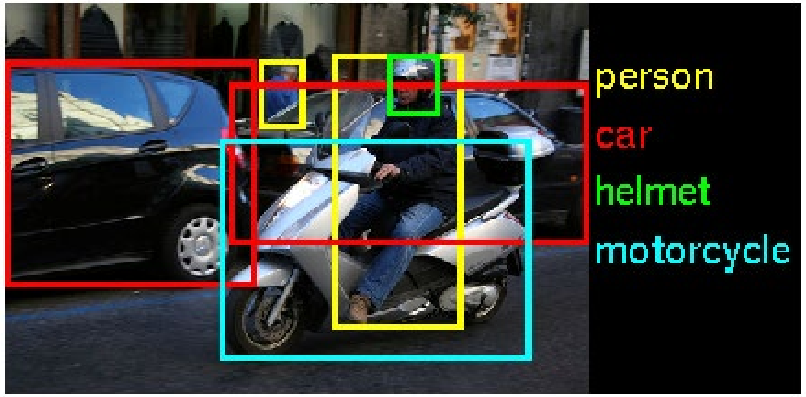
Physical Scene Understanding

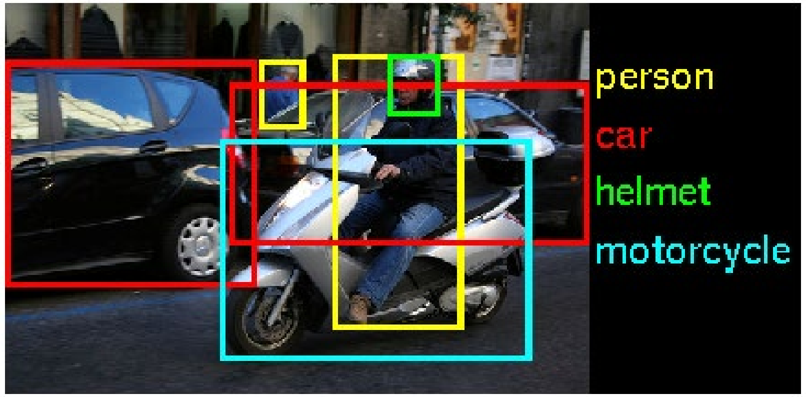
Jiajun Wu
MIT, Stanford

Scene Understanding

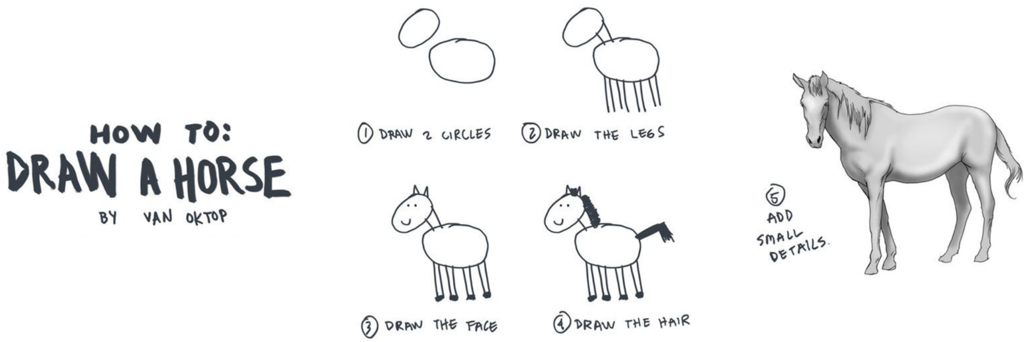


Recognition

Scene Understanding

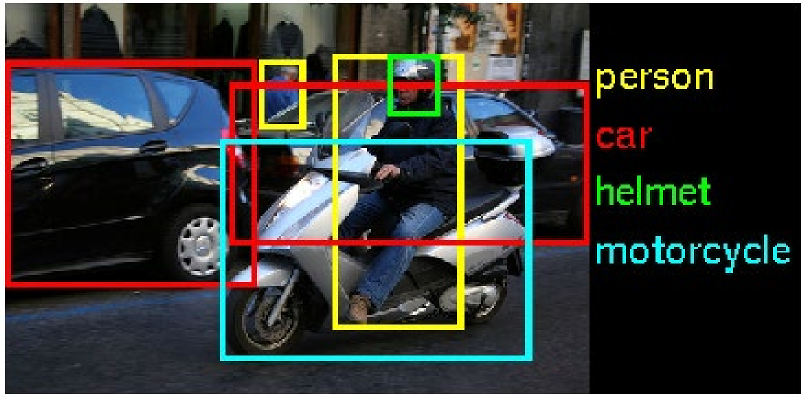


Recognition

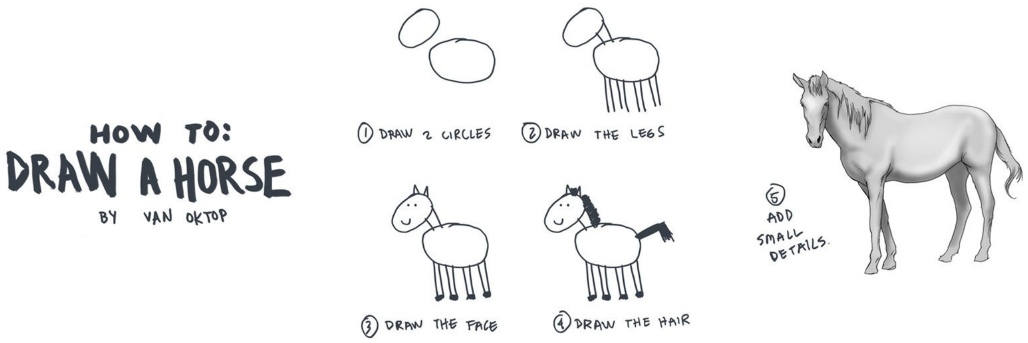


Interpretation

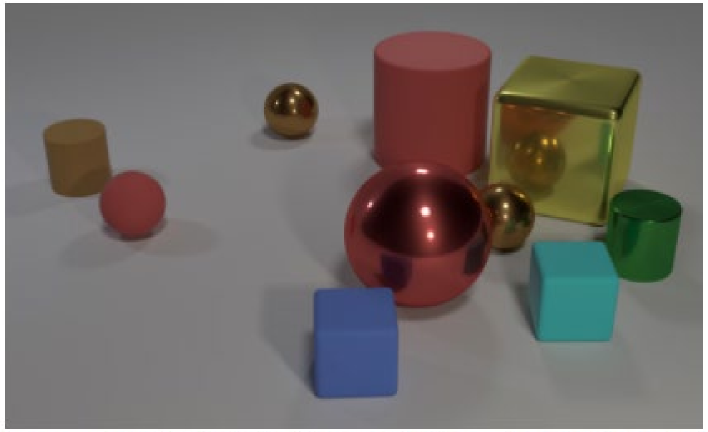
Scene Understanding



Recognition



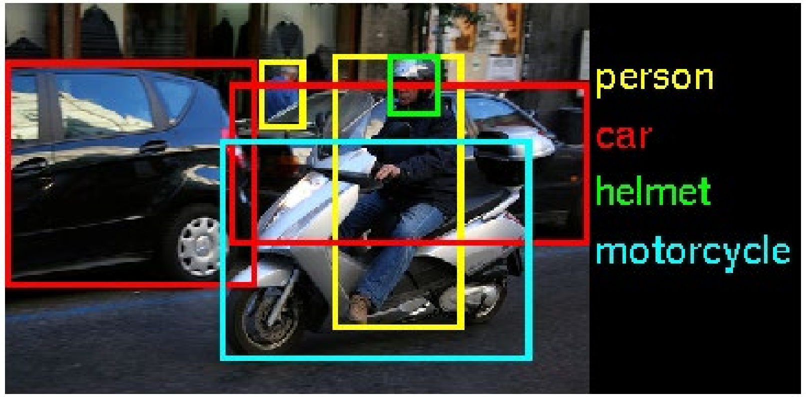
Interpretation



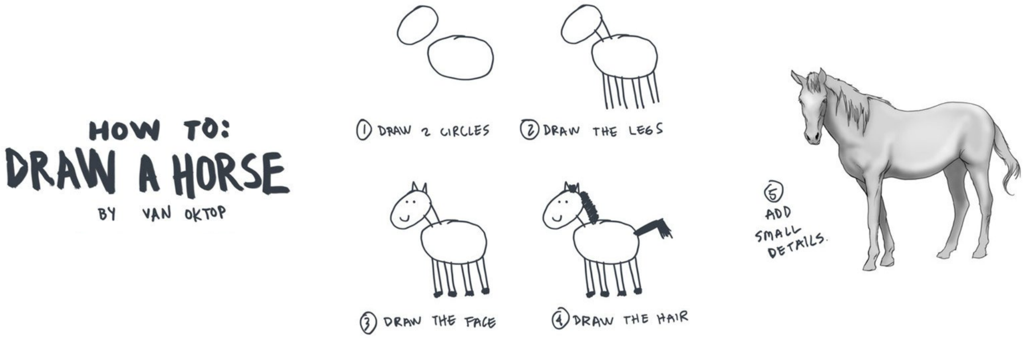
Q: Are there an equal number of large things and metal spheres?

Reasoning

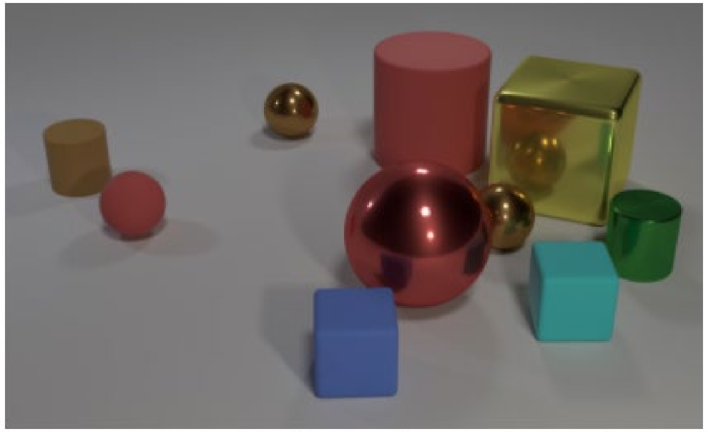
Scene Understanding



Recognition



Interpretation



Q: Are there an equal number of large things and metal spheres?

Reasoning

Generation

Physical Scene Understanding

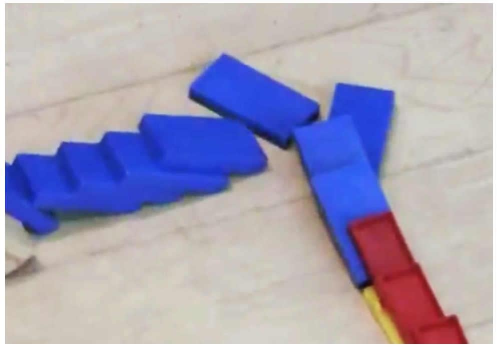
- What can we learn from this video?
  - 3D object shapes (geometry)

Physical Scene Understanding

- What can we learn from this video?
  - 3D object shapes (geometry)
  - Object properties (physics)
    - Masses / coefficients of frictions

Physical Scene Understanding

- What can we learn from this video?
  - 3D object shapes (geometry)
  - Object properties (physics)
    - Masses / coefficients of frictions
  - Physical events (interaction)

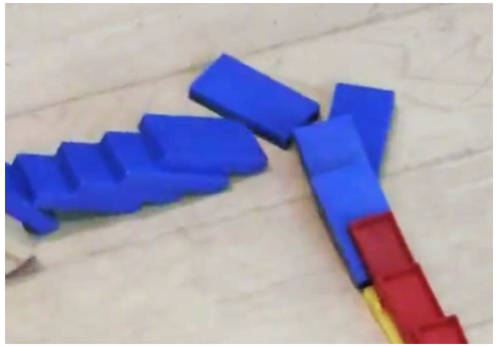


collisions

rolling

Physical Scene Understanding

- What can we learn from this video?
  - 3D object shapes (geometry)
  - Object properties (physics)
    - Masses / coefficients of frictions
  - Physical events (interaction)



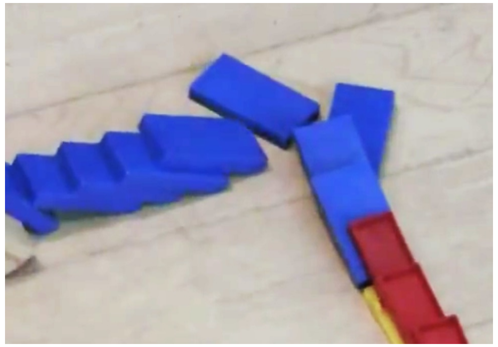
collisions

rolling

- Humans recover rich information from this short video.

Physical Scene Understanding

- What can we learn from this video?
  - 3D object shapes (geometry)
  - Object properties (physics)
    - Masses / coefficients of frictions
  - Physical events (interaction)

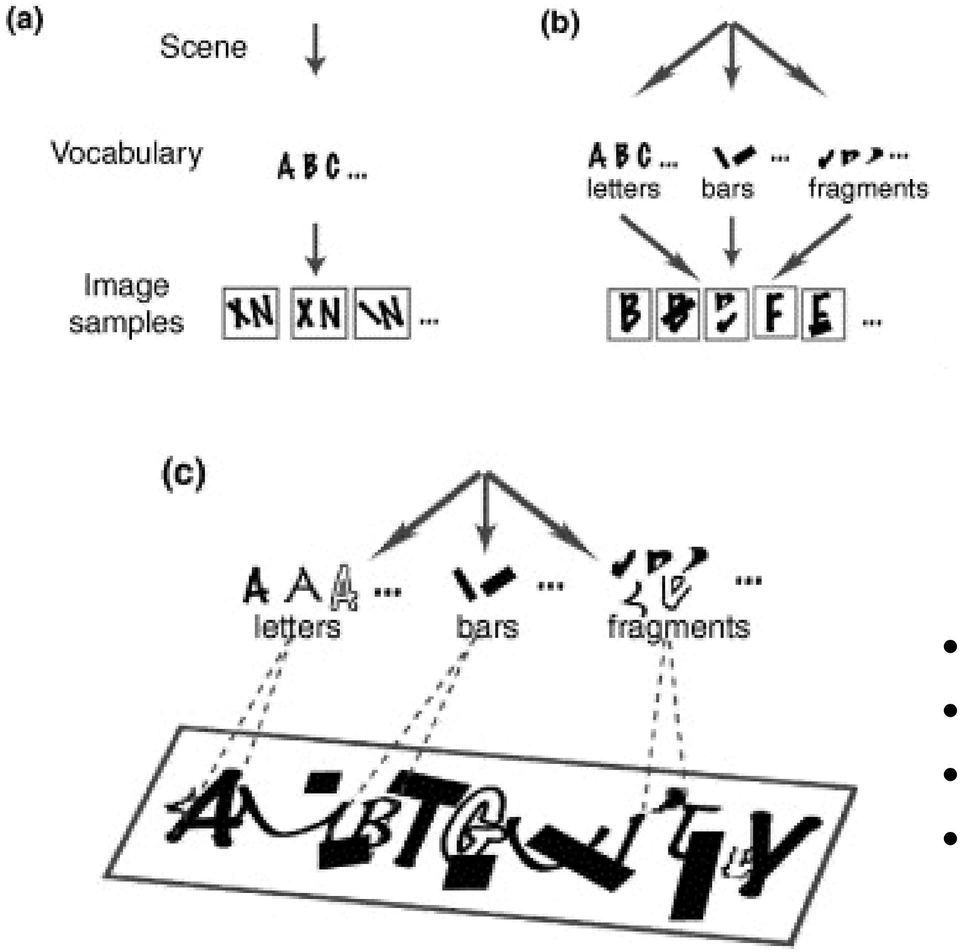


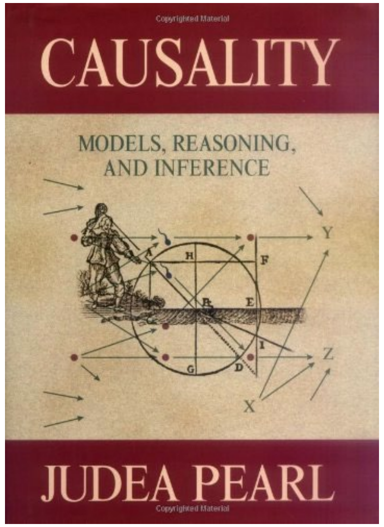
collisions

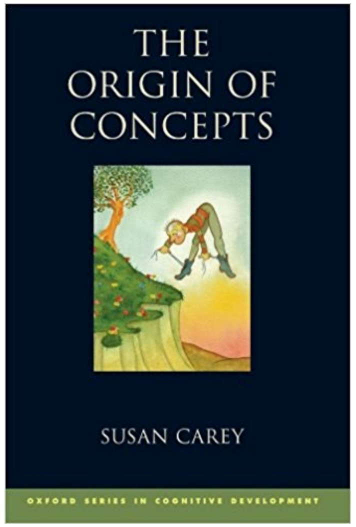
rolling

- Humans recover rich information from this short video.
- Generalization: Humans easily answer questions like
  - What will happen next?
  - What if … ?
  - How to … ?

Causal Models for Vision



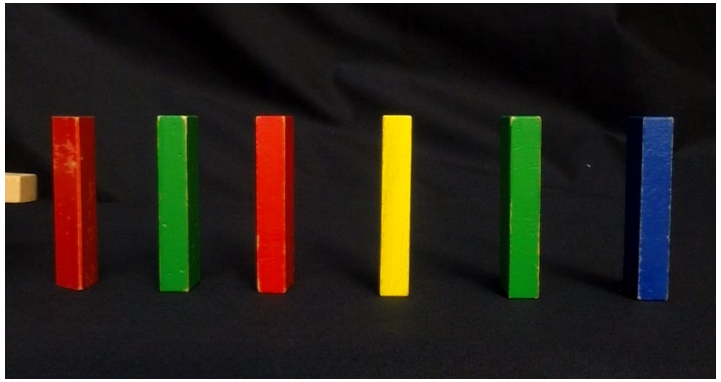




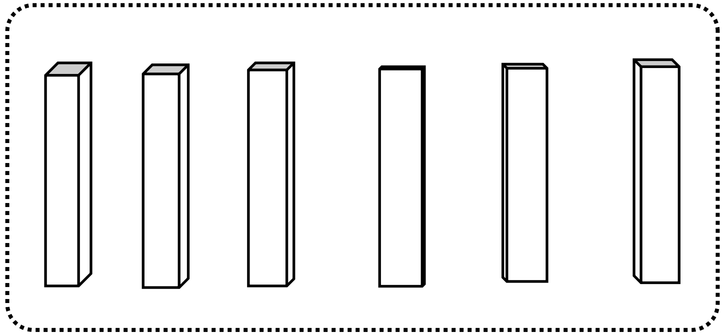
- Helmholtz. Treatise on Physiological Optics. 1867.
- Pearl. Causality. 2000.
- Carey. The Origin of Concepts. 2009.
- Yuille and Kersten. Vision as Bayesian inference: analysis by synthesis? Trends in Cognitive Science, 2006.

Modeling the Physical World

Image (t-1)

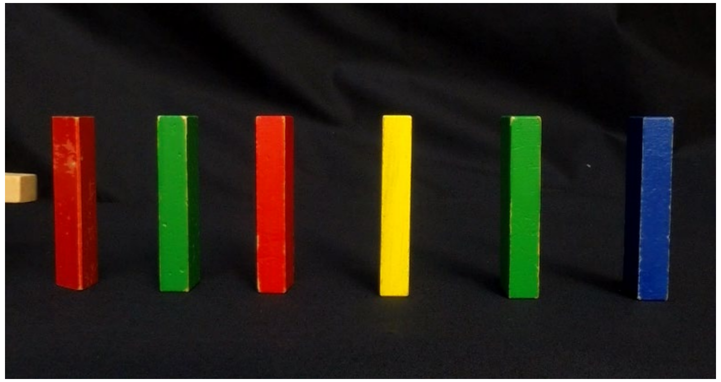


Modeling the Physical World

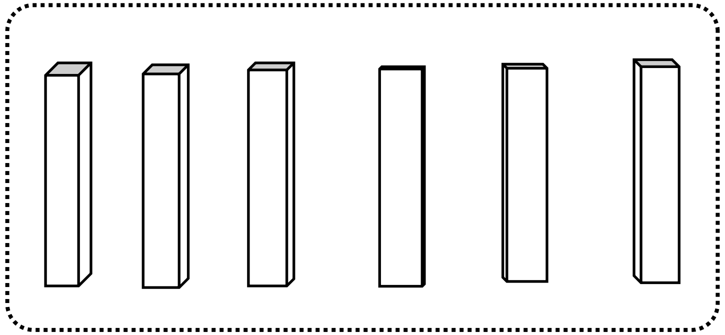


World state (t-1)

Image (t-1)

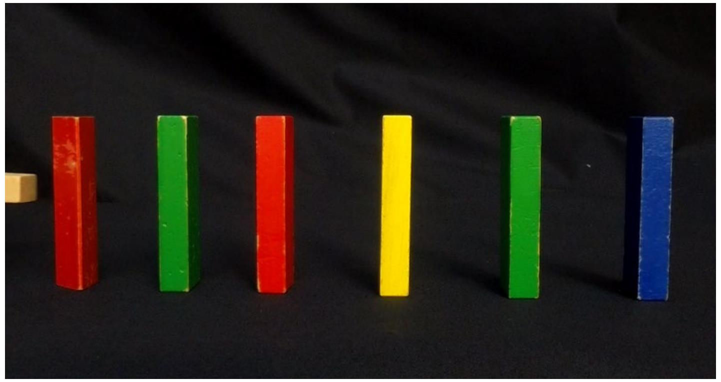


Modeling the Physical World



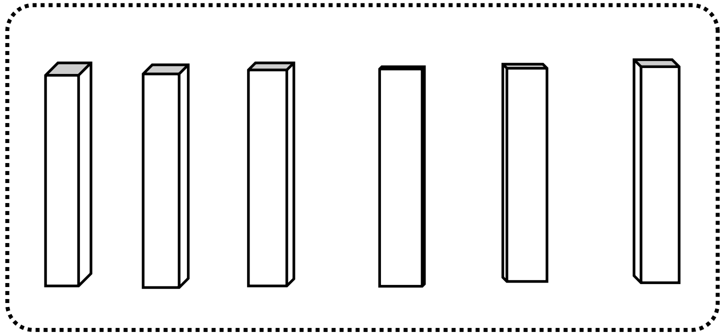
World state (t-1)

Image (t-1)



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

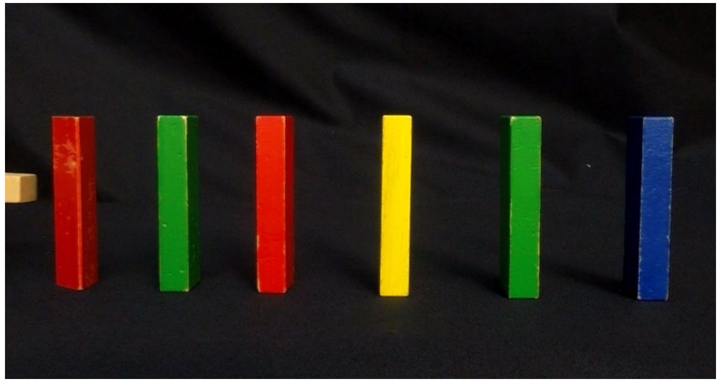
Modeling the Physical World



World state (t-1)

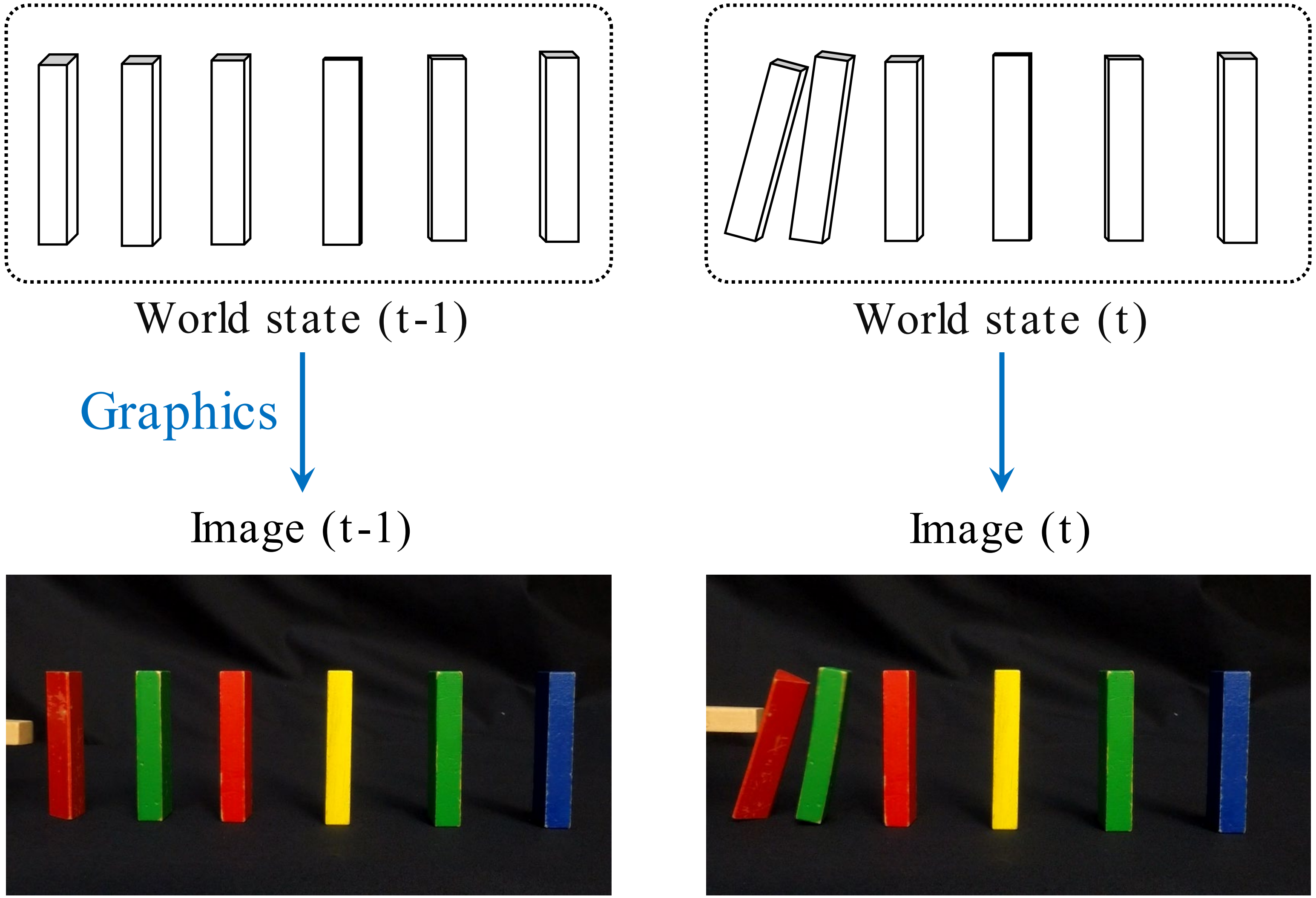
Graphics

Image (t-1)

