Physical Scene Understanding with Compositional Structure

Jiajun Wu (MIT CSAIL)

GAMES Webinar, Jan 10, 2019

Advisors and Collaborators

Josh Tenenbaum

Bill Freeman

Antonio Torralba

Pushmeet Kohli

Tianfan Xue	Andrew Owens	Michael Janner	David Zheng	Josh McDermott
Ilker Yildirim	Xiuming Zhang	Anurag Ajay	Shaoxiong Wang	Leslie Kaelbling
Joseph Lim	Zhoutong Zhang	Nima Fazeli	Xingyuan Sun	Alberto Rodriguez
Jun-Yan Zhu	Wenzhen Yuan	Maria Bauza	Qiujia Li	Kevin Murphy
Katie Bouman	Erika Lu	James Traer	Zhengjia Huang	Ted Adelson
Chengkai Zhang	Zhijian Liu	Hongyi Zhang	Yifan Wang	Russ Tedrake
Yunzhu Li	Chen Sun	Yuanming Hu	Shunyu Yao	Daniela Rus
Harry Hsu	Yuandong Tian	Tejas Kulkarni	Andrew Luo	Wojciech Matusik
Jiancheng Liu	Andrew Spielberg	Kevin Ellis	Yonglong Tian	Daniel Ritchie

What can we learn from this video?

Human Physical Scene Understanding

What can we see in this video?

I. Scene structure (**perception**)

- Object appearance (geometry, texture)
- Physical properties (e.g., mass)

II. Interactions and events (**physics**)

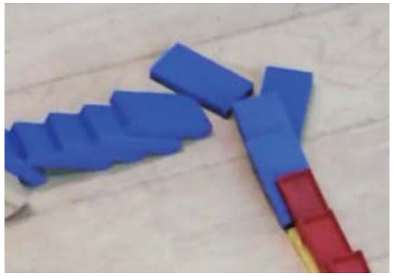
- Collision, rolling, etc.

III. Concepts and regularity (**reasoning**)

- Balls can roll, but not blocks
- Blocks are of the same size and shape
- Blocks are lined up in a row

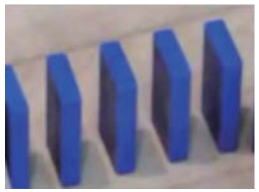
sphere
orange
plastic

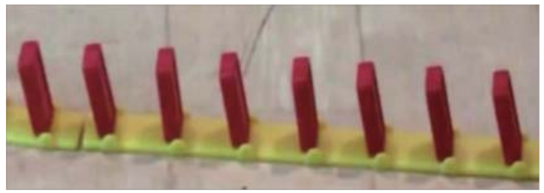
hedge
white/yellow
wooden



collisions

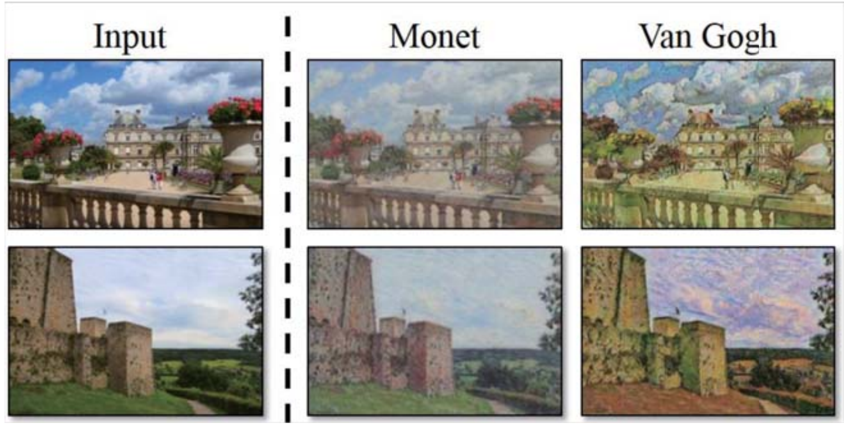
rolling





Current Machine Scene Understanding

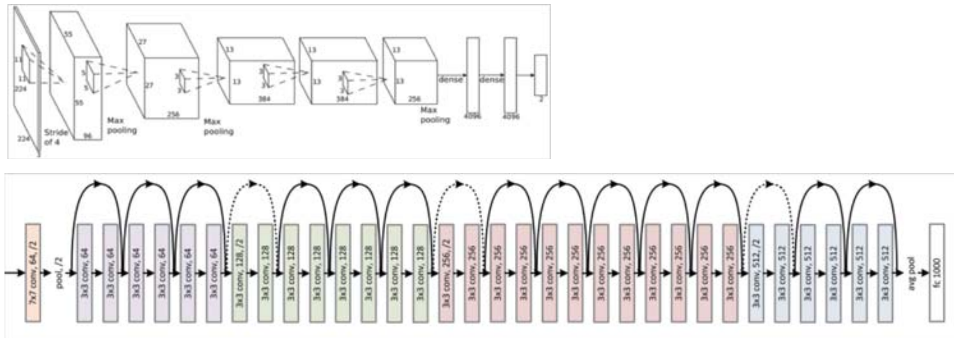
Recognition



Synthesis

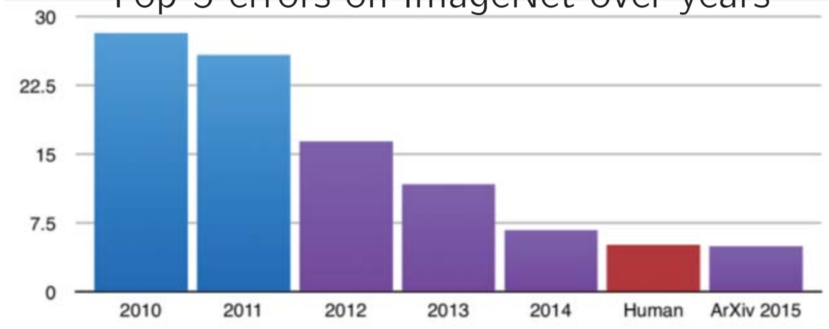
AlexNet

ResNet



Deep Learning Models

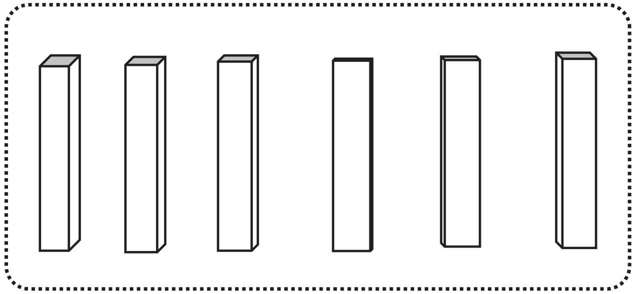
Top-5 errors on ImageNet over years



Performance

Image Credit: DeepLab, Chen et al., 2018; CycleGAN, Zhu et al., 2017

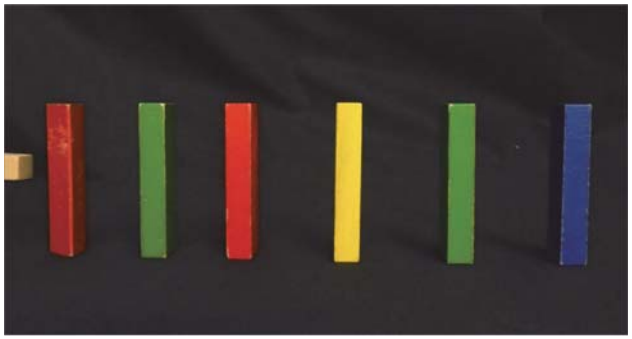
Modeling the Physical World



World State (t-1)

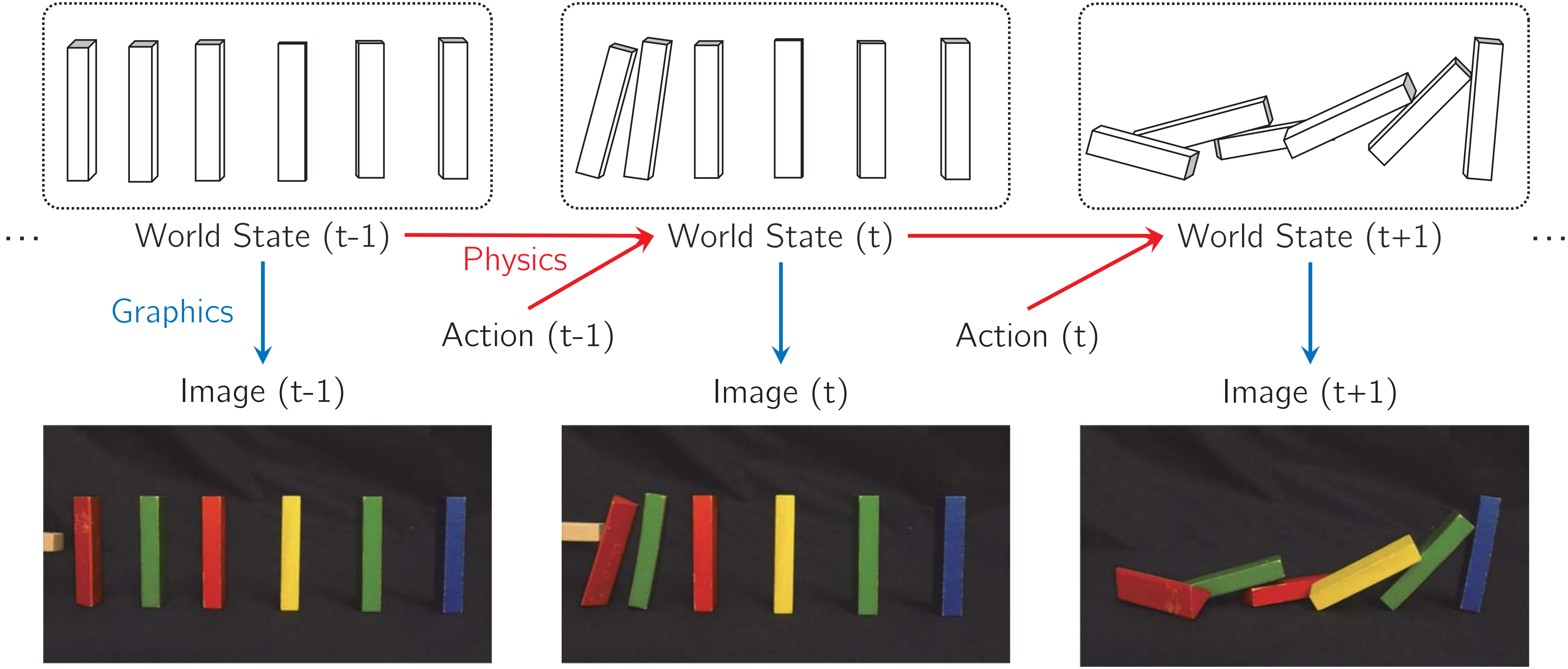
Graphics

Image (t-1)

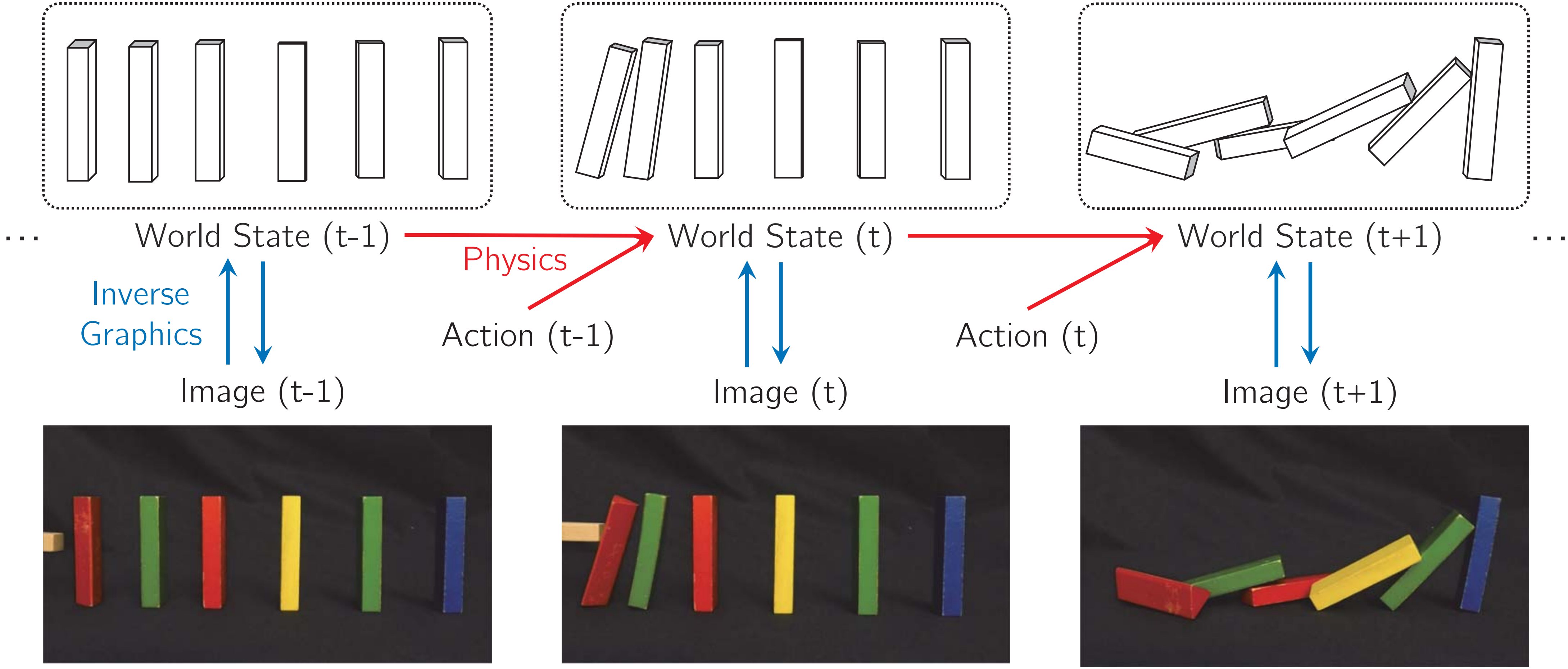


- Object Intrinsics
  - Geometry
  - Physical properties
- Object Extrinsics
  - Position
  - Velocity
- Scene Descriptions
  - Lighting
  - Camera parameters

