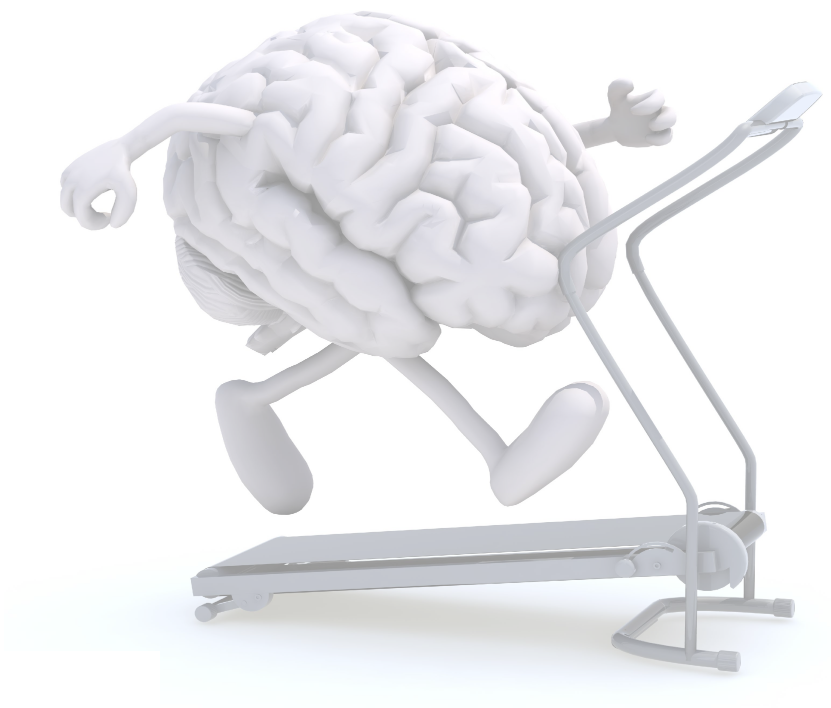


Embodied Language Grounding

Katerina Fragkiadaki

Carnegie Mellon University

Reward learning using natural language

Goal: place *the coca-cola to the right of the bowl*

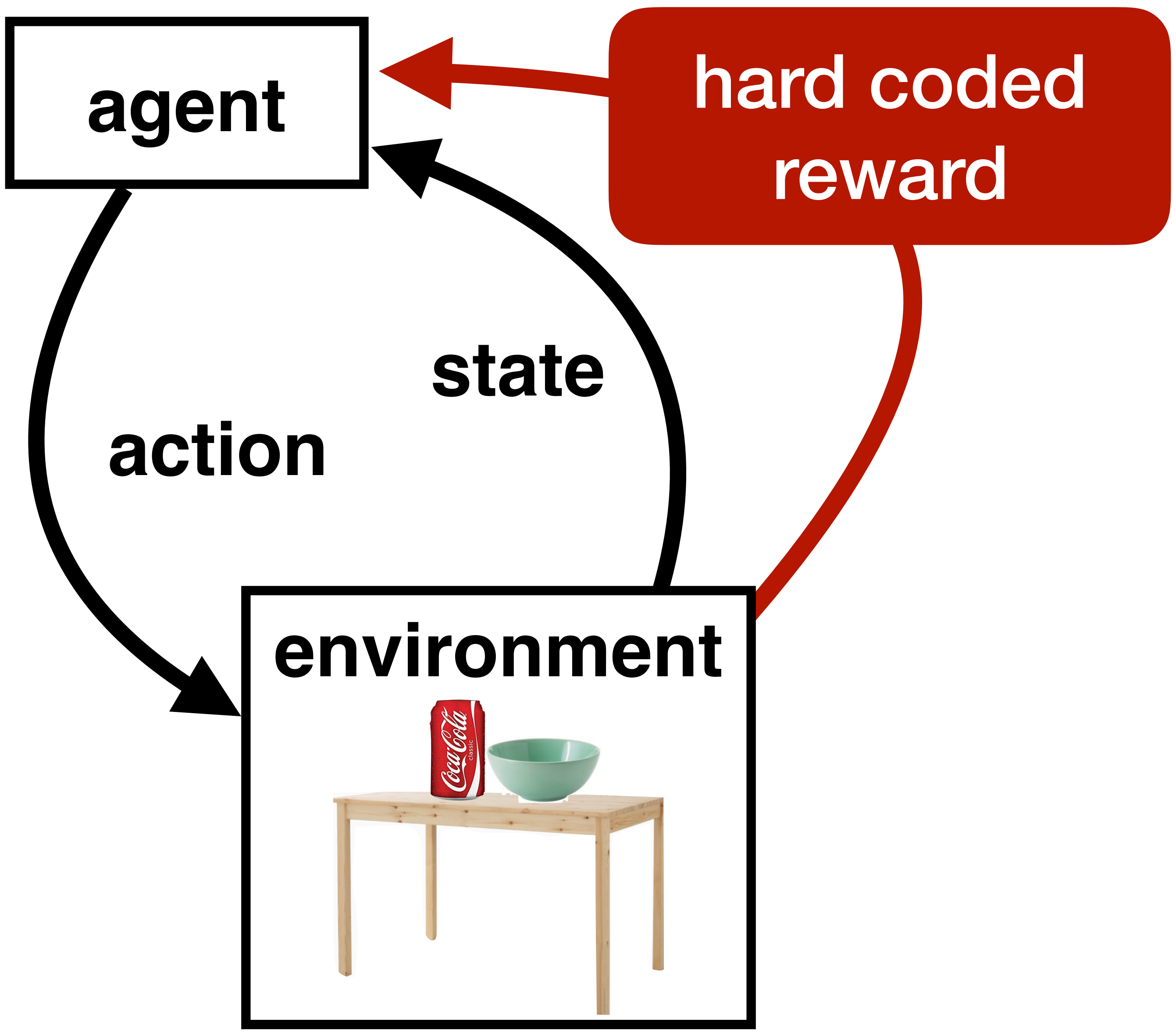
"Can is to the right of the bowl"

Use the learned visual detector to get rewards for policy learning

Tung et al. CVPR 2018

Reward learning using natural language

Goal: place *the coca-cola to the right of the bowl*

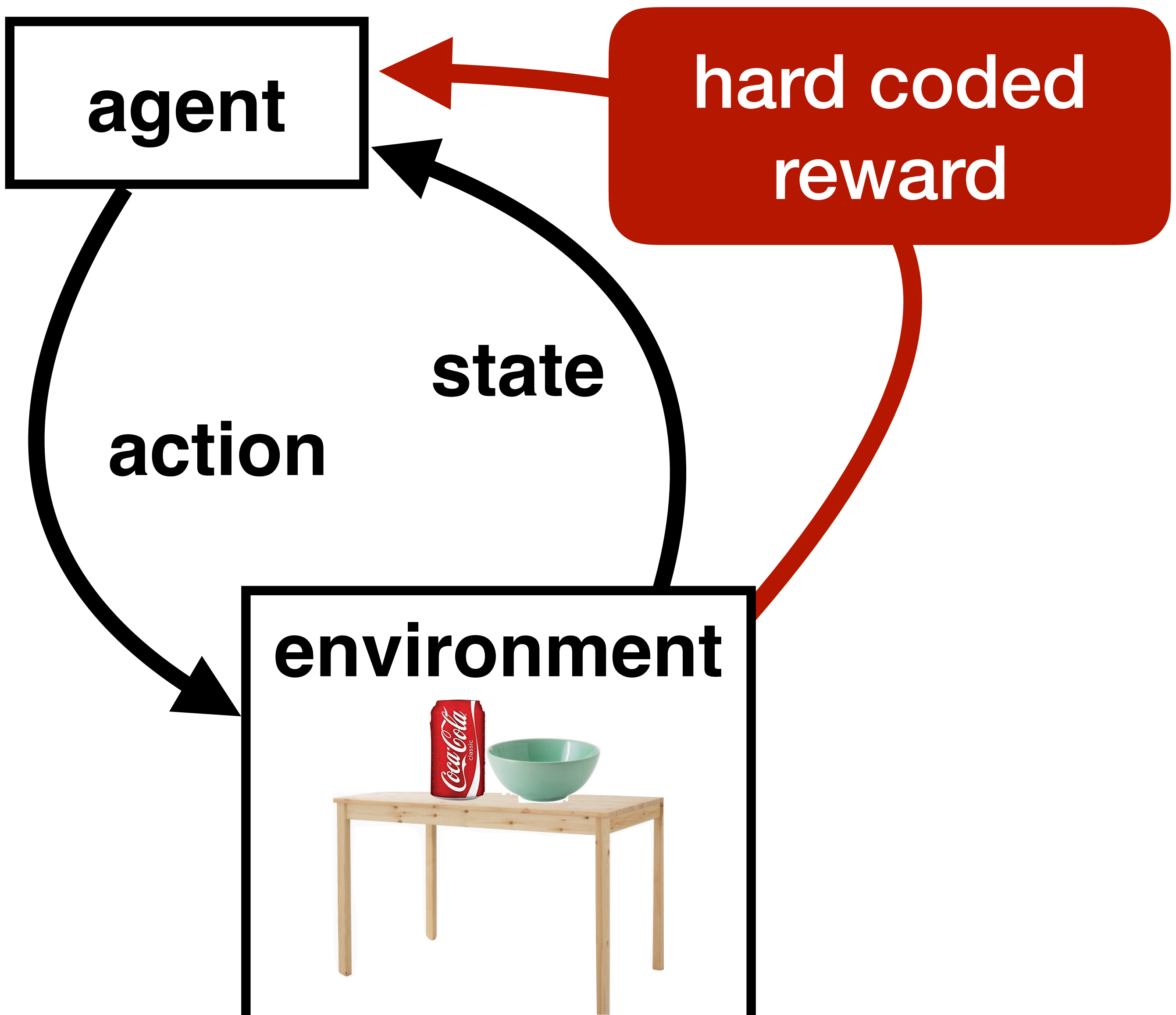


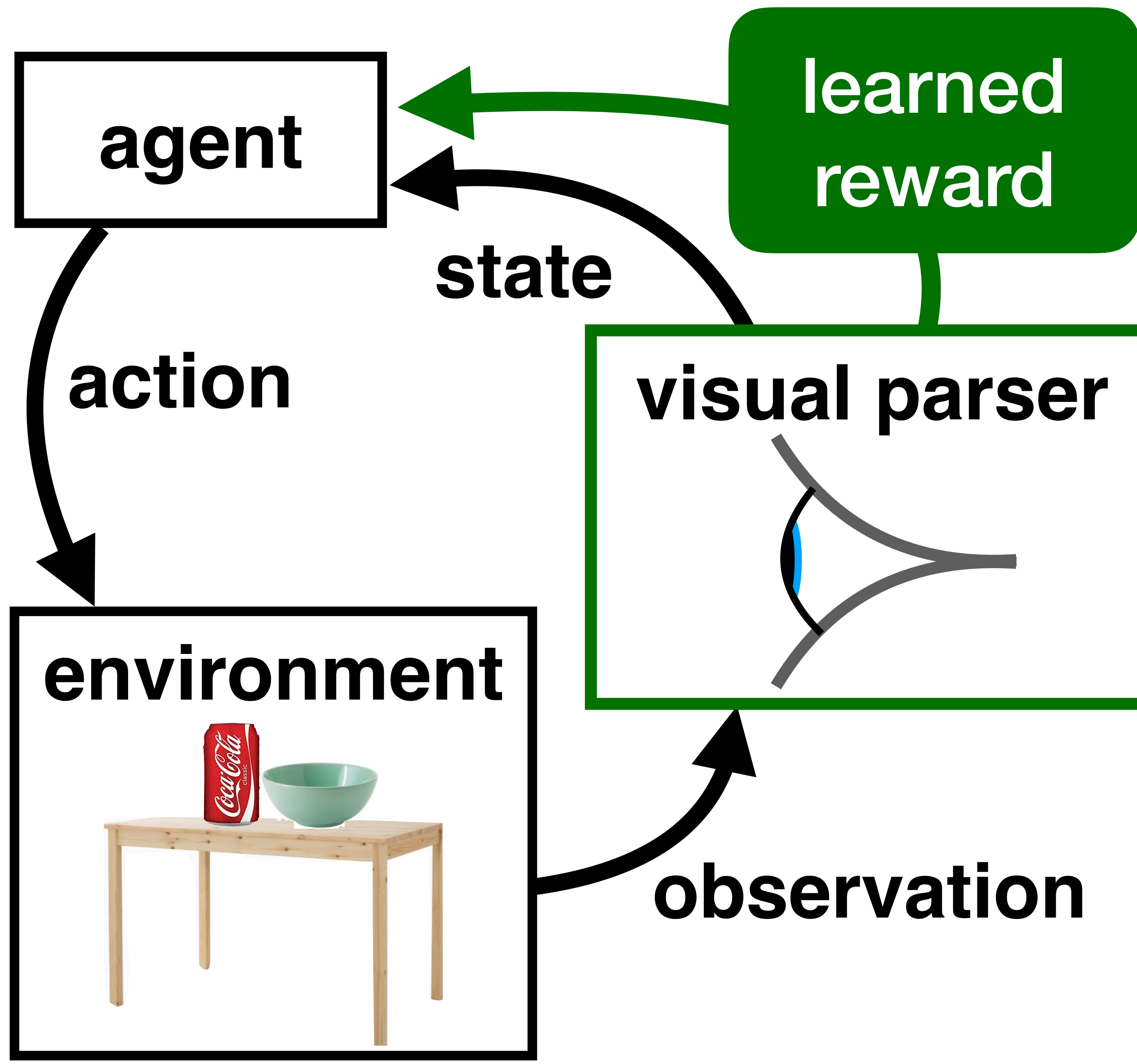
Manually code the reward in a simulated or instrumented environment

Tung et al. CVPR 2018

Reward learning using natural language

Goal: place *the coca-cola to the right of the bowl*





Manually code the reward in a simulated or instrumented environment

Learn to detect from an RGB image when the goal is achieved

Tung et al. CVPR 2018

Reward learning using natural language

"Can is to the right of the mug"

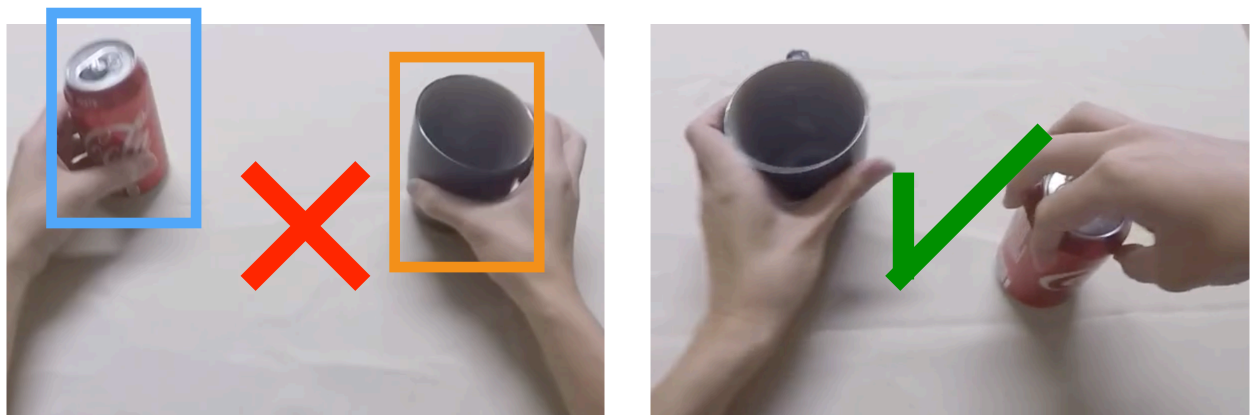
Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

Reward learning using natural language

"Can is to the right of the mug"

Reward learning using natural language

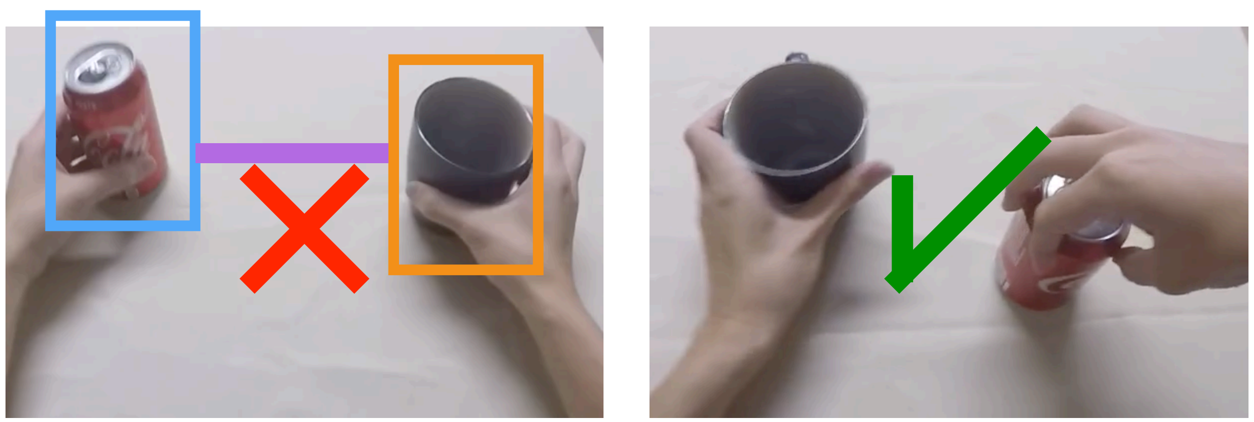
"Can is to the right of the mug"



Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

Reward learning using natural language

"*Can is to the right of the mug*"

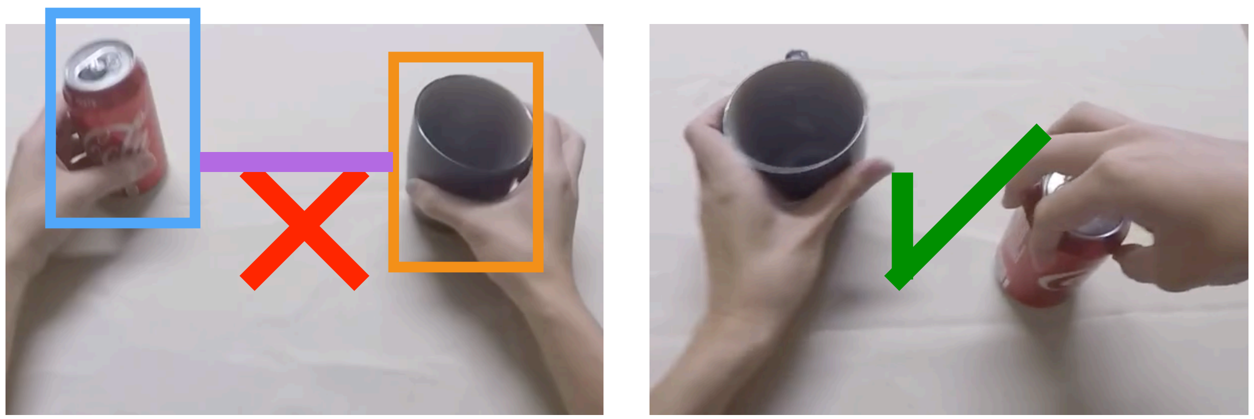


reward detector

Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

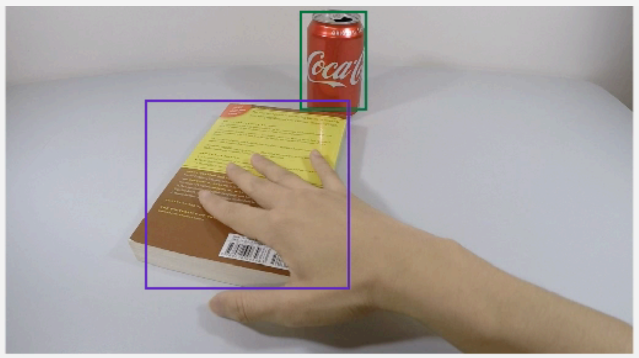
Reward learning using natural language

“Can is to the right of the mug”

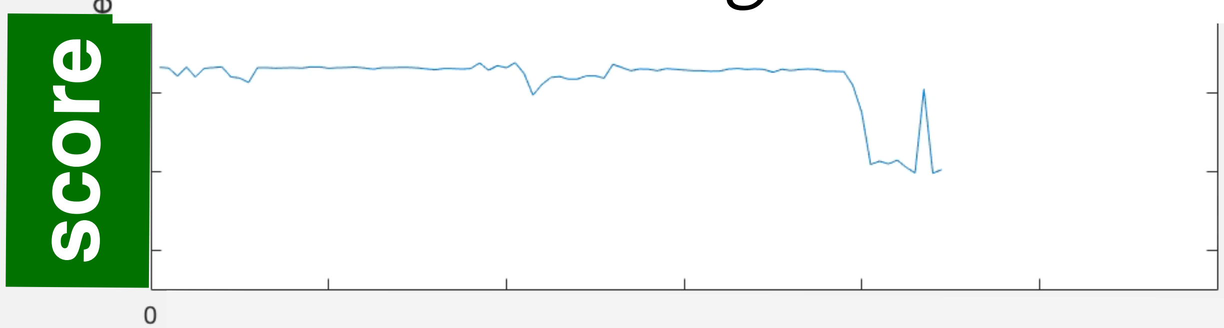


reward detector

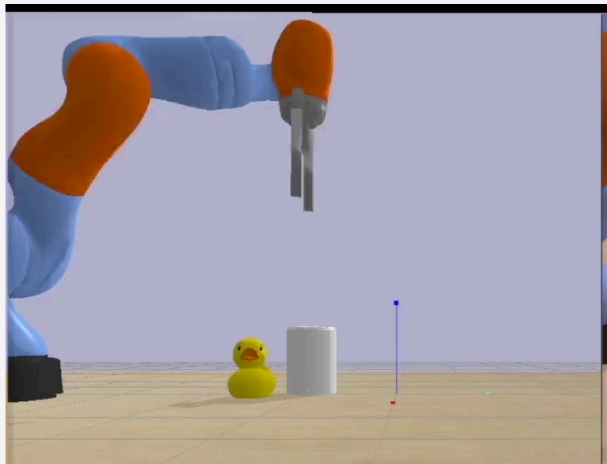
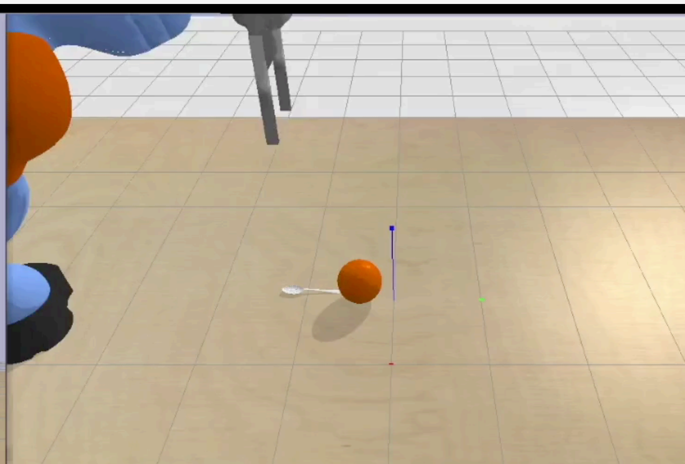
Learned reward detector



``can is to the right of the book”



Learned policy

Reward learning using natural language

"Can is to the right of the mug"

Our conclusions:

- the reward detector could not effectively generalize across camera placements
- could not provide shaped rewards
- could not discerne impossible goals for possible ones, e.g., "*the mug inside the coca cola*" versus "*the coca cola inside the mug*"

Tung et al. CVPR 2018

People can infer affordability of utterances.

- *“After wading barefoot in the lake, Erik used his shirt to dry his feet.”*
- *“After wading barefoot in the lake, Erik used his glasses to dry his feet.”*

Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning, Glenberg and Robertson, 2000

People can infer affordability of utterances.

- *“He used the newspaper to protect his face from the wind.”*
- *“He used the matchbox to protect his face from the wind.”*

Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning, Glenberg and Robertson, 2000

People can answer million questions regarding the described situation.

"He used the newspaper to protect his face from the wind."

- *How many free hands the man has?*
- *Is the newspaper in front or behind his eyes?*
- *Can the newspaper be a single page?*
- *Is he holding the newspaper?*
- *Is he lying on top of the newspaper?*
- *Is the newspaper protecting also his neck from the wind? His feet?*

Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning, Glenberg and Robertson, 2000

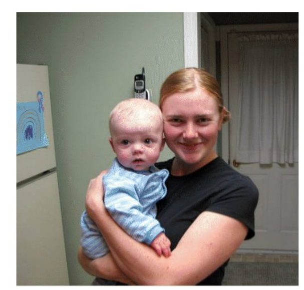
Computational models of language and vision

…cannot answer *basic* questions

Where is the child sitting?

fridge

arms



Where are the arms sitting? Can the fridge door close? Can a baby hold two bottles? Can a baby hold three bottles? Does a baby disappear when mom walks in front? Is mom or baby taller?

Computational models of language and vision

…cannot infer affordbility of language uttrances

- *“The bowl inside the cube”*
- *“The cube inside the bowl”*

Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning, Glenberg and Robertson, 2000

People can infer affordability of utterances.

- Words and phrases index to objects in the world or to prototypical symbols of those objects
- We derive affordance from those objects
- The derived affordances constrain the way ideas can be coherently combined

Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning, Glenberg and Robertson, 2000

Simulation Semantics

We understand utterances by simulating their content, using similar constucts to perception and control

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters , Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg

Language grounding to visual cues

2D boxes or 2D CNN activations do not have any affordability attached

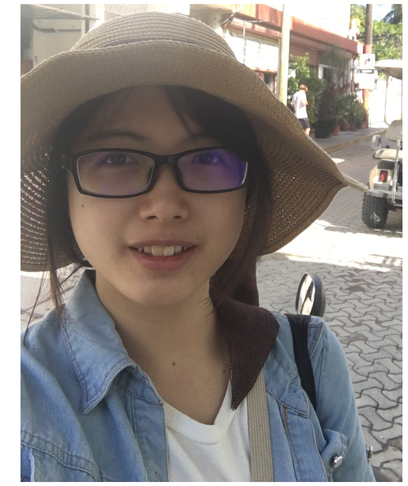
They are themselves **ungrounded** :-(

Affordandable visual representations

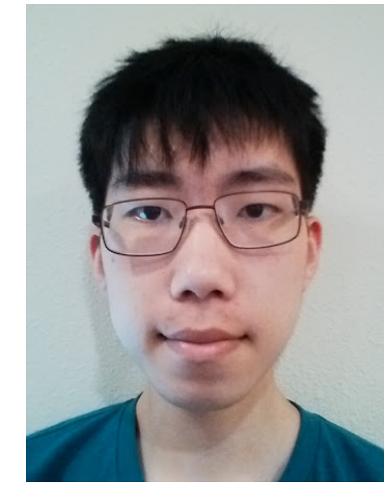
We seek visual feature representations to ground NL onto that obey basic spatial common sense constraints:

- Objects have 3D extent
- Objects do not interpenetrate in 3D
- Objects come in regular sizes
- Objects persist over time