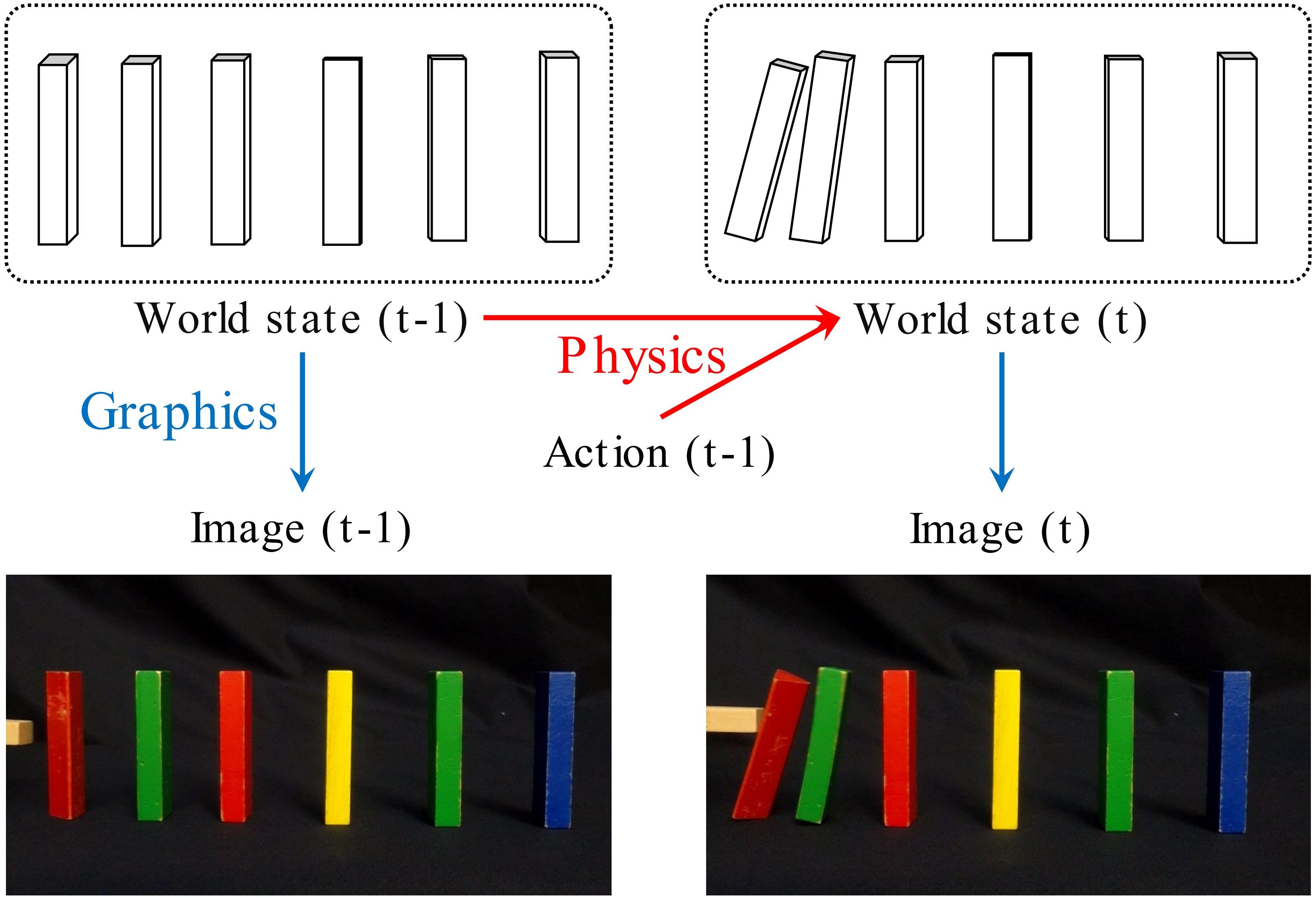


- Object Intrinsics
  - Geometry
  - Physical properties
- Object Extrinsics
  - Position
  - Velocity
- Scene Descriptions
  - Lighting
  - Camera parameters

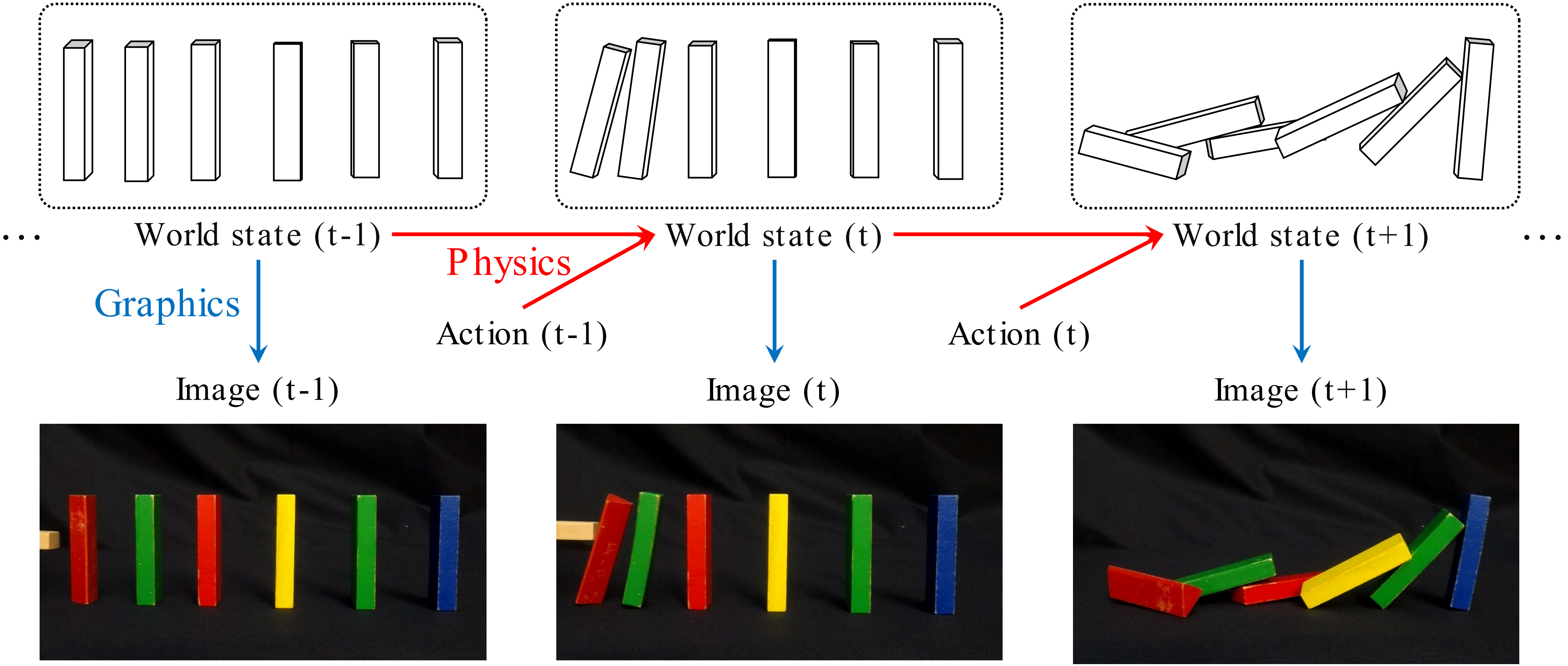
Modeling the Physical World



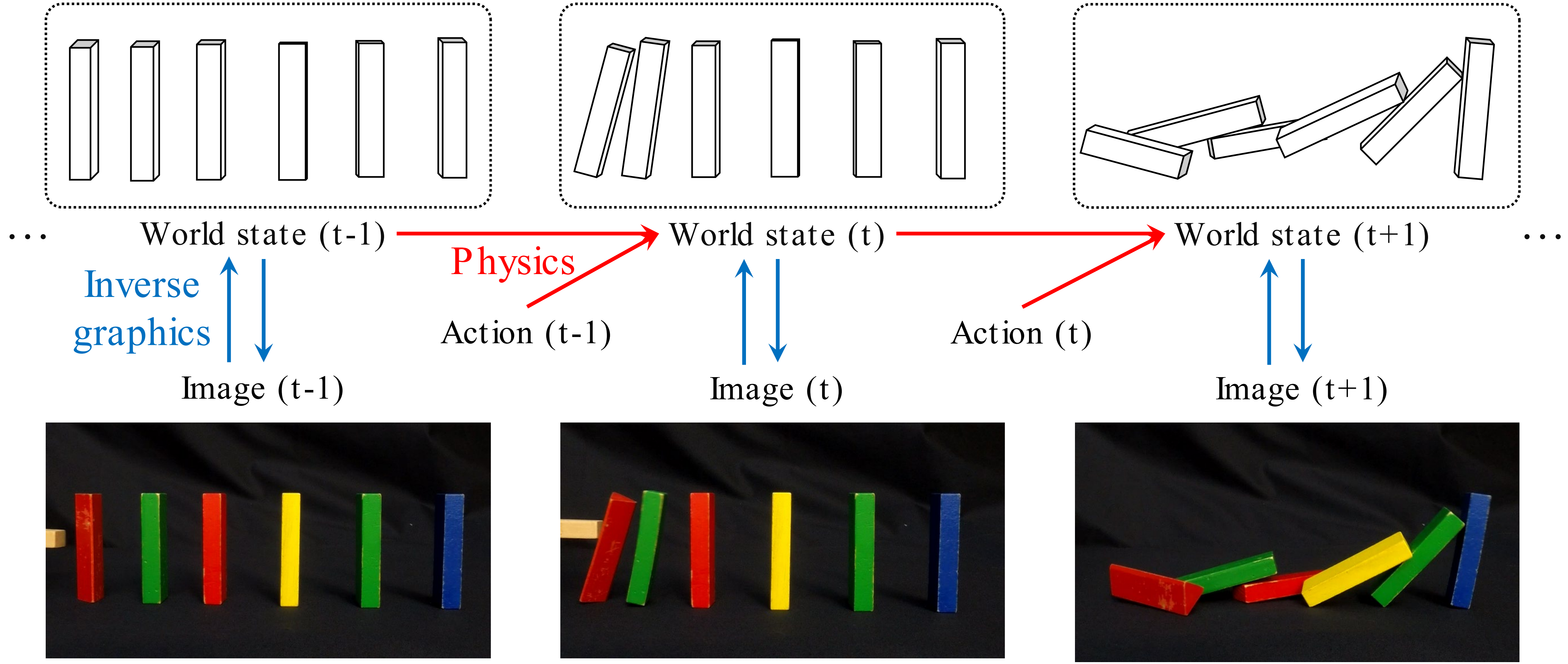
Modeling the Physical World



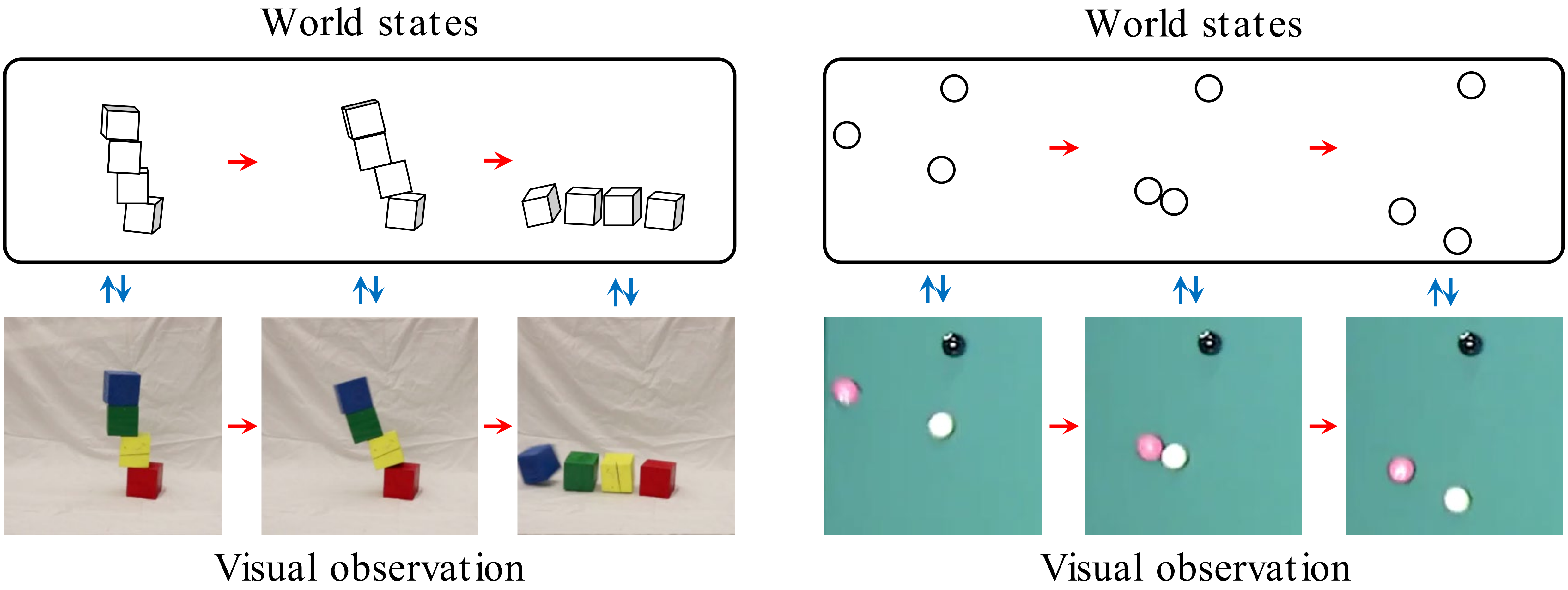
Modeling the Physical World



Modeling the Physical World

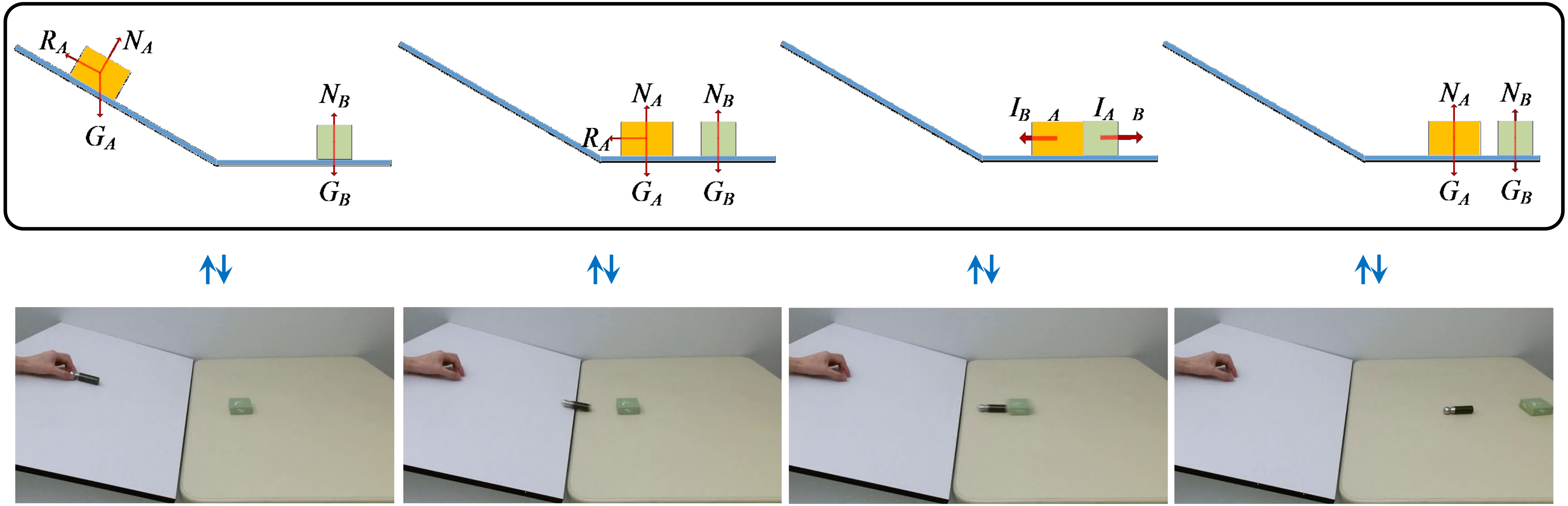


Physical World Representations are Universal



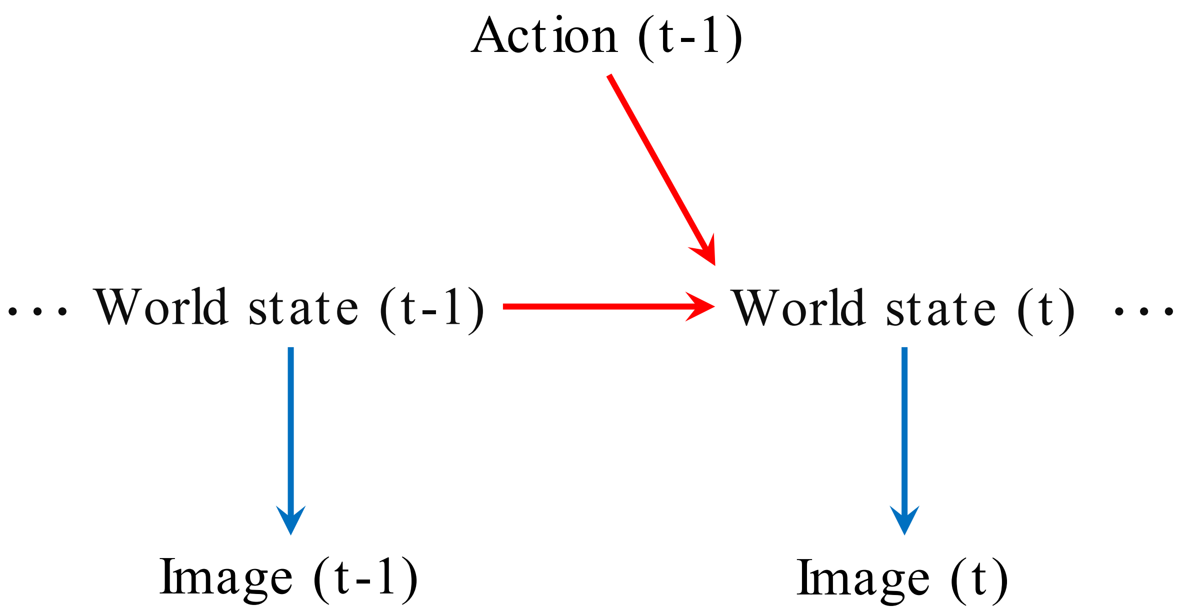
Physical World Representations are Universal

World states



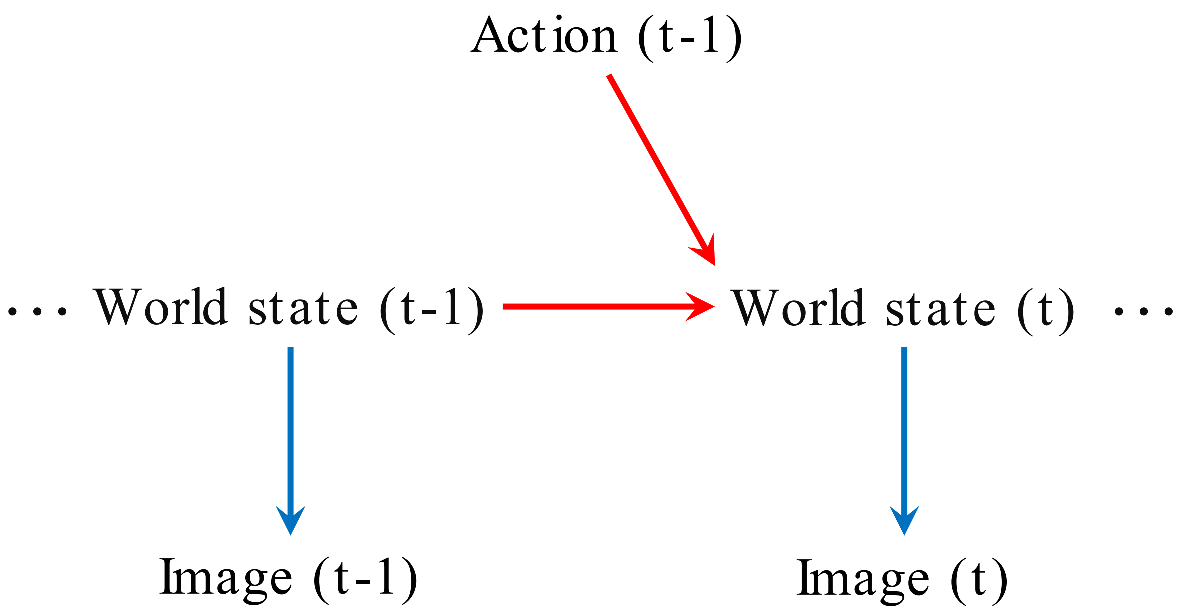
Visual observation

Approach I: Graphical Models



- Classical Estimation and Control, Graphical Models (HMMs, Bayes Nets)
  - **Pro:** Optimized for certain inference and learning algorithms
  - **Con:** Limited expressiveness

Approach I: Graphical Models, Simulation Engines



- Classical Estimation and Control, Graphical Models (HMMs, Bayes Nets)
  - **Pro:** Optimized for certain inference and learning algorithms
  - **Con:** Limited expressiveness
- Simulation (Graphics/ Physics) Engines, Probabilistic Programs
  - **Pro:** Flexible, rich representations
  - **Con:** Lacking efficient, general-purpose inference and learning algorithms

Approach II: End -to-End Deep Learning

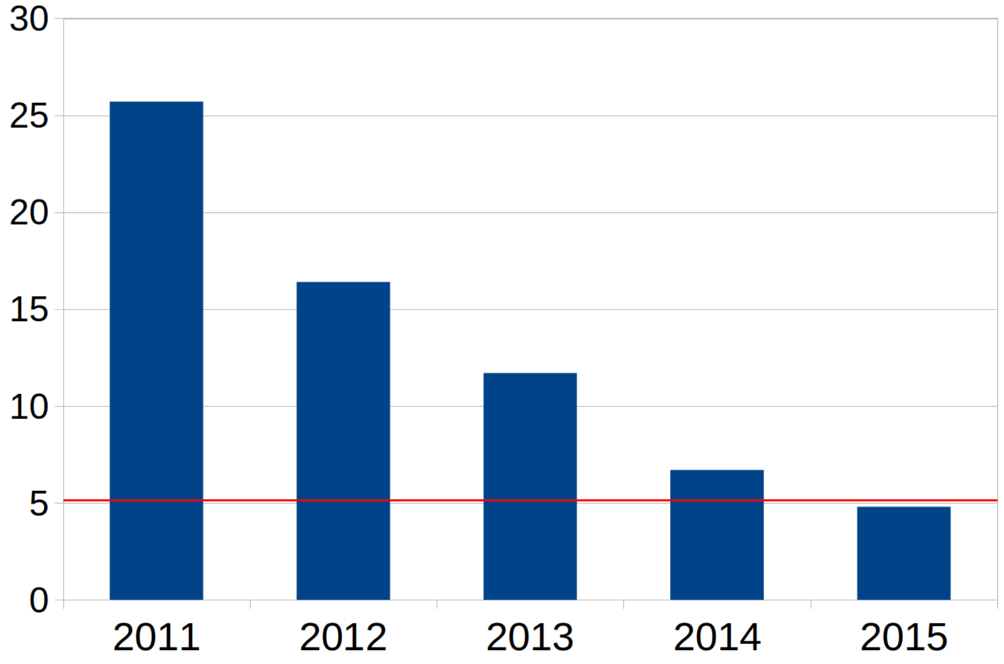
Action (t-1)

… World state (t-1)　　　World state (t) …

↑　　　↑

Image (t-1)　　　Image (t)

IMAGENET　Image Classification



Approach II: End -to-End Deep Learning

Action (t-1)

… World state (t-1)

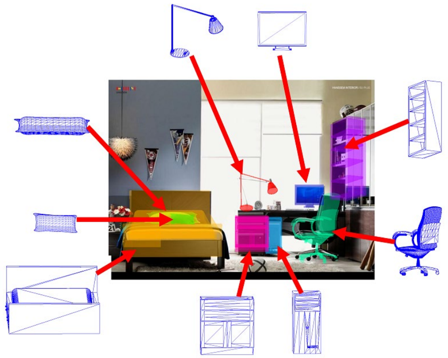
World state (t) …

Image (t-1)

Image (t)

IMAGENET — Image Classification

SHAPENET — Shape Modeling



ObjectNet3D — Shape Reconstruction

Approach II: End -to-End Deep Learning

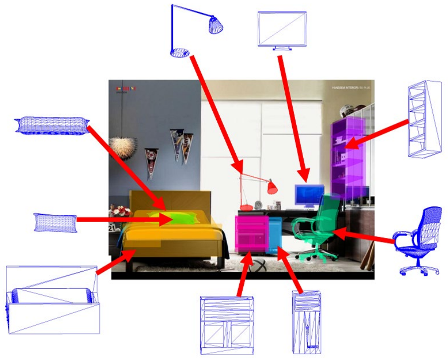
Action (t-1)

… World state (t-1) World state (t) …

Image (t-1) → Image (t)

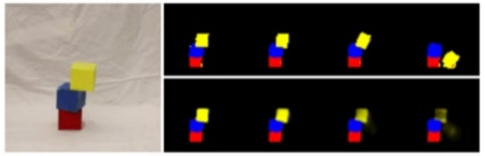
IMAGENET Image Classification

SHAPENET Shape Modeling



ObjectNet3D

Shape Reconstruction



PhysNet

State Prediction

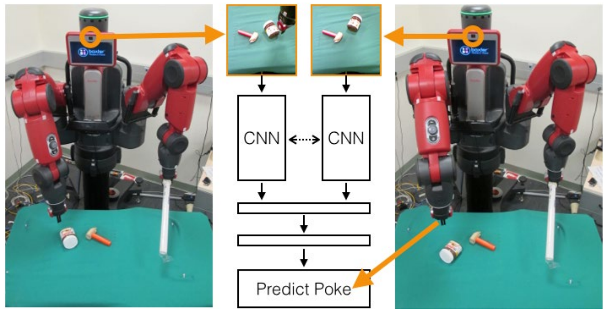
Approach II: End -to-End Deep Learning

Action (t-1)

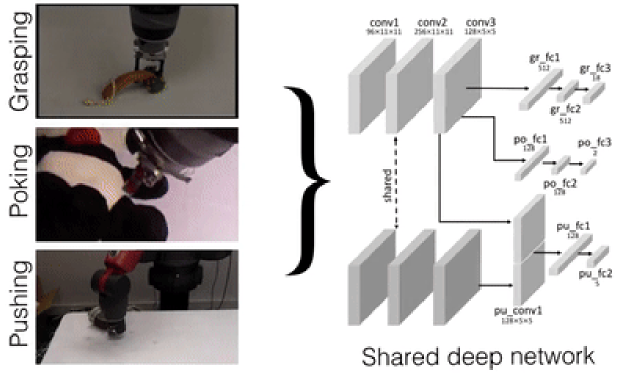
… World state (t-1) World state (t) …

Image (t-1) Image (t)