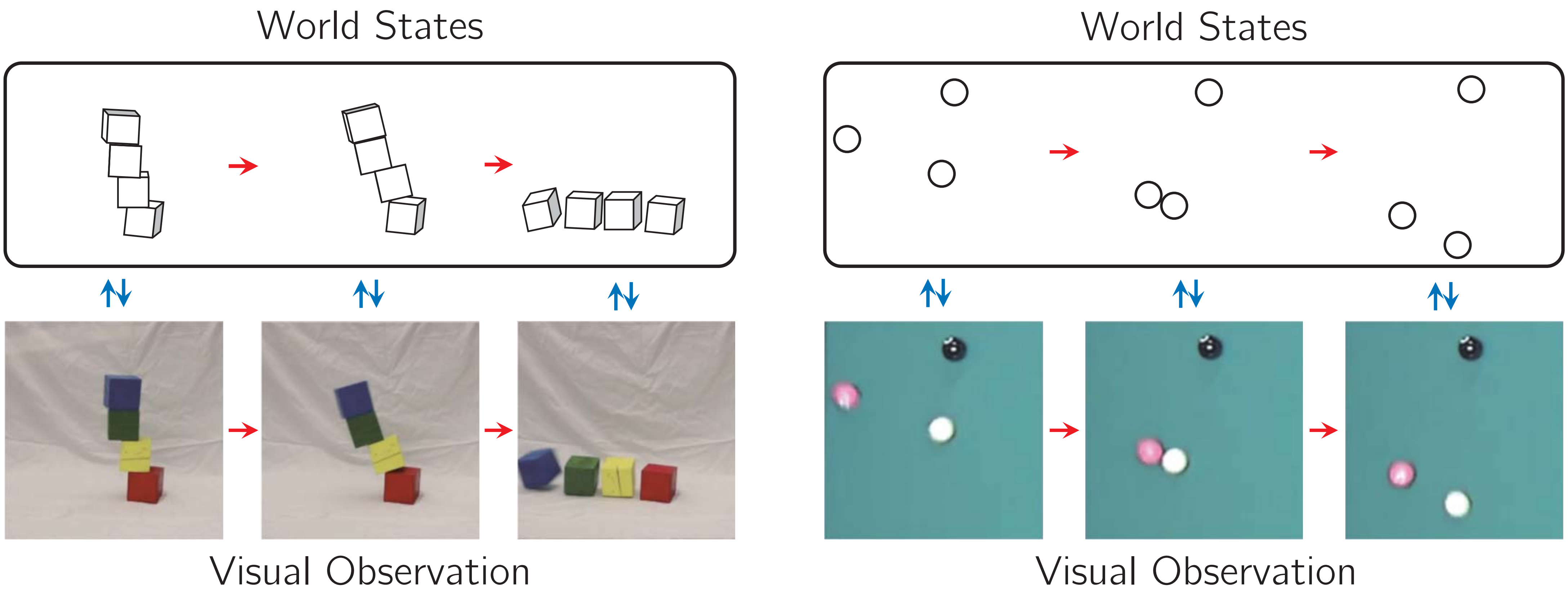
Modeling the Physical World



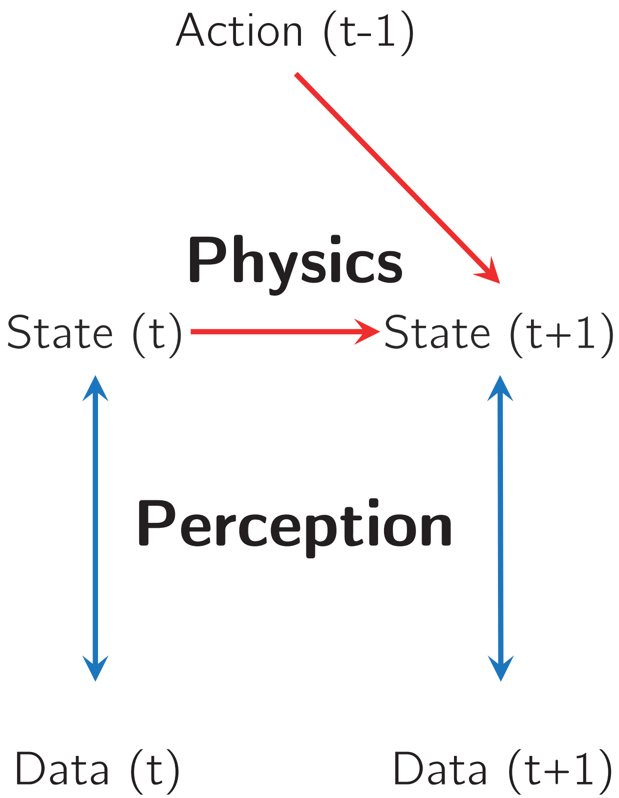
Modeling the Physical World



Physical World Representations are Universal



Cognitive Science Meets Machine Scene Understanding



Causal structure and cognitive science insights provide guidance on building machine scene understanding models:

- When and where to use top-down simulation engines vs. bottom-up neural networks?
- What training targets to use for neural networks?
- What intermediate representations to use?
- What training data to use?

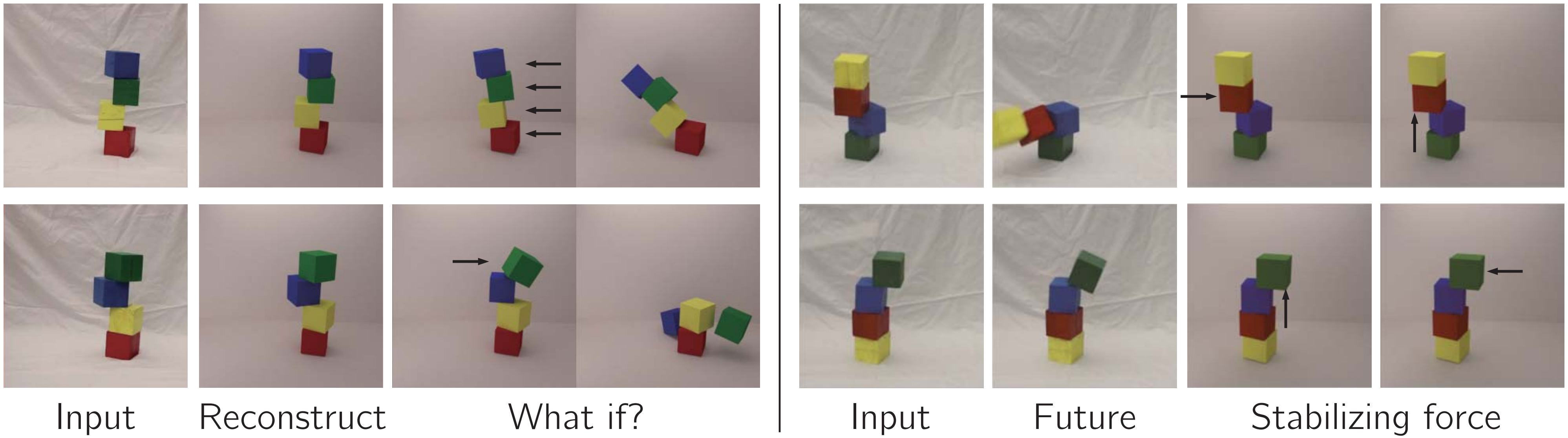
Research in machine intelligence helps to stimulate research in human cognition and neuroscience:

- Computational models for human behaviors;
- Algorithms and representations in the brain.

Learning to See Physics via Visual De-animation

Wu, Lu, Kohli, Freeman, Tenenbaum. NeurIPS'17

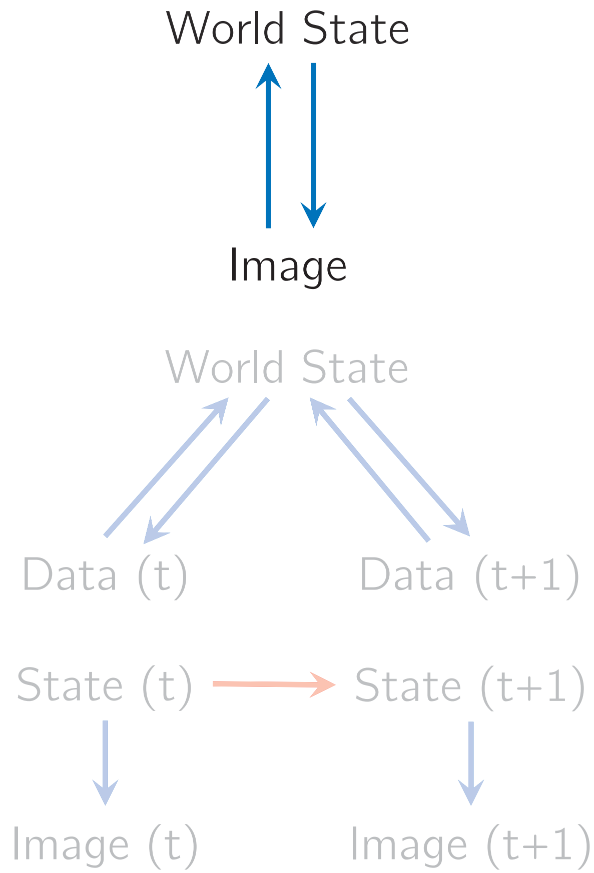
Learning to See Physics via Visual De-animation



Wu, Lu, Kohli, Freeman, Tenenbaum. NeurIPS'17

Physical Scene Understanding

- Learning to invert a graphics engine
- Learning to invert a physics engine
- Learning simulation engines themselves

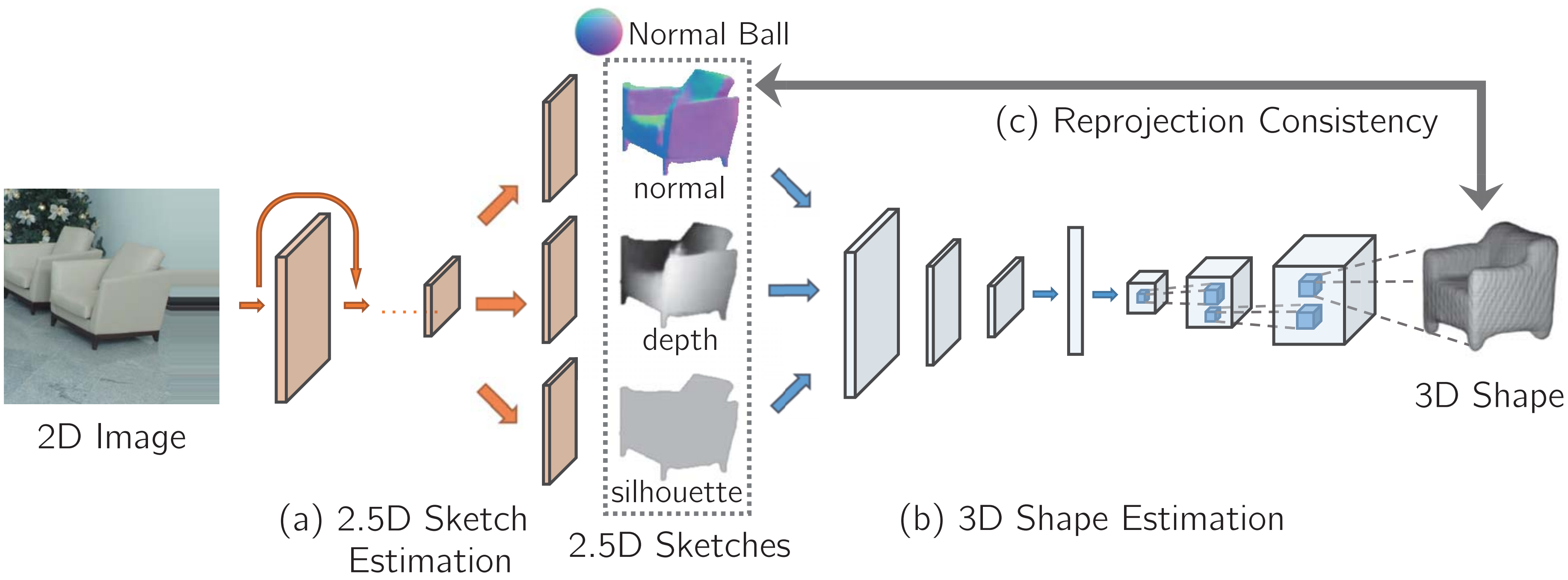


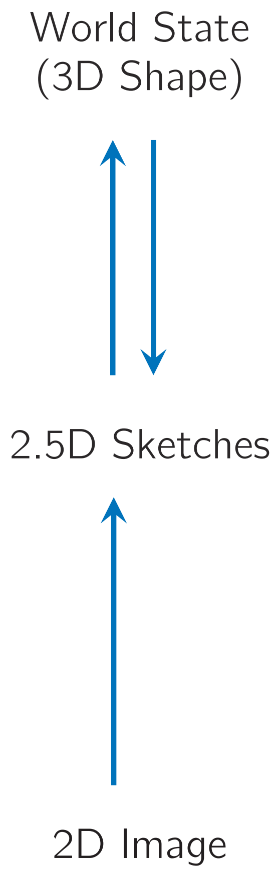
3D Reconstruction

Forward: image formation

Inverse: shape estimation

MarrNet: 3D Reconstruction via 2.5D Sketches





Wu*, Wang*, Xue, Sun, Freeman, Tenenbaum. NeurIPS'17

Comparisons

Images Ground truth DRC MarrNet

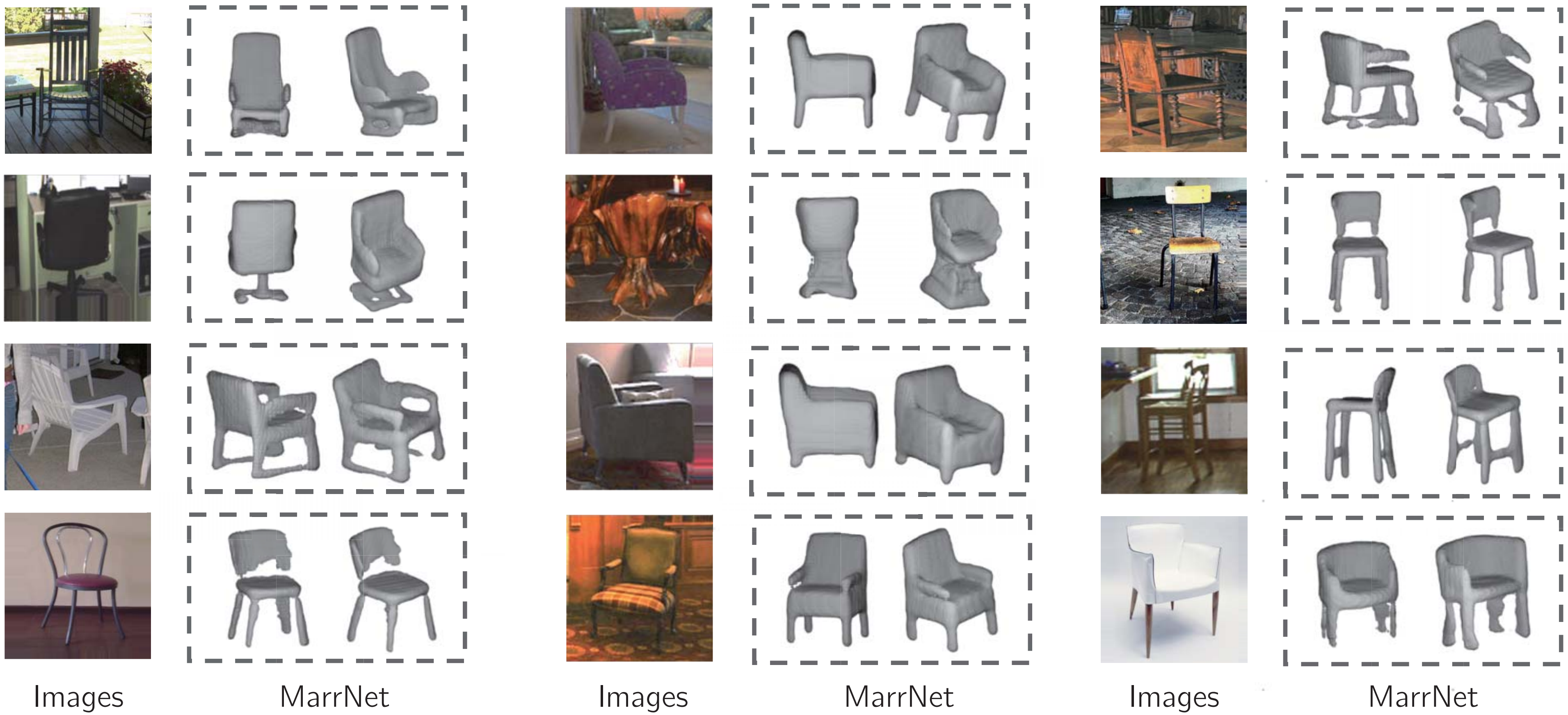
Methods	IoU
DRC 3D [CVPR '17]	0.34
MarrNet	0.38

