Carnegie Mellon
School of Computer Science

Deep Reinforcement Learning and Control

Perceptual front-ends in RL

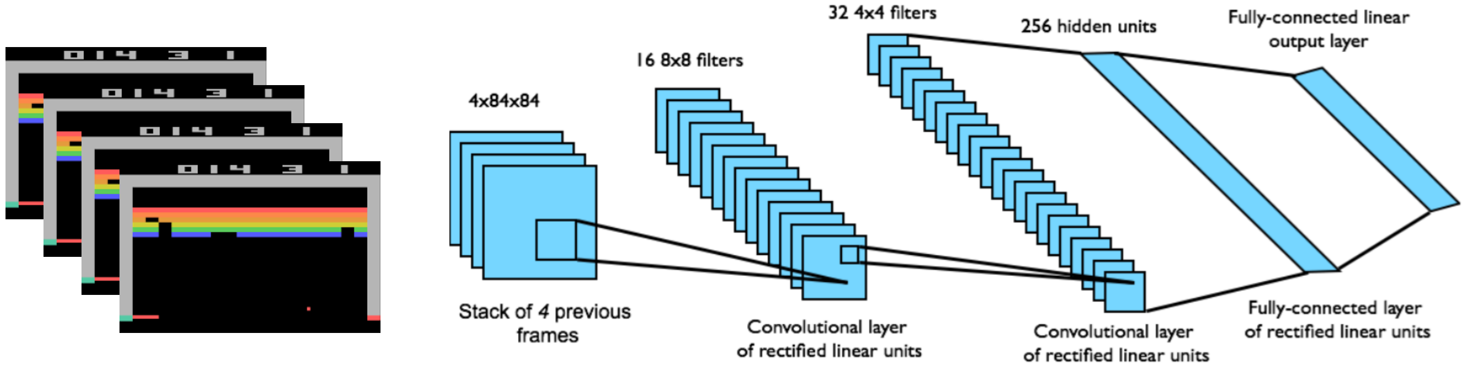
Katerina Fragkiadaki

This lecture: a close look into visual state representations for RL

- Consider what previous works use as perceptual front-end
- 3D aware feature representation

Visual frame concatenation

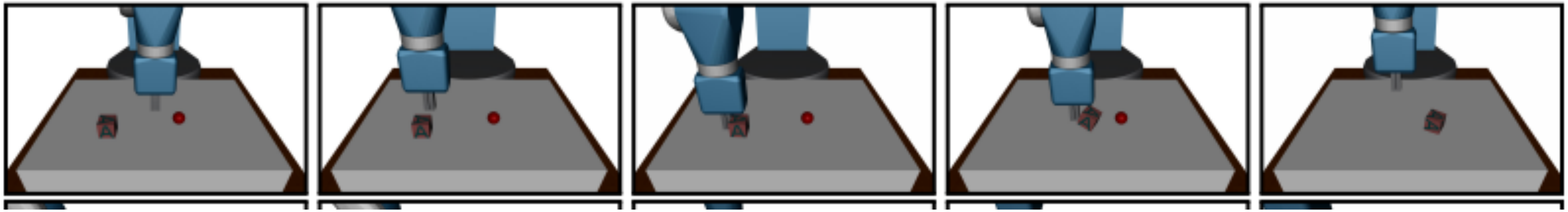
- k frame concatenation+2D convolutions



Learning to play atari games with deep reinforcement learning,2013

3D object/robot locations/poses

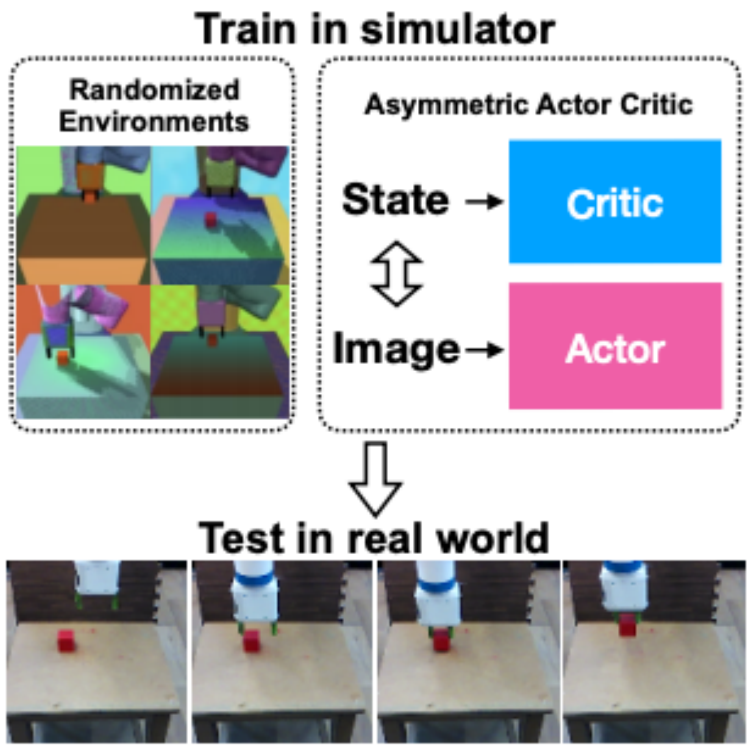
- Angles and velocities of all robot joints as well as 3D positions, rotations and velocities (linear and angular) of all objects



Hindsight experience replay

3D object/robot locations/poses

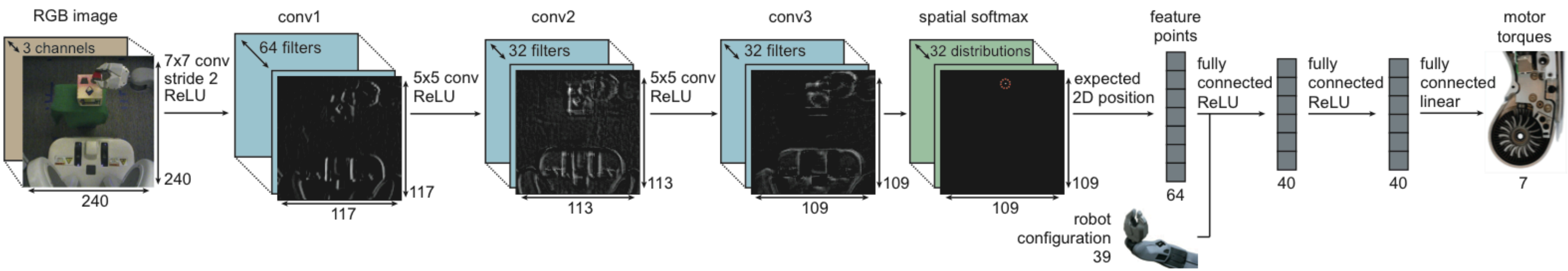
- Angles and velocities of all robot joints as well as 3D positions, rotations and velocities (linear and angular) of all objects for the critic
- Visual frame concatenation for the actor!
- Q: Why having different input for criti and actor is useful?



Asymmetric Actor Critic for Image-Based Robot Learning, Pinto et al.

Spatial Softmax

- frame concatenation as input
- tight bottleneck being the K 2D x,y coordinates of k keypoints



- For each feature map, ``flatten” it and compute a softmax
- Then take X and Y grid coordinates and compute the corresponding weighted averages
- Imposes a very tight bottleneck and avoids overfitting

End-to-end learning of visuomotor policies, Levine et al. 2015

There is something fundamentally unsatisfying about the perceptual front-ends used out there…

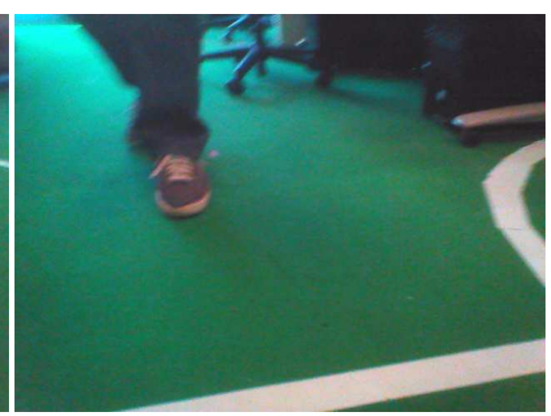
Internet Vision

Photos taken by people (and uploaded on the Internet)

Mobile (Embodied) Computer Vision

Photos taken by a NAO robot during a robot soccer game

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

Image Understanding as Inverse Graphics

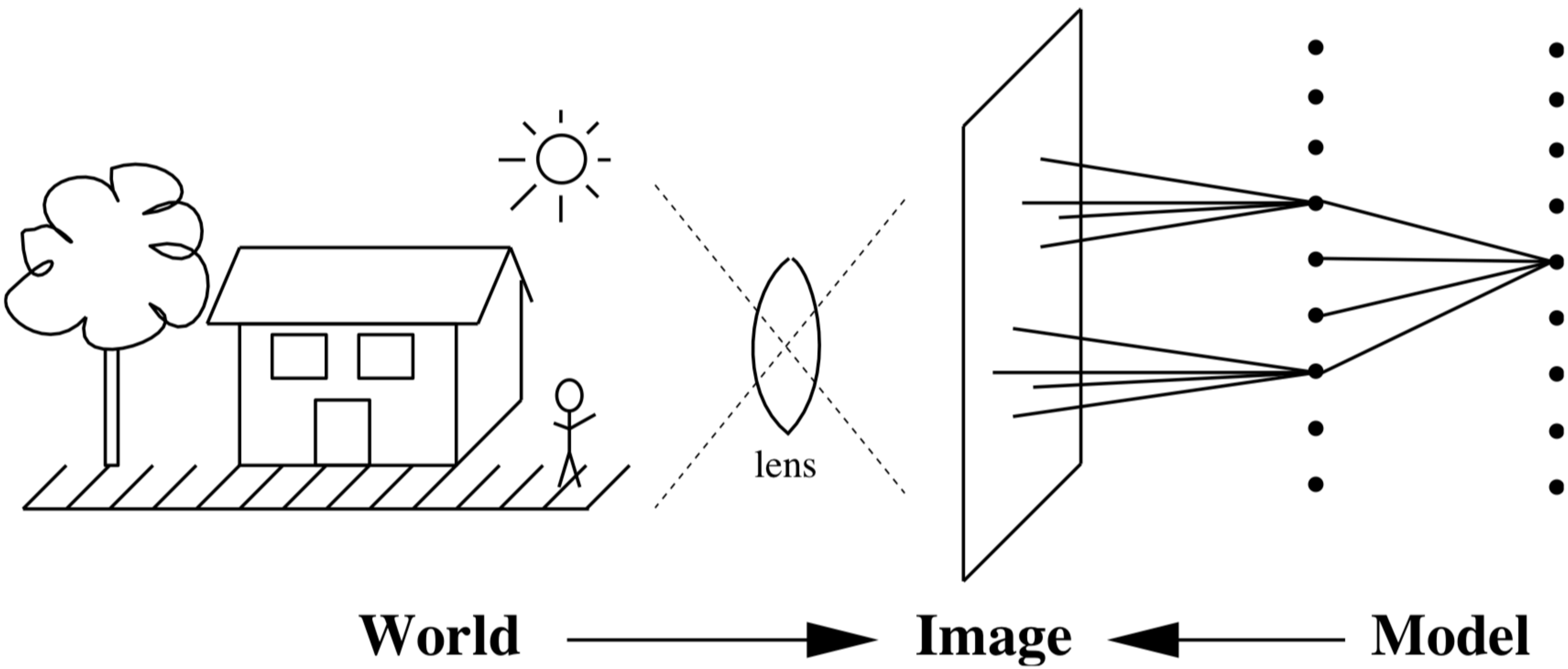
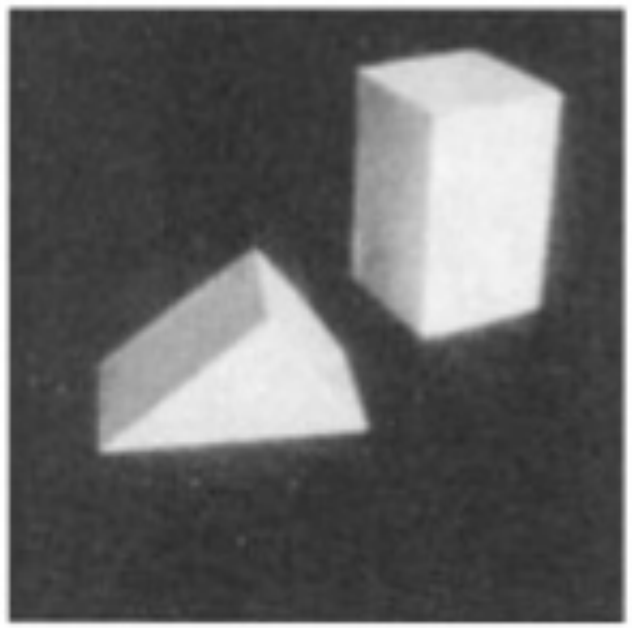


Image Understanding as Inverse Graphics

Blocks world

Larry Roberts



Input image

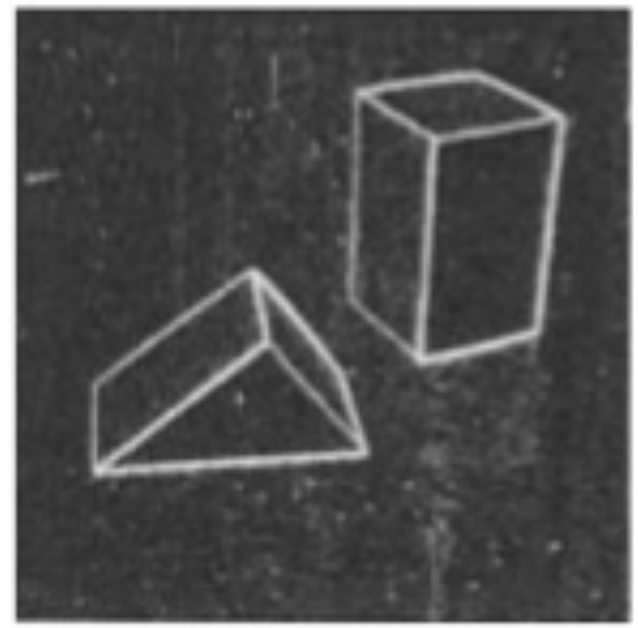
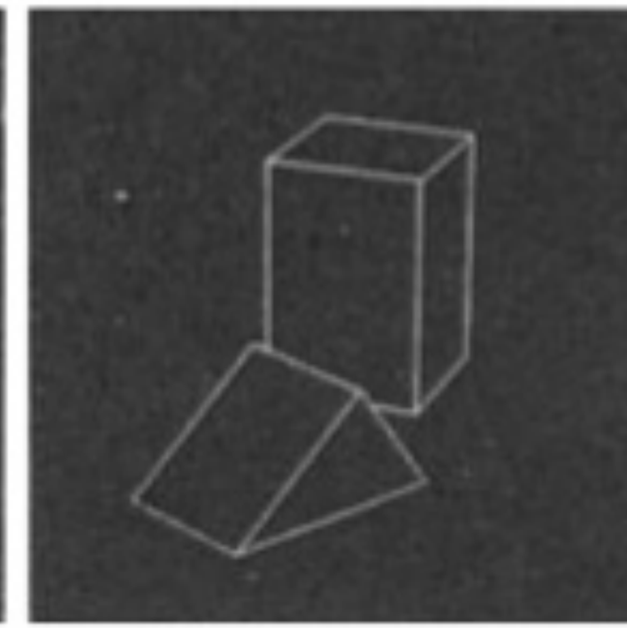