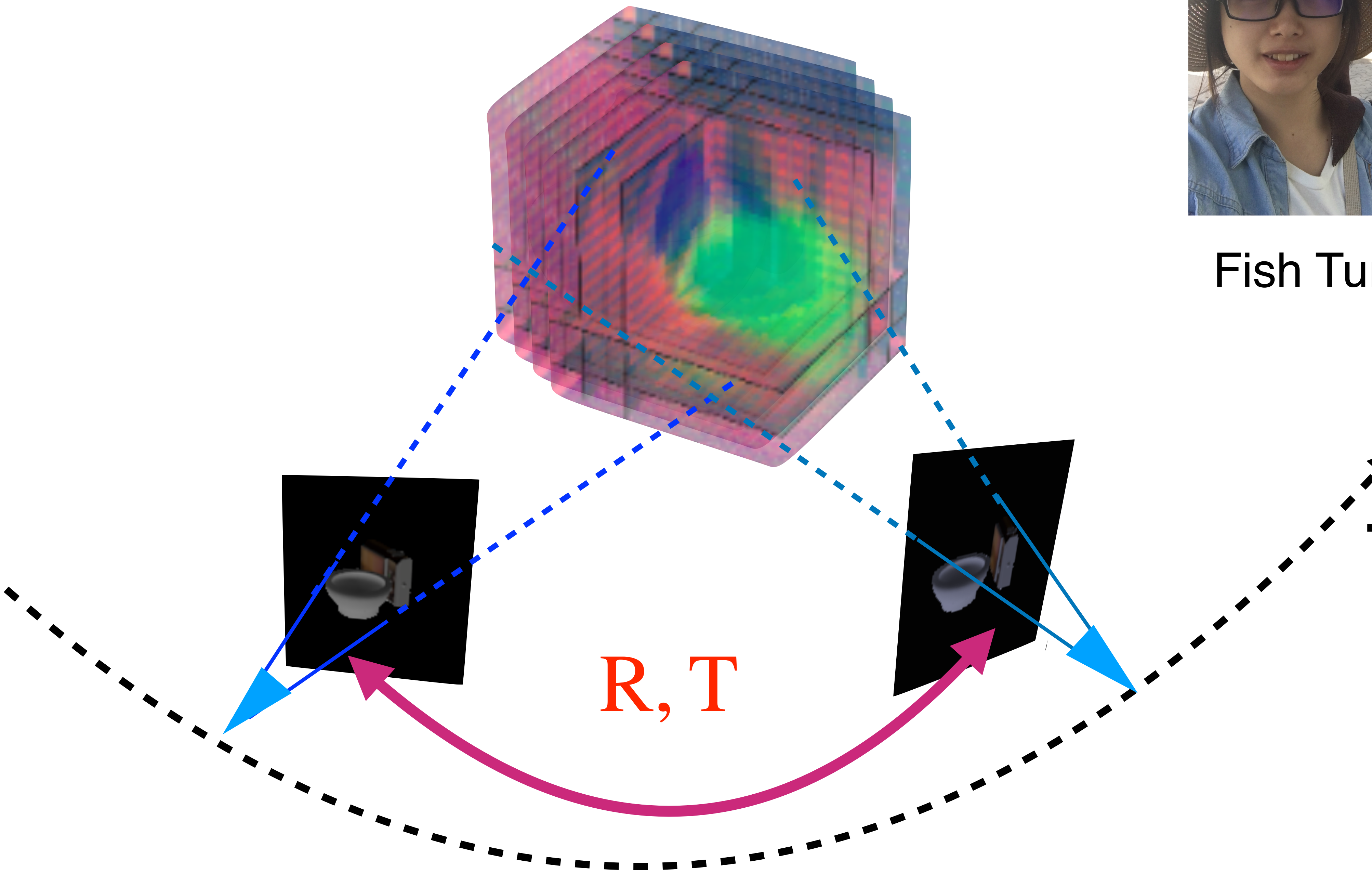
Geometry-Aware Recurrent Networks



Fish Tung



Ricson Chen

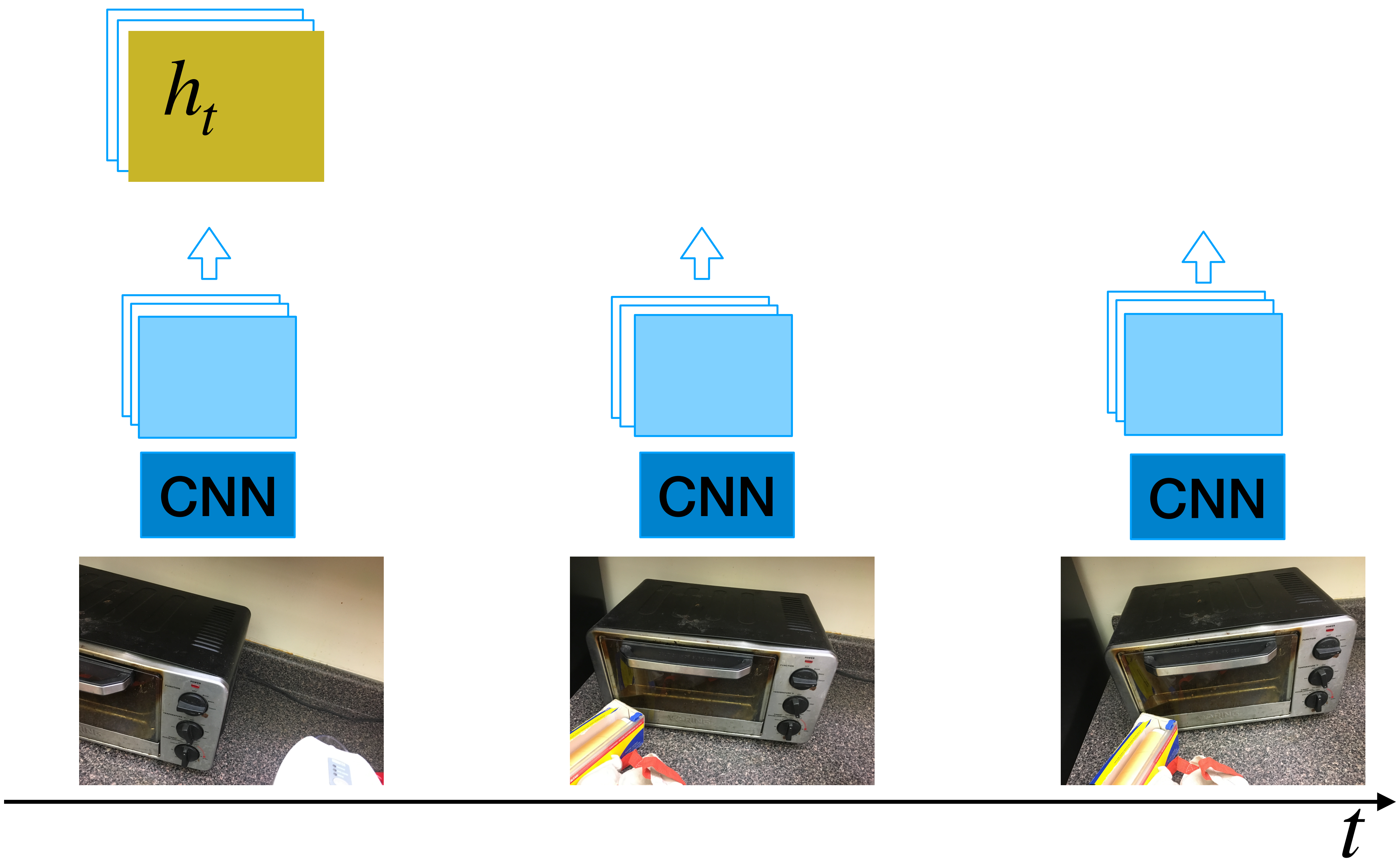


1. Hidden state: 3D feature maps
2. Egomotion-stabilized hidden state updates

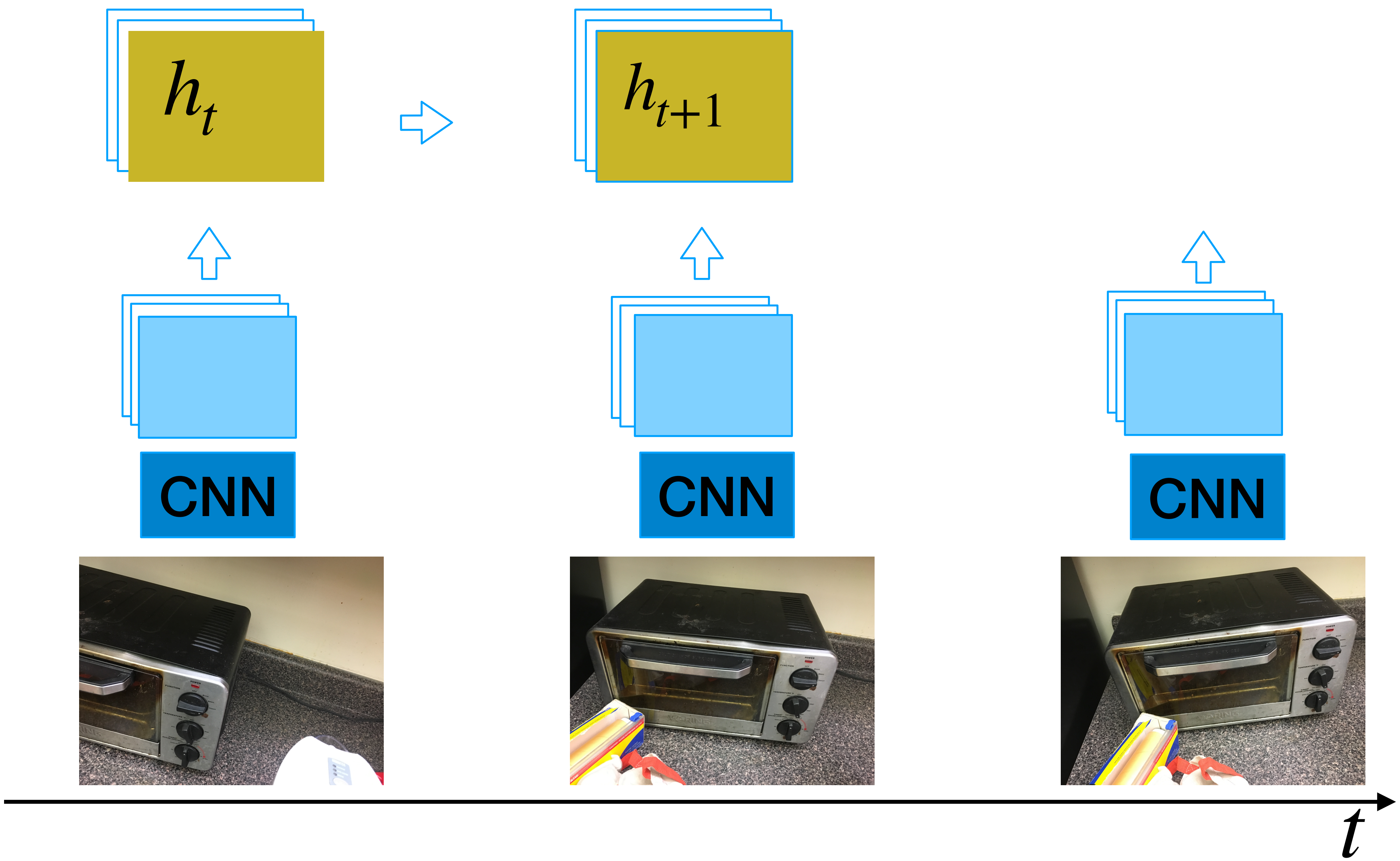
2D RNNs (conv-LSTMs/GRUs)

2D RNNs (conv-LSTMs/GRUs)

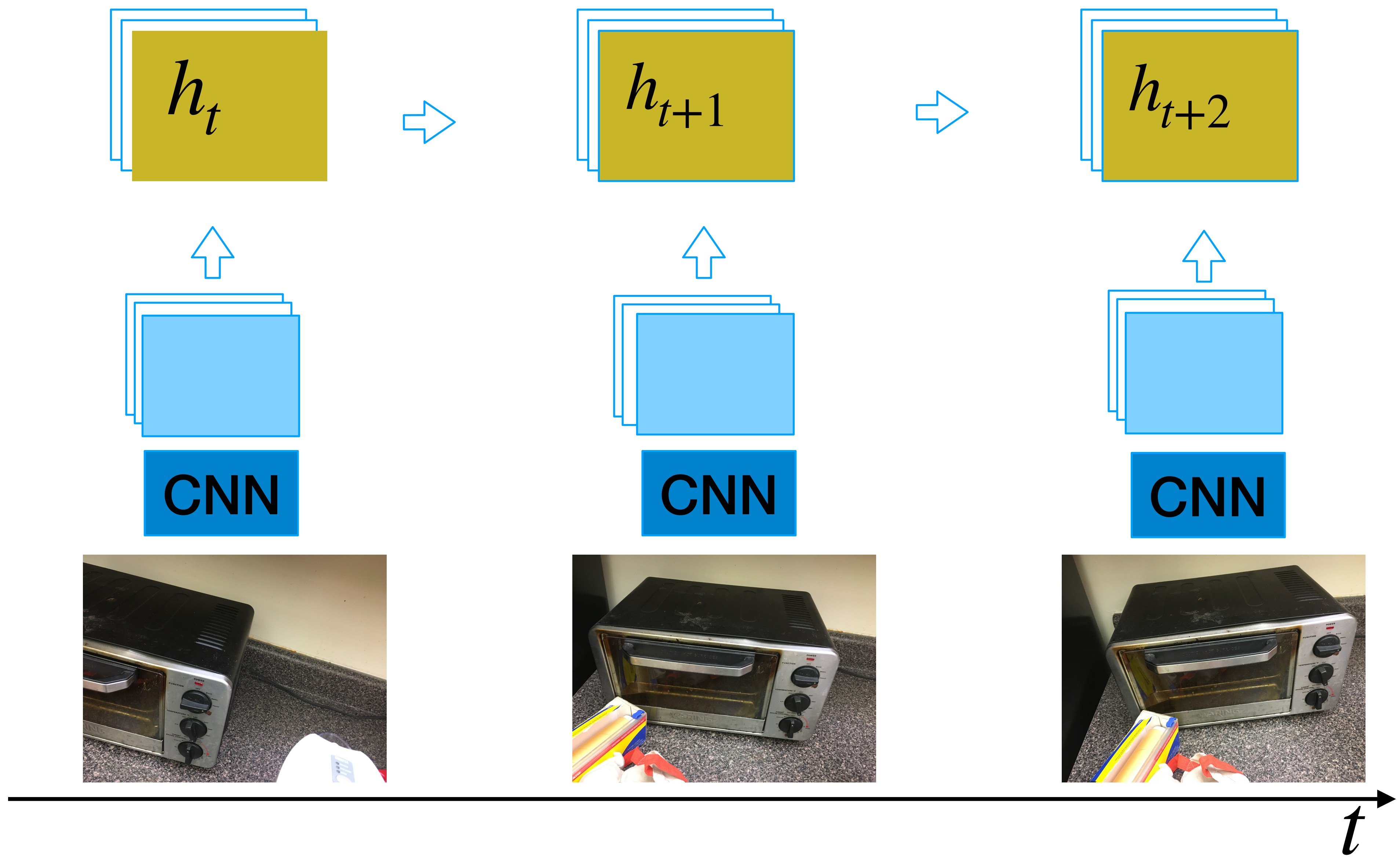
2D RNNs (conv-LSTMs/GRUs)



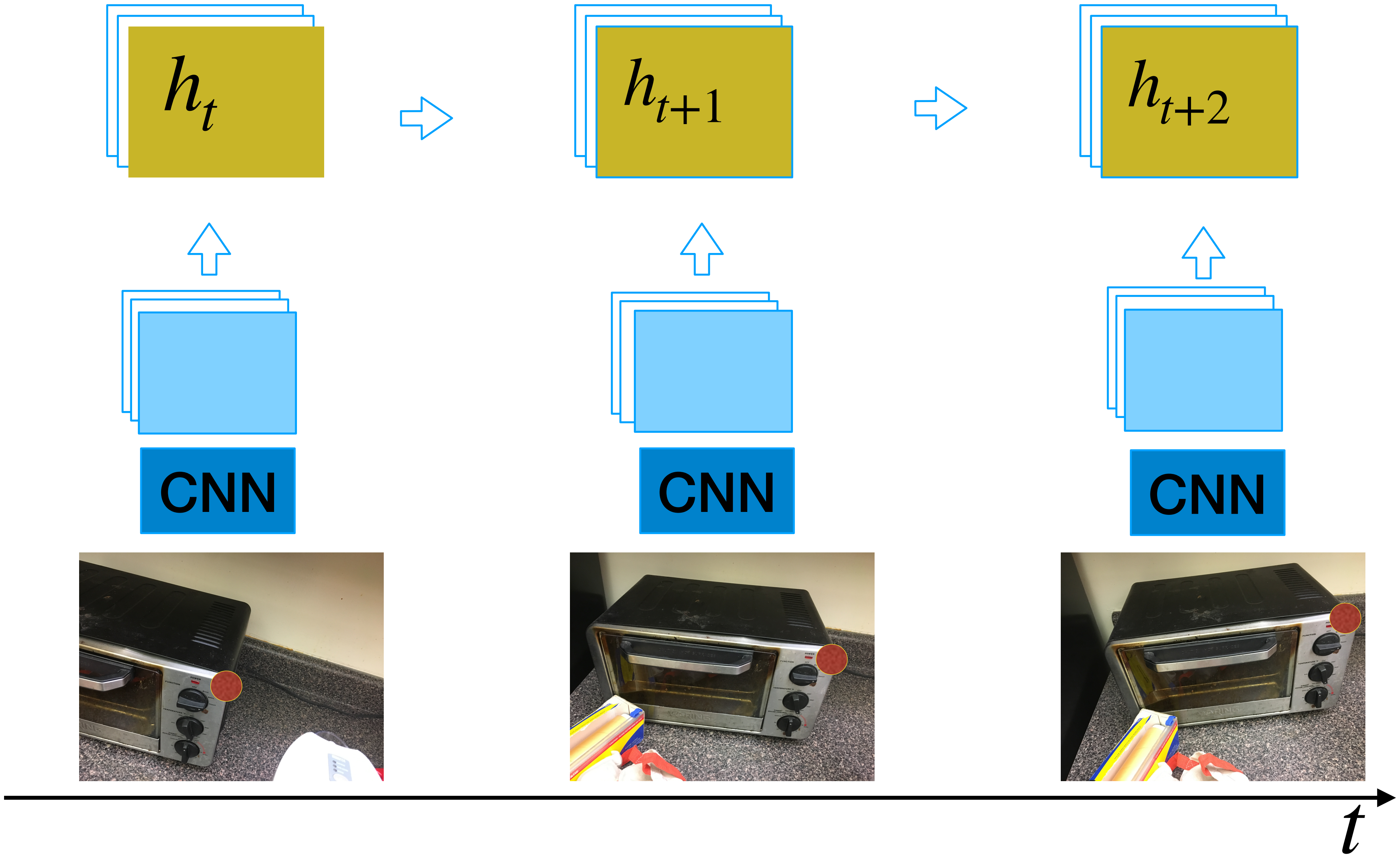
2D RNNs (conv-LSTMs/GRUs)



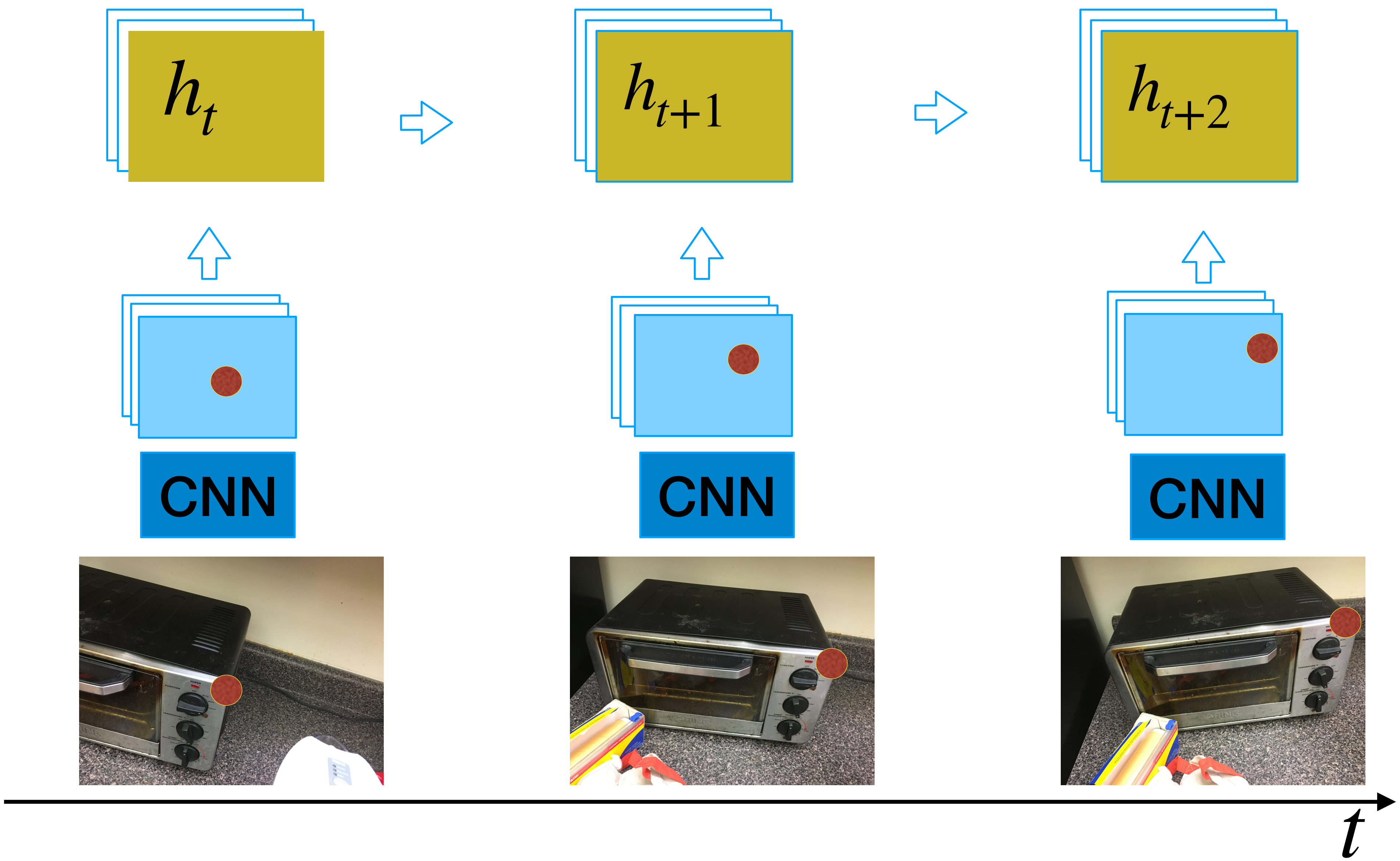
2D RNNs (conv-LSTMs/GRUs)



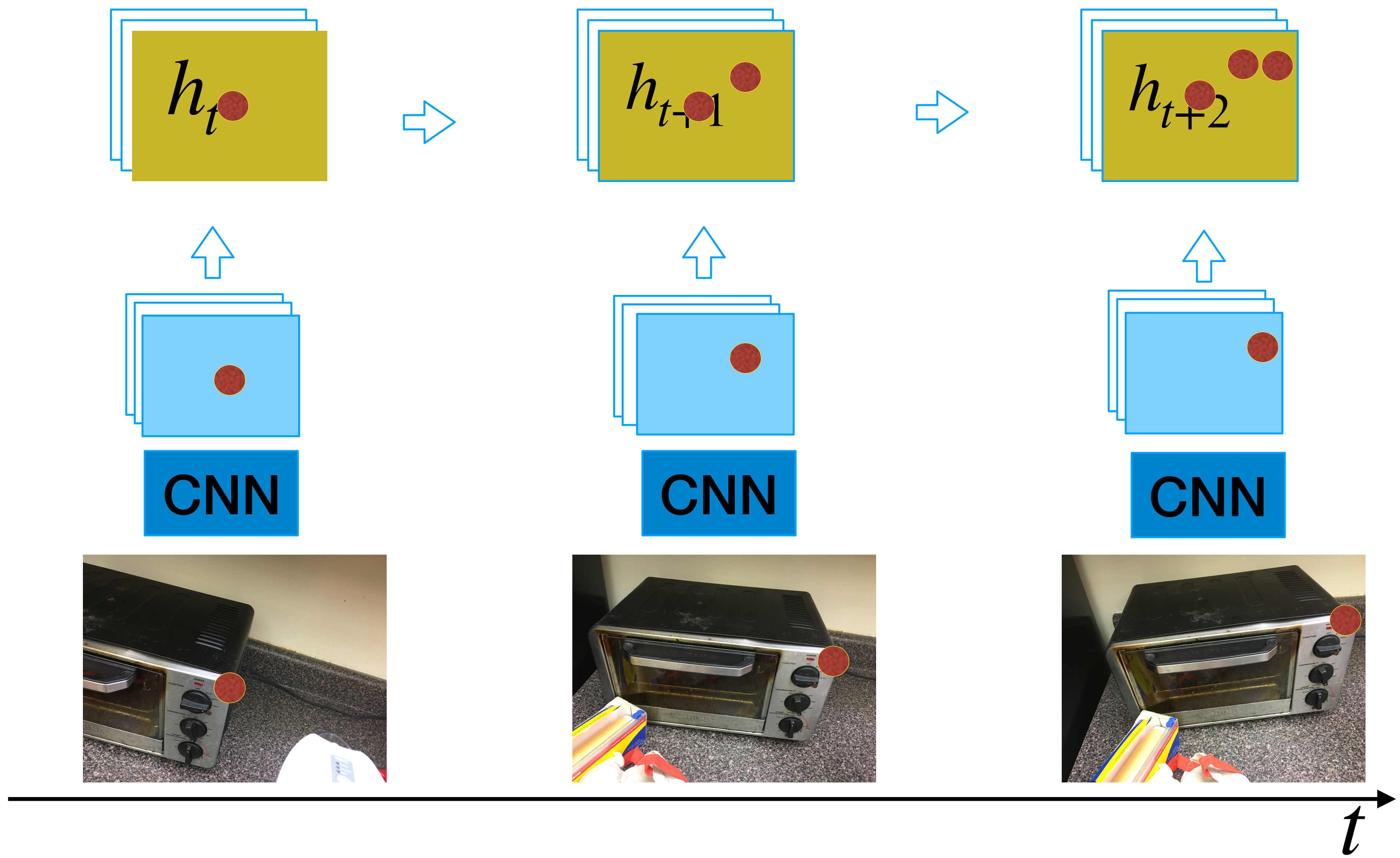
2D RNNs (conv-LSTMs/GRUs)



2D RNNs (conv-LSTMs/GRUs)

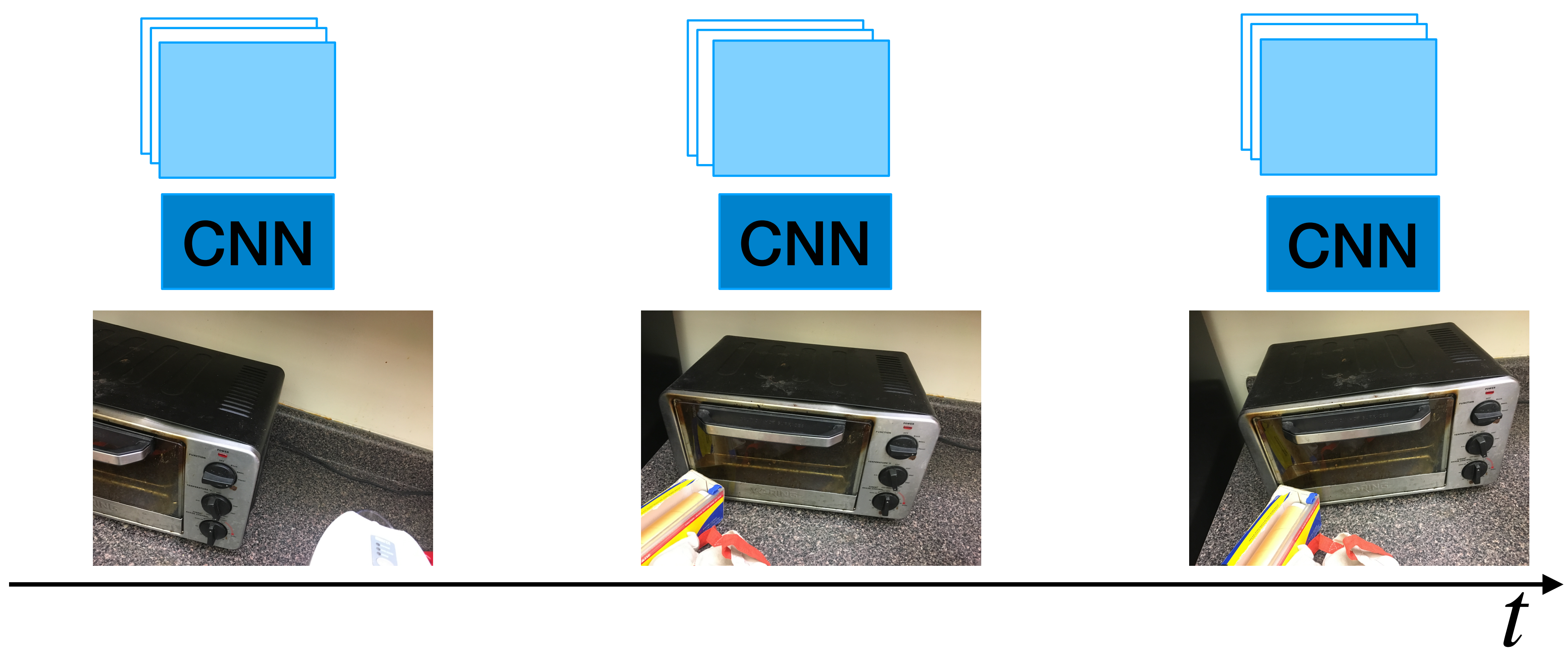


2D RNNs (conv-LSTMs/GRUs)

