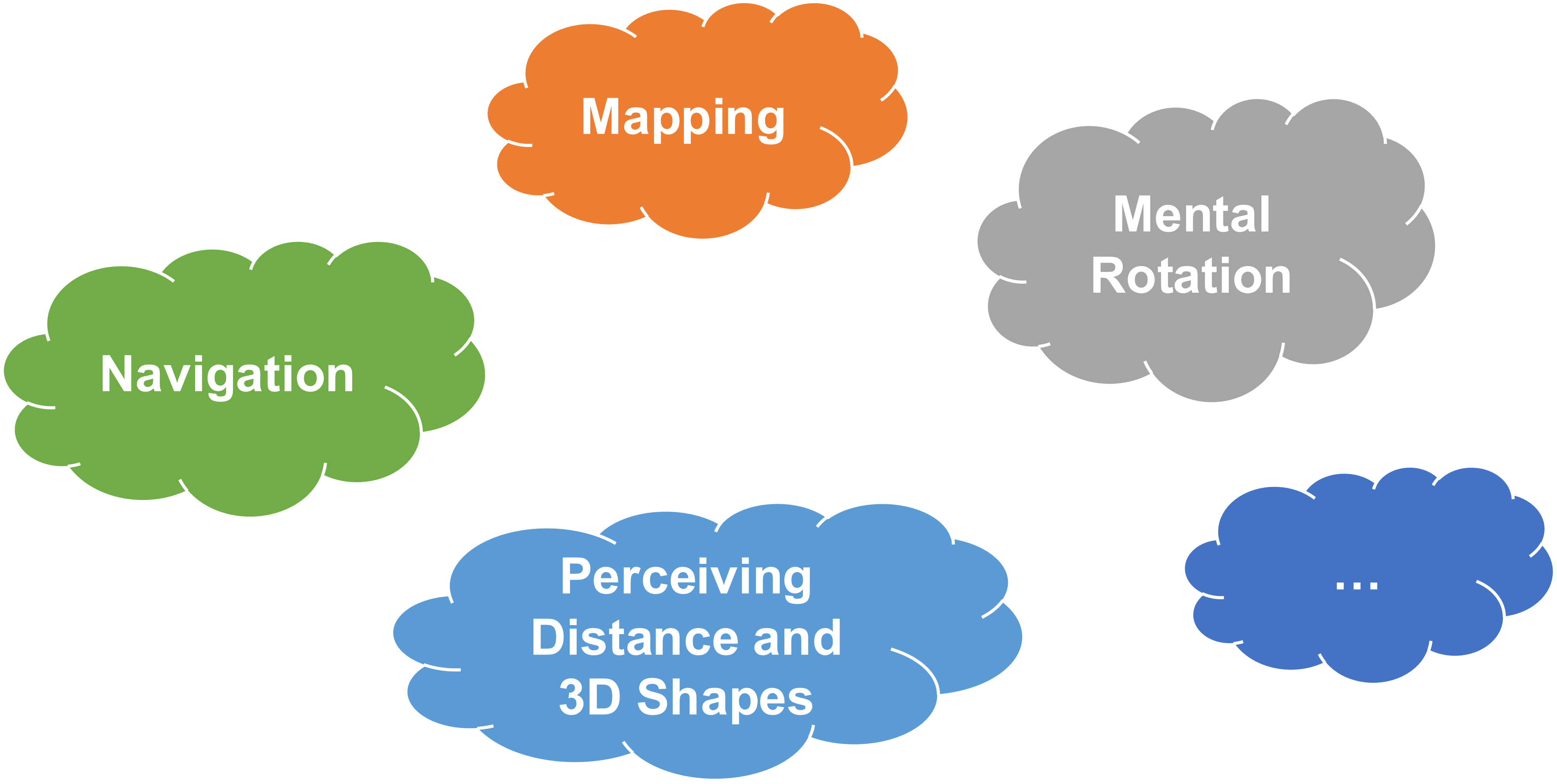
On Latent Abilities Underlying Spatial Intelligence

Qianqian Wang

MUSI Workshop @ ICCV

Oct 20, 2025

When We Talk About "Spatial Intelligence"…



But before any of that can happen…

- Before we can map or measure space, we have to *believe* that space — and the things within it — persist even when we're not looking

Object Permanence

"Peekaboo!"

Latent Ability 1: Understanding the world is persistent

Our world is not …

Movie "Everything Everywhere All at Once"

Our world is …

Latent Ability 2: The ability to update

The world is not static – it changes!   Our observation is always partial

"To Save Your Child Or Your Lawn Mower?"

Persistence and Update

Genie 3

Today's Talk

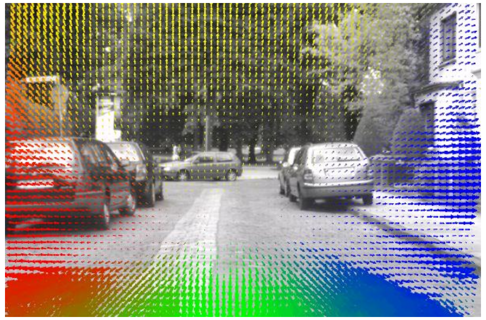
Persistence and Consistency → Motion and Structure

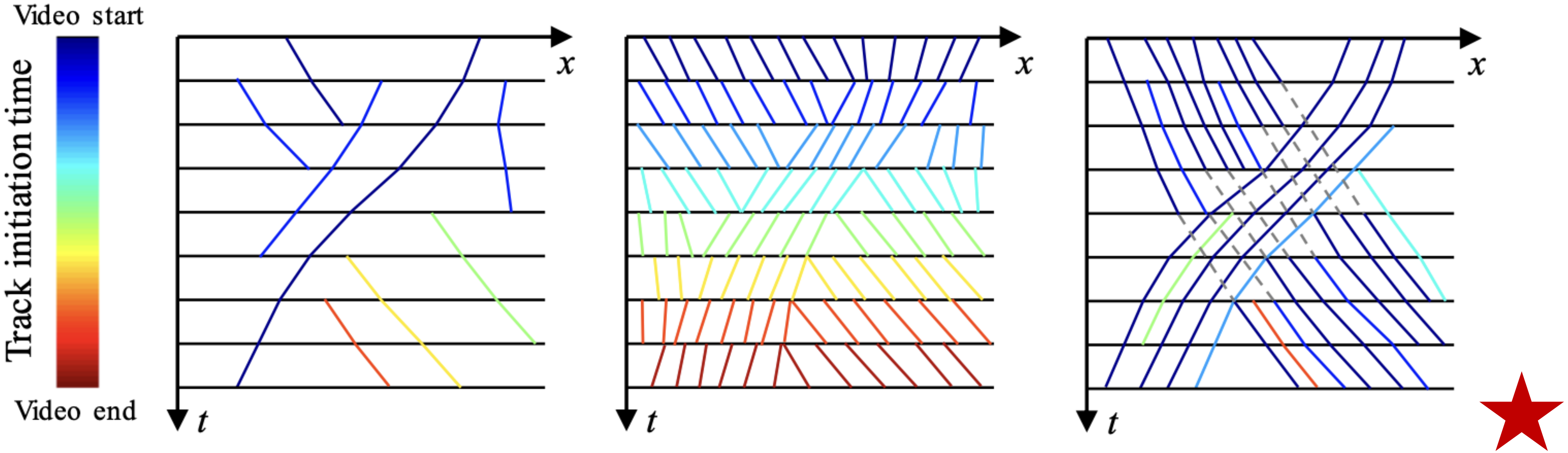
Wang et al. Tracking Everything Everywhere All at Once.
ICCV 2023 (**Best Student Paper**)

A Continuously-Updating 3D Perception Framework

Wang et al. Continuous 3D Perception with Persistent State.
CVPR 2025 (**Oral**)

Motion Estimation





Sparse feature tracking
(long but sparse)

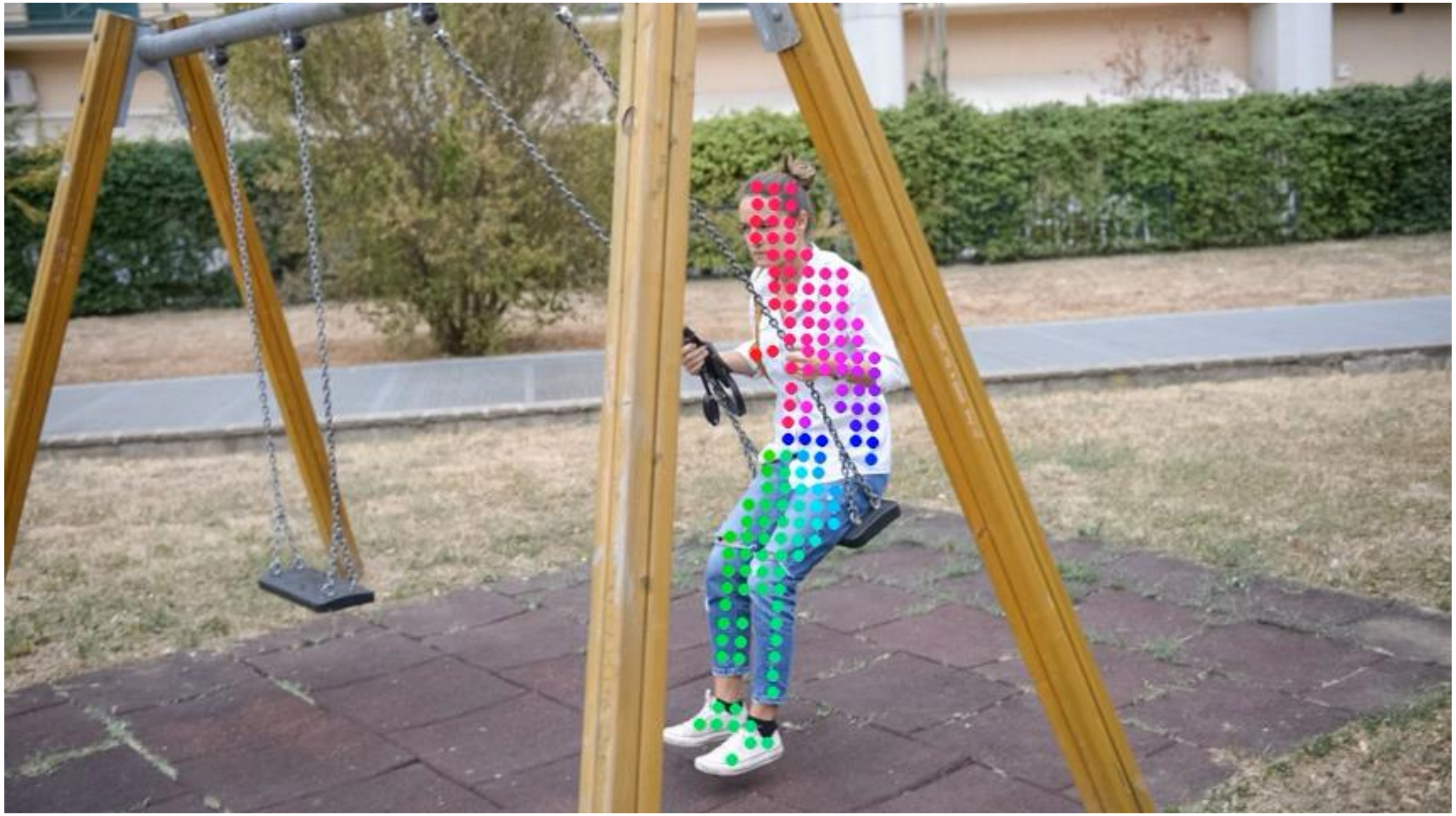
Optical flow
(dense but short)

Dense and long-range motion

Rubinstein, Michael, Ce Liu, and William T. Freeman. "Towards longer long-range motion trajectories." (2012).

Chaining Optical Flow for Long-Range Motion?

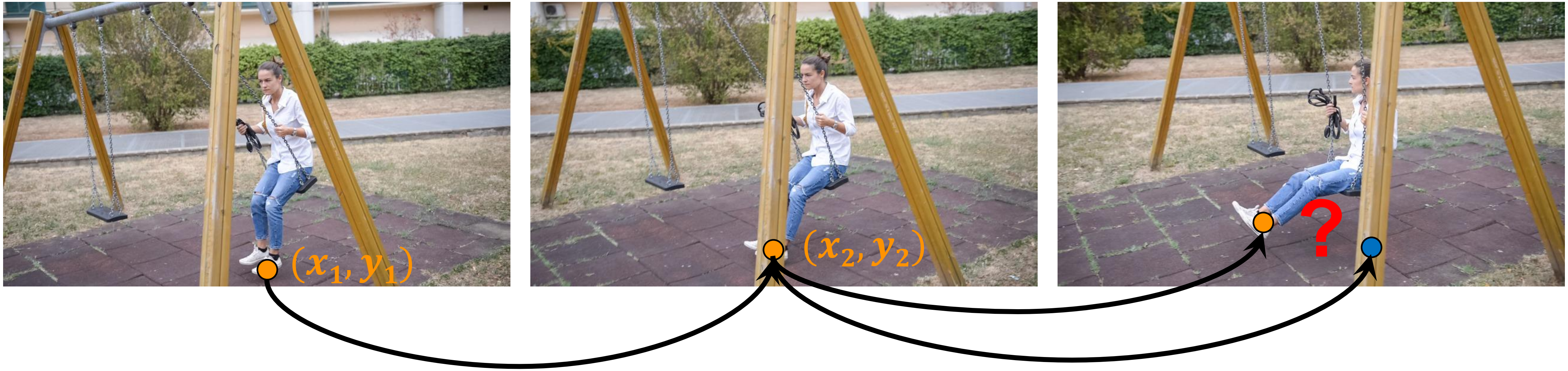
1→2→3→...→N



Challenge 1: Occlusion

Modeling motion in the 2D pixel space!

