Carnegie Mellon University
The Robotics Institute

Sequential Voting with Relational Box Fields for Active Object Detection

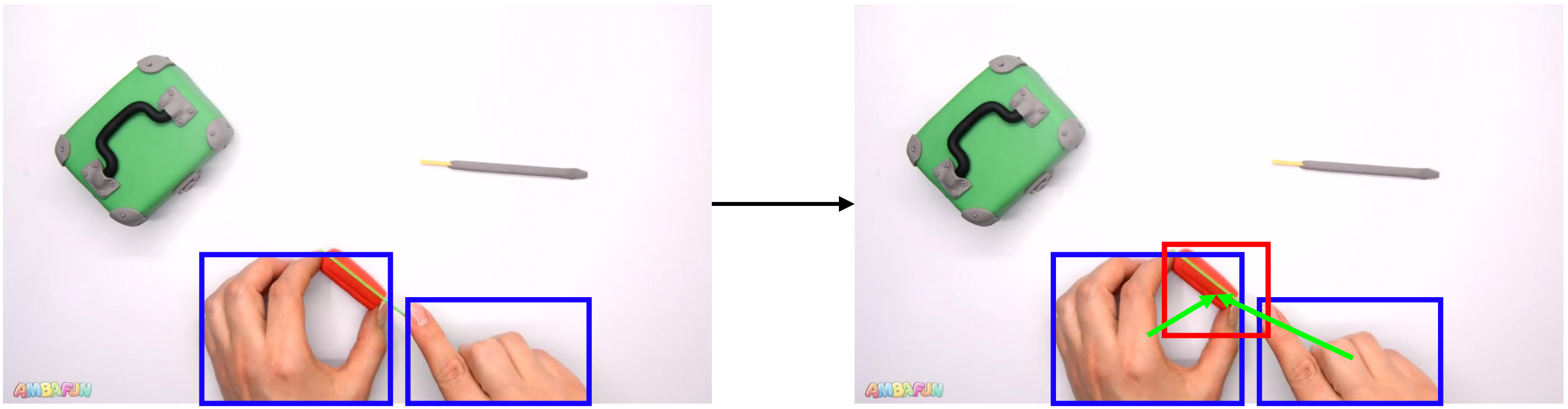
CVPR 2022

Qichen Fu, Xingyu Liu, Kris M. Kitani
Carnegie Mellon University

Paper & Code:
fuqichen1998.github.io/SequentialVotingDet/

Goal

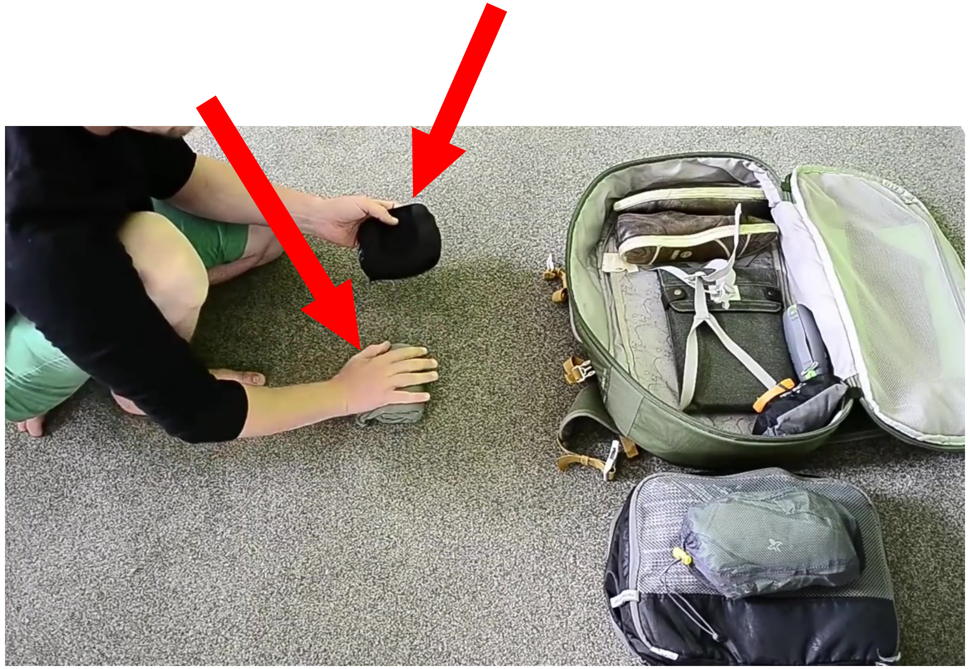
Detect the bounding box of the active object (red box), along with its correspondence (green arrow) to the human hands (green arrow).