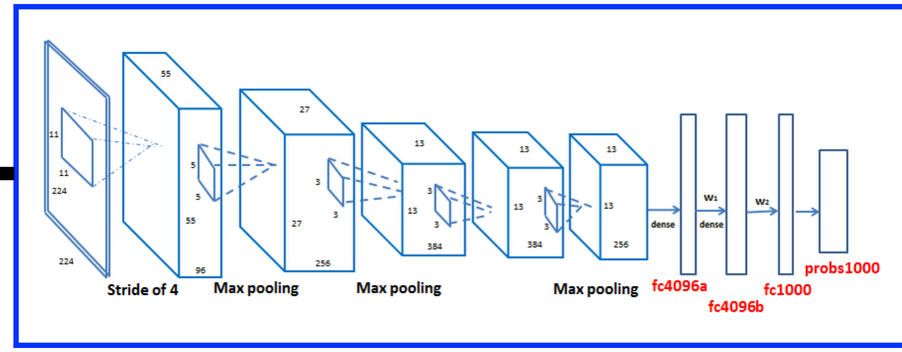
Image gradient



Computed 3D model rendered from a new viewpoint

Machine perception of Three-Dimensional solids, MIT 1965

3D Models are impossible and unecessary



Steering angle

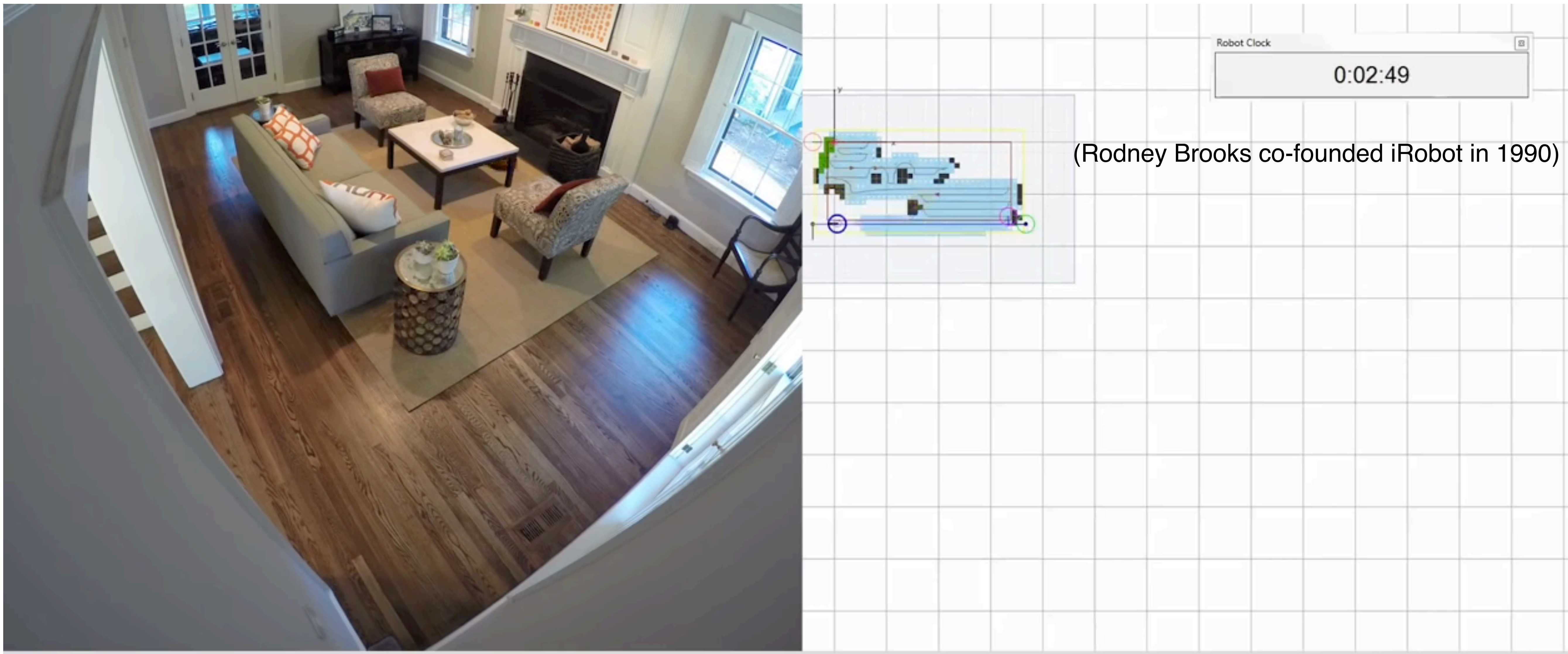
``Internal world models which are complete representations of the external environment, besides being impossible to obtain, are not at all necessary for agents to act in a competent manner."

``…(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)

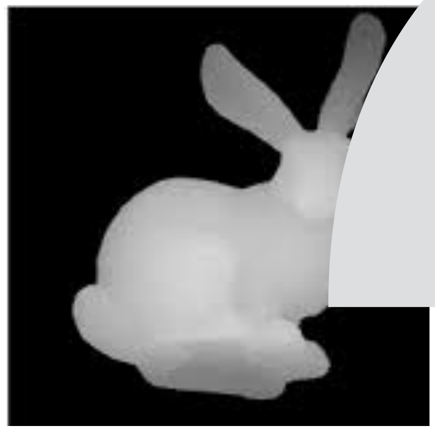
Internet and Mobile Perception have developed independently and have each made great progress

- 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

To 3D or not to 3D?

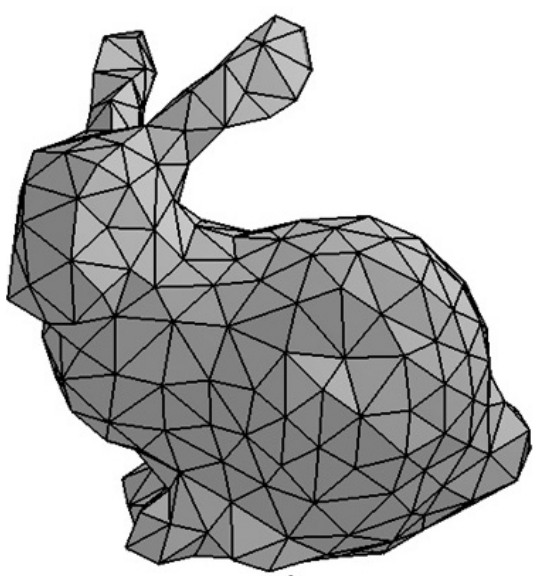
And if to 3D, what 3D representation to use?

depth map

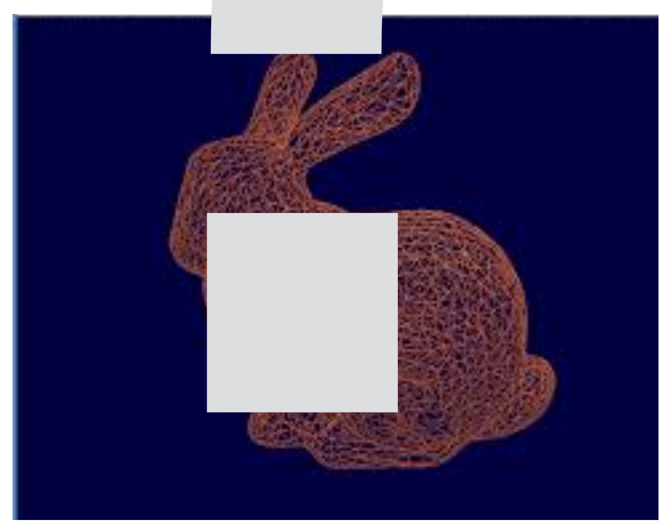


face normals

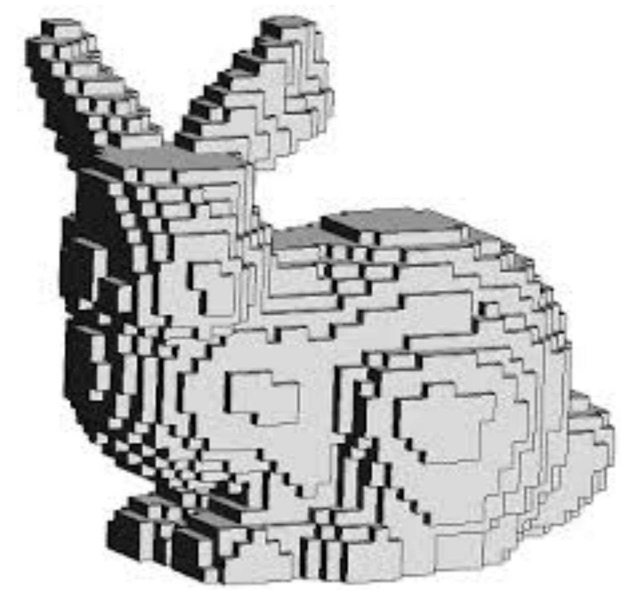
3d mesh



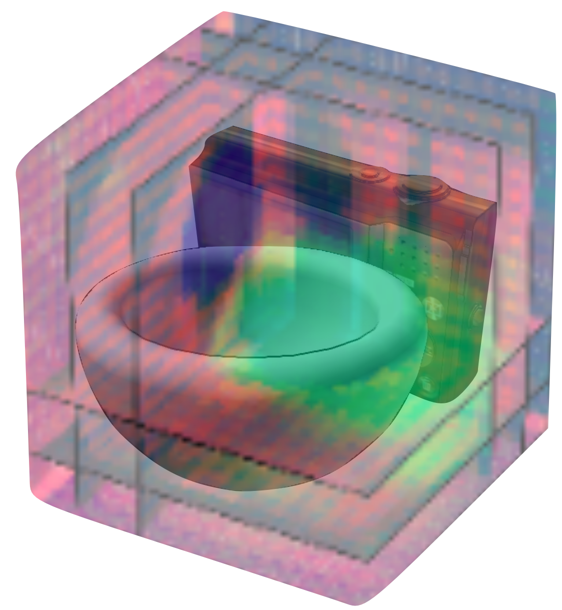
3d po cloud



3d voxel occupancy



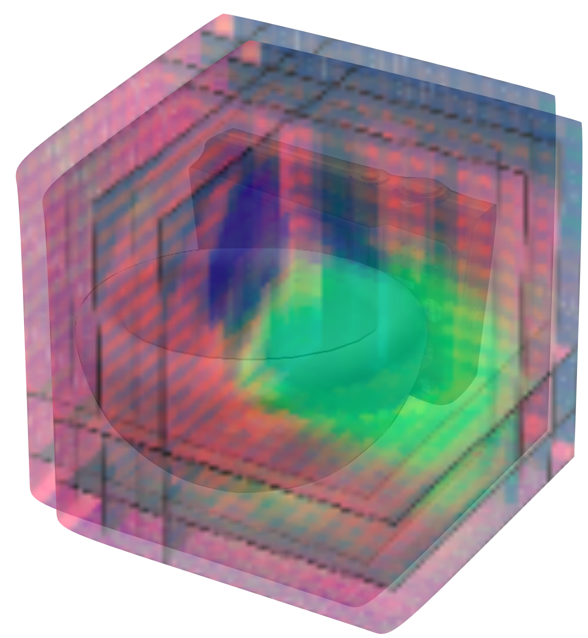
This talk: To 3D using 3D feature tensors



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

3 spatial dimensions, 1 feature dimension

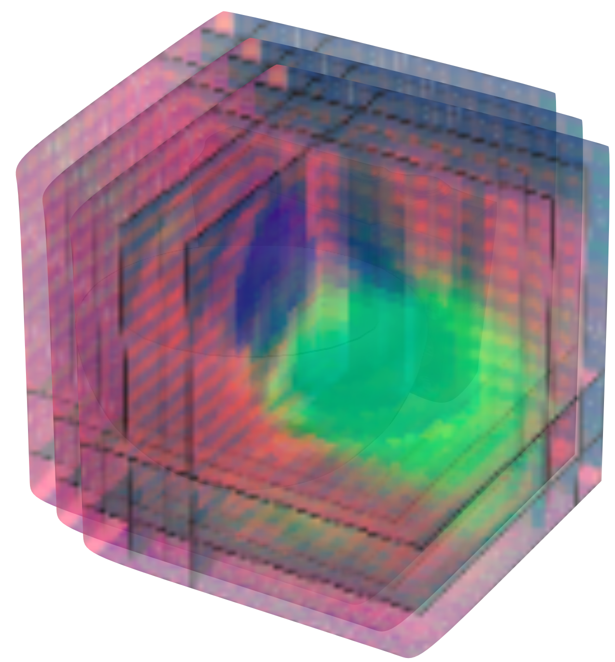
This talk: To 3D using 3D feature tensors



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

3 spatial dimensions, 1 feature dimension

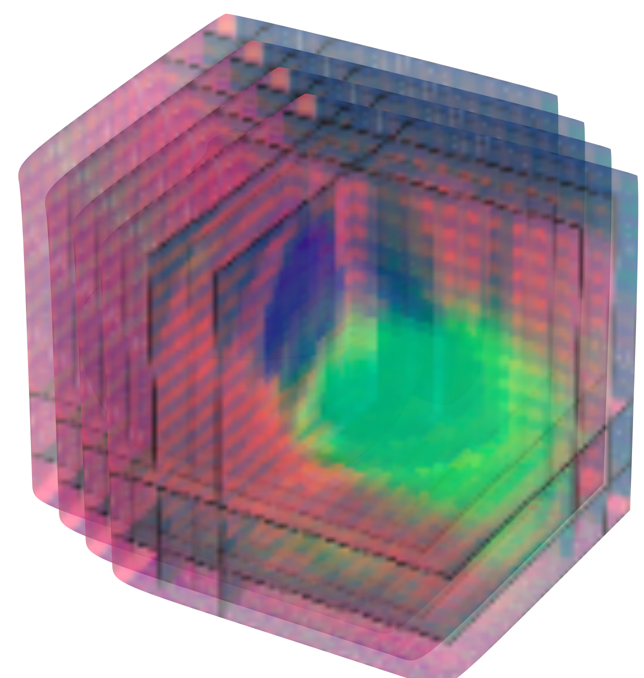
This talk: To 3D using 3D feature tensors