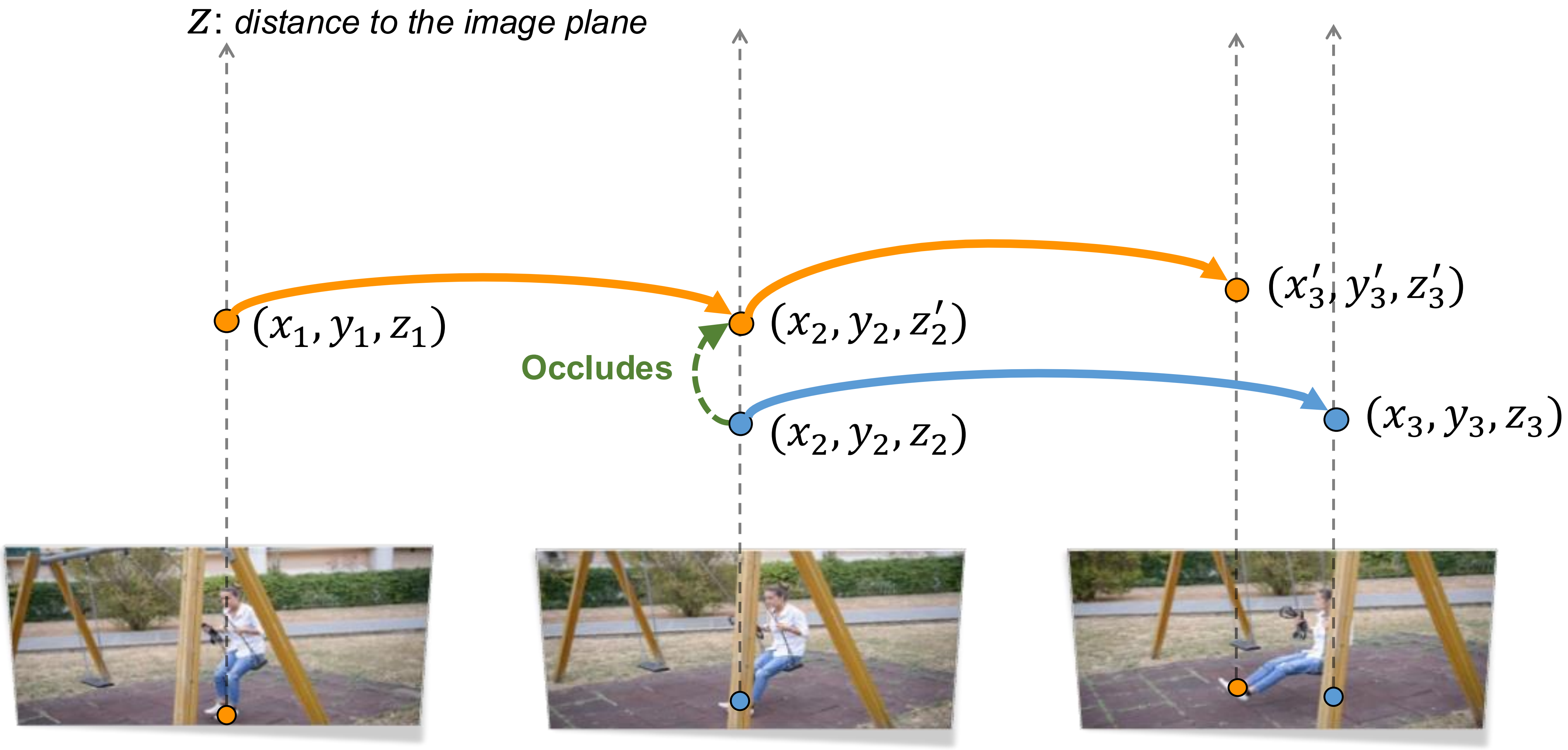
2D mapping function: $[x', y'] = \boldsymbol{f}([x, y])$



A point on the **swing set frame**
or on the **shoe**?

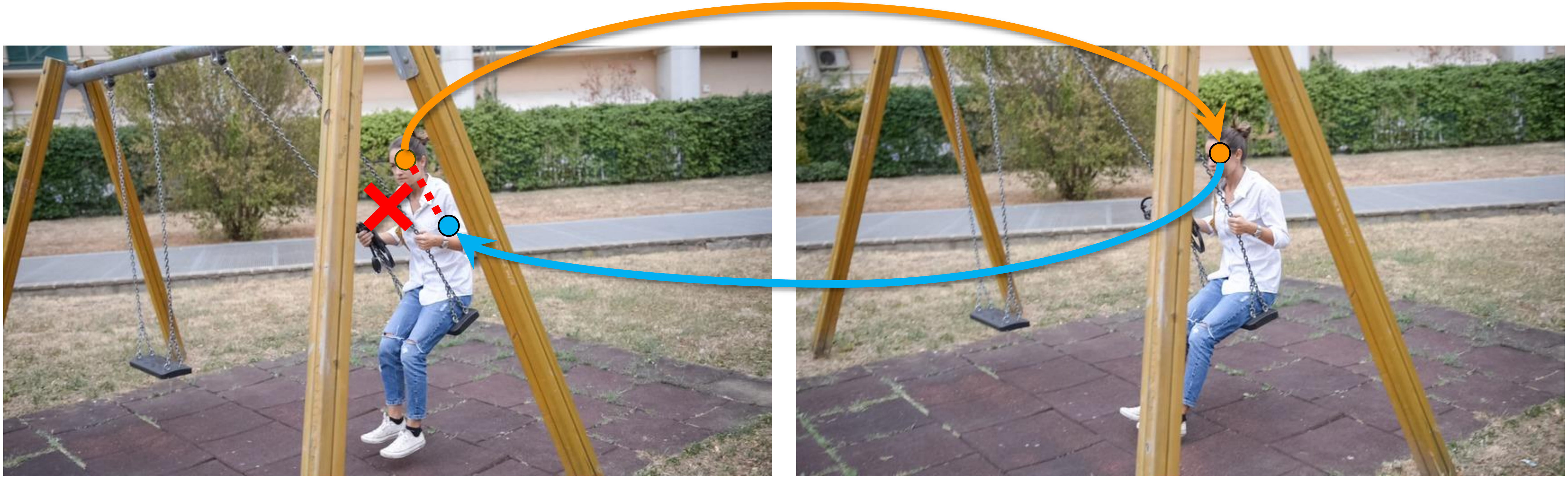
The World is 3D

We should model motion in 3D space



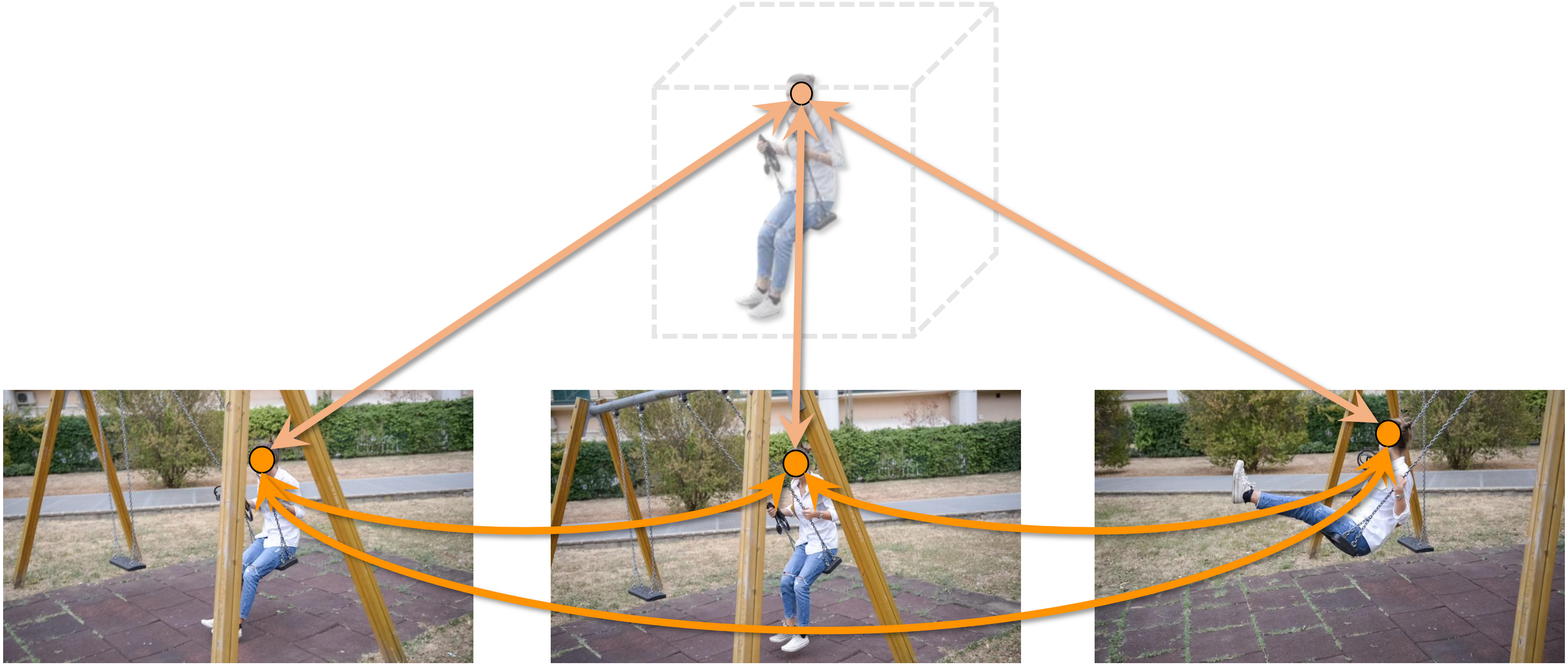
Challenge 2: No Guarantee of Cycle Consistency

$$\boldsymbol{f}_{j\to i}(\boldsymbol{f}_{i\to j}([x,y]_i)) \,!= [x,y]_i$$



Correspondences Are Cycle Consistent

Global Cycle Consistency



Key Insights

We need:

- A **3D representation**
- A representation that ensures **global cycle consistency**

OmniMotion

Test-Time optimization (per-video)