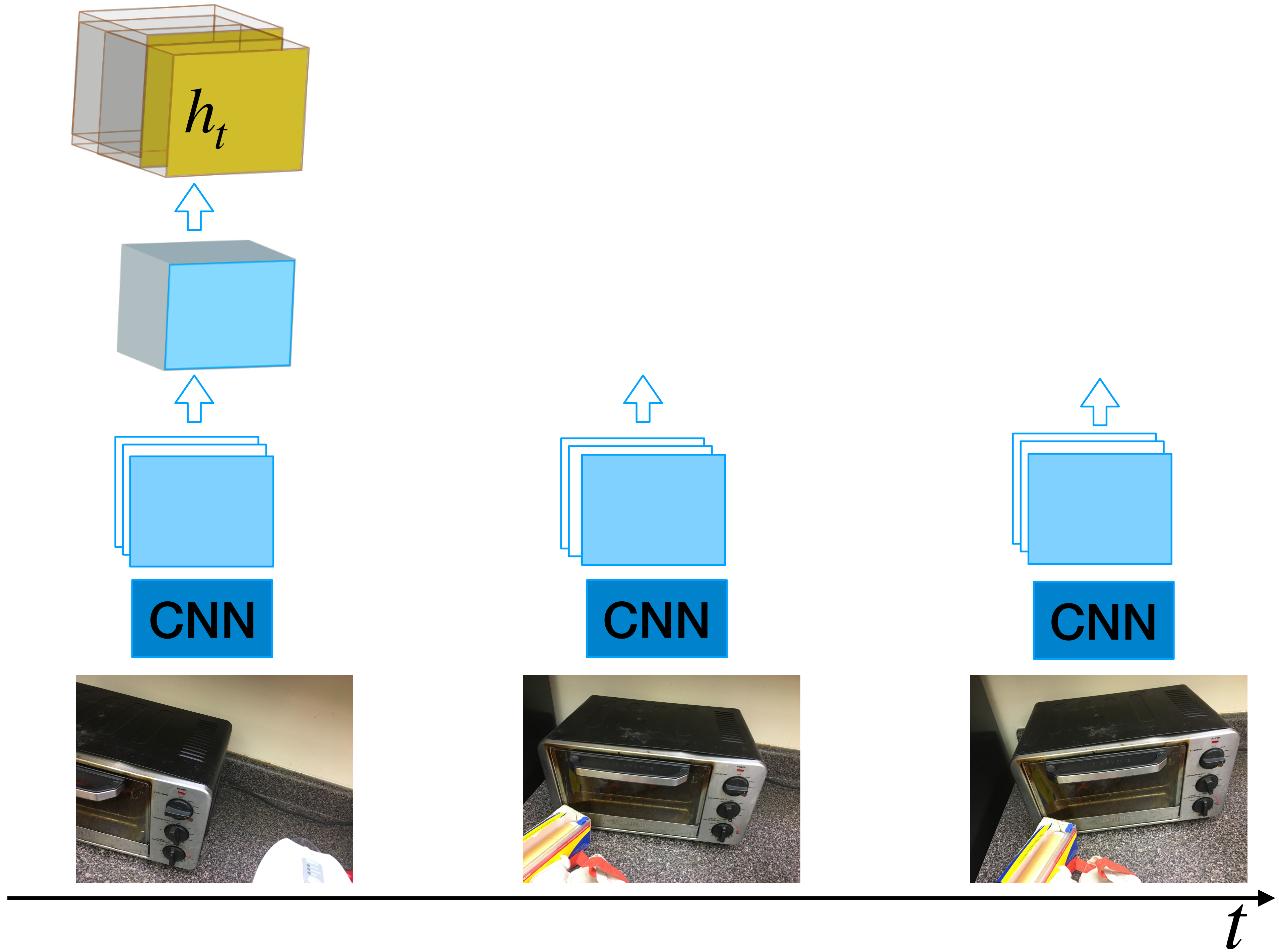


Geometry-Aware Recurrent Networks

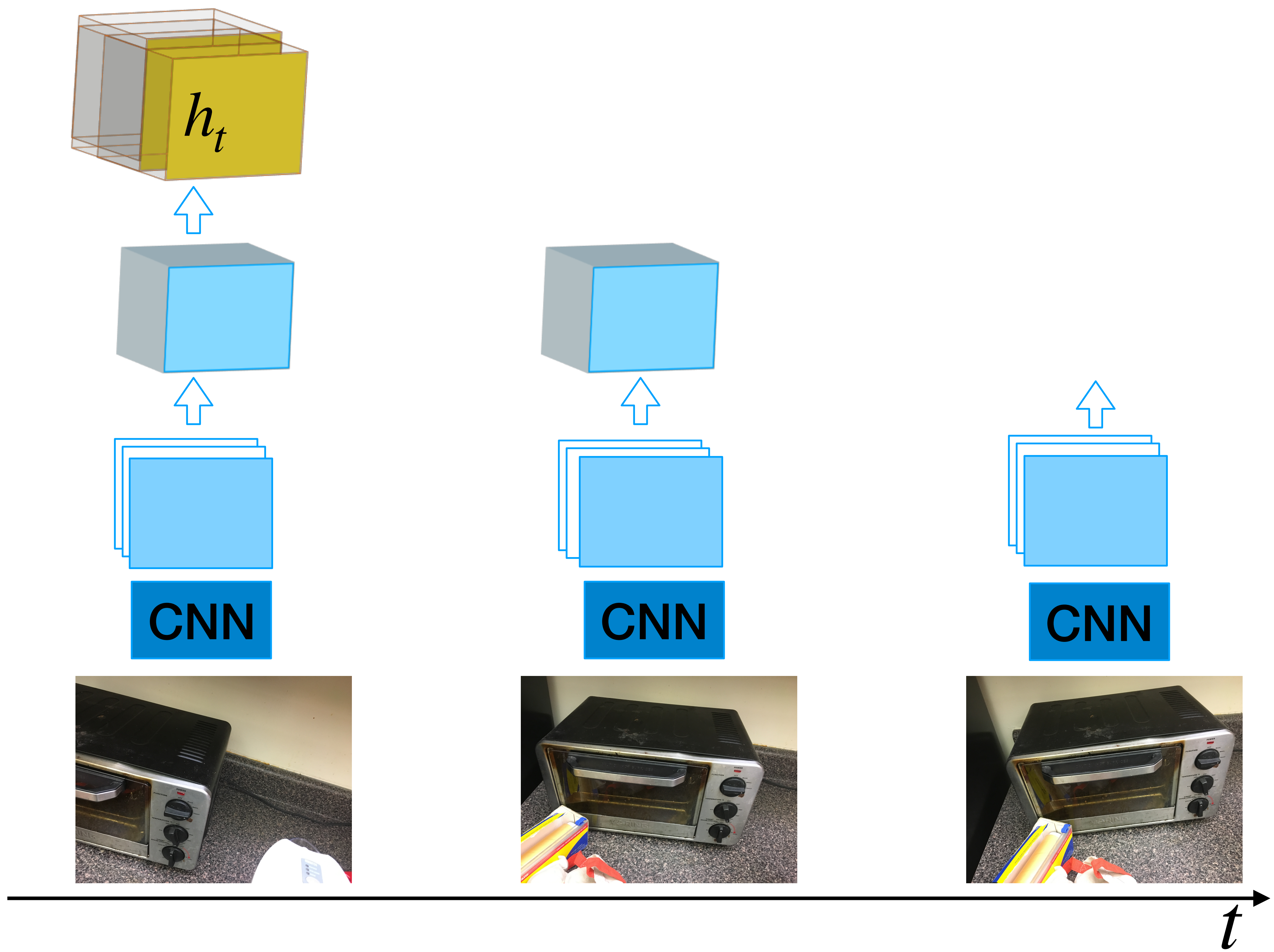
Geometry-Aware Recurrent Networks



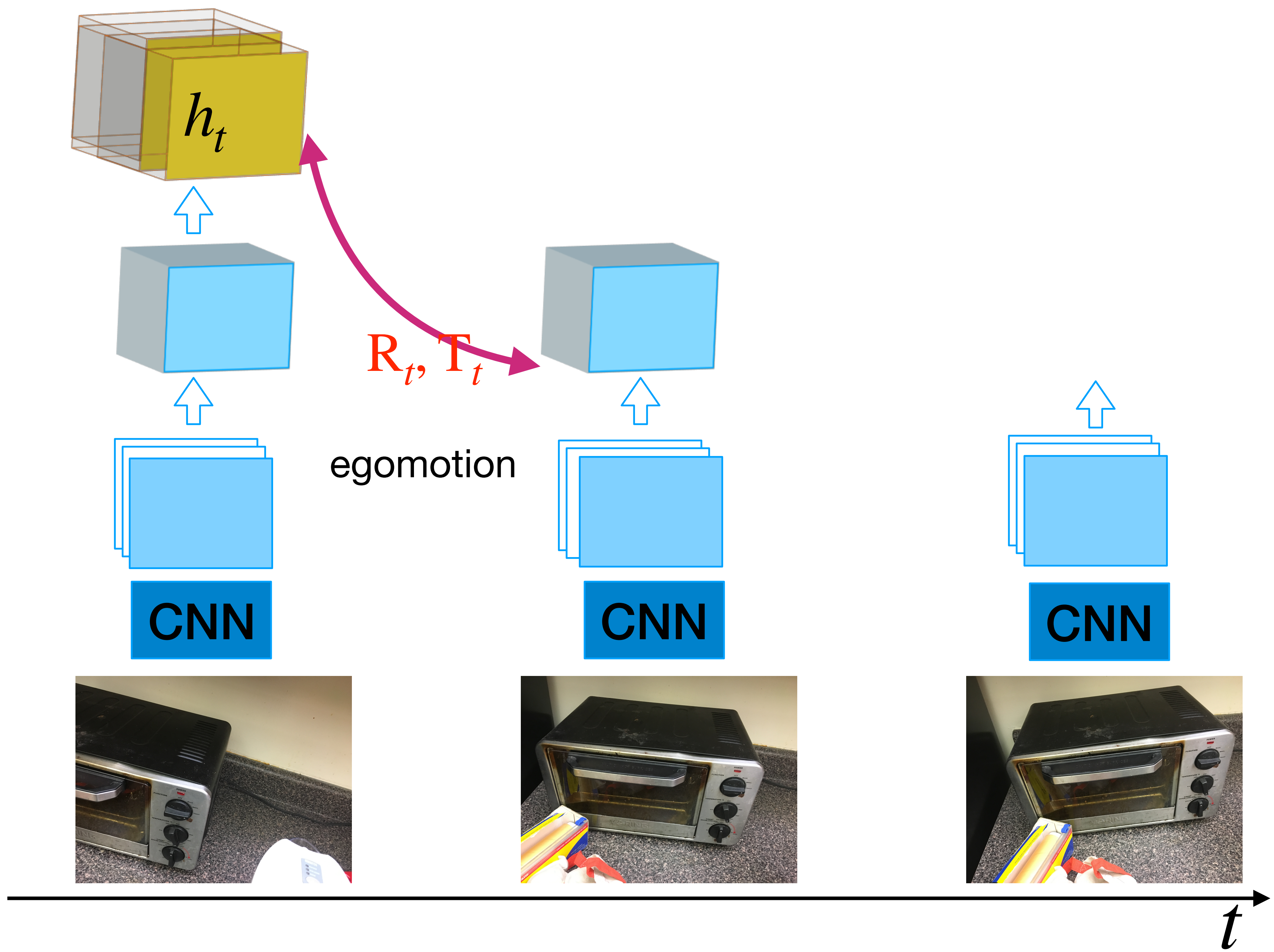
Geometry-Aware Recurrent Networks



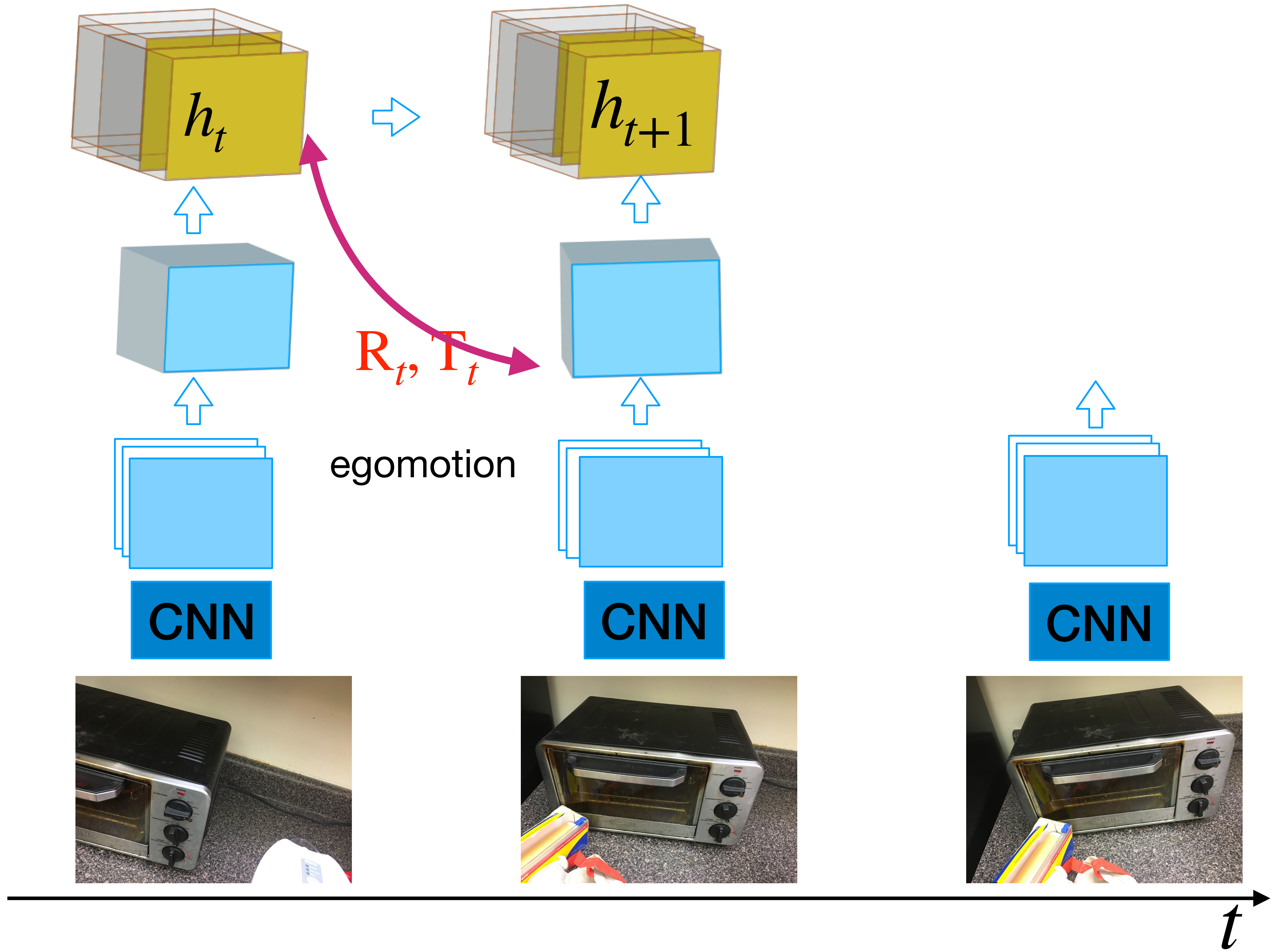
Geometry-Aware Recurrent Networks



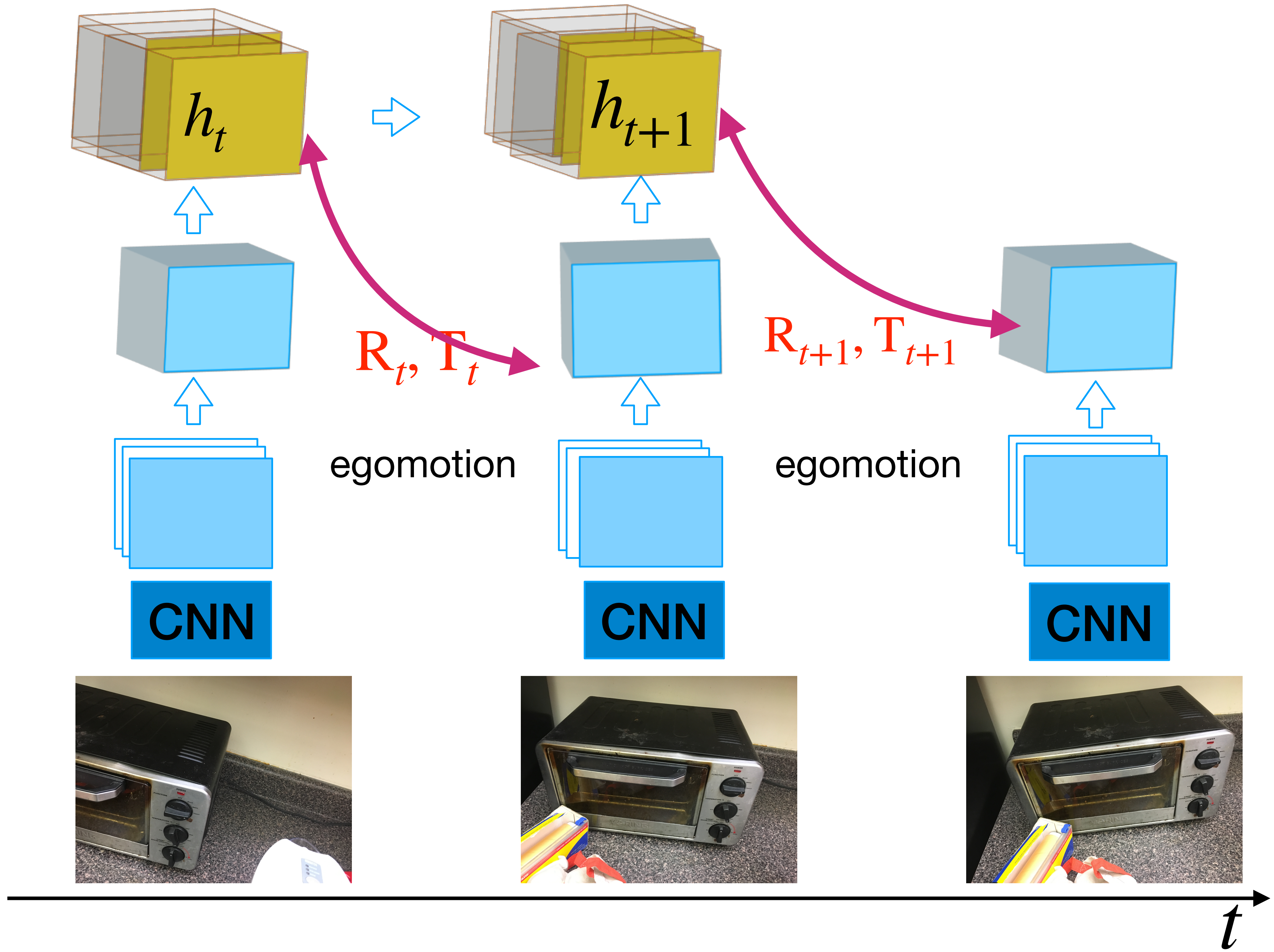
Geometry-Aware Recurrent Networks



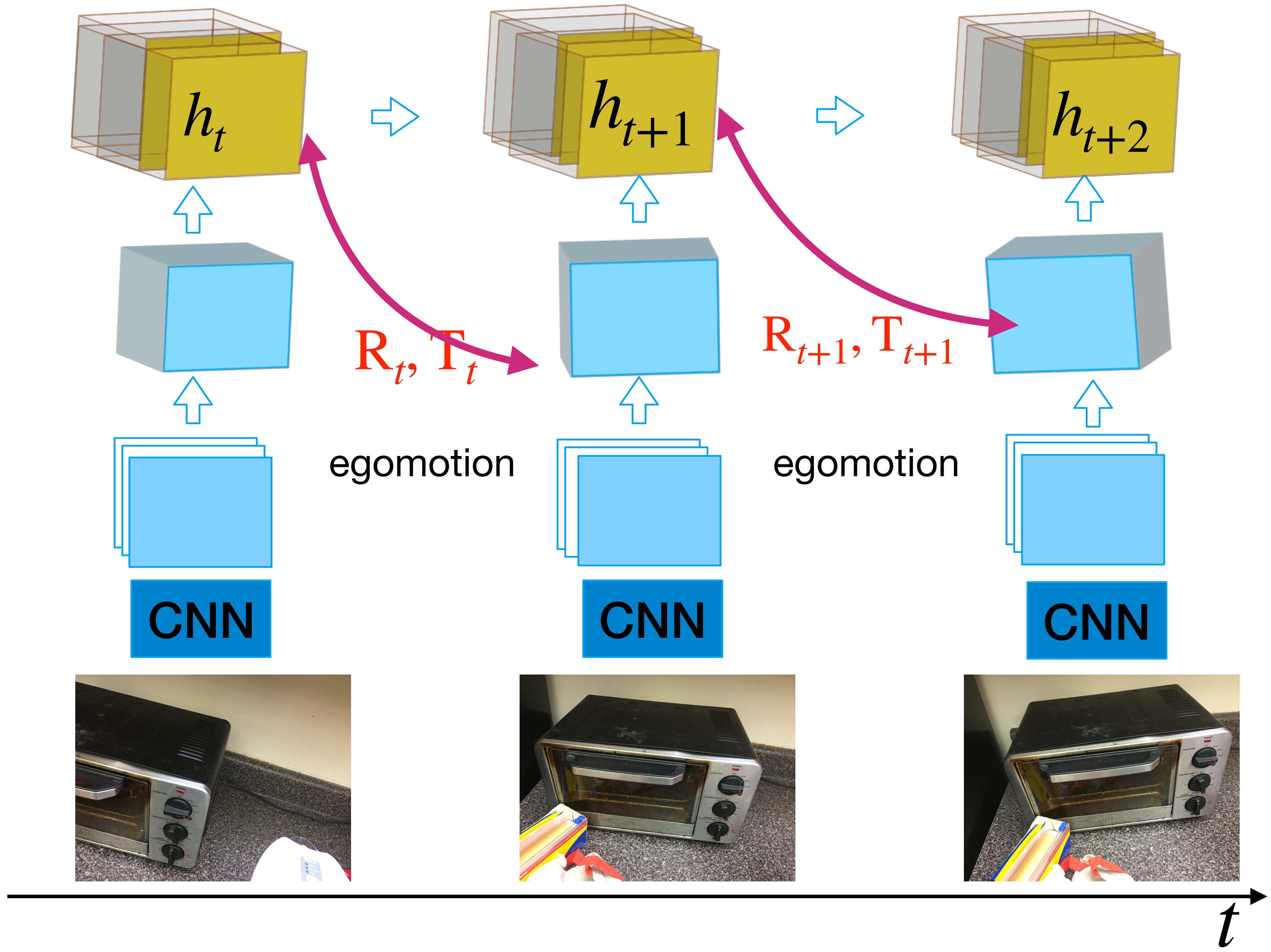
Geometry-Aware Recurrent Networks



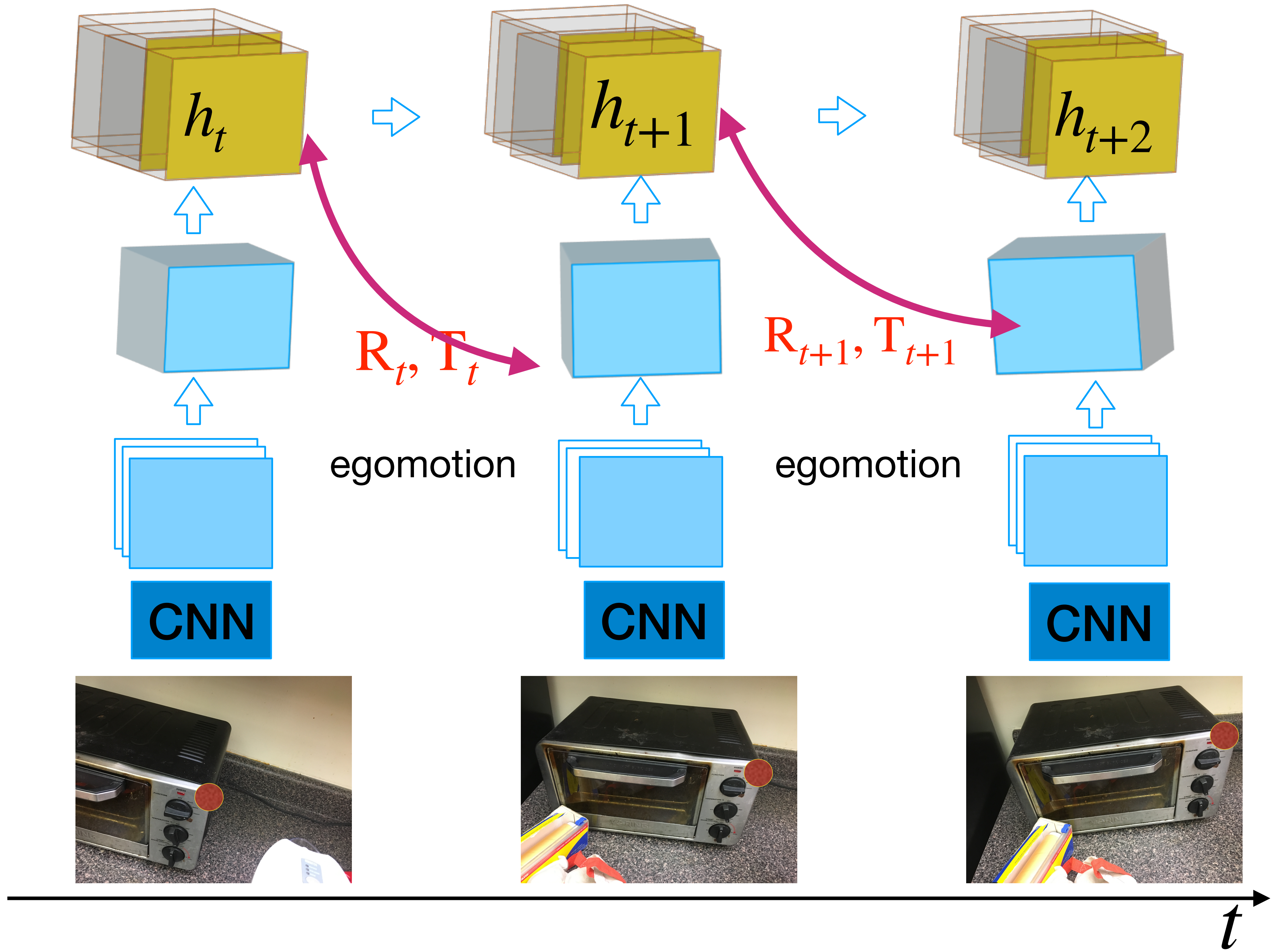
Geometry-Aware Recurrent Networks



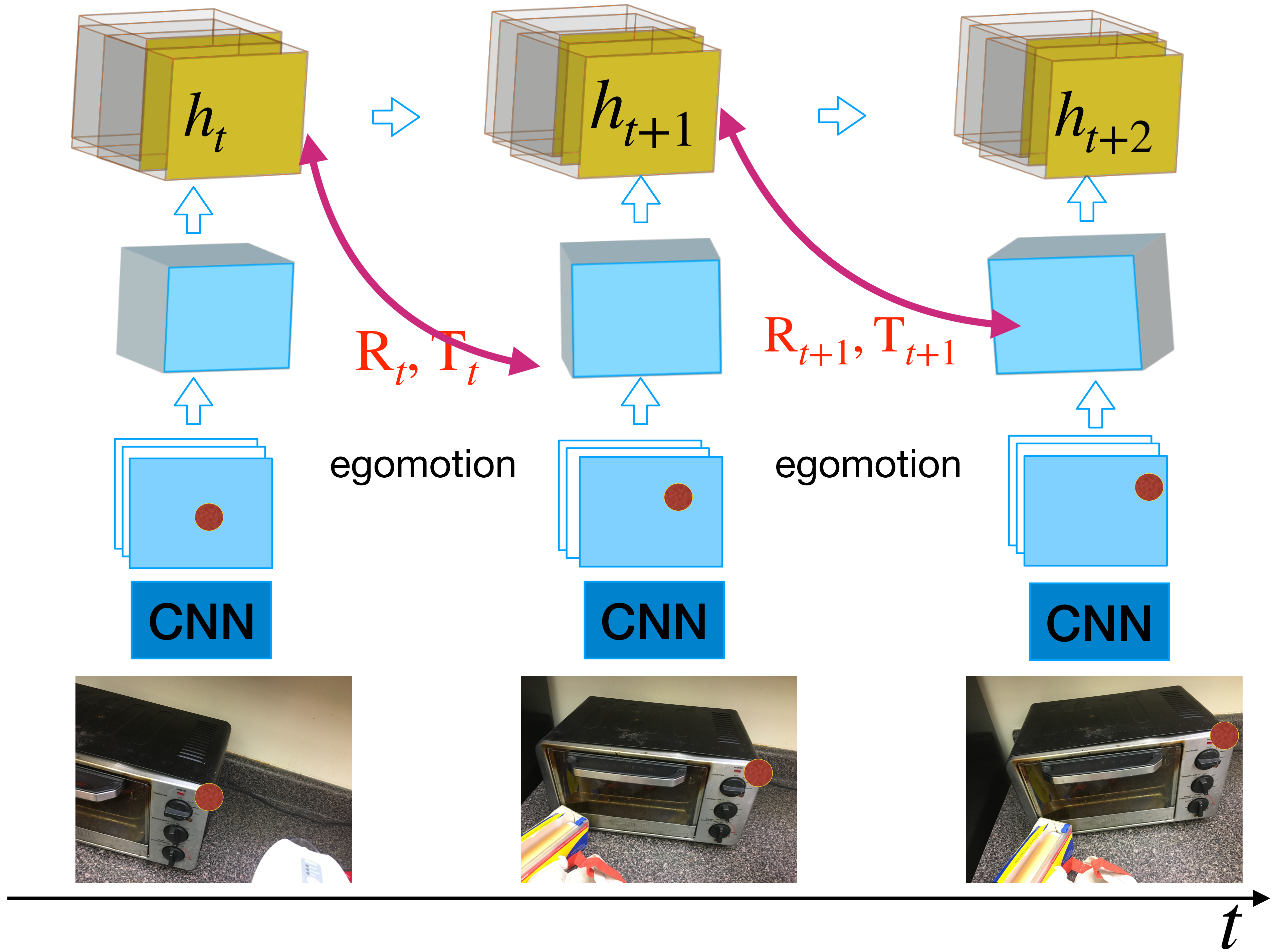
Geometry-Aware Recurrent Networks



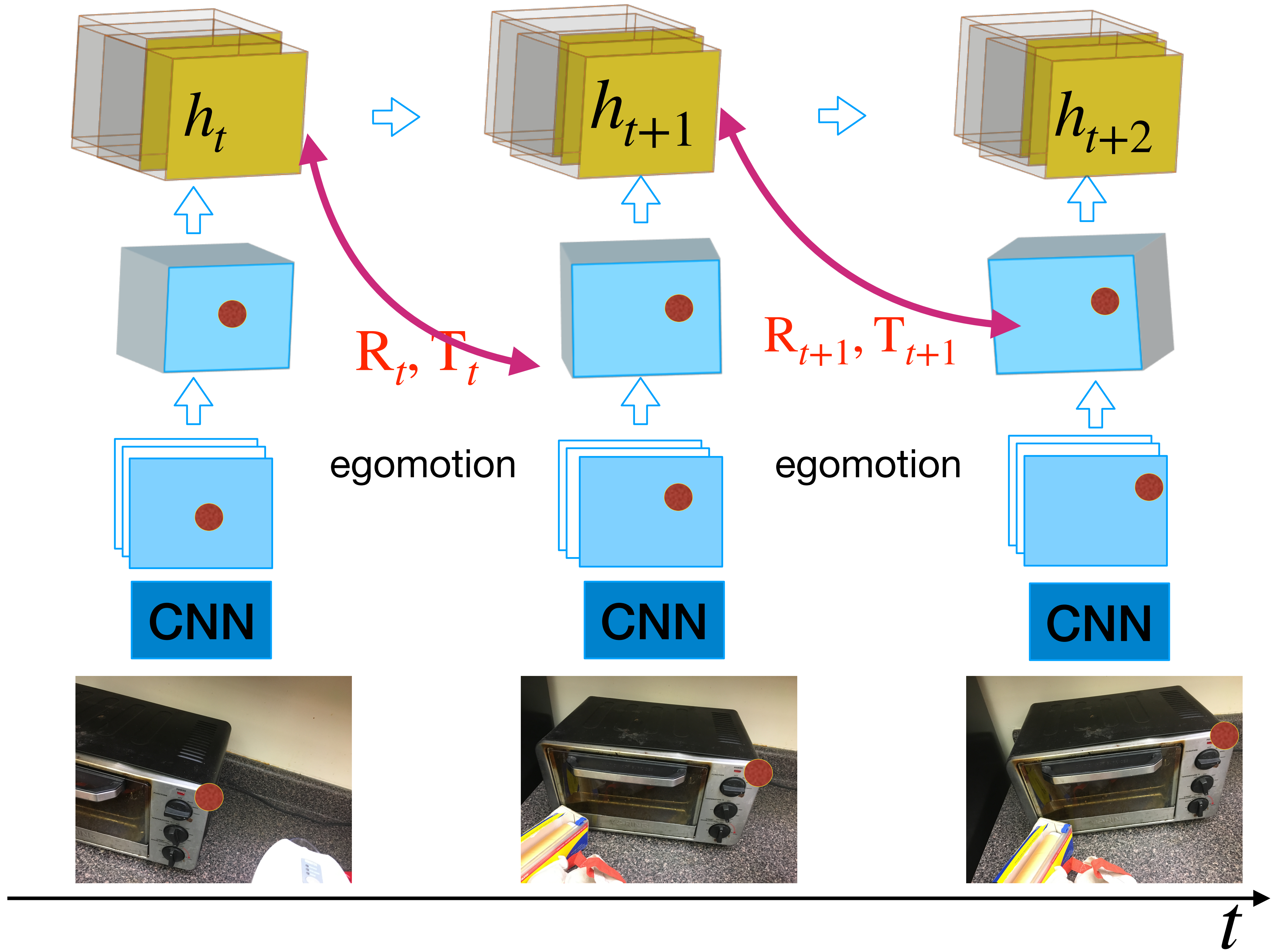
Geometry-Aware Recurrent Networks