Active Object Detection

Challenges

Natural **occlusions** caused by the hands during hand-object interactions



Motivation

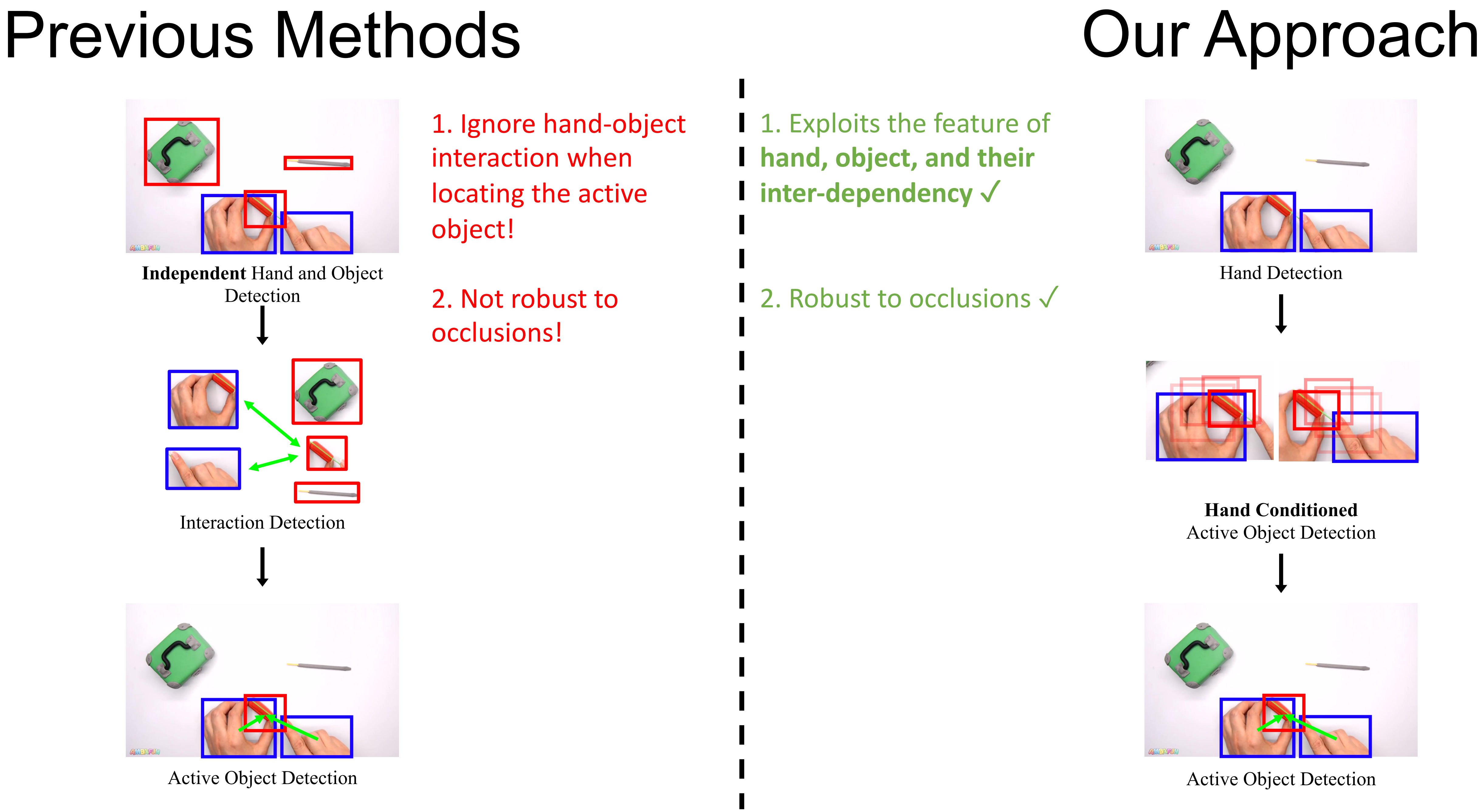
Despite occlusion, the appearance of the hand gives a strong **hint** about the location, shape, size, and pose of the active object.

