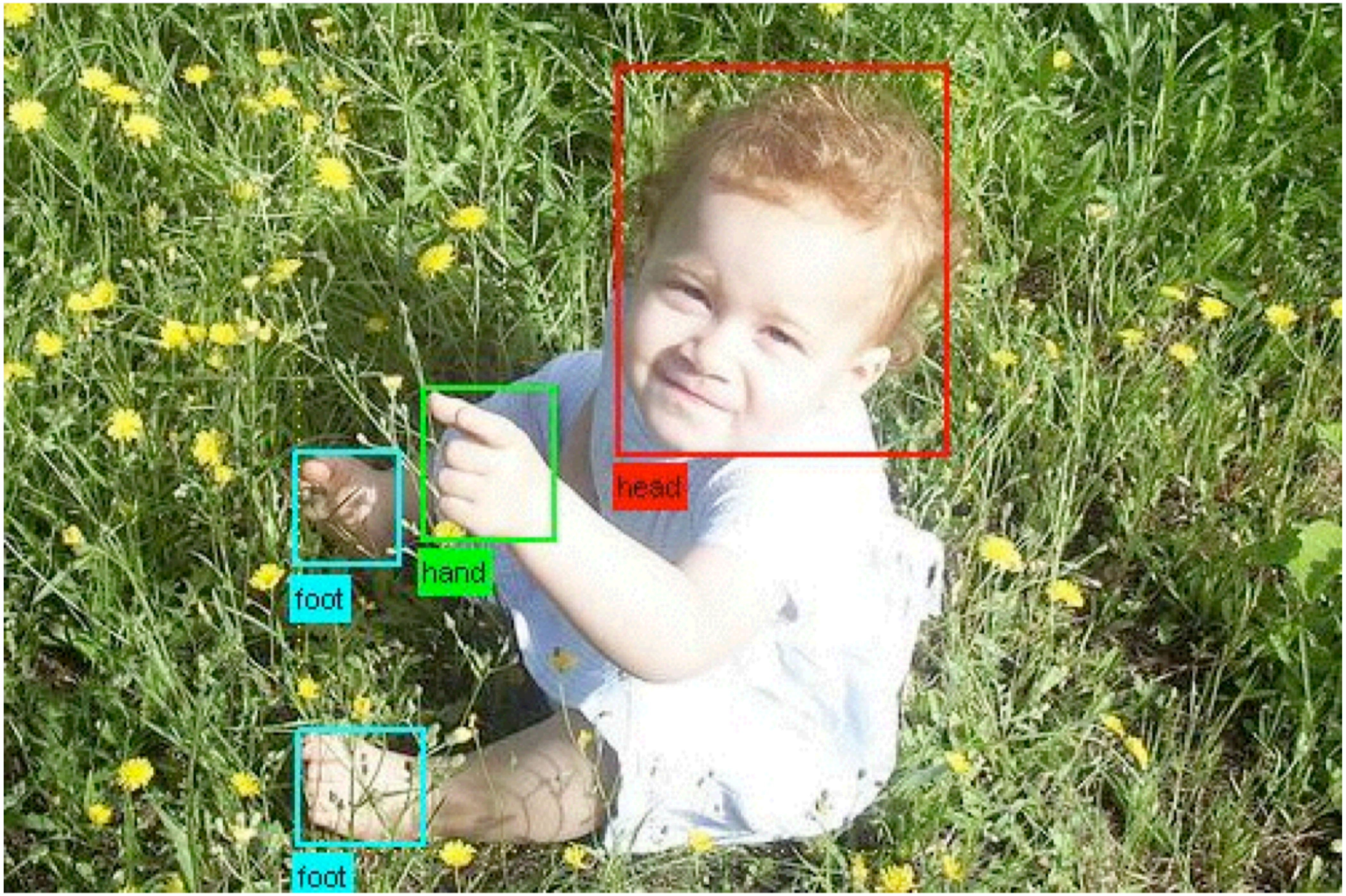
Geometry-Aware Deep Visual Learning

Katerina Fragkiadaki

How this talk fits the workshop

- We will discuss new neural architectures for video understanding and feature learning without human annotations
- We will still use SGD to train the models

What is the goal of computer vision?



label image pixels, detect and segment objects

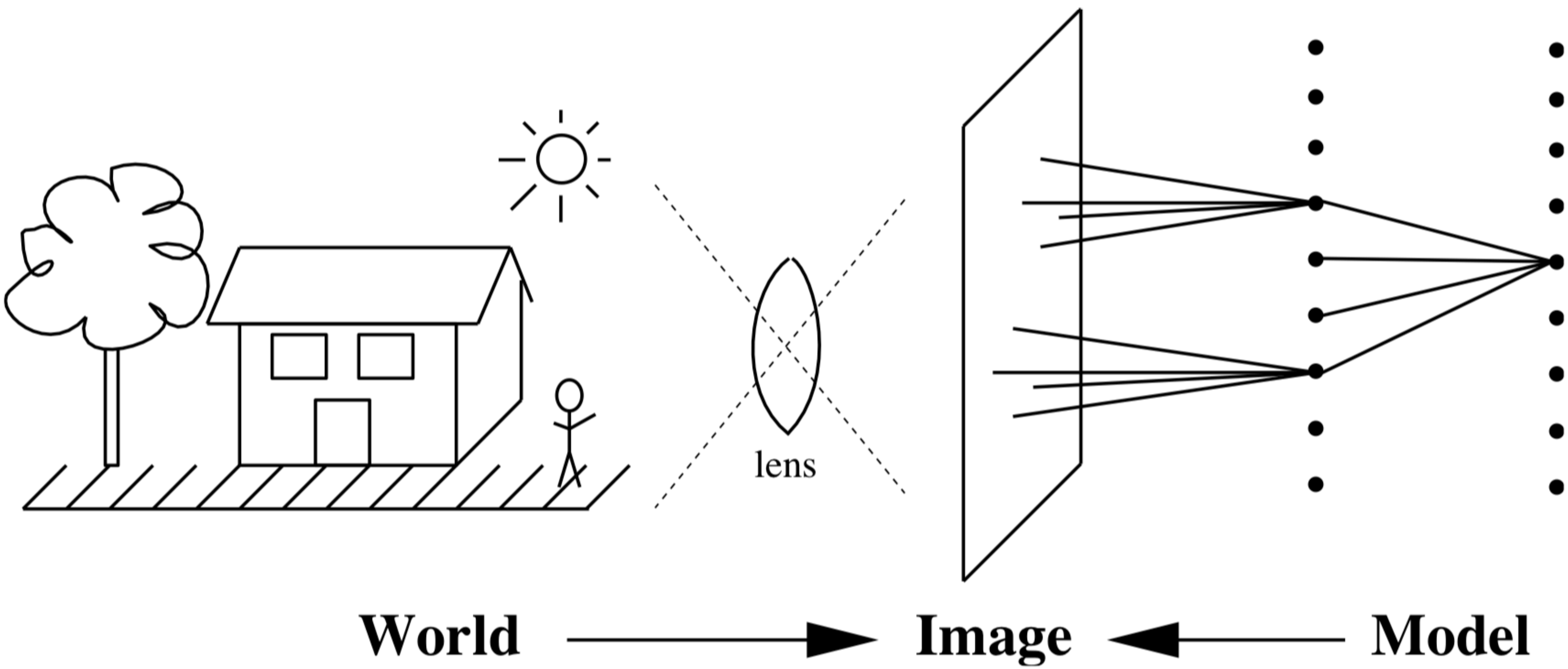
Image from Bruno Olshausen

label image pixels, detect and segment objects

K. He et al., MaskRCNN, 2017

Registration against known HD maps, 3D object detection, 3D motion forecasting

Image Understanding as Inverse Graphics



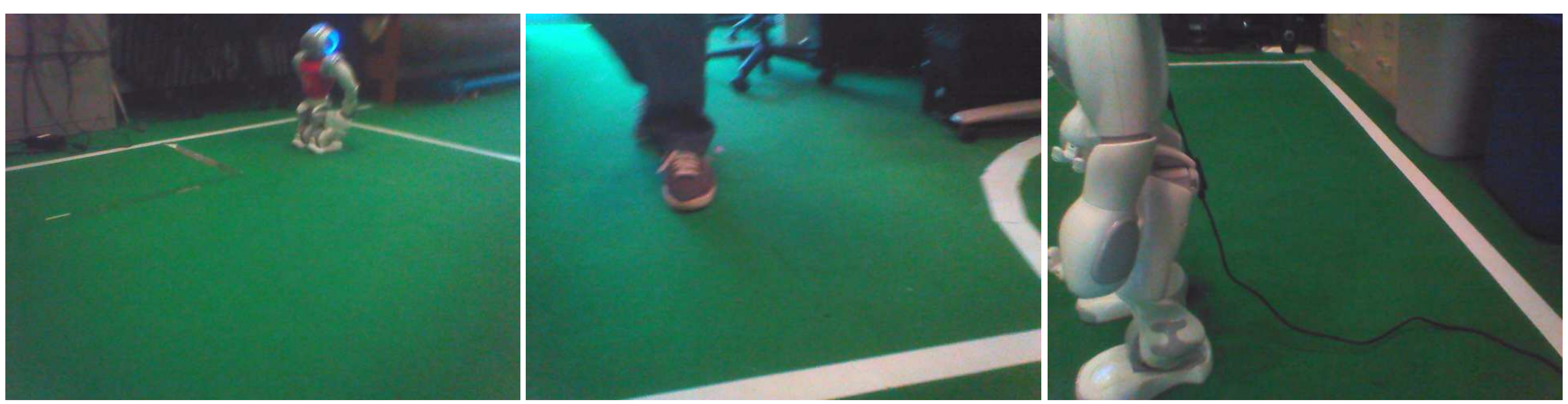
A reasonable answer: the goal of computer vision is task specific

Internet Vision

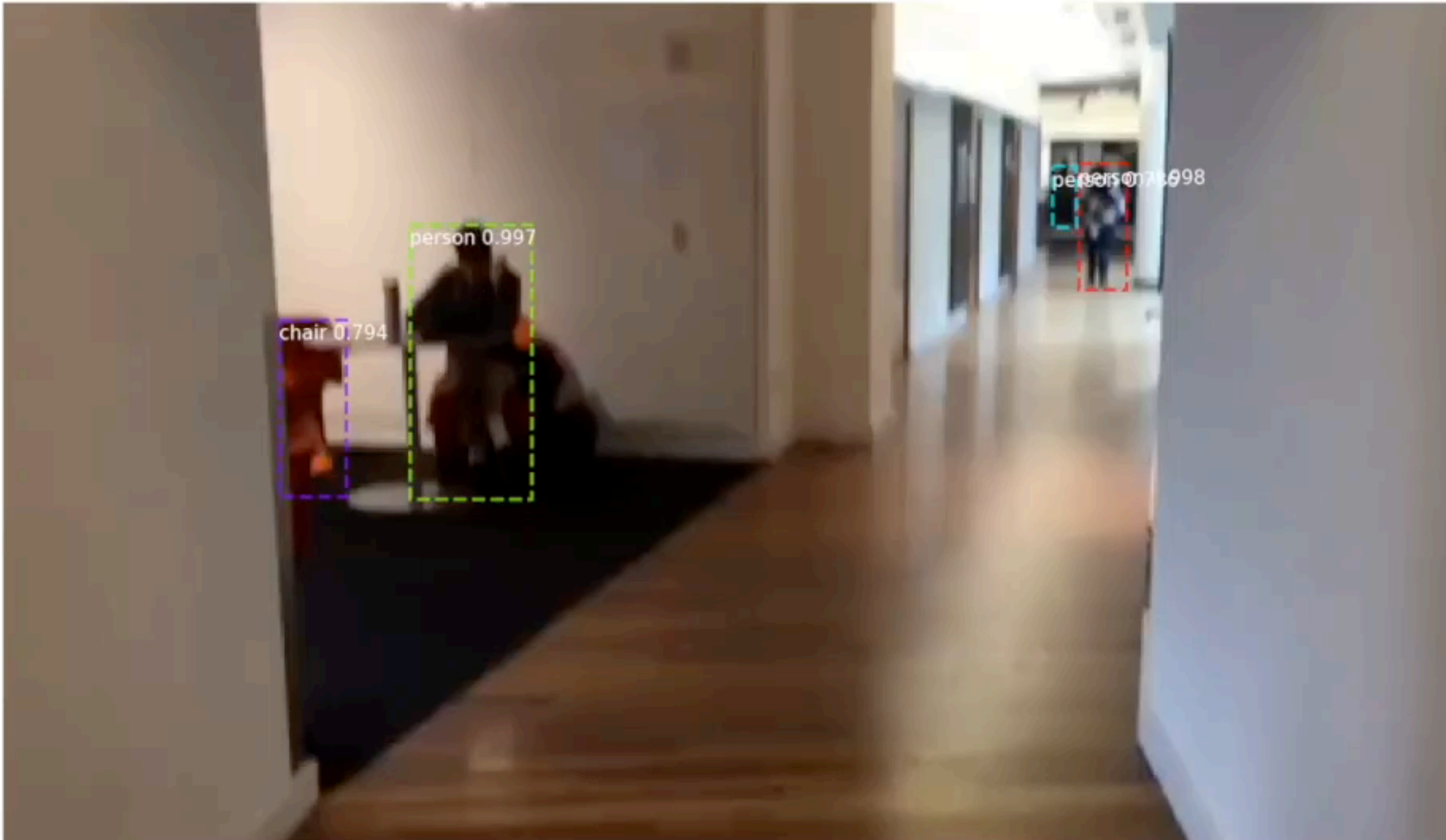
Photos taken by people (and uploaded on the Internet)

Mobile (Embodied) Computer Vision

Photos taken by a NAO robot during a robot soccer game



Our detectors may not work very well here…

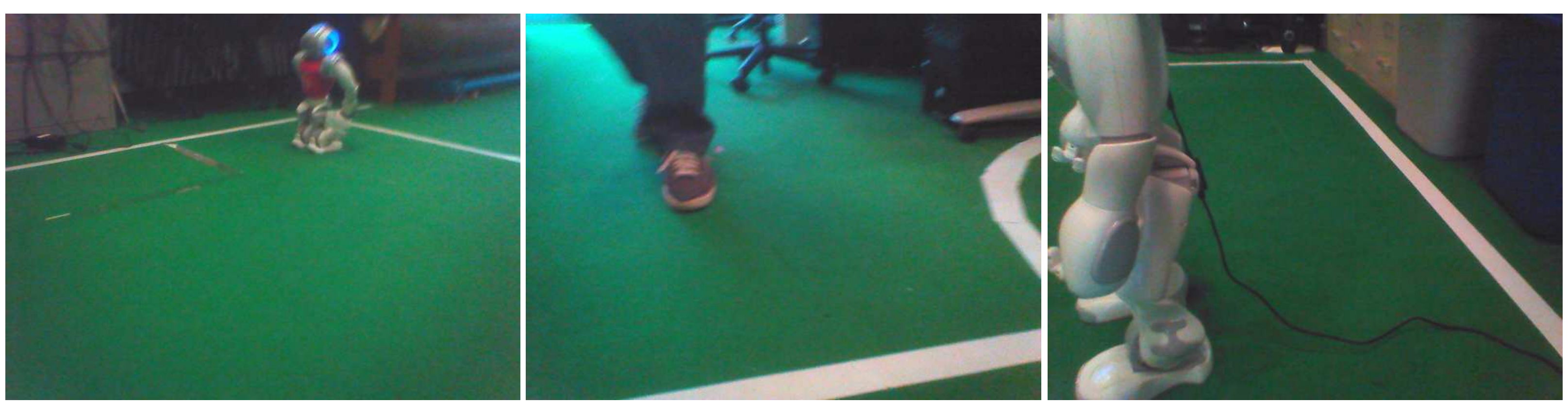

person 0.997
chair 0.794

Internet Vision

Photos taken by people (and uploaded on the Internet)

Mobile (Embodied) Computer Vision

Photos taken by a NAO robot during a robot soccer game



Our detectors may not work very well here…

Do we have more suitable models for this domain?

Why Embodied Computer Vision Matters

1. Agents that move around in the world, perceive the world and accomplish tasks is (close to) the goal of AI research
2. It *may* be the key towards unsupervised visual feature learning

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

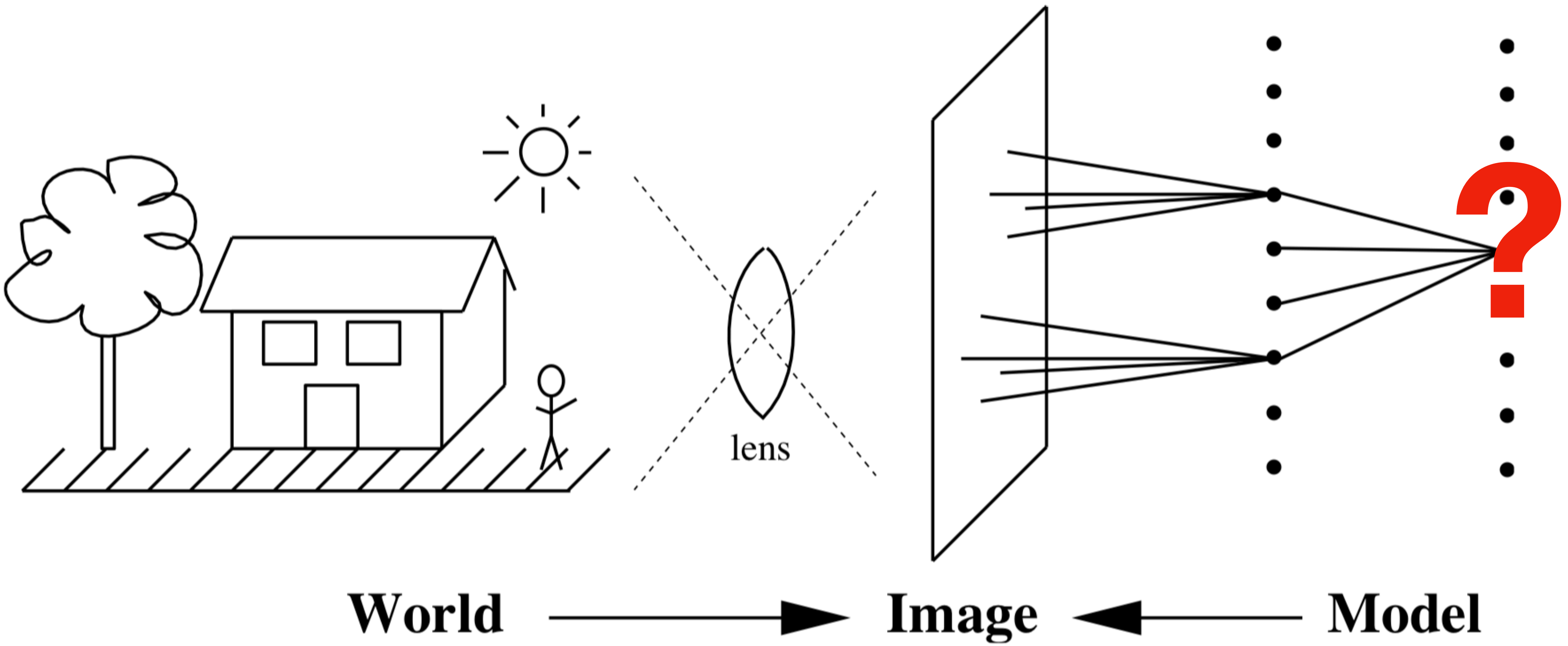
JJ Gibson

Ecological Approach to Visual Perception, Gibson, 1979

Internet and Mobile Perception have developed independently and have each made great progress

- Internet vision has trained great DeepNets for image labelling and object detection+segmentation
- Mobile computer vision has produced great SLAM (Simultaneous Localization and Mapping) methods

Image Understanding as Inverse Graphics

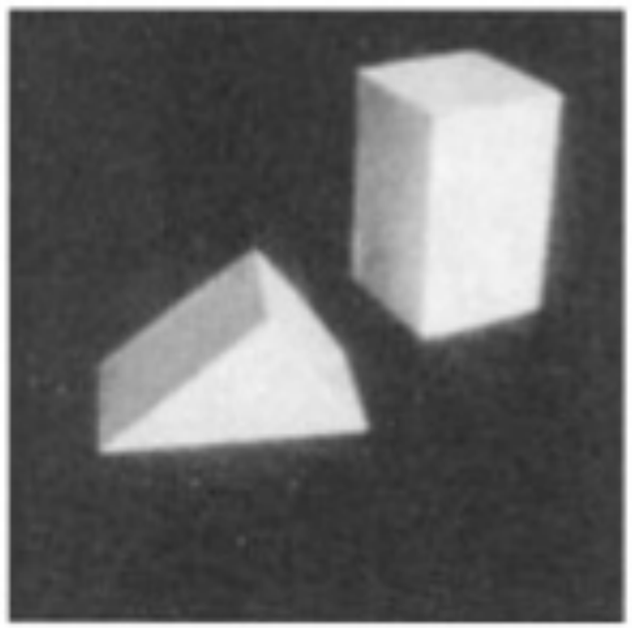


Should we be engineering a different model for every domain?

