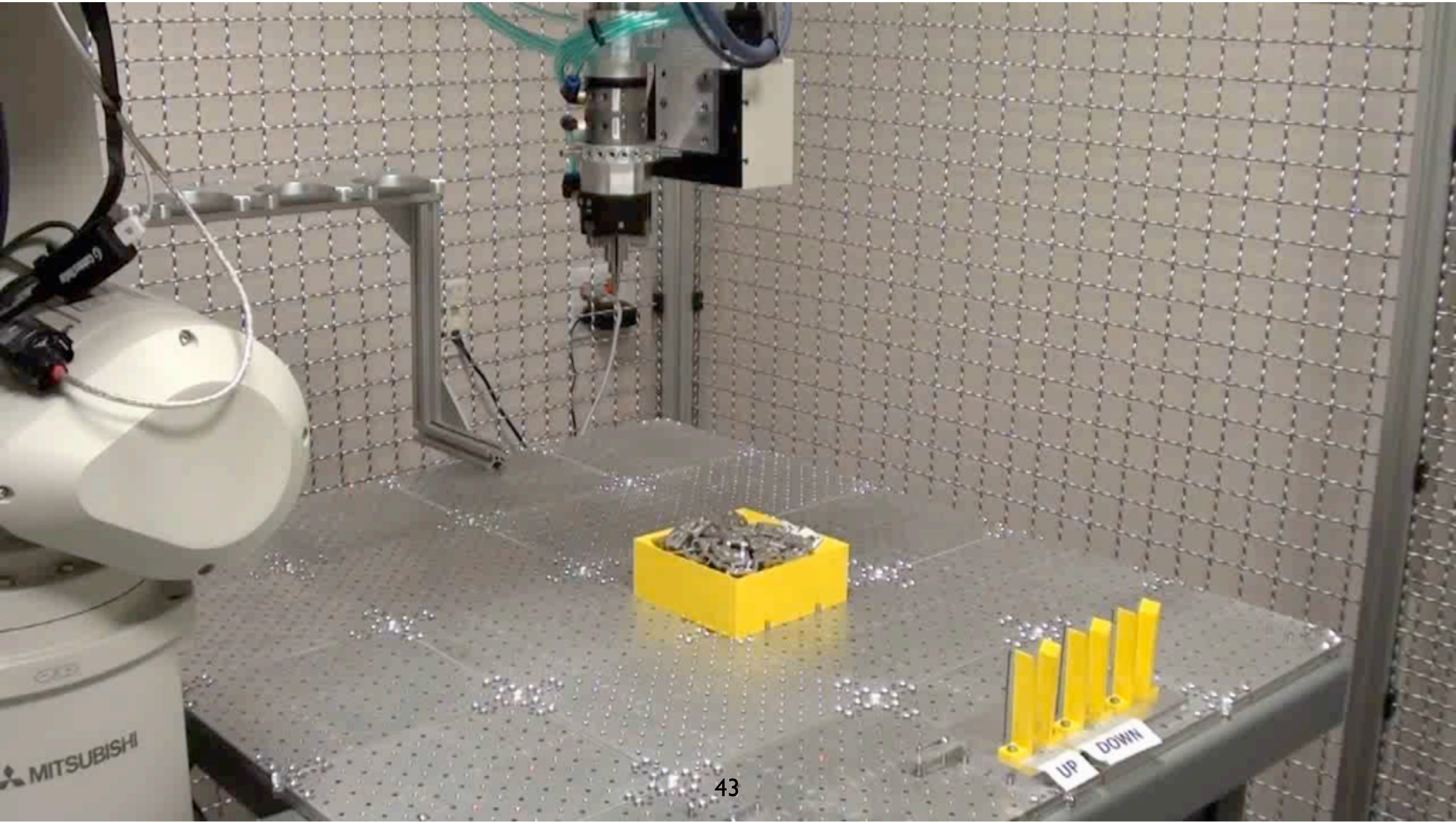
- 3D Visual Information (RGB-D Camera)
  - Voting-based Pose Estimation using Pair Features [ICRA'12, IROS'12]
  - Object Pose Tracking [IROS'13]

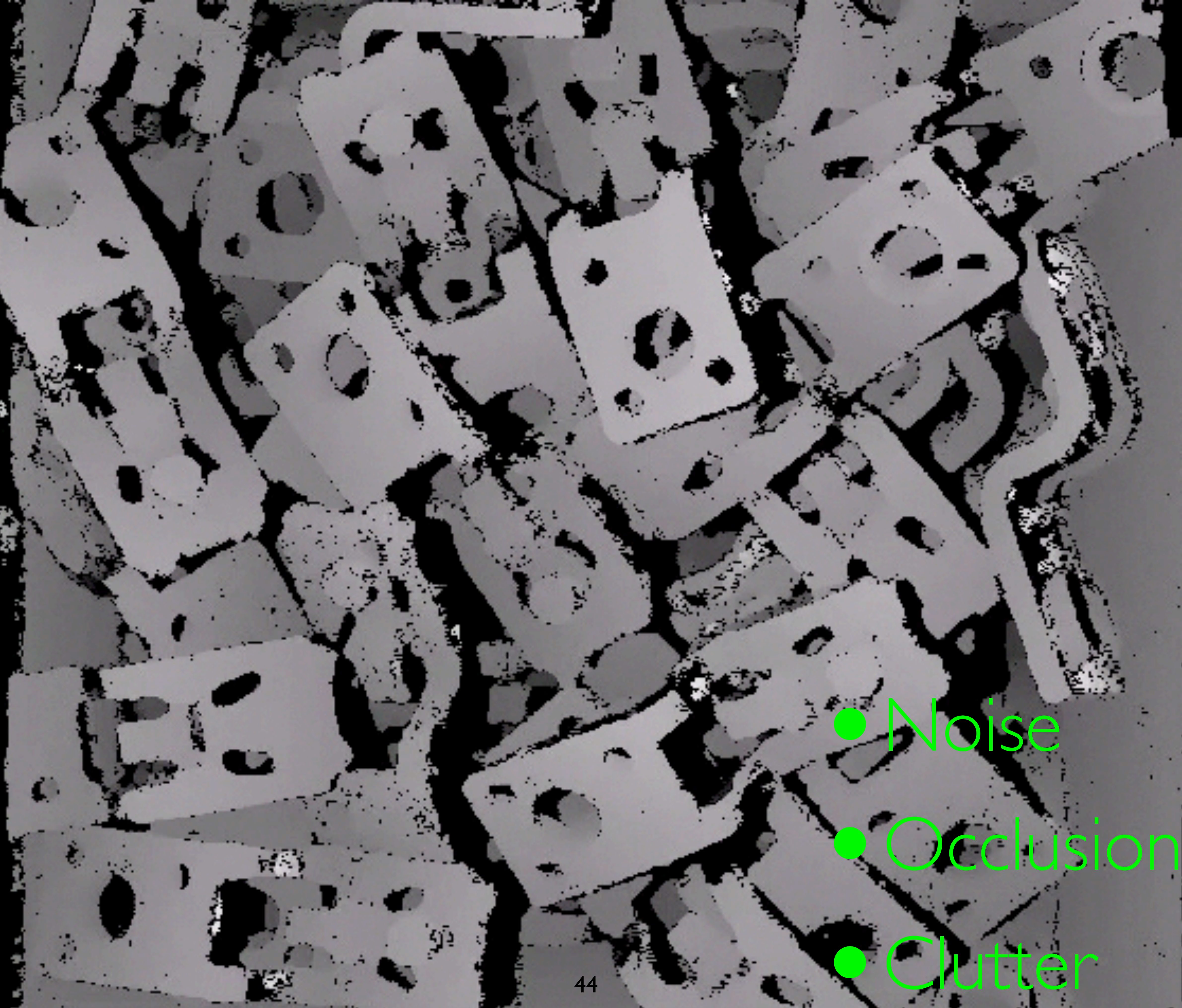
geometric

geometric + photometric

MITSUBISHI ELECTRIC
RESEARCH LABORATORIES

Overview



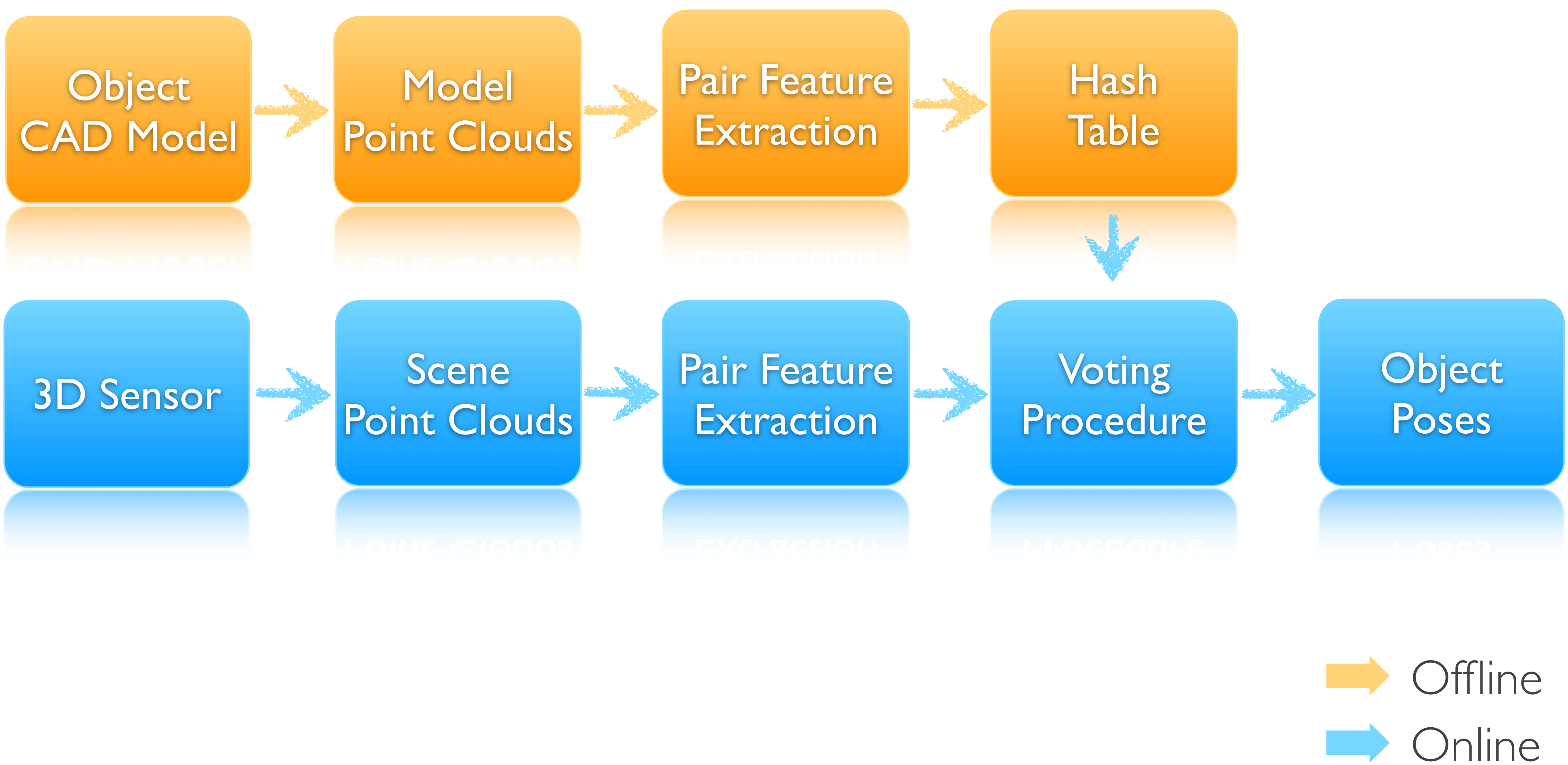


- Noise
- Occlusions
- Clutter

Contributions

- Exploiting objects' ***boundary*** information
  - B2B, S2B, and L2L features
- Better for **planar** objects
- **Sparser** primitives
- More **efficient**

Flowchart


Object
CAD Model
Model
Point Clouds
Pair Feature
Extraction
Hash
Table
3D Sensor
Scene
Point Clouds
Pair Feature
Extraction
Voting
Procedure
Object
Poses
Offline
Online