Q1: Is the person inside the yellow bounding box interacting with any object or other person in the image?

Q2: If an interaction is present, draw a bounding box on the object or person that the person in the given yellow bounding box is interacting with.

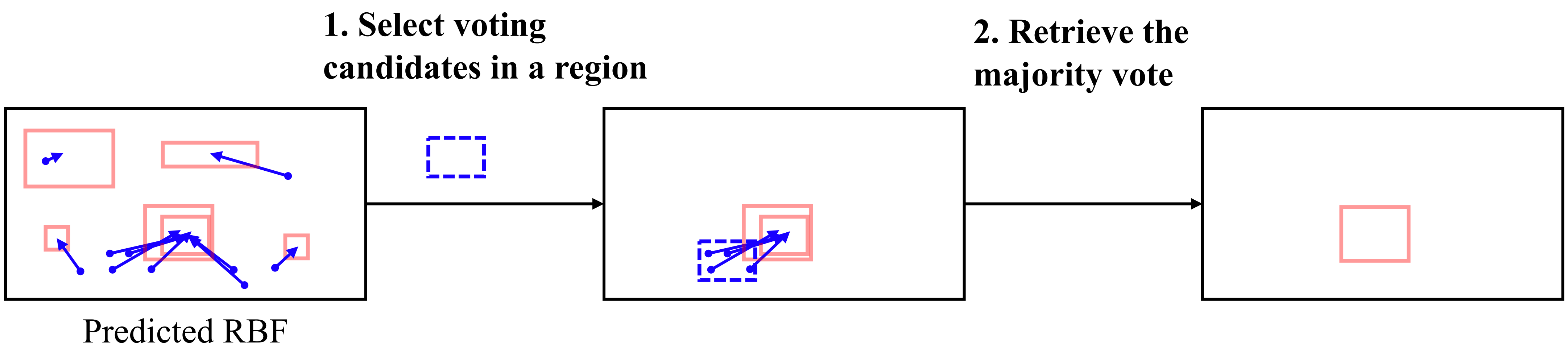
Multiple annotators' interactee estimates (orange) and the consensus ground truth (thick red).

Chen, Chao-Yeh, and Kristen Grauman. "Predicting the location of “interactees” in novel human-object interactions." Asian conference on computer vision. Springer, Cham, 2014.


Previous Methods
Our Approach
1. Ignore hand-object interaction when locating the active object!
2. Not robust to occlusions!
Independent Hand and Object Detection
Interaction Detection
Active Object Detection
1. Exploits the feature of hand, object, and their inter-dependency ✓
2. Robust to occlusions ✓
Hand Detection
Hand Conditioned Active Object Detection
Active Object Detection

Voting on Relational Box Field for Object Detection under **Occlusions**

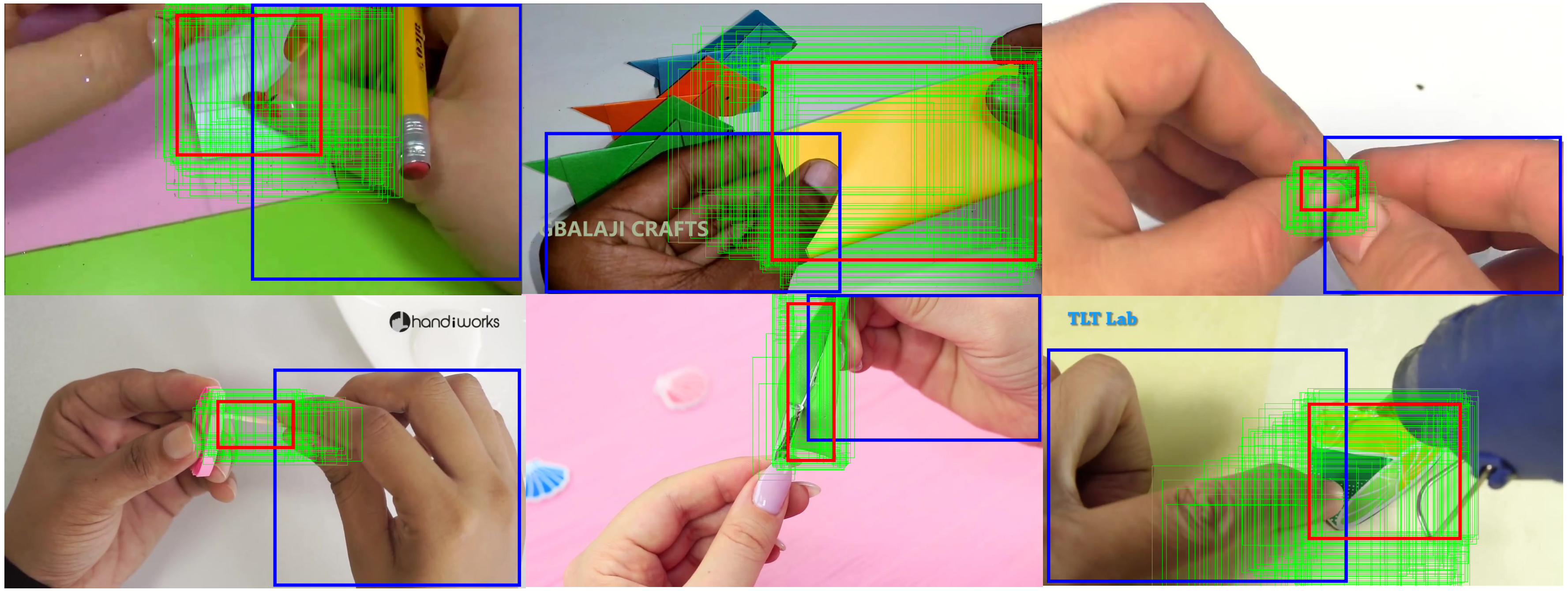
Relational Box Field (RBF): every pixel point to one estimated bounding box