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

3 spatial dimensions, 1 feature dimension

This talk: To 3D using 3D feature tensors

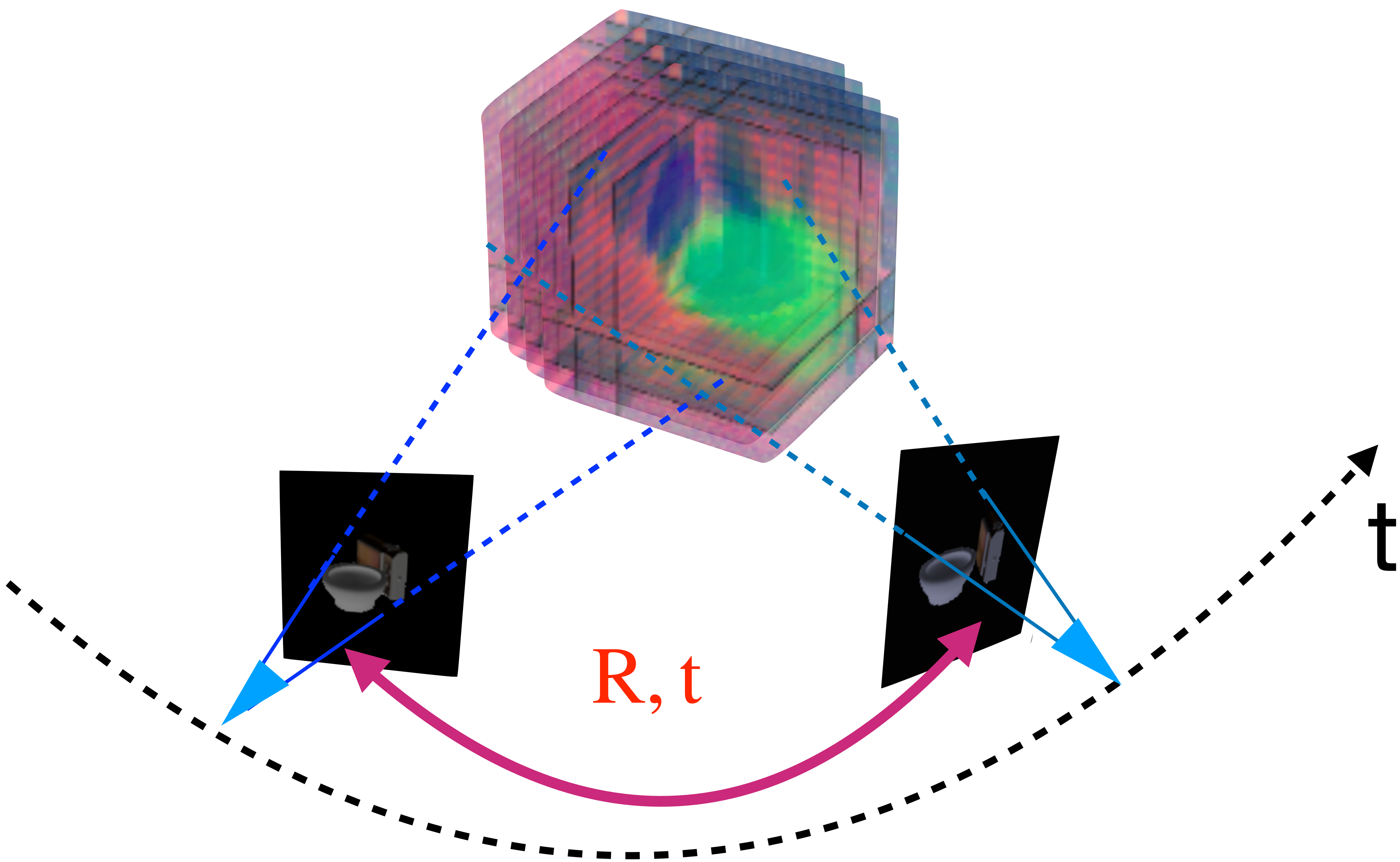


$H \times W \times D \times C$

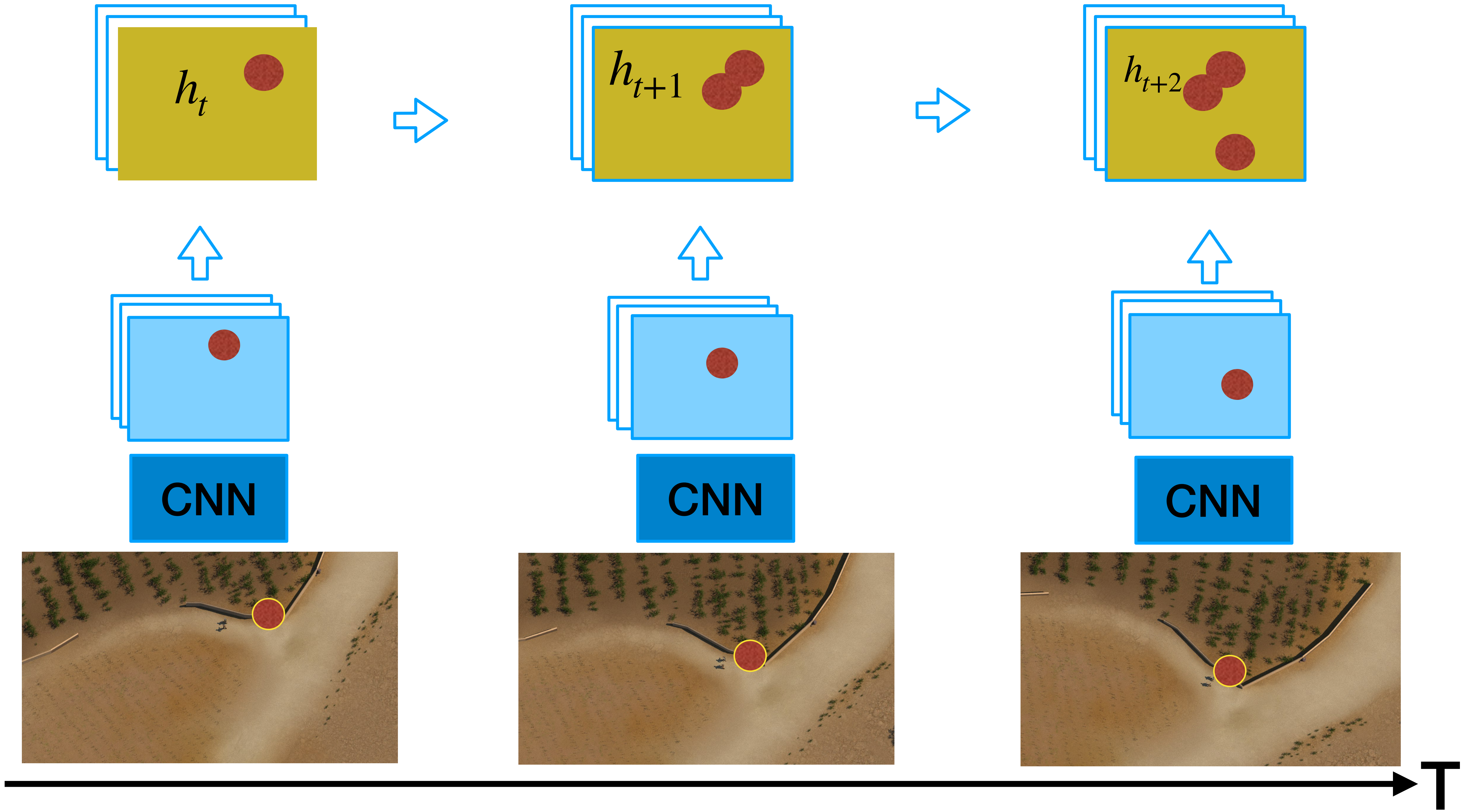
3 spatial dimensions, 1 feature dimension

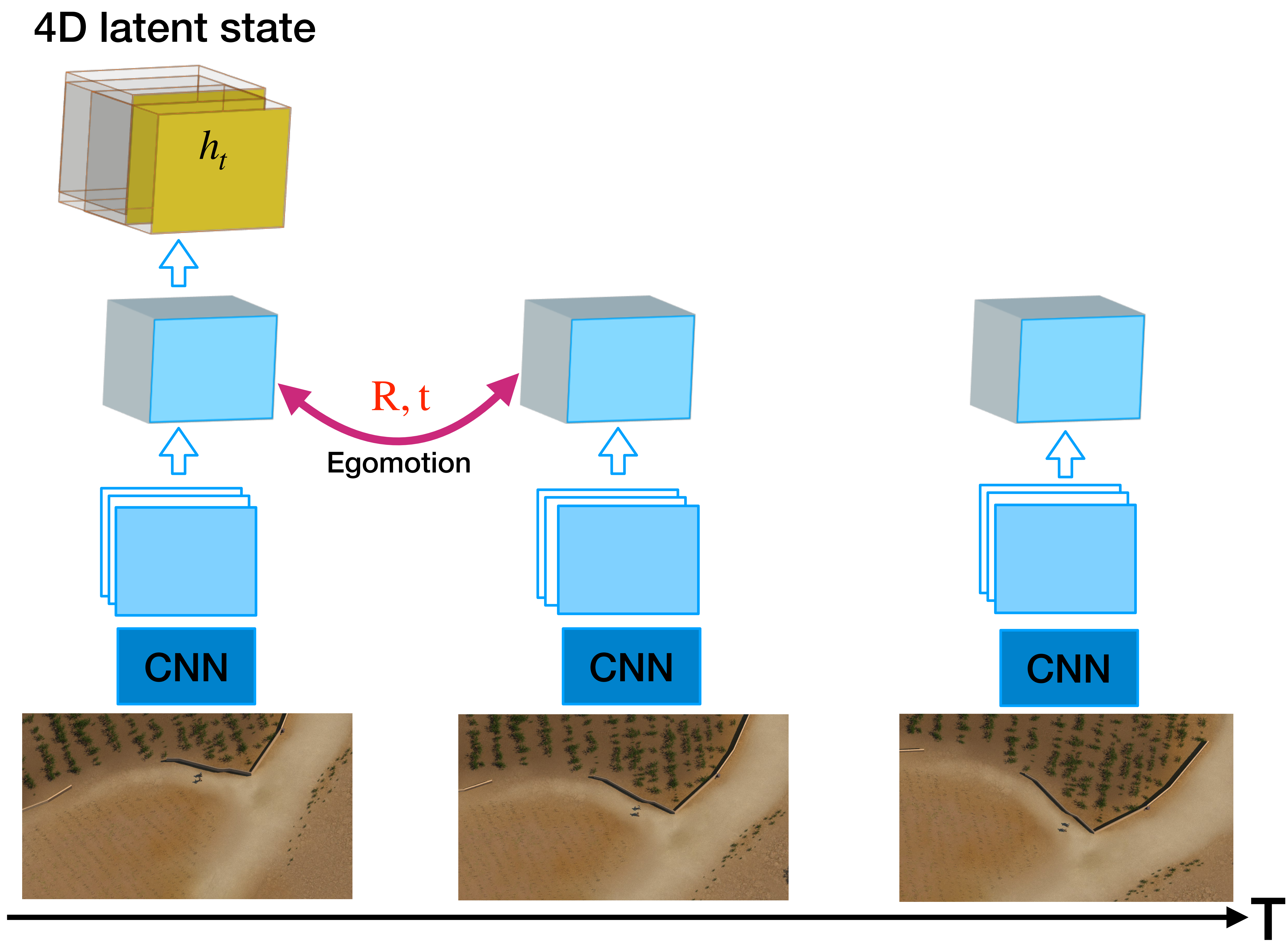
Geometry-Aware Recurrent Networks

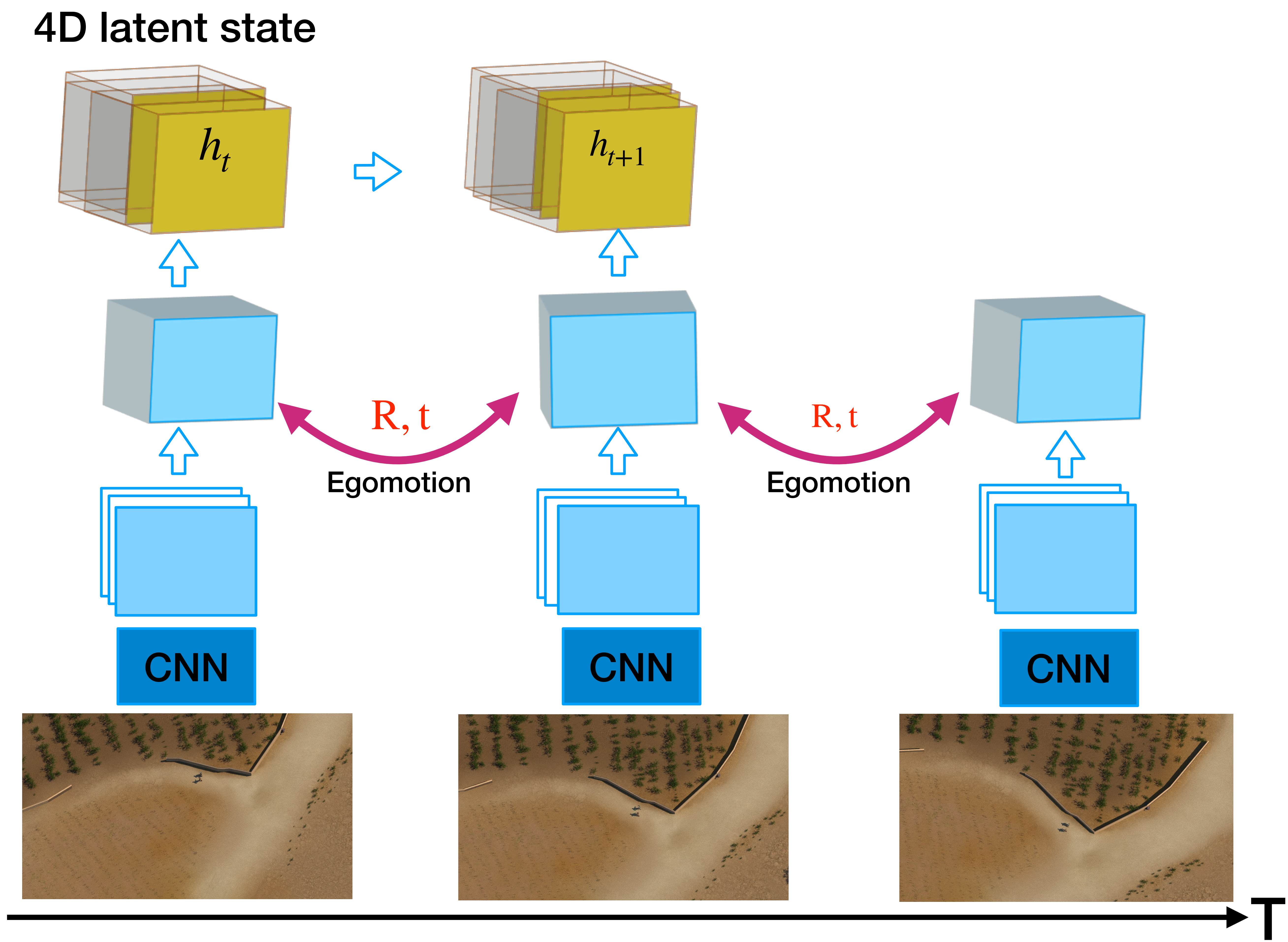
1. Hidden state: A 4D deep feature tensor, akin to a 3D (feature as opposed to pointcloud) map of the scene
2. Egomotion-stabilized hidden state updates

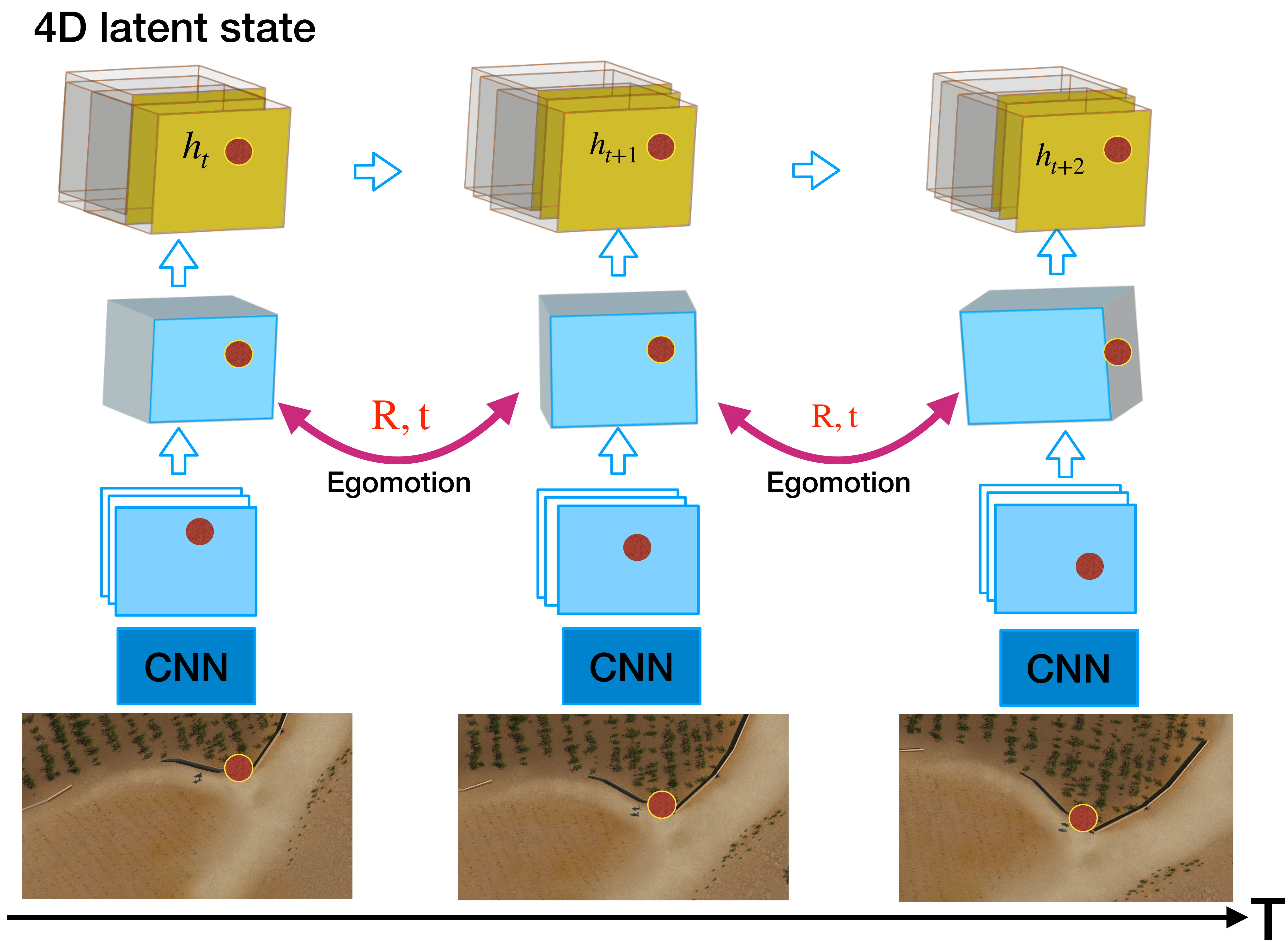


2D Recurrent networks, LSTMs, CONVLSTMs,..