Geometric Primitives

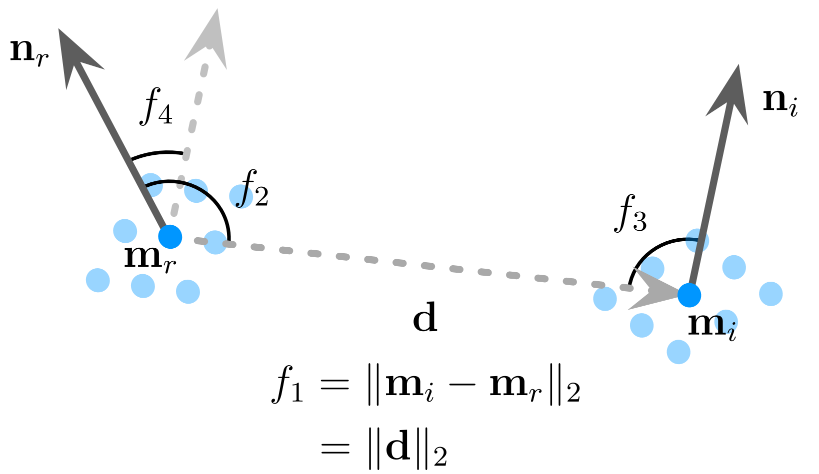
Surface Points

Boundary Points

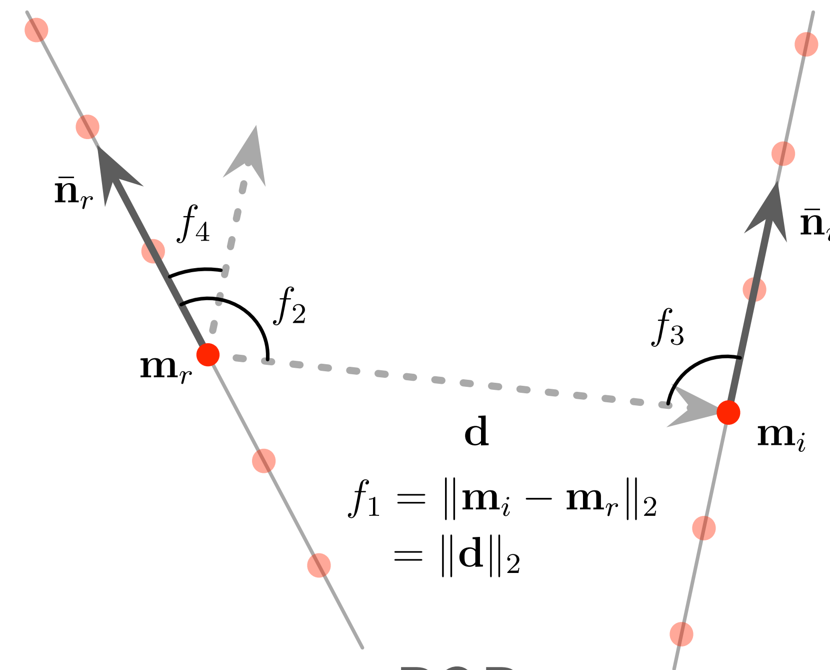
Surface & Boundary

Lines

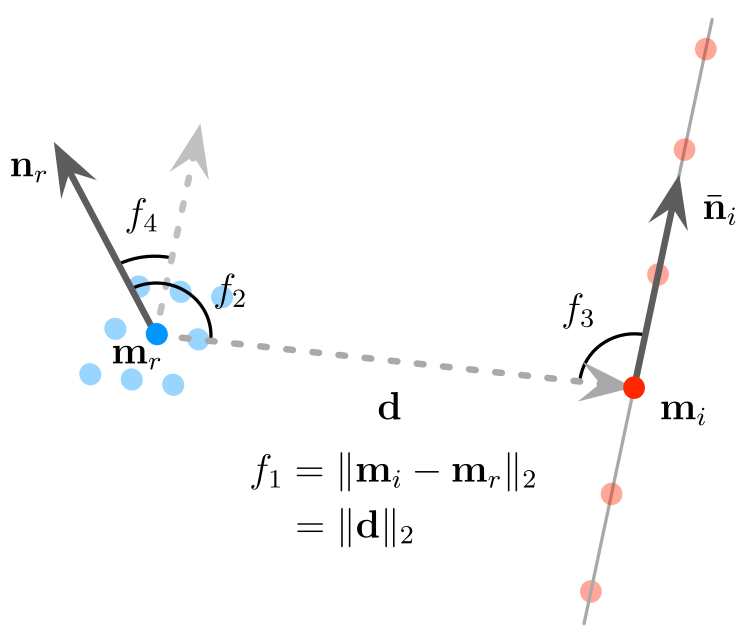
Pair Features



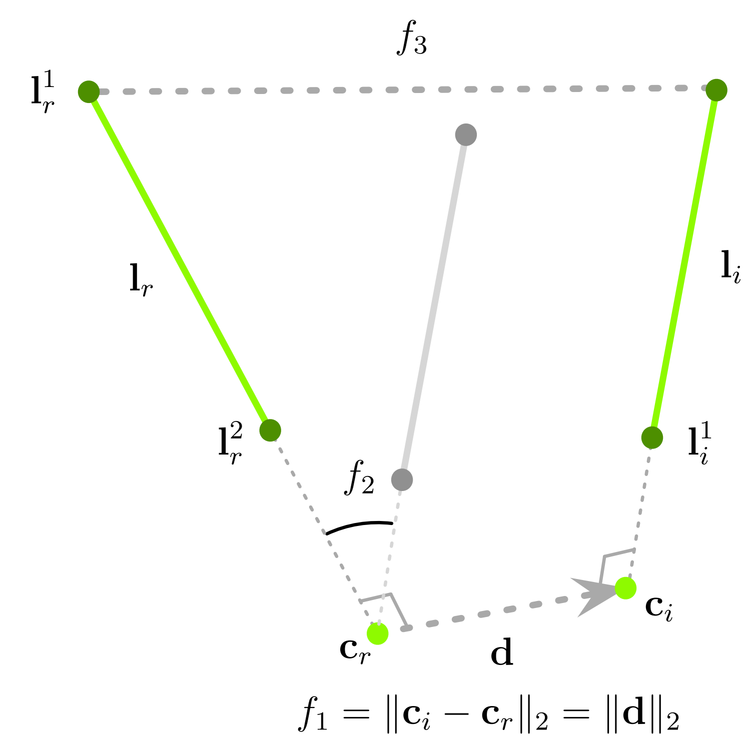
S2S (Drost et al.)



B2B

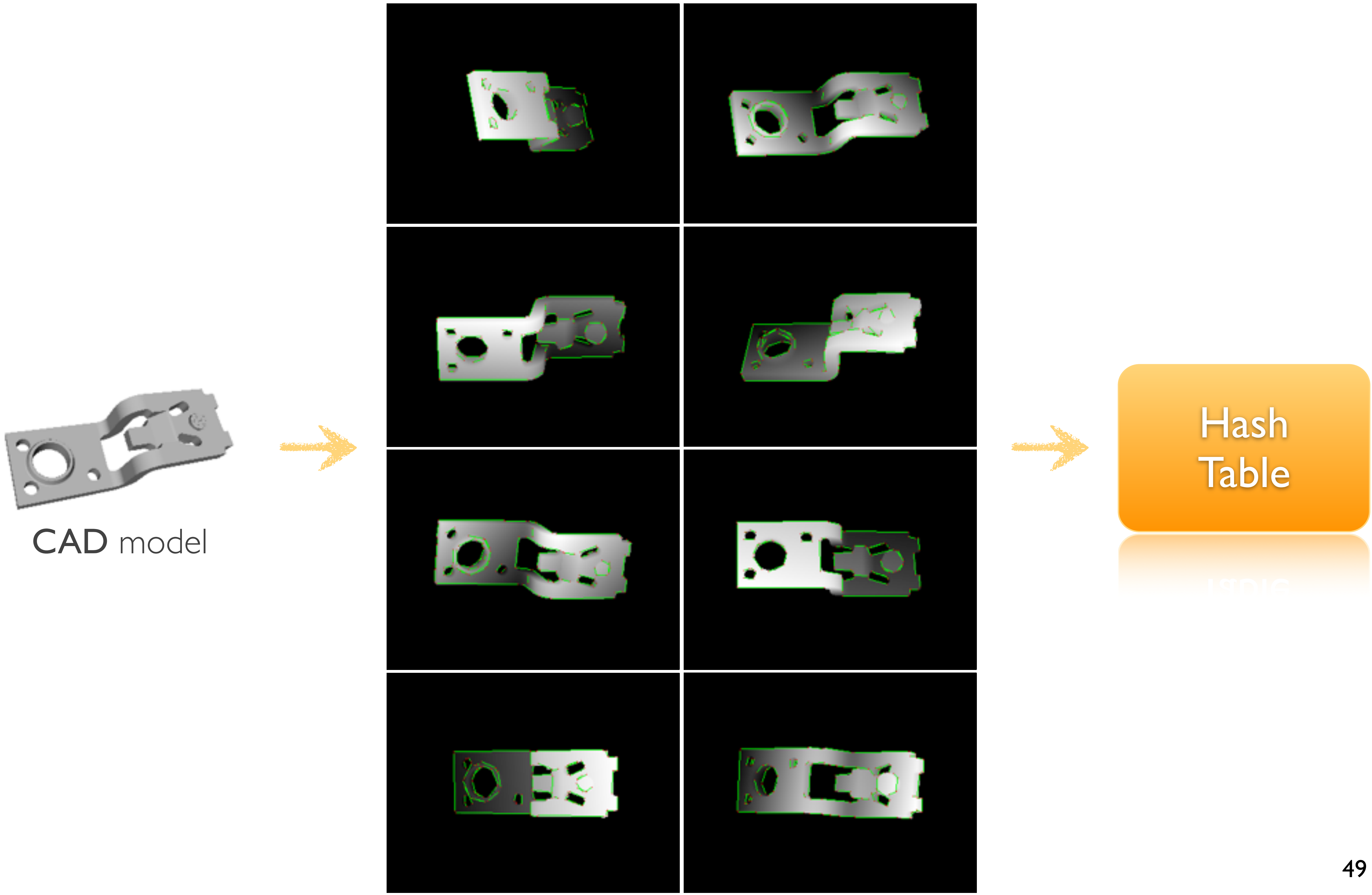


S2B

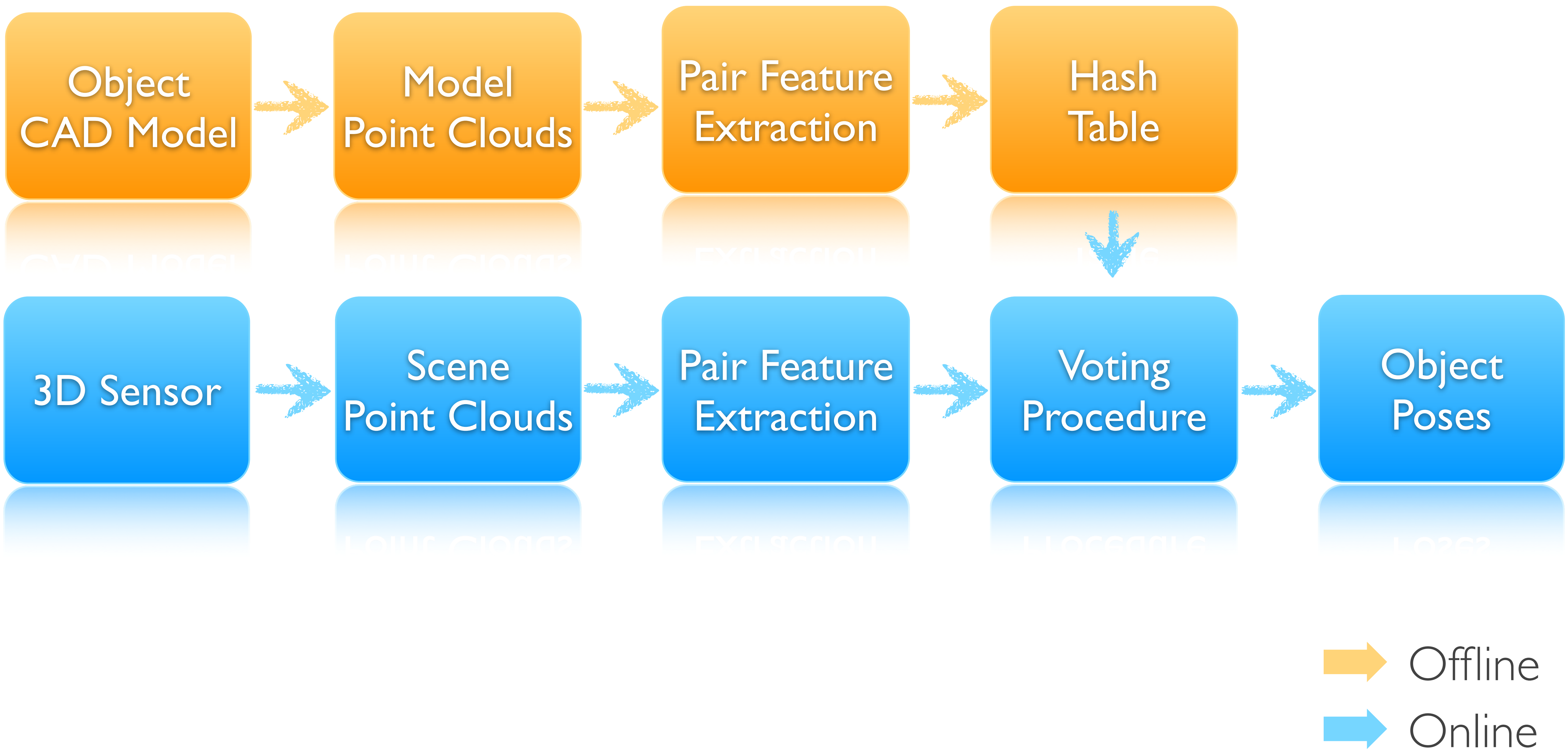


L2L

Object Learning



Flowchart


Object
CAD Model
Model
Point Clouds
Pair Feature
Extraction
Hash
Table
3D Sensor
Scene
Point Clouds
Pair Feature
Extraction
Voting
Procedure
Object
Poses
Offline
Online

