ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

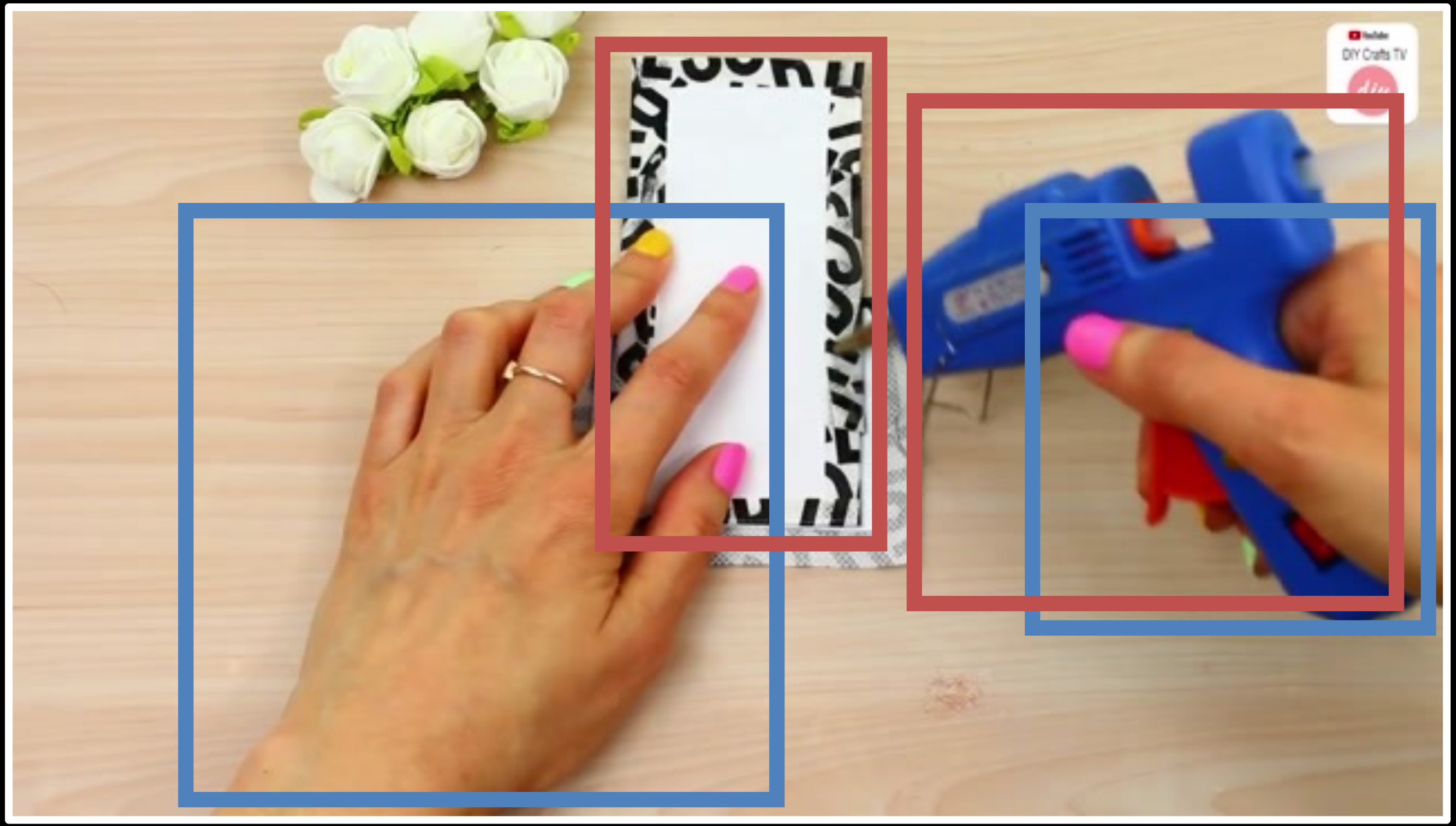
COHESIV: Contrastive Object and Hand Embeddings for Segmentation In Video

Dandan Shan*, Richard E.L. Higgins*, David F. Fouhey

University of Michigan

NeurIPS 2021

Previously

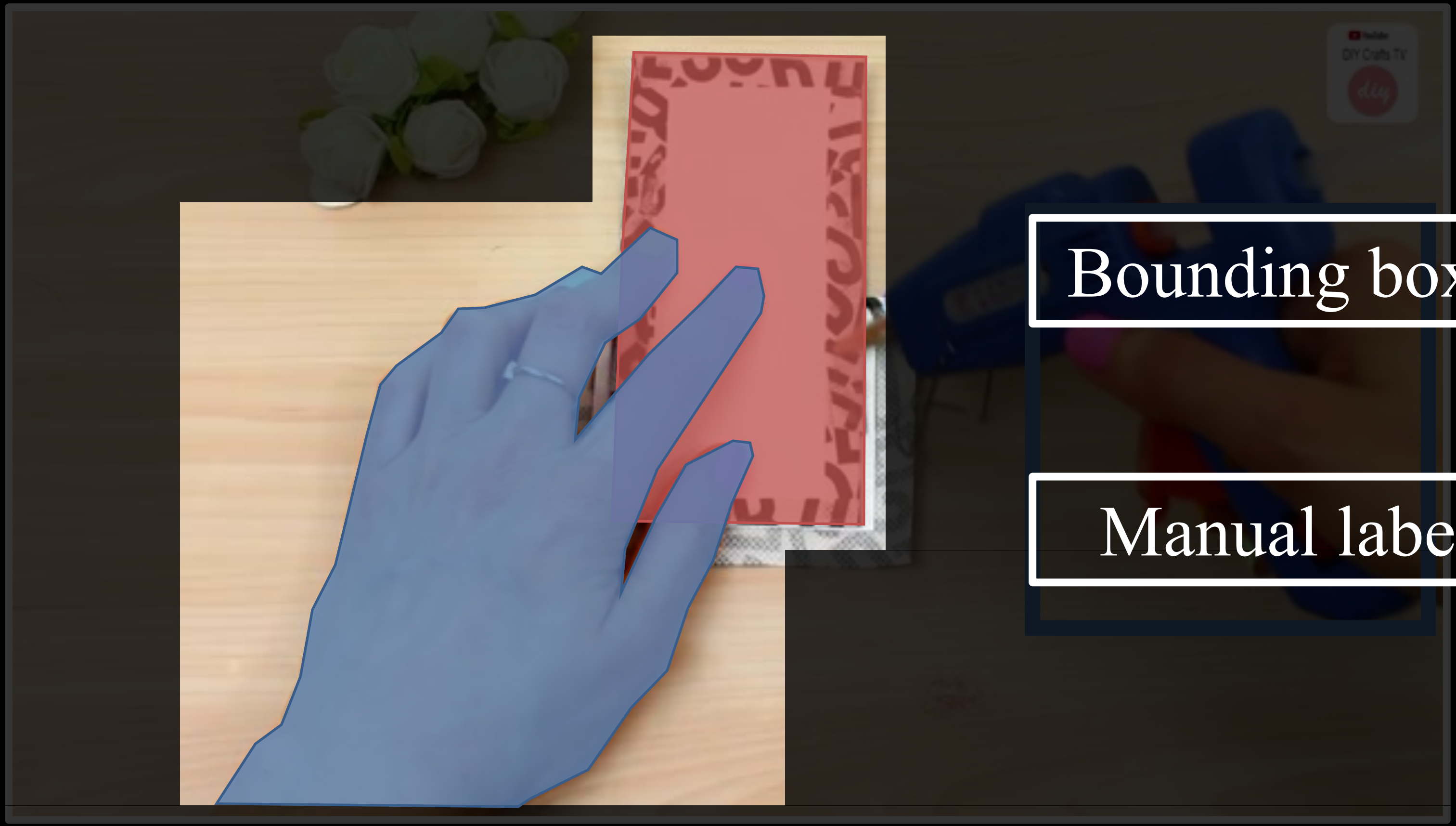


Previously

Bounding boxes

Manual labels

Our goals

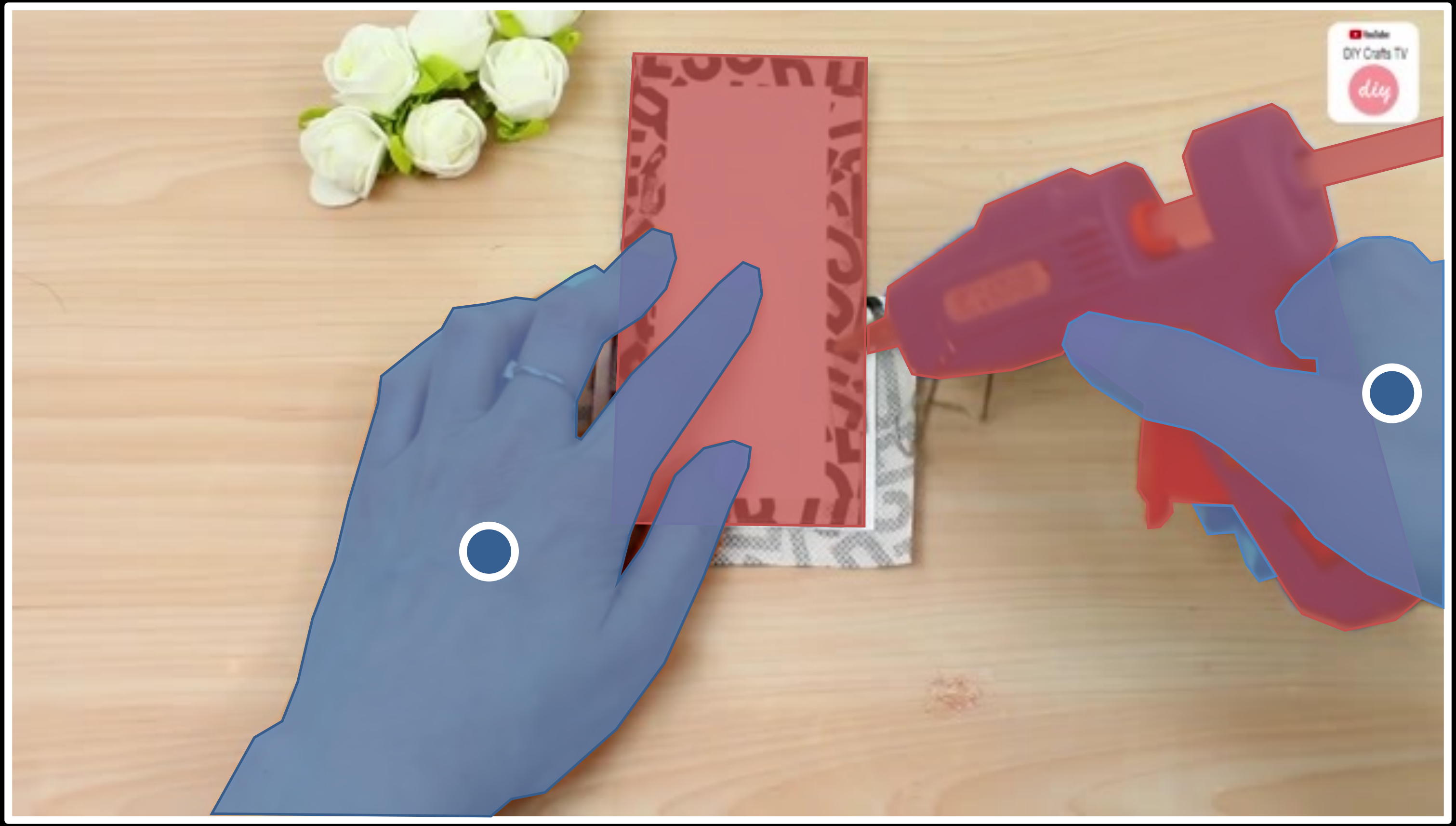


Bounding boxes → Pixels

Manual labels → Pseudo labels

The Problem

Segment hand and hand-held object.



Motion

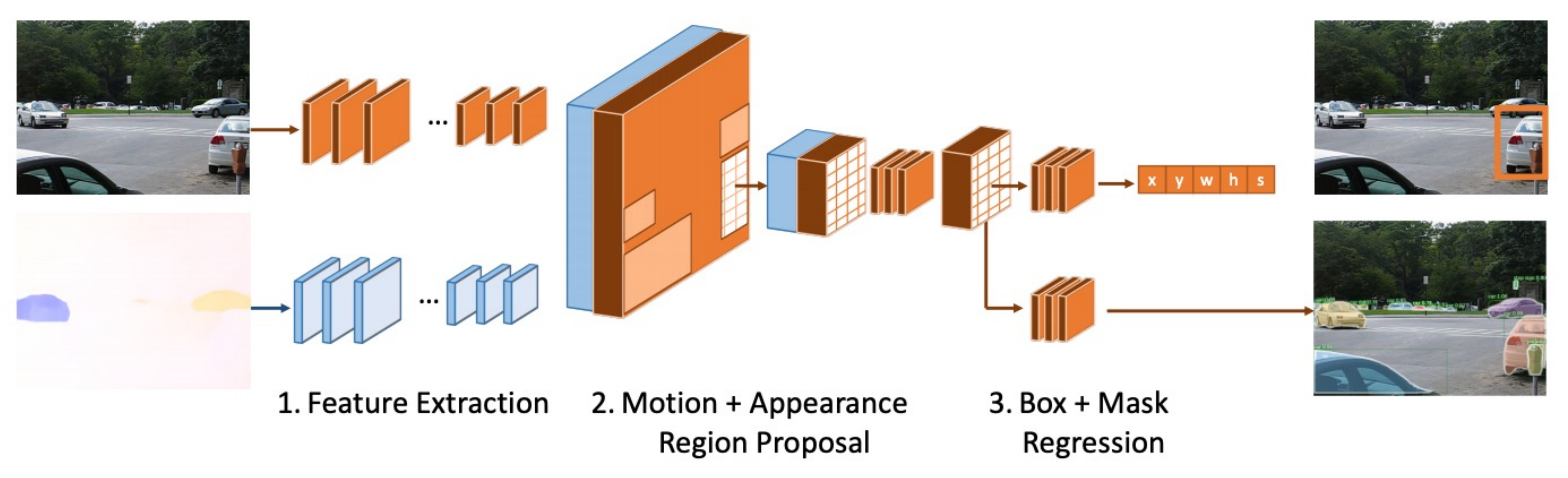
Common Fate

Common Fate in Gestalt Psychology (Wertheimer 1938): elements that are moving together tend to be perceived as a unified group.

Pic resource: Pinterest.com

Related Work

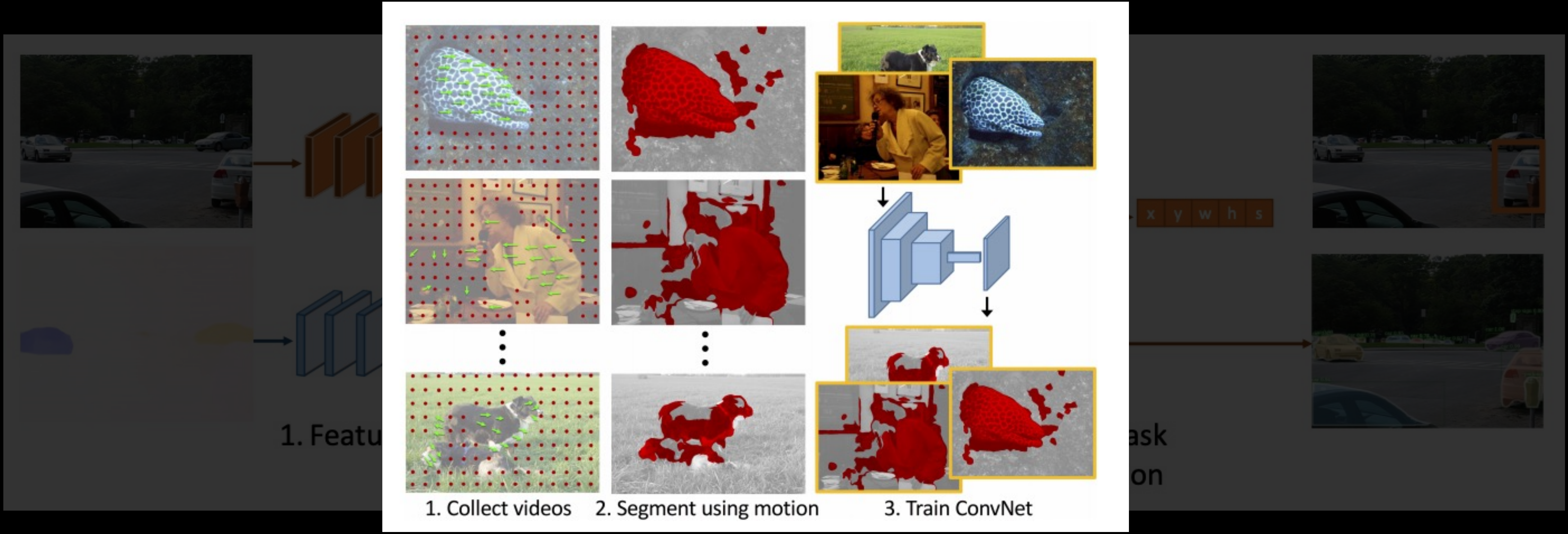
Need optical flow at train/test time.



Dave et al. Towards Segmenting Anything That Moves. CVPR 2019 Workshop.

