

Recent Trends in 3D Reconstruction for General Non-Rigid Scenes

State-Of-The-Art Report

Raza Yunus, Jan Eric Lenssen, Michael Niemeyer, Christian Rupprecht, Yiyi Liao, Christian Theobalt, Gerard Pons-Moll, Jia-Bin Huang, Vladislav Golyanik and Eddy Ilg

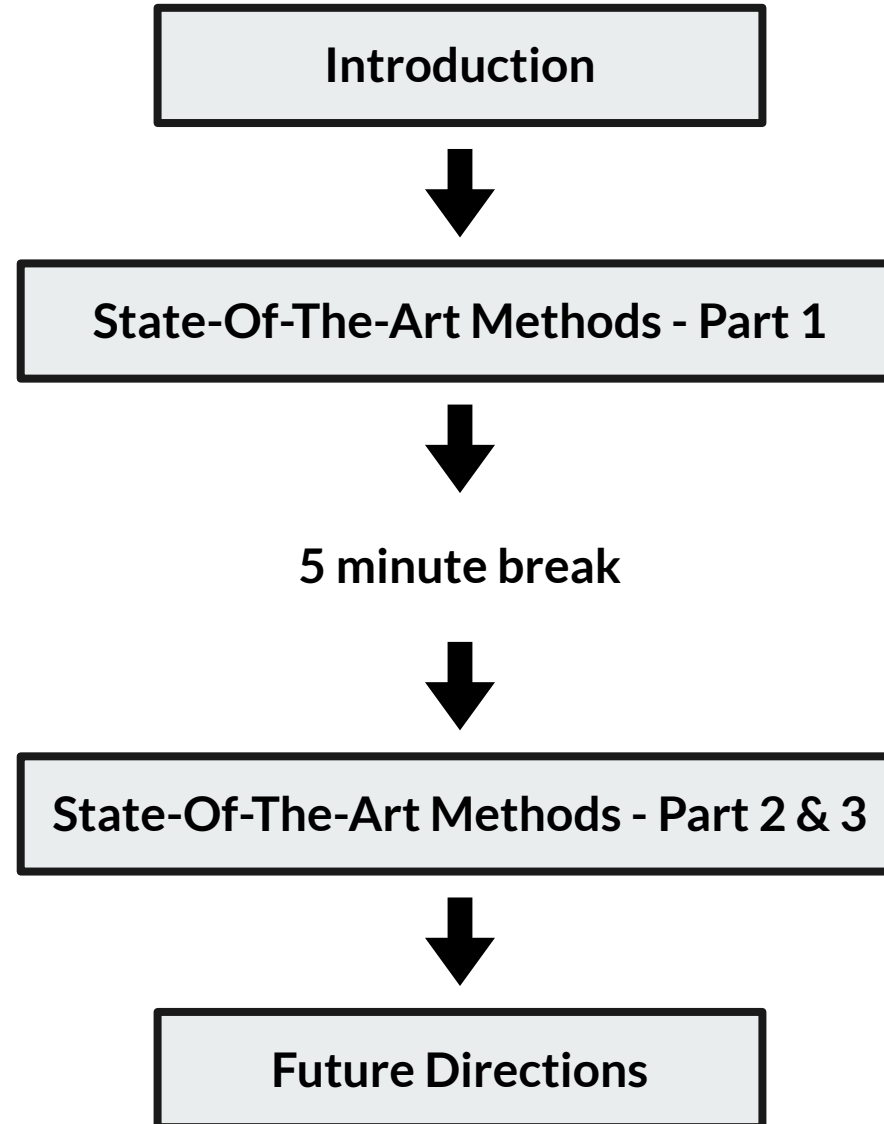
Presented by Raza Yunus

© Eurographics Conference 2024. All rights preserved.



The 45th Annual Conference of the European Association for Computer Graphics is organized by CYENS Centre of Excellence in collaboration with the University of Cyprus and the Cyprus University of Technology.

Talk Schedule



Motivation & Applications

The world is dynamic! Needs to be modelled in various applications.

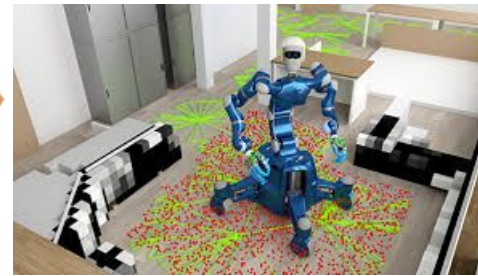
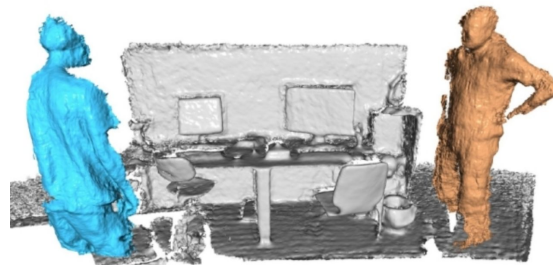
Motivation & Applications



Telepresence / VR

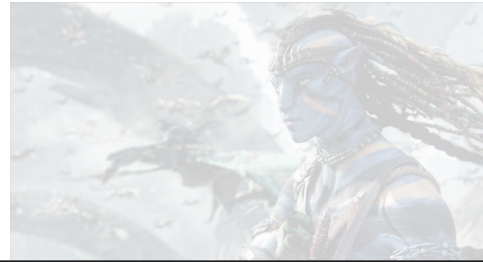


Movie & Gaming Industry



Robotics / AR

Motivation & Applications



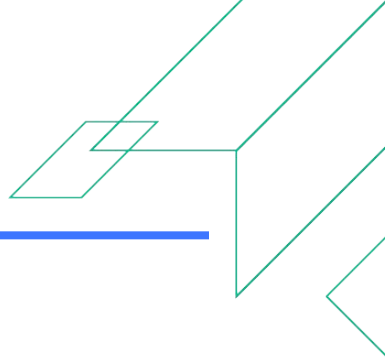
Recent advances are making non-rigid 3D reconstruction methods more and more powerful!



Robotics / AR

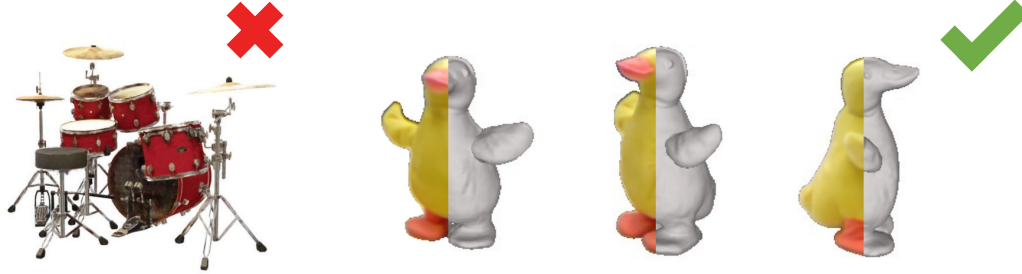
Scope

Recent Trends in 3D Reconstruction of General Non-Rigid Scenes



Scope

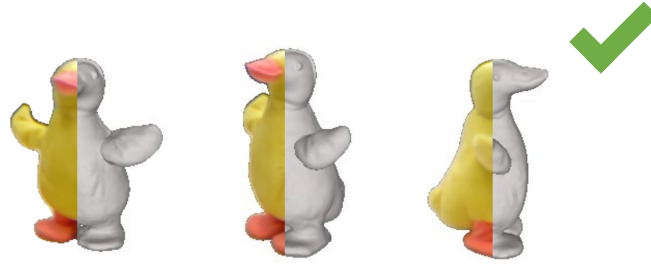
Recent Trends in 3D Reconstruction of General **Non-Rigid** Scenes



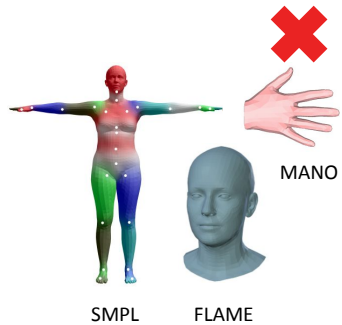
- Focus on methods that consider non-rigid deformations during reconstruction

Scope

Recent Trends in 3D Reconstruction of **General** Non-Rigid Scenes



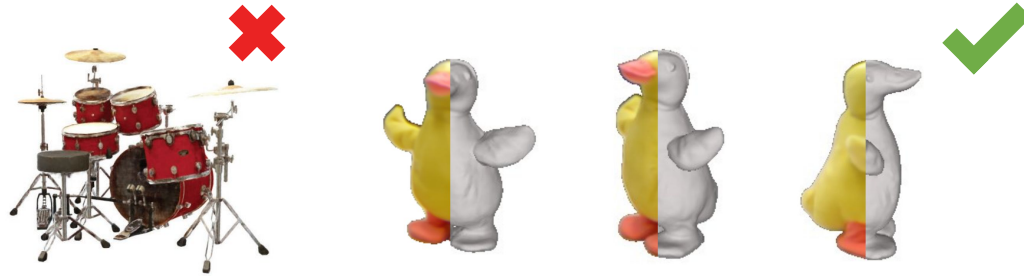
- Focus on methods that consider non-rigid deformations during reconstruction



- No domain-specific methods

Scope

Recent Trends in 3D Reconstruction of General Non-Rigid Scenes



→ Covers methods mostly from the last three years

→ We refer to older Eurographics STARs for a survey of earlier techniques:

- Focus on methods that consider non-rigid deformations during reconstruction



- No domain-specific methods

DOI: 10.1111/1471-1174.14107
EUROGRAPHICS 2022
D. Basci and G. Frazee
(Guest Editors)

Volume 43 (2022), Number 2
STAR - State of the Art Report

Advances in Neural Rendering

A. Tovar^{1*}, J. Thies^{2*}, B. Mikkelson¹, P. Srinivasan¹, E. Tretschk³, W. Yifan⁴, C. Laine⁵, V. Stamm⁶, R. Martin-Bruhl⁷, S. Lischke⁸, M. Stamm⁹, C. Theobald¹⁰, M. Nießner¹¹, G. Wetzstein¹², M. Zollhöfer¹³, V. Goltzani¹⁴

¹UPMC Informatics, ²Light for Synthetic Systems, ³Google Research, ⁴UT Dallas, ⁵Max Planck Institute Saarland Informatics Center, ⁶University of Tübingen, ⁷Technical University of Munich, ⁸Stanford University, ⁹Intel Labs Research, ¹⁰MIT, ¹¹Technical University of Munich, ¹²Stanford University, ¹³Technical University of Munich, ¹⁴University of Tübingen

Abstract
Synthesizing photo-realistic images and videos is at the heart of computer graphics and has been the focus of decades of research. Traditionally, synthetic images of a scene are generated using rendering algorithms such as rasterization or ray tracing, which take explicitly defined representations of geometry and material properties as input. Collectively, these input data define the actual scene and what is rendered, and are referred to as the scene representation (where a scene consists of one or more objects). Example scene representations are: triangle meshes with accompanying texture files, created by an artist; point clouds (e.g., from a depth sensor), semantic grids (e.g., from a CT scan), or implicit surfaces (functions (e.g., learned neural networks) that map a point in space to a color and a depth value). The reconstruction of such a scene representation from observation data (e.g., monocular images) and/or other learning to create algorithms for synthesizing images from real-world observations. Neural rendering is a step forward towards the goal of synthesizing photo-realistic images and videos content. In recent years, we have seen immense progress in this field through hundreds of publications that show different ways to digitize the real world content into the rendering pipeline. This state-of-the-art report on advances in neural rendering focuses on methods that combine classical rendering principles with learned 3D scene representations, often used to represent scene geometry. A key advantage of these methods is that they can be used to design, enabling applications such as novel view synthesis or virtual reality. In addition to methods that handle static scenes, we review neural scene representations for modeling non-rigidly deforming objects and scene editing and composition. While any of these approaches are scene-specific, we also discuss techniques that generalize across object classes and can be used for generative tasks. In addition to reviewing these state-of-the-art methods, we provide an overview of fundamental concepts and definitions used in this current literature. We conclude with a discussion on open challenges and social implications.

1. Introduction
Synthesis of photorealistic and photo-realistic images and videos is one of the fundamental goals of computer graphics. During the last decades, methods and representations have been developed to represent the image formation model of real cameras, including the handling of complex materials and global illumination. These methods are based on the laws of physics and simulate the light transport from light sources to the virtual camera for synthesis. In this way, all physical parameters of the scene have to be known for the rendering process. These parameters, the exact, explicit information about the scene geometry and material properties such as reflectivity or opacity. Given this information, rendering ray tracing

DOI: 10.1111/1471-1174.14108
EUROGRAPHICS 2022
A. Basci and C. Theobald
(Guest Editors)

COMPUTER GRAPHICS Forum
Volume 43 (2022), Number 2
STAR - State of the Art Report

State of the Art in Dense Monocular Non-Rigid 3D Reconstruction

Erdi Tretschk^{1*}, Norami Kainuma^{2*}, Mallikarjun B³, Rishabh Dabral⁴, Adam Korytkowski⁵, Bernhard Egger⁶, Marc Habermann⁷, Pascal Fua⁸, Christian Theobald⁹, Vladislav Goltzani¹⁰

¹Max Planck Institute for Informatics, ²Sociedad Informatica Centre, ³University of Tübingen, ⁴Technical University of Munich, ⁵University of Tübingen, ⁶University of Tübingen, ⁷University of Tübingen, ⁸University of Tübingen, ⁹University of Tübingen, ¹⁰University of Tübingen

Abstract
3D reconstruction of deformable (or non-rigid) scenes from a set of monocular 2D image observations is a long-standing and actively researched area of computer vision and graphics. It is an ill-posed inverse problem, since—without additional prior assumptions—it permits infinitely many solutions leading to various projections to the input 2D images. Non-rigid reconstruction is a fundamental building block for downstream applications like robotics, AR/VR, or virtual camera control. The key advantage of using monocular cameras is their compactness and availability in the real world as well as their ease of use compared to more sophisticated camera setups such as stereo or multi-view cameras. This survey focuses on state-of-the-art methods for dense non-rigid 3D reconstruction of various deformable objects and complete scenes from monocular video or sets of monocular images. It reviews the fundamentals of 3D reconstruction and deformation modeling for 2D image observations. We then start from general methods that handle arbitrary scenes and make only a few prior assumptions and proceed towards methods making stronger assumptions about the observed objects and types of deformations (e.g., known forces, handles, hands, and animals). A significant part of this STAR is also devoted to classification and a high-level comparison of the methods, as well as an overview of the datasets for training and evaluation of the discussed techniques. We conclude by discussing open challenges in the field and the social aspects associated with the usage of the reviewed methods.

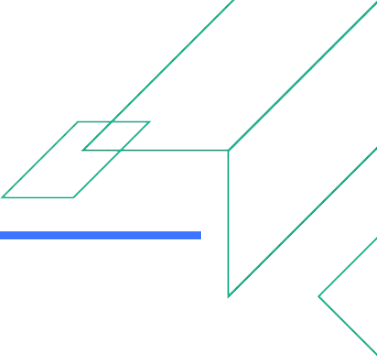
1. Introduction
Humans can close one eye, look around, and get a far sense of their surroundings in terms of the 3D geometry, appearance, and even deformation (Eckhardt, 1971). Nevertheless, designing computational methods that densely reconstruct a dynamic scene in 3D using a single monocular camera remains a challenging task that is far from solved, as this STAR shows.



Tewari et al. (2022)

Tretschk et al. (2023)

State-of-the-Art Report



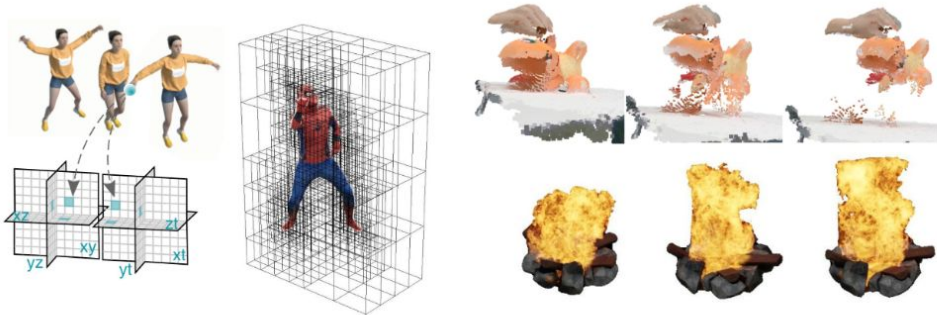
State-of-the-Art Report

Over 150 methods divided into four categories

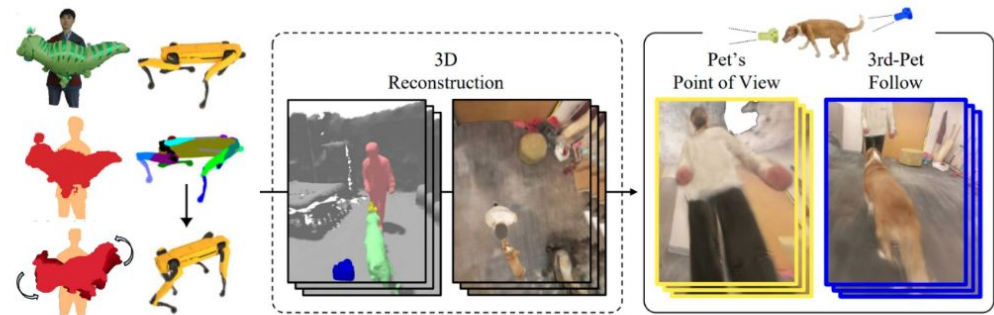


State-of-the-Art Report

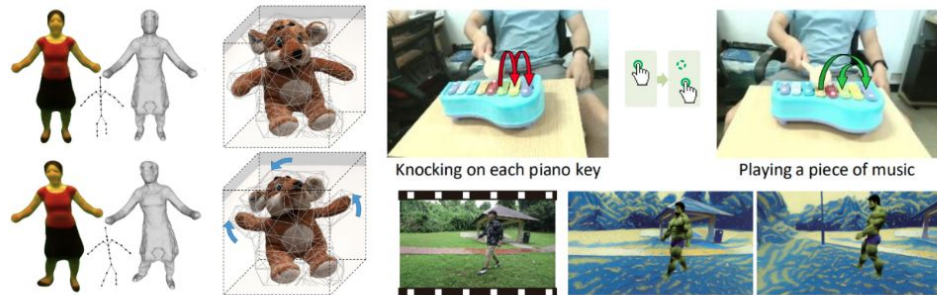
Over 150 methods divided into four categories



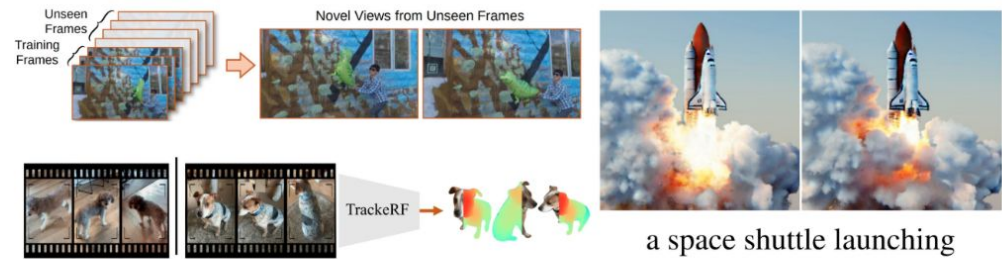
3D Non-Rigid Reconstruction and View Synthesis



Decompositional Scene Analysis



Editability and Control



Generalizable and Generative Modeling

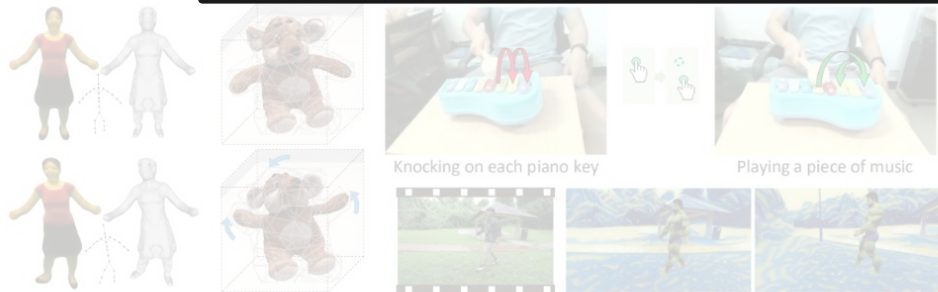
State-of-the-Art Report

Over 150 methods divided into four categories



3D

For this talk, we will look at three main trends from these four categories

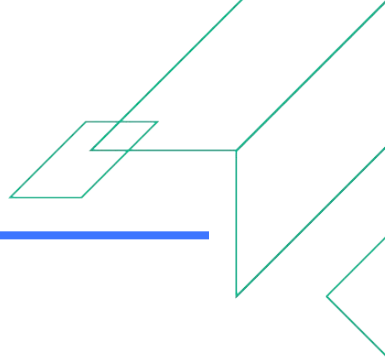


Editability and Control



Generalizable and Generative Modeling

Trends



Trends

1. Speed and Quality Advancements



Trends

1. Speed and Quality Advancements
2. Handling of Large Deformations / Long-Term 3D Correspondences



Trends

1. Speed and Quality Advancements
2. Handling of Large Deformations / Long-Term 3D Correspondences
3. Modelling Articulated Motion for General Objects

Trends

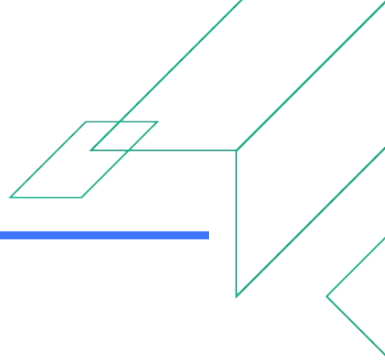
- Speed and Quality Advancements

- **First, let's have a brief look at the different aspects of non-rigid 3D reconstruction**

- Modelling Articulated Motion for General Objects

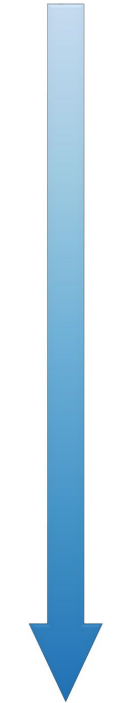
Task

Non-Rigid 3D Reconstruction and View Synthesis

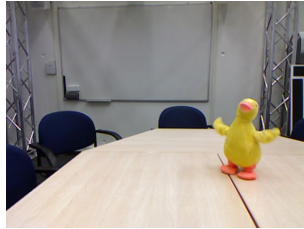


Task

Non-Rigid 3D Reconstruction and View Synthesis

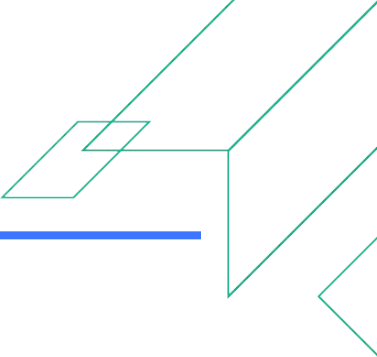


Time

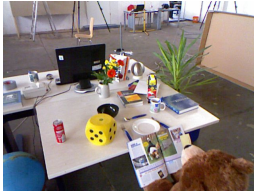


Observations

Sensors and Capture Settings



Sensors and Capture Settings

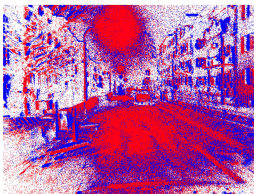


RGB



Depth

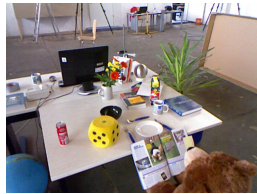
- Passive Depth
- Structured Depth
- Time-of-Flight



Event

Sensor Types

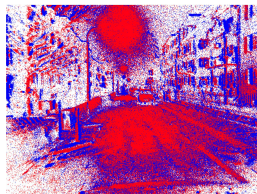
Sensors and Capture Settings



RGB

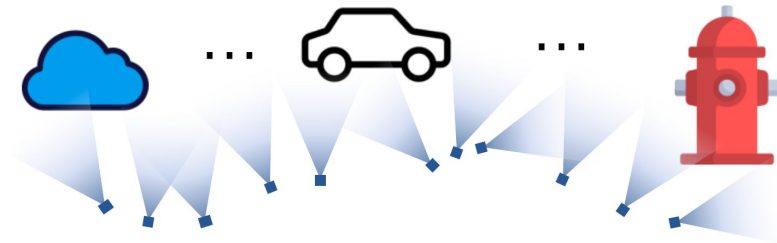


Depth



Event

Sensor Types



Monocular

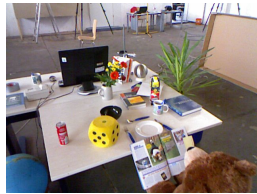


Multi-view

Capture Settings

- Passive Depth
- Structured Depth
- Time-of-Flight

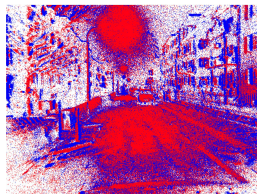
Sensors and Capture Settings



RGB



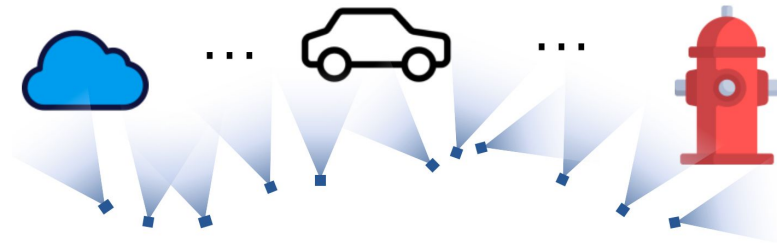
Depth



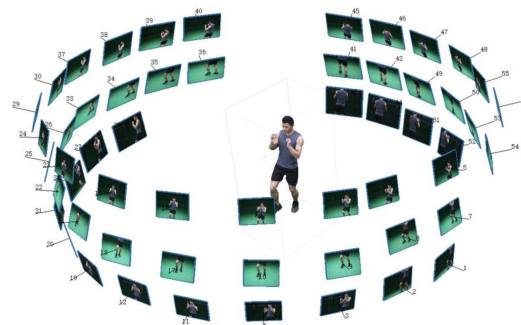
Event

Sensor Types

- Passive Depth
- Structured Depth
- Time-of-Flight

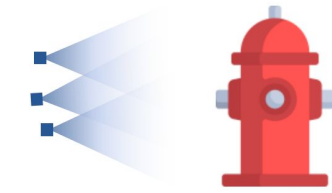


Monocular

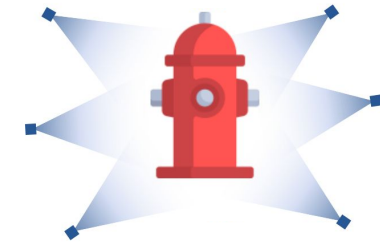


Multi-view

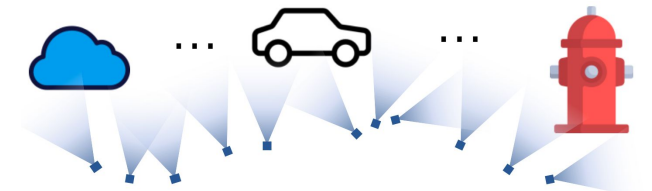
Capture Settings



Forward Facing



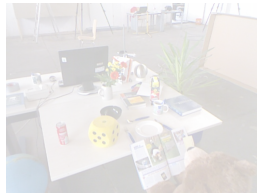
360 Degree



Freeform

Capture Trajectories

Sensors and Capture Settings

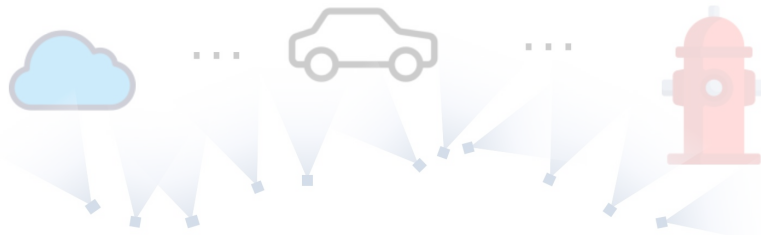


RGB



Event

Sensor Types



Forward Facing



Multi-view

Capture Settings



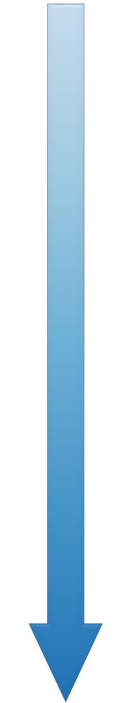
Freeform

Capture Trajectories

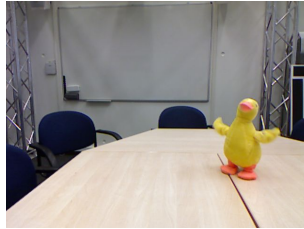
- Determine the difficulty of reconstruction and influence quality
- More prior information is required when observations are sparse

Task

Non-Rigid 3D Reconstruction and View Synthesis



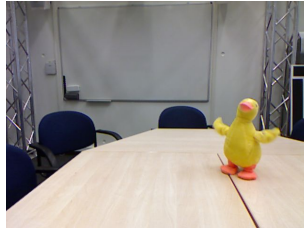
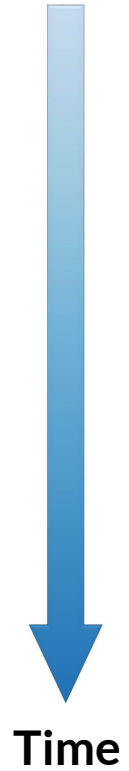
Time