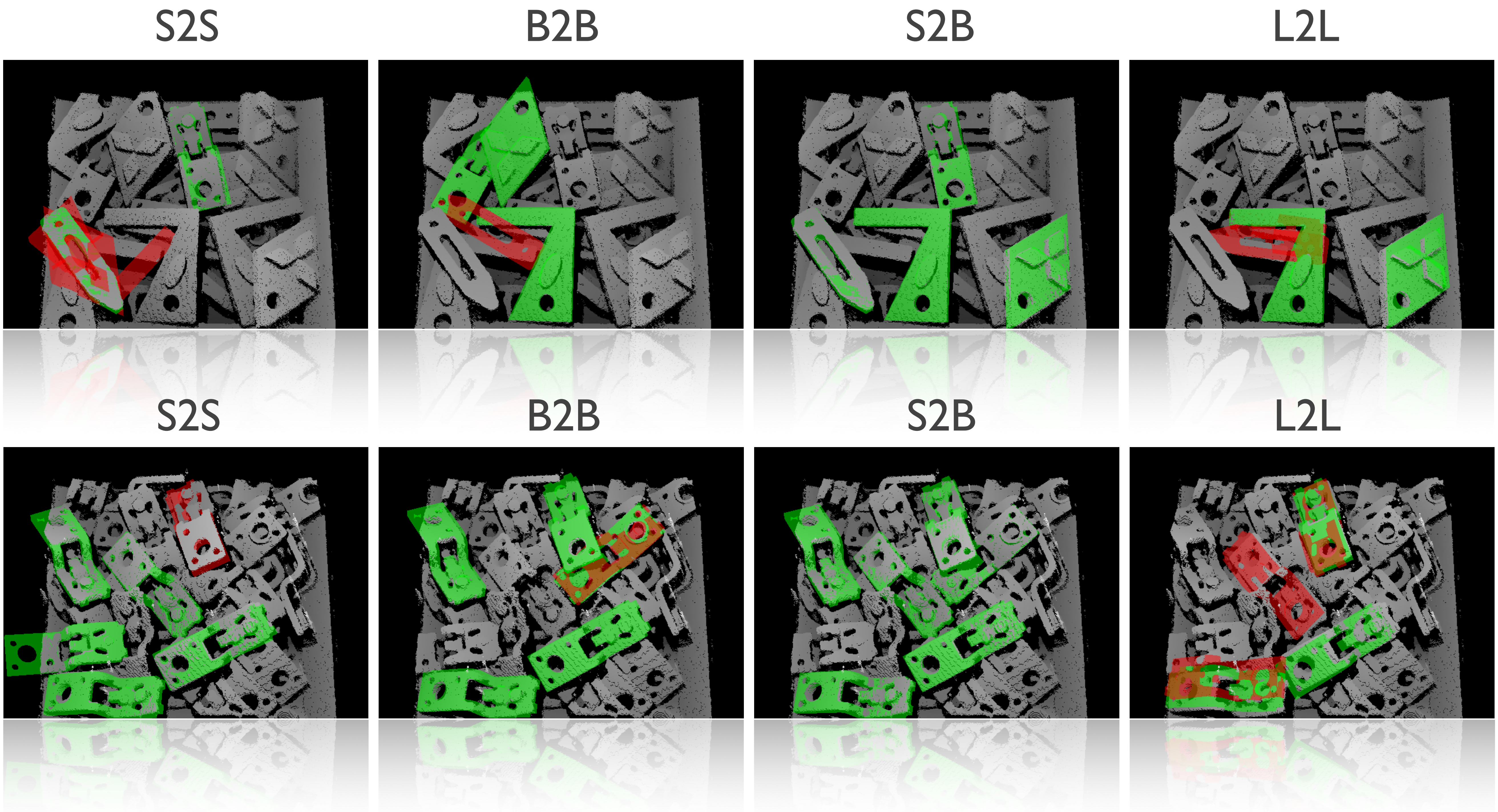
Why Voting?

- Low dimensional pair features: 3D or 4D
- One scene pair feature → Many model pair features
  - Self symmetric regions
  - Noise
  - Background clutter
- Voting procedure to overcome the *ambiguities*
- Maximum votes → the most likely pose hypothesis

Real Scan

True Positives

False Positives



Processing Time

TABLE I

AVERAGE NUMBERS OF PAIR FEATURES IN THE SYNTHETIC SCENE DATASET AND RELATIVE PROCESS TIME.

Feature	**Number of Features**	**Relative Process Time**†
S2S [23]	23040000 ($= 4800 \times 4800$)	3.21
B2B	2616953 ($\approx 1618 \times 1618$)	**1.00**
S2B	7689280 ($\approx 4800 \times 1602$)	1.20
L2L	121058 ($\approx 348 \times 348$)	1.03

† The fastest method, *B2B*, is shown as one.

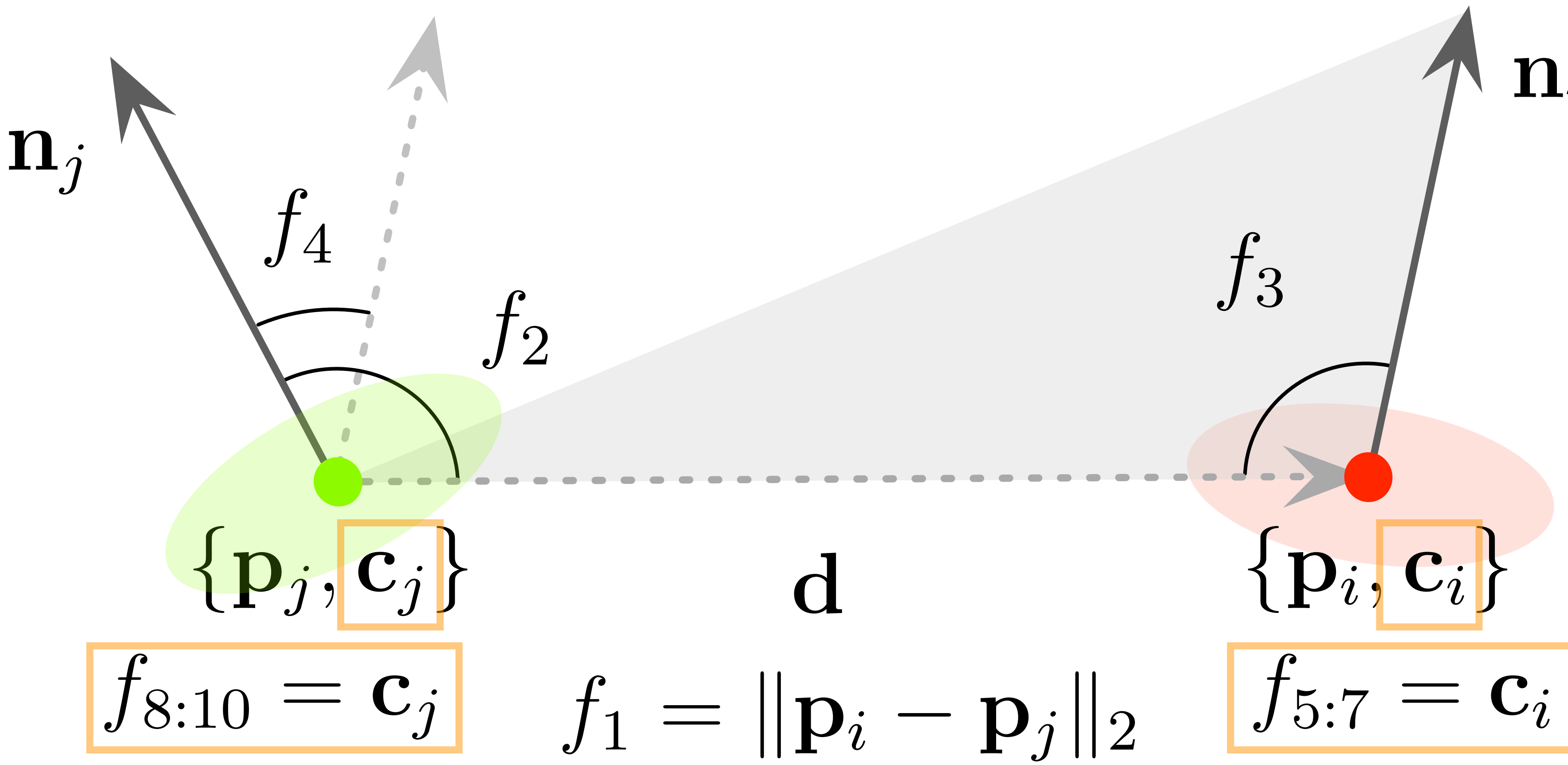
- Our pair features are sparser and faster.

Exploiting Color Info.

- Industrial parts
  - Low texture or textureless
  - Boundary information is useful

- Daily objects
  - Rich color and texture information
  - Exploit both color and depth information

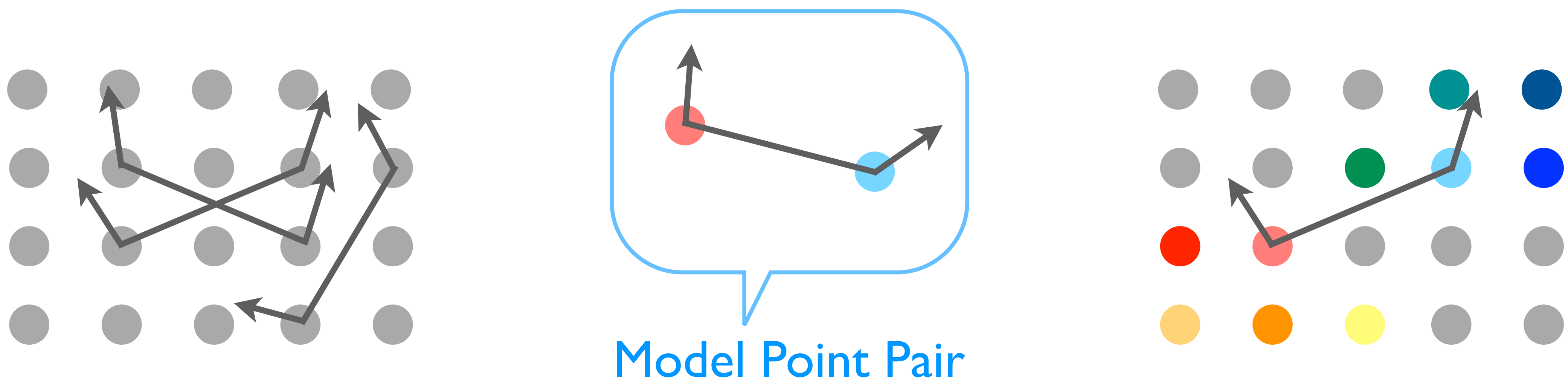
Color Point Pair Feature



$$\begin{aligned} f_1 &= \|\mathbf{p}_i - \mathbf{p}_j\|_2 \\ &= \|\mathbf{d}\|_2 \end{aligned}$$

- **PPF** (Drost et al.): 4 dimensional
- **CPPF** (proposed): 10 dimensional

Why Color Points?



To **prune** unnecessary feature matching

- **Point Pair Feature**
  - Objects having rich variations in surface normals
  - **Inefficient** for planar or self-symmetric objects
  - **False matching** from background clutter
- **Color Point Pair Feature**
  - **Prune** potentially false matches based on **color similarity**
  - HSV color space
  - More **efficient** because unnecessary votes are skipped

Parallel Implementation

- parallel NVIDIA Thrust lib
  - reduction
  - counting
  - partition
  - binary search
  - sorting
  - …