Related Work

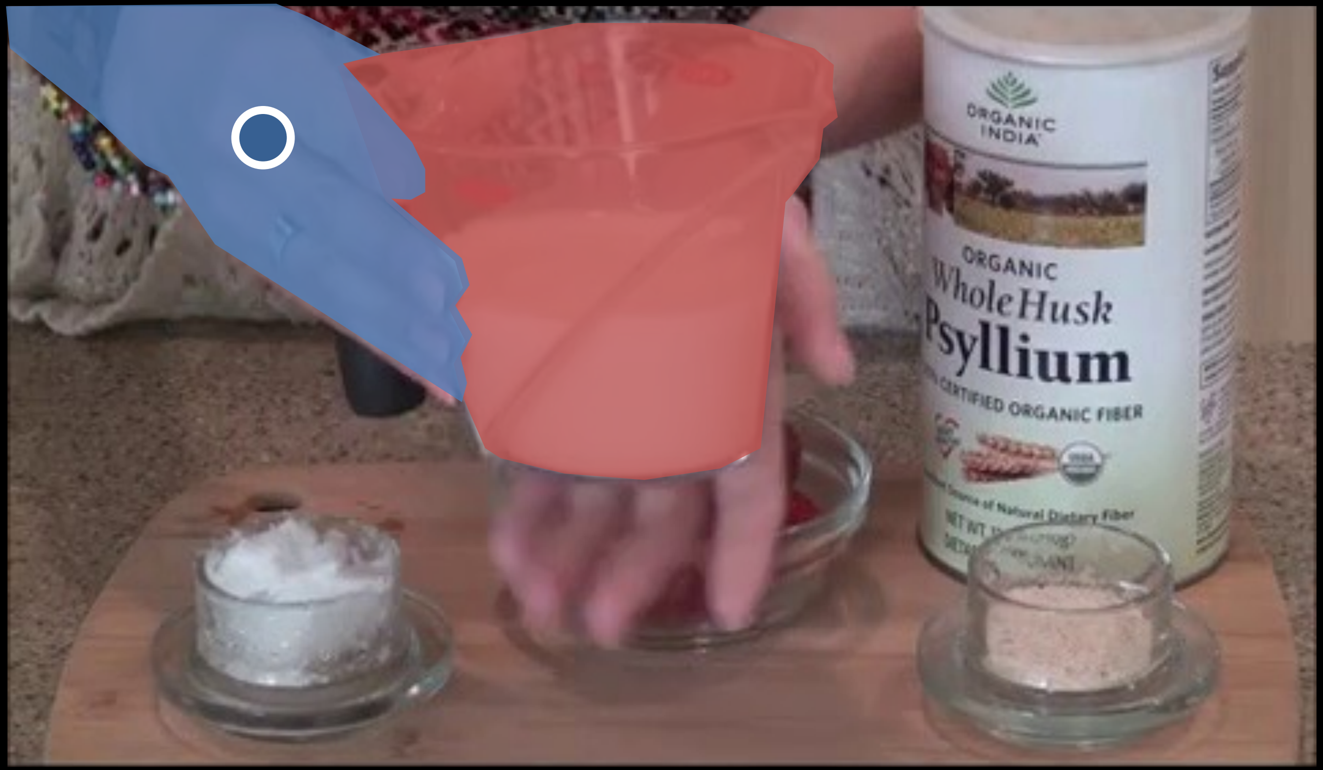
Use motion cues for feature learning.



Pathak et al. Learning Features by Watching Objects Move. CVPR 2017.

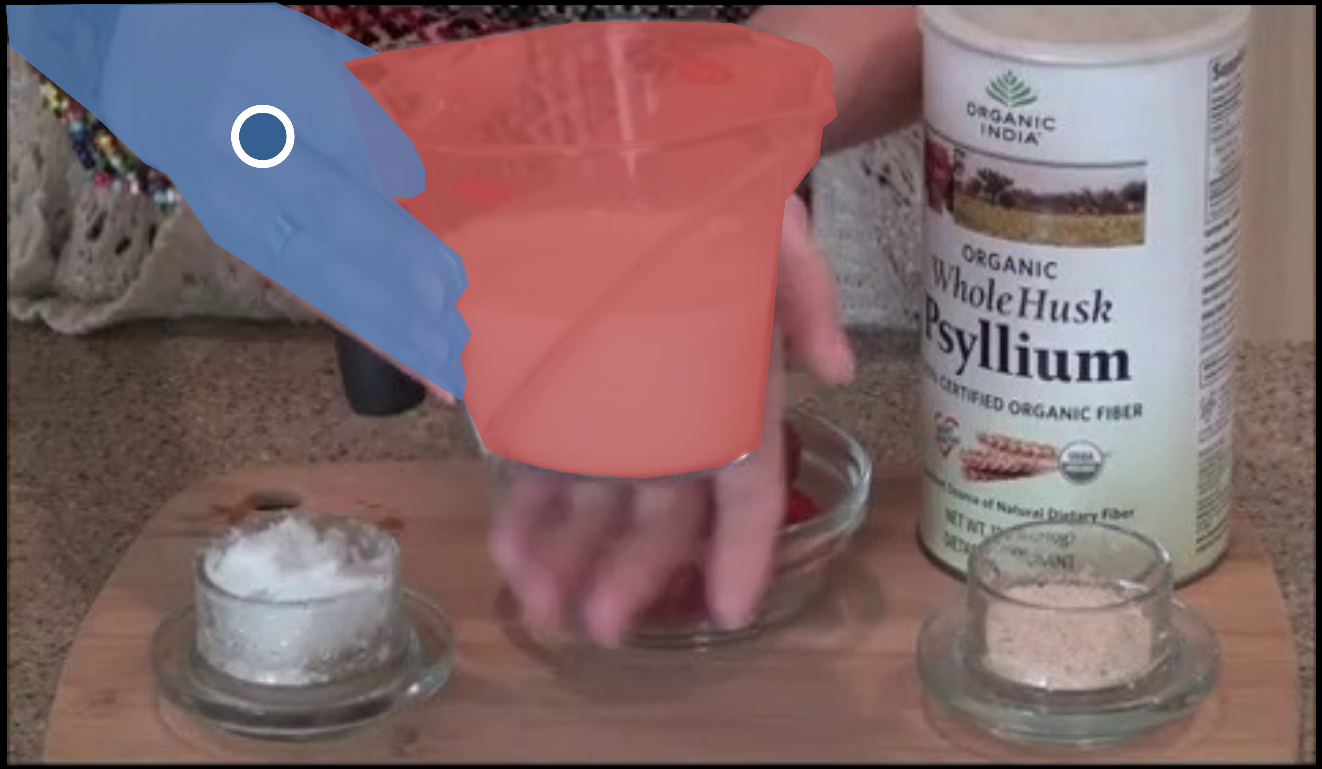
Problem Setup

- Task: learn from motion to segment hand and hand-held object in image.



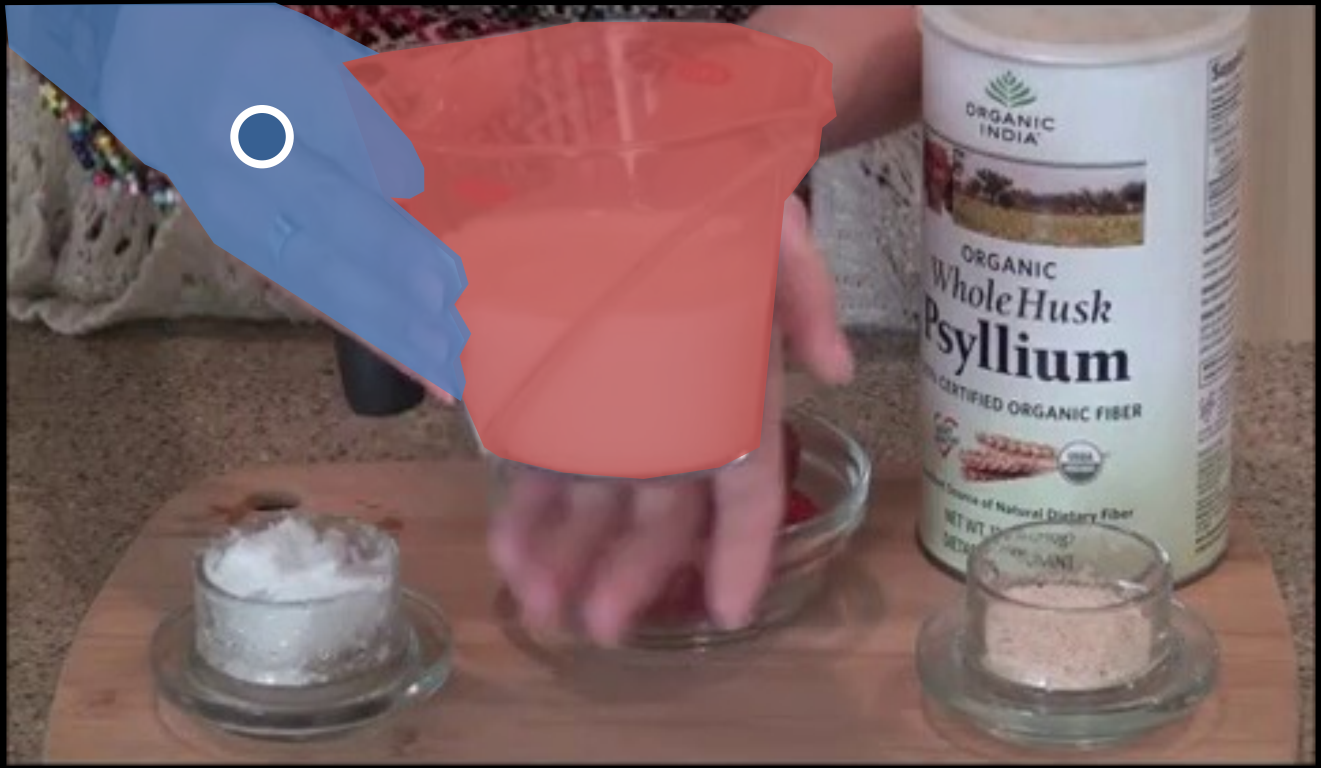
Problem Setup

- Task: learn from motion to segment hand and hand-held object in image.
- Training time: use pseudo-labels from motion for learning.



Problem Setup

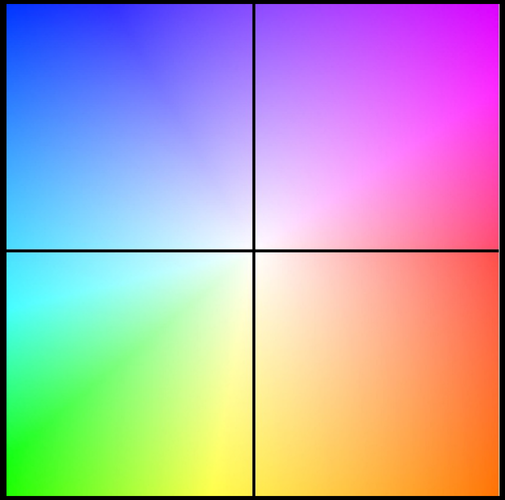
- Task: learn from motion to segment hand and hand-held object in image.
- Training time: use pseudo-labels from motion for learning.
- Test time: only input RGB+(x, y) to get prediction.

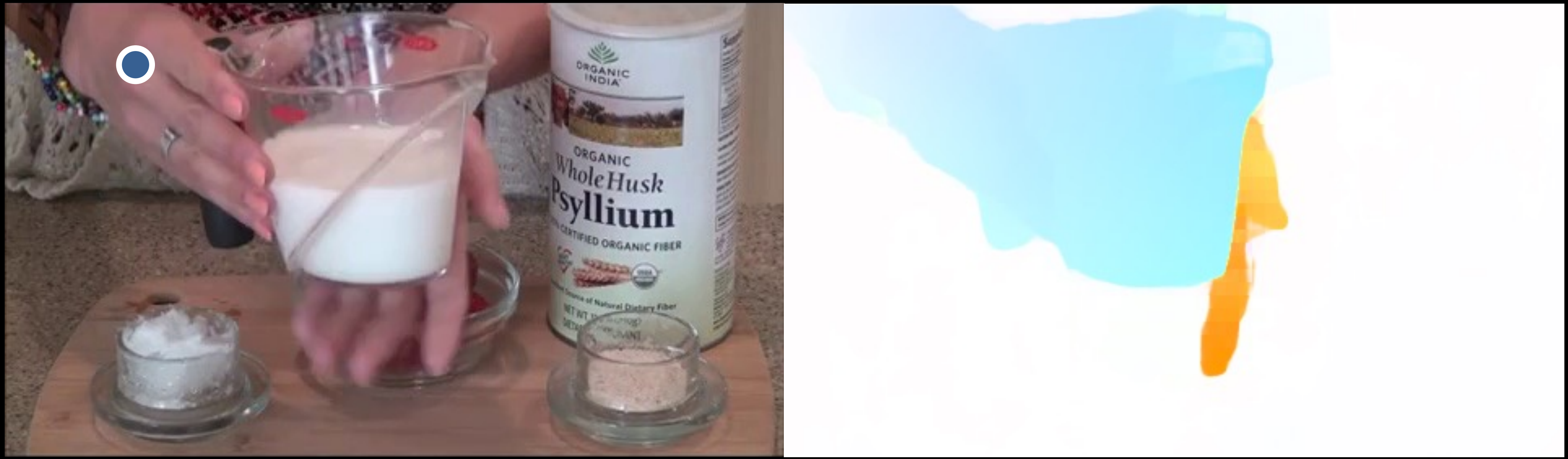


Optical Flow

- Motion of pixels between frames.

Flow Field Color Coding

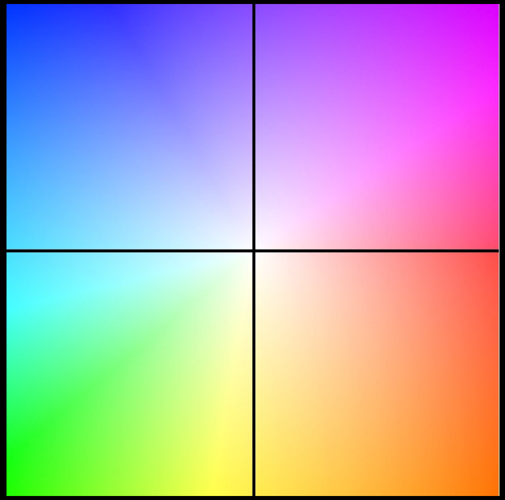




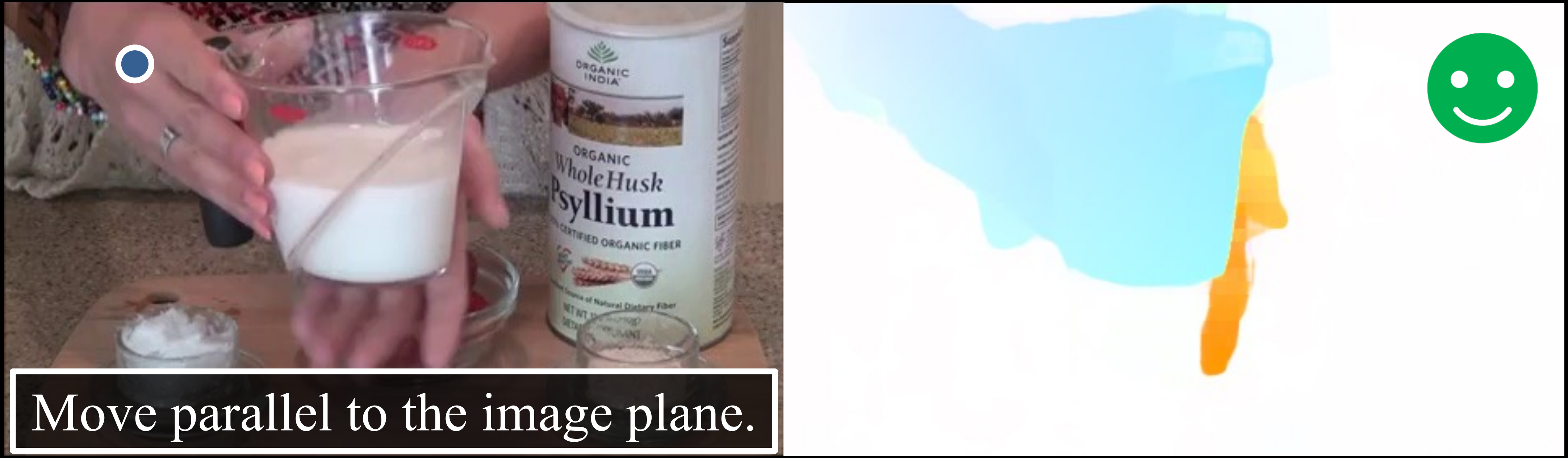
Teed et al. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. ECCV 2020.

Optical Flow

Flow Field Color Coding



- Motion of pixels between frames.
- Work well on in-plane motion!



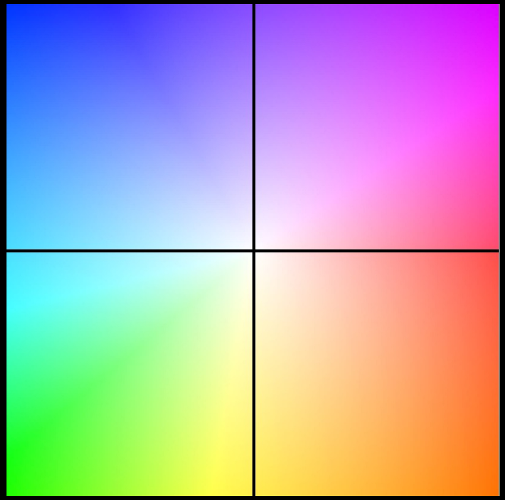
Move parallel to the image plane.

