Embodied Visual Recognition

Katerina Fragkiadaki

Carnegie Mellon University







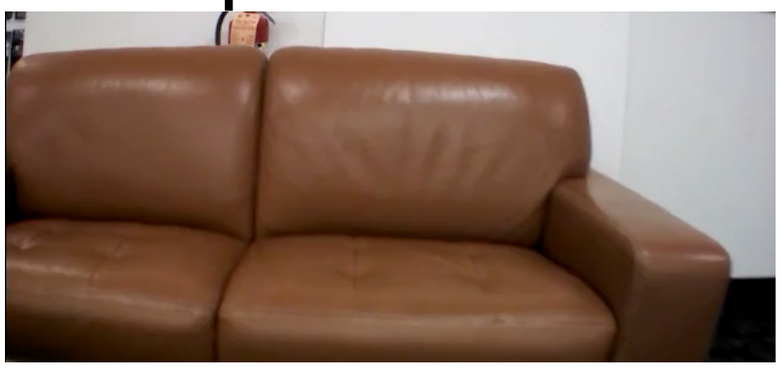

Internet Vision



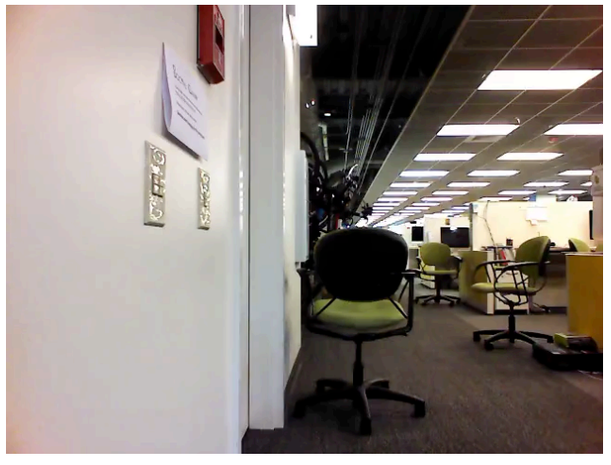
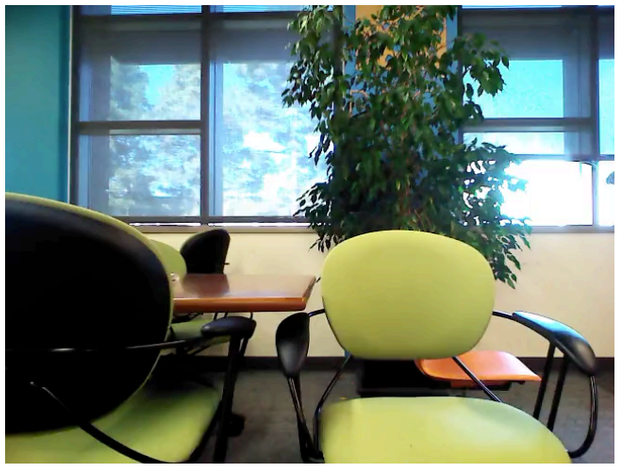


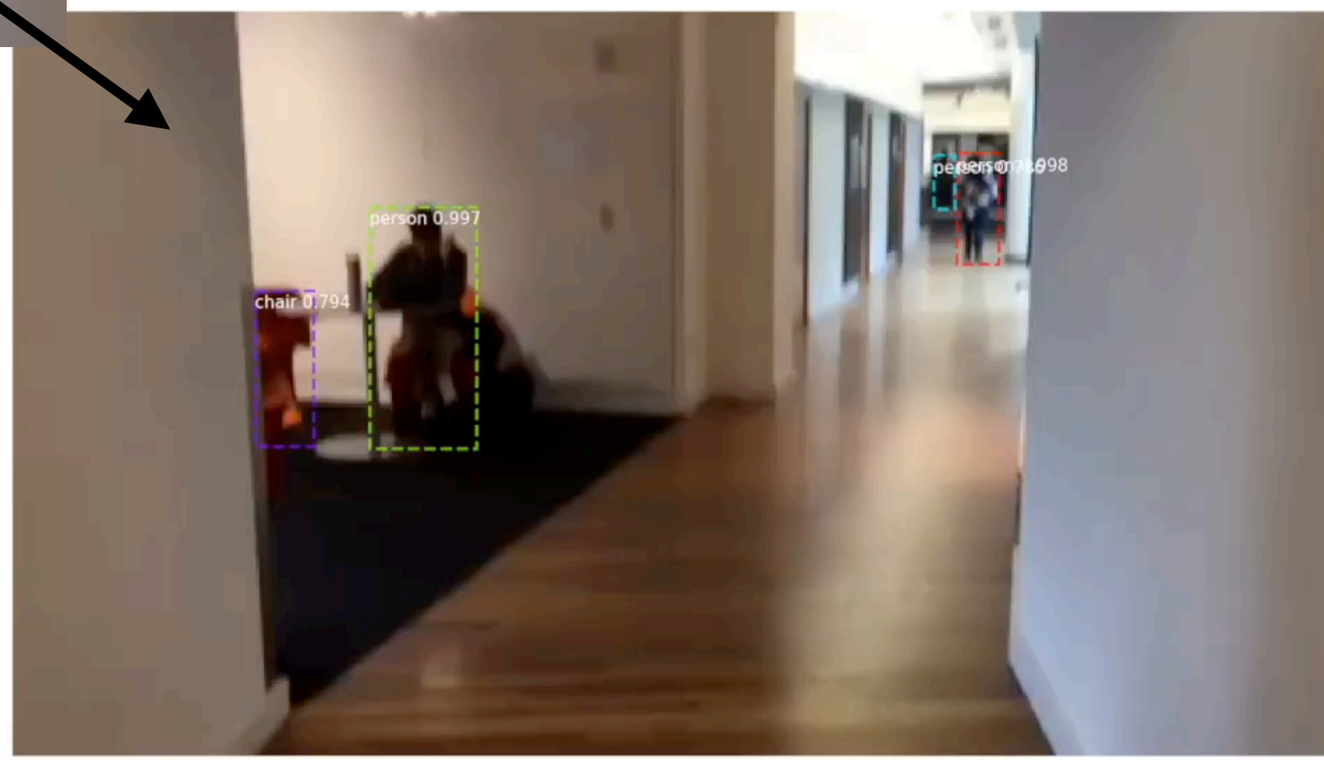


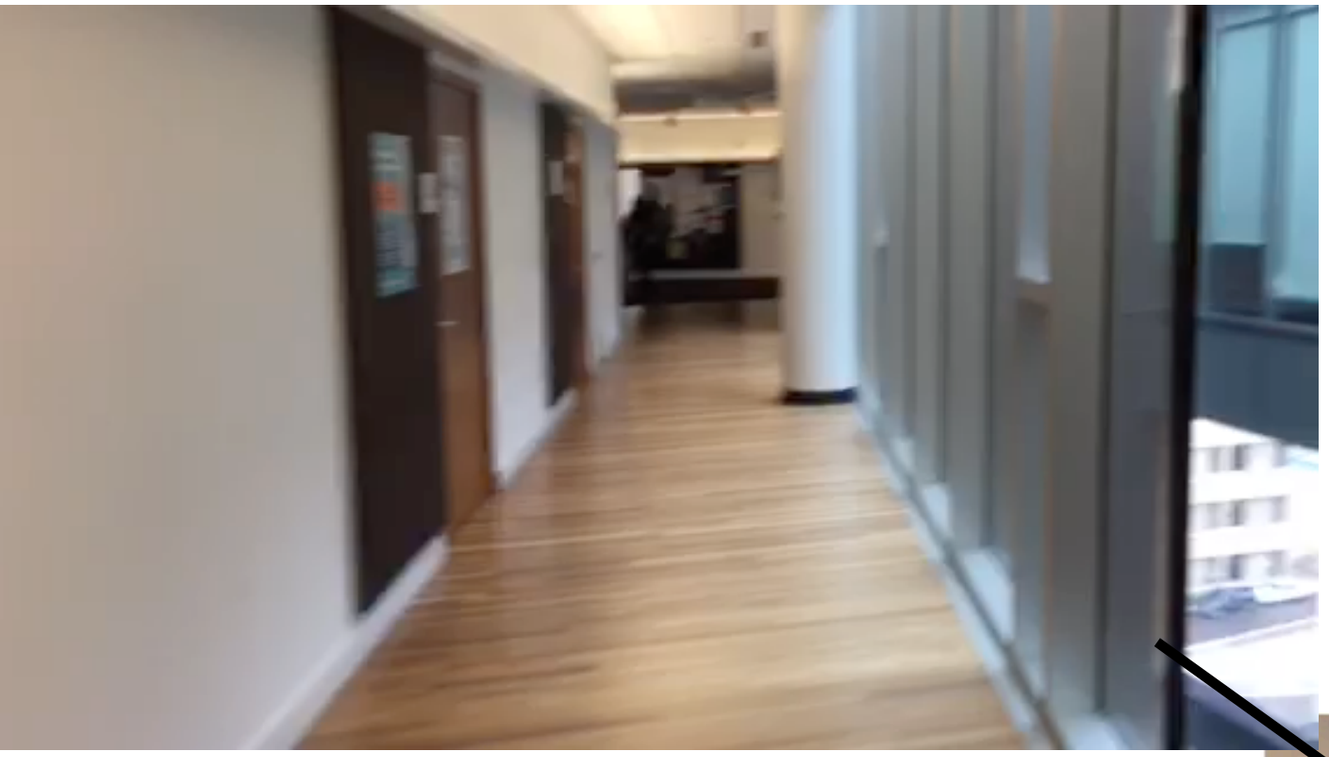

Mobile (Embodied) Computer Vision

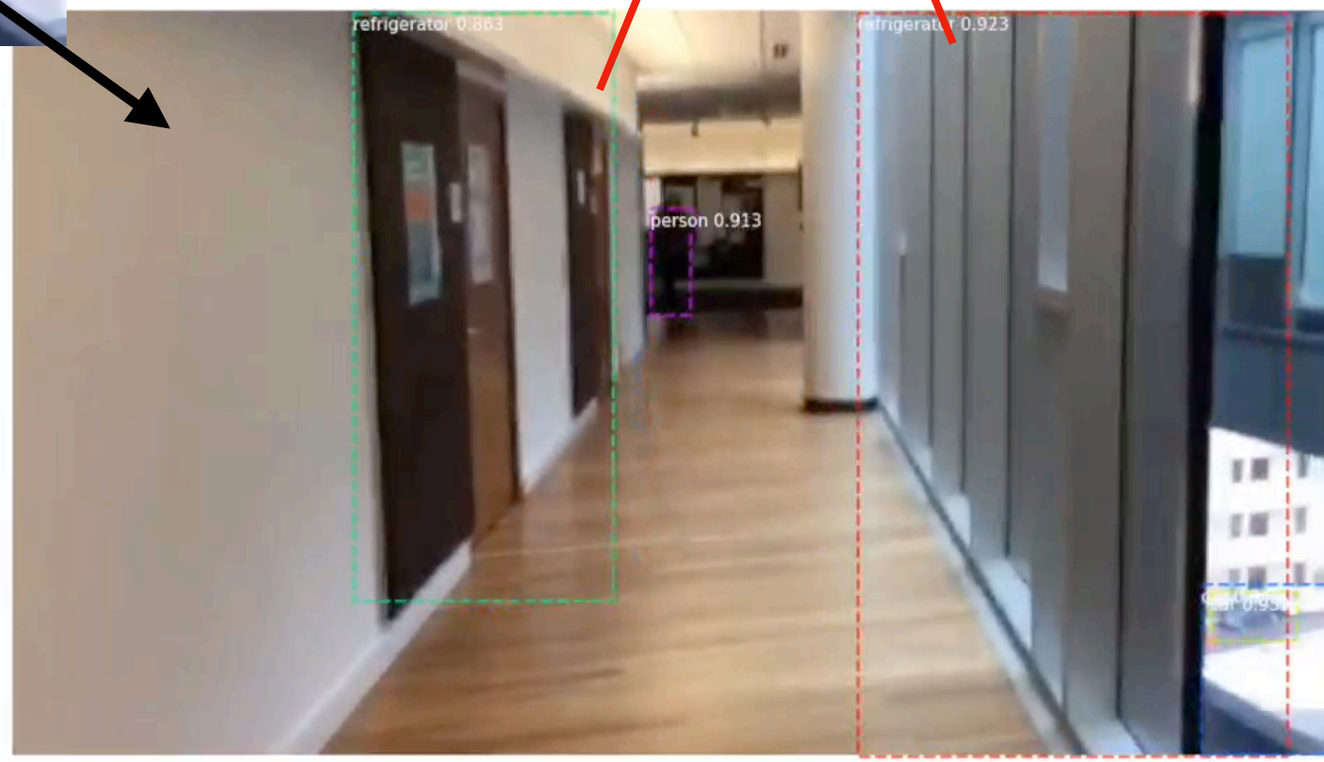


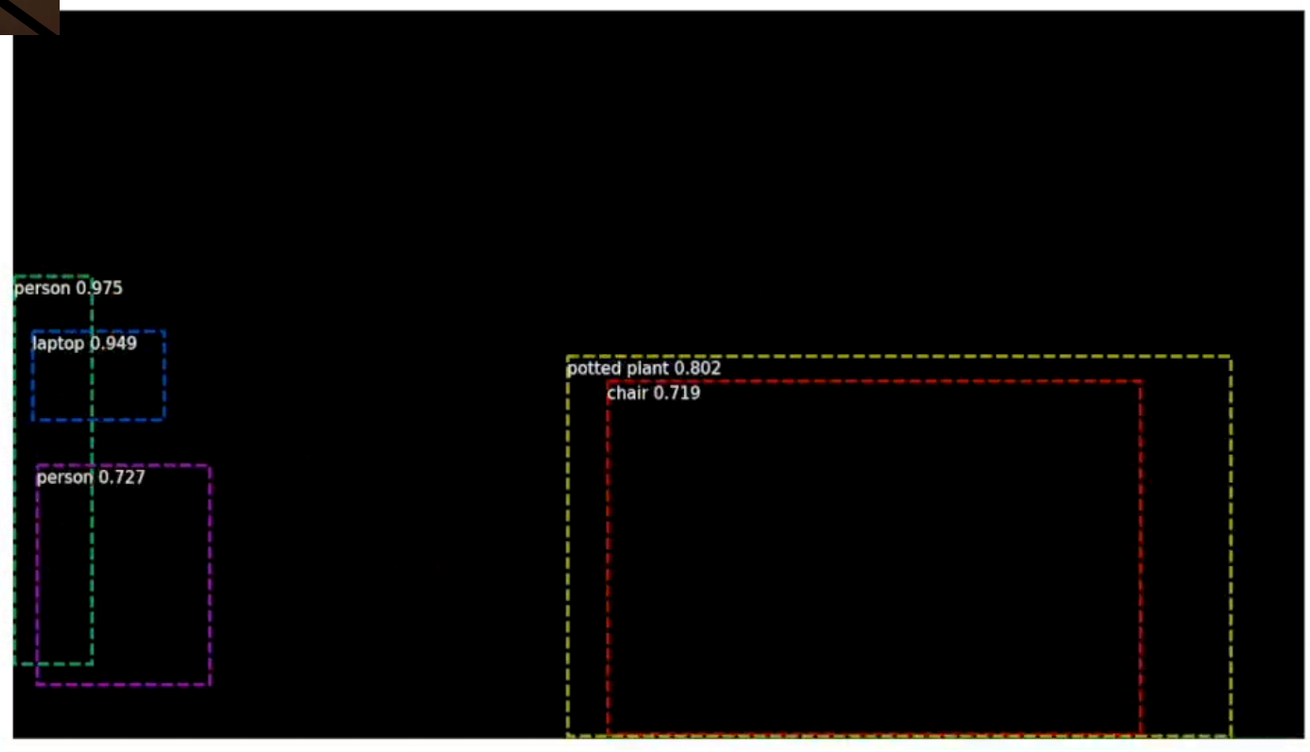




person 0.997
chair 0.794



refrigerators ?




person 0.975
laptop 0.949
person 0.727
potted plant 0.802
chair 0.719

2D CNNs do not have common sense

- No object permanence
- Objects ``move" with camera motion
- Objects change size during zoom in/ zoom out
- Objects are not in perspective

Embodied visual recognition

- Neural architectures for video recognition under arbitrary camera motion

Embodied visual recognition

- Neural architectures for video recognition under arbitrary camera motion (*what we can do for embodied vision*)

Embodied visual recognition

- Neural architectures for video recognition under arbitrary camera motion (*what we can do for embodied vision*)
- Learning image representations supervised by moving and watching objects move

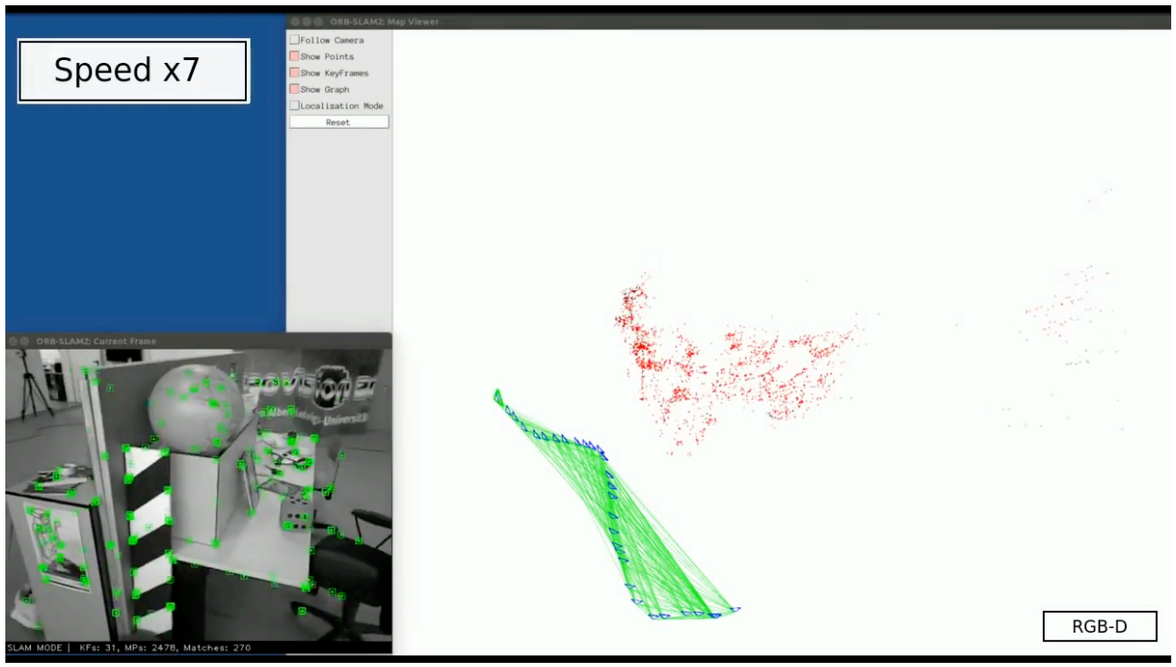
Embodied visual recognition

- Neural architectures for video recognition under arbitrary camera motion (*what we can do for embodied vision*)
- Learning image representations supervised by moving and watching objects move (*what embodiment can do for us*)

Embodied visual recognition

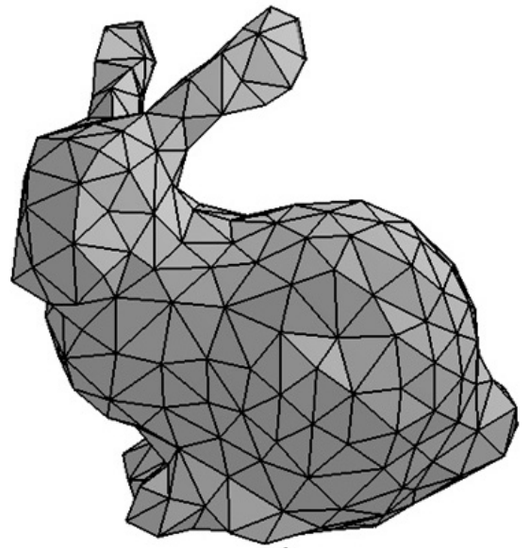
- Neural architectures for video recognition under arbitrary camera motion (*what we can do for embodied vision*)
- Learning image representations supervised by moving and watching objects move (*what embodiment can do for us*)

3D representations

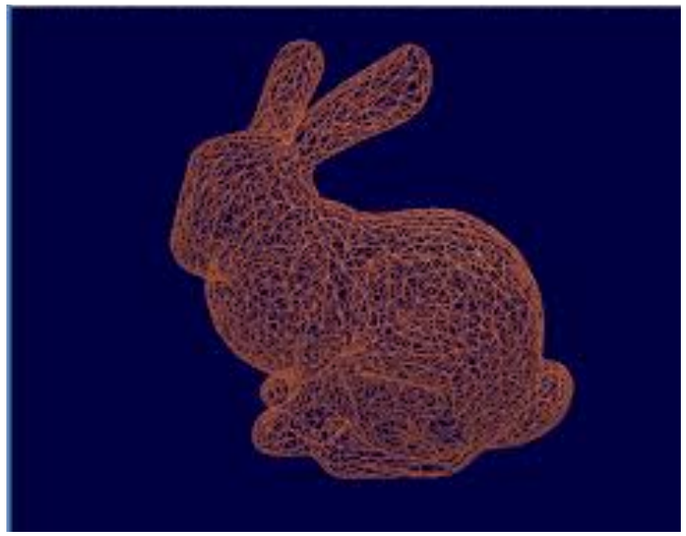


3D representations

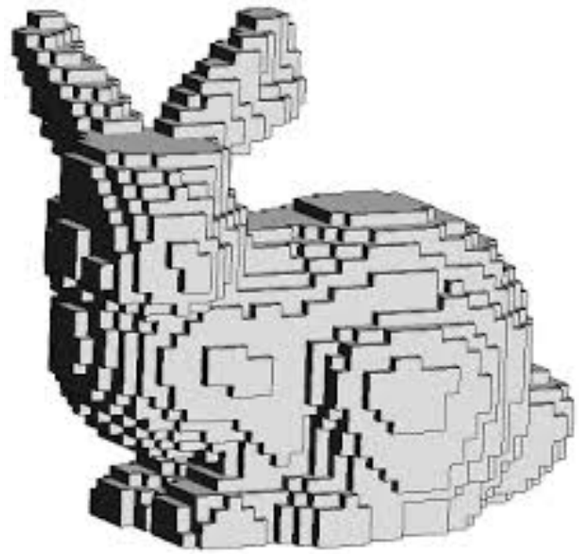
3D mesh



3D pointcloud



3D voxel occupancy



- …ask too much: high level of 3D details that may be impossible to obtain
- …ask too little: information about semantics of the objects is not captured

3D representations

3D mesh

3D pointcloud

3D voxel occupancy

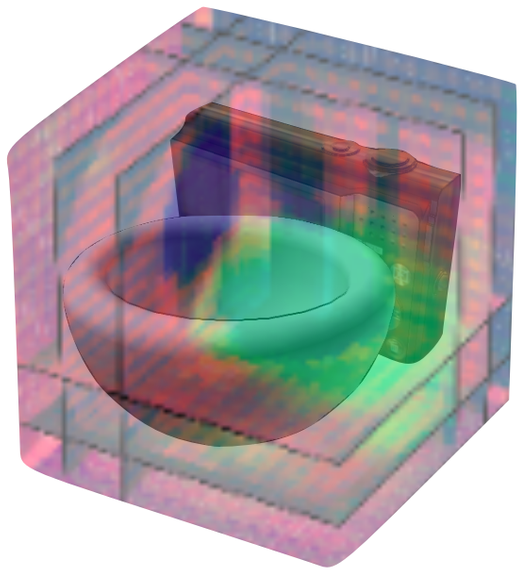
``Internal world models which are complete representations of the external environment, besides being impossible to obtain, are not at all necessary for agents to act in a competent manner.”