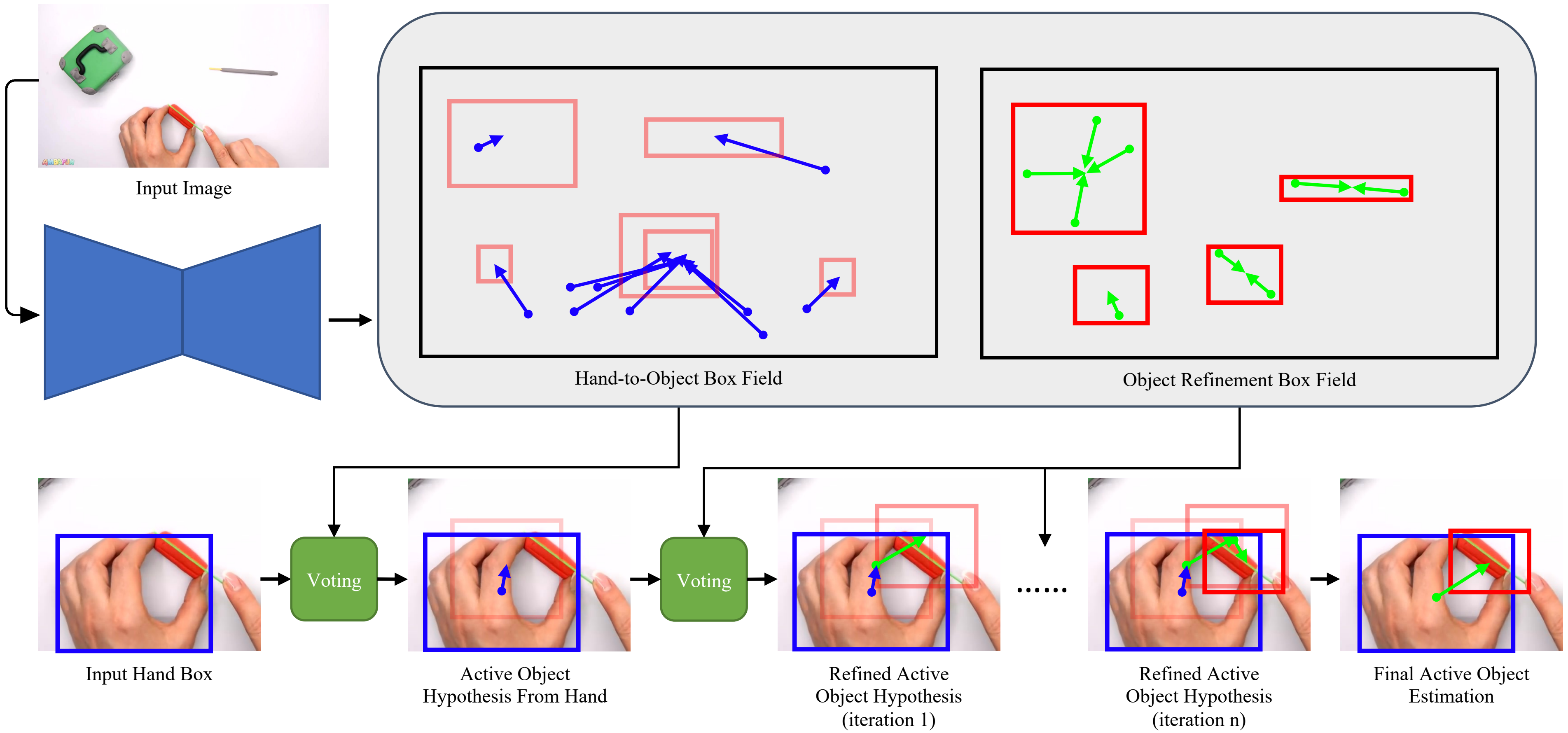
Voting Examples

Green Box: voting candidates from pixels belong to the hand (blue box)

Red Box: the majority vote of the active object



Method Overview



Quantitative Results on 100DOH Dataset

Method	Backbone	# Params	Hand Source	AP^{50}_{hand}	AP^{75}	AP^{50}	AP^{25}
Simple Baseline	R101	47M	FasterRCNN	89.59	28.15	44.73	47.57
100DOH Detector	DLA34	47M	FasterRCNN	89.59	28.50	46.95	51.80
PPDM	R50	21M	CenterNet	89.64	26.89	45.80	53.04