Image Understanding as Inverse Graphics

Blocks world

Larry Roberts



Input image

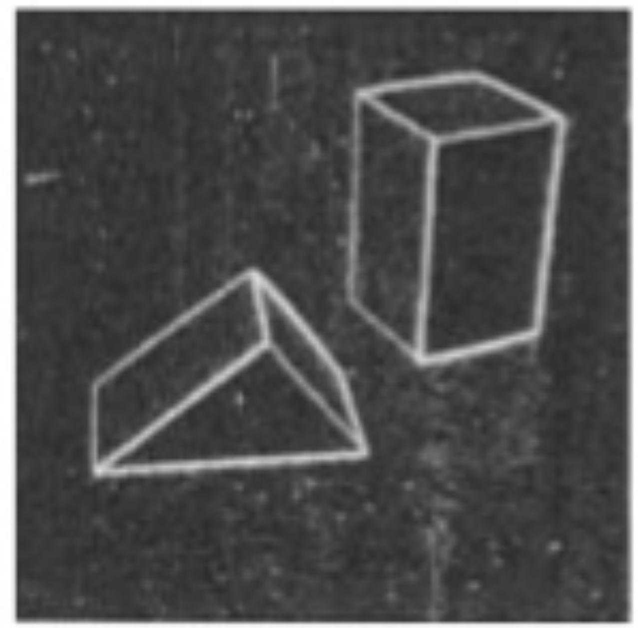
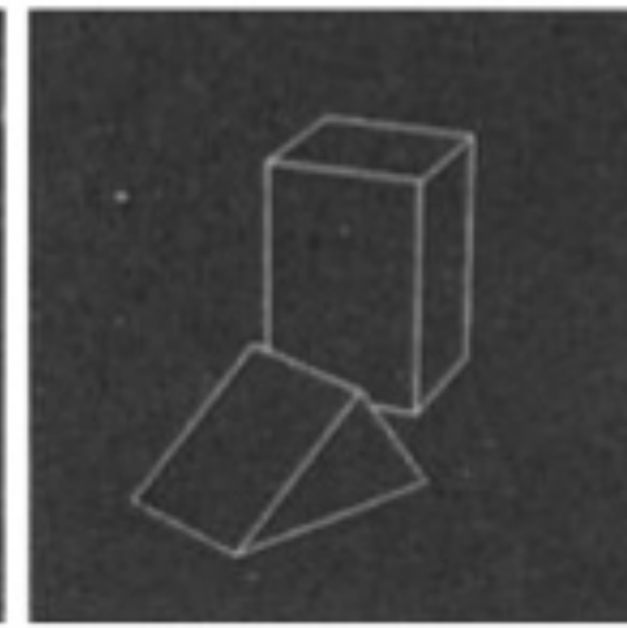


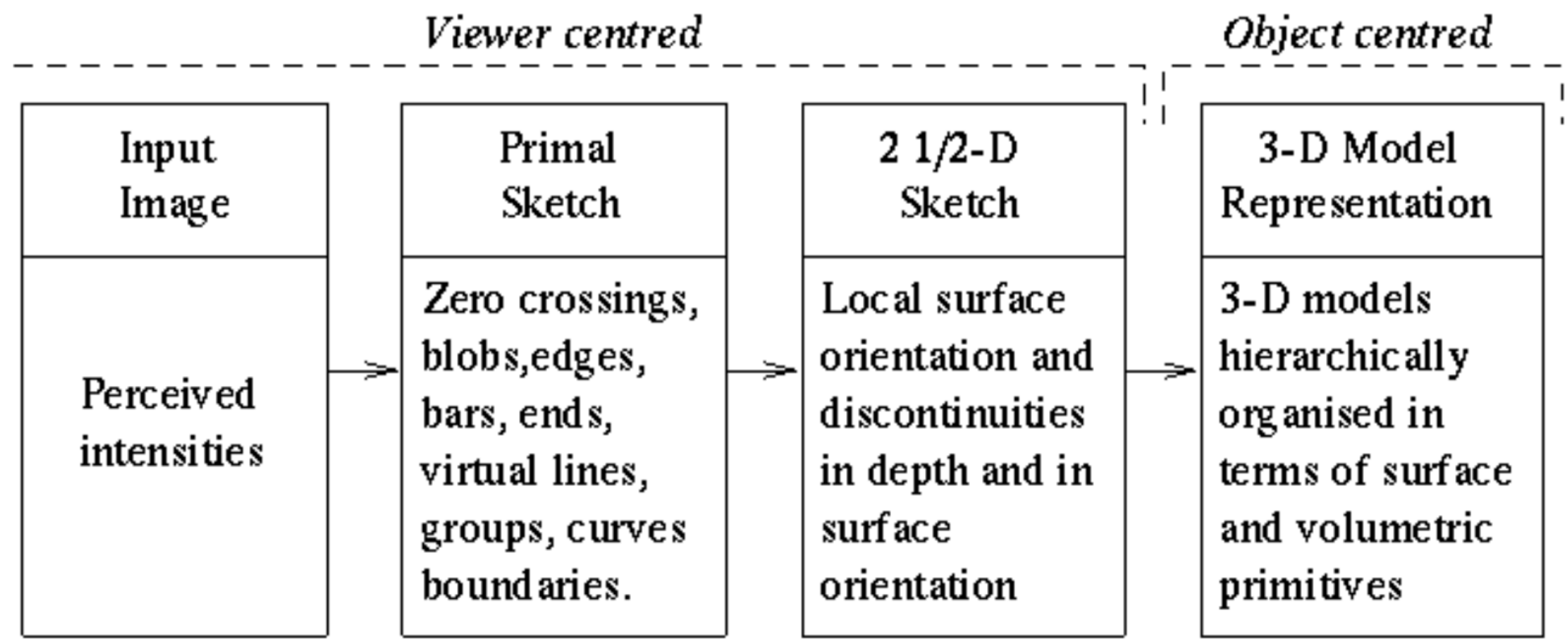
Image gradient



Computed 3D model rendered from a new viewpoint

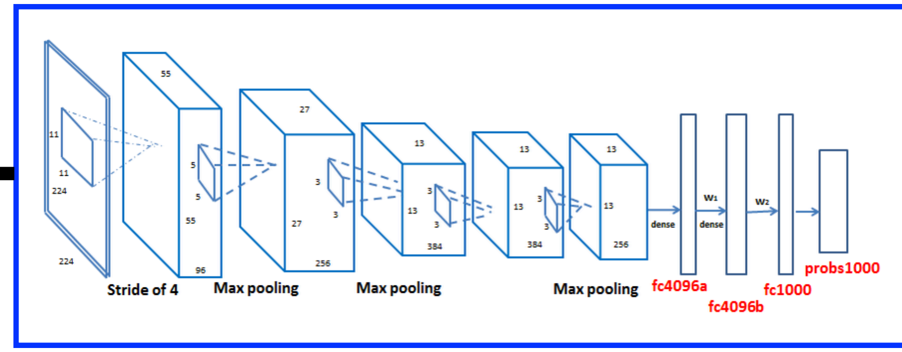
Machine perception of Three-Dimensional solids, MIT 1965

Image Understanding as Inverse Graphics



David Marr 1982

3D Models are impossible and unecessary



Steering angle

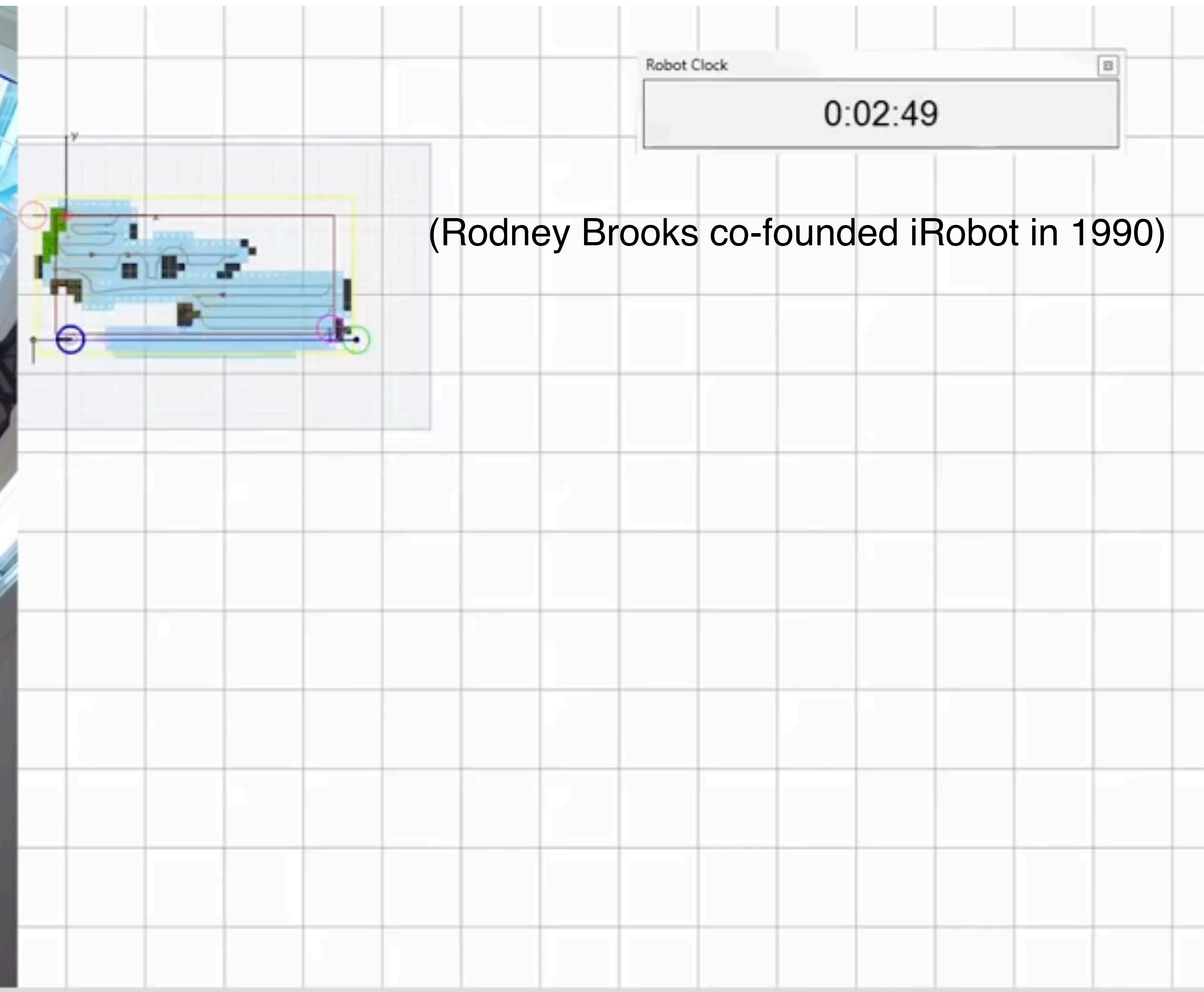
``Internal world models which are complete representations of the external environment, besides being impossible to obtain, are not at all necessary for agents to act in a competent manner."

``…(1) eventually computer vision will catch up and provide such world models—-I don't believe this based on the biological evidence presented below, or (2) complete objective models of reality are unrealistic and hence the methods of Artificial Intelligence that rely on such models are unrealistic."

"*Intelligence without reason*", IJCAI, Rodney Brooks (1991)

25 years later

iRobot vacuum cleaner is building a map!

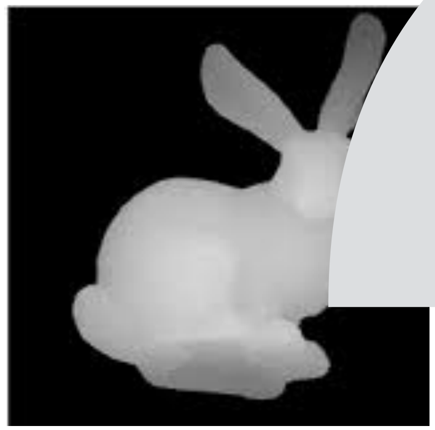


(Rodney Brooks co-founded iRobot in 1990)

To 3D or not to 3D?

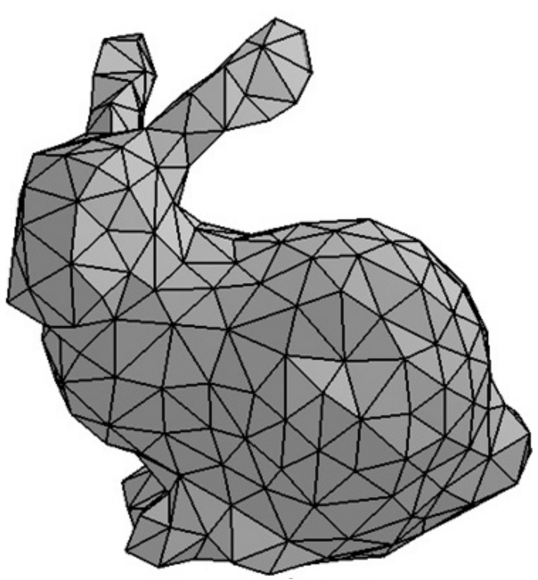
And if to 3D, what 3D representation to use?

depth map

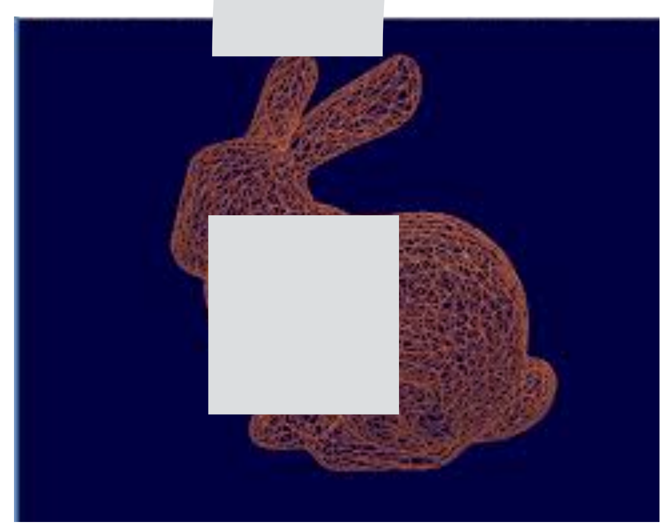


face normals

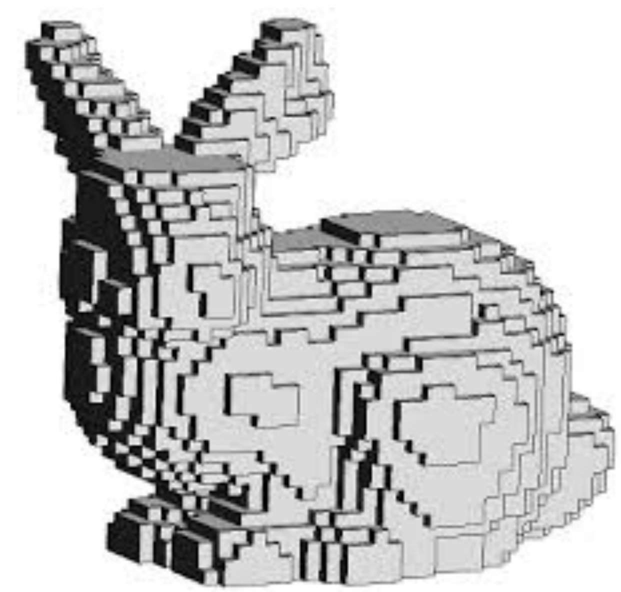
3d mesh



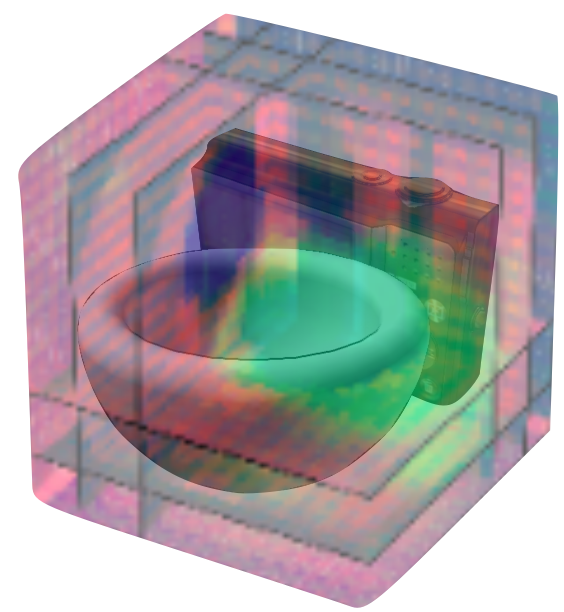
3d po loud



3d voxel occupancy



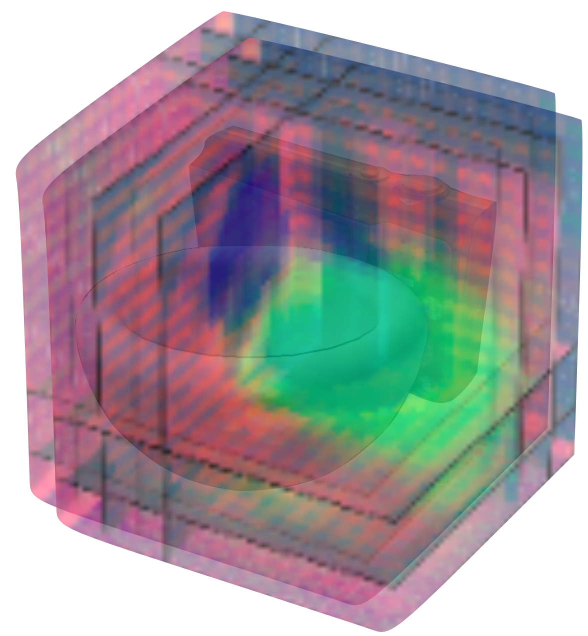
This talk: To 3D using 3D feature tensors



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

3 spatial dimensions, 1 feature dimension

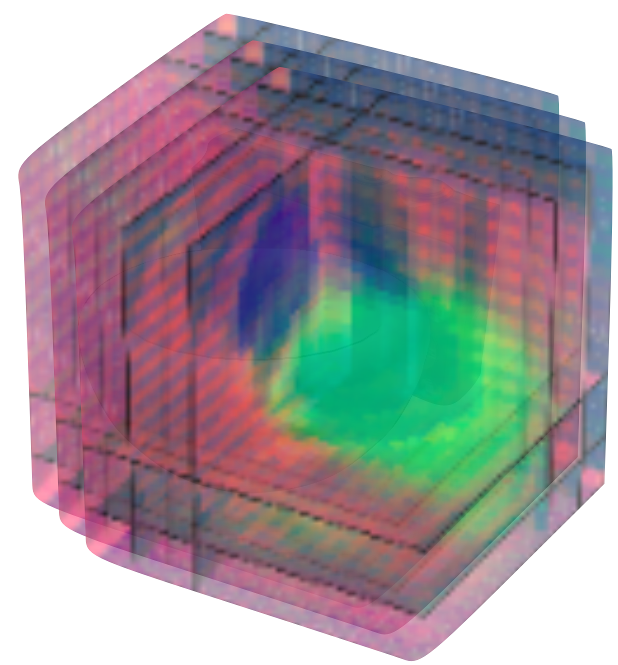
This talk: To 3D using 3D feature tensors