Observations

Task

Non-Rigid 3D Reconstruction and View Synthesis



Observations



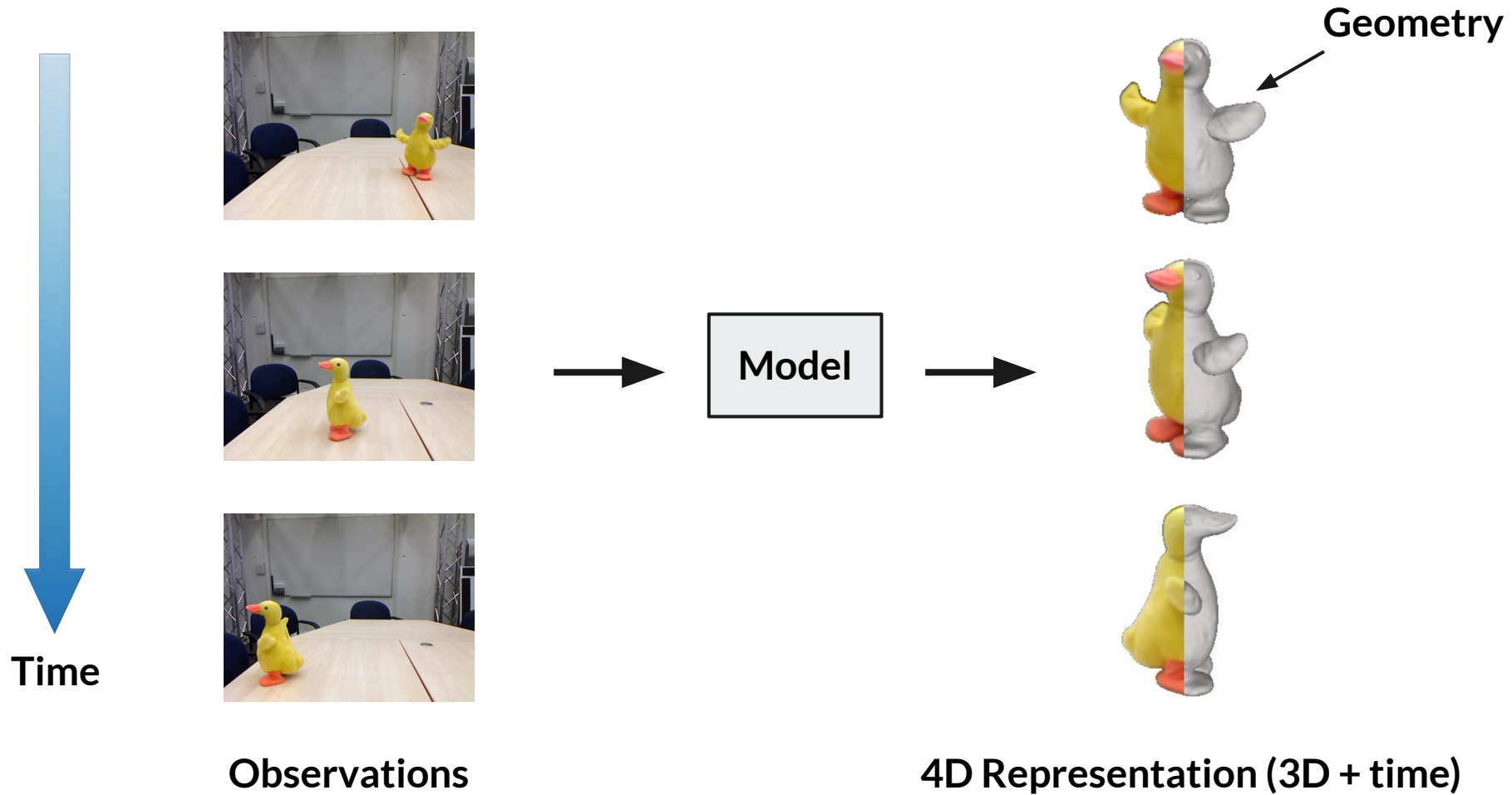
Model



4D Representation (3D + time)

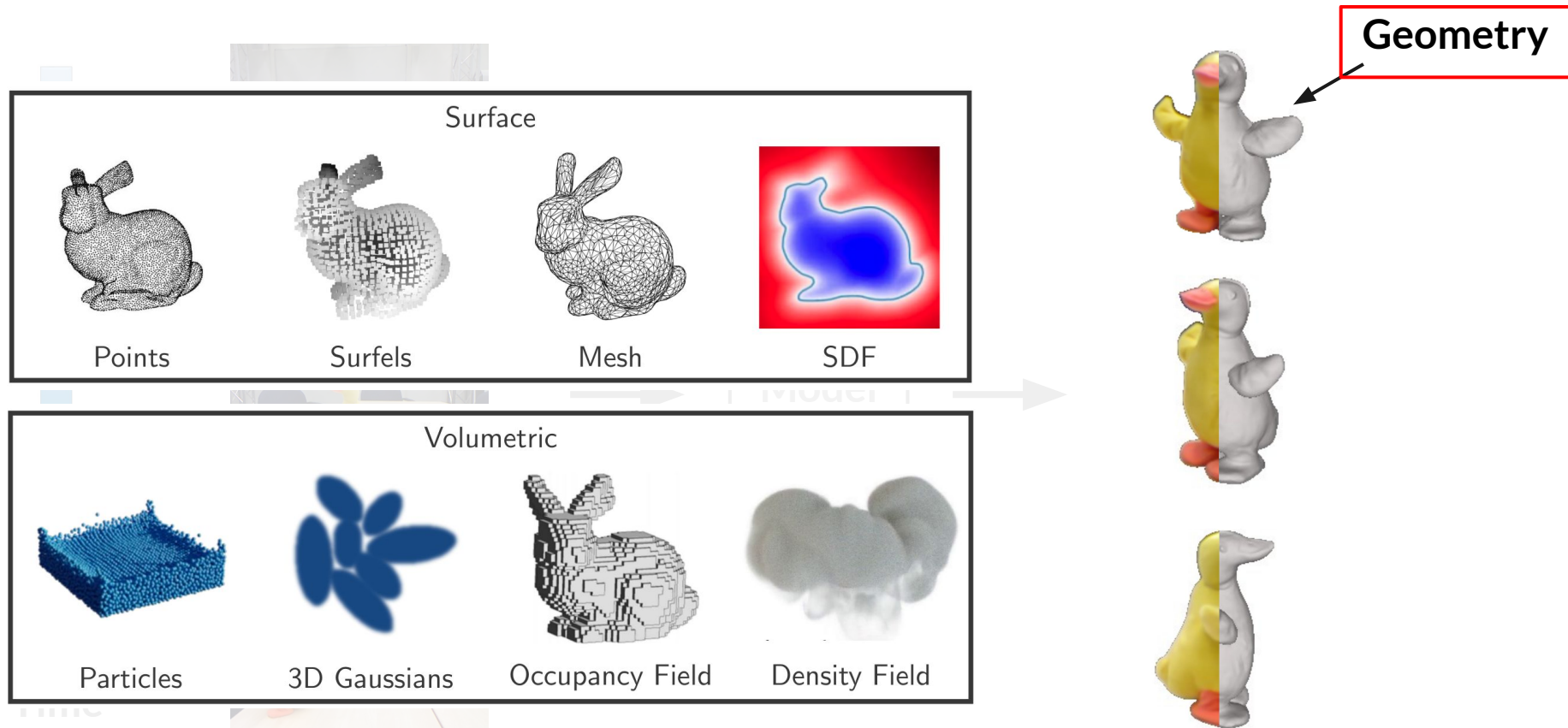
Task

Non-Rigid 3D Reconstruction and View Synthesis



Task

Non-Rigid 3D Reconstruction and View Synthesis

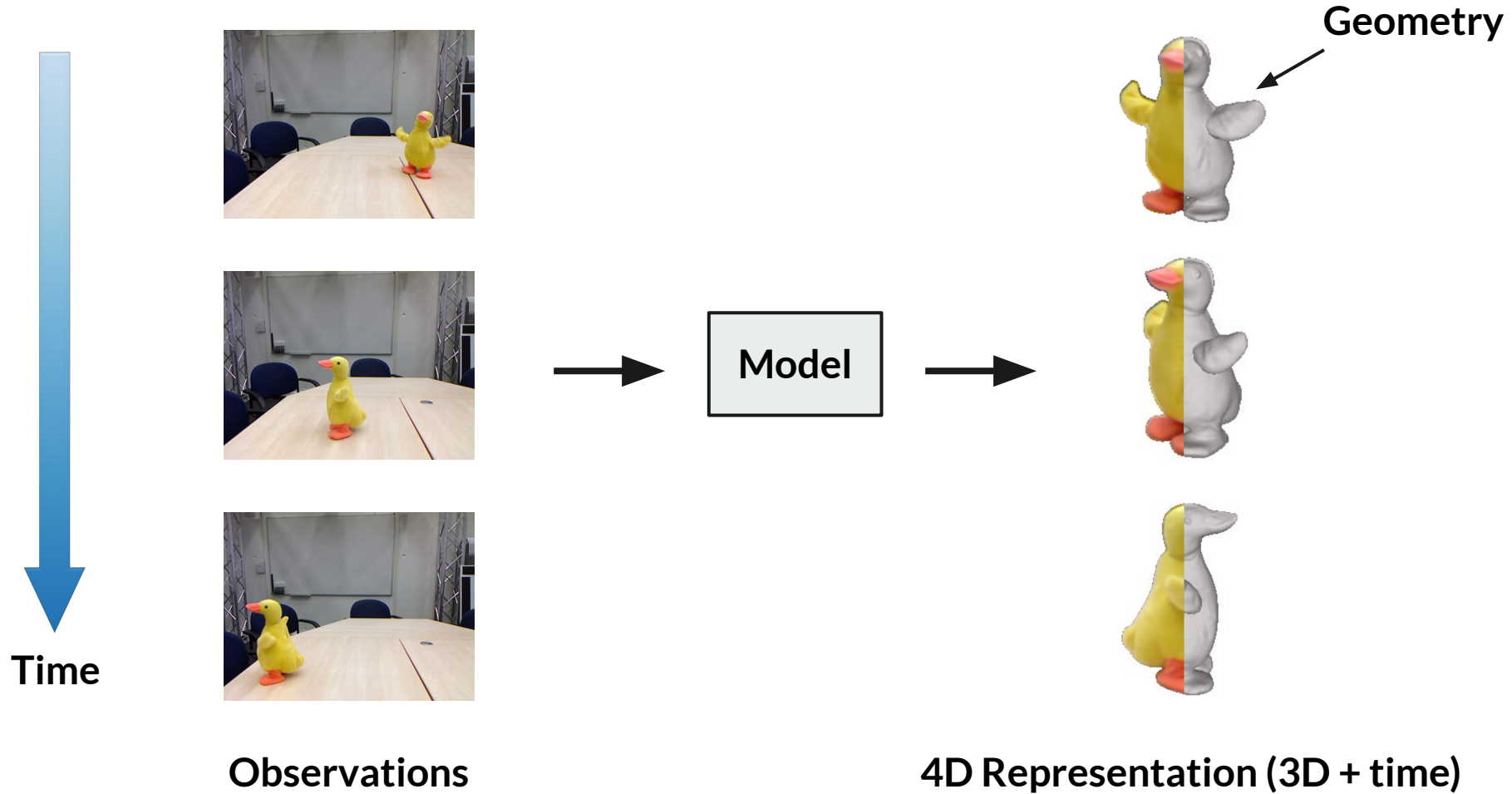


Observations

4D Representation (3D + time)

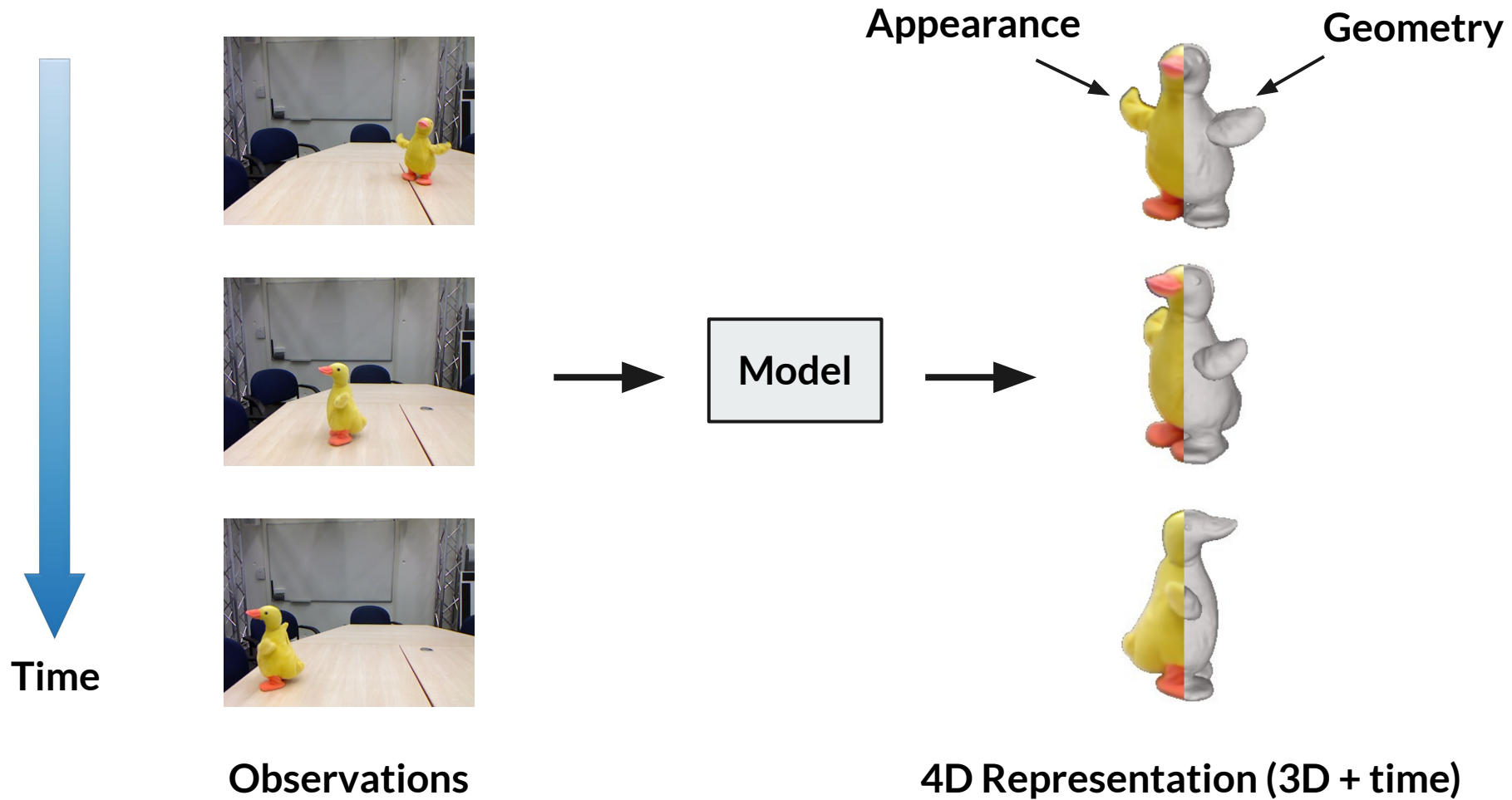
Task

Non-Rigid 3D Reconstruction and View Synthesis




Task

Non-Rigid 3D Reconstruction and View Synthesis



Task

Non-Rigid 3D Reconstruction and View Synthesis


$$= \int_{\mathcal{S}} \left(\begin{array}{c} \text{(b) Light Visibility} \\ \times \\ \text{(c) Direct Illumination} \\ + \\ \text{(d) Indirect Illumination} \end{array} \right) \times \begin{array}{c} \text{(e) BRDF} \end{array} d\omega_i$$

Interaction between environment lighting and surface materials!

Appearance


Geometry



4D Representation (3D + time)

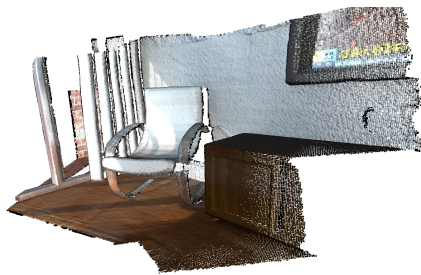
Task

Non-Rigid 3D Reconstruction and View Synthesis


$$= \int_{\mathcal{S}} \left(\begin{array}{c} \text{(b) Light Visibility} \\ \times \\ \text{(c) Direct Illumination} \\ + \\ \text{(d) Indirect Illumination} \end{array} \right) \times \begin{array}{c} \text{(e) BRDF} \\ d\omega_i \end{array}$$

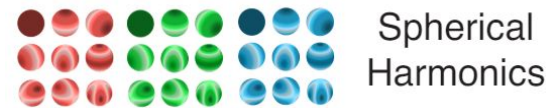
Interaction between environment lighting and surface materials!

Simplifications:



Lambertian Surface Model

$$(x, y, z, \theta, \phi) \rightarrow \begin{array}{c} \text{|||} \\ \text{|||} \\ \text{|||} \\ \hline F_{\Theta} \end{array} \rightarrow (RGB\sigma)$$

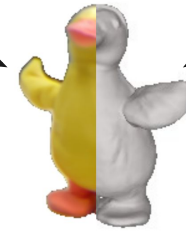


Spherical Harmonics

View-Dependent Outgoing Irradiance

Appearance

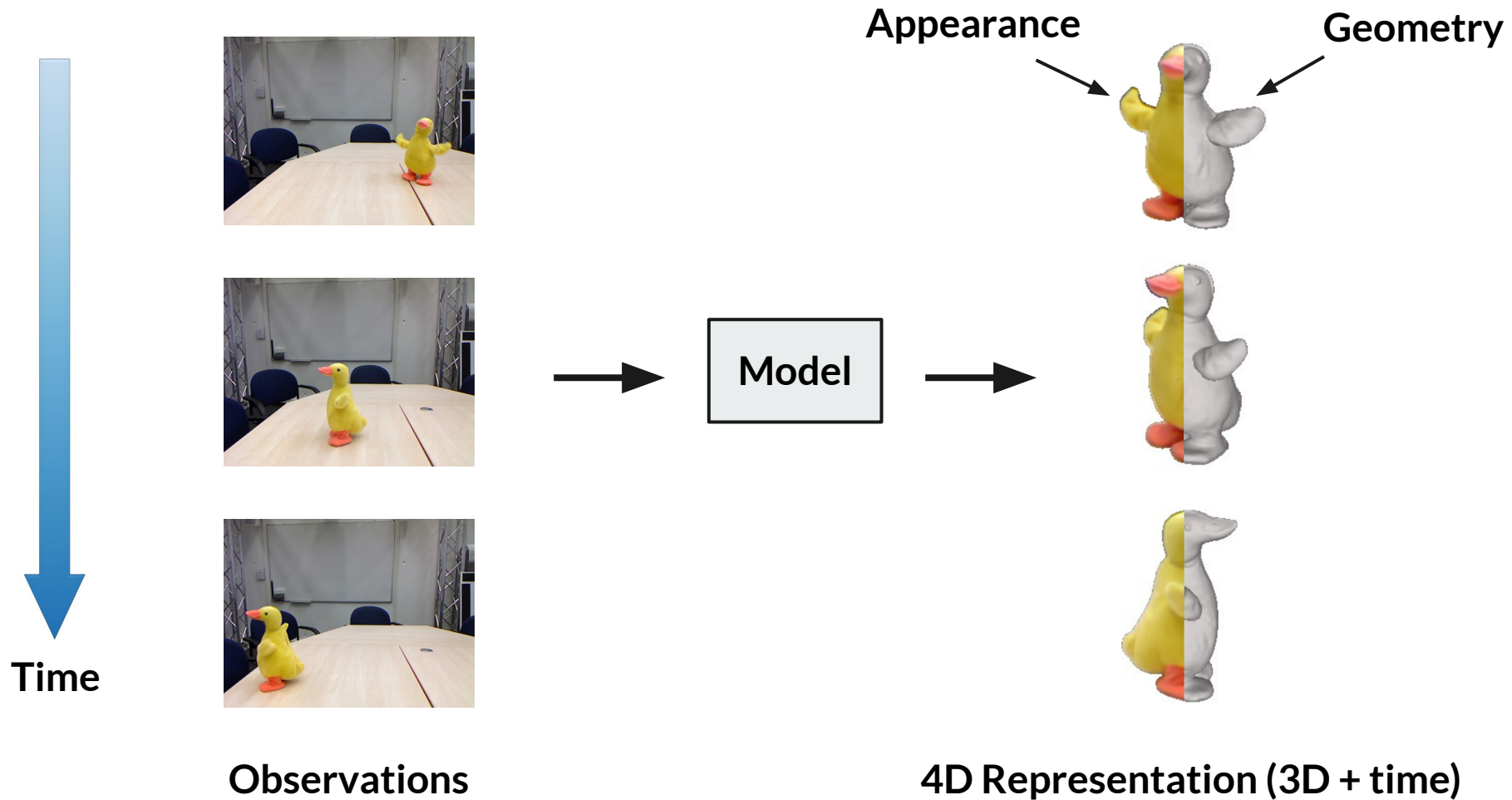
Geometry



4D Representation (3D + time)

