H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions

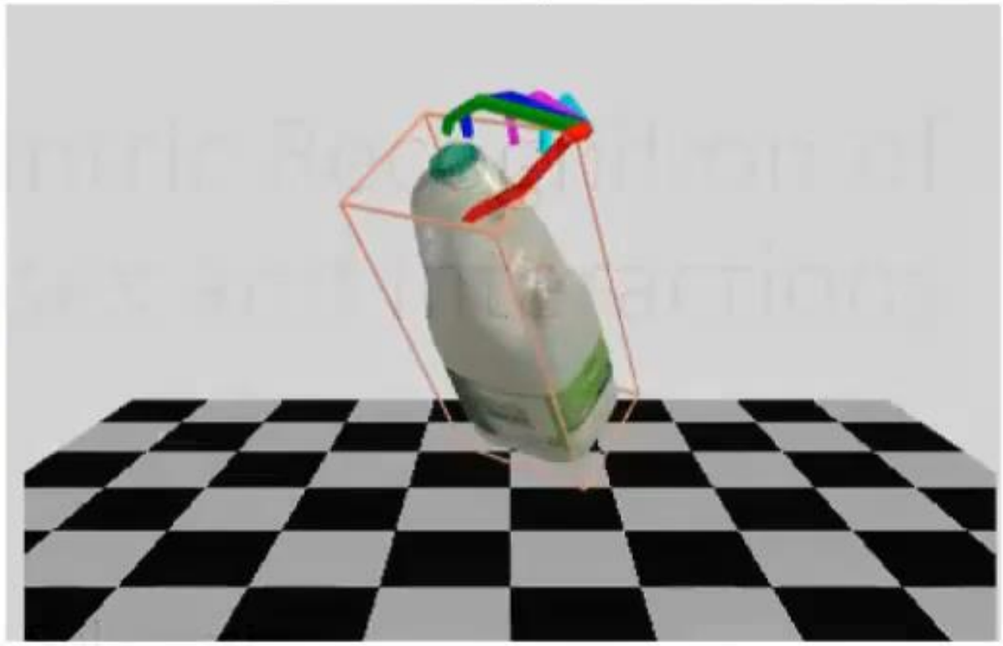
Buğra Tekin

Microsoft MR & AI Lab, Zürich

Hand & Object Interactions

2D Hand + Object Pose

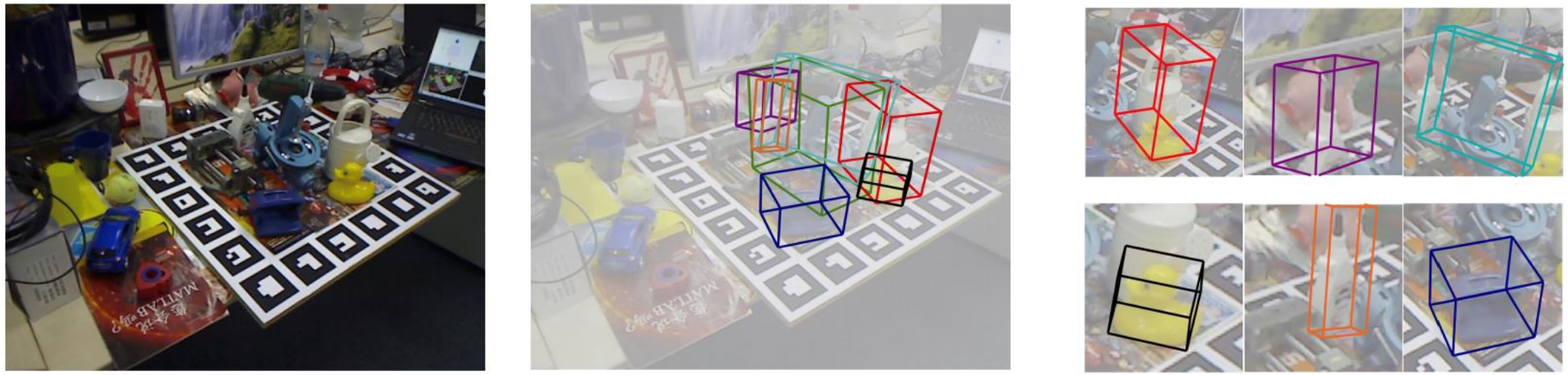
3D Hand + Object Pose



Per-frame predictions

Outline

1. Object Understanding

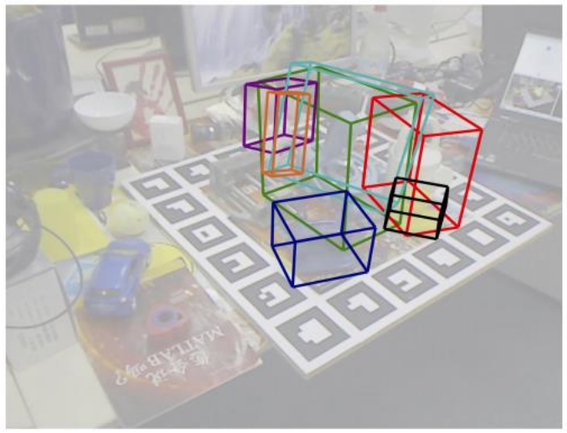


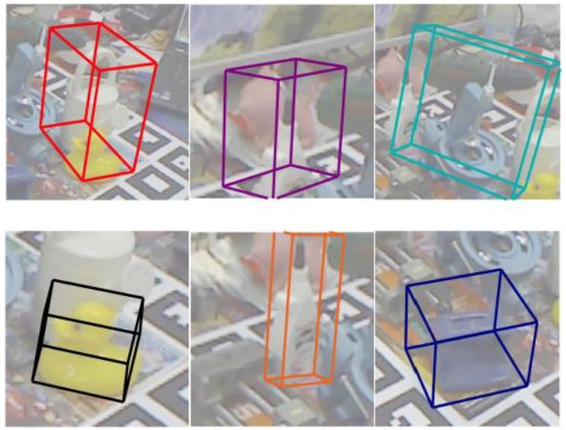
2. Hand + Object Understanding

1. B. Tekin, S. Sinha, P. Fua, “Real-time Seamless Single Shot 6D Object Pose Prediction”, **CVPR’18**
2. B. Tekin, F. Bogo, M. Pollefeys, “H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions”, **CVPR’19**