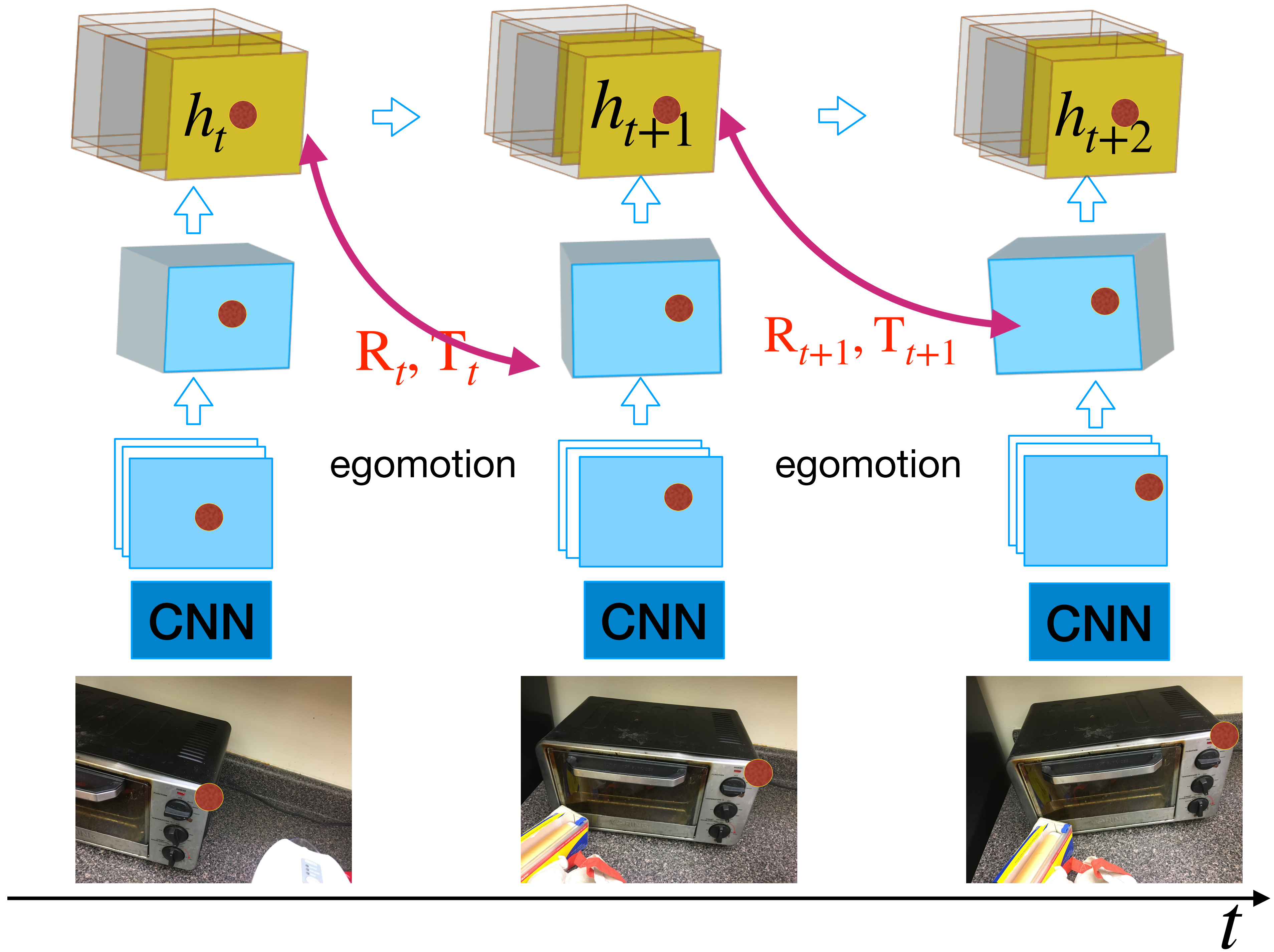
Geometry-Aware Recurrent Networks



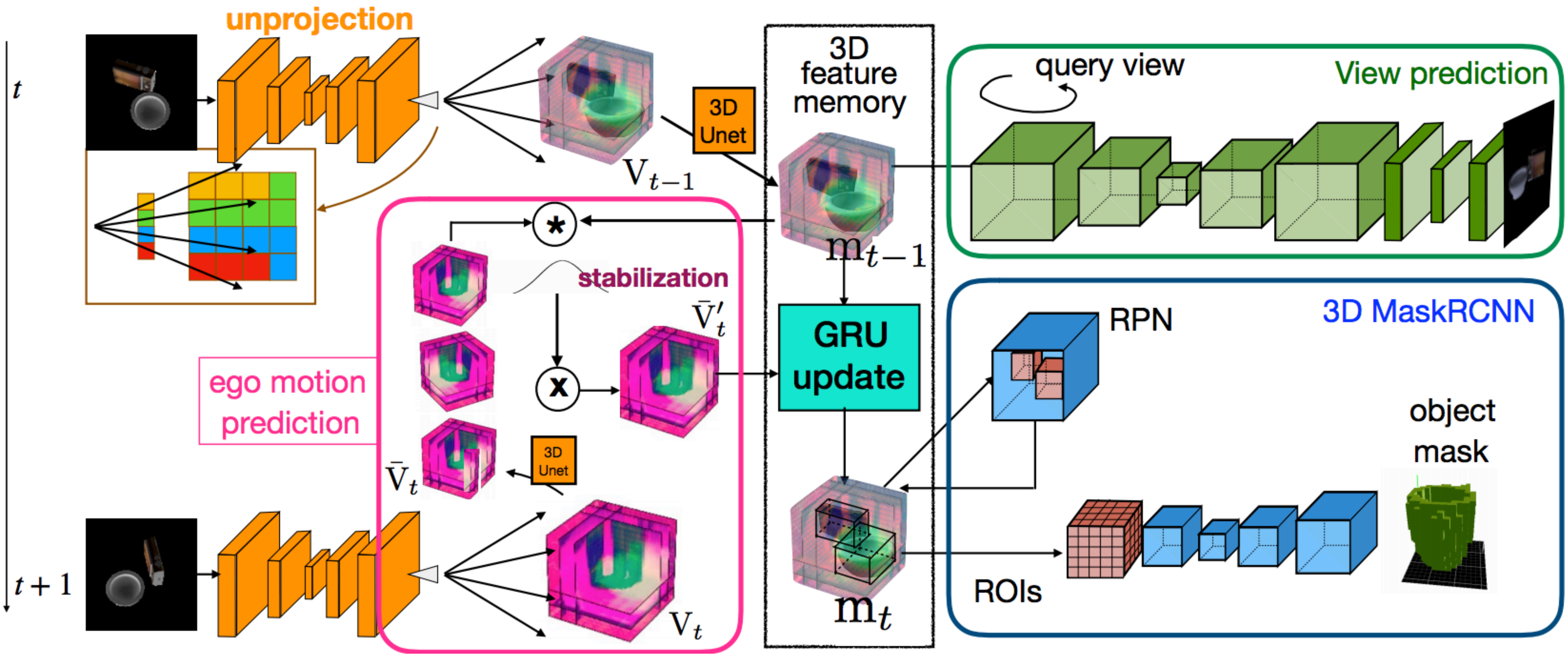
Geometry-Aware Recurrent Networks



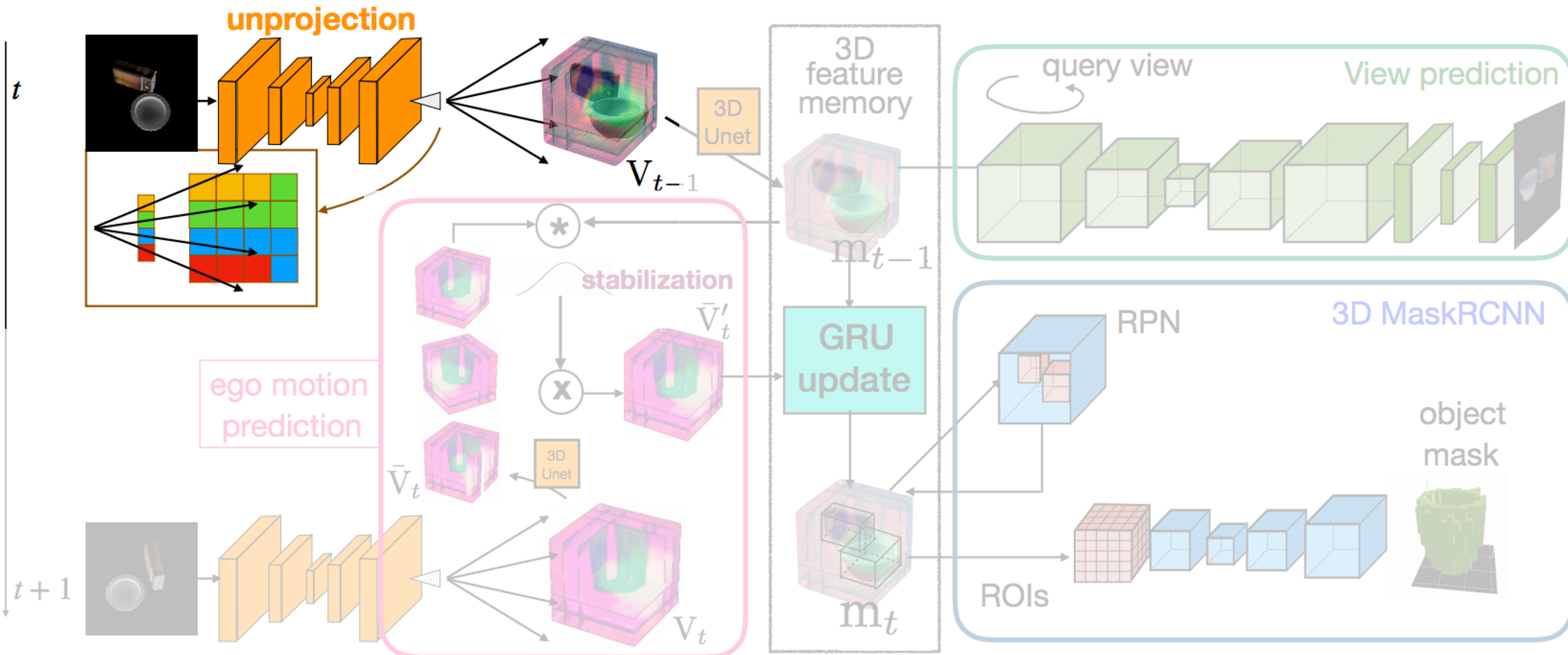
Geometry-Aware Recurrent Networks



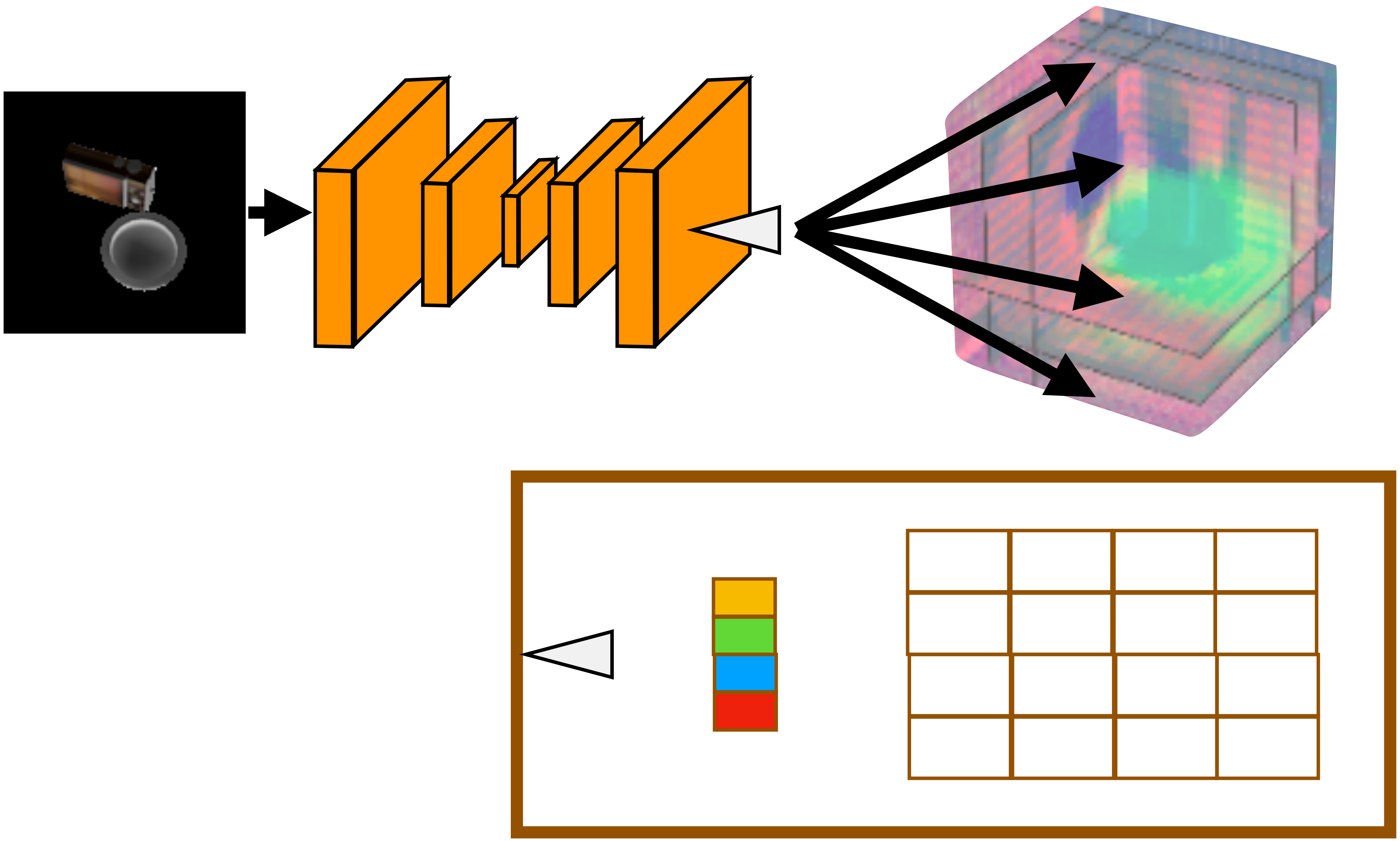
Architecture



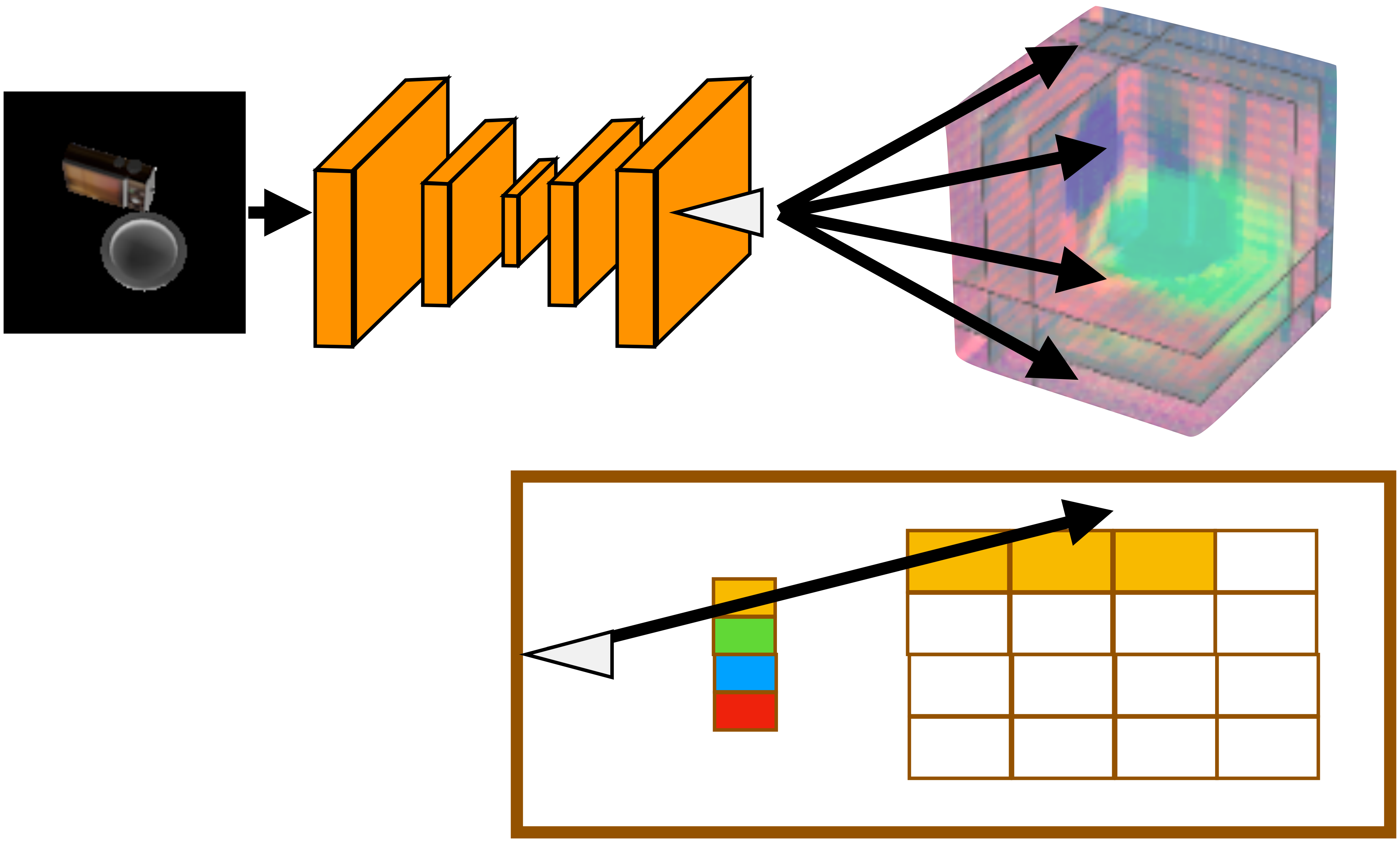
Architecture



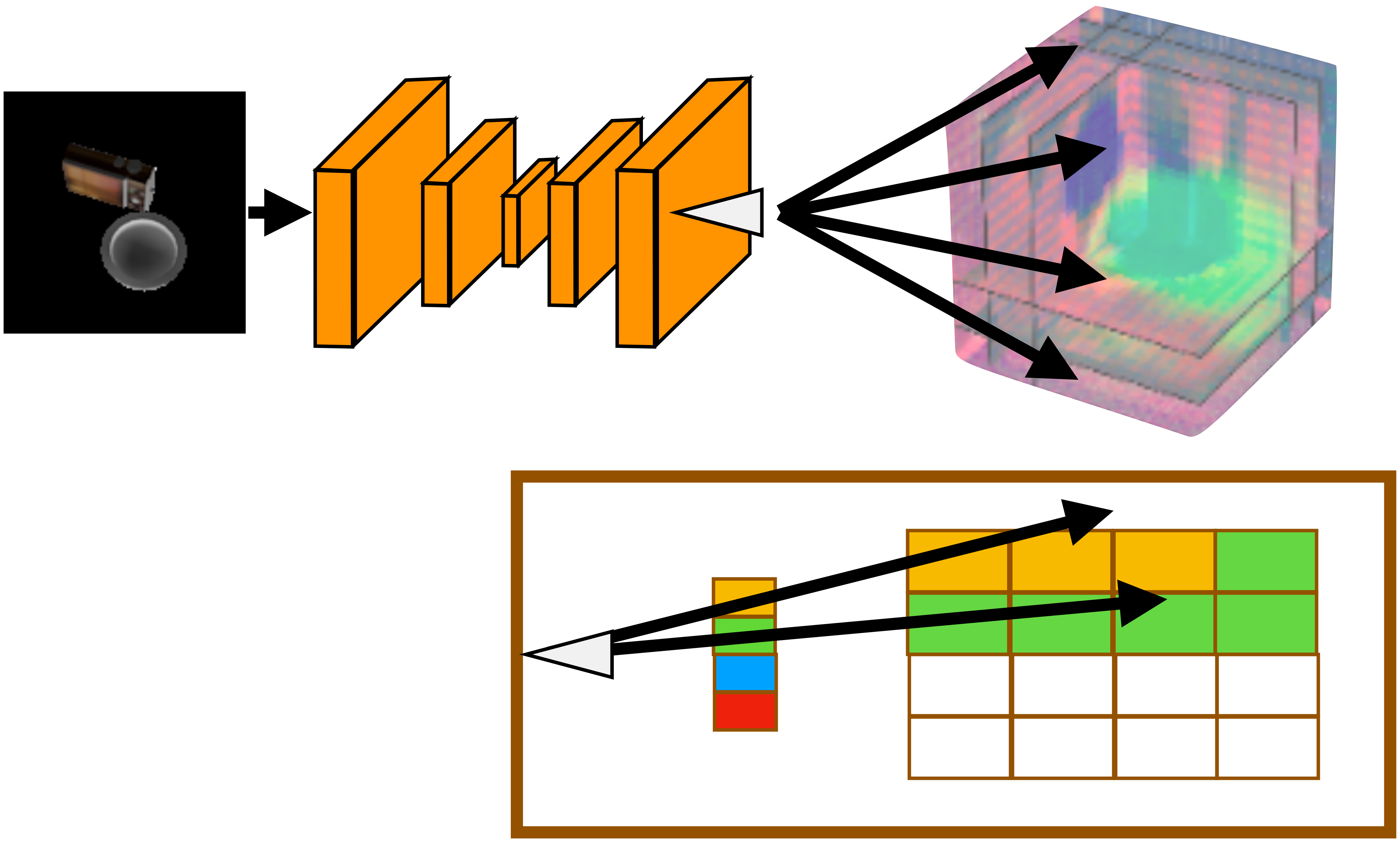
Unprojection (2D to 3D)



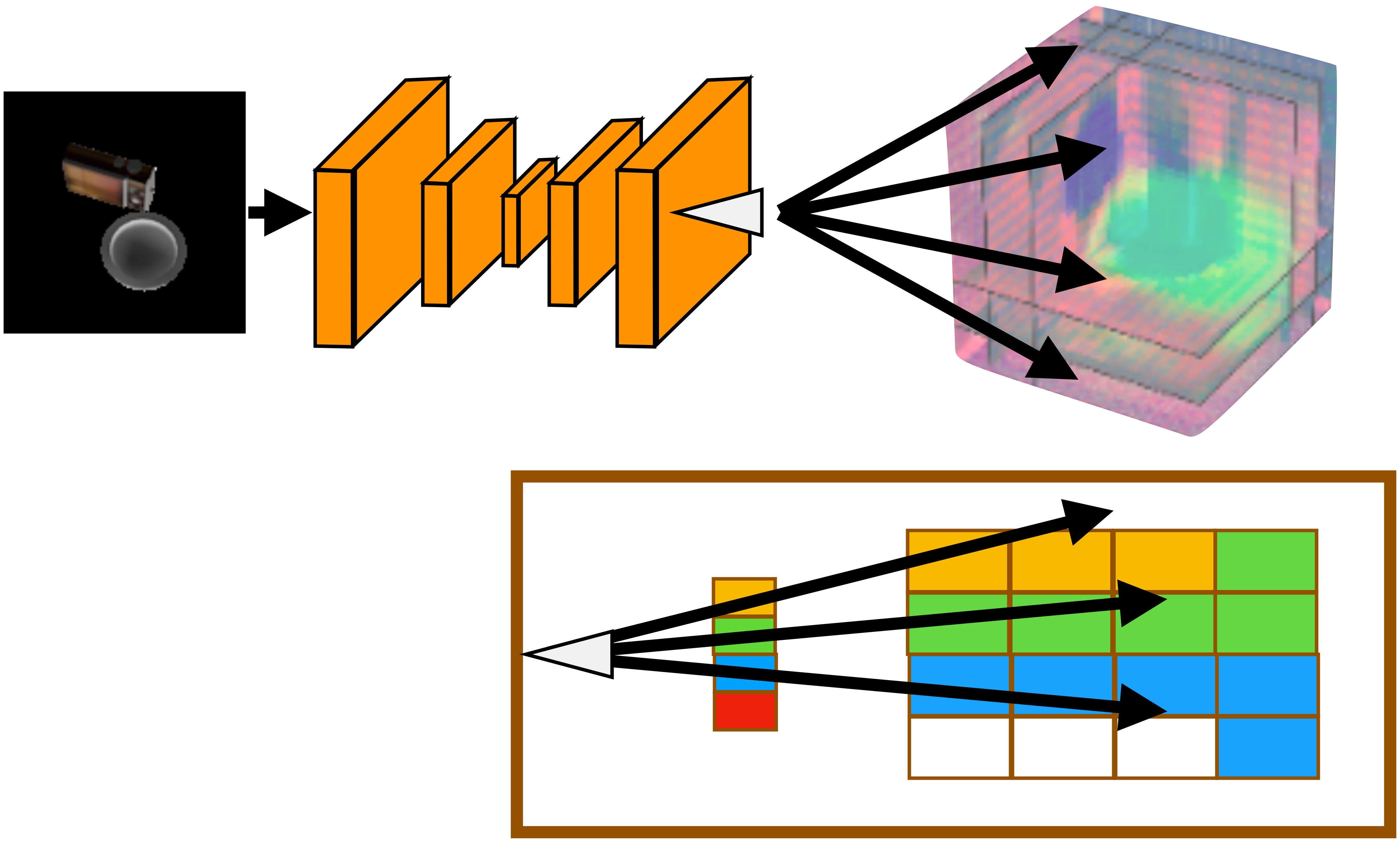
Unprojection (2D to 3D)



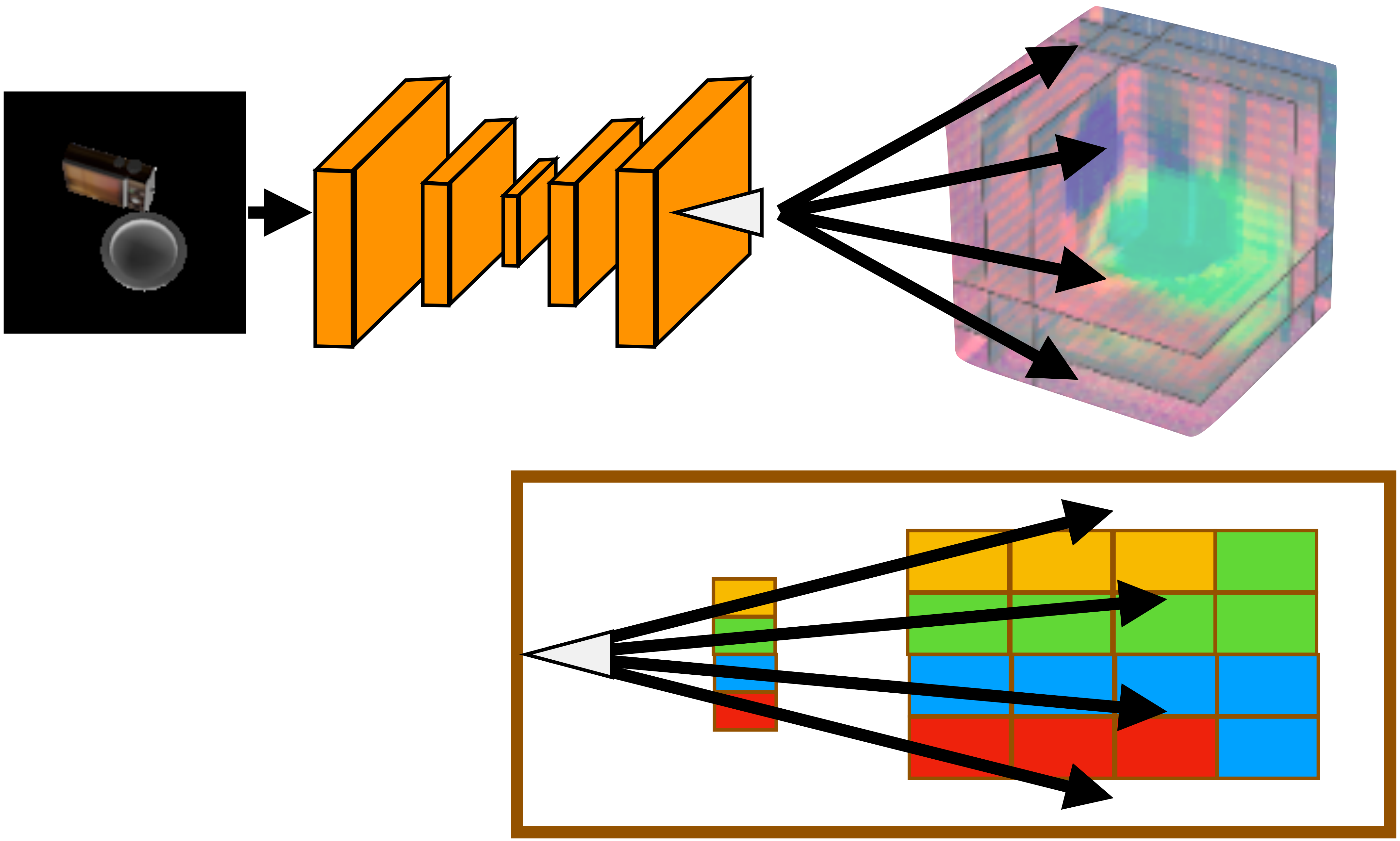
Unprojection (2D to 3D)



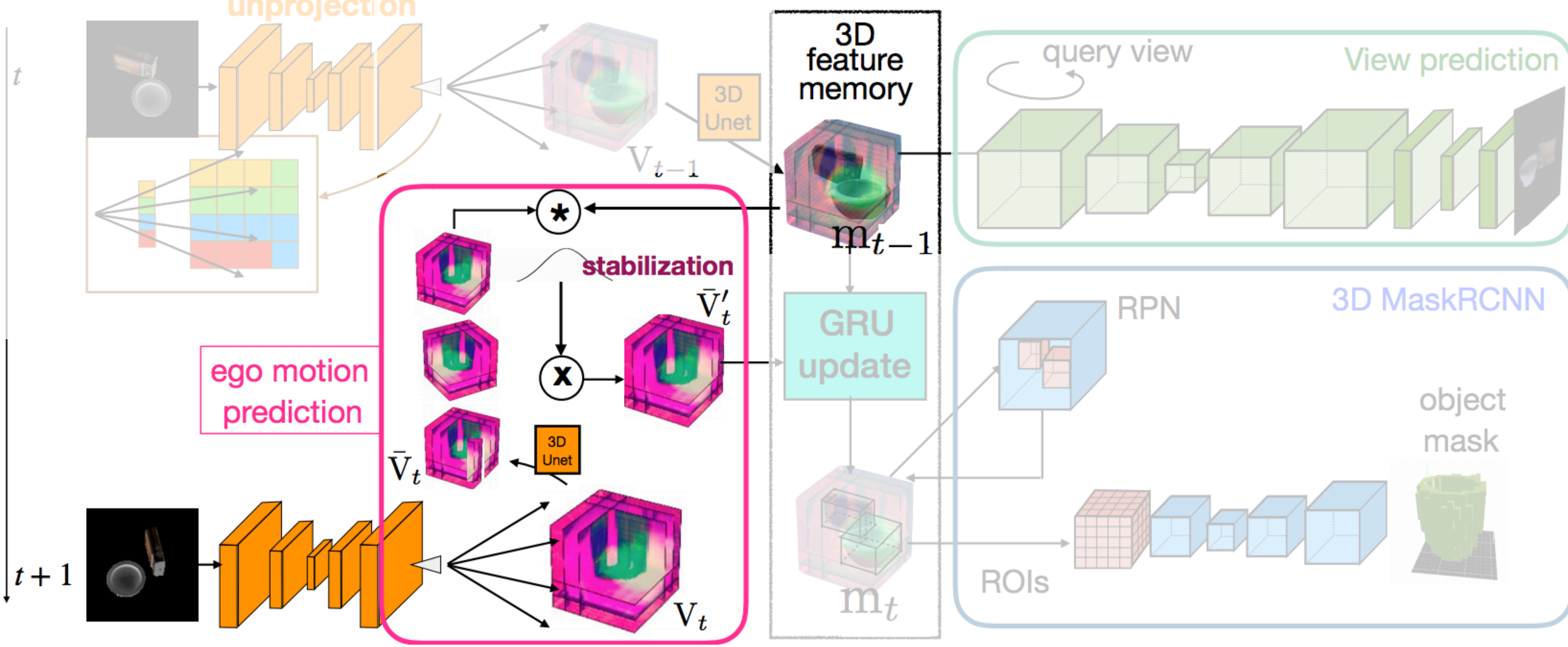
Unprojection (2D to 3D)



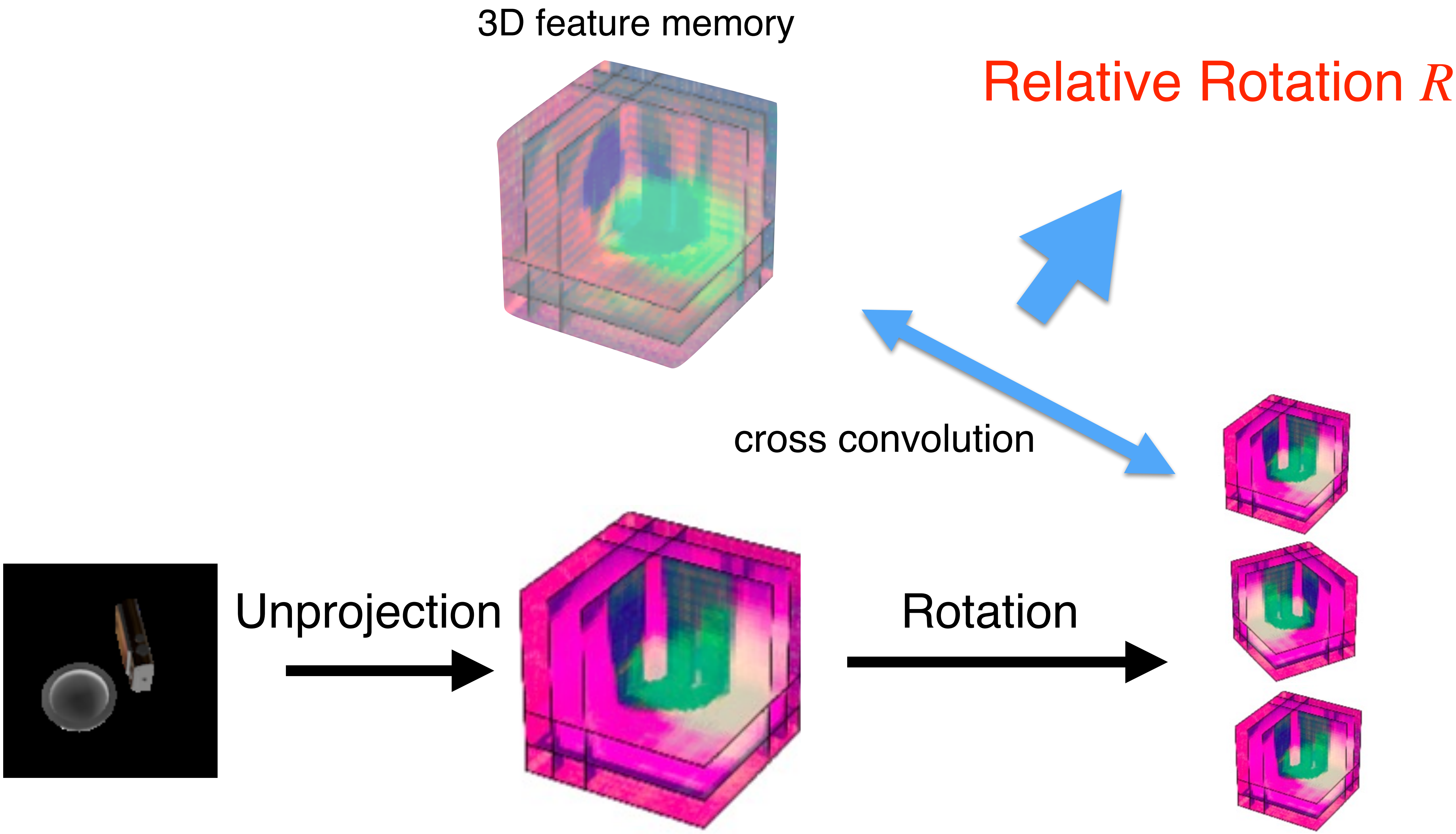
Unprojection (2D to 3D)