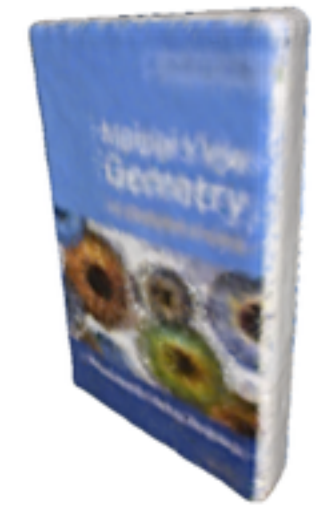
Test Objects


Clorox


Flash


Kuka Mug

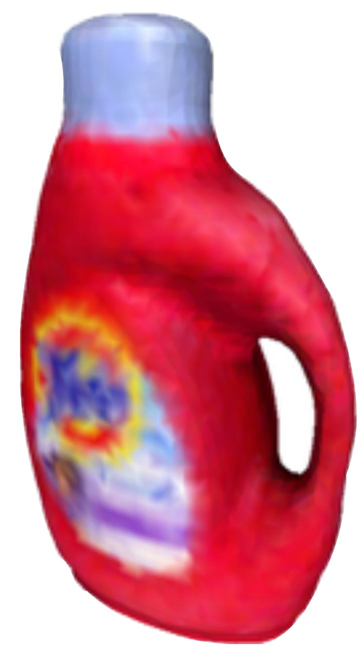

Milk


MVG Book


Orange Juice


Pringles

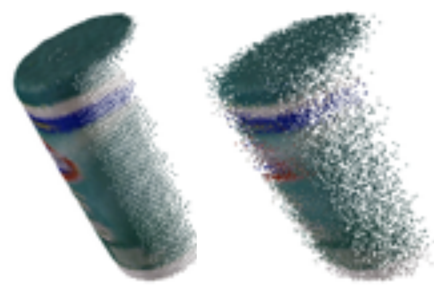

Starbucks Mug


Tide


Wrench

Performance Evaluation

Dataset



Gaussian noise

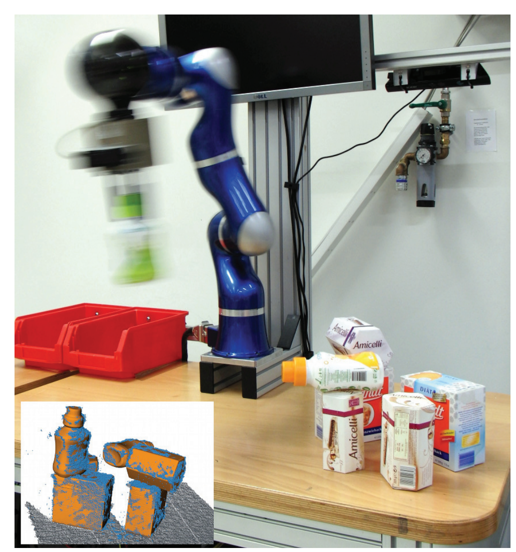
MOSI

SOMI

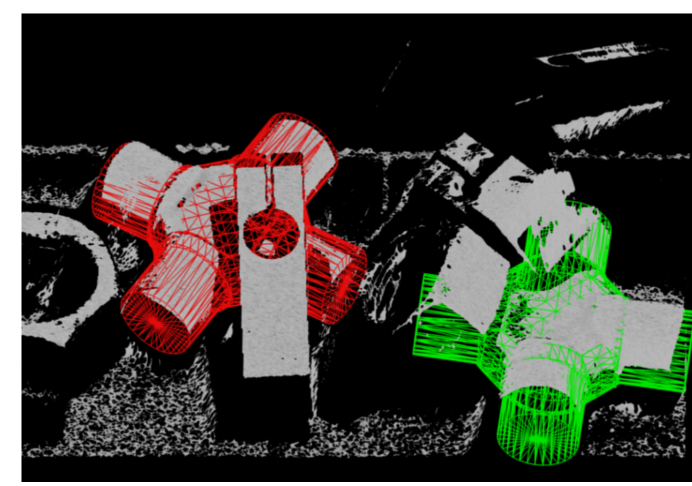
Real cluttered scenes

Compared Approaches

Hinterstoisser et al., PAMI'11
with/without ICP

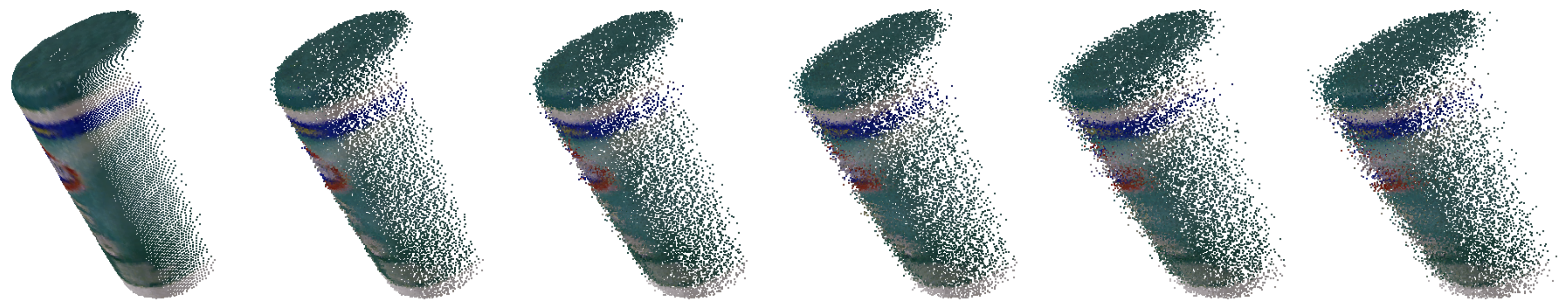


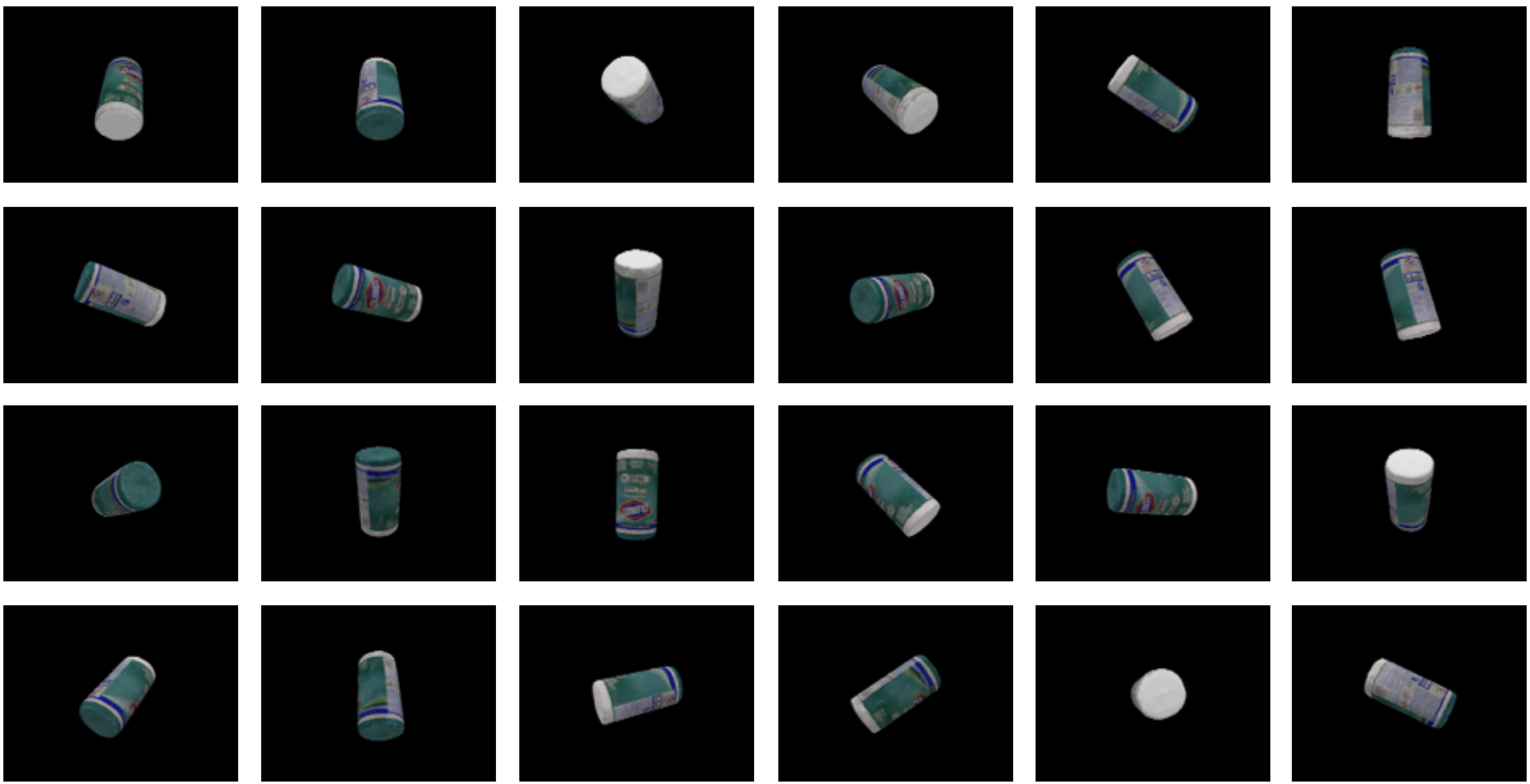
Papazov et al., IJRR'12



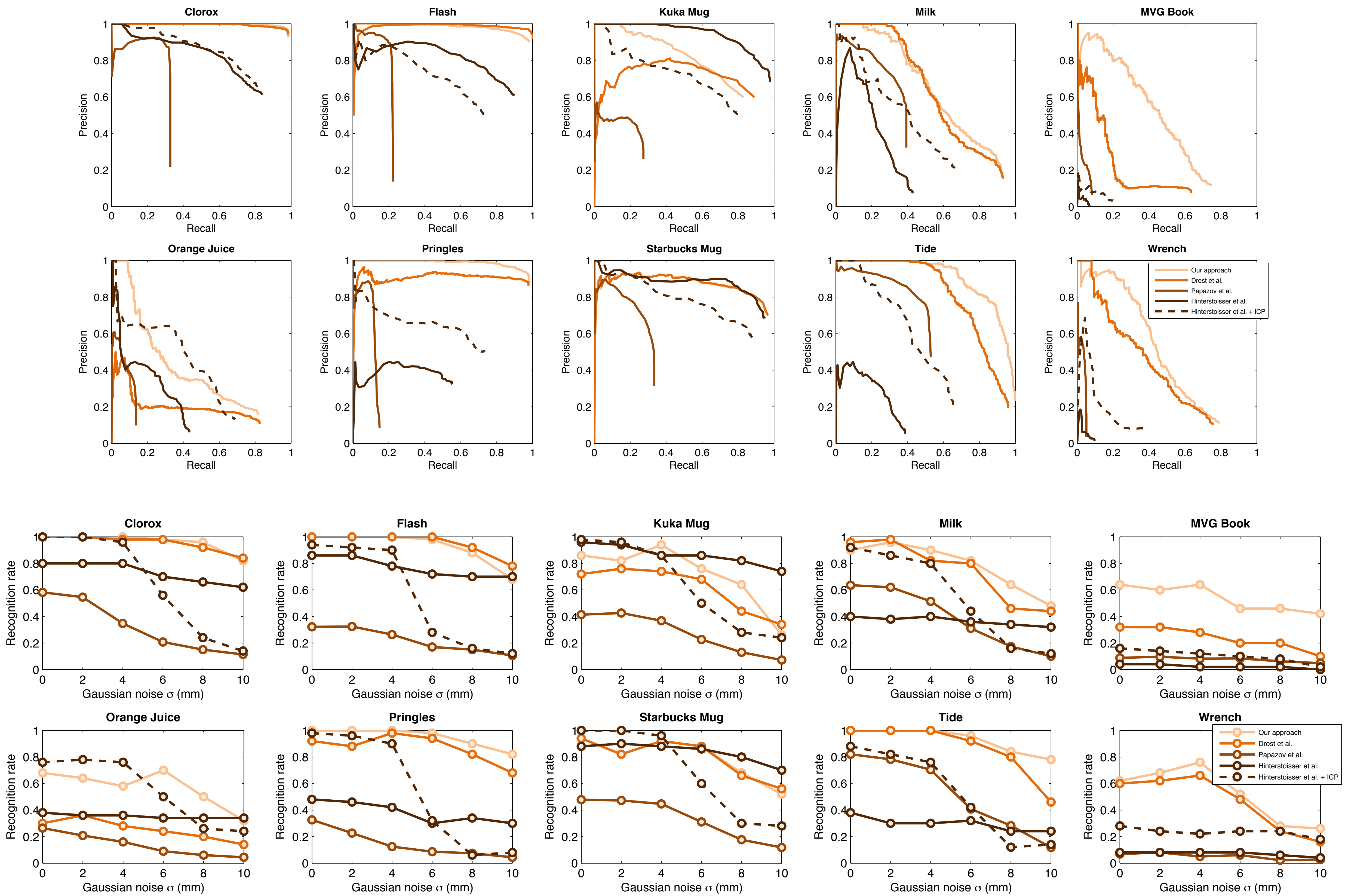
Drost et al., CVPR'10

Dataset: Gaussian noise





Results: Gaussian noise



Cluttered Scenes

Pose Estimation Results

[1]
[2]
[3]
[4]
[5]

Hinterstoisser et al. without [1] and with ICP [2], Papazov et al. [3], Drost et al. [4], and ours [5]


Tide
Tide

Approaches

- 2D Visual Information (Monocular Camera)
  - Combining Keypoint and Edge Features [ICRA'10, ICRA'11, IJRR'12]
  - Extending to Textureless Objects [IROS'12]

- 3D Visual Information (RGB-D Camera)
  - Voting-based Pose Estimation using Pair Features [ICRA'12, IROS'12]
  - **Object Pose Tracking** [IROS'13]


real-time

Motivations

- Posterior p.d.f. as a set of weighted particles
- Slow frame rate due to a serial likelihood evaluation of particles
- Inherently parallel algorithm
- each particle weight update is independent of other updates

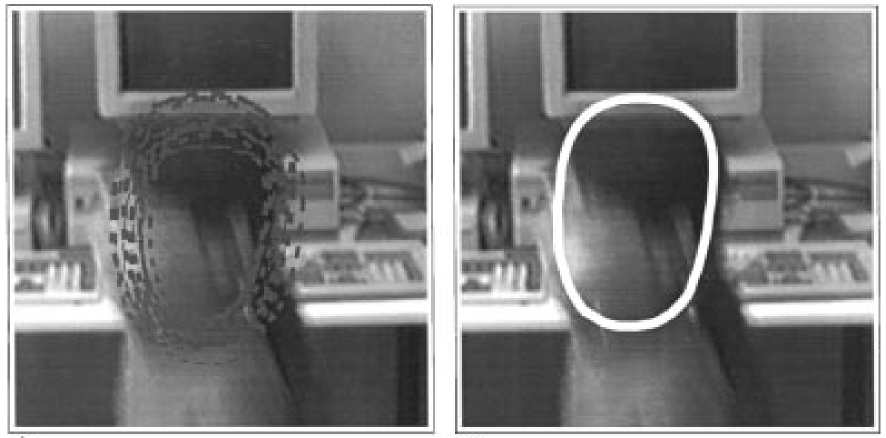
To **parallelize** the time-consuming likelihood evaluation