OmniMotion

- **Complete (Any-to-Any)**
- **Handling Occlusion**
- **Globally Consistent**

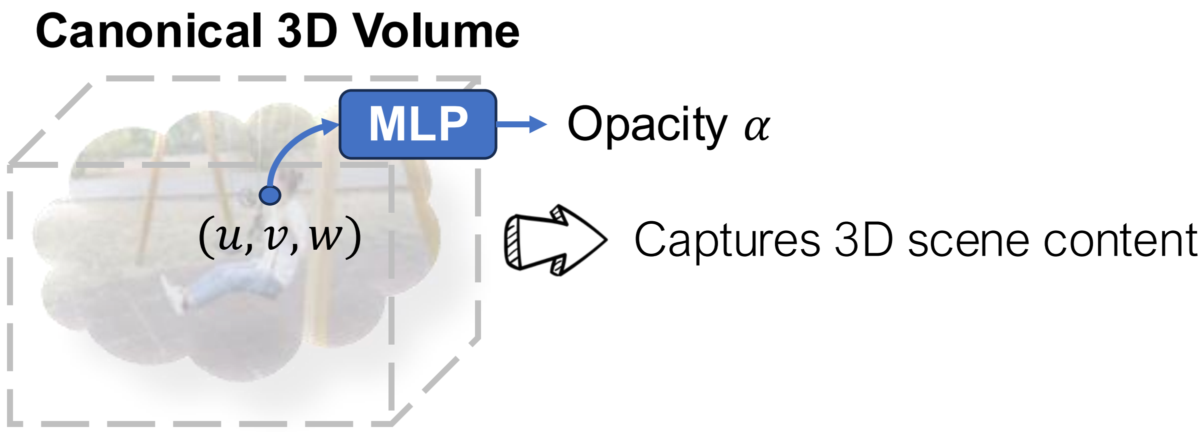
Query Frame

Target Frames

Visible   Occluded

● Visible + Occluded

OmniMotion: The Motion Representation

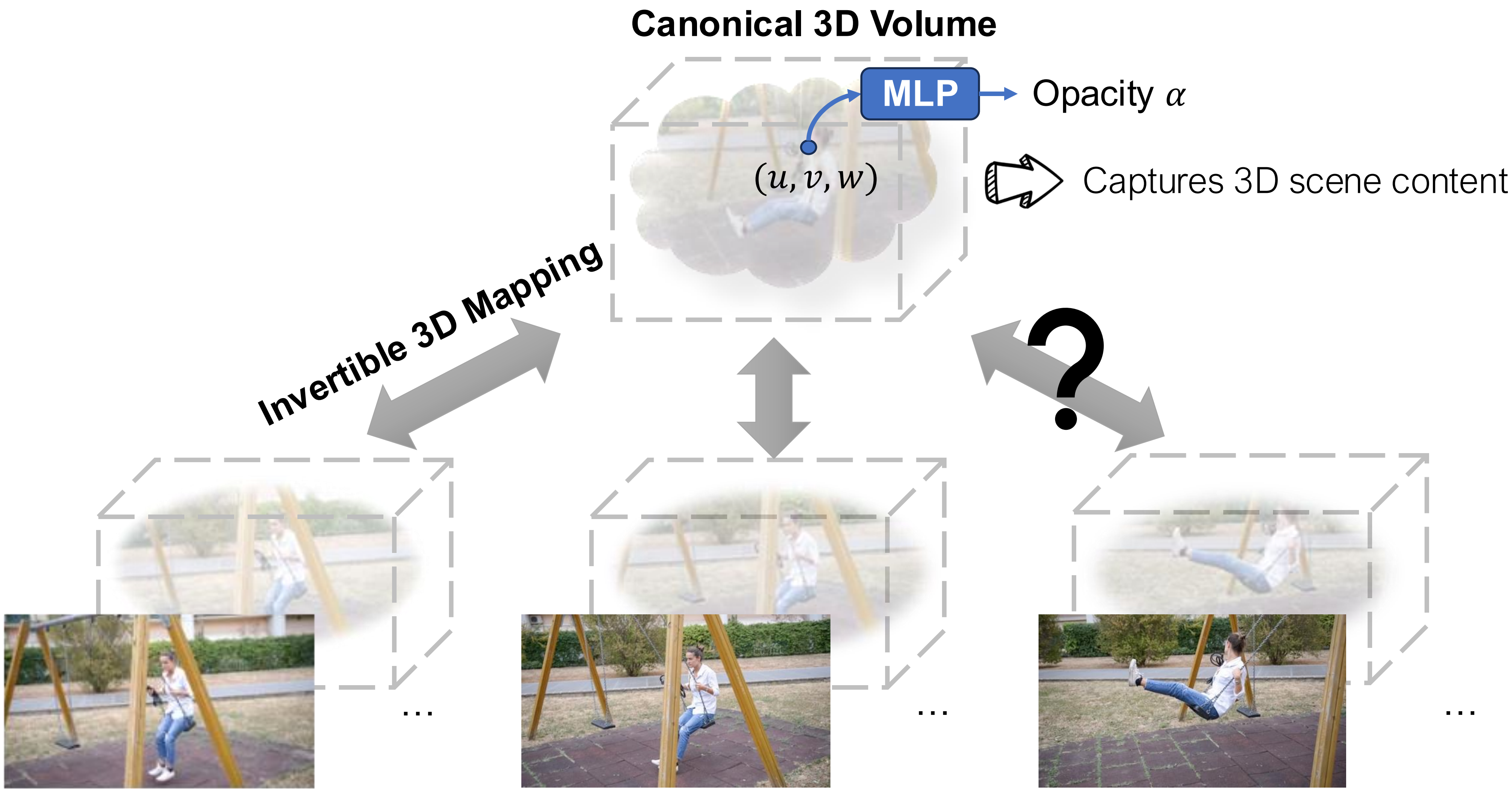


…

…

…

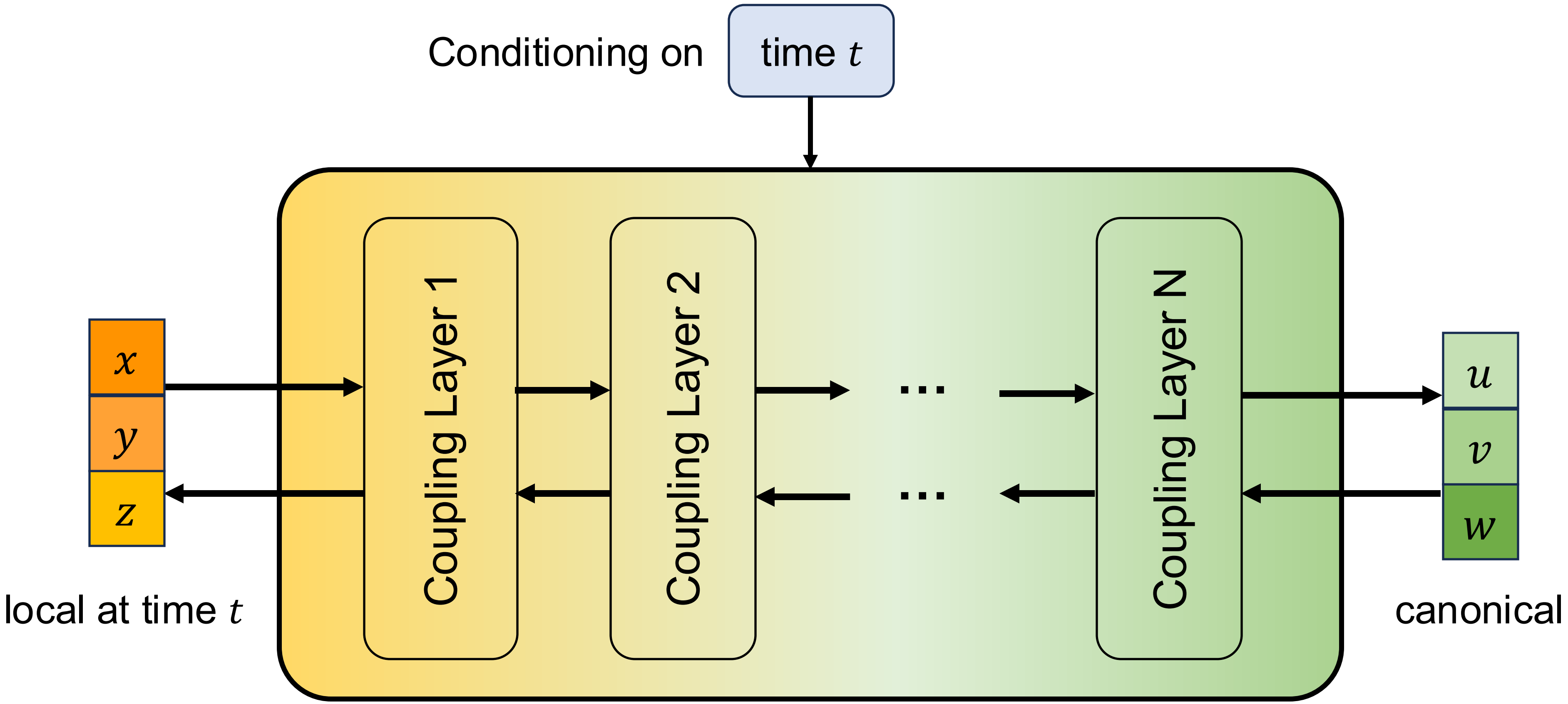
OmniMotion: The Motion Representation



Invertible 3D Mapping

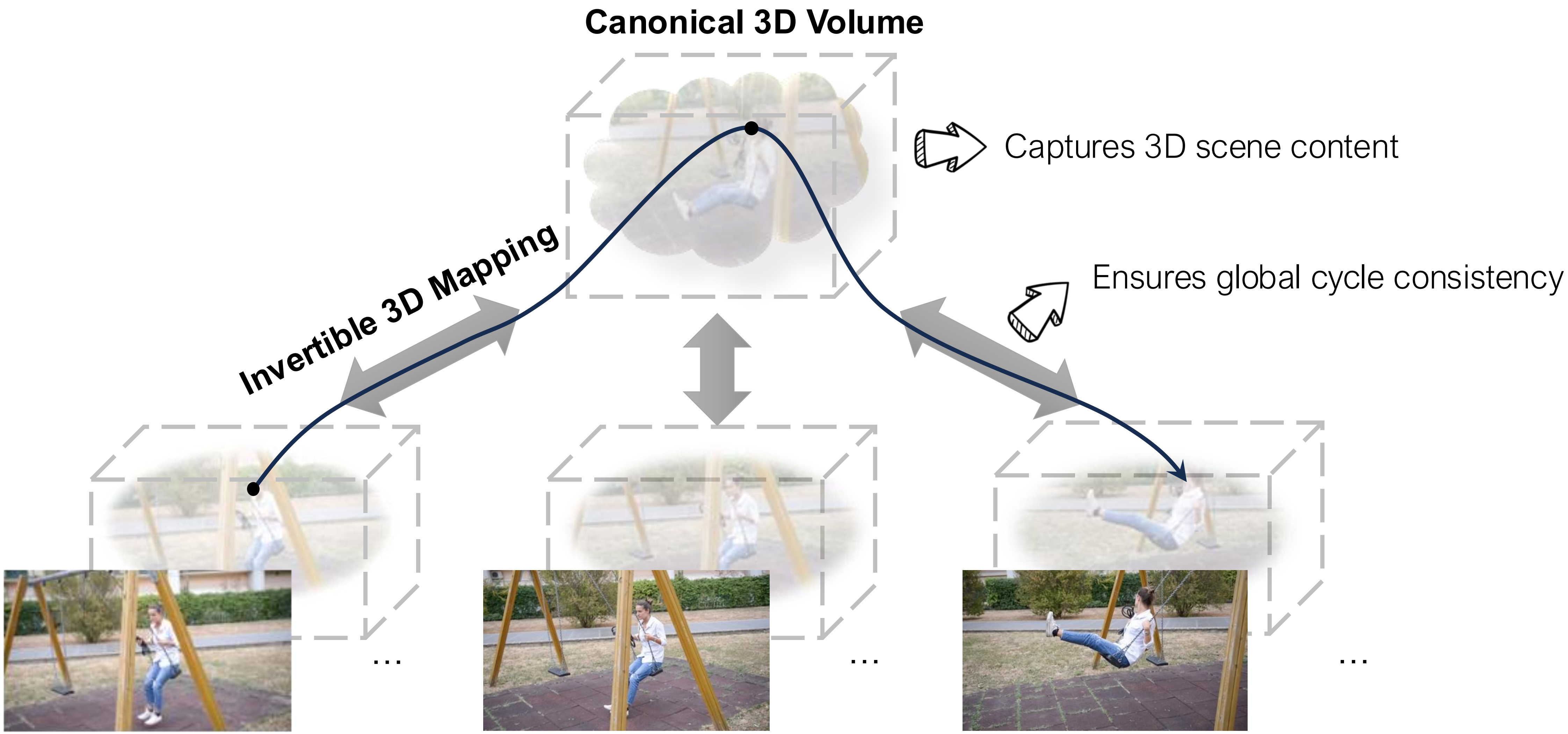
Invertible Neural Networks

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x});\ \boldsymbol{x} = \boldsymbol{f}^{-1}(\boldsymbol{y})$$

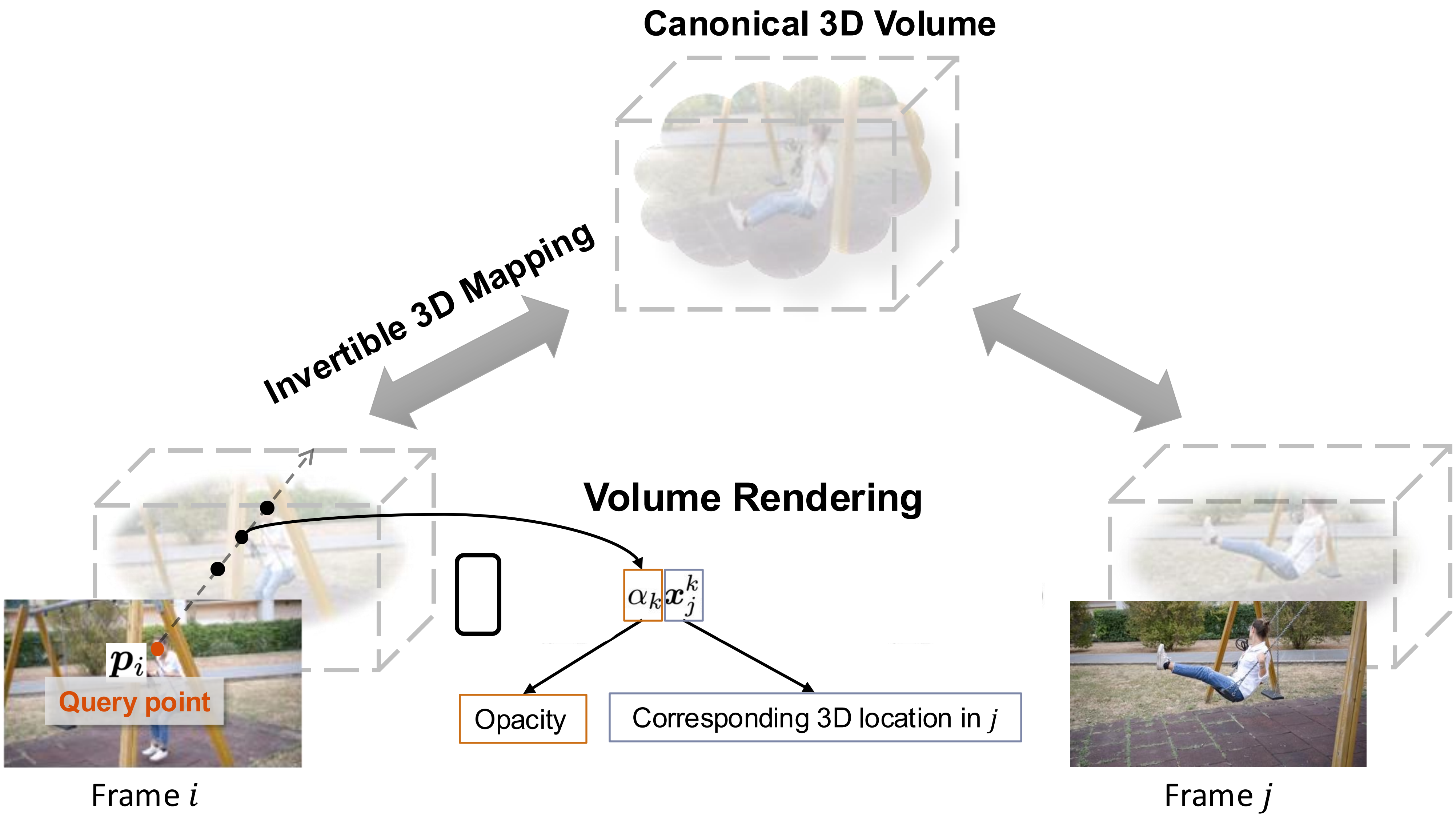


Dinh et al. "Density Estimation using Real NVP." ICLR 2017

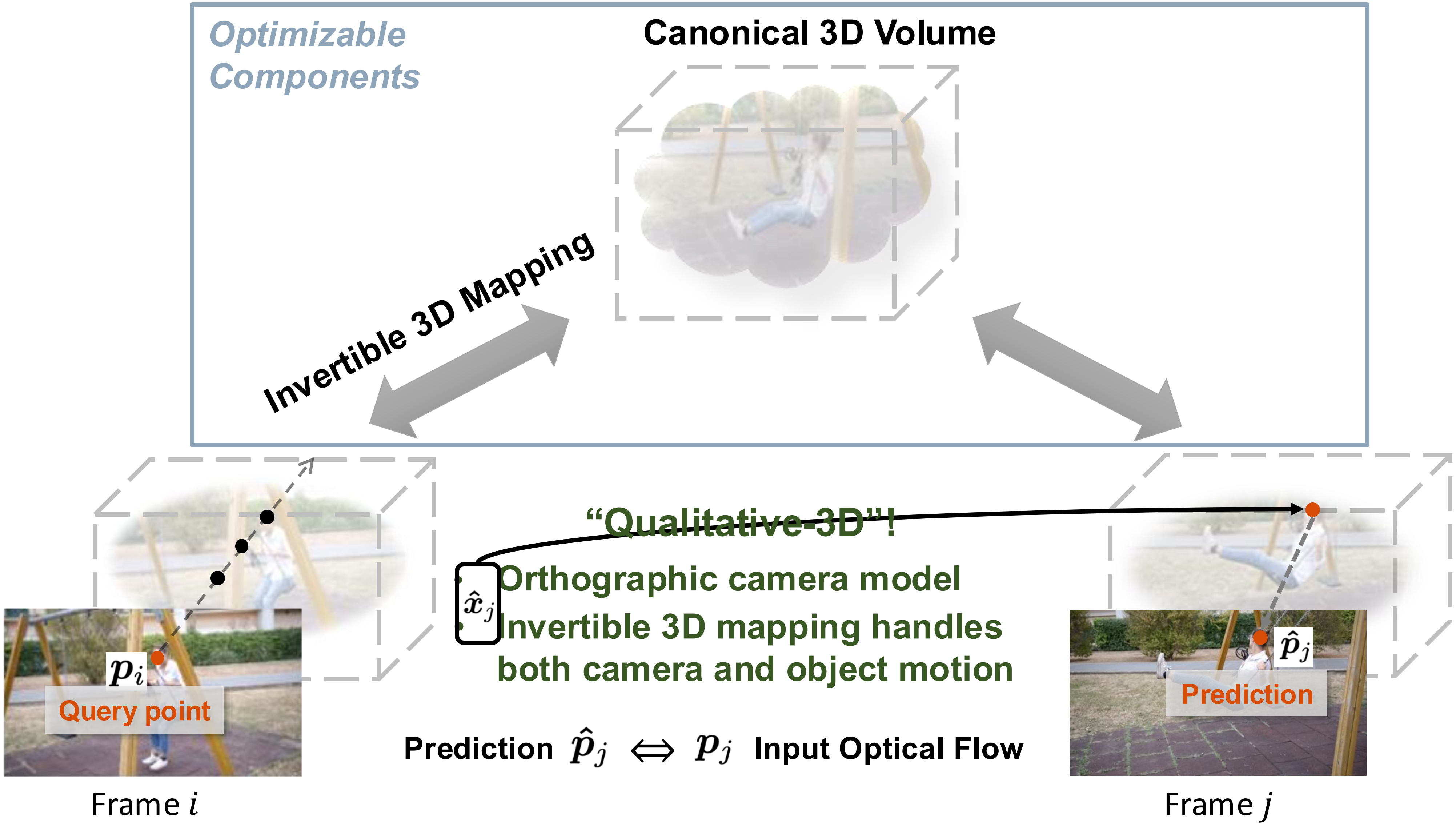
OmniMotion: The Motion Representation



How to Compute 2D Motion?