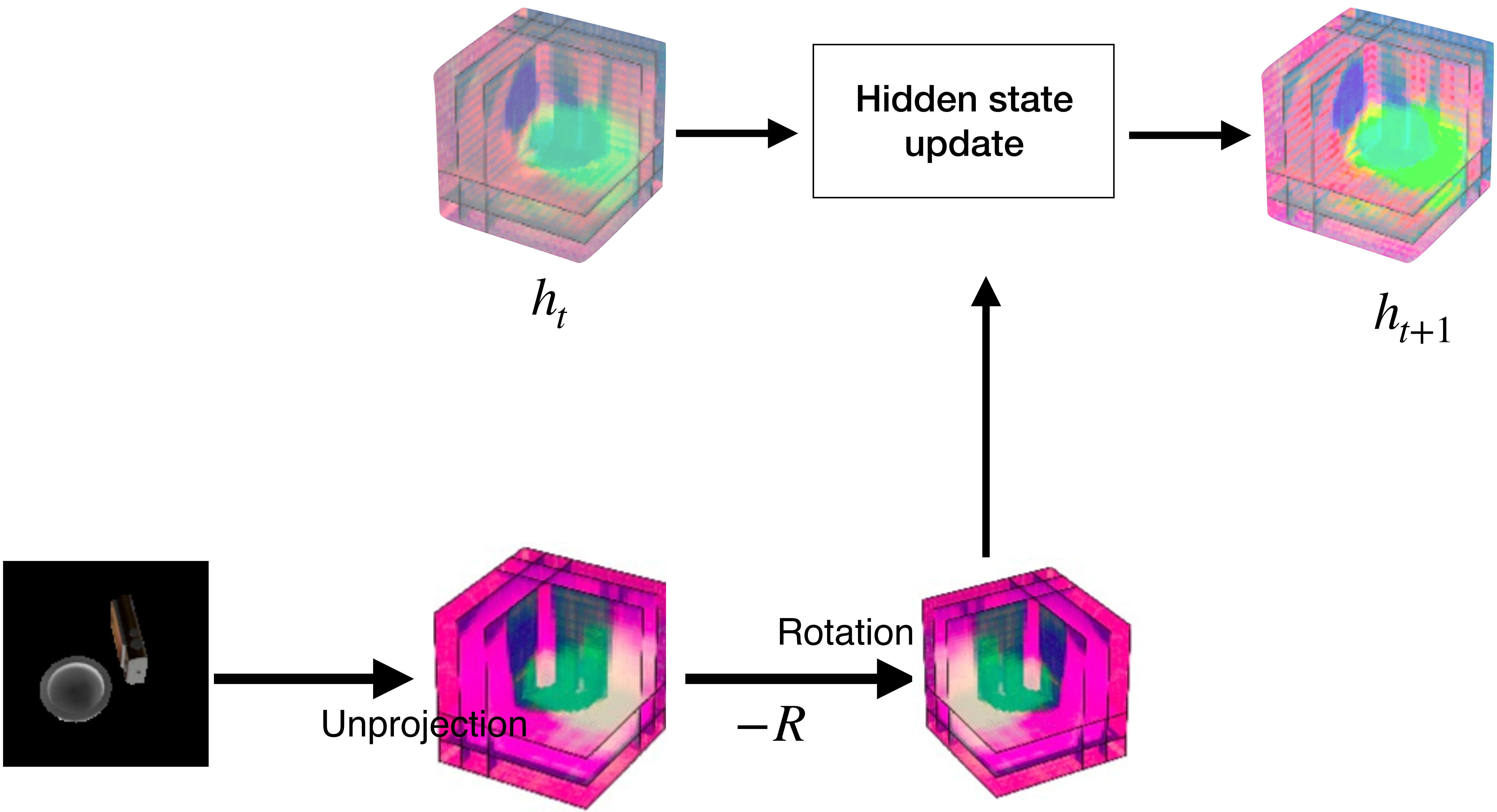
Architecture



Egomotion-stabilized memory update



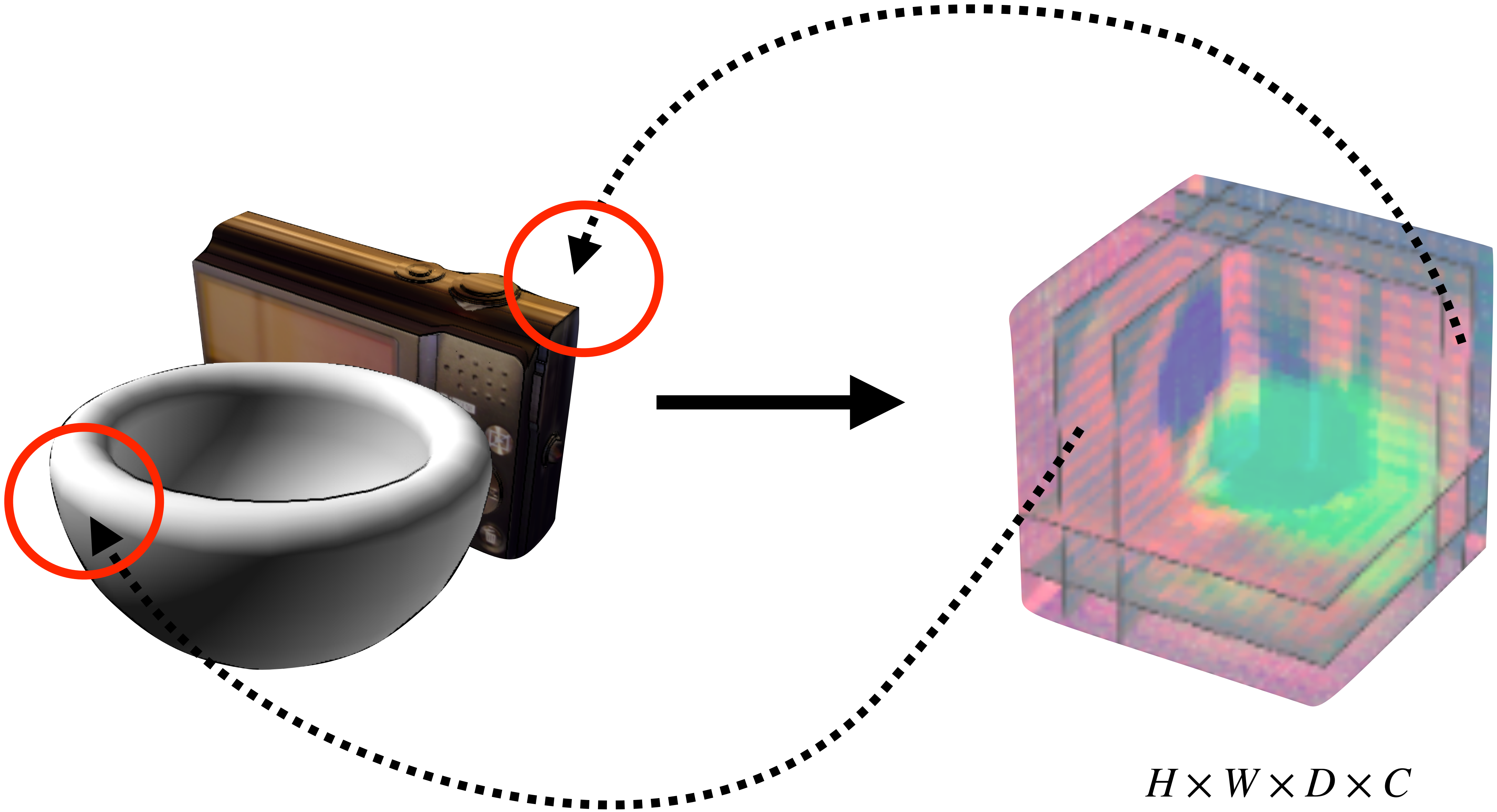
Egomotion-stabilized memory update



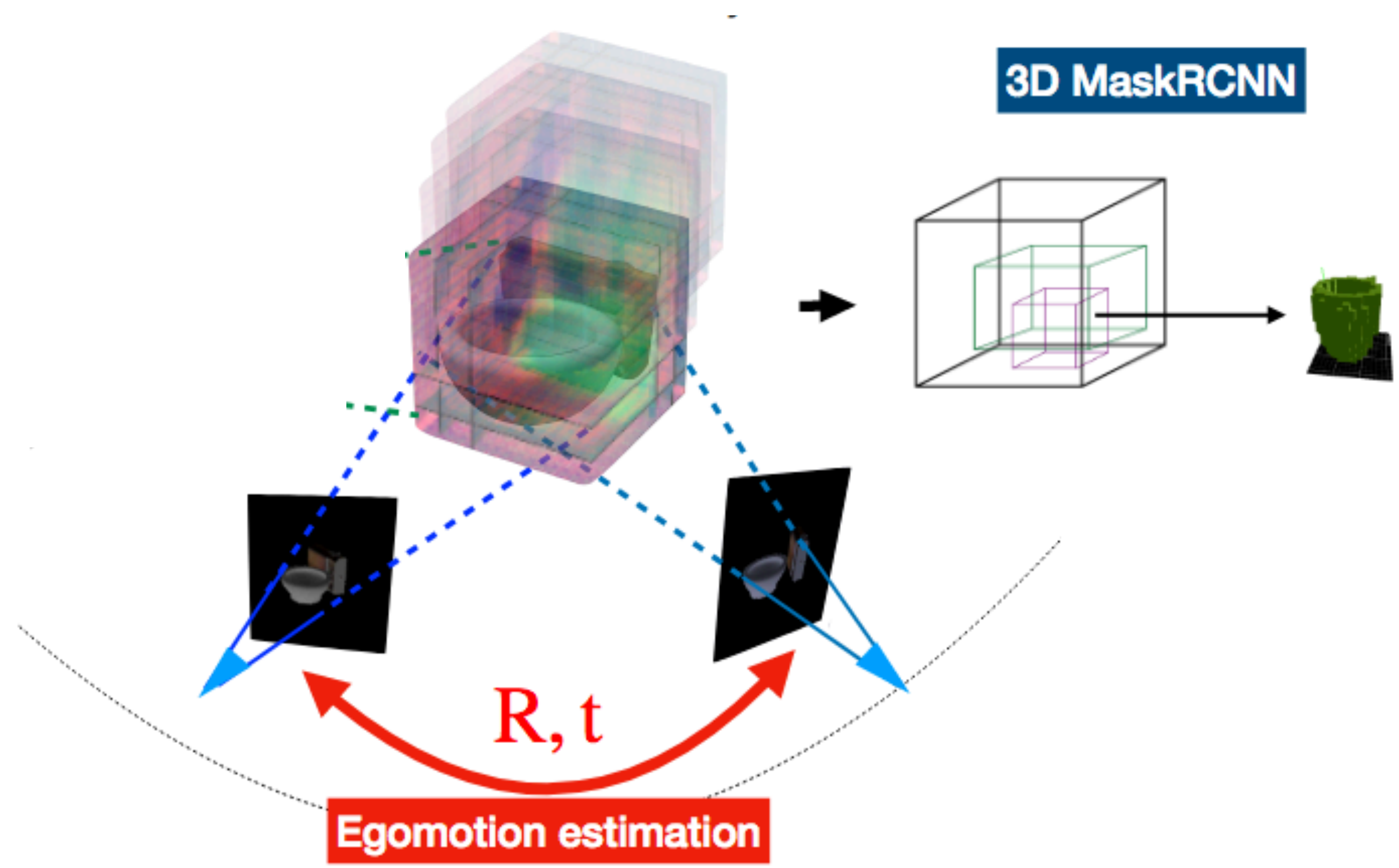
Geometry-Aware Recurrent Networks (GRNNs)

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

Geometry-Aware Recurrent Networks (GRNNs)

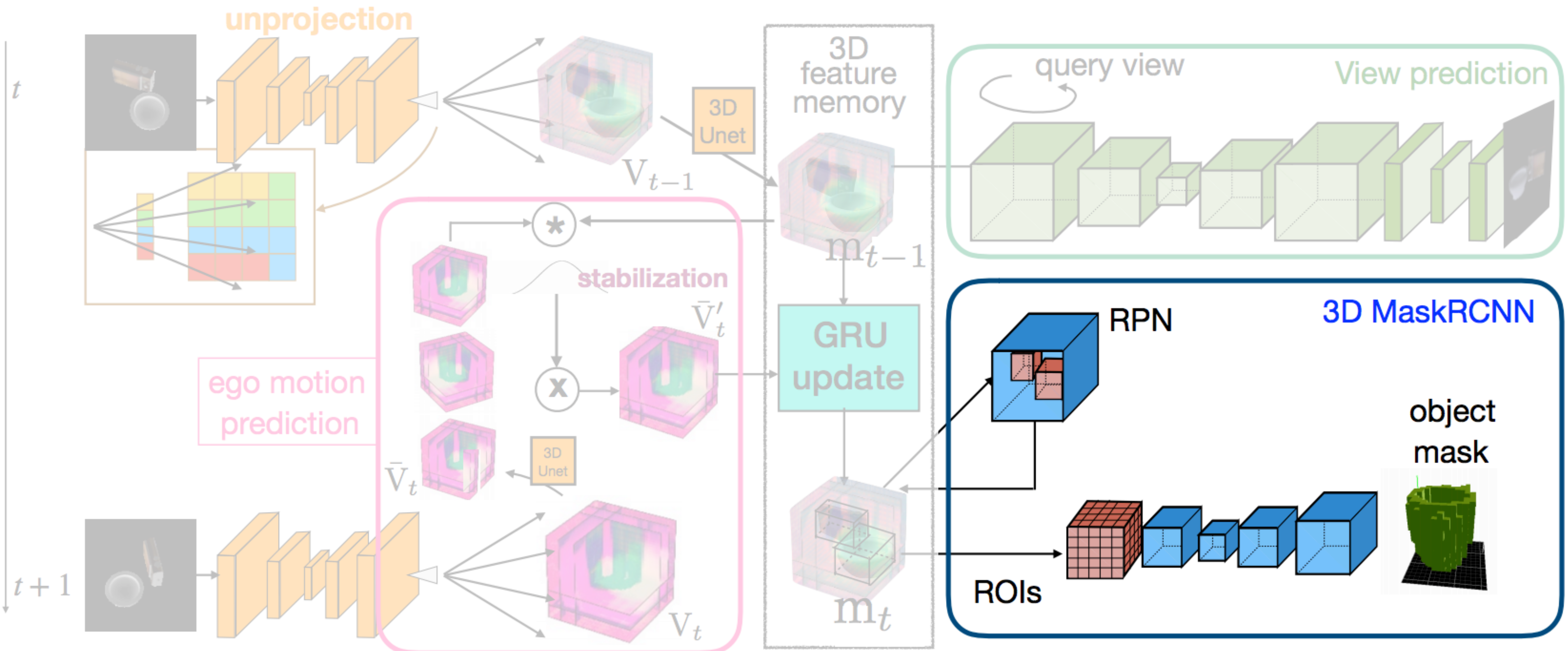


Training GRNNs



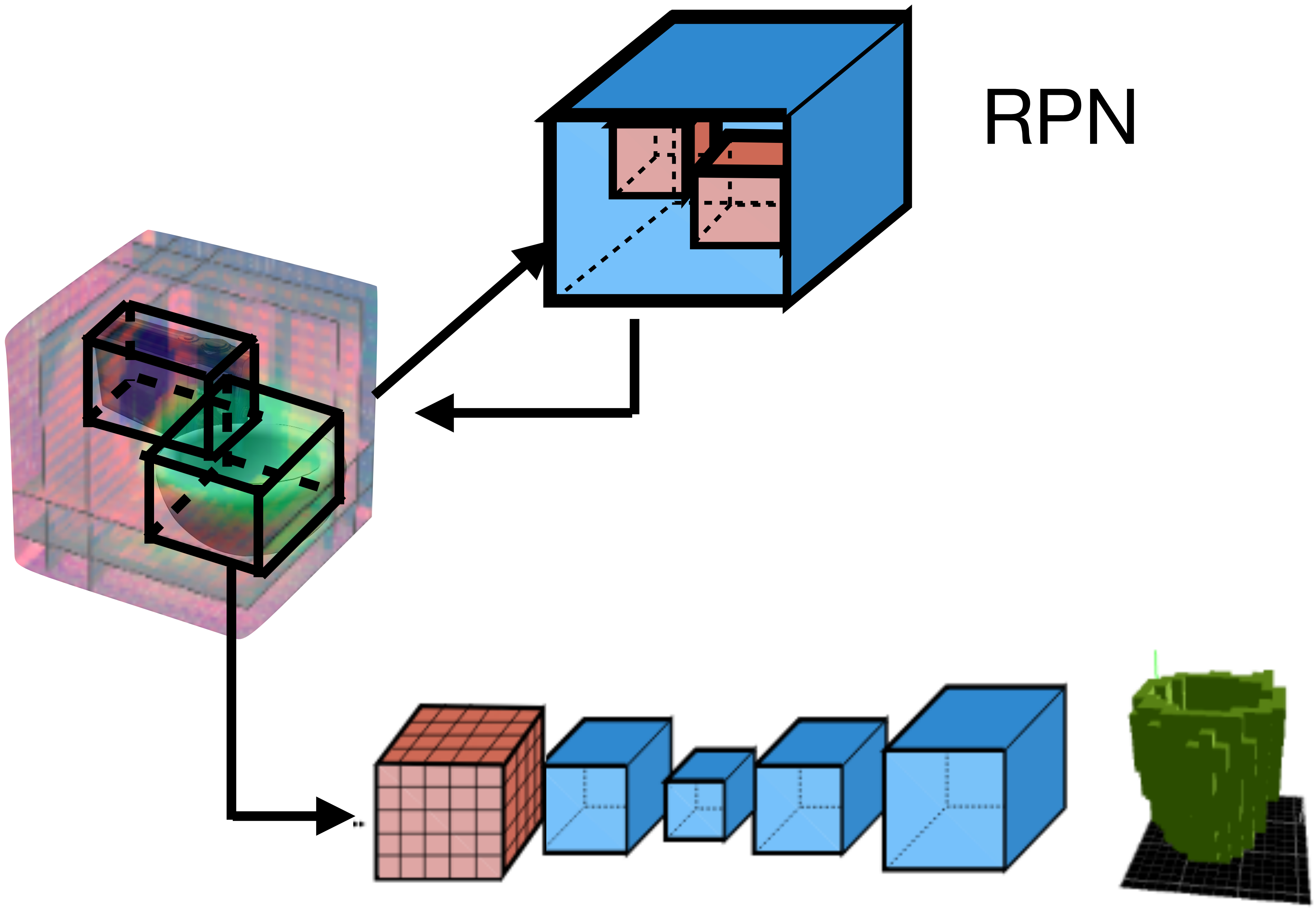
1. Supervised for 3D object detection
2. Self-supervised for view prediction

Architecture



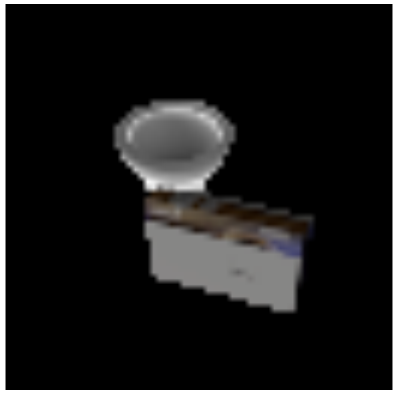
3D Object Detection

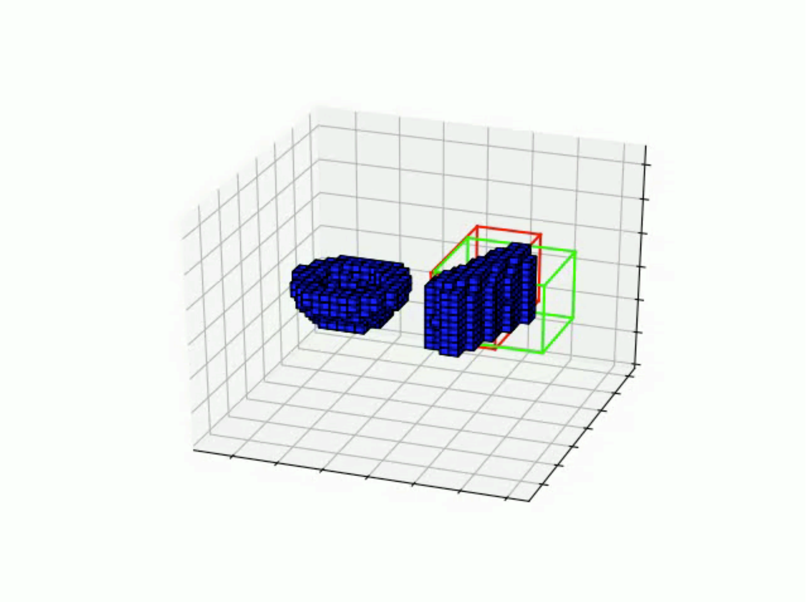
Input: the 3D latent feature map
Output: 3D boxes and segmentations for the objects



Results - 3D object detection

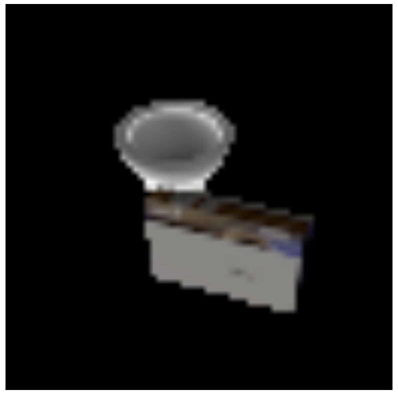
of input views

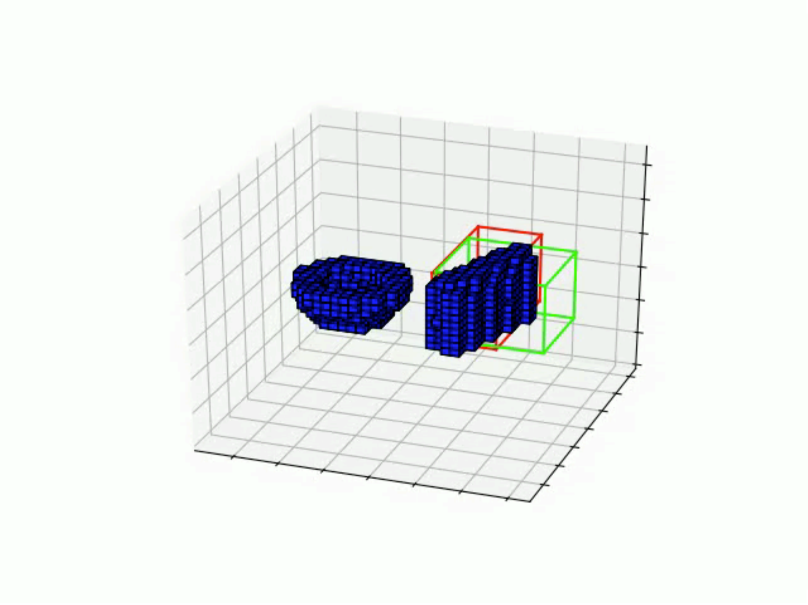
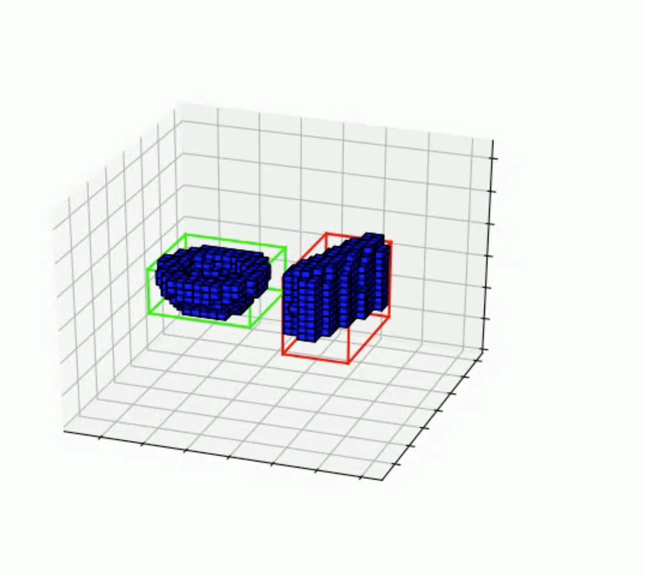




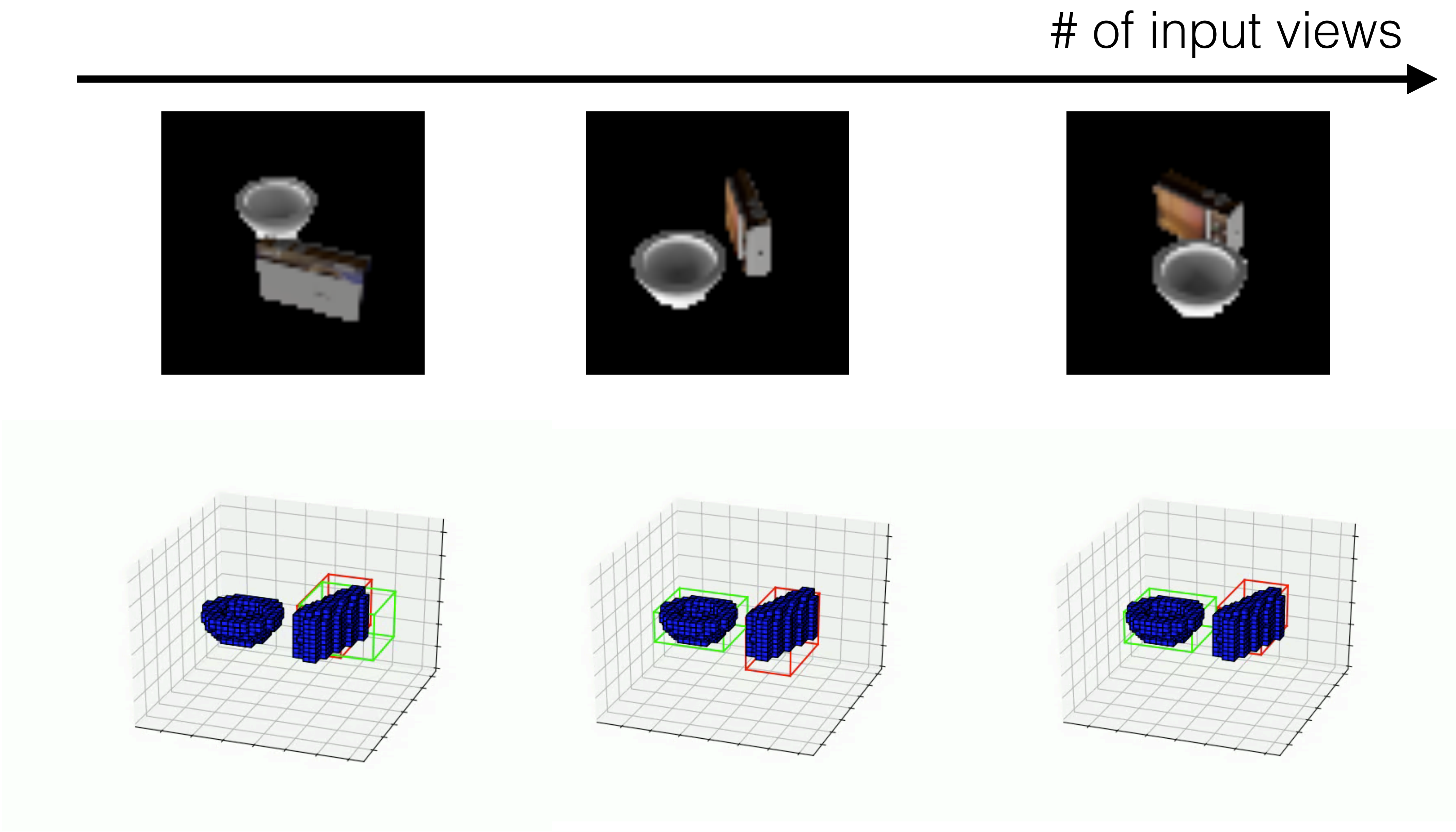
Results - 3D object detection

of input views

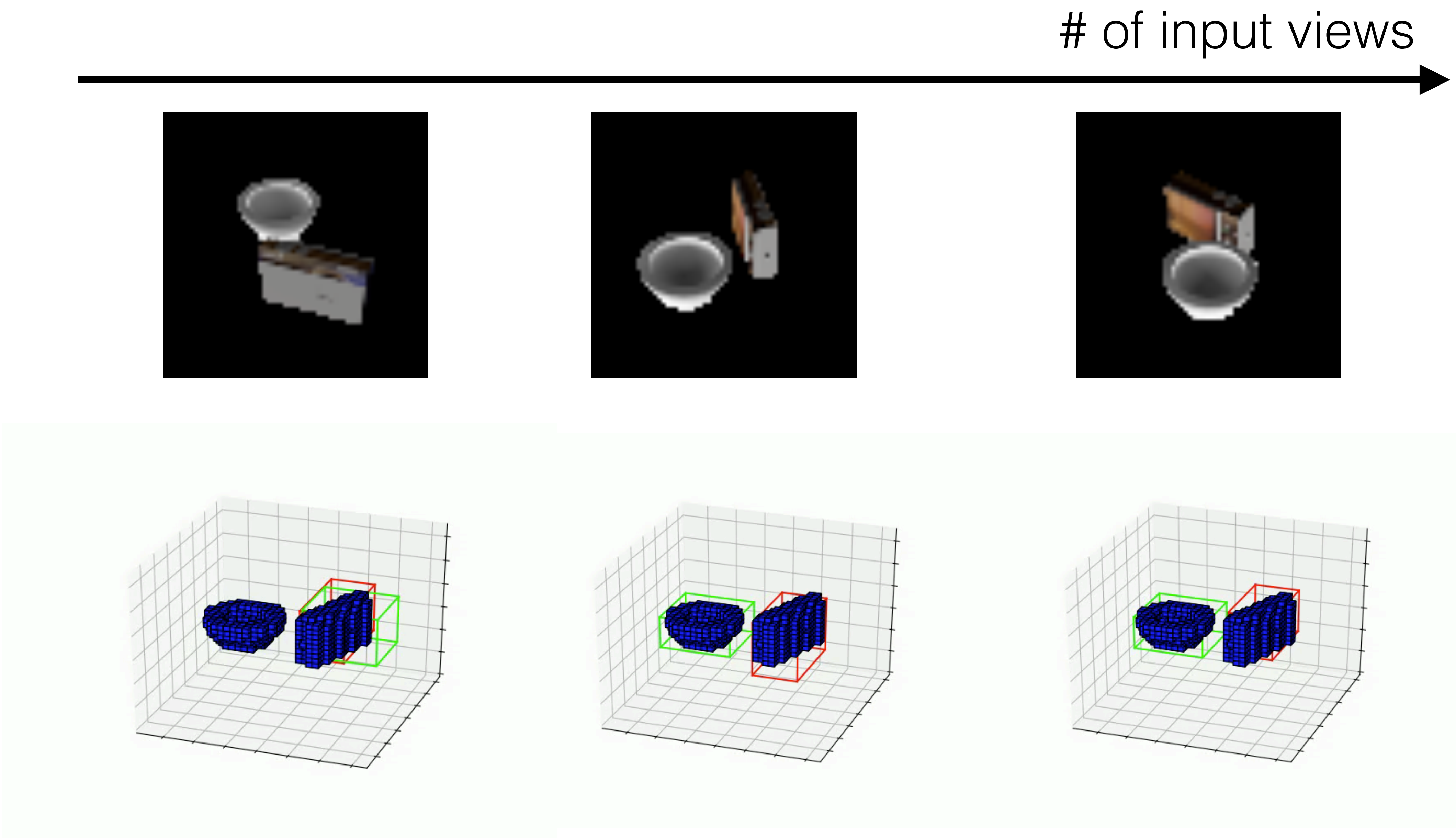
 

Results - 3D object detection

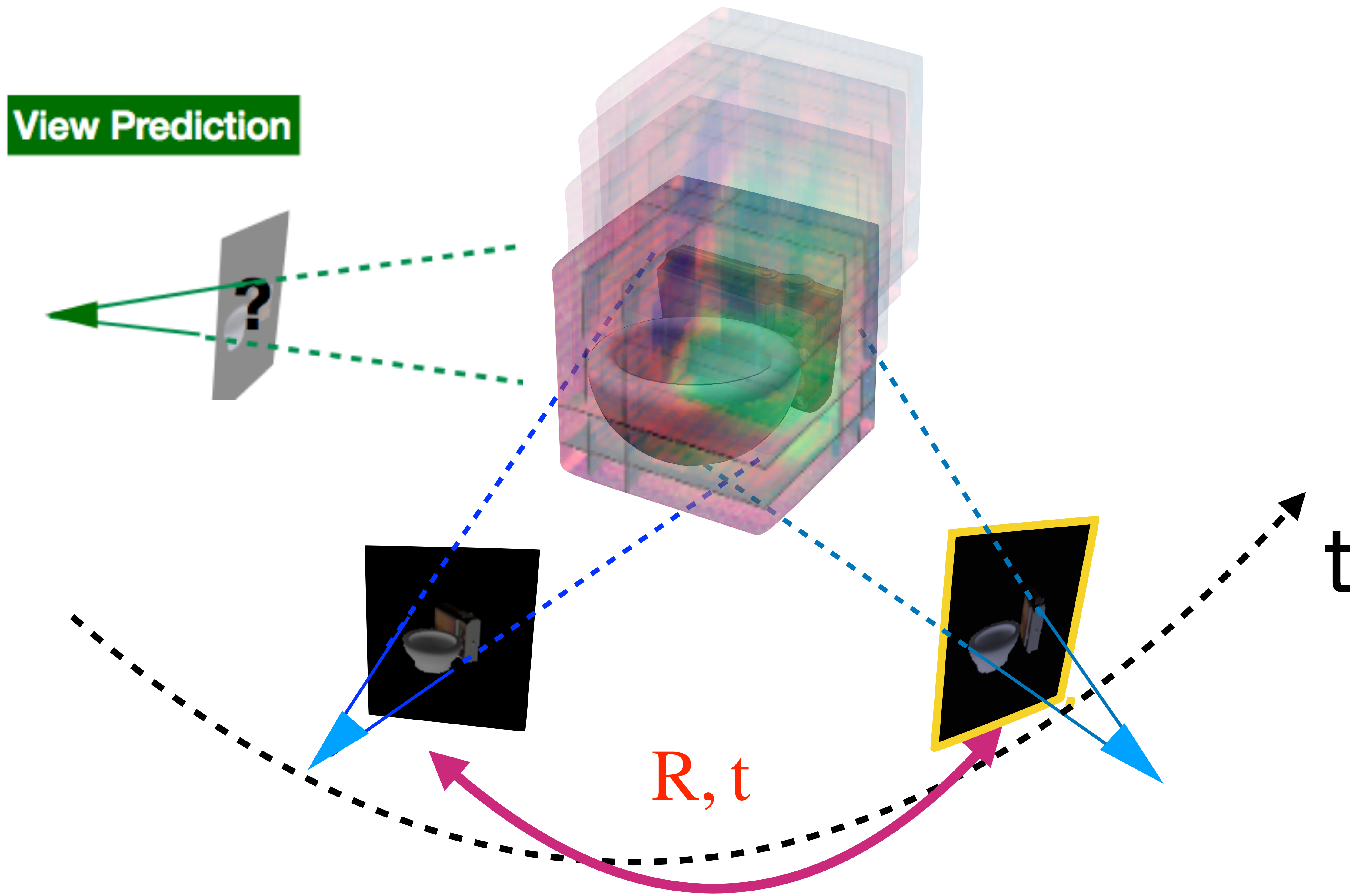


Object permanence emerges



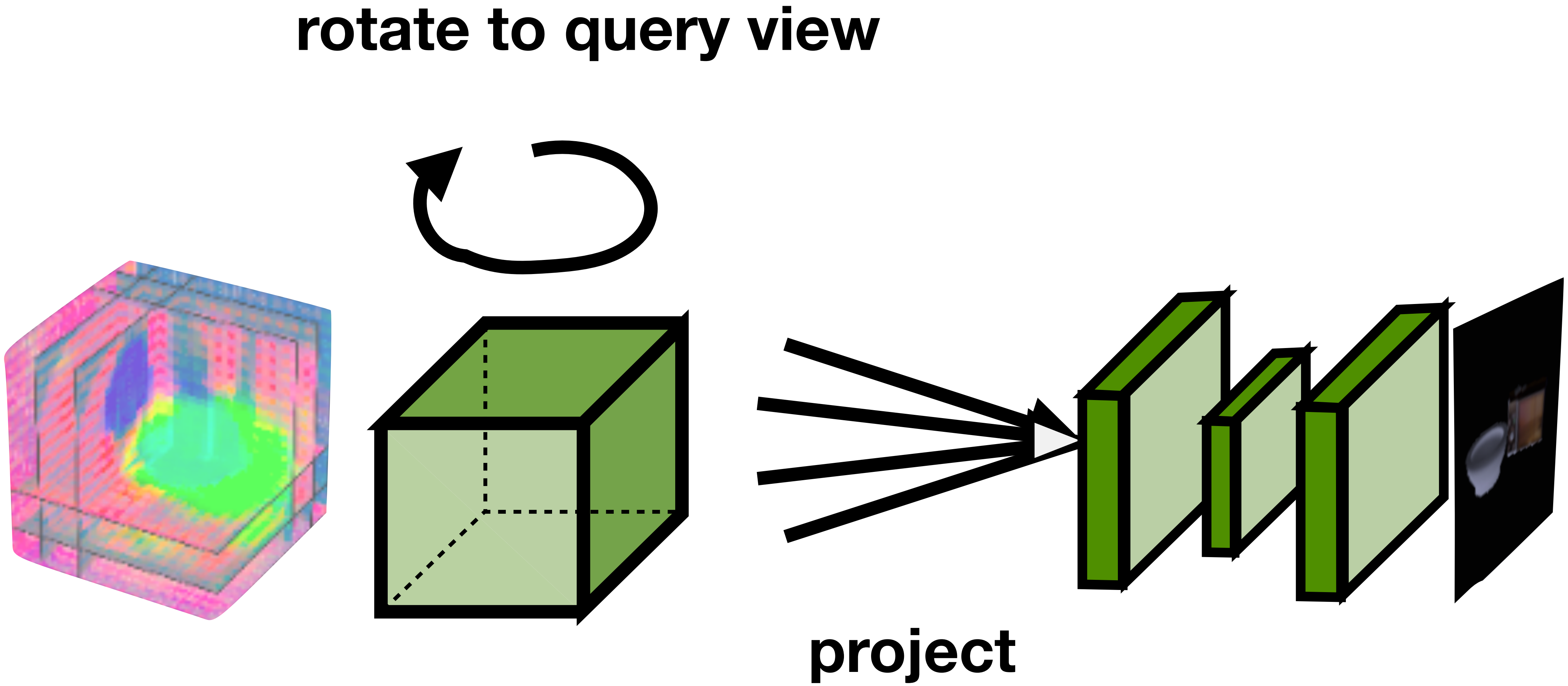
- Objects persist over time, objects have 3D extent