Modeling actions with deep networks

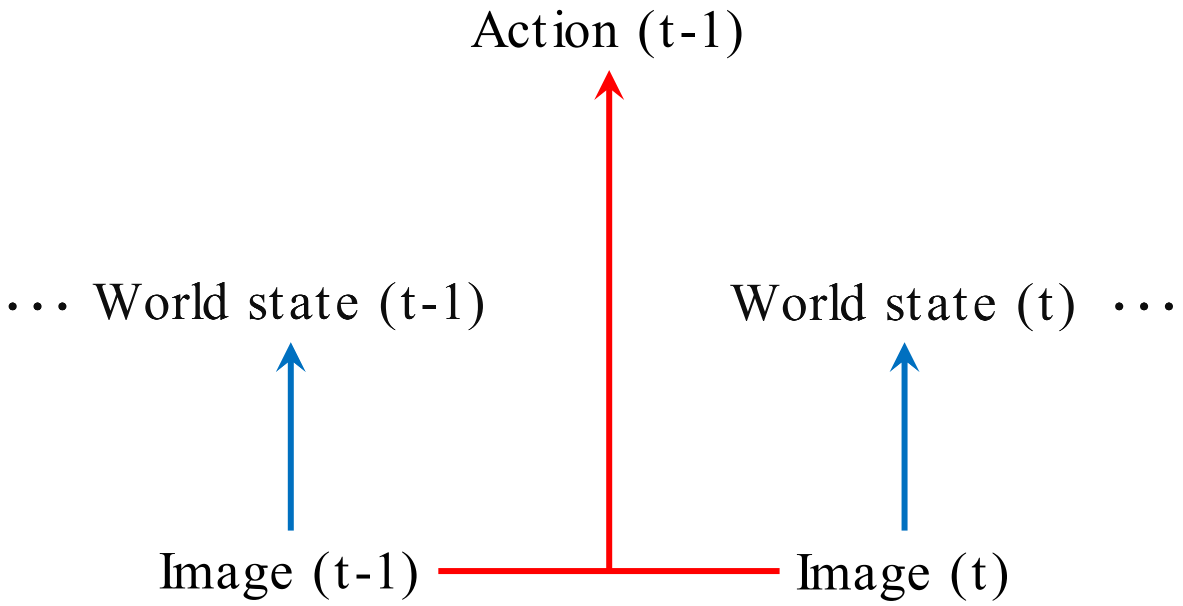


Learning to poke by poking, NIPS'16



Learning to push by grasping, ICRA'17
Learning to fly by crashing, IROS'17

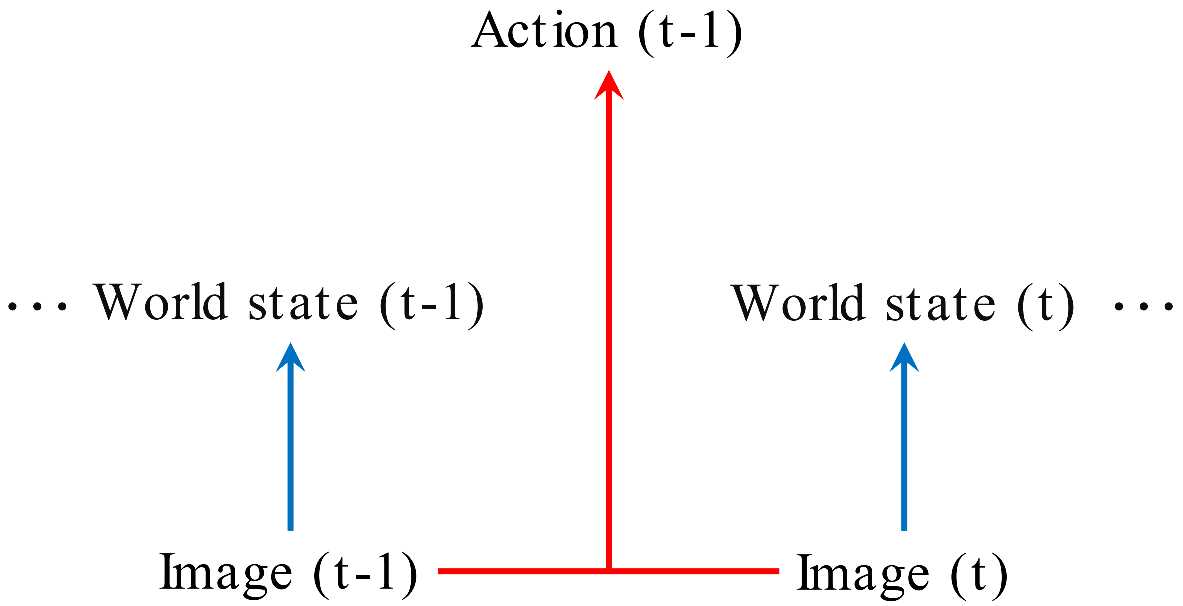
Approach II: End -to-End Deep Learning



Pros

- Efficient, model-agnostic inference
- Scales with large labeled datasets

Approach II: End -to-End Deep Learning

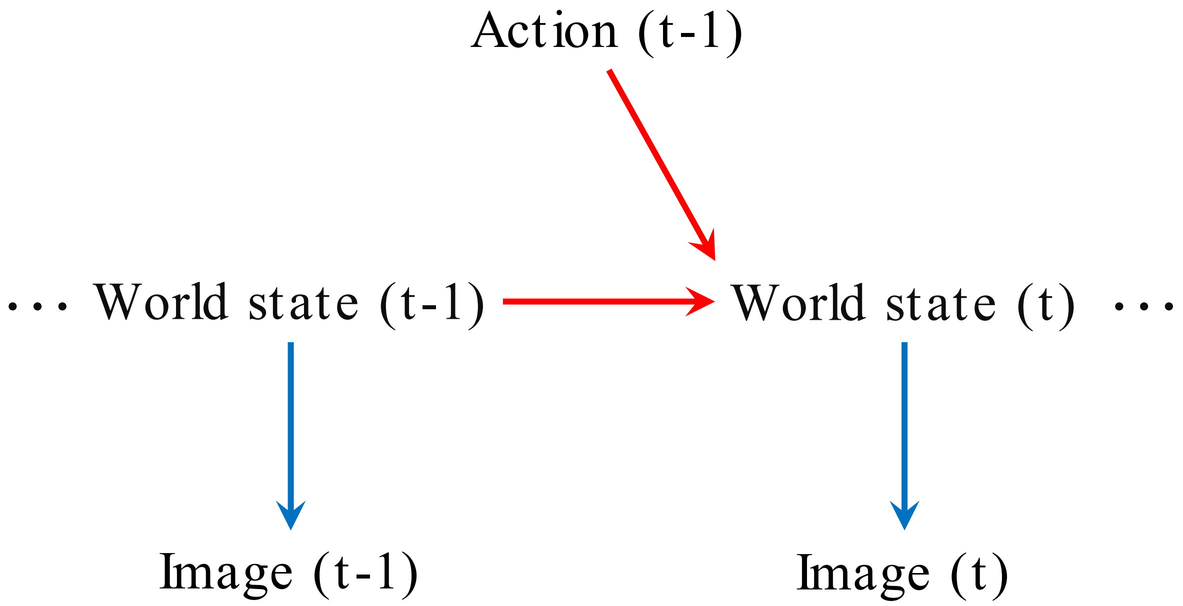


Pros

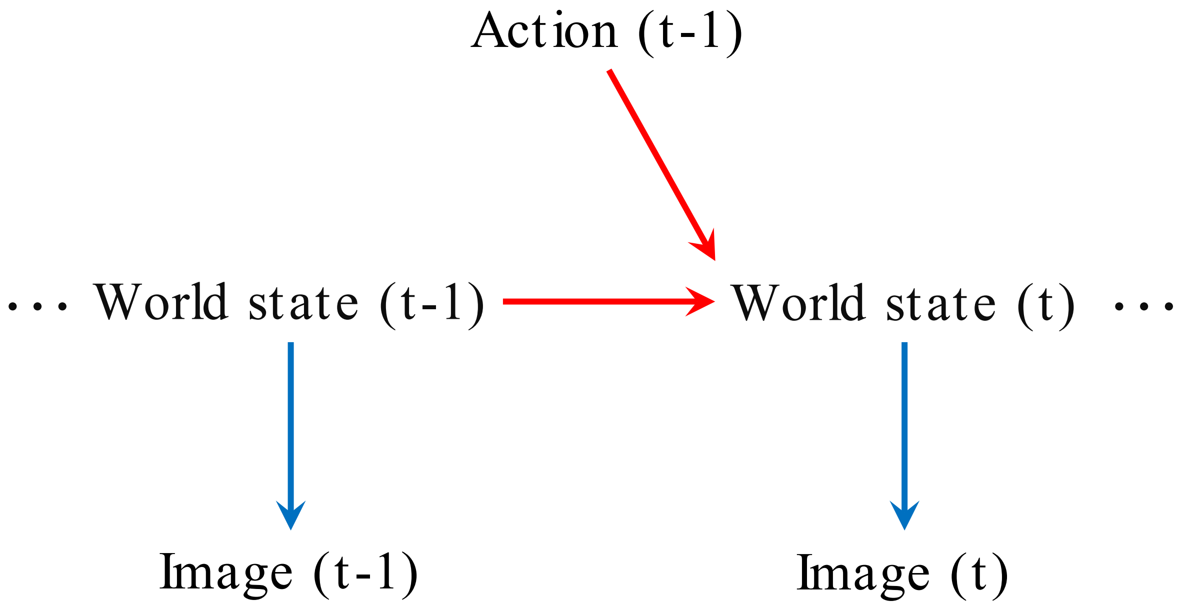
- Efficient, model-agnostic inference
- Scales with large labeled datasets

- Generalization?
  - Limited in generalizing outside training

Leveraging Causal Structure to Combine the Best of Both



Leveraging Causal Structure to Combine the Best of Both

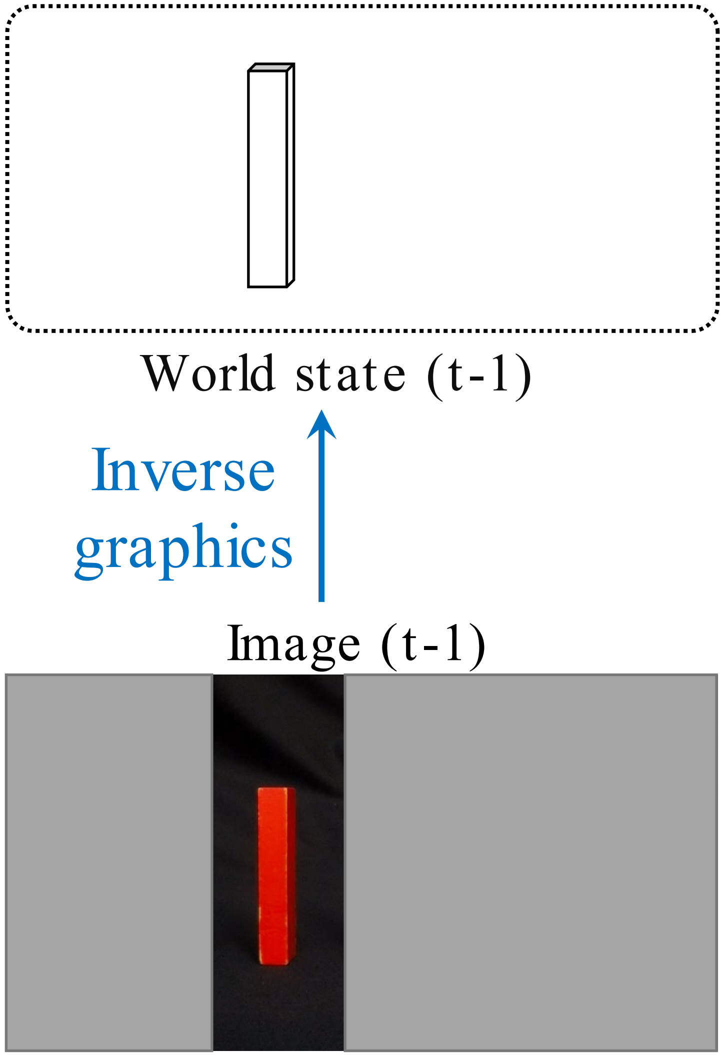


Key Idea: Conditional Independence

Provides guidance on combining neural networks with simulation engines.

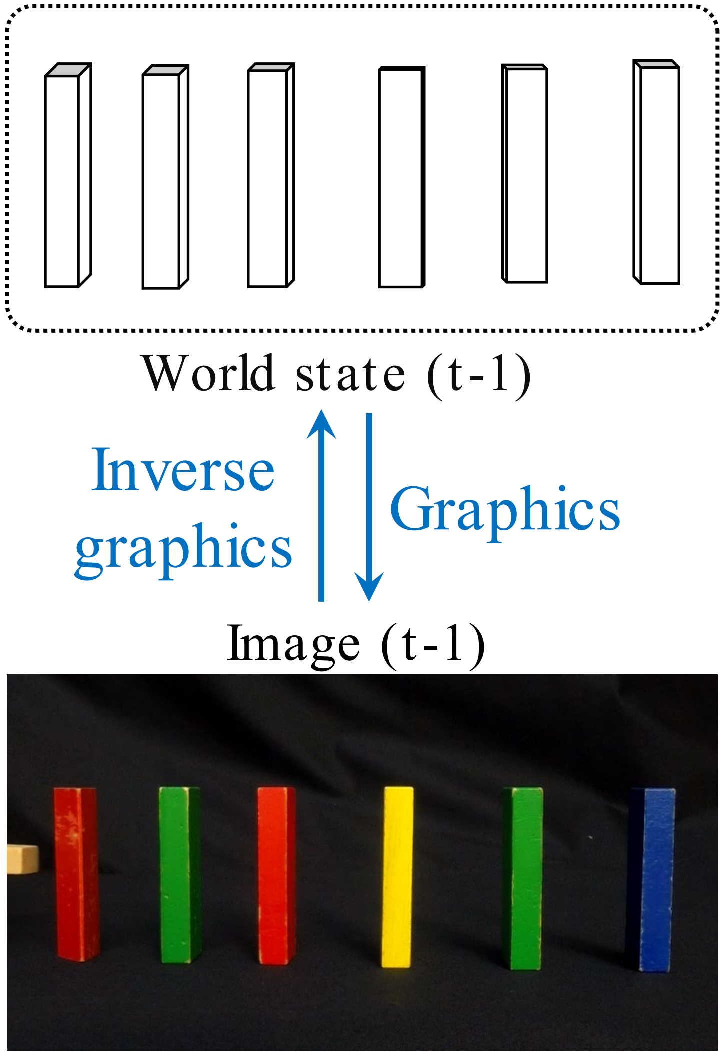
- When and where to use simulation engines vs. neural networks?
- What training targets to use for neural networks?
- What intermediate representations to use in the neural networks?
- What training data to use for neural nets?

Outline



- Single Object
  - 3D Shape [NIPS'17]
  - Intrinsic Images [NIPS'17]

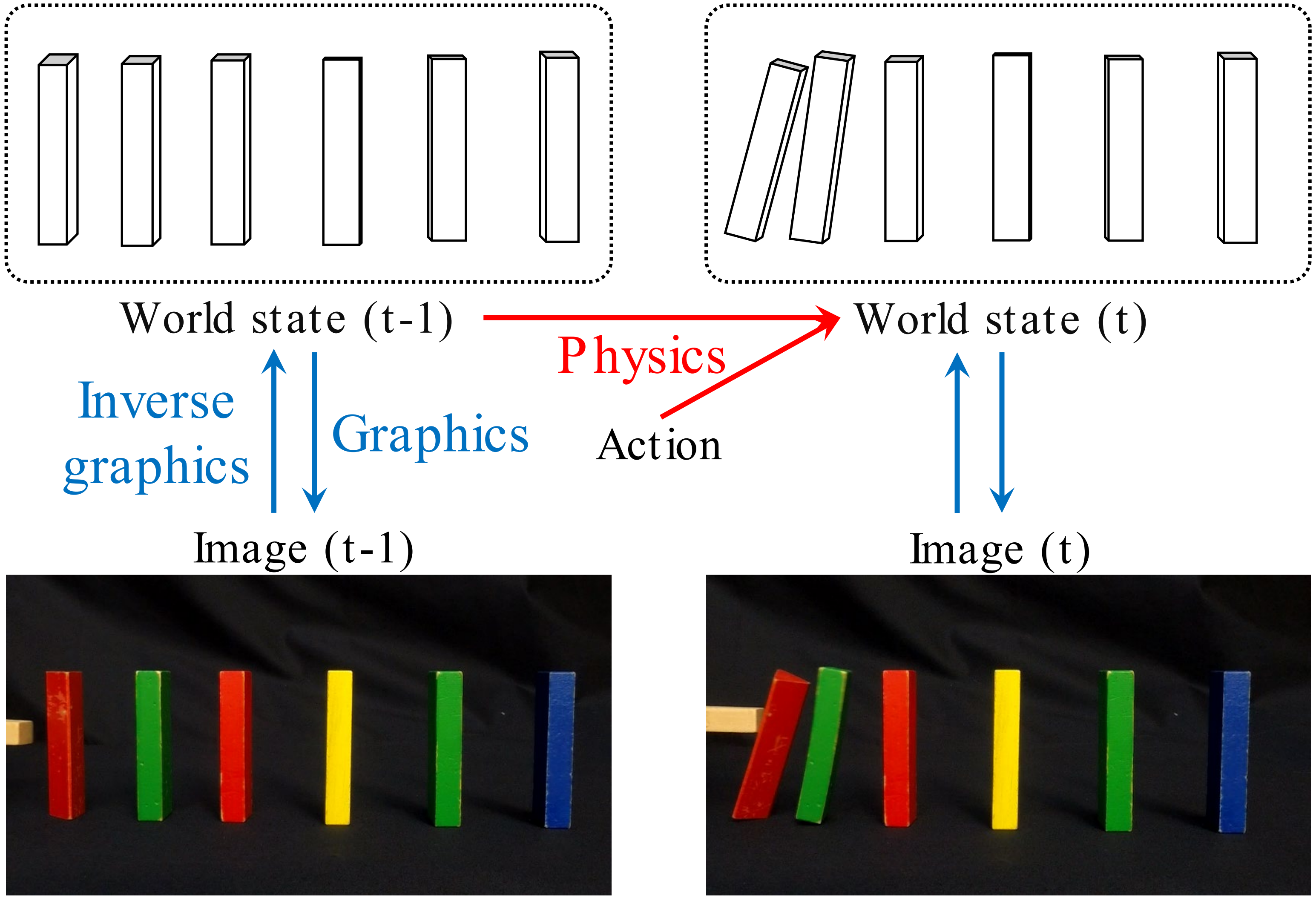
Outline



- Single Object
  - 3D Shape [NIPS'17]
  - Intrinsic Images [NIPS'17]

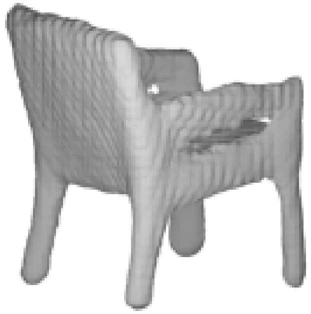
- Static Scene
  - Scene de-rendering [CVPR'17]

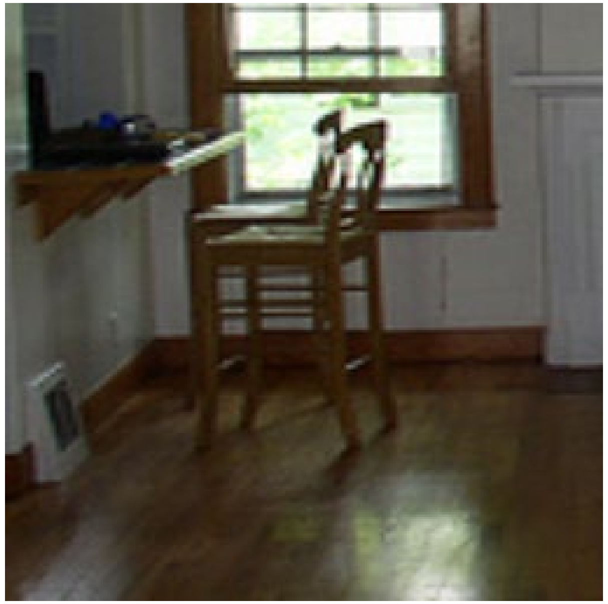
Outline



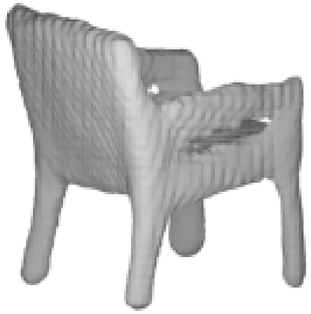
- Single Object
  - 3D Shape [NIPS'17]
  - Intrinsic Images [NIPS'17]
- Static Scene
  - Scene de-rendering [CVPR'17]
- Scene Dynamics
  - Perception + Physics [NIPS'17]
  - Multi-Modal Learning (V + A) [ICCV'17, NIPS'17]

Goal: Single Image 3D Reconstruction



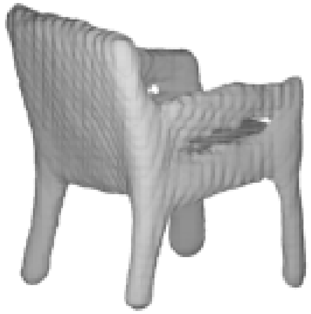


Current Approaches



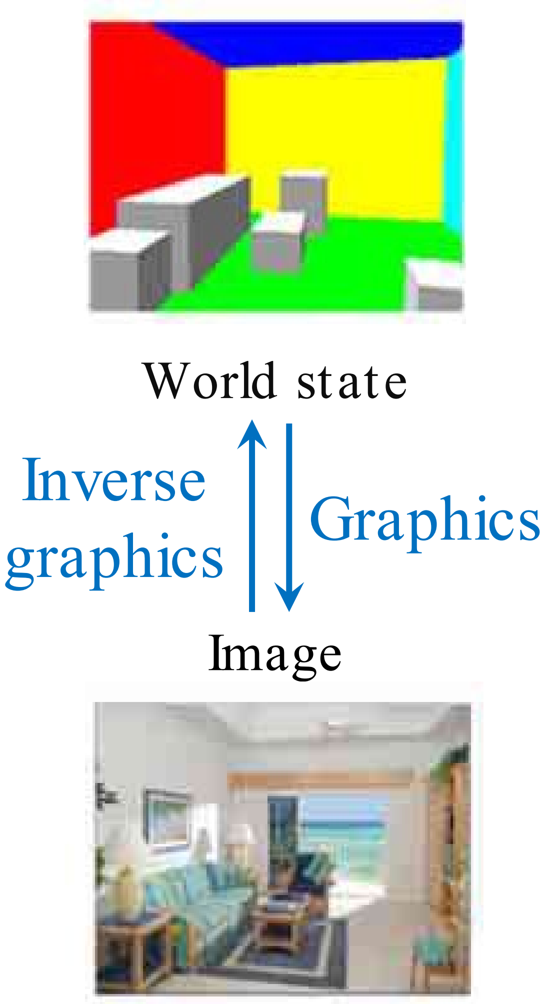
- Reconstruction with neural nets
  - 3D-R2N2 [ECCV'16]
  - TL-Network [ECCV'16]
  - HSP [3DV'17]
  - …

Current Approaches

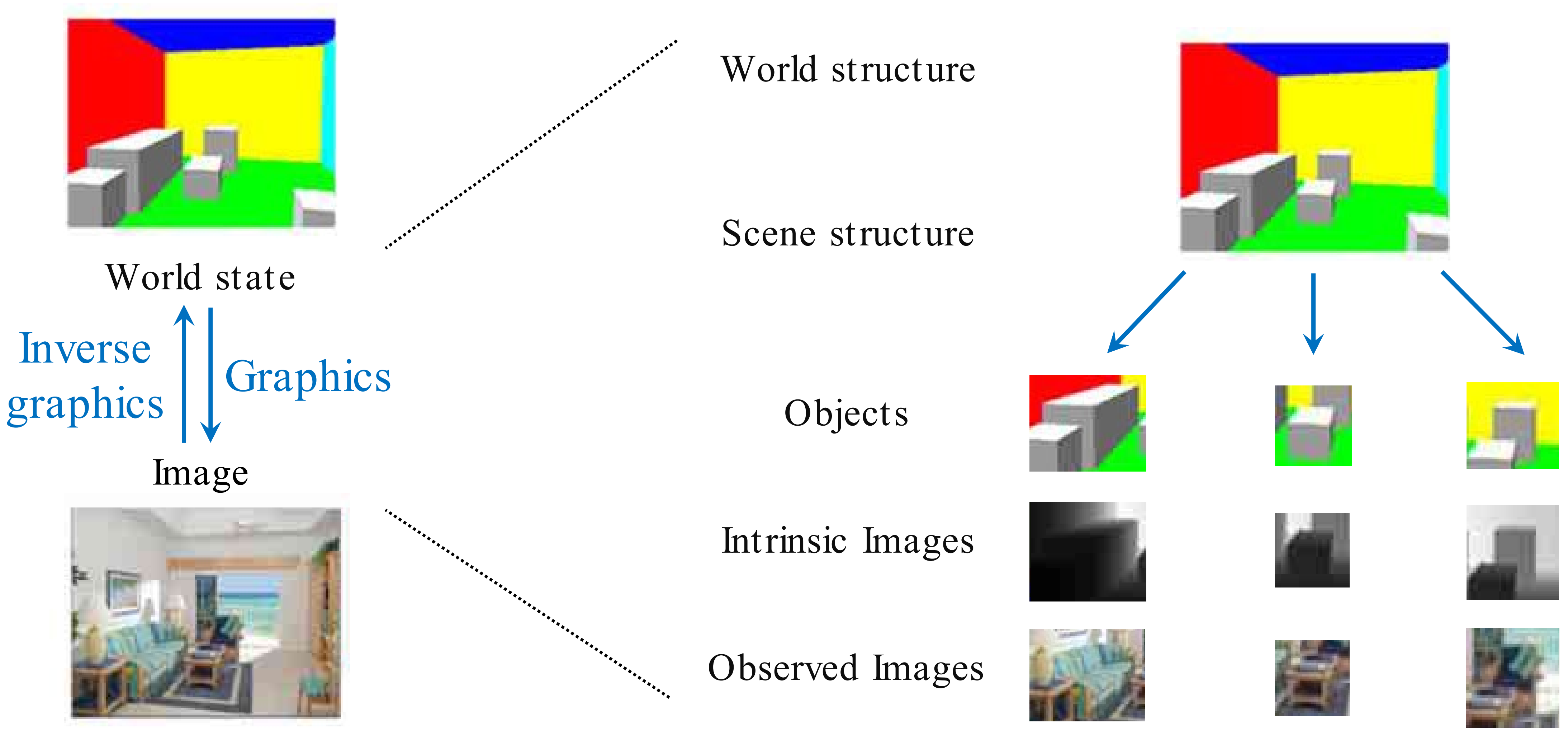


- Reconstruction with neural nets
  - 3D-R2N2 [ECCV'16]
  - TL-Network [ECCV'16]
  - HSP [3DV'17]
  - …
- Reprojection consistency
  - Unsupervised Learning of 3D Structure from Images [NIPS'16]
  - Perspective Transformer Net [NIPS'16]
  - DRC [CVPR'17]
  - …