HOTR	R101	51M	DETR	90.26	29.30	49.27	**57.80**
Ours	R101	48M	FasterRCNN	89.59	**29.90**	**53.02**	57.15
Simple Baseline	R101	47M	Ground Truth	100	34.51	44.68	52.35
Ours	R101	48M	Ground Truth	100	**40.05**	**54.82**	**64.86**

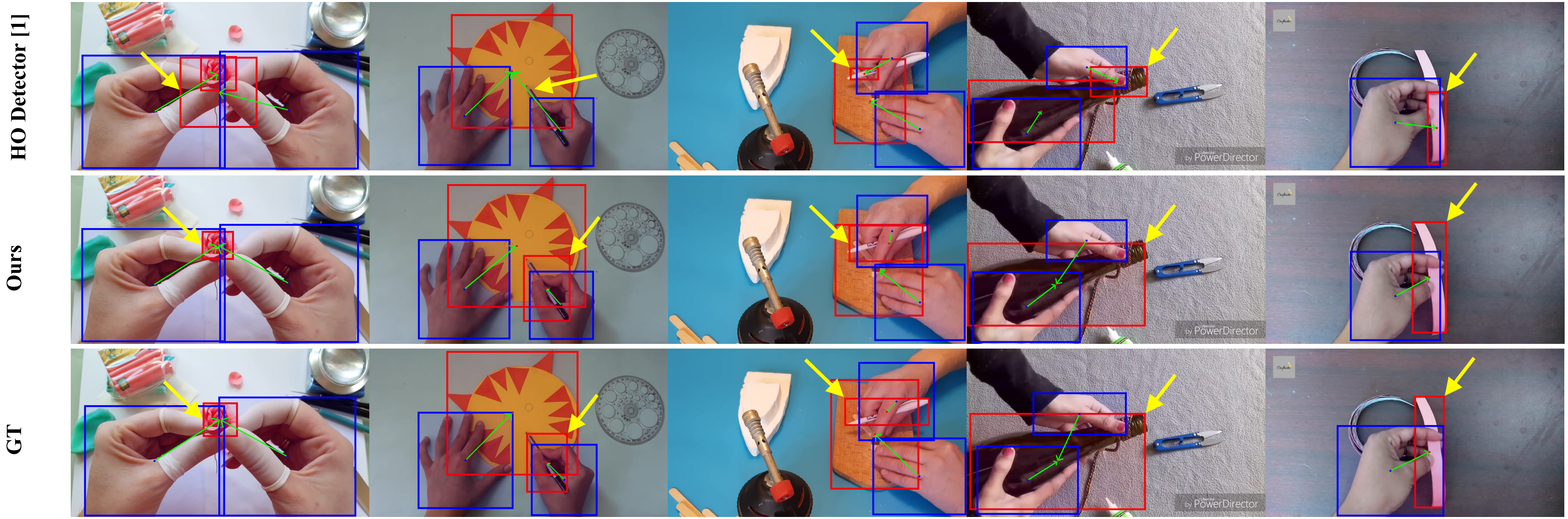
Method	Recall (IoU ∈ [.25,.5))	Recall (IoU ∈ [.5,.75))	Recall (IoU ∈ [.75,.1))
100DOH Detector	68.68	63.22	78.57
PPDM	53.24	53.45	64.29
HOTR	71.69	68.10	71.43
Ours	**77.22**	**78.45**	**100** (14 samples)
	low occlusion	medium occlusion	high occlusion