Intersection over Union (IoU)

	DRC 3D	MarrNet	GT
DRC 3D	50	26	17
MarrNet	74	50	42
GT	83	58	50

Percentages of users that preferred the left approach to the top one

Results on PASCAL 3D+

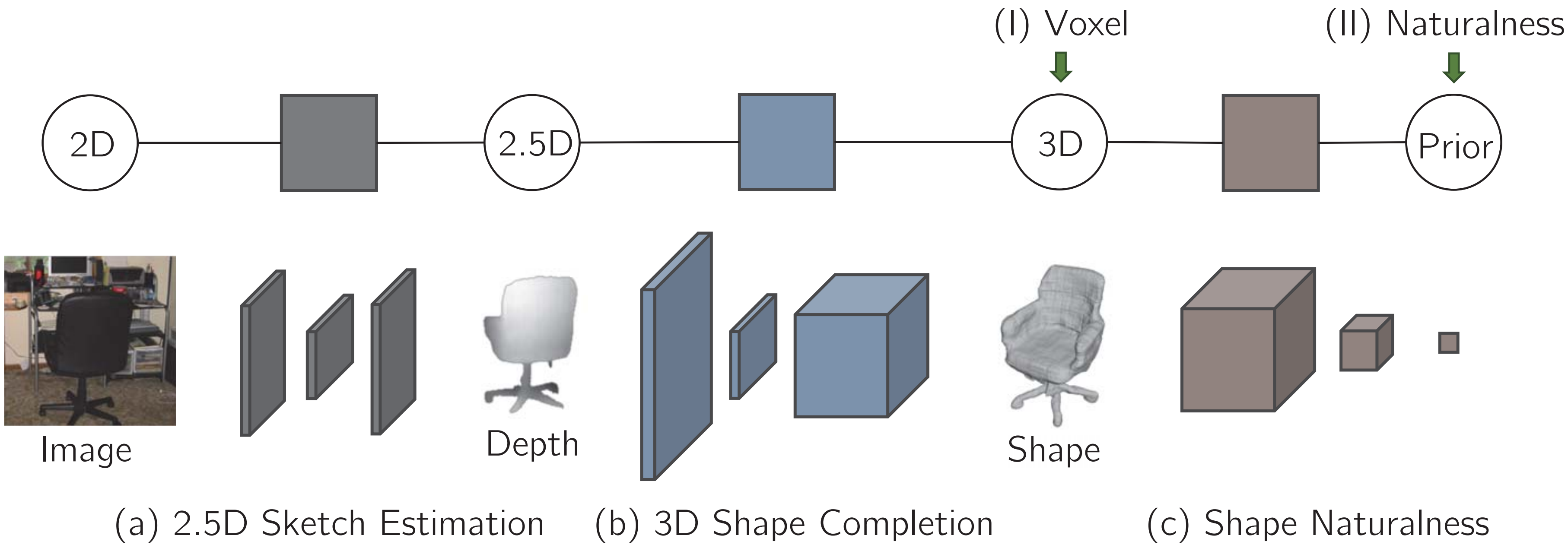


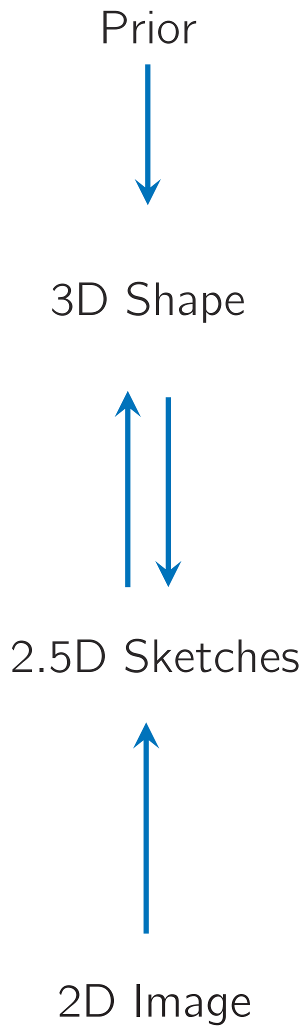
Wu*, Wang*, Xue, Sun, Freeman, Tenenbaum. NeurIPS'17

Results on IKEA

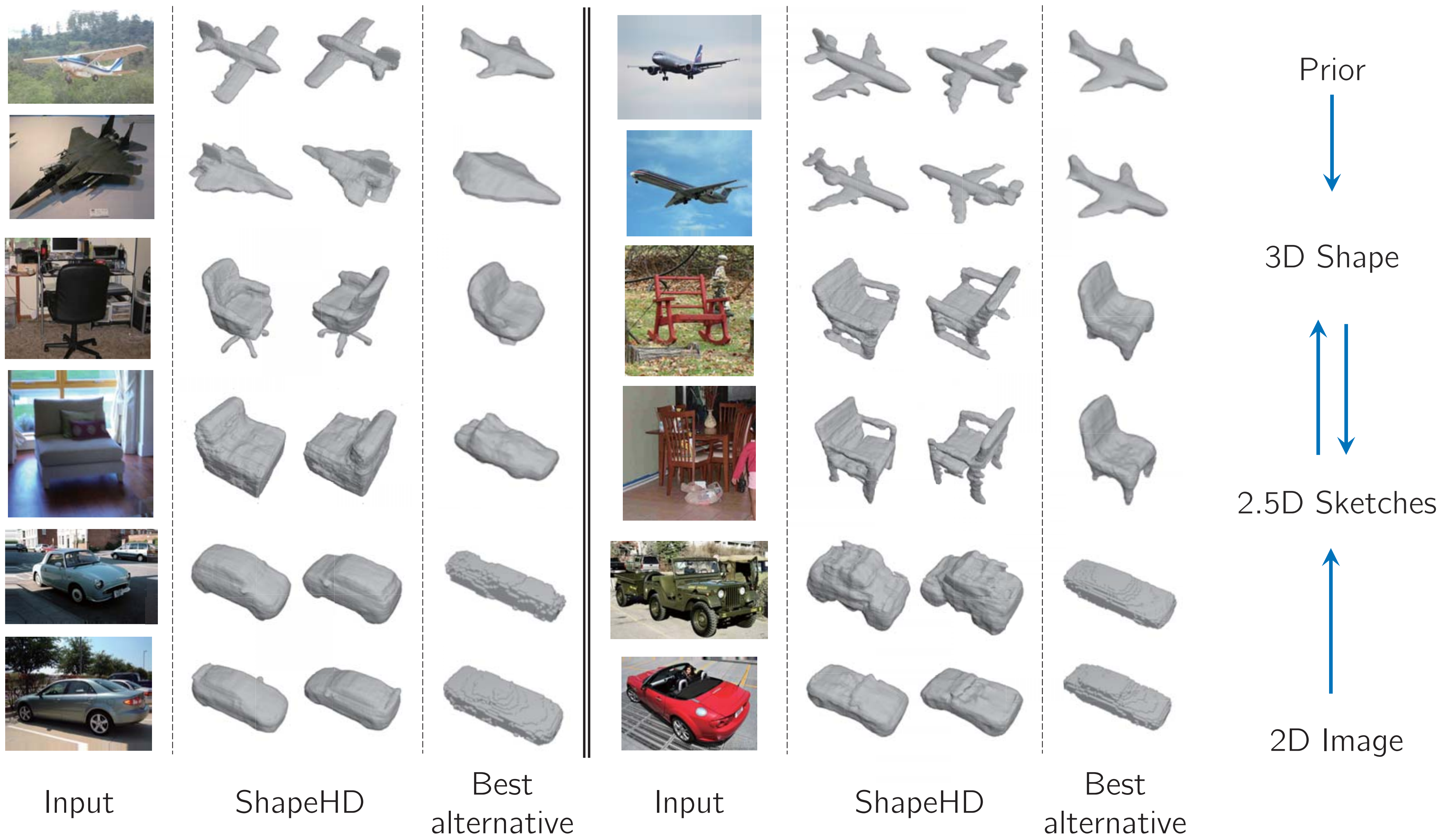
Images Ground truth MarrNet Images Ground truth MarrNet

Ext. I: Incorporating Priors





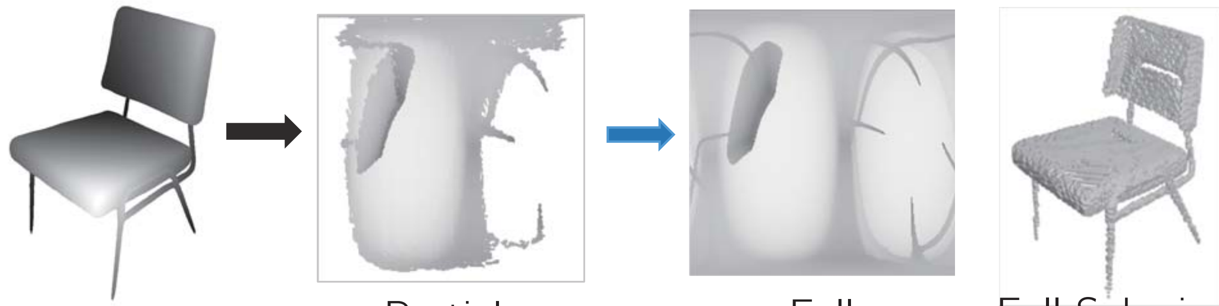
Wu*, Zhang*, Zhang, Zhang, Freeman, Tenenbaum. ECCV'18


Input
ShapeHD
Best alternative
Input
ShapeHD
Best alternative
Prior
3D Shape
2.5D Sketches
2D Image

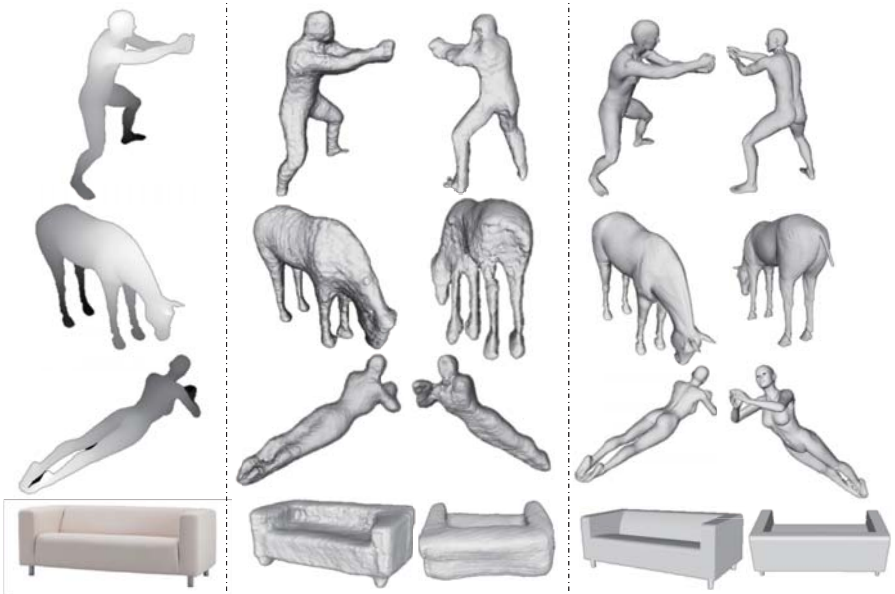
Ext. II: Generalization to Unseen Classes

Trained on chairs, planes, cars
Tested on tables

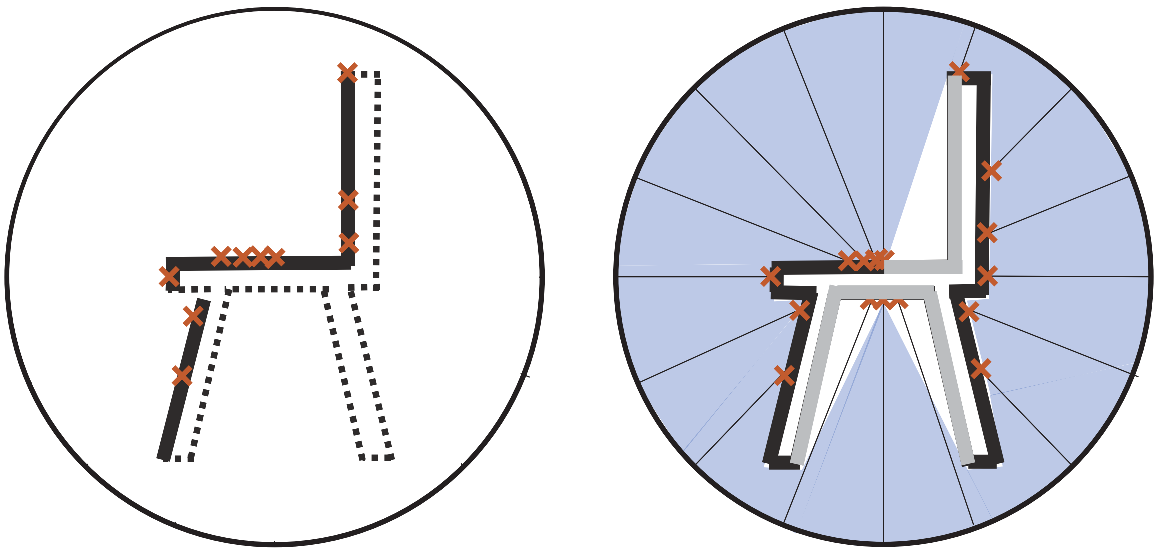
Input DRC (UCB) Our results



Depth D Partial Spherical Map Full Spherical Map Full Spherical Map (3D)