View prediction



Learning spatial common sense with geometry-aware RNNs, Tung et al. 2019

View prediction

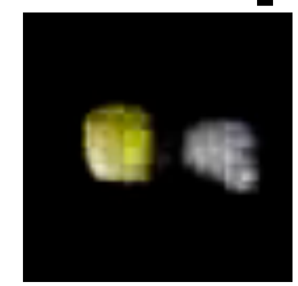


Results - view prediction

Input views

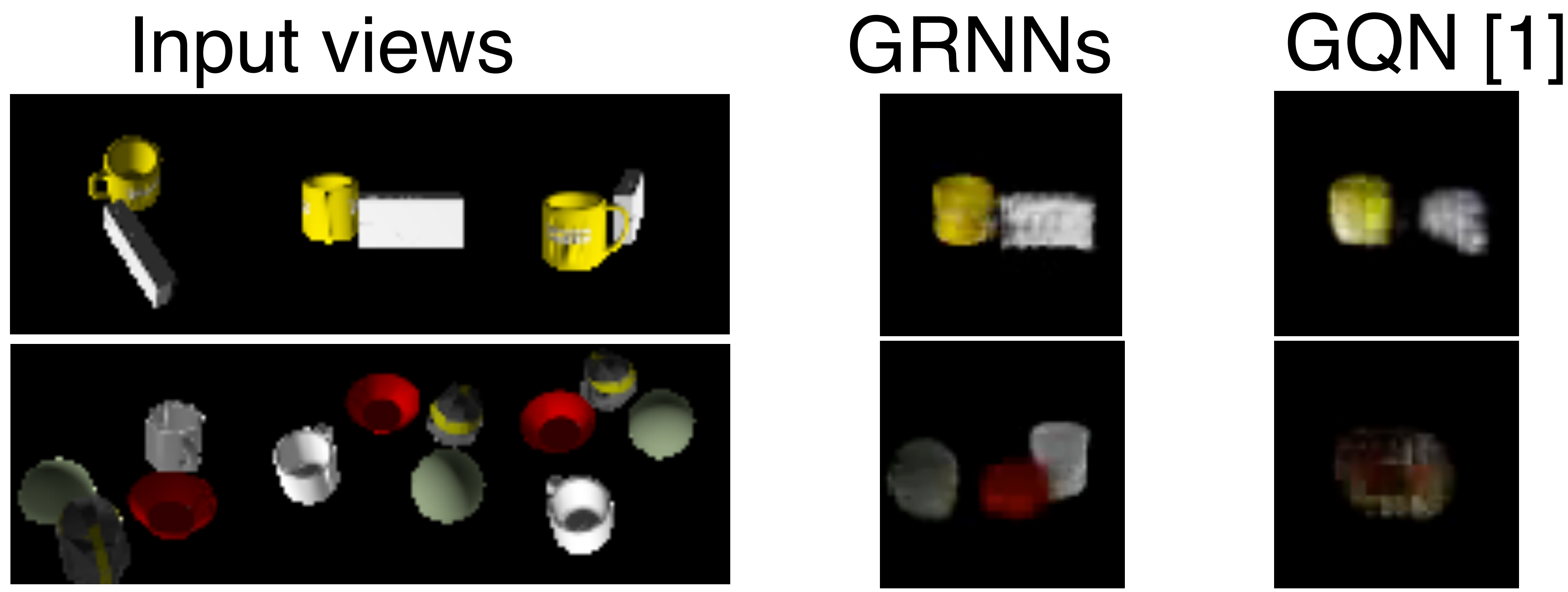
GRNNs

GQN [1]



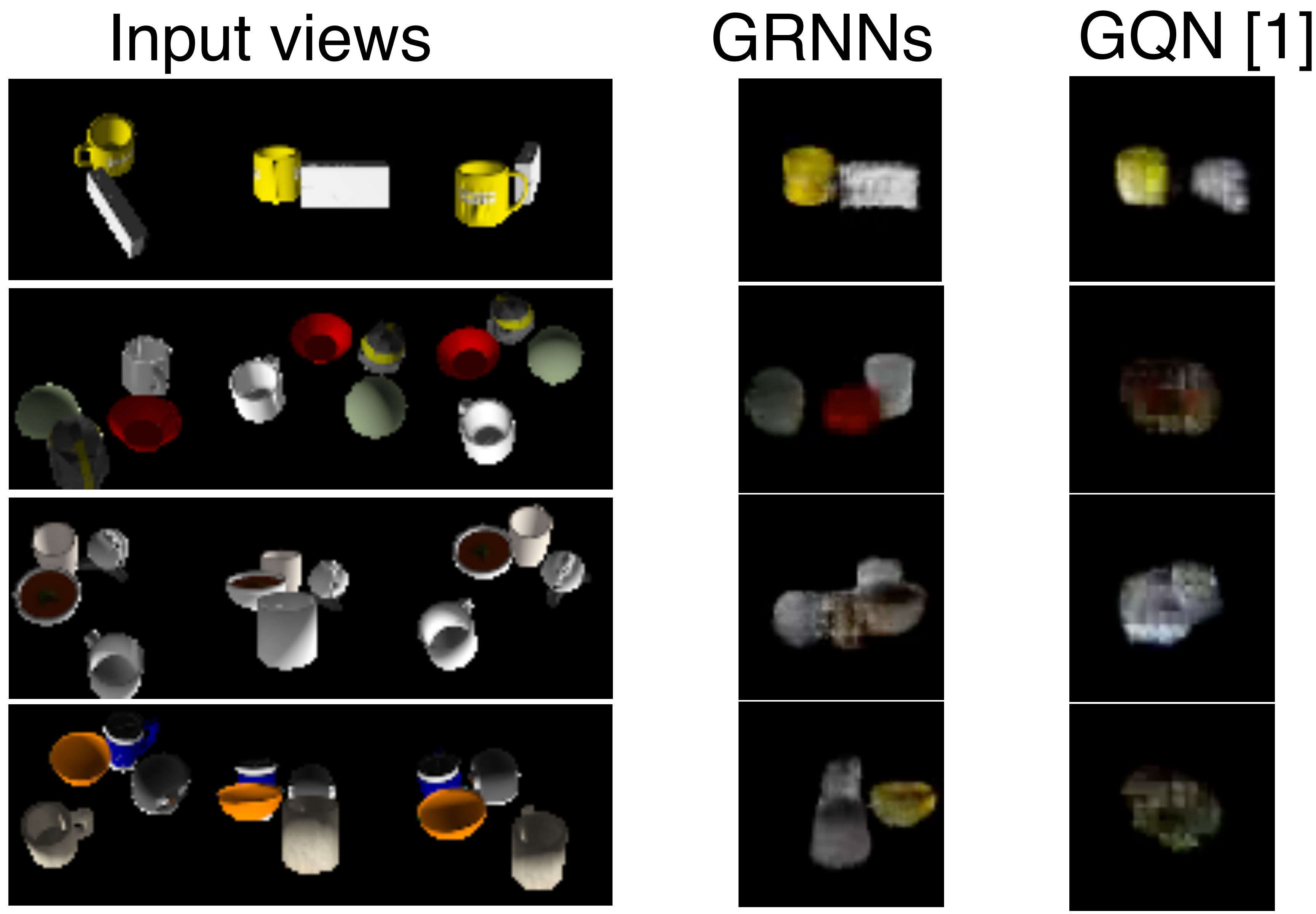
1. *Neural scene representation and rendering Dee*

Results - view prediction



1. Neural scene representation and rendering Deep

Results - view prediction

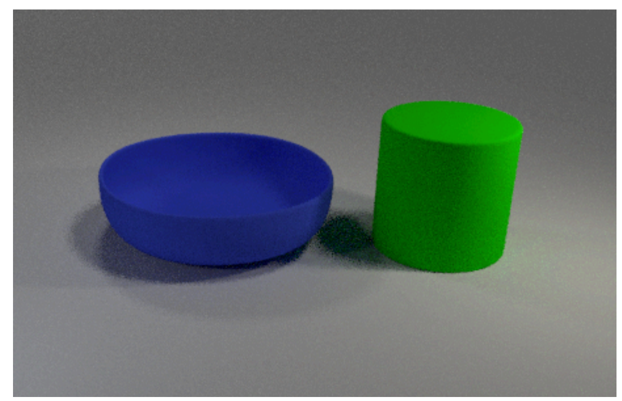


1. *Neural scene representation and rendering Dee*

Embodied language grounding

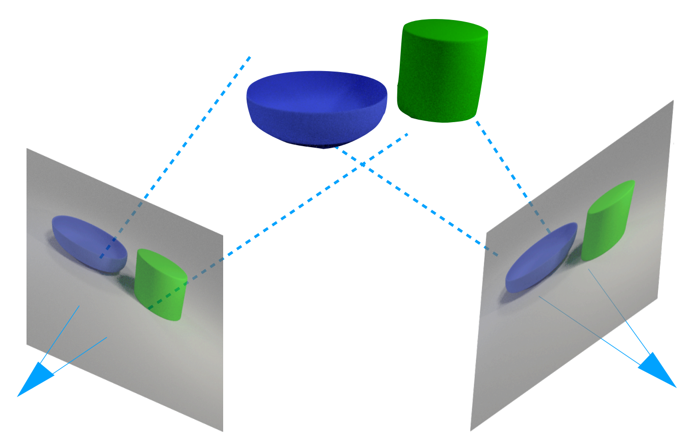
Learn to associate natural language utterances with 3D feature representations of the scene described.

“The green rubber cylinder is on the right of the blue bowl”



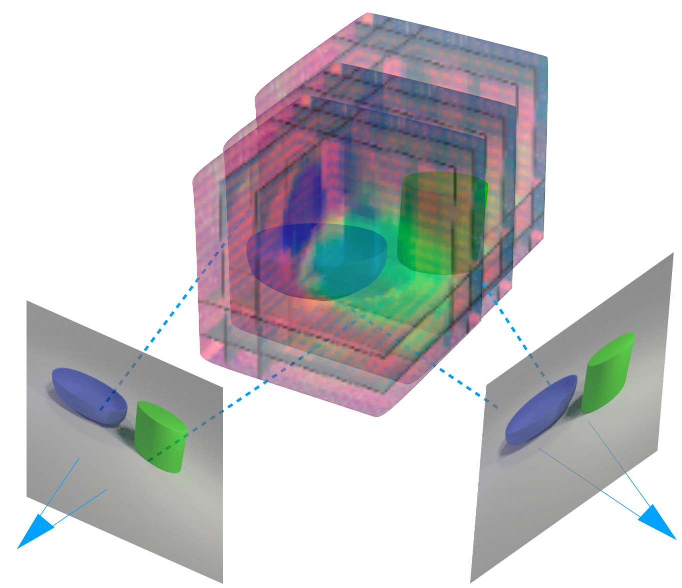
1. We consider an embodied agent that can see a scene from multiple viewpoints

"The green rubber cylinder is on the right of the blue bowl"



1. We consider an embodied agent that can see a scene from multiple viewpoints

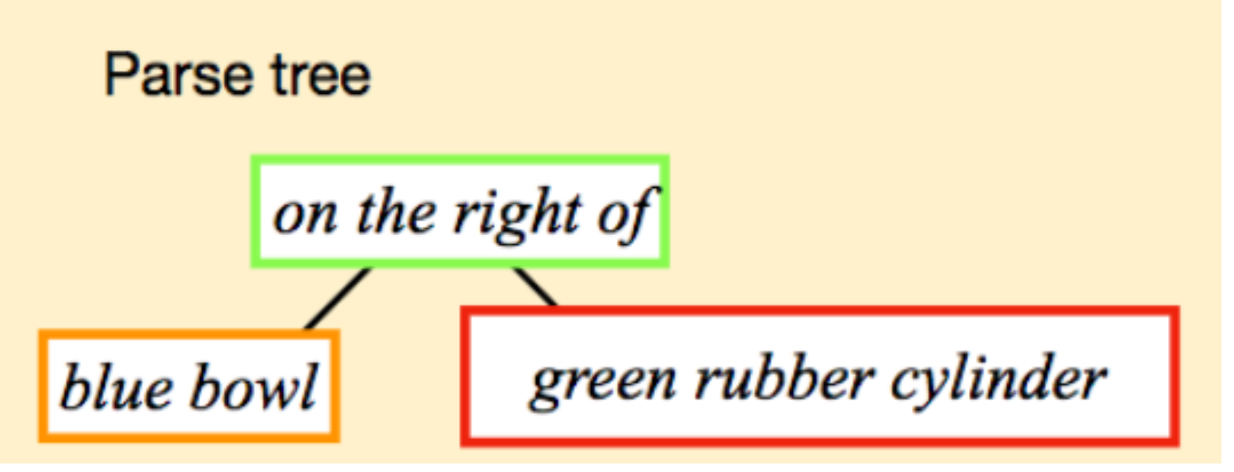
“The green rubber cylinder is on the right of the blue bowl”



2. Our agent learns to map an RGB image to a set of 3D feature maps by training GRNNs to predict views

“The green rubber cylinder is on the right of the blue bowl”

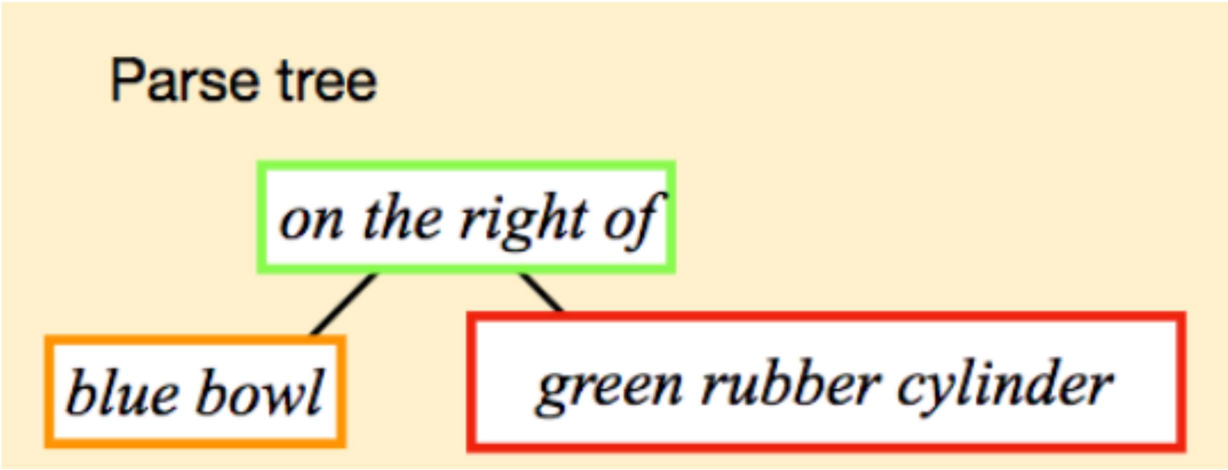
“The green rubber cylinder is on the right of the blue bowl”



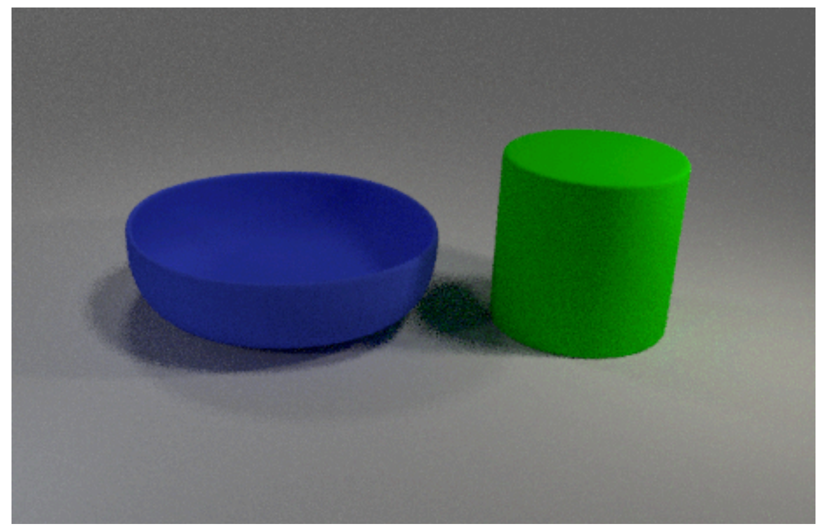
“The green rubber cylinder is on the right of the blue bowl”

Where:

What:

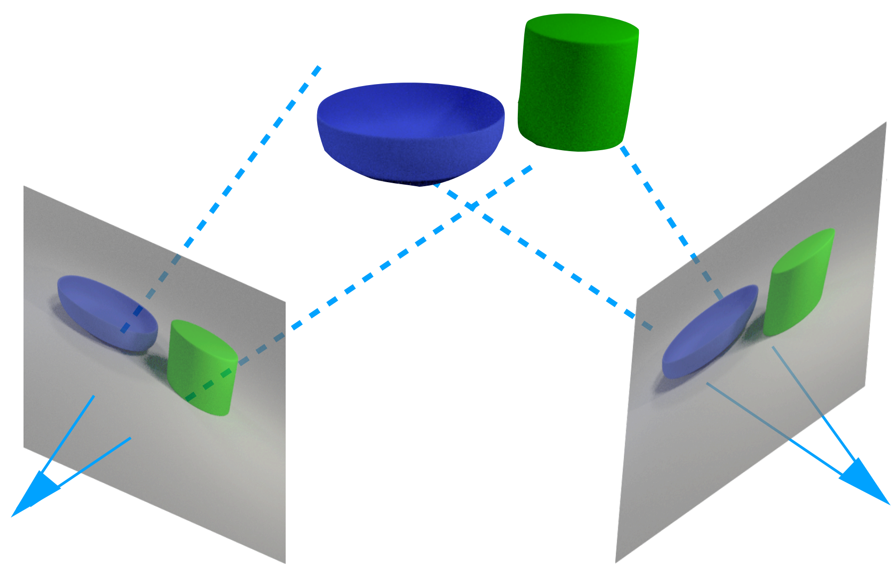


“The green rubber cylinder is on the right of the blue bowl”



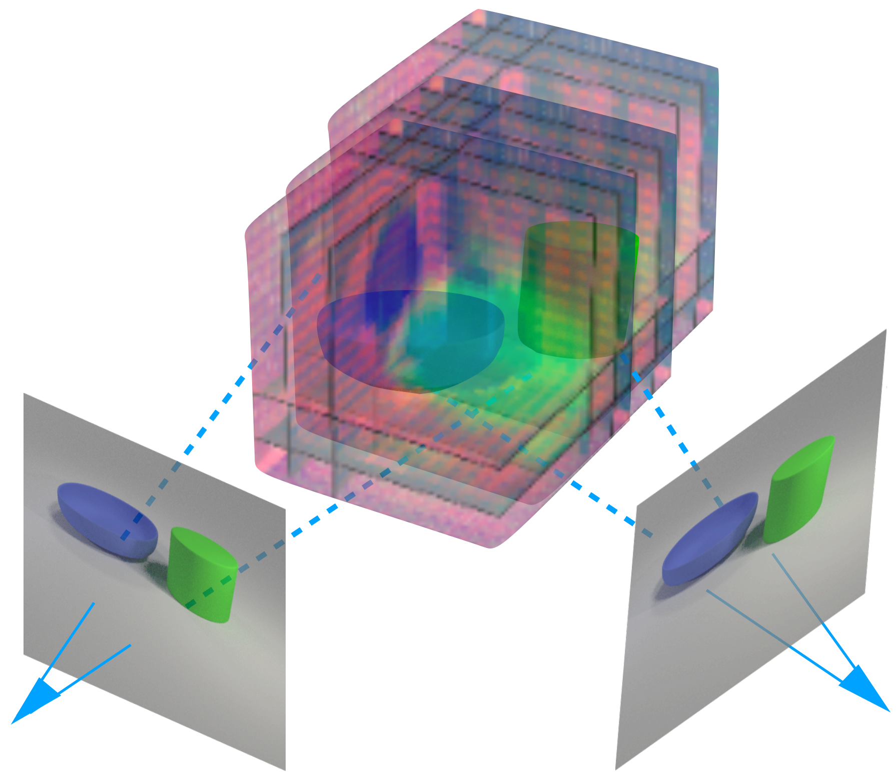
1. We consider an embodied agent that can see a scene from multiple viewpoints

“The green rubber cylinder is on the right of the blue bowl”



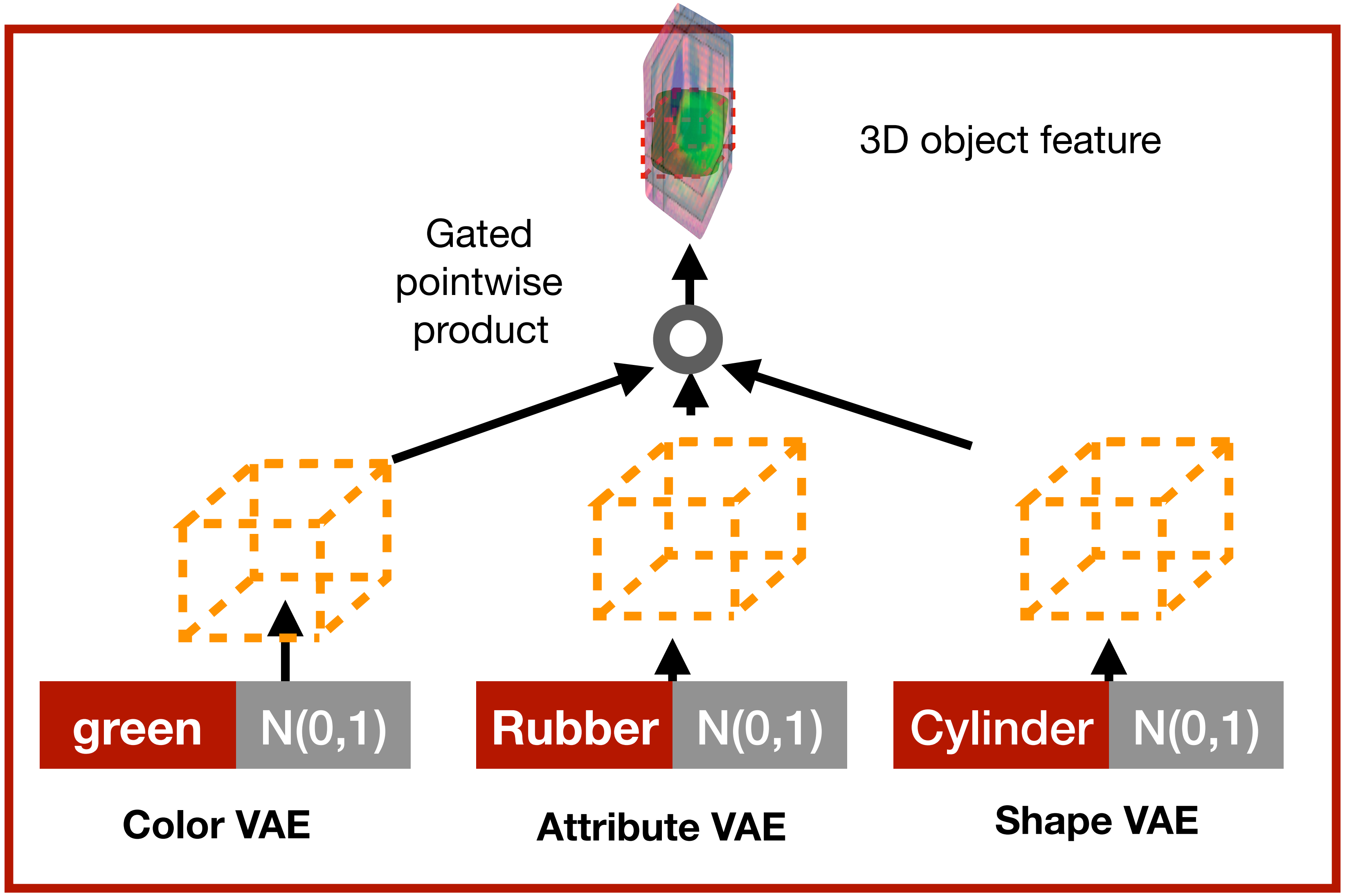
1. We consider an embodied agent that can see a scene from multiple viewpoints

"The green rubber cylinder is on the right of the blue bowl"



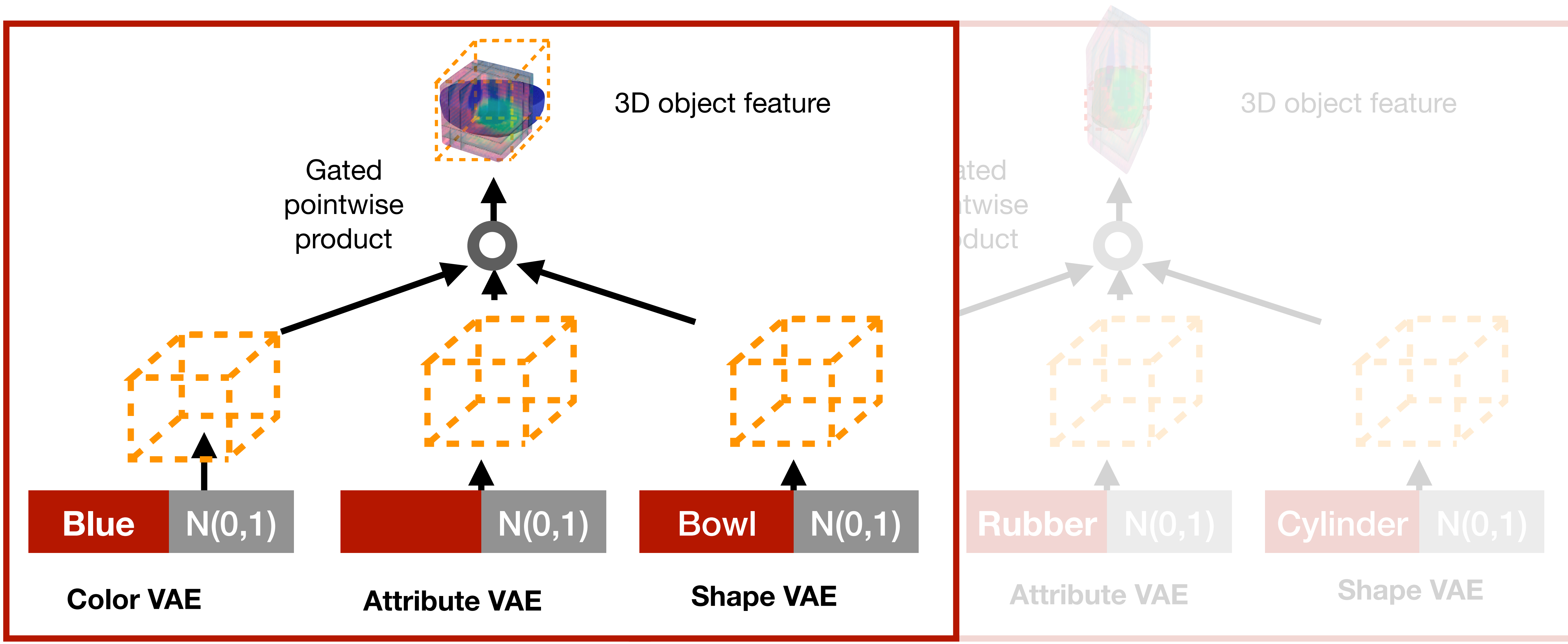
2. Our agent learns to map an RGB image to a set of 3D feature maps by training GRNNs to predict views

*“The **green rubber cylinder** is on the right of the blue bowl”*



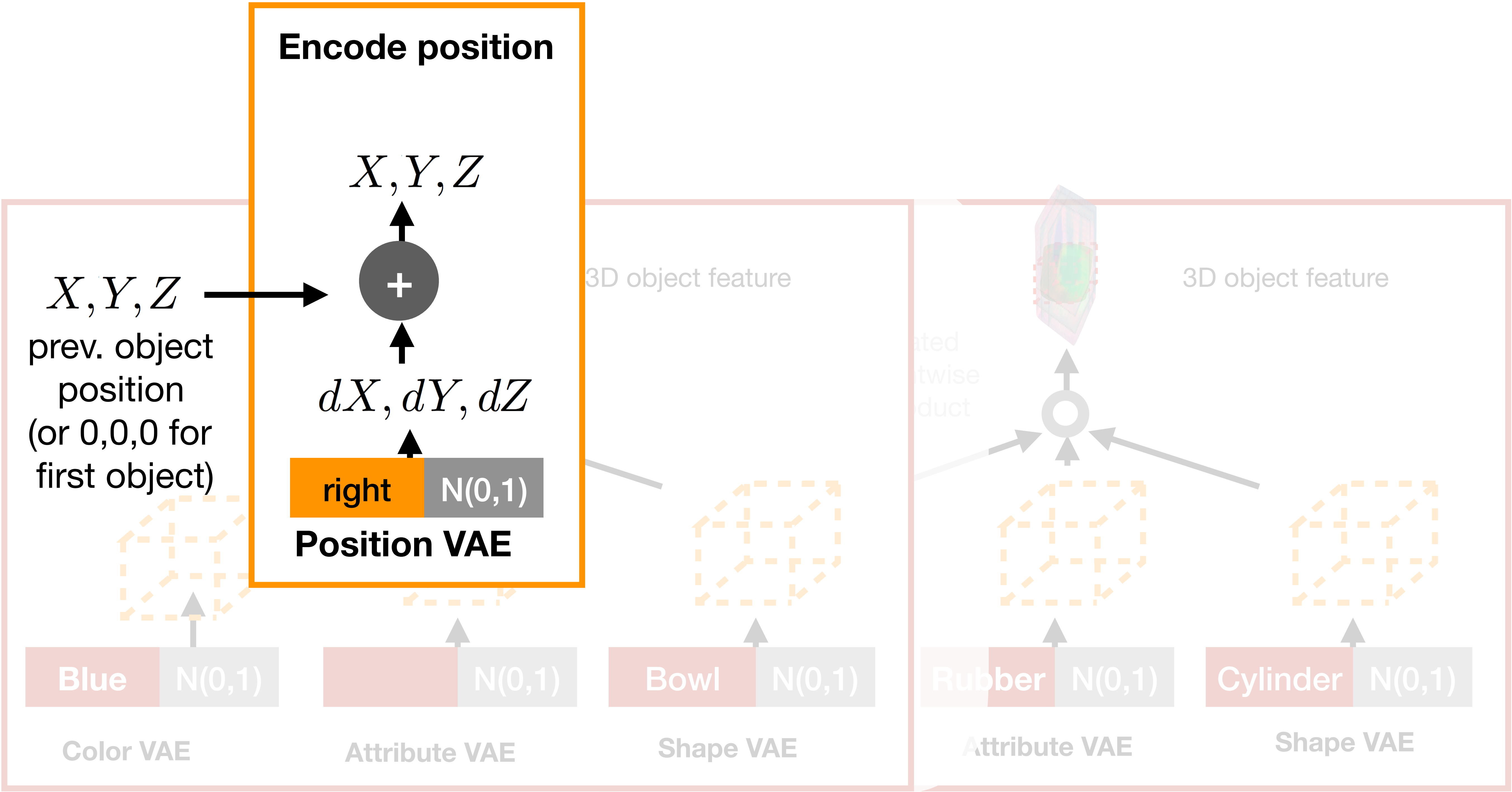
3. Our agent maps noun phrases to object-centric 3D feature maps