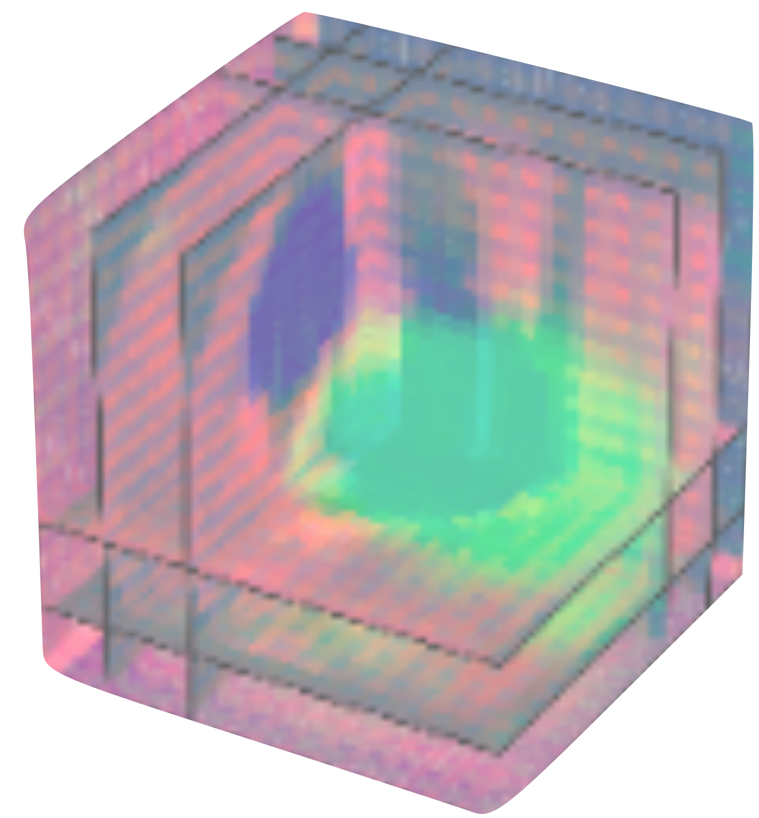



4D latent state
h_t
R, t
Egomotion
CNN
CNN
CNN
T


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

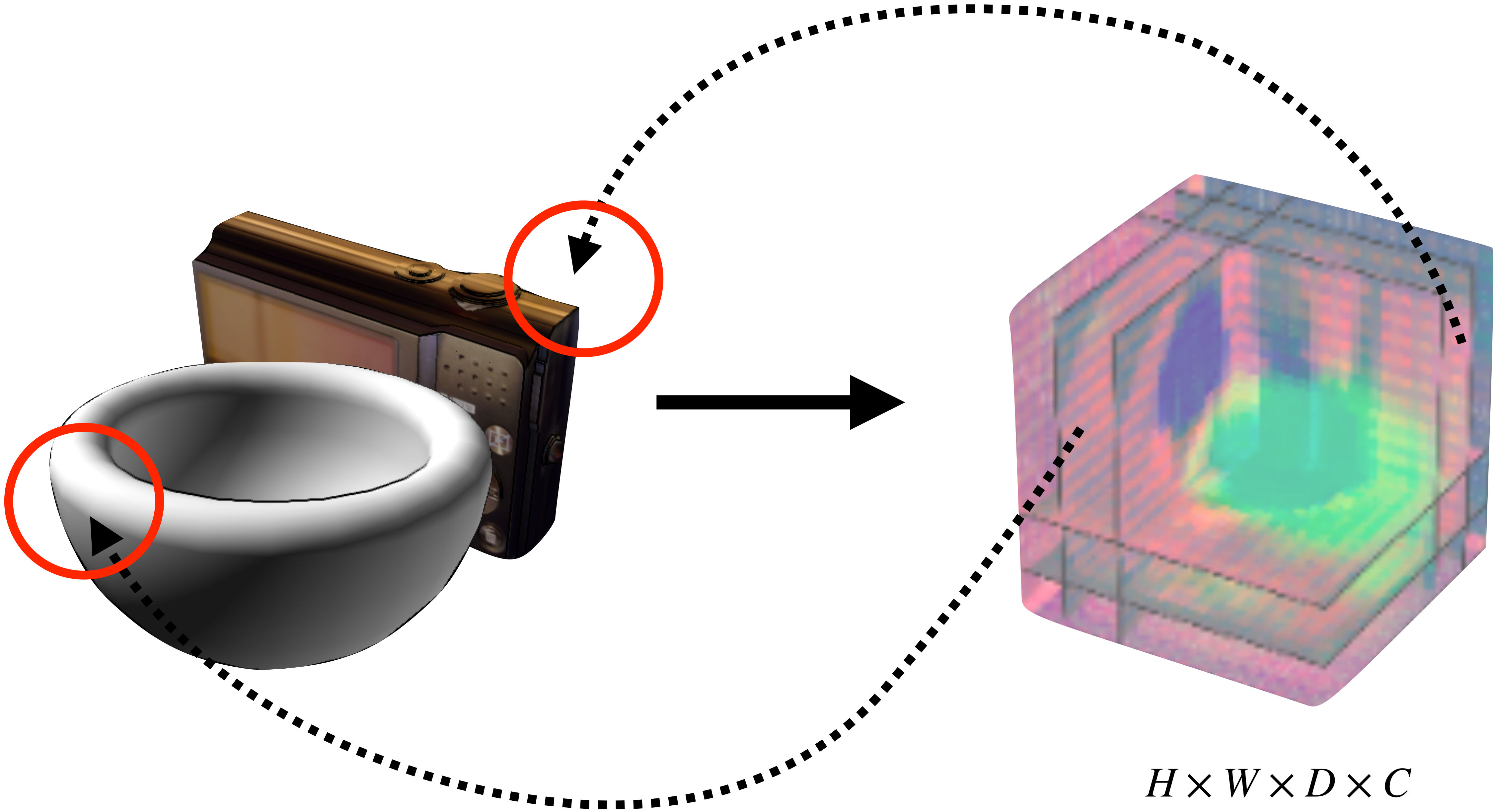

4D latent state
h_t
h_{t+1}
h_{t+2}
R, t
Egomotion
R, t
Egomotion
CNN
CNN
CNN
T

Geometry-Aware Recurrent Networks (GRNNs)

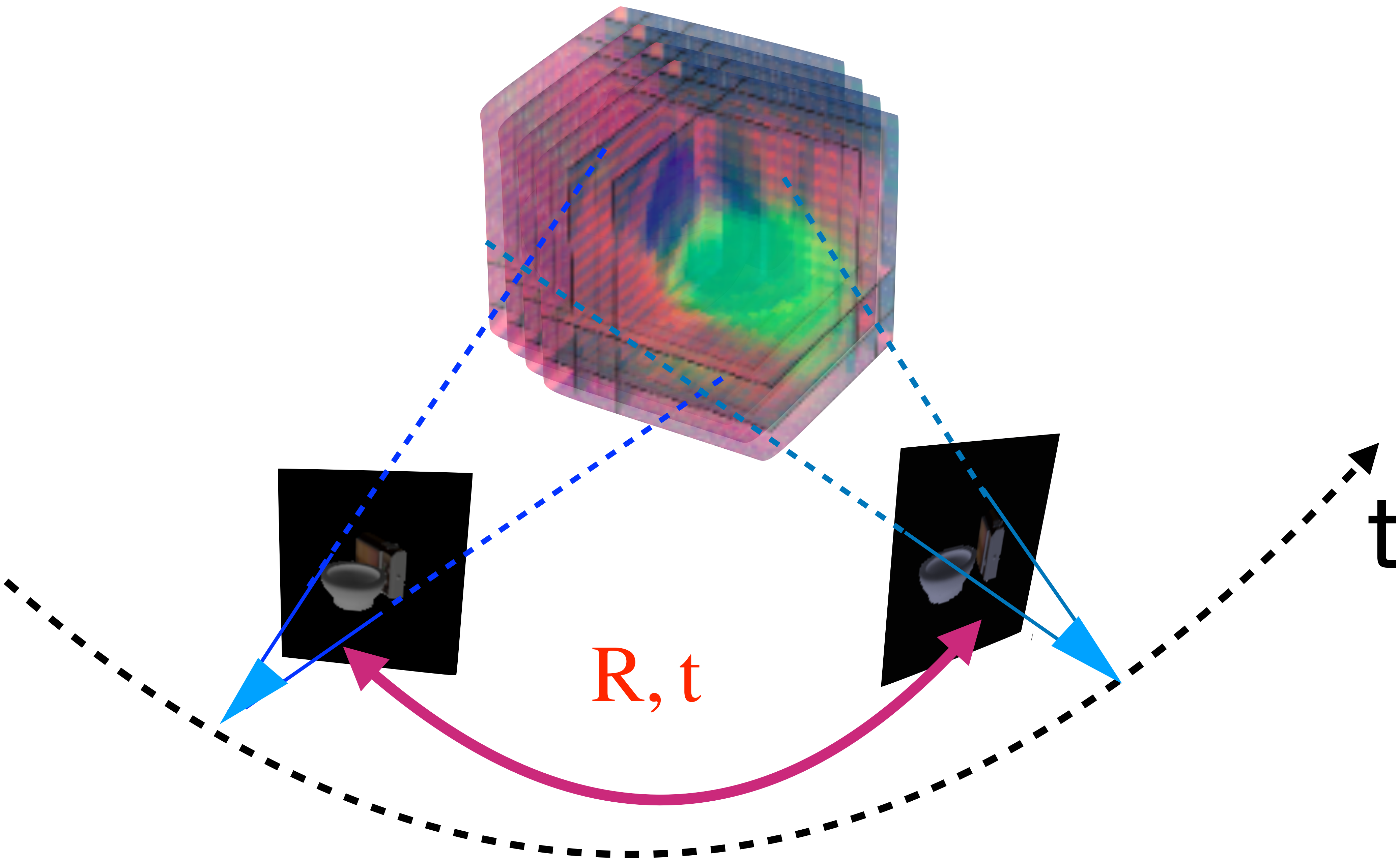


$H \times W \times D \times C$

Geometry-Aware Recurrent Networks (GRNNs)

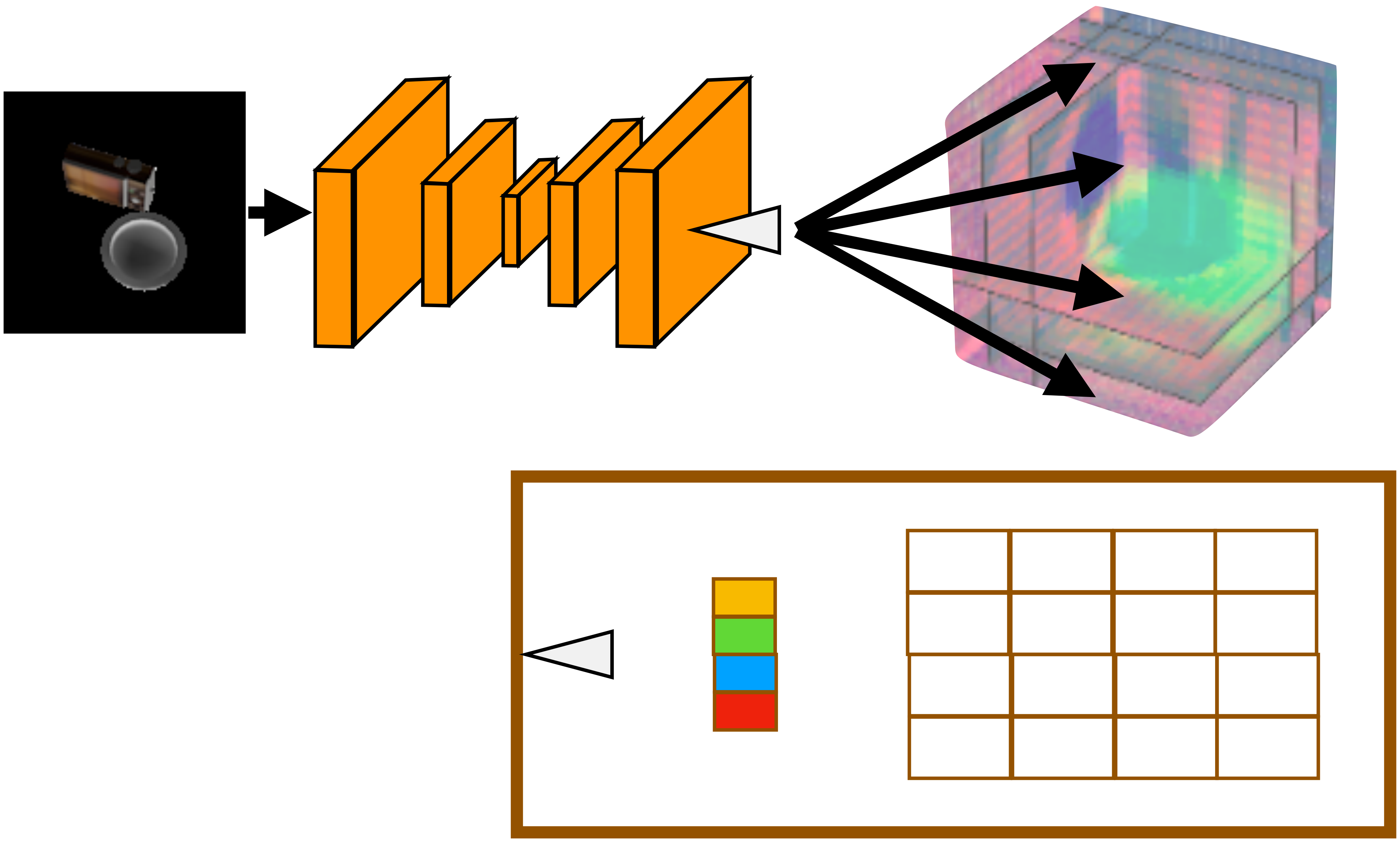


GRNNs

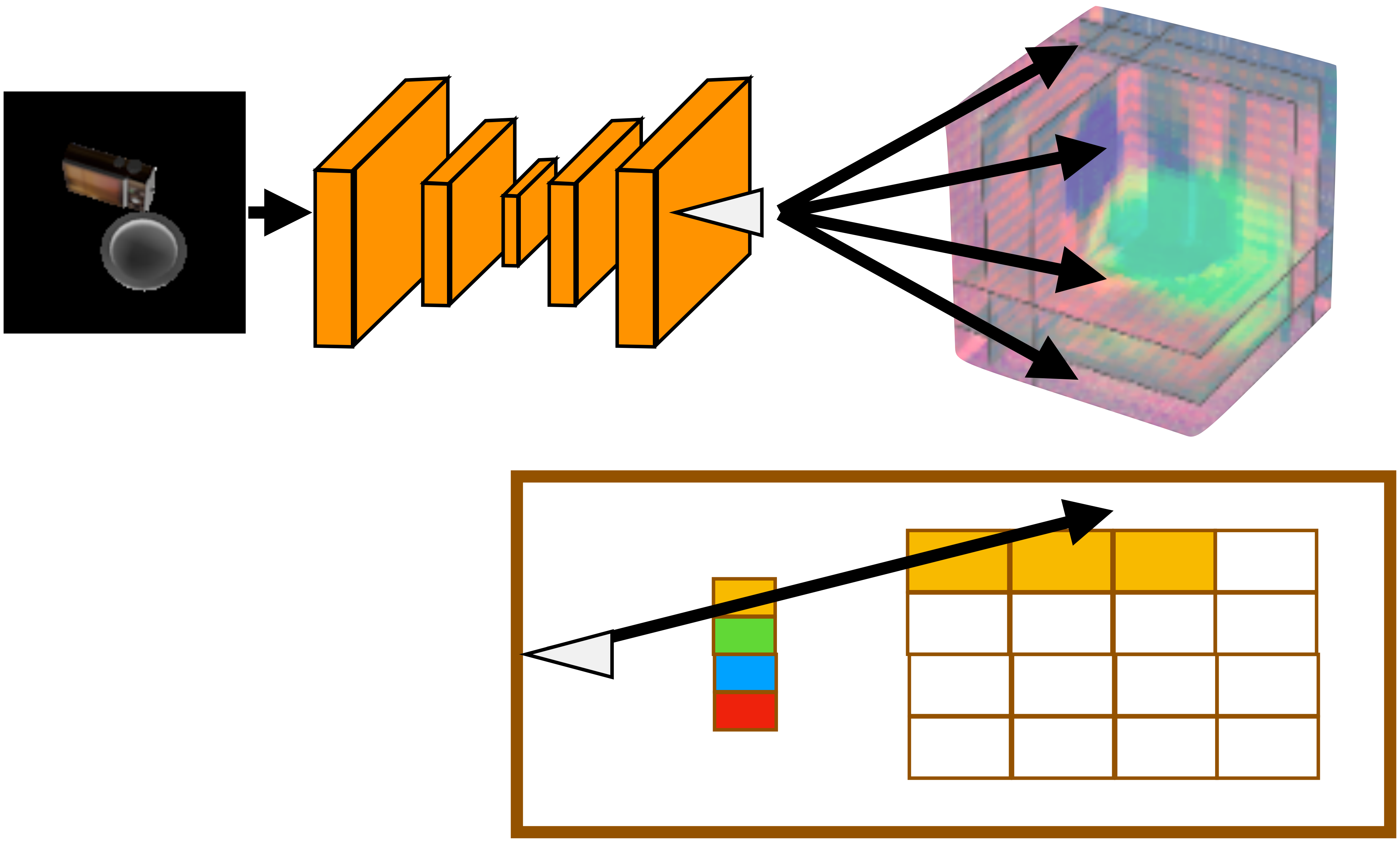


- A set of differentiable neural modules to learn to go from 2D to 3D and back
- A lot of SLAM ideas into the neural modules