Contributions

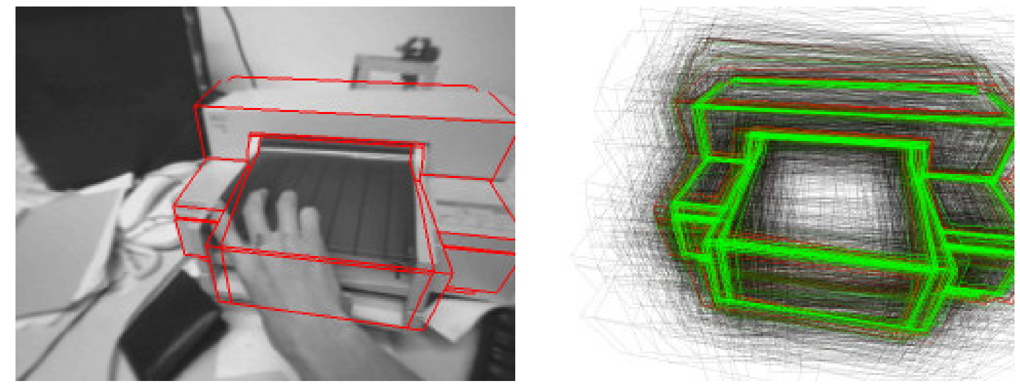
- Rich features from RGB-D channels (colors, points, normals)
- Frame Buffer Object (FBO) in OpenGL & CUDA OpenGL interoperability
- Multiple object rendering

Related Work

Edges

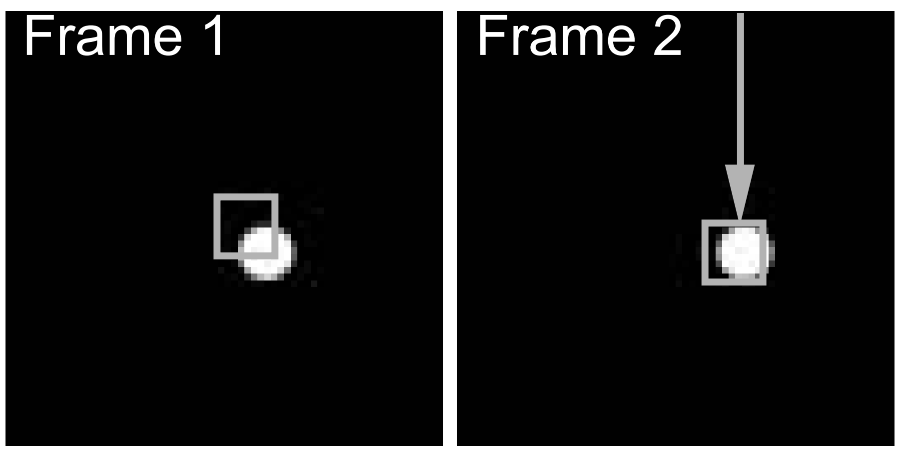


[Isard, IJCV'98]
Condensation in 2D



[Klein, BMVC'06]
Fast PF using GPU shader

Intensity

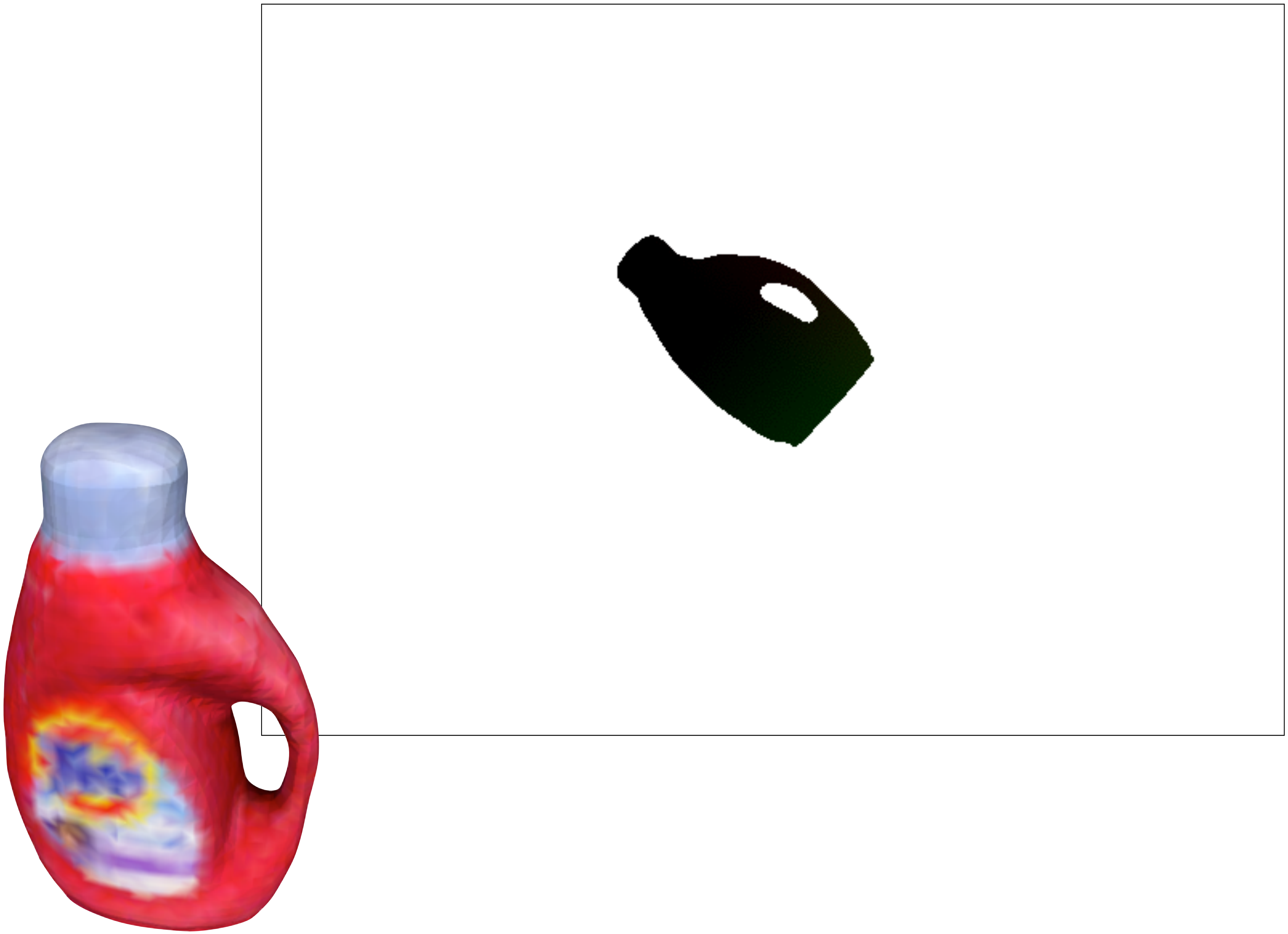


[Montemayor, SIGGRAPH'04]
Simple PF on GPU

[Azad, ICRA'11]
Fast PF using CUDA

Employ rich features: *depth*, *normals*, and *color*

Likelihood Evaluation



Likelihood Evaluation

RGB-D scene

rendered object

a point in $\mathbf{M}_t$

$$p(\mathbf{Z}_t | \mathbf{X}_t^{(n)}, \mathbf{M}_t) = \prod_{(i,j)\in\mathcal{A}} p(\mathbf{z}_t^{(i)} | \mathbf{X}_t^{(n)}, \mathbf{m}_t^{(j)})$$

n-th pose

a point in $\mathbf{Z}_t$

$$\mathcal{A} = \{(i,j) | \, \mathrm{proj}(\mathrm{x}(\mathbf{z}_t^{(i)})) = \mathrm{proj}(\mathbf{X}_t^{(n)} \cdot \mathrm{x}(\mathbf{m}_t^{(j)}))\}$$

Distance functions

$$d_e(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} \|\mathbf{x}_1 - \mathbf{x}_2\| & \text{if } \|\mathbf{x}_1 - \mathbf{x}_2\| \leq \tau \\ 1 & \text{otherwise} \end{cases}$$

Euclidean distance

$$d_n(\mathbf{n}_1, \mathbf{n}_2) = \frac{\cos^{-1}(\mathbf{n}_1^\top \mathbf{n}_2 - 1)}{\pi}$$

Normal distance

$$d_c(\mathbf{c}_1, \mathbf{c}_2) = \|\mathbf{c}_1 - \mathbf{c}_2\|$$

Color distance

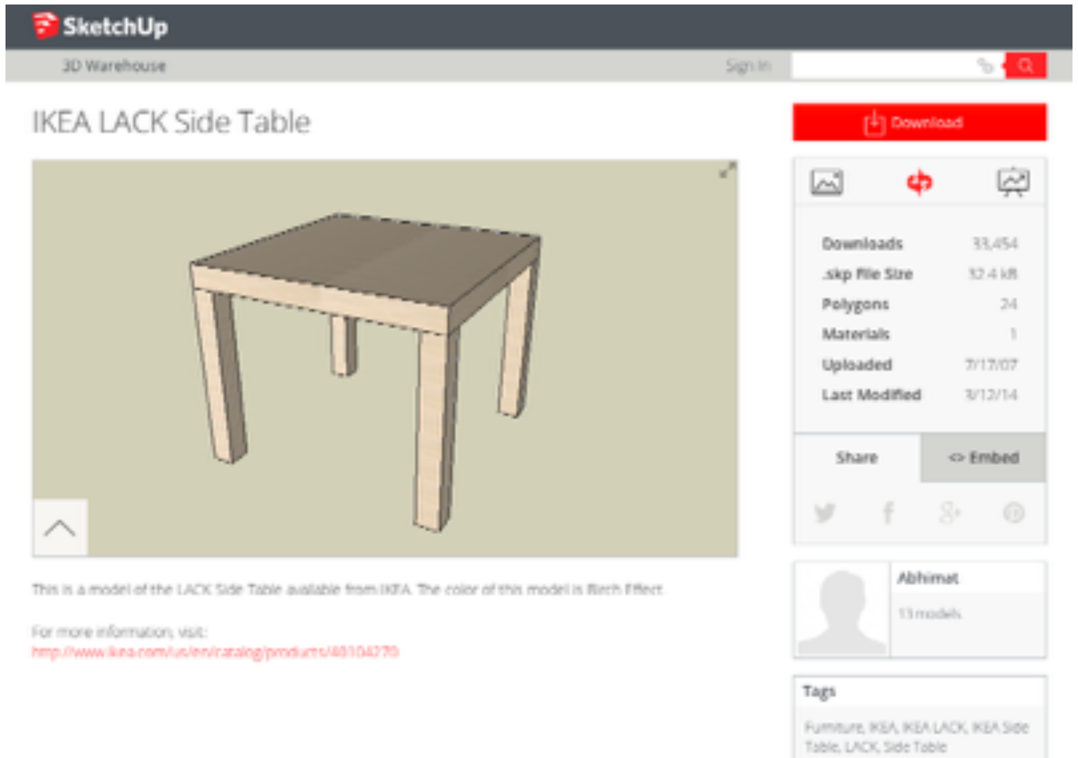
Likelihood Evaluation

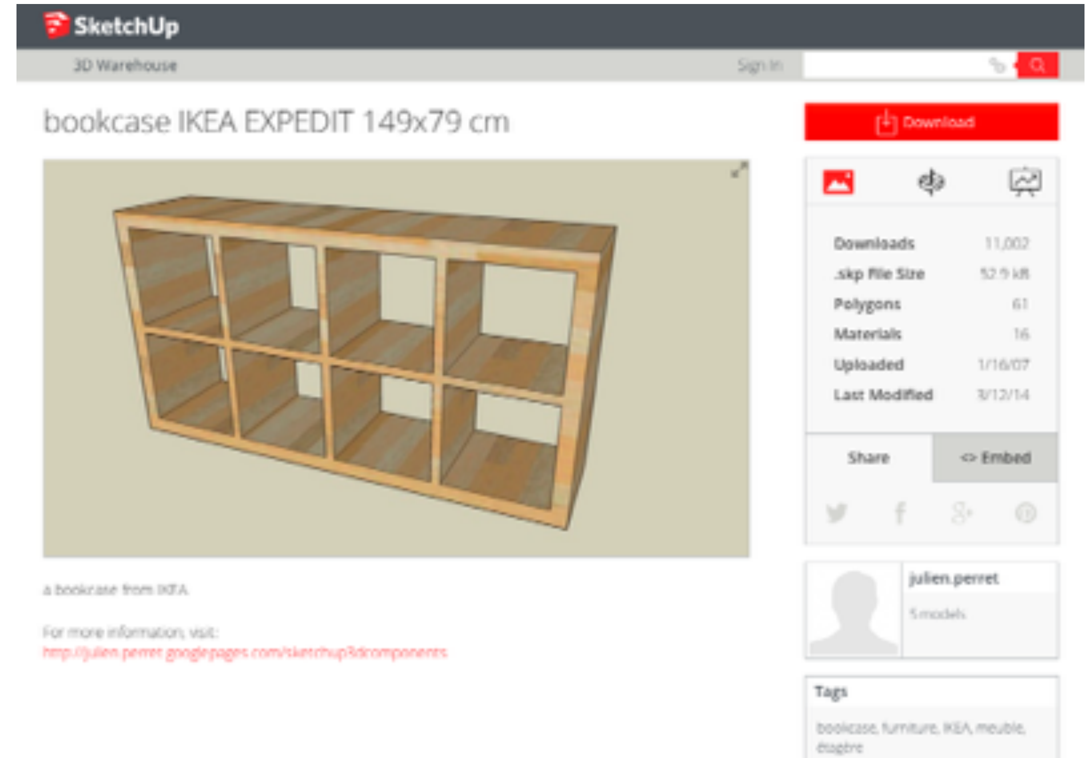
$$p(\mathbf{Z}_t|\mathbf{X}_t^{(n)}, \mathbf{M}_t) = \prod_{(i,j)\in\mathcal{A}} p(\mathbf{z}_t^{(i)}|\mathbf{X}_t^{(n)}, \mathbf{m}_t^{(j)})$$