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

3 spatial dimensions, 1 feature dimension

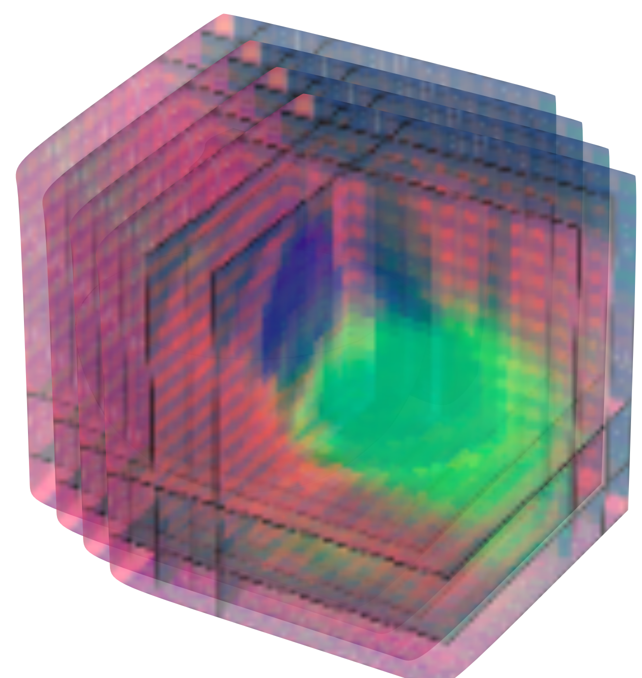
This talk: To 3D using 3D feature tensors



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

3 spatial dimensions, 1 feature dimension

This talk: To 3D using 3D feature tensors

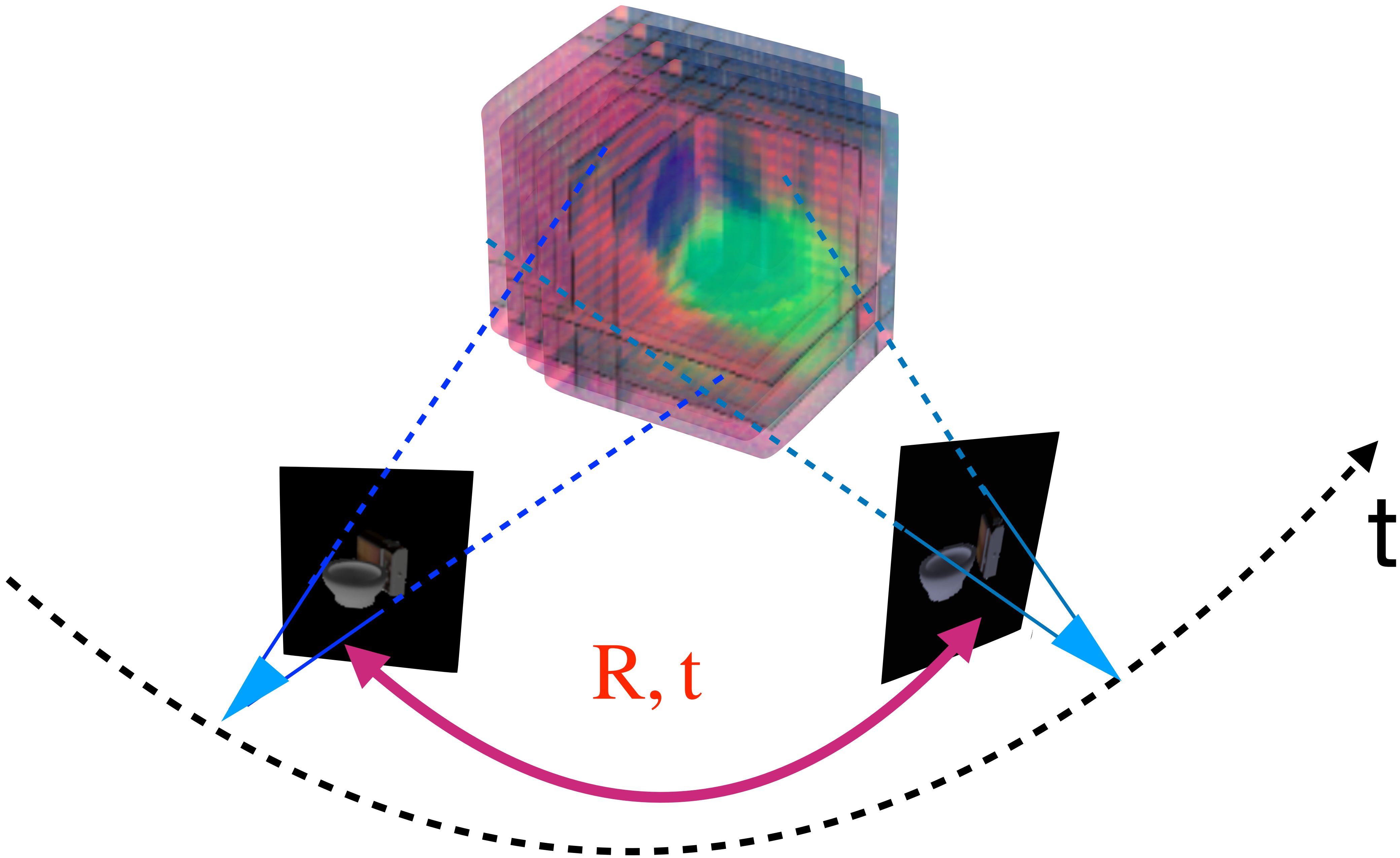


$H \times W \times D \times C$

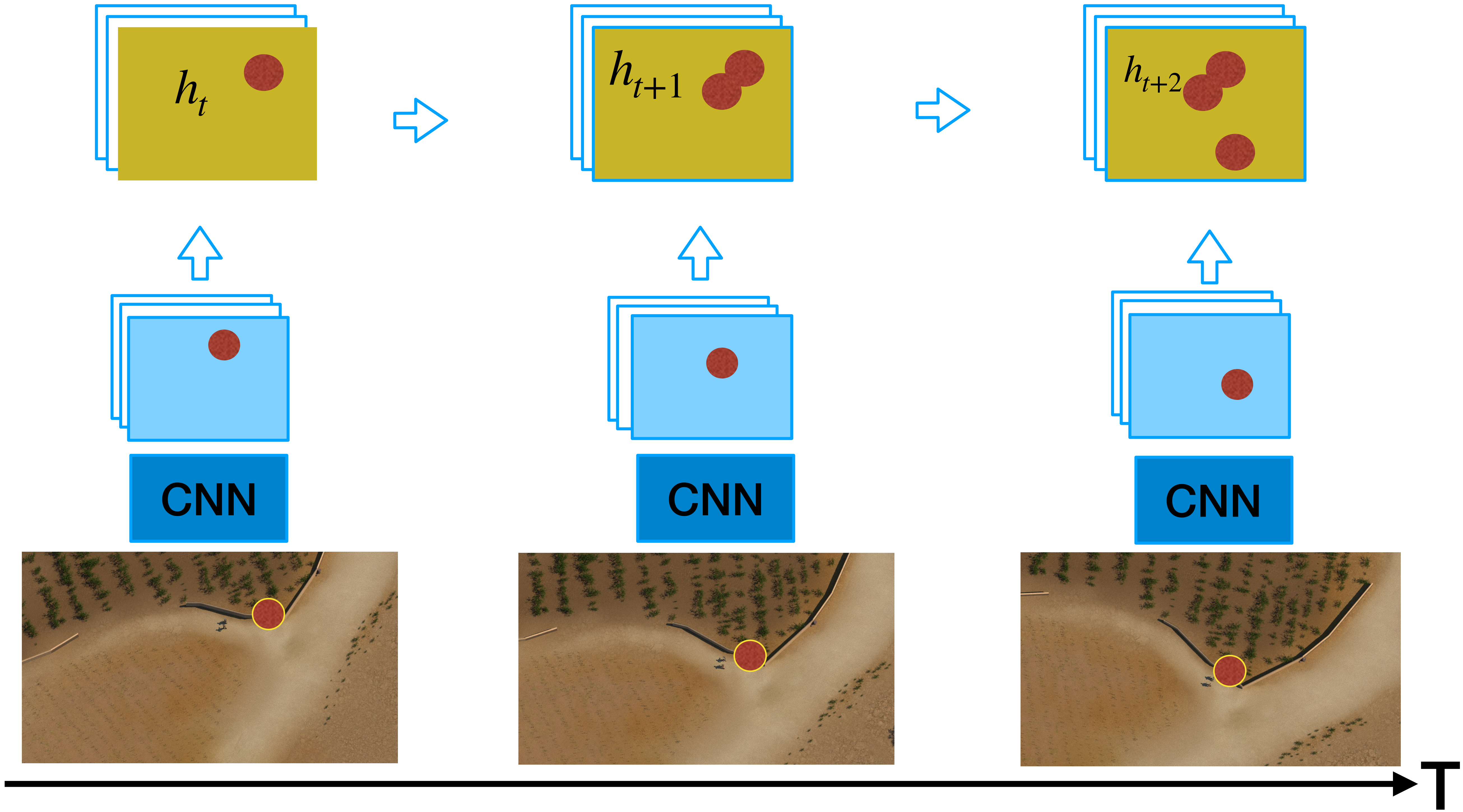
3 spatial dimensions, 1 feature dimension

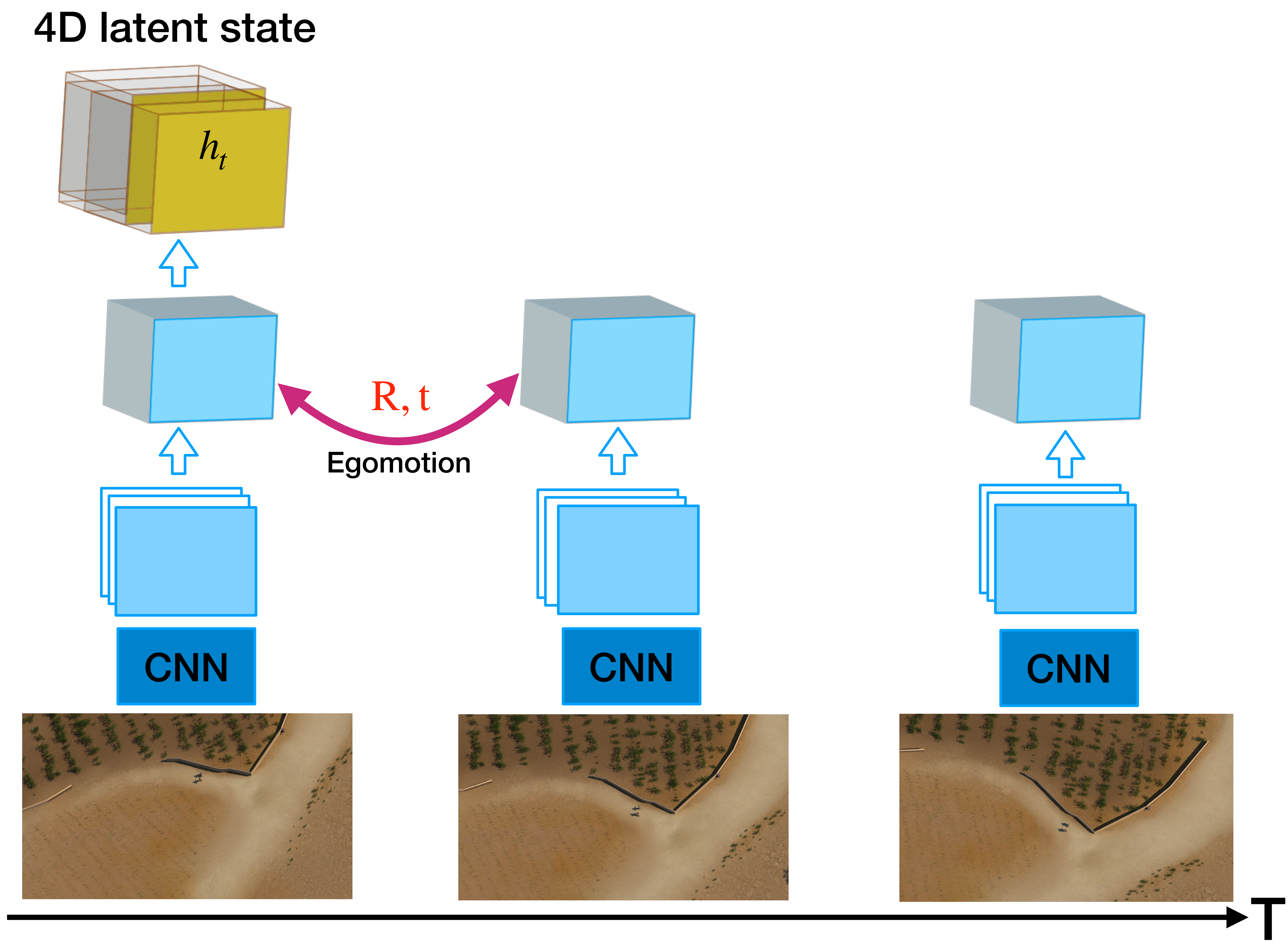
Geometry-Aware Recurrent Networks

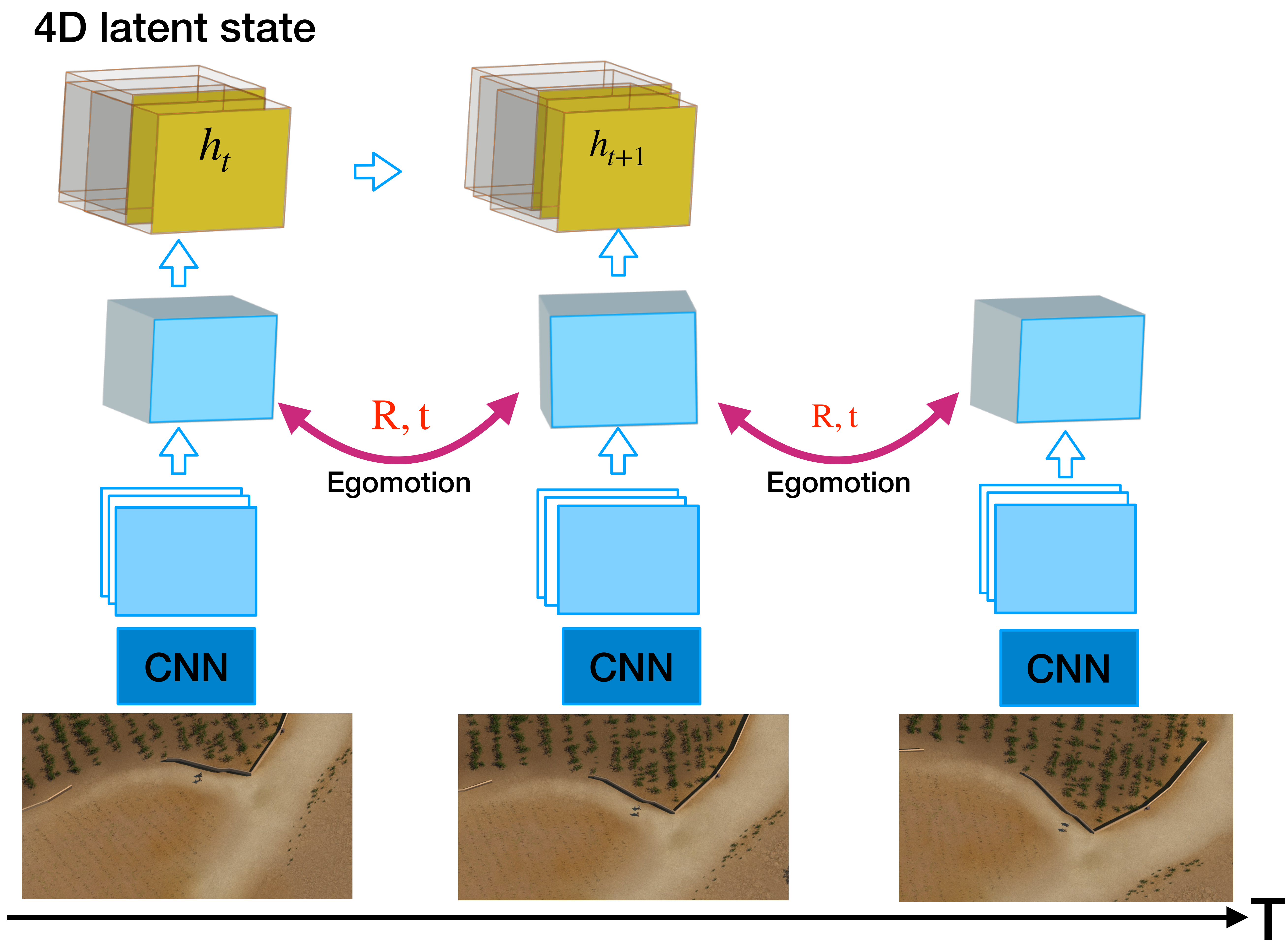
1. Hidden state: A 4D deep feature tensor, akin to a 3D (feature as opposed to pointcloud) map of the scene
2. Egomotion-stabilized hidden state updates

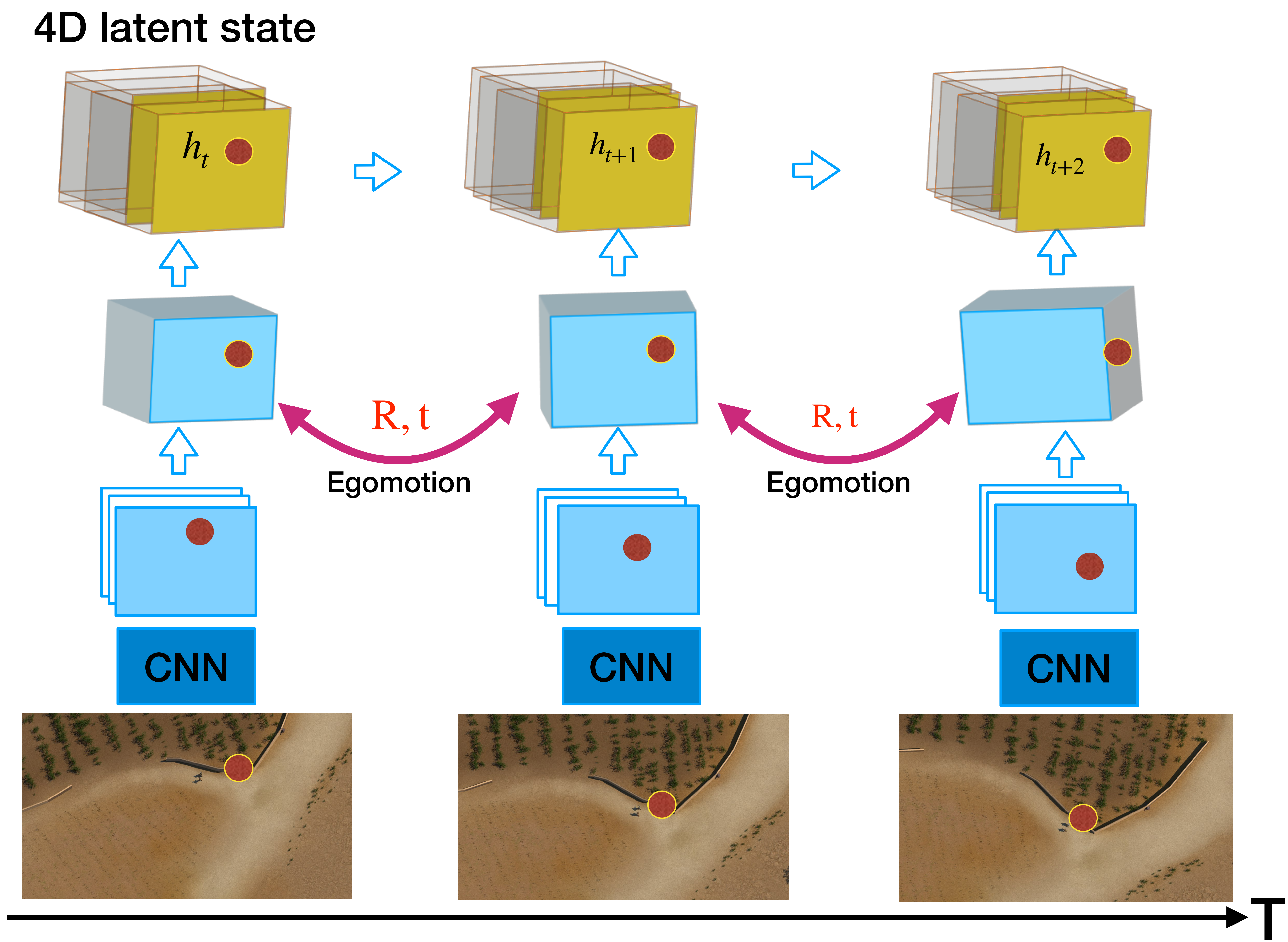


2D Recurrent networks, LSTMs, CONVLSTMs,..