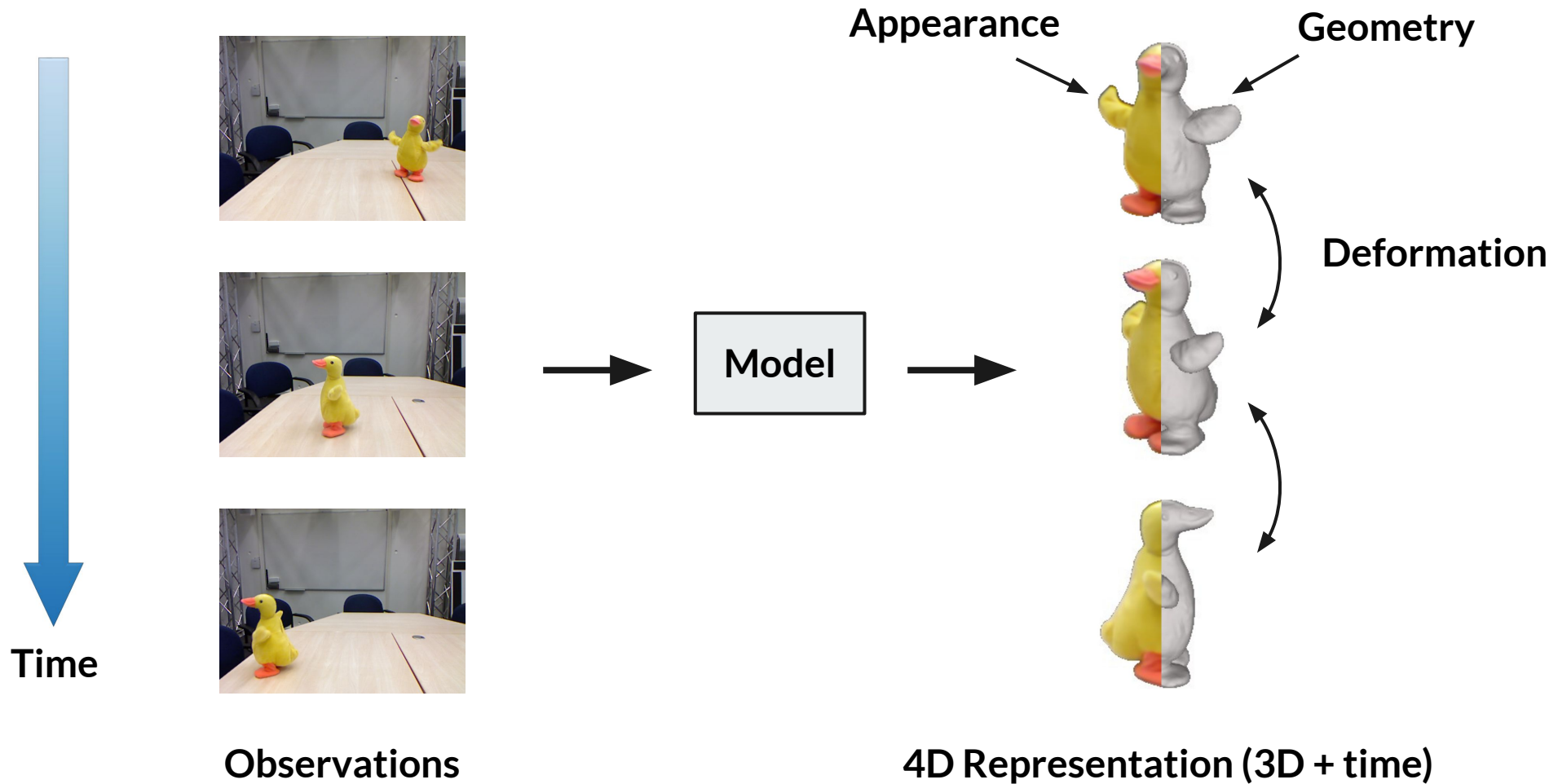
Task

Non-Rigid 3D Reconstruction and View Synthesis



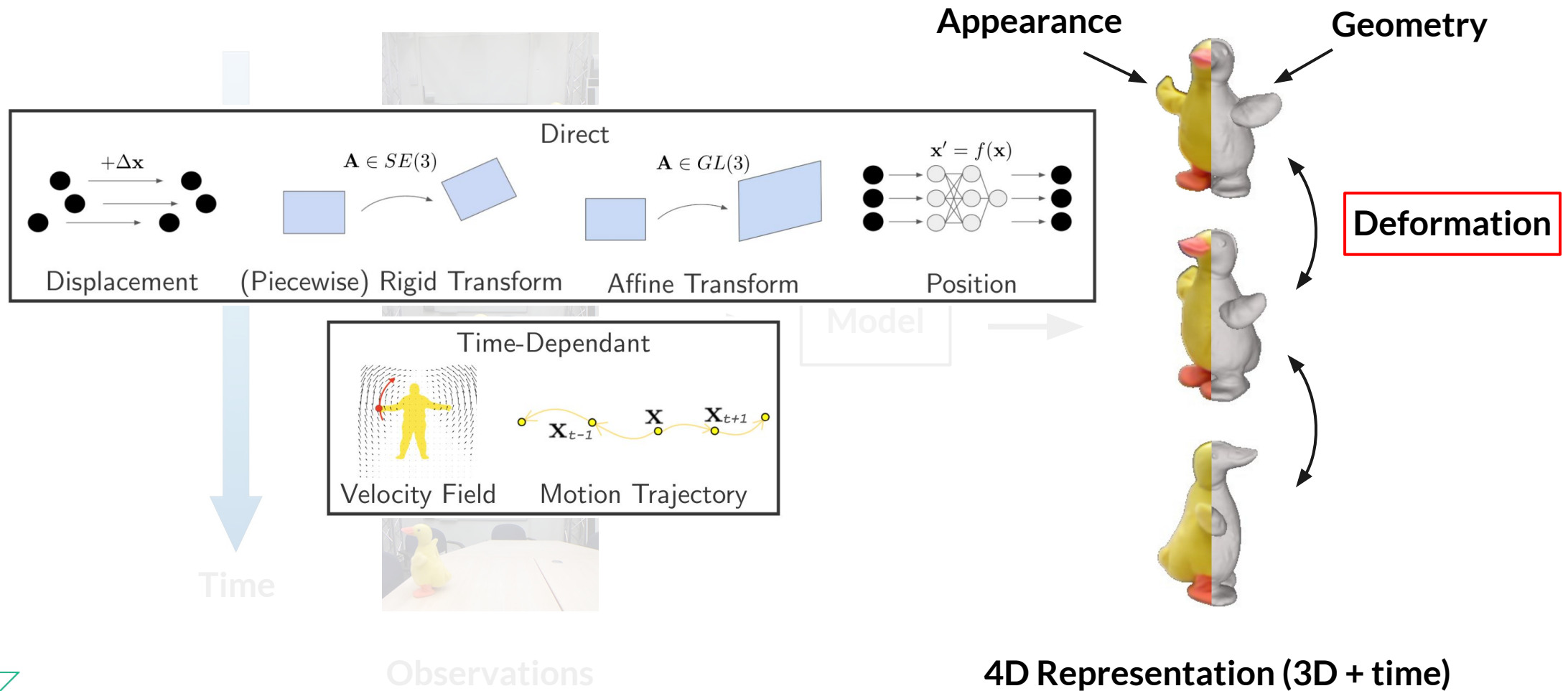
Task

Non-Rigid 3D Reconstruction and View Synthesis



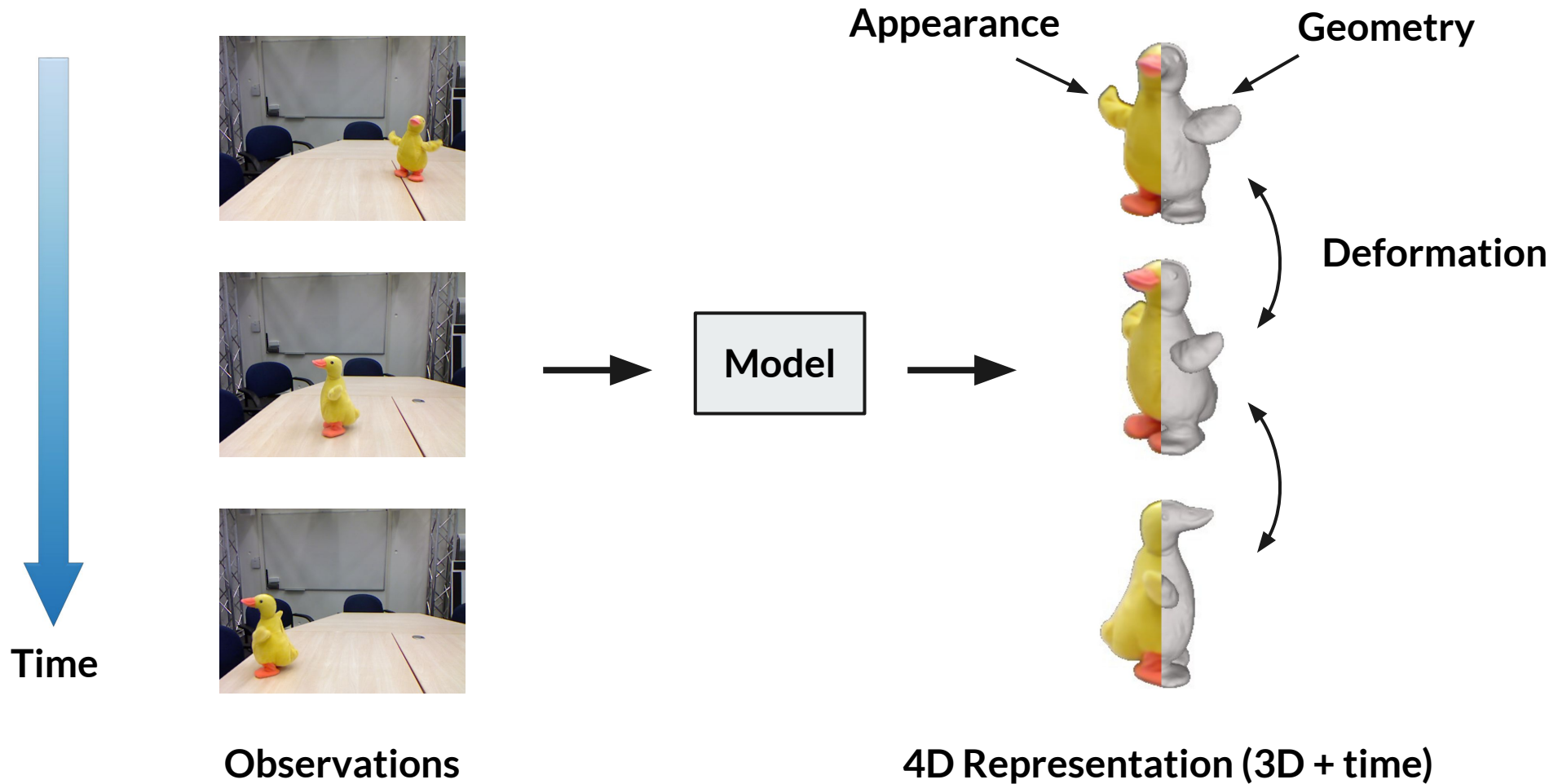
Task

Non-Rigid 3D Reconstruction and View Synthesis



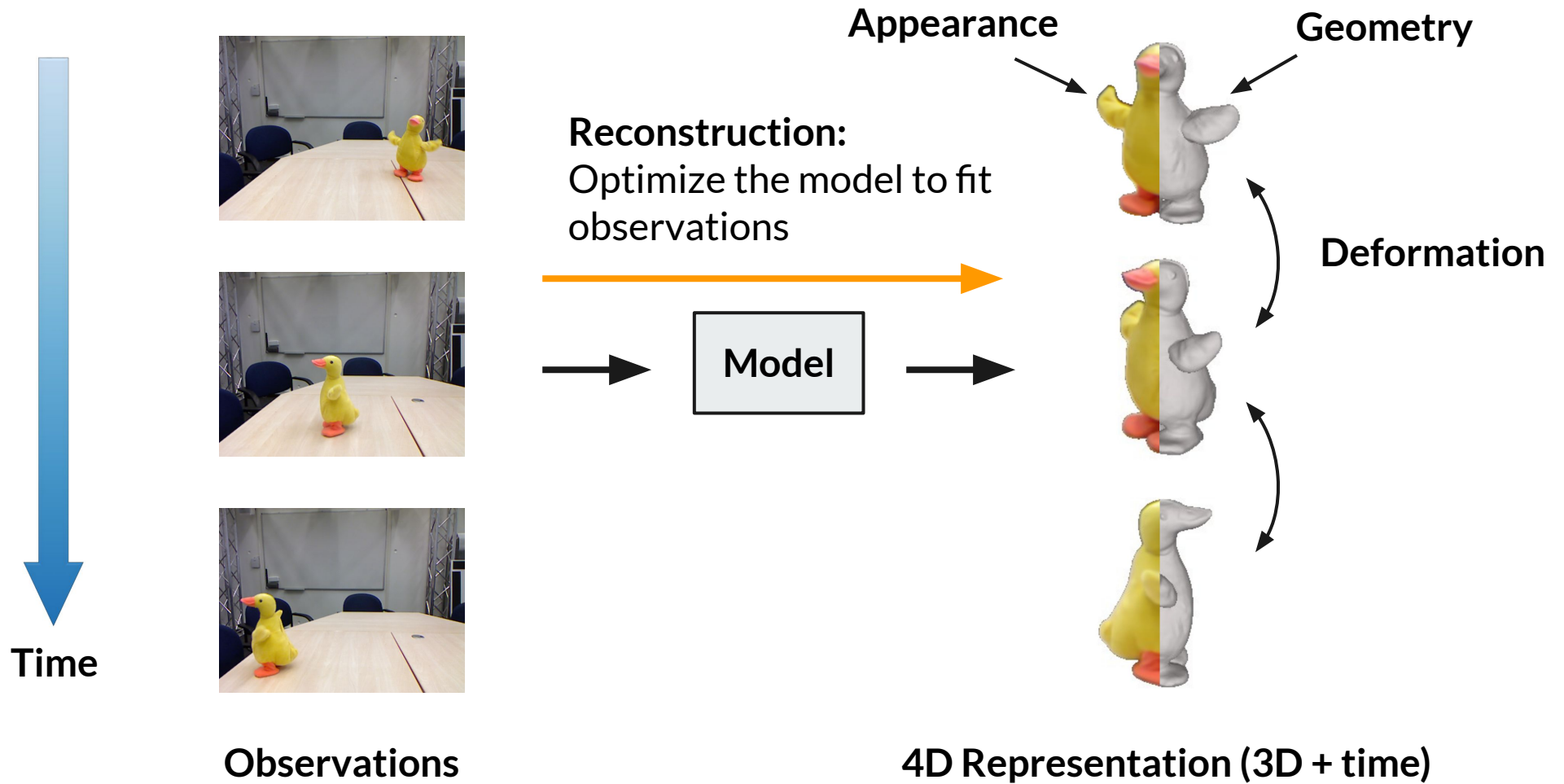
Task

Non-Rigid 3D Reconstruction and View Synthesis



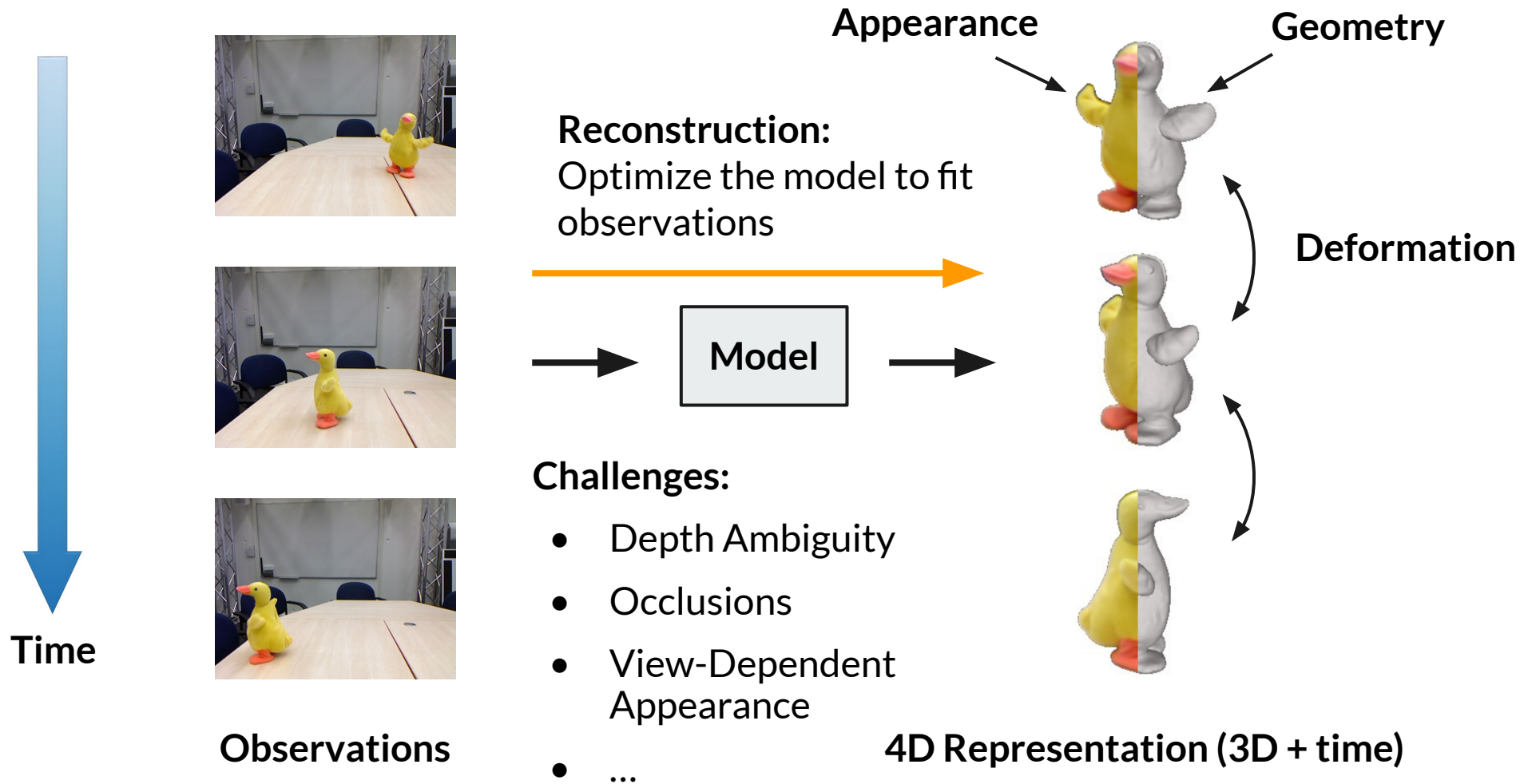
Task

Non-Rigid 3D Reconstruction and View Synthesis



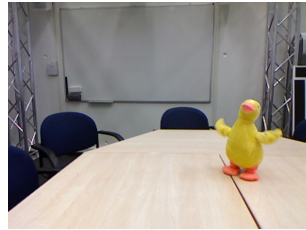
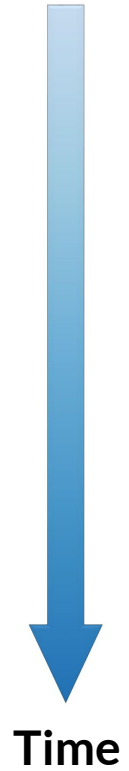
Task

Non-Rigid 3D Reconstruction and View Synthesis

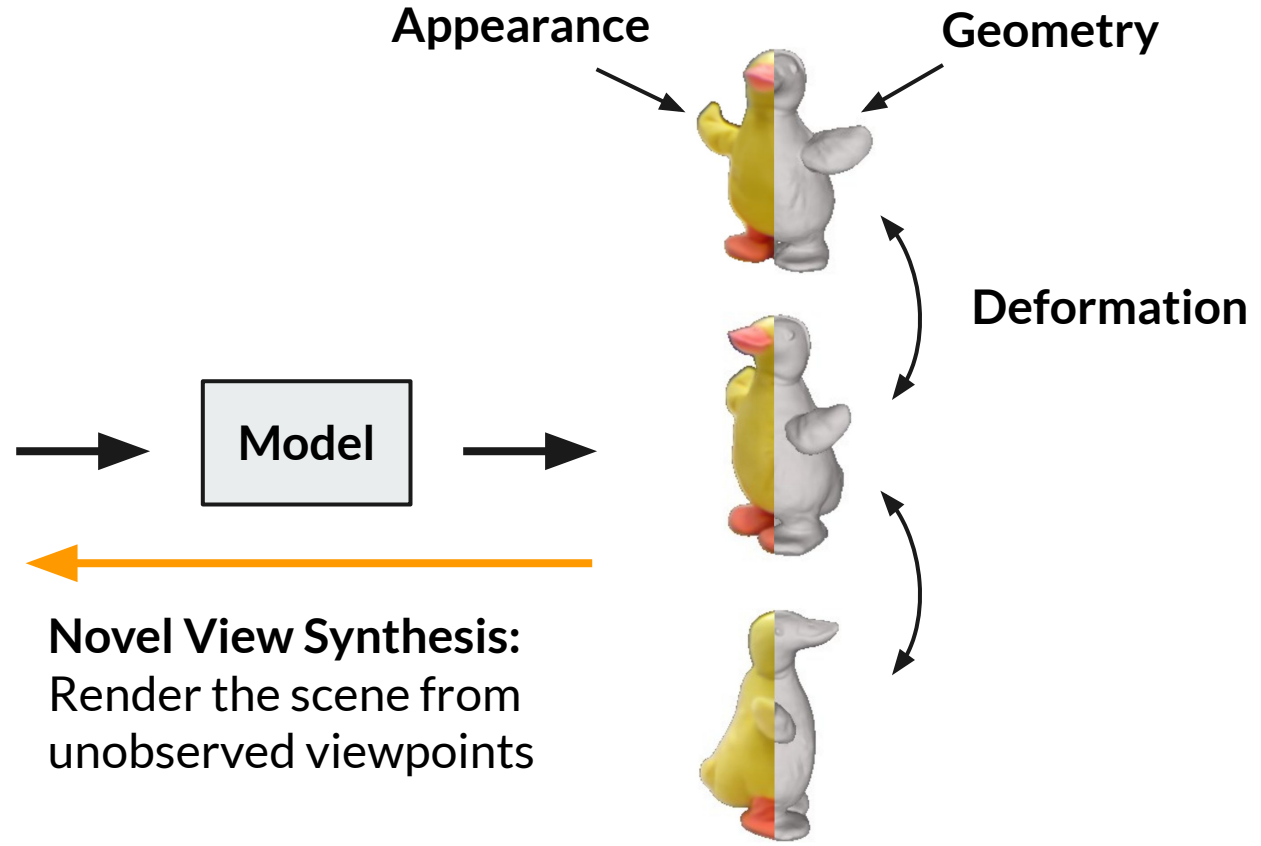


Task

Non-Rigid 3D Reconstruction and View Synthesis



Observations

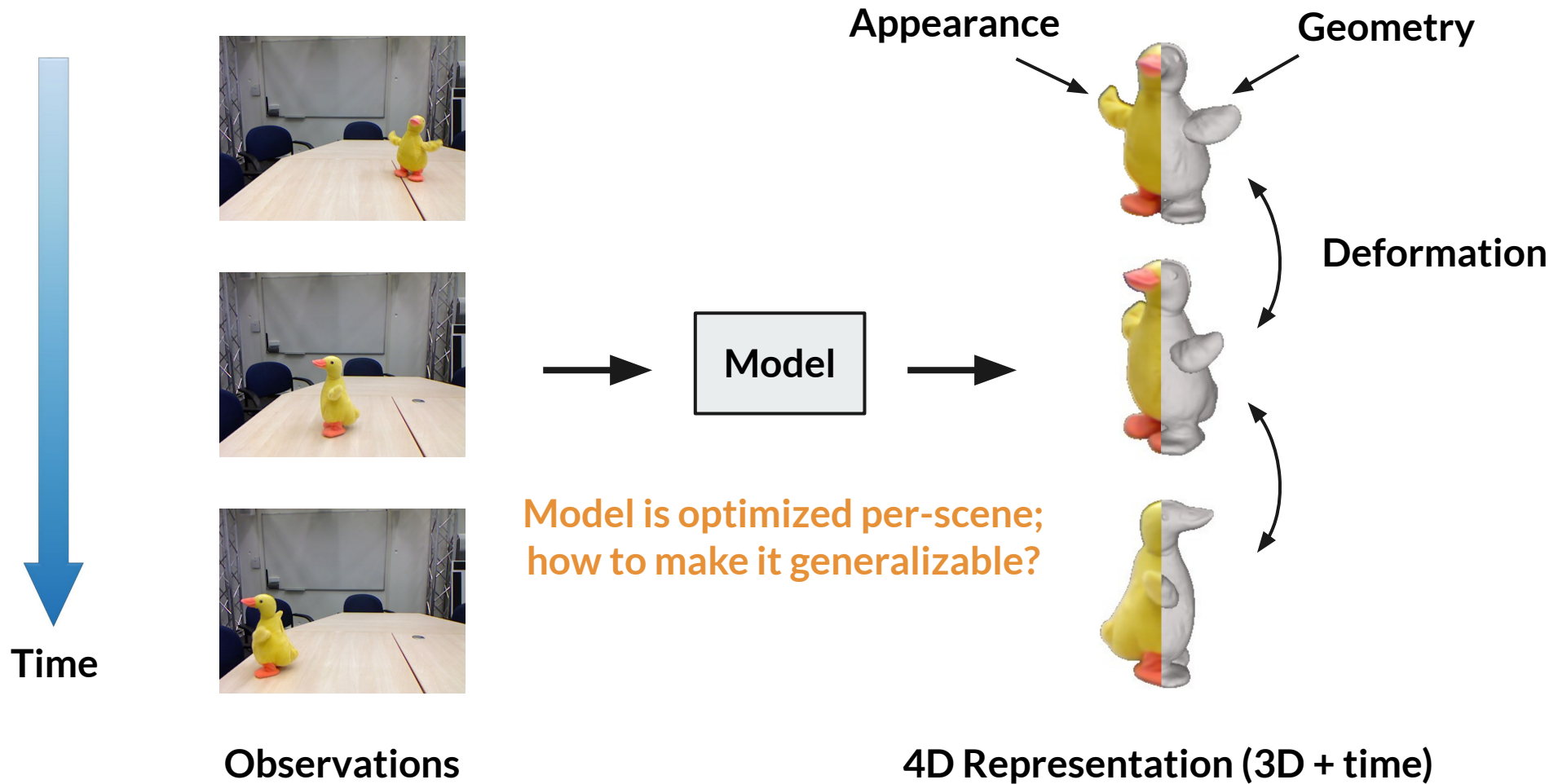


Novel View Synthesis:
Render the scene from
unobserved viewpoints

4D Representation (3D + time)

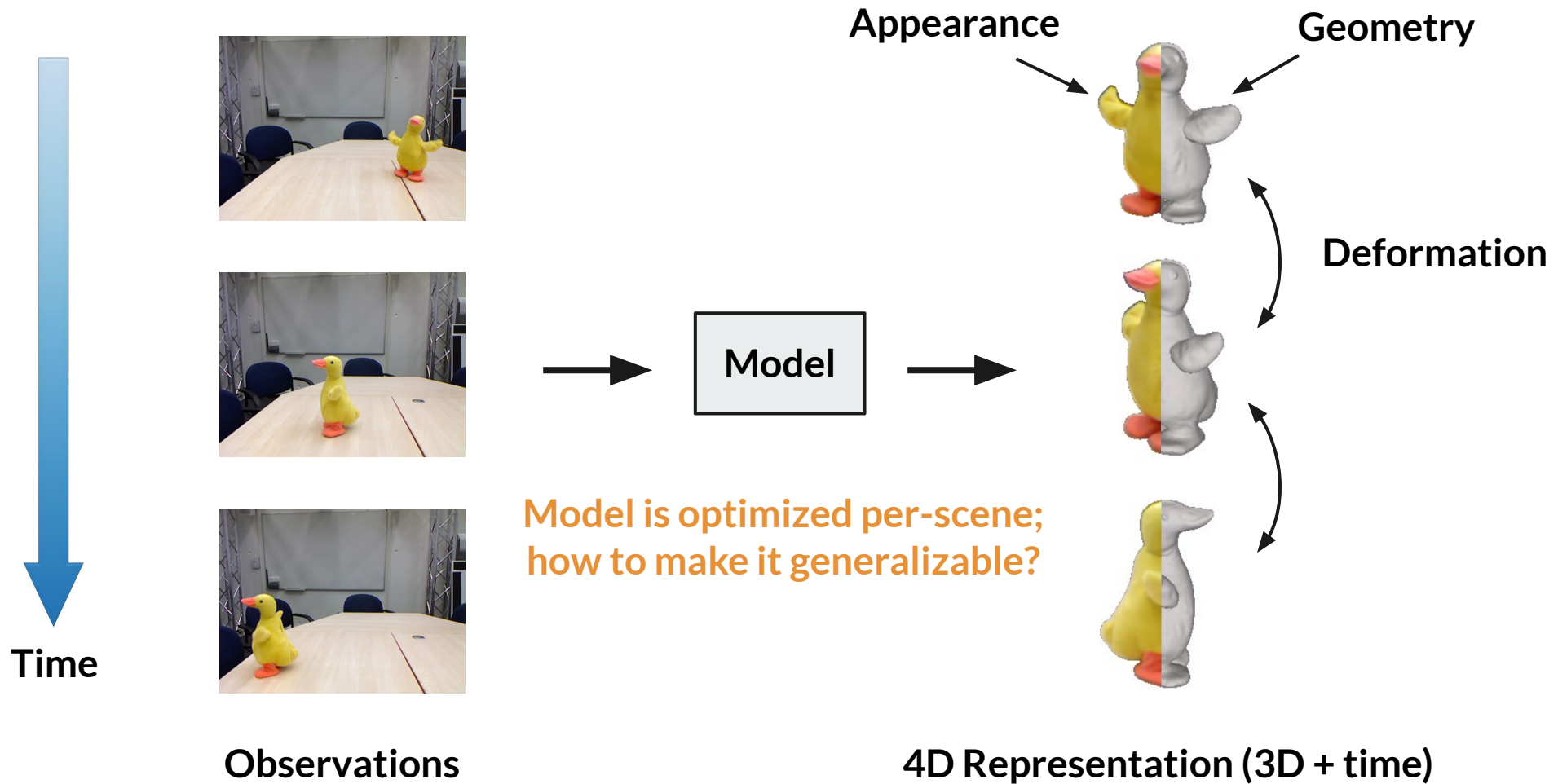
Task

Non-Rigid 3D Reconstruction and View Synthesis



Task

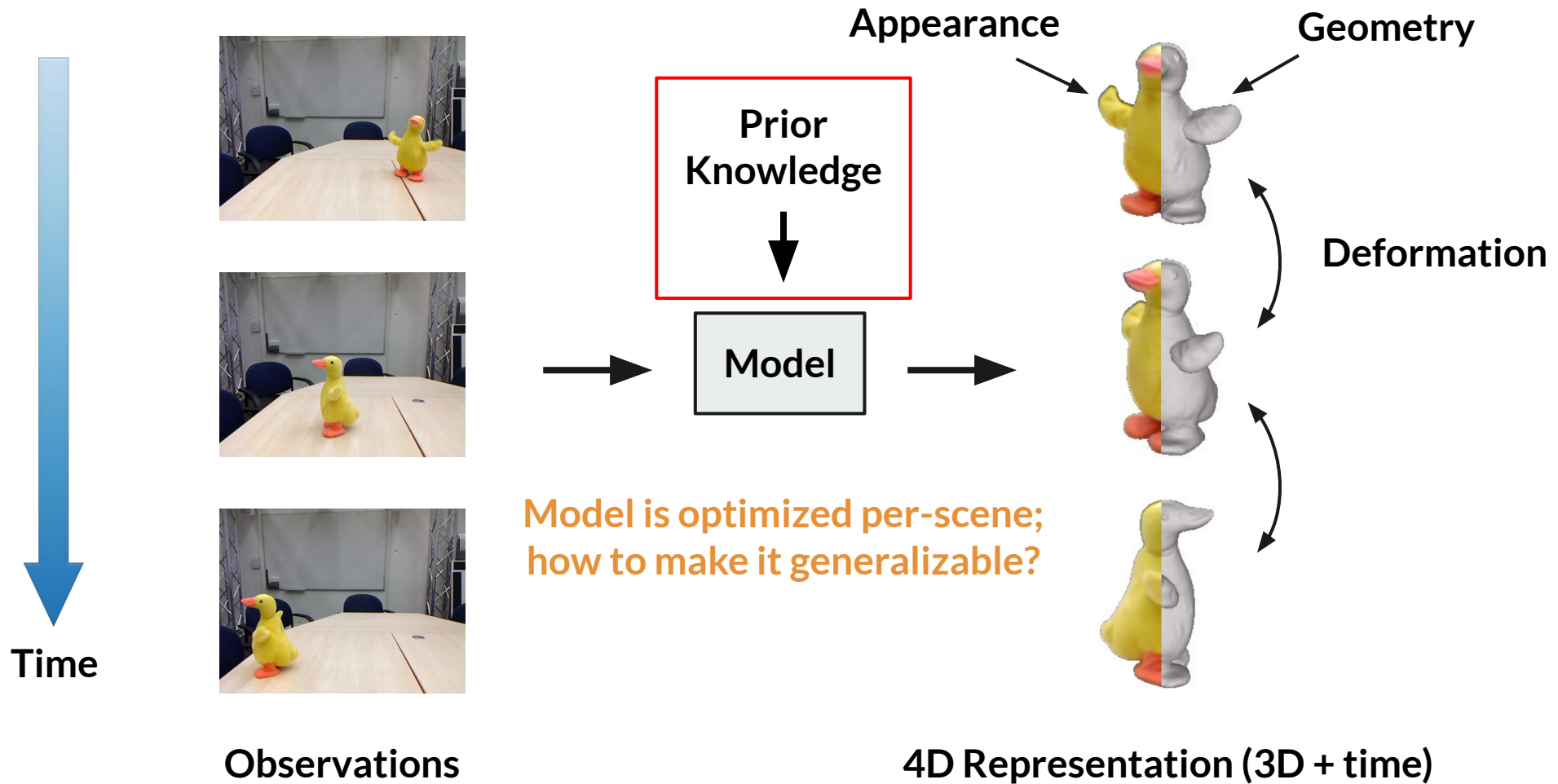
Non-Rigid 3D Reconstruction and View Synthesis



Also, how to get a better reconstruction when observations are sparse?

Task

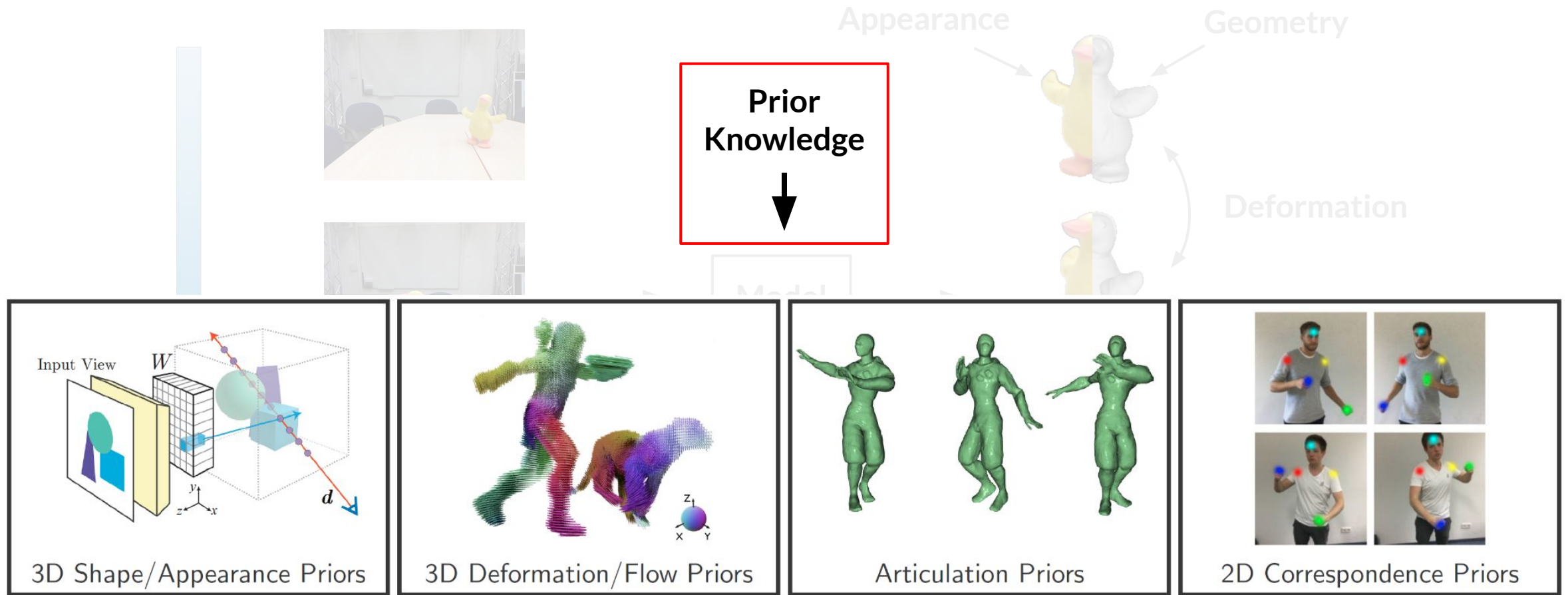
Non-Rigid 3D Reconstruction and View Synthesis



Also, how to get a better reconstruction when observations are sparse?