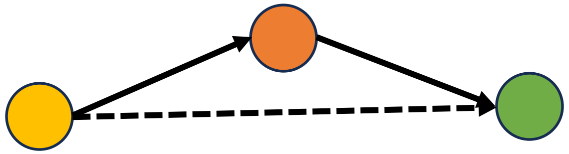
How To Optimize?

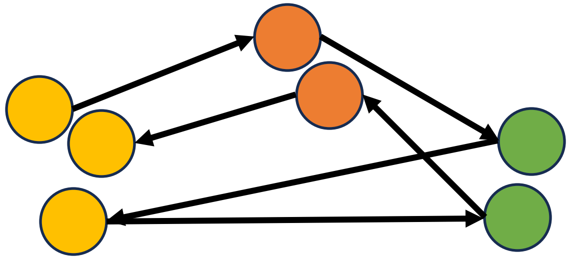


Zachary and Deng. "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow." ECCV 2020.

How Do We Improve upon Input Optical Flow?

Built-In cycle consistency guarantee!

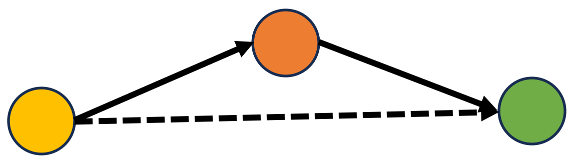




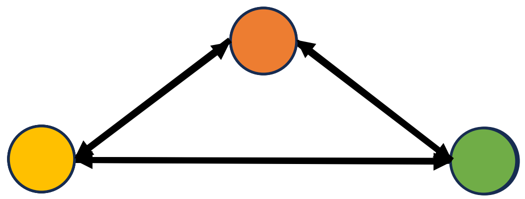
Connect short-ranged motion

How Do We Improve upon Input Optical Flow?

Built-In cycle consistency guarantee!



Connect short-ranged motion

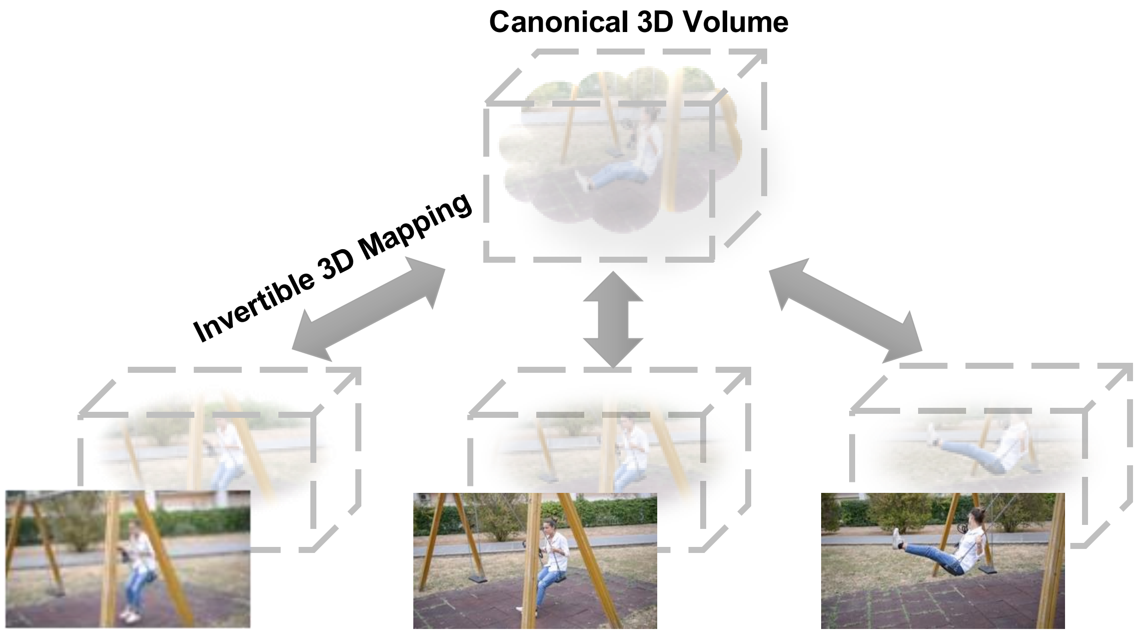


Consolidate inconsistent motion

Connection to Classical 3D Reconstruction

OmniMotion

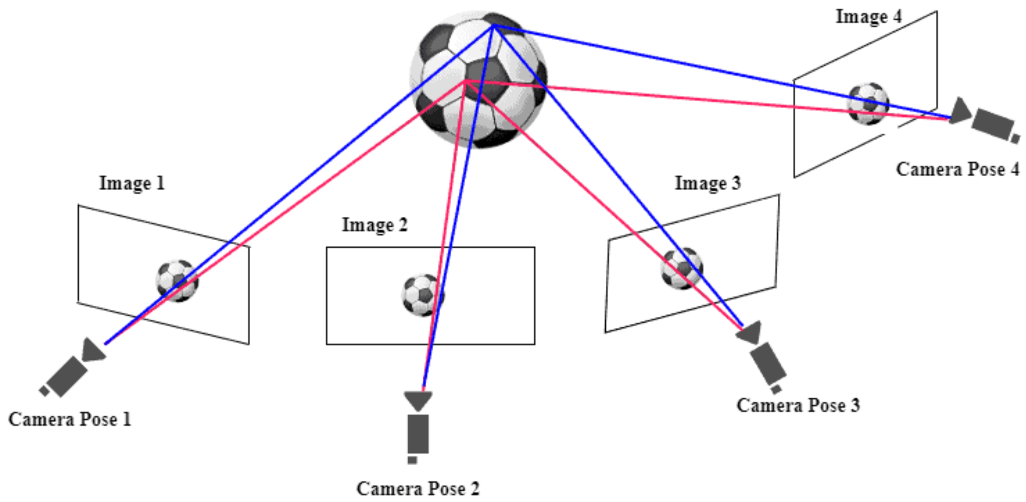
(For 2D Tracking)



Bundle Adjustment

[Triggs et al. ICCV'99]

(For Static 3D Reconstruction)



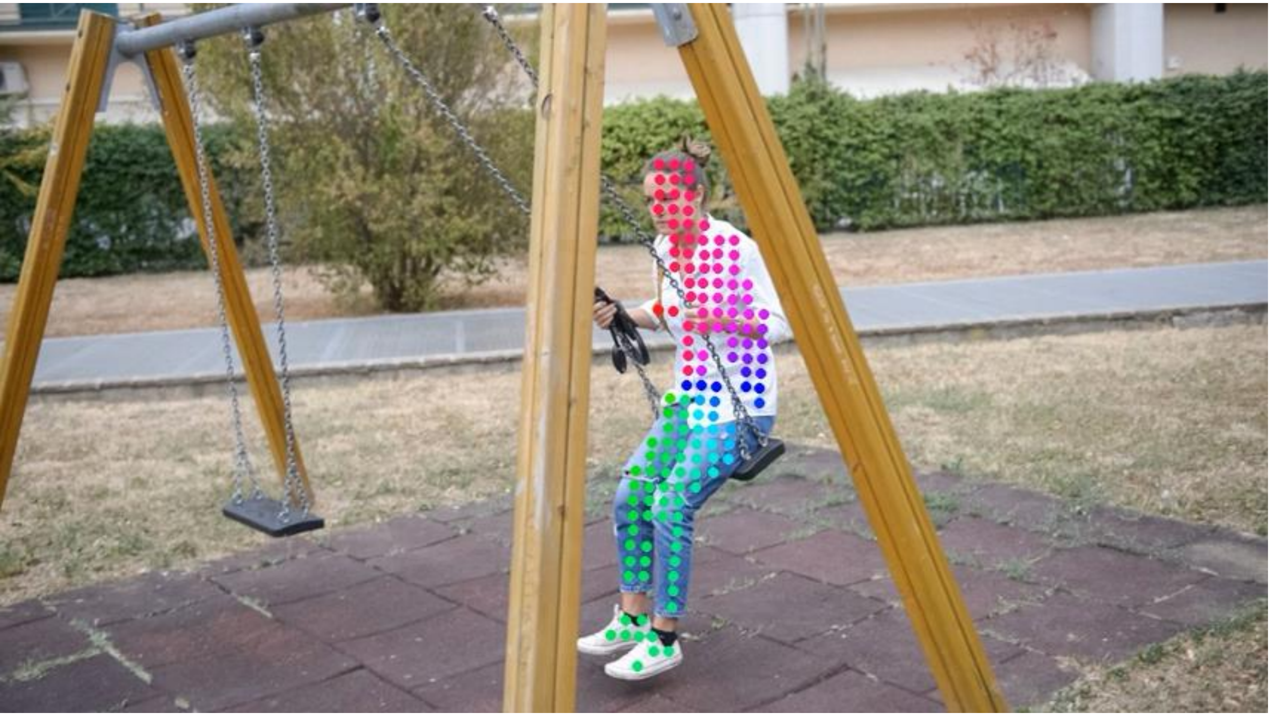
Invertible 3D Mapping = A neural network that subsumes both camera and object motion

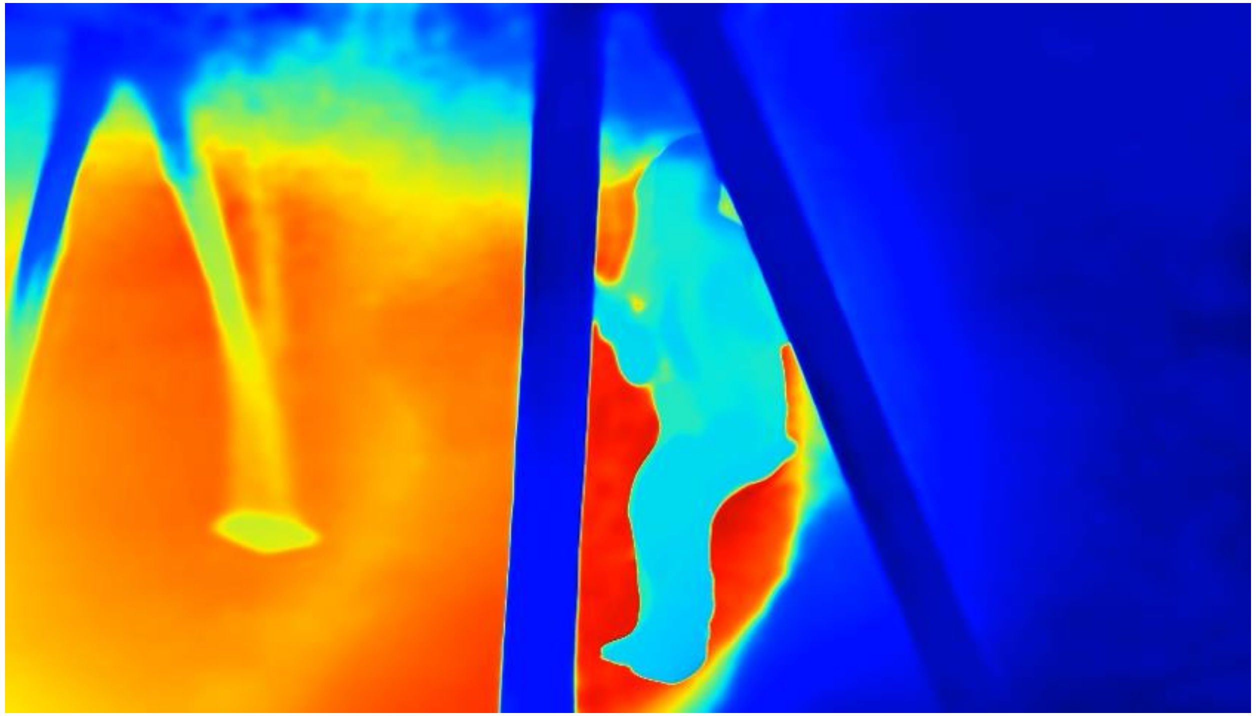
Invertible 3D Mapping = SE(3) Camera motion