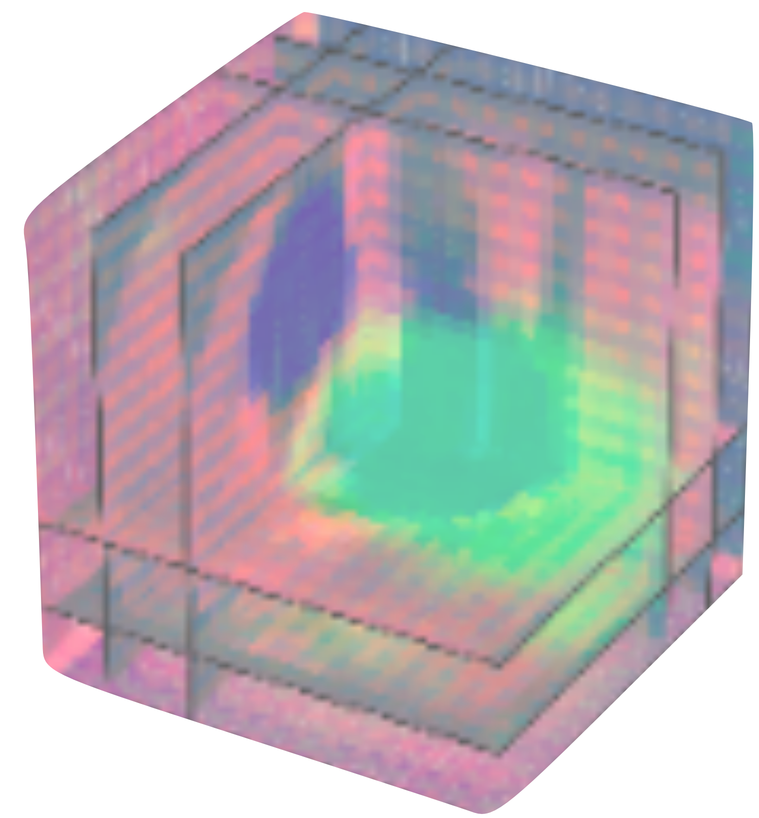

4D latent state
h_t
R, t
Egomotion
CNN
CNN
CNN
T


4D latent state
h_t
h_{t+1}
R, t
Egomotion
R, t
Egomotion
CNN
CNN
CNN
T



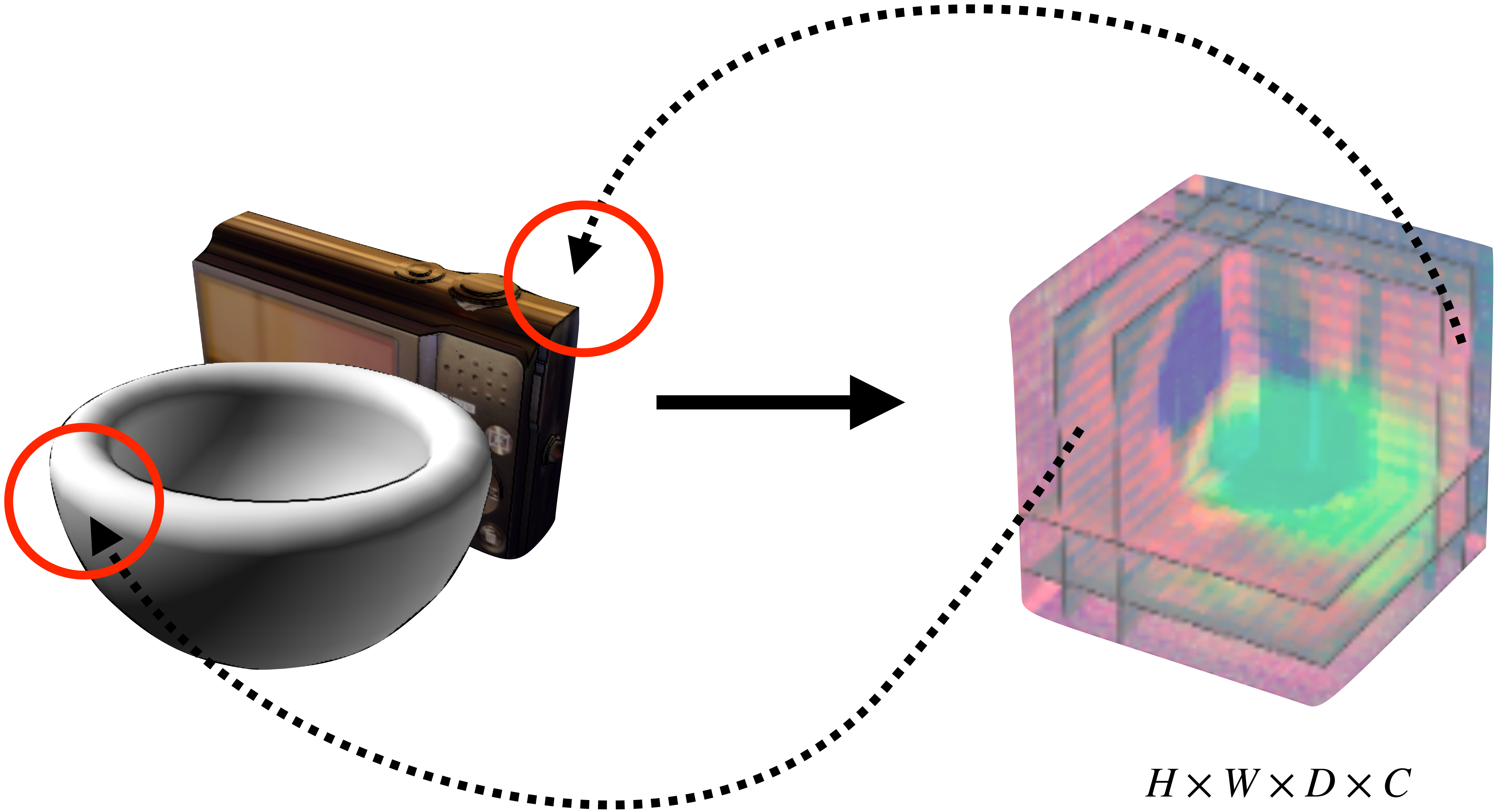
4D latent state
h_t
h_{t+1}
h_{t+2}
R, t
Egomotion
R, t
Egomotion
CNN
CNN
CNN
T

Geometry-Aware Recurrent Networks (GRNNs)

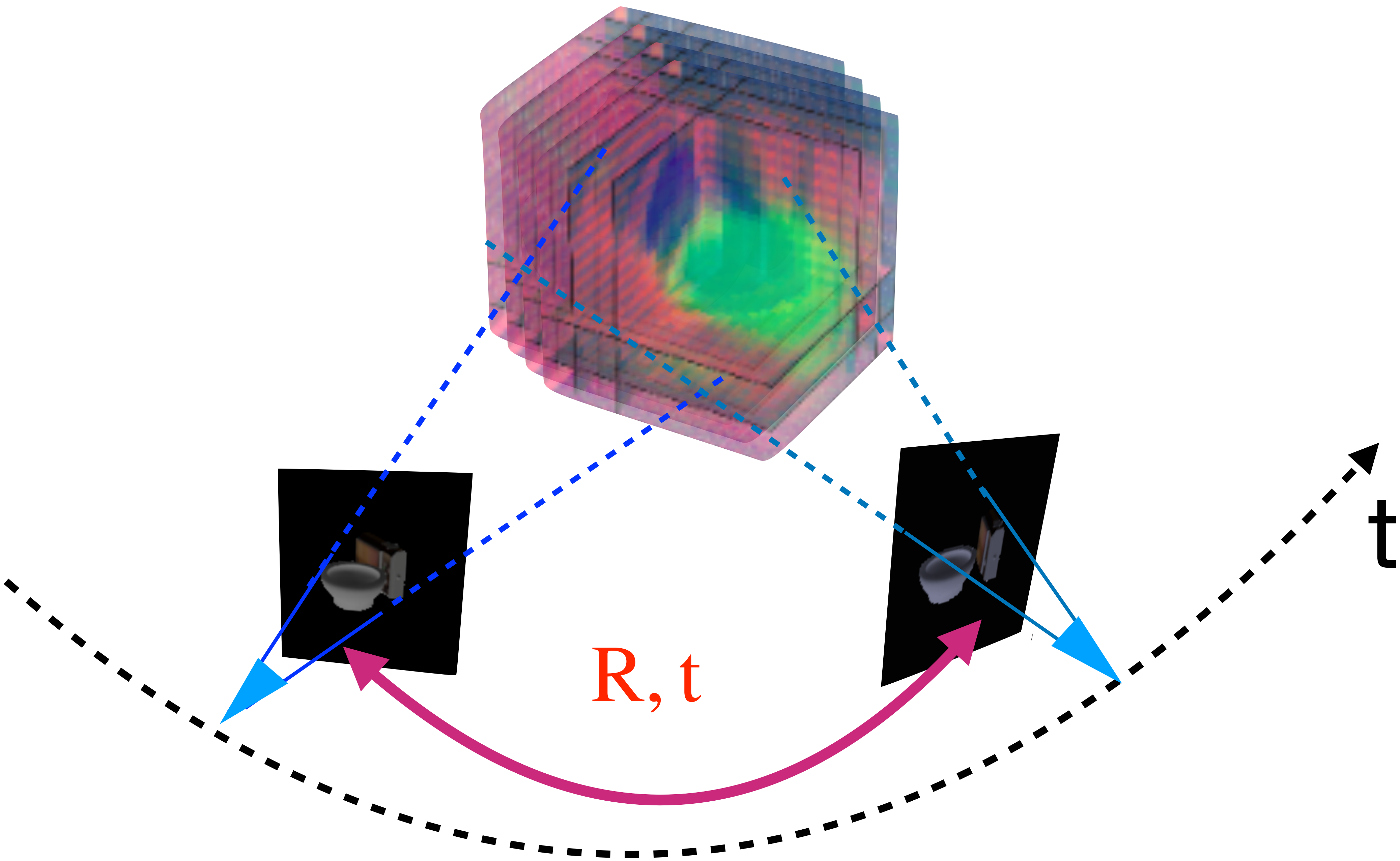


$H \times W \times D \times C$

Geometry-Aware Recurrent Networks (GRNNs)

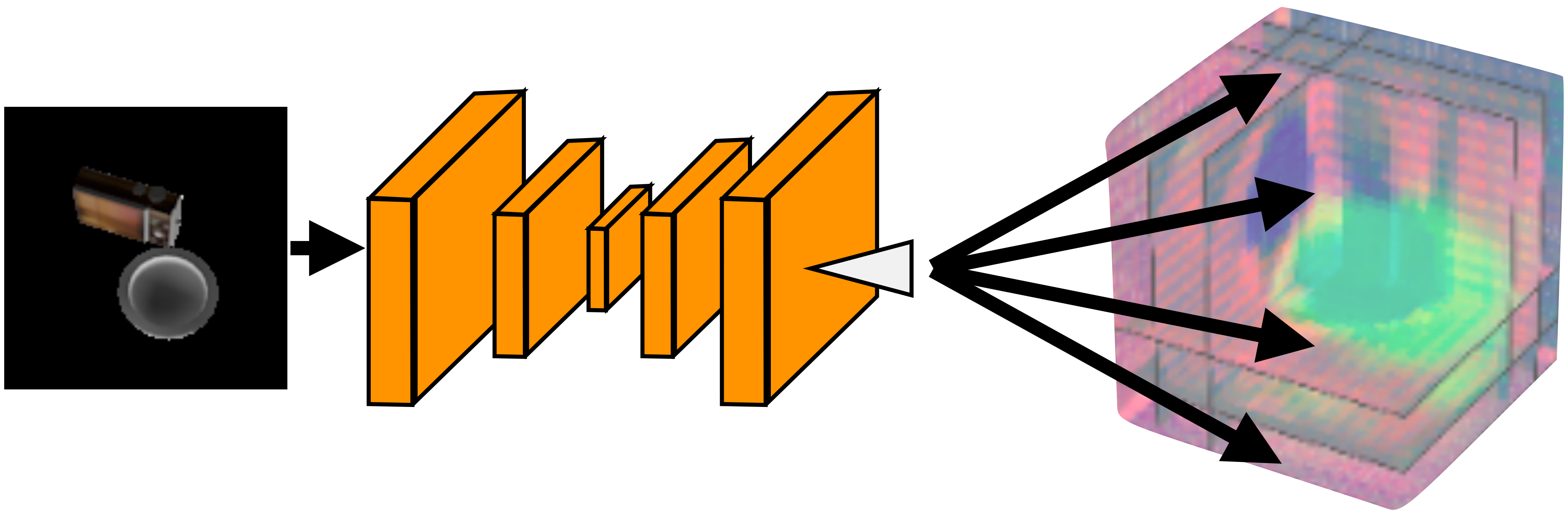


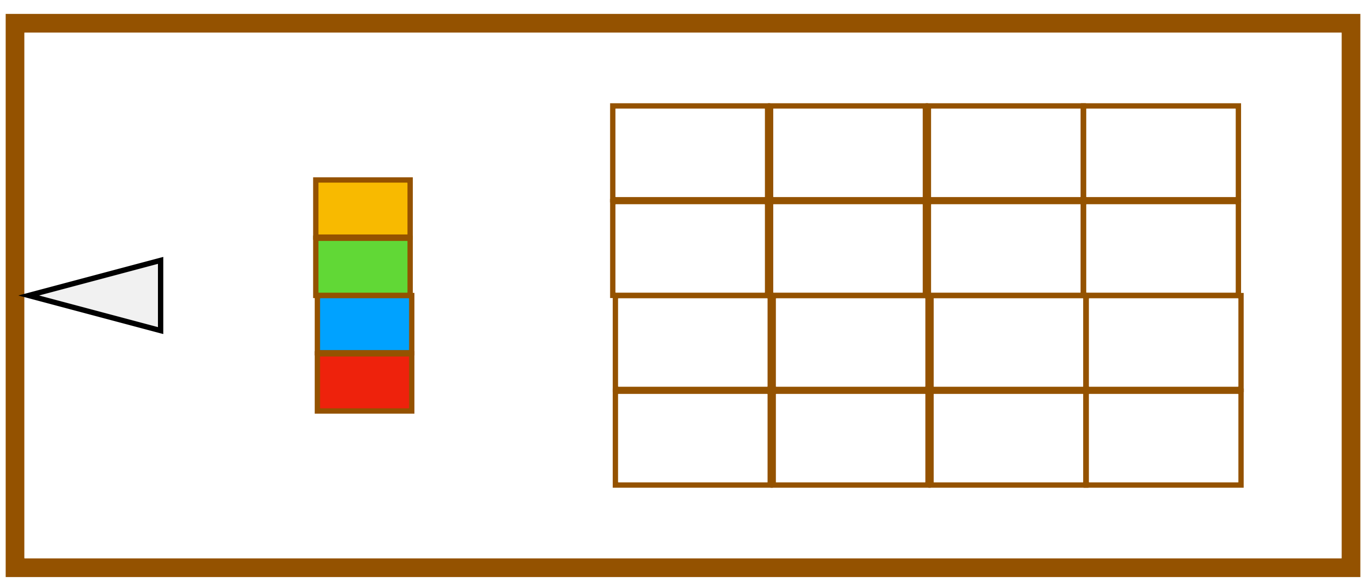
GRNNs



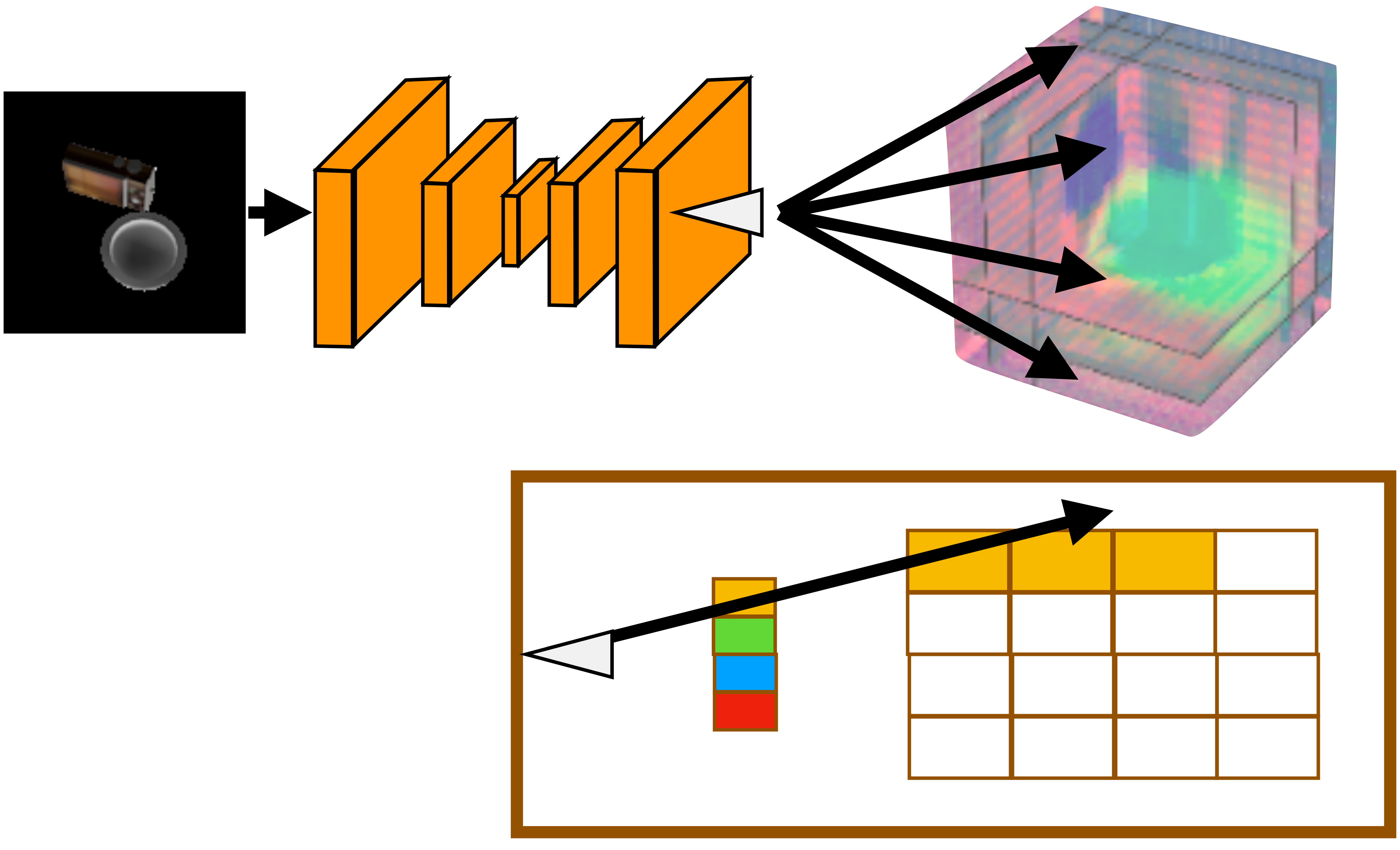
- A set of differentiable neural modules to learn to go from 2D to 3D and back
- A lot of SLAM ideas into the neural modules