Intelligence without reason, IJCAI, Rodney Brooks (1991)

- …ask too much: high level of 3D details that may be impossible to obtain
- …ask too little: information about semantics of the objects is not present

To 3D or not to 3D?

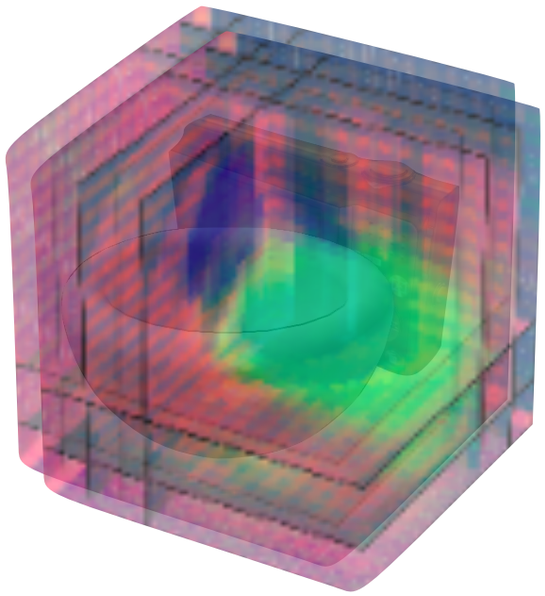
3D feature maps



$$H \times W \times D \times C$$

3 spatial dimensions, 1 feature dimension

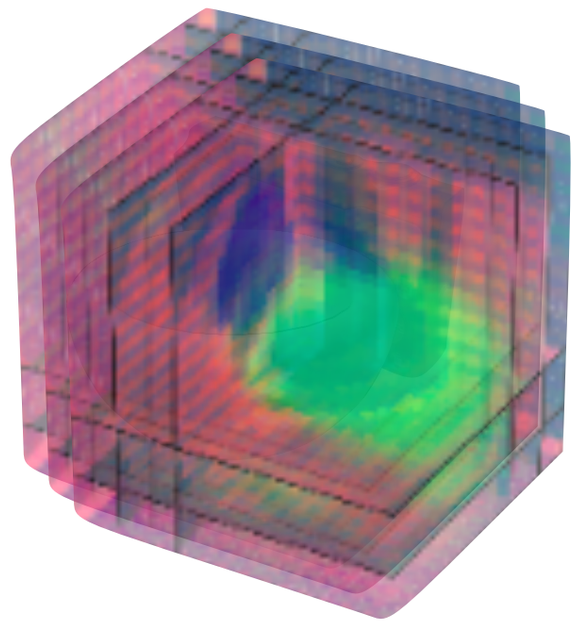
3D feature maps



$H \times W \times D \times C$

3 spatial dimensions, 1 feature dimension

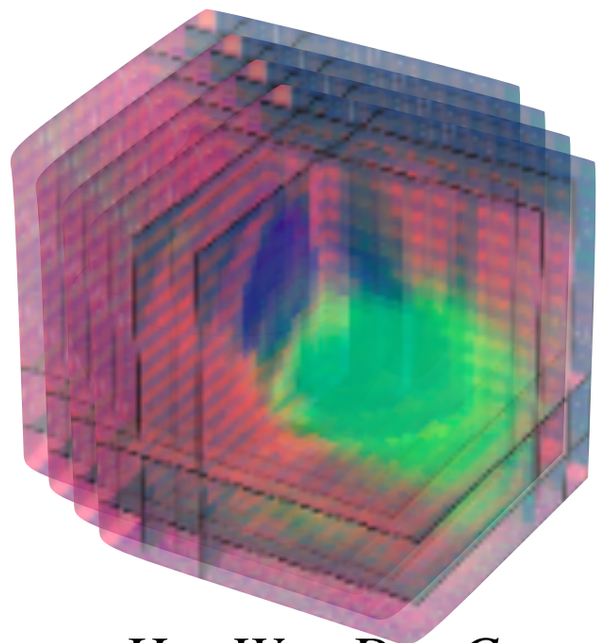
3D feature maps



$H \times W \times D \times C$

3 spatial dimensions, 1 feature dimension

3D feature maps

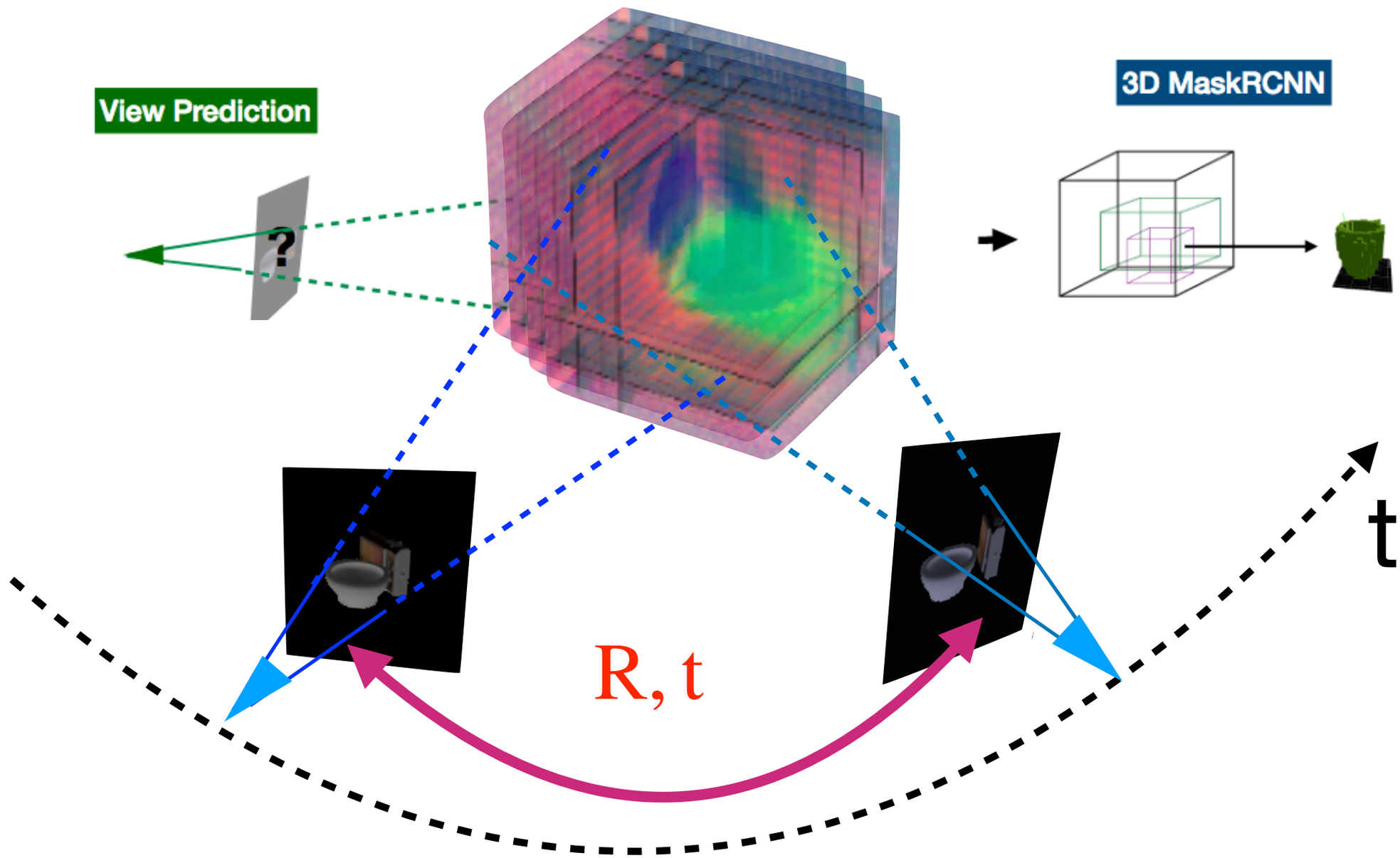


$H \times W \times D \times C$

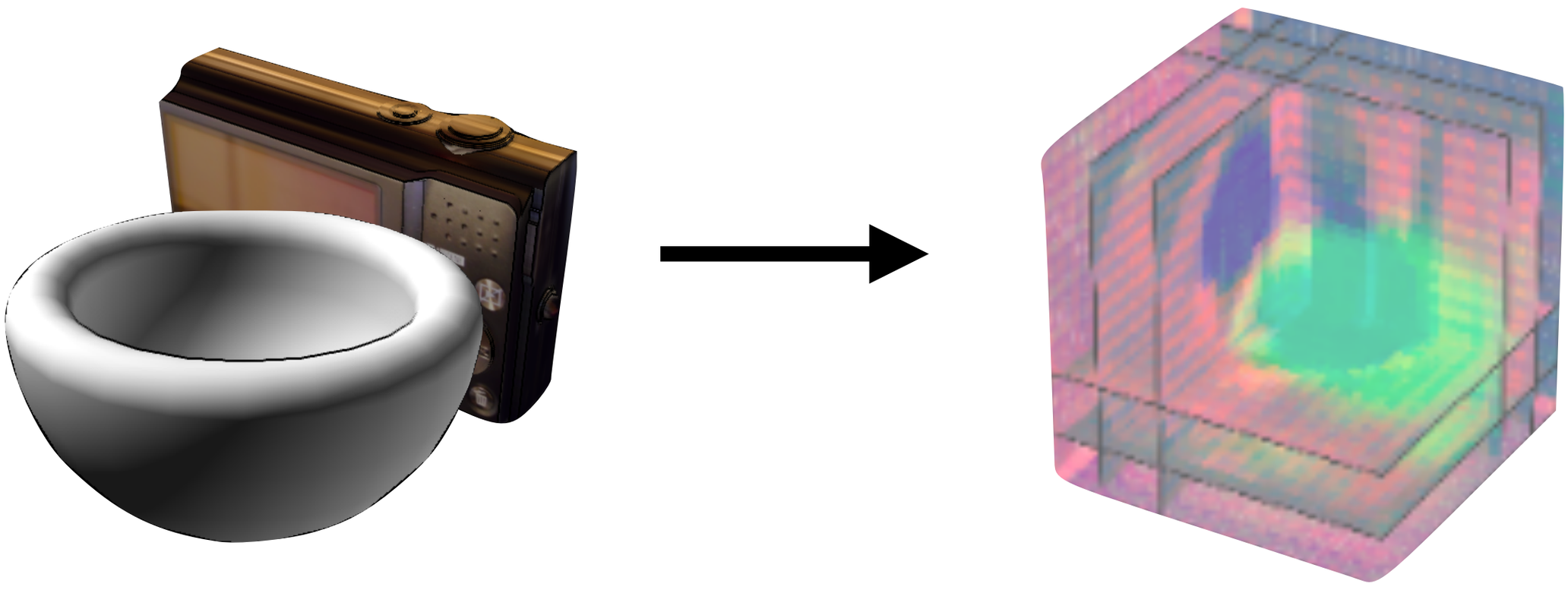
3 spatial dimensions, 1 feature dimension

Geometry-Aware Recurrent Networks

1. Hidden state: geometrically consistent 3D feature maps
2. Egomotion-stabilized hidden state updates

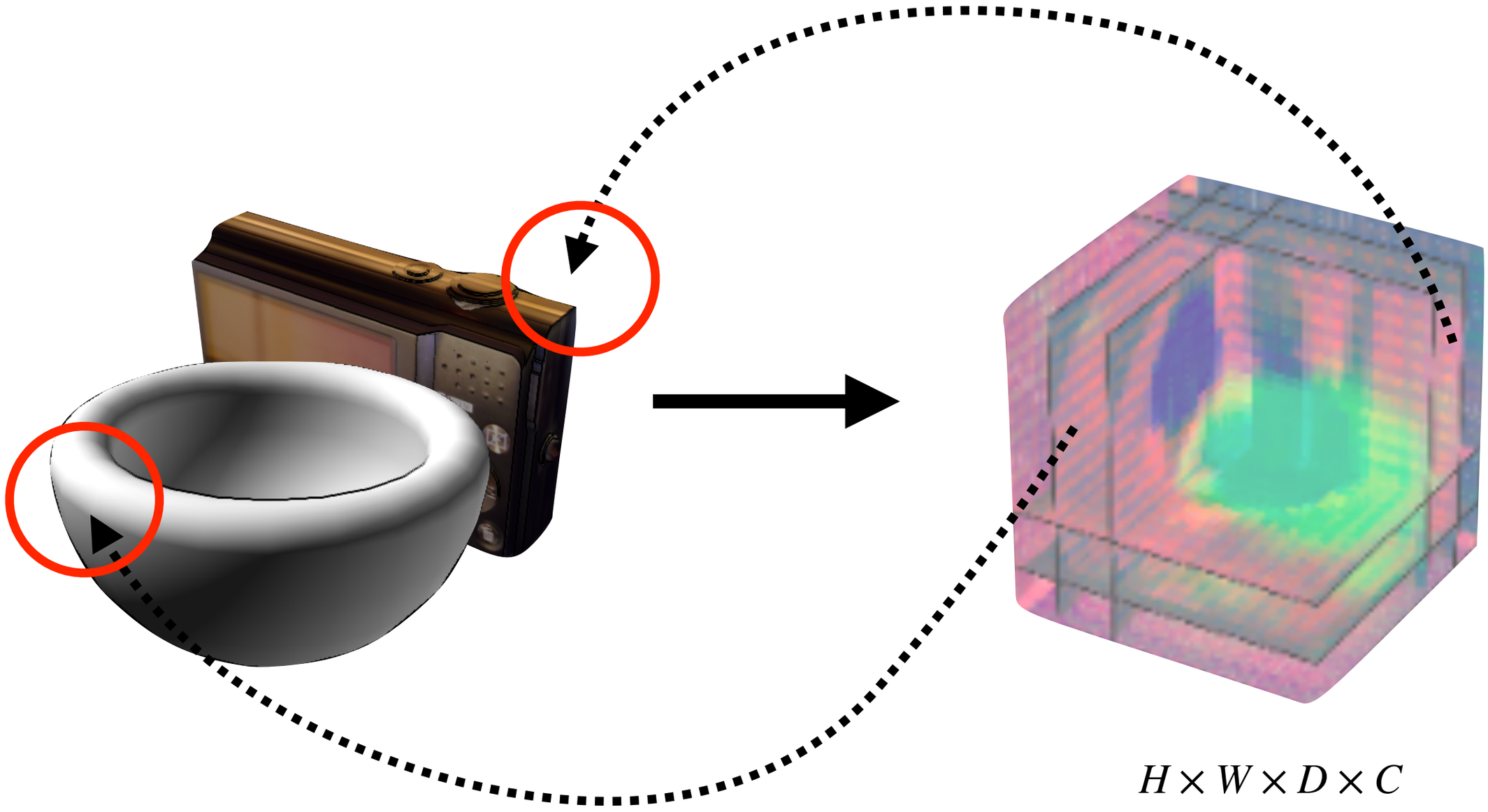


Geometry-Aware Recurrent Networks (GRNNs)



$H \times W \times D \times C$

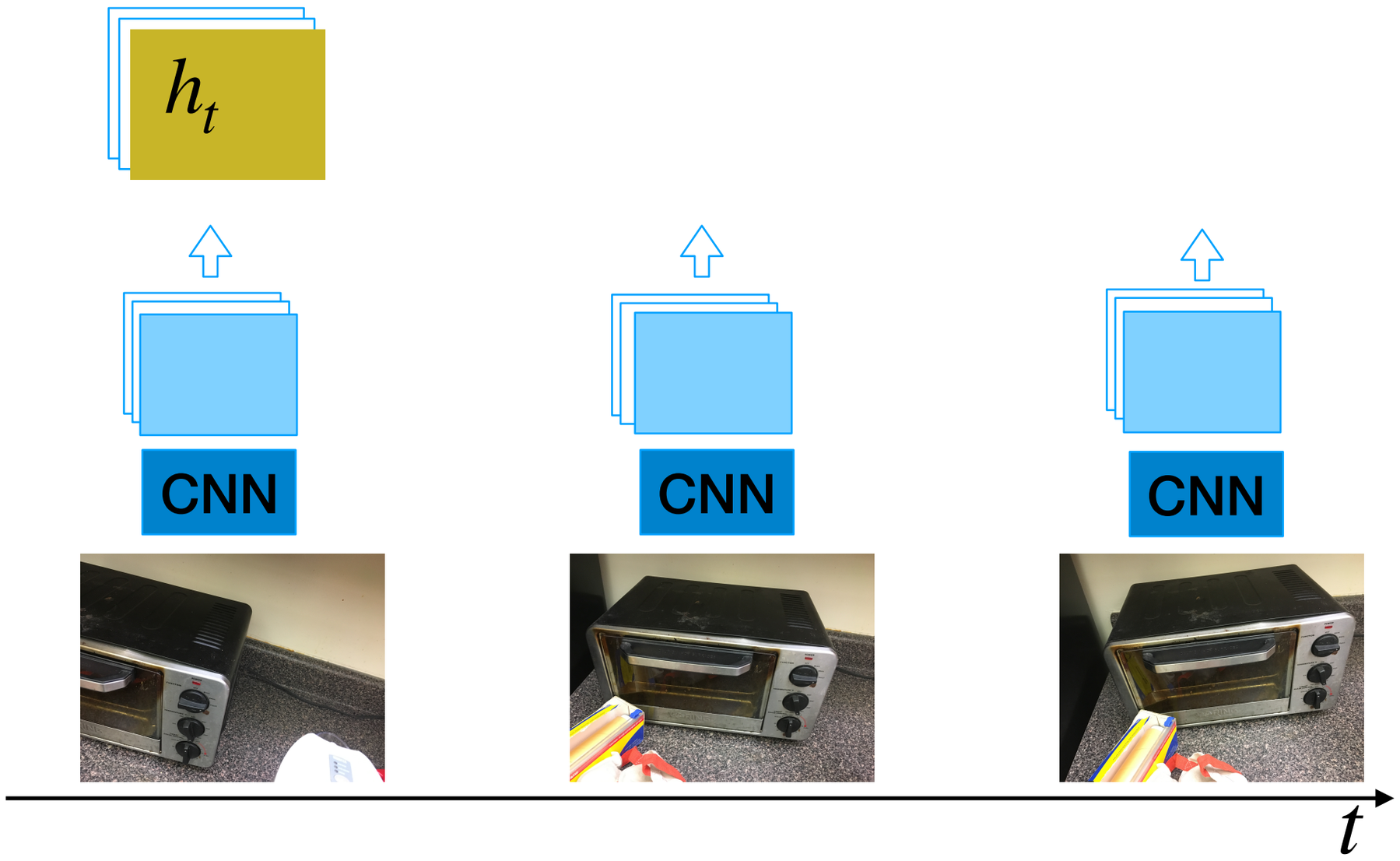
Geometry-Aware Recurrent Networks (GRNNs)