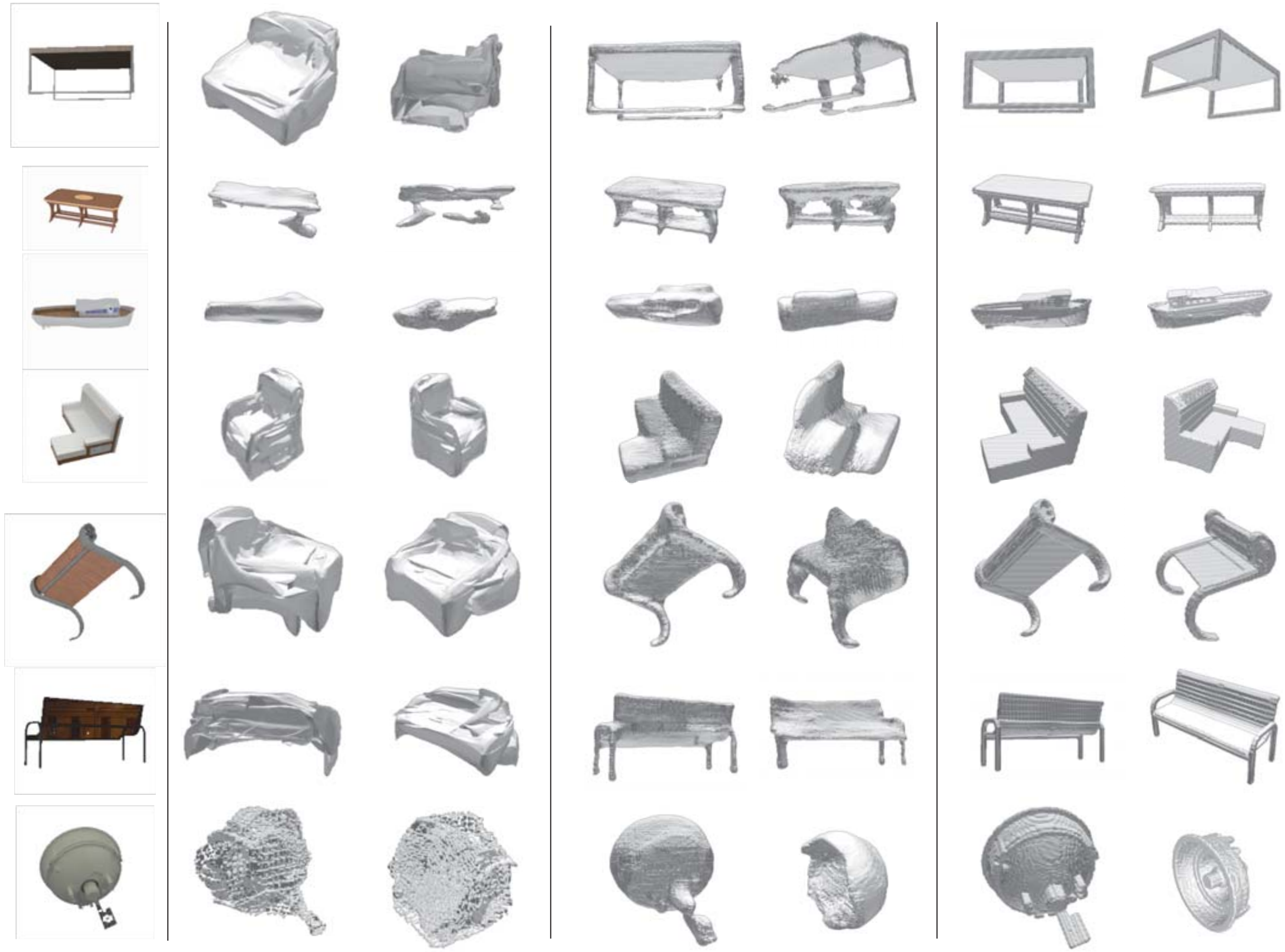


Input Ours Ground Truth



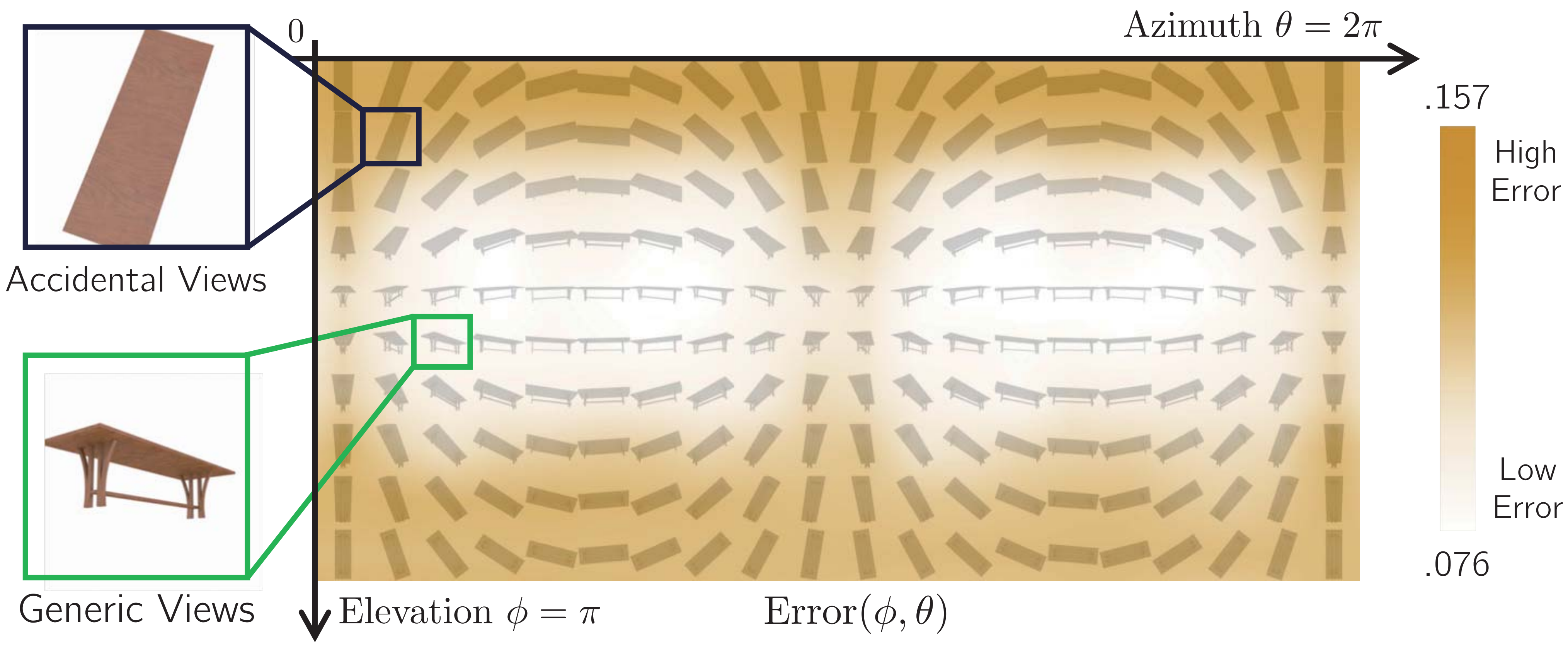
Zhang*, Zhang*, Zhang, Tenenbaum, Freeman, Wu. NeurIPS'18

Generalization to Novel Classes (Table, Boat, Sofa, Bench, Lamp)



Input Best Baseline Ours Ground Truth

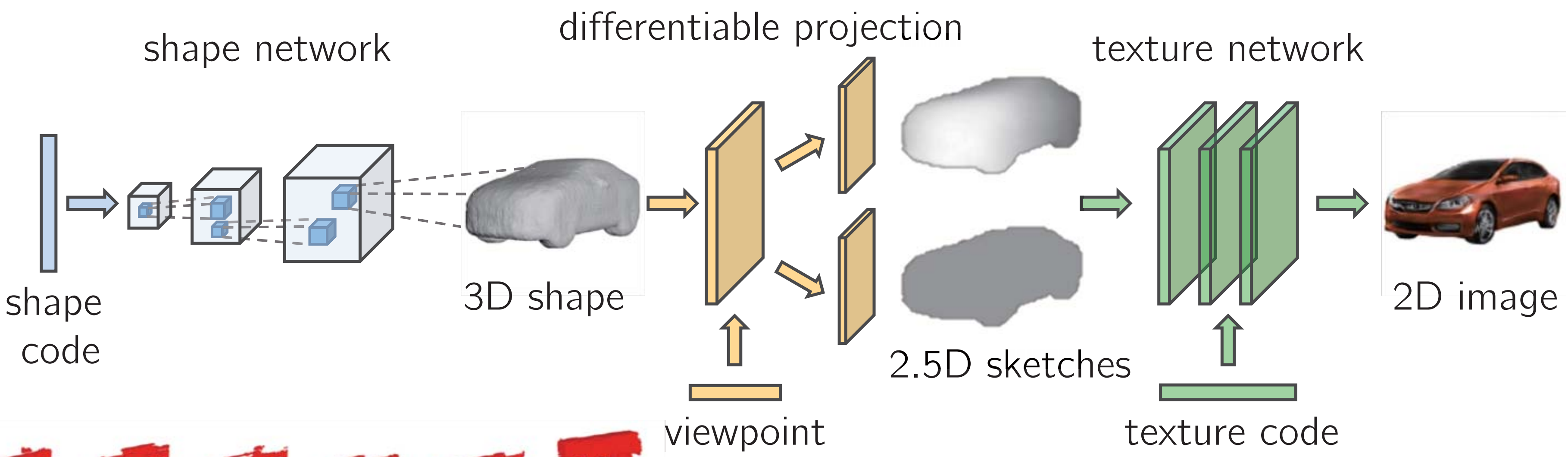
Canonical Viewpoints in Generalization

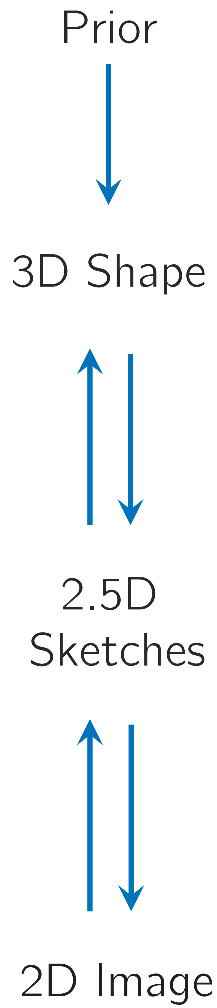


Accidental views → large errors

Palmer, Rosch, Chase. Atten. Perform. 1981

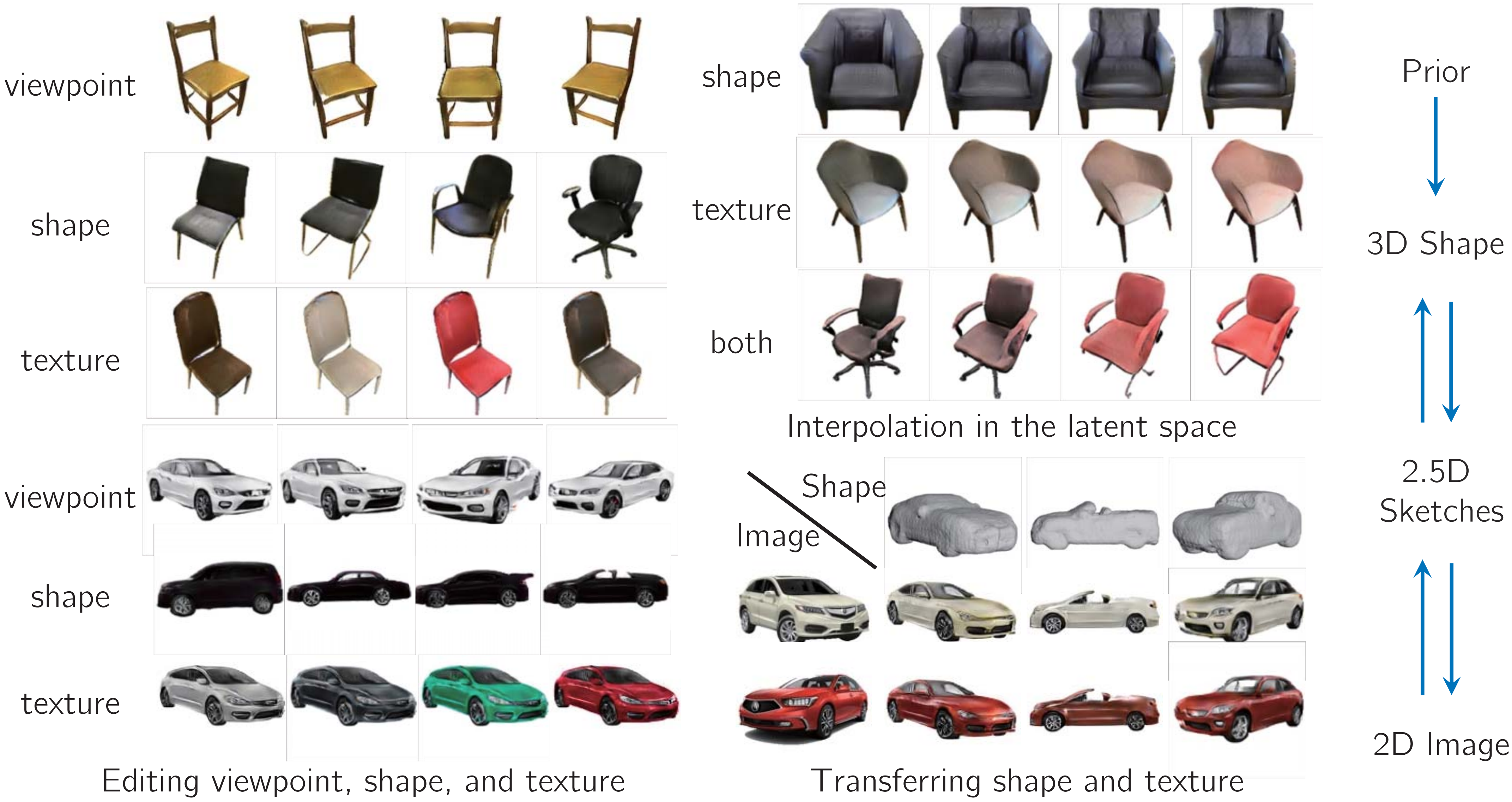
Top-Down Shape and Texture Synthesis





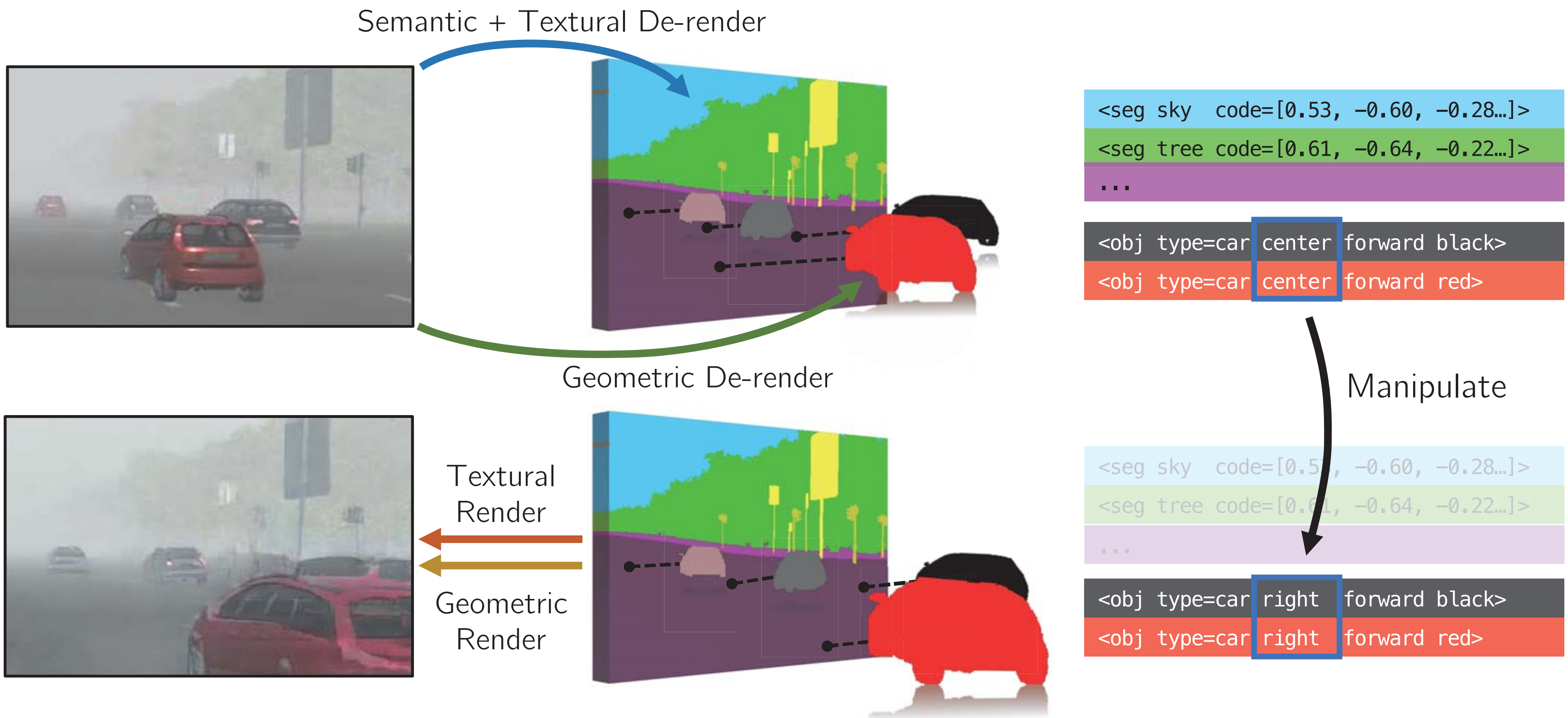
Wu*, Zhang*, Xue, Freeman, Tenenbaum. NeurIPS'16