Both leverage the underlying structure of the world — consistency and persistence

https://www.baeldung.com/cs/computer-vision-bundle-adjustment

Structure Emerges from Tracking



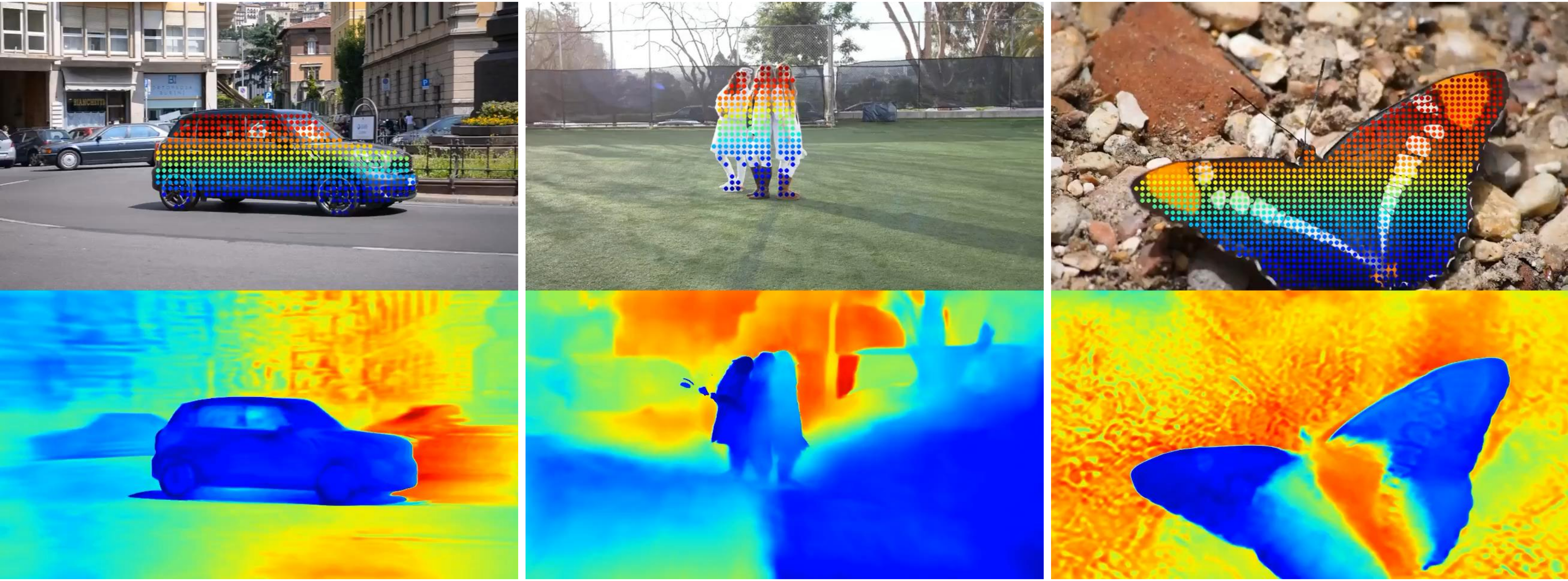


visible occluded

near far

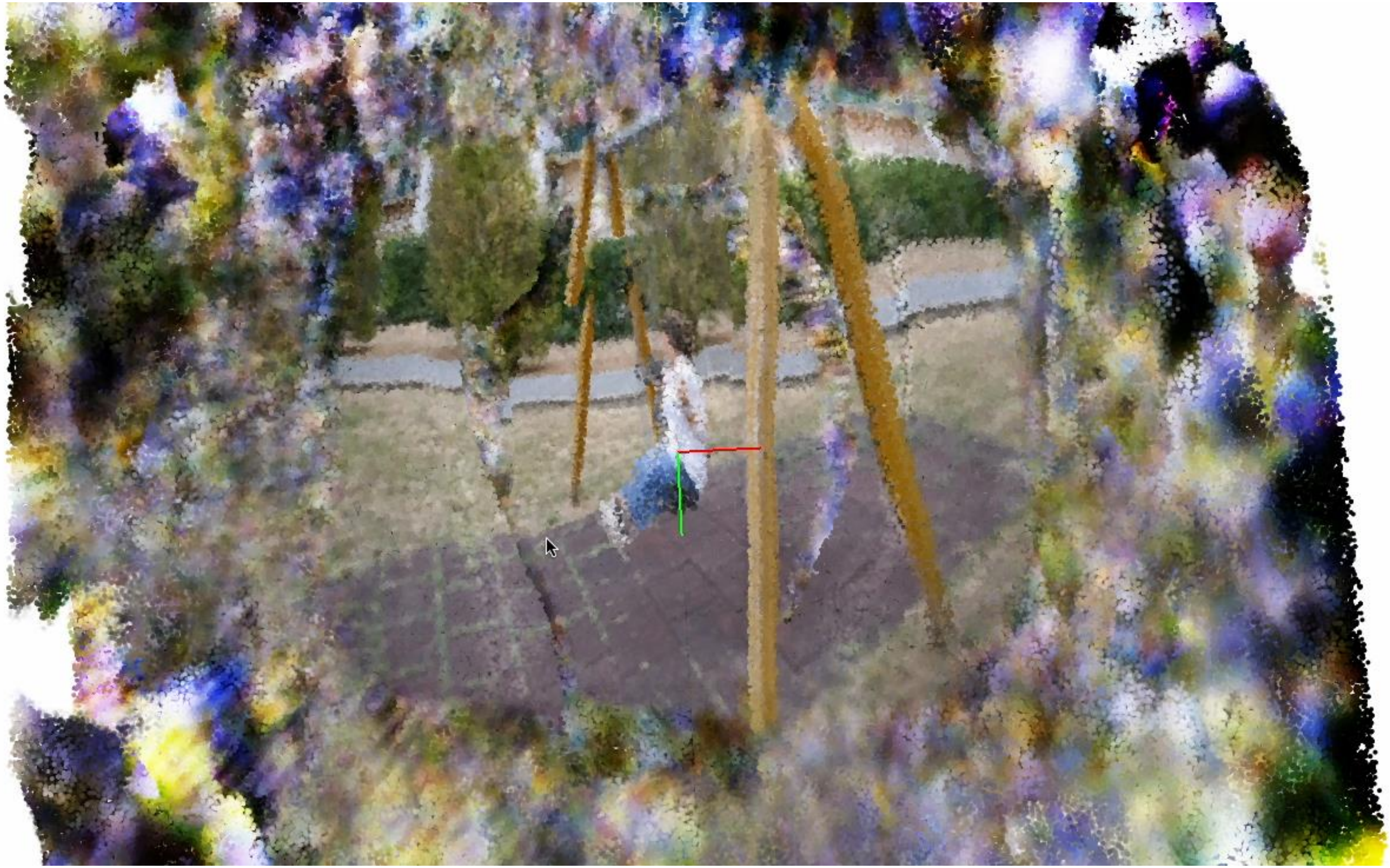
Pseudo Depth

Structure Emerges from Tracking



No explicit 3D supervision or input!

A Visualization of Canonical 3D Volume



Summary

- Consistency and persistence can give rise to motion and even pseudo geometry understanding
- However, limitations exist:
  - Per-Video optimization is slow, offline and not scalable
  - Bijection can be overly restrictive
  - The optimization is highly non-convex and ill-conditioned

Open Question: How to learn an online, feed-forward system that preserves consistency, without being overly restrictive?

Today’s Talk

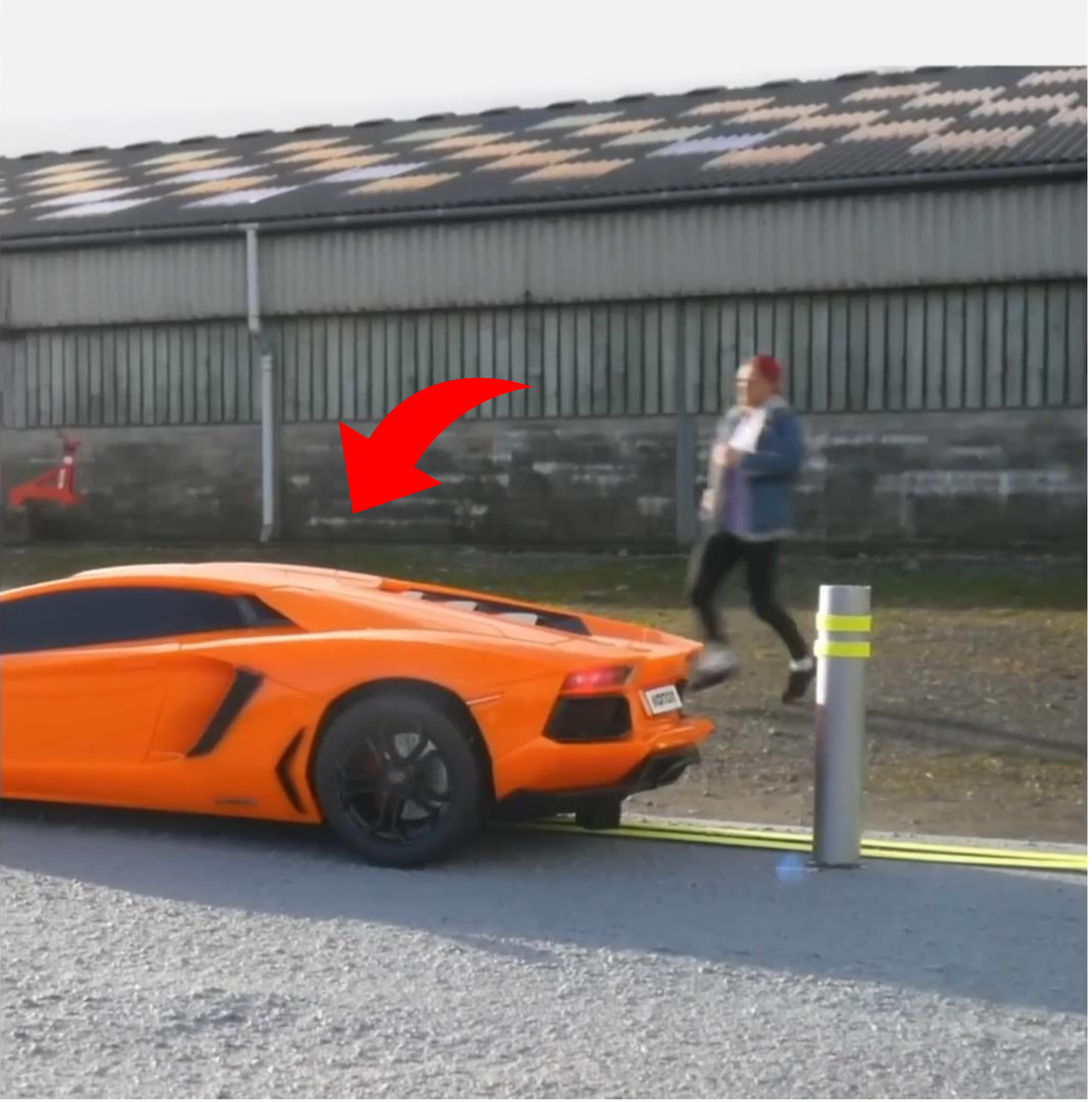
Persistence and Consistency → Motion and Structure

Wang et al. Tracking Everything Everywhere All at Once.
ICCV 2023 (**Best Student Paper**)

A Continuously-Updating 3D Perception Framework

Wang et al. Continuous 3D Perception with Persistent State.
CVPR 2025 (**Oral**)

How Do We Perceive the Visual World?



https://www.youtube.com/watch?v=JG-m9Ypc7Is

We see the world through our past experience

Data-Driven Priors

How Do We Perceive the Visual World?

Data-Driven Priors

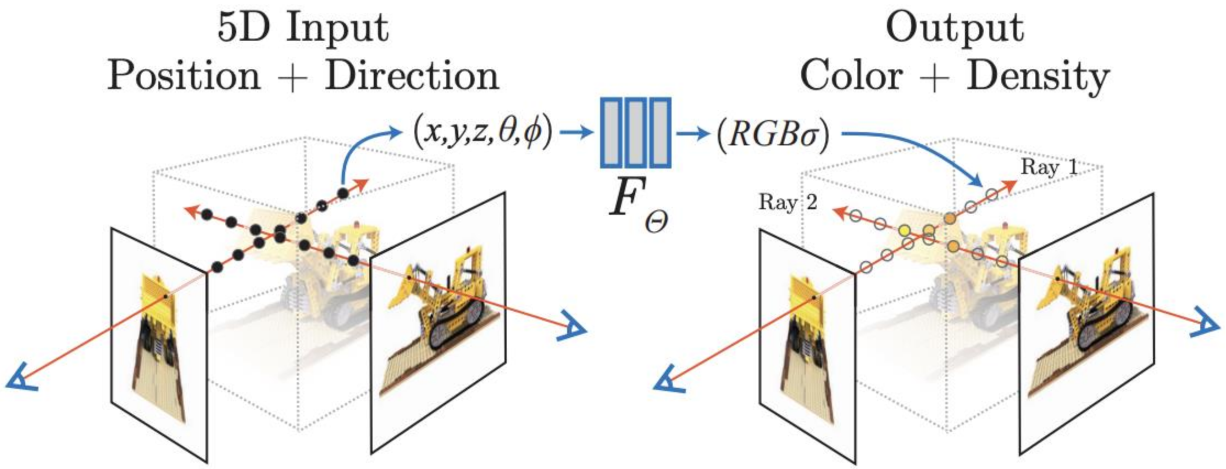
Online, Continuous Update

Efficient Accurate

https://www.youtube.com/watch?v=JG-m9Ypc7Is

Prior Art: *Tabula Rasa* Reconstruction

SfM / SLAM



NeRF
[Mildenhall et al. ECCV'20]

Data-Driven Priors

Not learning from past experience

Prior Art: *Tabula Rasa* Reconstruction

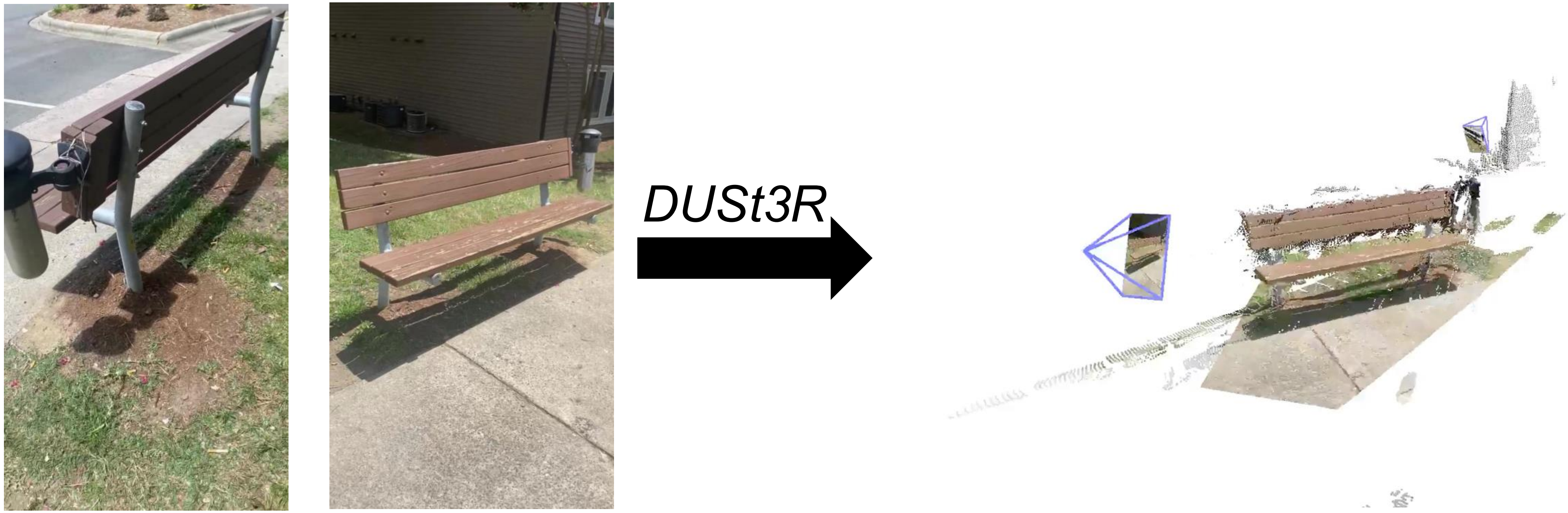
Do not work in under-constrained settings