Unprojection (2D to 3D)

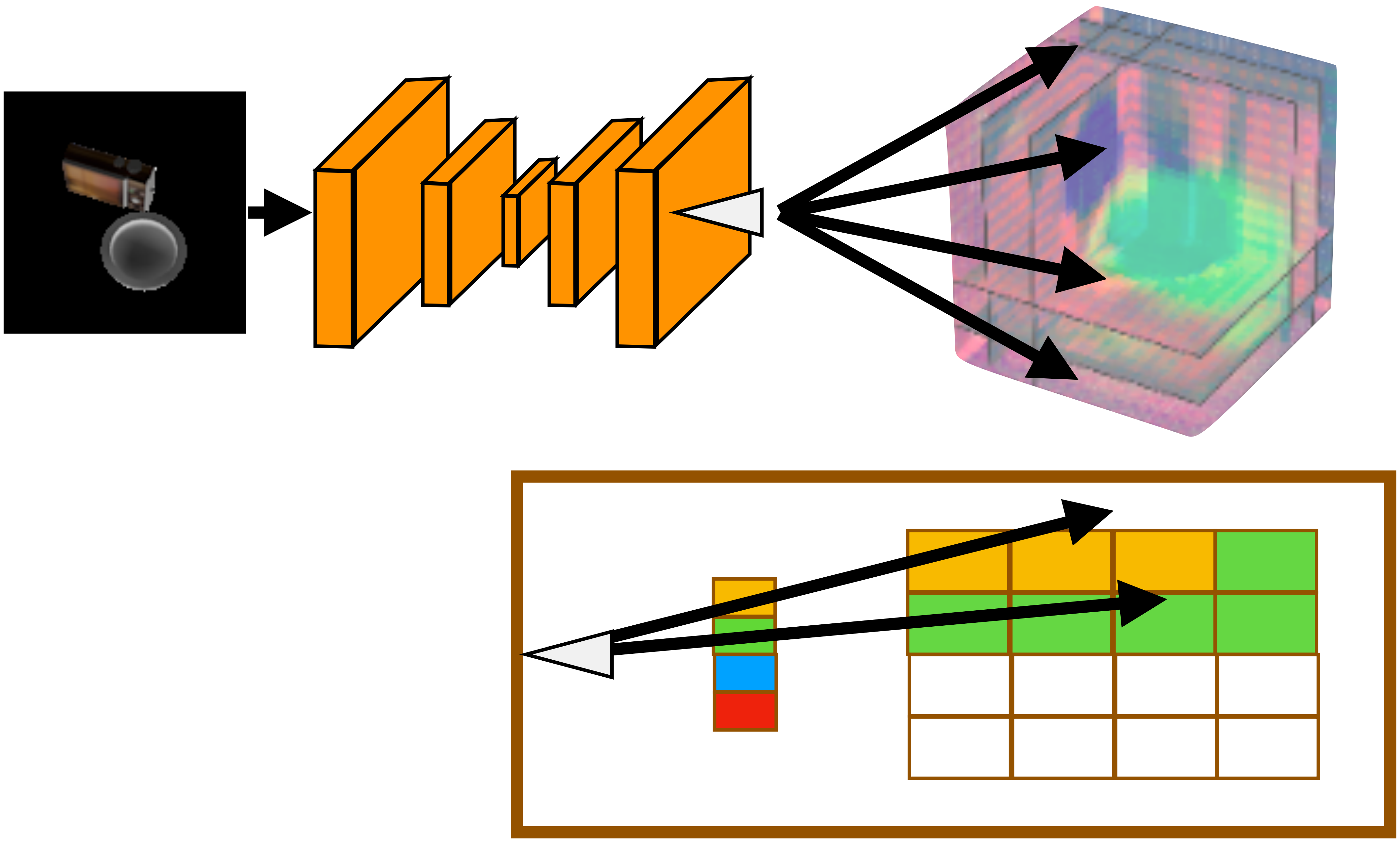




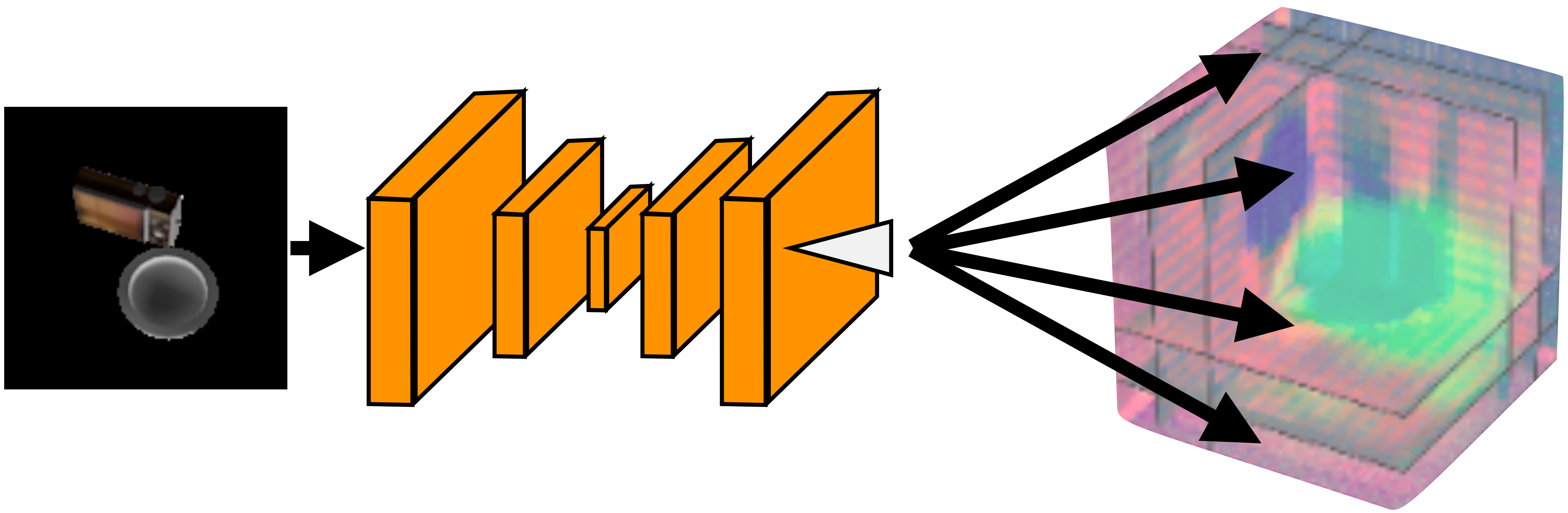
Unprojection (2D to 3D)

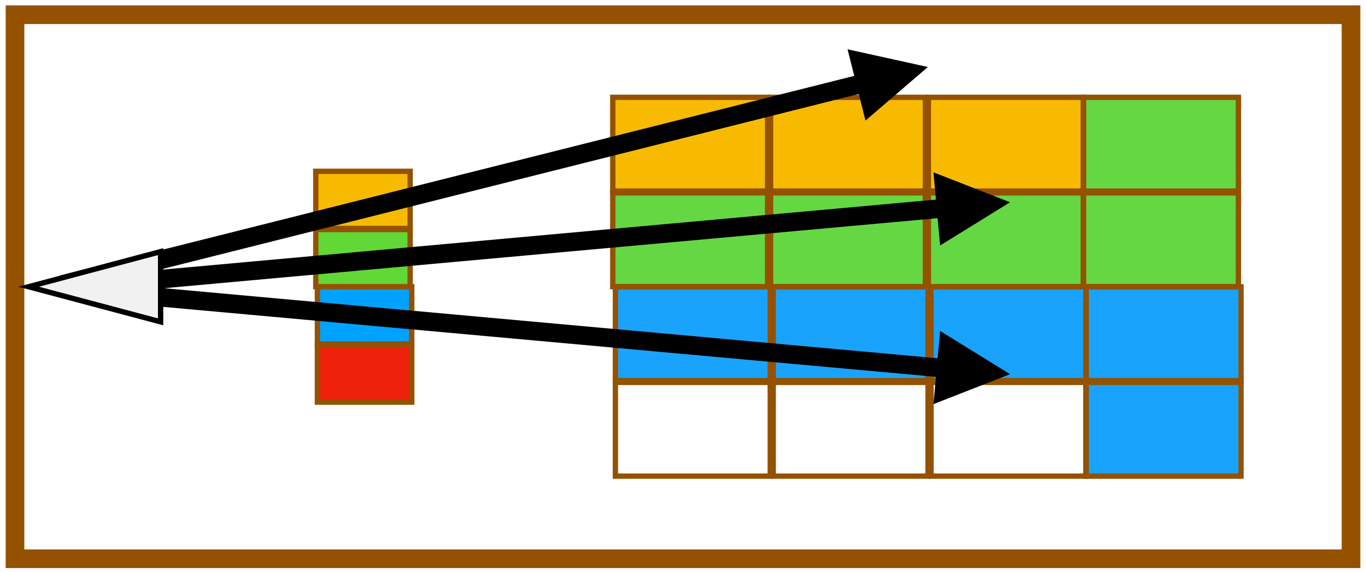


Unprojection (2D to 3D)

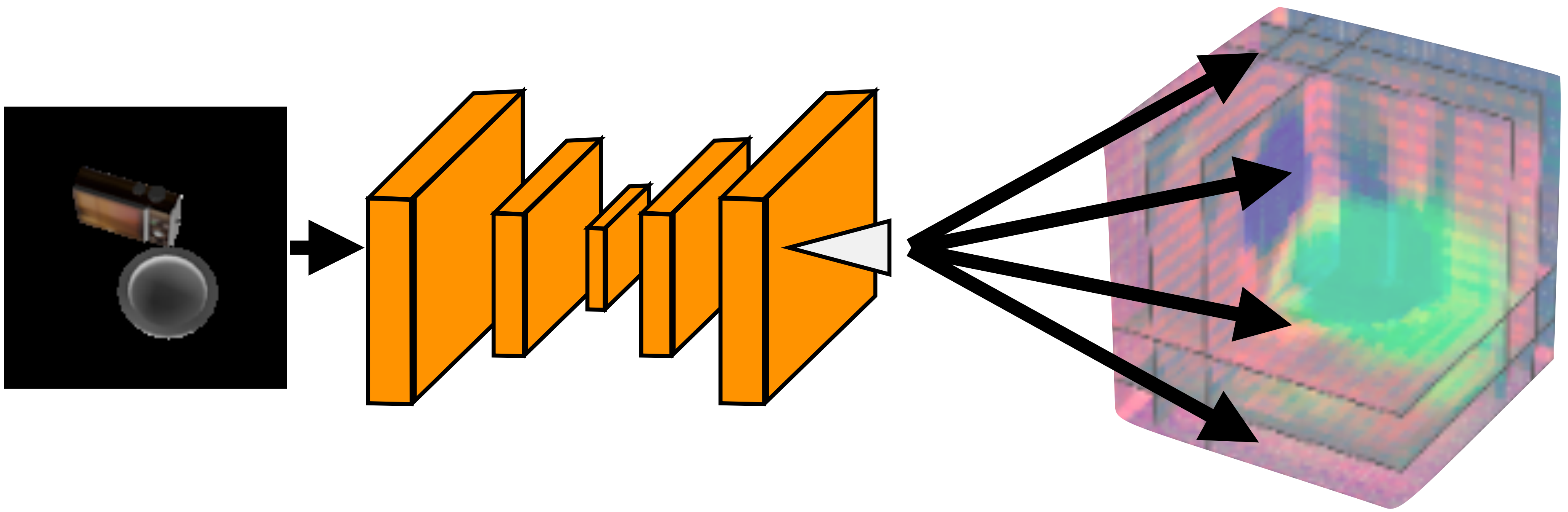


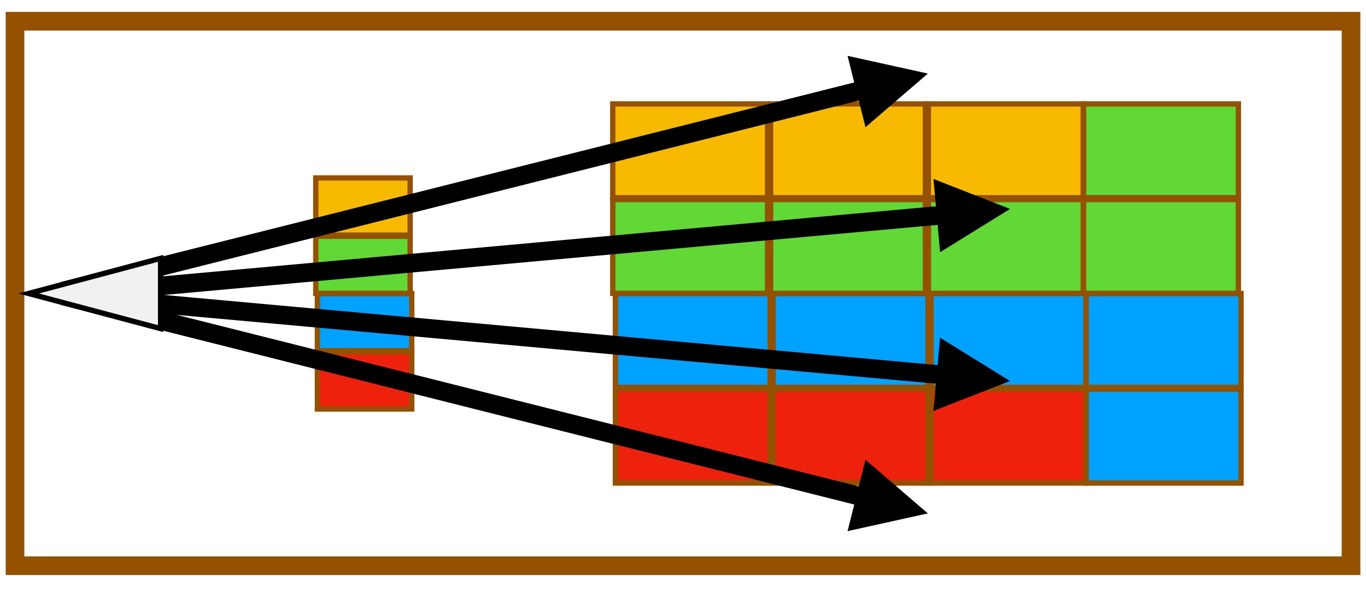
Unprojection (2D to 3D)



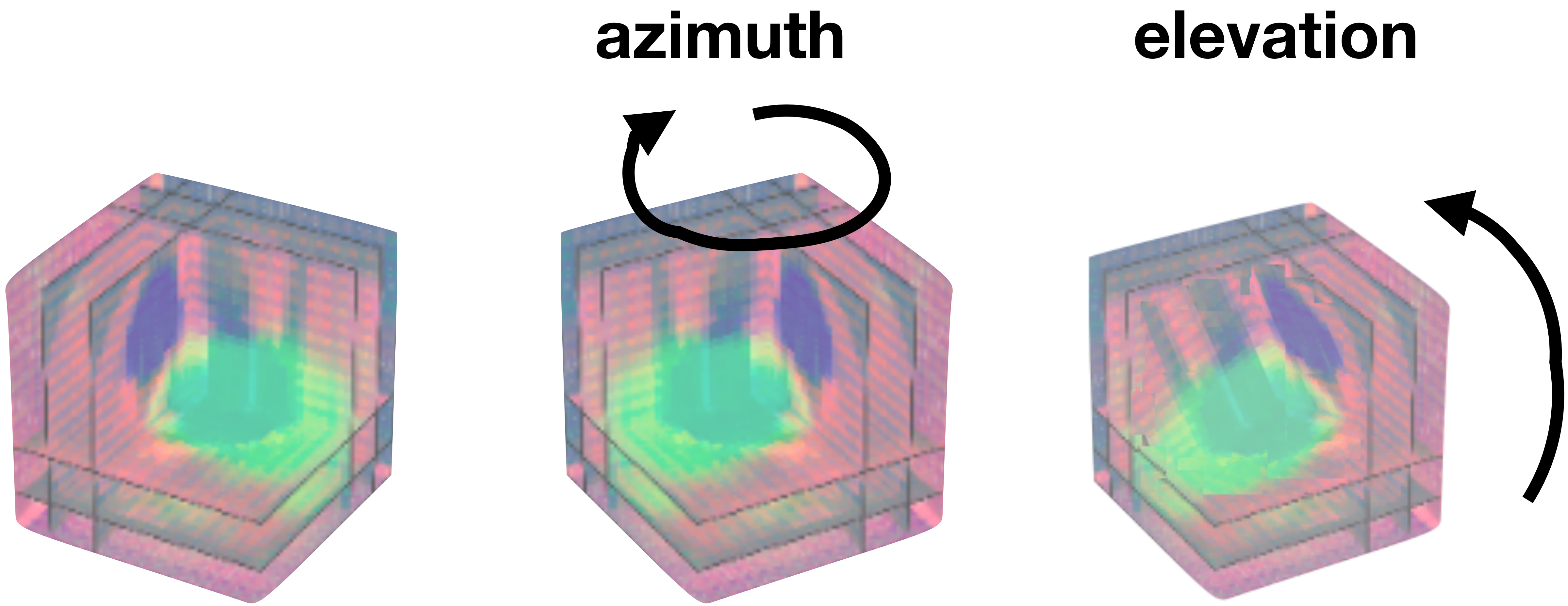


Unprojection (2D to 3D)