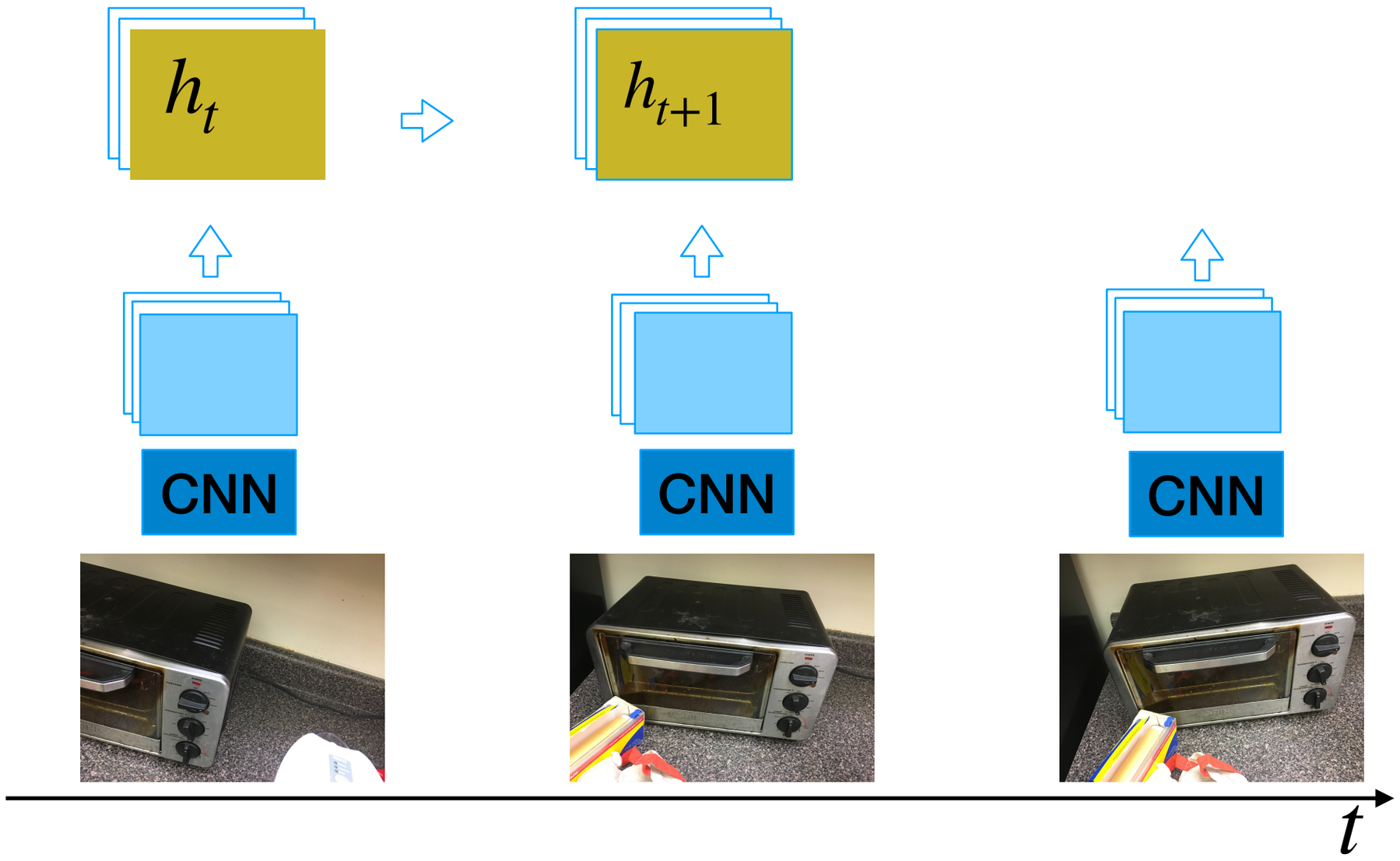
2D RNNs (conv-LSTMs/GRUs)

2D RNNs (conv-LSTMs/GRUs)

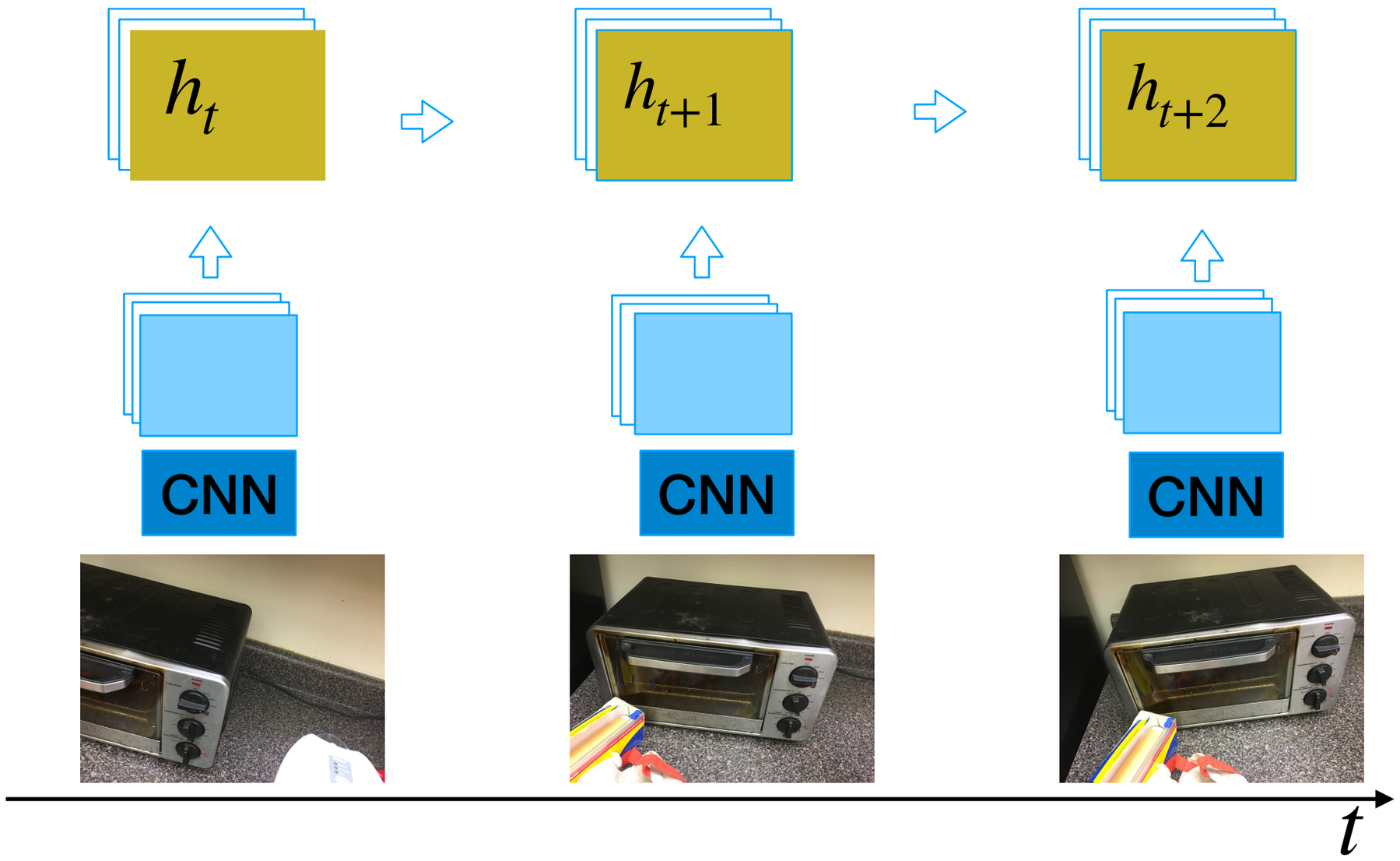
2D RNNs (conv-LSTMs/GRUs)



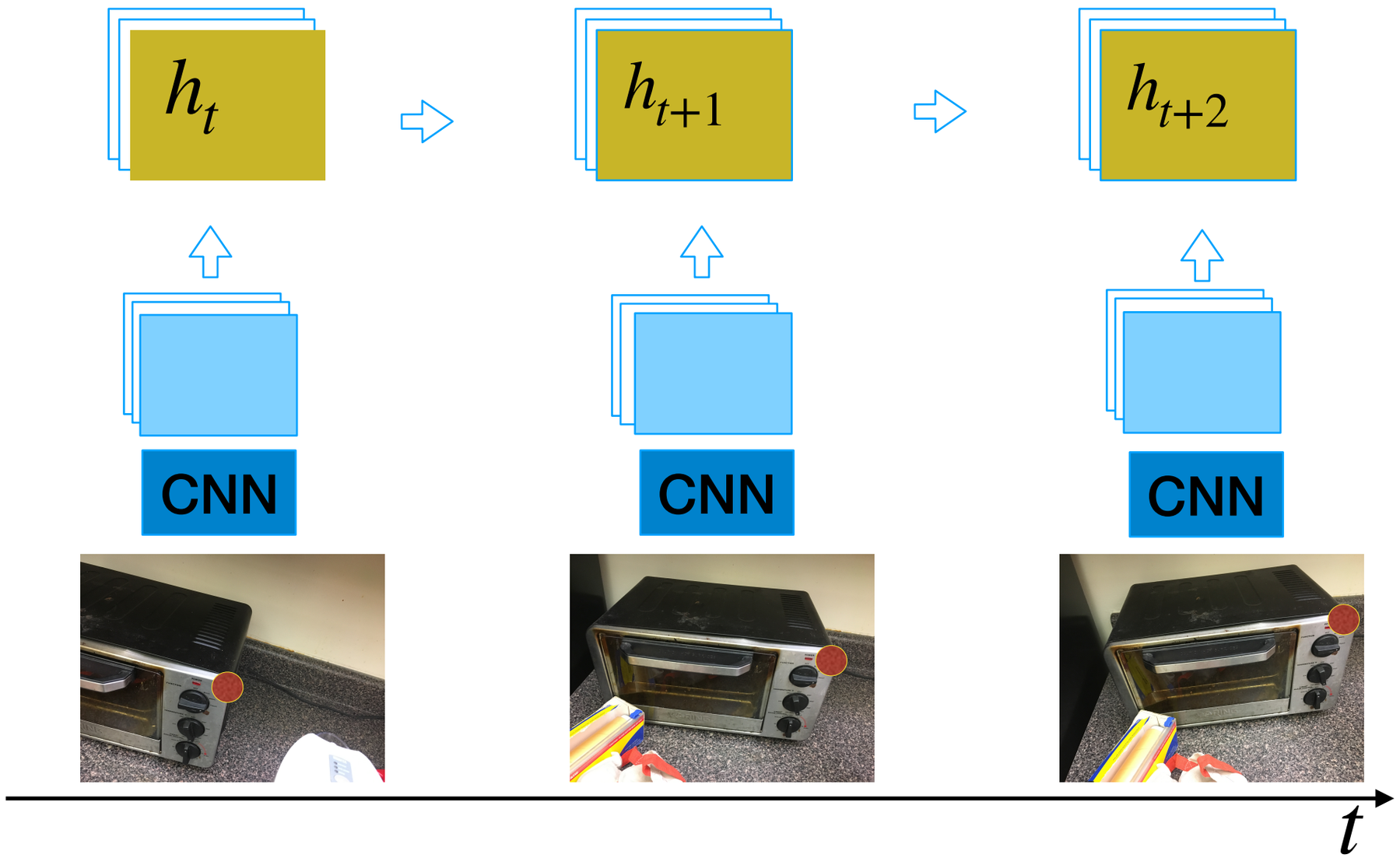
2D RNNs (conv-LSTMs/GRUs)



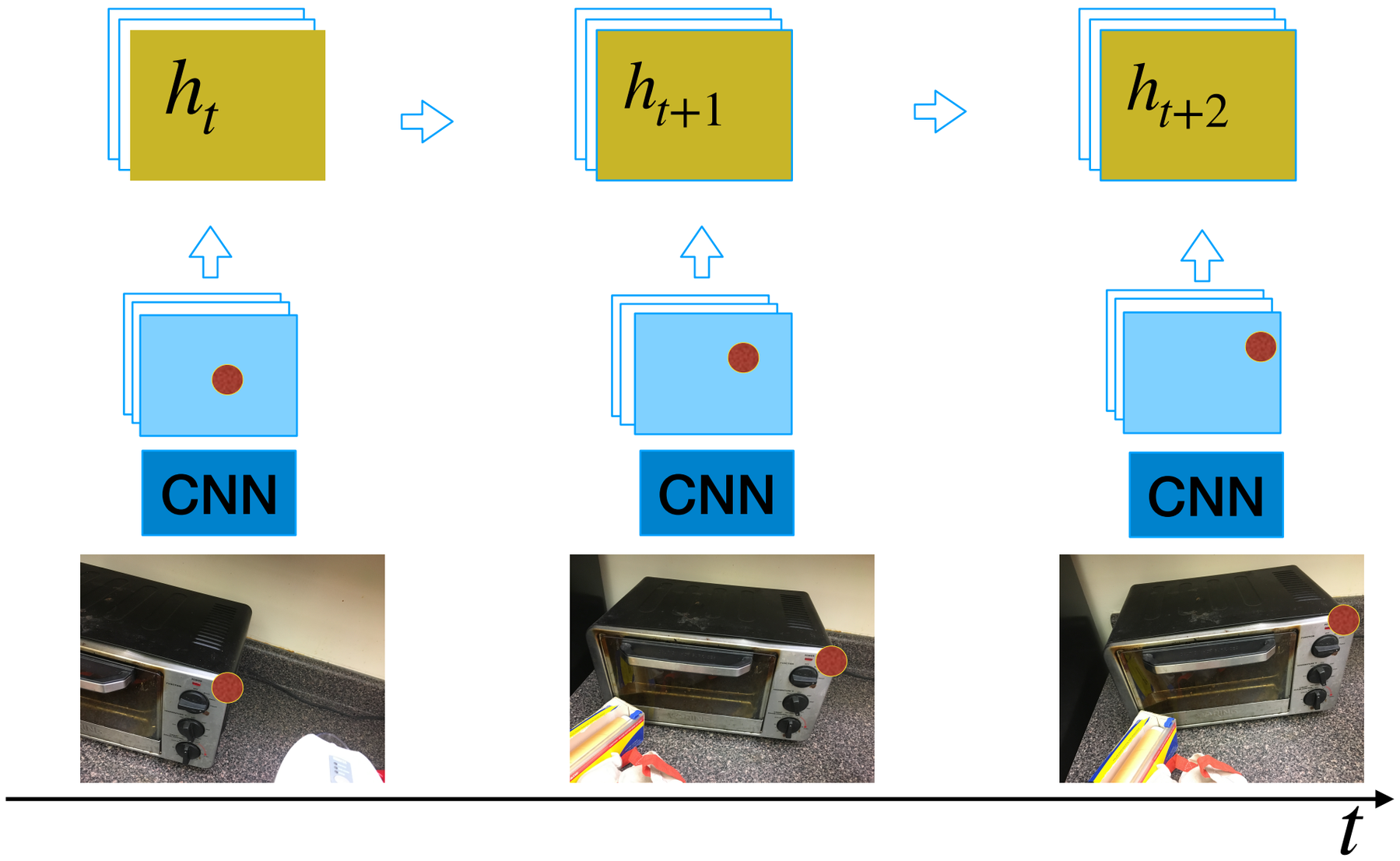
2D RNNs (conv-LSTMs/GRUs)



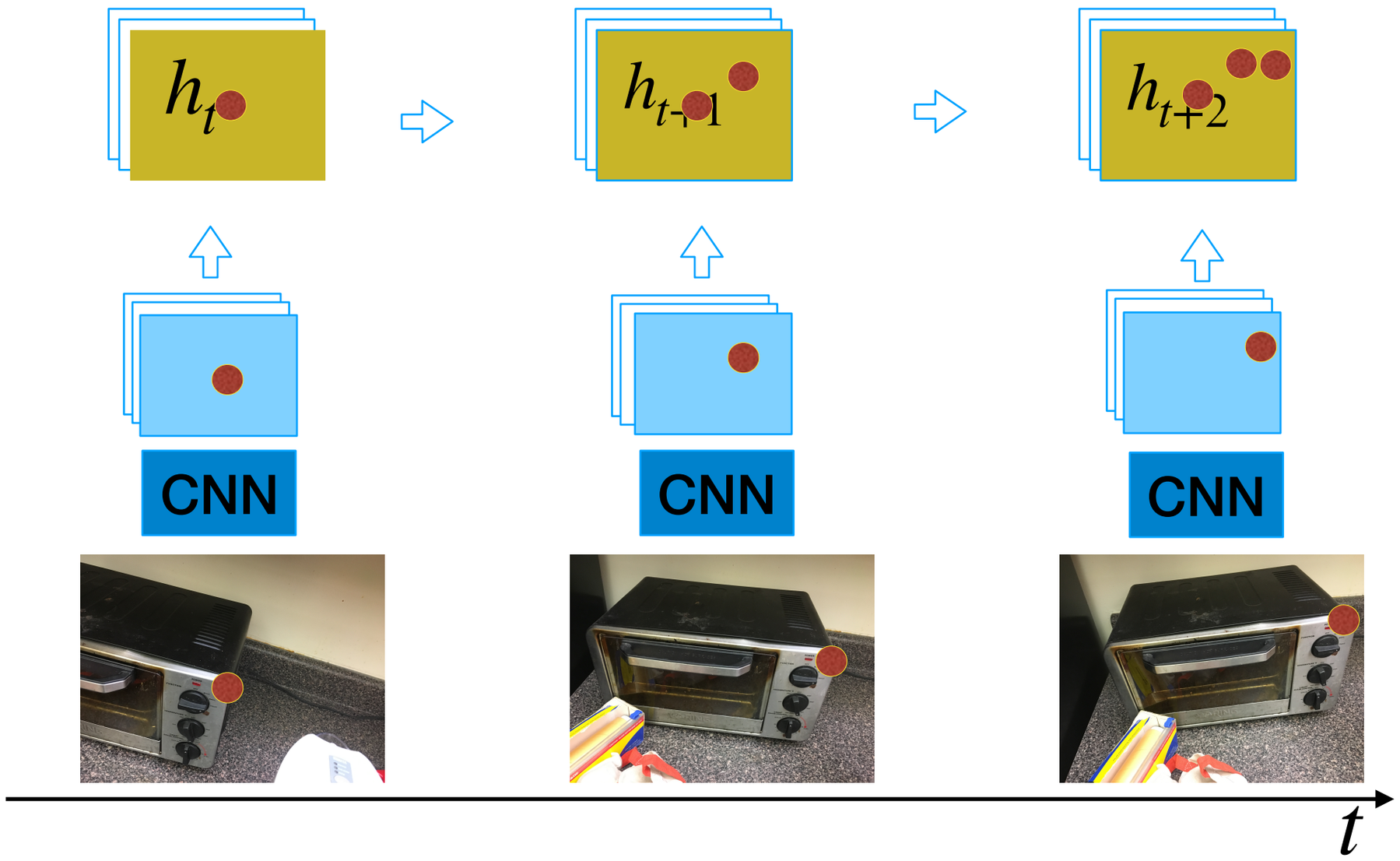
2D RNNs (conv-LSTMs/GRUs)



2D RNNs (conv-LSTMs/GRUs)



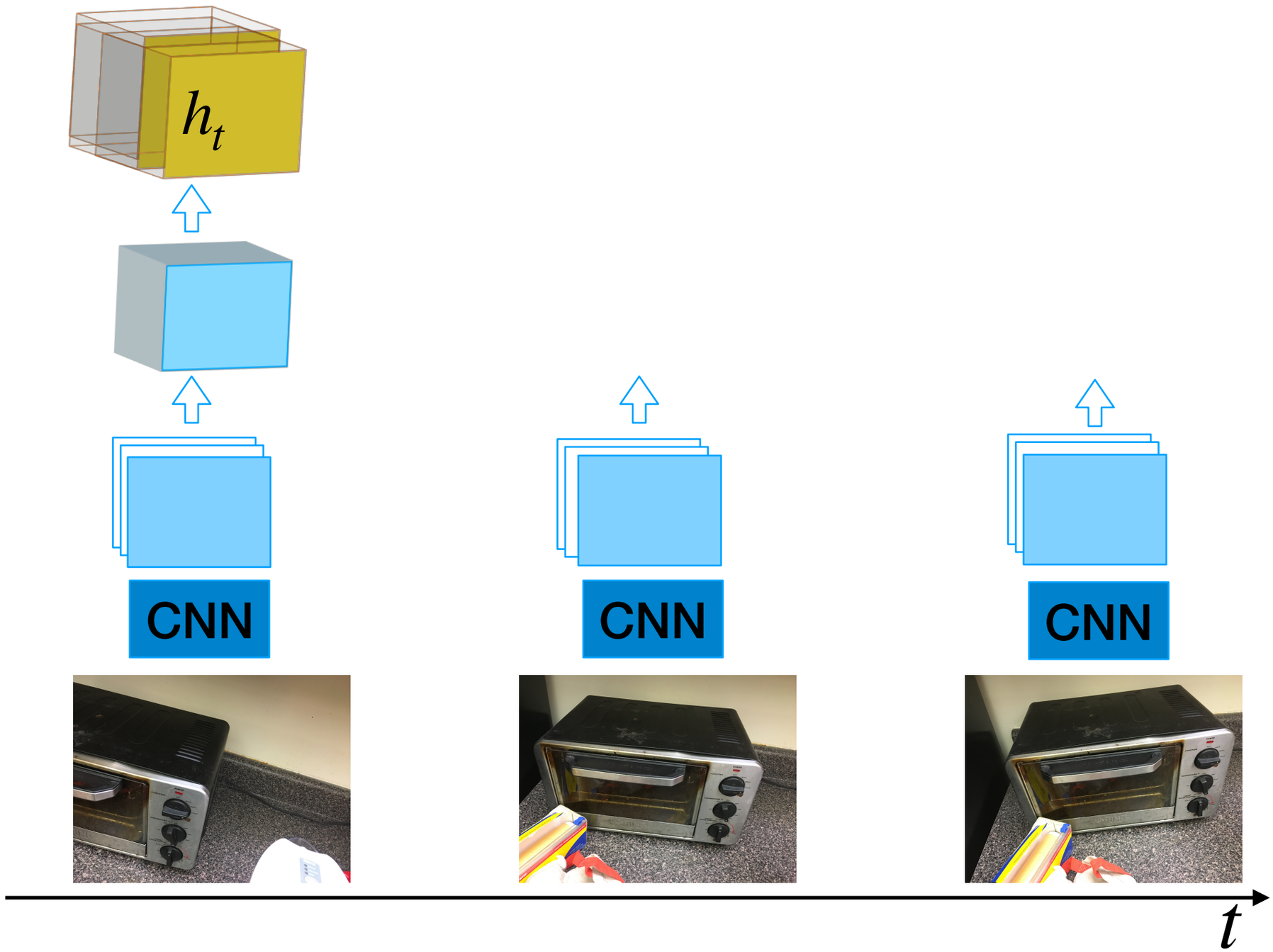
2D RNNs (conv-LSTMs/GRUs)