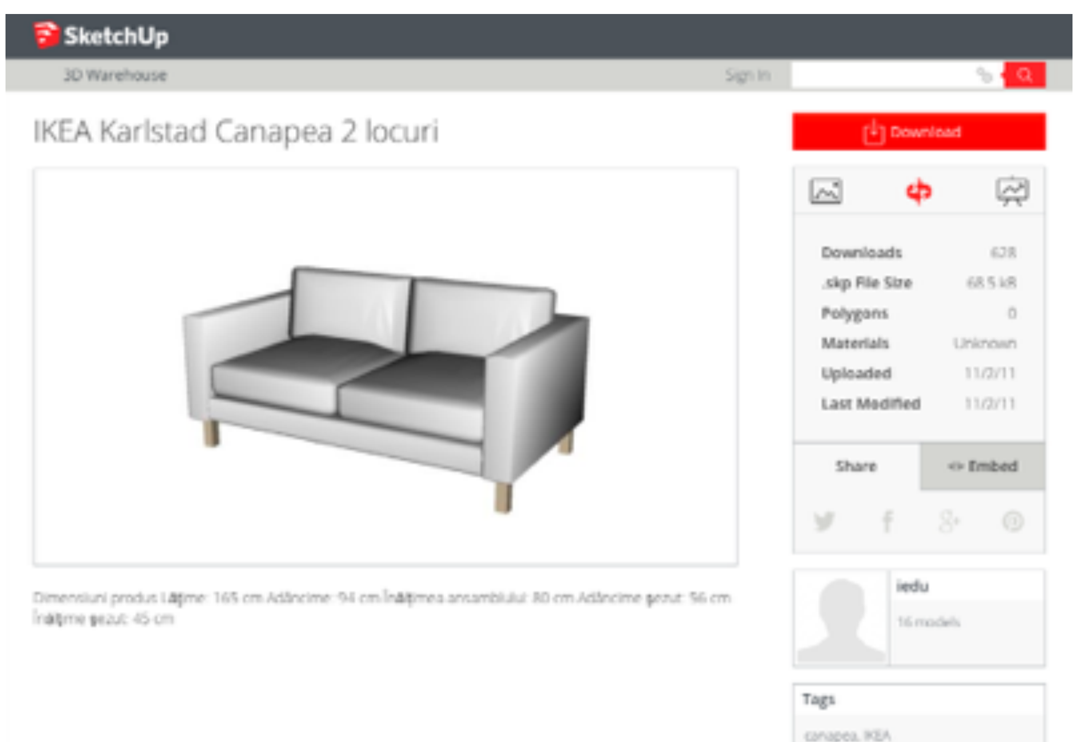
$$\begin{aligned} p(\mathbf{z}_t^{(i)}|\mathbf{X}_t^{(n)}, \mathbf{m}_t^{(j)}) = &\exp^{-\lambda_e \cdot d_e(\mathrm{x}(\mathbf{z}_t^{(i)}),\ \mathbf{X}_t^{(n)} \cdot \mathrm{x}(\mathbf{m}_t^{(j)}))} \\ &\cdot \exp^{-\lambda_n \cdot d_n(\mathrm{n}(\mathbf{z}_t^{(i)}),\ \mathbf{X}_t^{(n)} \cdot \mathrm{n}(\mathbf{m}_t^{(j)}))} \\ &\cdot \exp^{-\lambda_c \cdot d_c(\mathrm{c}(\mathbf{z}_t^{(i)}),\ \mathrm{c}(\mathbf{m}_t^{(j)}))} \end{aligned}$$