Zhu, Zhang, Zhang, Wu, Torralba, Tenenbaum, Freeman. NeurIPS'18



Zhu, Zhang, Zhang, Wu, Torralba, Tenenbaum, Freeman. NeurIPS'18

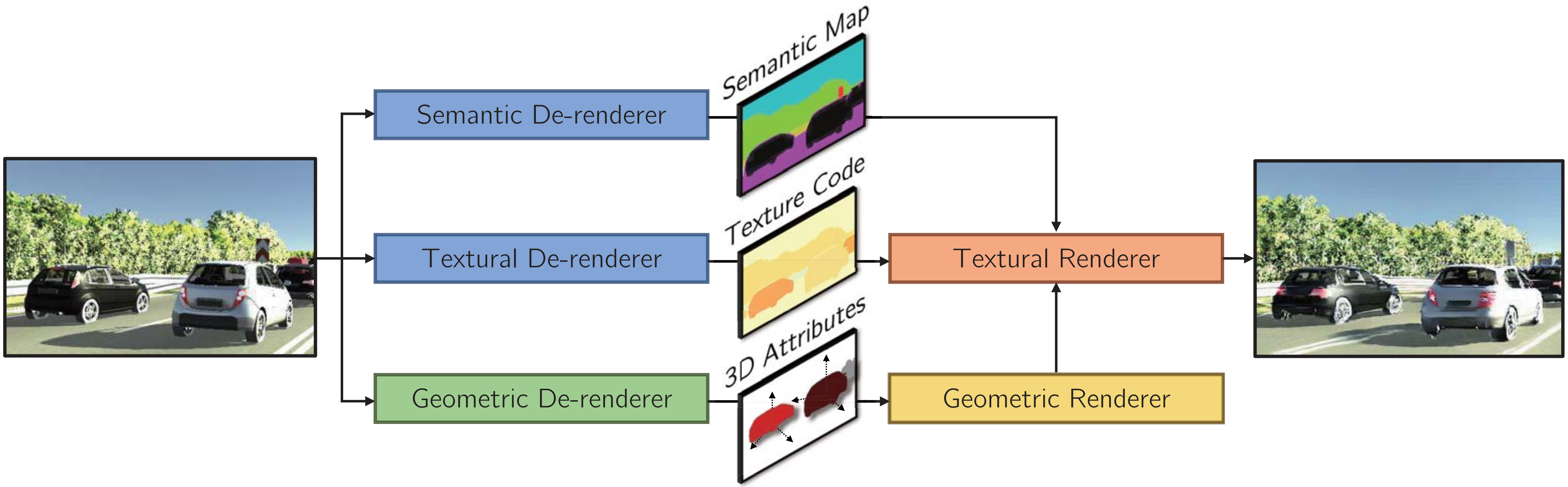
Ext. IV: Extension to Scenes



Goal: Recovering a structured, 3D-aware scene representation.

The structured representation allows re-rendering and editing the image.

3D Disentangled Scene Representation



Disentangled model for the scene's semantics, texture, and object geometry and 6DOF pose.

Yao*, Hsu*, Zhu, Wu, Torralba, Freeman, Tenenbaum. NeurIPS'18

Image Editing on Virtual KITTI

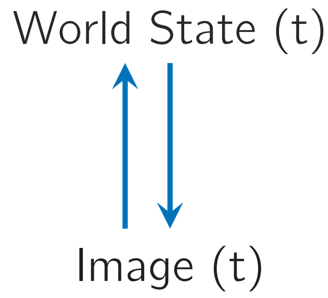
Yao*, Hsu*, Zhu, Wu, Torralba, Freeman, Tenenbaum. NeurIPS'18

Image Editing on CityScapes (Real Images)

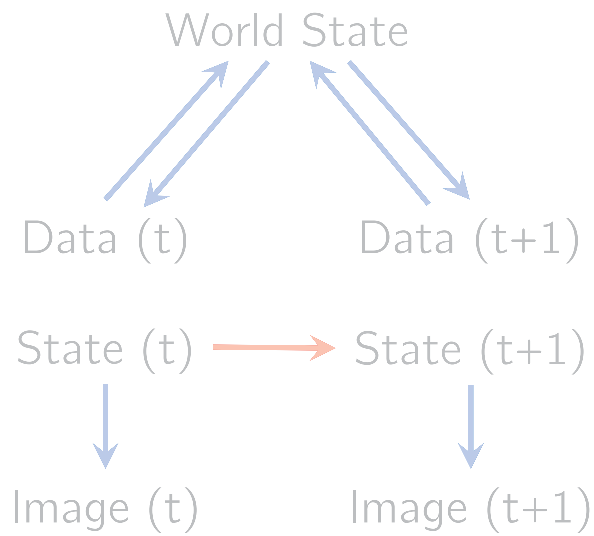
Yao*, Hsu*, Zhu, Wu, Torralba, Freeman, Tenenbaum. NeurIPS'18

Physical Scene Understanding

- Learning to invert a graphics engine
  - Inferring fine object geometry
  - Learning structured shape representations (shape + texture)
  - Beyond single object, learning scene representations



- Learning to invert a physics engine



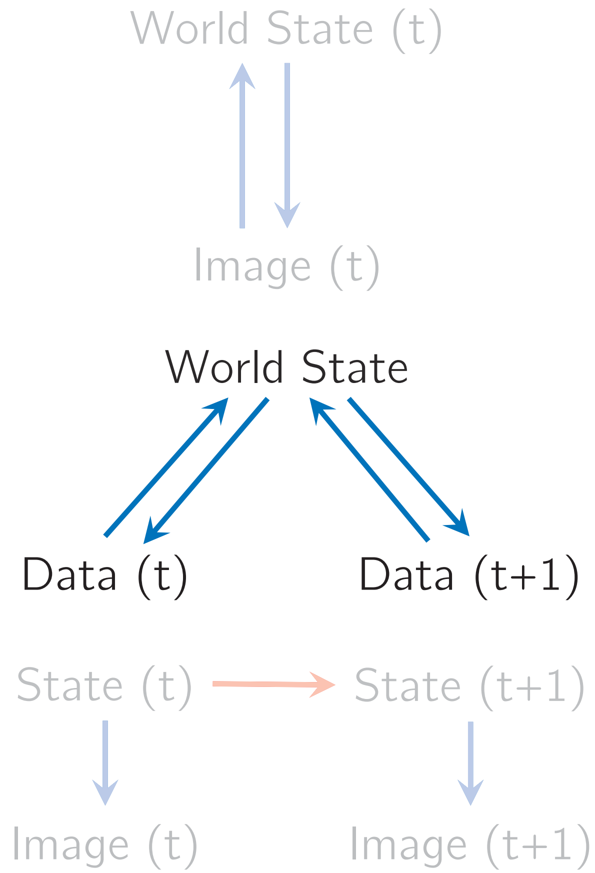
- Learning simulation engines themselves

Physical Scene Understanding

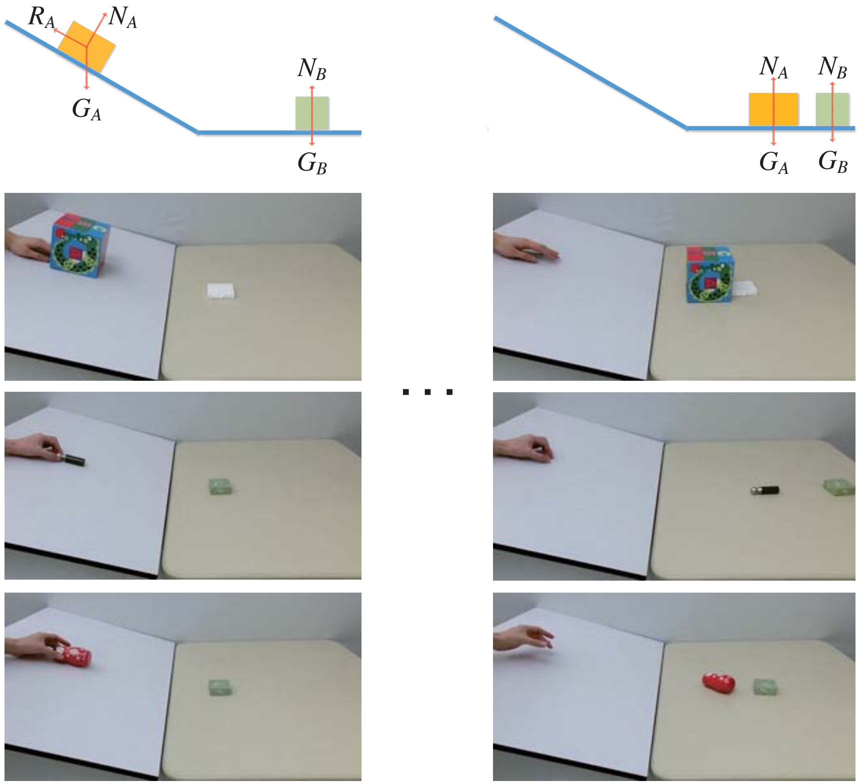
- Learning to invert a graphics engine
  - Inferring fine object geometry
  - Learning structured shape representations (shape + texture)
  - Beyond single object, learning scene representations