Task

Non-Rigid 3D Reconstruction and View Synthesis

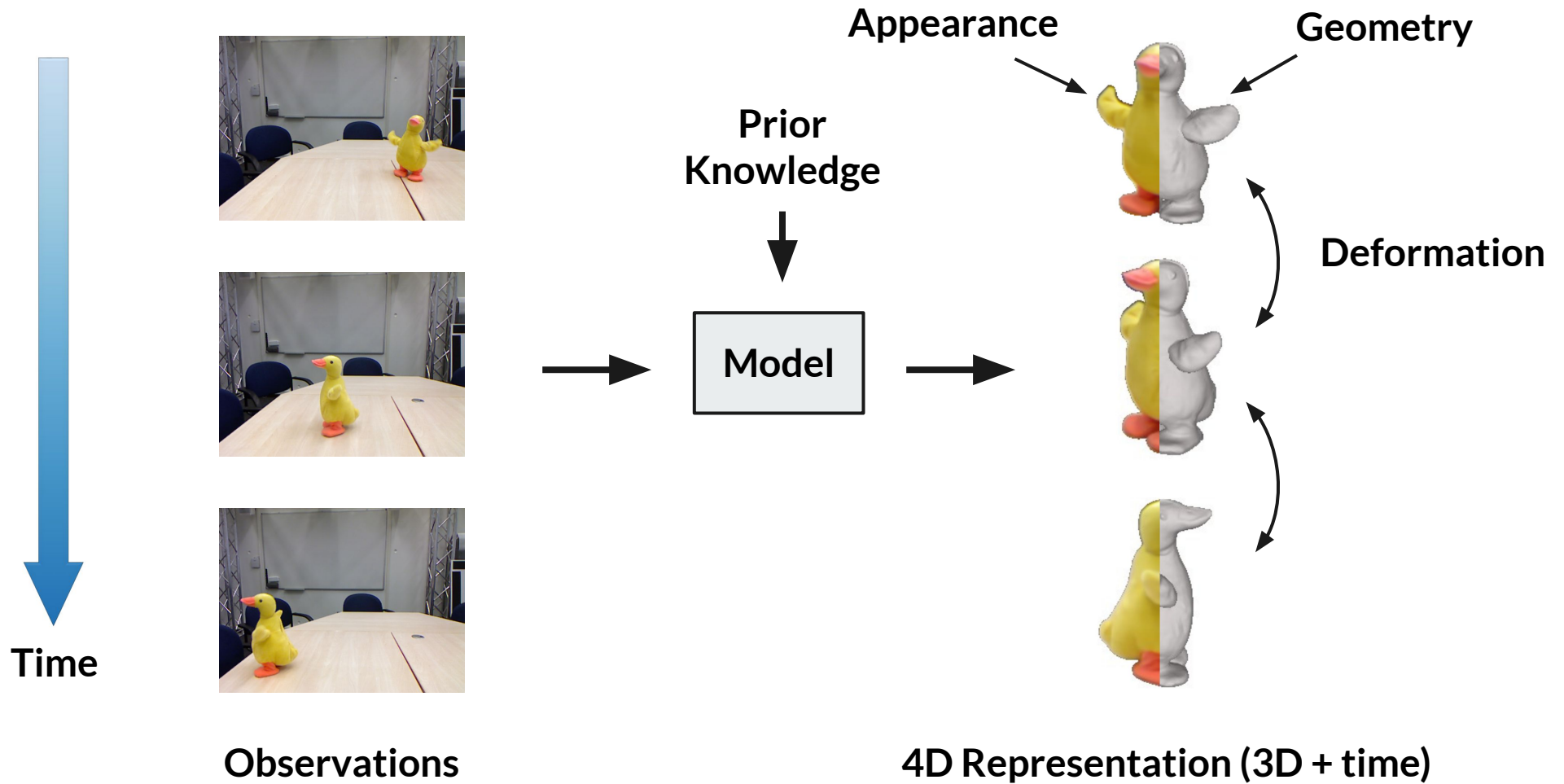


Observations

4D Representation – 3D + time

Task

Non-Rigid 3D Reconstruction and View Synthesis



Task

Non-Rigid 3D Reconstruction and View Synthesis



Trends

1. Speed and Quality Advancements
2. Handling of Large Deformations / Long-Term 3D Correspondences
3. Modelling Articulated Motion for General Objects

Trends

1. Speed and Quality Advancements

2. Handling of Large Deformations / Long-term 3D correspondences

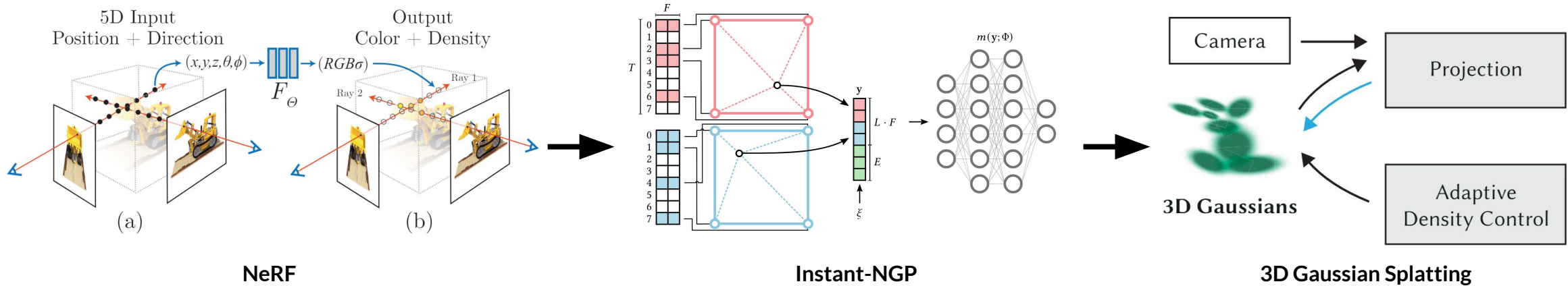
3. Modelling Articulated Motion for General Objects



Speed and Quality Advancements

Seminal Works in 3D Rigid Reconstruction and View Synthesis

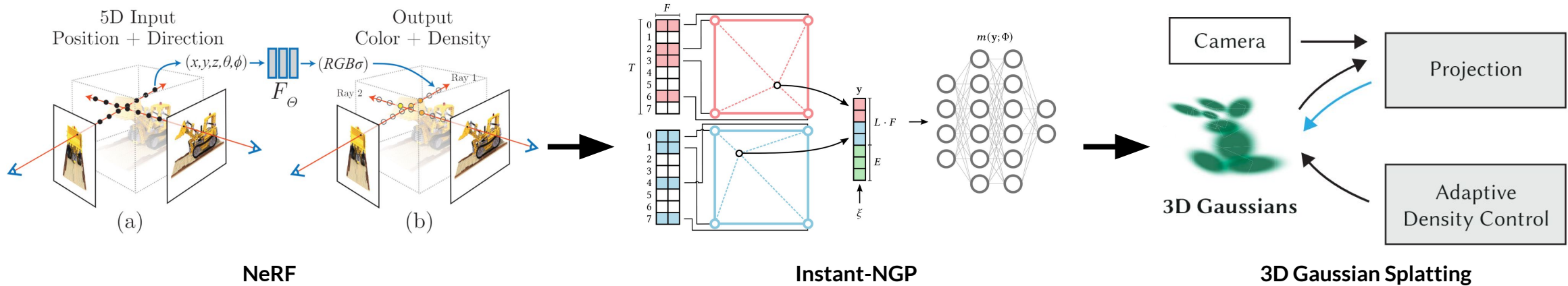
Quality or speed advancements in non-rigid setting follows the advancements in rigid setting:



Speed and Quality Advancements

Seminal Works in 3D Rigid Reconstruction and View Synthesis

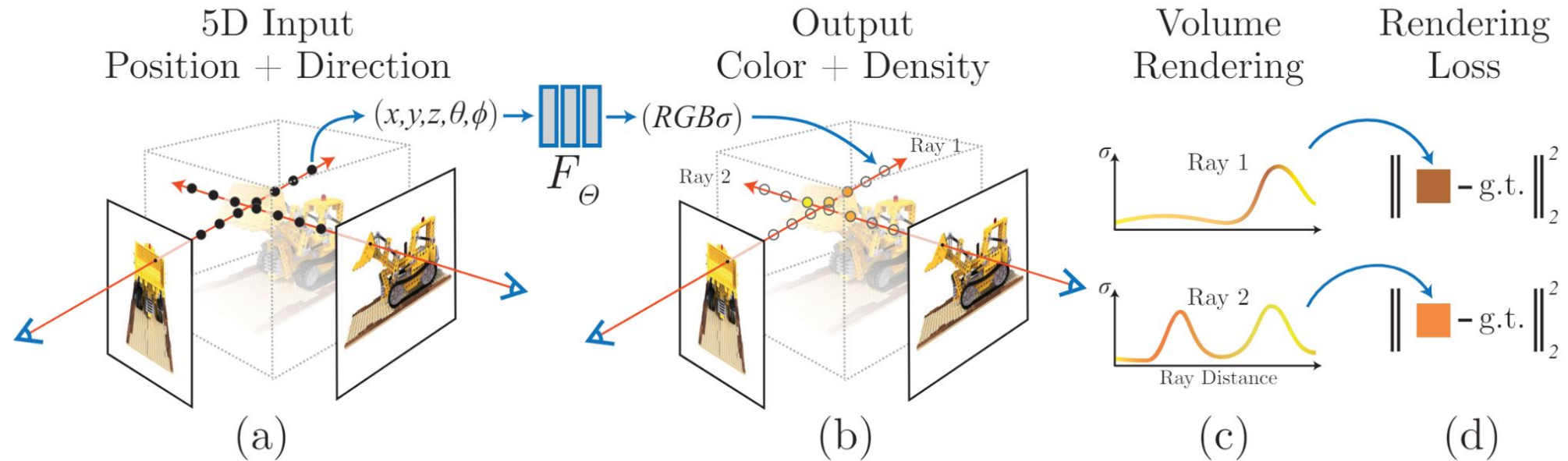
Quality or speed advancements in non-rigid setting follows the advancements in rigid setting:



Let's see how these rigid setting advancements have been adapted to the non-rigid setting in recent years

Speed and Quality Advancements

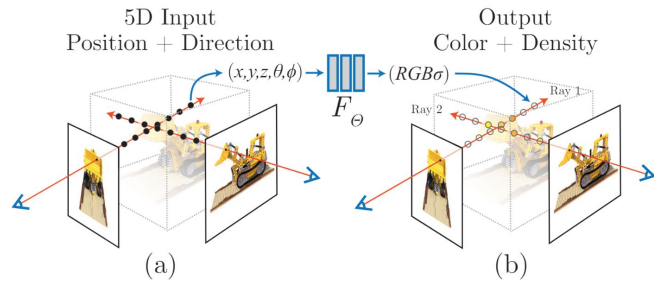
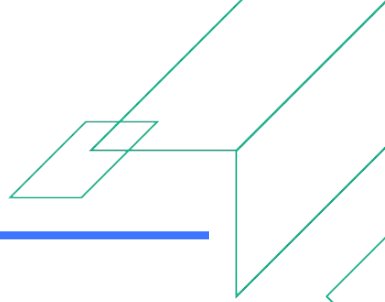
Photo-realistic View Synthesis: Neural Scene Representations



NeRF

Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations

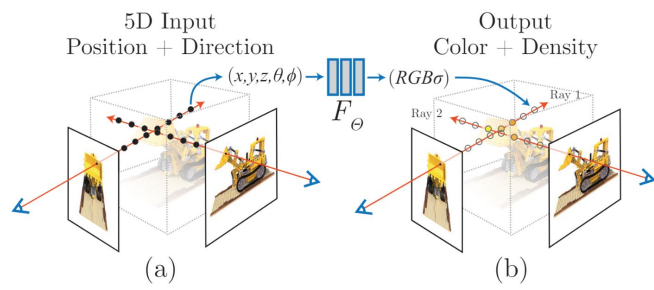


NeRF

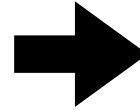


Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations

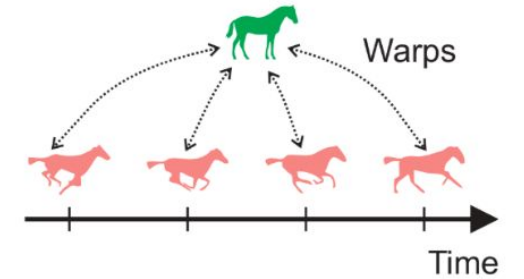


NeRF



Deformable NeRF

Global Canonical Model

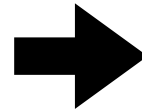
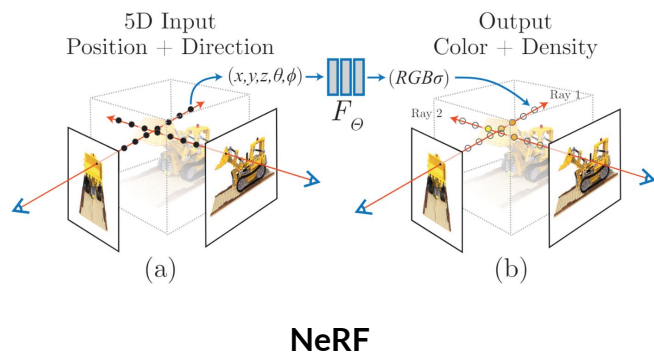


Nerfies



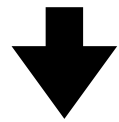
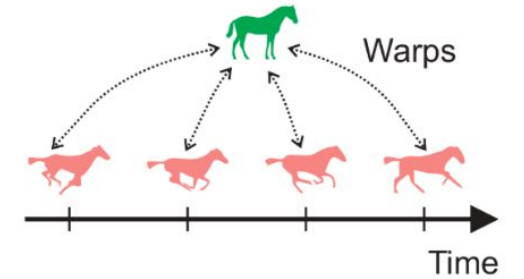
Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



Deformable NeRF

Global Canonical Model

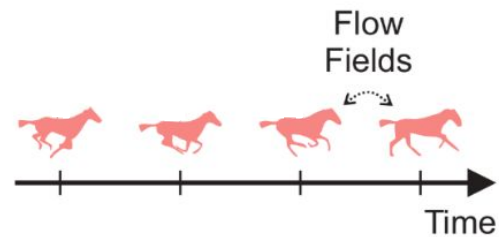


Space-Time NeRF

Individual Frame



Nerfies

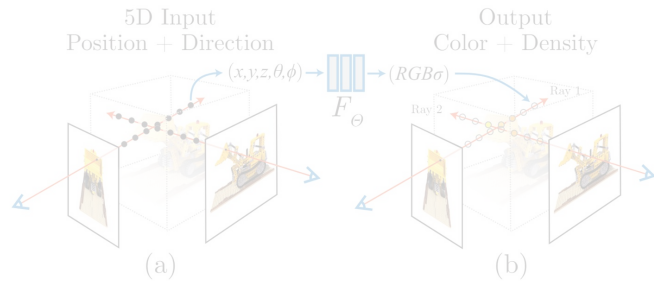
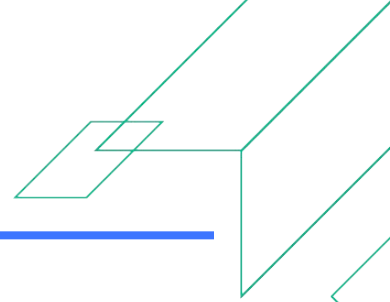


Neural Scene Flow Fields

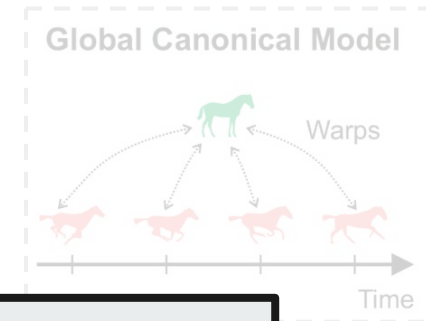


Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



Deformable NeRF



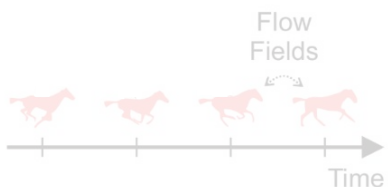
- Further advancements mostly with hybrid representations
- A few advancements were also seen in quality with purely neural scene representations



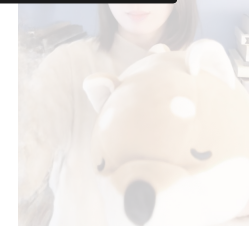
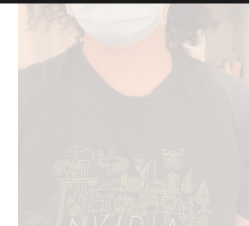
Space-Time NeRF



Individual Frame



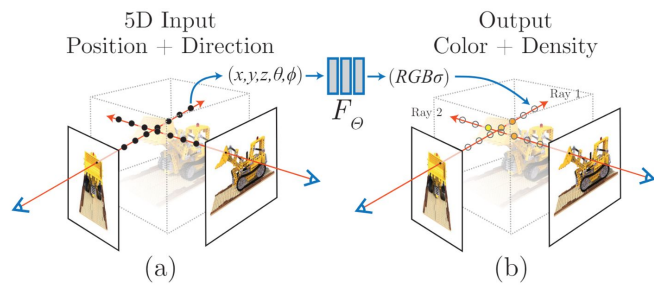
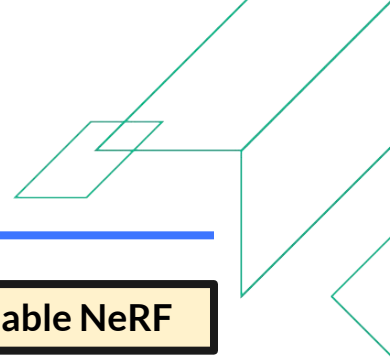
Neural Scene Flow Fields



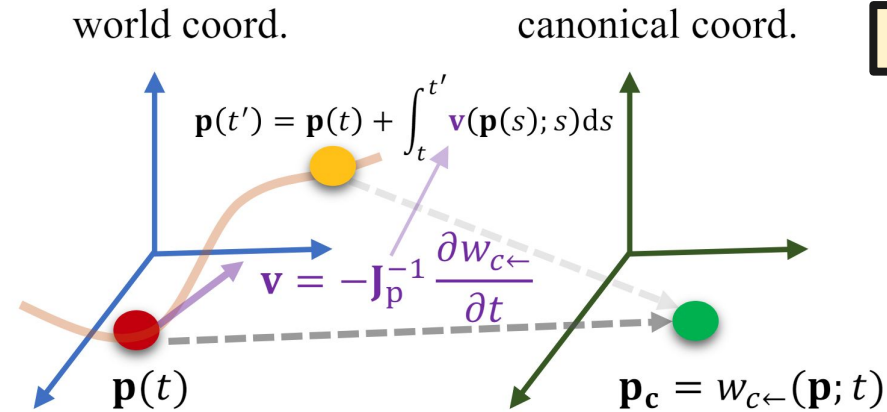
Nerfies

Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations

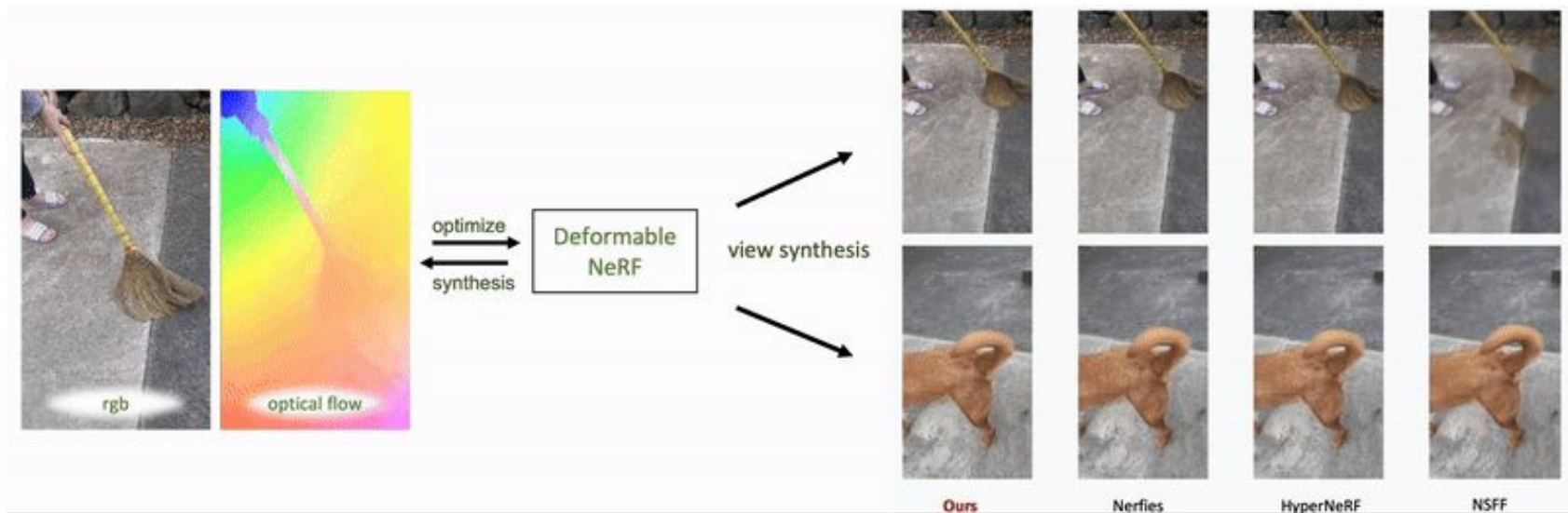


NeRF



Deformable NeRF

Optical Flow Supervision:

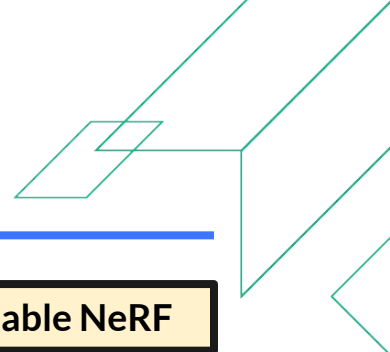


FSD-NeRF

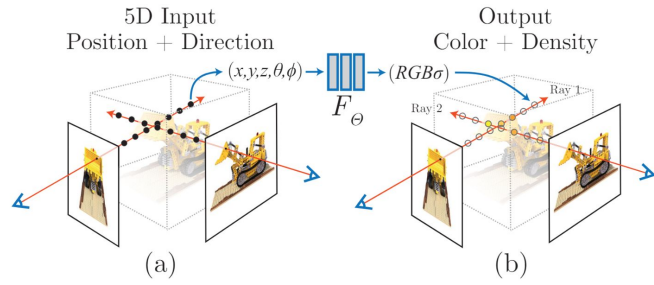


Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations

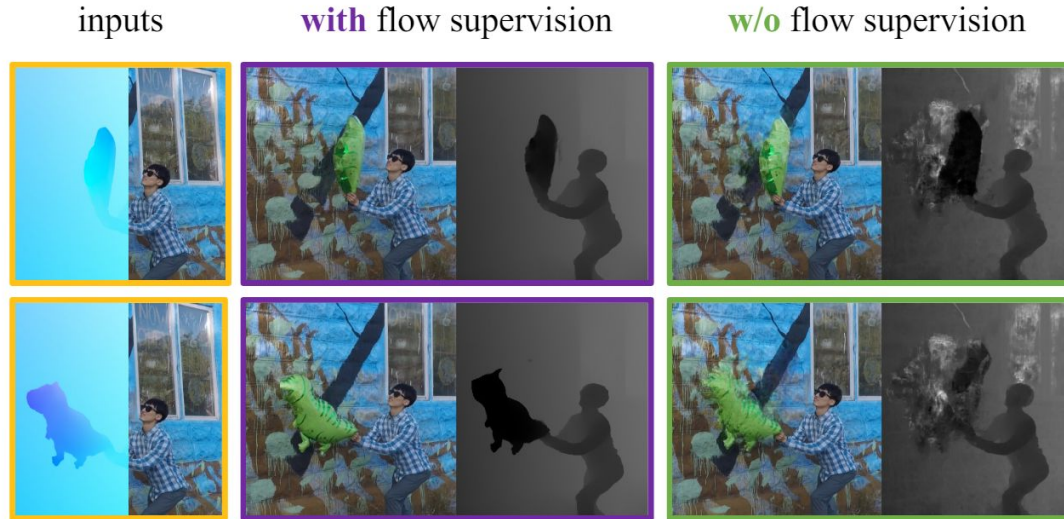


Deformable NeRF



NeRF

Optical Flow
Supervision:



FSD-NeRF



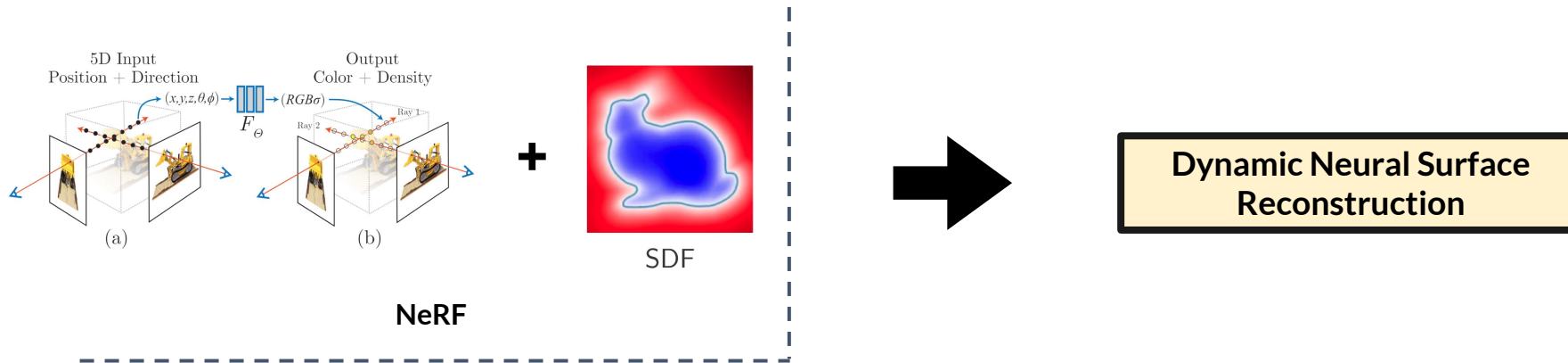
Speed and Quality Advancements

High-fidelity Geometry: Neural Scene Representations

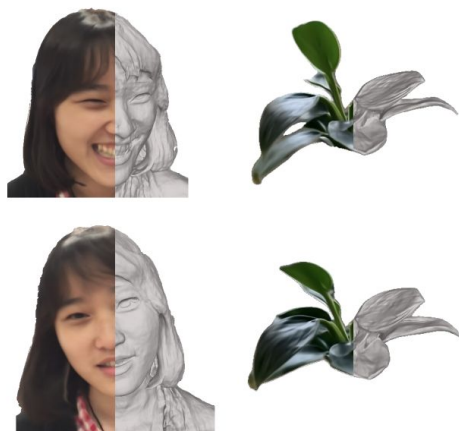


Speed and Quality Advancements

High-fidelity Geometry: Neural Scene Representations



- RGB-D with mask



NDR

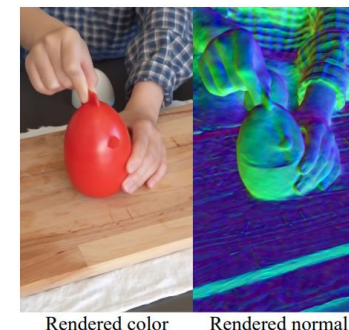
- RGB with mask and mesh proxy



Input Image Input View Novel View

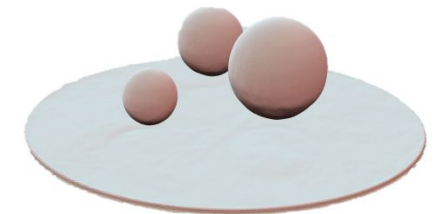
Unbiased4D

- RGB only



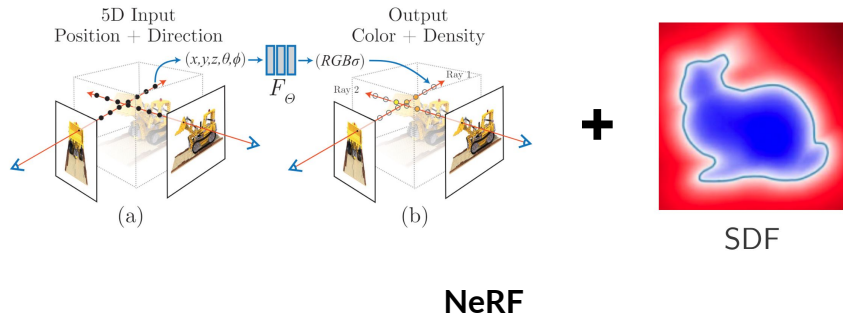
Rendered color Rendered normal

4DRegSDF



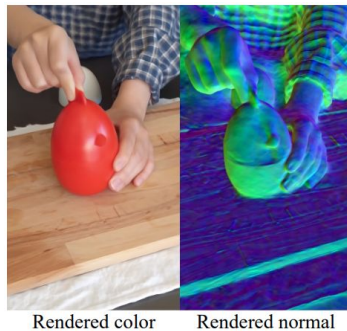
Speed and Quality Advancements

High-fidelity Geometry: Neural Scene Representations

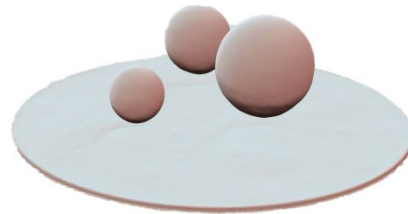


Dynamic Neural Surface Reconstruction

- RGB only



4DRegSDF



$$\sum (\text{curvature}_1 = \text{curvature}_2) \quad : \text{enforce local rigidity}$$

Total variation of curvature

$$\sum (\text{curvature}_1 \rightarrow \text{curvature}_2) \quad : \text{limit unnecessary kinks}$$