Single Image

Moving Objects

Prior Art: Learning-Based 3D Methods

Learning rich data-driven priors about the 3D world



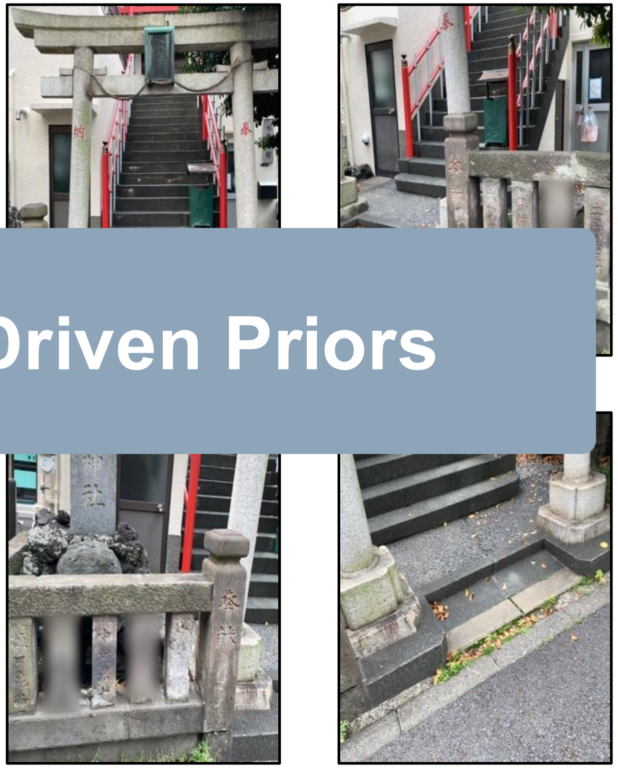
Online, Continuous Update

Only works for a pair of images

Wang et al. “DUSt3R: Geometric 3D Vision Made Easy.” CVPR 2024

Online Framework for 3D Perception

- Reconstructing 3D scenes from **few observations**



Input: sparse photo collections

Data-Driven Priors + **Online, Continuous Update**

Online Framework for 3D Perception

- Reconstructing 3D scenes from **few observations**
- Inferring unseen regions **beyond observations**

Input View

Online Framework for 3D Perception

- Reconstructing 3D scenes from **few observations**
- Inferring unseen regions **beyond observations**
- **Continuously updating** the reconstruction with more observations

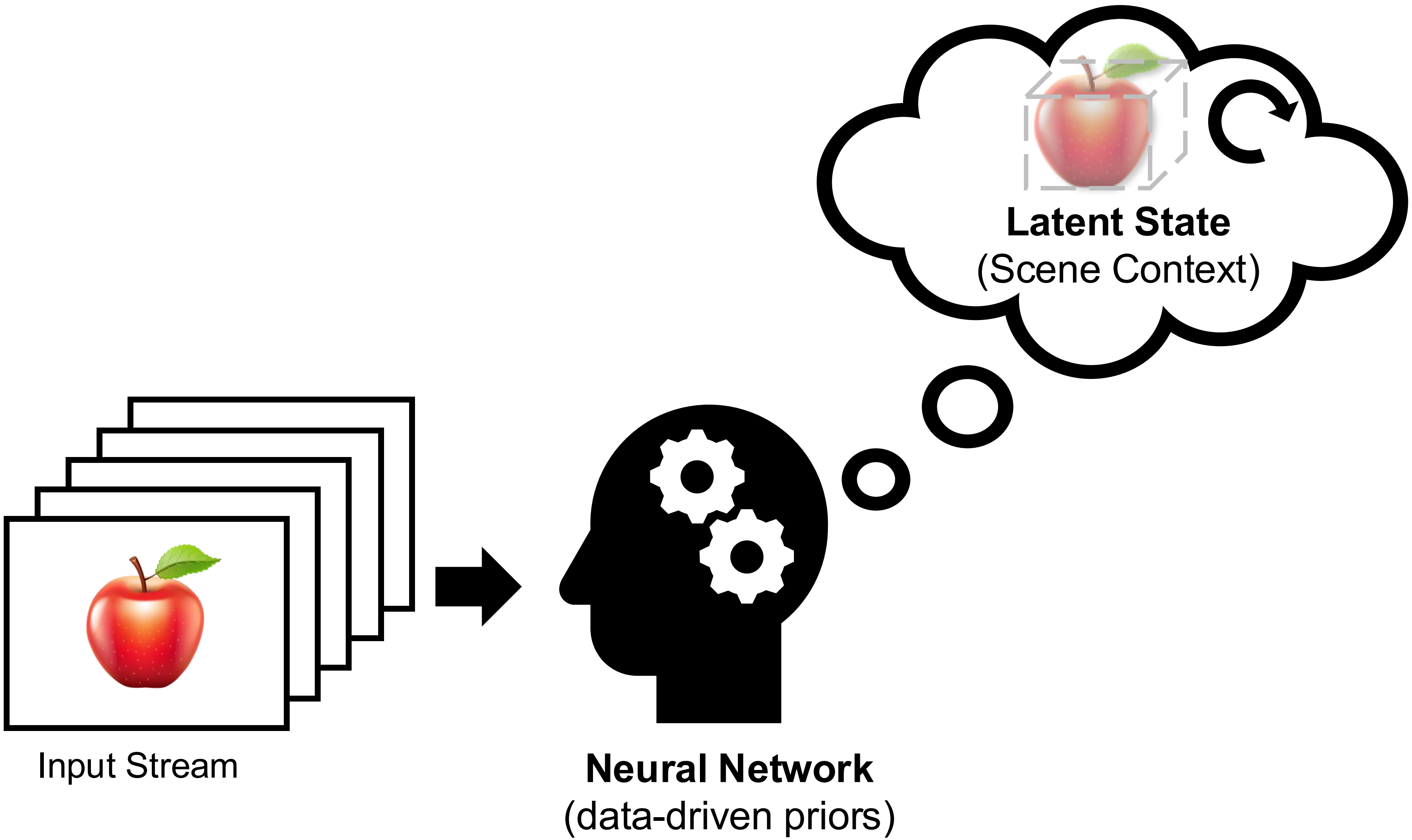
Static Scenes

Input stream

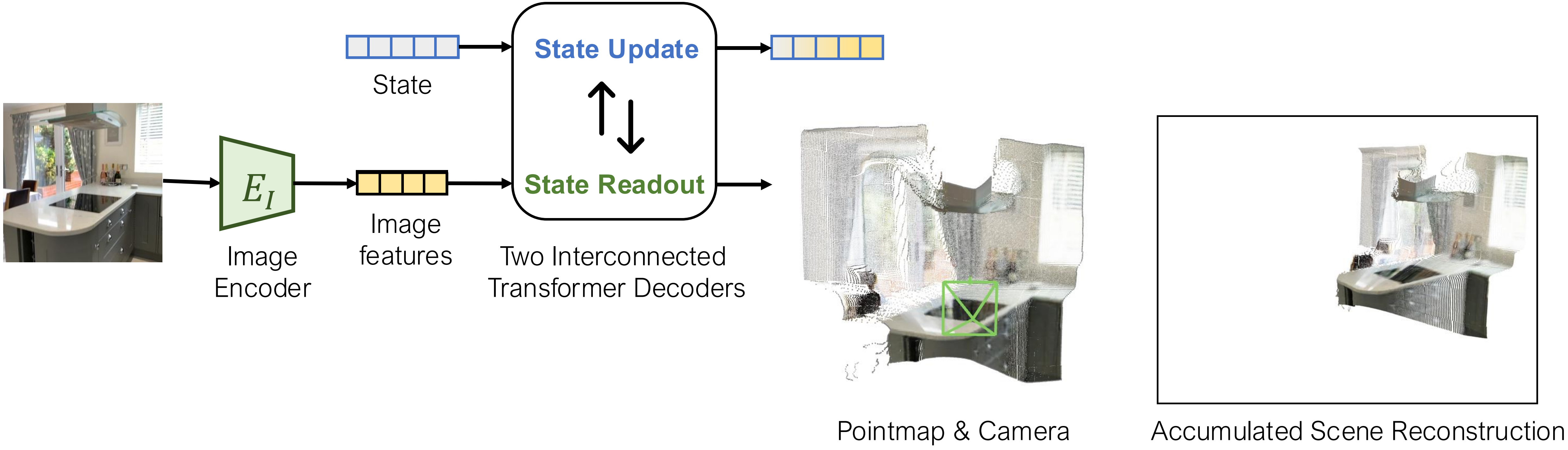
Dynamic Scenes

Input stream

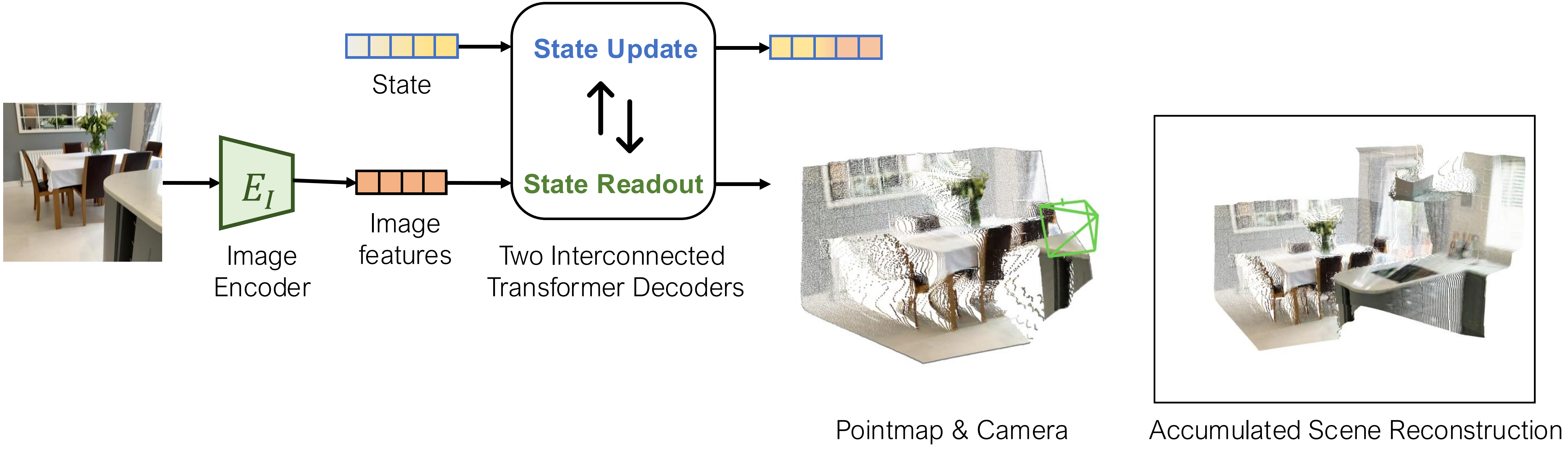
Key Idea


Latent State
(Scene Context)
Input Stream
Neural Network
(data-driven priors)