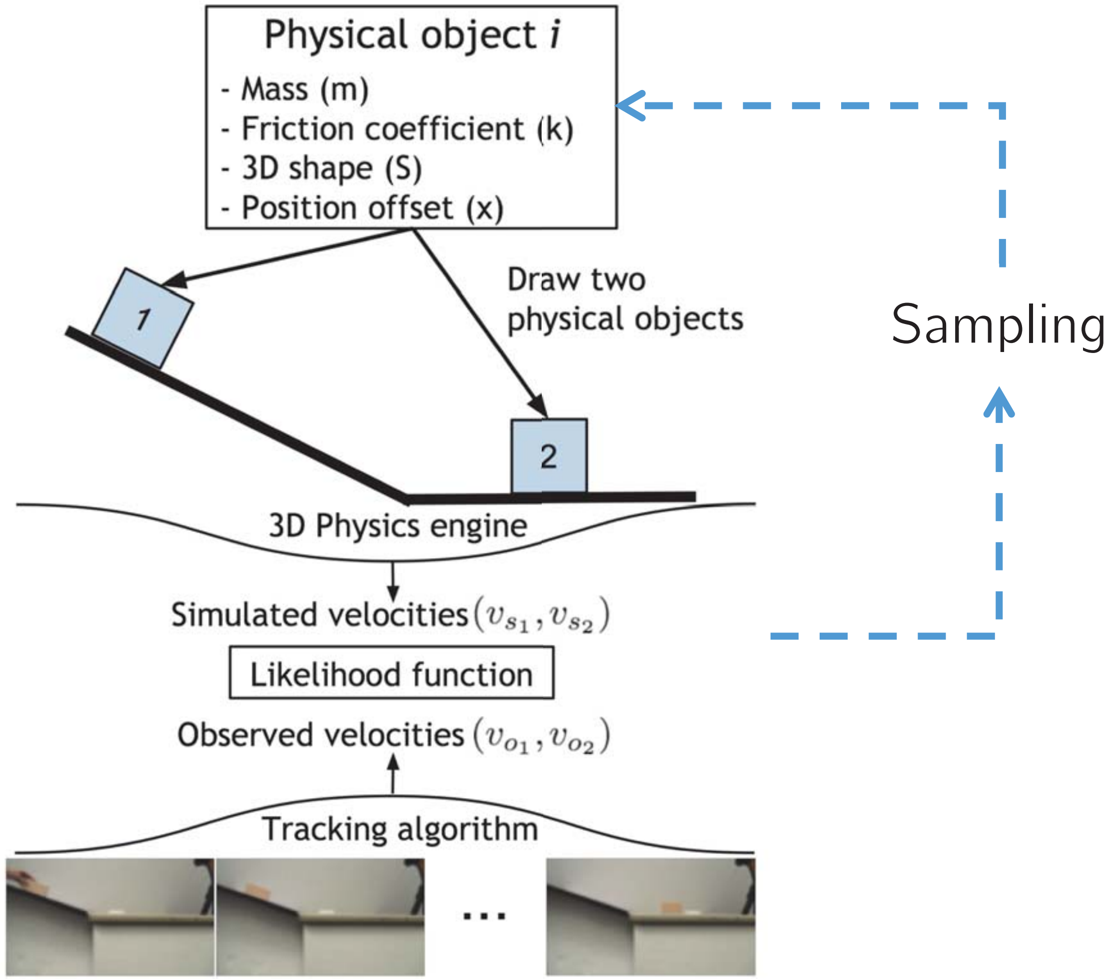
- Learning to invert a physics engine

- Learning simulation engines themselves



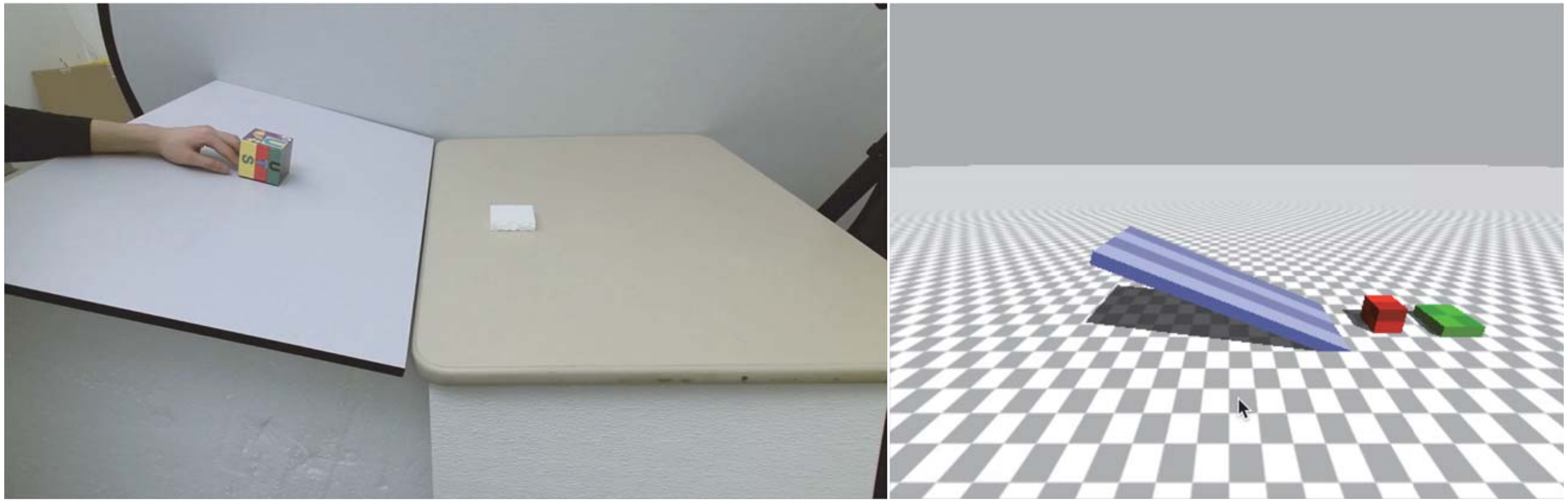
Galileo





Wu*, Yildirim*, Lim, Freeman, Tenenbaum. NeurIPS'15

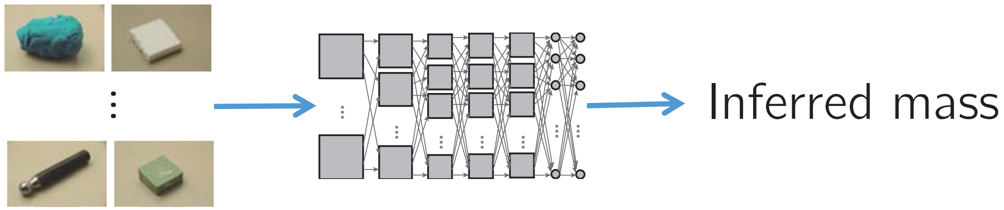
Results

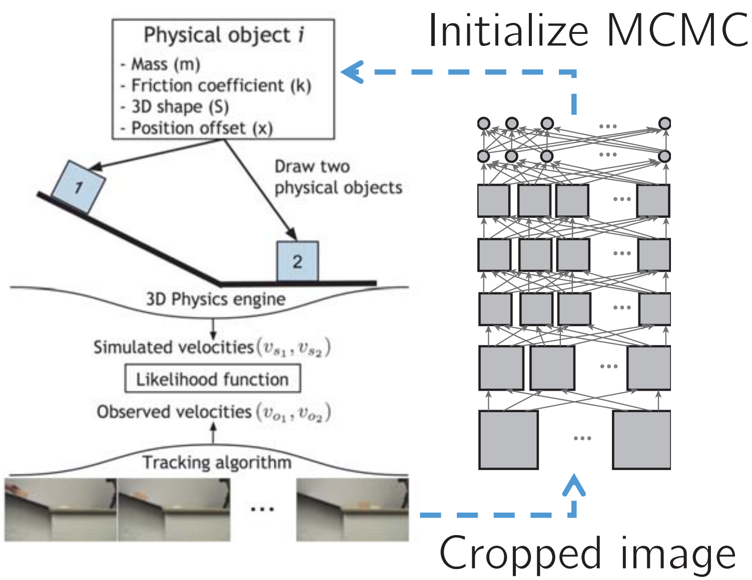


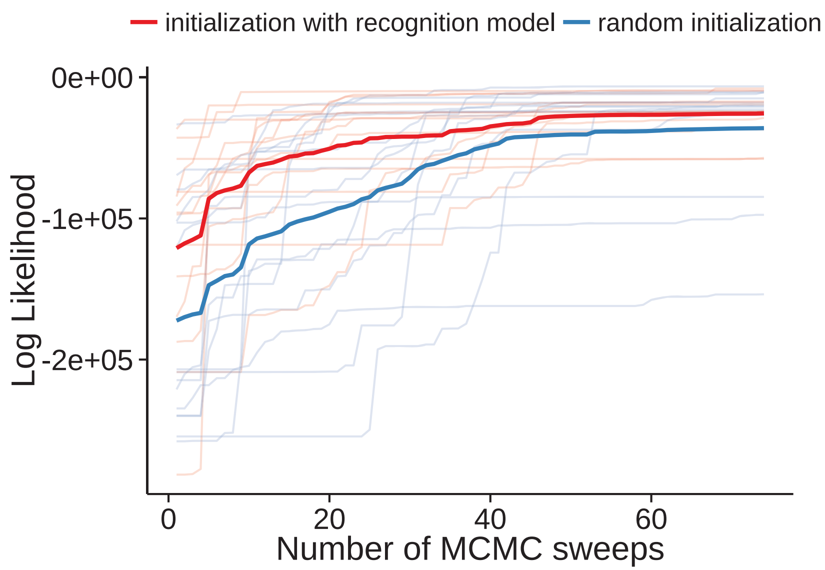
Wu*, Yildirim*, Lim, Freeman, Tenenbaum. NeurIPS'15

Generative + Recognition Model

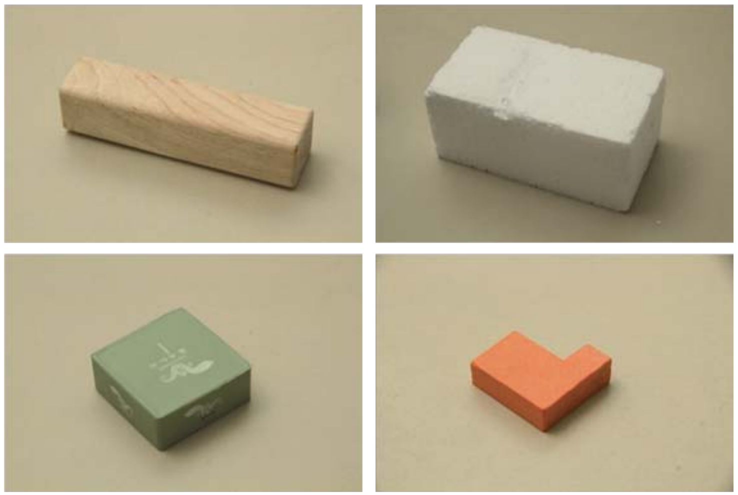
If the model has prior knowledge like humans do...





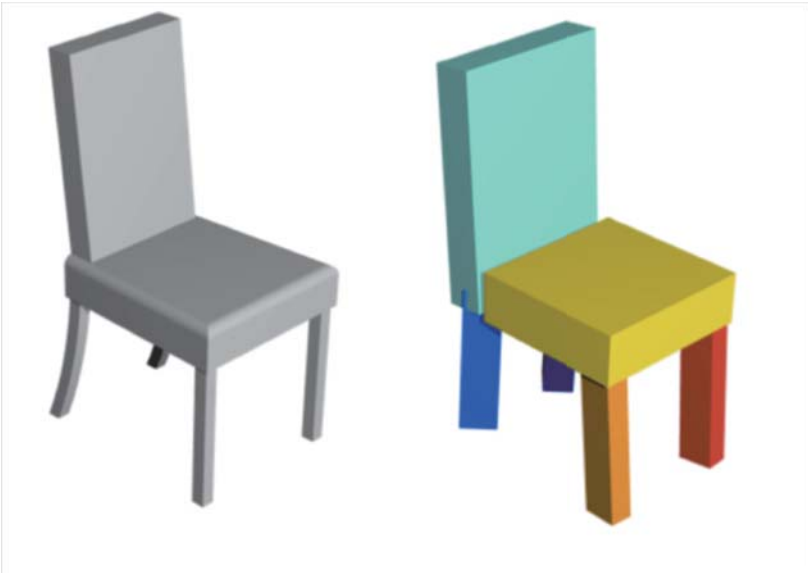


We've seen...



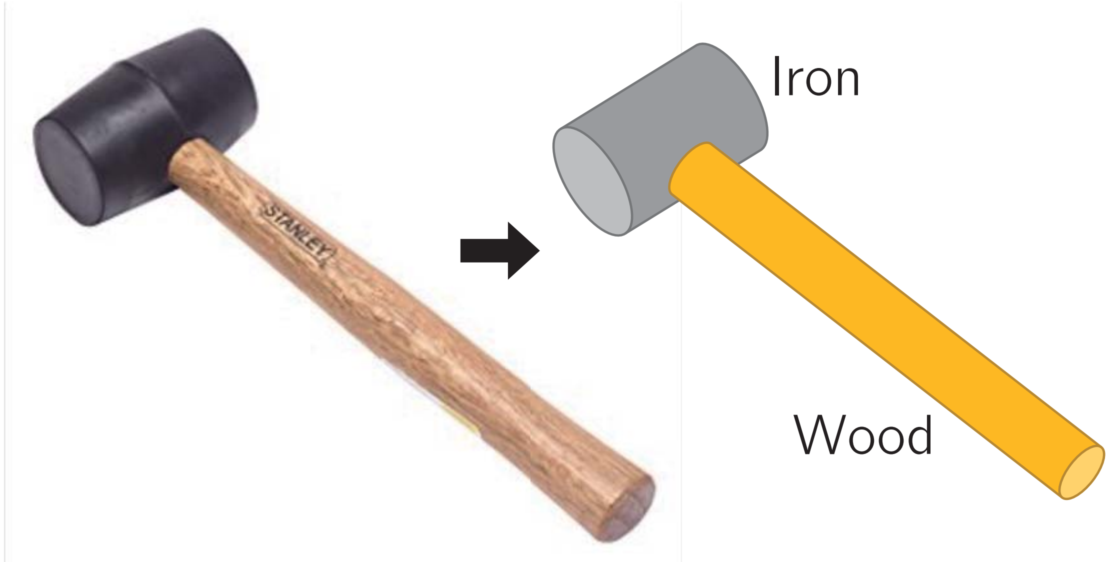
What about?

Learning Shape Abstractions



Tulsiani, Su, Guibas, Efros, Malik. CVPR'17

Physical Primitive Decomposition

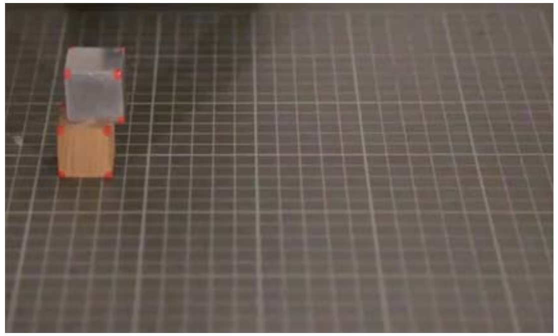


Liu, Freeman, Tenenbaum, Wu. ECCV'18

Appearance + Physics

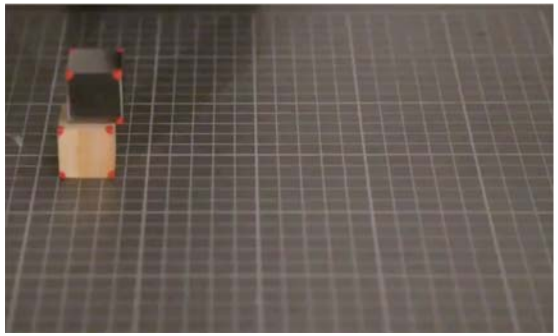
Aluminum
(2.87g/ml)

Oak
(0.67g/ml)



Steel
(7.74g/ml)

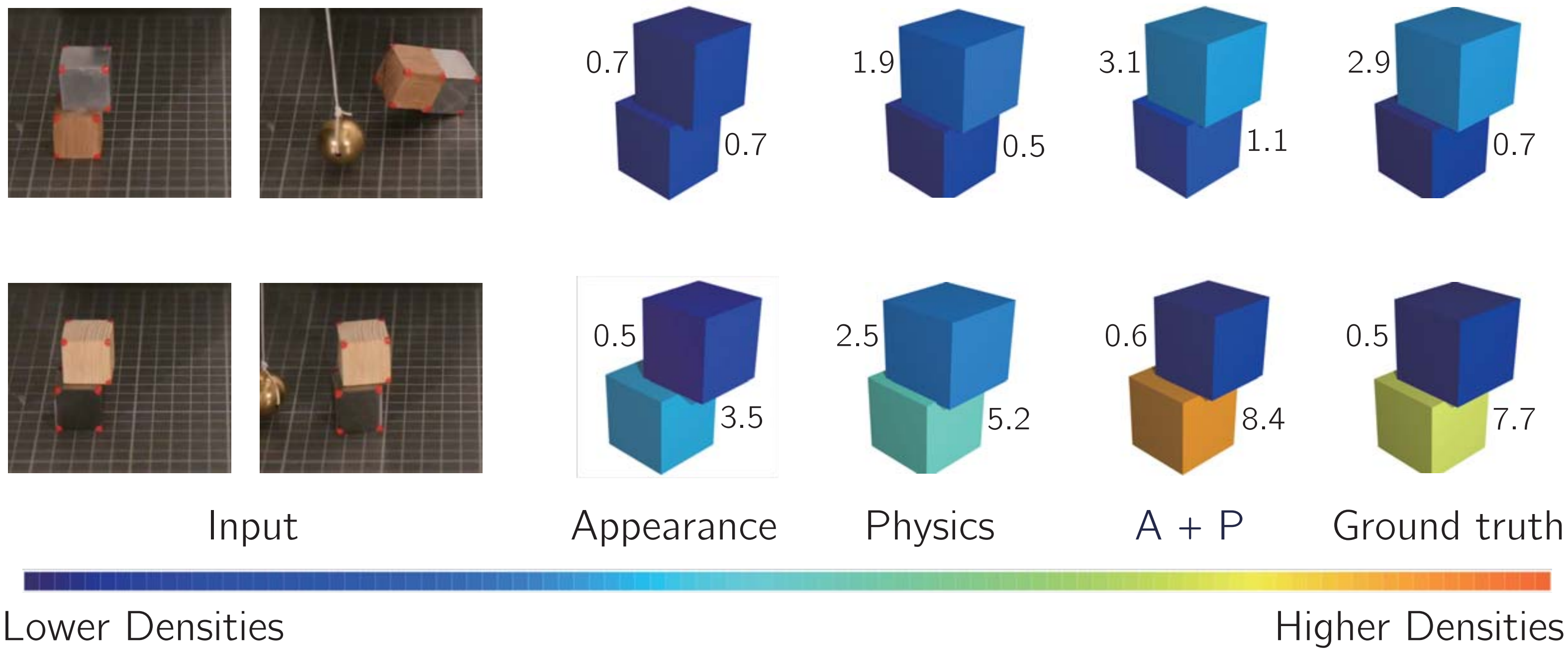
Pine
(0.48g/ml)



Visual Appearance
(Very Similar)

Physics Trajectory
(Very Different)

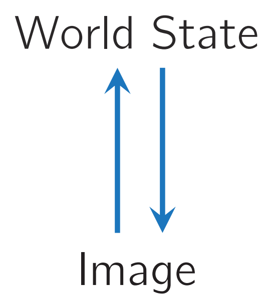
Physical Primitive Decomposition



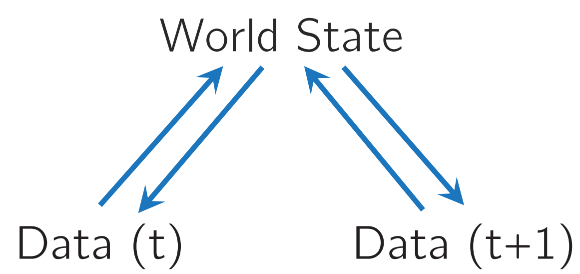
Liu, Freeman, Tenenbaum, Wu. ECCV'18

Physical Scene Understanding

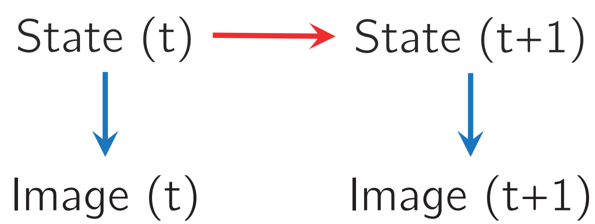
- Learning to invert a graphics engine
  - Inferring fine object geometry
  - Learning structured shape representations (shape + texture)
  - Beyond single object, learning scene representations



- Learning to invert a physics engine
  - Inferring object physical properties
  - Joint modeling of object shape and physics



- Learning simulation engines themselves

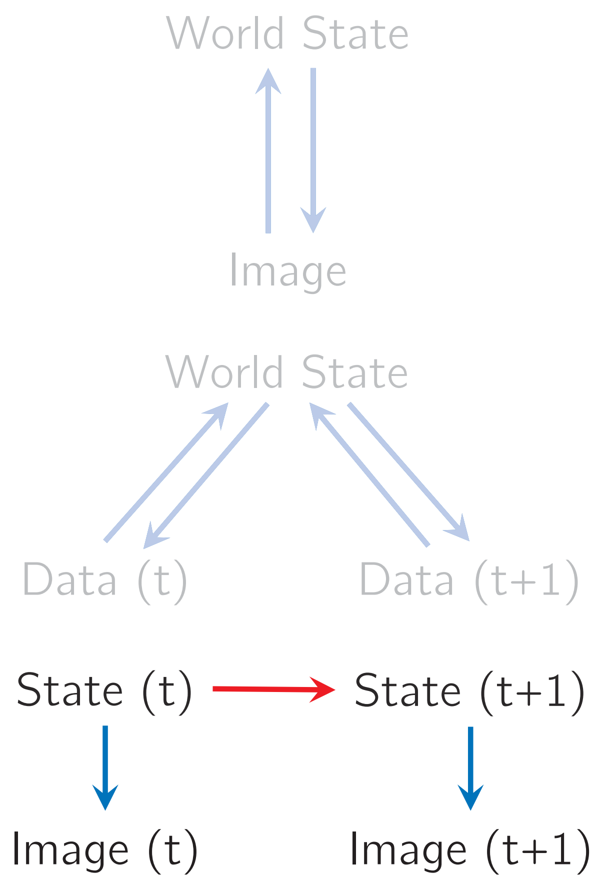


Physical Scene Understanding

- Learning to invert a graphics engine
  - Inferring fine object geometry
  - Learning structured shape representations (shape + texture)
  - Beyond single object, learning scene representations