Teed et al. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. ECCV 2020.

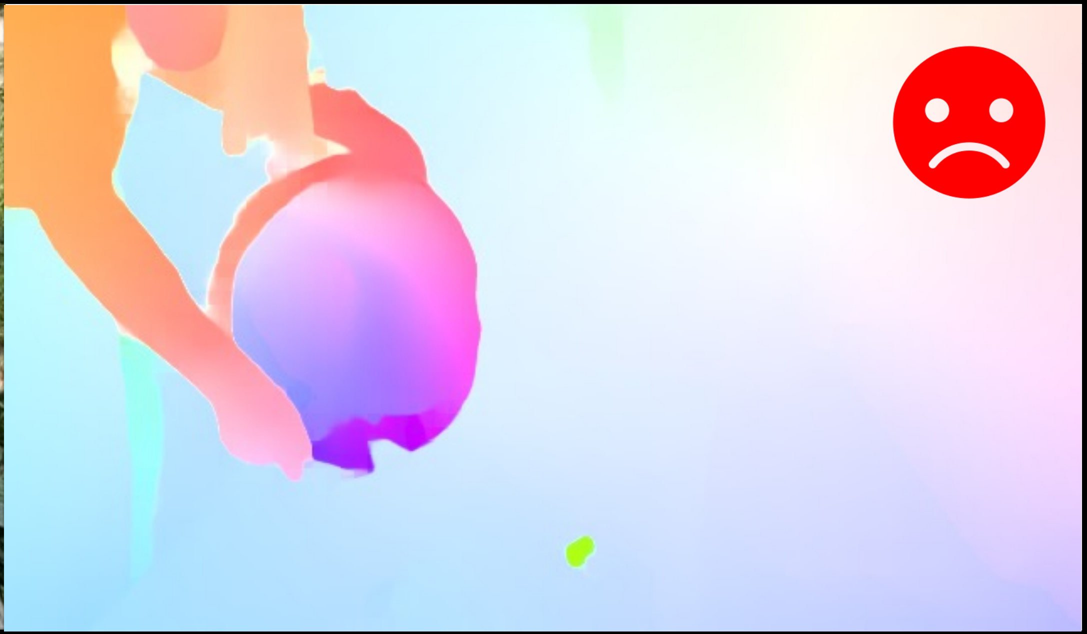
Optical Flow

- Motion of pixels between frames.
- Work well on in-plane motion!
- Out-of-plane motion is not simple!

Flow Field Color Coding



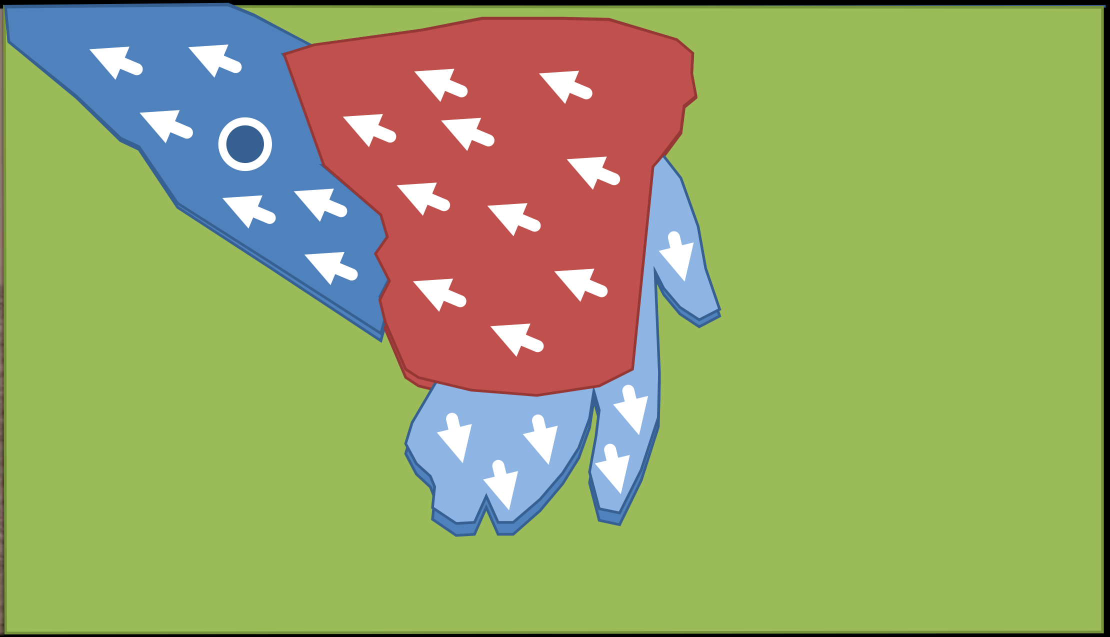
Rotate towards/away the camera.



Teed et al. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. ECCV 2020.

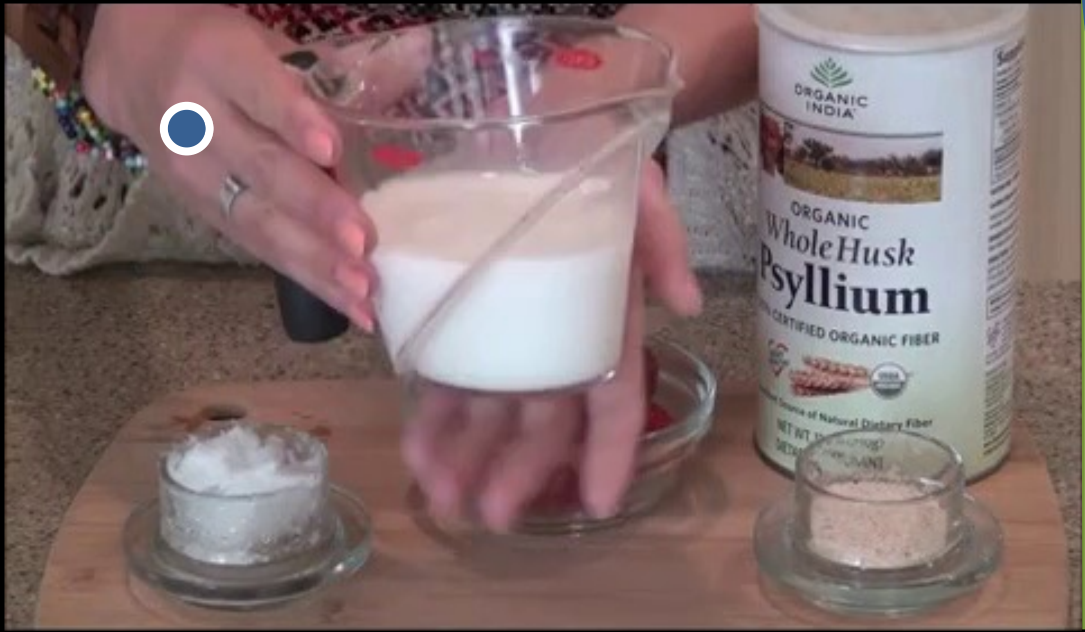
Responsibility

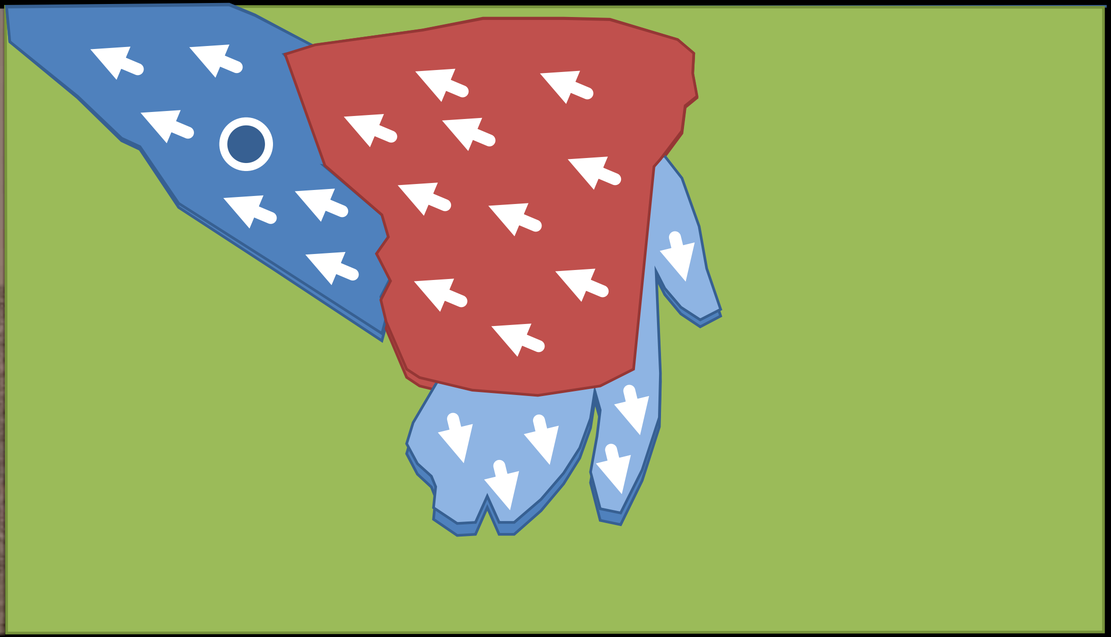
Responsibility



Responsibility

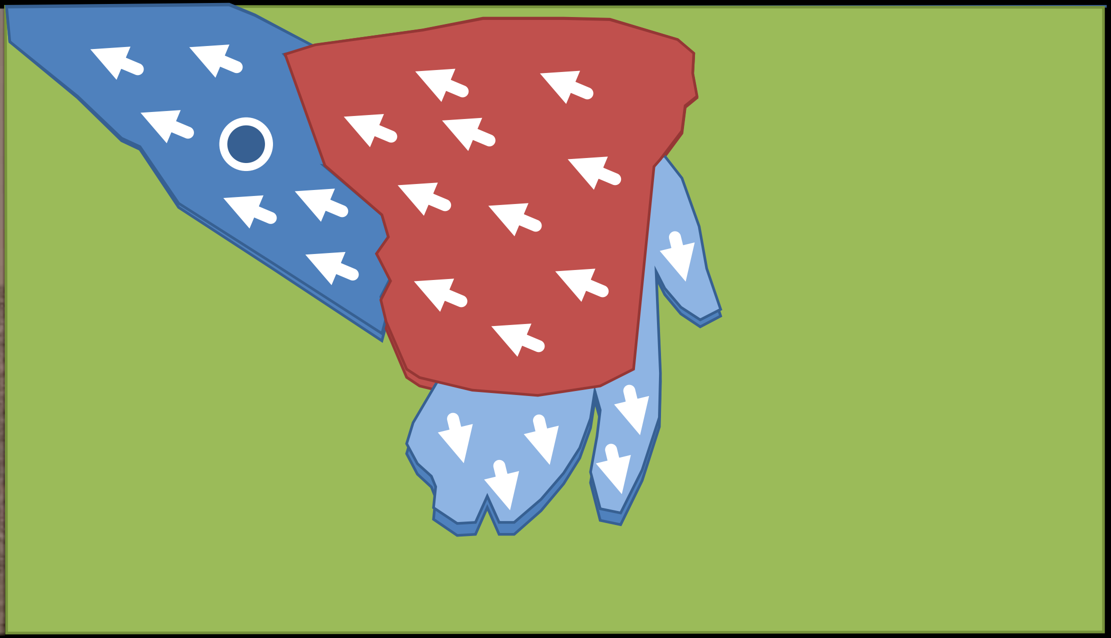
- How well does a hand explain the motion?





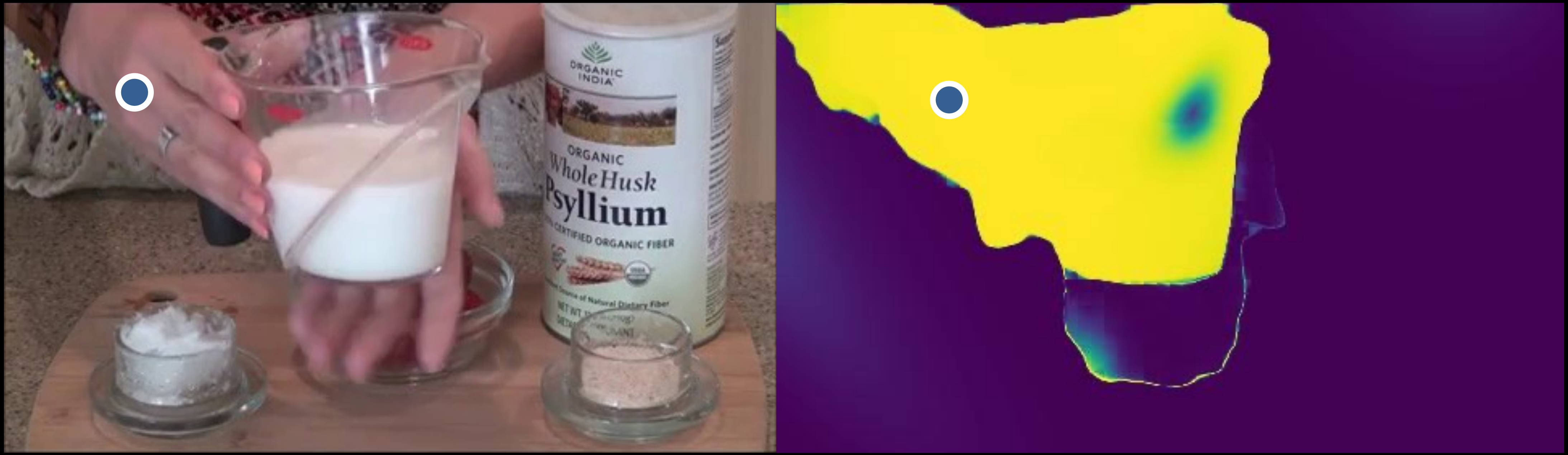
Responsibility

- How well does a hand explain the motion?
- Idea: one hand's responsibility for a pixel is how well that hand explains the pixel's motion compared to other hands and the background.