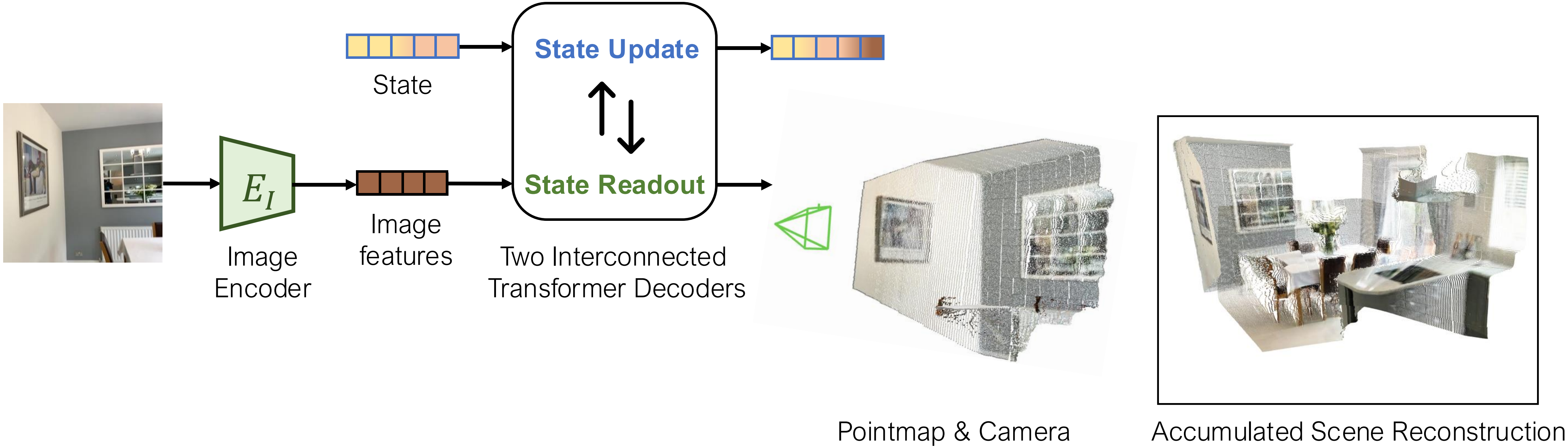
Our Approach: CUT3R



Our Approach: CUT3R



Our Approach: CUT3R



Flexible: Static & Dynamic Scenes; Videos & Unstructured Photo Collections

Online Reconstruction for Static Scenes

View 1

View 2

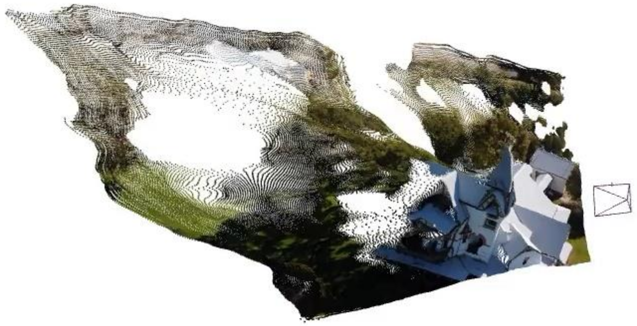
Online Reconstruction for Static Scenes

View 1

View 2

Online Reconstruction for Static Scenes

View 1



View 2

Online Reconstruction for Dynamic Scenes

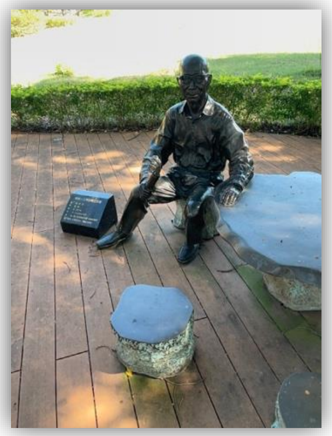
Input video

Online Reconstruction for Dynamic Scenes

Input video

Online Reconstruction for Dynamic Scenes

Online Reconstruction for Photo Collections



1

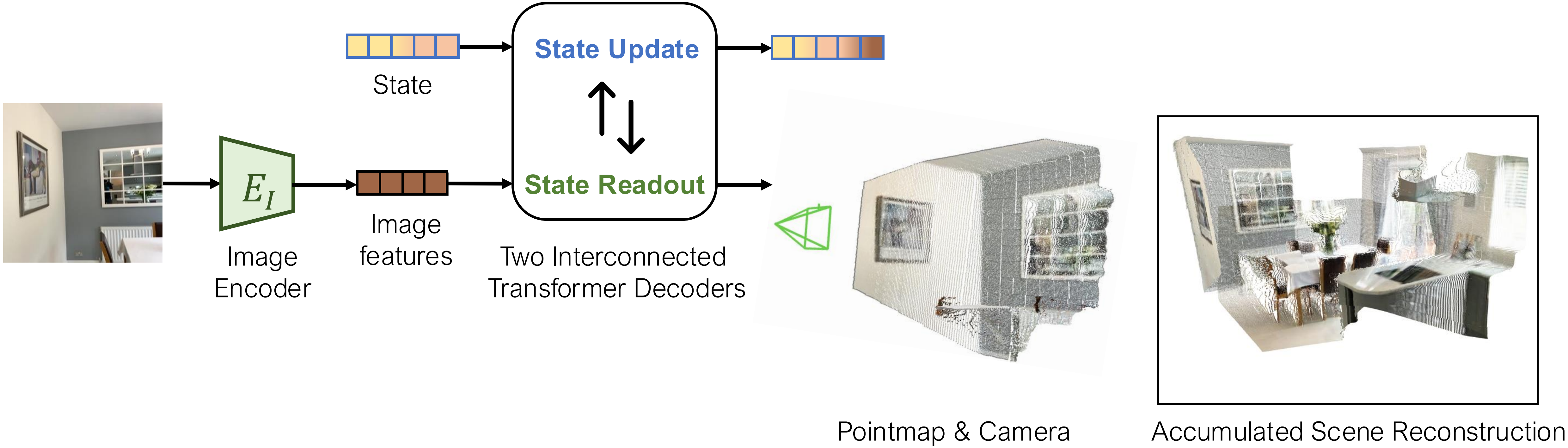
2

3

4

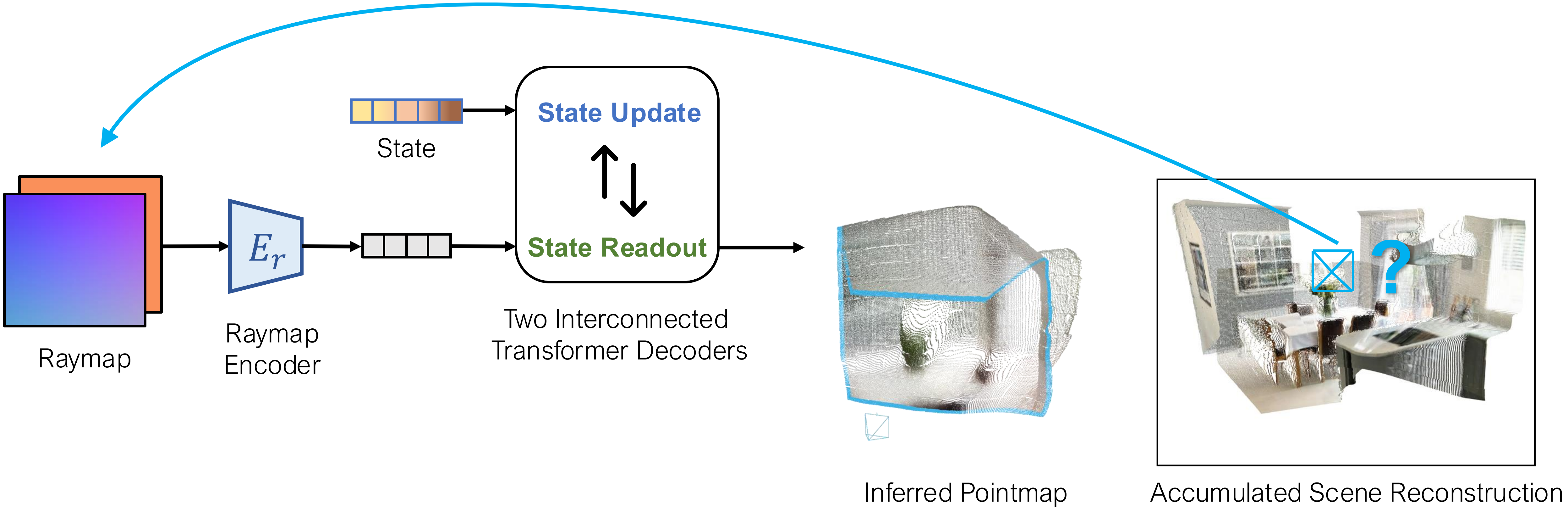
Input images

Our Approach: CUT3R



Flexible: Static & Dynamic Scenes; Videos & Unstructured Photo Collections

What's Inside the State?



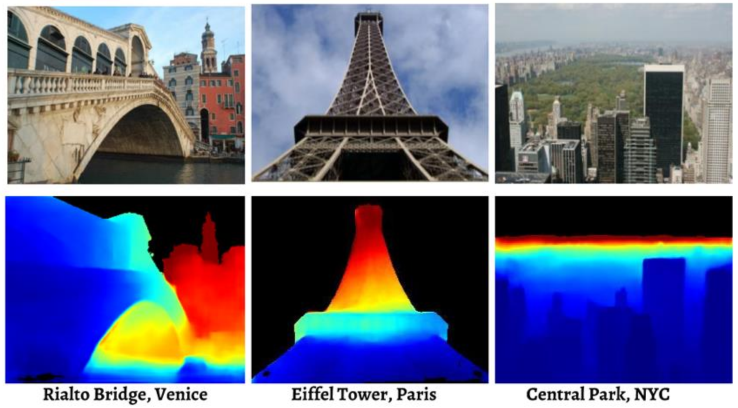
Inferring New Structures

Scene Reconstruction with Inferred Pointmap

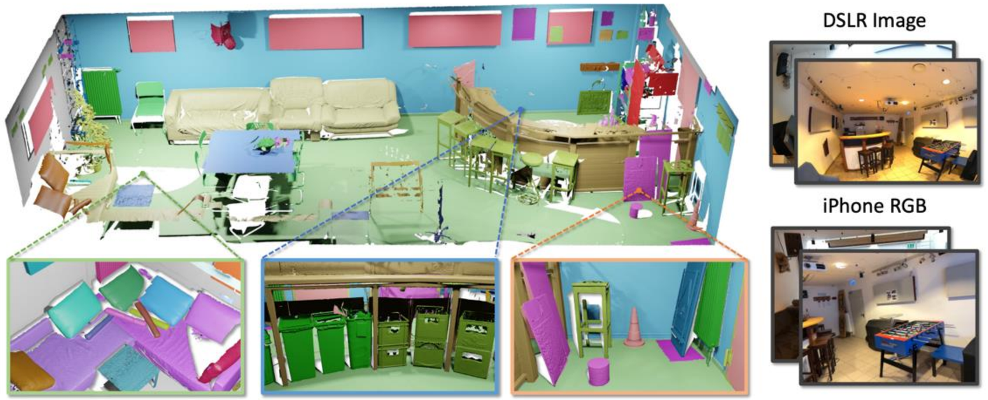
Inferring New Structures

Large-Scale Training on Diverse Datasets

ARKitScenes



MegaDepth

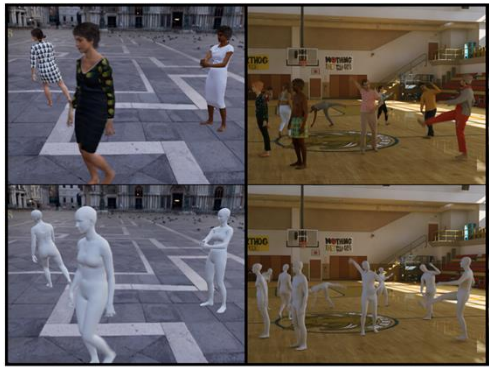


ScanNet++

Waymo Dataset

TartanAir

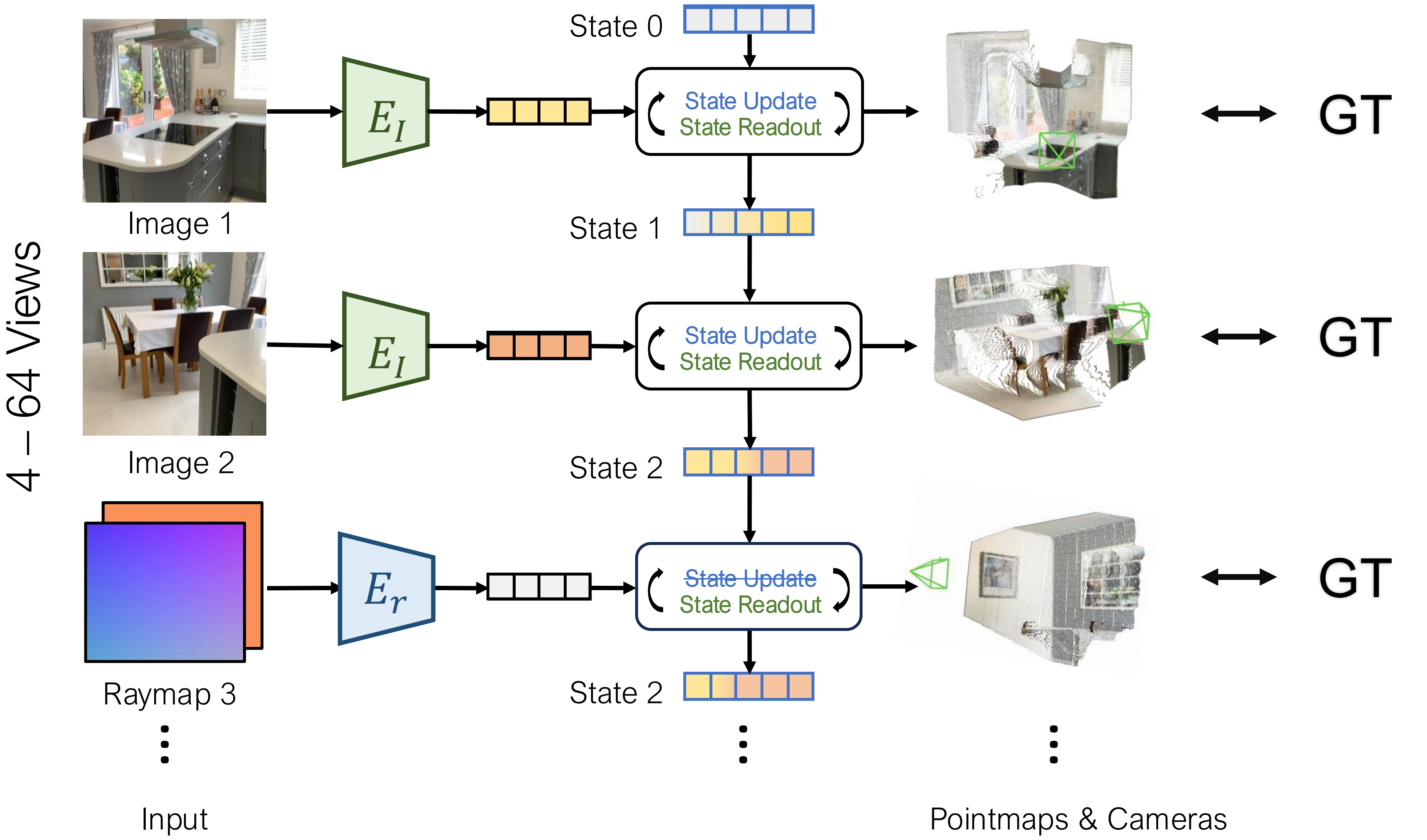
CO3D v2



BEDLAM

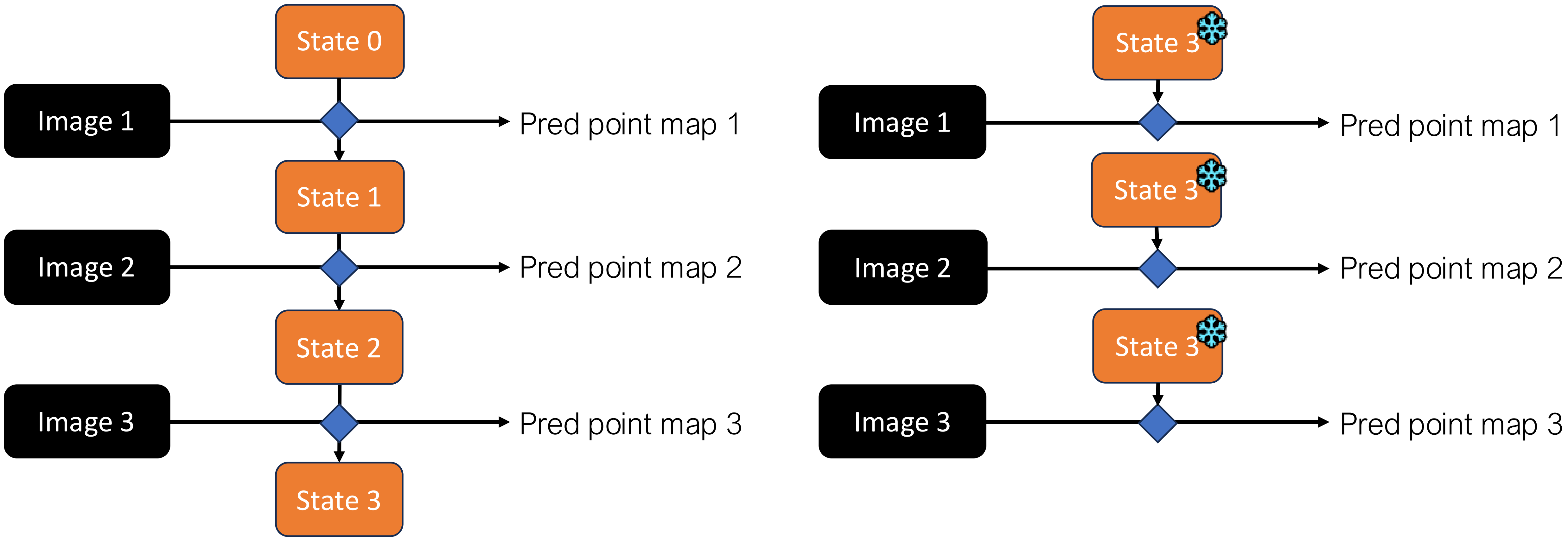
32 Datasets, ~12M images, ∞ sequences

Training



4 – 64 Views
Image 1
Image 2
Raymap 3
E_I
E_I
E_r
State 0
State 1
State 2
State 2
State Update
State Readout
State Update
State Readout
~~State Update~~
State Readout
GT
GT
GT
Input
Pointmaps & Cameras