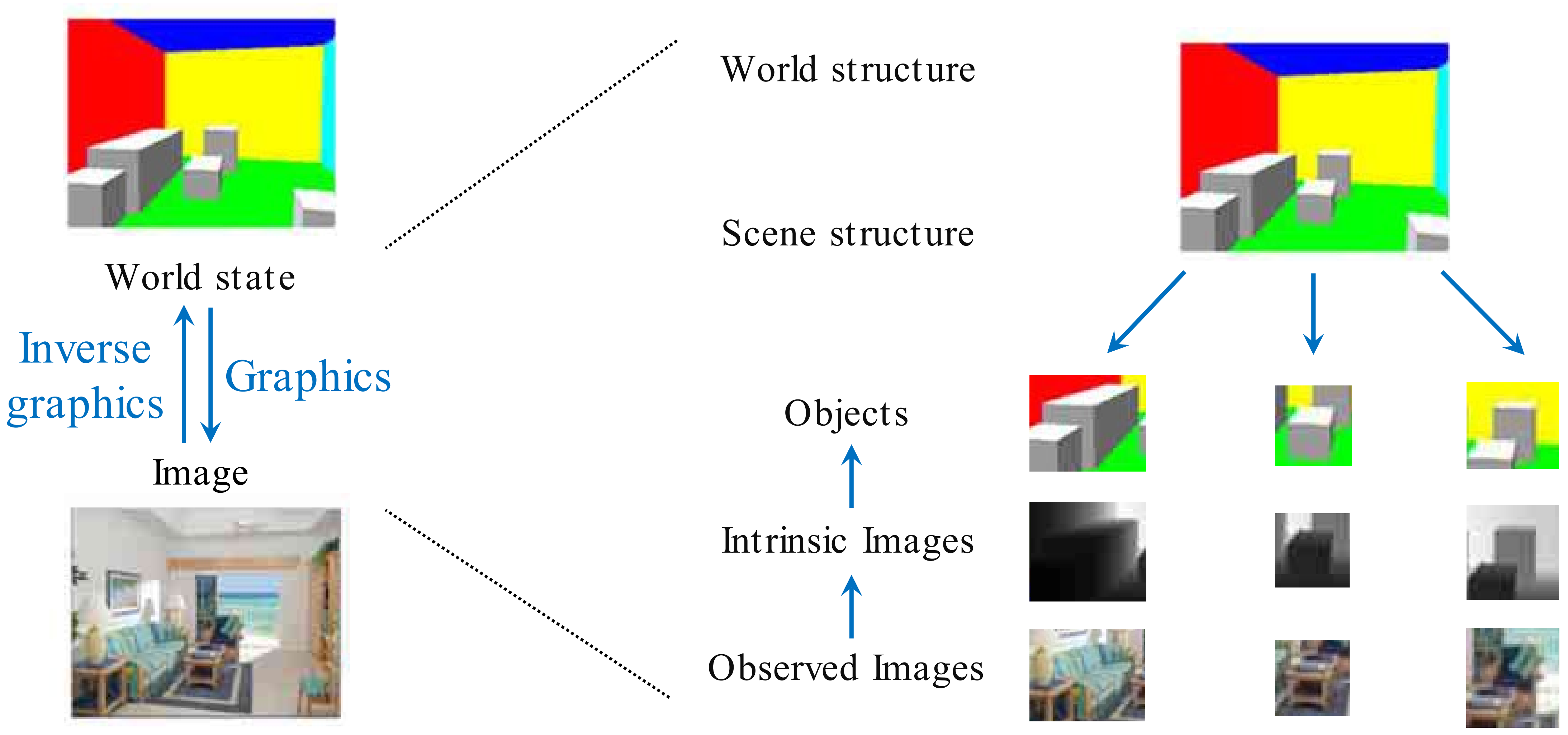
How Computer Graphics Helps Computer Vision



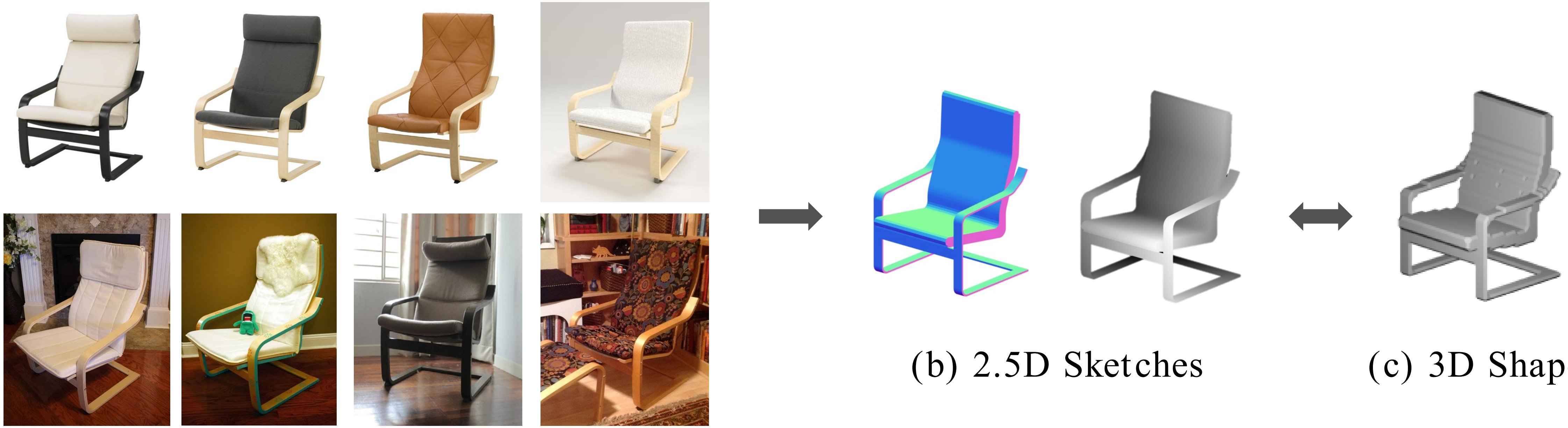
How Computer Graphics Helps Computer Vision



How Computer Graphics Helps Computer Vision



2.5D Sketches as an Intermediate Representation

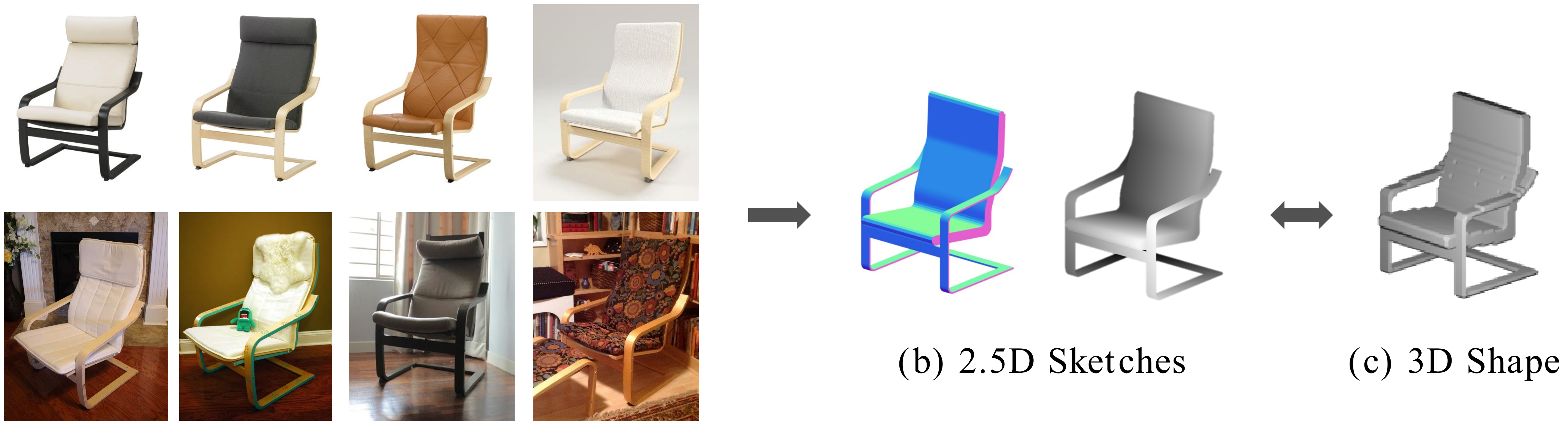


(a) Images

(b) 2.5D Sketches

(c) 3D Shape

2.5D Sketches as an Intermediate Representation

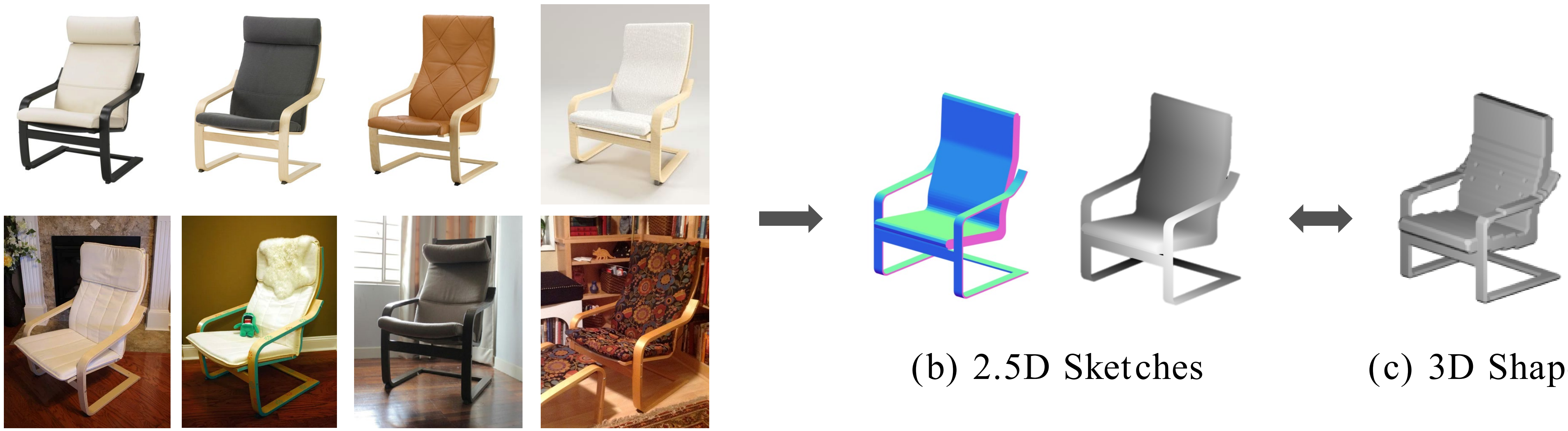


(a) Images (b) 2.5D Sketches (c) 3D Shape

Inquiry had to do with feature-based recognition, how to separate figure from ground, how to extract and interpret a 'form' or 'figure', how much analysis could be done in a data-driven or bottom-up way, and how much needed top-down influences.

Marr. Vision. 1980

2.5D Sketches as an Intermediate Representation



(a) Images

(b) 2.5D Sketches

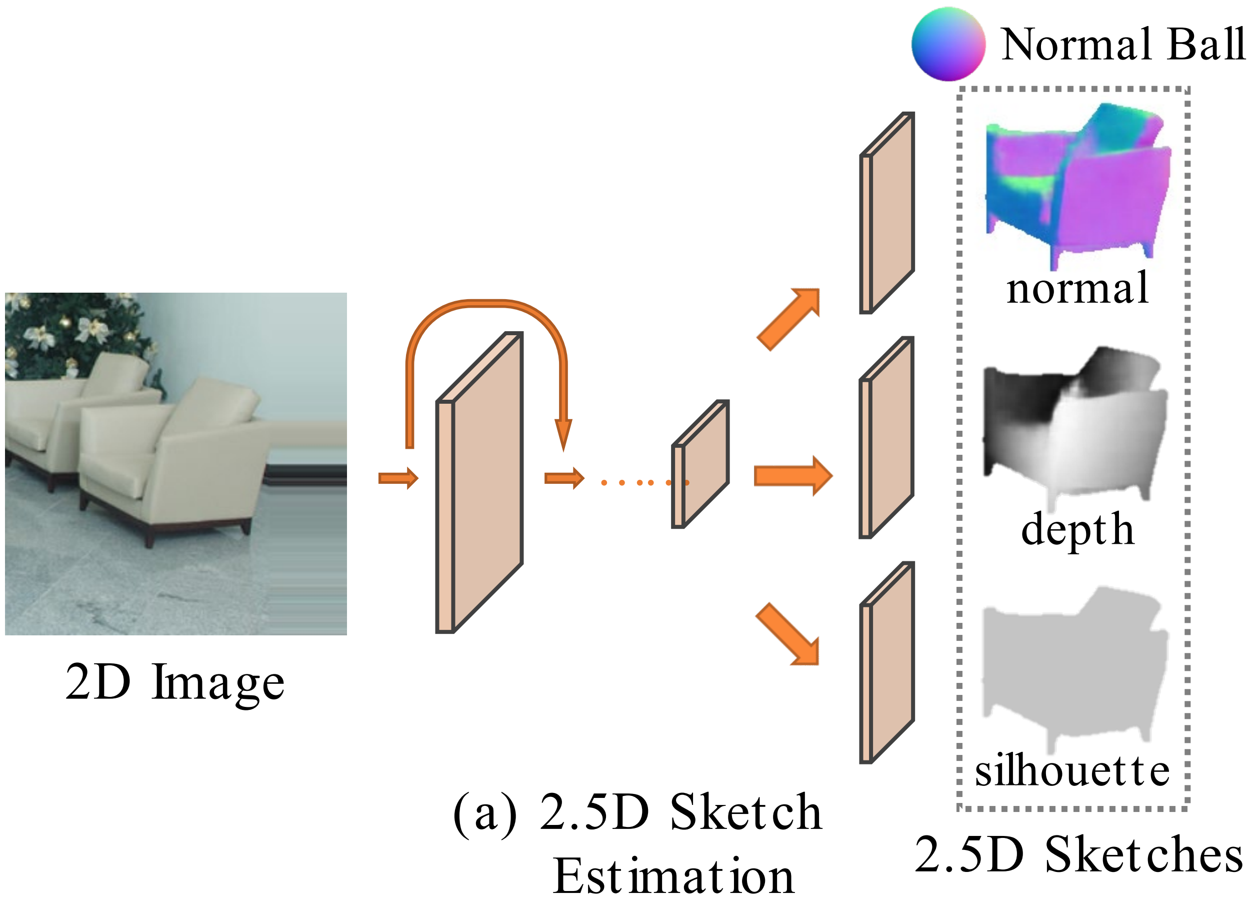
(c) 3D Shape

Inquiry had to do with feature-based recognition, how to separate figure from ground, how to extract and interpret a ‘form’ or ‘figure’, how much analysis could be done in a data-driven or bottom-up way, and how much needed top-down influences.

All this type of thinking was dramatically swept away by the idea of the 2.5D sketch.

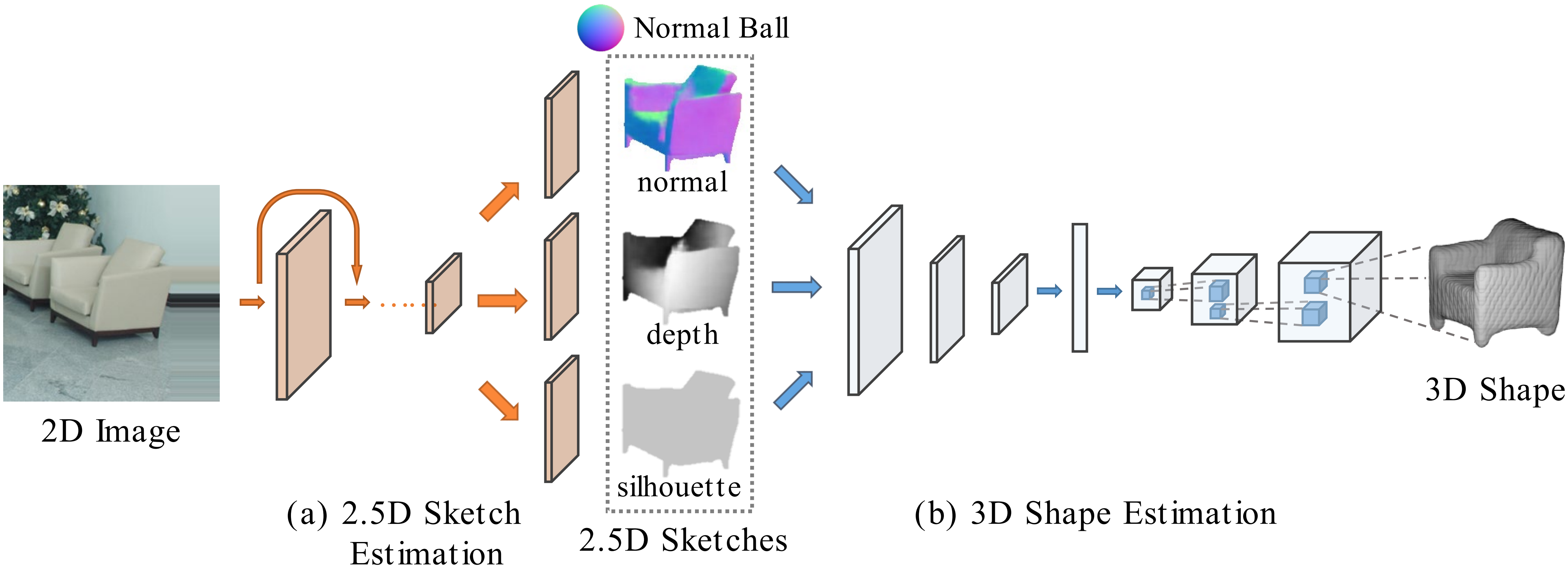
Marr. Vision. 1980

MarrNet : 3D Reconstruction via 2.5D Sketches



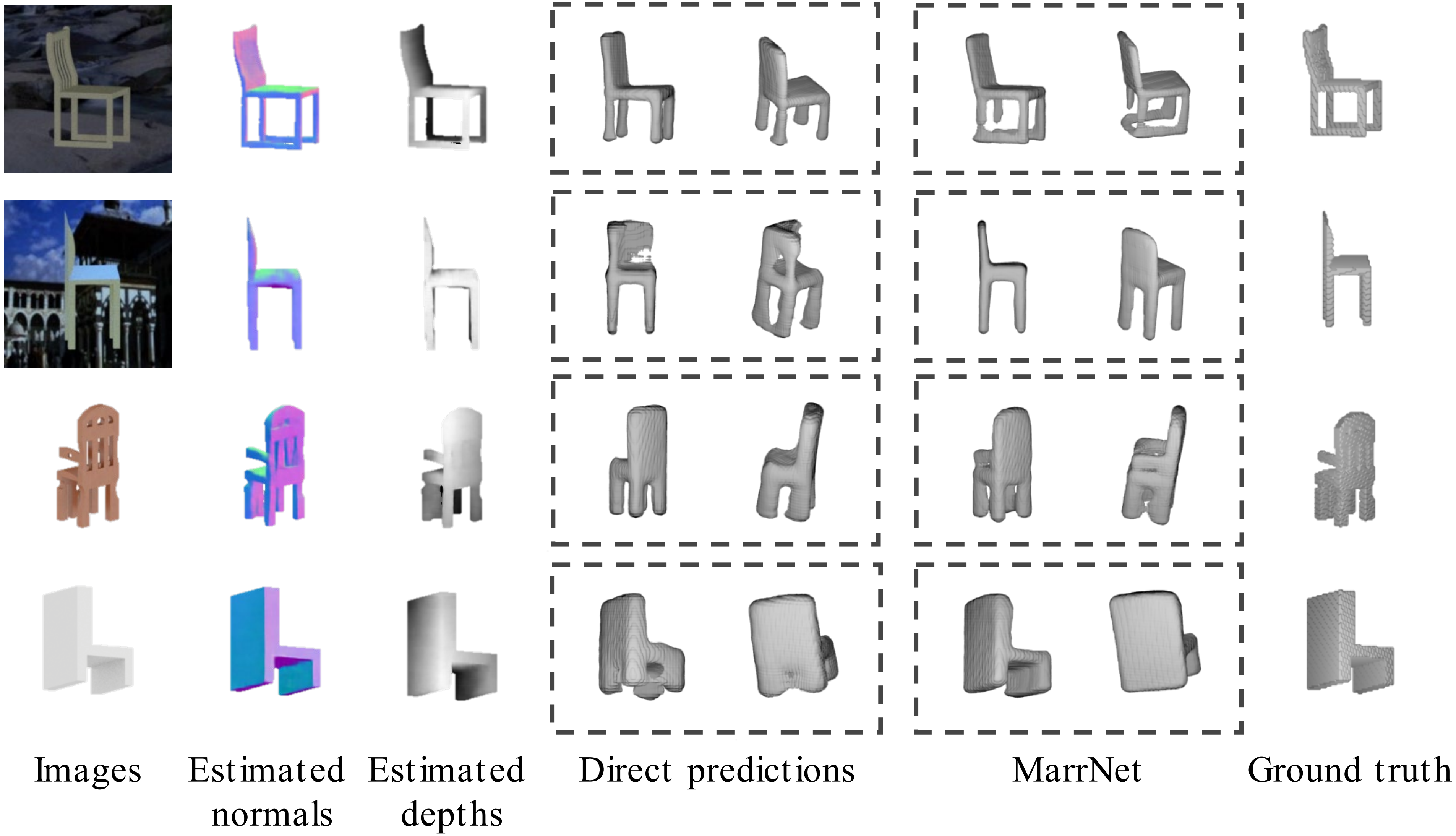
Wu*, Wang*, Xue, Sun, Freeman, Tenenbaum. NIPS'17