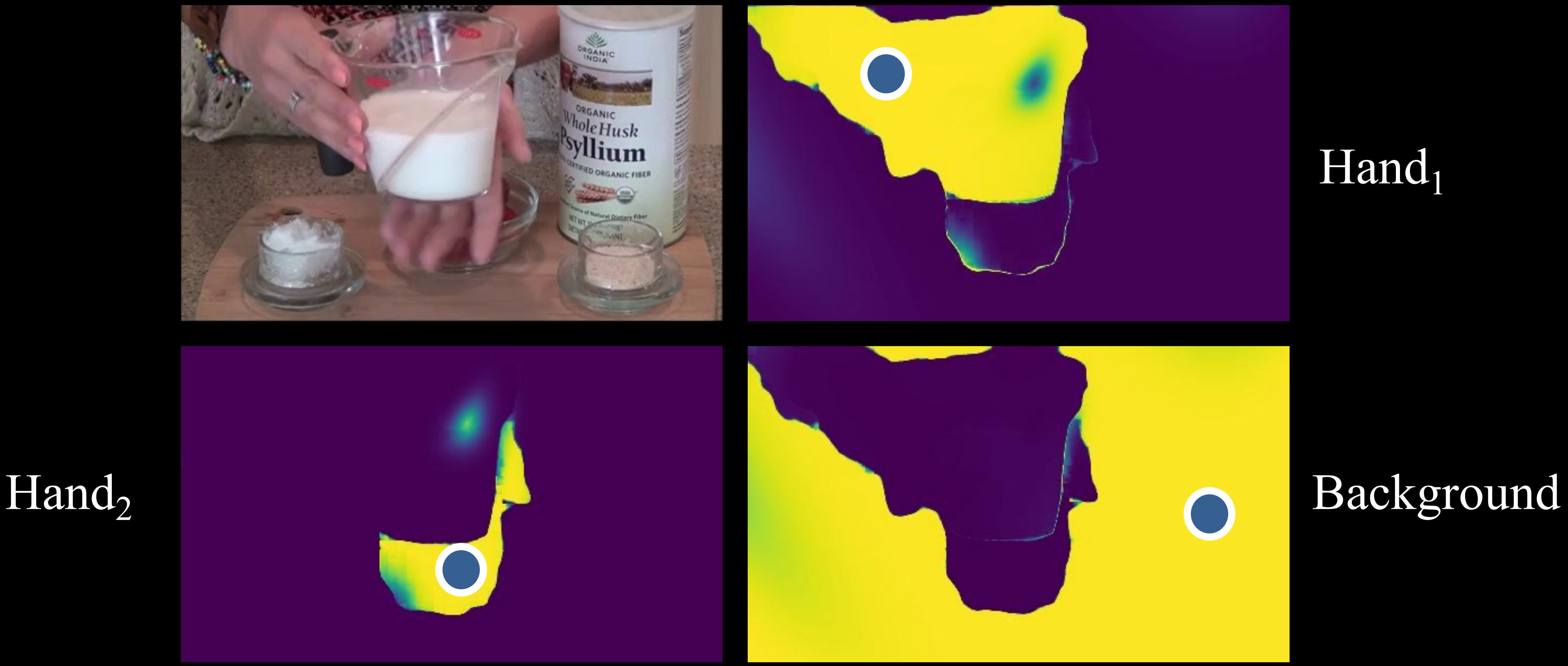


Responsibility

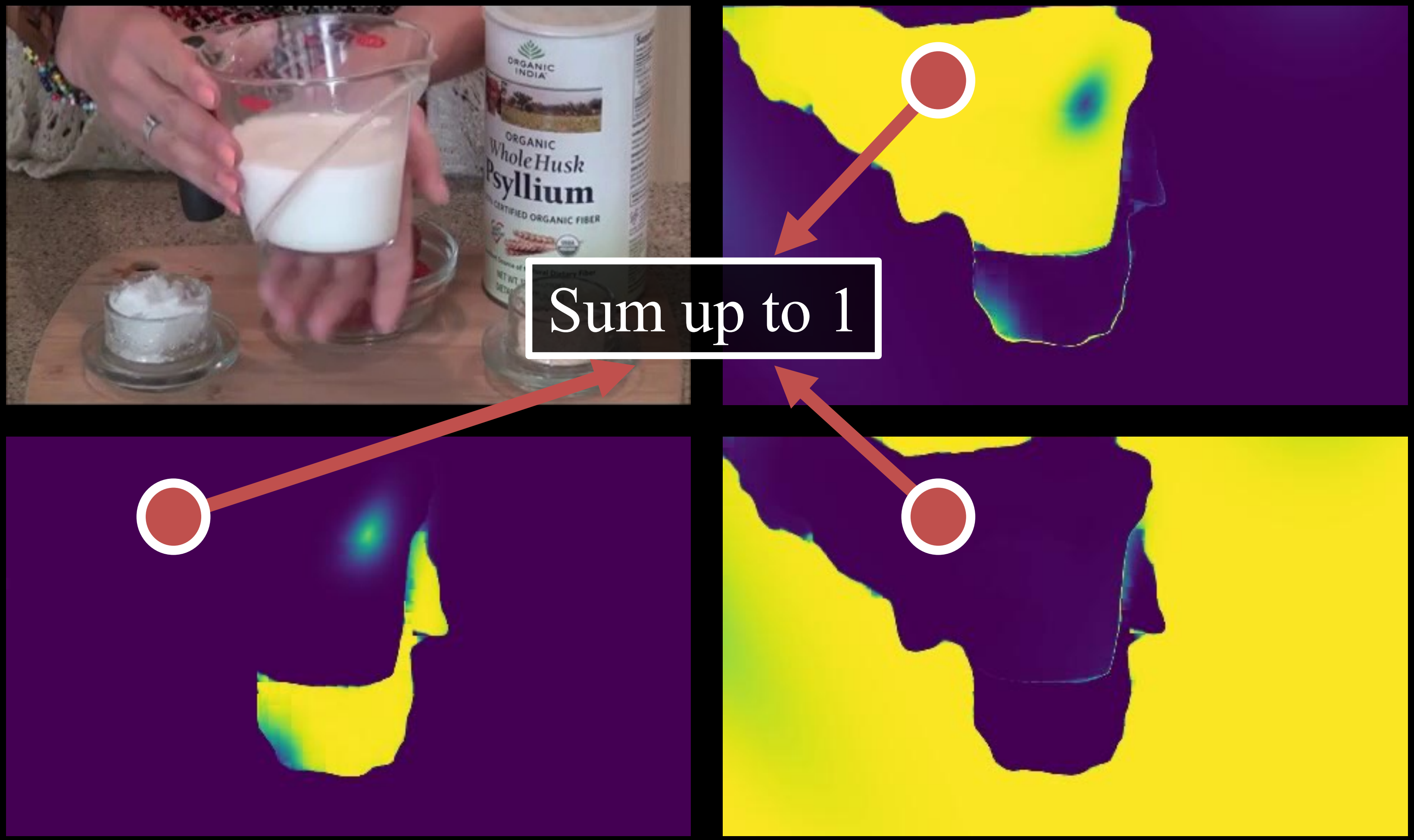
- How well does a hand explain the motion?
- Idea: one hand’s responsibility for a pixel is how well that hand explains the pixel’s motion compared to other hands and the background.



Responsibility



Responsibility



Sum up to 1

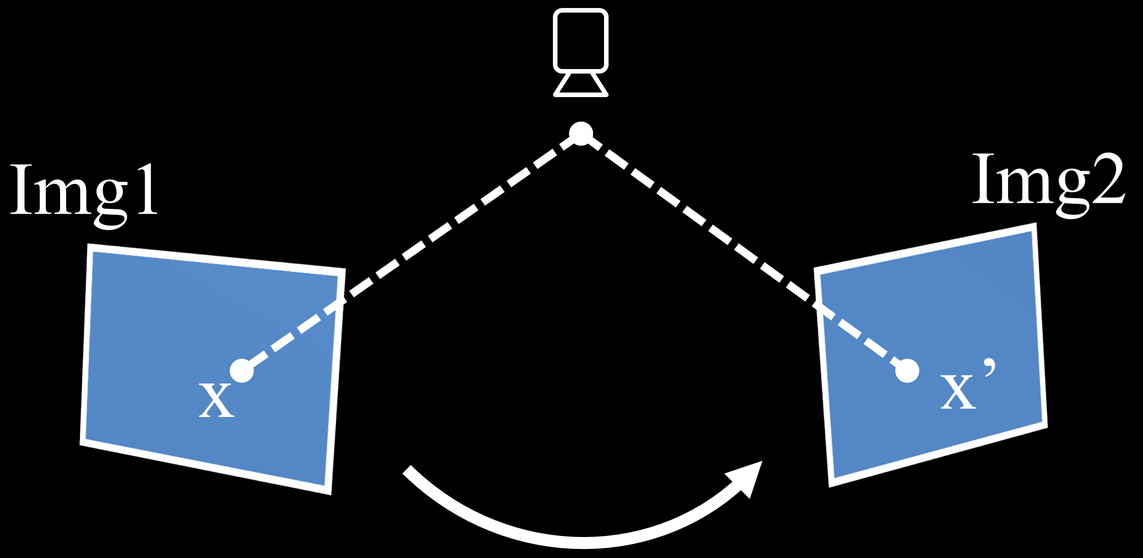
Hand$_1$

Hand$_2$

Background

Homography

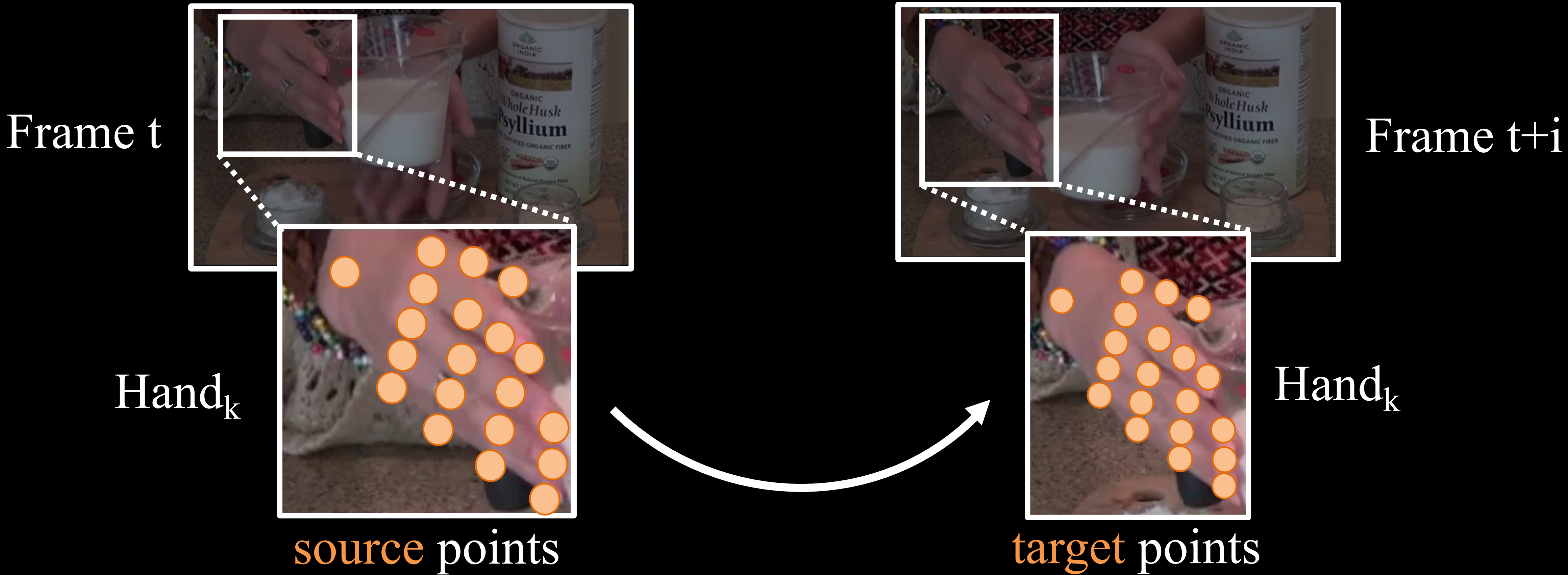
Planar homography relates the transformation between two planes.



Homography M

Generate Responsibility

- Fit a Homography M_k for hand$_k$ using source and target points.



Frame t

Hand$_k$

source points

Frame t+i

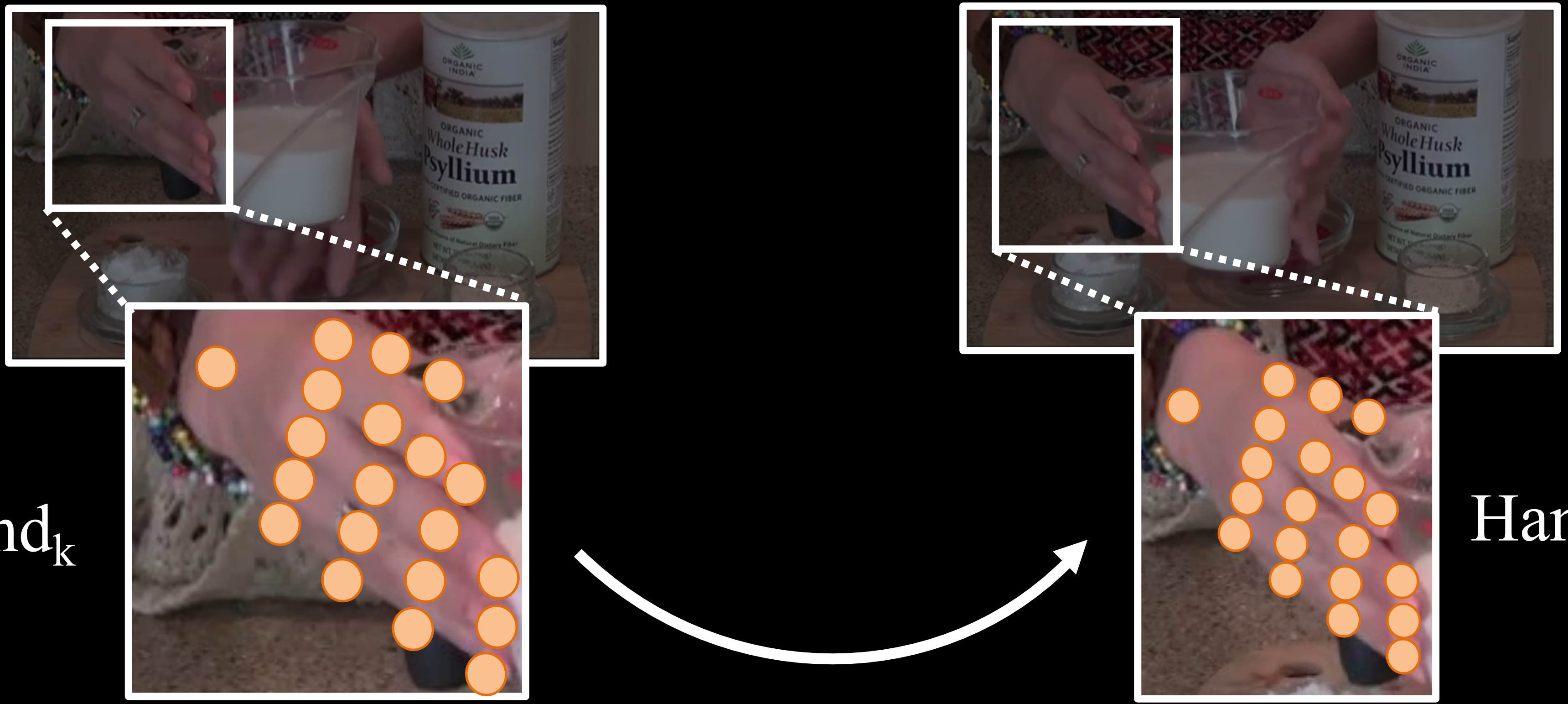
Hand$_k$

target points

Rong et al. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. ICCVW 2021.

Generate Responsibility

- Fit a Homography M_k for hand_k using source and target points.



Frame t

Hand_k

source points

M_k

target points

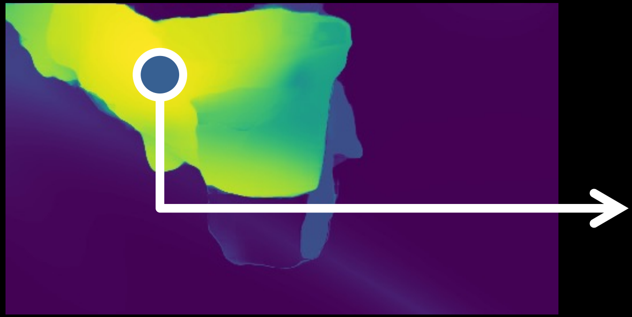
Frame t+i

Hand_k

Rong et al. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. ICCVW 2021.

Generate Responsibility

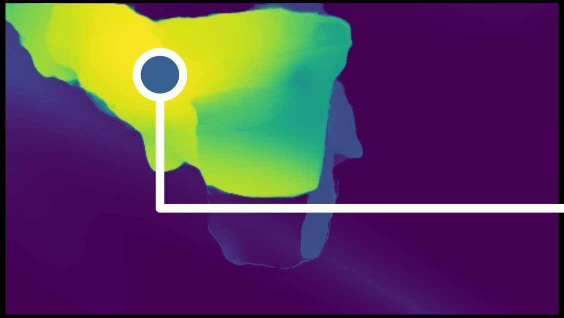
- Fit a Homography M_k for hand$_k$ using source and target points.
- Calculate responsibility using Softmax.



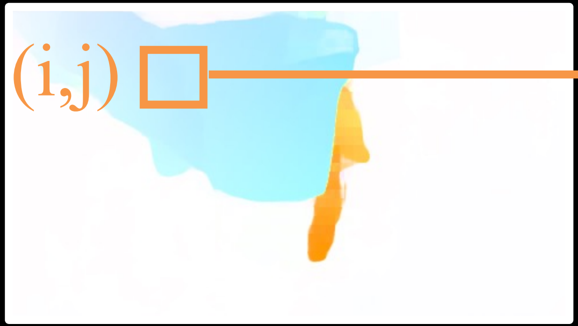
$$= \frac{\exp_t(-d_k(\boldsymbol{o}))}{\exp_t(-d_{BG}(\boldsymbol{o})) + \sum_{k'=1}^{N} \exp_t(-d_{k'}(\boldsymbol{o}))}$$

Generate Responsibility

- Fit a Homography M_k for $hand_k$ using source and target points.
- Calculate responsibility using Softmax.



$$= \frac{\exp_t(-d_k(\boldsymbol{o}))}{\exp_t(-d_{BG}(\boldsymbol{o})) + \sum_{k'=1}^{N} \exp_t(-d_{k'}(\boldsymbol{o}))}$$