*“The green rubber cylinder is on the right of the **blue bowl**”*



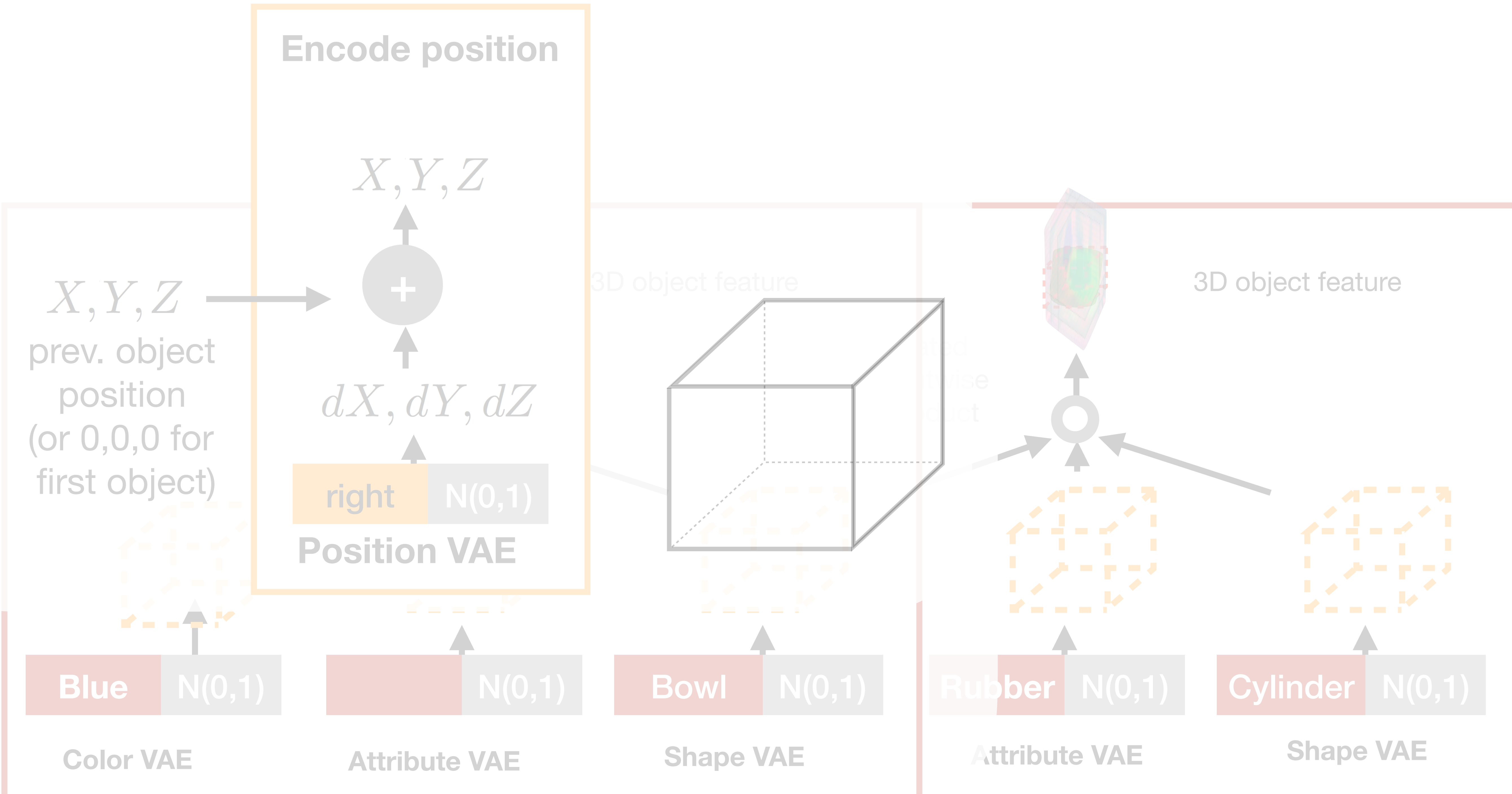
3. Our agent maps noun phrases to object-centric 3D feature maps

“The green rubber cylinder is *on the right of* the blue bowl”



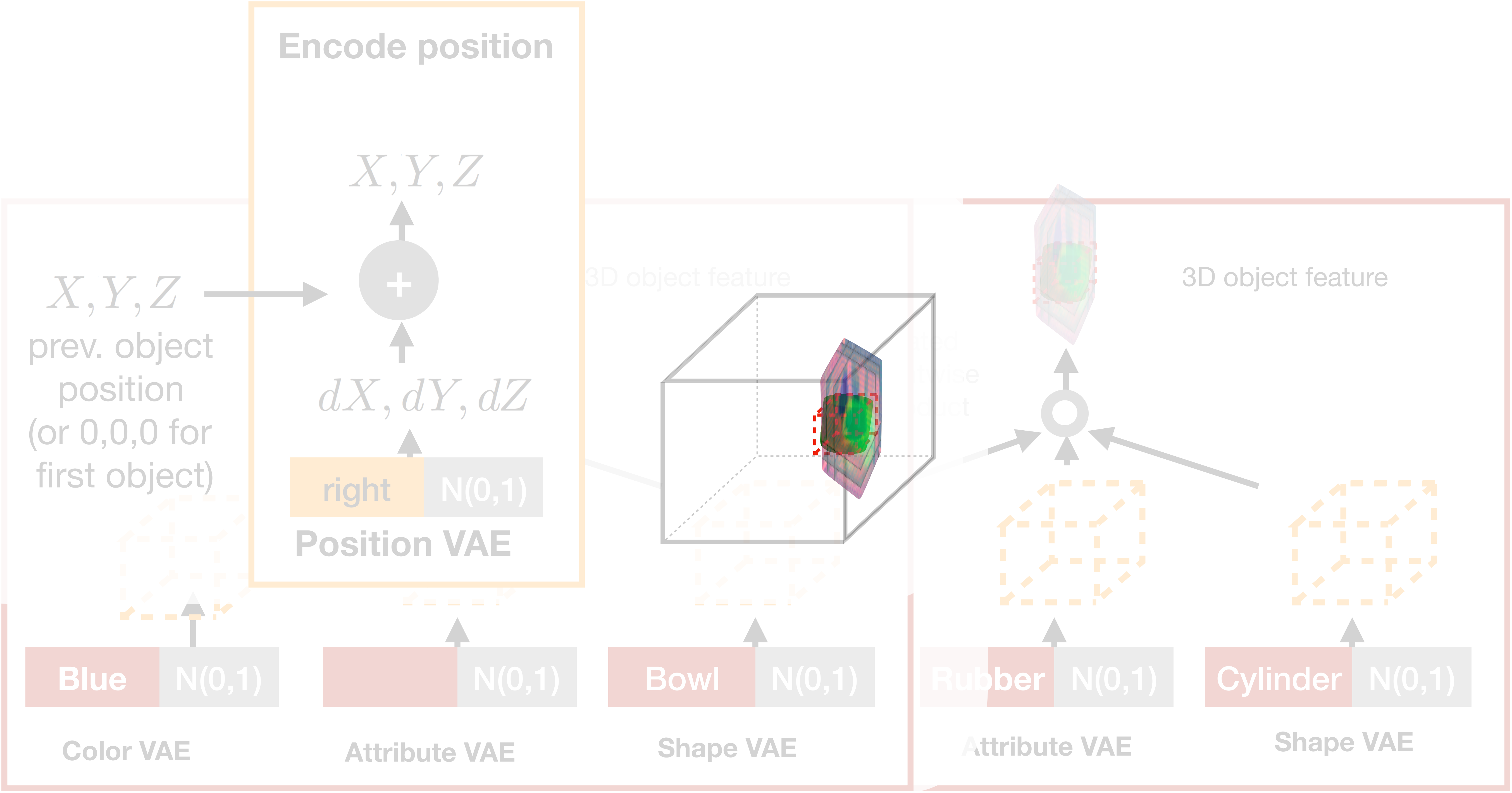
4. Our agent maps spatial expressions to relative 3D offsets

“The green rubber cylinder is on the right of the blue bowl”



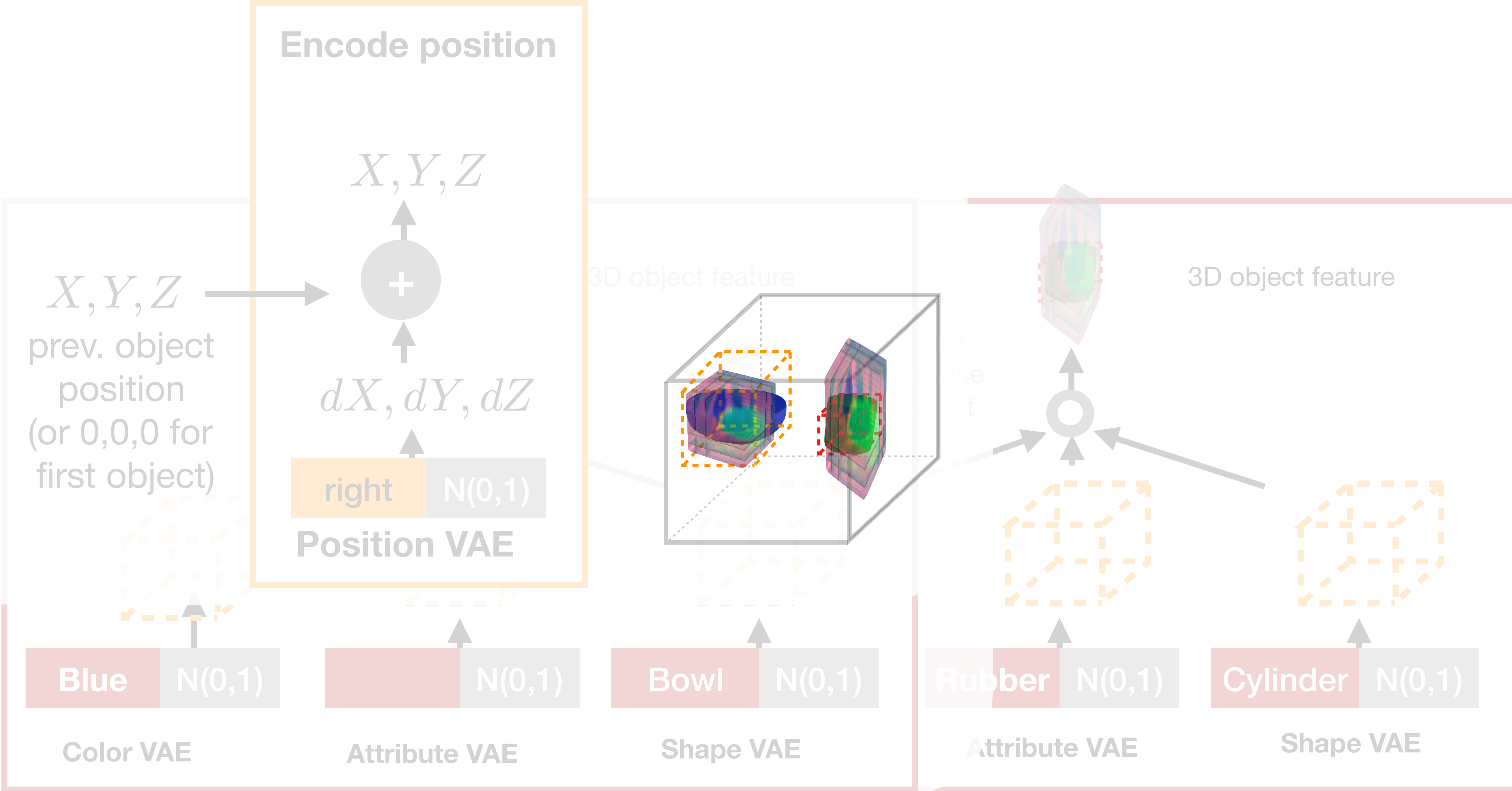
5. Our agent populates a 3D canvas with the predicted object tensors adn their relative offsets

"The green rubber cylinder is on the right of the blue bowl"



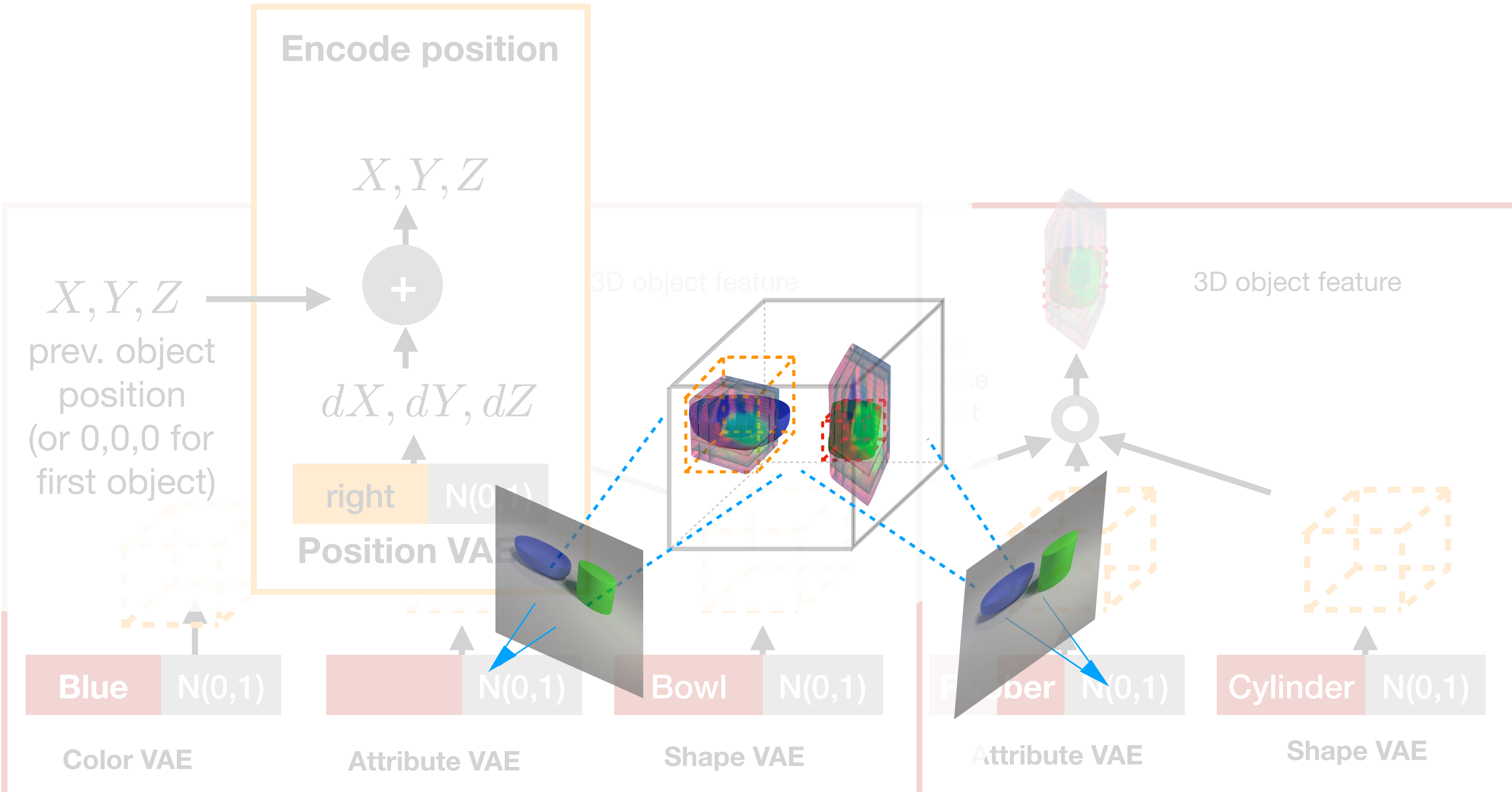
5. Our agent populates a 3D canvas with the predicted object tensors adn their relative offsets

“The green rubber cylinder is on the right of the blue bowl”



5. Our agent populates a 3D canvas with the predicted object tensors adn their relative offsets

“The green rubber cylinder is on the right of the blue bowl”

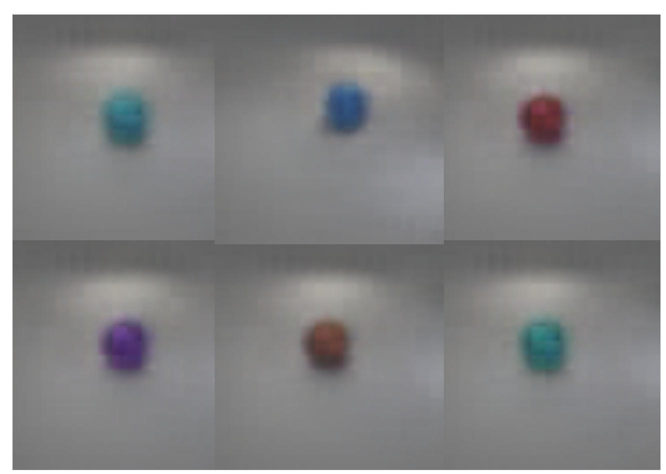


6. The generated canvas when projected should match the RGB image views

Multimodality in appearance

cylinder

sphere



Multimodality in spatial arrangements

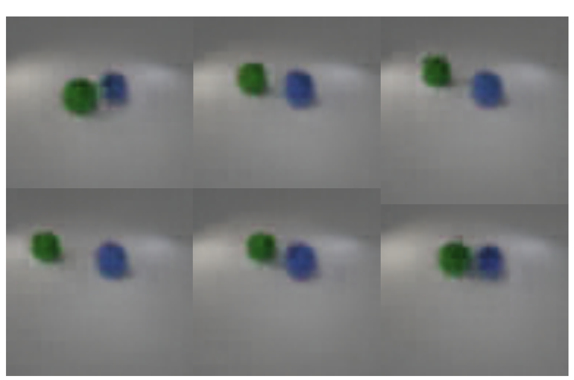
“red sphere **front left** of blue cylinder”

View Angle 1

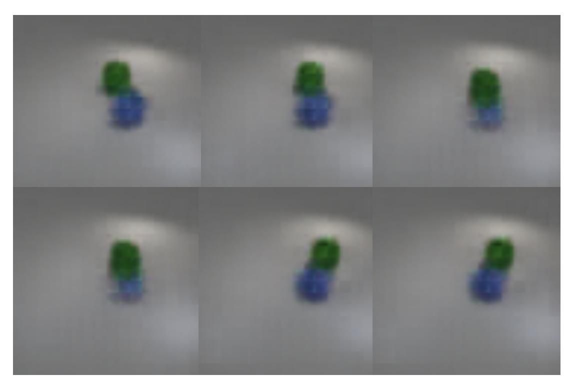
View Angle 2

View Angle 3

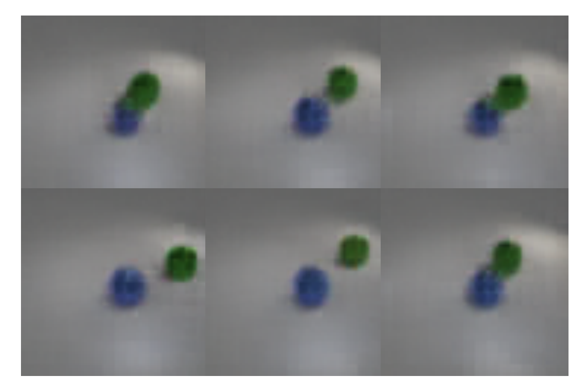
“green sphere t**o the left behind** of blue sphere”



View Angle 1



View Angle 2



View Angle 3

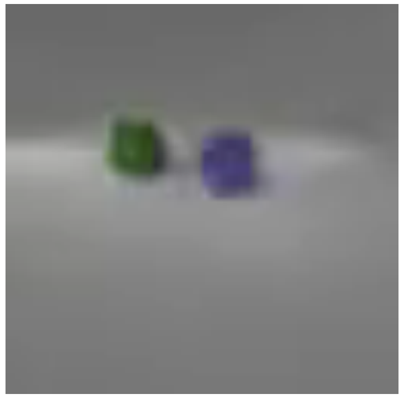
Scene imagination

“cyan sphere to the left of red cube”

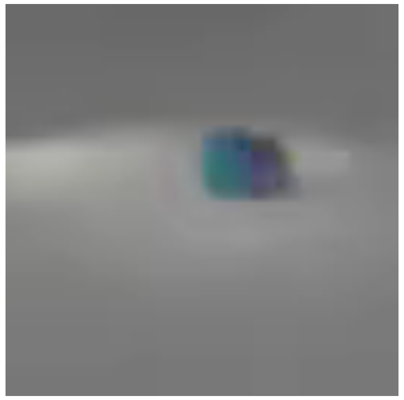
“red cylinder to the front of red sphere to the left-front of blue sphere”

“cyan cylinder to the left of red sphere to the front of green sphere”

“blue sphere to the left front of green cube”



“cyan cylinder to the front of yellow cube”



“cyan cylinder to the left front of yellow sphere to the behind of green sphere to the front of blue sphere to the front of gray cylinder to the behind of red sphere”

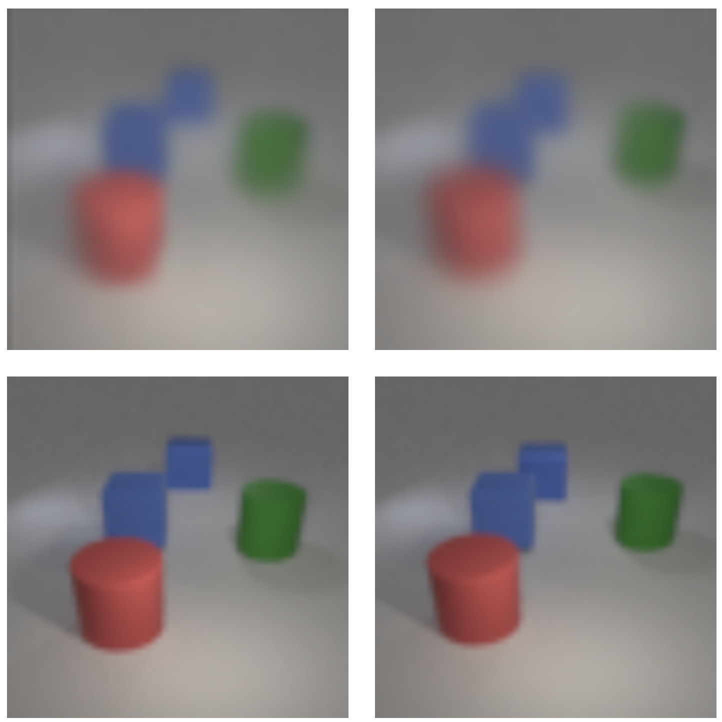
Scene imagination

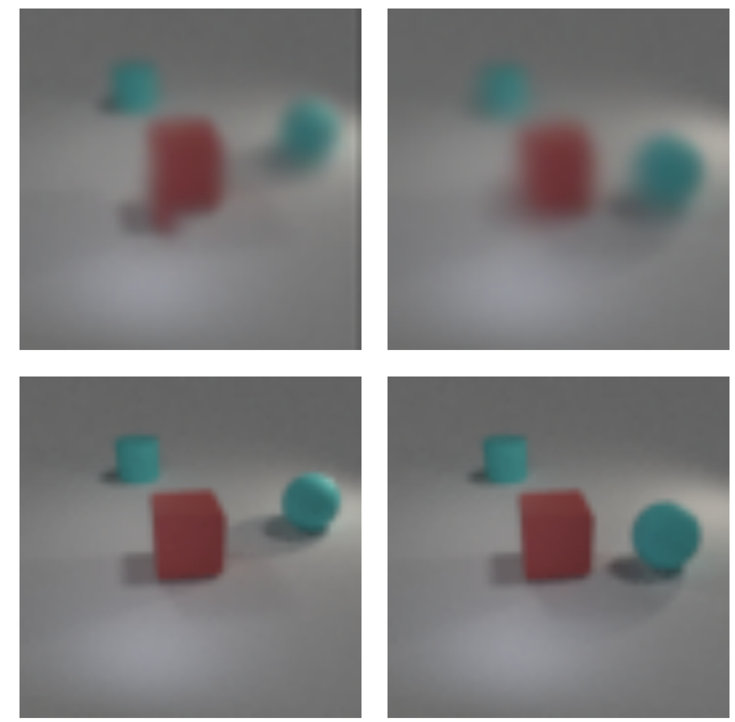
“Red Rubber Cylinder to the left front of Blue Rubber Cube to the left front of Green Rubber Cylinder to right front of Blue Rubber Cube”

“Red Rubber Cube to the left front of the Blue Rubber Sphere to the right front of Cyan Metal Cylinder”

Neural rendering

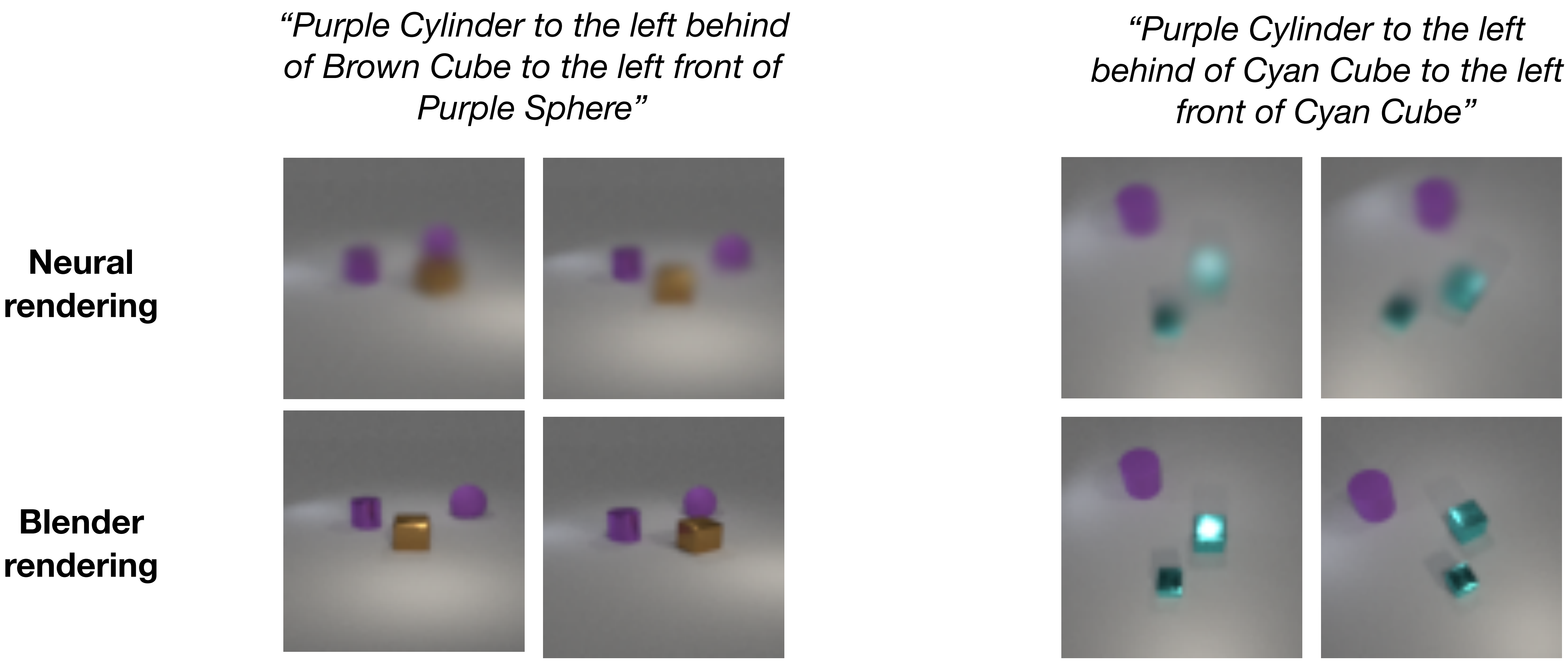
Blender rendering





- **Neural rendering**: project the 3D feature maps using our learned project+RGB decoder neural module
- **Blender rendering**: use the object-centric 3D feature maps to retrieve nearest 3D mesh neighbors from a training set, then arrange the retrieved meshes based on predicted 3D spatial offsets

Scene imagination



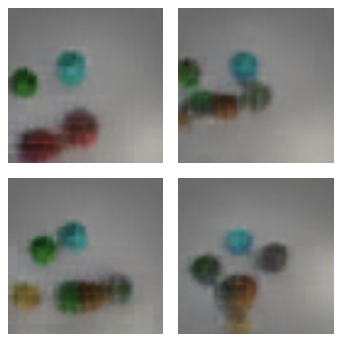
- **Neural rendering**: project the 3D feature maps using our learned project+RGB decoder neural module
- **Blender rendering**: use the object-centric 3D feature maps to retrieve nearest 3D mesh neighbors from a training set, then arrange the retrieved meshes based on predicted 3D spatial offsets

Grounding arbitrarily long utterances

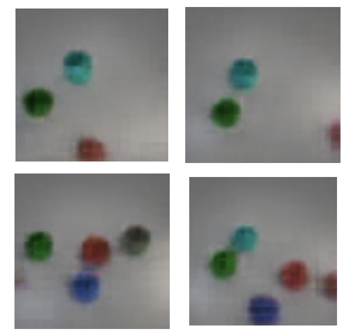
"yellow sphere to the left front of green sphere to the left behind of blue sphere to the left front of blue cylinder to the left behind of red cube to the left front of gray cube"

Top View

IOU > 0.1



Object Out of Camera View



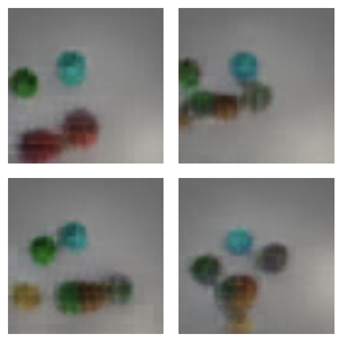
IOU= 0

Grounding arbitrarily long utterances

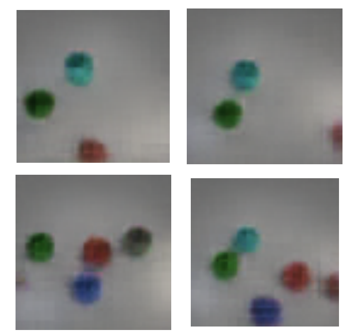
“yellow sphere to the left front of green sphere to the left behind of blue sphere to the left front of blue cylinder to the left behind of red cube to the left front of gray cube”

Top View

IOU > 0.1



Object Out of Camera View

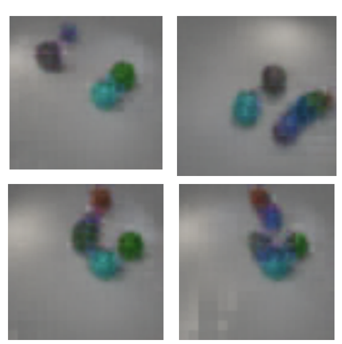


IOU= 0

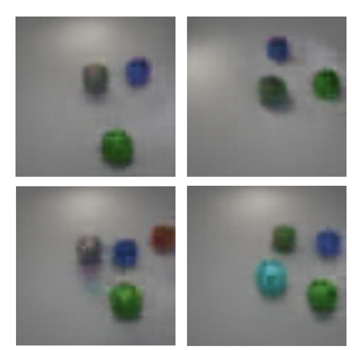
“gray sphere to the left front of blue sphere to the left front of red sphere to the left behind of cyan sphere to the left behind of green sphere”