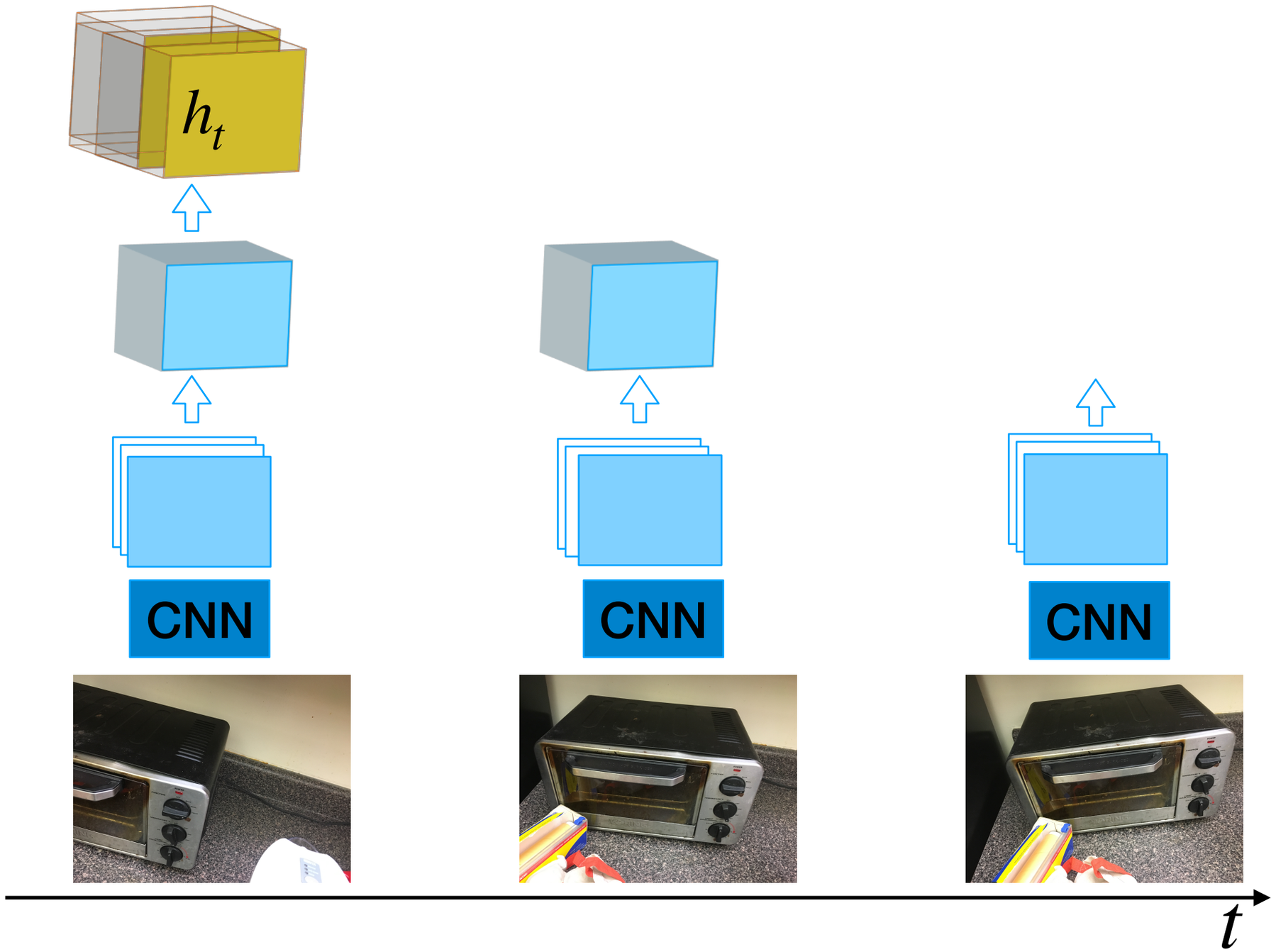
Geometry-Aware Recurrent Networks

Geometry-Aware Recurrent Networks

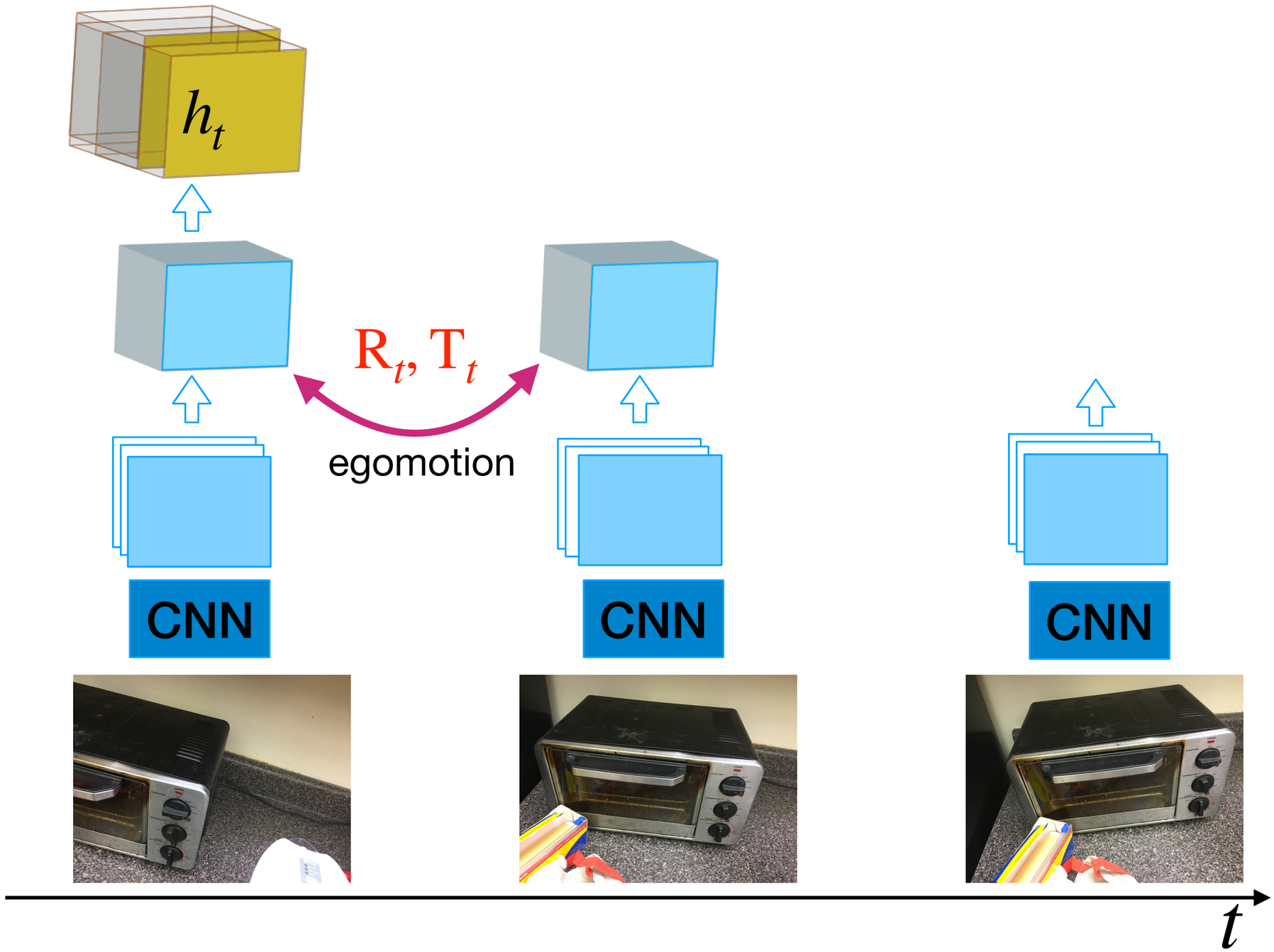
Geometry-Aware Recurrent Networks



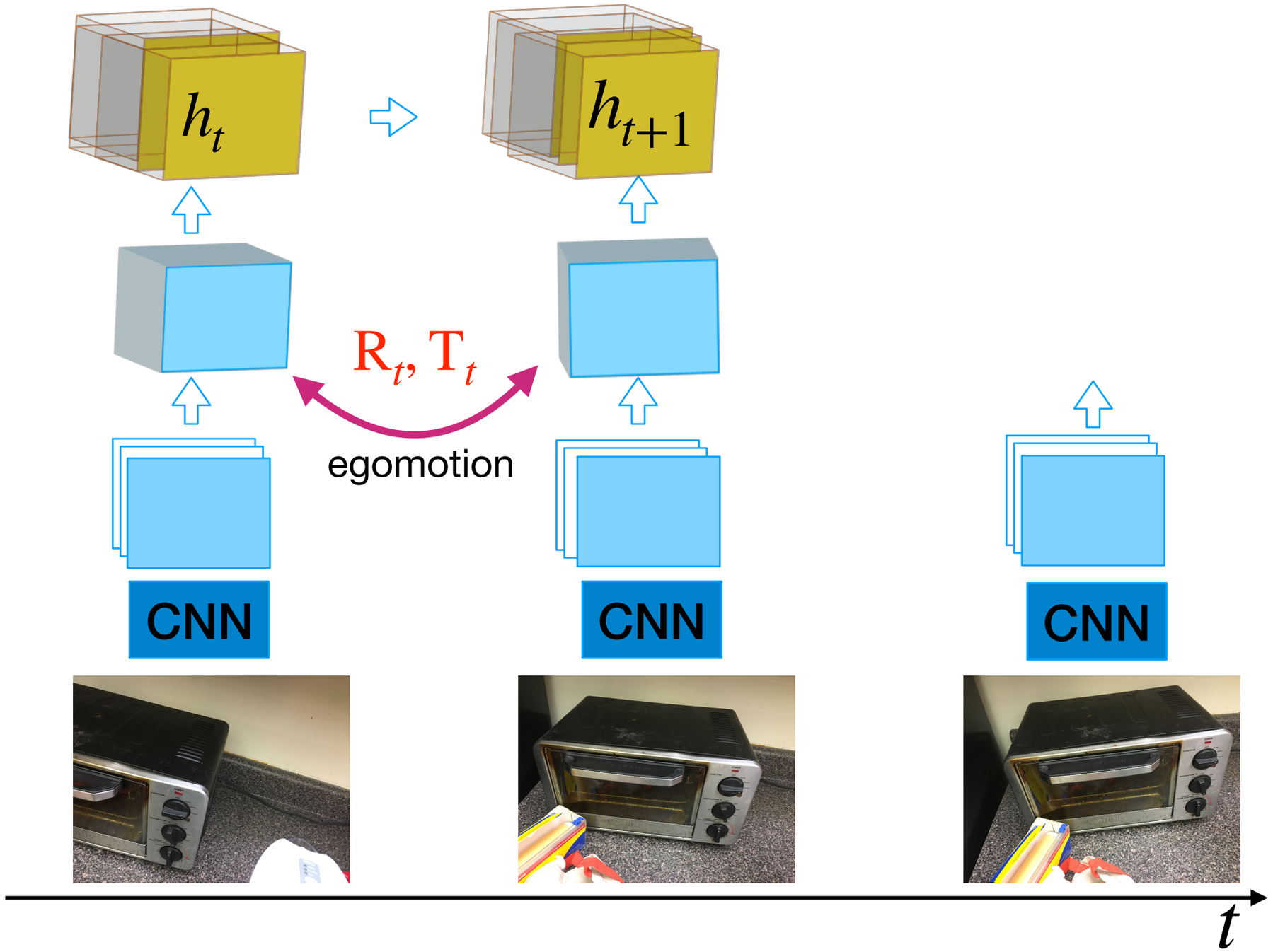
Geometry-Aware Recurrent Networks



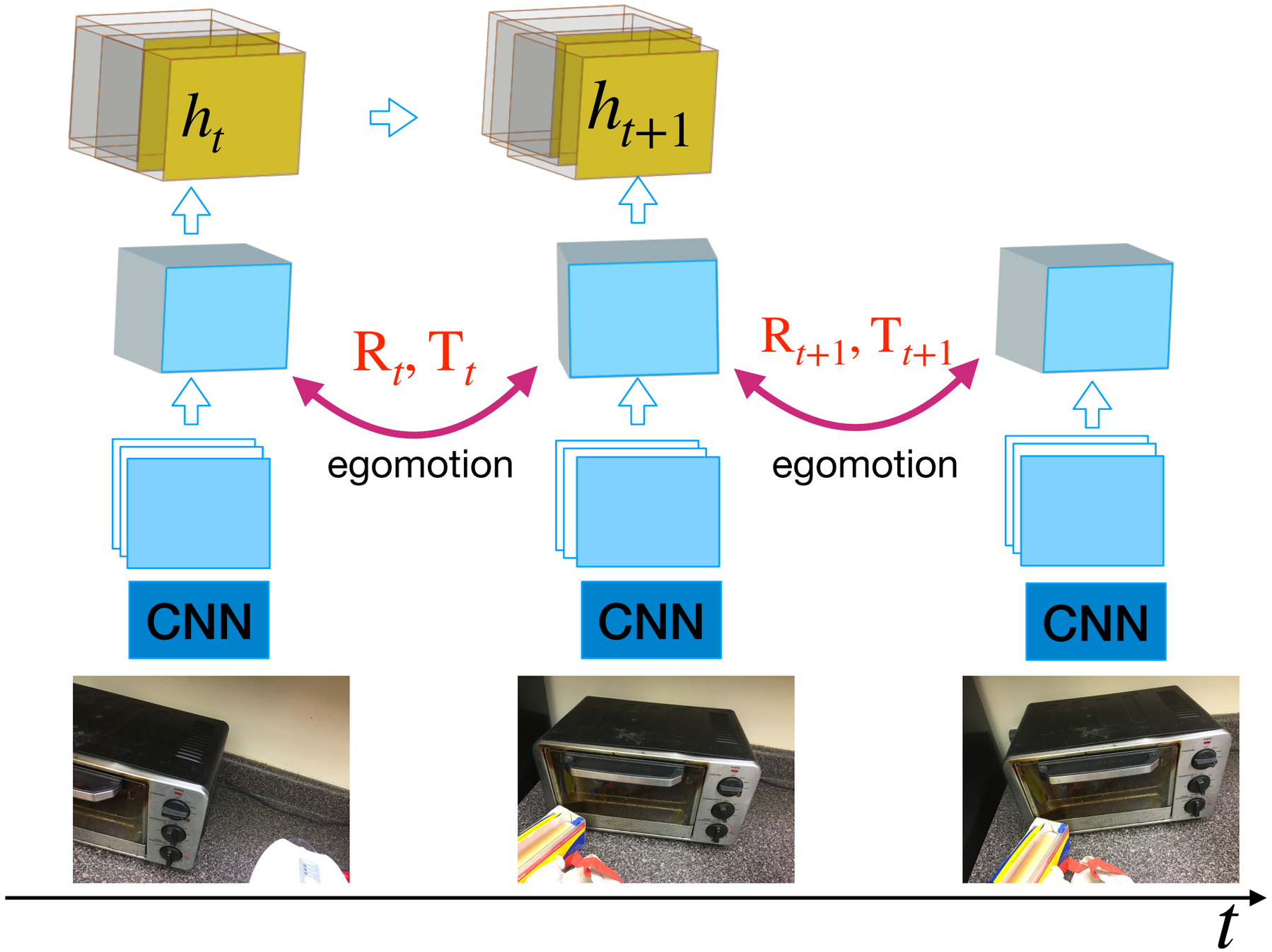
Geometry-Aware Recurrent Networks



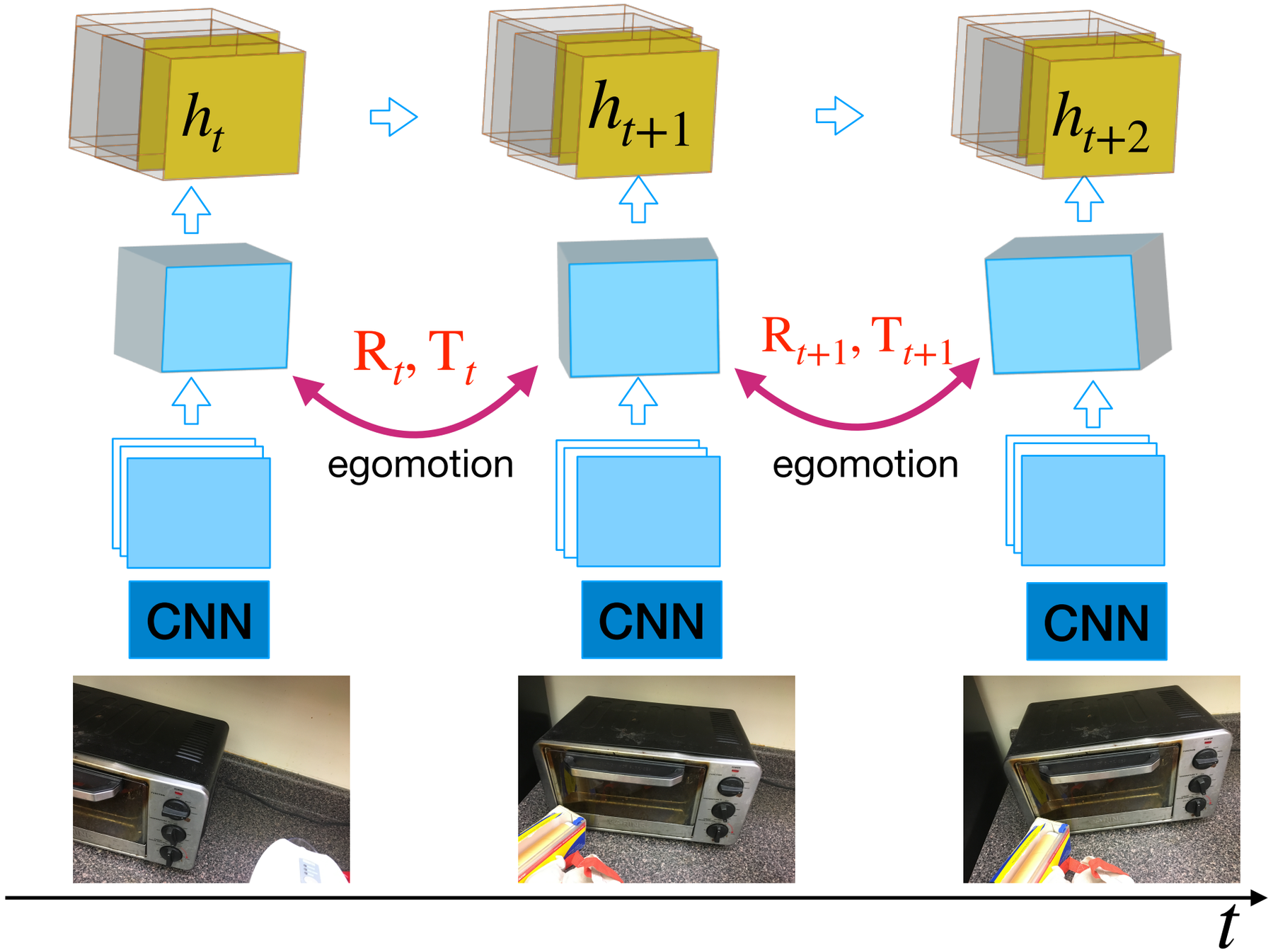
Geometry-Aware Recurrent Networks



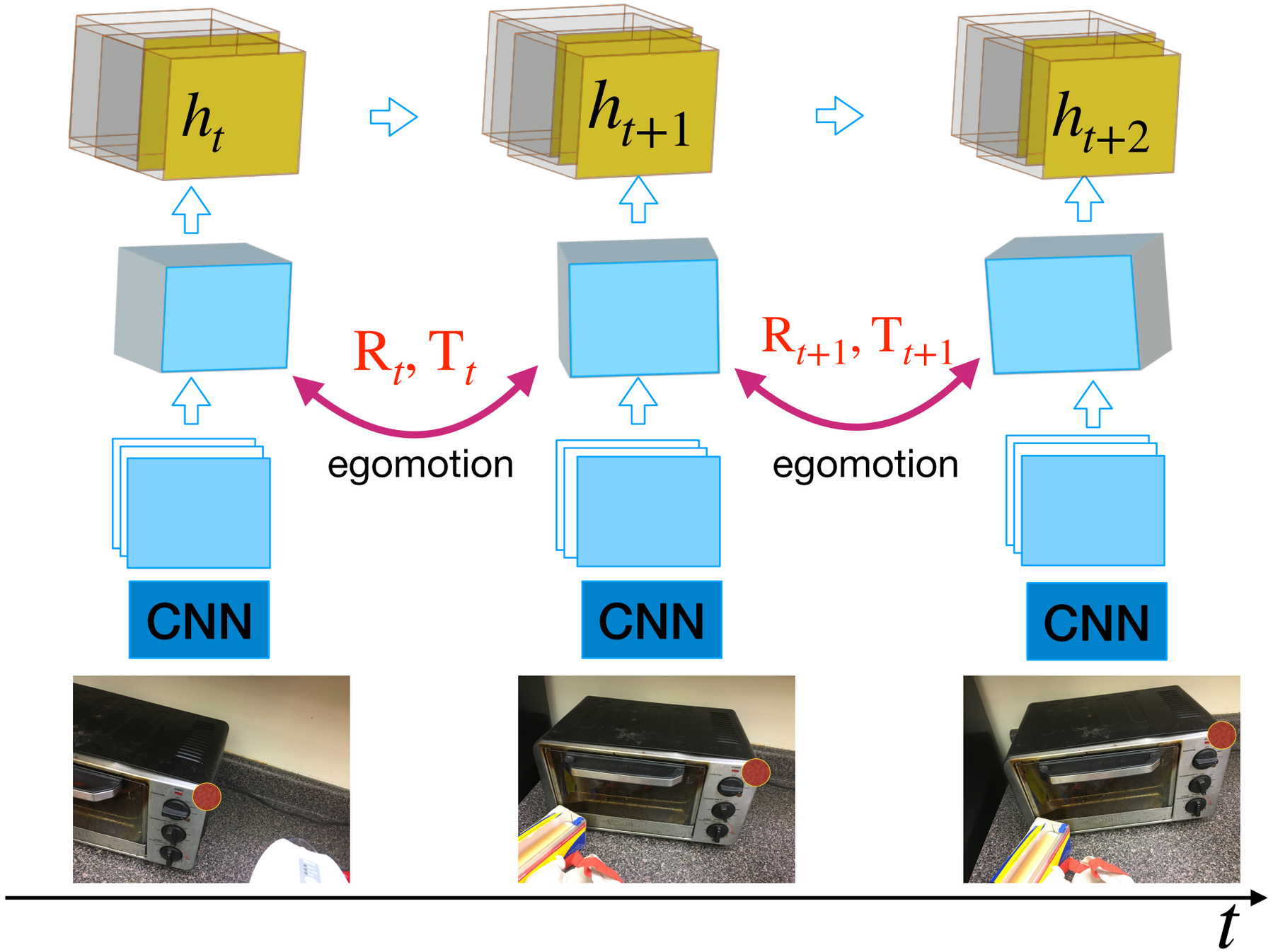
Geometry-Aware Recurrent Networks



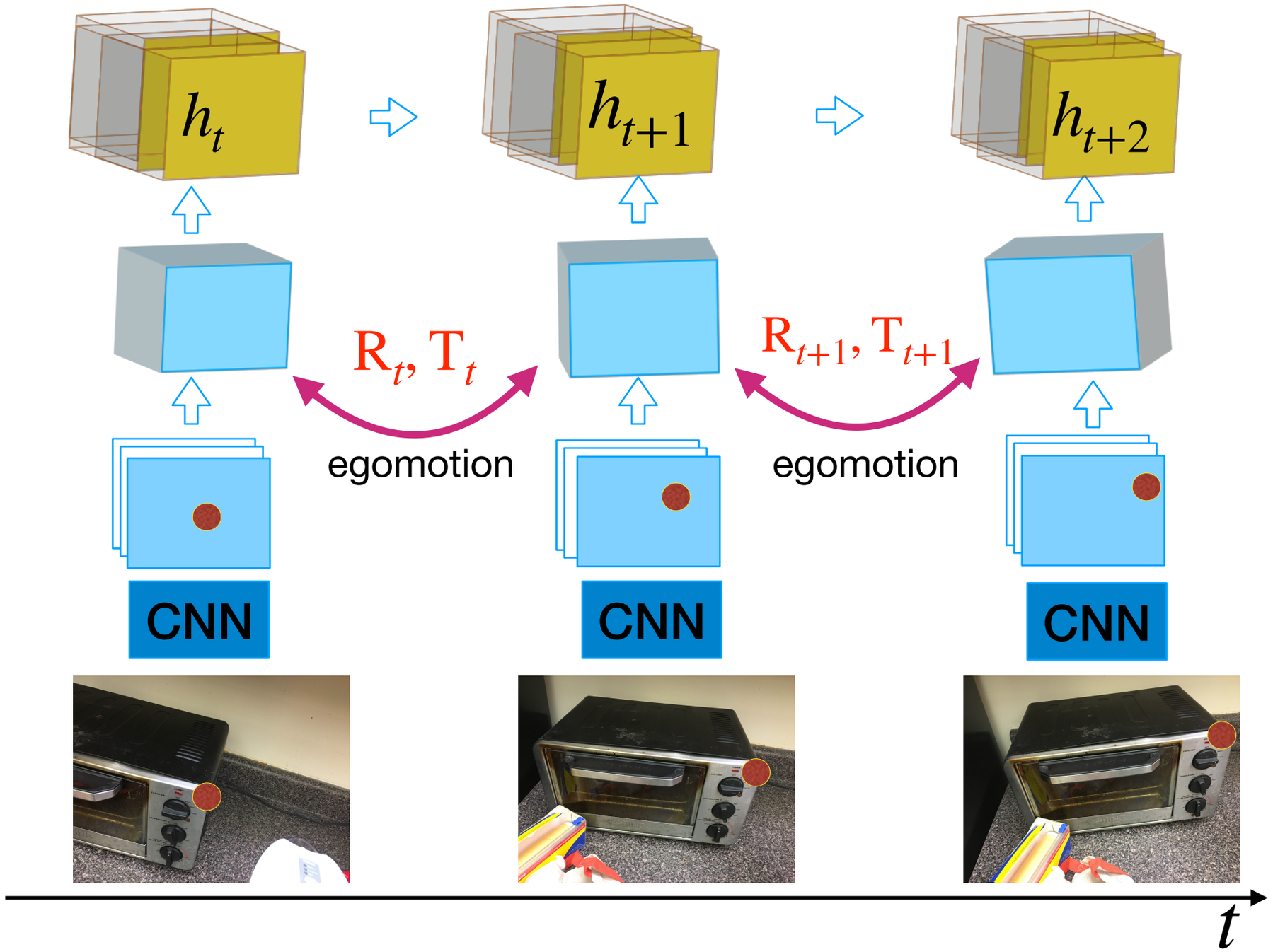
Geometry-Aware Recurrent Networks



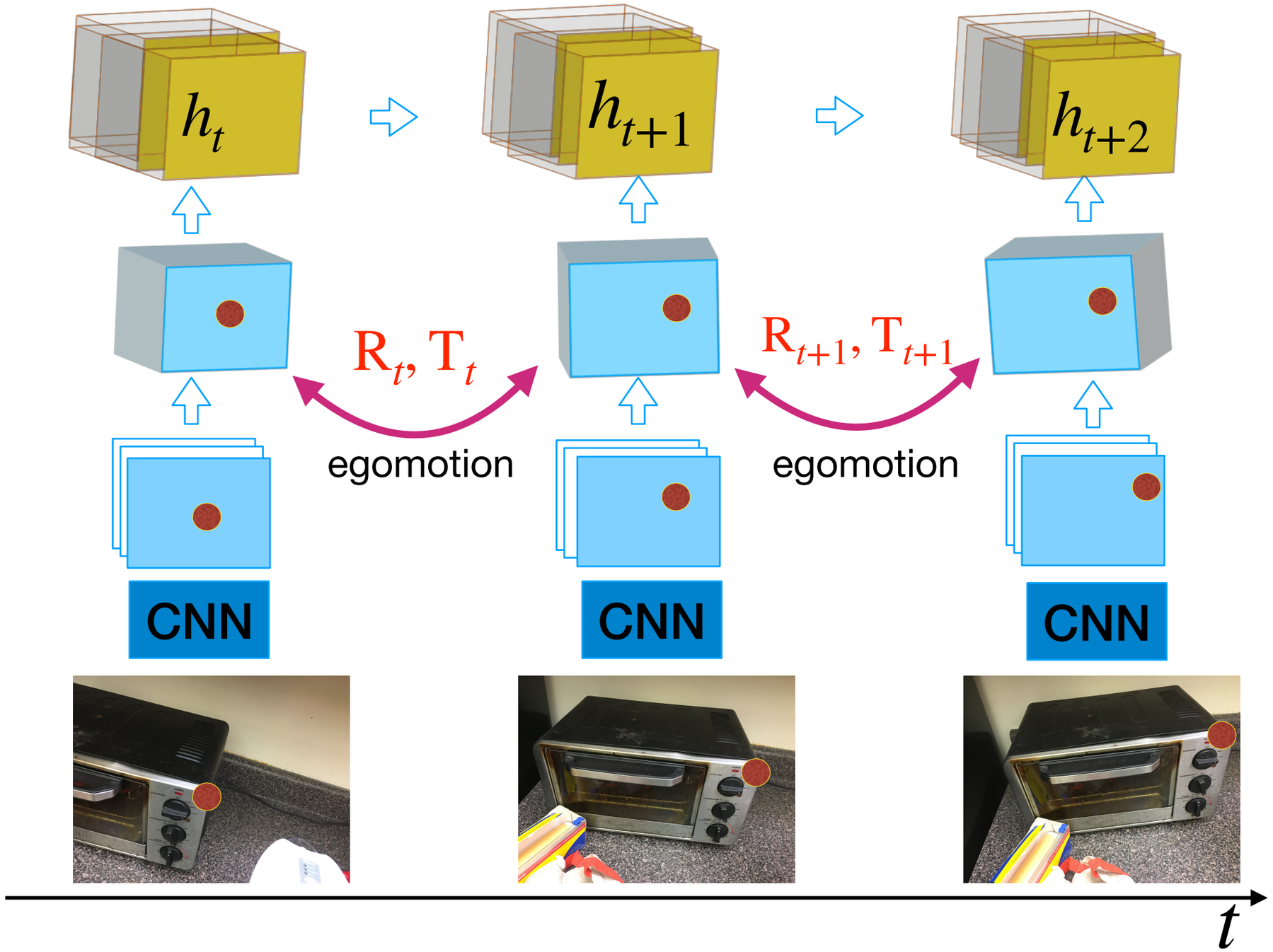
Geometry-Aware Recurrent Networks