Outline

1. Object Understanding

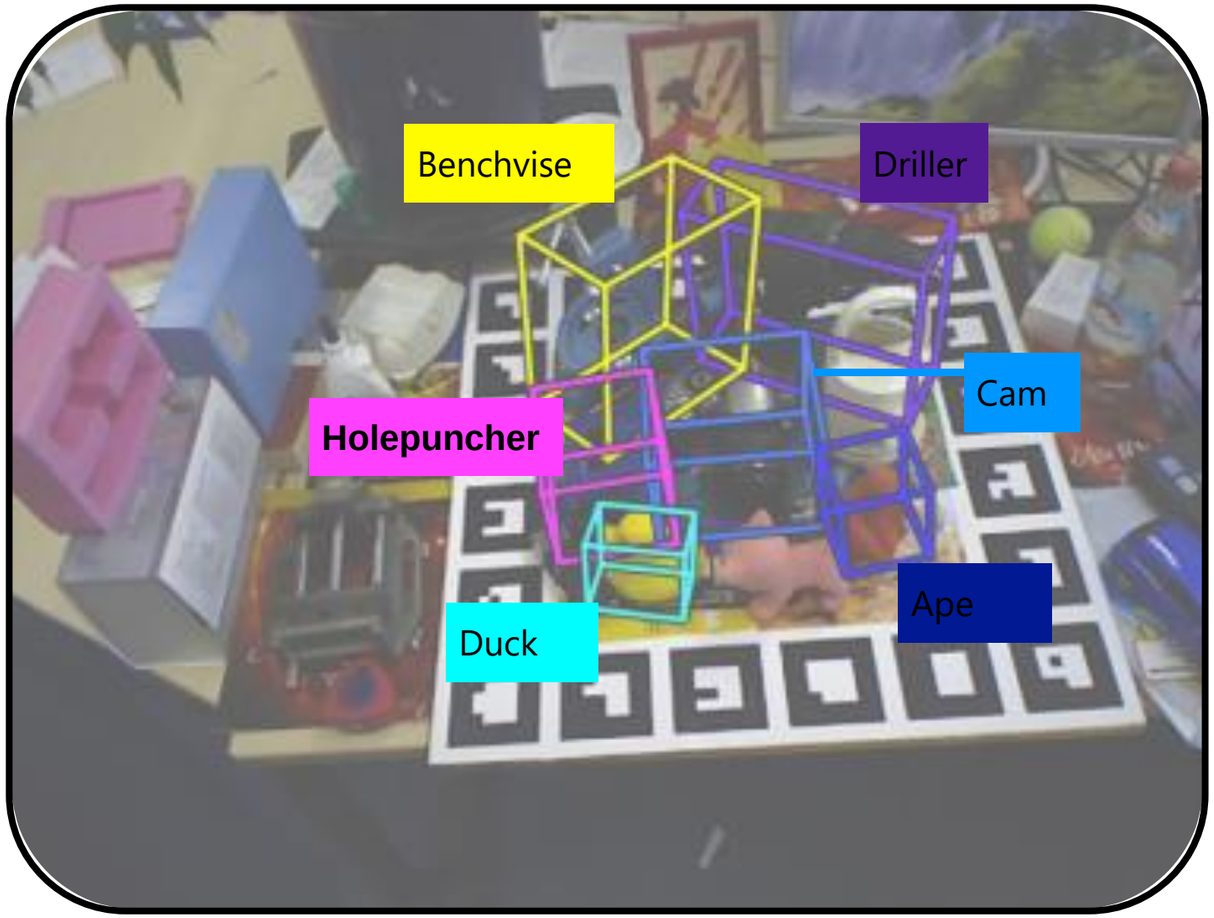




1. B. Tekin, S. Sinha, P. Fua, “Real-time Seamless Single Shot 6D Object Pose Prediction”, CVPR’18

Microsoft

Real-time Seamless Single Shot 6D Object Pose Prediction

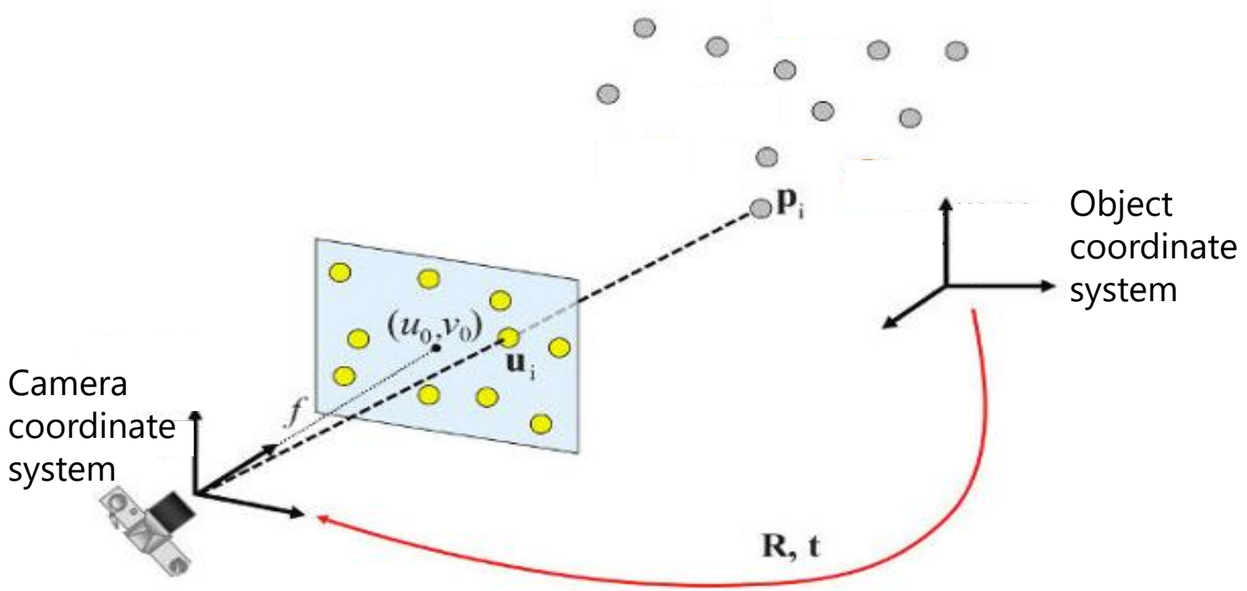


Buğra Tekin, Sudipta Sinha, Pascal Fua, “Real-time Seamless Single Shot 6D Object Pose Prediction”, CVPR’18

Microsoft

6D Object Pose Estimation

Estimating the pose (rotation and translation) of an object instance in the camera coordinate system