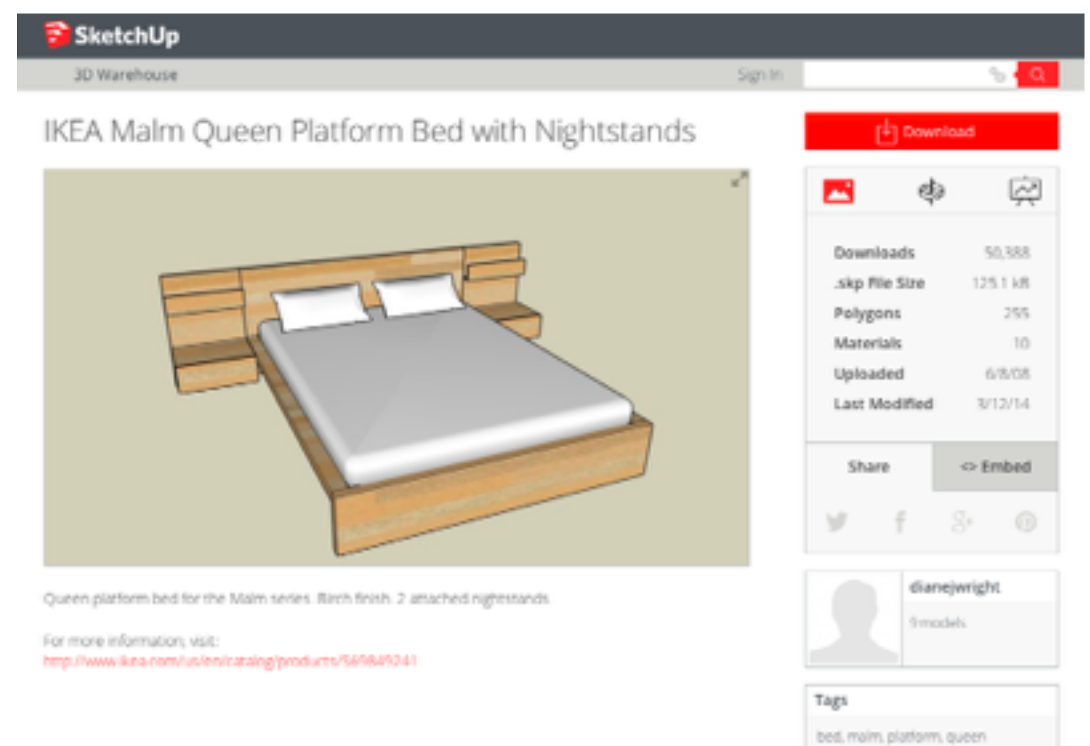

Ours
PCL tracking
Ours
PCL tracking

3D models on the Web

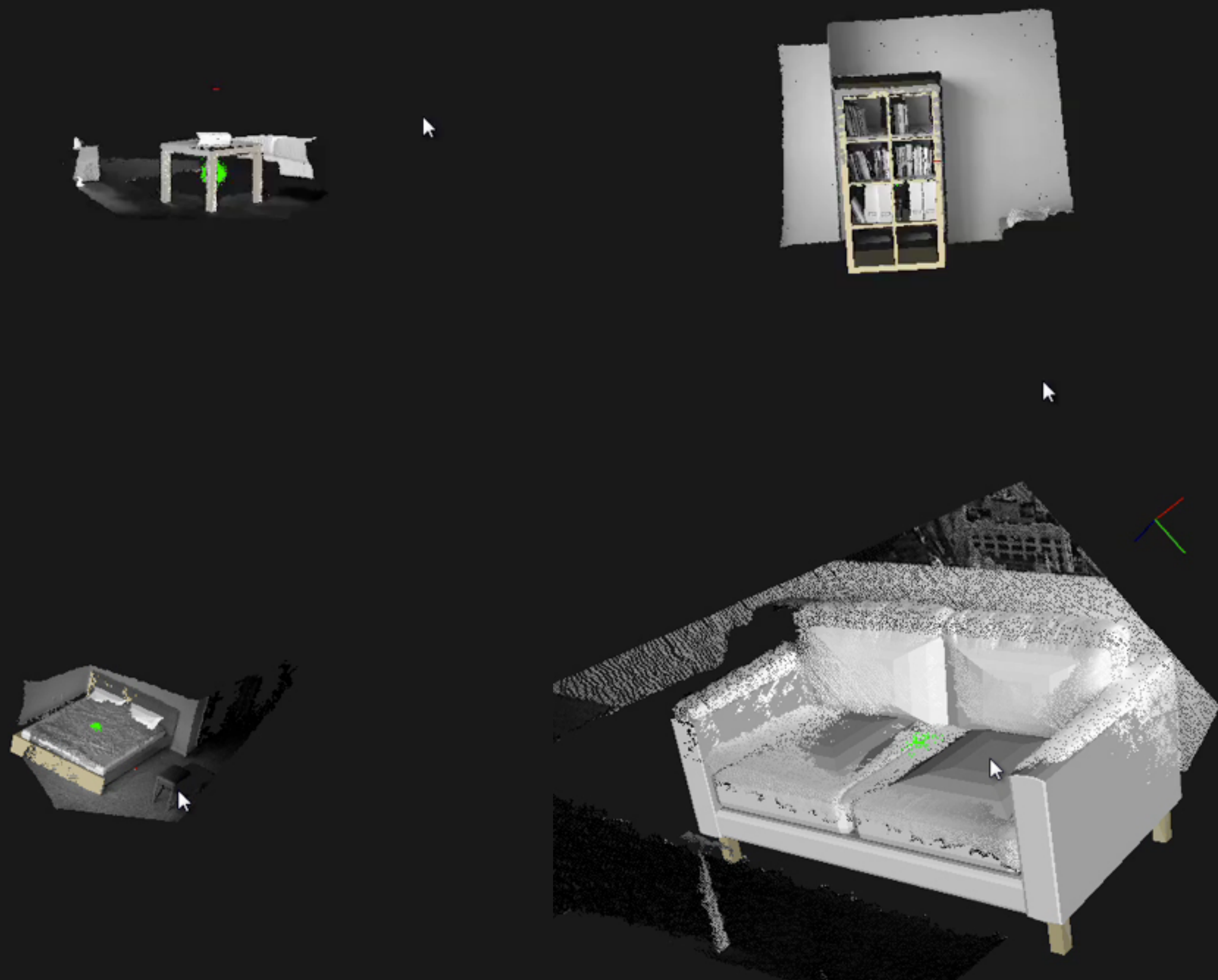


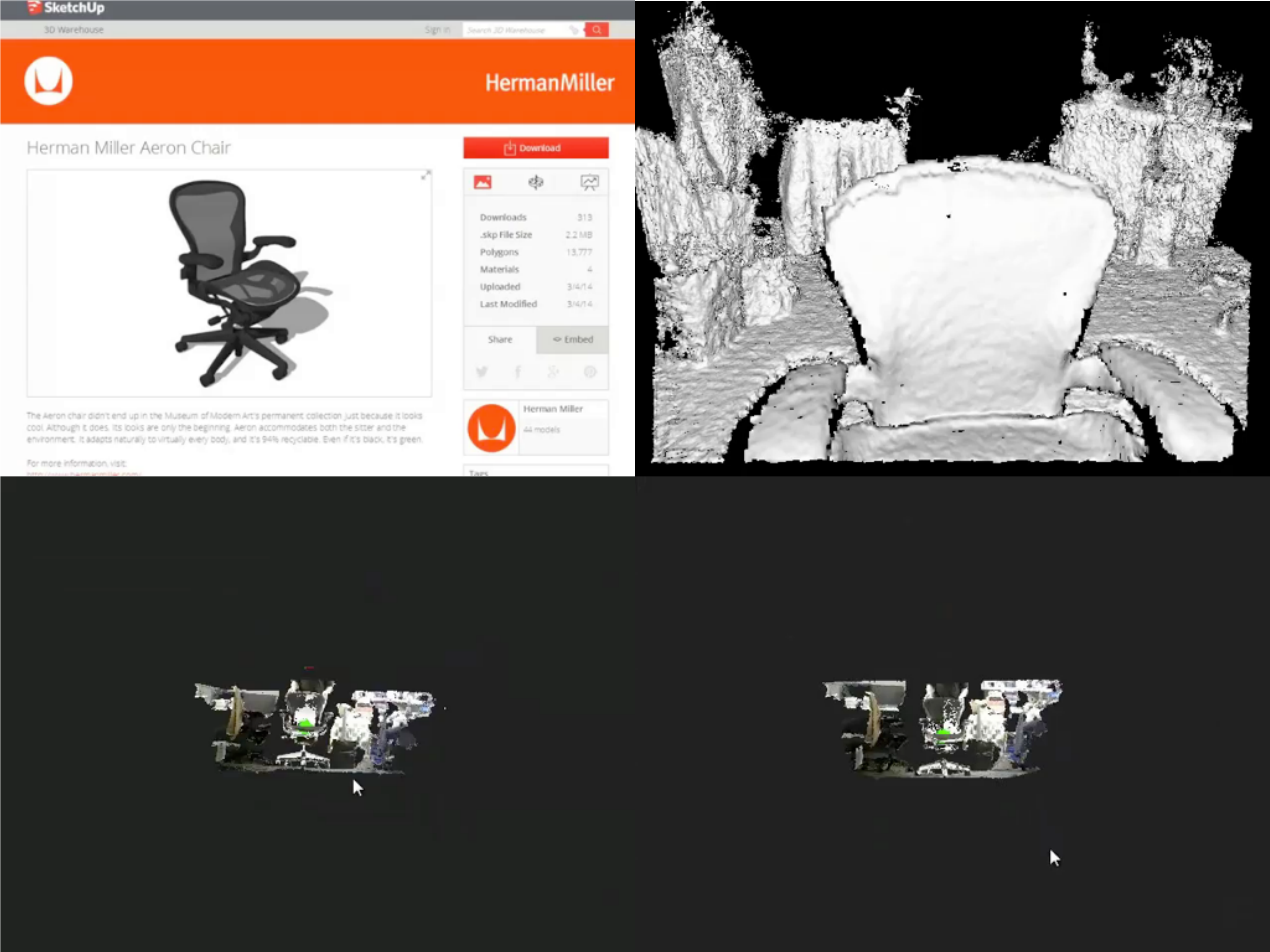




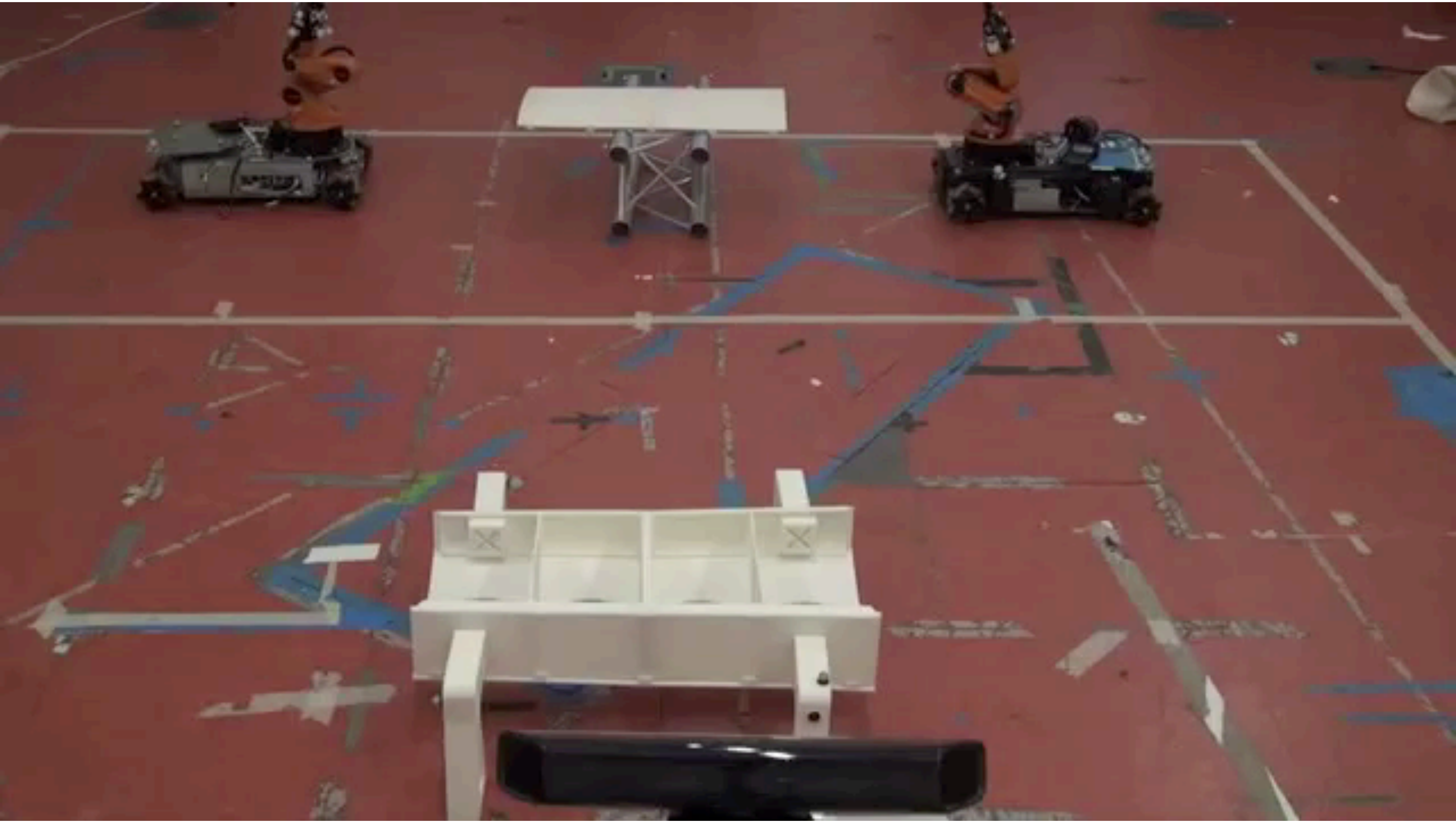


real-time




SketchUp
3D Warehouse
Sign in
HermanMiller
Herman Miller Aeron Chair
Download
Downloads 313
.skp File Size 2.2 MB
Polygons 13,777
Materials 4
Uploaded 3/4/14
Last Modified 3/4/14
Share
Embed
Herman Miller
44 models
The Aeron chair didn't end up in the Museum of Modern Art's permanent collection just because it looks cool. Although it does. Its looks are only the beginning. Aeron accommodates both the sitter and the environment. It adapts naturally to virtually every body, and it's 94% recyclable. Even if it's black, it's green.
For more information, visit:

Robotic Assembly



Conclusions

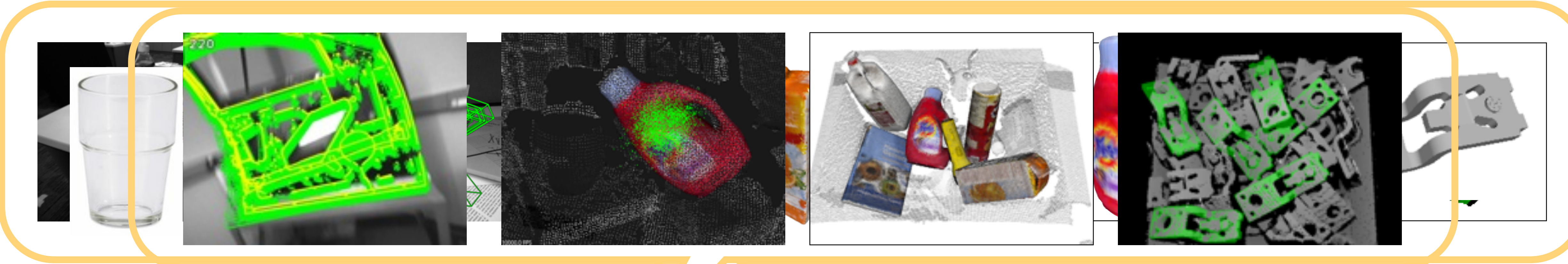
- Contributed toward robust object perception in unstructured environments

Four challenges

- object perception regardless of the degree of texture
- highly cluttered backgrounds
- object discontinuities
- real-time constraints

- combined photometric and geometric features
- multiple pose hypotheses frameworks
- combined pose estimation and tracking
- parallelized on GPU

Revisit Thesis Statement



- To close the loop between the geometric era of early computer vision and the currently dominating appearance age, both photometric and geometric features need to be considered.
- The combination of these features enables object perception algorithms not only to be more effective but also to handle an increased spectrum of objects.
- Two theoretical frameworks using multiple pose hypotheses based on combined features are contributed in this thesis.
- These new frameworks are robust to significant clutter and occlusions, and are therefore efficacious solutions for visual object perception in unstructured environments.

Future Work

- Object model adaptation
- Object modeling
- Multi-object tracking
- Scalable object perception
- Object categorization

Thank You


Henrik Christensen
Gatech


Dieter Fox
UW


Irfan Essa
Gatech


Jim Rehg
Gatech


Anthony Yezzi
Gatech


Alex Trevor
Gatech


Yuichi Taguchi
MERL


Ming-Yu Liu
MERL


Srikumar Ramalingam
MERL


Oncel Tuzel
MERL


Ross Knepper
MIT


Mehmet Dogar
MIT

GM

BOEING

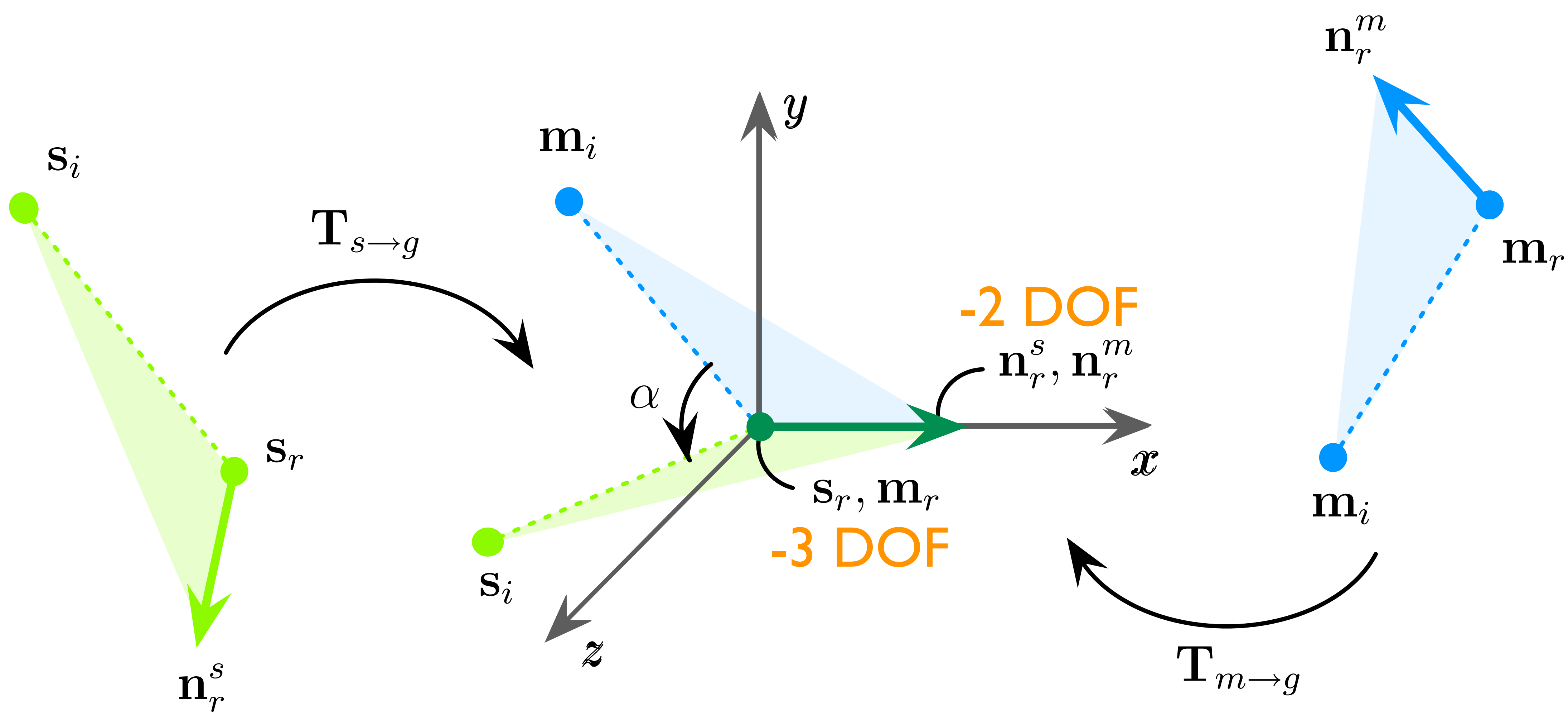
MITSUBISHI ELECTRIC
RESEARCH LABORATORIES

pcl gsoc '12

KOREA FOUNDATION
FOR ADVANCED STUDIES KFAS

Backup Slides

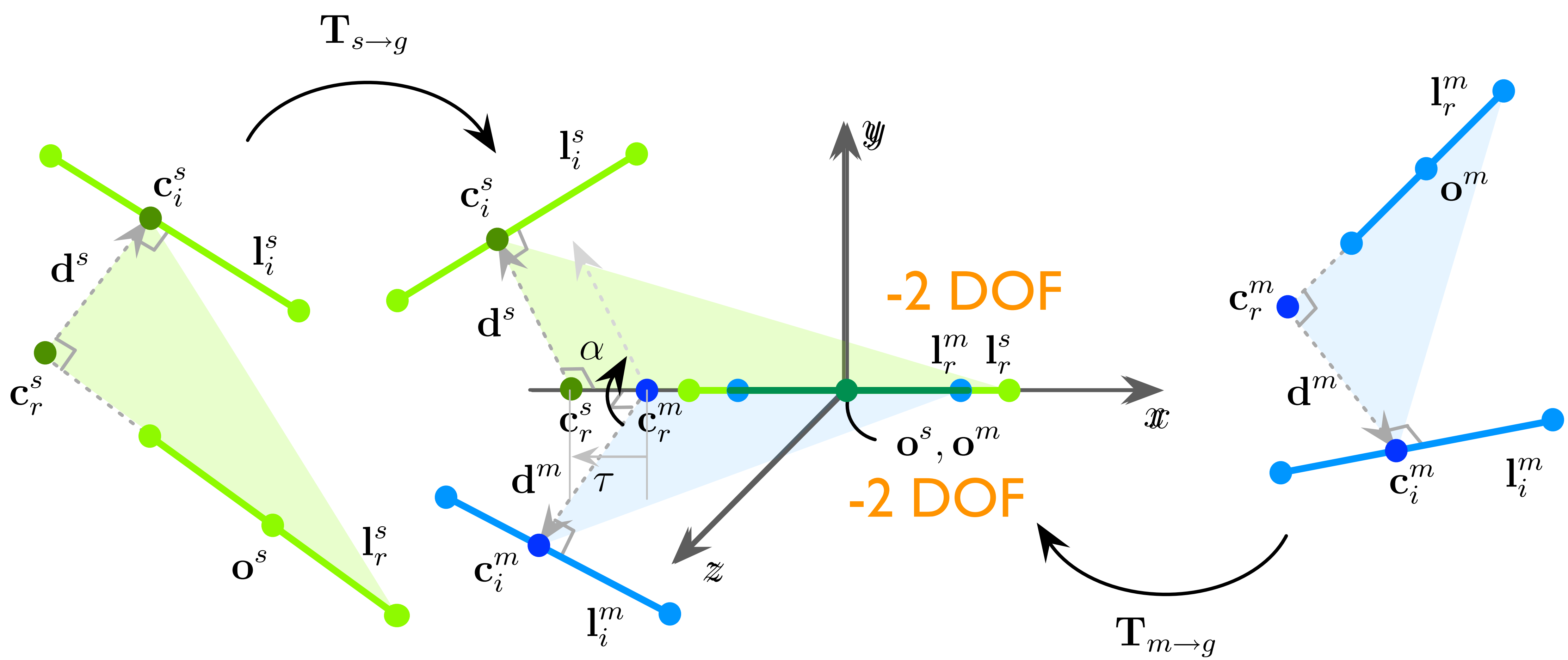
Voting Scheme I



$$\mathbf{s}_i = \boxed{\mathbf{T}_{s\to g}^{-1}\mathbf{R}_{\mathbf{x}}(\alpha)\mathbf{T}_{m\to g}}\mathbf{m}_i$$