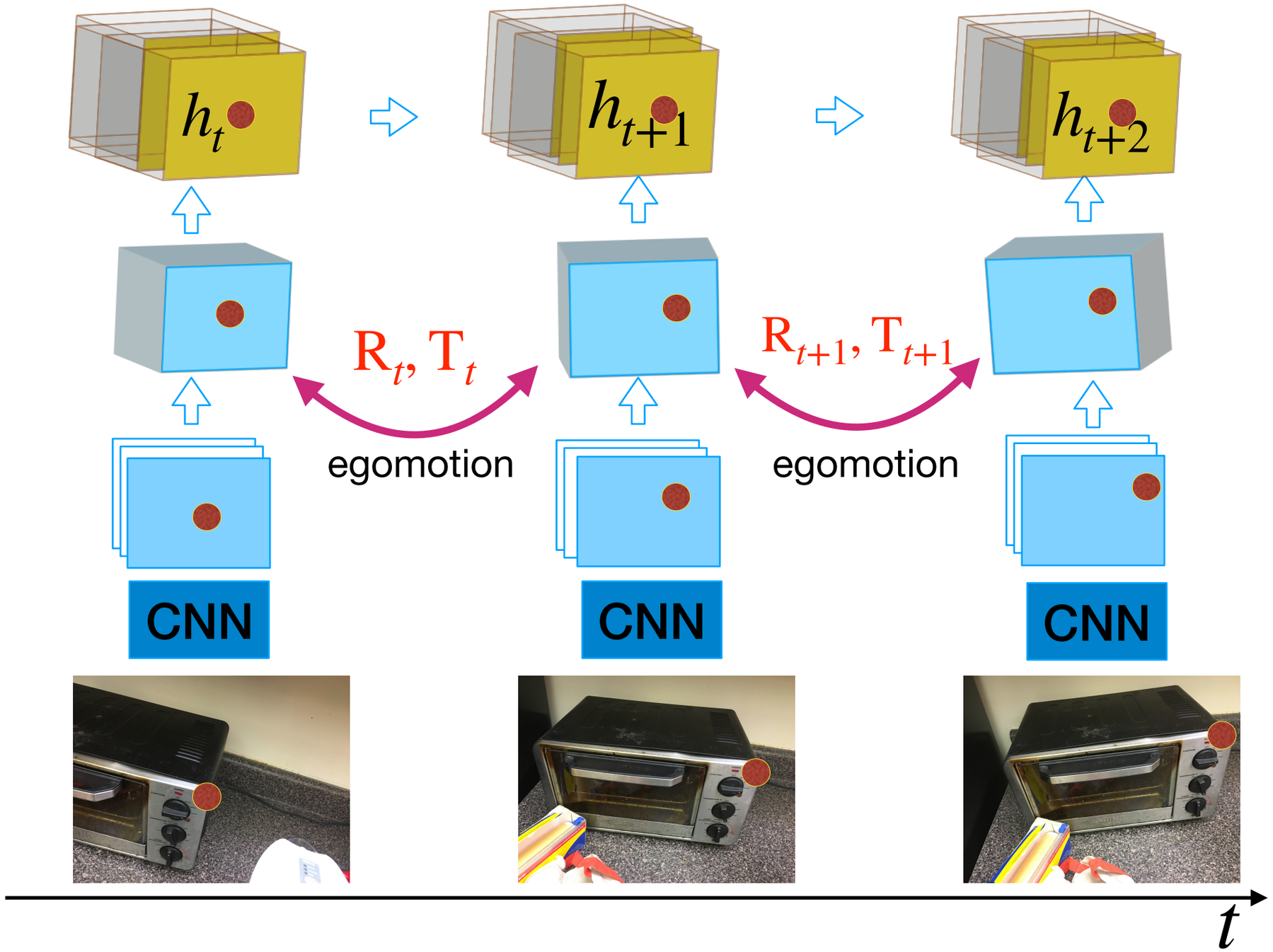
Geometry-Aware Recurrent Networks



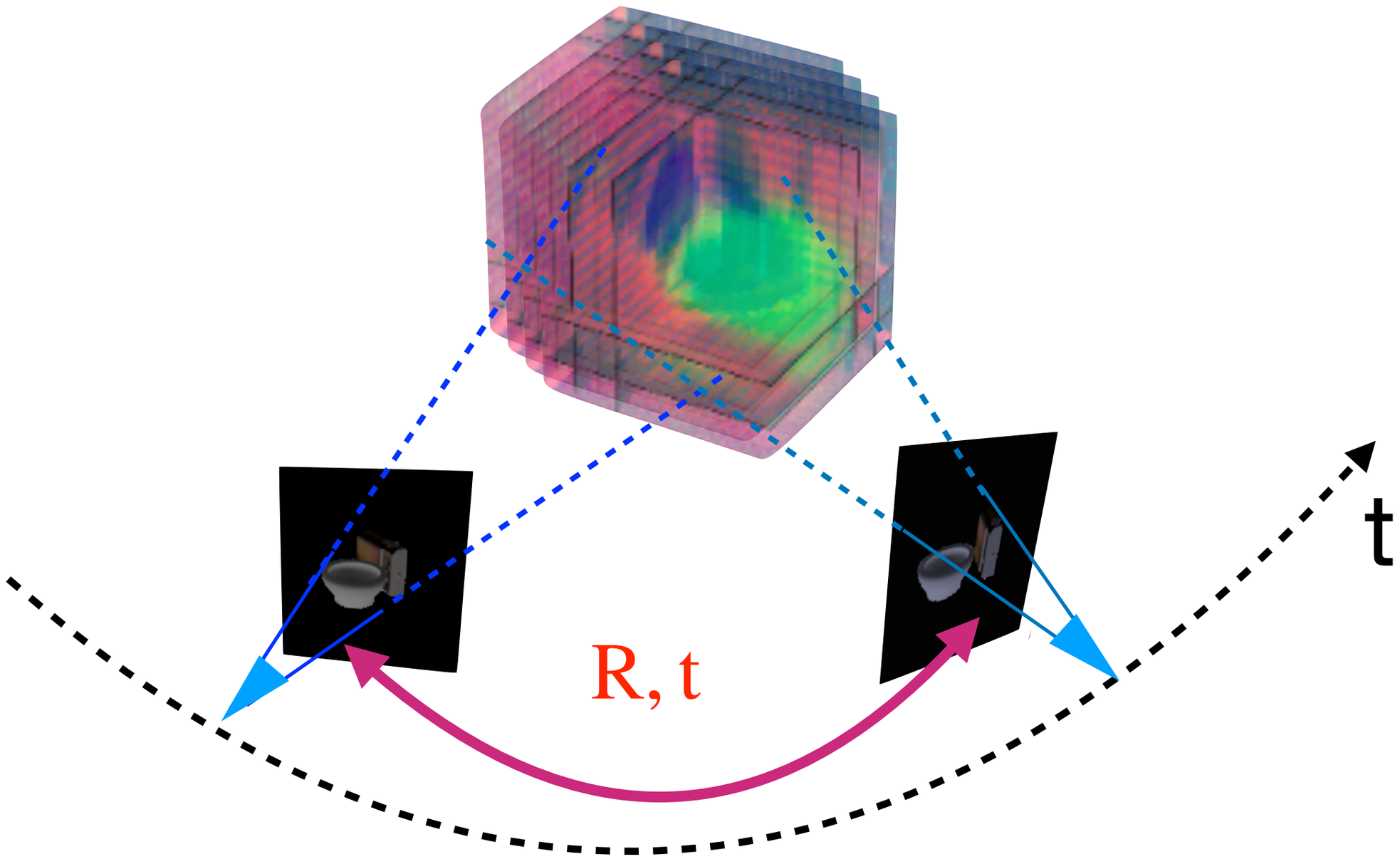
Geometry-Aware Recurrent Networks



Geometry-Aware Recurrent Networks

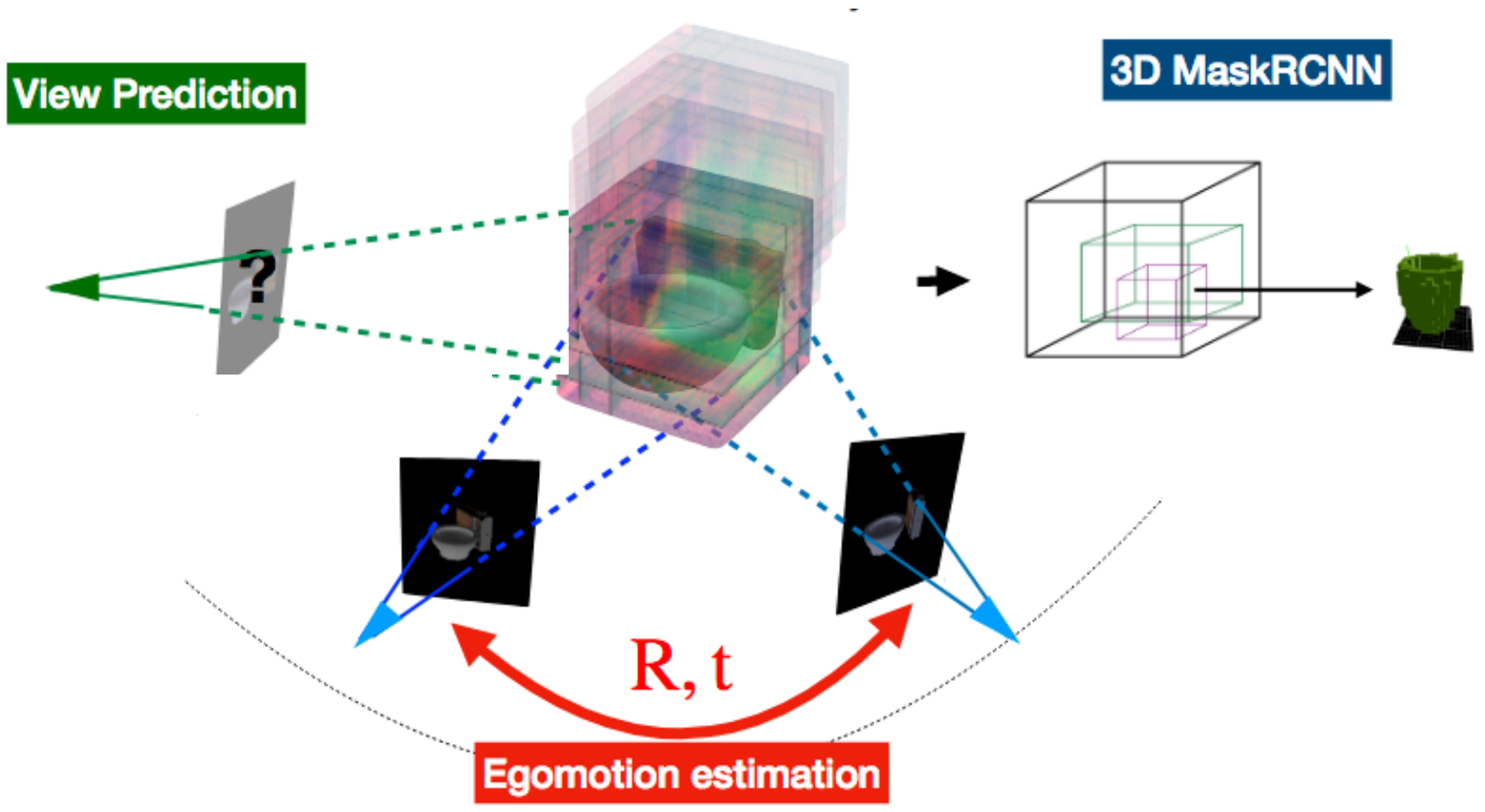


GRNNs



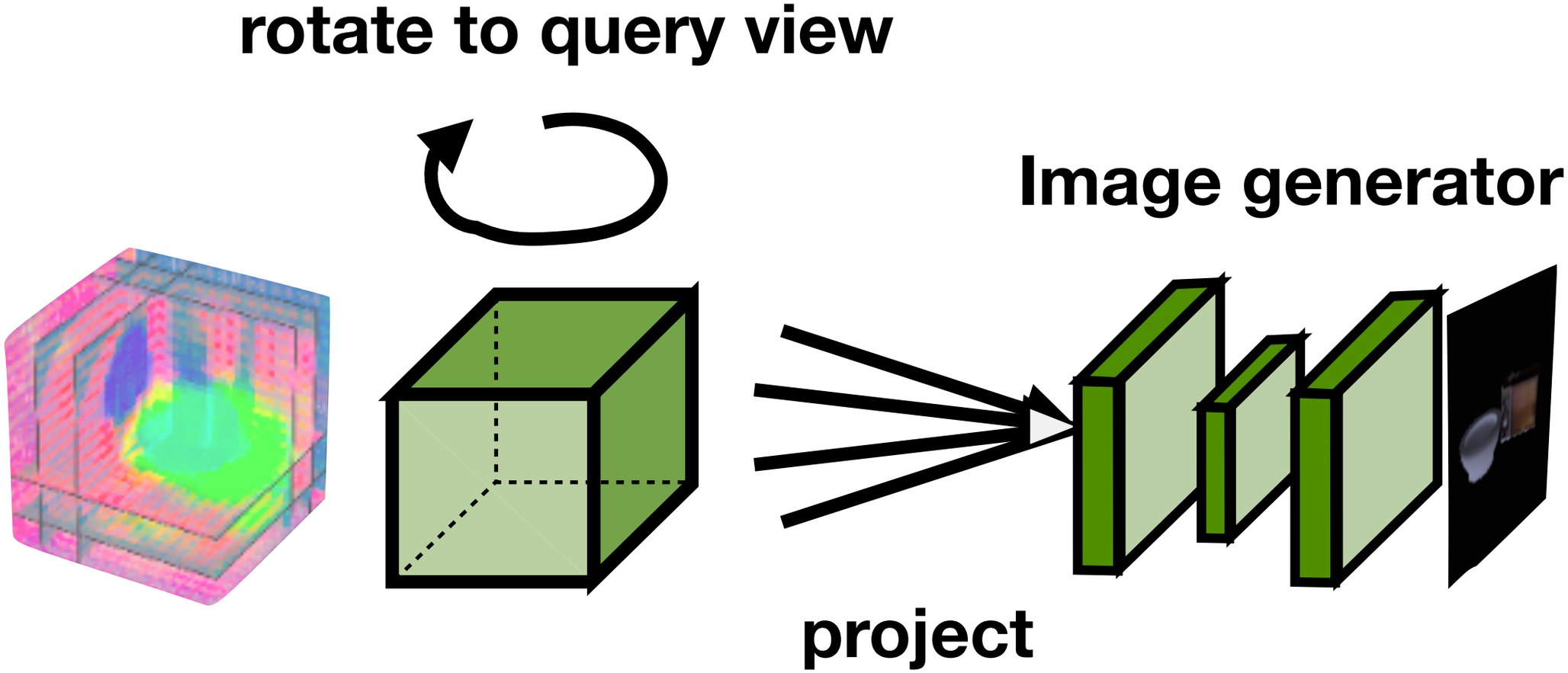
- A set of differentiable neural modules to learn to go from 2D to 3D and back
- A lot of SLAM ideas into the neural modules

Training GRNNs



1. Self-supervised via predicting images the agent will see under novel viewpoints
2. Supervised for 3D object detection

Image generation



Results - view prediction

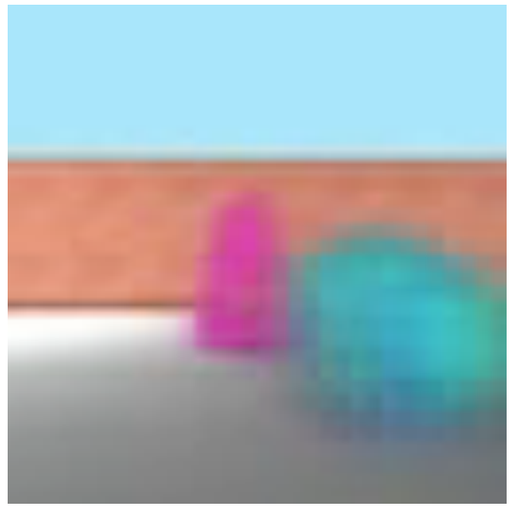