Quantitative Results on MECCANO Dataset

Method	Backbone	Finetune	AP^{75}	AP^{50}	AP^{25}
100DOH Detector	R101	✗	-	11.17	-
Ours	R101	✗	**9.09**	**16.61**	**23.97**
100DOH Detector	R101	✓	-	20.18	-
Ours	R101	✓	**12.99**	**26.25**	**34.88**

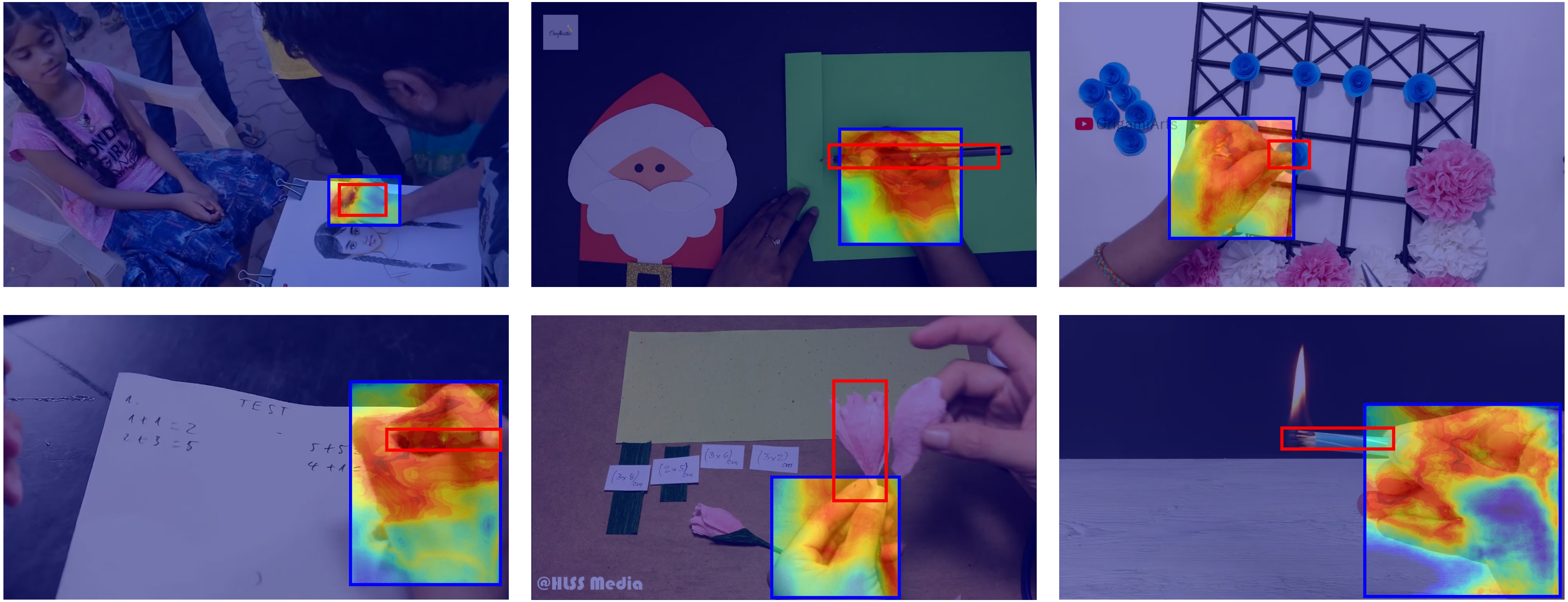
Trained on 100DOH,
test on MEECANO.

Comparison with HO Detector [1]



[1] Shan, Dandan, et al. "Understanding human hands in contact at internet scale." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

The correlation between pixel-wise predictions and the final prediction



Paper & Code:

fuqichen1998.github.io/SequentialVotingDet/