MarrNet : 3D Reconstruction via 2.5D Sketches

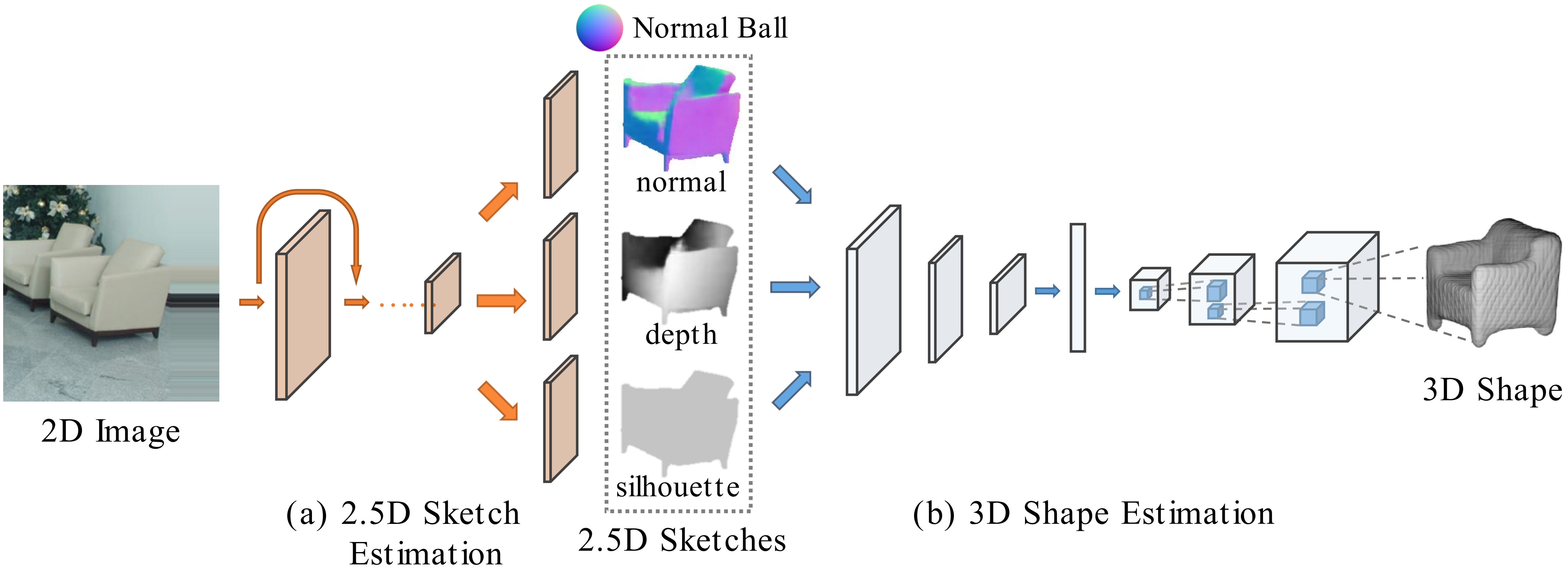


Wu*, Wang*, Xue, Sun, Freeman, Tenenbaum. NIPS'17

Results on ShapeNet

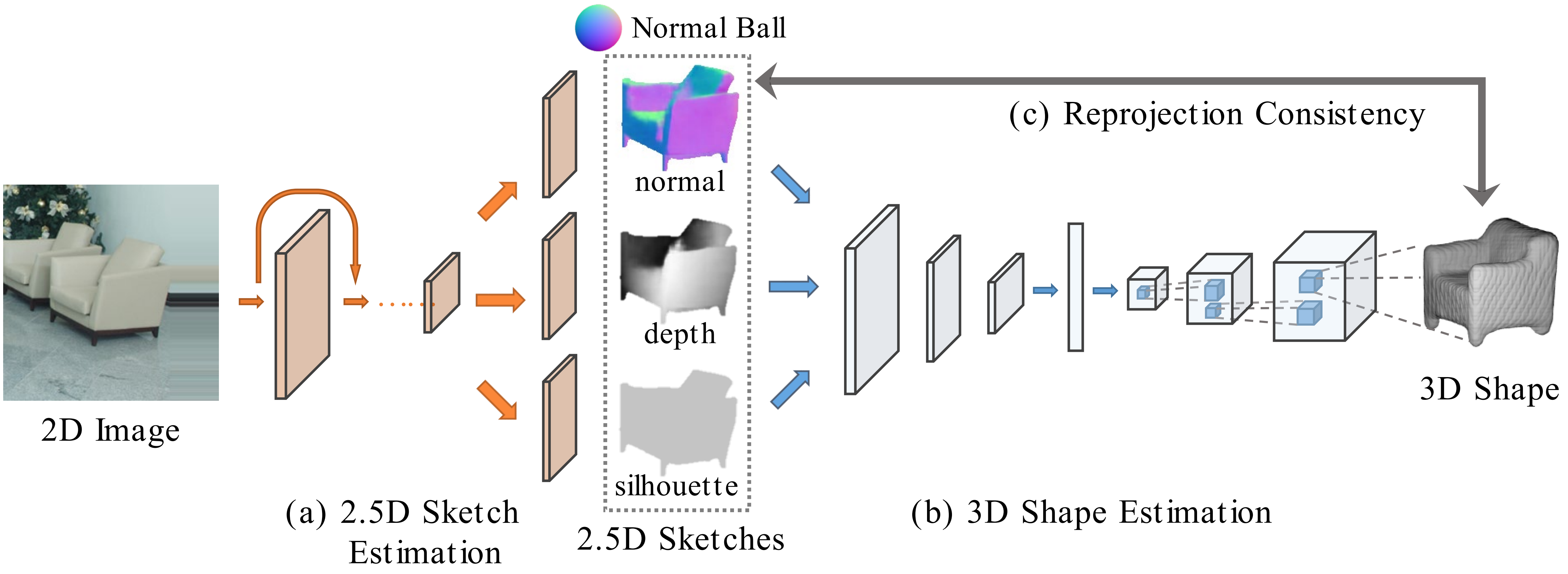


MarrNet : 3D Reconstruction via 2.5D Sketches



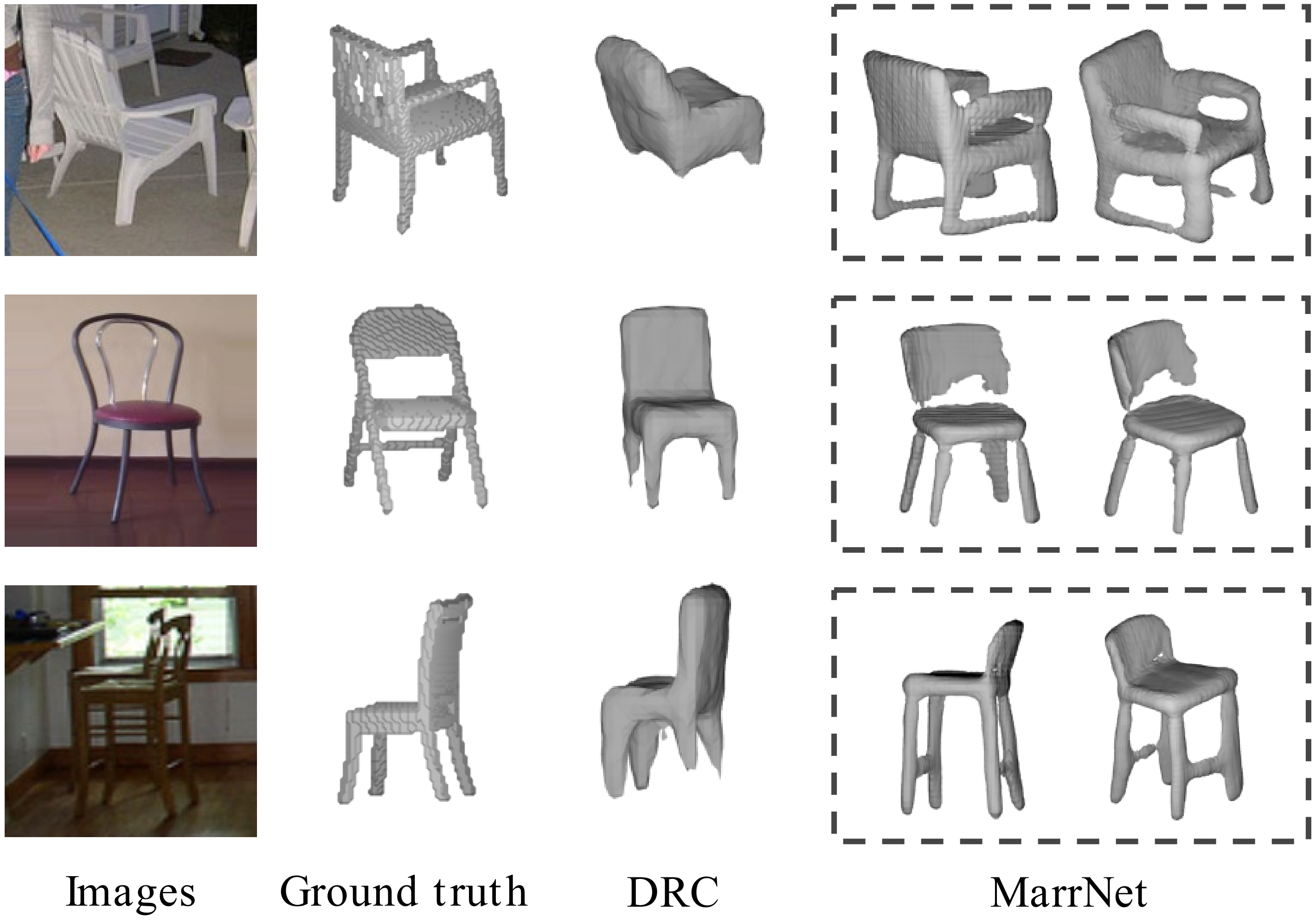
Wu*, Wang*, Xue, Sun, Freeman, Tenenbaum. NIPS'17

MarrNet : 3D Reconstruction via 2.5D Sketches



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

Comparisons on PASCAL 3D+

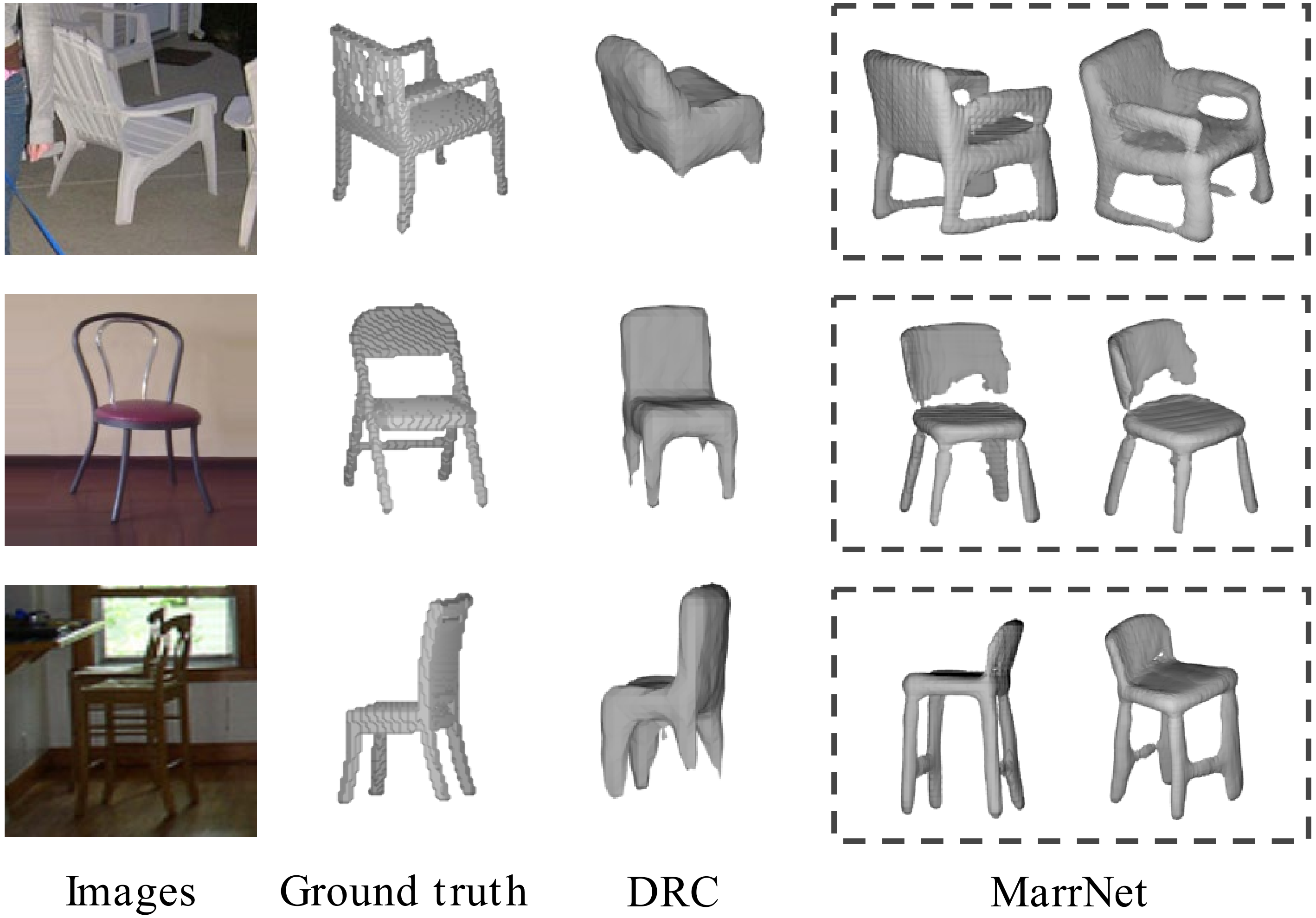


Comparisons on PASCAL 3D+

Methods	IoU
DRC [CVPR '17]	0.34
MarrNet	0.38

Intersection over Union (IoU)

Comparisons on PASCAL 3D+



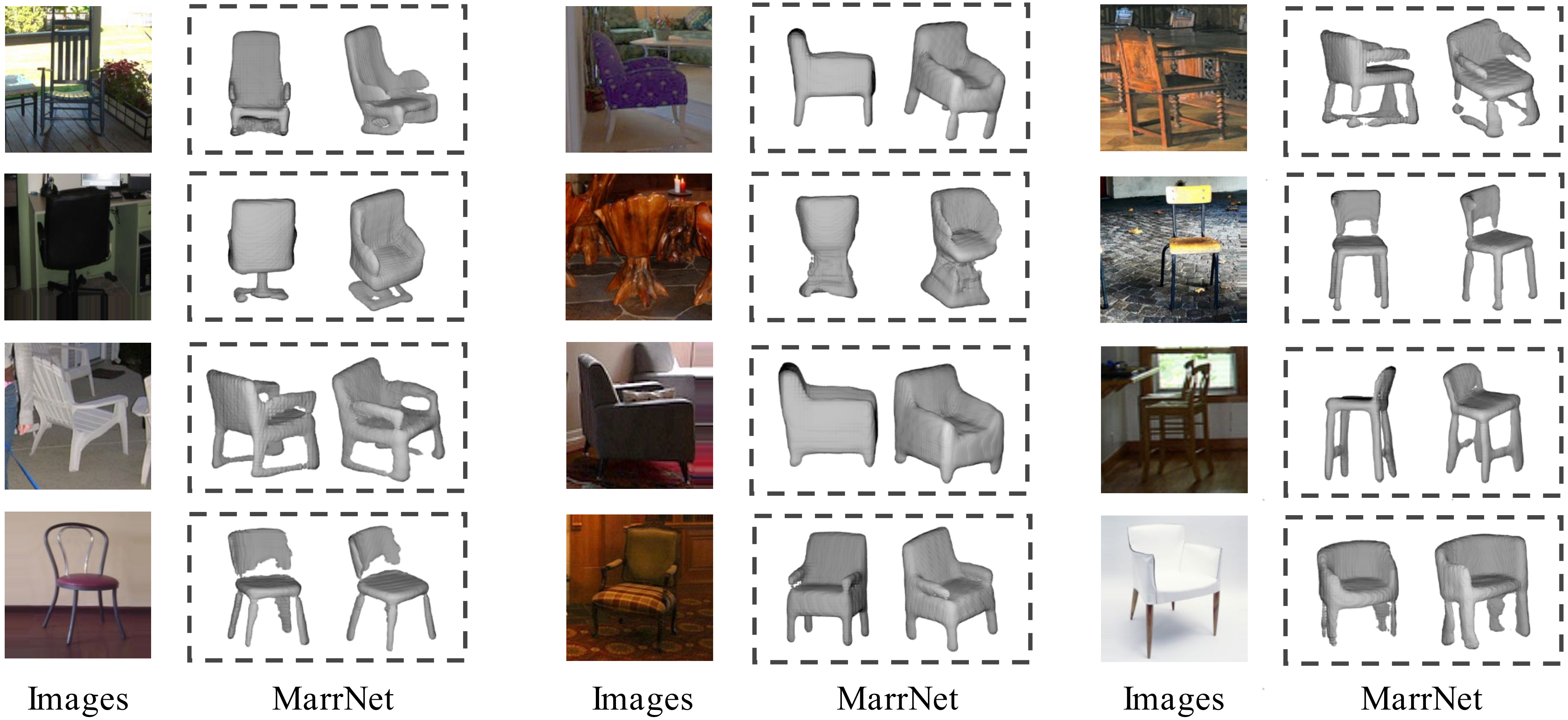
Methods	IoU
DRC [CVPR '17]	0.34
MarrNet	0.38

Intersection over Union (IoU)

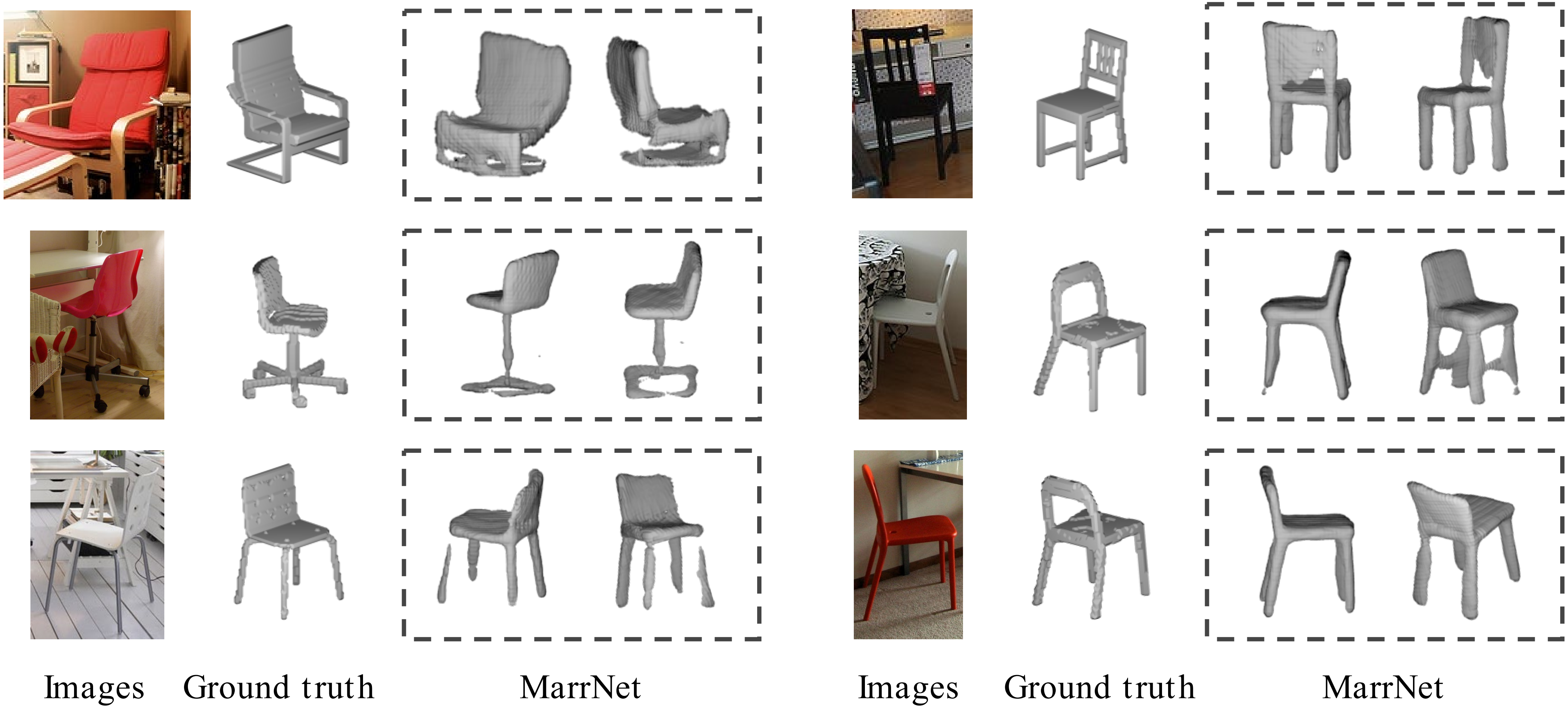
	DRC	MarrNet	GT
DRC	50	26	17
MarrNet	74	50	42
GT	83	58	50

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

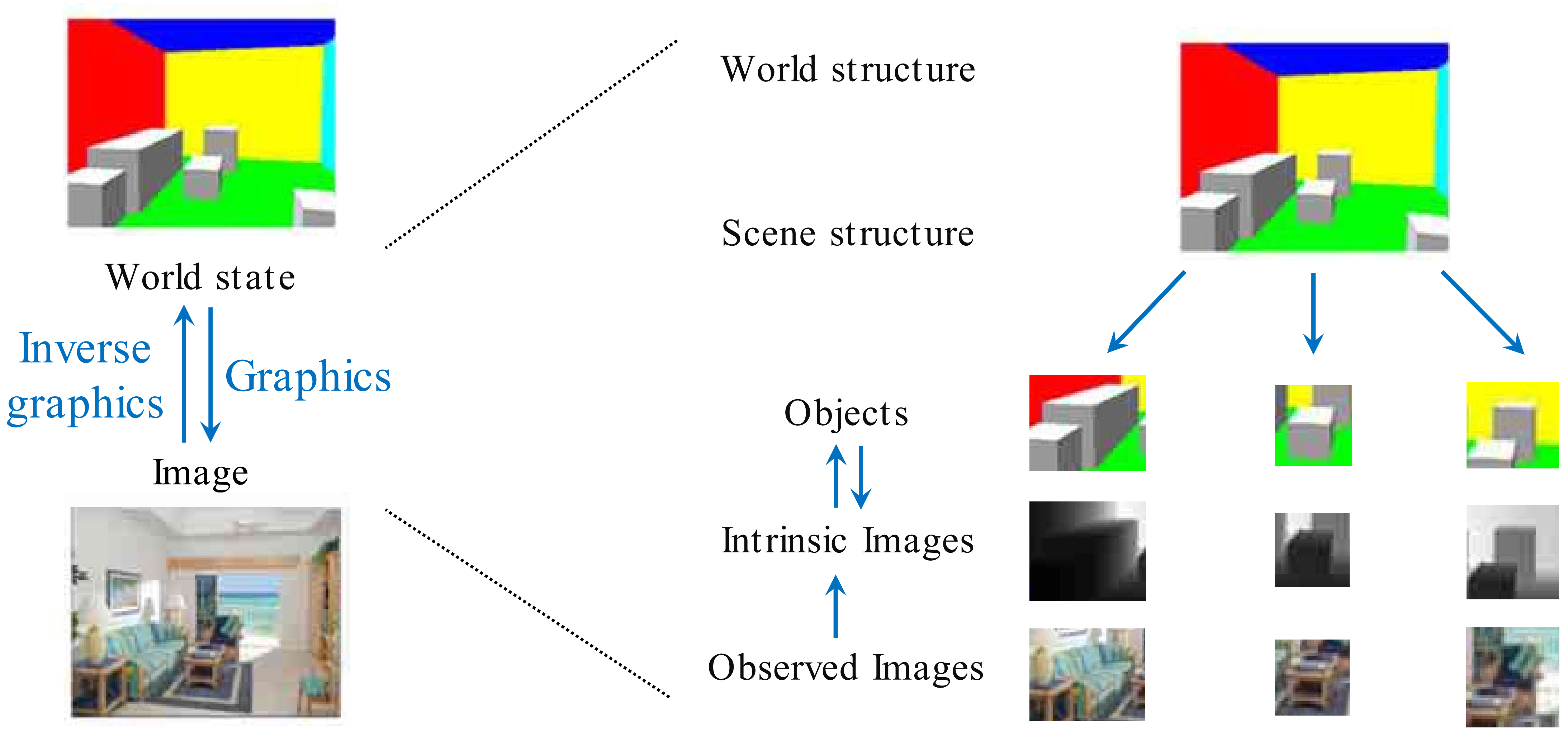
Results on PASCAL 3D+



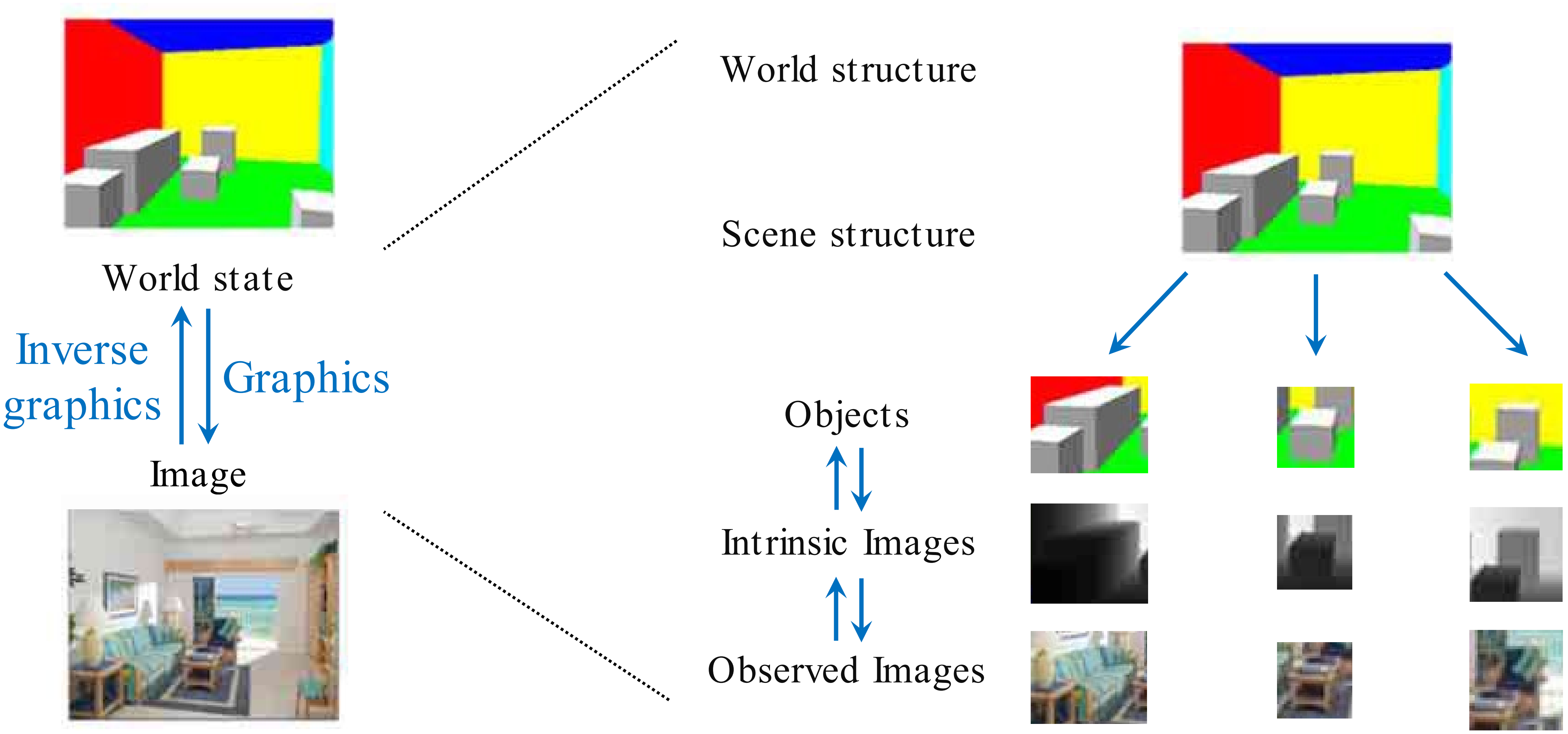
Results on IKEA



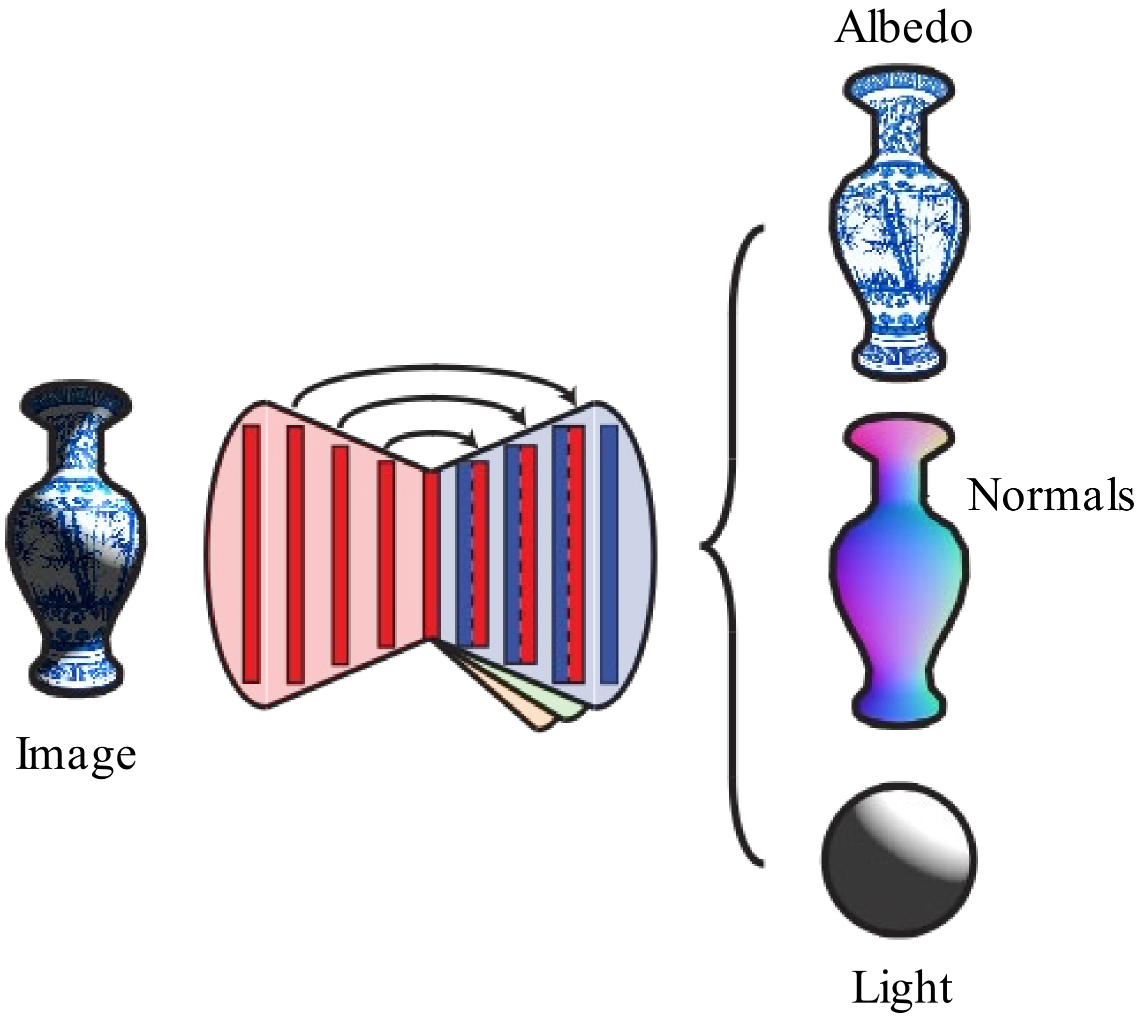
How Computer Graphics Helps Computer Vision