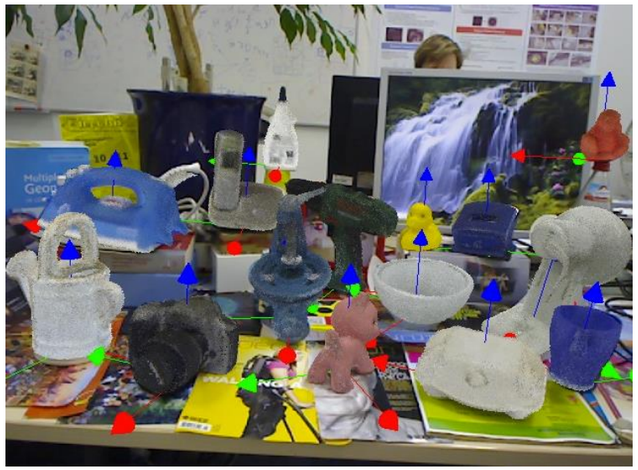


Microsoft

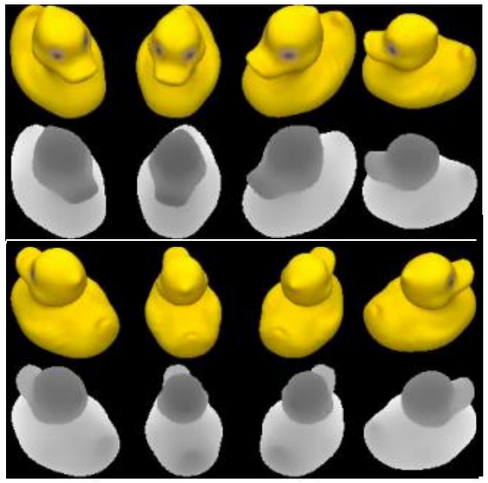
Challenges

Variation in illumination

Occlusion



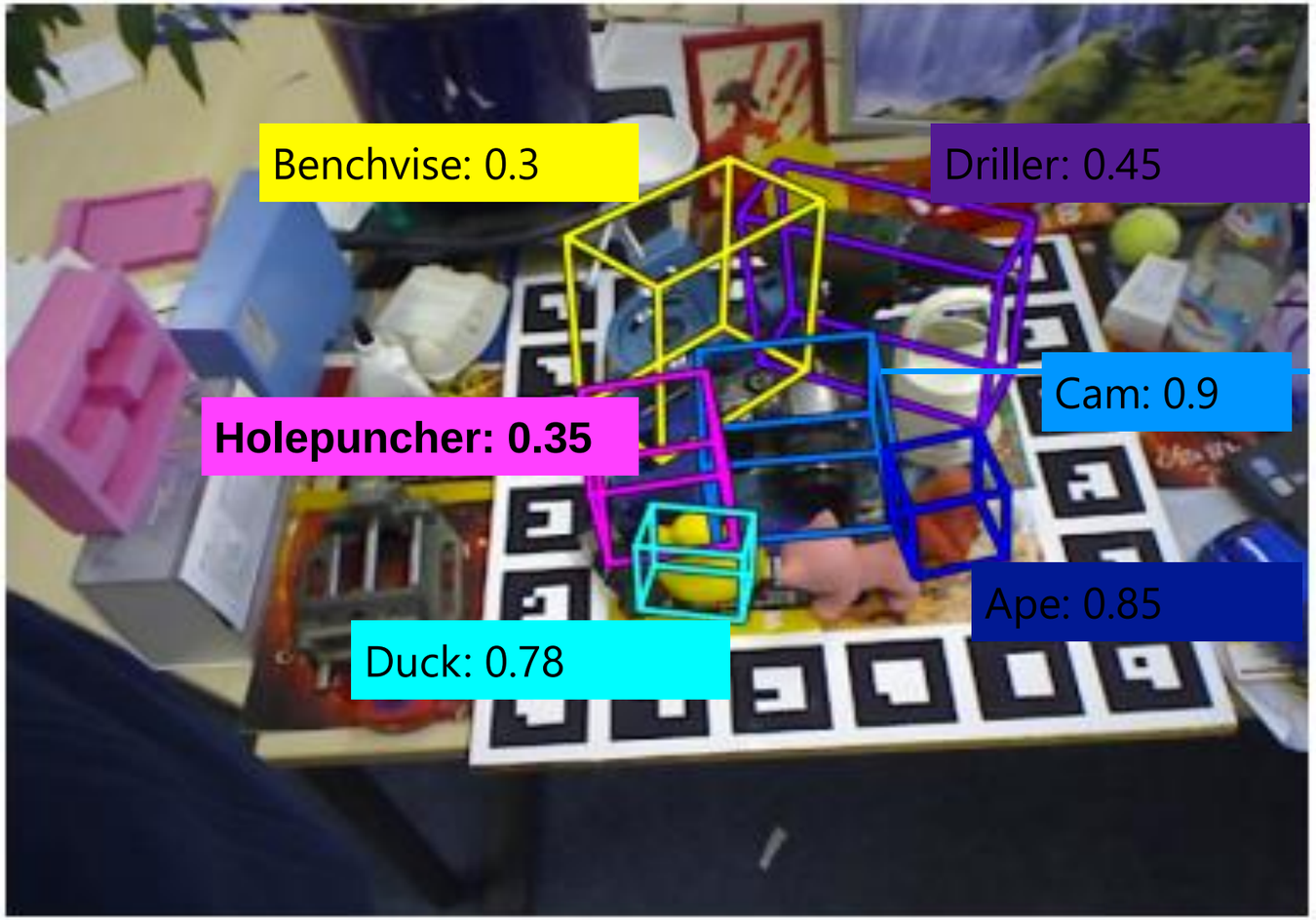
Clutter



Variation in viewpoint

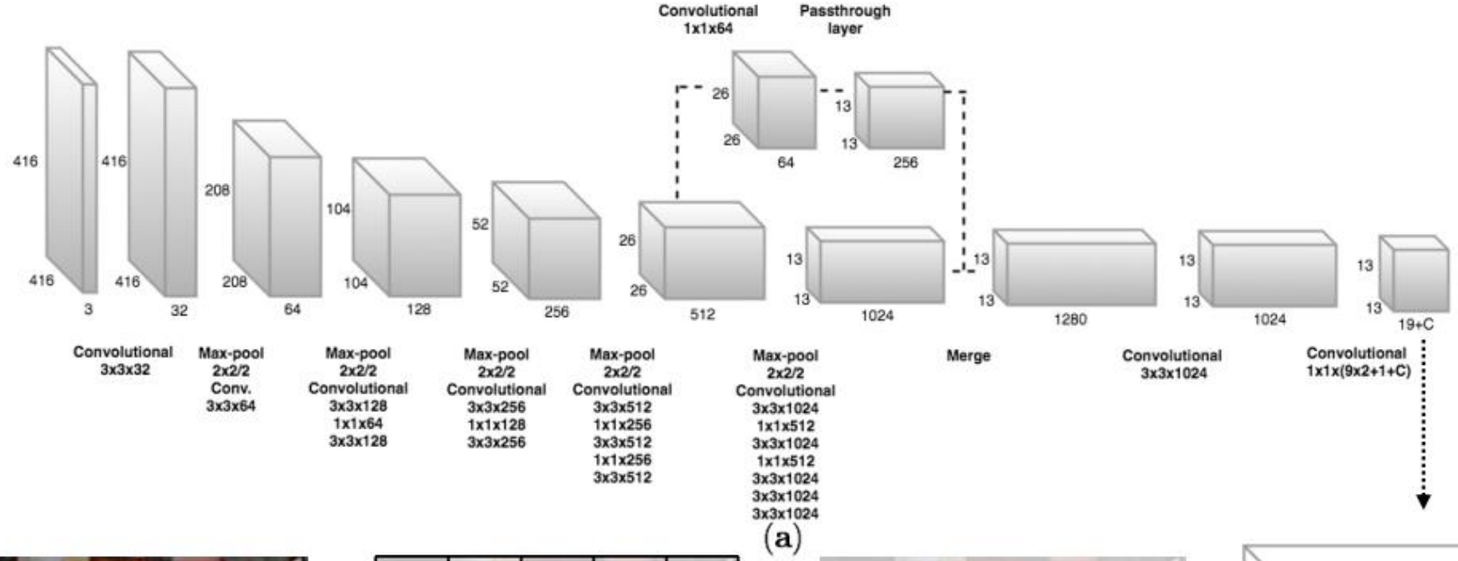
Microsoft

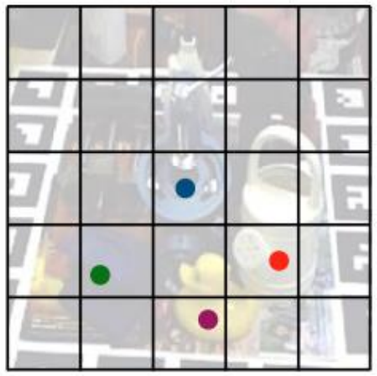
Our Goal

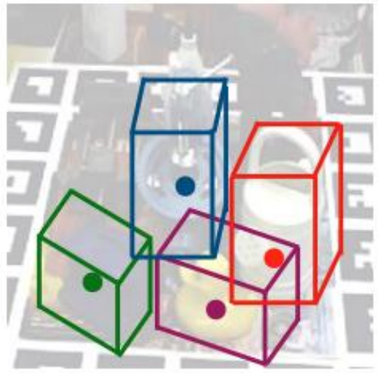


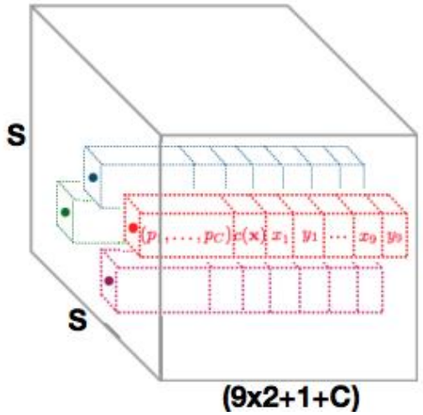
Predict the object pose, the class and the detection confidence on the full image with a CNN in a single shot