Top View

IOU > 0.1



Object Out of Camera View

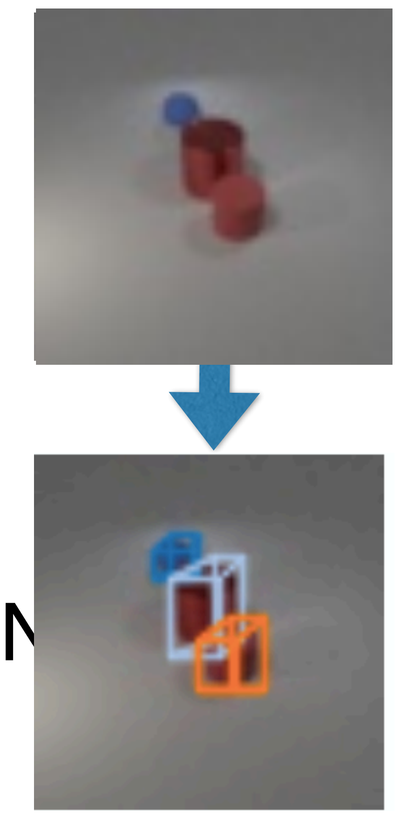


IOU= 0

3D referential object detection

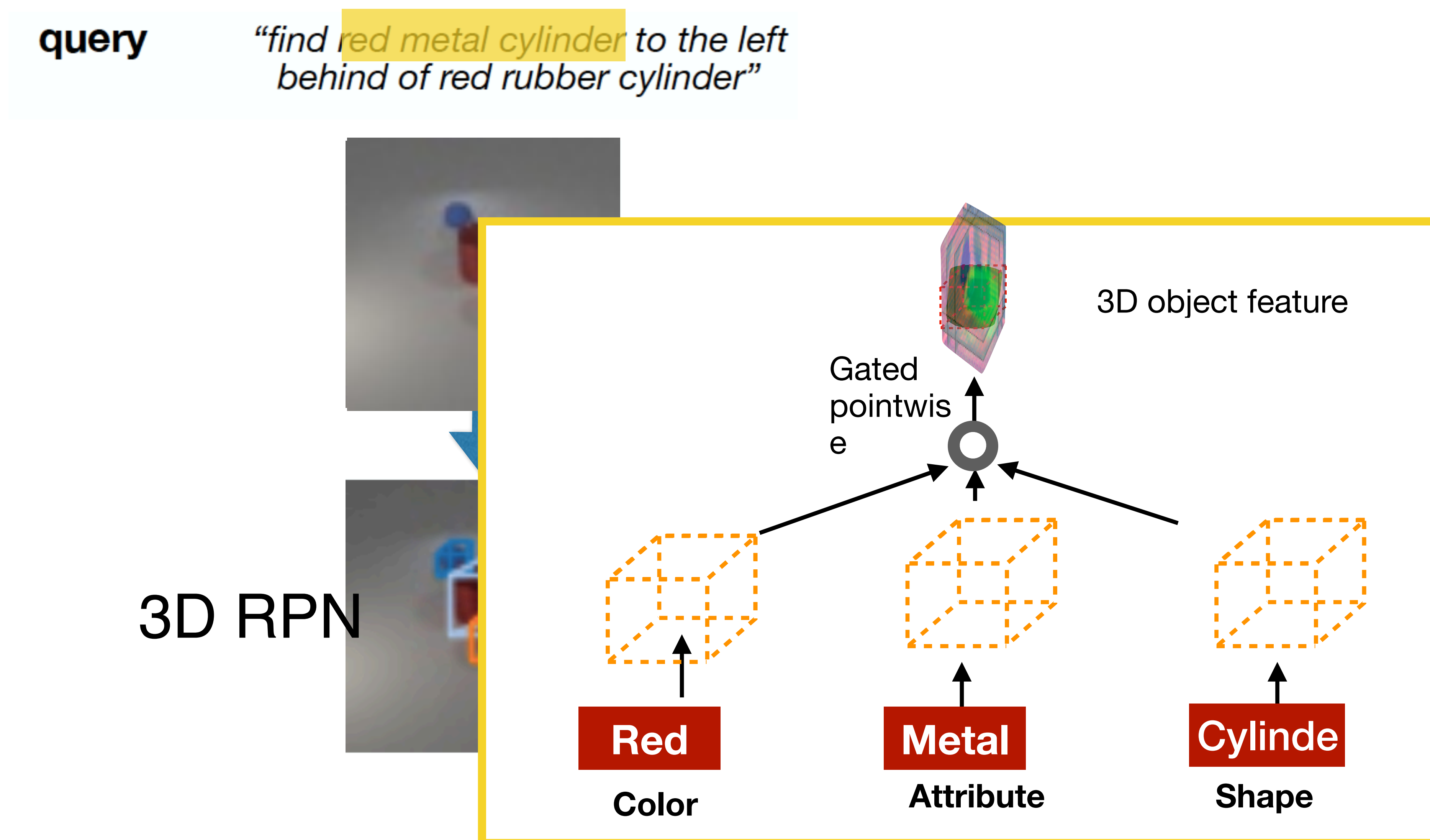
3D referential object detection

query *"find red metal cylinder to the left behind of red rubber cylinder"*



3D RPN

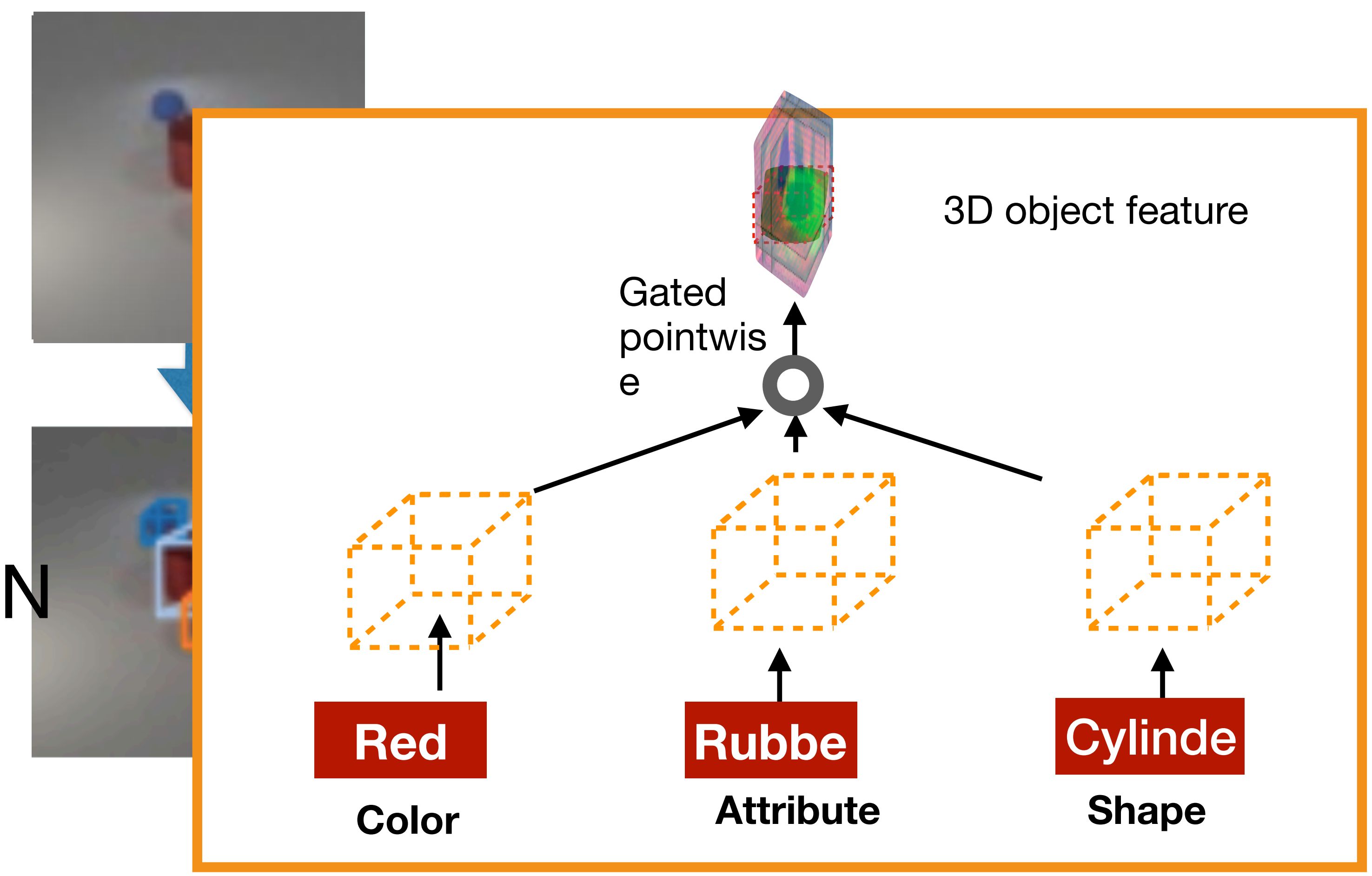
3D referential object detection



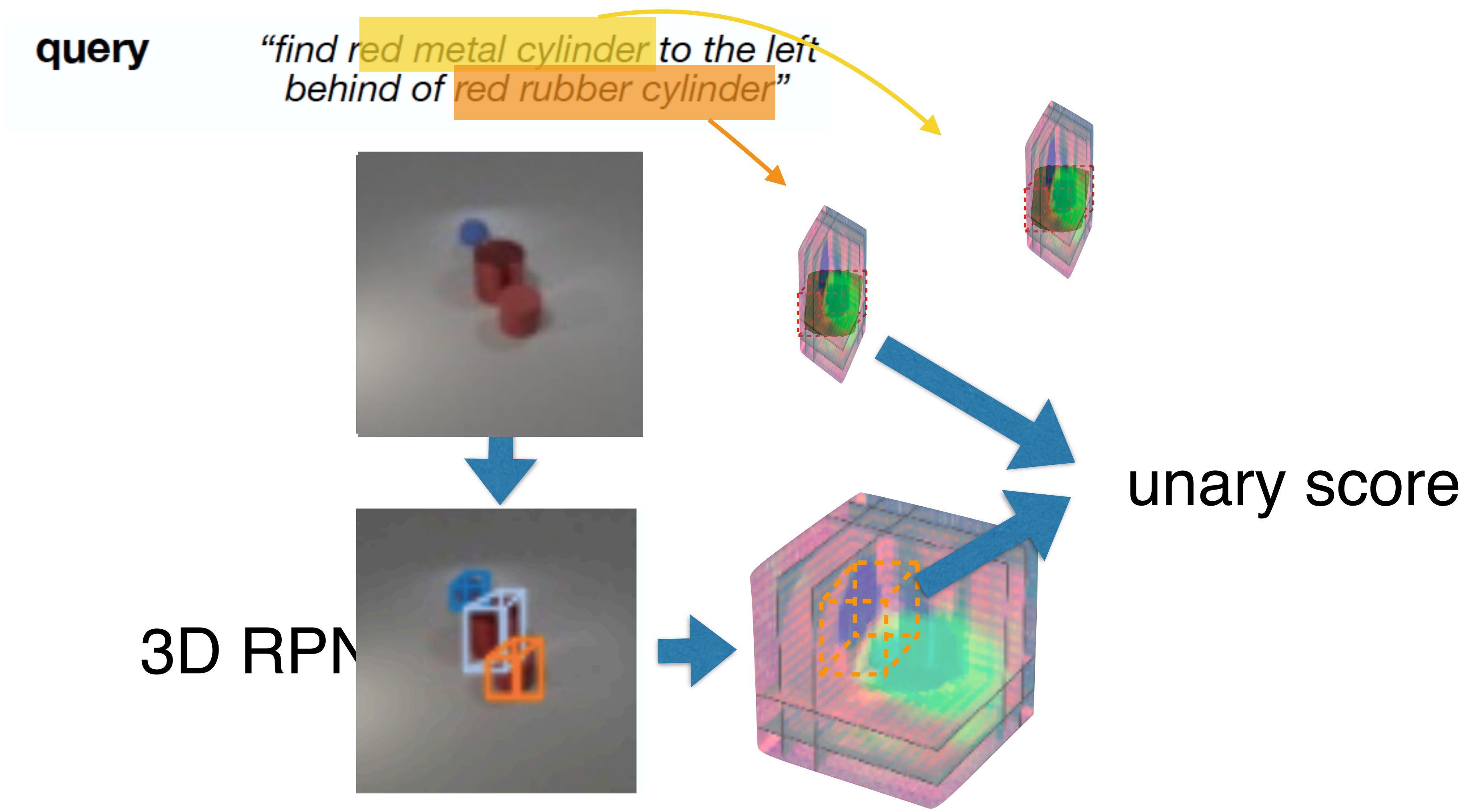
3D referential object detection

query *"find red metal cylinder to the left behind of red rubber cylinder"*

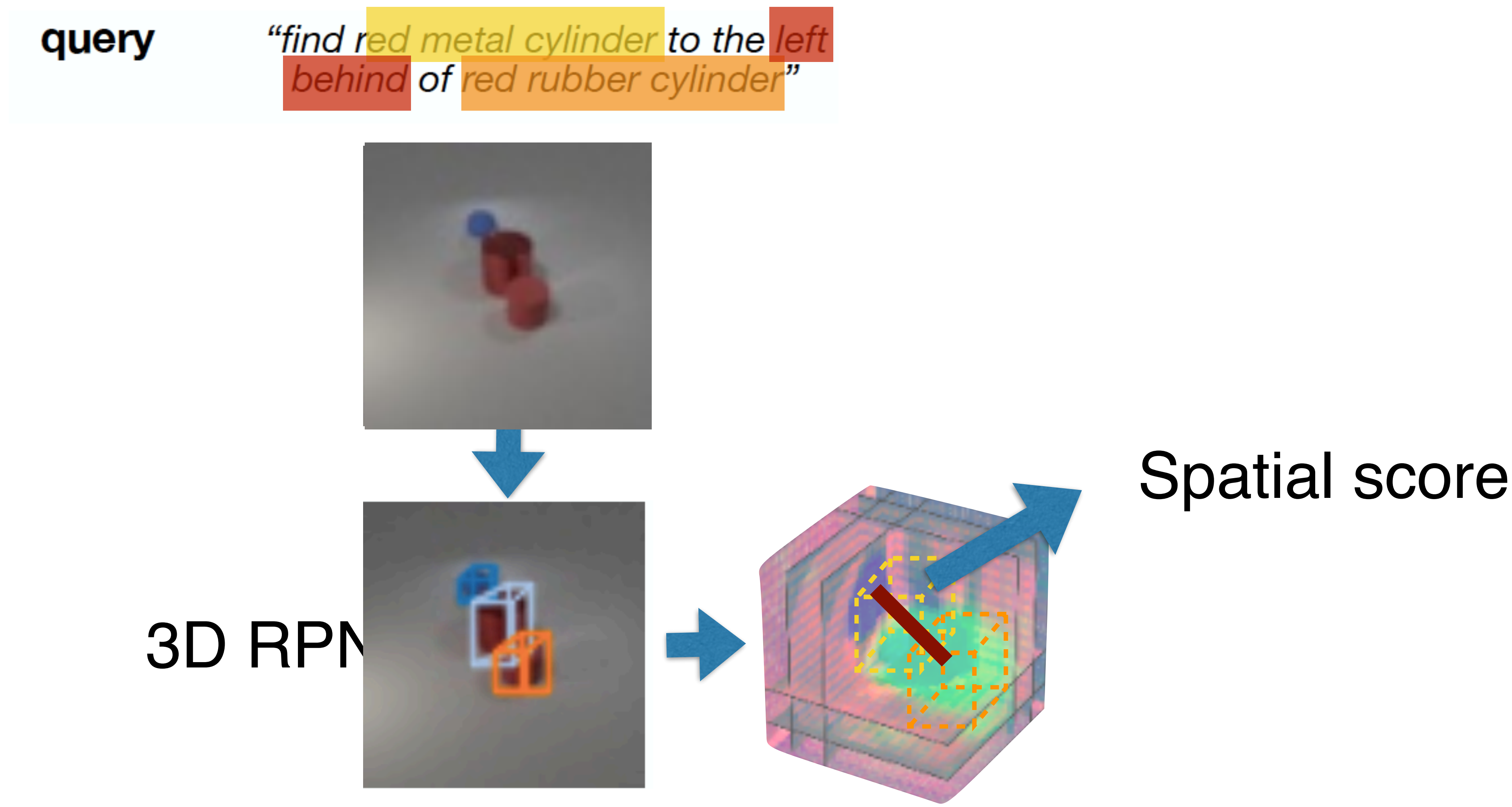
3D RPN



3D referential object detection



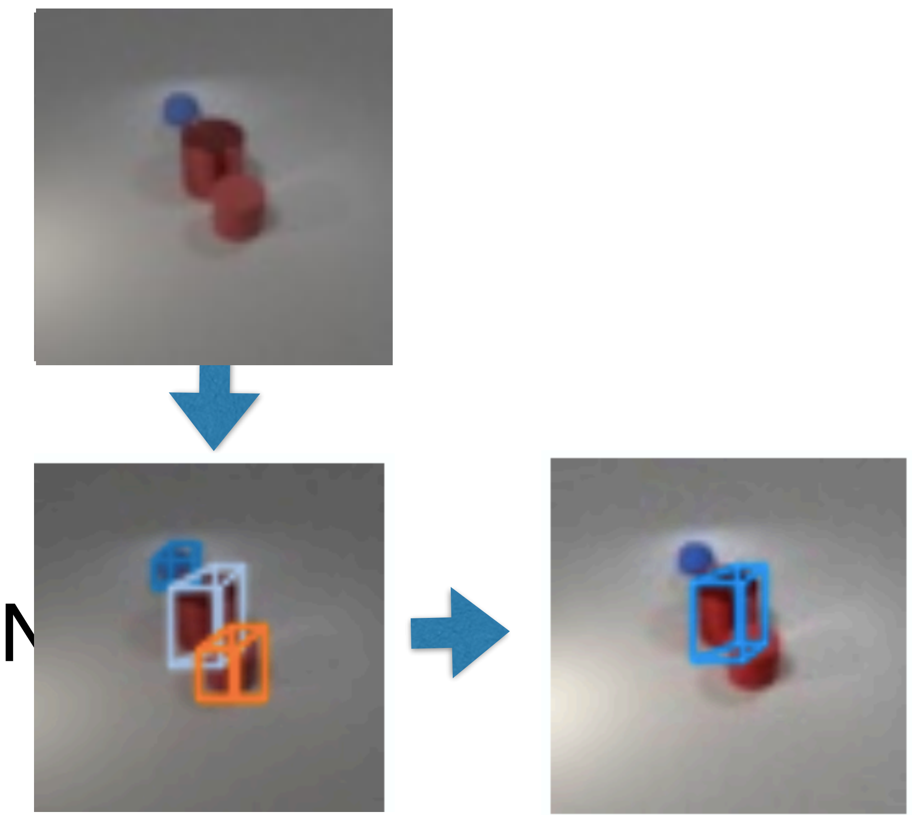
3D referential object detection



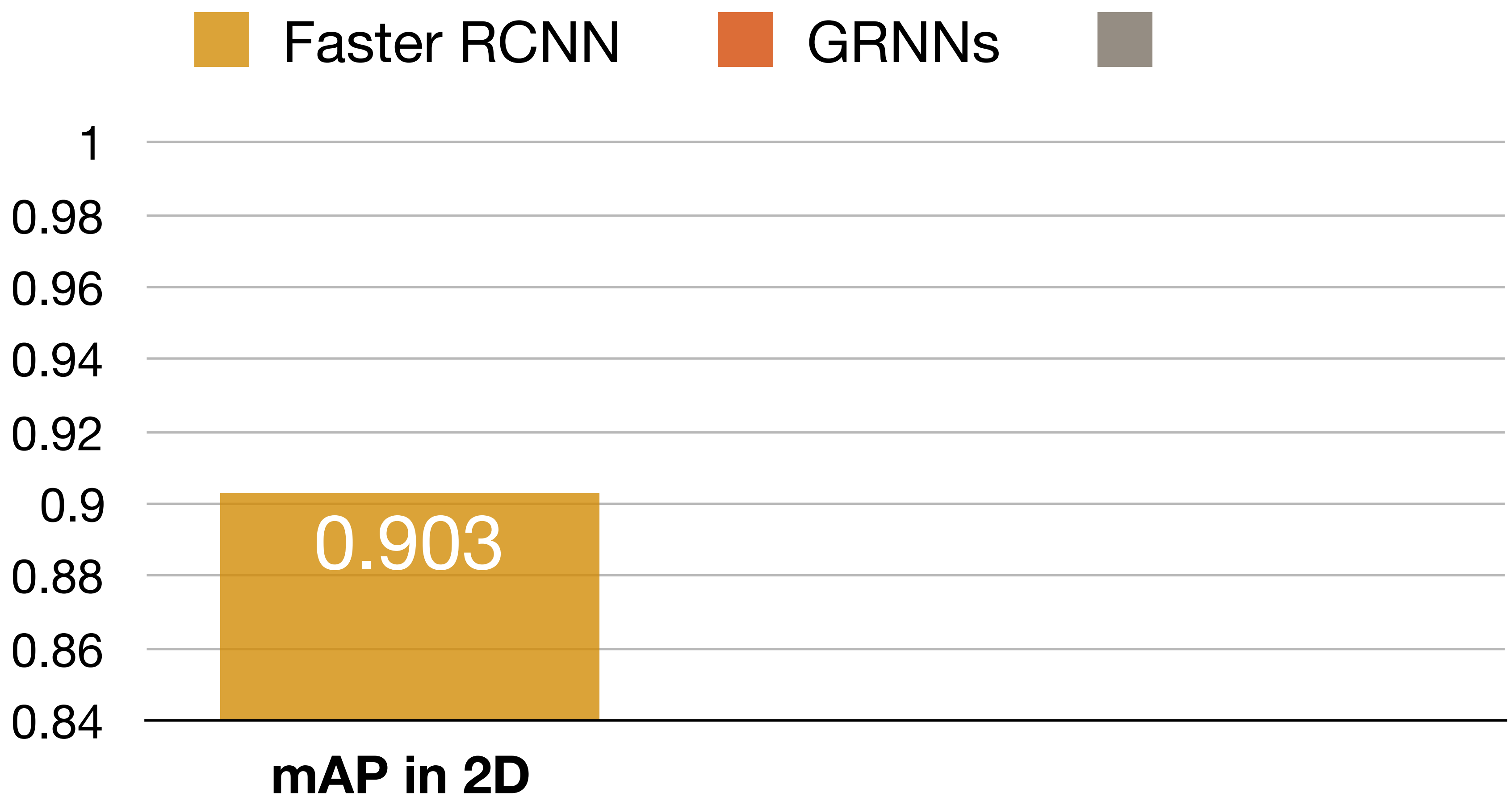
3D referential object detection

query *"find red metal cylinder to the left behind of red rubber cylinder"*

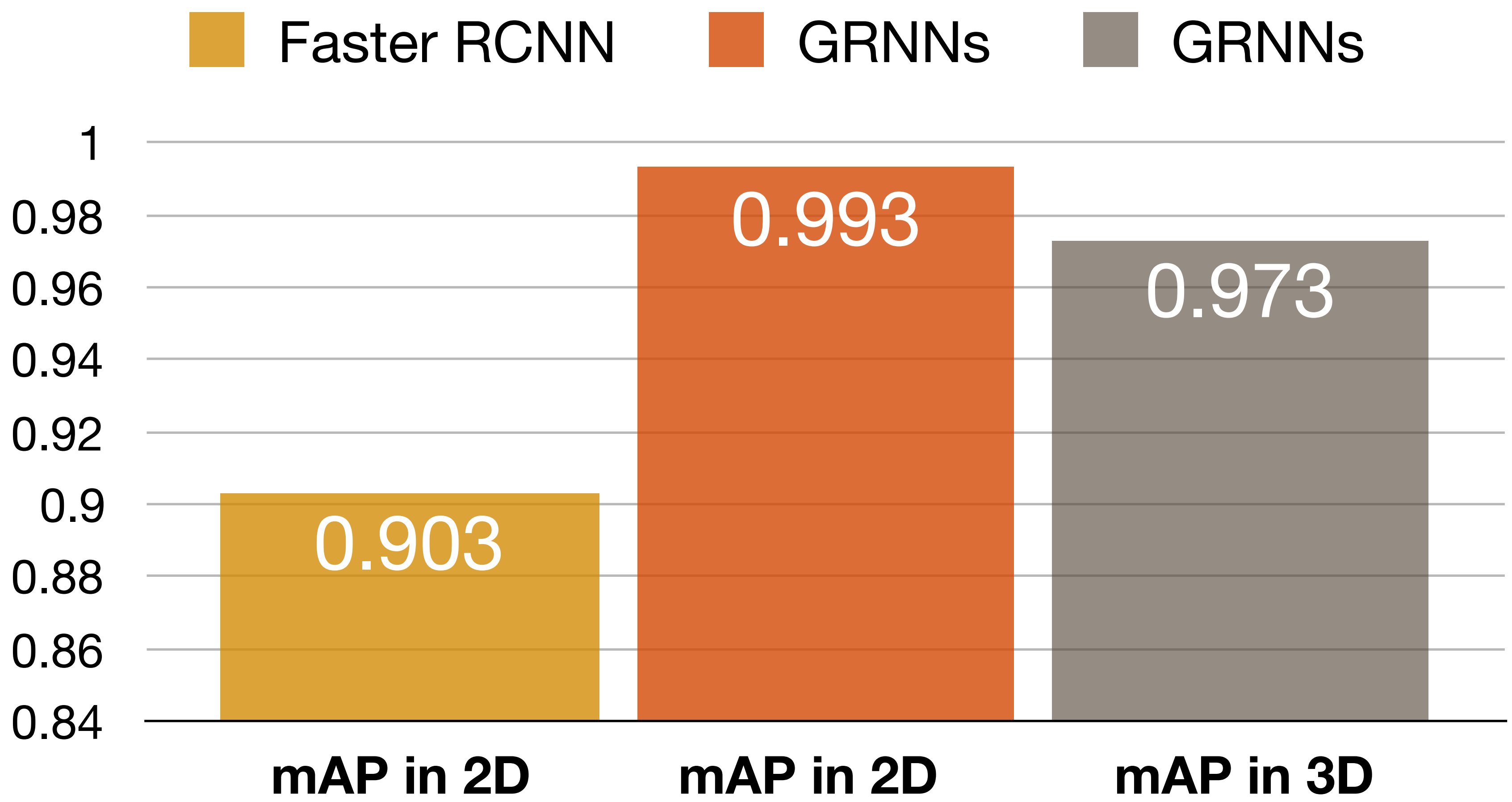
3D RPN



3D referential object detection

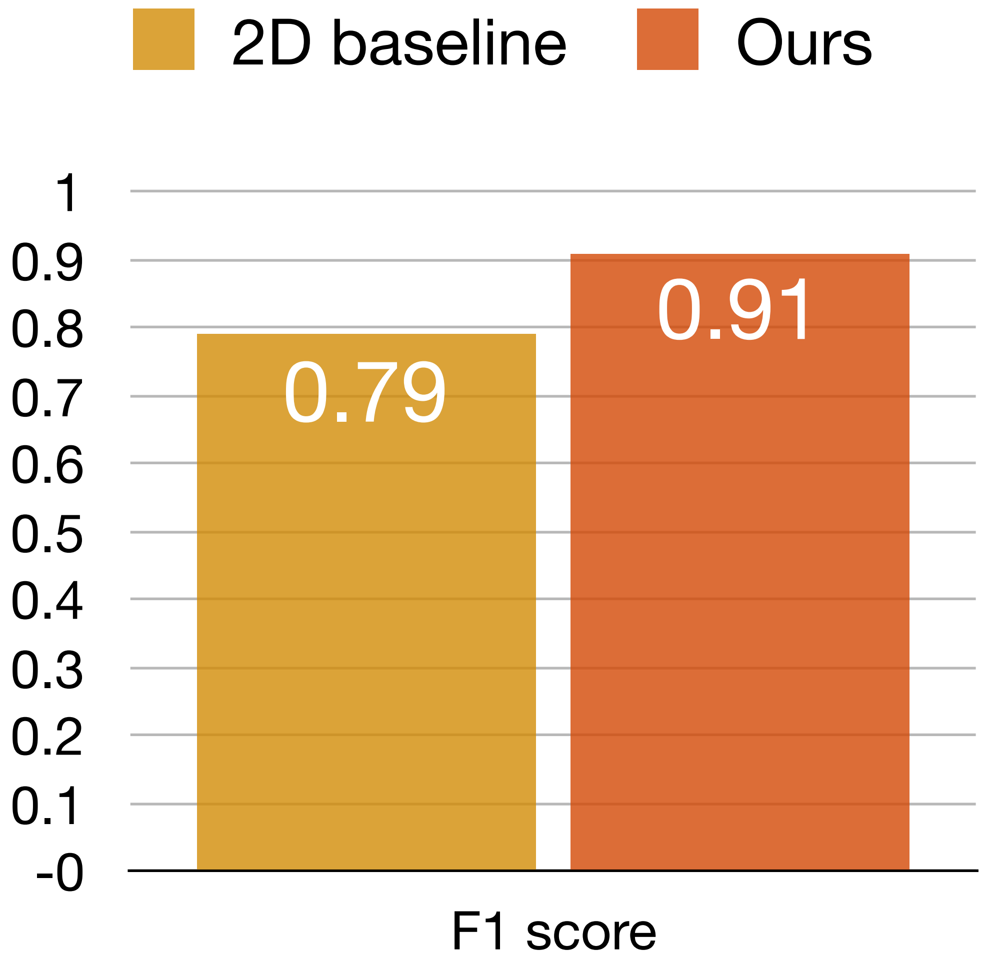


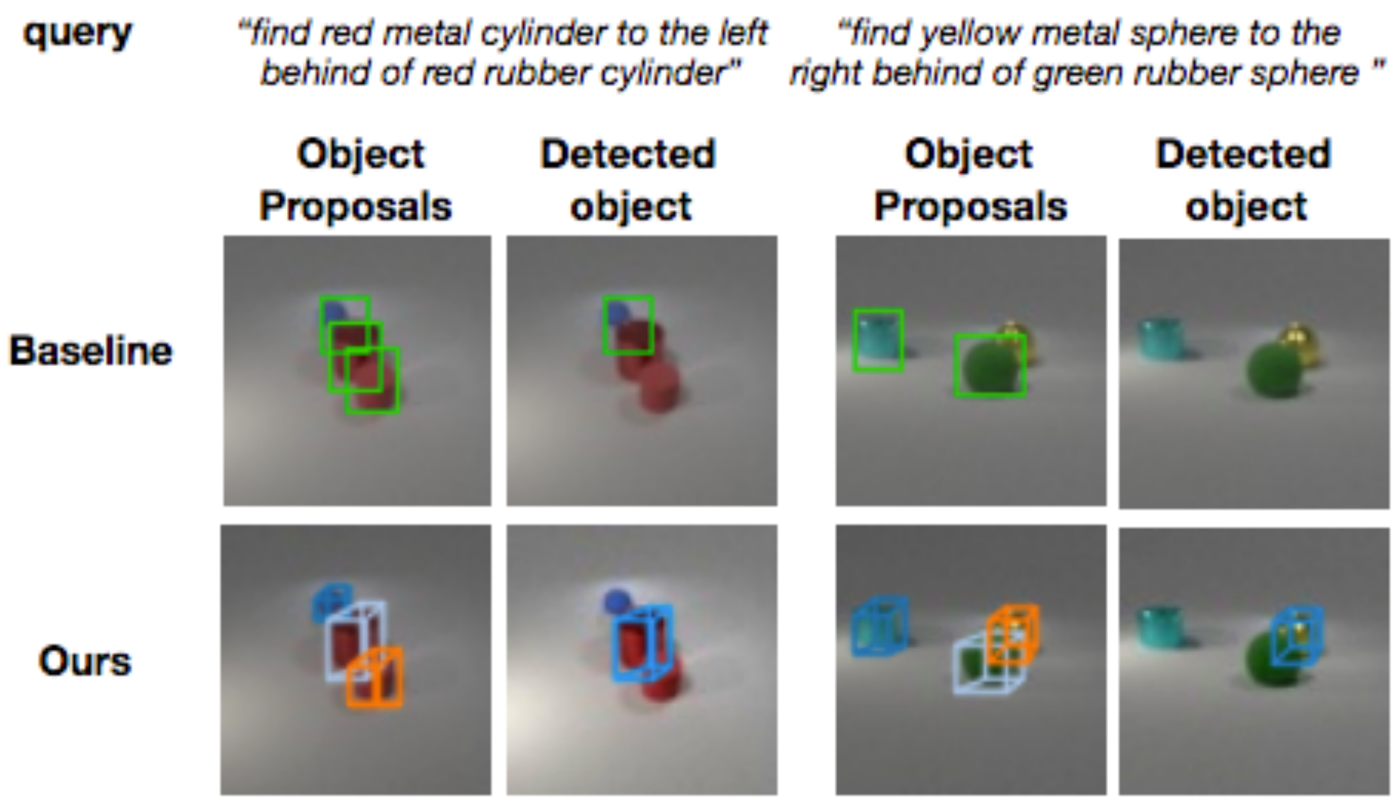
Object region proposals



3D referential object detection

F1 score for detecting spatial referential expression

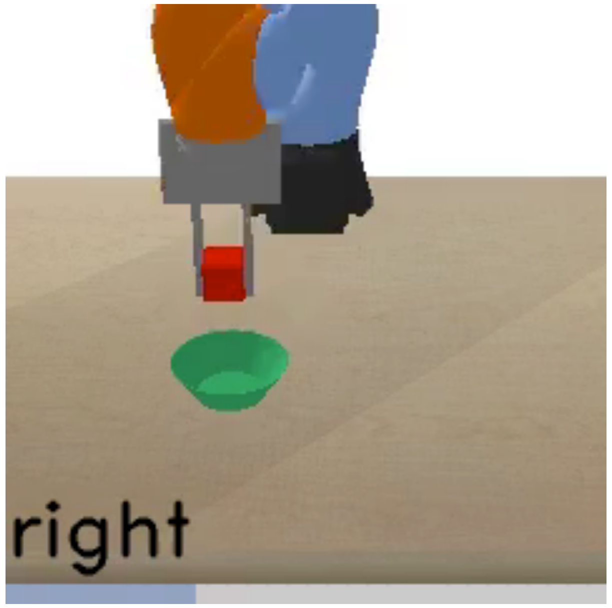




Ronghang Hu et al. Modeling relationships in referential expressions with compositional modular networks

Instruction Following

``put the cube on the right of the bowl"



1. Referential 3D object detection
2. Goal generation: Predict relative 3D desired location for the object
3. Use LQR with Euclidean distance of current to goal location as the cost.

Grounding Language on 3D visual feature representations

- Objects have regular sizes: object size is disentangled from the camera viewpoint
- Objects have 3D extent
- Objects do not interpenetrate in 3D: during iterative scene generation we can detect 3D intersection and continue sampling valid configurations
- Objects persist over time

Next steps

- Grounding action descriptions
- Use intuitive physics and dynamics beyond static spatial constraints

ML
MACHINE LEARNING
DEPARTMENT

THE ROBOTICS INSTITUTE

Thank you

Mihir Prabhudesai

Fish Tung

Syed Javed

Adam Harley

Max Sieb

- Embodied language grounding, Prabhudesai et al., arxiv
- Reward Learning from Narrated Demonstrations , Tung et al., CVPR 2018