Absolute curvature of SDF

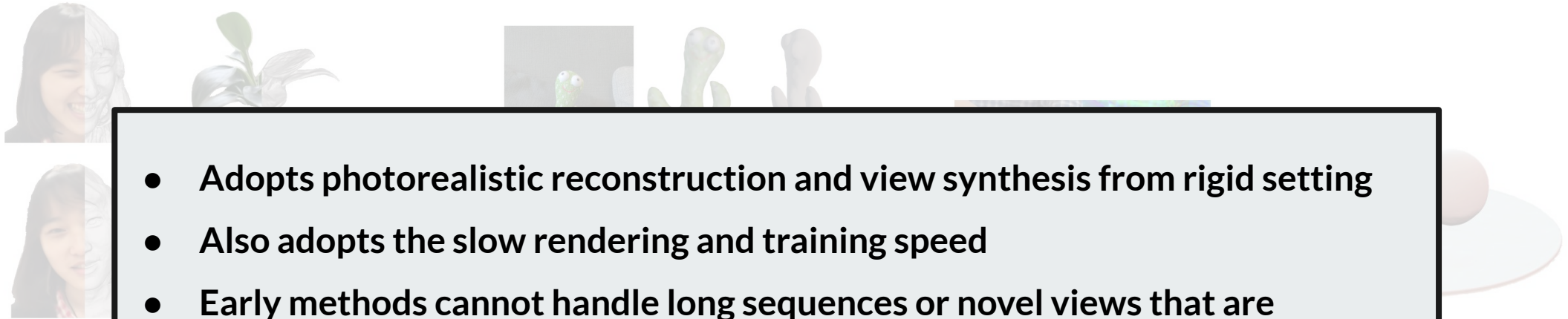
$$\sum (\text{gradient}_1 \rightarrow \text{gradient}_2) \quad : \text{make gradient of SDF valid}$$

Eikonal loss

Speed and Quality Advancements

High-fidelity Geometry: Neural Scene Representations

- Extensions regarding additional inputs, surface reconstruction, model improvements, etc. are usually seen with pure neural fields first



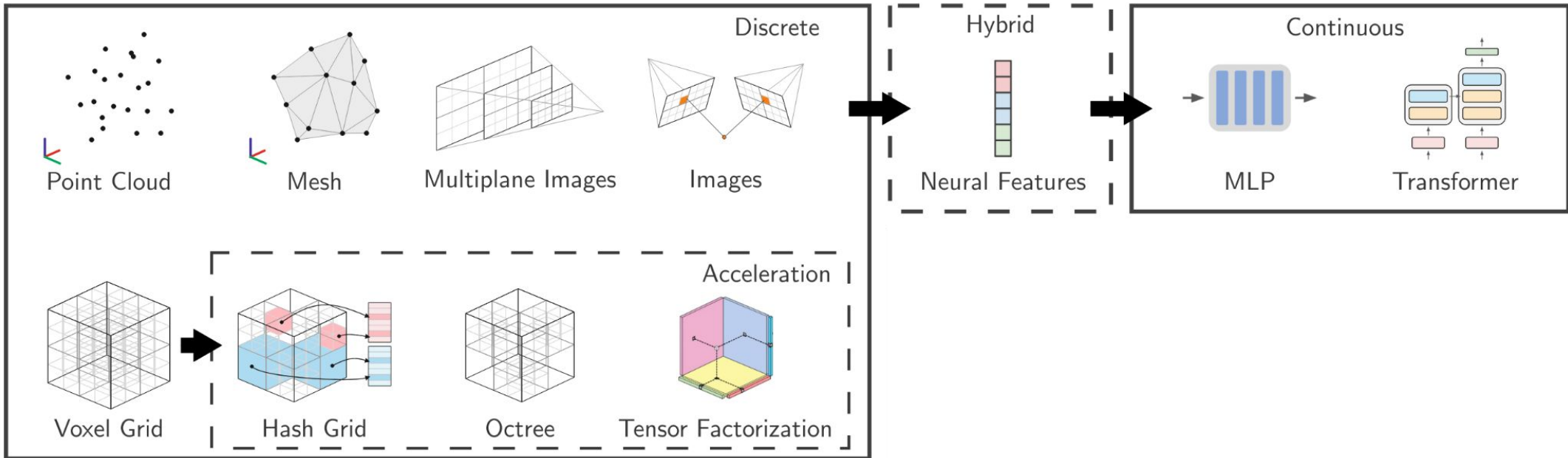
- **Adopts photorealistic reconstruction and view synthesis from rigid setting**
- **Also adopts the slow rendering and training speed**
- **Early methods cannot handle long sequences or novel views that are significantly different than training views**

- RGB-D with mask
- RGB with mask and mesh proxy
- RGB only

Surface Reconstruction

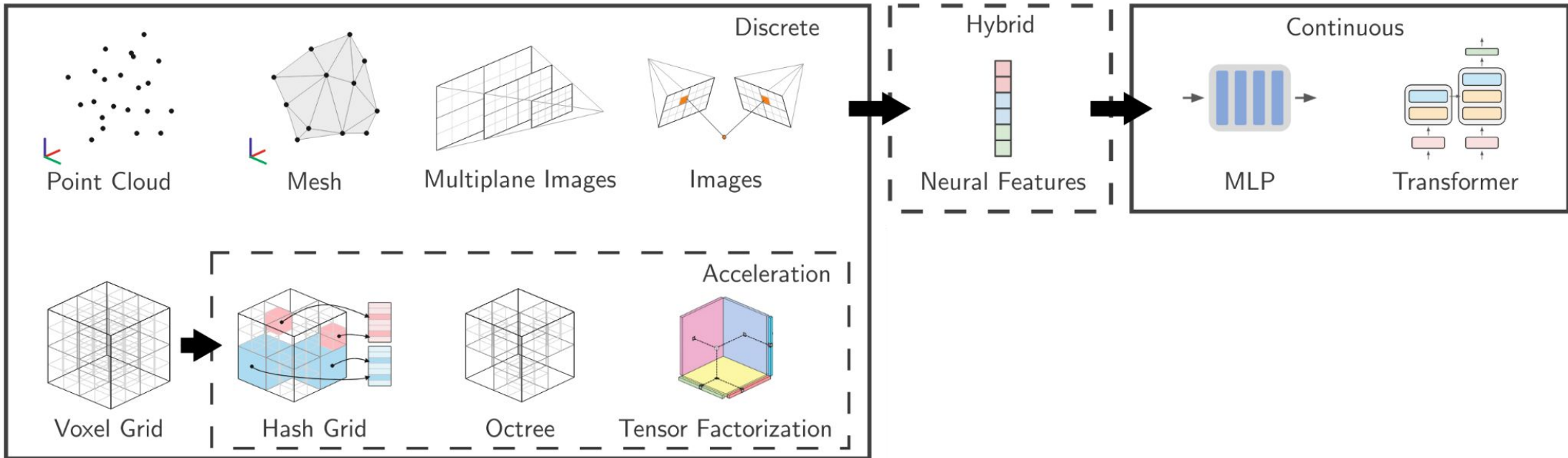
Speed and Quality Advancements

3D Scene Representations



Speed and Quality Advancements

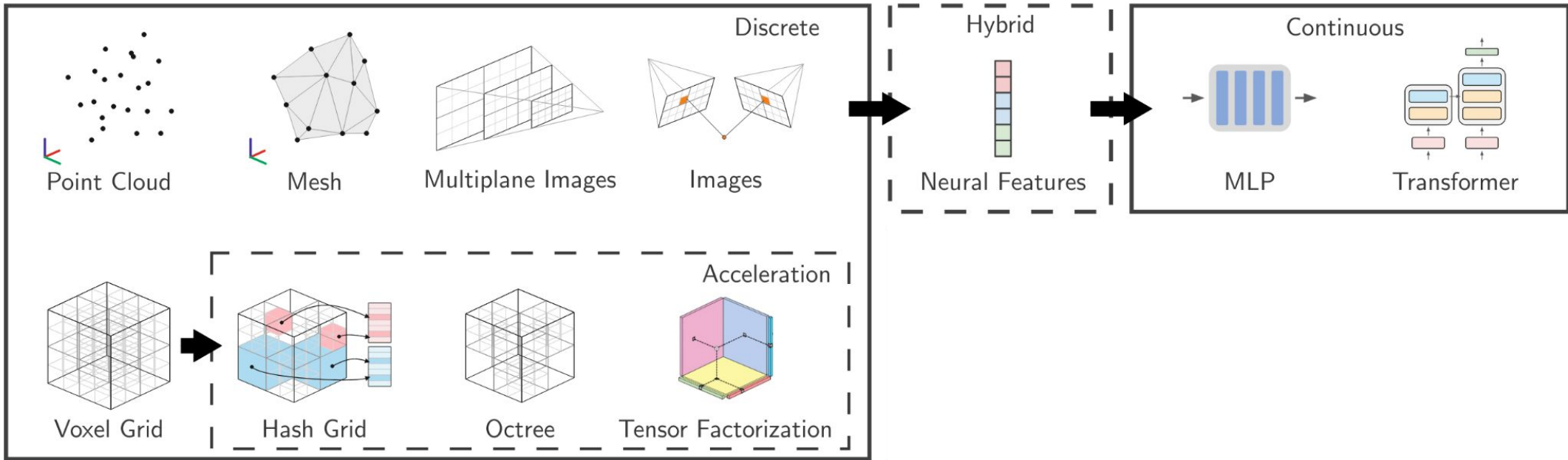
3D Scene Representations



$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta),$$

Speed and Quality Advancements

3D Scene Representations

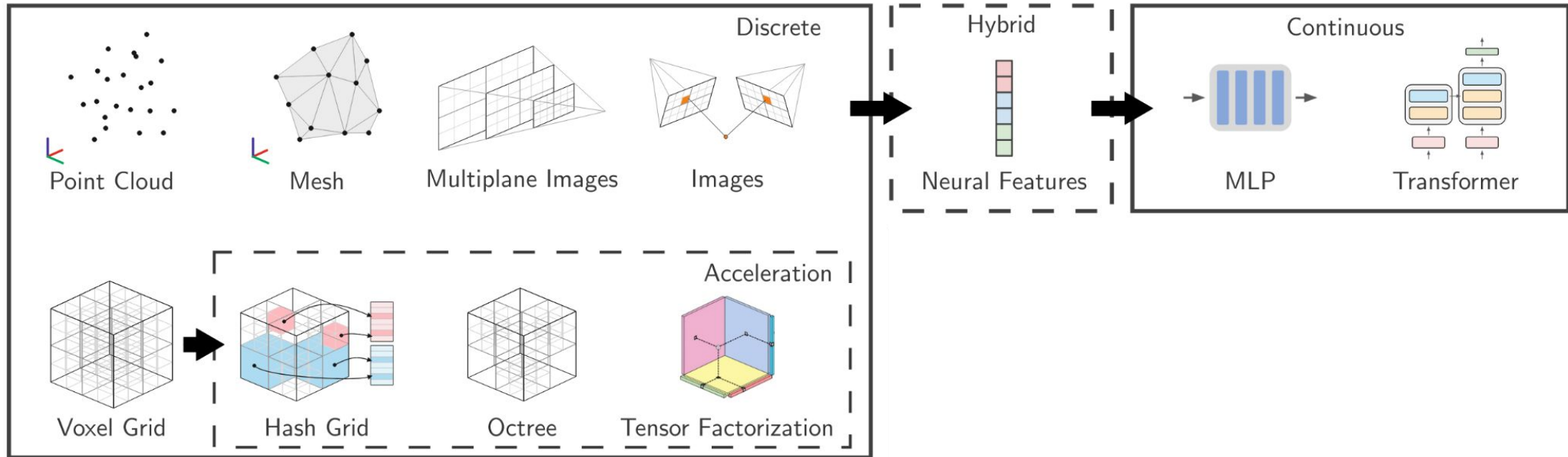


$\mathbf{x} \in \mathbb{R}^3$ are the 3D coordinates

$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta),$$

Speed and Quality Advancements

3D Scene Representations

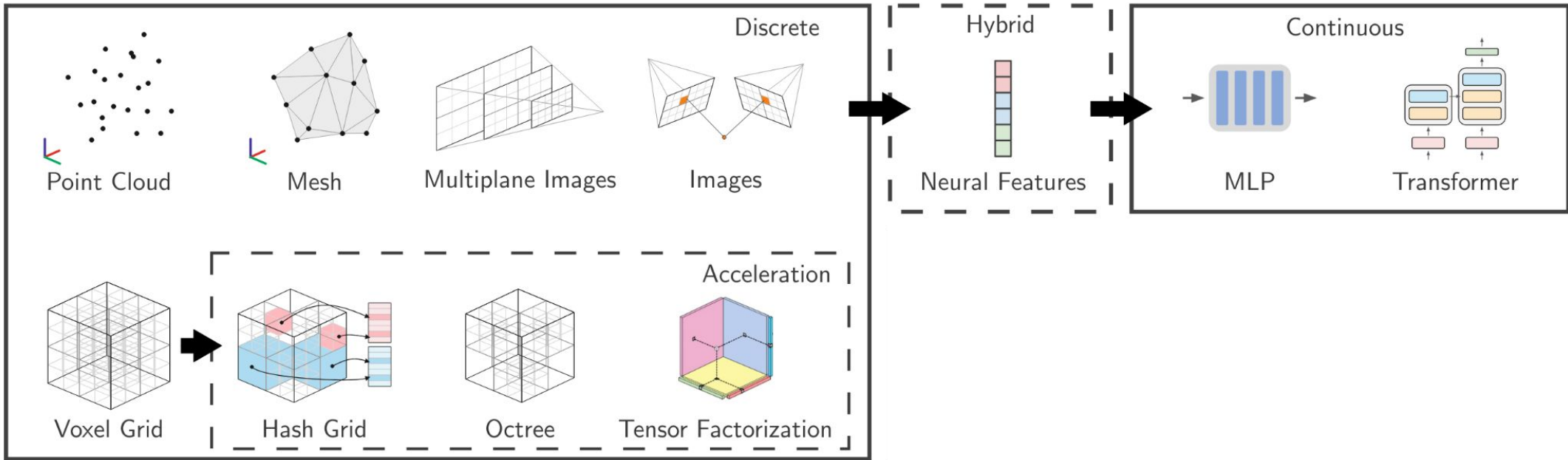


$\mathbf{x} \in \mathbb{R}^3$ are the 3D coordinates

$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta),$ \mathcal{H} are optional additional inputs (e.g. view direction)

Speed and Quality Advancements

3D Scene Representations



$\mathbf{x} \in \mathbb{R}^3$ are the 3D coordinates

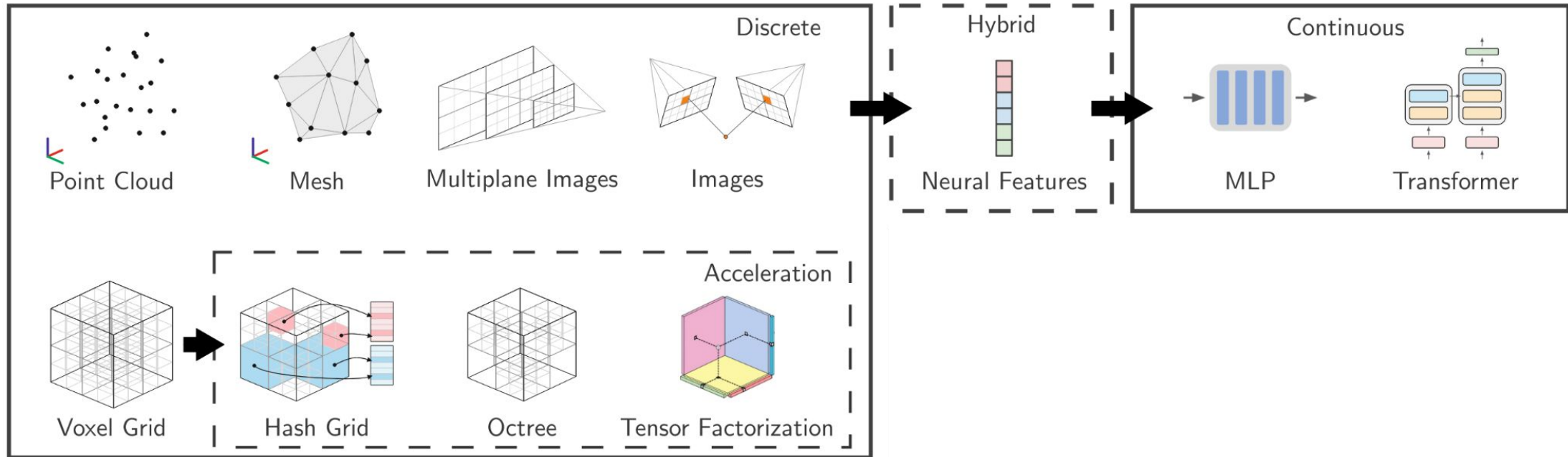
\mathcal{H} are optional additional inputs (e.g. view direction)

$\boldsymbol{\theta}$ stores the scene information

$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \boldsymbol{\theta}),$

Speed and Quality Advancements

3D Scene Representations



$\mathbf{x} \in \mathbb{R}^3$ are the 3D coordinates

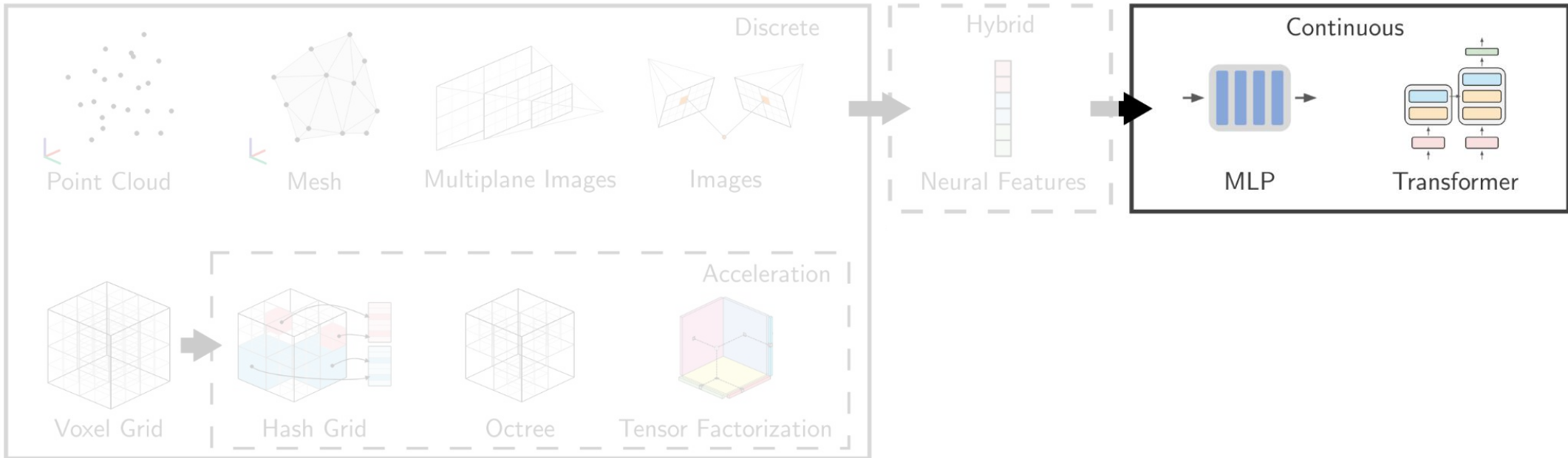
\mathcal{H} are optional additional inputs (e.g. view direction)

$\boldsymbol{\theta}$ stores the scene information

\mathbf{y} represents any scene property (e.g. geometry, colour, deformation, etc.)

Speed and Quality Advancements

3D Scene Representations

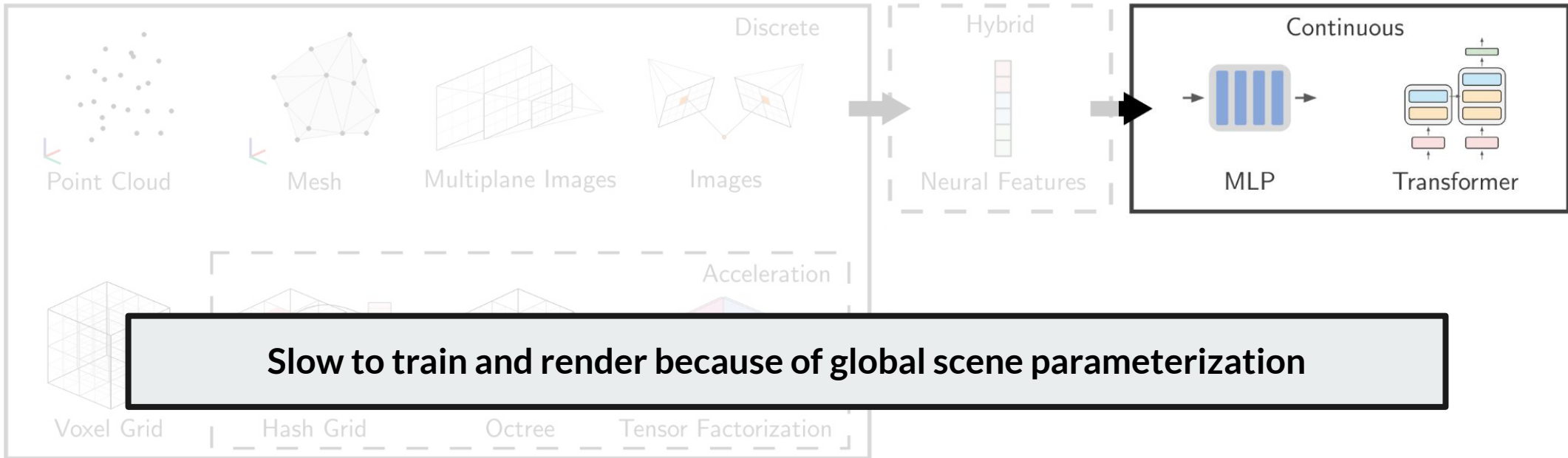


$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \boldsymbol{\theta}),$$

$\boldsymbol{\theta}$ is stored in network parameters
 ρ is an MLP or Transformer

Speed and Quality Advancements

3D Scene Representations

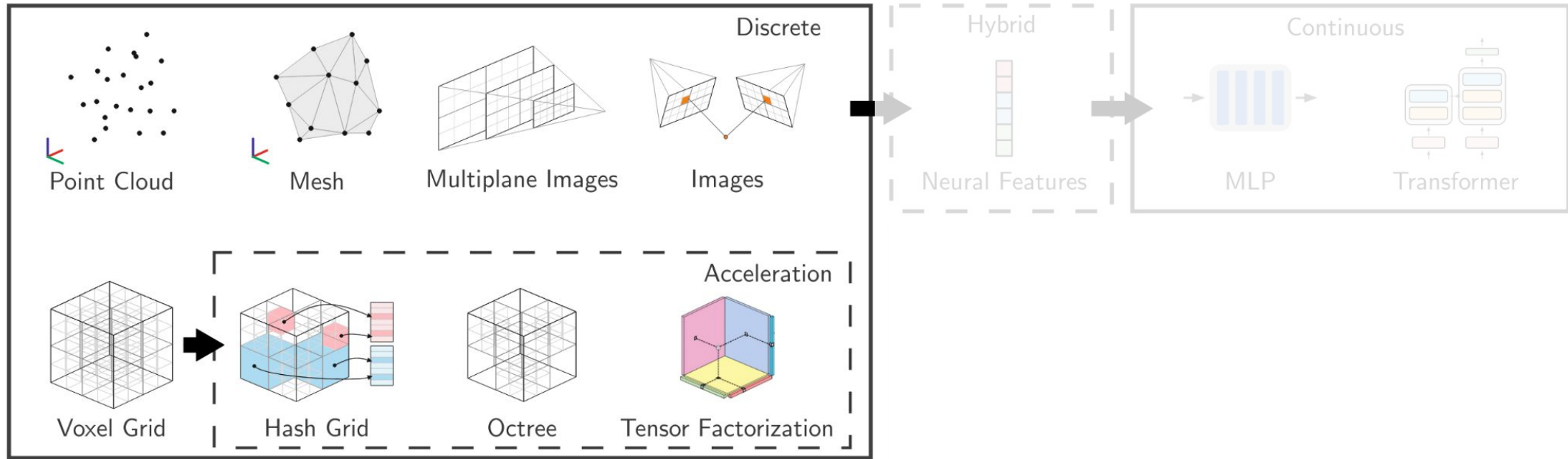


$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \boldsymbol{\theta}),$$

$\boldsymbol{\theta}$ is stored in network parameters
 ρ is an MLP or Transformer

Speed and Quality Advancements

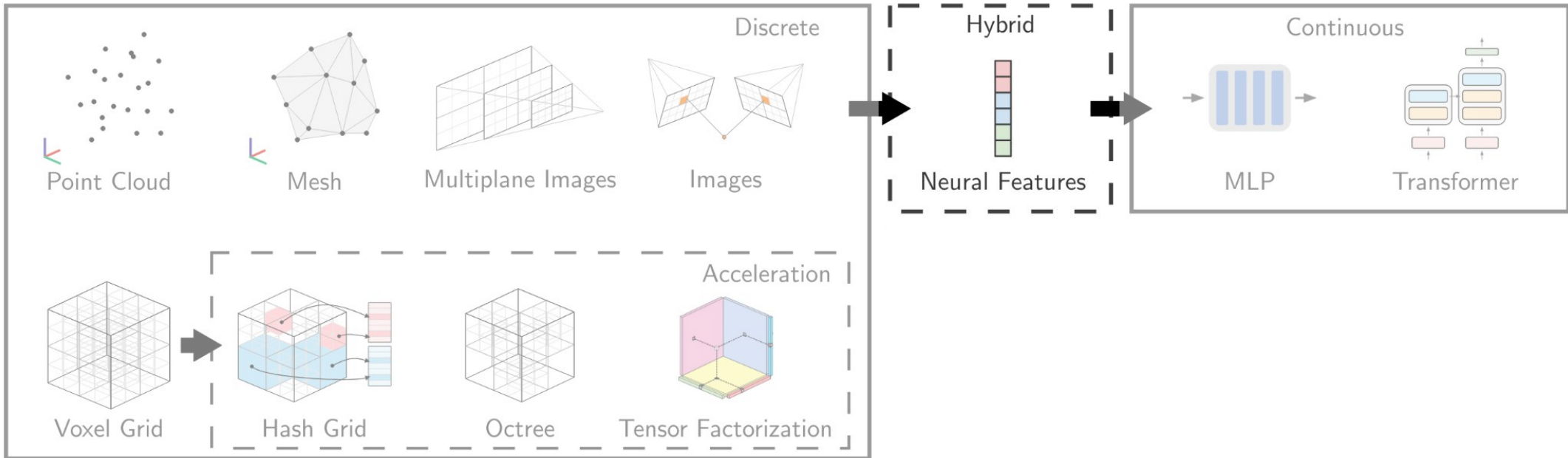
3D Scene Representations



$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \boldsymbol{\theta}),$
 $\boldsymbol{\theta}$ is stored at discretely defined nodes
 ρ interpolates the scene information for any continuous 3D point

Speed and Quality Advancements

3D Scene Representations

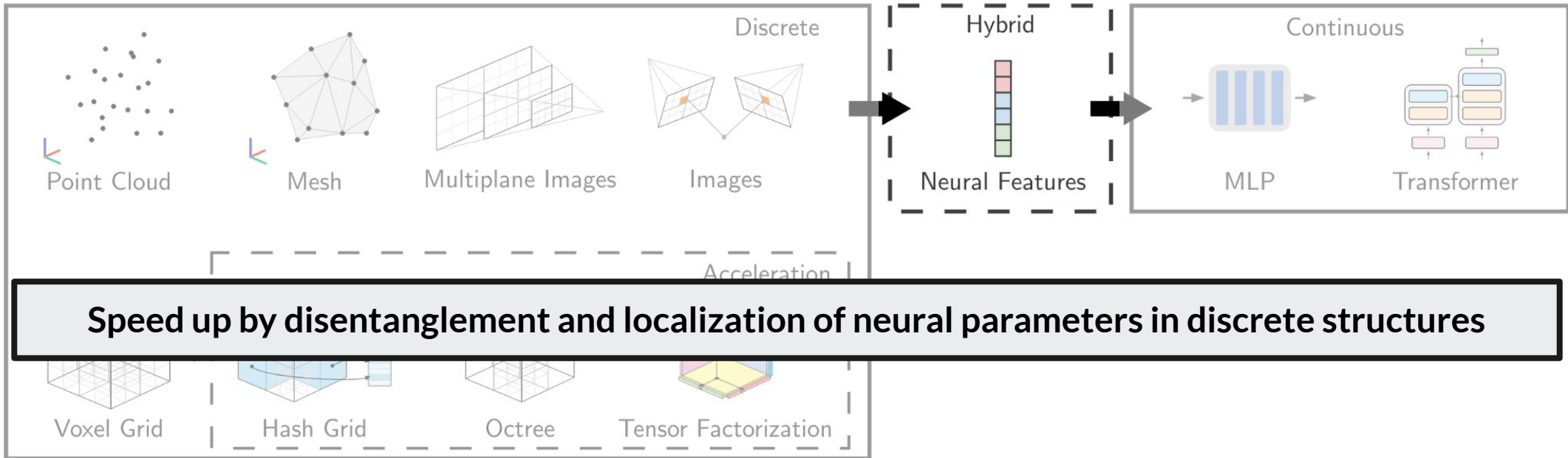


$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \boldsymbol{\theta}),$

- $\boldsymbol{\theta}$ are neural features stored in a discrete structure
- ρ defines interpolation of discrete information followed by network query