of input views

Results - view prediction

of input views

Results - view prediction

of input views

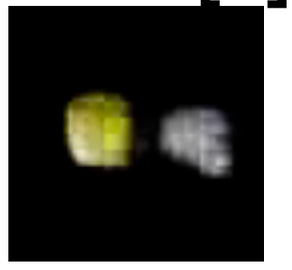


Results - view prediction

Input views

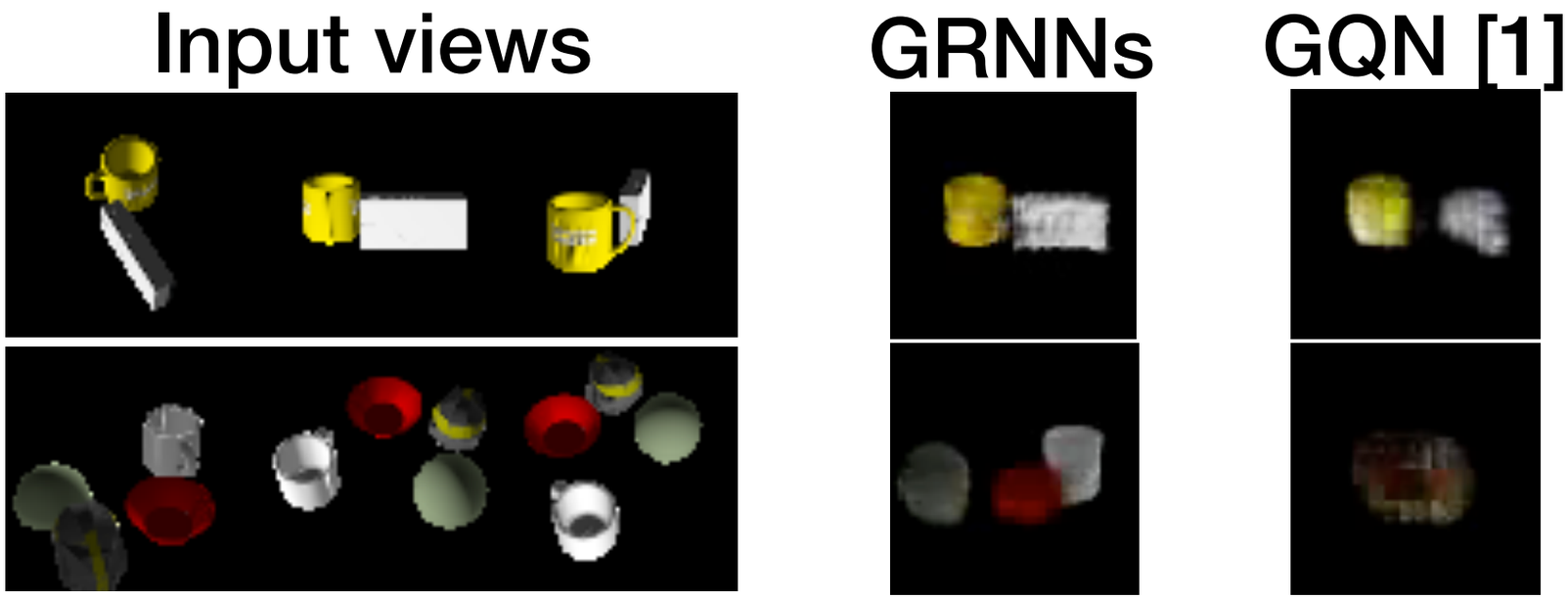
GRNNs

GQN [1]



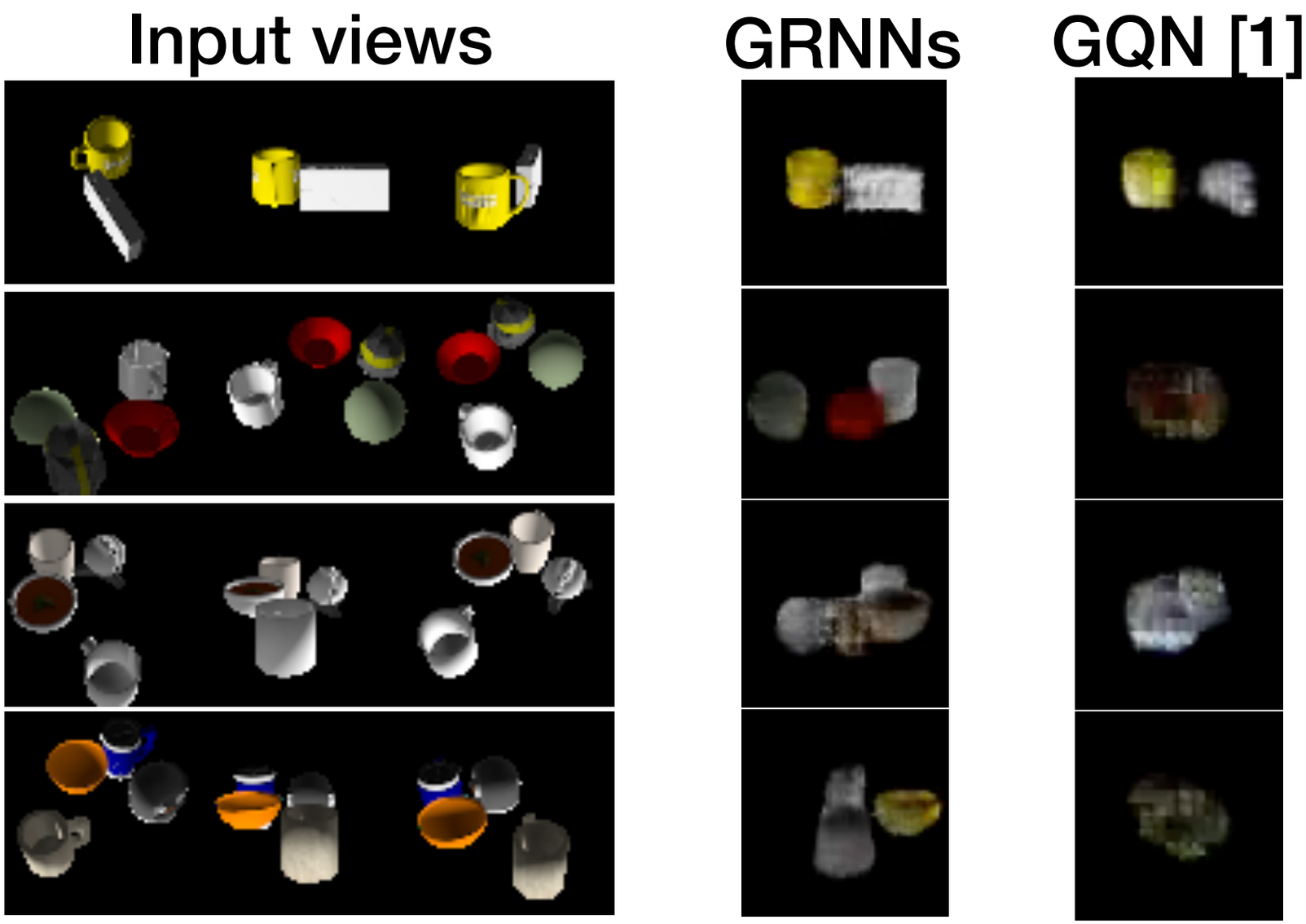
1. *Neural scene representation and rendering* DeepMind, Science, 2018

Results - view prediction



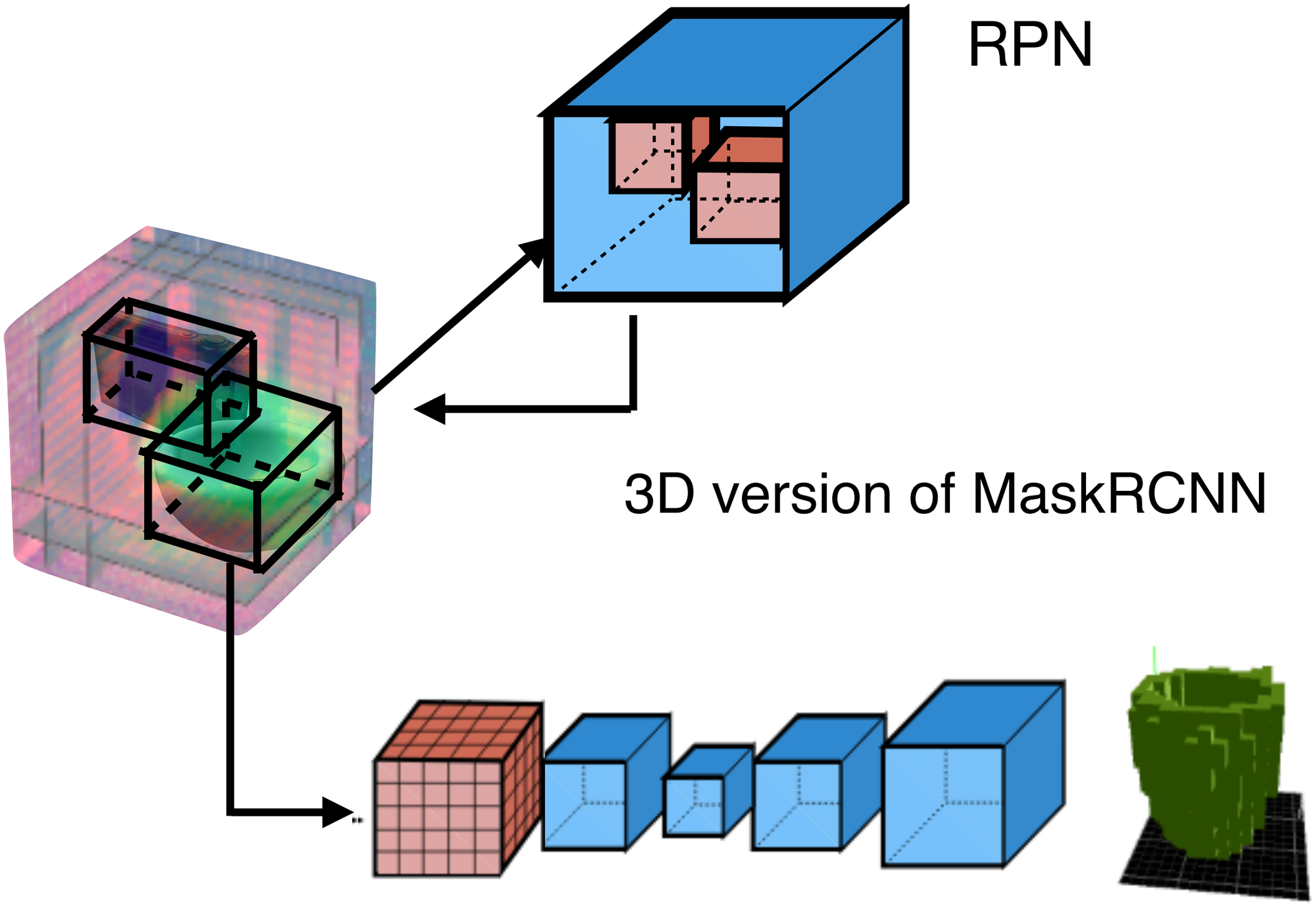
1. *Neural scene representation and rendering* DeepMind, Science, 2018

Results - view prediction



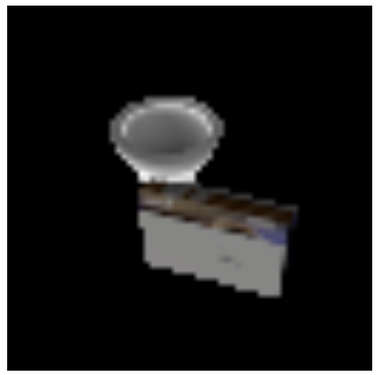
1. *Neural scene representation and rendering* DeepMind, Science, 2018