How Computer Graphics Helps Computer Vision

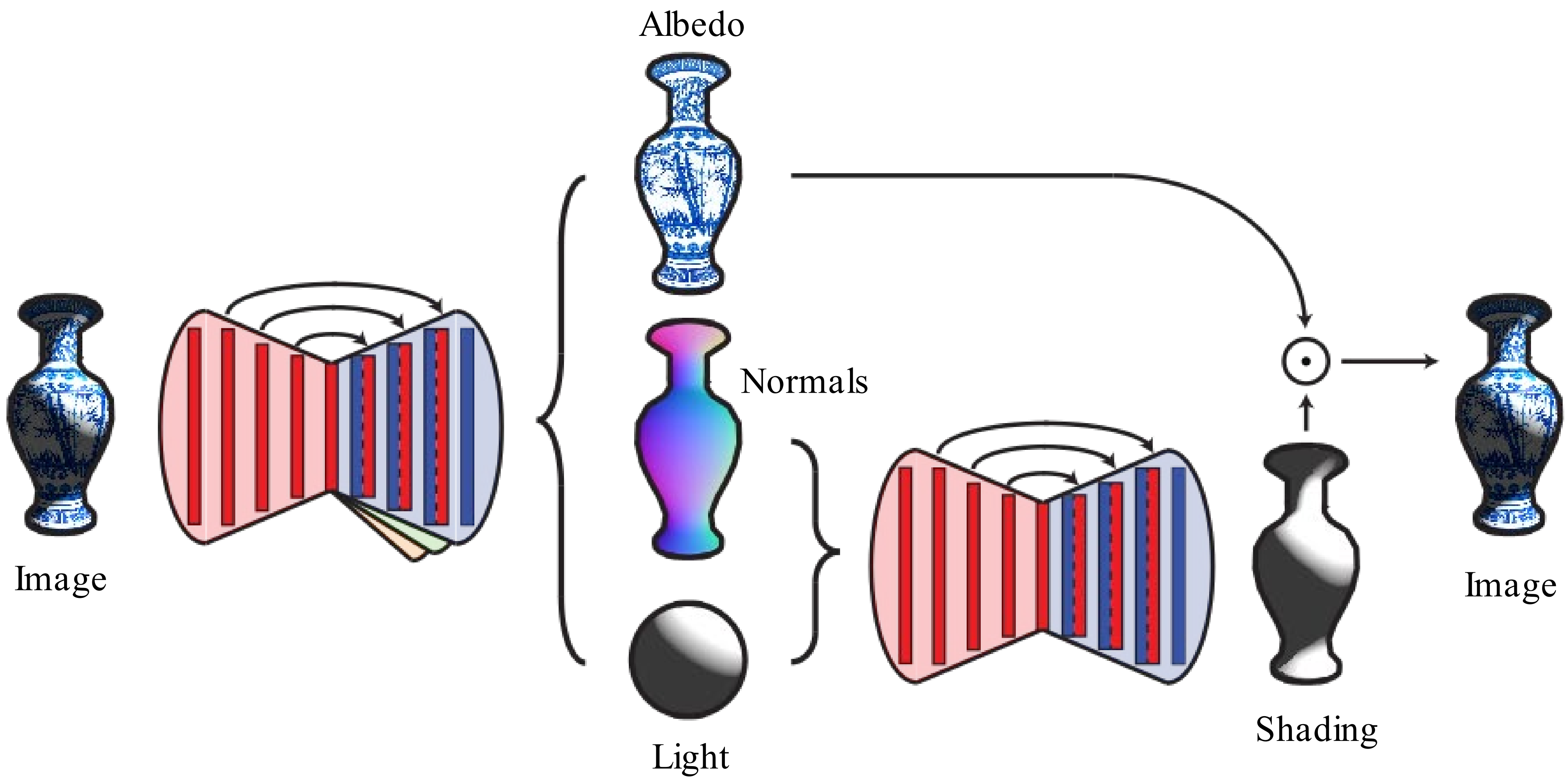


Self-Supervised Intrinsic Image Decomposition



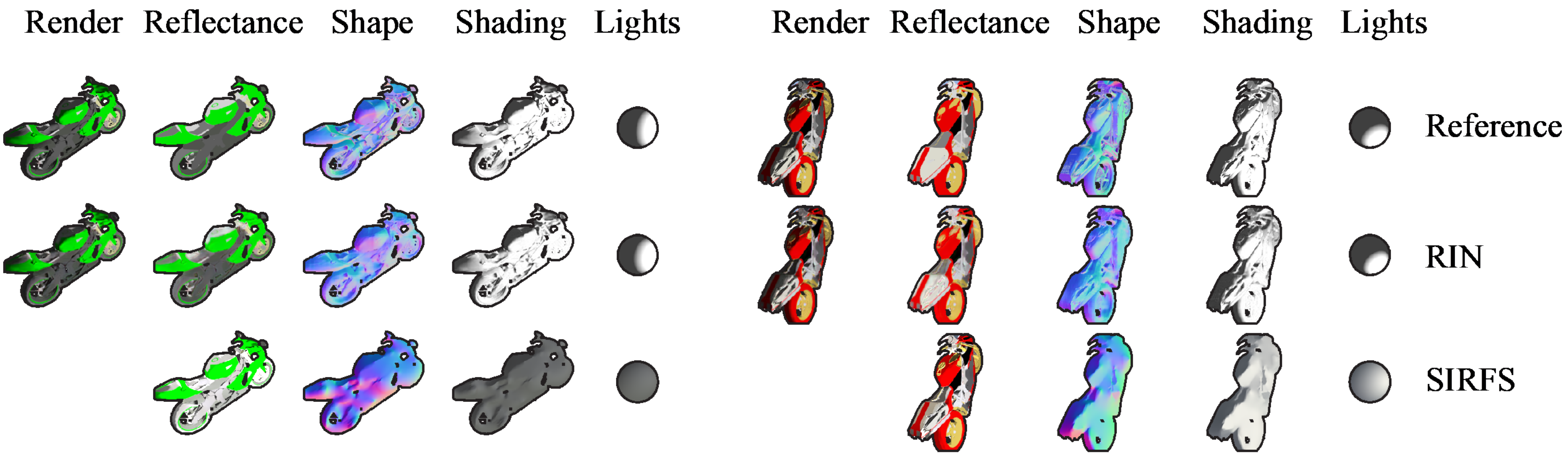
Janner, Wu, Kulkarni, Yildirim, Tenenbaum. NIPS'17

Self-Supervised Intrinsic Image Decomposition

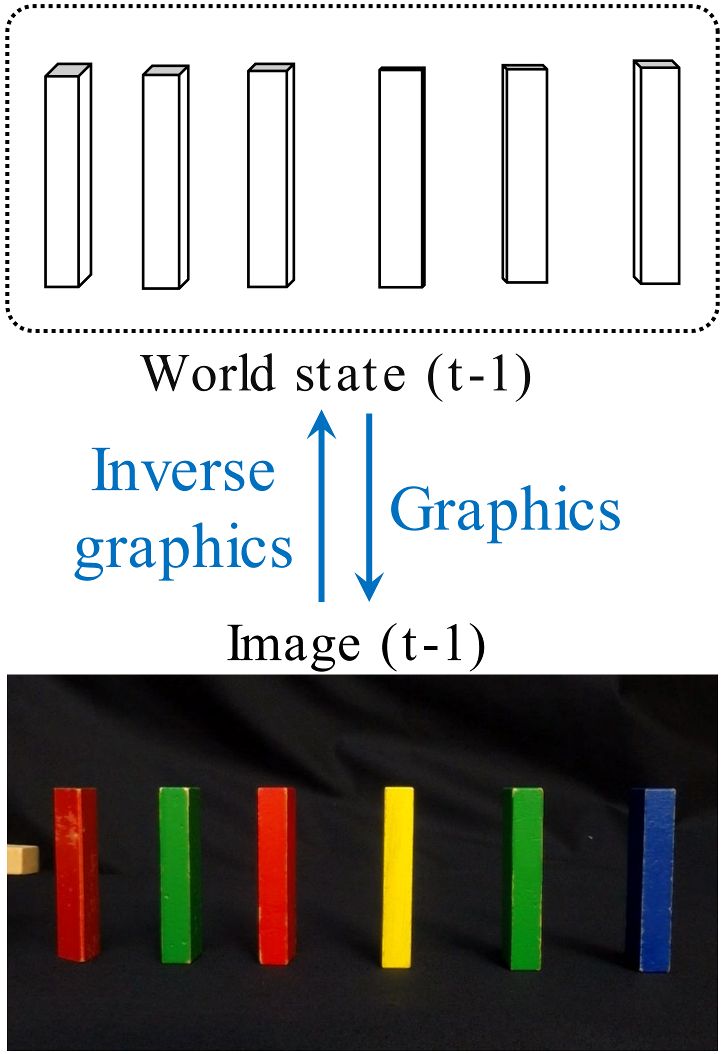


Janner, Wu, Kulkarni, Yildirim, Tenenbaum. NIPS'17

Results on Intrinsic Image Decomposition

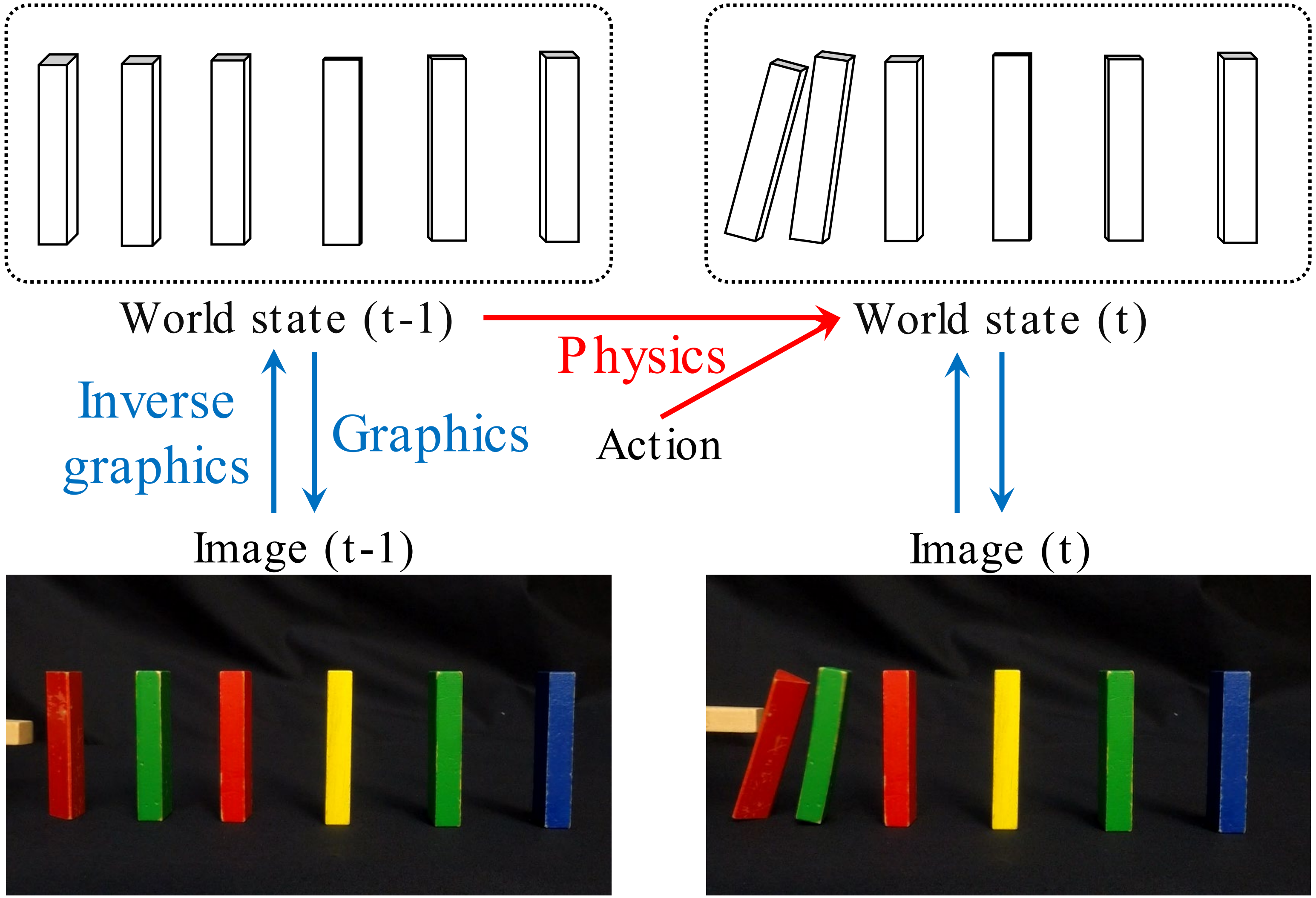


Outline



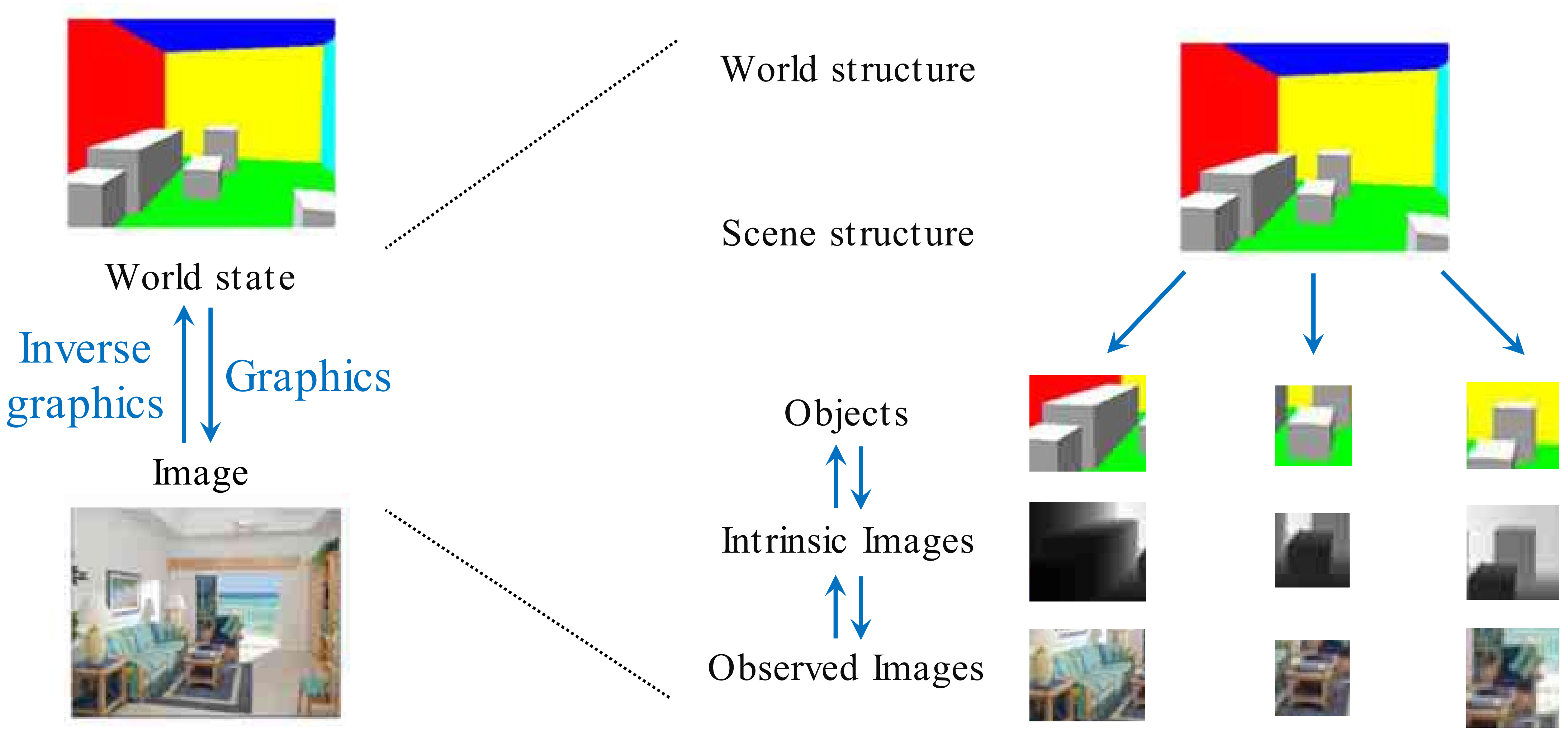
- Single Object
  - 3D Shape [NIPS'17]
  - Intrinsic Images [NIPS'17]
- Static Scene
  - Scene de-rendering [CVPR'17]

Outline

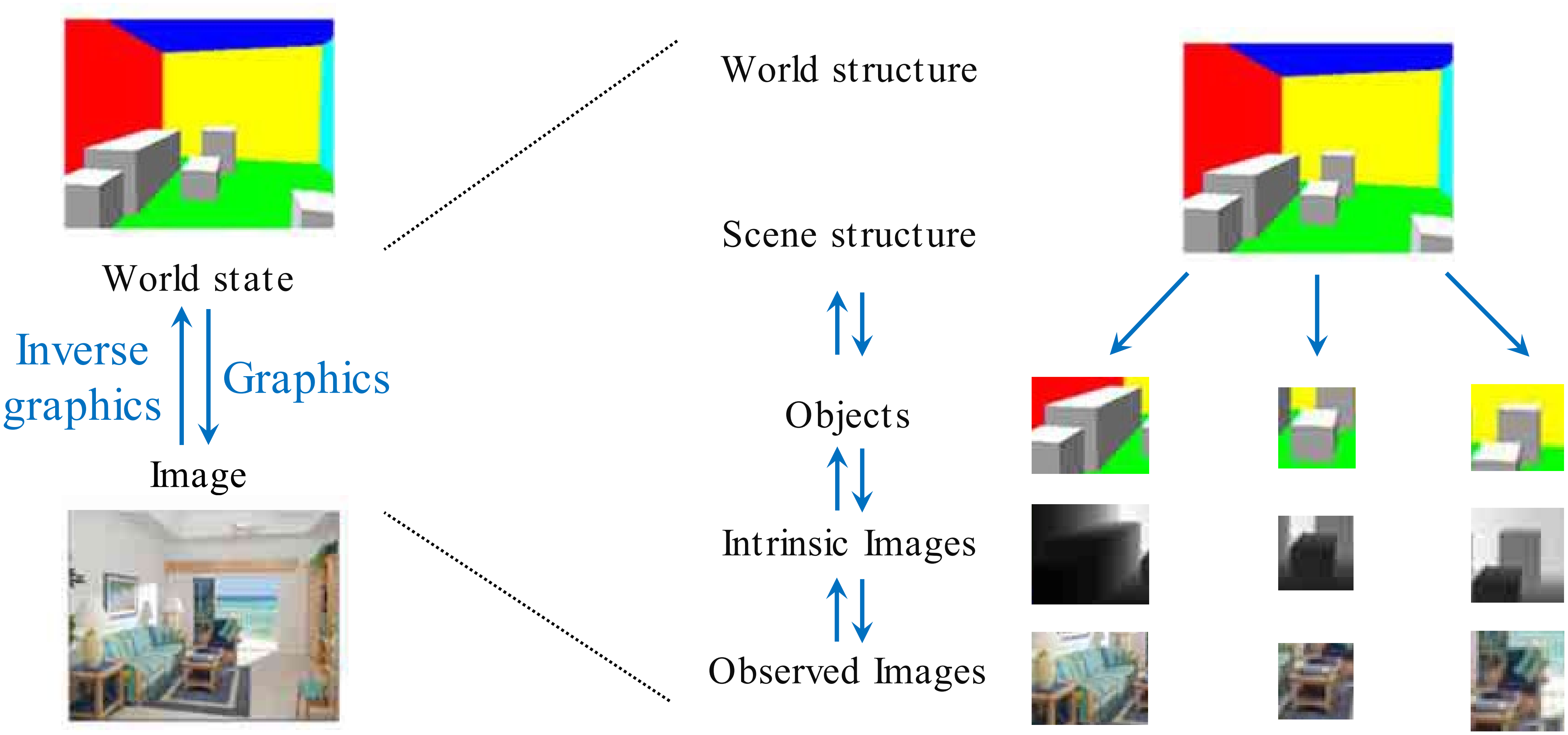


- Single Object
  - 3D Shape [NIPS'17]
  - Intrinsic Images [NIPS'17]
- Static Scene
  - Scene de-rendering [CVPR'17]
- Scene Dynamics
  - Perception + Physics [NIPS'17]
  - Multi-Modal Learning (V + A) [ICCV'17, NIPS'17]

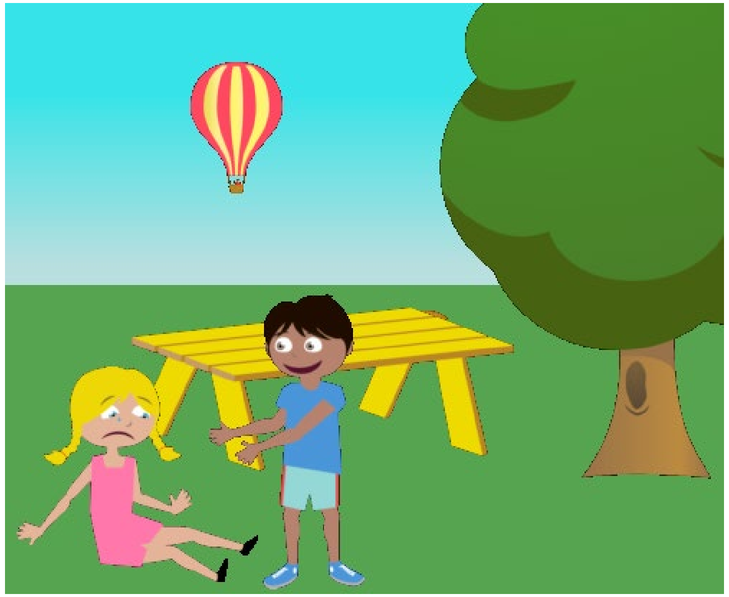
How Computer Graphics Helps Computer Vision



How Computer Graphics Helps Computer Vision

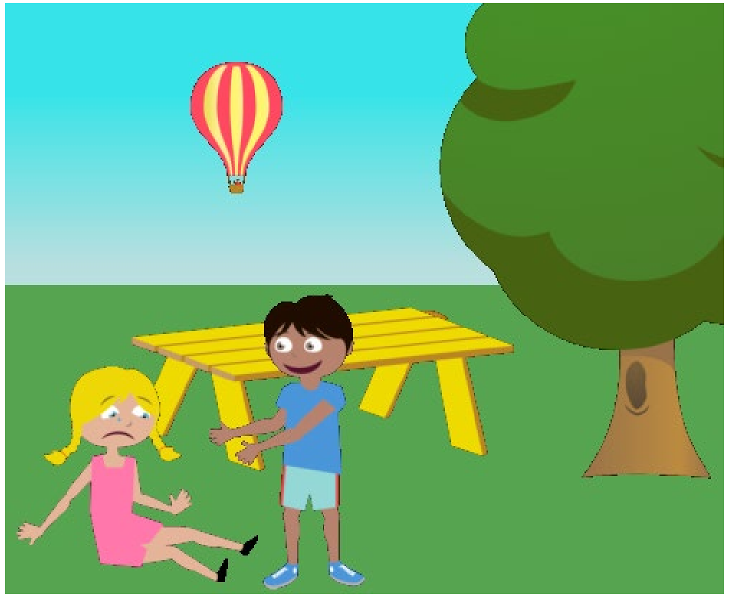


From Objects to Scenes (Scene De -rendering)



Wu, Tenenbaum, Kohli. CVPR’17

From Objects to Scenes (Scene De -rendering)

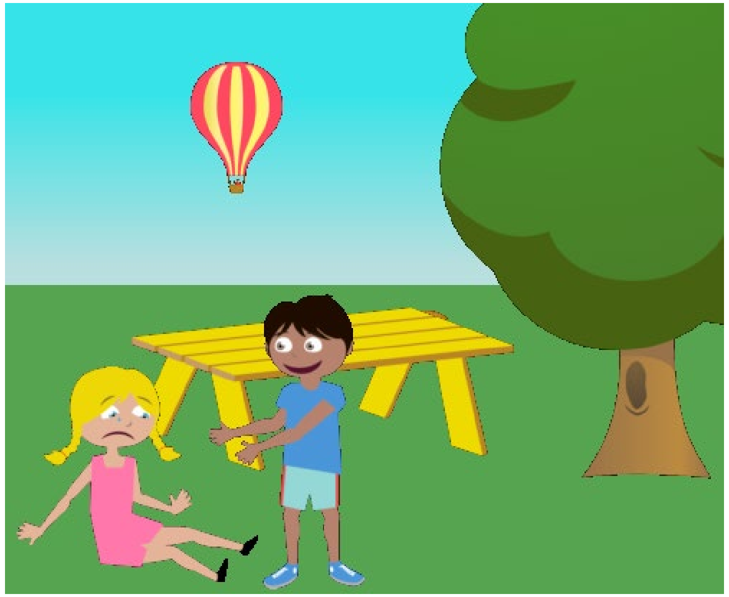


De-render

Render

Wu, Tenenbaum, Kohli. CVPR'17

From Objects to Scenes (Scene De -rendering)