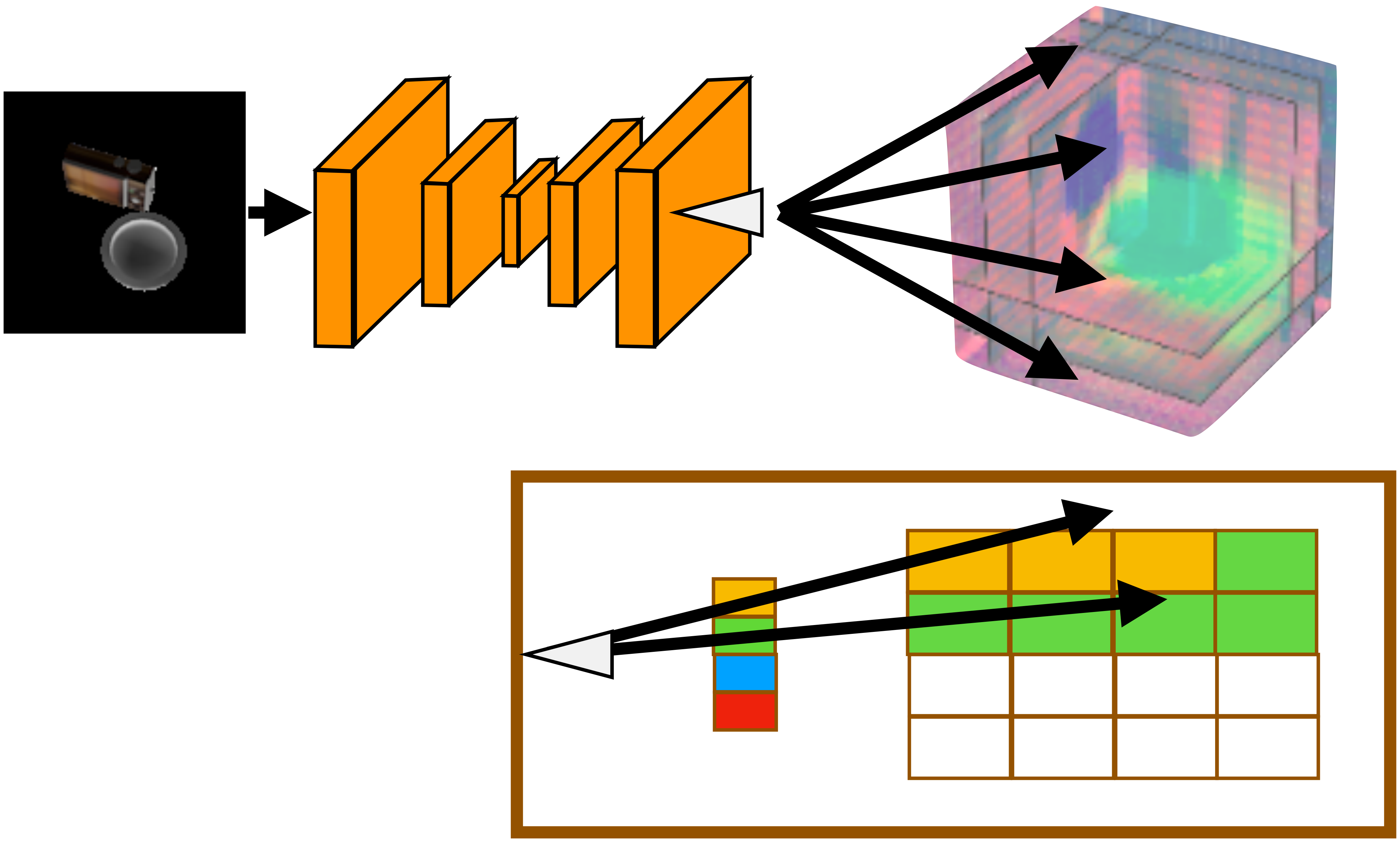
Unprojection (2D to 3D)



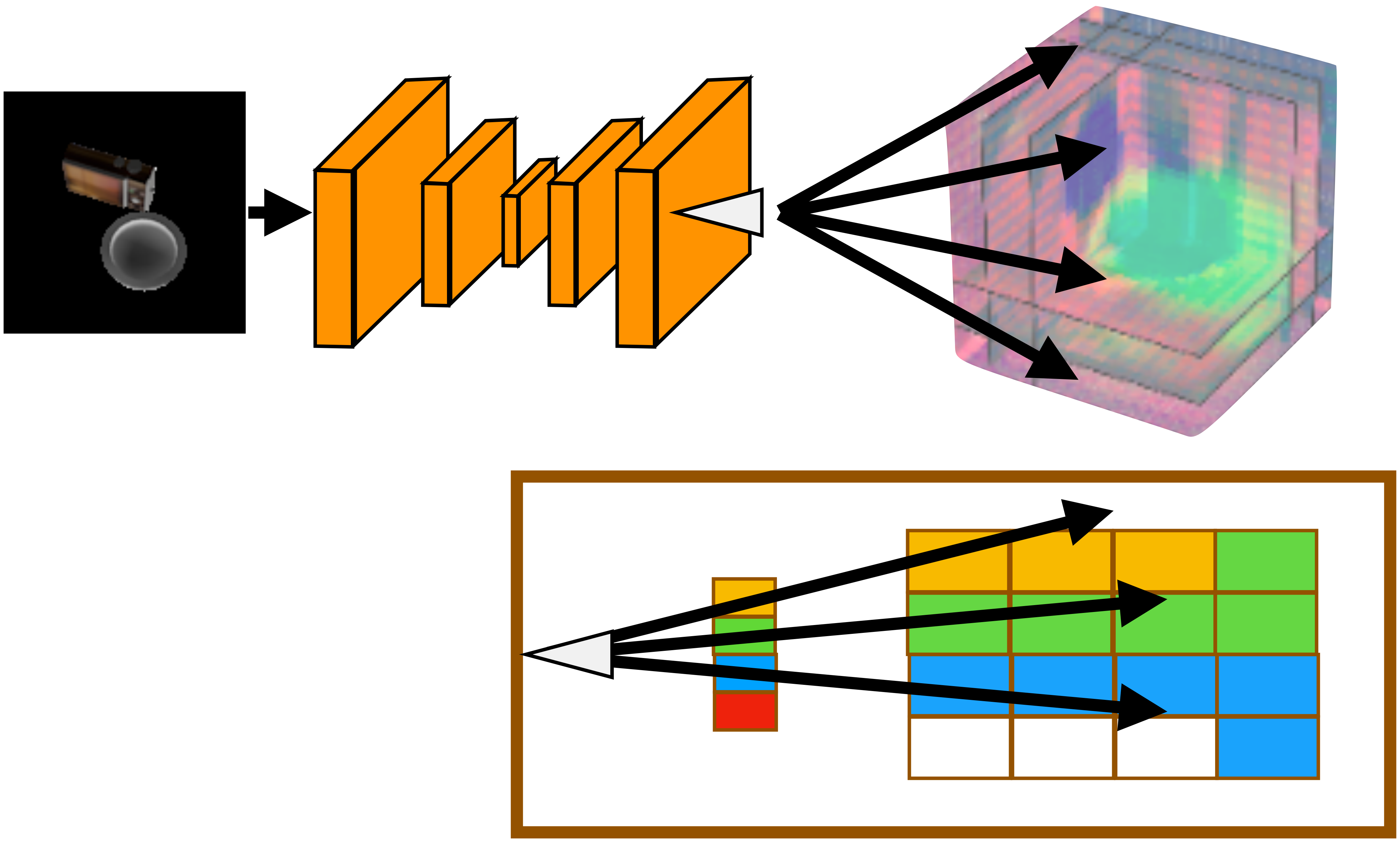
Unprojection (2D to 3D)



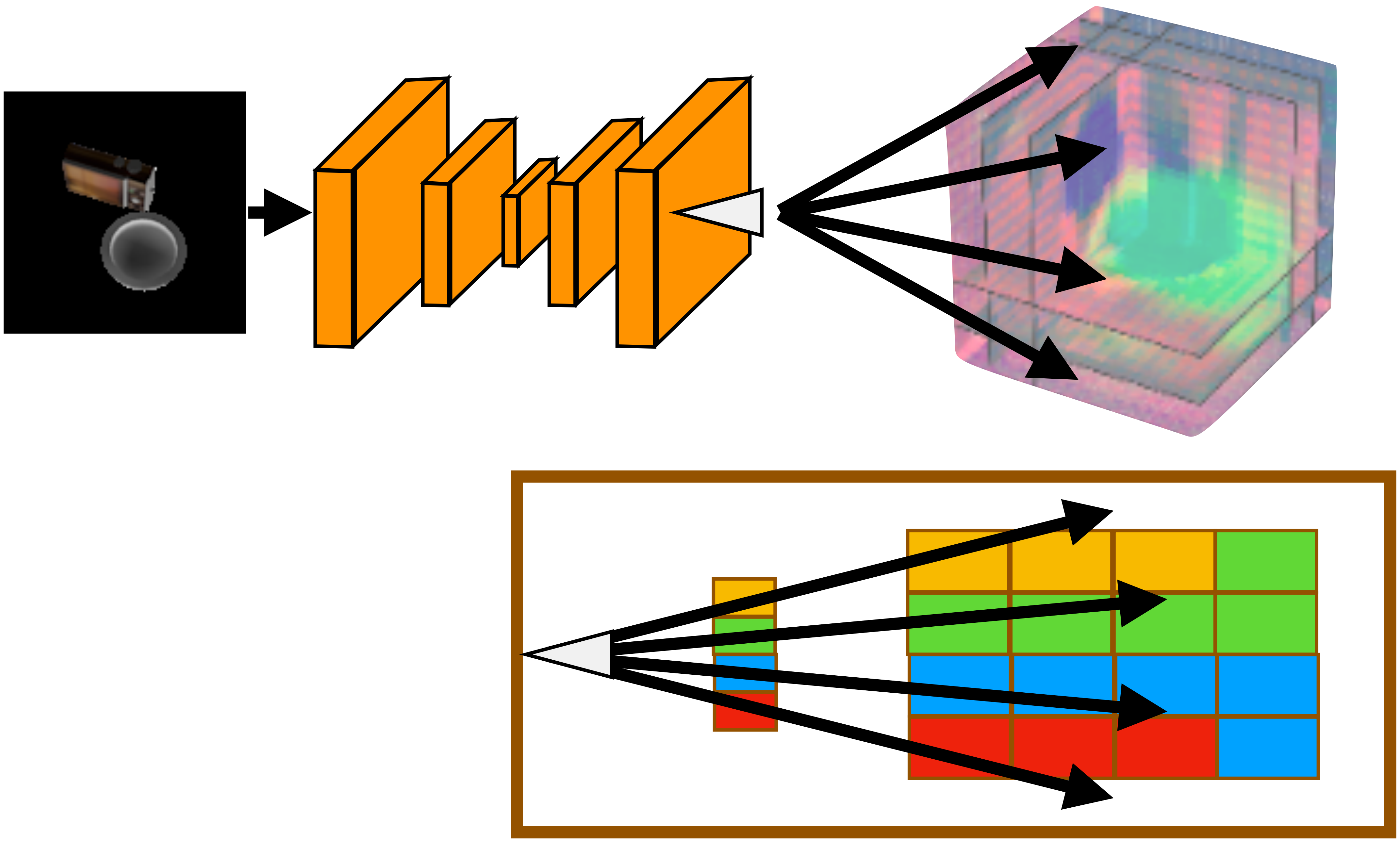
Unprojection (2D to 3D)



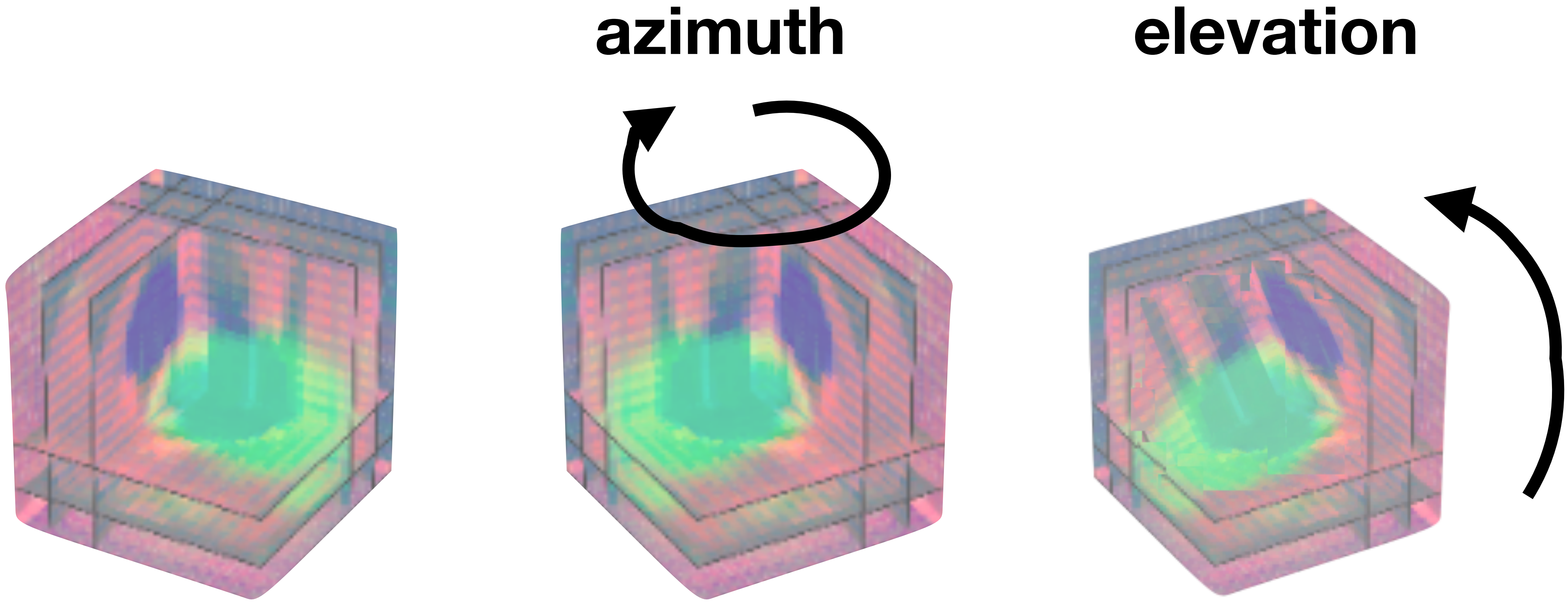
Unprojection (2D to 3D)