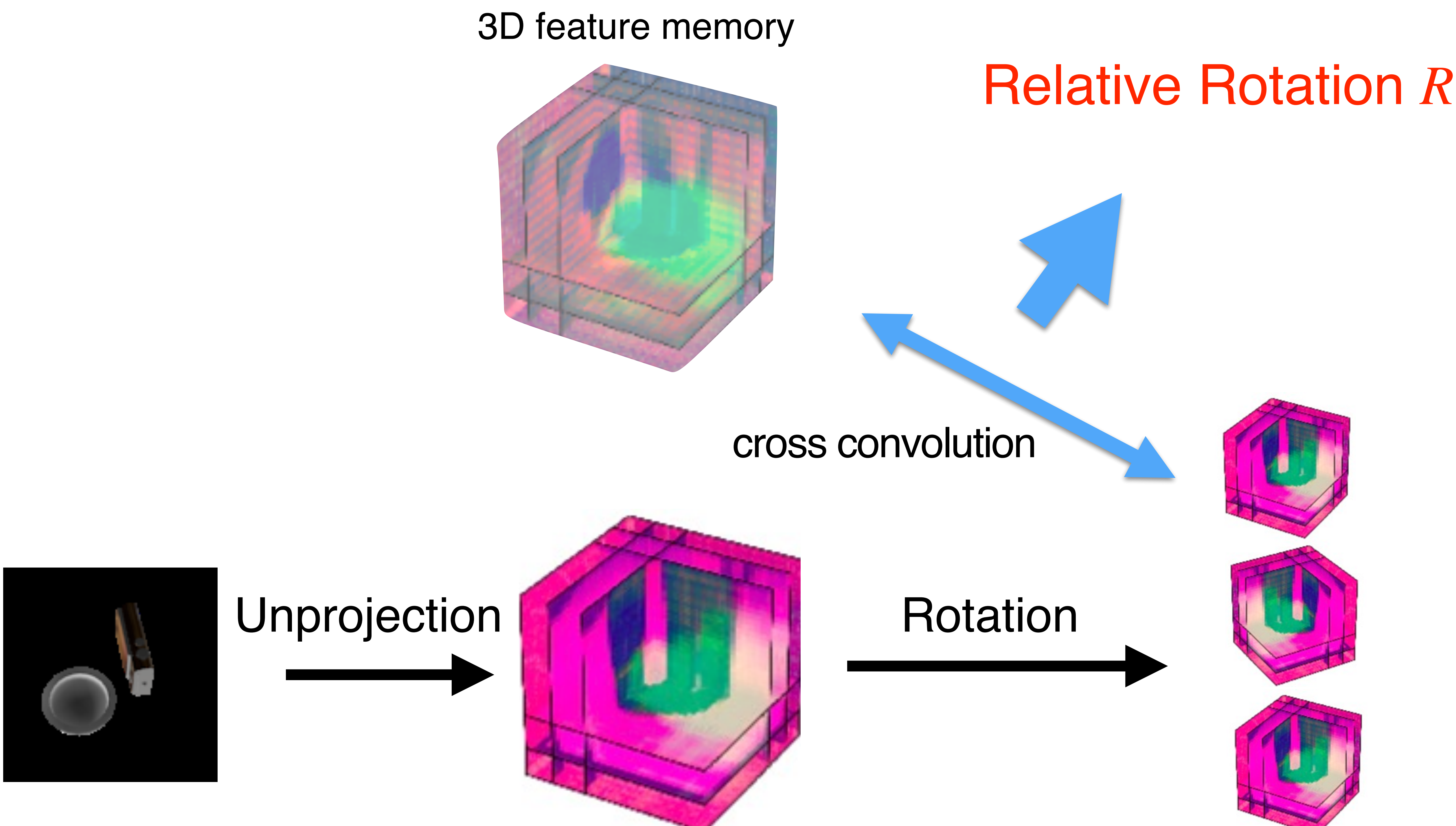




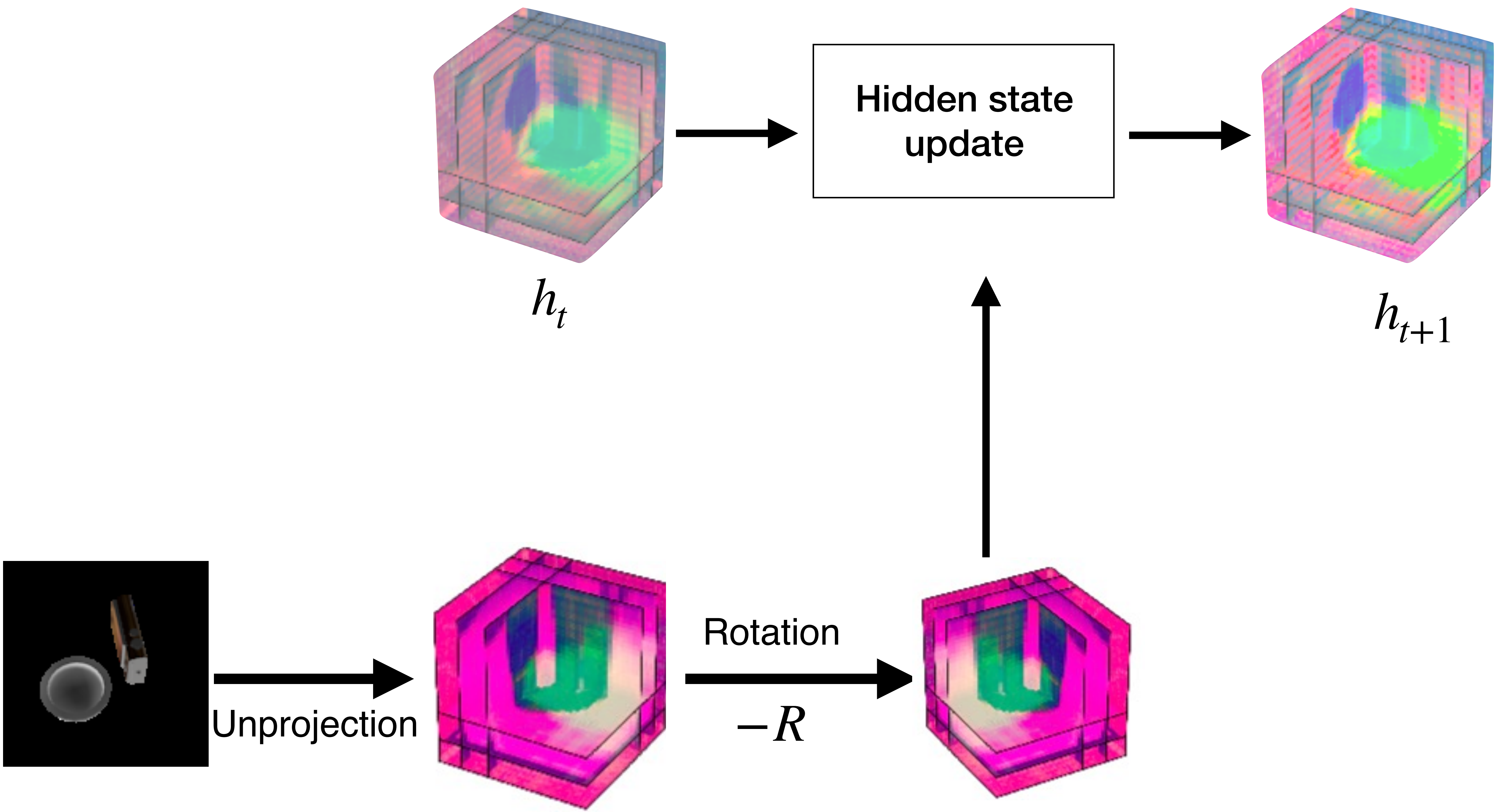
Rotation



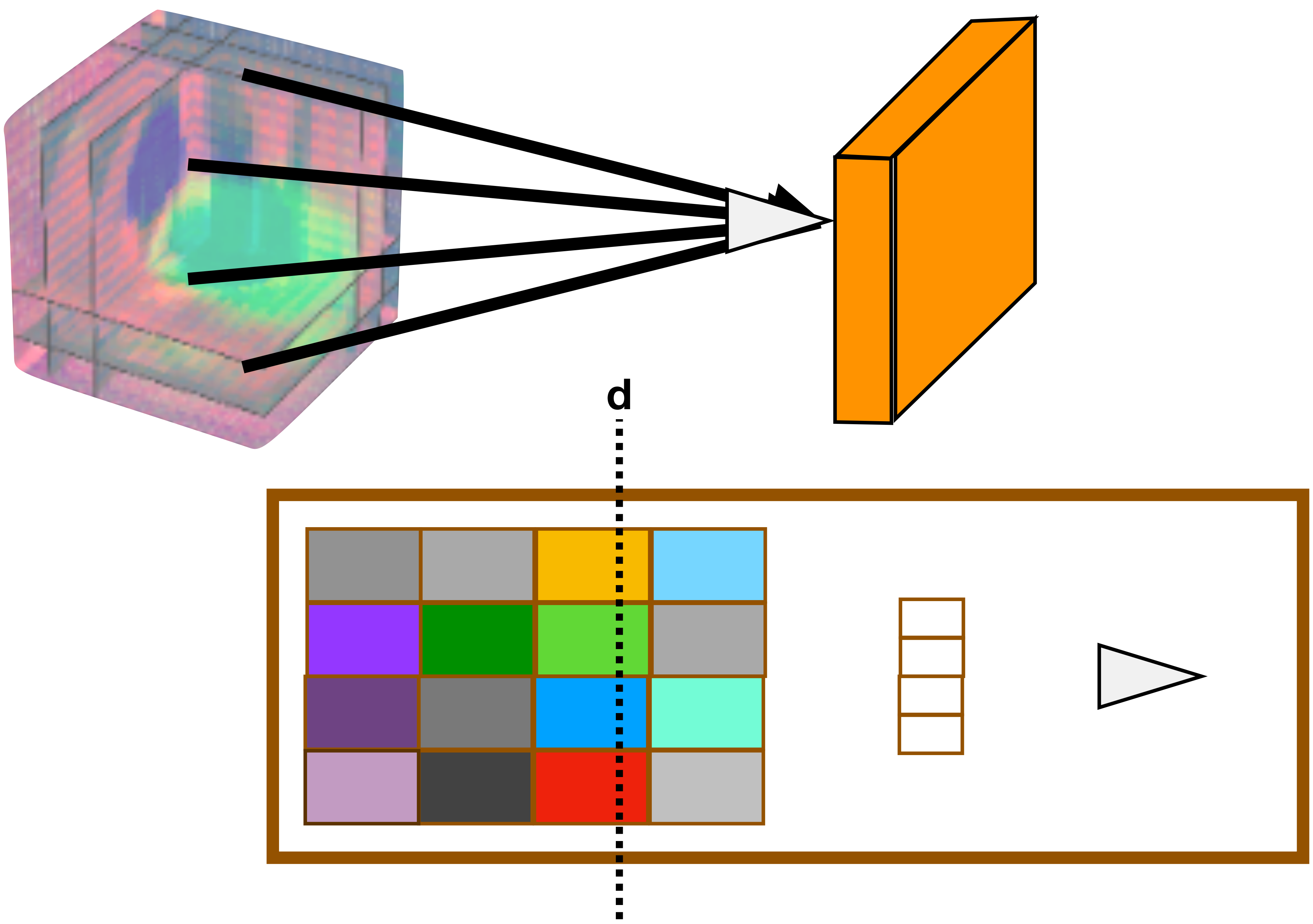
Egomotion-stabilized memory update



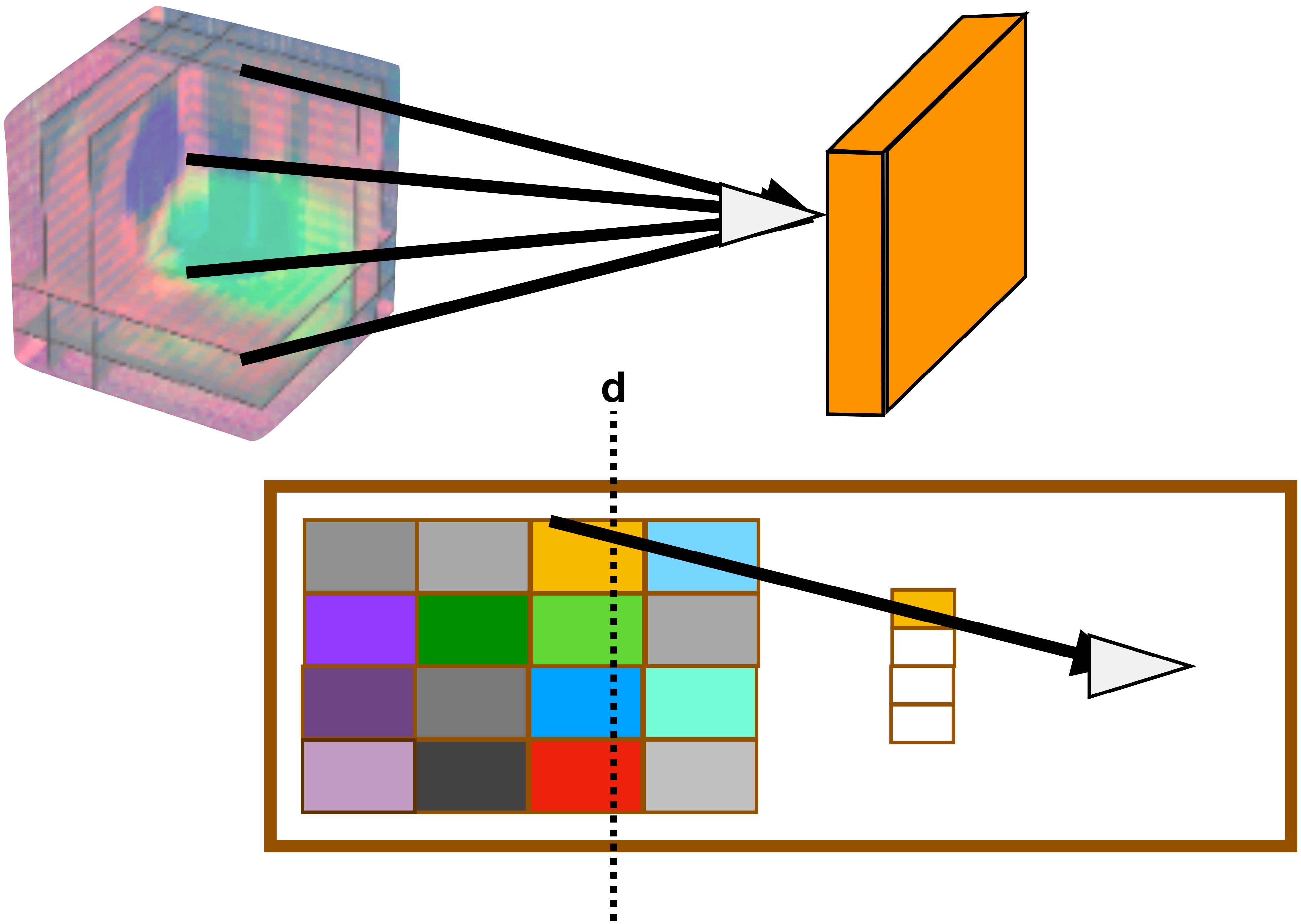
Egomotion-stabilized memory update



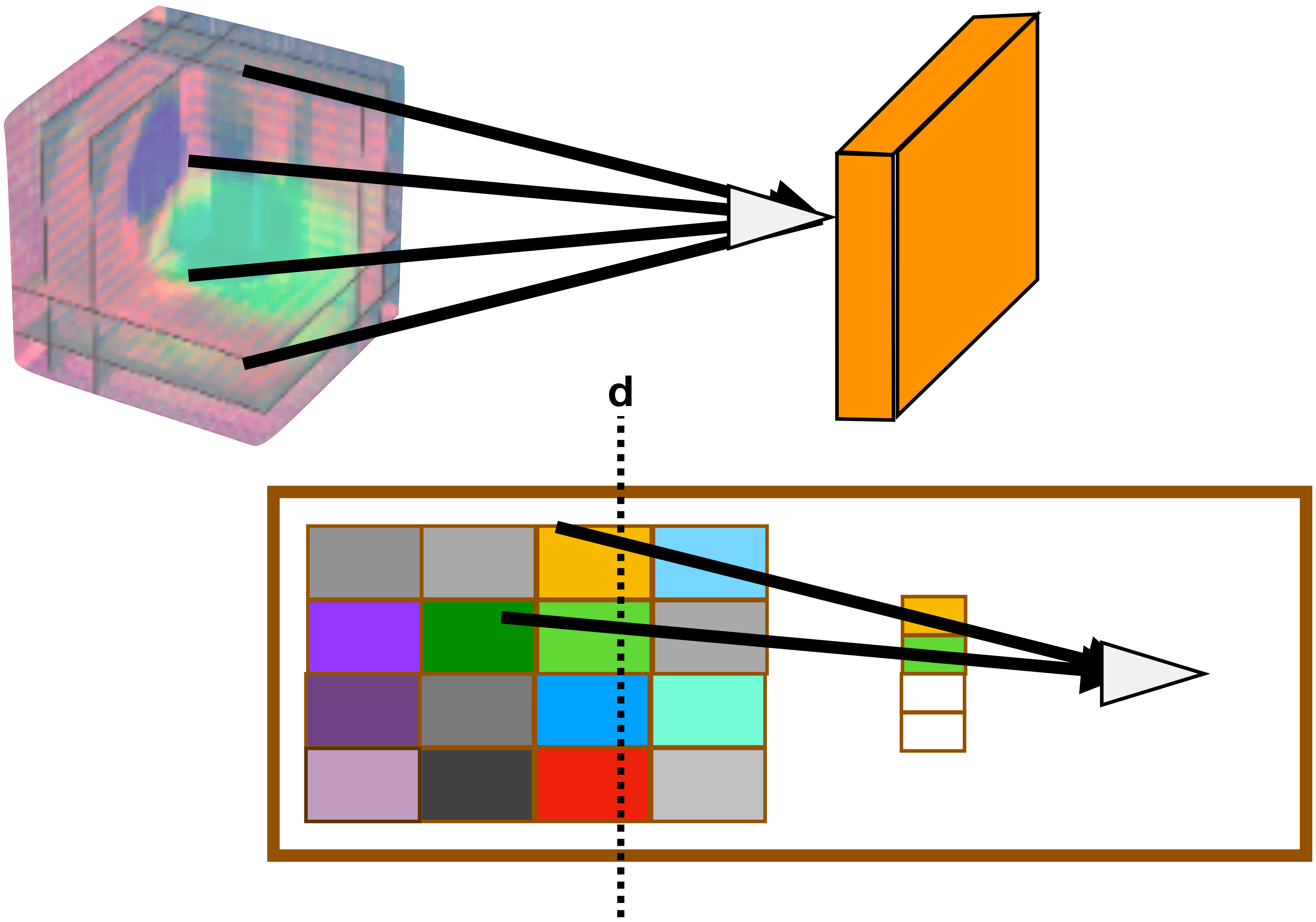
Projection (3D to 2D)



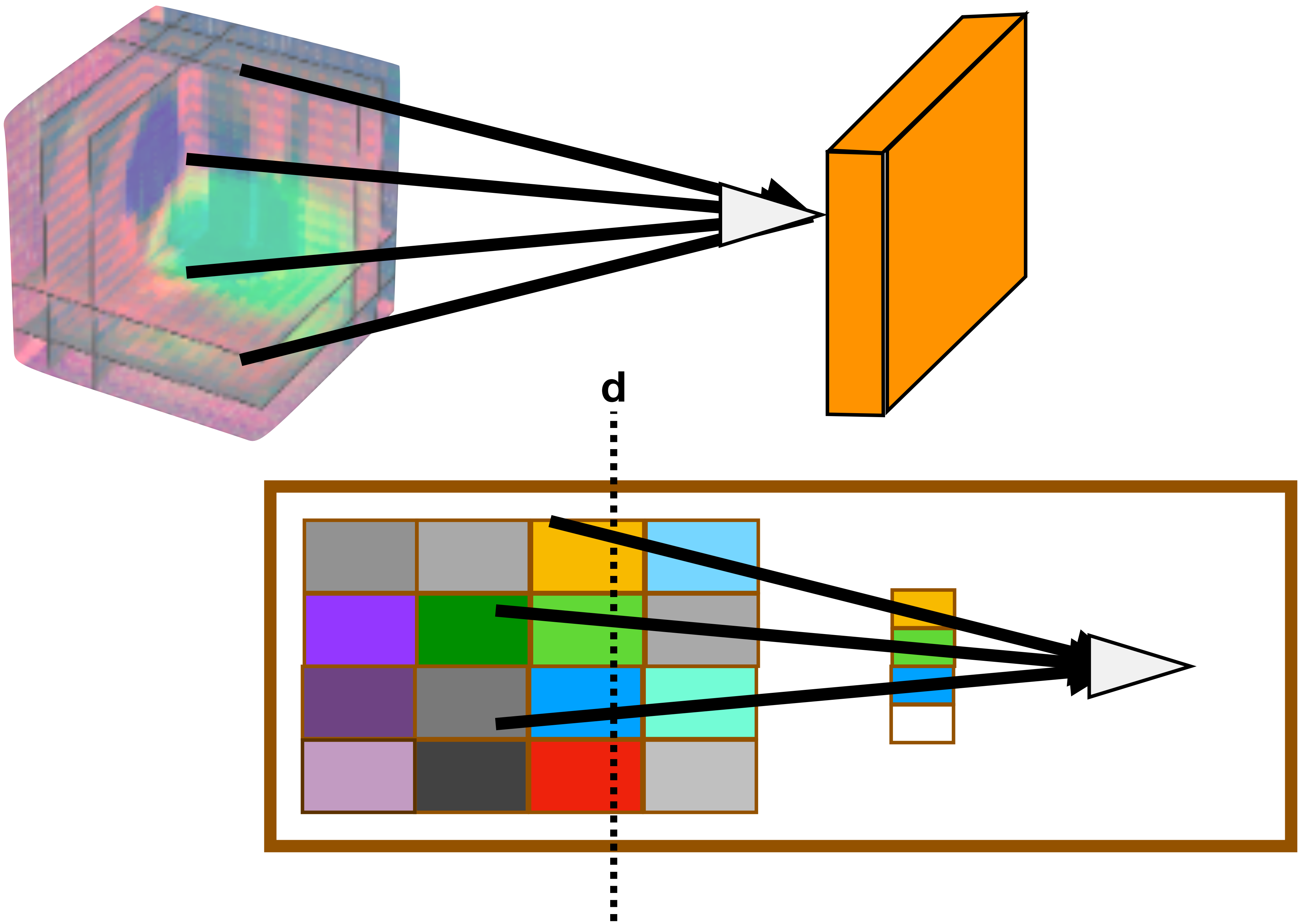
Projection (3D to 2D)