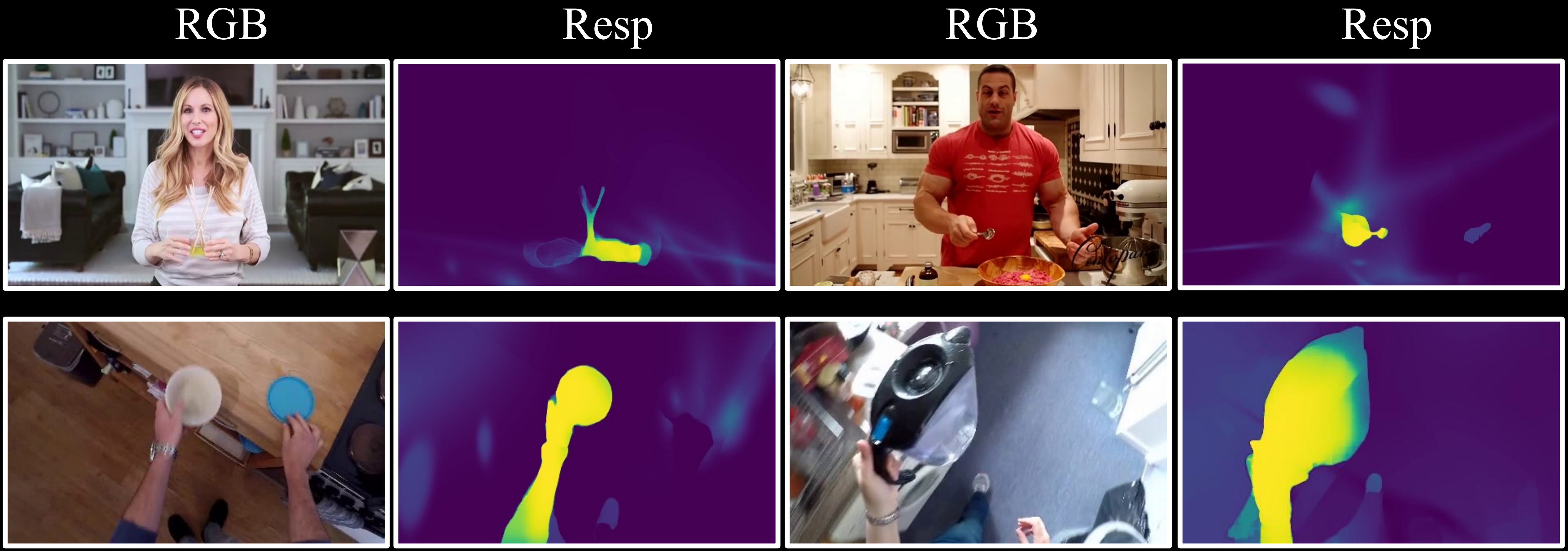


Optical Flow

Model Pred

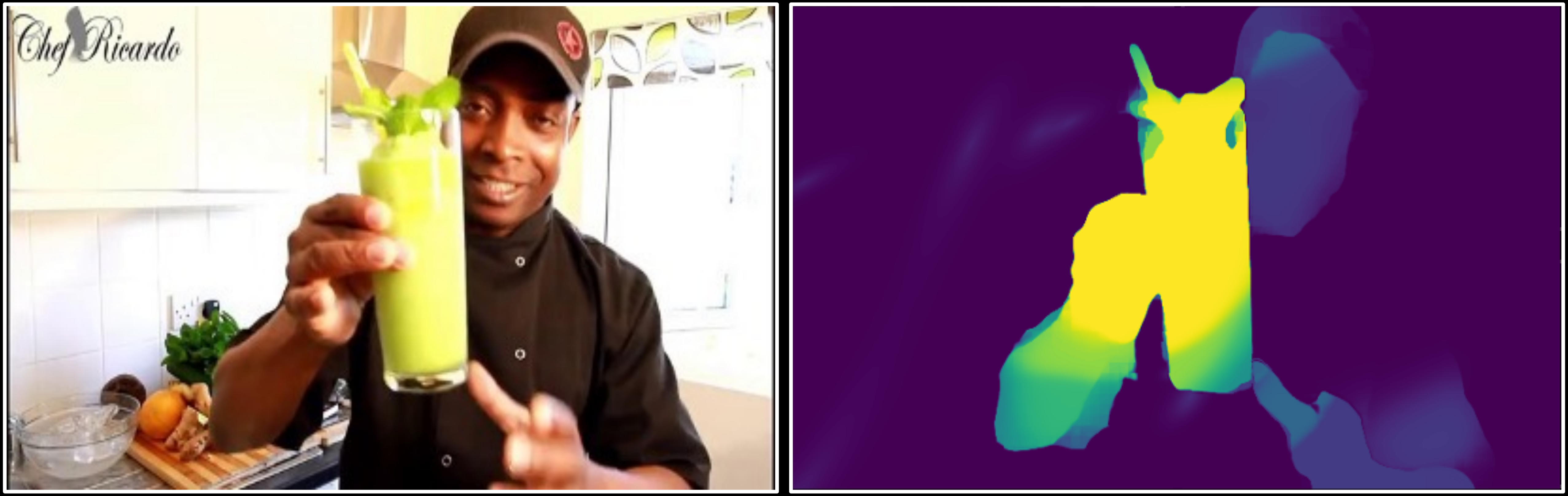
$$d_k(o) = ||[i,j]^T + \boxed{Oi_{,j}} - proj(\boxed{Mk[i,j,1]^T}) ||^2$$

Responsibility Visualization



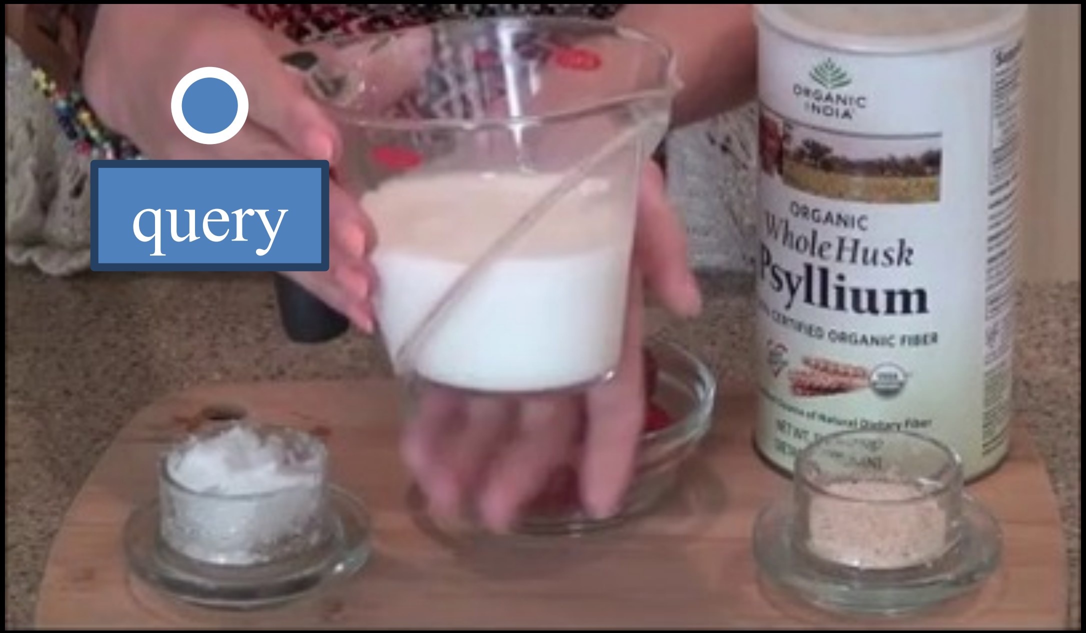
EPIC-Kitchens, Damen et al. 2018, 2020. / 100 Days of Hands, Shan et al. 2020.

Responsibility Visualization

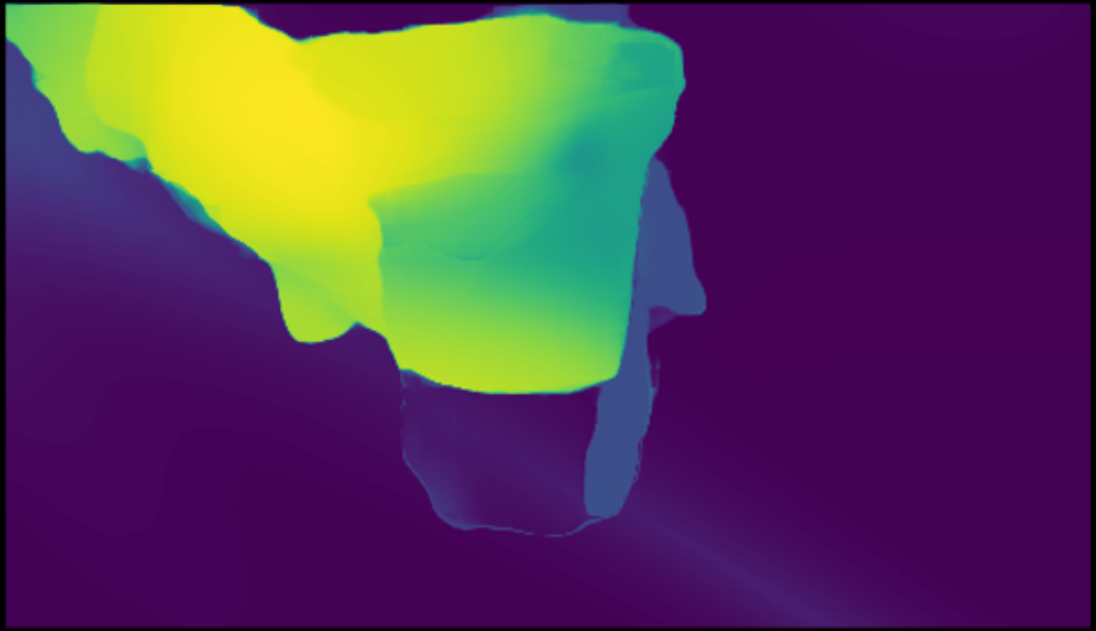


Training COHESIV Model

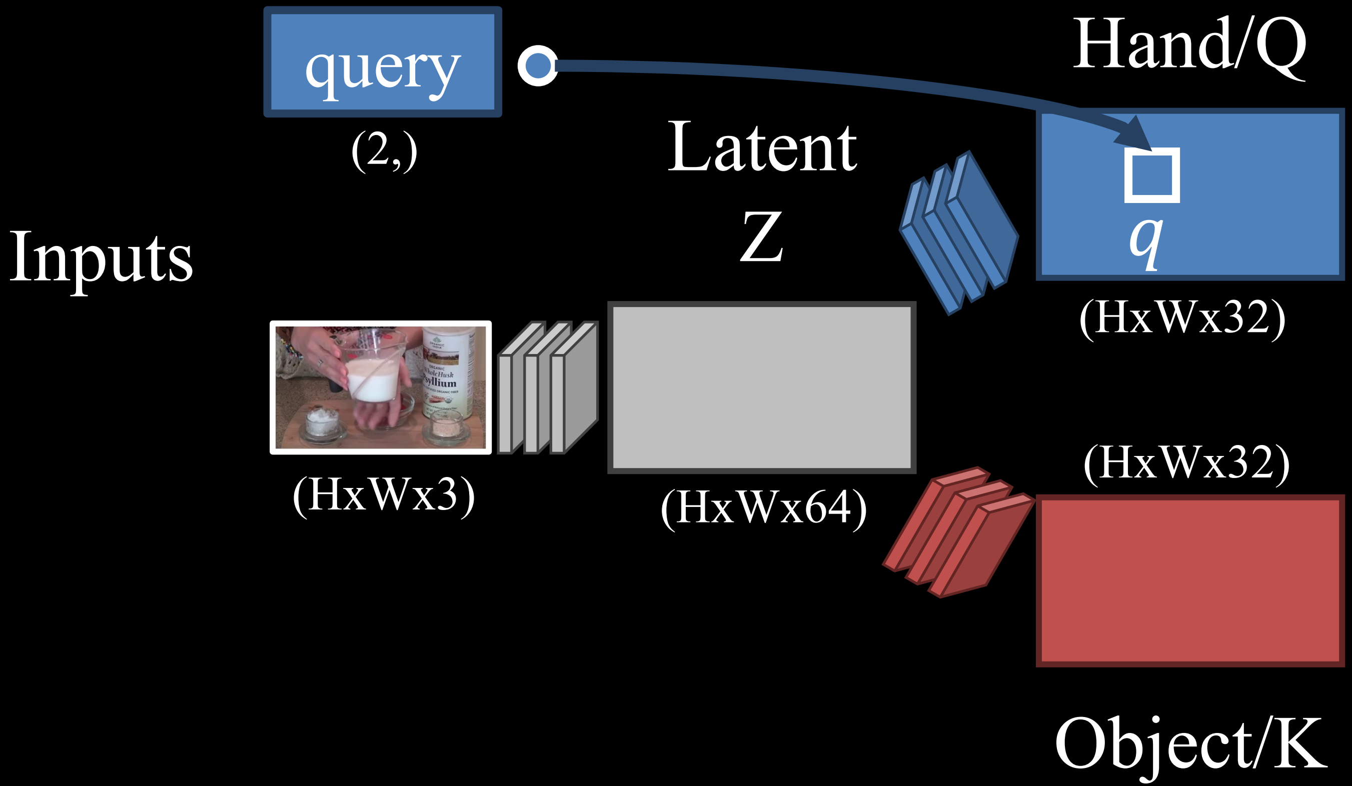
Input: Image + Query

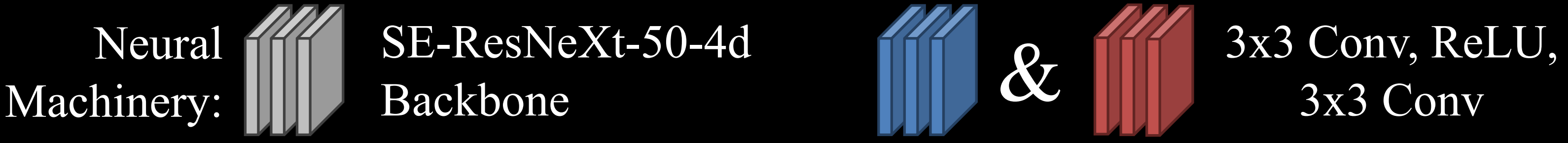


Desired Output

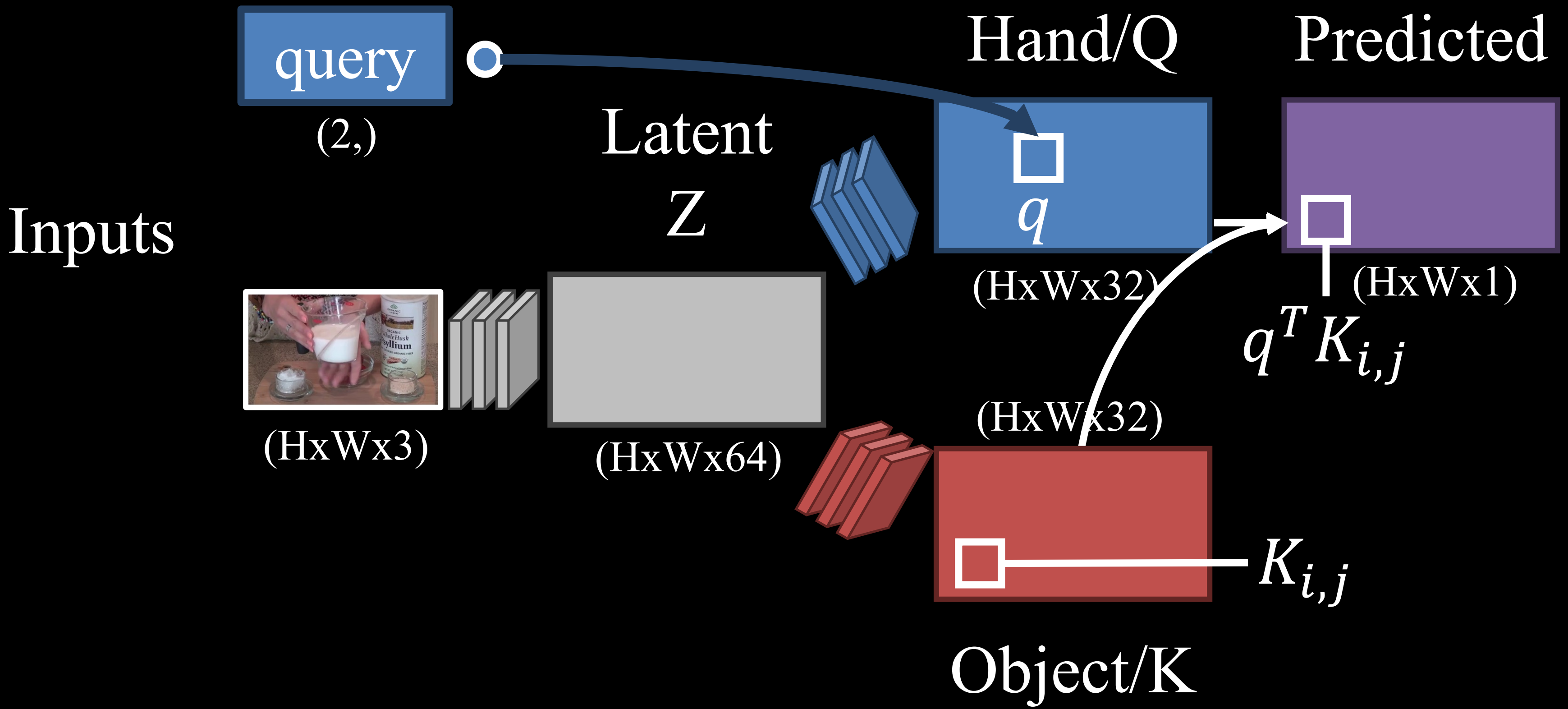


COHESIV Model – Learning





COHESIV Model – Learning



Inputs

query

(2,)

(HxWx3)

Latent

Z

(HxWx64)

Hand/Q

q

(HxWx32)

Predicted

(HxWx1)

$q^T K_{i,j}$

(HxWx32)