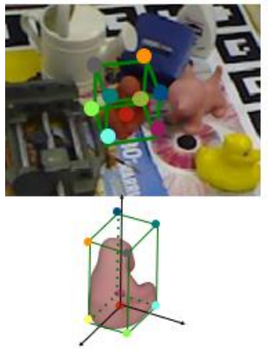
Our Approach









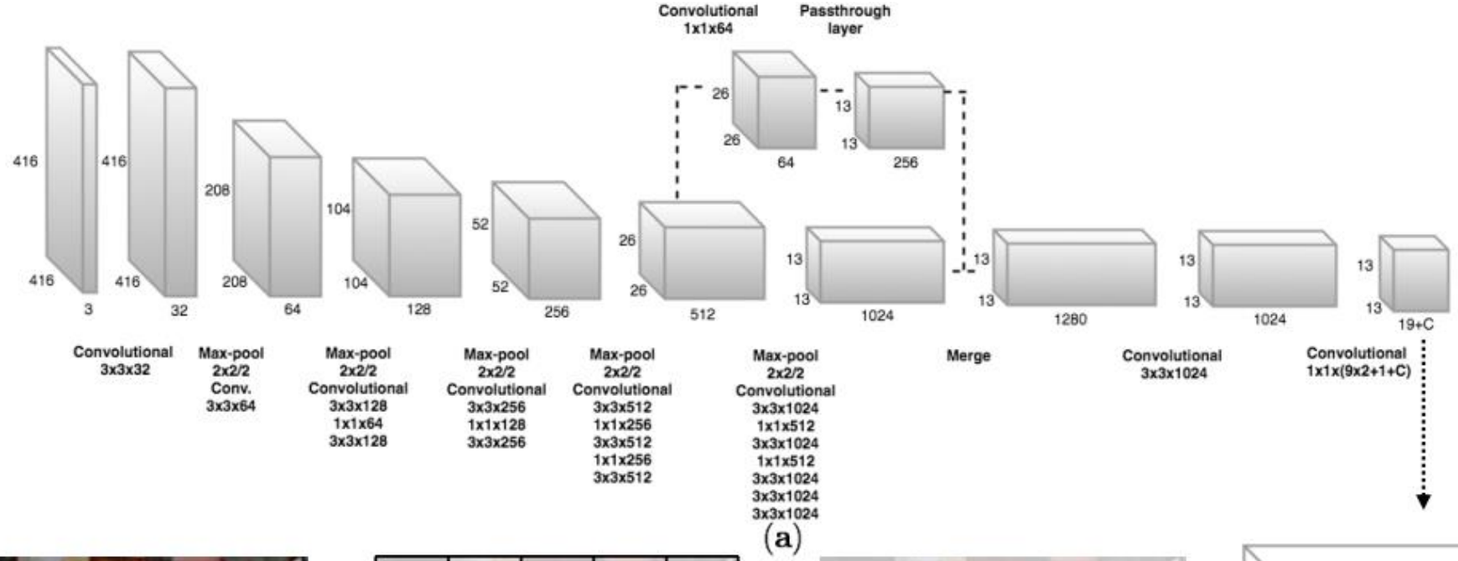


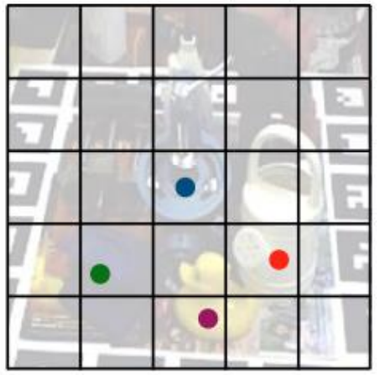
PNP

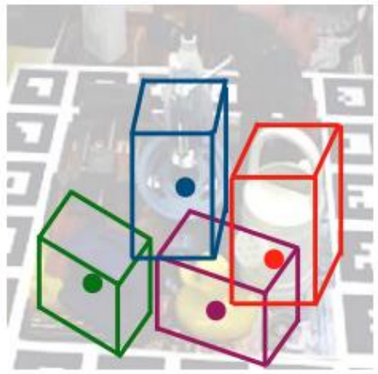
[R, t]

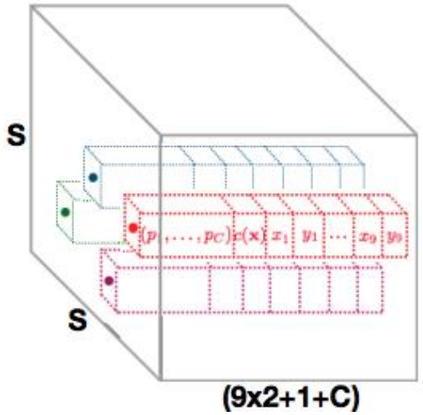
Microsoft

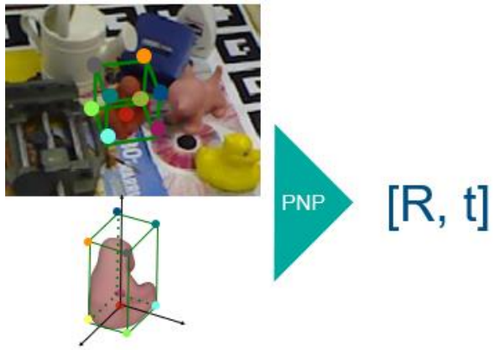
Our Approach











- Predict relative **coordinates of the 8 2D corner points**
- Also predict the **confidence and class probabilities**
- At inference time, recover the pose of the object with **PnP**

Microsoft

Sequence “Ape”

Sequence “Benchvise”

Sequence “Cat”

Quantitative Results

- Evaluation metric: Percentage of correctly estimated poses

	Method	Avg PCP
w/ Refinement	Brachmann et al., CVPR'16	73.7
	Rad & Lepetit, ICCV'17	89.3
w/o Refinement	Brachmann et al., CVPR'16	69.5
	Rad & Lepetit, ICCV'17	83.9
	OURS	**90.4**

Quantitative Results

- Evaluation metric: Percentage of correctly estimated poses

	Method	Avg PCP
w/ Refinement	Brachmann et al., CVPR'16	73.7
	Rad & Lepetit, ICCV'17	89.3
w/o Refinement	Brachmann et al., CVPR'16	69.5
	Rad & Lepetit, ICCV'17	83.9
	OURS	**90.4**

- Runtime efficiency

Method	FPS/object	Refinement runtime
Brachmann et al., CVPR'16	2 fps	100 ms/object
Rad & Lepetit, ICCV'17	3 fps	21 ms/object
Kehl et al., ICCV'17	10 fps	24 ms/object
OURS	50 fps	-

Real-time

Method	2D projection metric	Speed
416 × 416	89.71	94 fps
480 × 480	90.00	67 fps
544 × 544	90.37	50 fps
608 × 688	90.65	43 fps

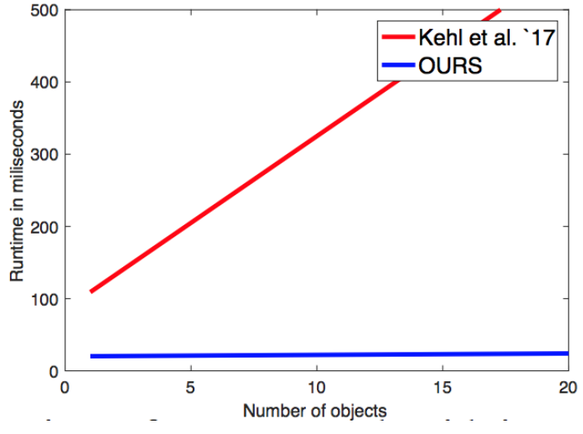
94 fps speed

Quantitative Results

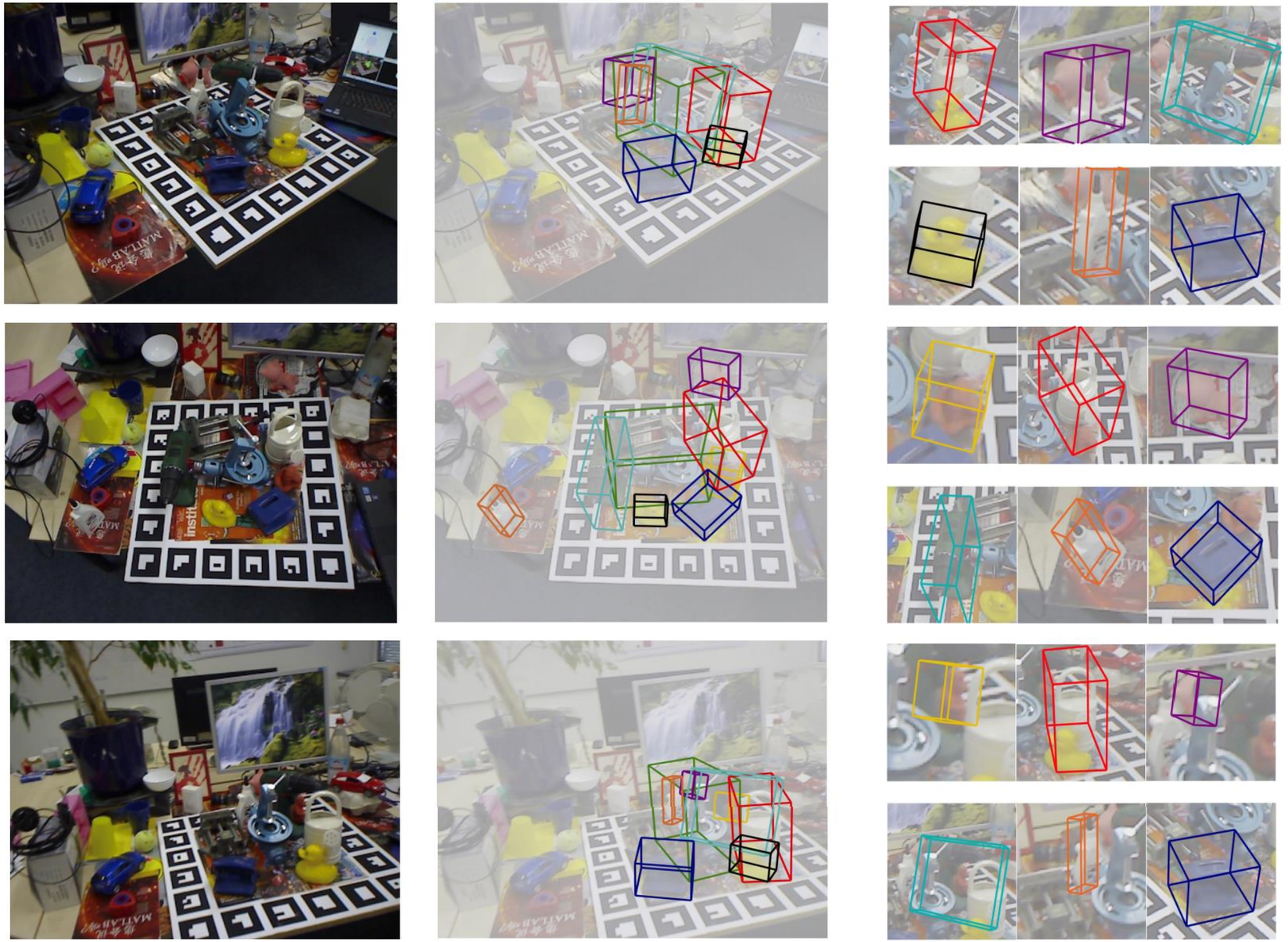
- Evaluation metric: Percentage of correctly estimated poses