Speed and Quality Advancements

3D Scene Representations



$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \boldsymbol{\theta}),$$

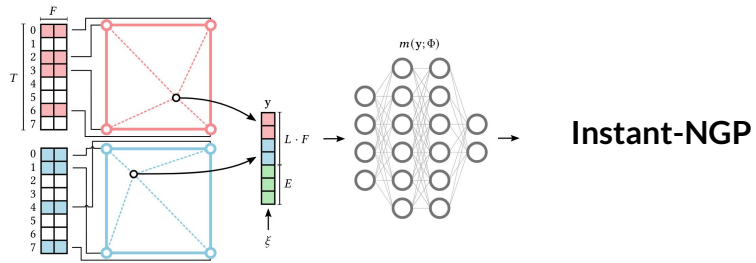
$\boldsymbol{\theta}$ are neural features stored in a discrete structure

ρ defines interpolation of discrete information followed by network query

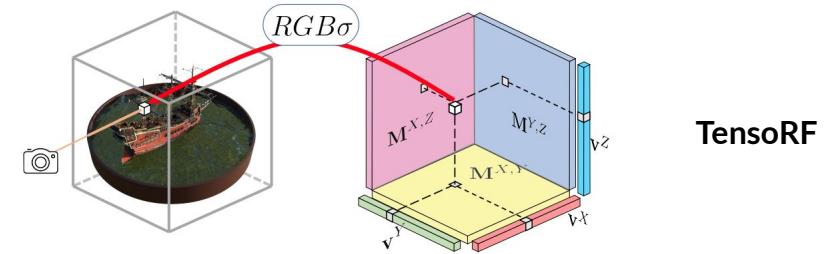
Speed and Quality Advancements

Seminal Hybrid Scene Representations for Rigid Setting

Voxel Grid



Planar Factorization



Points

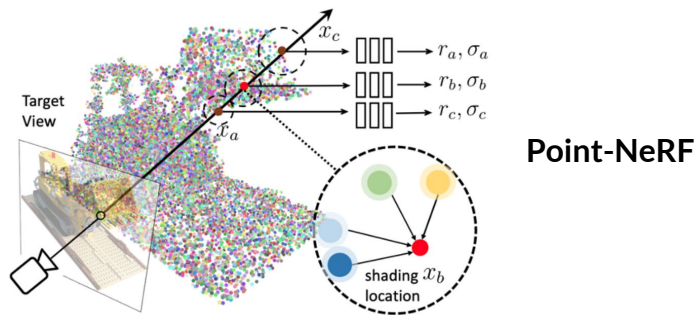
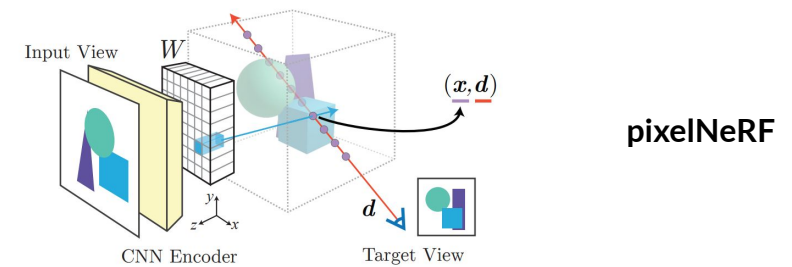
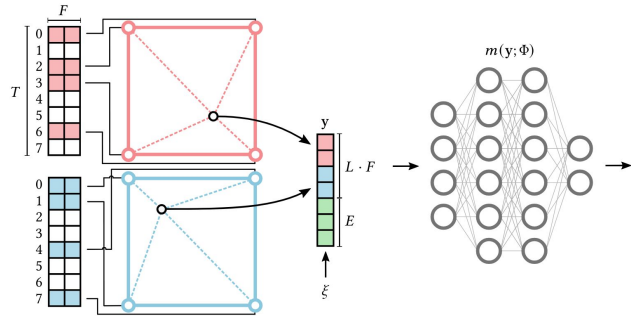
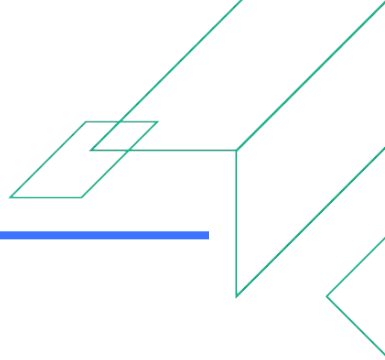


Image-based



Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations



Voxel Grid (Instant-NGP)

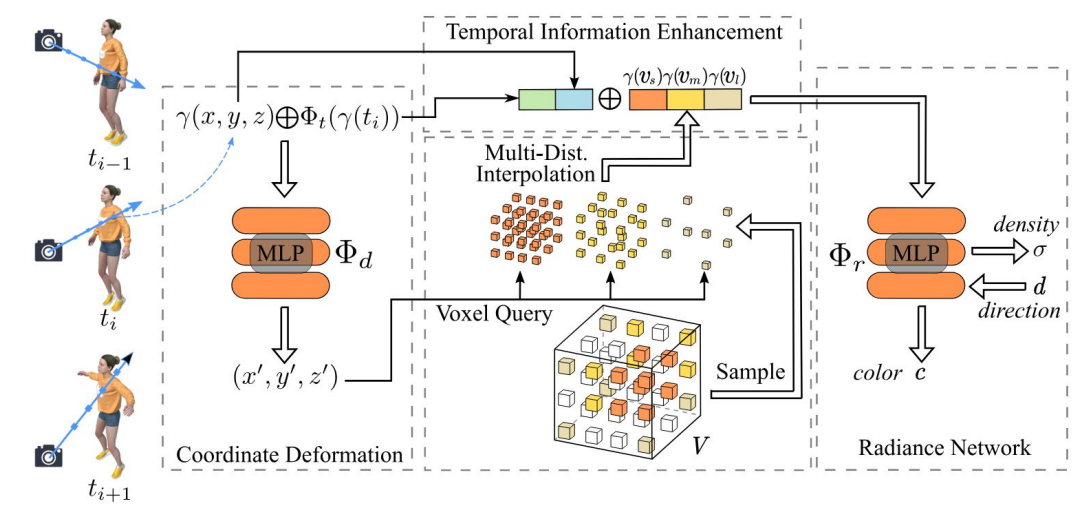
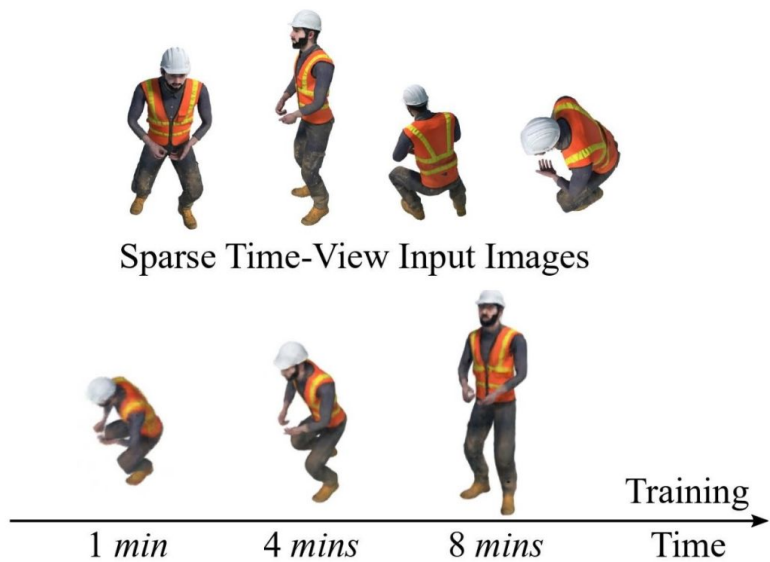
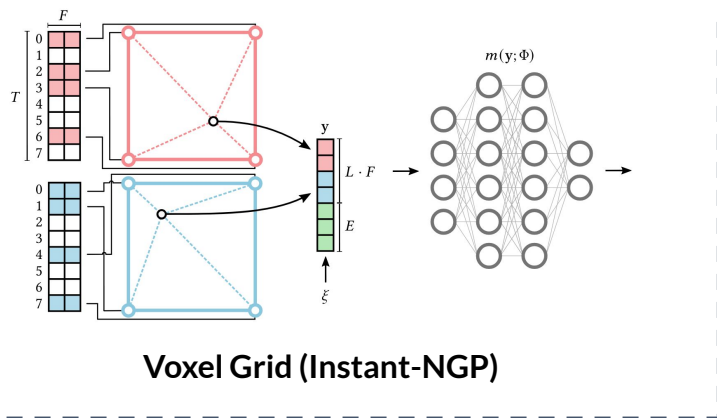


Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations

Dynamic Voxel NeRF

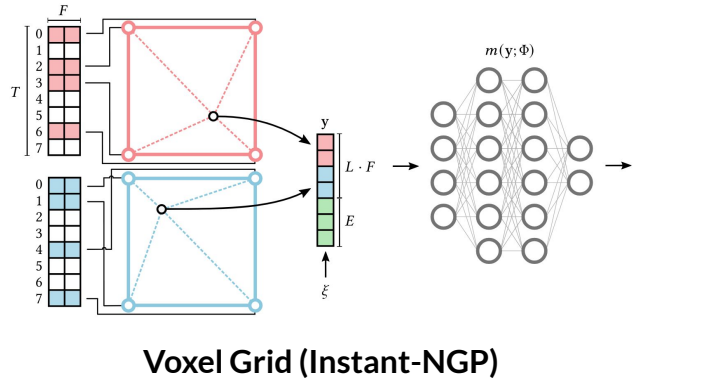
- Both the deformation field and canonical space are parameterized by MLPs with voxel grids
- Very light deformation MLP for fast training
- Canonical radiance field enhanced through temporal embeddings



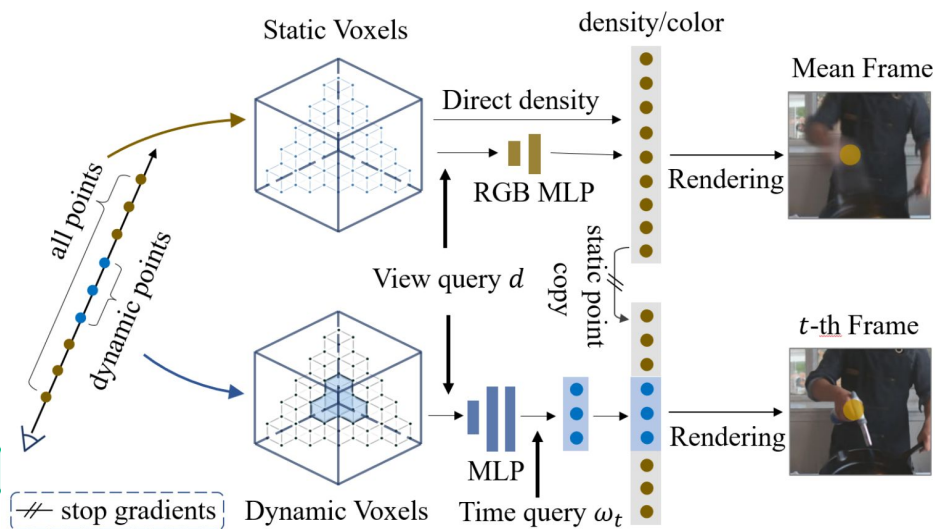
Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations

Dynamic Voxel NeRF



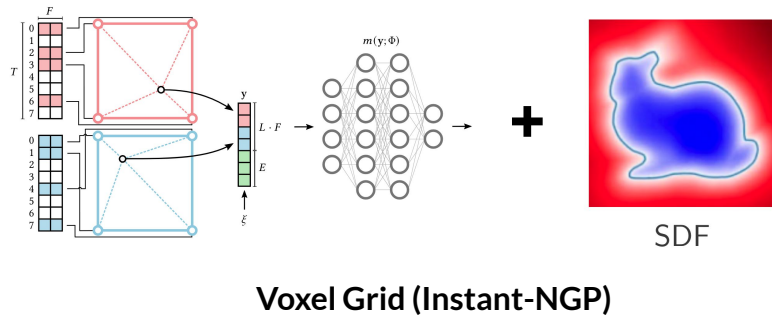
- Uses static-dynamic decomposition for better scalability and motion modelling capacity
- Lightweight static model for fast training from multi-view videos and near real-time rendering



MixVoxels

Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations



Dynamic Neural Surface Reconstruction

- Online per-frame optimization from multiple views
- Speeds up surface reconstruction while retaining high-quality

Novel view synthesis *Geometry reconstruction*



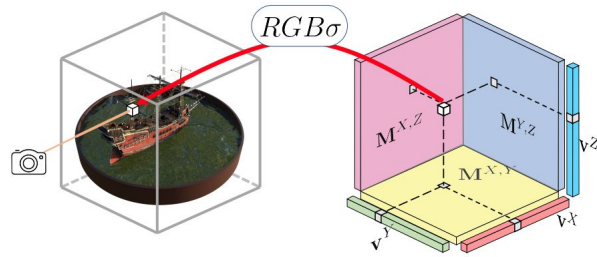
Reference images



NeuS2

Speed and Quality Advancements

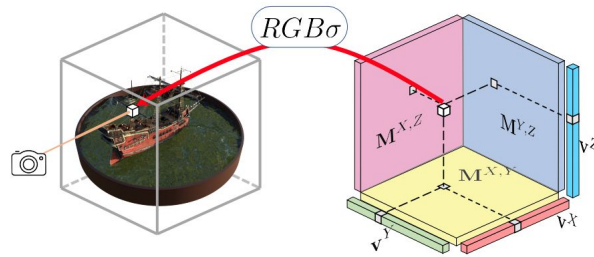
Faster Training and Rendering: Hybrid Neural Scene Representations



Planar Factorization (TensorRF)

Speed and Quality Advancements

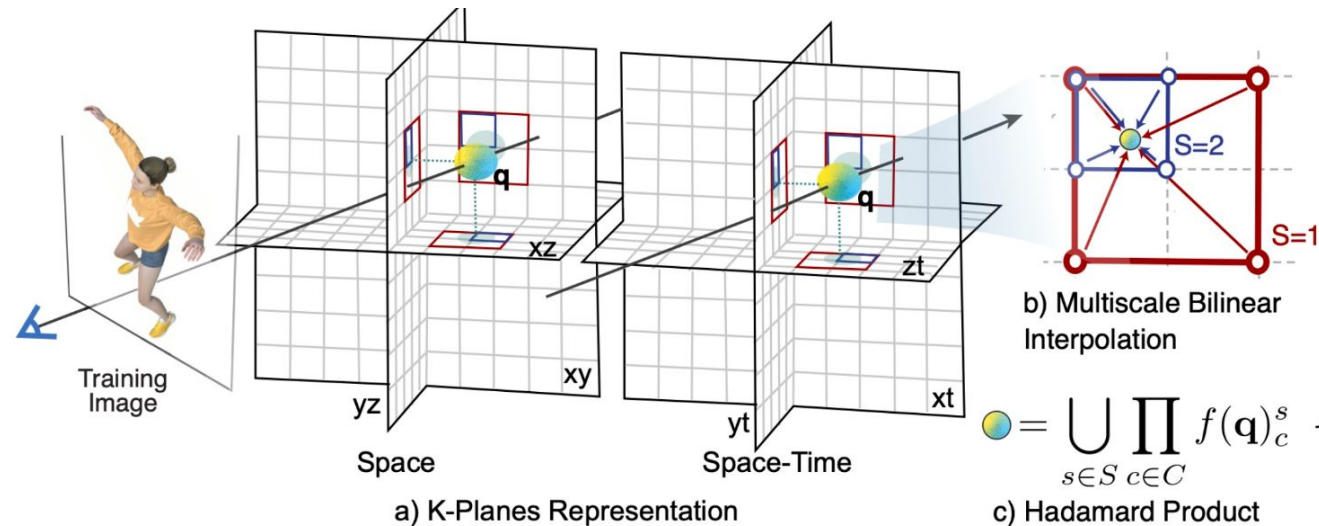
Faster Training and Rendering: Hybrid Neural Scene Representations



Planar Factorization (TensorRF)

Compaction of voxel grid
for memory efficiency

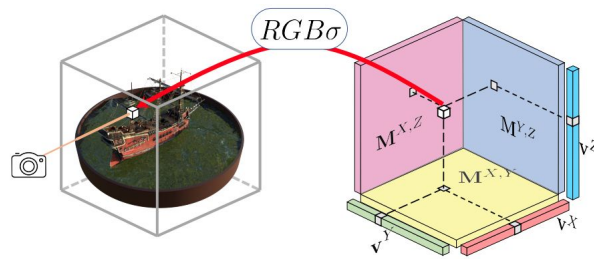
Space-Time Planar NeRF



K-Planes

Speed and Quality Advancements

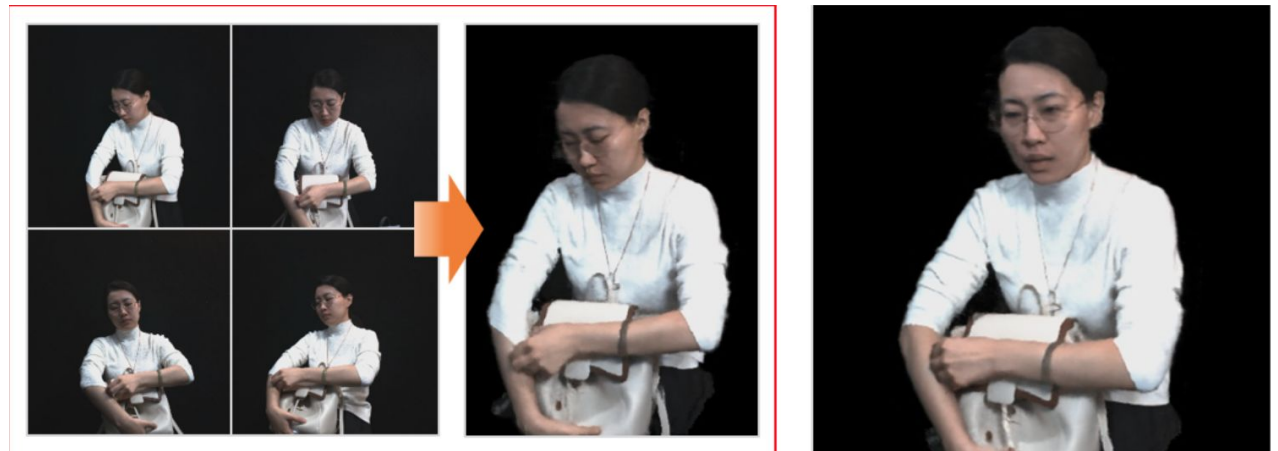
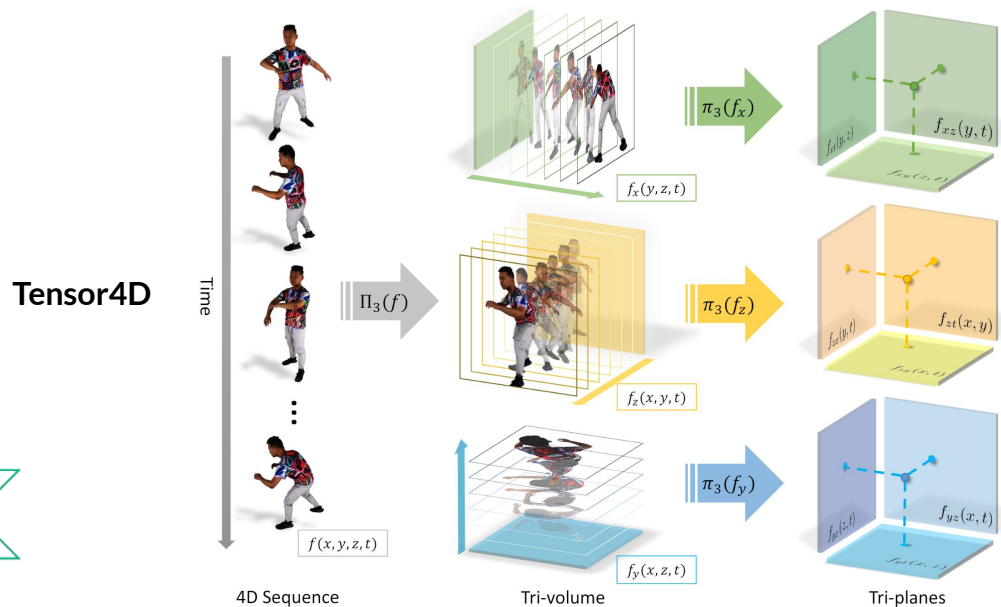
Faster Training and Rendering: Hybrid Neural Scene Representations



Planar Factorization (TensorRF)

Space-Time Planar NeRF

- Coarse-to-fine hierarchical decomposition policy
- Results in 9 planes which can model finer details
- High-fidelity reconstruction from sparse multi-views



Speed and Quality Advancements

High-quality Rendering: Hybrid Neural Scene Representations

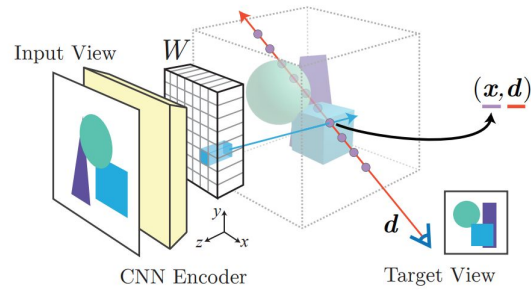


Image-based (pixelNeRF)

Speed and Quality Advancements

High-quality Rendering: Hybrid Neural Scene Representations

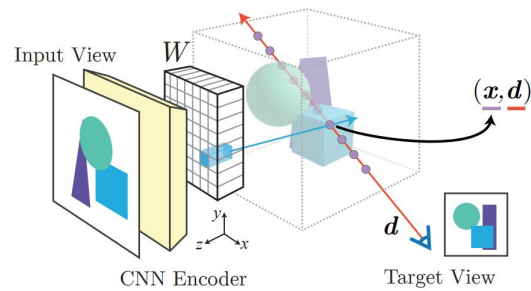
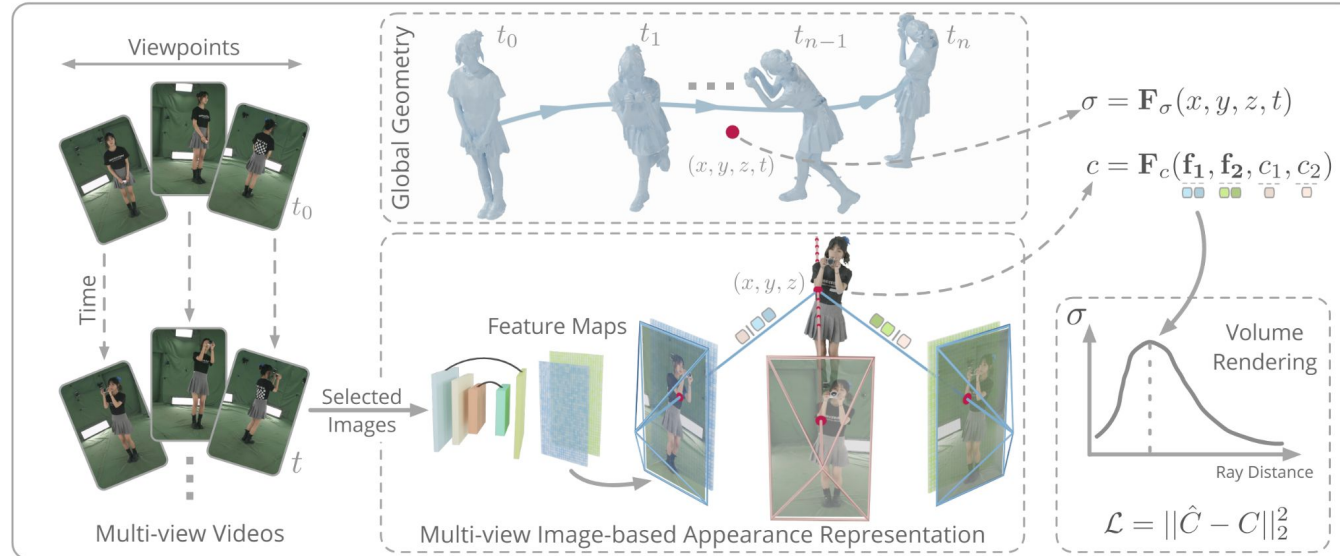


Image-based (pixelNeRF)

Image-based Dynamic NeRF

- Image features for fine appearance details
- Aggregated from multiple views
- High-resolution rendering possible with high resolution training images, upto 4K!



Im4D



4K4D

Speed and Quality Advancements

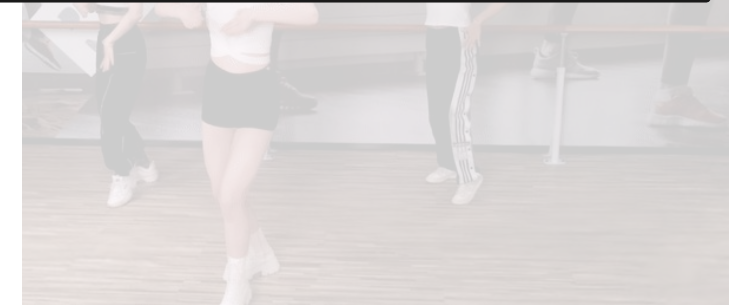
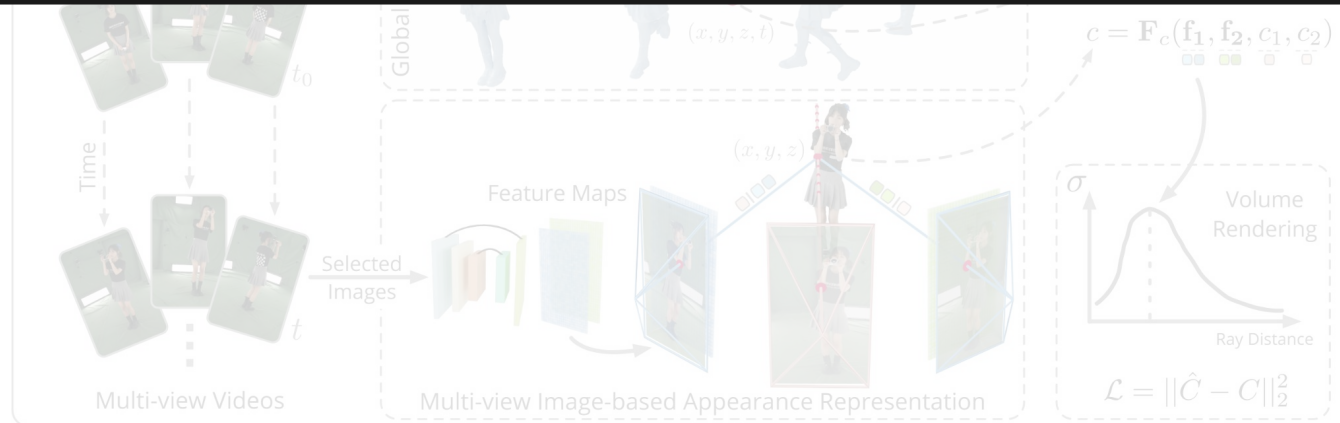
High-quality Rendering: Hybrid Neural Scene Representations

Image-based Dynamic NeRF

Discriminative image features for high-quality

Planar NeRF for

- Speed-up by disentanglement and localization of neural parameters in discrete structures
- Reconstruction time down from hours to minutes
- Fast rendering times
- Higher resolution renders possible with fine appearance details



4K4D

Im4D

Speed and Quality Advancements

High-quality Rendering: Hybrid Neural Scene Representations

Image-based Dynamic NeRF

Discriminative image features for high-quality

Planar NeRF for

- Speed-up by disentanglement and localization of neural parameters in discrete structures
- Reconstruction time down from hours to minutes
- Fast rendering times
- Higher resolution renders possible with fine appearance details

Rendering is fast for hybrid representations but seldom real-time, which brings us to the next major breakthrough!