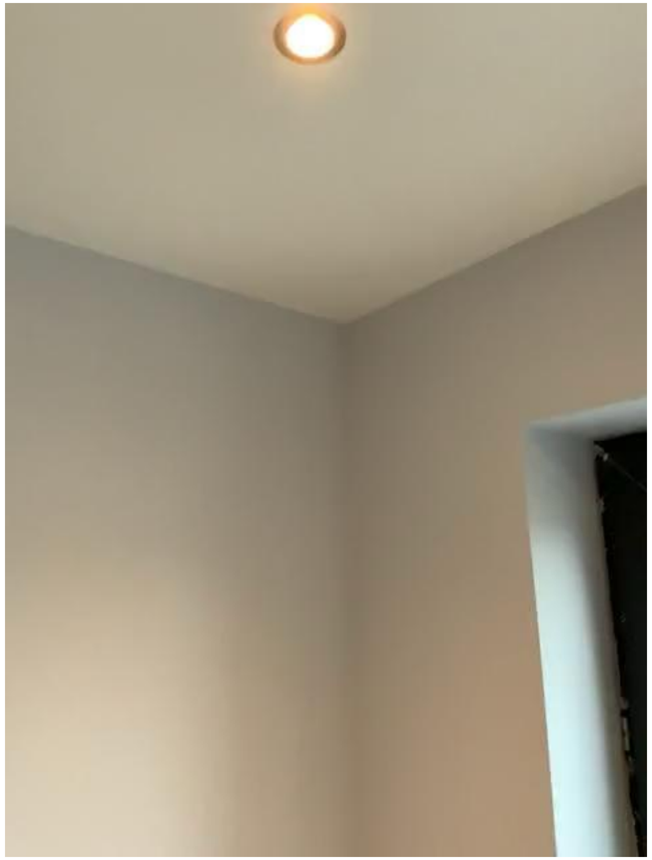
State Update Analysis



Streaming

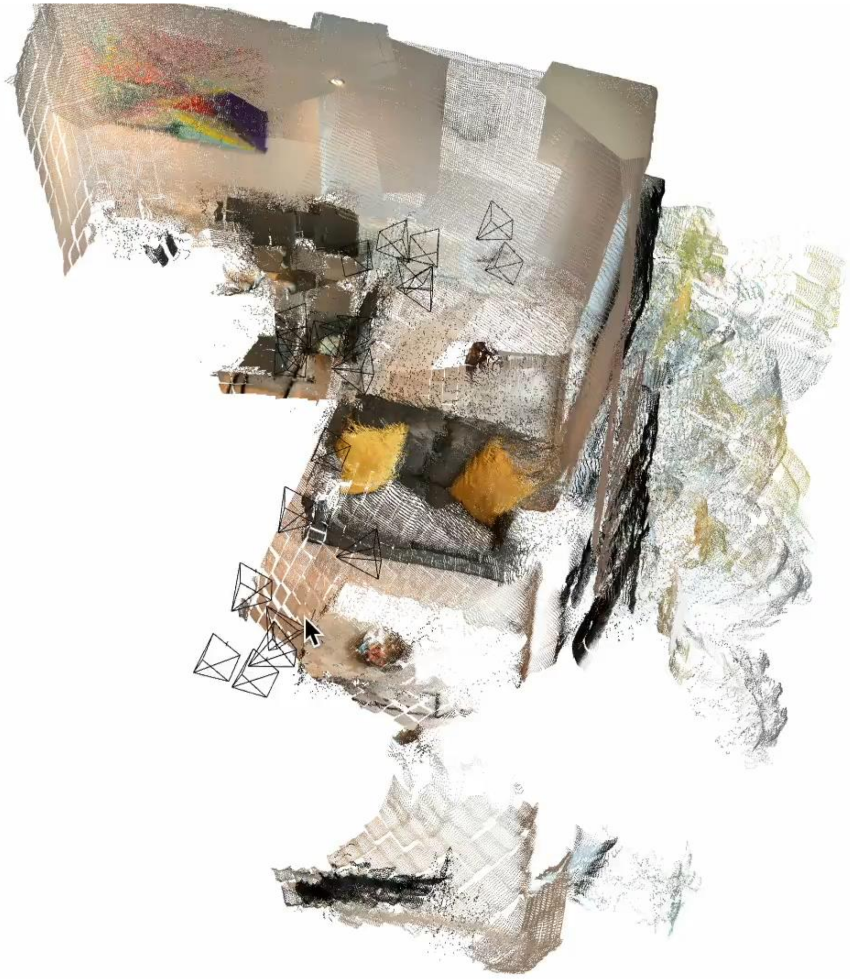
Revisiting

Streaming vs. Revisiting



Streaming

Streaming vs. Revisiting



revisiting

A Visual Illusion Example

Start – 3D chair?

2D painting

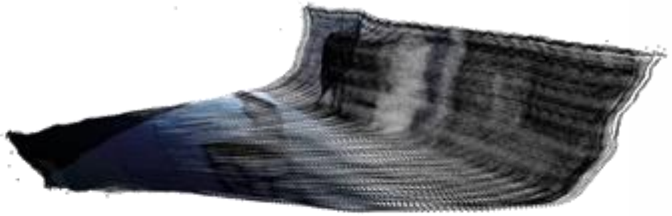
A Visual Illusion Example

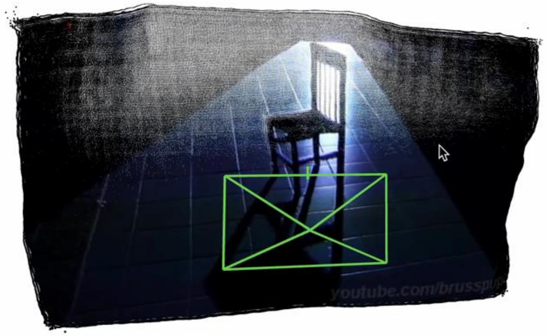
Start – 3D chair?

2D painting

End – 2D painting!

A Visual Illusion Example





Start – 3D chair

2D painting

End – 2D painting

Summary

- From the belief that the world persists emerges the understanding of motion and structure
- Spatial intelligence requires both data-driven priors and the ability to update continuously online

On Multi-Modal Spatial Intelligence

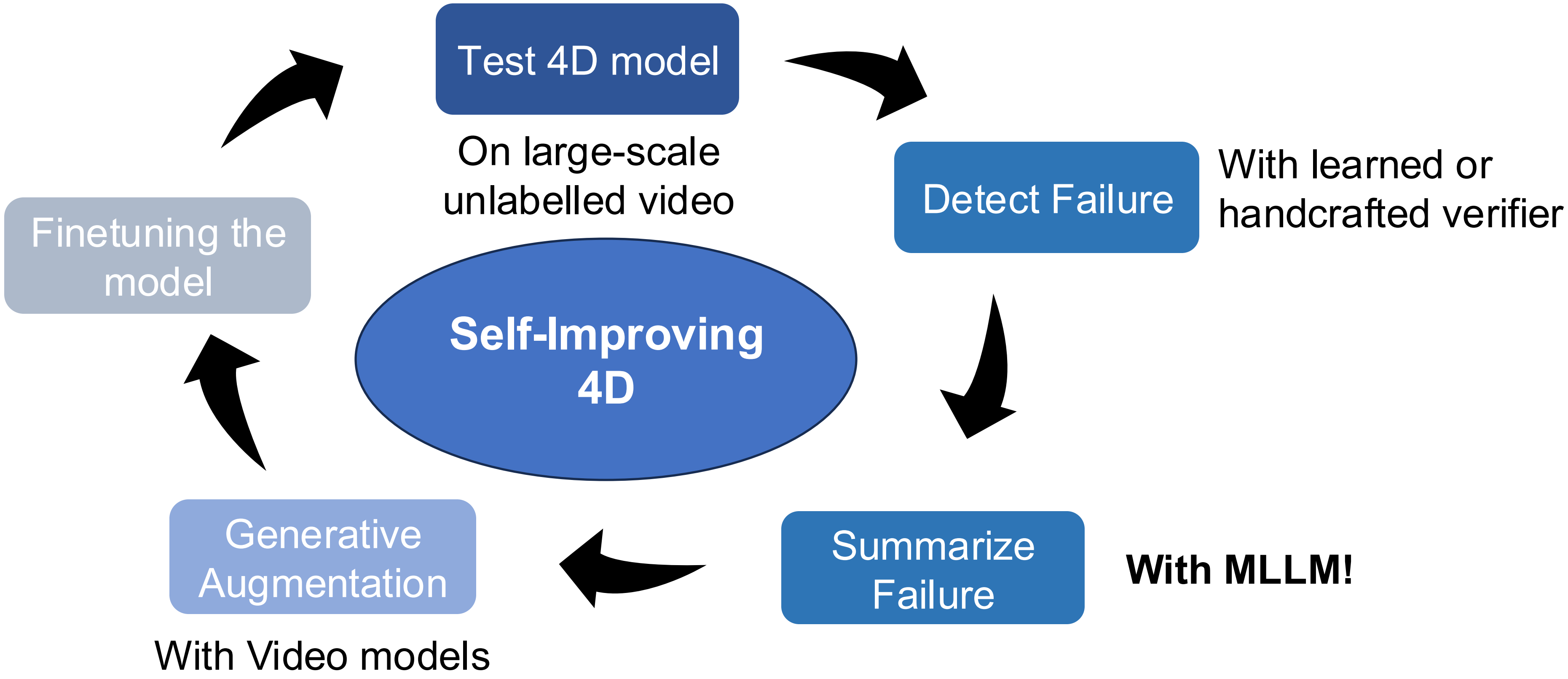
- Spatial intelligence doesn't need MLLMs

Squirrel scatter hoarding

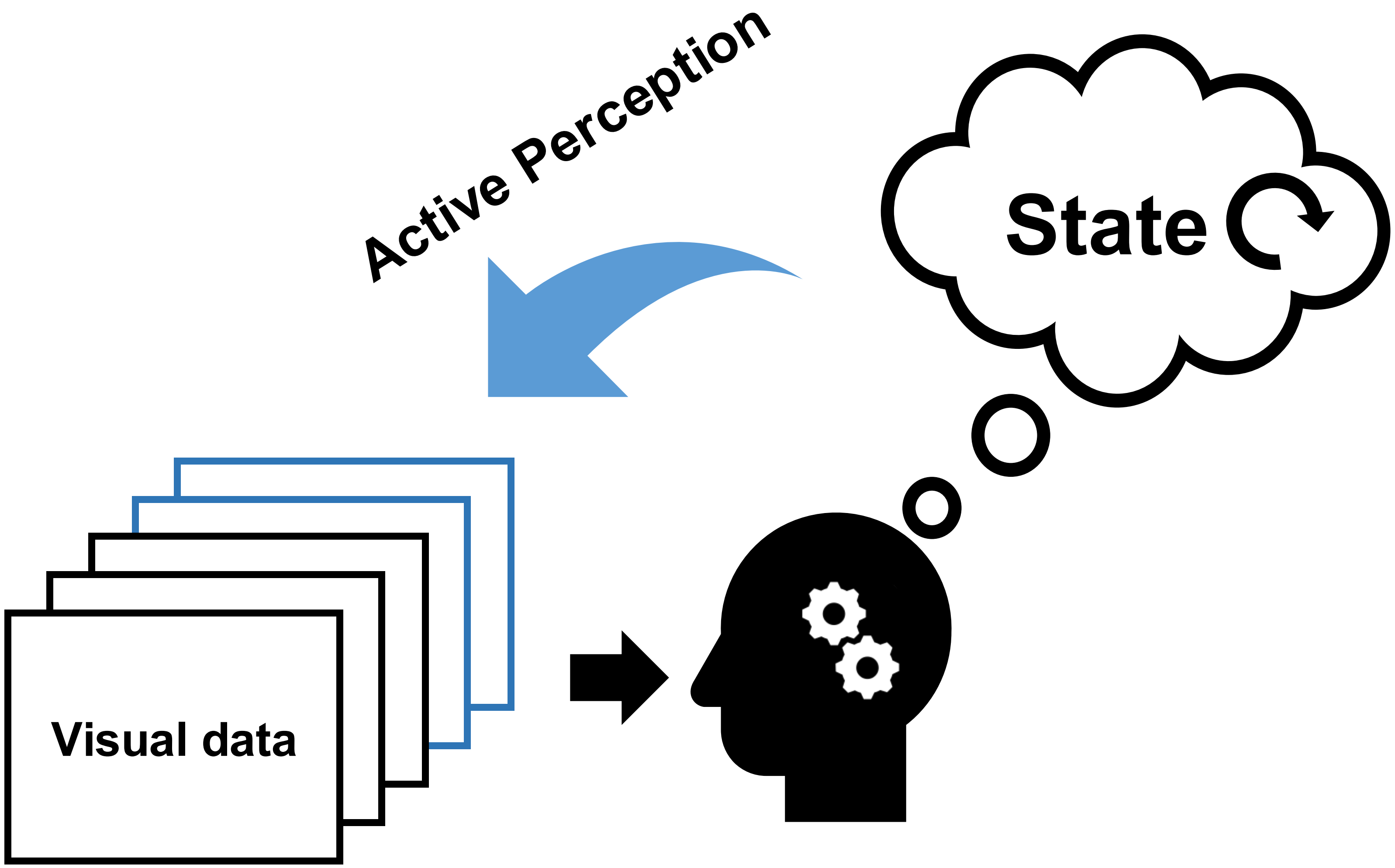
Still "multi-modal":
- Vision
- Audition
- Olfaction
- Touch
- …

On Multi-Modal Spatial Intelligence

- But MLLMs can help us build spatial intelligence
  - concepts and common-sense knowledge from large-scale multimodal data
  - an interface for communication between humans and machines



Spatial Intelligence in Active Settings



Collaborators


Noah Snavely

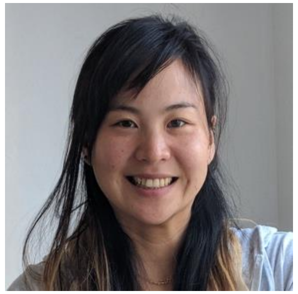

Bharath Hariharan


Zhengqi Li


Aleksander Holynski


Yen-Yu Chang


Ruojin Cai


Angjoo Kanazawa


Alexei A. Efros


Yifei Zhang

and many more…)

Thank you!