	Method	Avg PCP
w/ Refinement	Brachmann et al., CVPR'16	73.7
	Rad & Lepetit, ICCV'17	89.3
w/o Refinement	Brachmann et al., CVPR'16	69.5
	Rad & Lepetit, ICCV'17	83.9
	OURS	**90.4**

- Runtime efficiency

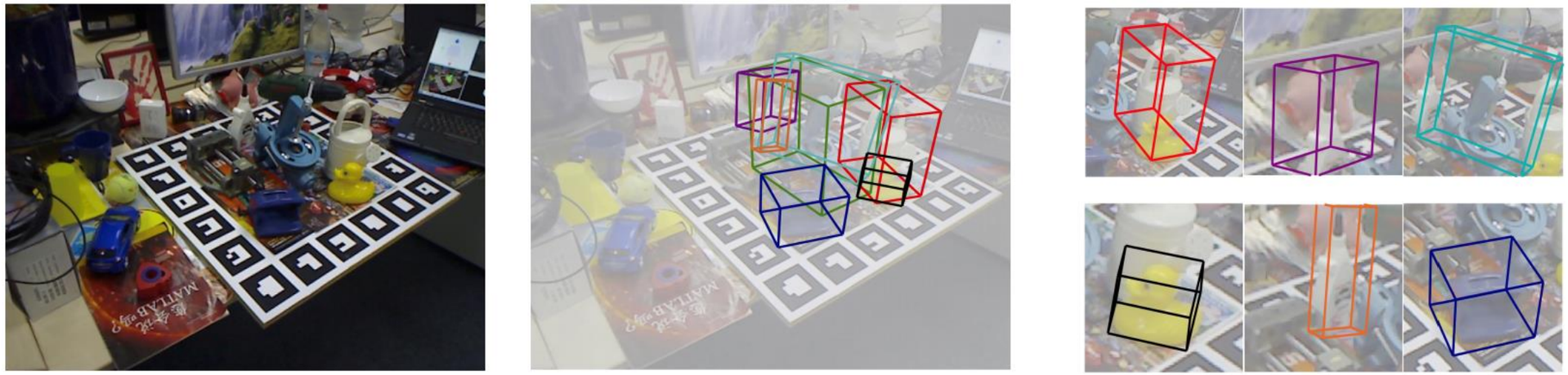


Multi-Object 6D Pose Estimation



Outline

1. Object Understanding



2. Hand + Object Understanding

1. B. Tekin, S. Sinha, P. Fua, “Real-time Seamless Single Shot 6D Object Pose Prediction”, CVPR’18

2. B. Tekin, F. Bogo, M. Pollefeys, “H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions”, CVPR’19

Microsoft

H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions

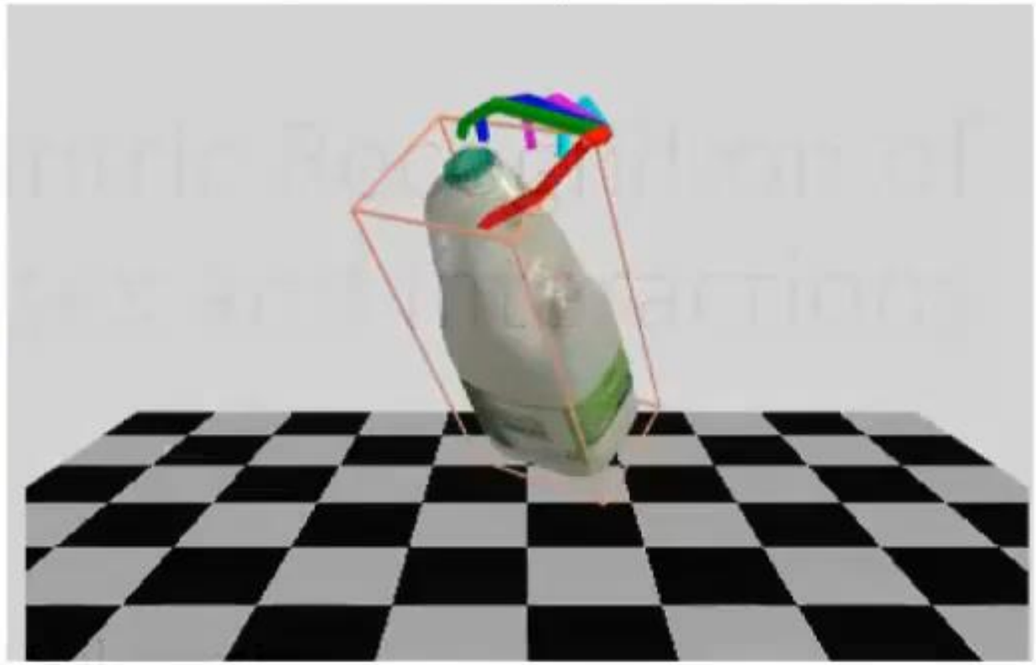
B. Tekin, F. Bogo, M. Pollefeys, “H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions”, **CVPR’19**