- Learning to invert a physics engine
  - Inferring object physical properties
  - Joint modeling of object shape and physics

- Learning simulation engines themselves



Key Features on Dynamics Modeling

- Depending on visual content
- Modeling uncertainty

Visual Dynamics

- Two temporally-consecutive frames

$P(I_2|I_1)$

$P(I_2|I_1)$: Probabilistic distribution of the second frame conditioned on the first frame

- Prediction

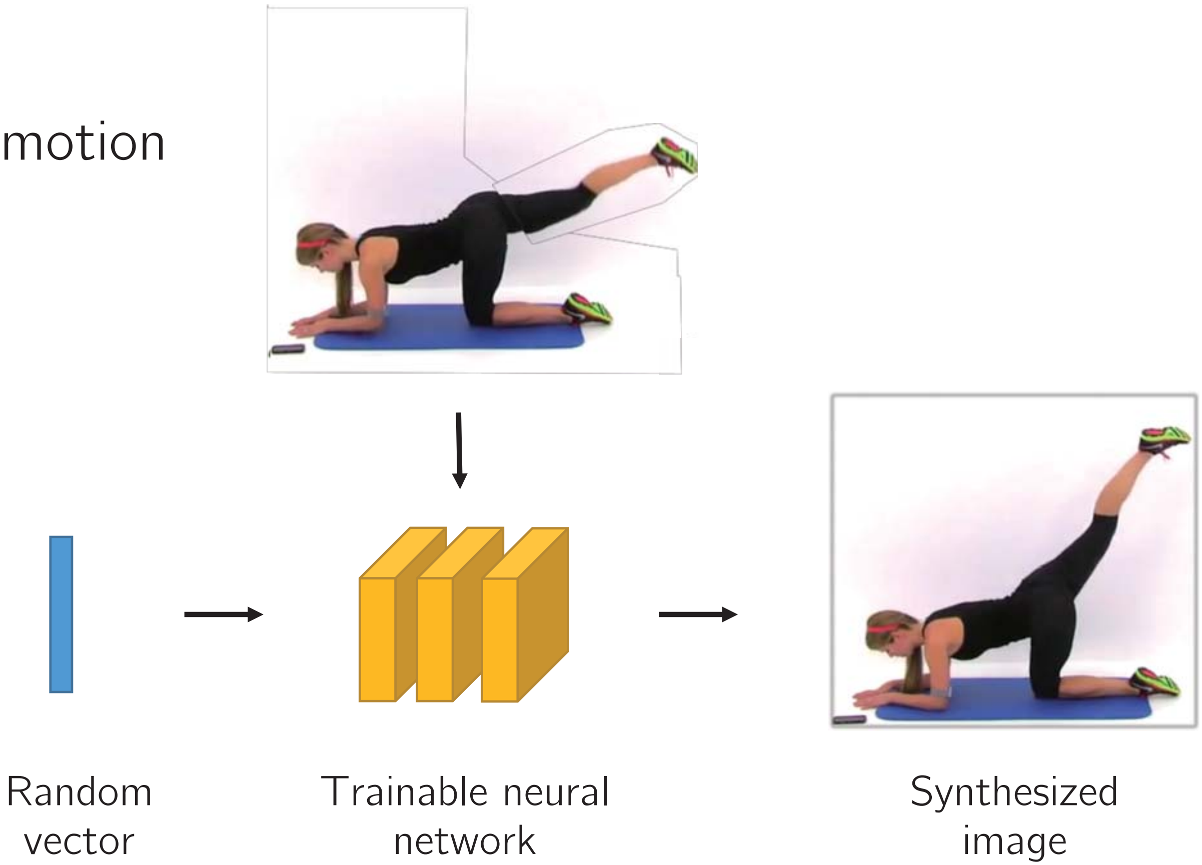
Input

Sampled frames

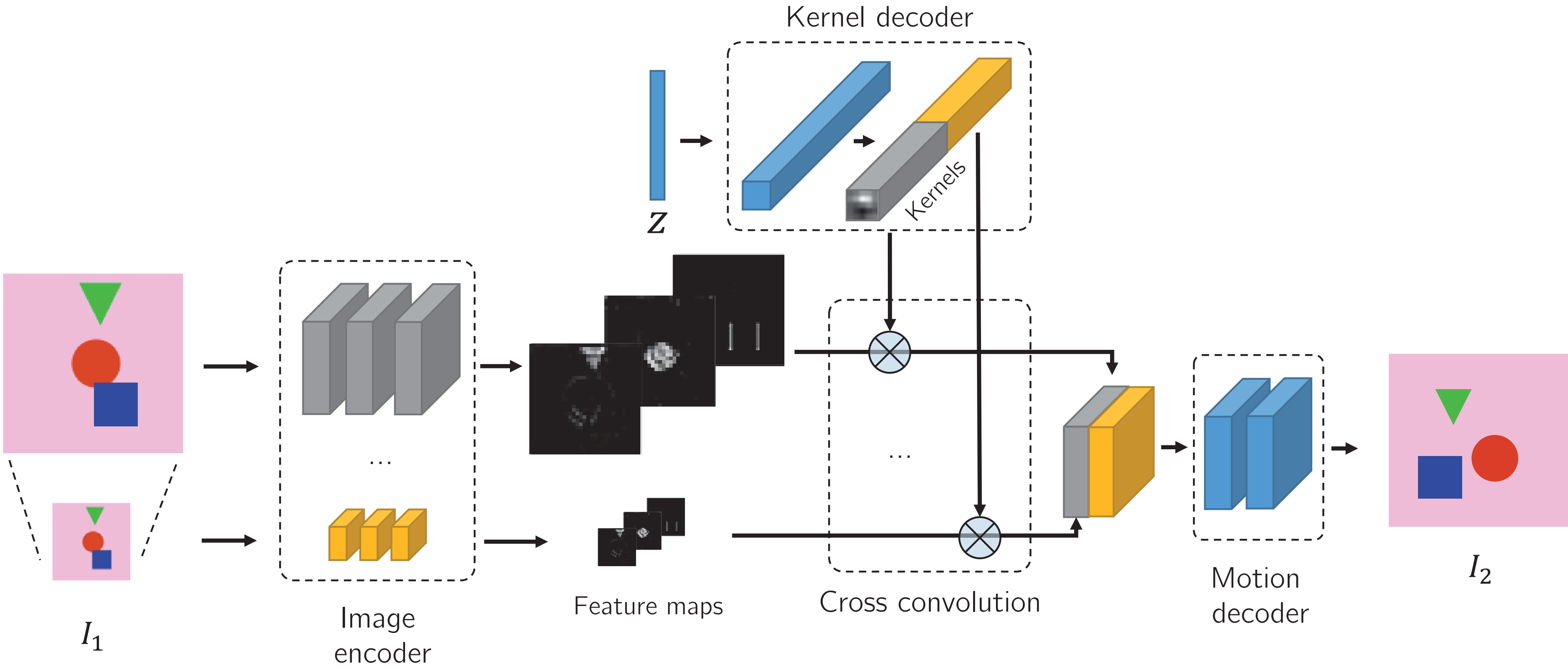
Xue*, Wu*, Bouman, Freeman. NeurIPS'16, TPAMI'18

Decomposing Objects into Independently Movable Parts

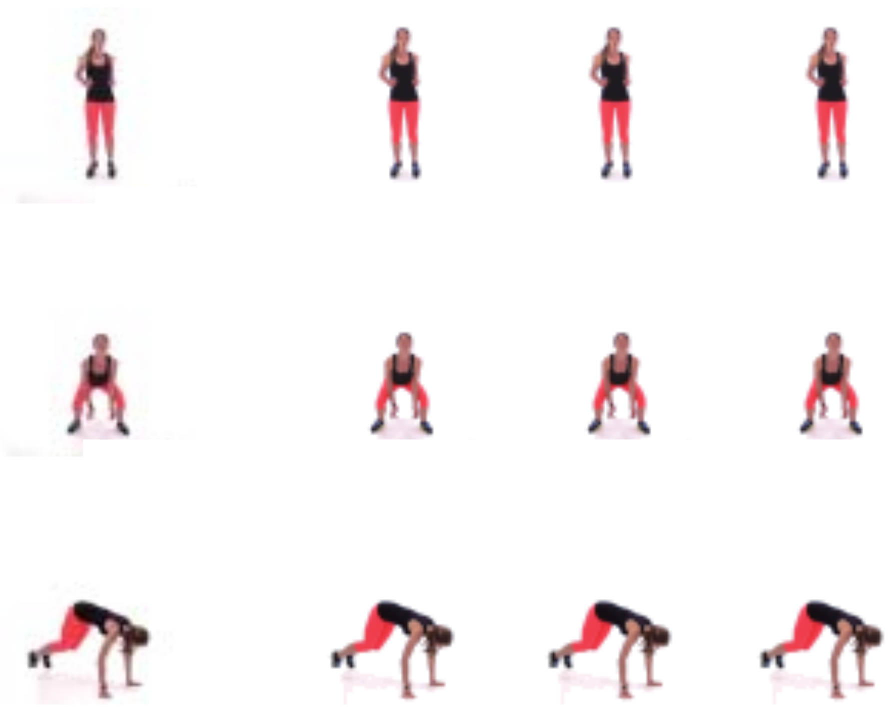
- Identify movable segments
- Model their dynamics
- Combine the sampled motion