4K4D

Multi-view Videos

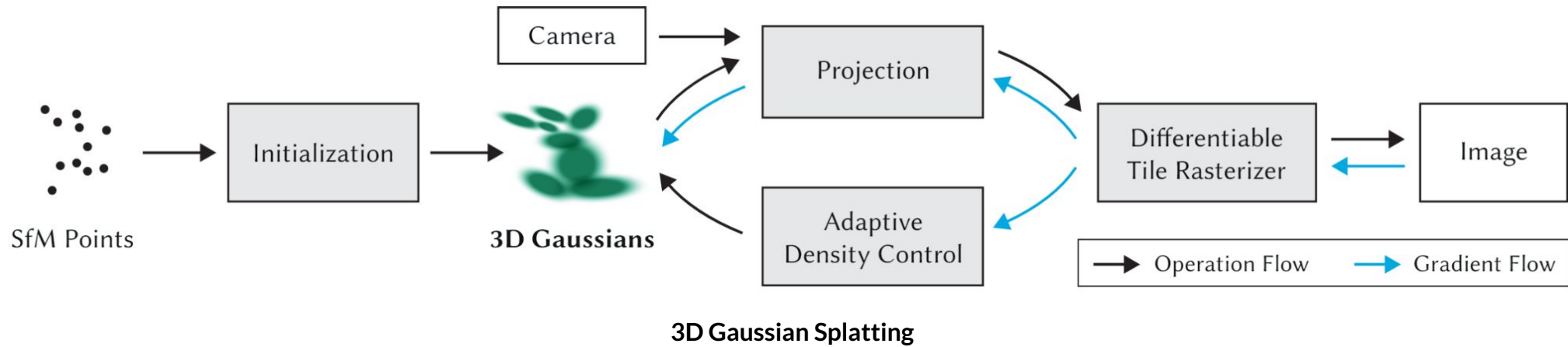
Multi-view Image-based Appearance Representation

$$\mathcal{L} = \|\hat{C} - C\|_2^2$$

Im4D

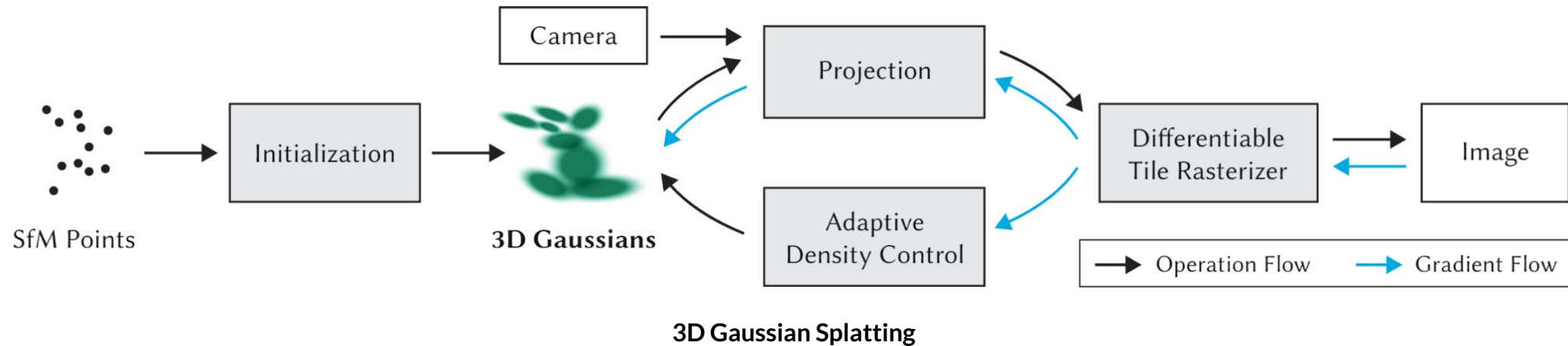
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



Speed and Quality Advancements

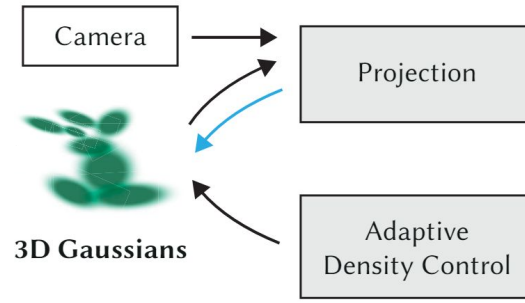
Real-time Rendering: 3D Gaussian Splatting



- Each 3D Gaussian stores position, rotation, scale and spherical harmonics coefficients, which are optimized from images using a fast tile-based rasterizer
- Much faster than volume rendering, enabling real-time performance

Speed and Quality Advancements

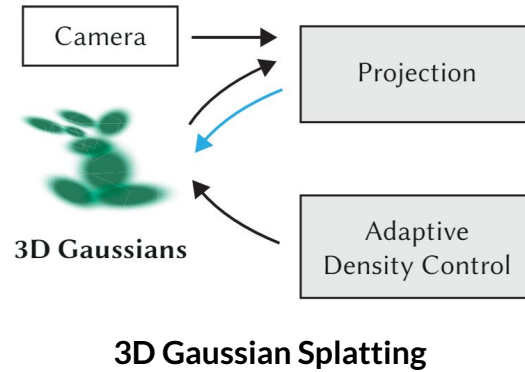
Real-time Rendering: 3D Gaussian Splatting



3D Gaussian Splatting

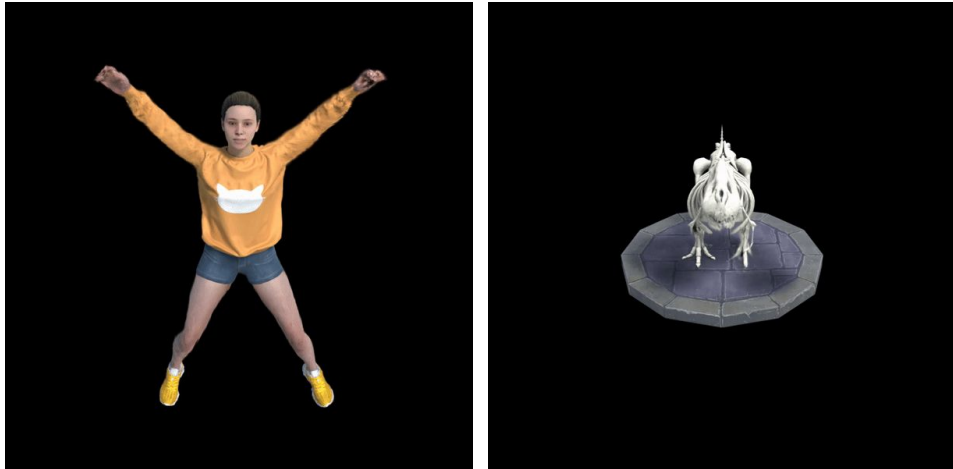
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting

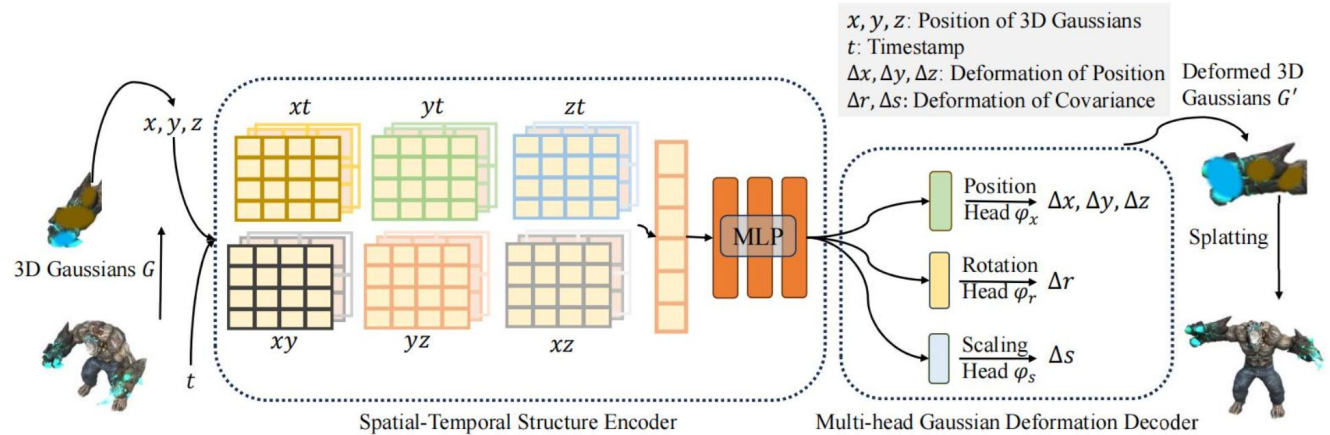


Deformable 3D Gaussian Splatting

- Deform position, rotation and scale of canonical Gaussians to fit each time-step
- Upto 80 FPS rendering speed



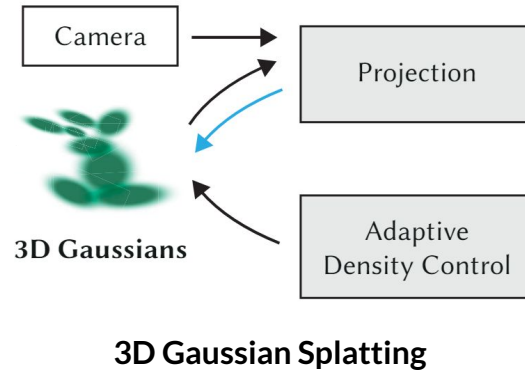
Deformable3DGS



4D-GS

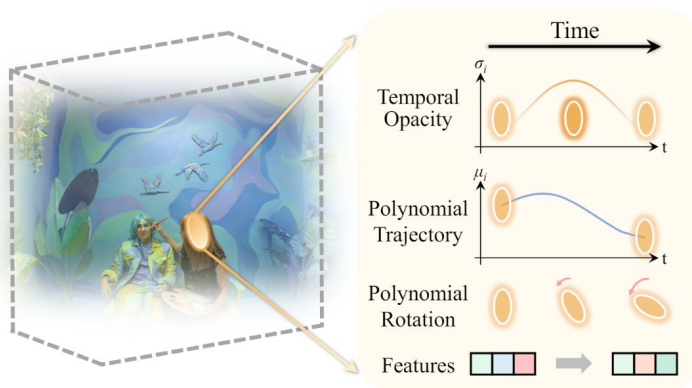
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting

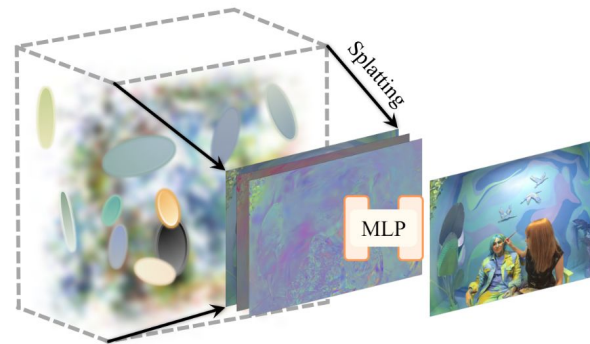


Space-time 3D Gaussian Splatting

- Extra 1D Gaussian added to 3D Gaussians
- Features instead of SHs, with an MLP to convert them into an RGB image after splatting
- 8K video rendering at 66 FPS!



(a) Spacetime Gaussians

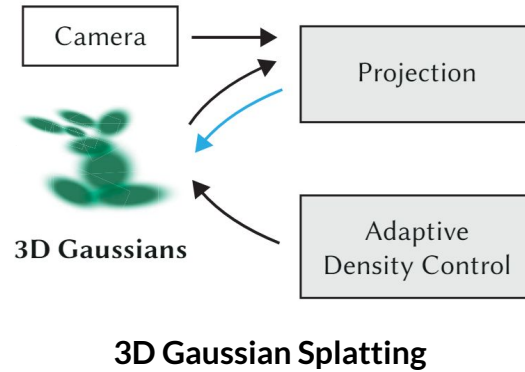


(b) Feature Splatting and Rendering



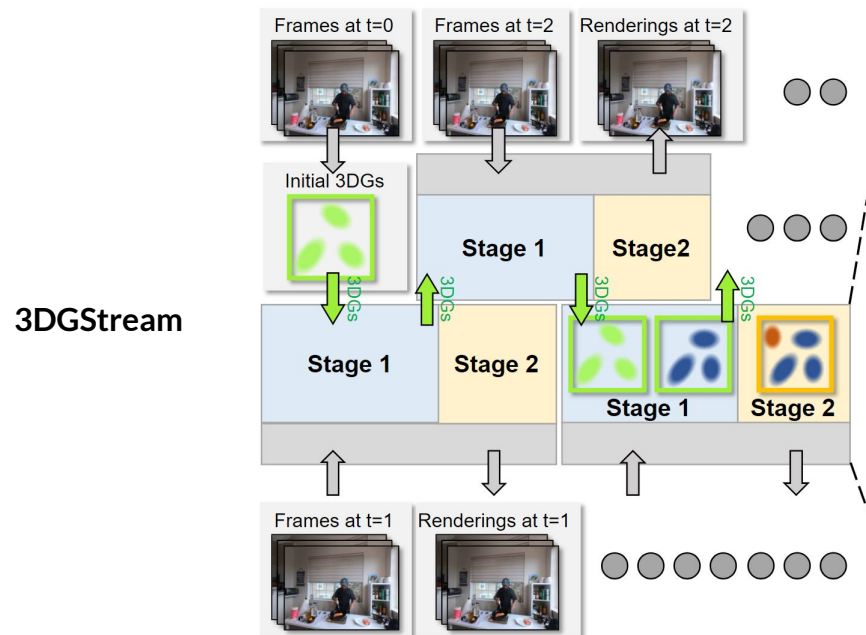
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



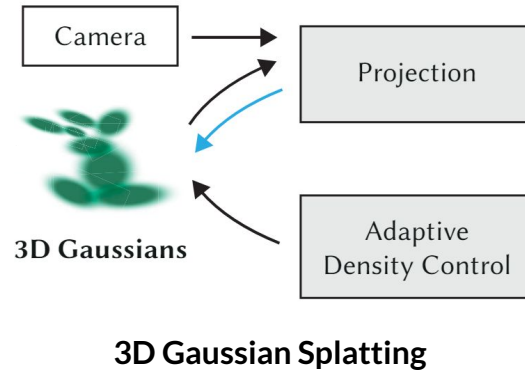
Streamable 3D Gaussian Splatting

- Online Reconstruction from multi-view videos
- Multi-resolution neural hash-grid as a cache for transformation
- Additional frame-specific Gaussian spawning



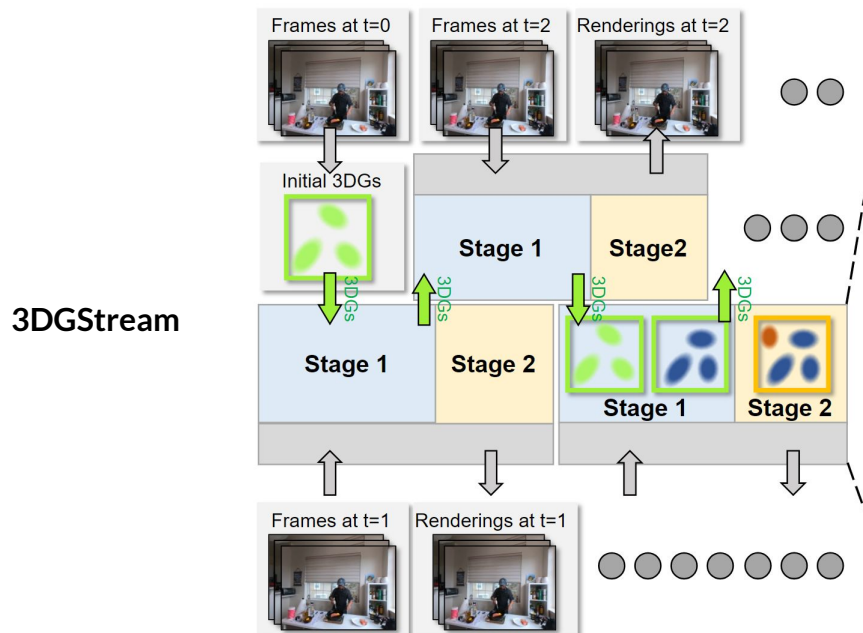
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



Streamable 3D Gaussian Splatting

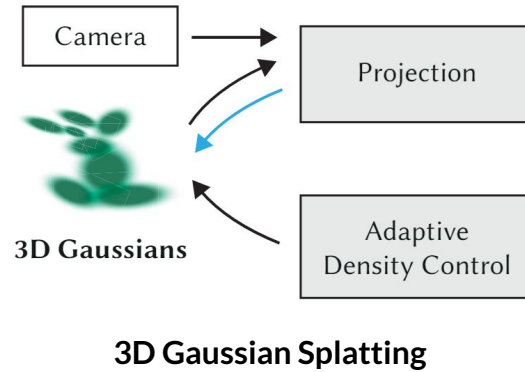
- Online Reconstruction from multi-view videos
- Multi-resolution neural hash-grid as a cache for transformation
- Additional frame-specific Gaussian spawning



Reconstruction in 12 seconds with up to 200 FPS rendering speed!

Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting

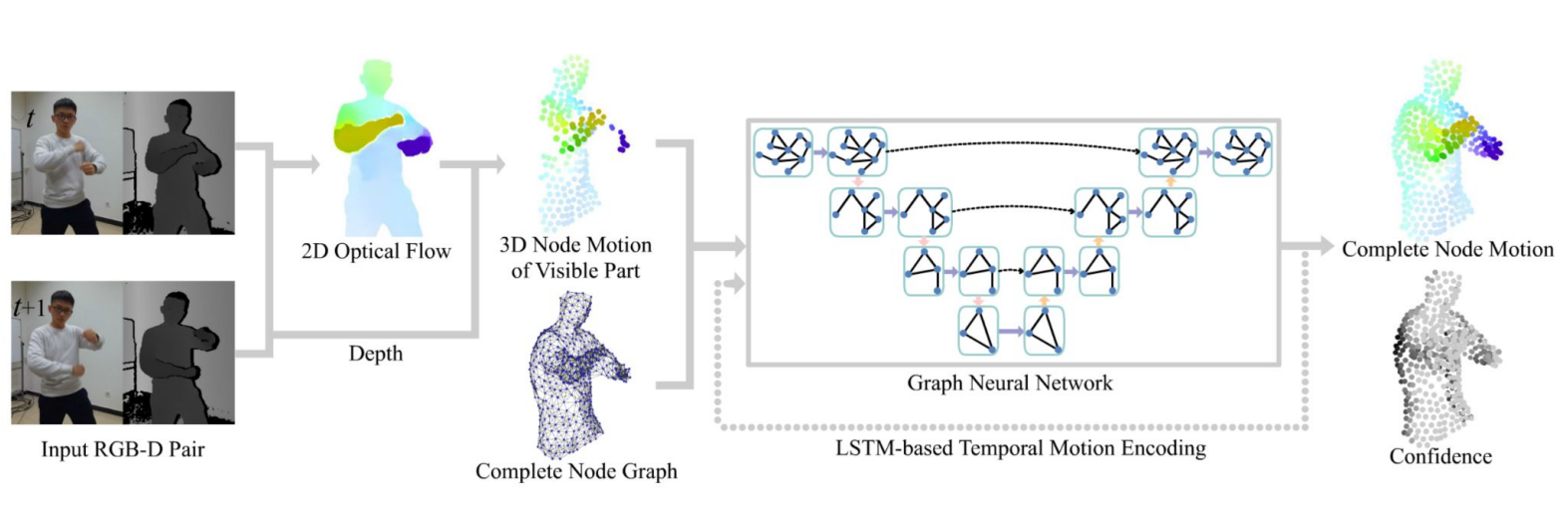


Even though it is getting close, real-time reconstruction is still only possible using classical representations

Speed and Quality Advancements

Real-time Reconstruction: Classical Representations

- Registers RGB-D frames into a canonical TSDF grid
- Uses a mesh-based deformation graph to track deformation of canonical frame to each timestep
- Pre-trains a GNN to predict motion of occluded regions from the visible motion
- Geometry only!



Occlusion Fusion

Live Demo



Trends

1. Speed and Quality Advancements

2. Handling of Large Deformation and Long-term 3D correspondences

5 minute break!

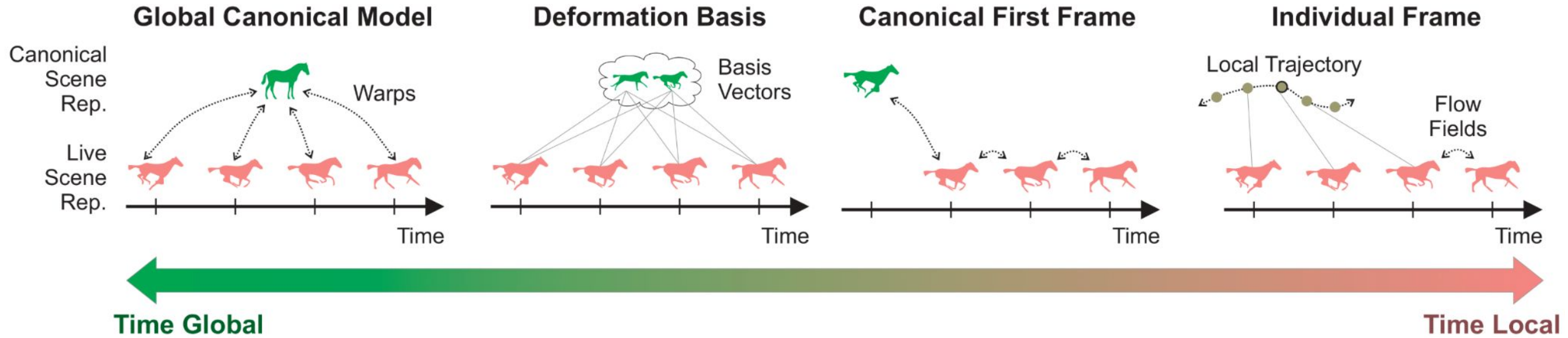
3. Modelling Articulated Motion for General Objects

Trends

1. Speed and Quality Advancements
2. Handling of Large Deformations / Long-Term 3D Correspondences
3. Modelling Articulated Motion for General Objects

Large Motion vs. Long-term 3D Correspondences

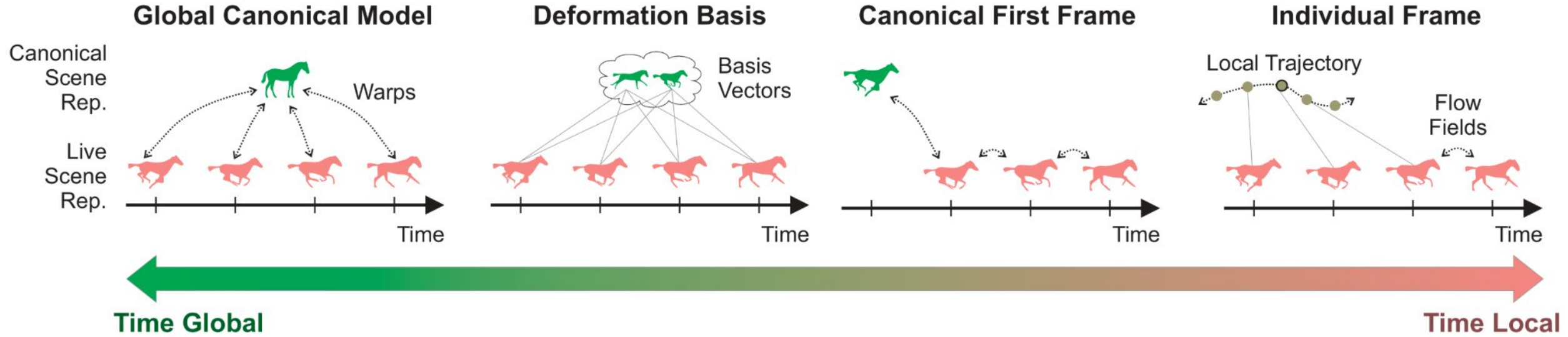
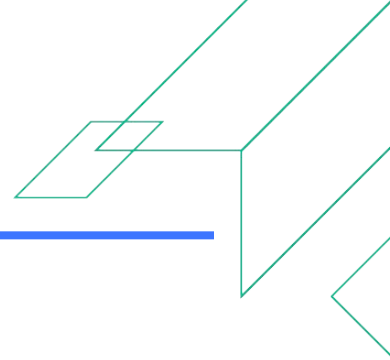
Spatio-Temporal Modelling



Design choice determines the trade-off between time consistency and large motion modelling!

Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

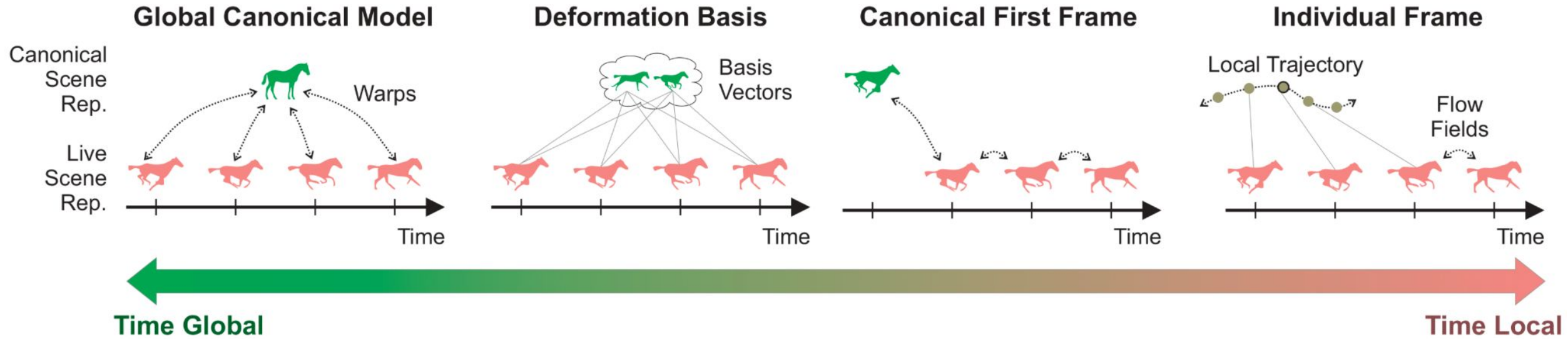
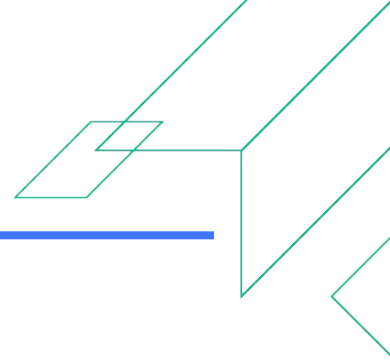


Time consistency enables applications like 3D editing and virtual asset creation



Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

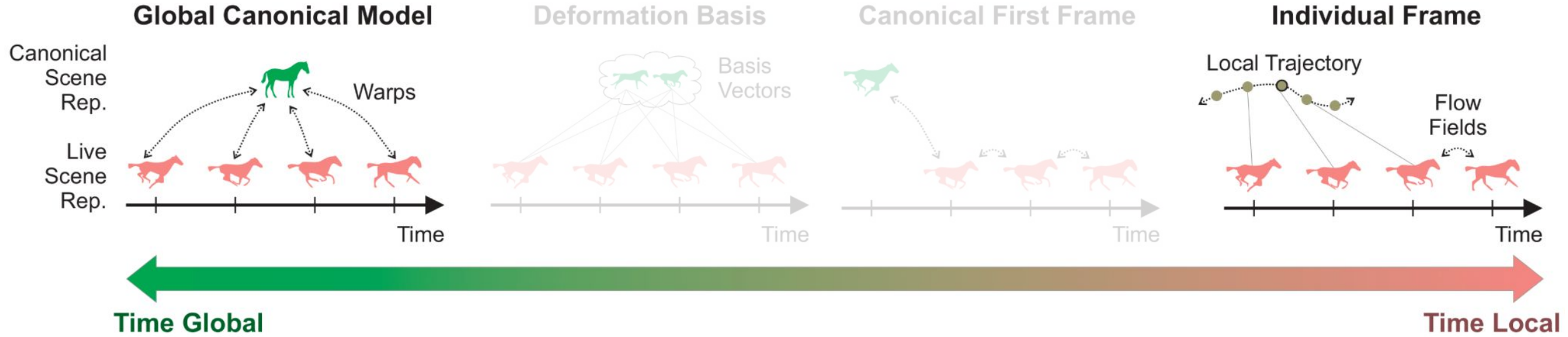
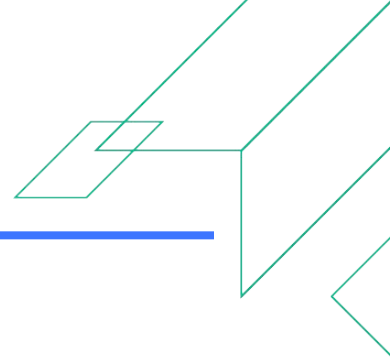


Time consistency enables applications like 3D editing and virtual asset creation

But we don't want to compromise on motion modelling

Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



- Global

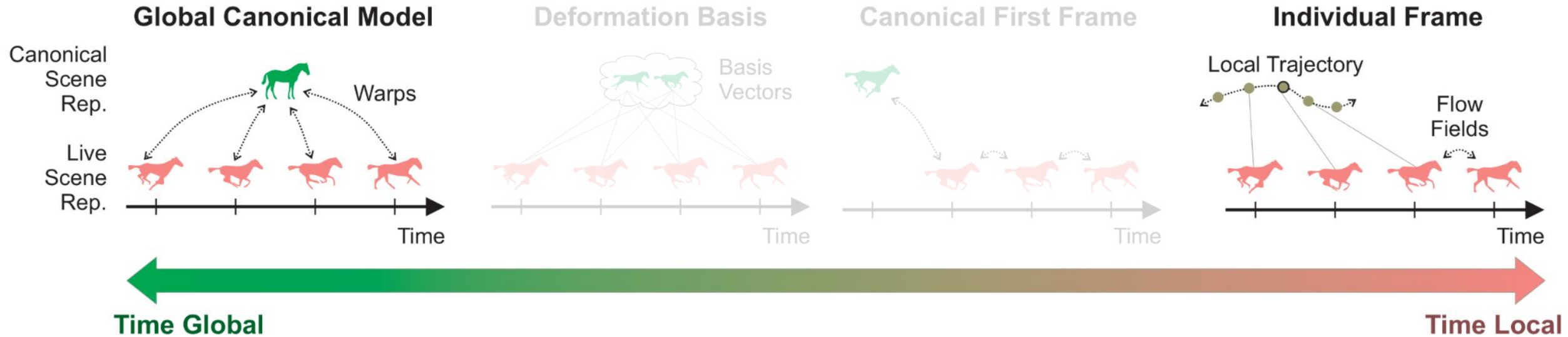
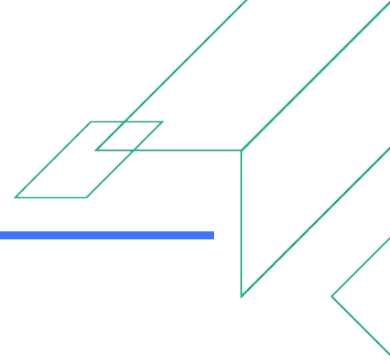
Temporal Correspondences