Microsoft

H+O

2D Hand+Object Pose

3D Hand+Object Pose



Per-frame predictions

Hand

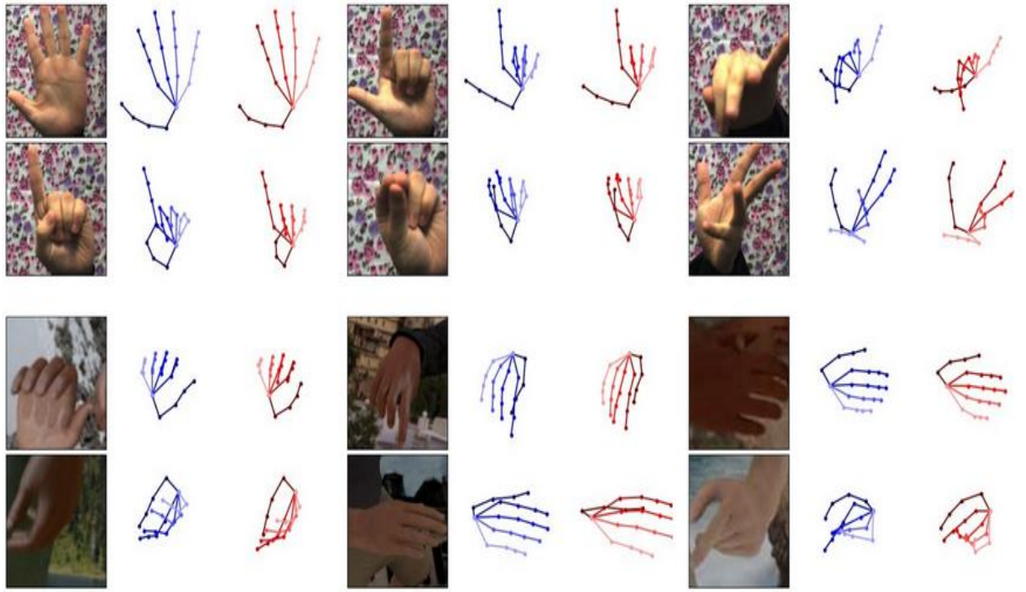
Zimmermann & Brox, ICCV’17

Müller et al., CVPR’18

Spurr et al., CVPR’18

Iqbal et al., ECCV’18

…



Object

Brachmann et al., CVPR’16

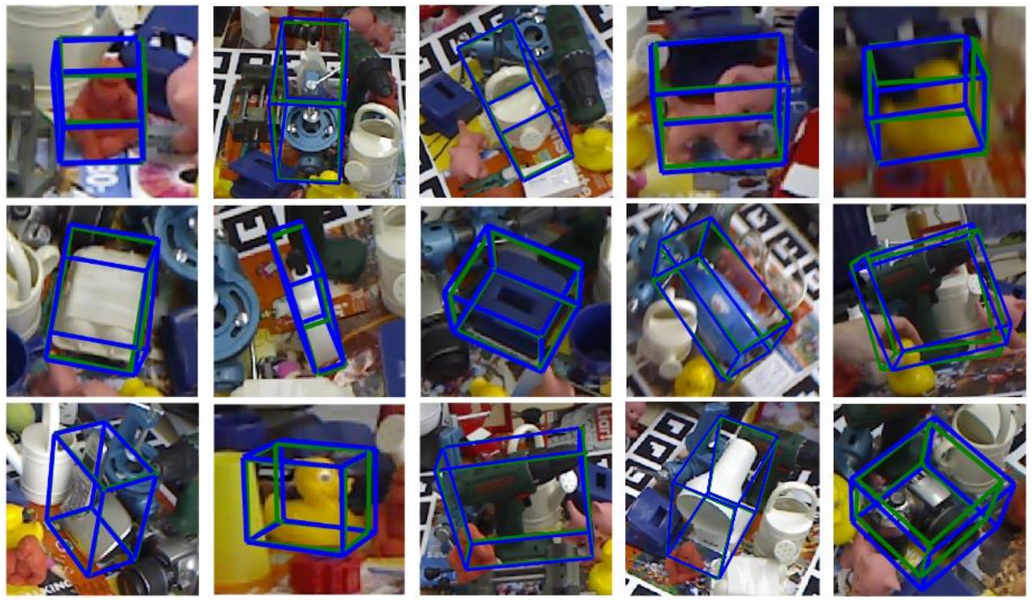
Rad & Lepetit, ICCV’17

Kehl et al., ICCV’17

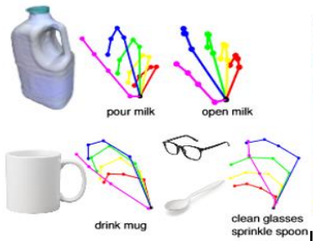
Tekin et al., CVPR’18

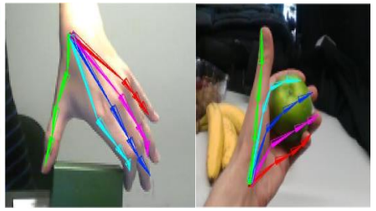
Sundermeyer et al., ECCV’18

…



Hands in Interaction with Objects

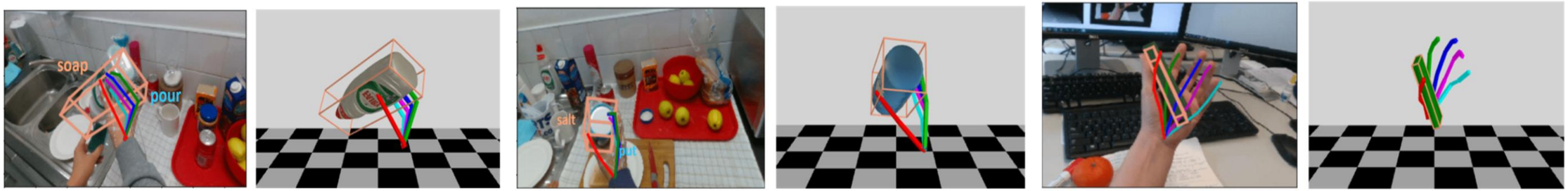




	RGB	Action Recognition	Object Pose	Hand Pose
Hamer et al., ICCV'09				●
Oikonomidis et al., ICCV'11			●	●
Sridhar et al., ECCV'16			●	●
Tzionas et al, IJCV'16			●	●
Müller et al., ICCV'17				●
Garcia-Hernando et al., CVPR'18		●		●
Hasson et al. CVPR'19	●		●	●
OURS	●	●	●	●

Microsoft

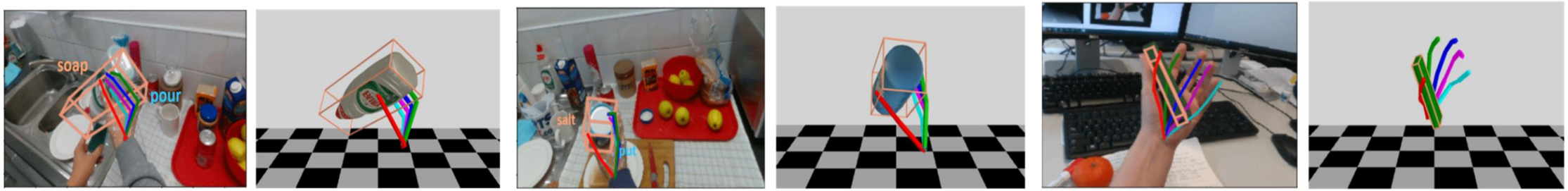
Hand+Object



1. Unified framework for recognizing 3D hand and object poses and interactions, by simultaneously solving 4 tasks

    a) 3D hand pose estimation
    b) 6D object pose estimation
    c) Object recognition
    d) Activity recognition