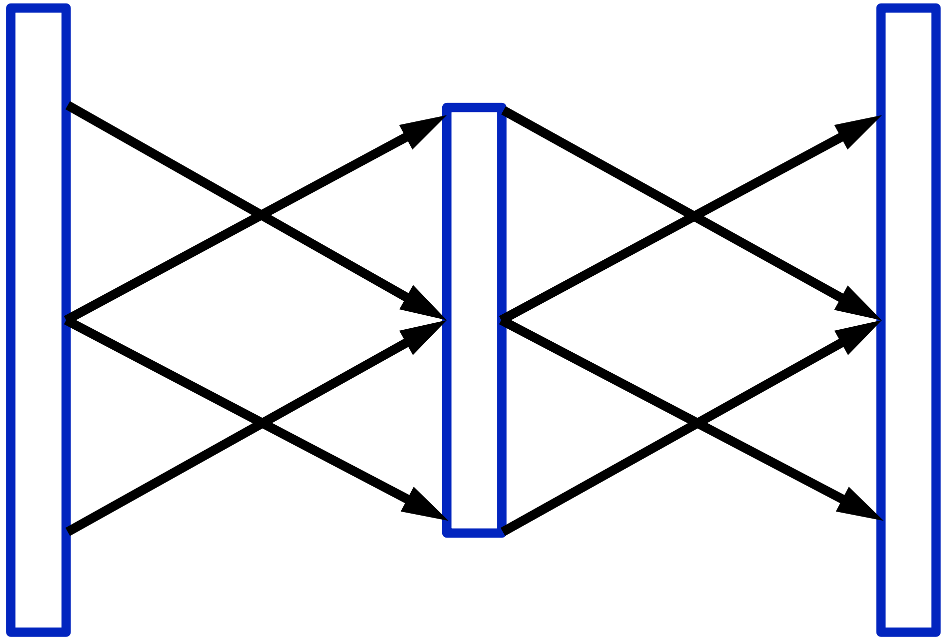
De-render

Render

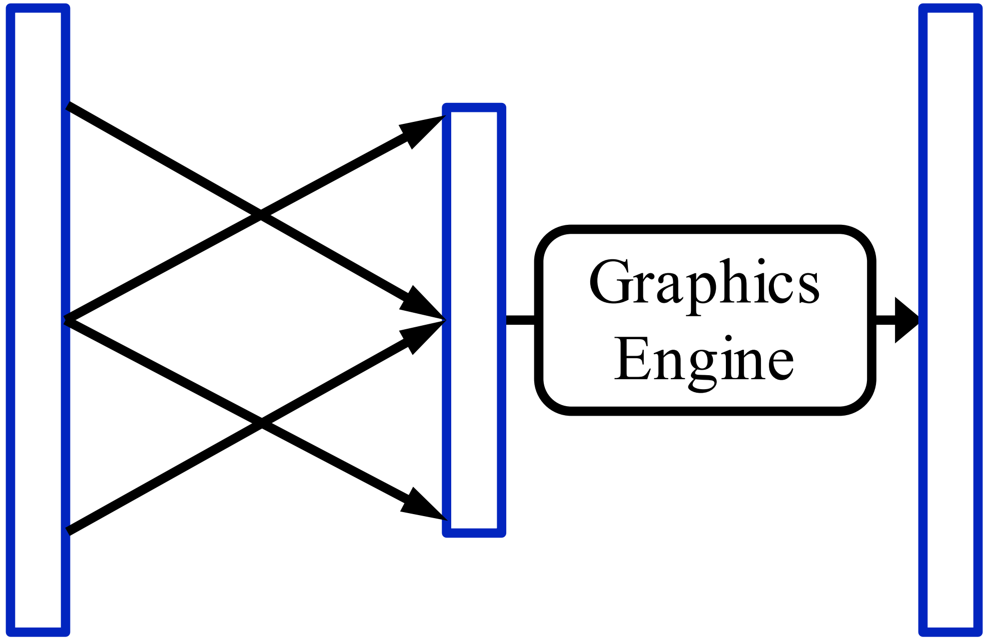
<objects>
<balloon: right>
<bench: yellow>
<tree: right>
<boy: stand happy>
<girl: sit sad>
</objects>

Wu, Tenenbaum, Kohli. CVPR’17

Generalized Encoding-Decoding Structure



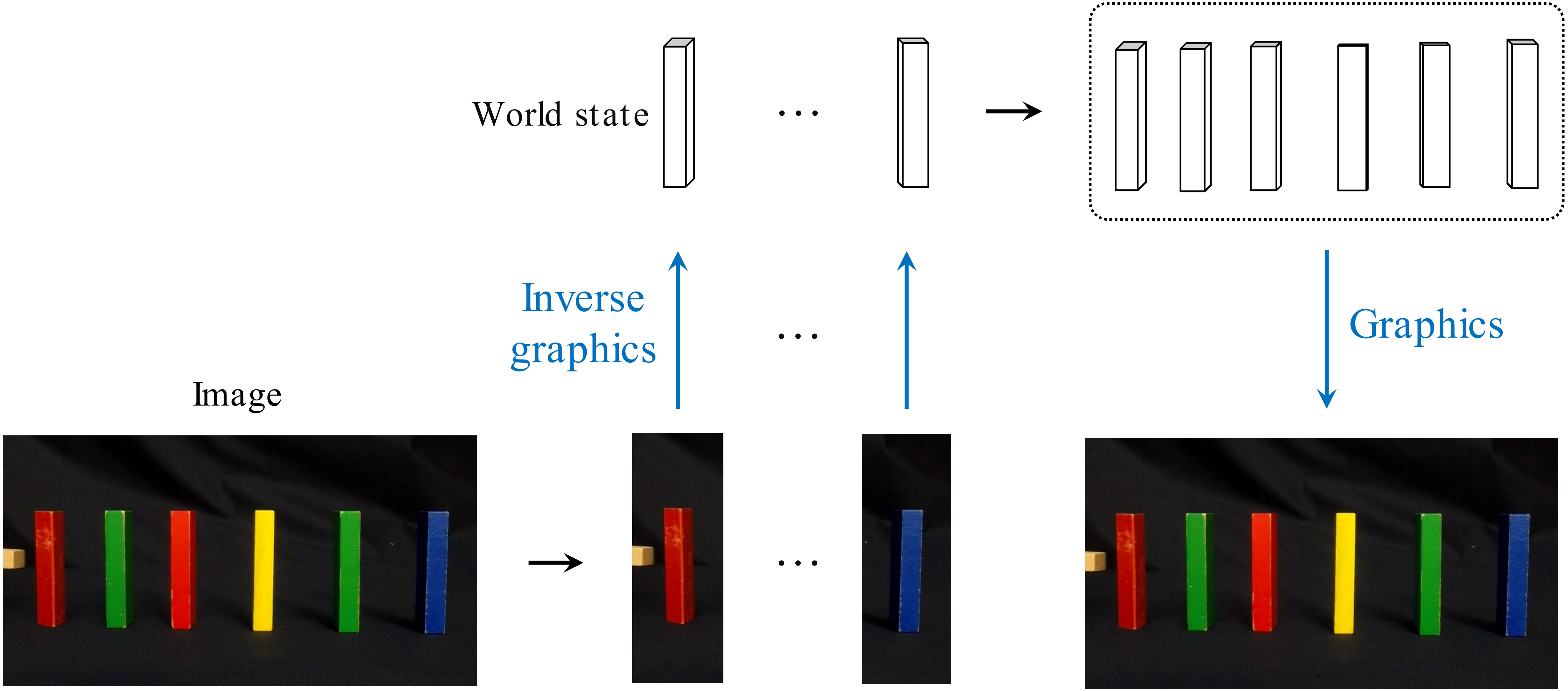
(a) A standard autoencoder



(b) A generalized autoencoder

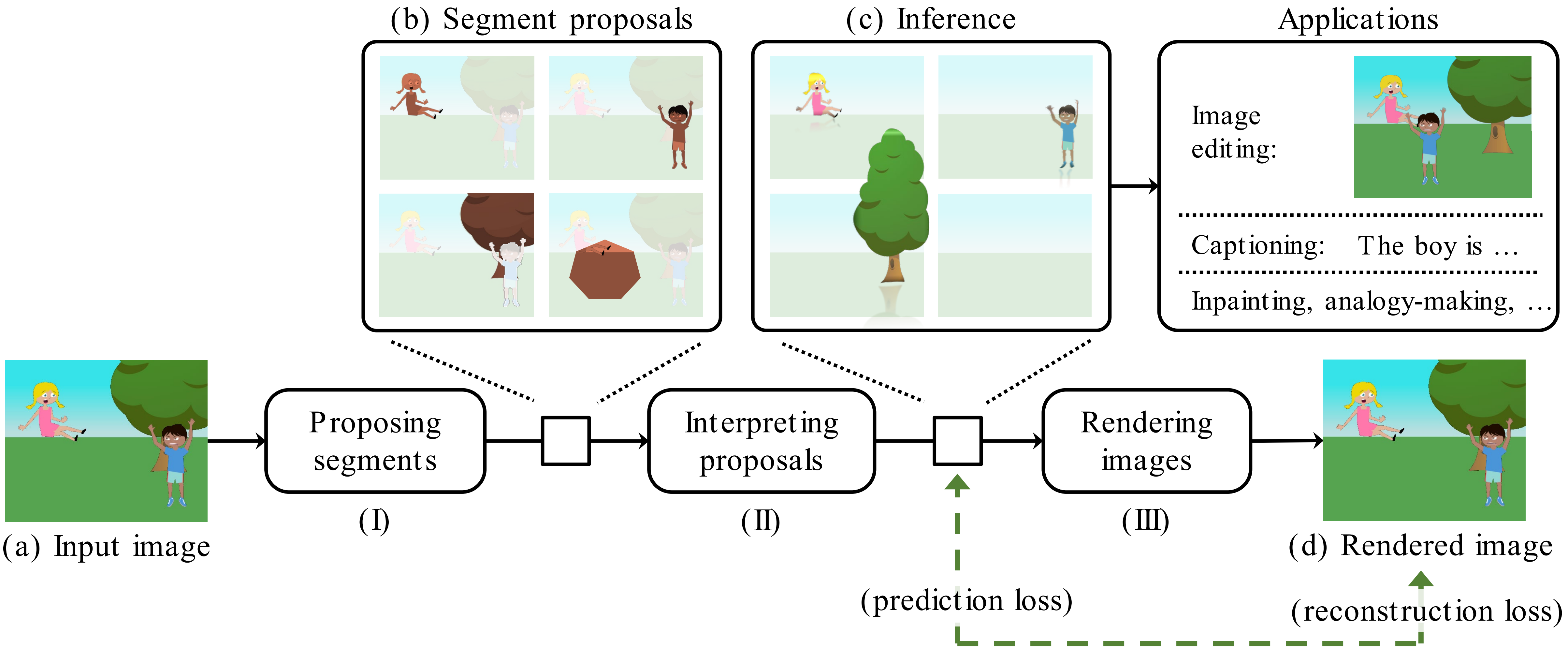
Wu, Tenenbaum, Kohli. CVPR'17

Scene De-rendering



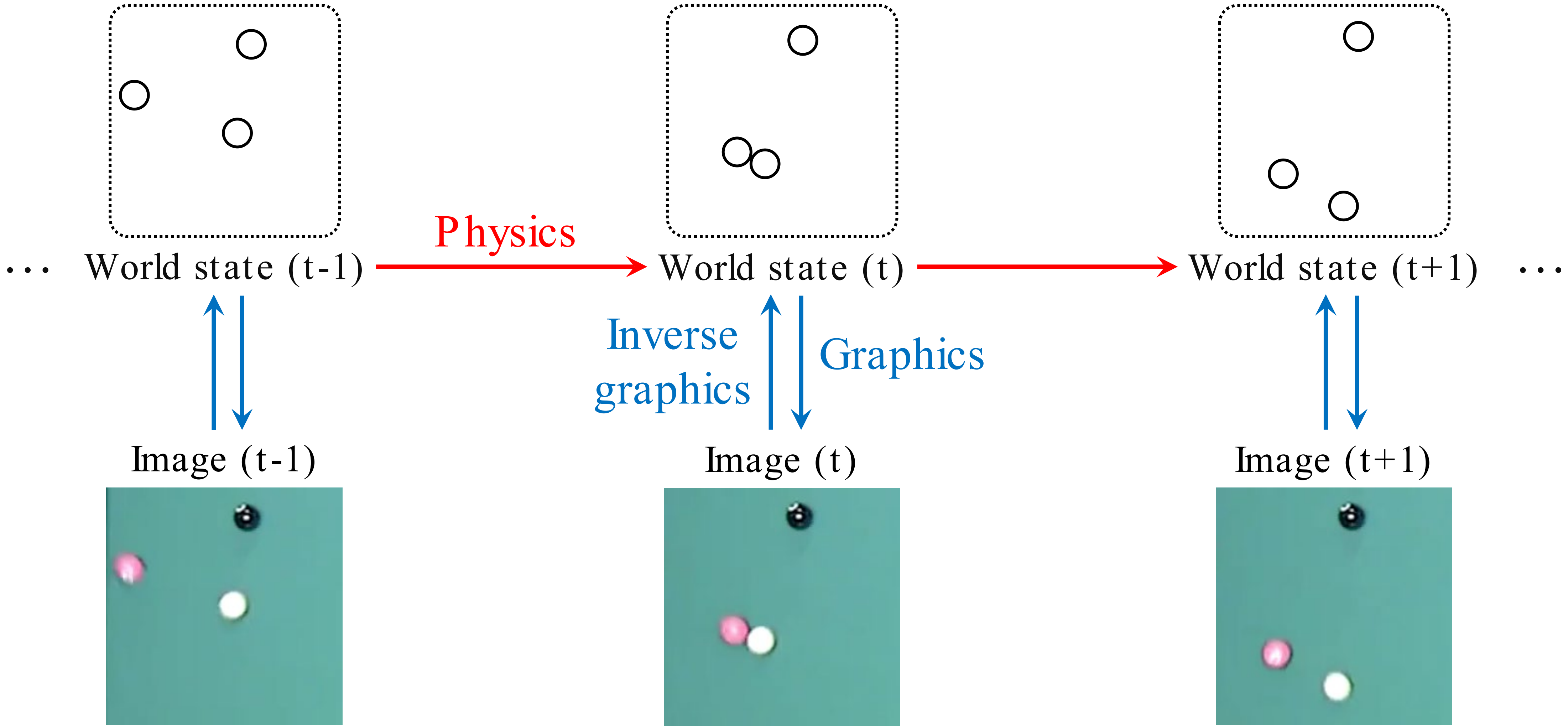
Wu, Tenenbaum, Kohli. CVPR'17

Model Details



Wu, Tenenbaum, Kohli. CVPR'17

Learning to See Physics via Visual De -animation

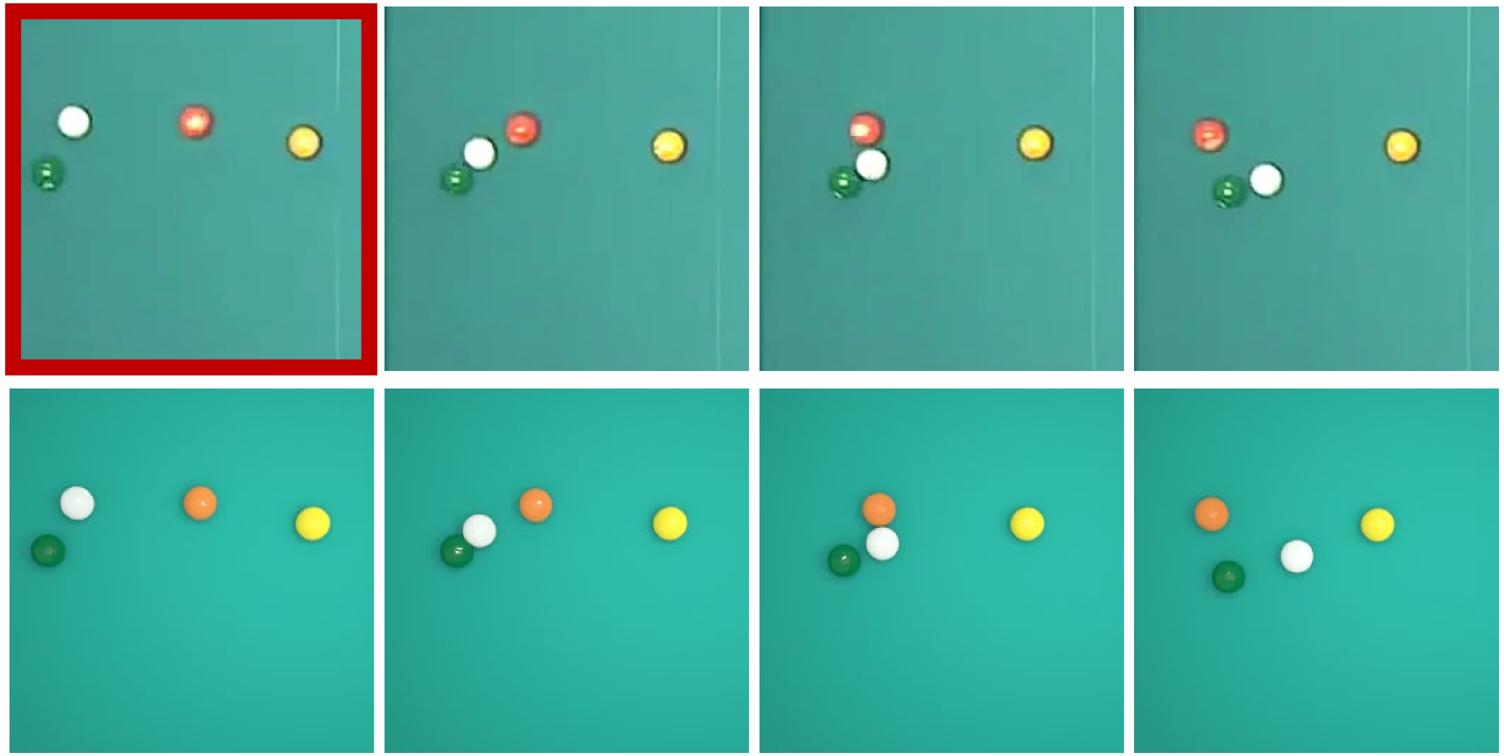


Wu, Lu, Kohli, Freeman, Tenenbaum. NIPS'17

Learning to See Physics via Visual De -animation

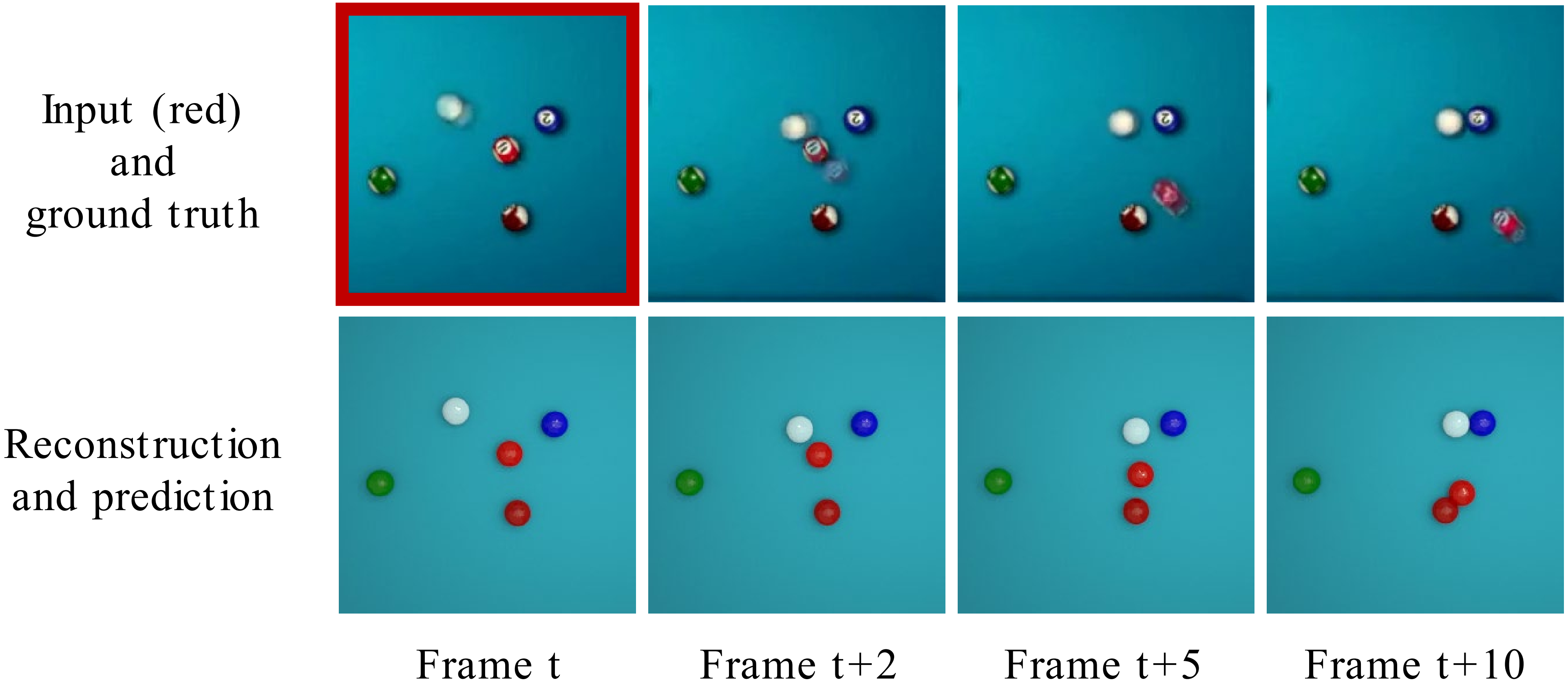
Input (red)
and
ground truth

Reconstruction
and prediction

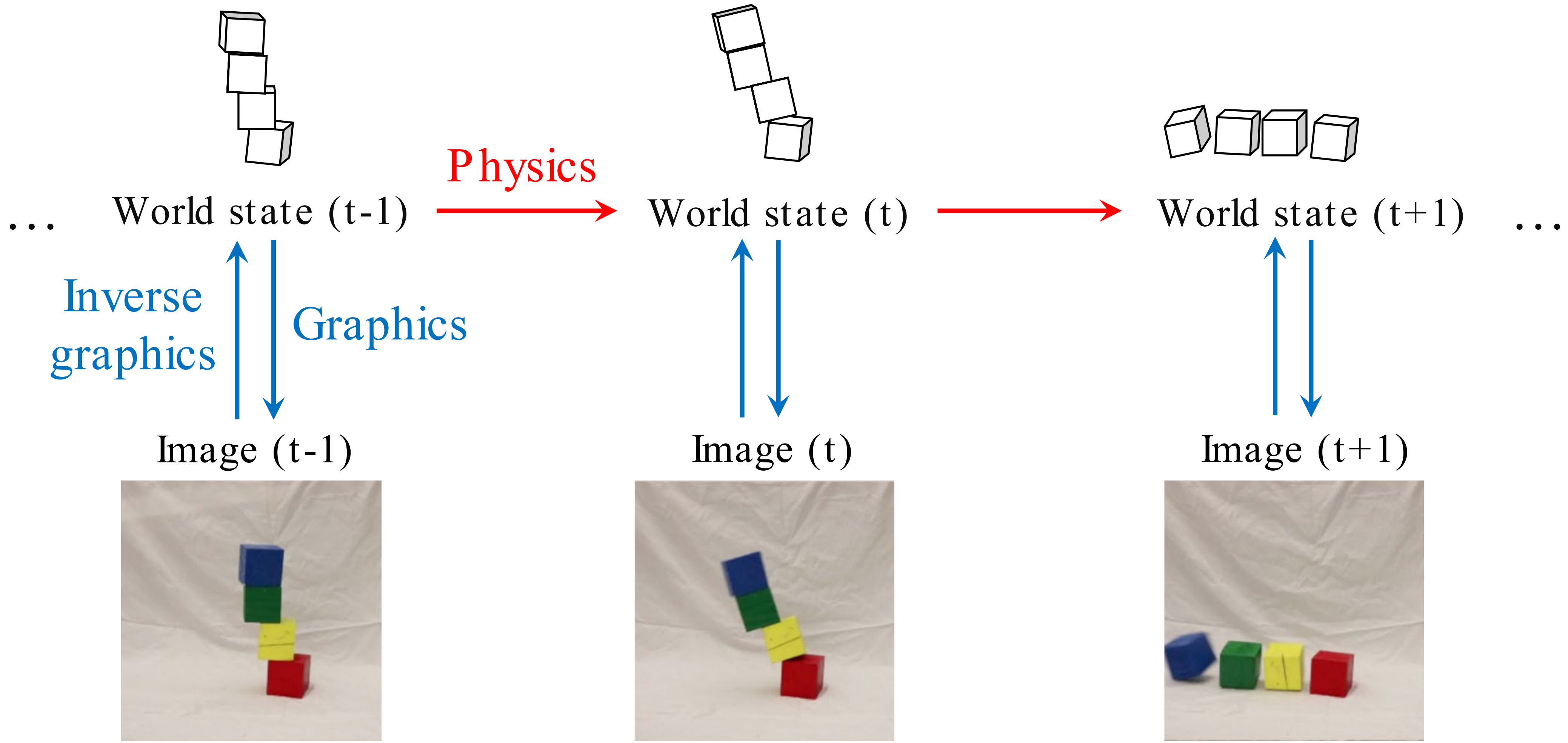


Frame t | Frame t+2 | Frame t+5 | Frame t+10

Learning to See Physics via Visual De -animation

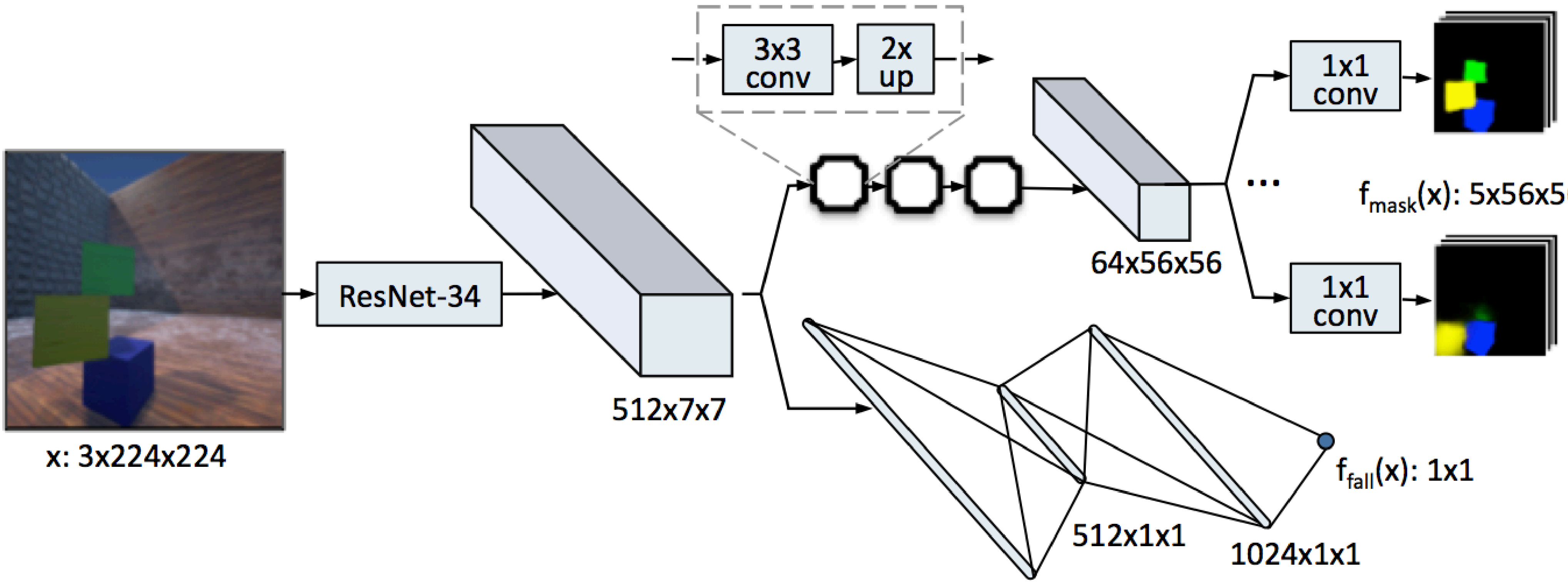


Learning to See Physics via Visual De -animation



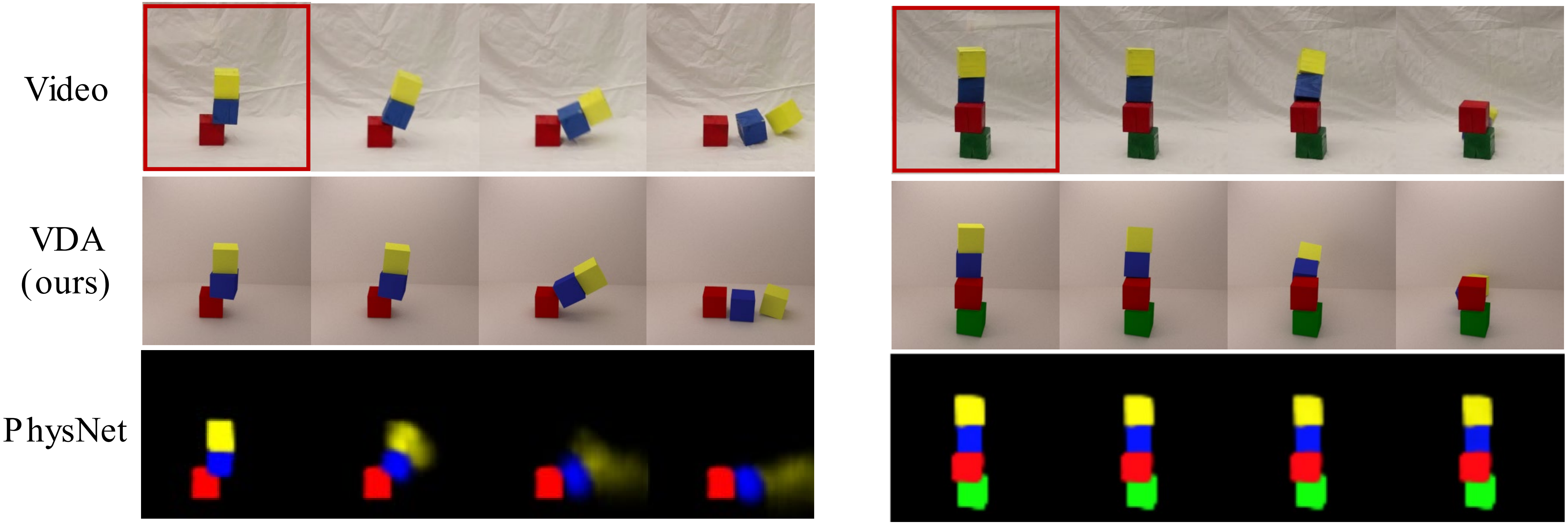
Wu, Lu, Kohli, Freeman, Tenenbaum. NIPS’17