- 2D accumulator space: $(\mathbf{m}_r, \alpha)$
- *S2S*, *B2B*, and *S2B* share the same transform

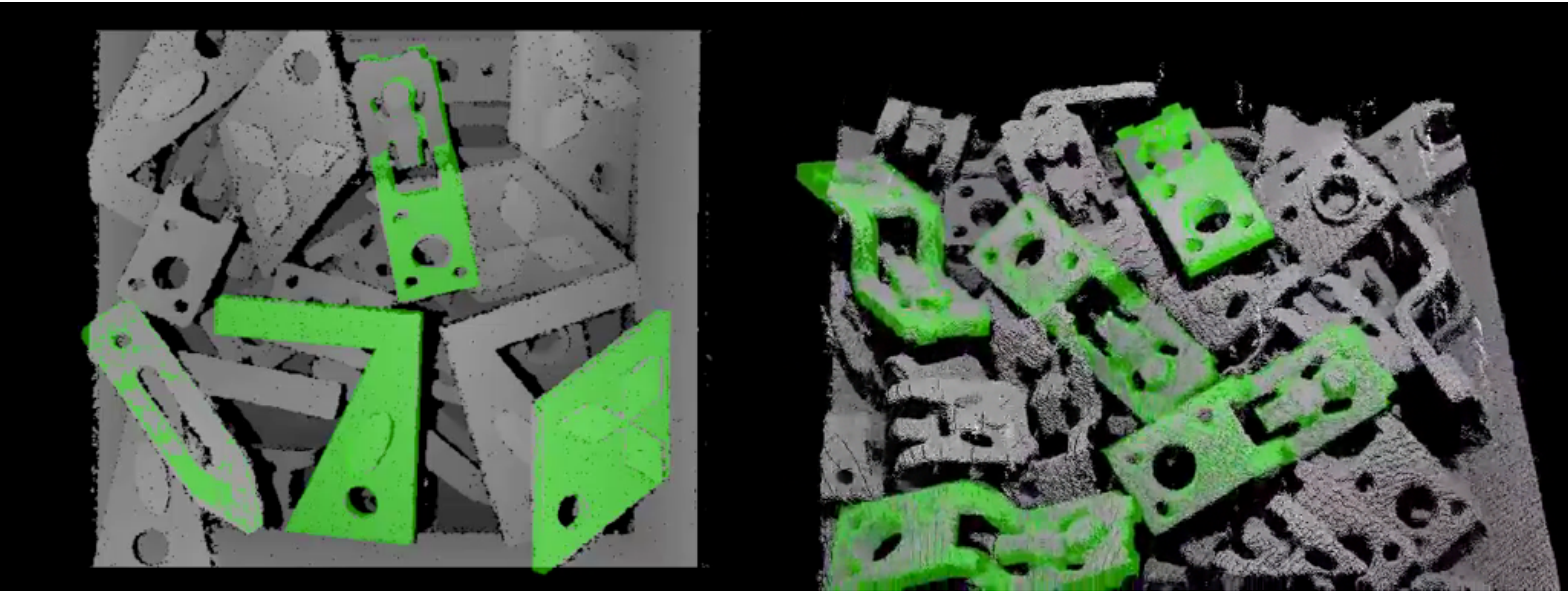
Voting Scheme II



$$\mathbf{l}_i^s = \boxed{\mathbf{T}_{s\to g}^{-1}\mathbf{T}_{\mathbf{x}}(\tau)\mathbf{R}_{\mathbf{x}}(\alpha)\mathbf{T}_{m\to g}}\mathbf{l}_i^m$$

- 3D accumulator space: $(\mathbf{o}^m, \alpha, \tau)$

Real Scan: *S2B* feature

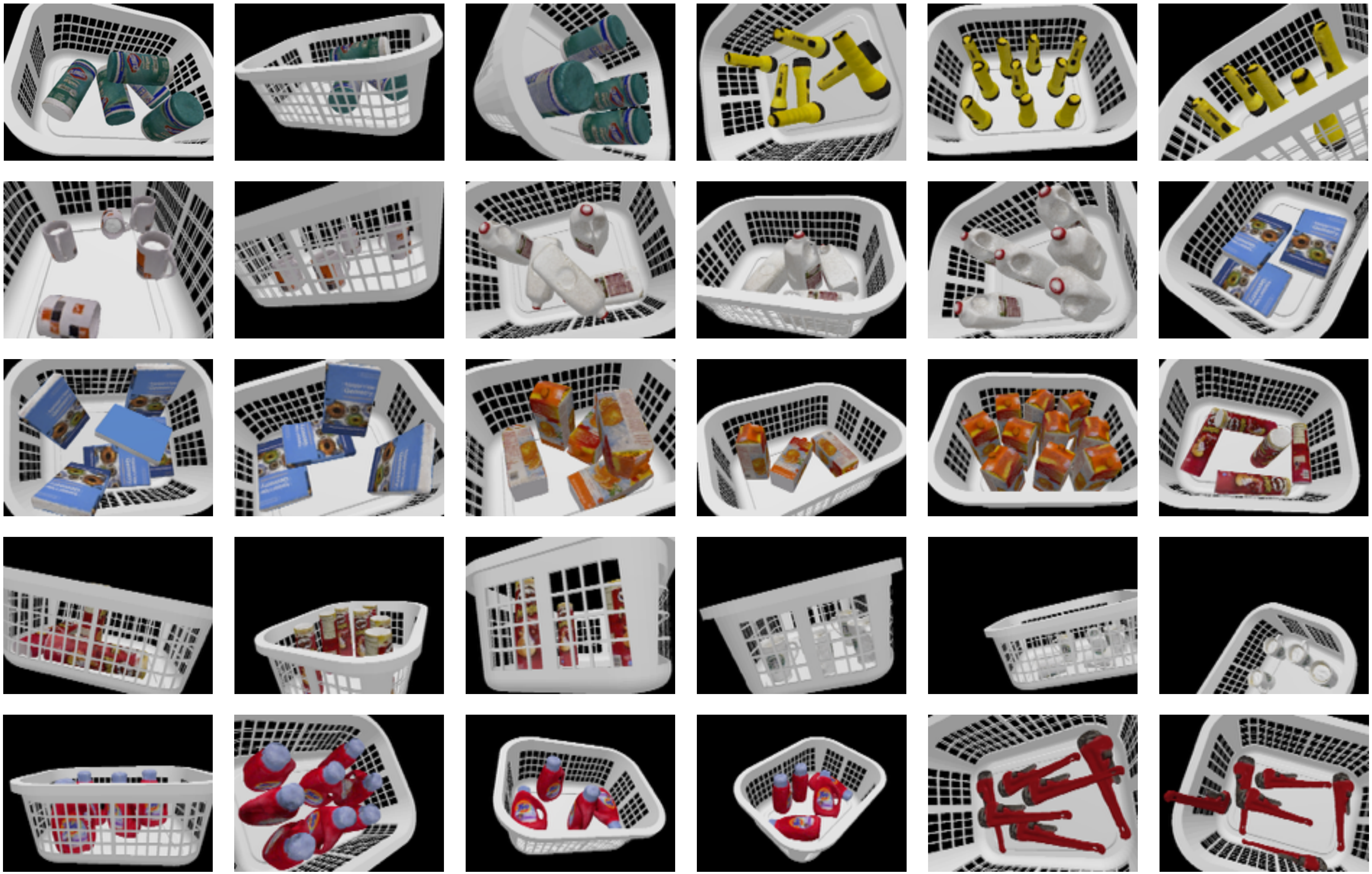


Dataset: MOSI

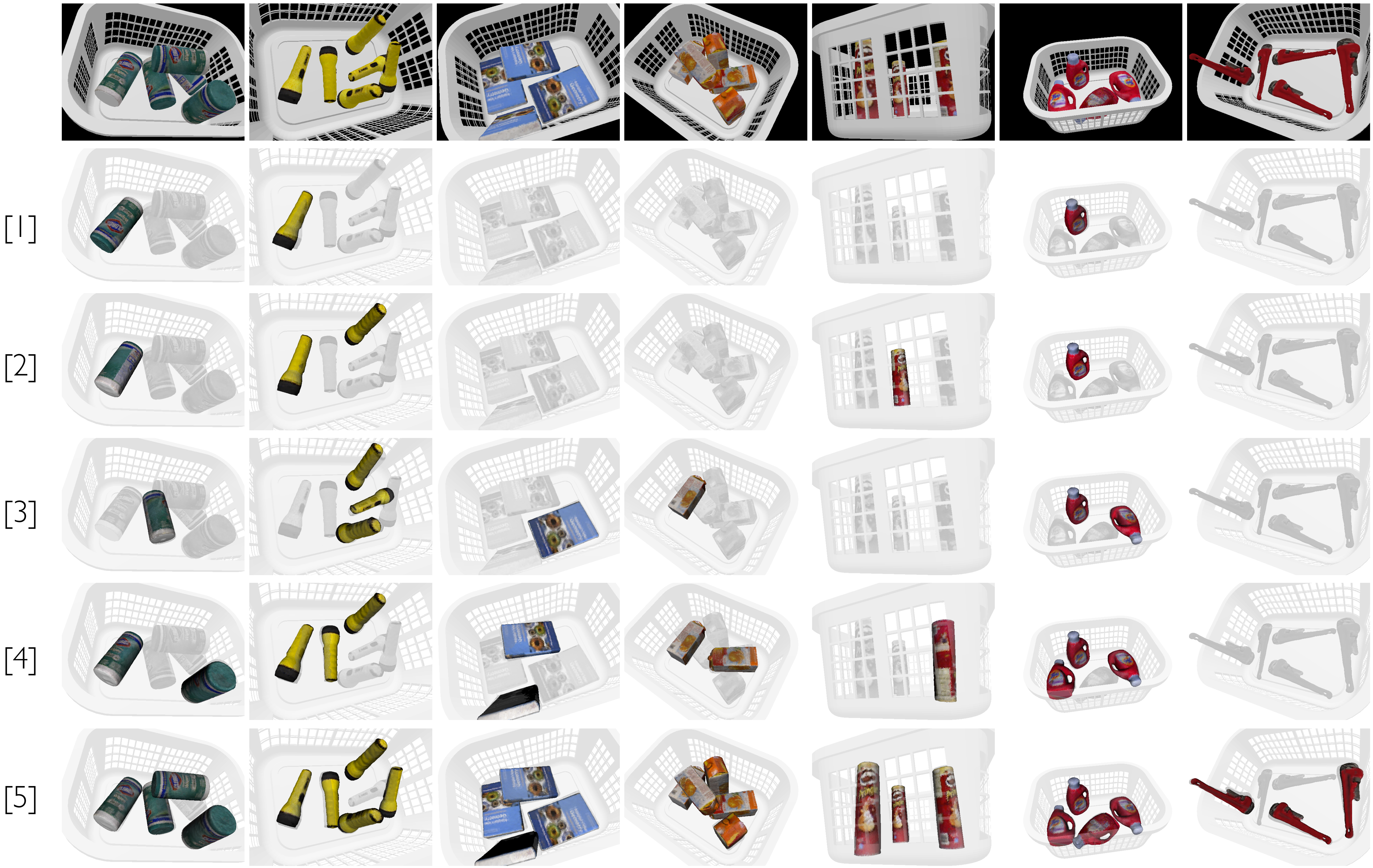
Results: MOSI

Hinterstoisser et al. without [1] and with ICP [2], Papazov et al. [3], Drost et al. [4], and ours [5]

Dataset: SOMI



Results: SOMI



Hinterstoisser et al. without [1] and with ICP [2], Papazov et al. [3], Drost et al. [4], and ours [5]