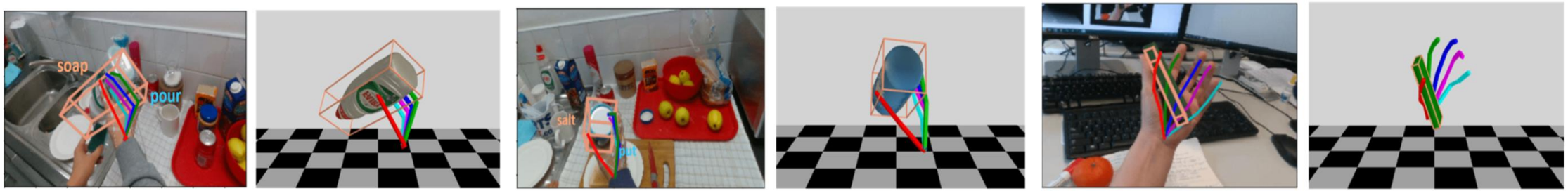
Hand+Object



1. Unified framework for recognizing 3D hand and object poses and interactions, by simultaneously solving 4 tasks

2. A single-shot deep architecture that jointly solves for articulated and rigid pose estimation

Hand+Object

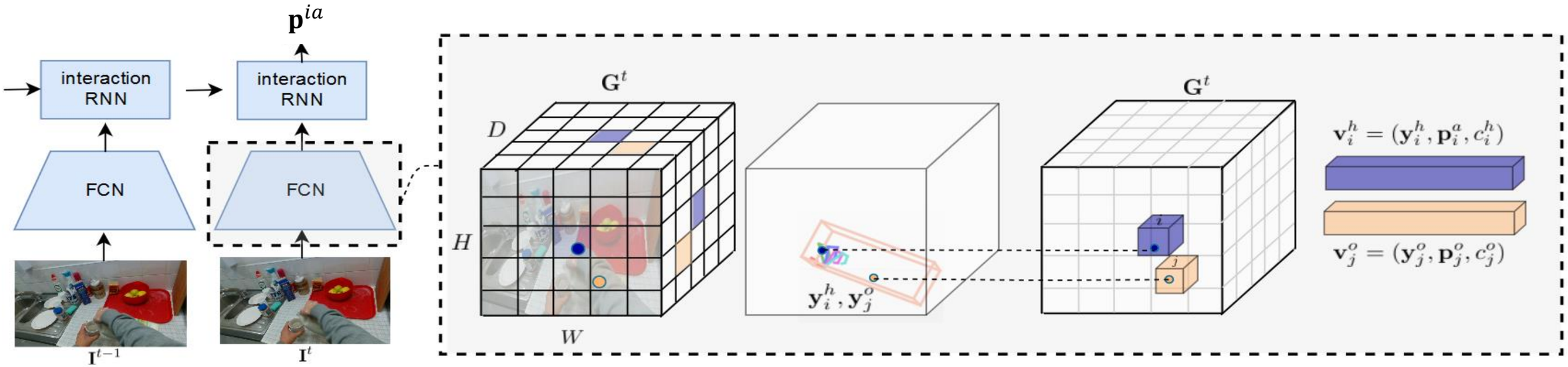


1. Unified framework for recognizing 3D hand and object poses and interactions, by simultaneously solving 4 tasks

2. A single-shot deep architecture that jointly solves for articulated and rigid pose estimation

3. A temporal model to explicitly model interactions in 3D between hands and objects

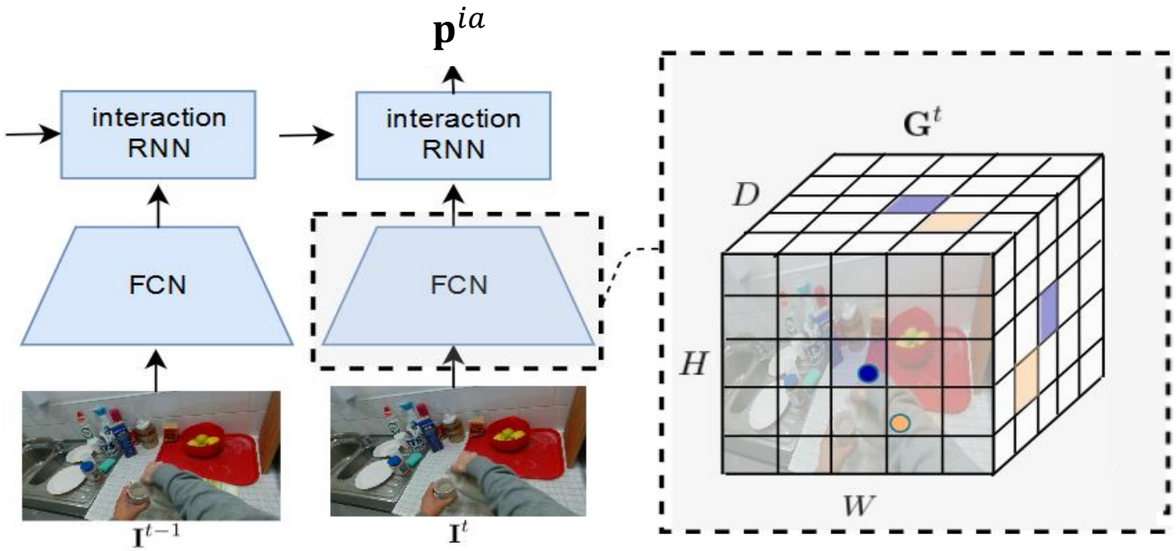
Hand+Object



Unified Egocentric Scene Understanding

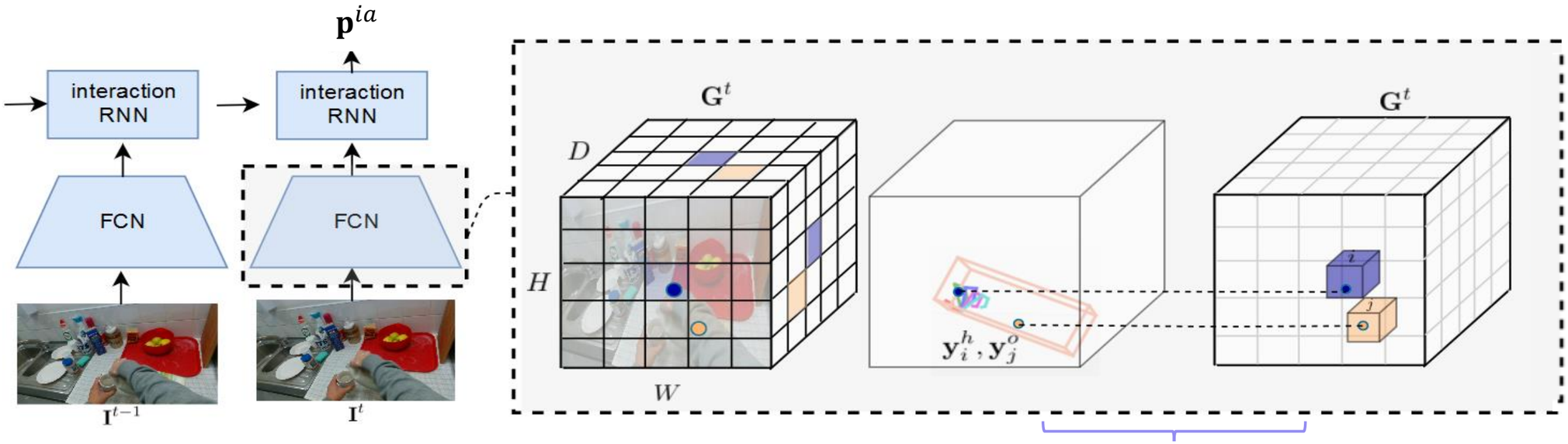
Microsoft

Hand+Object



store 3D locations
in a 3D grid

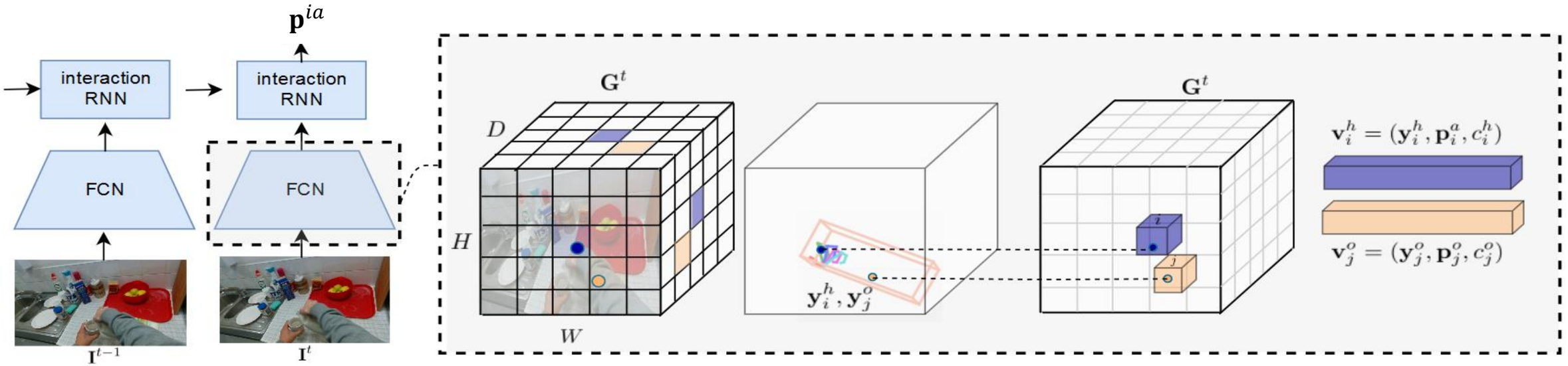
Hand+Object



3D hand skeleton
and 3D object BB
at the corresponding cell

Microsoft

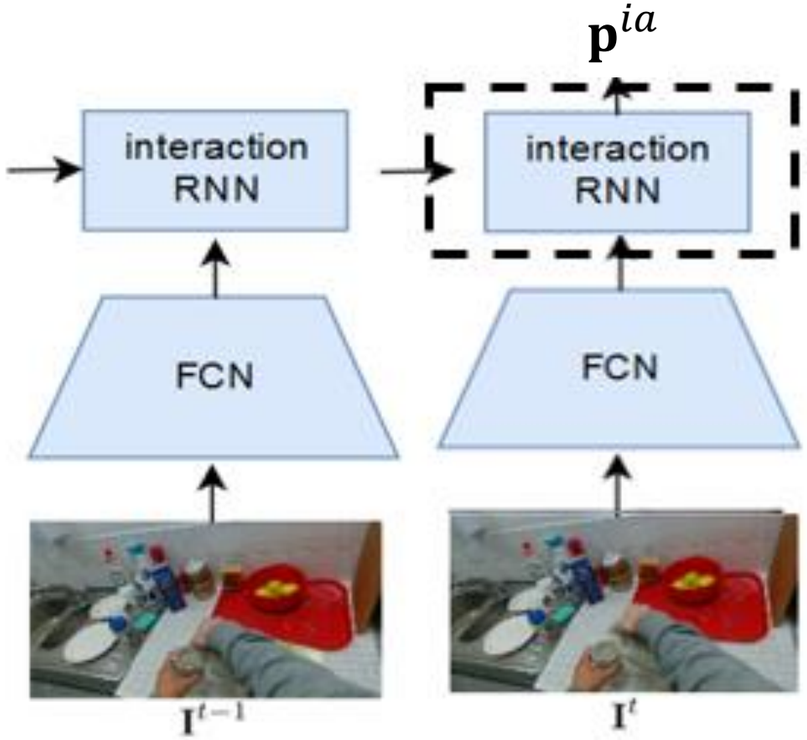
Hand+Object



encode 3D poses,
estimation confidence,
object and activity classes

Microsoft

Hand+Object



$f_\phi(g_\theta(\hat{\boldsymbol{y}}^h, \hat{\boldsymbol{y}}^o))$

explicitly model hand-object interactions, directly in 3D

Santoro et al., "A simple neural network module for relational reasoning", NIPS'17.

Microsoft

Ablation Studies

Method	Model	Action Accuracy (%)
Hernando et al., CVPR'18	Ground-truth Hand Pose	87.45
	Ground-truth Object Pose	74.45
	Ground-truth Hand + Object Pose	91.97
OURS	SINGLE-IMAGE	85.56
	HAND POSE	89.47
	OBJECT POSE	85.71