PhysNet

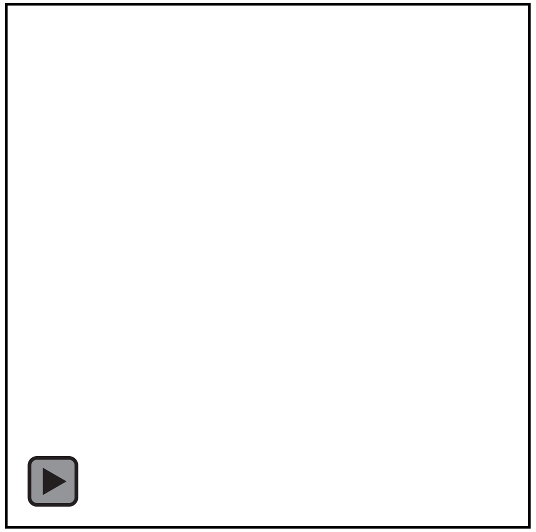


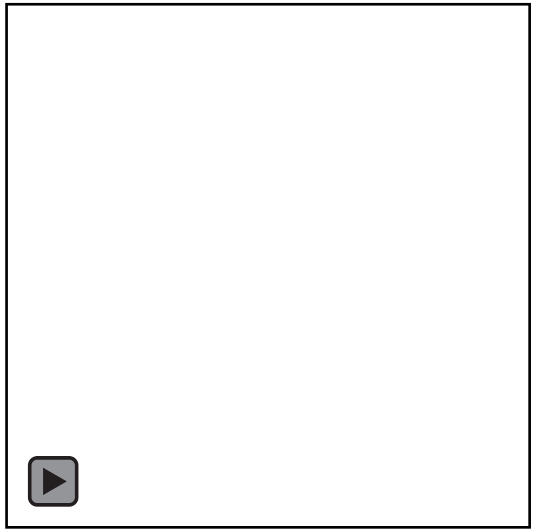
Lerer, Gross, Fergus. ICML'16

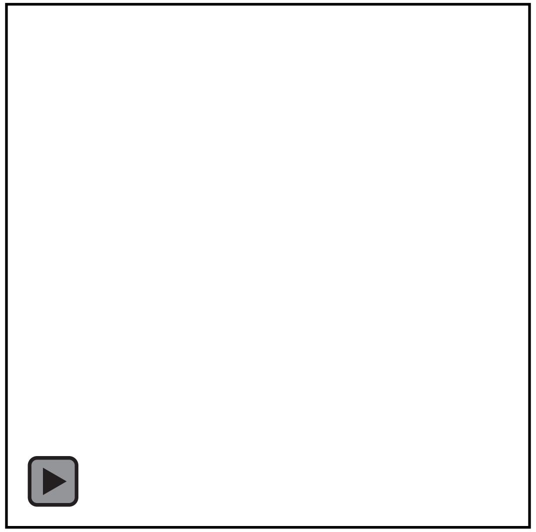
Comparing with PhysNet

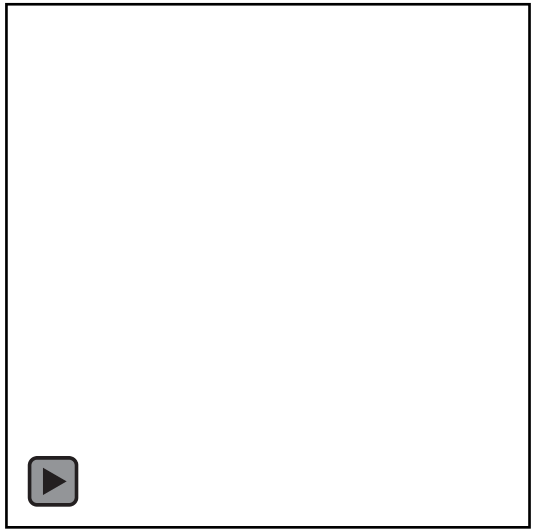


Learning to See Physics via Visual De -animation

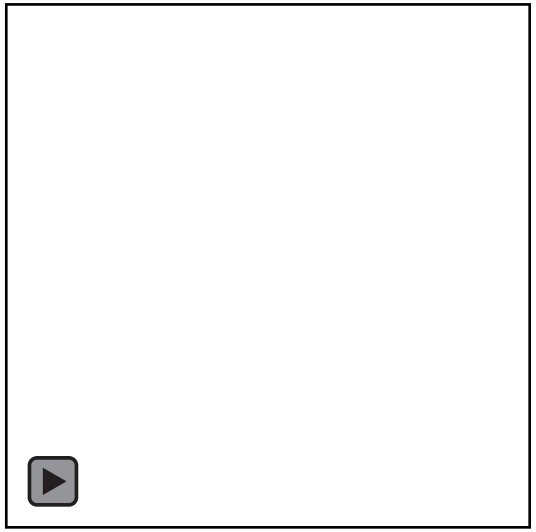
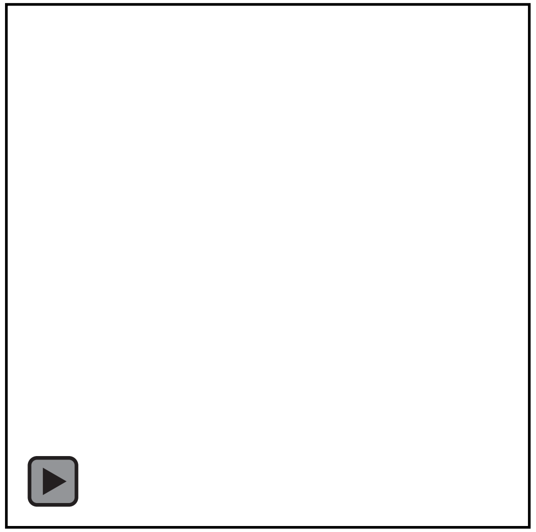
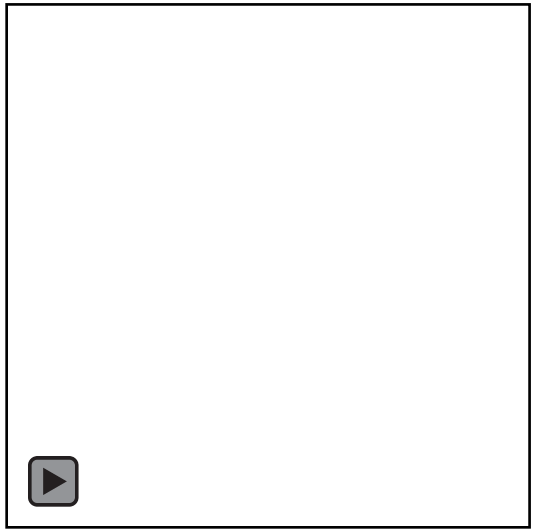
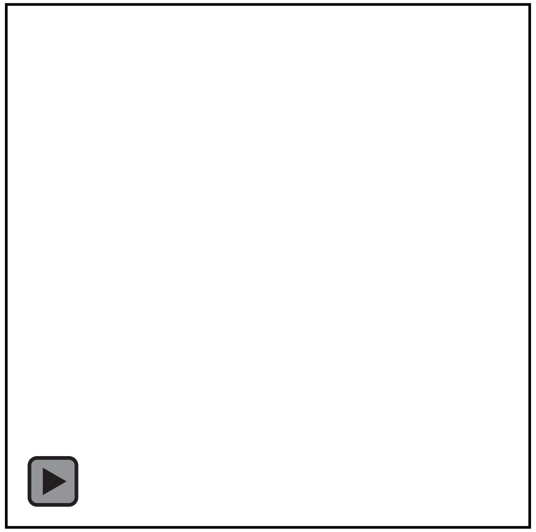




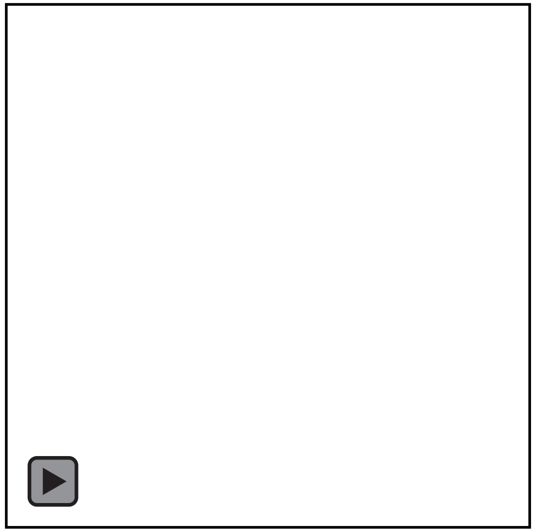


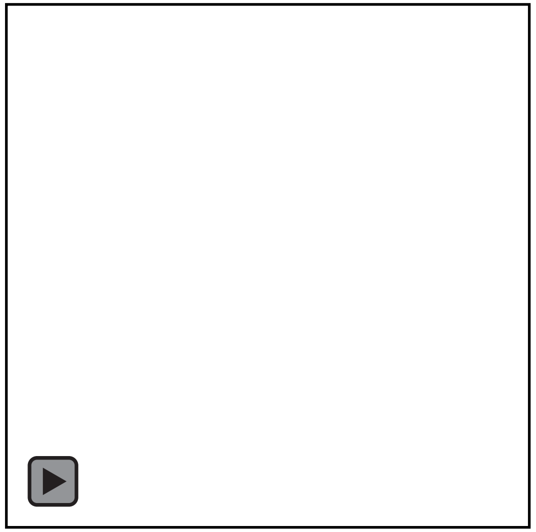


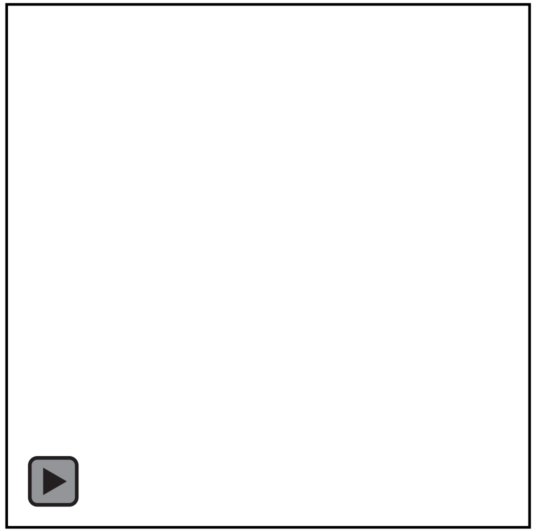
Learning to See Physics via Visual De -animation

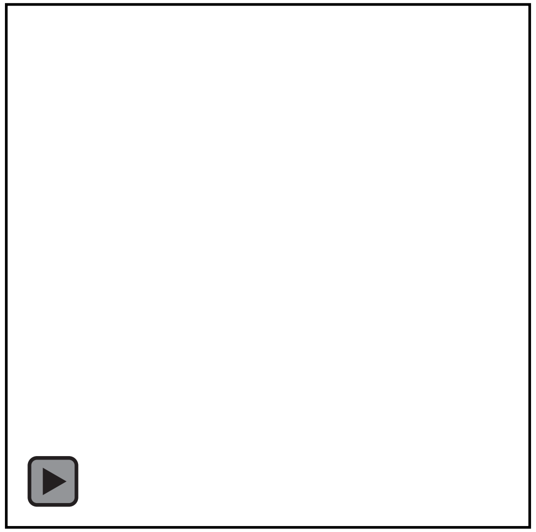
   

Learning to See Physics via Visual De-animation









Features

- Fast (<10ms)

Features

- Fast (<10ms)
- Accurate

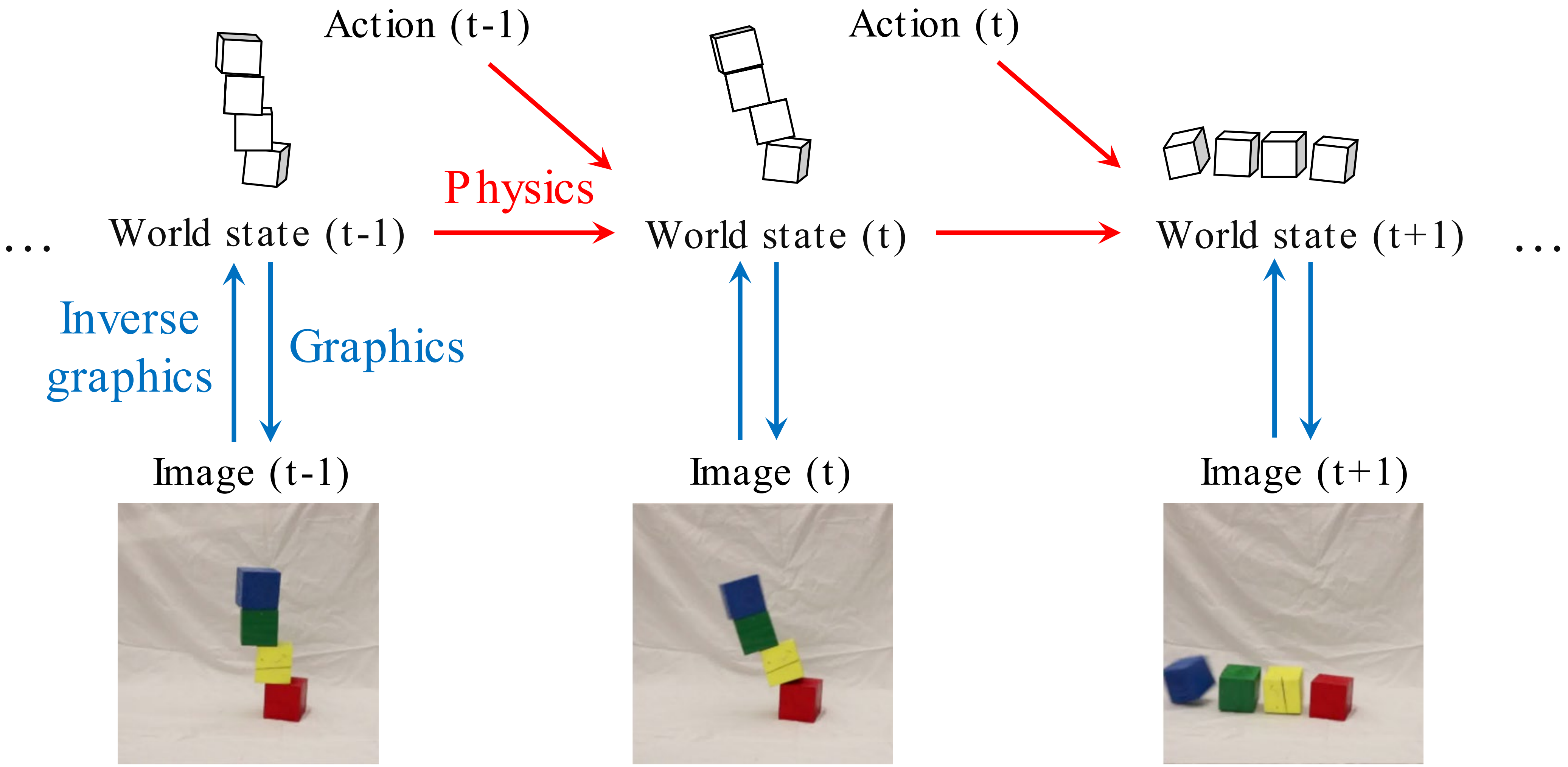
Methods	2 Blocks	3 Blocks	4 Blocks	Mean
Ours	**75**	**76**	**73**	**75**
PhysNet	66	66	**73**	68
GoogleNet	70	70	70	70
Chance	50	50	50	50

Features

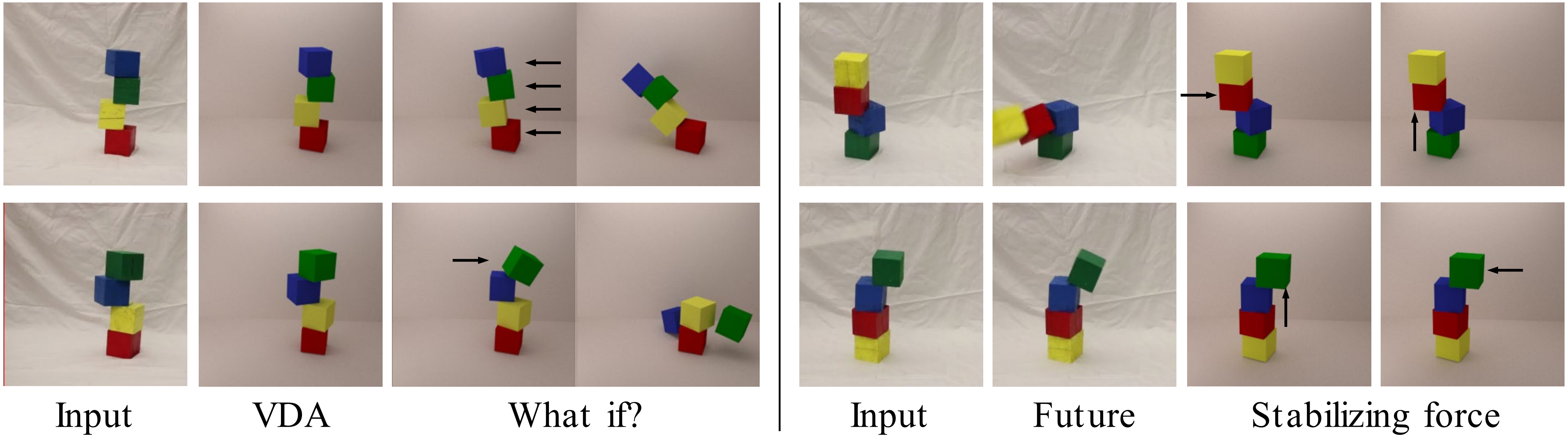
- Fast (<10ms)
- Accurate
- Rich: easily generalize to answer questions
  - 'What happens if? …' (external perturbation)

Methods	2 Blocks	3 Blocks	4 Blocks	Mean
Ours	**75**	**76**	**73**	**75**
PhysNet	66	66	**73**	68
GoogleNet	70	70	70	70
Chance	50	50	50	50

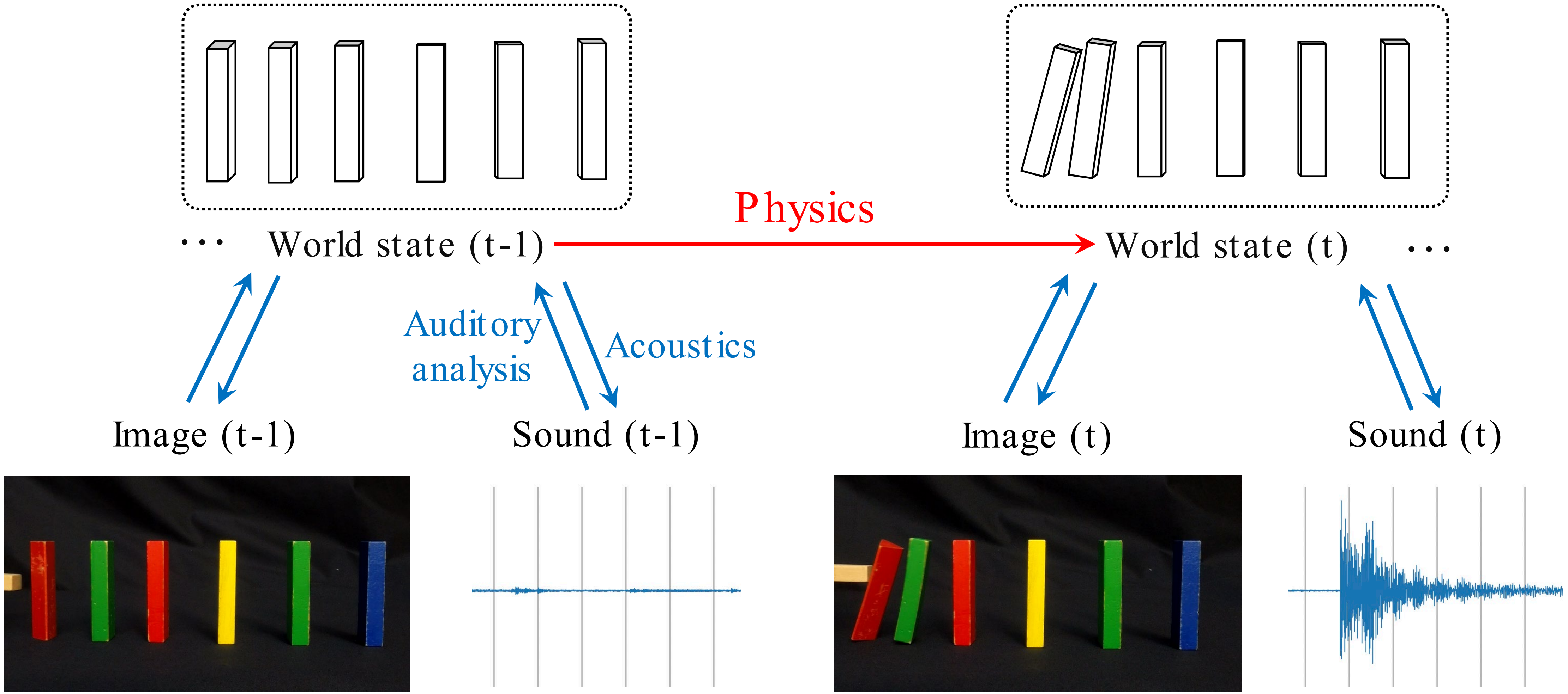
Learning to See Physics via Visual De -animation



Generalization



Modeling Multi -Modal Data



Zhang*, Wu*, Li, Huang, Traer, McDermott, Tenenbaum, Freeman. ICCV'17, NIPS'17

Physical Scene Understanding

Goal

- Explaining and reasoning about data

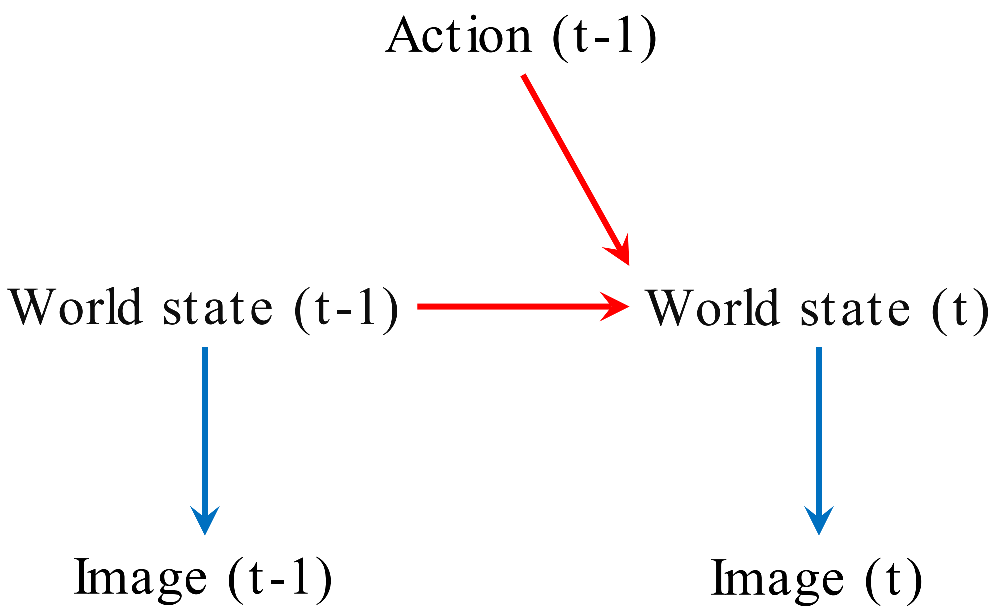
Approach

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

Advantages

Combining forward simulation engines and deep recognition networks.

- 1. Allowing learning with little or no supervision.
- 2. Offering rich generalization power.



Future Work

- How many shape details do we need?
    - Graphics and Vision vs. Robotics

Future Work

- How many shape details do we need?
  - Graphics and Vision vs. Robotics

- What are the right representational formats for 3D shape?
  - Voxels, Meshes, Point clouds, Procedures …

Future Work

- How many shape details do we need?
  - Graphics and Vision vs. Robotics

- What are the right representational formats for 3D shape?
  - Voxels, Meshes, Point clouds, Procedures …

- Can these representations be learned from externally observable data, or internally generated simulations? What has to be wired in?

Future Work

- How many shape details do we need?
  - Graphics and Vision vs. Robotics
- What are the right representational formats for 3D shape?
  - Voxels, Meshes, Point clouds, Procedures …
- Can these representations be learned from externally observable data, or internally generated simulations? What has to be wired in?
- Physics and 3D vision for more general shapes and scenes
  - Can we generalize the learned shape prior to unseen object categories?

Physical Scene Understanding