Layered Cross-Convolutional Networks



Results on Real Videos

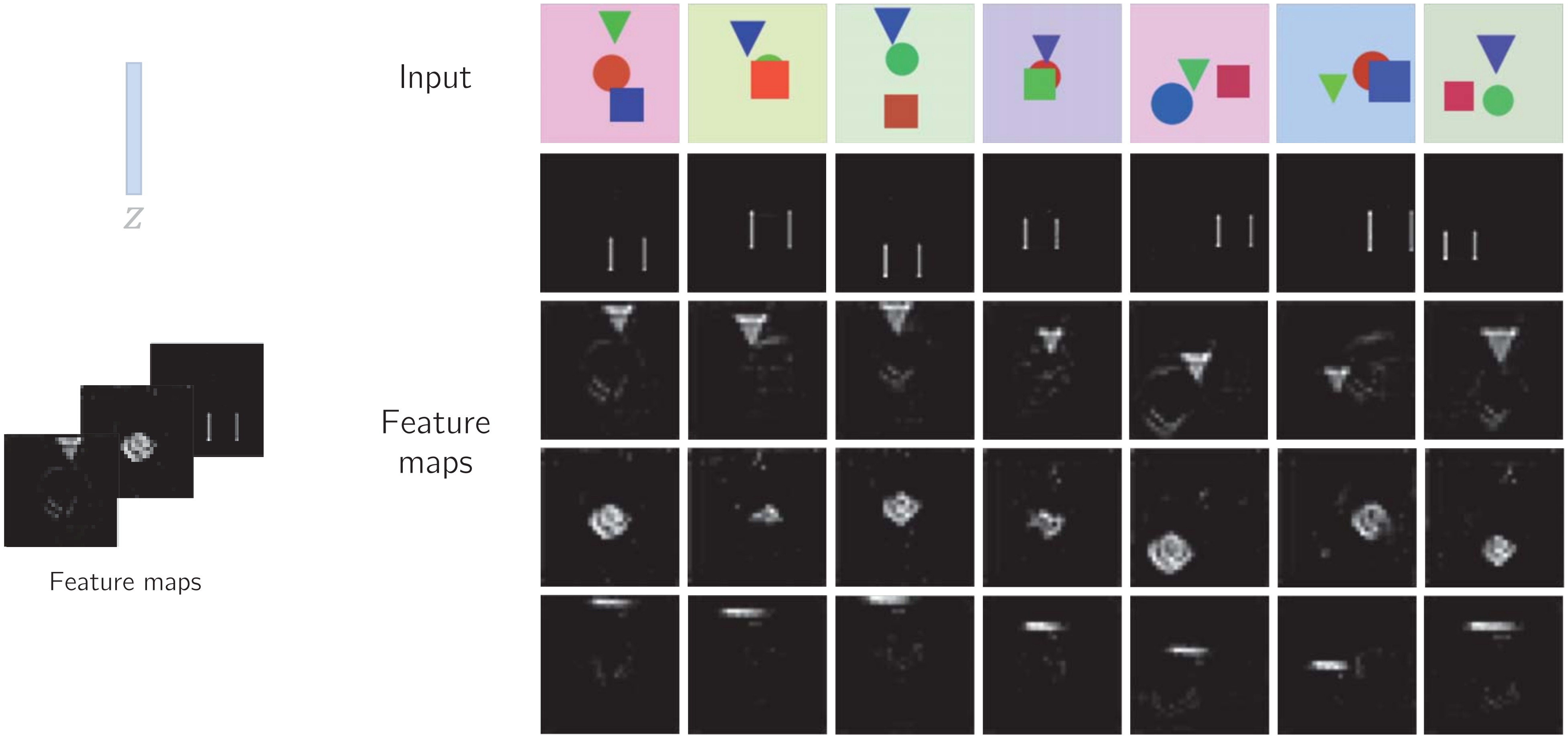


Input

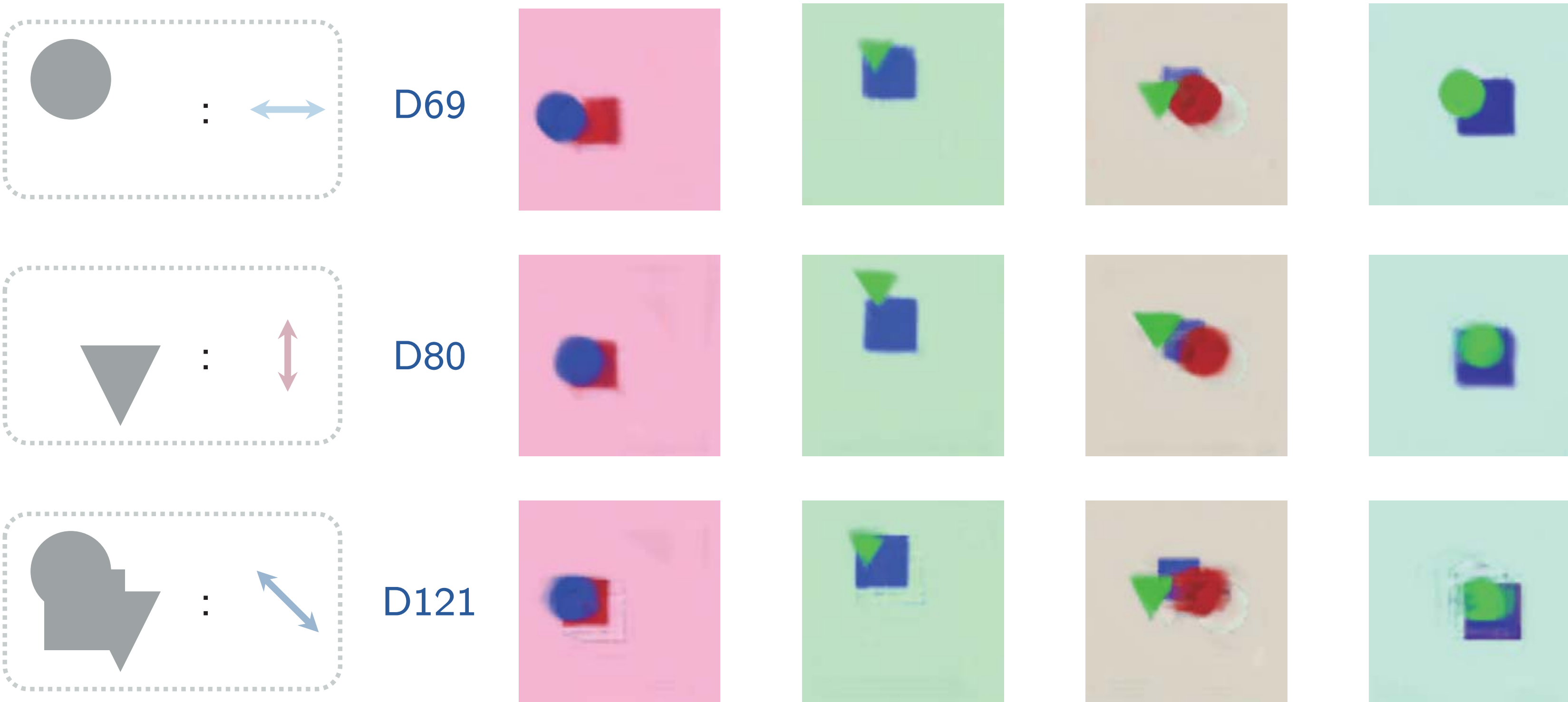
Synthesized next frames

	% of synthetic labeled as real
Transfer flow	25.5
Ours	31.3

Visualize learned features



Interpretable Latent Representations



Ext. I: Unsupervised Structure Discovery

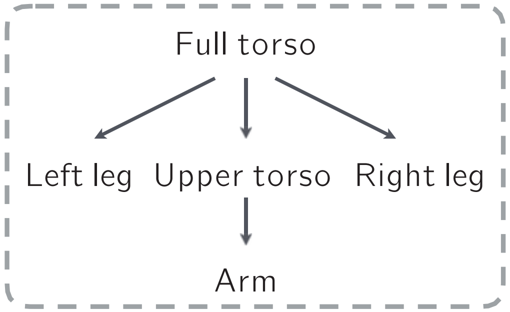
(a) Pairs of images (b) Global motions of parts (c) Local motions of parts

Xu*, Liu*, Sun, Murphy, Freeman, Tenenbaum, Wu. ICLR'19

Ext. I: Unsupervised Structure Discovery

Input frame | Flow field | (a) Full torso | (b) Upper torso | (c) Arm | (d) Right leg | (e) Left leg

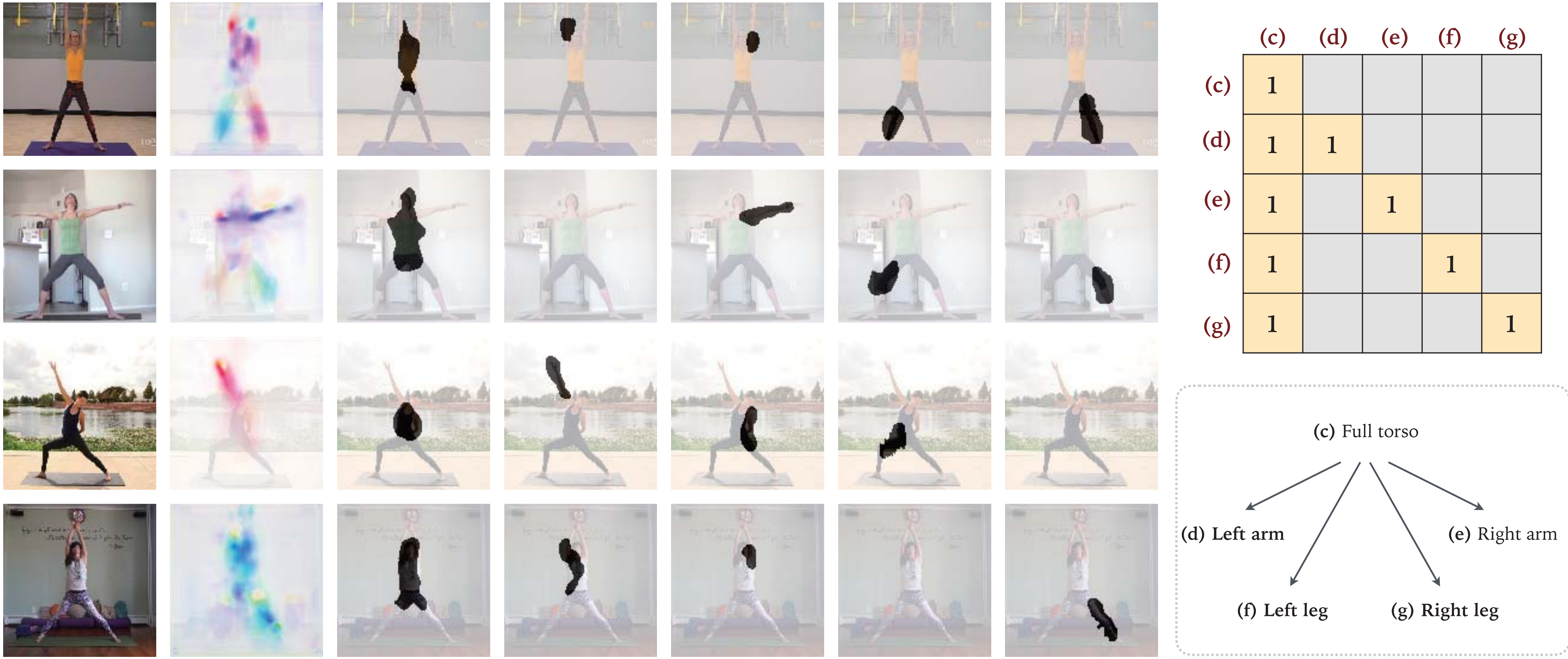
	(a)	(b)	(c)	(d)	(e)
(a)	1				
(b)	1	1			
(c)		1	1		
(d)	1			1	
(e)	1				1



Hierarchical tree structure

Xu*, Liu*, Sun, Murphy, Freeman, Tenenbaum, Wu. ICLR'19

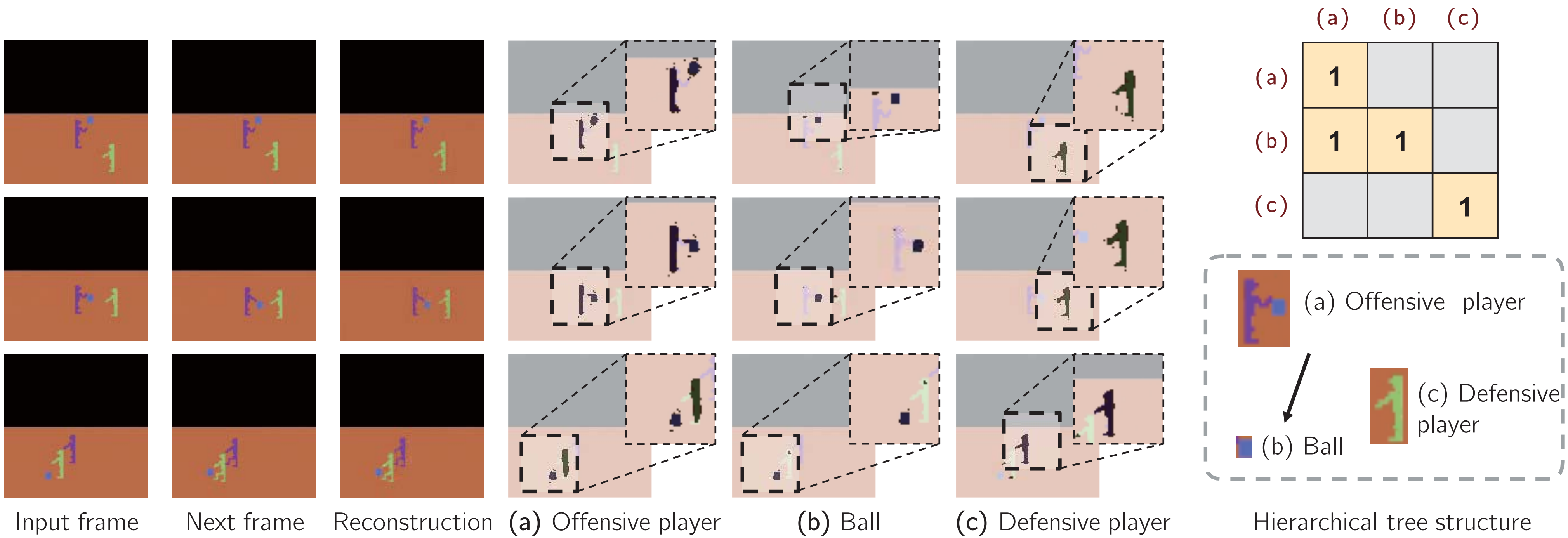
Ext. I: Unsupervised Structure Discovery



(a) Input frame **(b)** Flow field **(c)** Full torso **(d)** Left arm **(e)** Right arm **(f)** Left leg **(g)** Right leg **(h)** Hierarchical tree structure

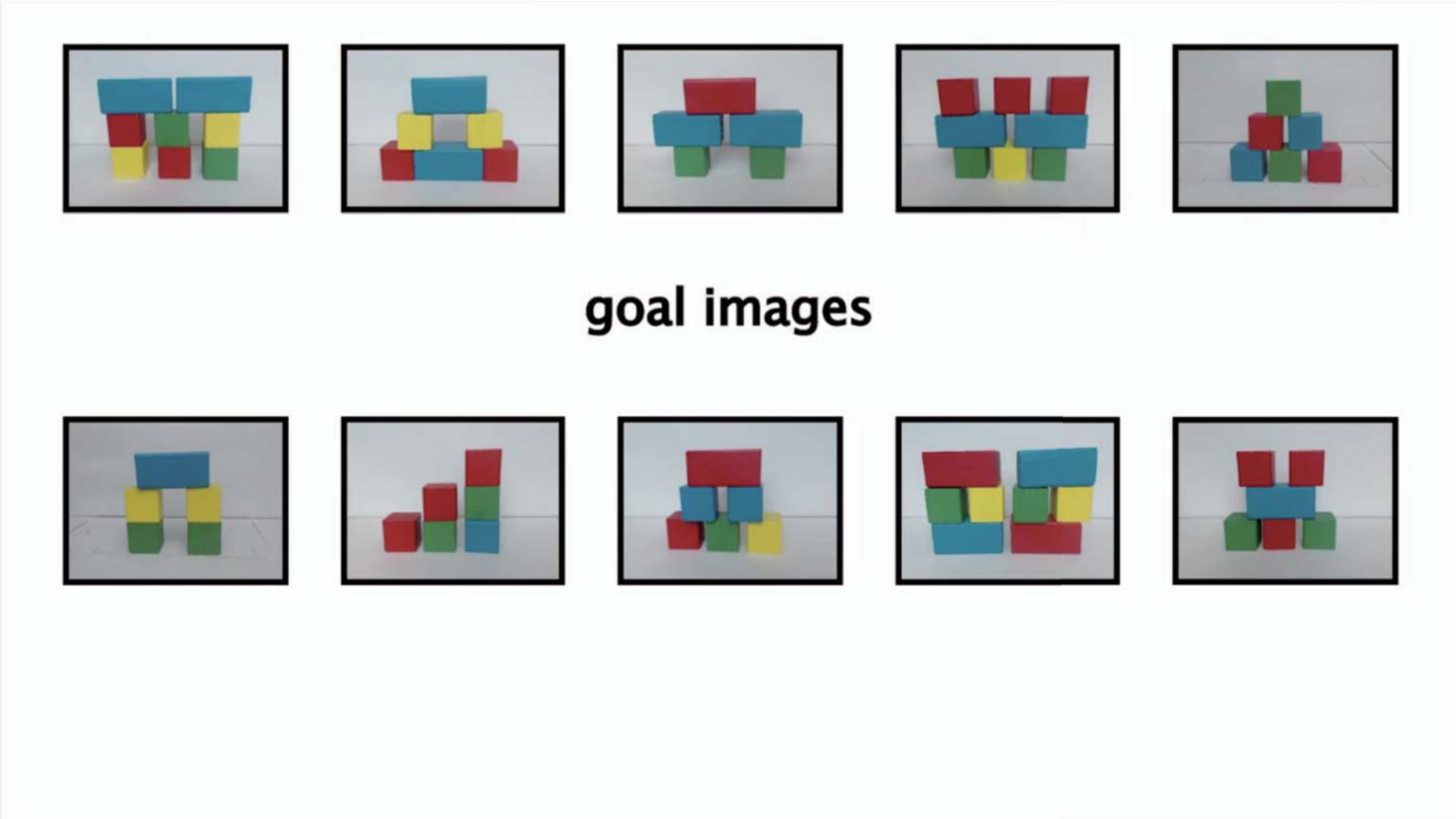
Xu*, Liu*, Sun, Murphy, Freeman, Tenenbaum, Wu. ICLR'19

Ext. I: Unsupervised Structure Discovery



Xu*, Liu*, Sun, Murphy, Freeman, Tenenbaum, Wu. ICLR'19

Ext II: Planning and Control



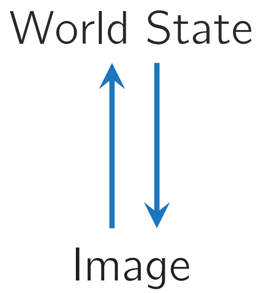
Janner, Levine, Freeman, Tenenbaum, Finn, Wu. ICLR'19

Ext II: Planning and Control

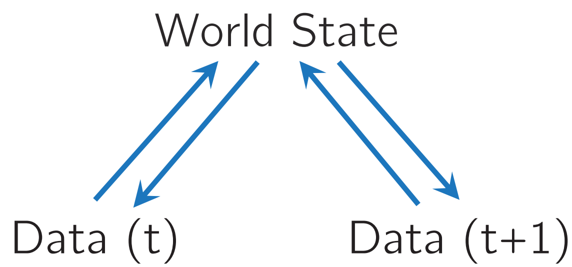
Li, Wu, Tedrake, Tenenbaum, Torralba. ICLR'19

Physical Scene Understanding

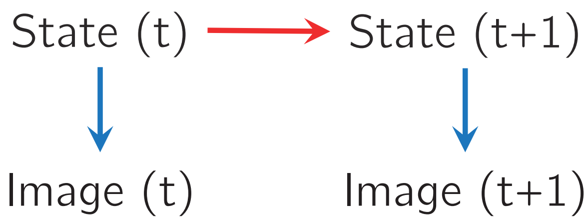
- Learning to invert a graphics engine
  - Inferring fine object geometry
  - Learning structured shape representations (shape + texture)
  - Beyond single object, learning scene representations



- Learning to invert a physics engine
  - Inferring object physical properties
  - Joint modeling of object shape and physics



- Learning simulation engines themselves
  - Learning object dynamics in the pixel space
  - Modeling object dynamics for control



Physical Scene Understanding with Compositional Structure

Goal

- Explaining and reasoning about data

Approach

- Leveraging causal structure to integrate generative, forward models with efficient inference algorithms.

Advantages

- 1. Guiding and facilitating model design.
- 2. Allowing learning with little or no supervision.
- 3. Offering rich generalization power.