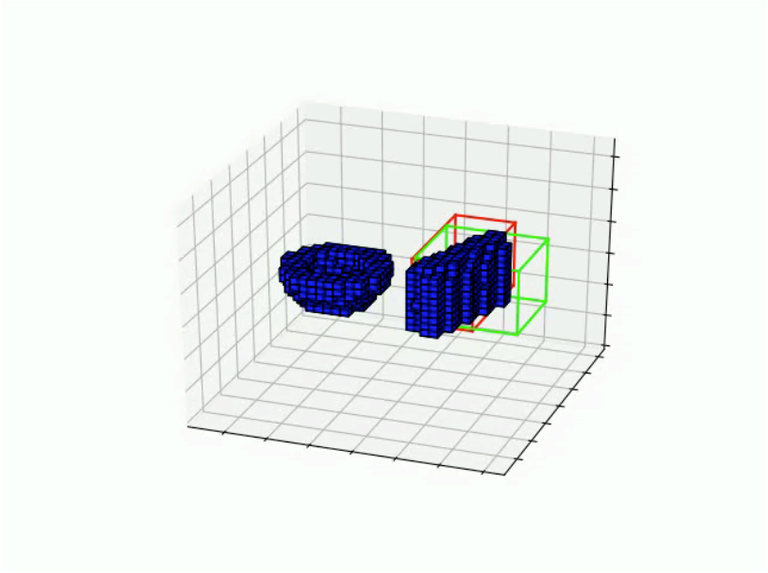
3D Object Detection



Results - 3D object detection

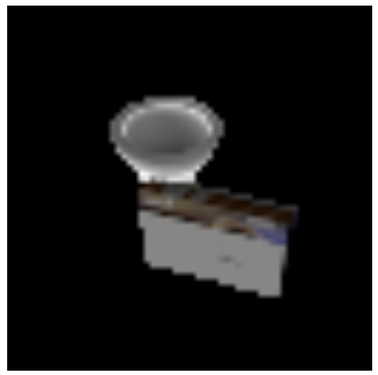
of input views

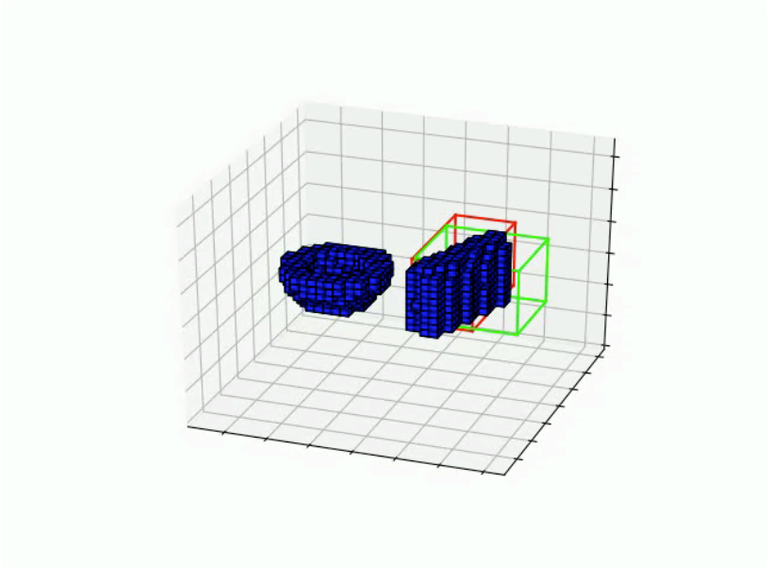


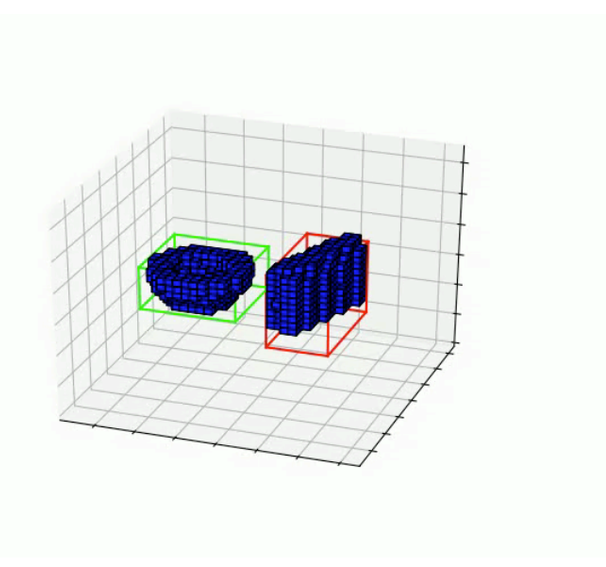


Results - 3D object detection

of input views

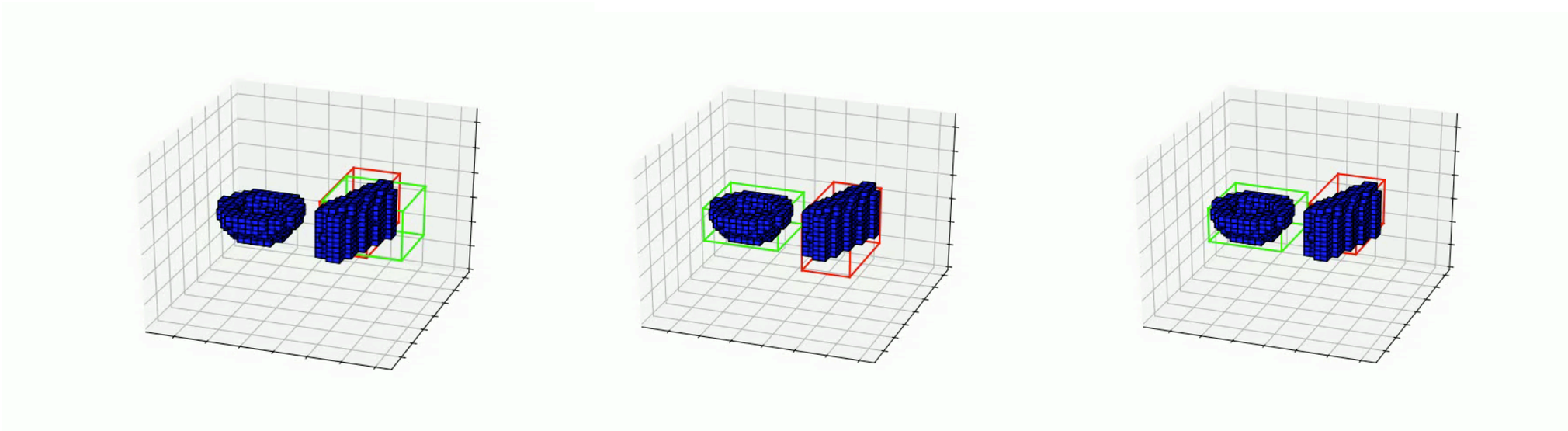




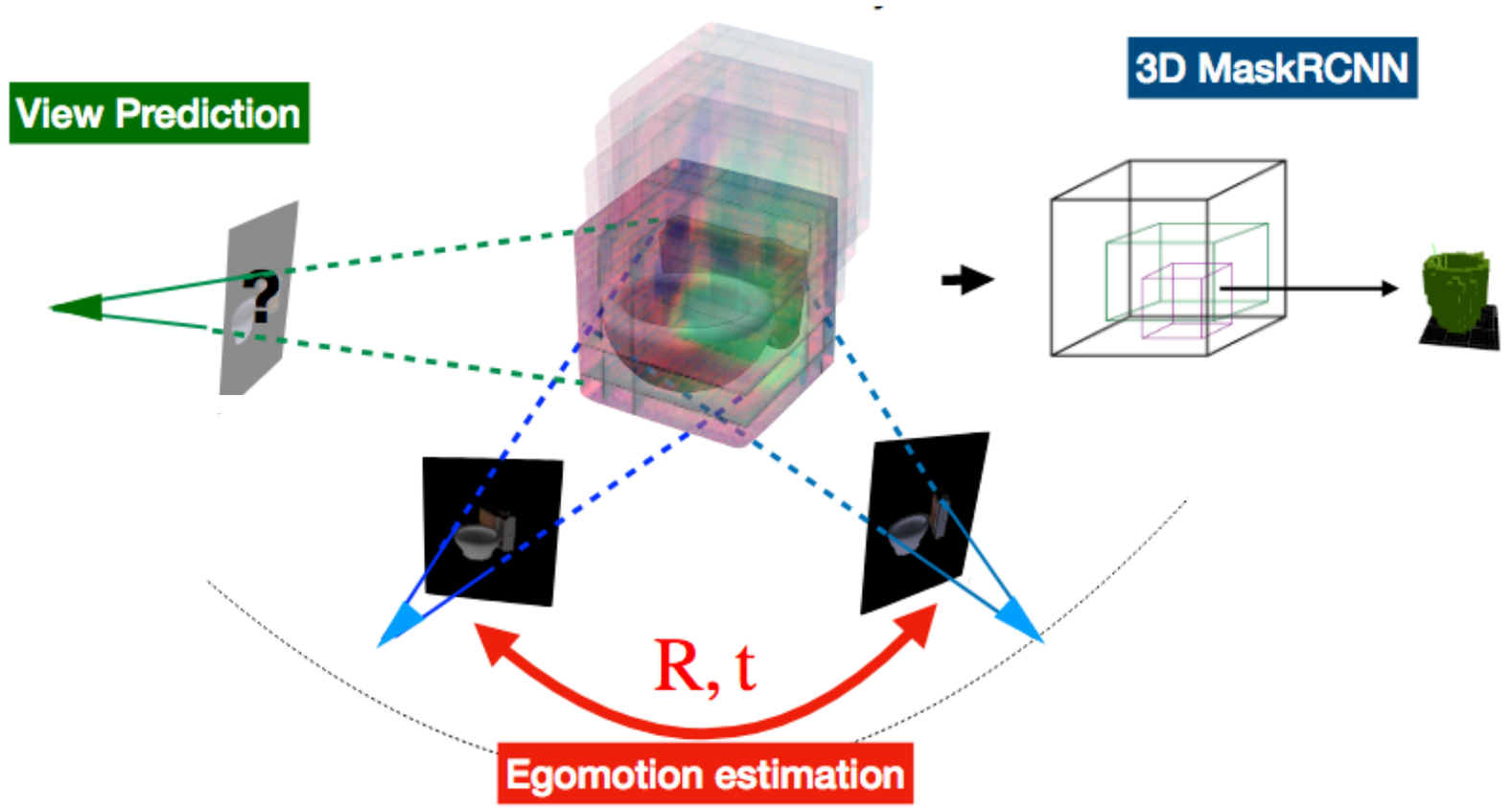


Results - 3D object detection

of input views



GRNNs



- Differentiable SLAM for better space-aware deep feature learning
- Generative model of scenes with a 3D bottleneck when trained from view prediction
- Generalize better than 2D models

Embodied visual recognition

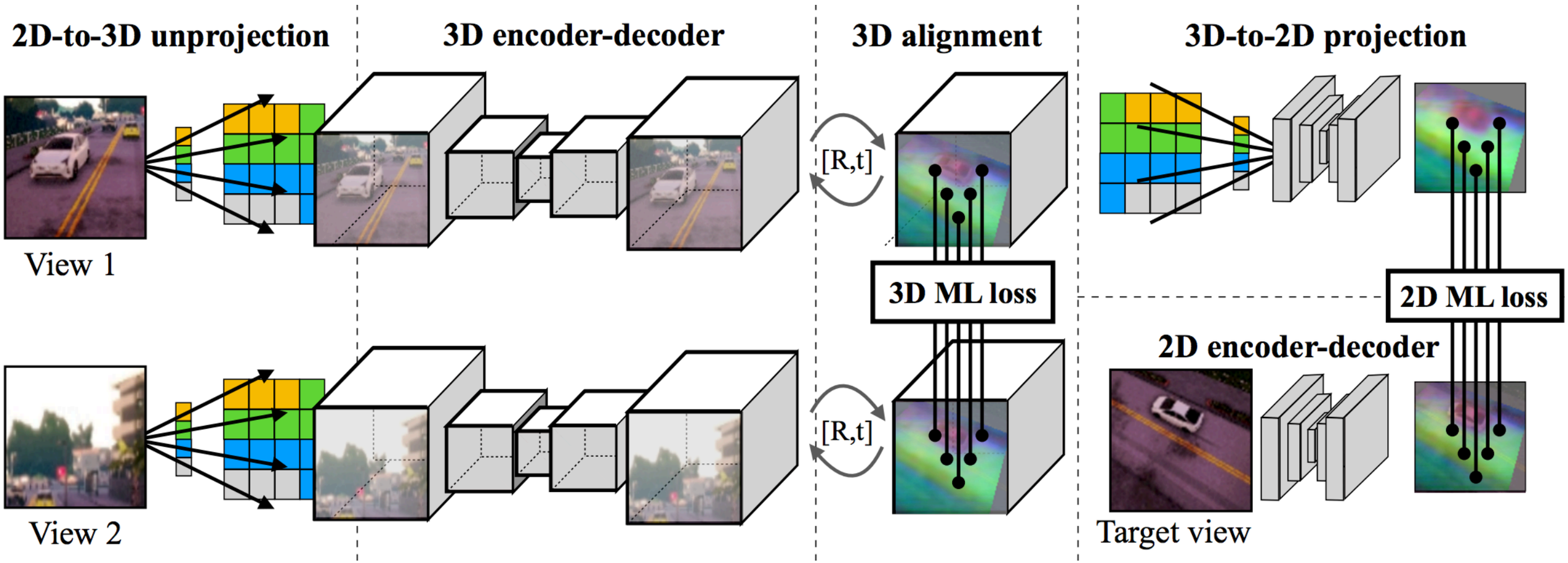
- Neural architectures for video recognition under arbitrary camera motion
- Learning image representations supervised by moving and watching objects move

Embodied visual recognition

- Can view prediction work beyond the toy simulation worlds we have just showed?
- Can view prediction learn features useful for object detection?