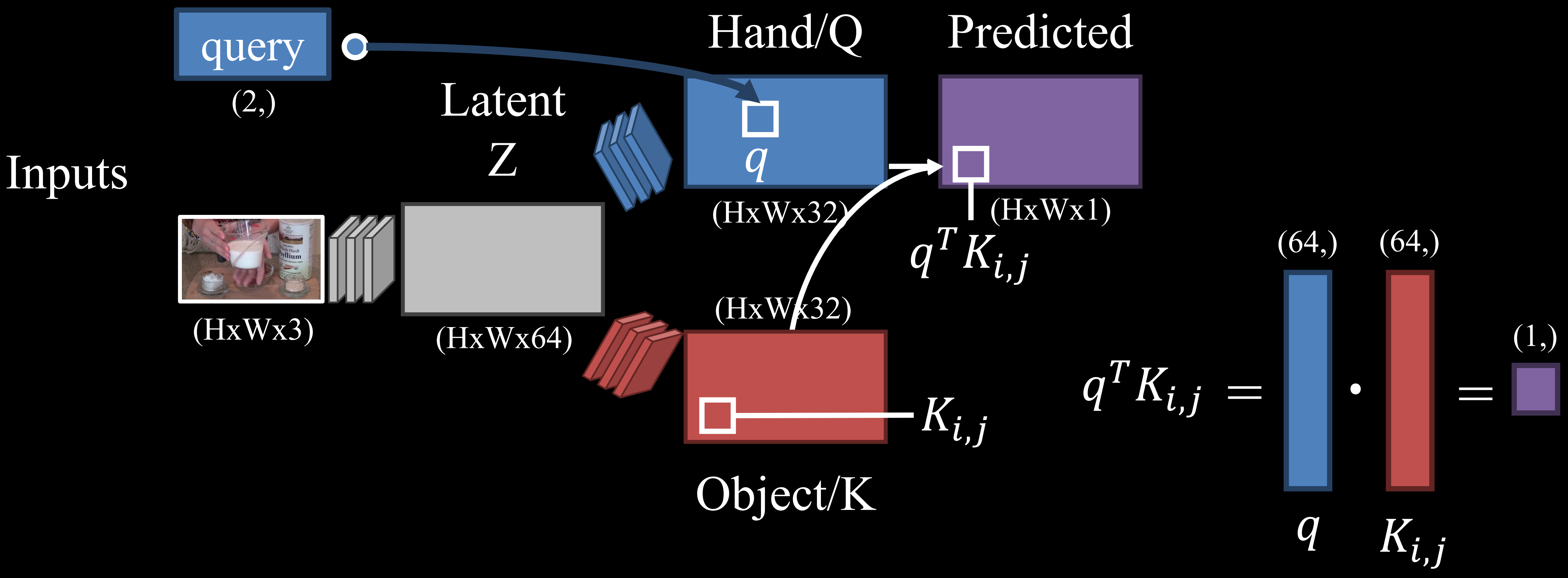
$K_{i,j}$

Object/K

Neural Machinery: SE-ResNeXt-50-4d Backbone

& 3x3 Conv, ReLU, 3x3 Conv

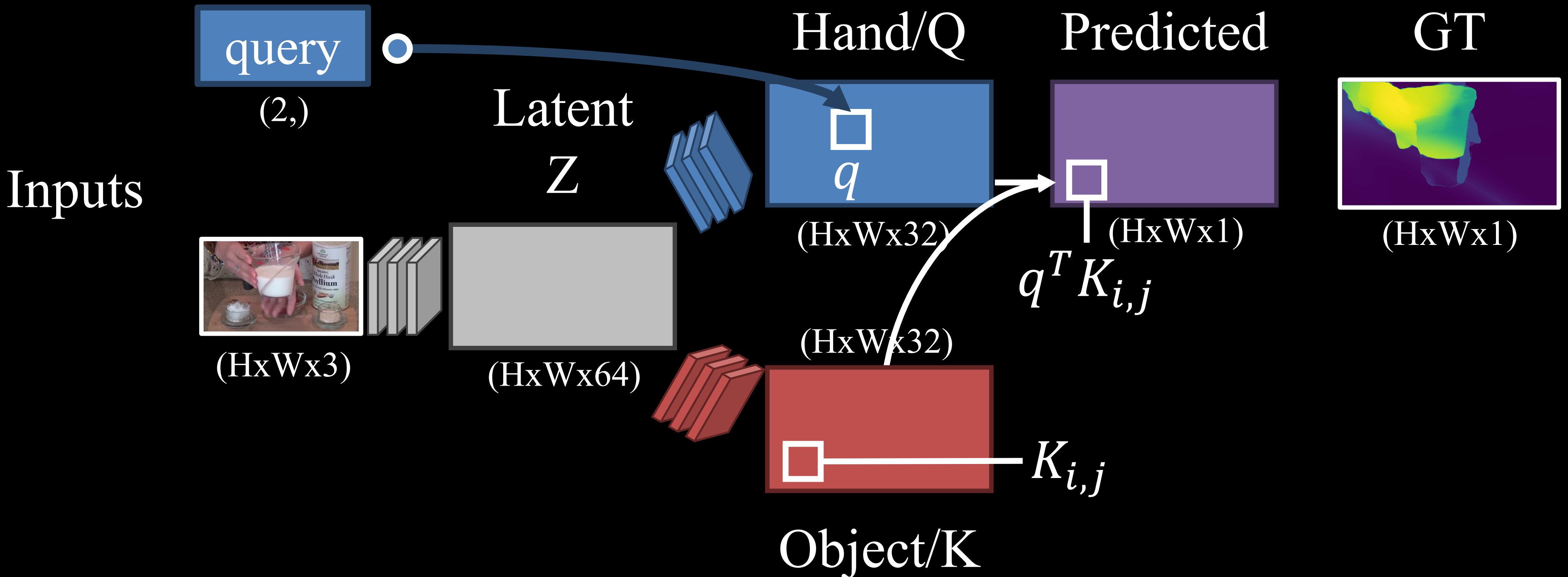
COHESIV Model – Learning



Neural Machinery: SE-ResNeXt-50-4d Backbone

& 3x3 Conv, ReLU, 3x3 Conv

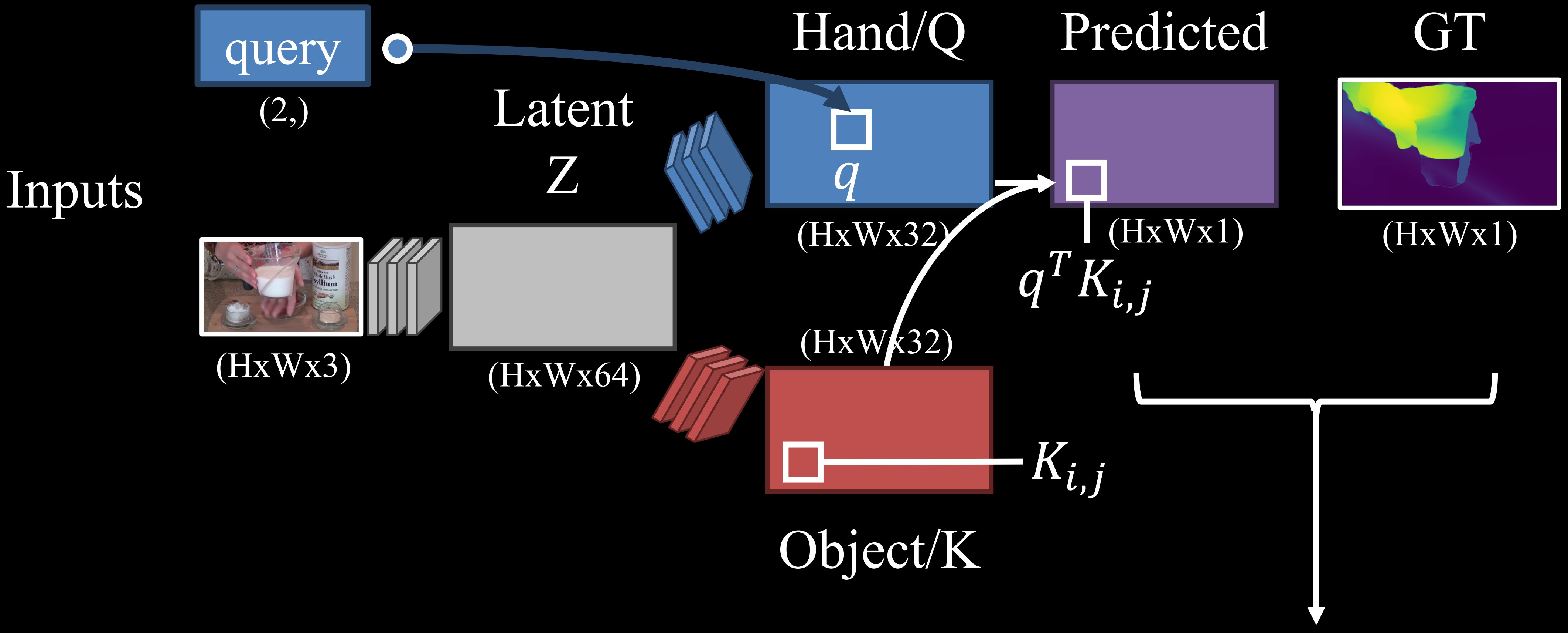
COHESIV Model – Learning



Neural Machinery: SE-ResNeXt-50-4d Backbone

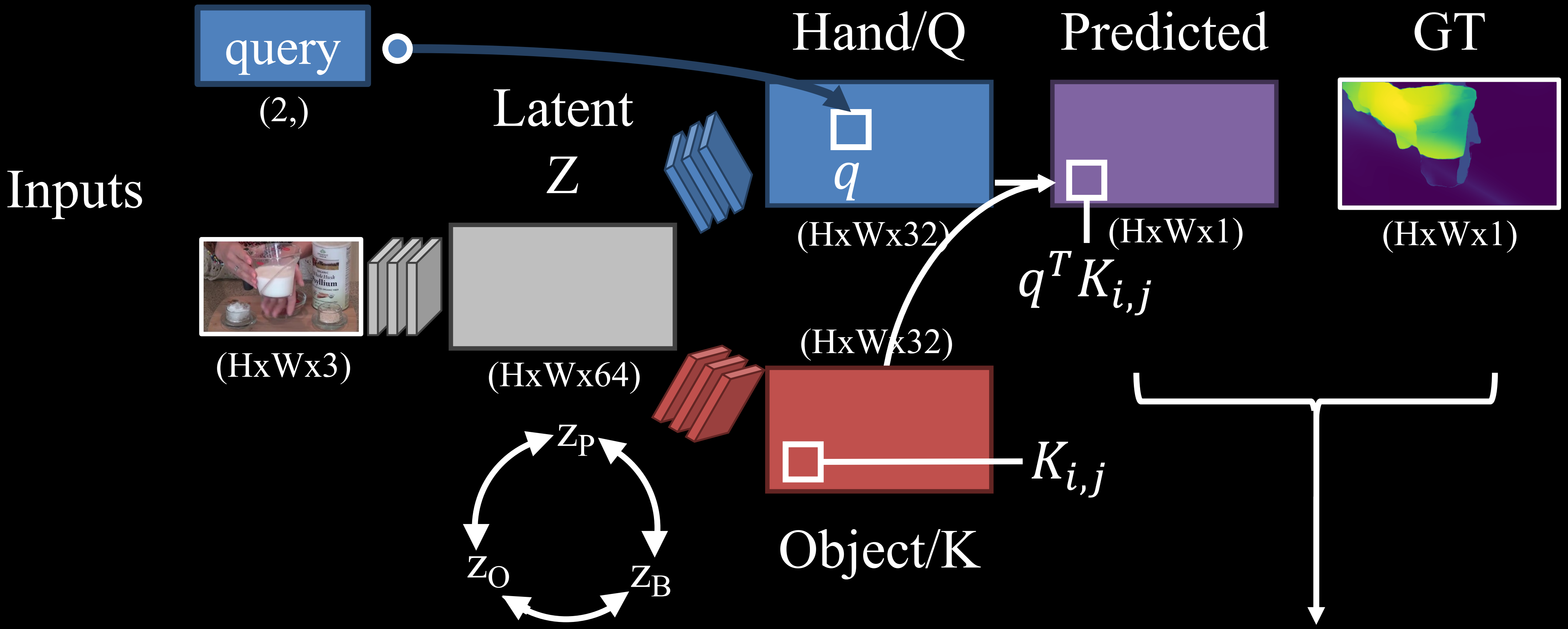
& 3x3 Conv, ReLU, 3x3 Conv

COHESIV Model – Learning



Smooth L1 Loss

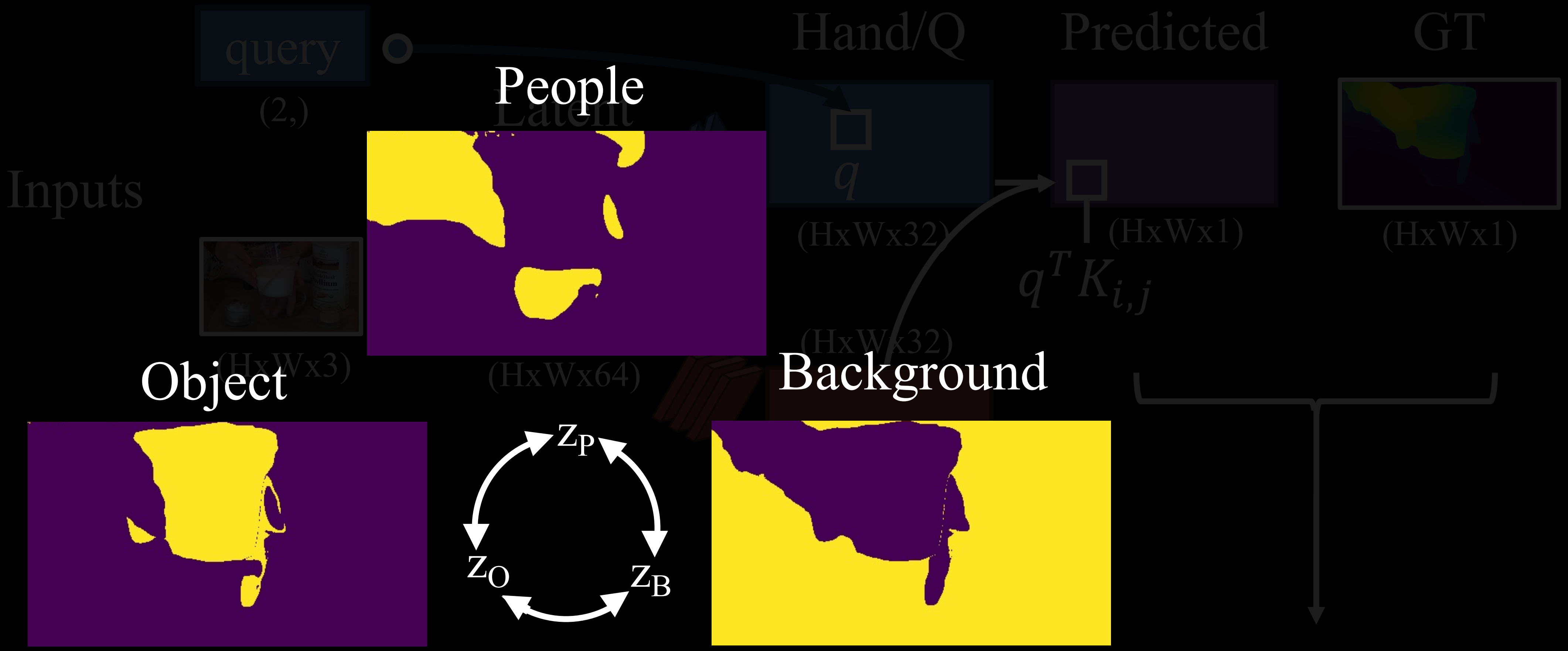
COHESIV Model – Learning



Contrastive Loss

Smooth L1 Loss

COHESIV Model – Learning

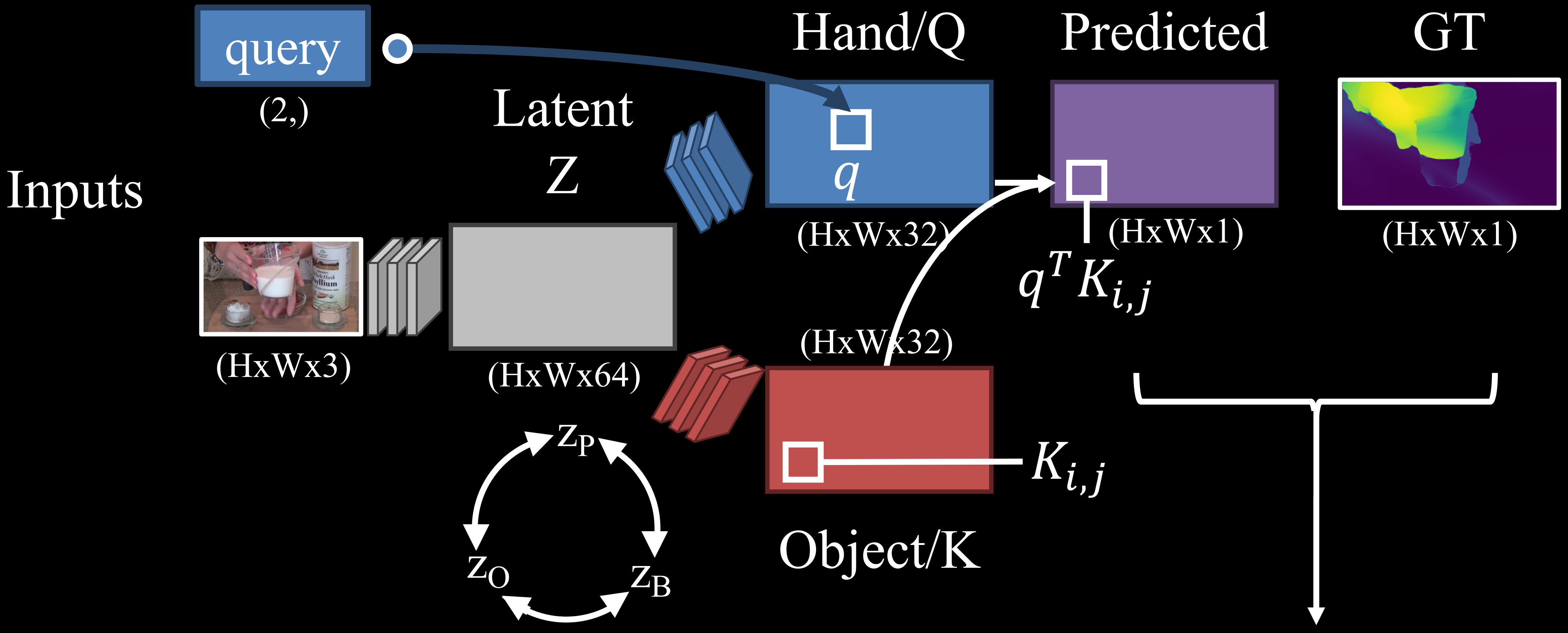


Contrastive Loss

Smooth L1 Loss

Ternaus et al. TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. Arxiv 2018.
Van Gansbeke et al. Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals. ICCV 2021.