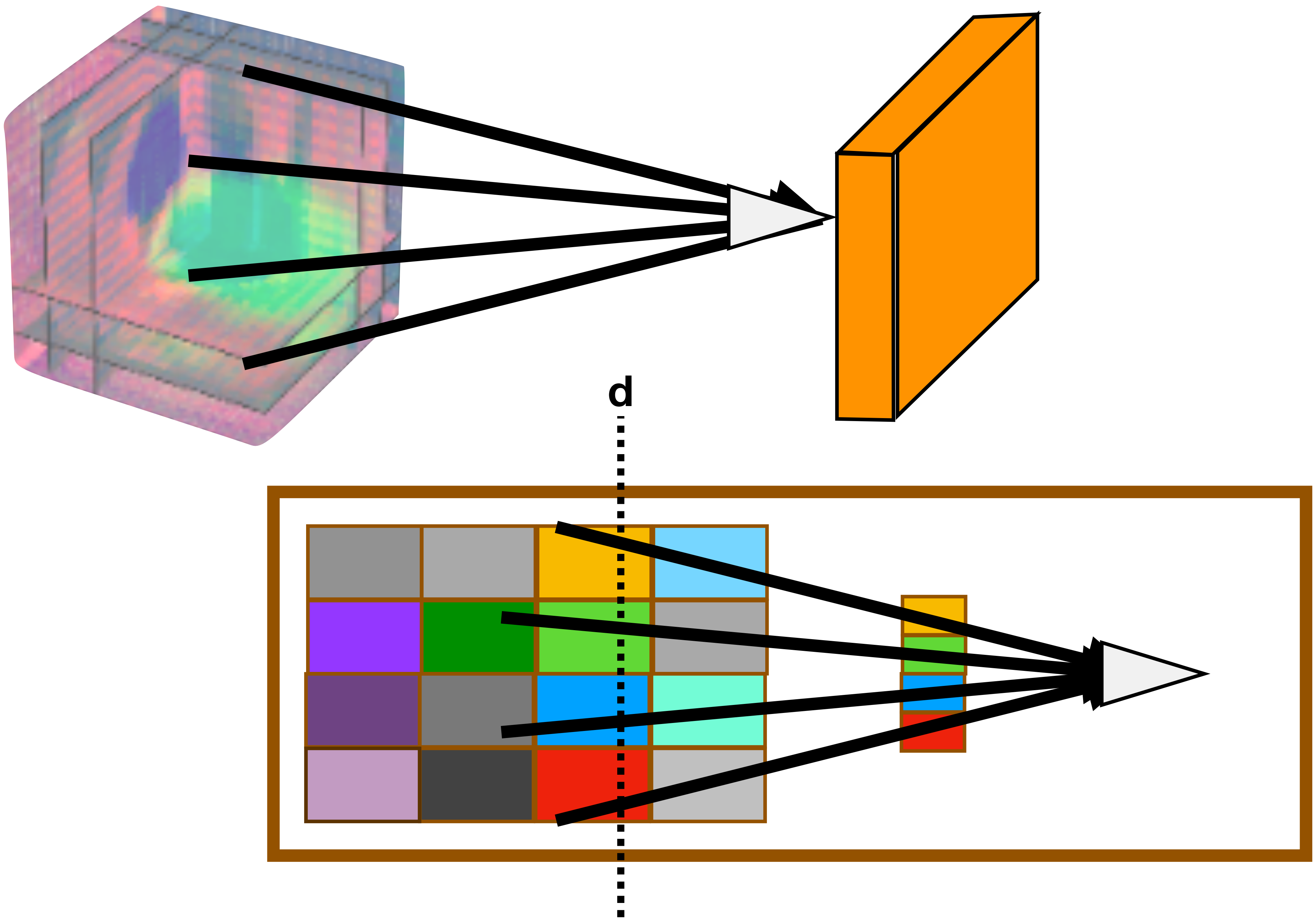
Projection (3D to 2D)



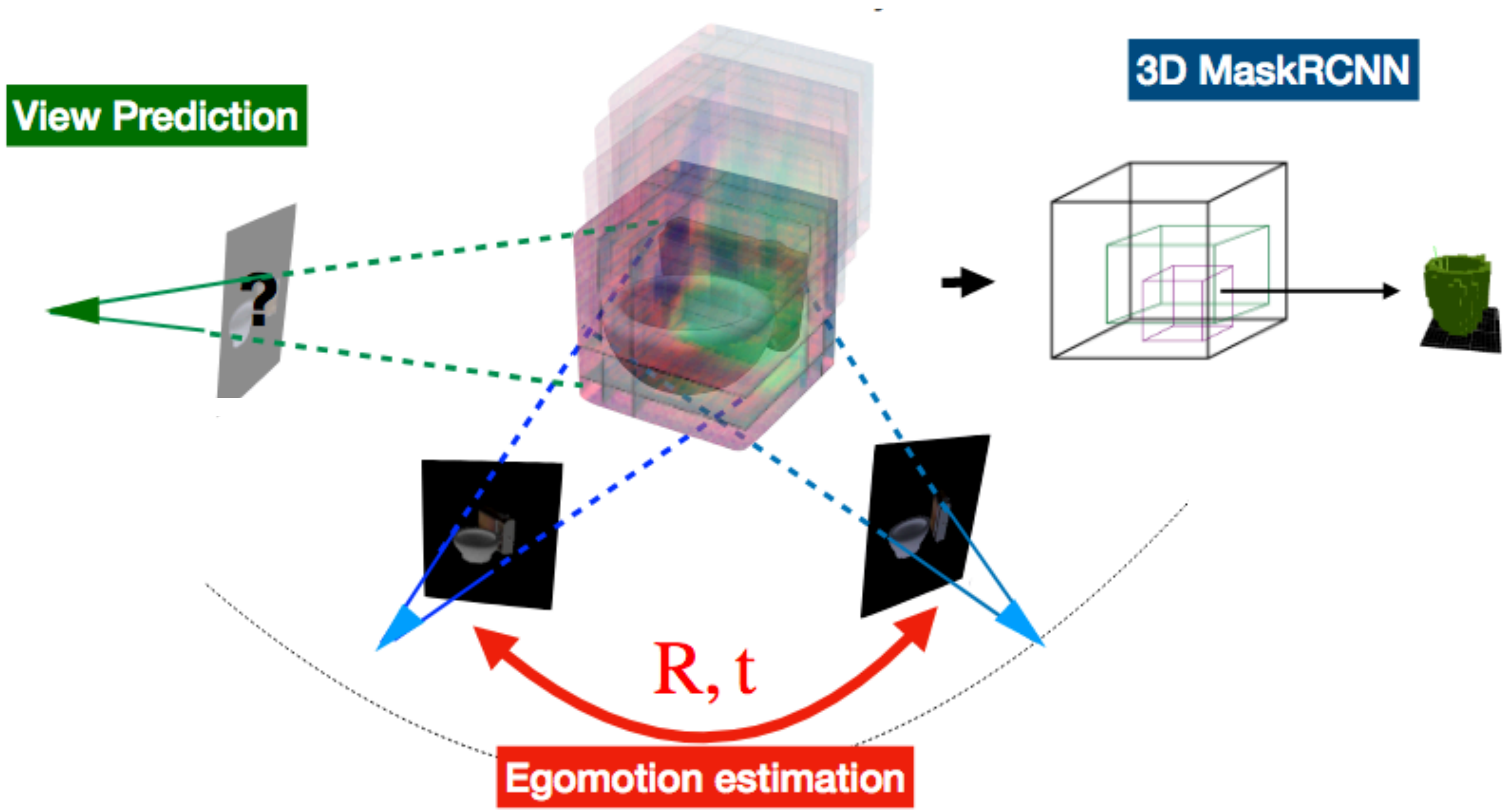
Projection (3D to 2D)



Projection (3D to 2D)

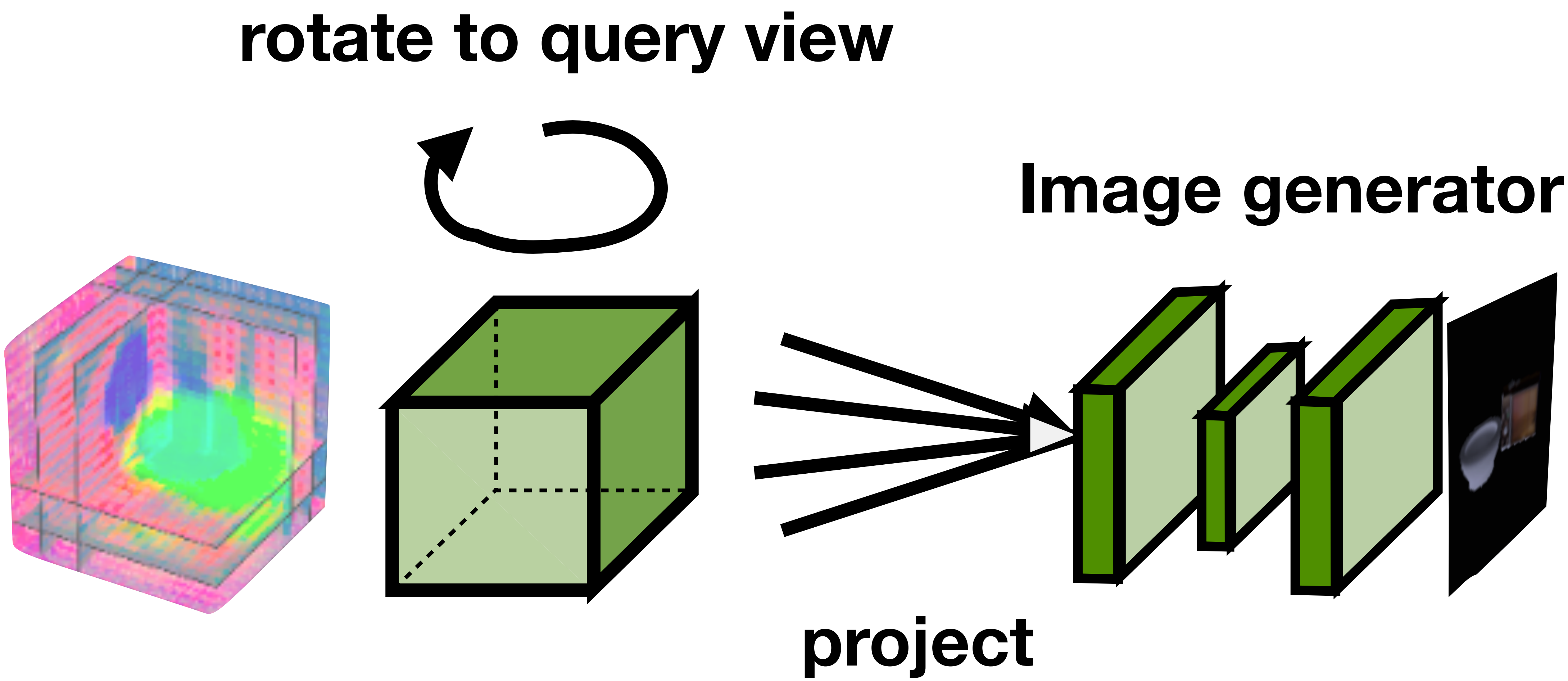


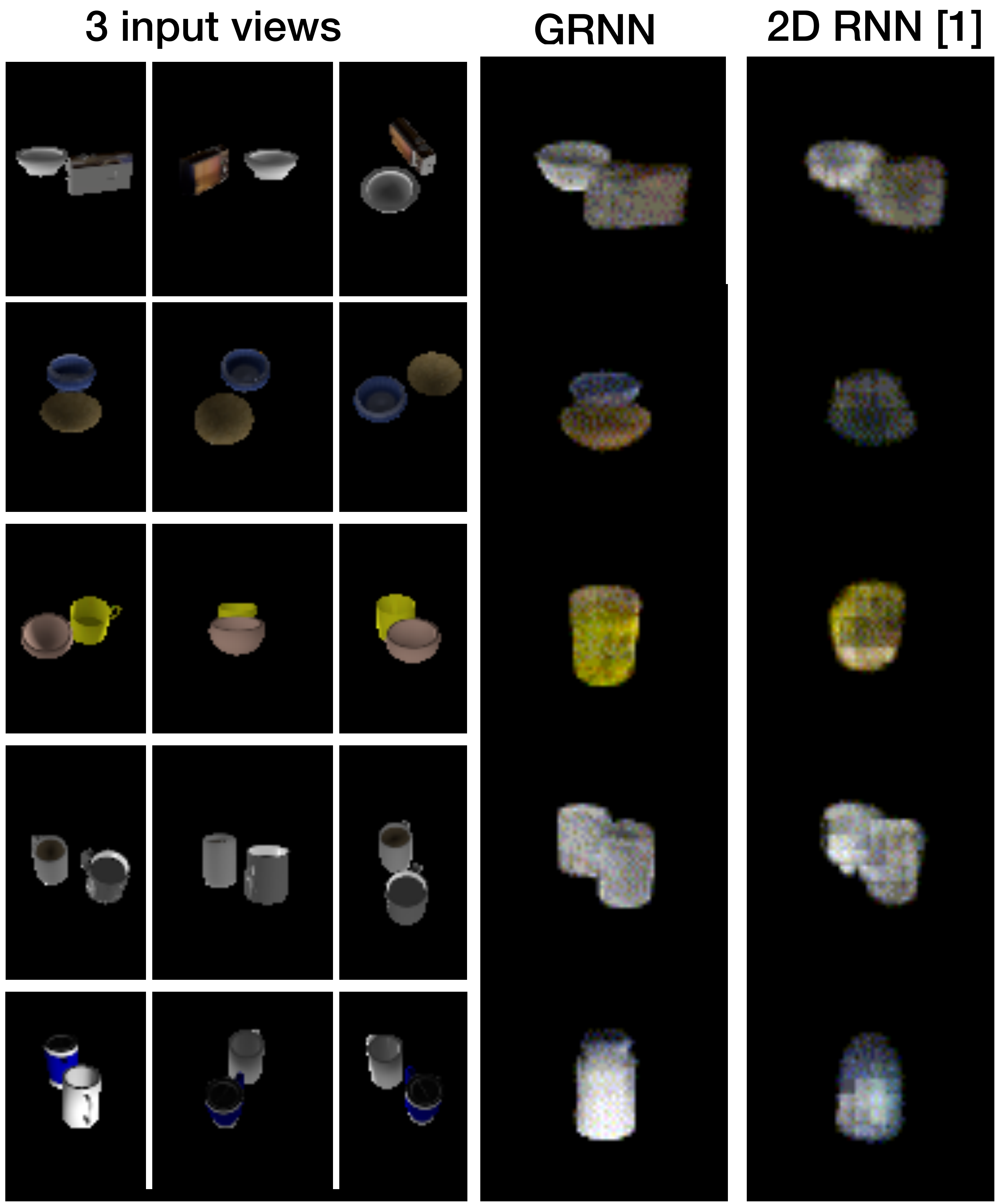
Training GRNNs



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

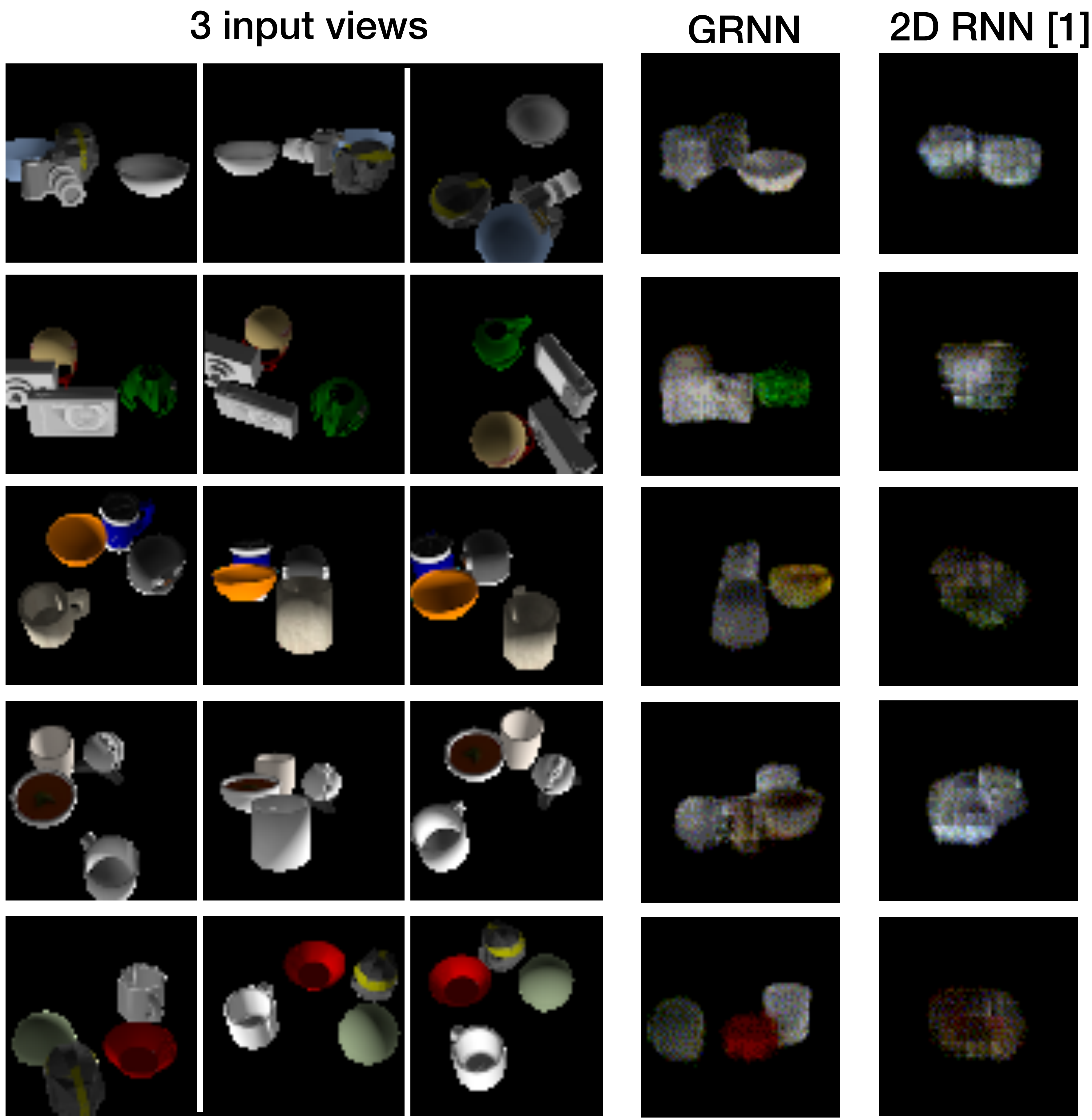
Image generation





[1] *Neural scene representation and rendering* DeepMind, Science, 2018

Testing on scenes with more objets than train time



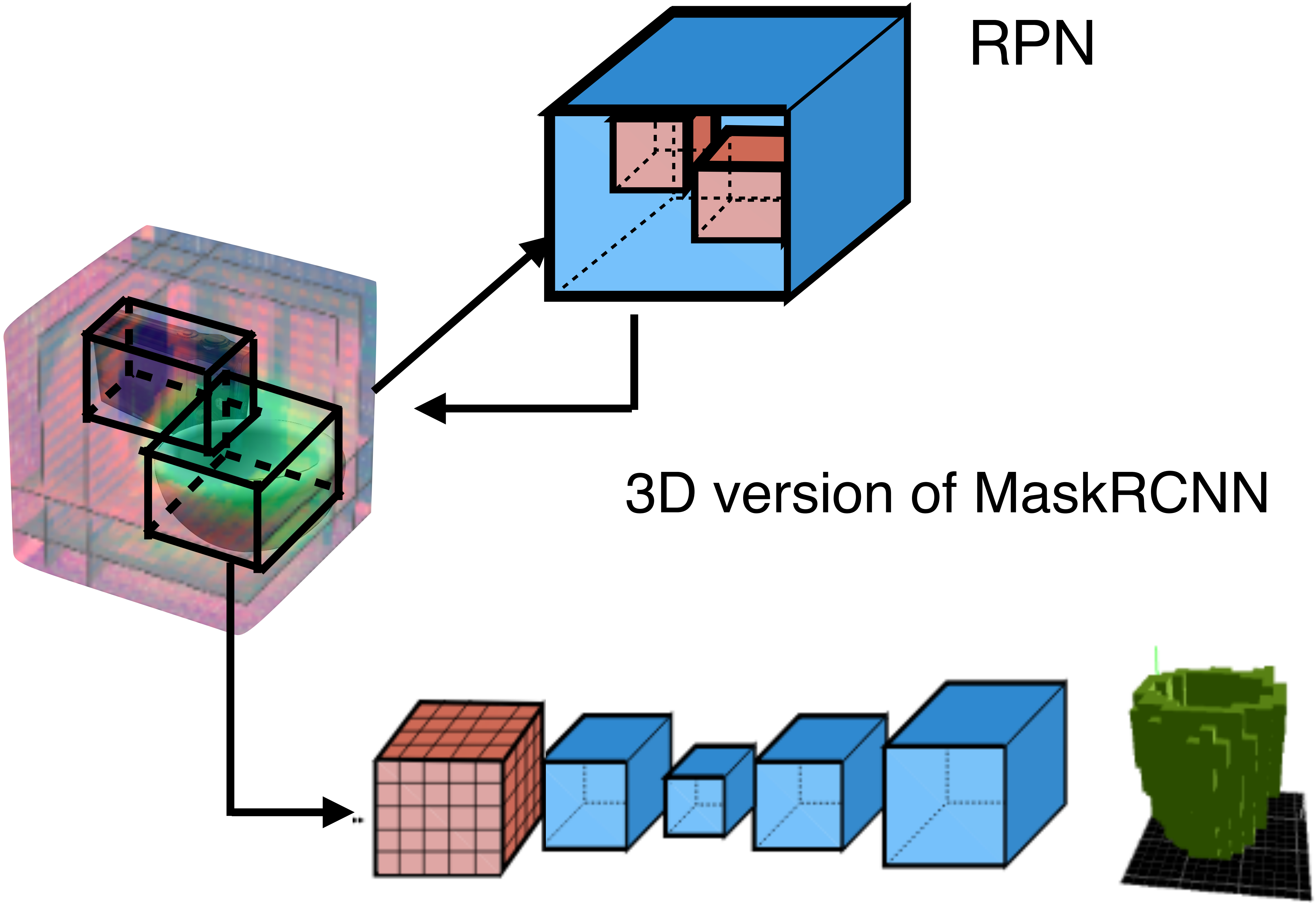
[1] *Neural scene representation and rendering* DeepMind, Science, 2018