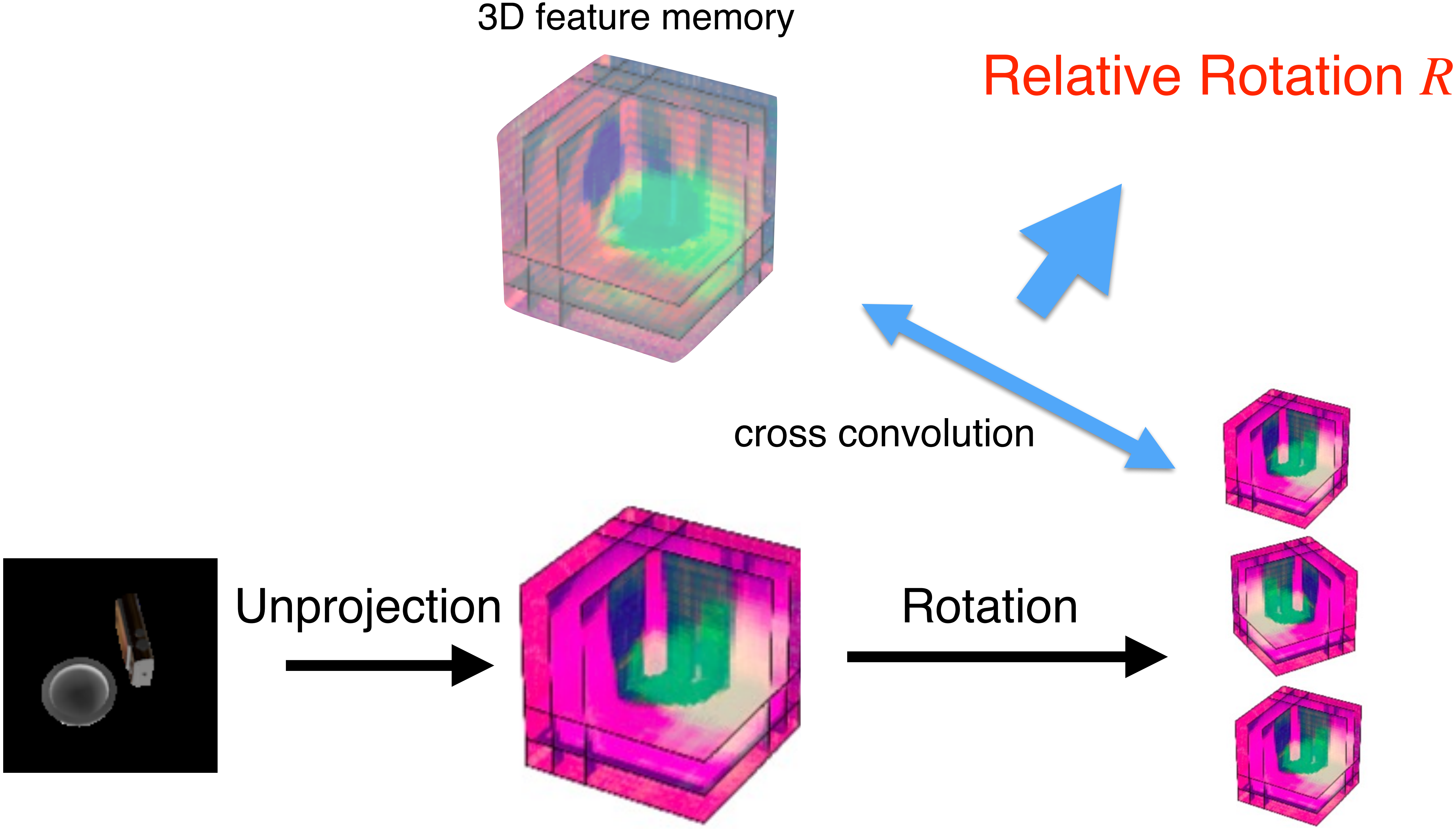
Unprojection (2D to 3D)



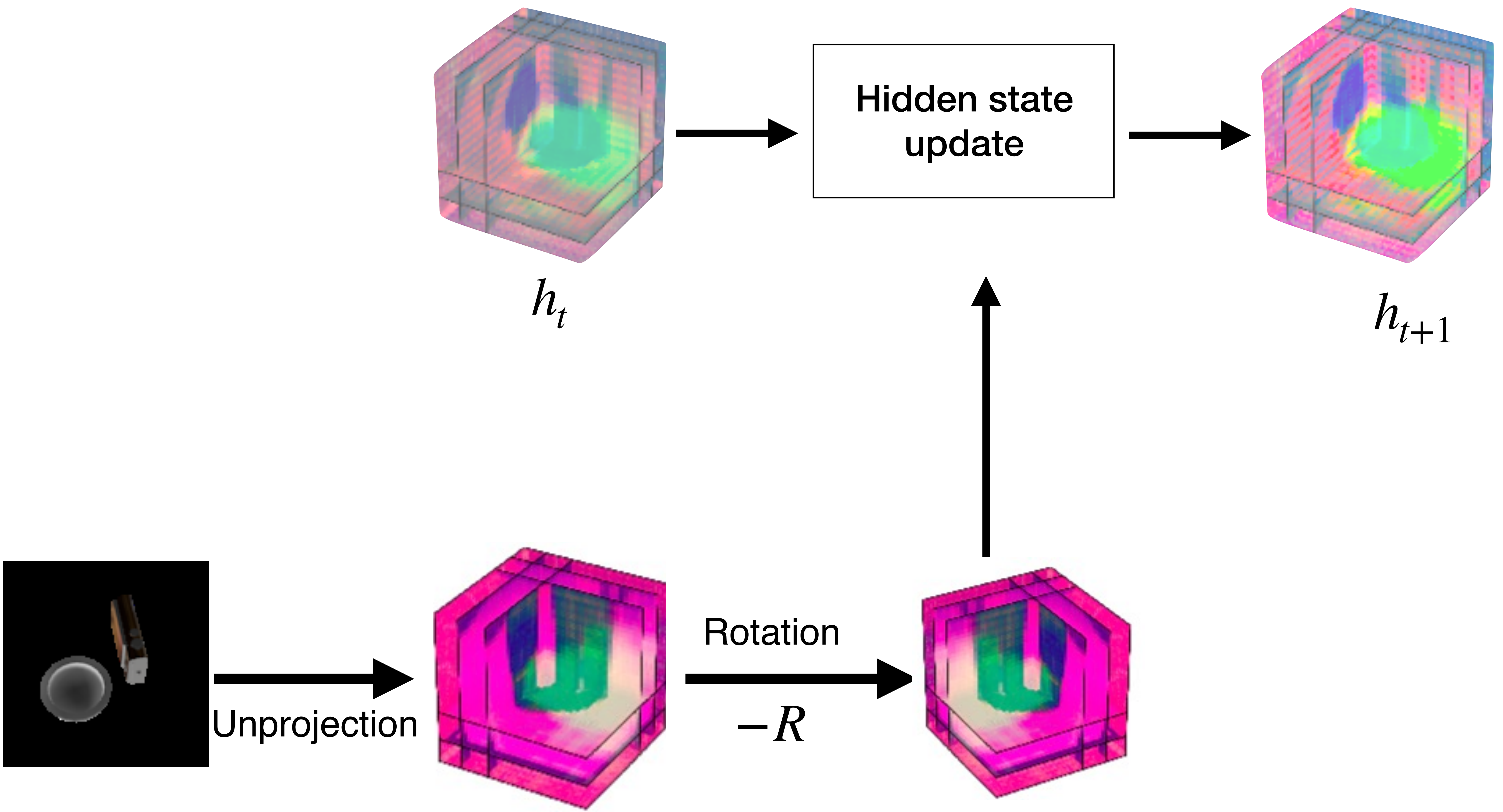
Rotation



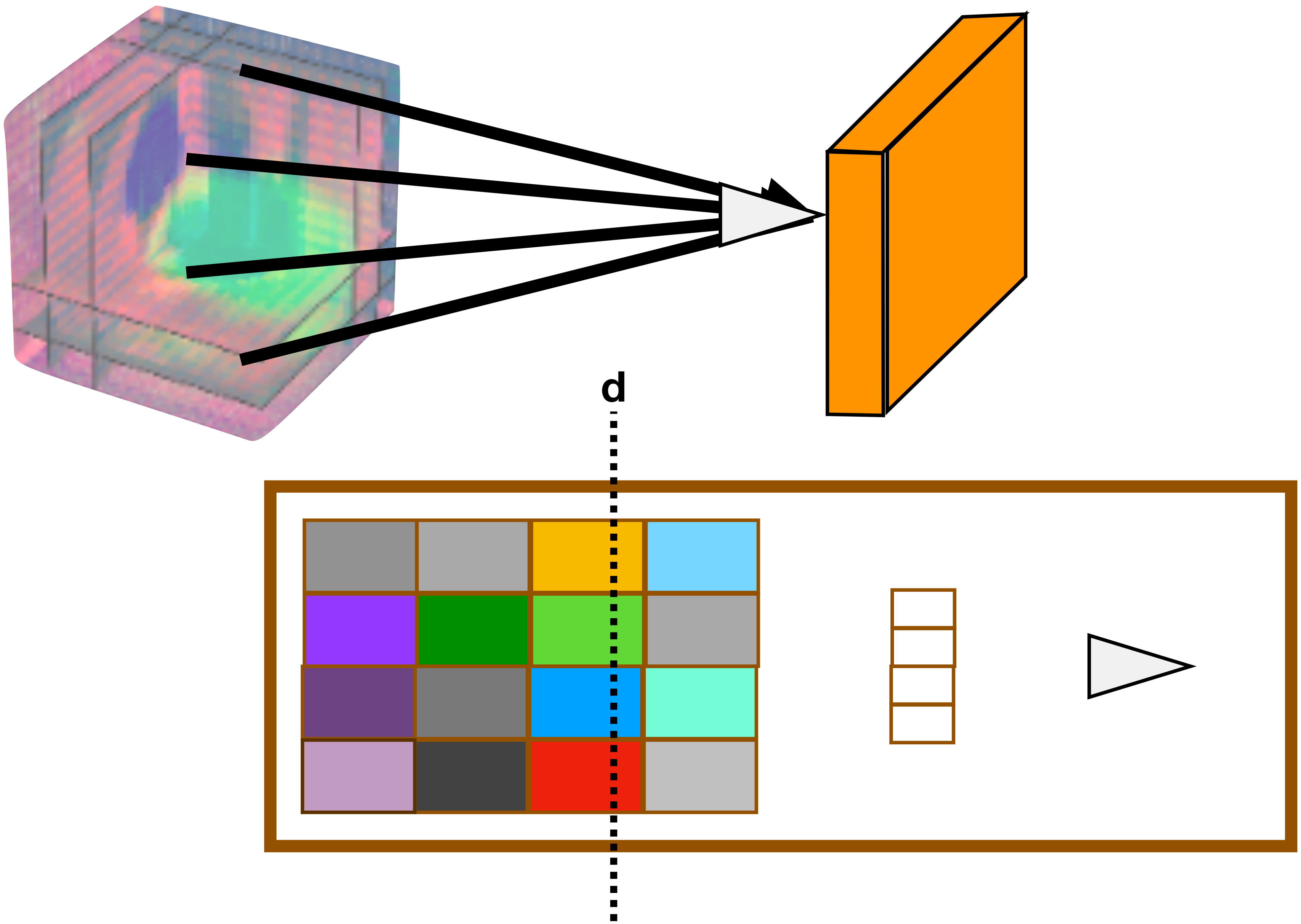
Egomotion-stabilized memory update



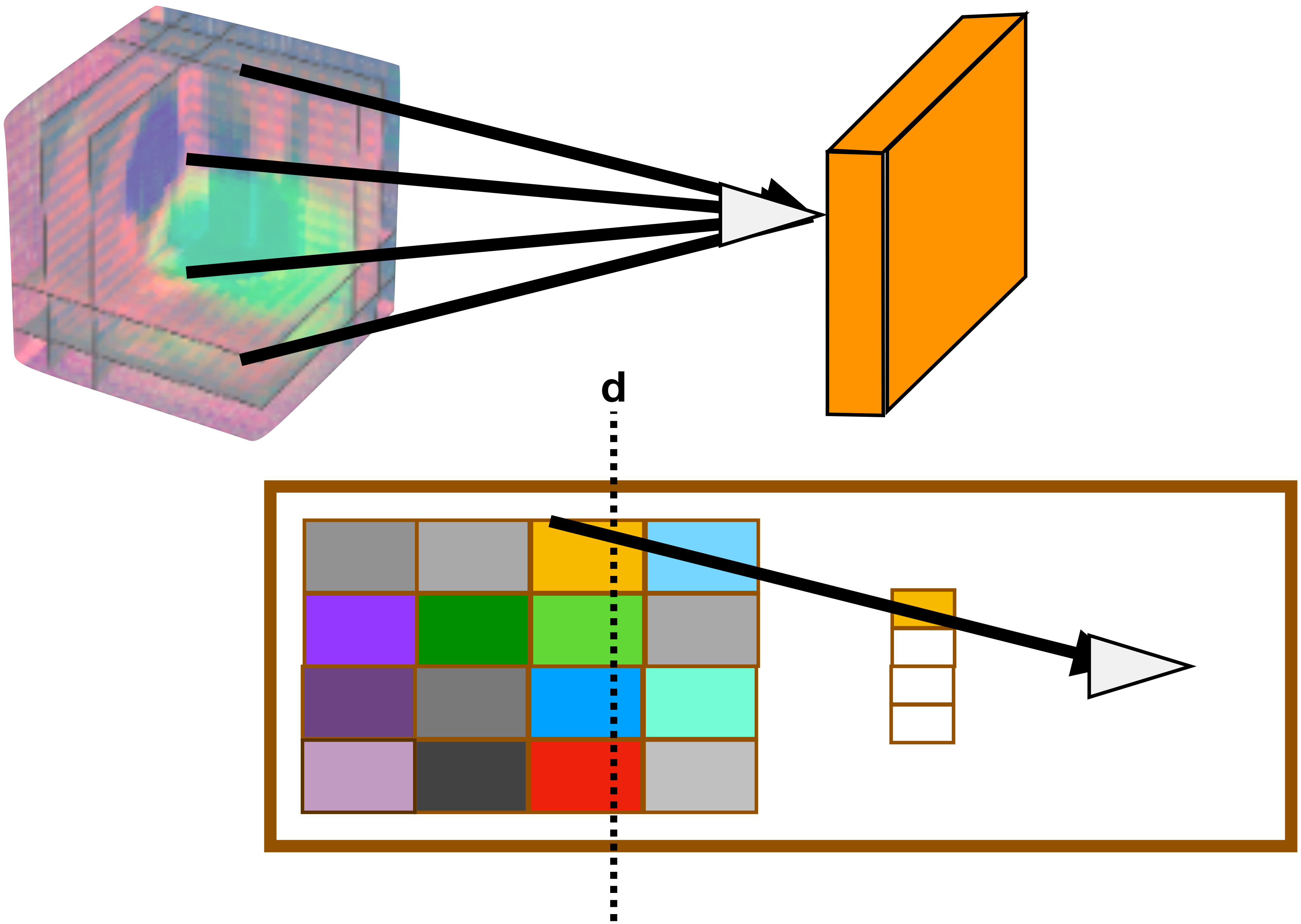
Egomotion-stabilized memory update



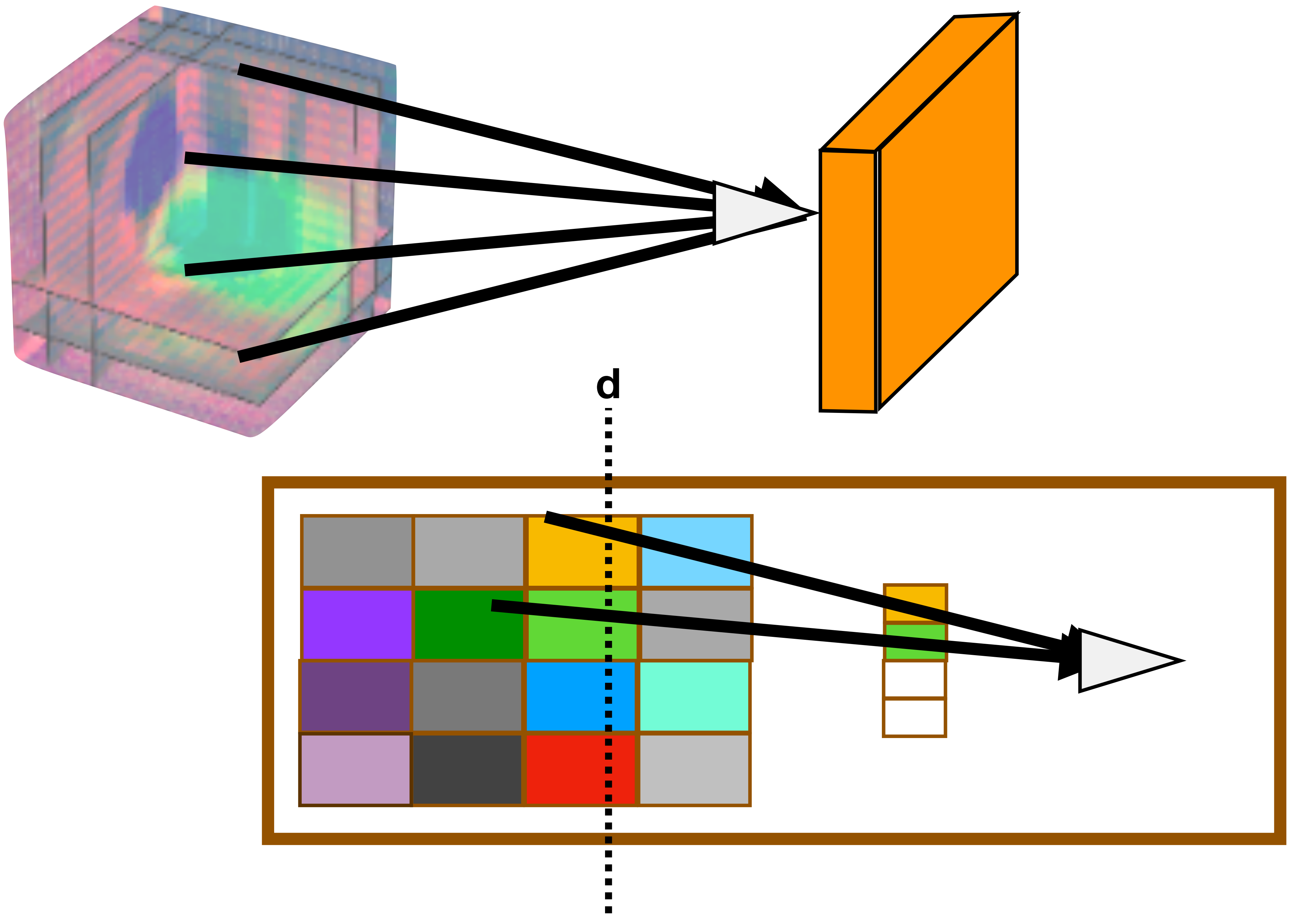
Projection (3D to 2D)