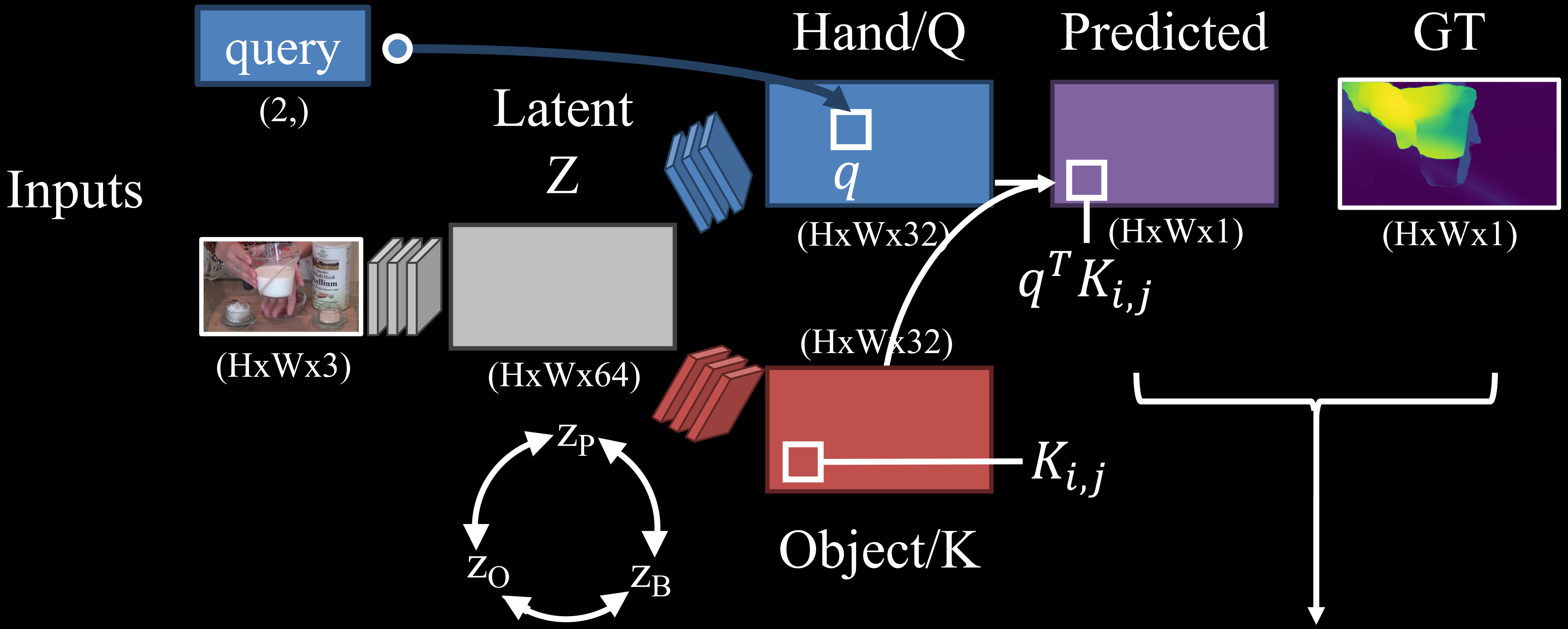
COHESIV Model – Learning



Contrastive Loss

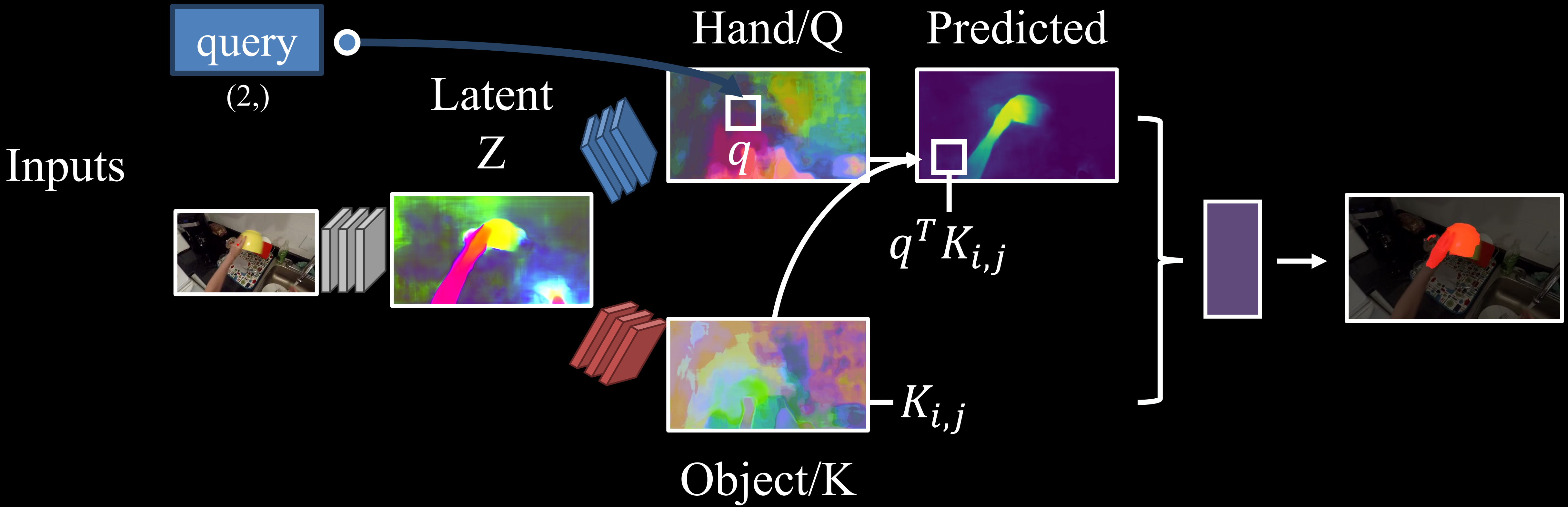
Smooth L1 Loss

COHESIV Model – Learning



Total Loss = Contrastive Loss + Smooth L1 Loss

COHESIV Model – Inference



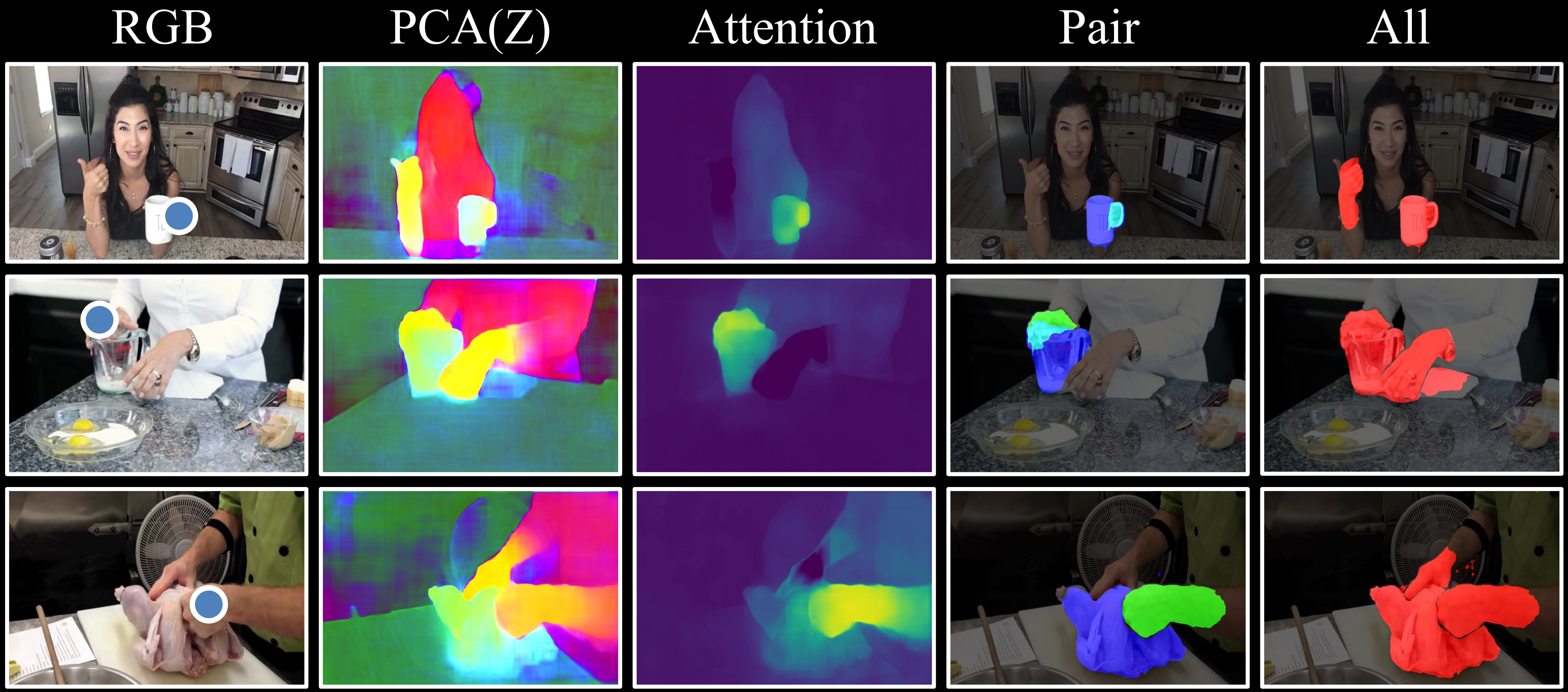
- Z has some per-pixel category-level information
- Q, K enable hand-specific information

Video Datasets

	100DOH	EPICK
#clips	88,153	28,982
#train	97,312	23,212
#val	482	438
#test (eval)	1,124	1,170

EPIC-Kitchens, Damen et al. 2018, 2020. / 100 Days of Hands, Shan et al. 2020.

Qualitative Results (100DOH)



RGB

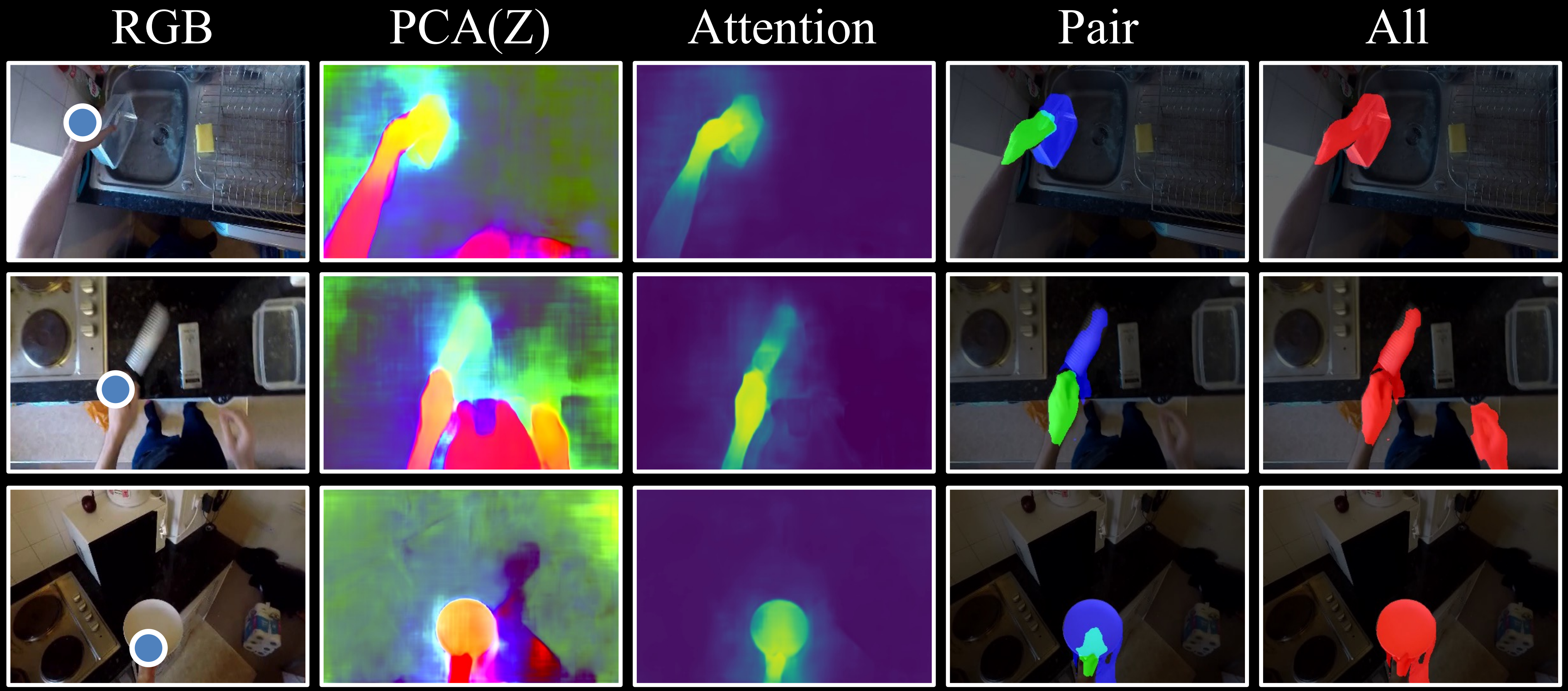
PCA(Z)

Attention

Pair

All

Qualitative Results (EPICK)



RGB

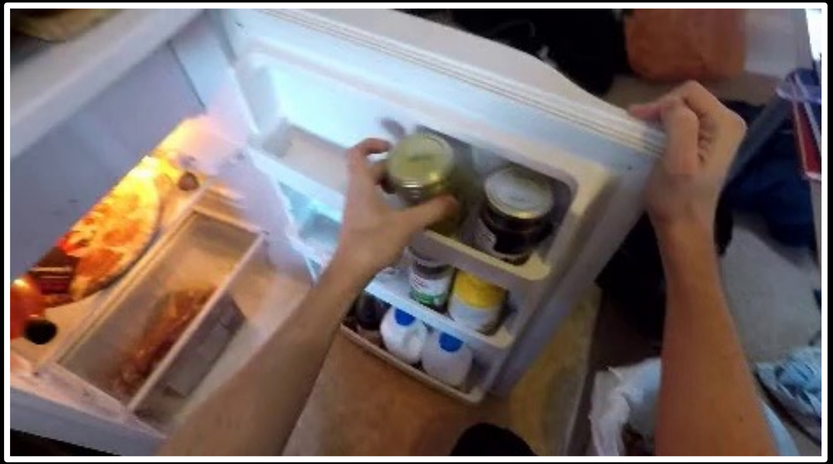
PCA(Z)

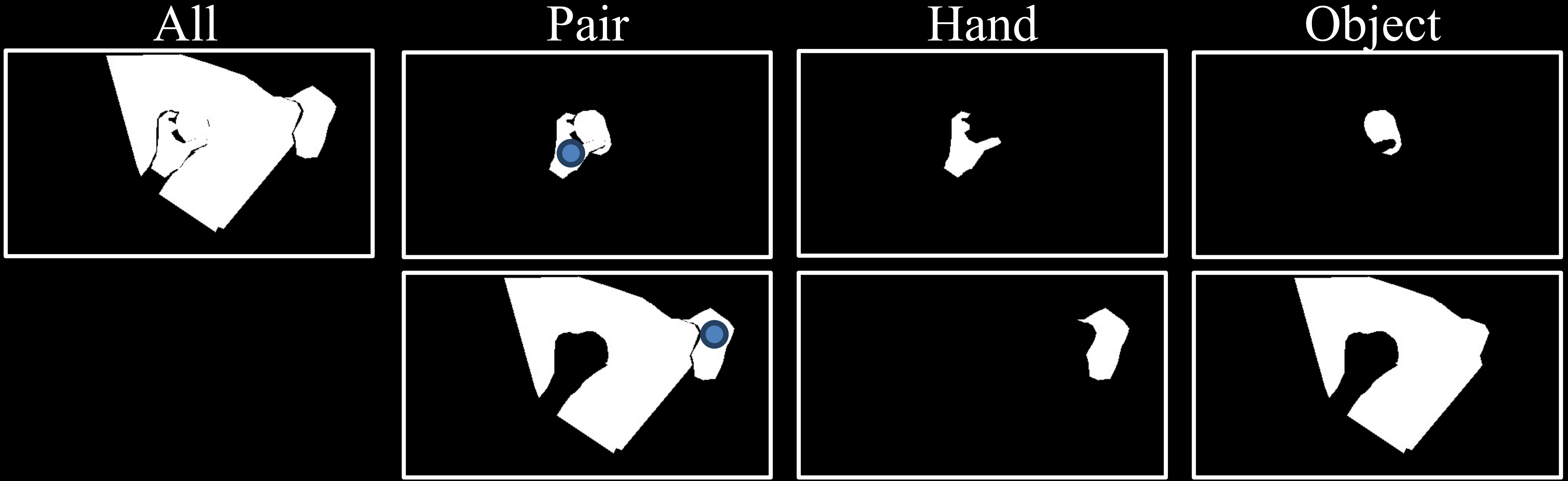
Attention

Pair

All

Evaluation Tasks





Quantitative Results - Baselines

Metric: mean intersection over union (mIoU) compared to GT.