View prediction

geometry-aware RNN

2D RNN [1]

3D Object Detection



Results - 3D object detection

detection	2DRNN-gtego-gtd	GRNN-gtego-gtd
$\text{mAP}^{d}_{0.75}$	0.025	**0.348**
$\text{mAP}^{d}_{0.50}$	0.323	**0.977**
$\text{mAP}^{d}_{0.33}$	0.653	**0.991**

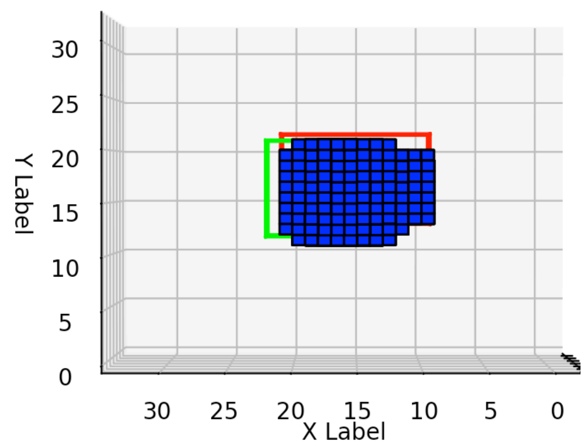
segmentation	2DRNN-gtego-gtd	GRNN-gtego-gtd
$\text{mAP}^{m}_{0.75}$	0.000	**0.110**
$\text{mAP}^{m}_{0.50}$	0.006	**0.378**
$\text{mAP}^{m}_{0.33}$	0.104	**0.545**

3D object detection

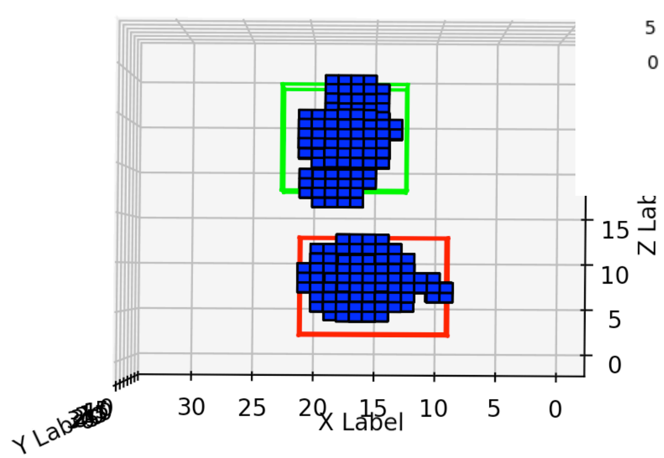
input views

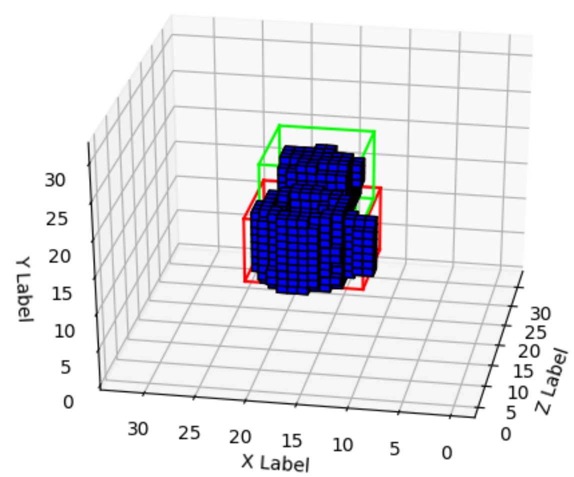
predicted boxes

front-view



bird-view

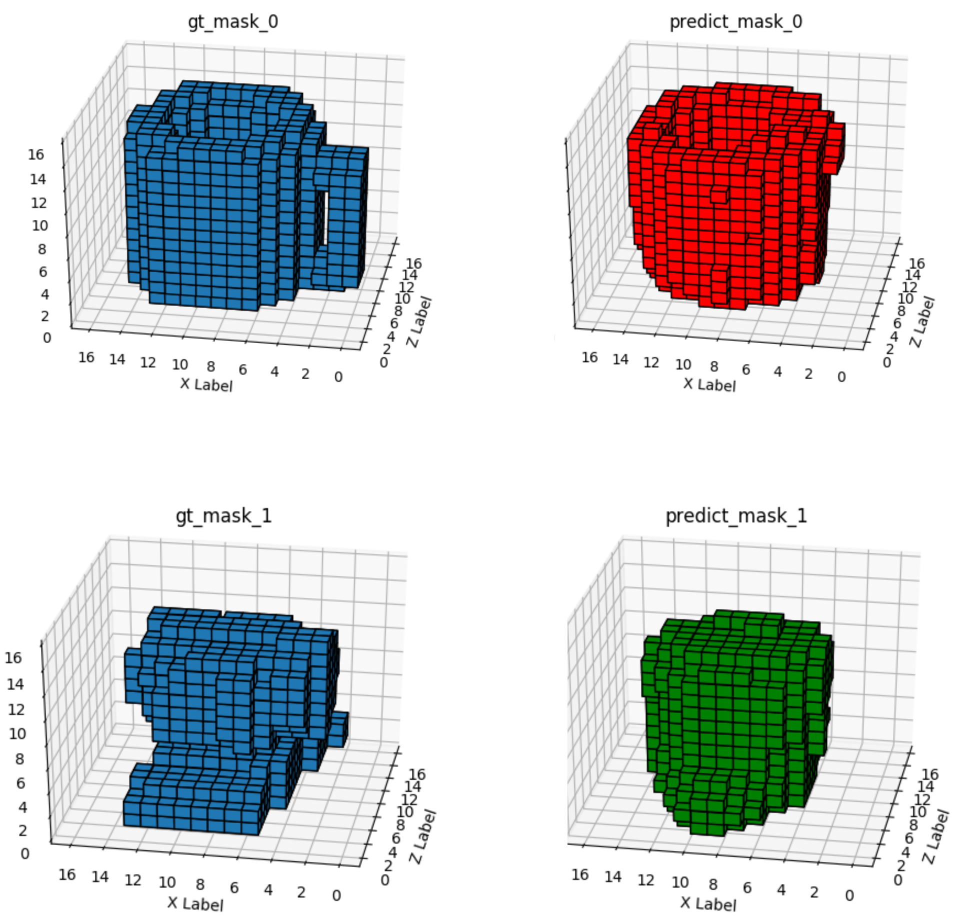




predicted segmentations

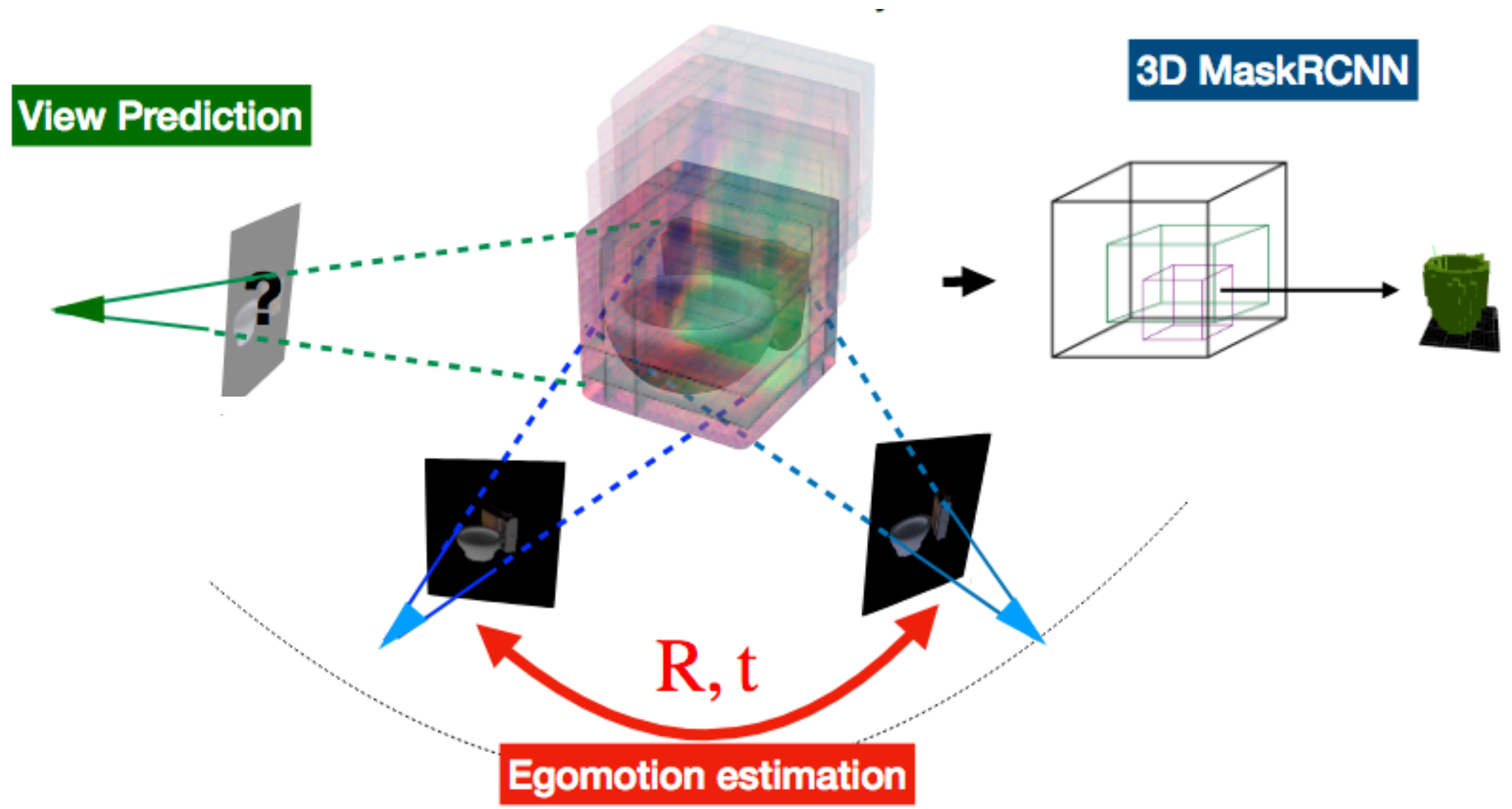
gt

prediction



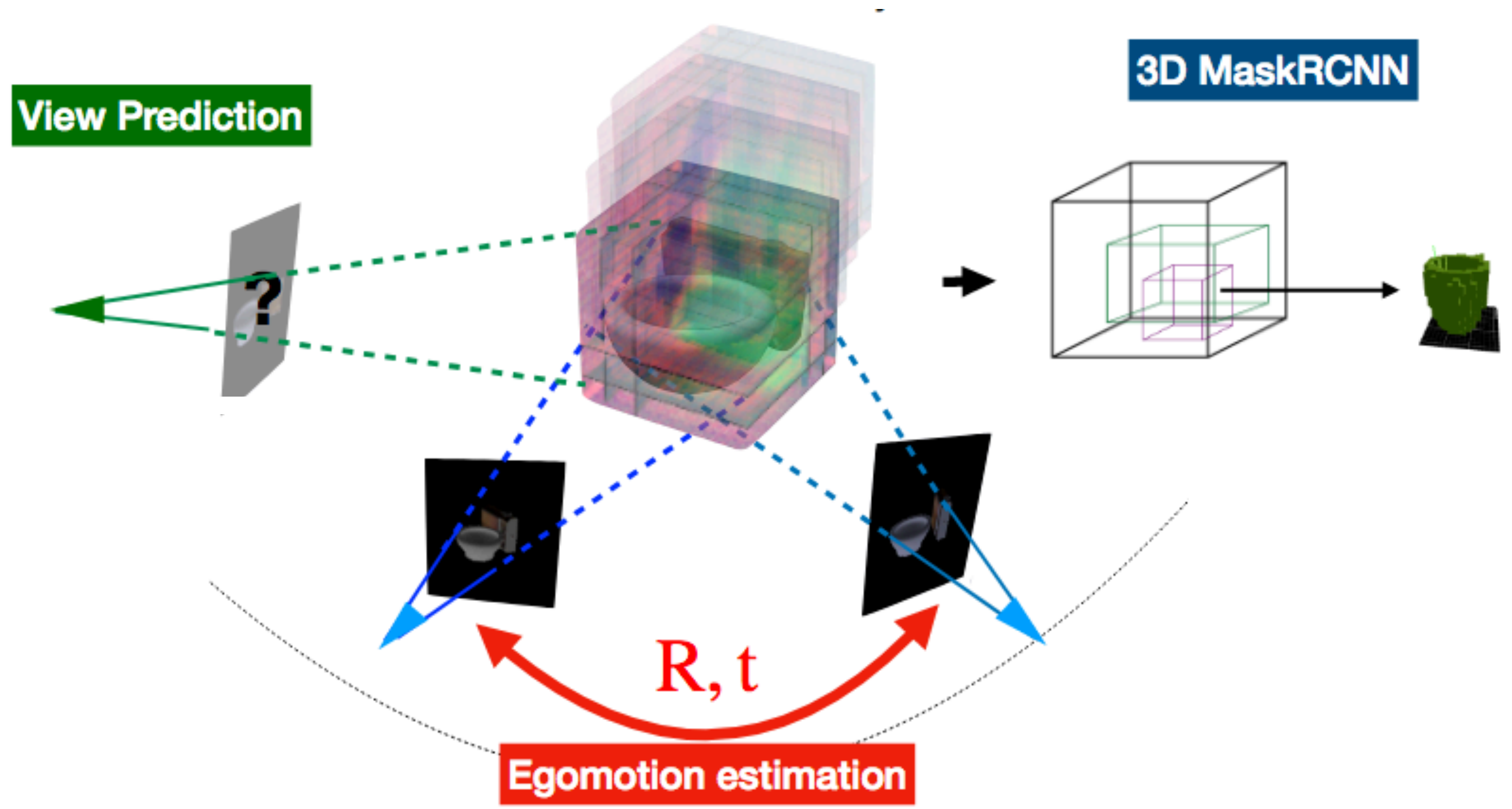
Objects detections learn to perist in time, they do not switch on and off from frame to frame

GRNNs



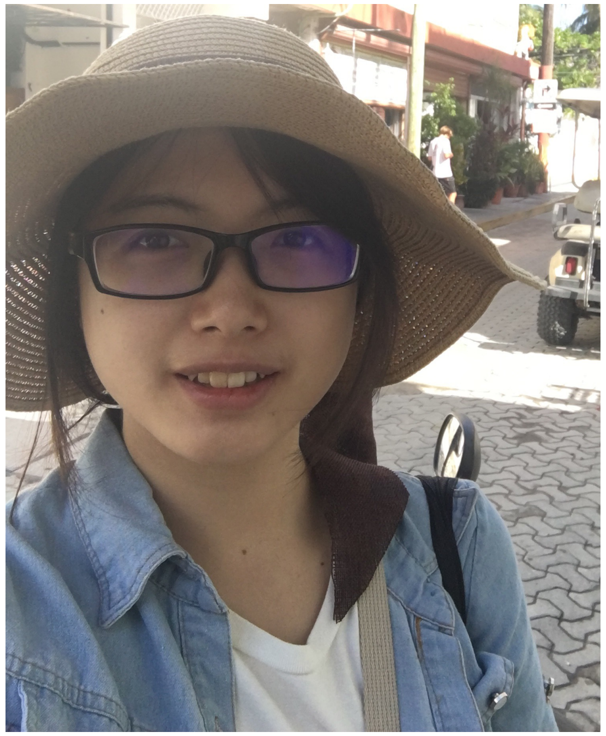
- Differentiable SLAM for better space-aware deep feature learning
- Generative model of scenes with a 3D bottleneck when trained from view prediction
- Generalize better than 2D models

What's next?

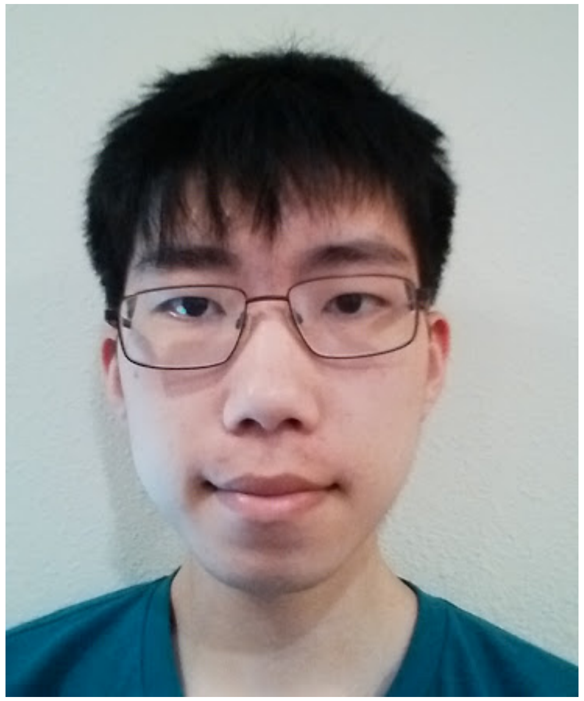


- Use GRNNs for tracking, dynamics, learning, perceptual front-end for RL, robotic learning

Thank you!



Fish Tung



Ricson Chen

Ziyan Wang

- Learning spatial common sense with geometry-aware recurrent networks, F. Tung, R. Cheng, K.F., arxiv
- Geometry-Aware Recurrent Neural Networks for Active Visual Recognition , R. Cheng, Z. Wang, K.F., NIPS 2018