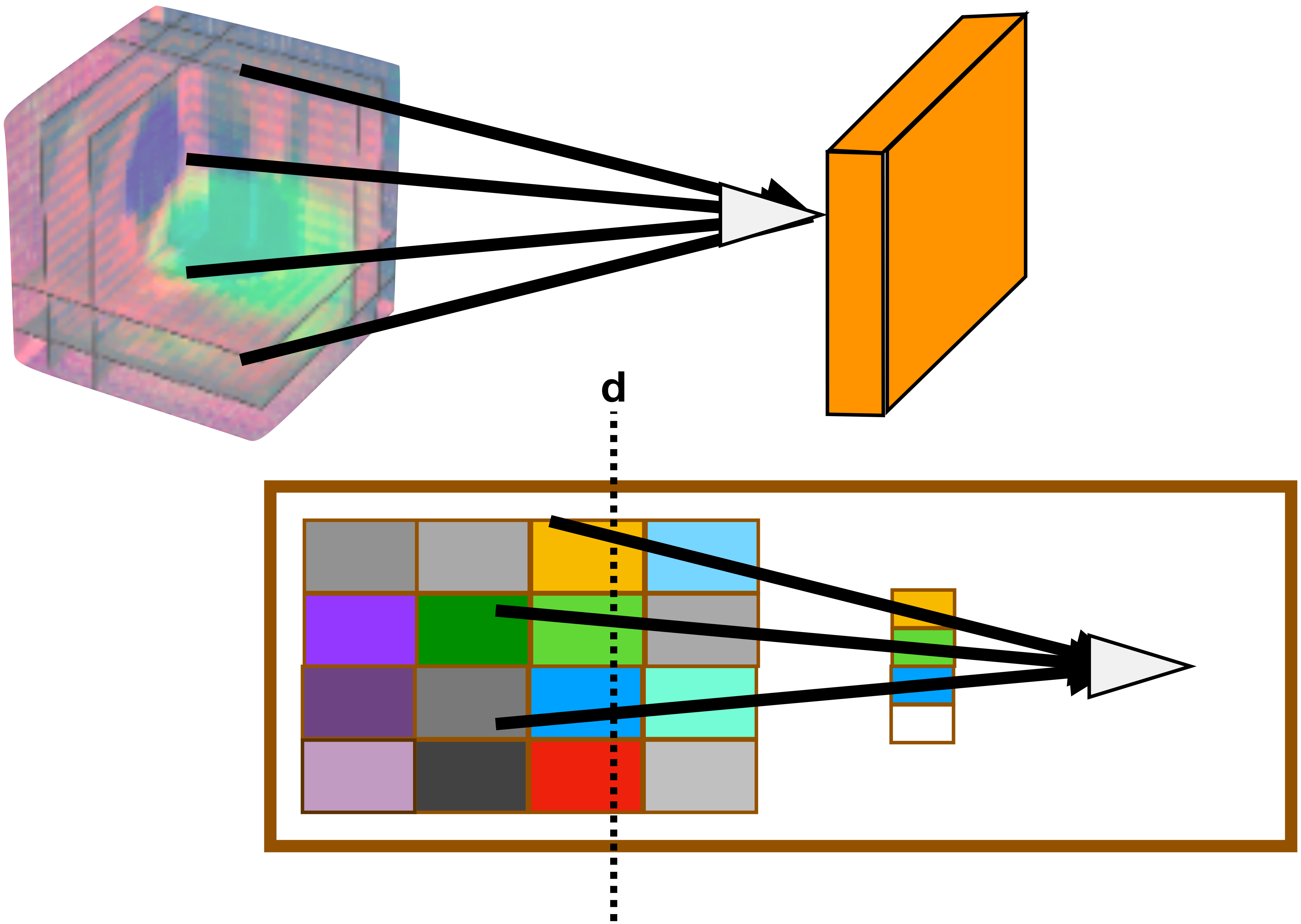
Projection (3D to 2D)



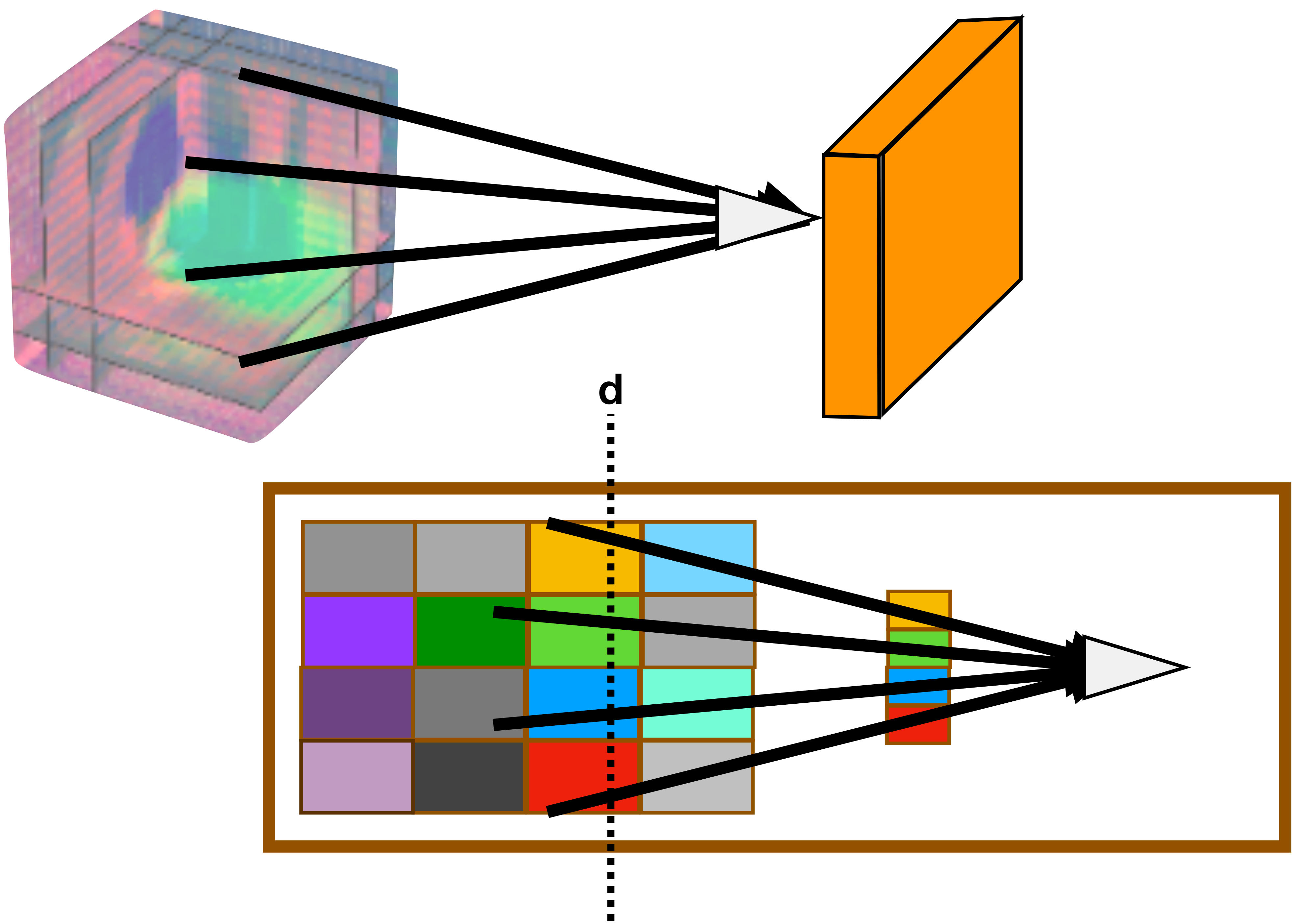
Projection (3D to 2D)



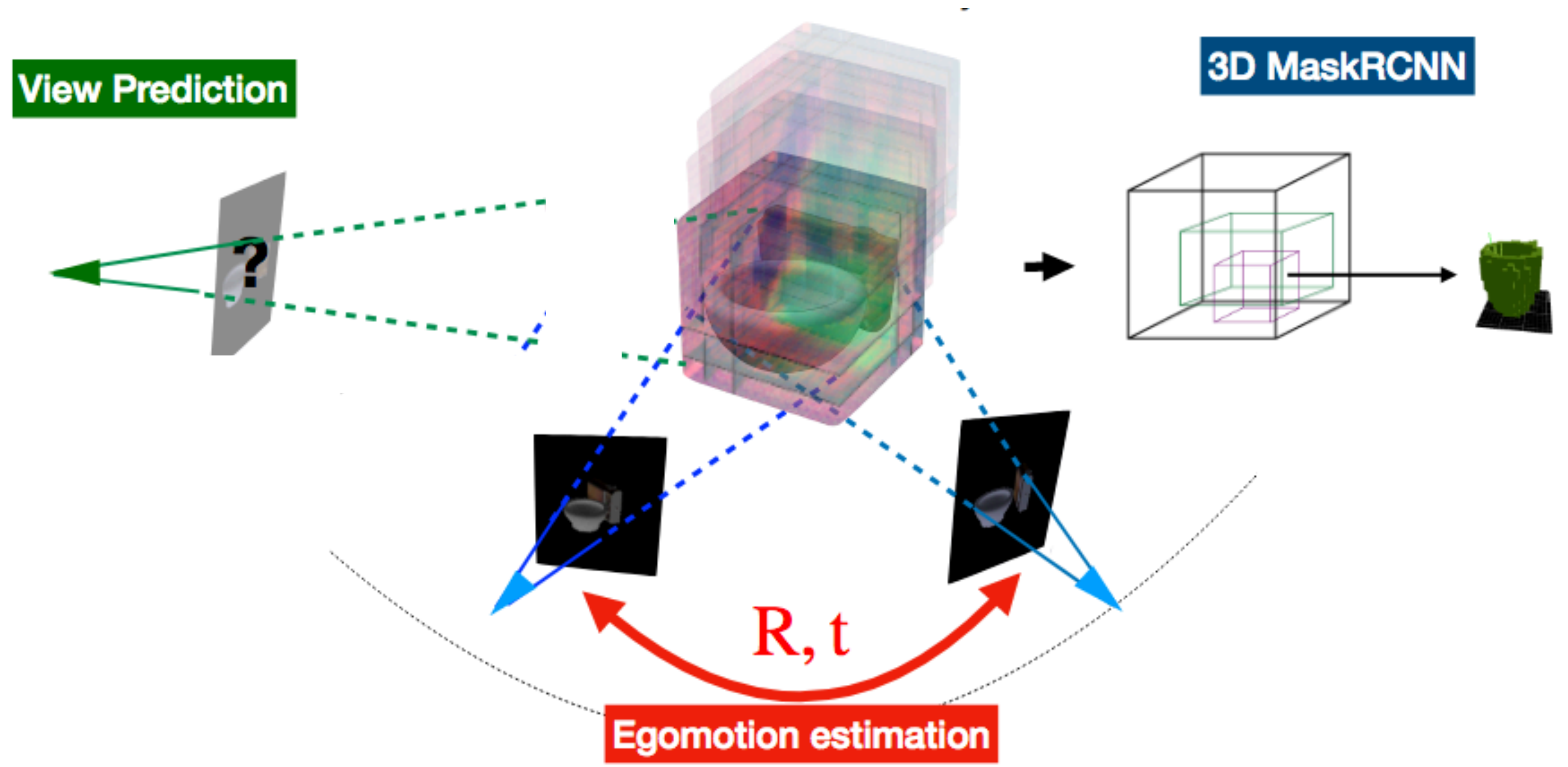
Projection (3D to 2D)



Projection (3D to 2D)