	HAND + OBJECT POSE	94.73
	HAND + OBJECT POSE + INTERACT	**96.99**

← Jointly using hand and object poses improves activity recognition

Ablation Studies

Method	Model	Action Accuracy (%)
Hernando et al., CVPR'18	Ground-truth Hand Pose	87.45
	Ground-truth Object Pose	74.45
	Ground-truth Hand + Object Pose	91.97
OURS	SINGLE-IMAGE	85.56
	HAND POSE	89.47
	OBJECT POSE	85.71
	HAND + OBJECT POSE	94.73
	HAND + OBJECT POSE + INTERACT	**96.99**

Network	HP error	OP error
HAND ONLY	16.15	-
OBJECT ONLY	-	28.27
HAND + OBJECT	16.87	25.54
HAND + OBJECT + INTERACT	**15.81**	**24.89**

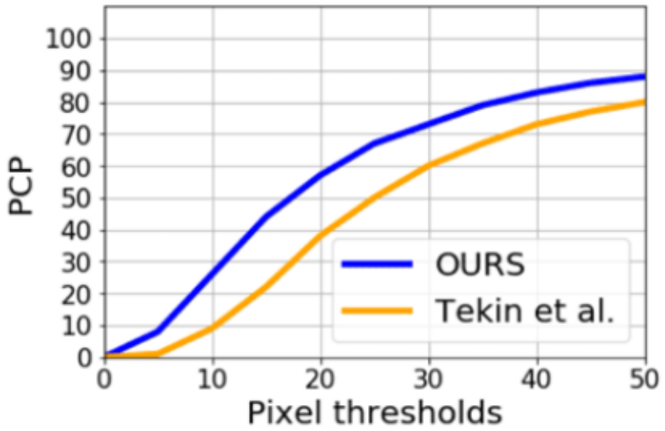
← Unified framework yields better accuracy than task-dedicated individual networks

Microsoft

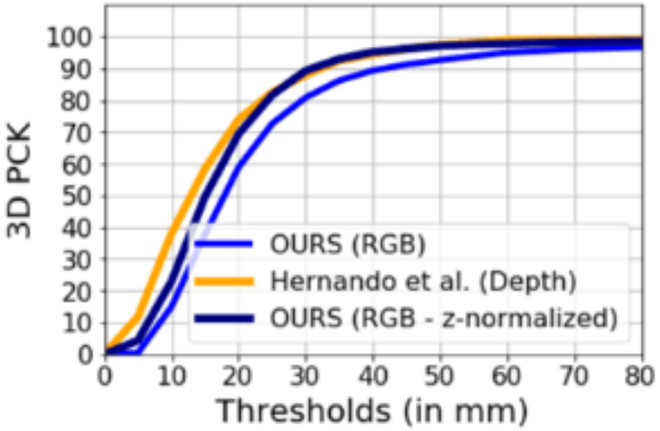
State-of-the-Art Comparison

Model	Input modality	Accuracy
Two-stream-color	Color	61.56
Two-stream-flow	Color	69.91
Two-stream-all	Color	75.30
Joule-color	Color	66.78
HON4D	Depth	70.61
Novel View	Depth	69.21
Joule-depth	Depth	60.17
GT-Pose + Gram Matrix	Depth	32.22
GT-Pose + Lie Group	Depth	69.22
GT-Pose + LSTM	Depth	72.06
OURS - HP	Color	62.54
OURS - HP + AC	Color	74.20
OURS - HP + AC + OC	Color	**82.43**

Activity Recognition



Object pose estimation



Hand pose estimation

FPHA dataset

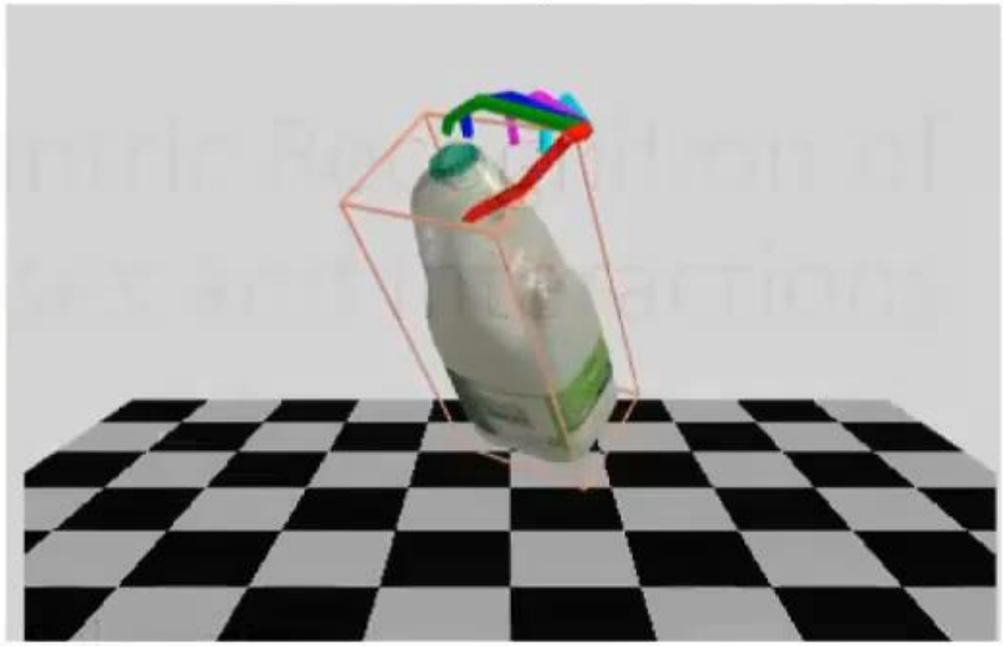
Garcia-Hernando et al., "First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations", CVPR'18.

Microsoft

Thanks!

2D Hand + Object Pose

3D Hand + Object Pose



Per-frame predictions

H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions

Microsoft