Yes, with a change of the loss function…

View-contrastive prediction



View-contrastive prediction

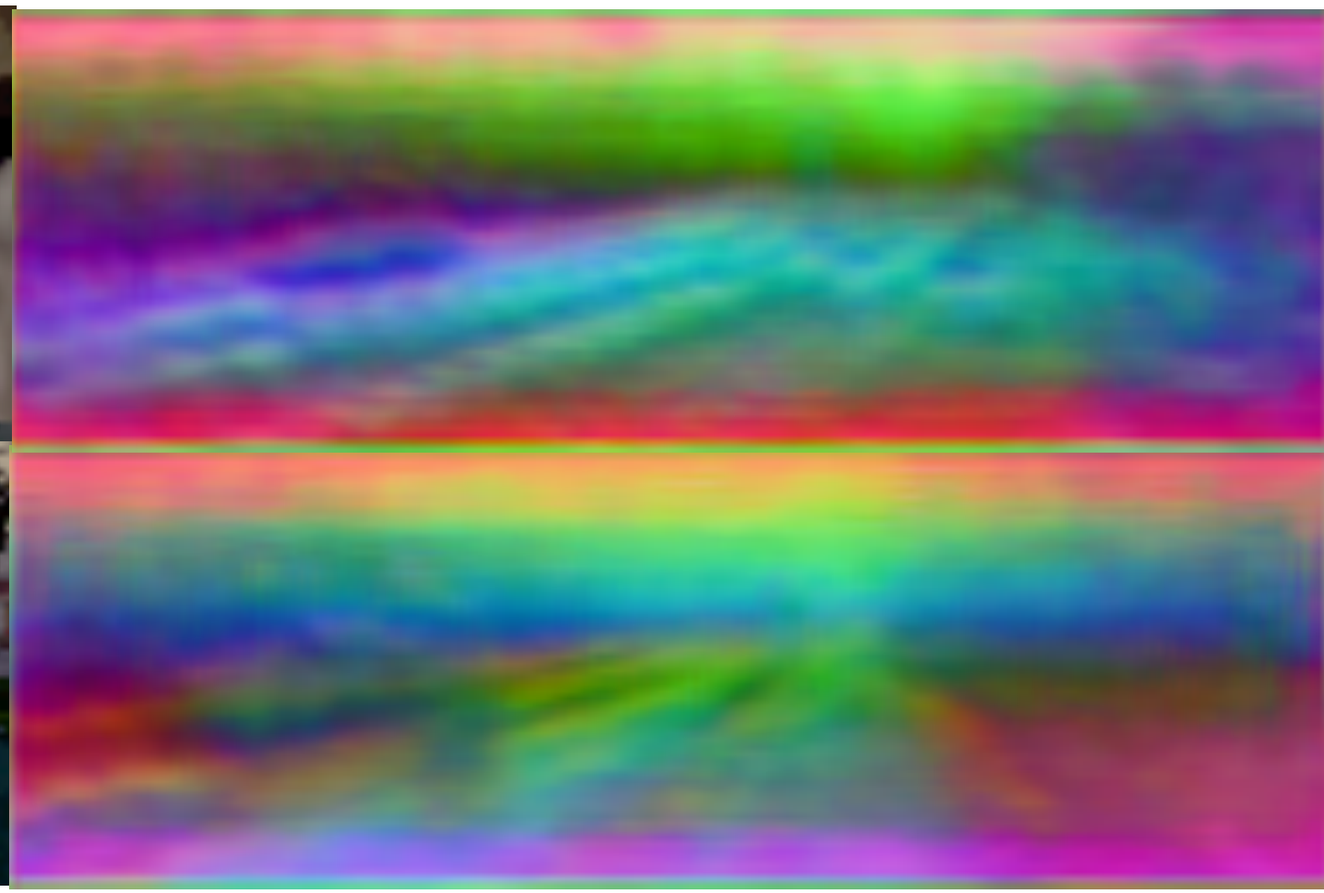
Target view

RGB estimates

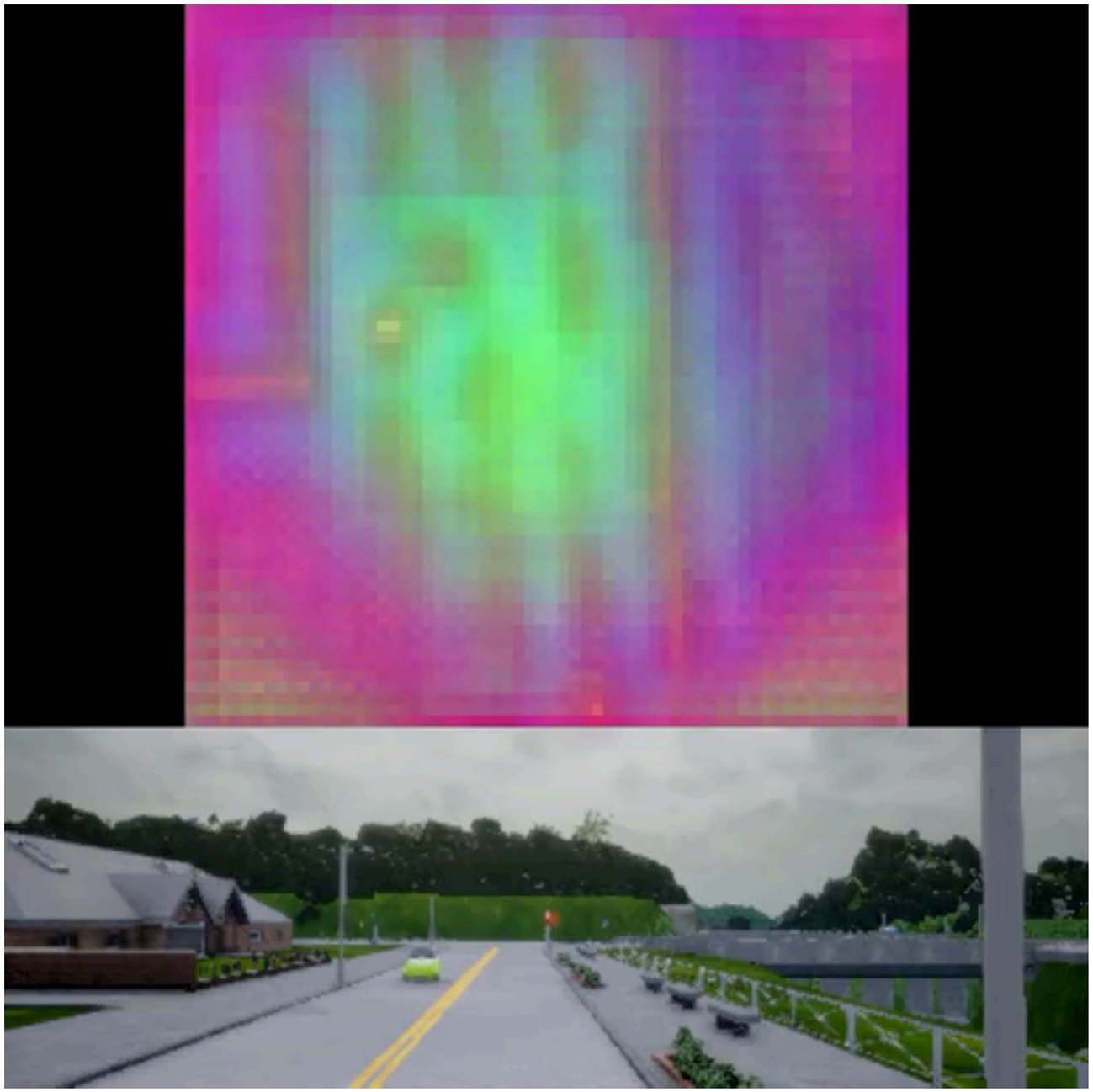
View-contrastive prediction

Target view

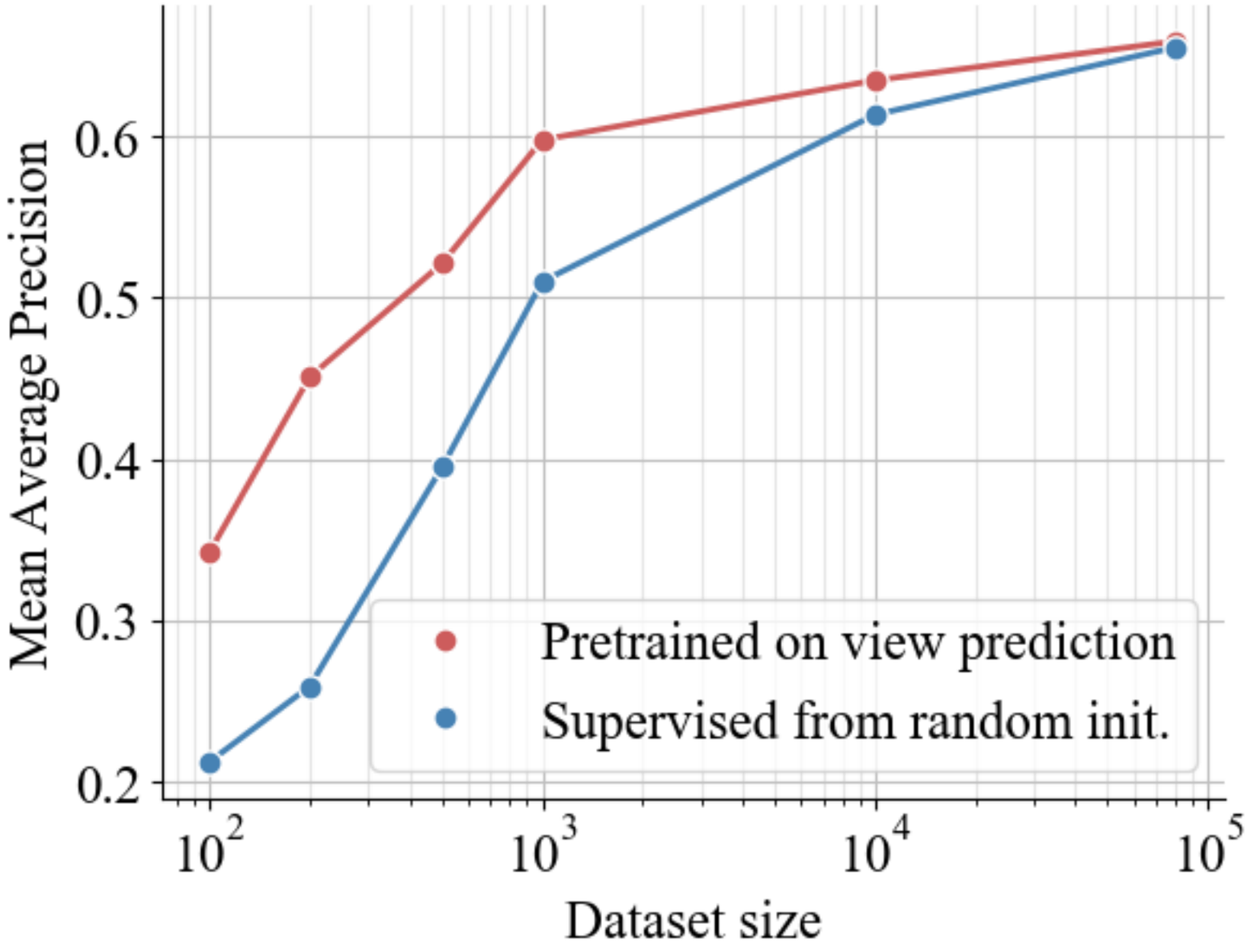
Embeddings

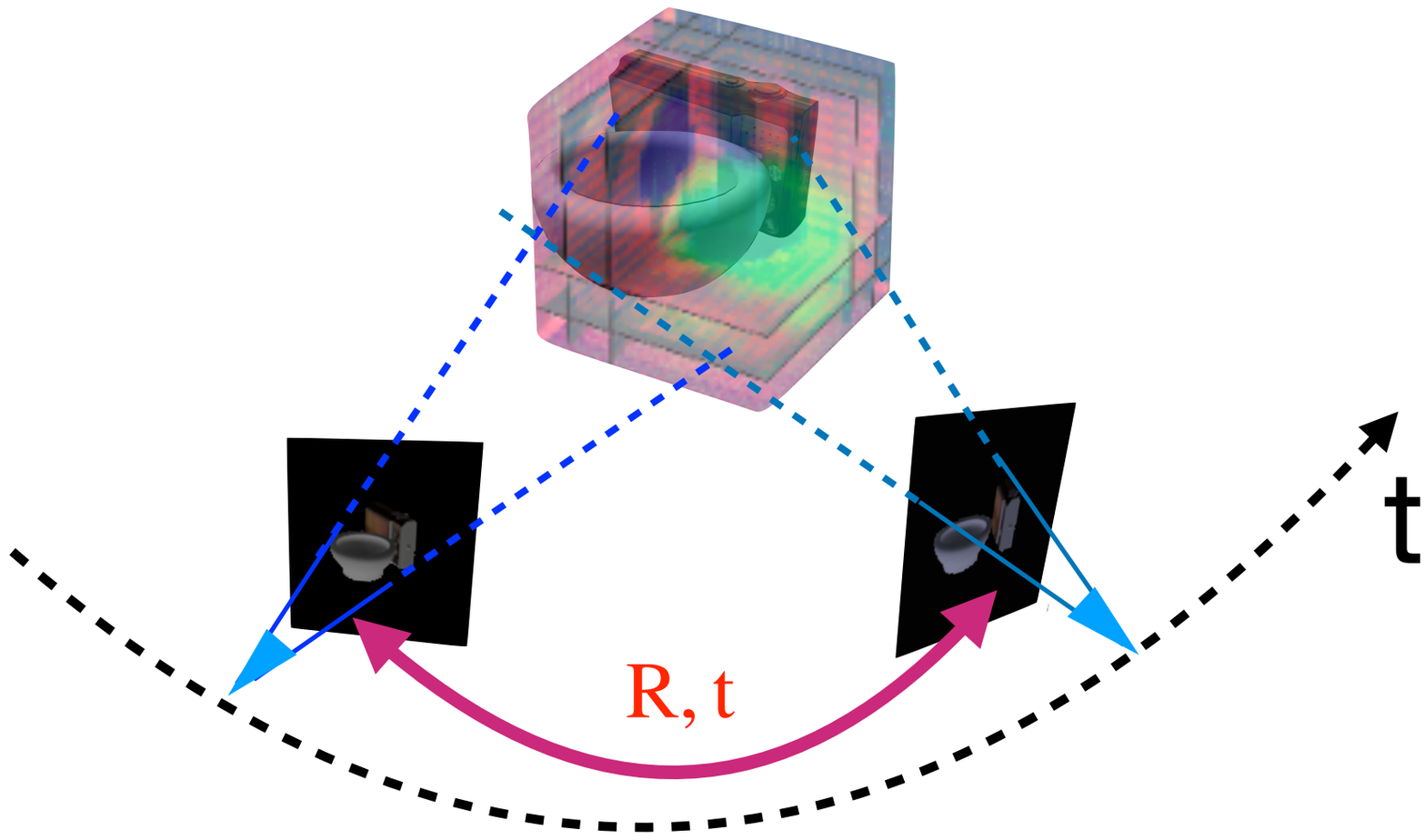


View-contrastive prediction

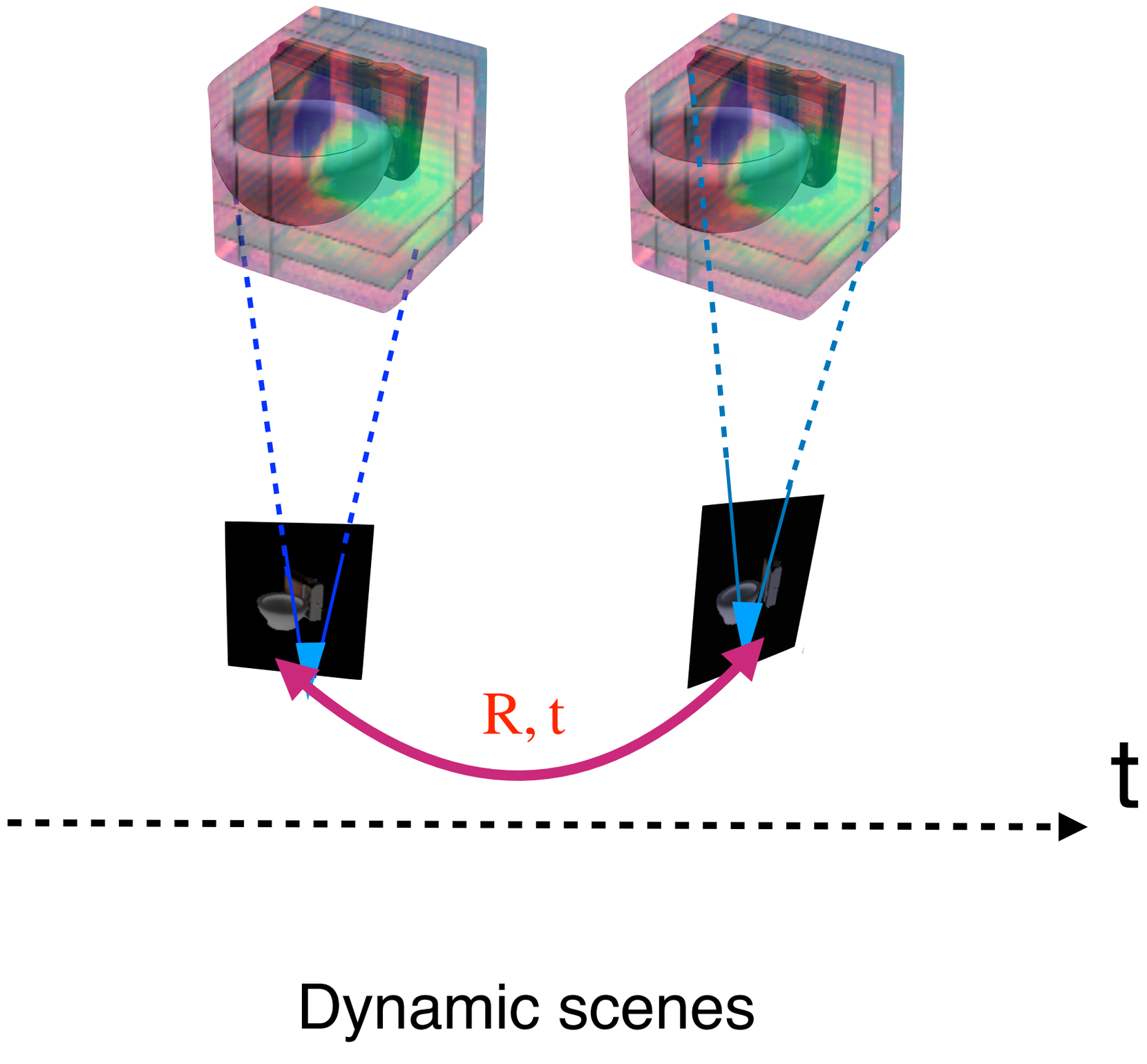


Semi-supervised learning of 3D object detection



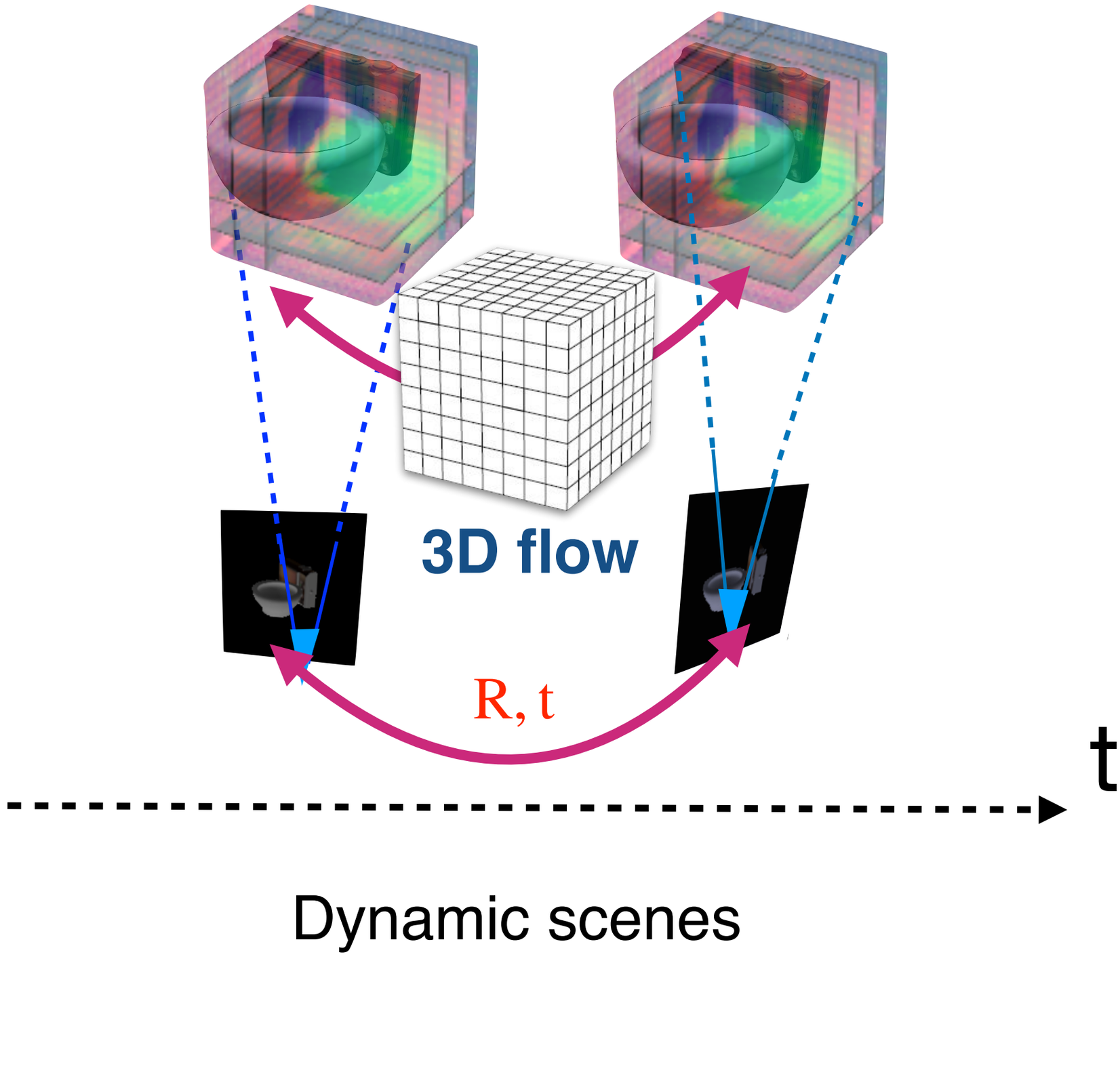


Static scenes



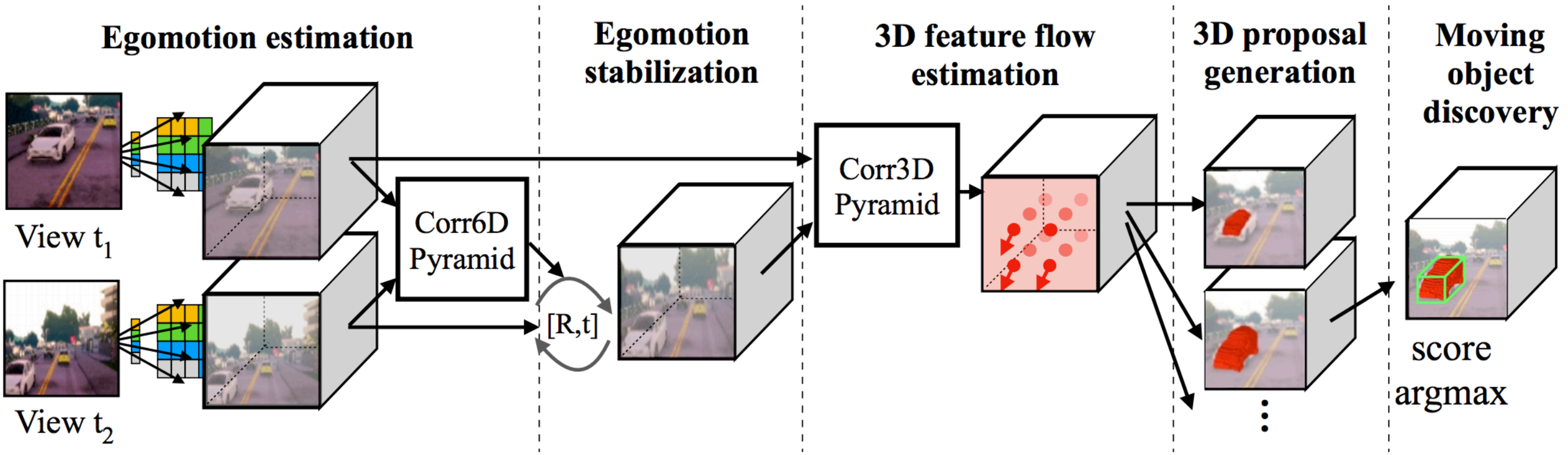
Dynamic scenes

3D imagination flow



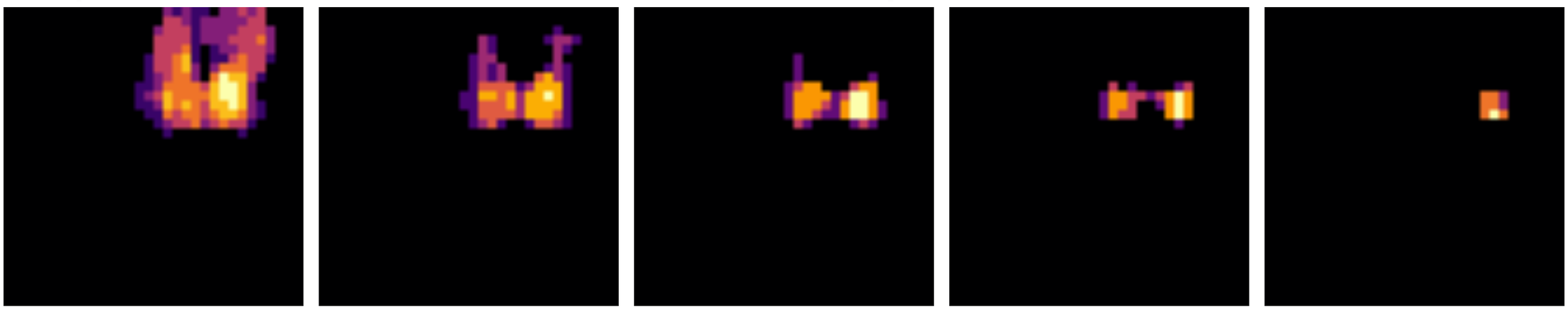
Dynamic scenes

3D object discovery

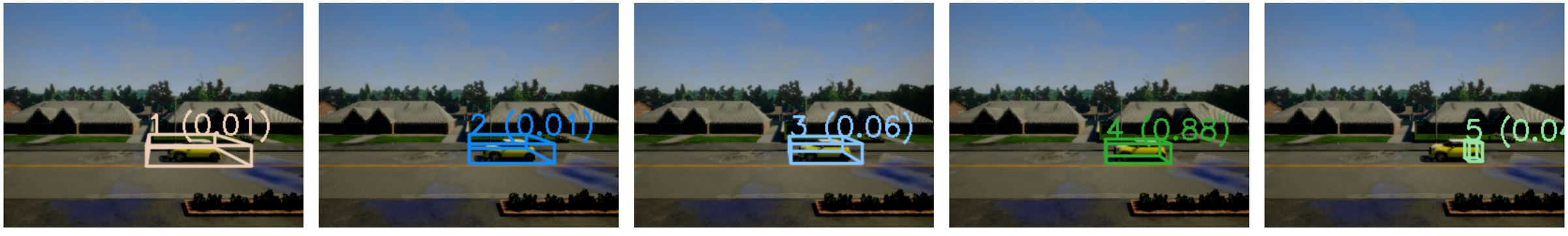


Results - unsupervised moving object segmentation

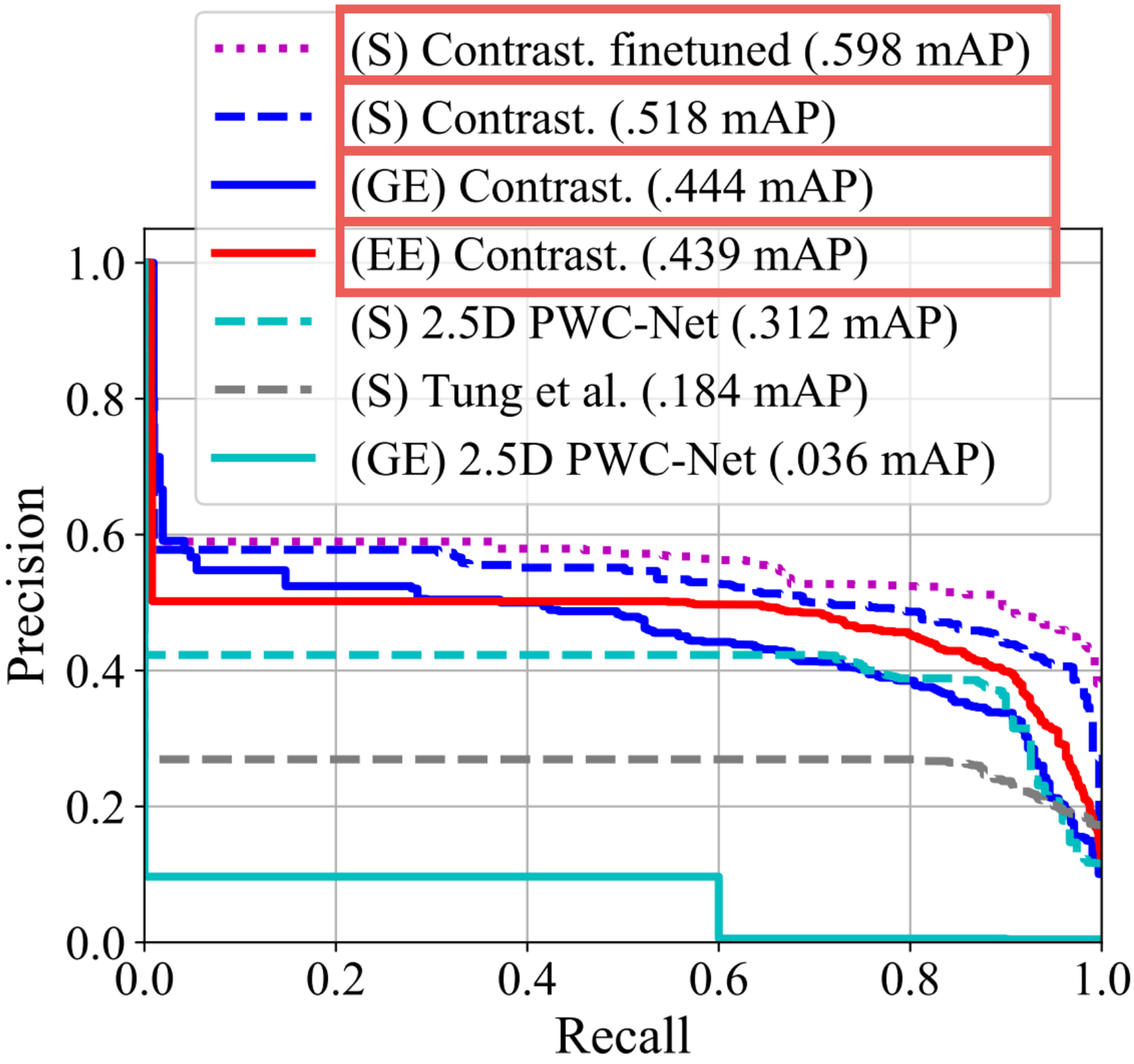
Top-down view of connected components in 3D flow field



Corresponding object boxes, with center-surround scores



Results - unsupervised moving object segmentation



Conclusion

Embodiment is the problem and the solution to visual recognition and common sense learning

Conclusion

`` *We must perceive in order to move, but we must also move in order to perceive”*

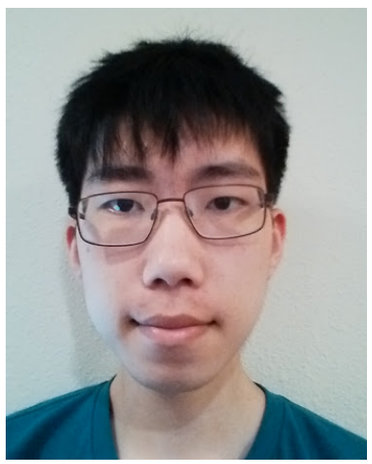
JJ Gibson

“If we figure out how to do 3D perception correctly, no one will use 2D again, same way when color TV was invented no one used black and white”

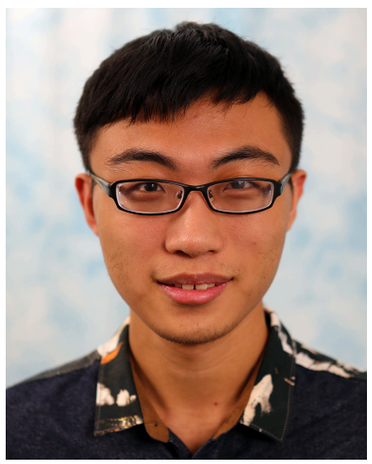
Yaser Sheikh

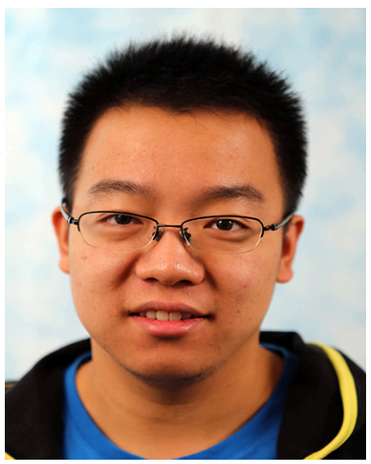
Thank you!


Fish Tung


Ricson Chen


Adam Harley


Fangyu Li


Xian Zhou


Shrinidhi K. Lakshmikanth

- **Learning spatial common sense with geometry-aware recurrent networks**, Tung et al., CVPR 2019,
- **Embodied View-Contrastive 3D Feature Learning**, Harley et al., arxiv