	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV								
Flow/RAFT								
Saliency								
Superv. Box								

Supervised Box: Dandan Shan et al. CVPR 2020. / RAFT: Zachary Teed and Jia Deng, ECCV 2020. / Saliency: Ting Zhao and Xiangqian Wu, CVPR 2019.

Quantitative Results - Baselines

Metric: mean intersection over union (mIoU) compared to GT.

	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	51.4	46.1	53.6	29.1	42.0	41.2	59.4	19.6
Flow/RAFT								
Saliency								
Superv. Box								

Supervised Box: Dandan Shan et al. CVPR 2020. / RAFT: Zachary Teed and Jia Deng, ECCV 2020. / Saliency: Ting Zhao and Xiangqian Wu, CVPR 2019.

Quantitative Results - Baselines

Metric: mean intersection over union (mIoU) compared to GT.

	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	51.4	46.1	53.6	29.1	42.0	41.2	59.4	19.6
Flow/RAFT	29.3	21.5	12.9	12.1	15.4	11.9	6.2	6.6
Saliency	25.2	20.1	8.6	17.0	21.6	15.9	6.0	11.7
Superv. Box								

Supervised Box: Dandan Shan et al. CVPR 2020. / RAFT: Zachary Teed and Jia Deng, ECCV 2020. / Saliency: Ting Zhao and Xiangqian Wu, CVPR 2019.

Quantitative Results - Baselines

Metric: mean intersection over union (mIoU) compared to GT.

	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	51.4	46.1	53.6	29.1	42.0	41.2	59.4	19.6
Flow/RAFT	29.3	21.5	12.9	12.1	15.4	11.9	6.2	6.6
Saliency	25.2	20.1	8.6	17.0	21.6	15.9	6.0	11.7
Superv. Box	56.9	47.0	56.5	34.9	54.3	44.8	53.8	34.4

Supervised Box: Dandan Shan et al. CVPR 2020. / RAFT: Zachary Teed and Jia Deng, ECCV 2020. / Saliency: Ting Zhao and Xiangqian Wu, CVPR 2019.

Quantitative Results - Baselines

Metric: mean intersection over union (mIoU) compared to GT.

	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	51.4	46.1	53.6	29.1	42.0	41.2	59.4	19.6
Flow/RAFT	29.3	21.5	12.9	12.1	15.4	11.9	6.2	6.6
Saliency	25.2	20.1	8.6	17.0	21.6	15.9	6.0	11.7
Superv. Box	56.9	47.0	56.5	34.9	54.3	44.8	53.8	34.4

Supervised Box: Dandan Shan et al. CVPR 2020. / RAFT: Zachary Teed and Jia Deng, ECCV 2020. / Saliency: Ting Zhao and Xiangqian Wu, CVPR 2019.

Quantitative Results - Ablations

Metric: mean intersection over union (mIoU) compared to GT.

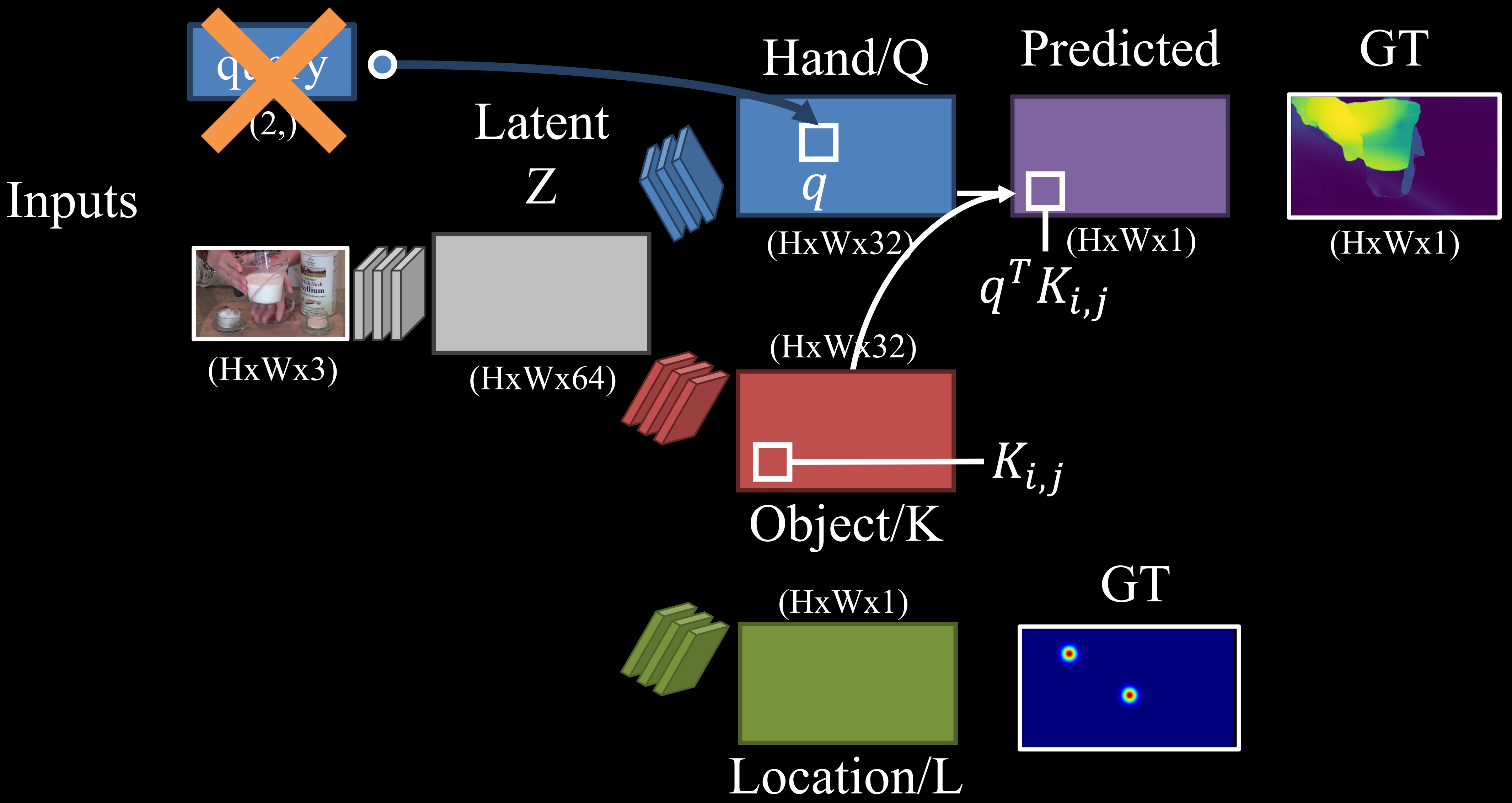
	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	51.4	46.1	53.6	29.1	42.0	41.2	59.4	19.6
Attention-Only								
Embedding-Only								

Quantitative Results - Ablations

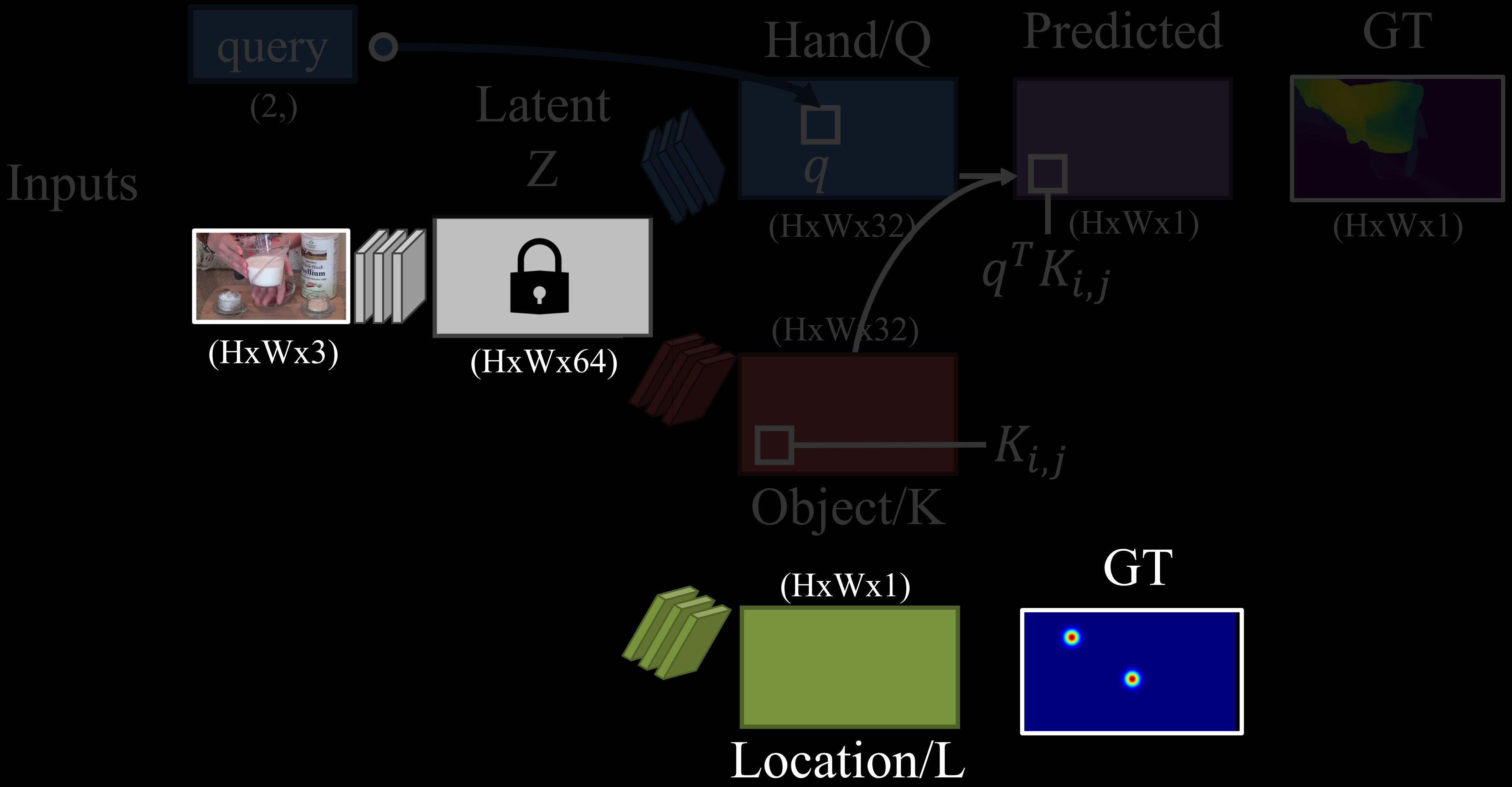
Metric: mean intersection over union (mIoU) compared to GT.

	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	51.4	46.1	53.6	29.1	42.0	41.2	59.4	19.6
Attention-Only	42.8	40.0	-	-	38.1	37.8	-	-
Embedding-Only	25.7	18.3	13.2	22.9	30.0	20.8	24.6	14.4

Extension: hand location prediction branch



Extension: hand location prediction branch



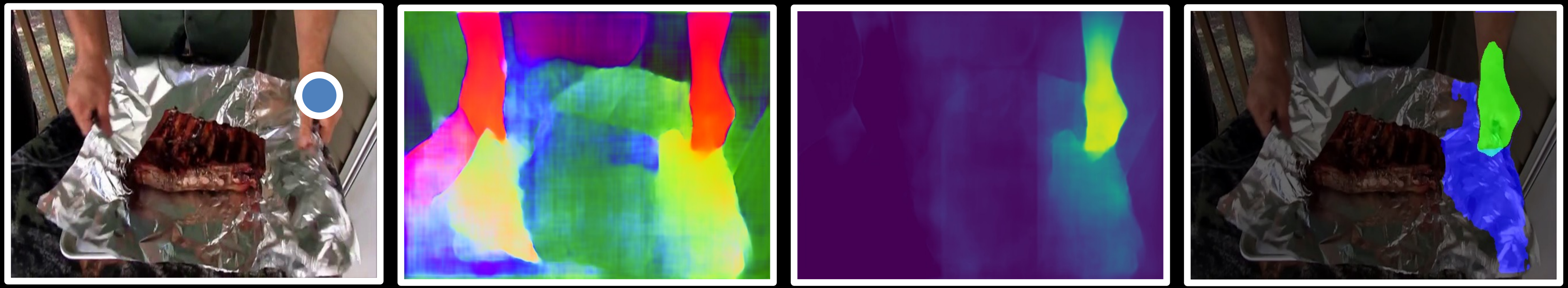
Quantitative Results - Ablations

Metric: mean intersection over union (mIoU) compared to GT.

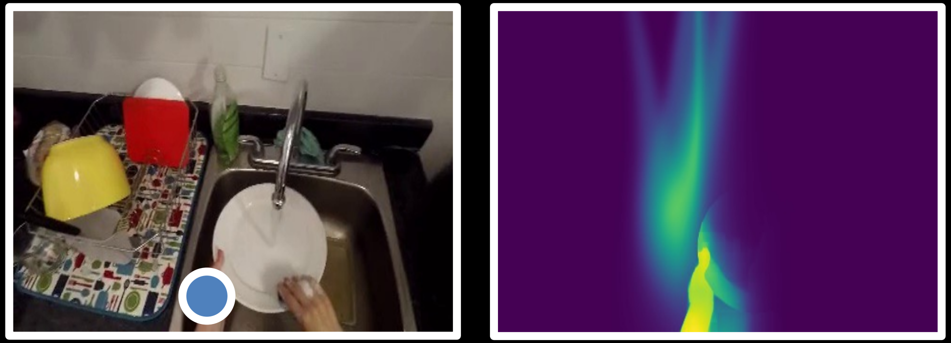
	100DOH				EPICK			
	All	Pair	Hand	Object	All	Pair	Hand	Object
COHESIV	51.4	46.1	53.6	29.1	42.0	41.2	59.4	19.6
Attention-Only	42.8	40.0	-	-	38.1	37.8	-	-
Embedding-Only	25.7	18.3	13.2	22.9	30.0	20.8	24.6	14.4
w/ Predicted location	47.7	42.8	47.8	28.1	40.0	38.6	55.1	19.4

Outstanding Issues

Large non-rigid object:

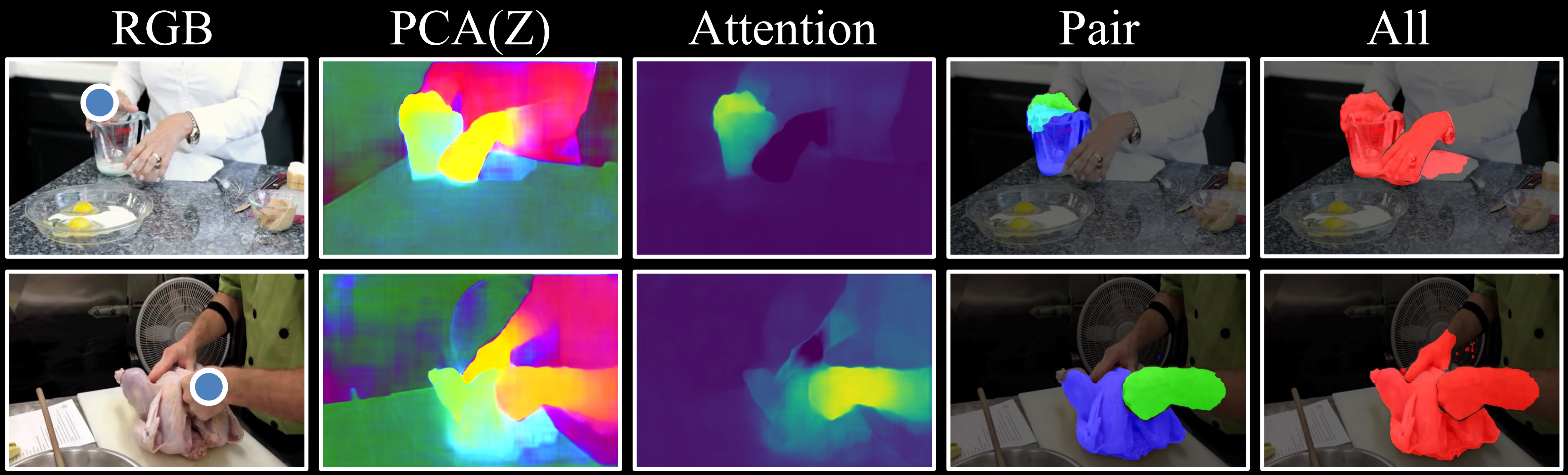


Bad responsibility:



Summary

- Responsibility map
- Hand-queried contact region segmentation
- **COHESIV**: contrastive + attention



RGB PCA(Z) Attention Pair All