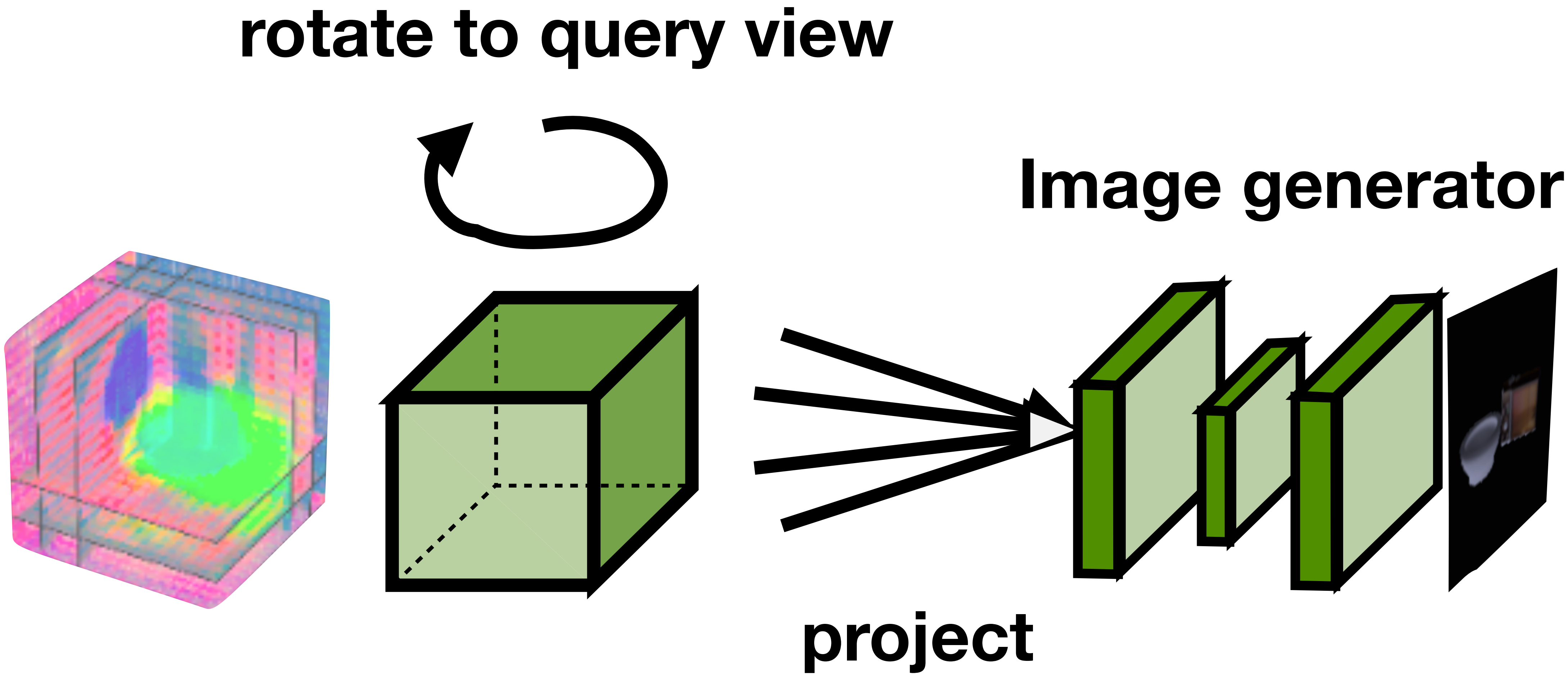


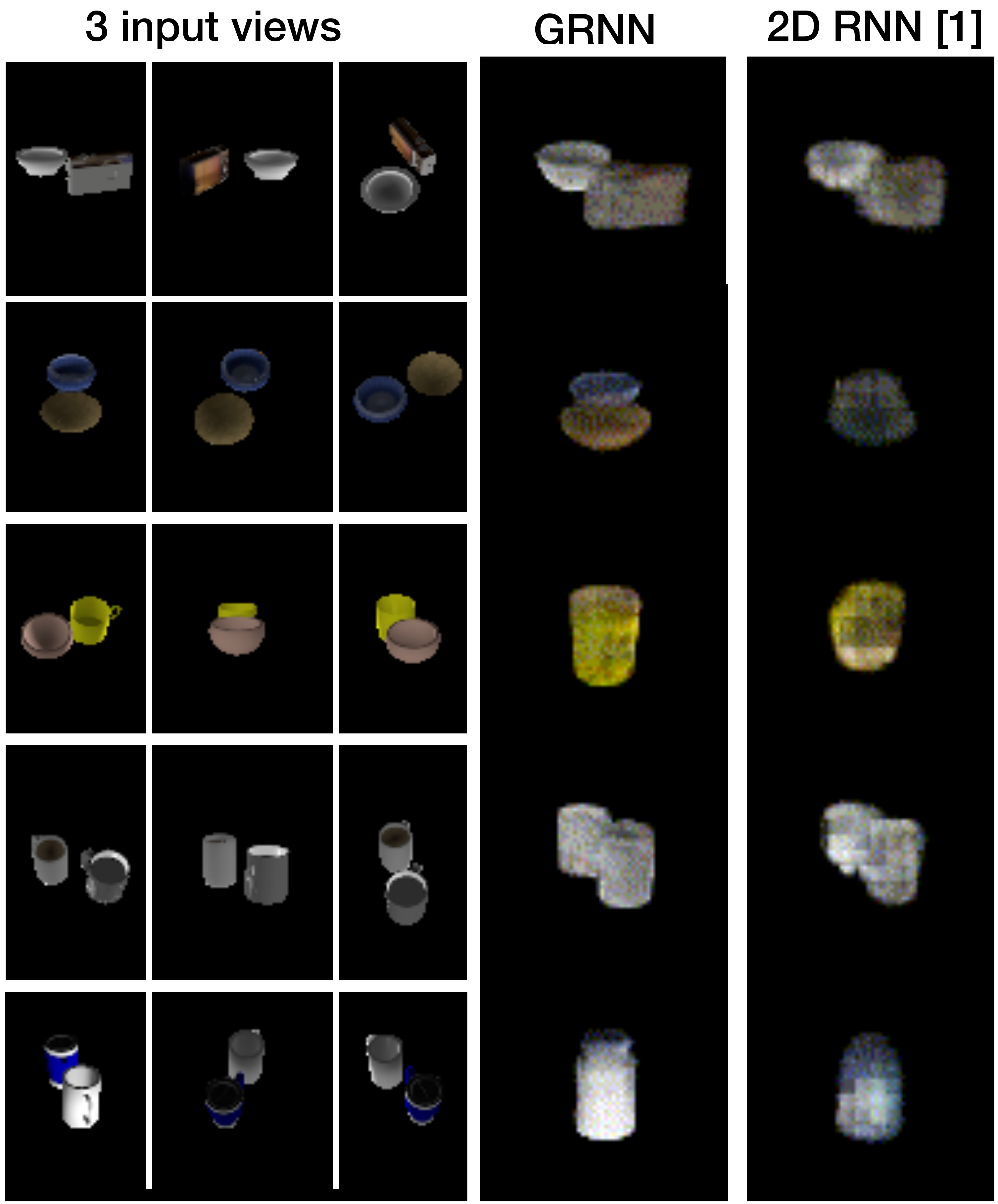
Training GRNNs



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

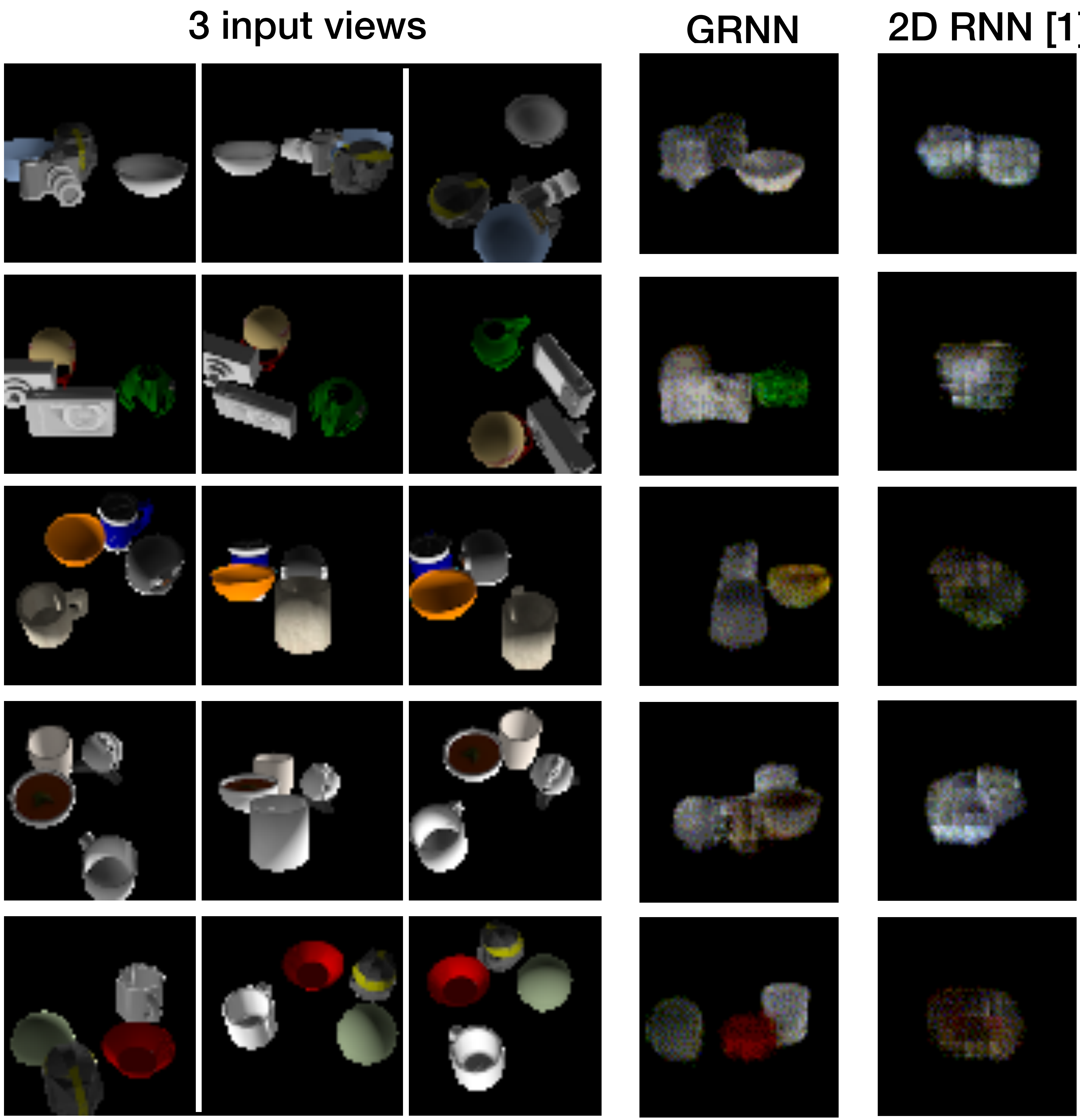
Image generation





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

Testing on scenes with more objets than train time



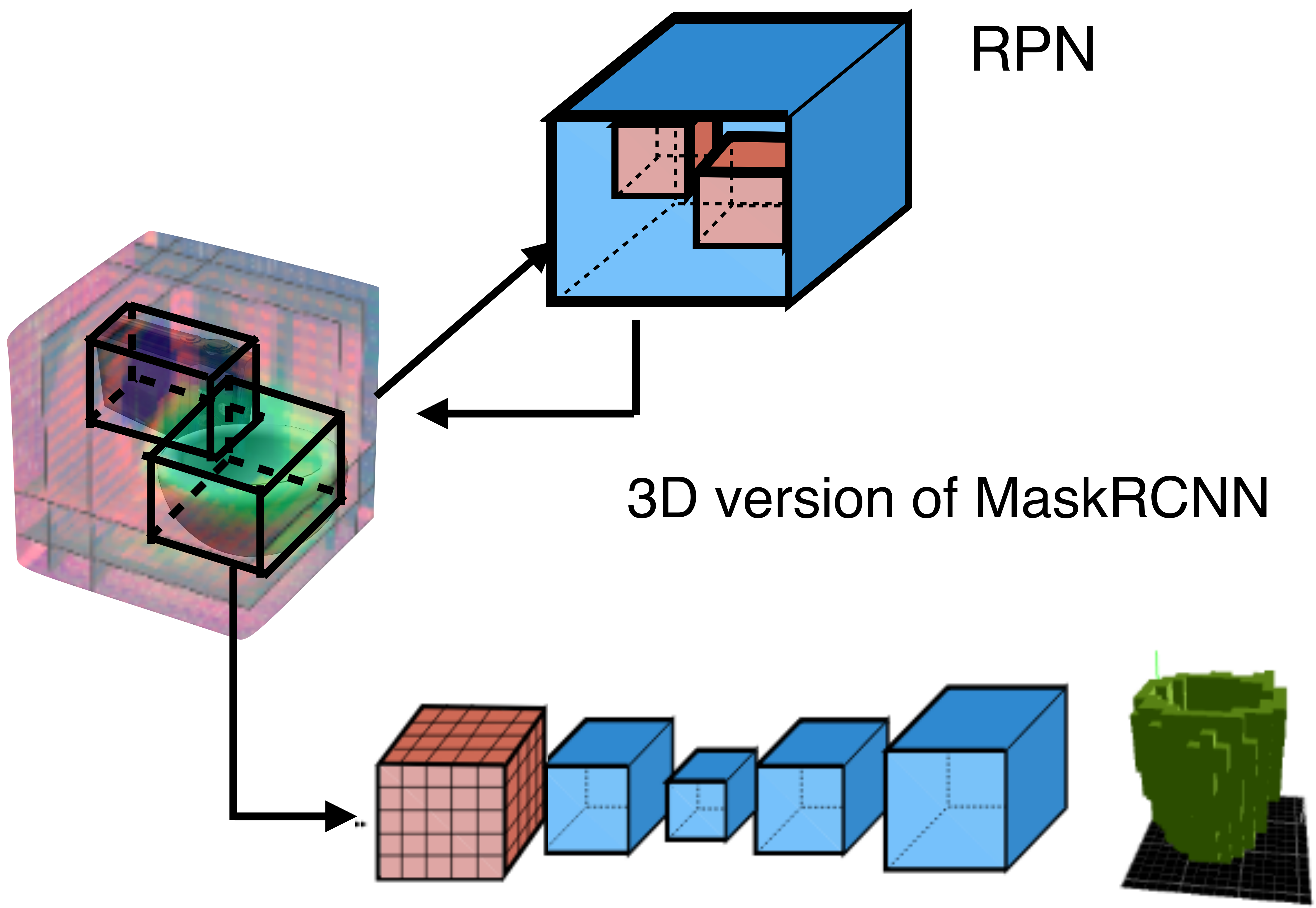
[1] *Neural scene representation and rendering* DeepMind, Science, 2018

View prediction

geometry-aware RNN

2D RNN [1]

3D Object Detection

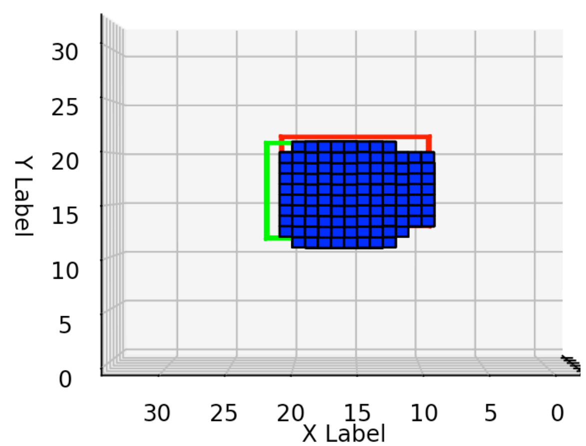


3D object detection

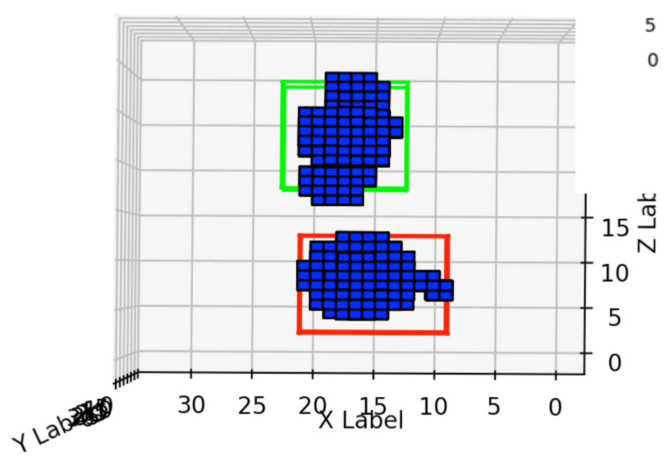
input views

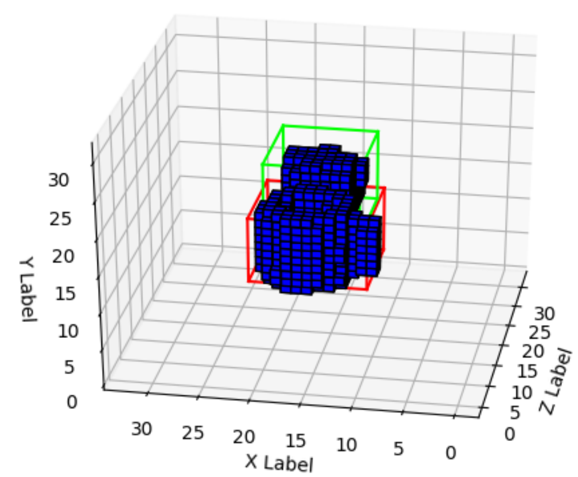
predicted boxes

front-view



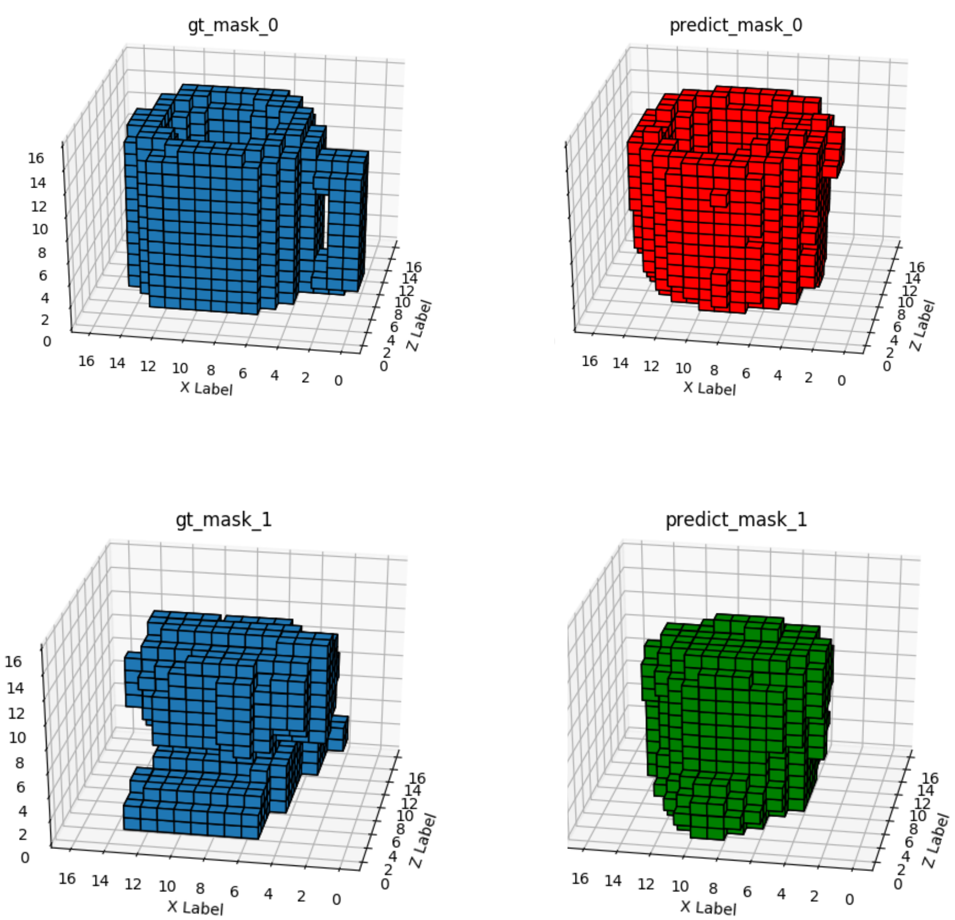
bird-view





predicted segmentations

gt **prediction**



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

A dream

Use the latent hidden map of GRNNs to learn models of Physics of the world, and build agents with persistent models of the world scene, not hostages of 2d projections