- Optional and local



Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



- Global
- Restricted

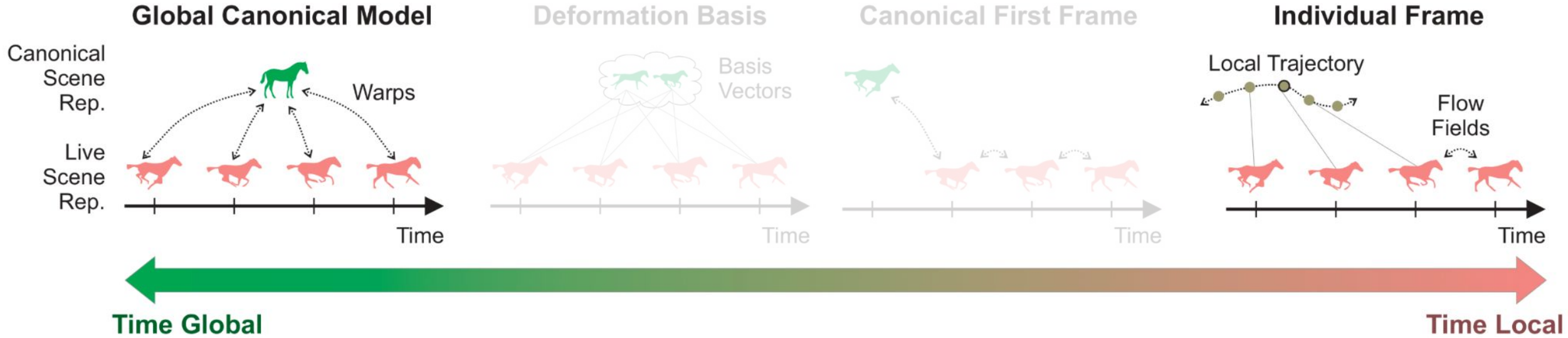
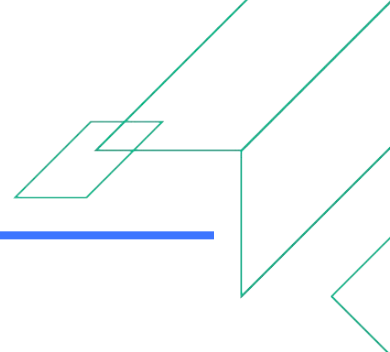
Temporal Correspondences
Motion Modelling

- Optional and local
- Empirically larger



Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



- Global
- Restricted
- Difficult; cannot handle discontinuities

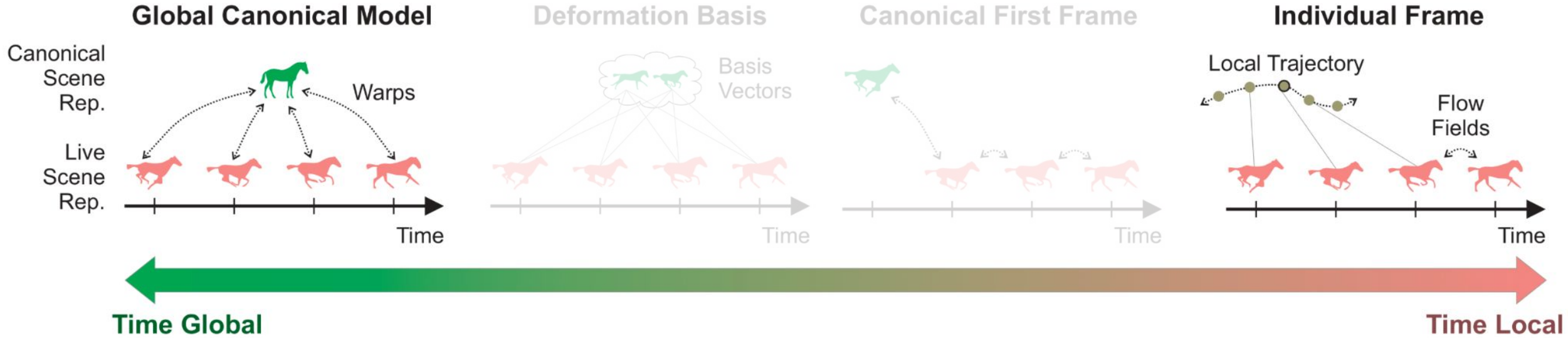
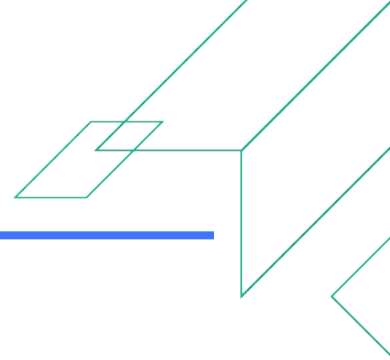
Temporal Correspondences
Motion Modelling
Topology Changes

- Optional and local
- Empirically larger
- Straightforward



Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



- Global
- Restricted
- Difficult; cannot handle discontinuities
- Difficult; canonical model is time-independent

Temporal Correspondences

Motion Modelling

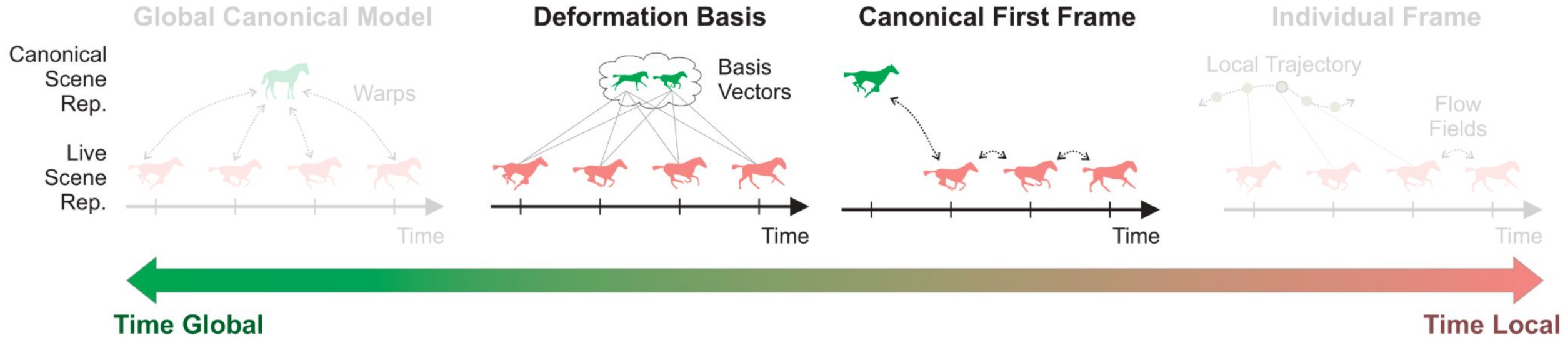
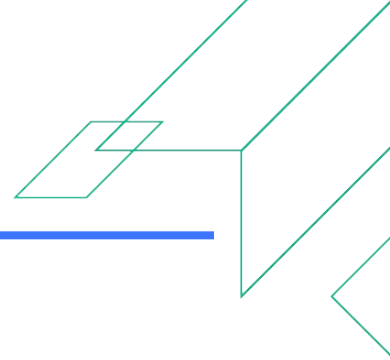
Topology Changes

Appearance Changes

- Optional and local
- Empirically larger
- Straightforward
- Straightforward

Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

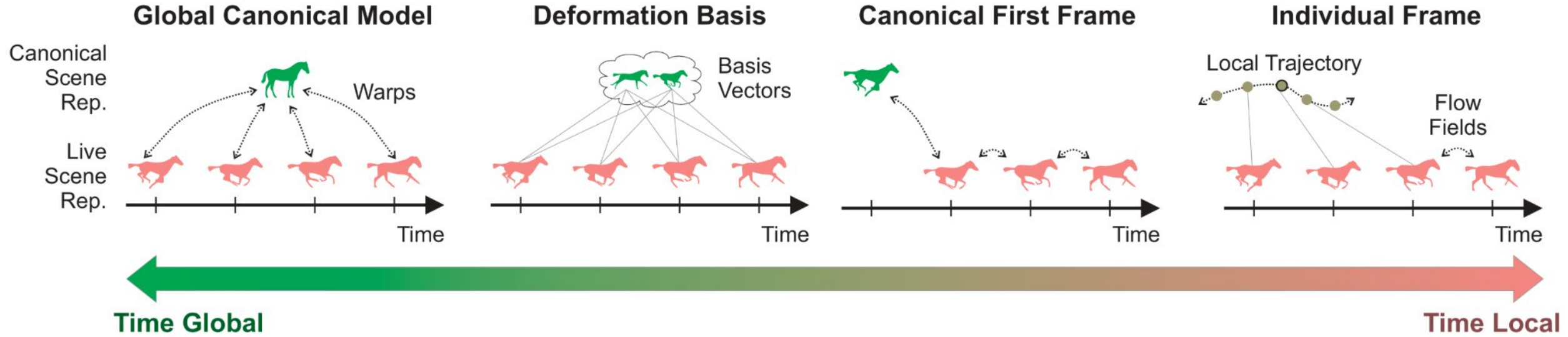
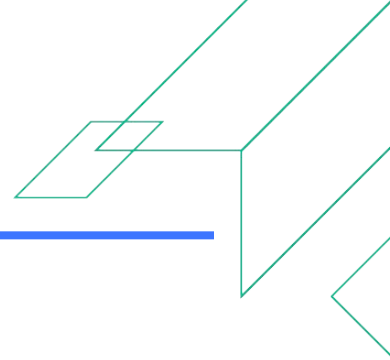


Trade-offs to balance the best of both worlds!



Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



Let's look at some improvements for each type of modelling in recent years

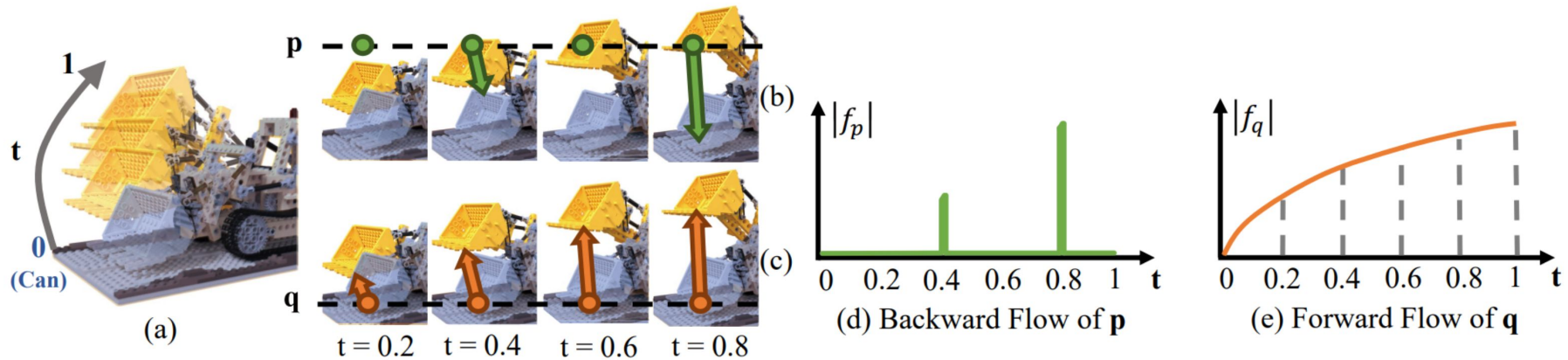


Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Forward Flow Modelling:

- Deformations are modelled from canonical to live frame for smooth and continuous motion model learning
- Enabled by a voxel-based canonical field for discrete forward warping
- Give point trajectories for each point in canonical space



ForwardFlowDNeRF

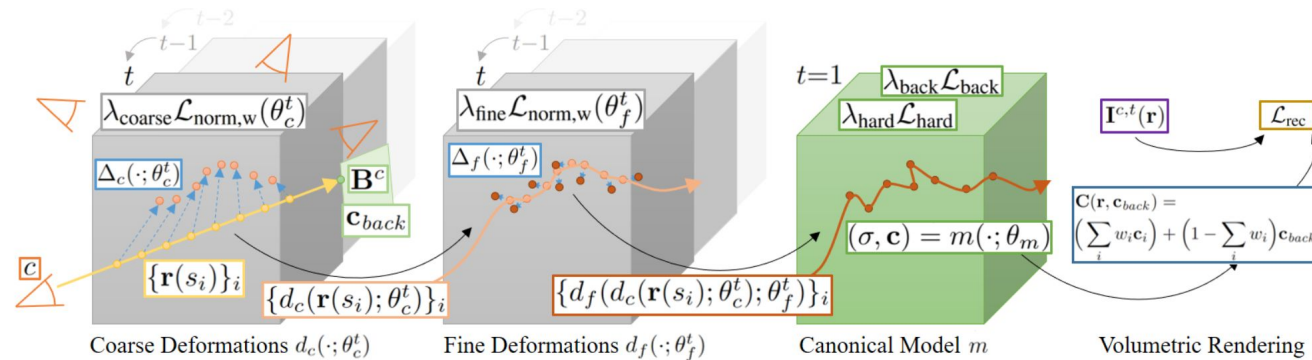


Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Time-Consistent Canonical Modelling:

- Builds a canonical model from the first frame of multiview videos and fixes it
- Online reconstruction of next timesteps
- Hard constraint on time-consistency of canonical model, thus improving temporal correspondences while handling large motion through coarse-to-fine deformations



ScenerFlow

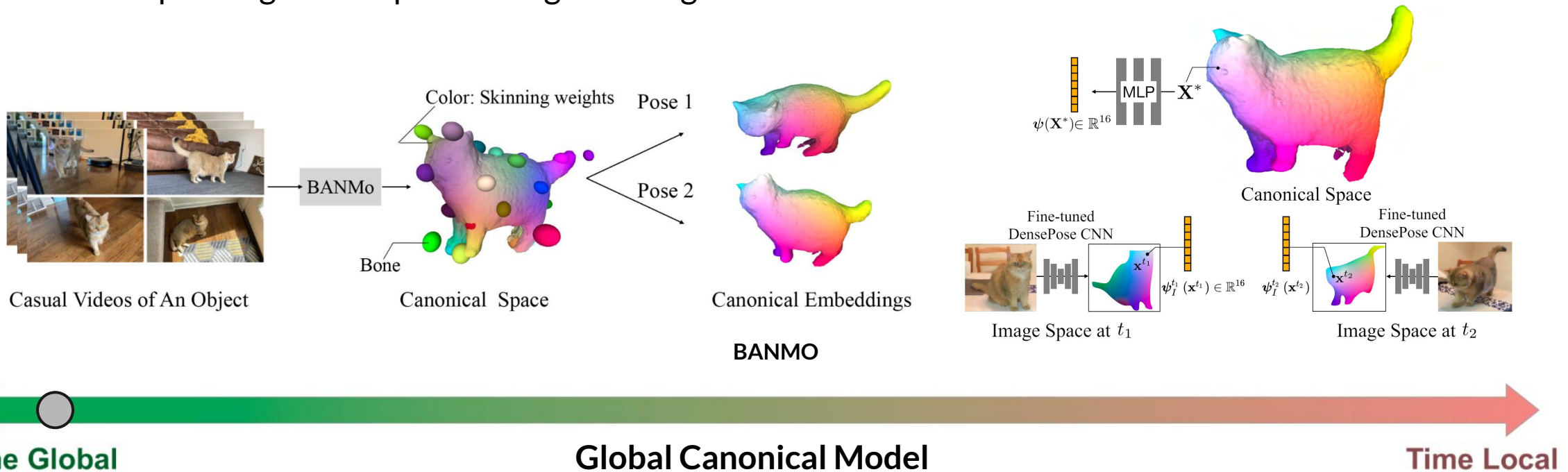


Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Canonical Feature Embeddings:

- Shares canonical space over multiple videos of an object
- 2D DensePose features are distilled into the 3D canonical model as embeddings
- Enforcement of 3D canonical embeddings to match 2D DensePose features in each corresponding view improves long-term registration

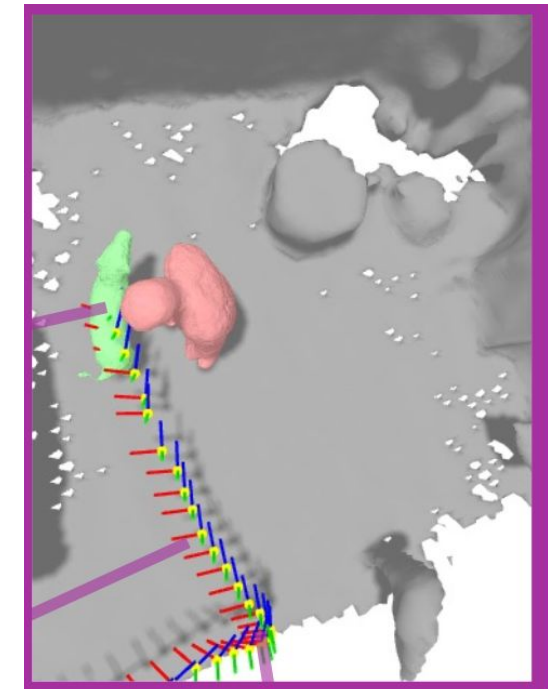
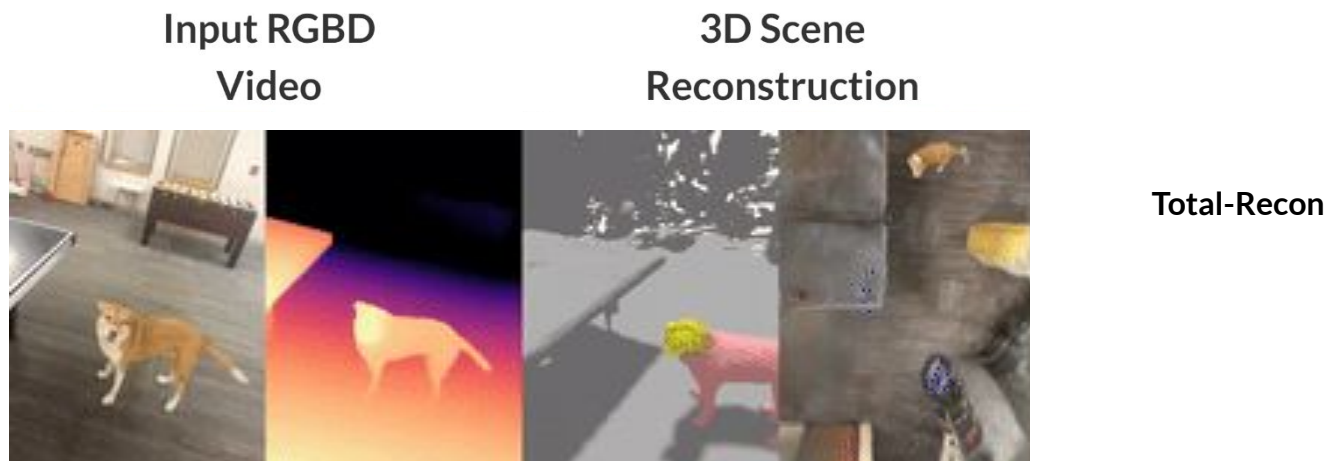


Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Decomposed Motion Modelling:

- Decomposes object motion into root pose and residual motion
- Simpler motion modelling allows it to scale to longer scenes
- Takes RGB-D input and needs root-pose initialization with PoseNet

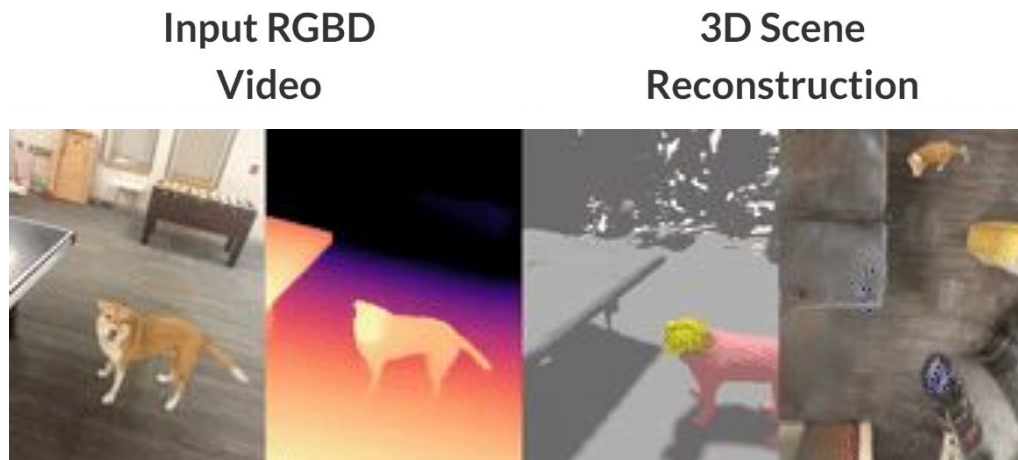


Large Motion vs. Long-term 3D Correspondences

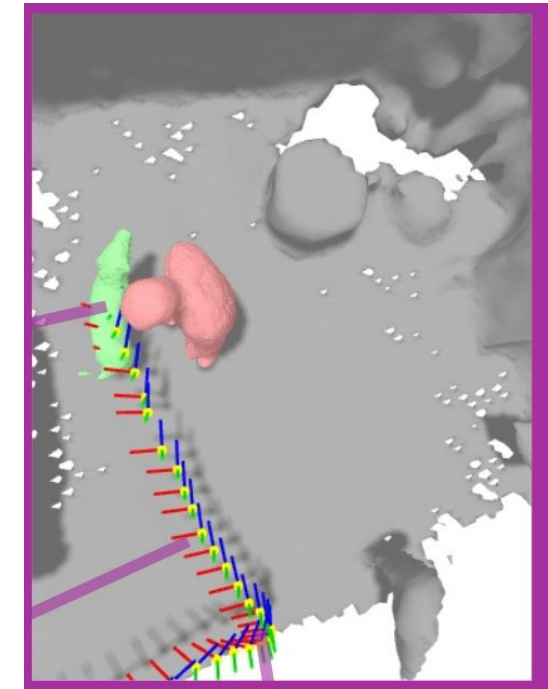
Spatio-Temporal Modelling

Decomposed Motion Modelling:

- Decomposes object motion into root pose and residual motion
- Simpler motion modelling allows it to scale to longer scenes
- Takes RGB-D input and needs root-pose initialization with PoseNet



Total-Recon



Scales up to minute-long RGB-D videos with large motion!

Time Global

Global Canonical Model

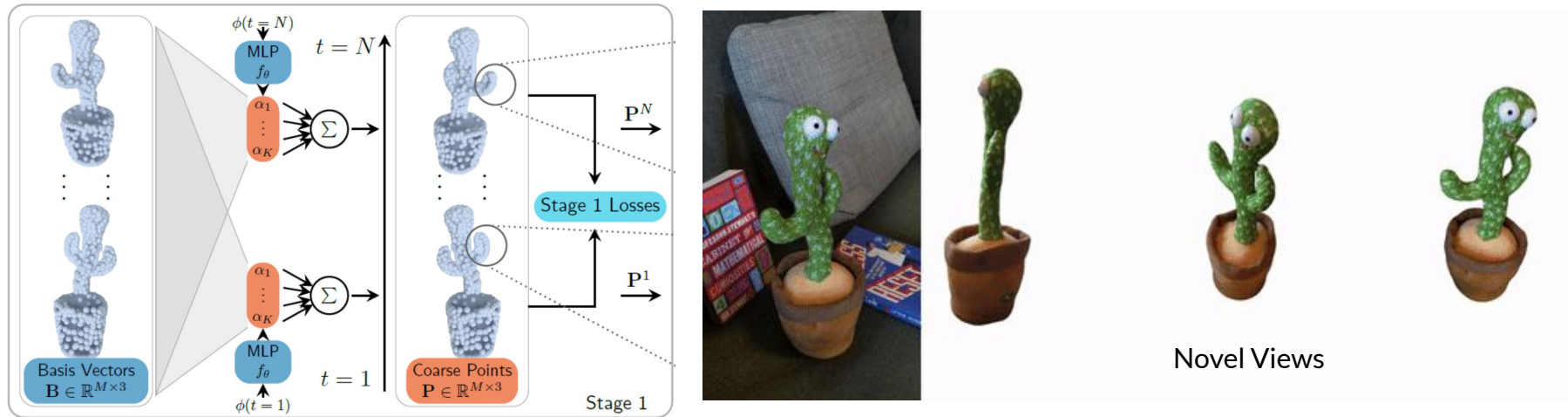
Time Local

Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Low-rank Deformation Template:

- Shared point template for each frame, automatically giving temporal correspondences
- Generated by low-rank basis, thus forcing information sharing
- Models complex motion while providing regularization for challenging novel views



Neural Parametric Gaussians

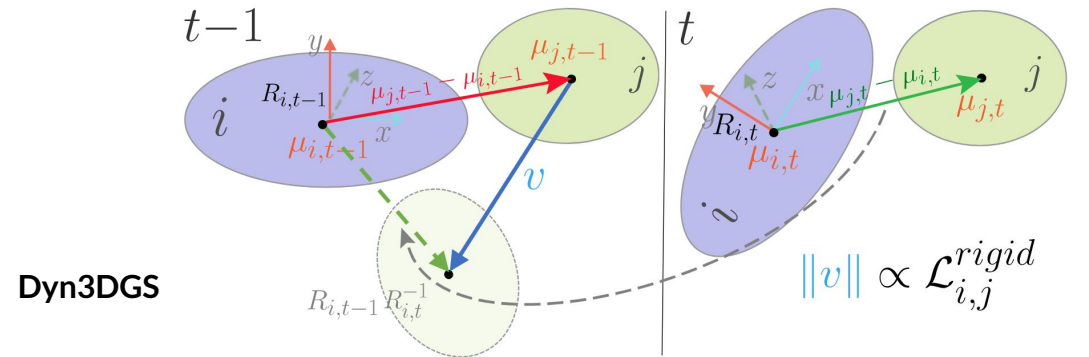


Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Per-frame Canonical Model Optimization:

- Rotation and position of canonical Gaussians are optimized for each timestep from last timestep, giving dense 6-DOF trajectories
- Models long-range motion but trajectories can drift over time
- Multi-view supervision required and surface rigidity losses introduced to tackle this

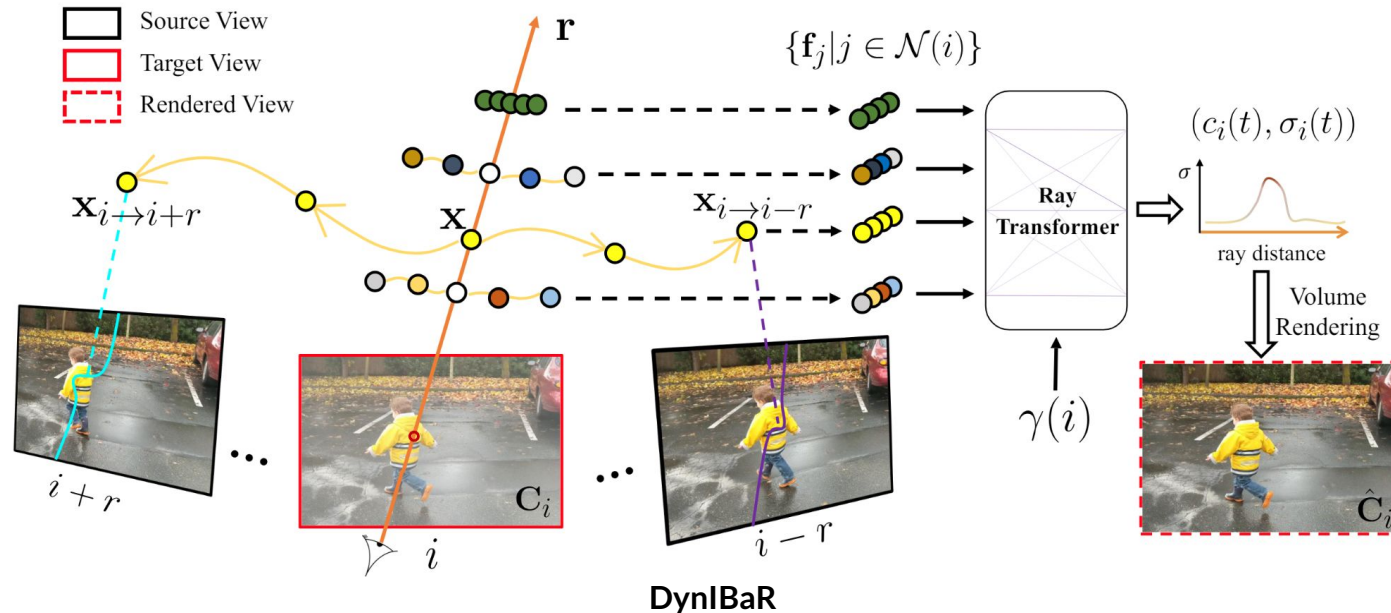


Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Motion Trajectory Modelling:

- Per-frame hybrid representation which takes in image features aggregated over time
- Motion trajectories allows information aggregation from a greater temporal neighbourhood



Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

Motion Trajectory Modelling:

- Per-frame hybrid representation which takes in image features aggregated over time
- Motion trajectories allows information aggregation from a greater temporal neighbourhood
- Improves time consistency while modelling free-form motion



DynIBaR



Trends

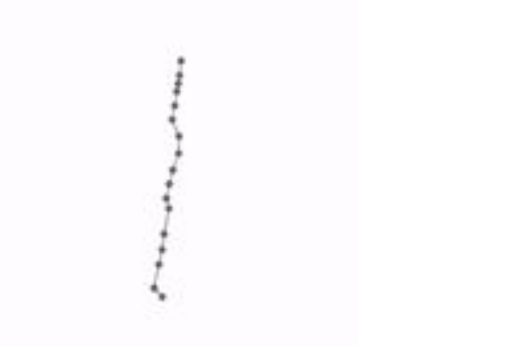
1. Speed and Quality Advancements
- 2. Handling of Large Deformations / Long-Term 3D Correspondences**
3. Modelling Articulated Motion for General Objects

Trends

1. Speed and Quality Advancements
2. Handling of Large Deformations / Long-term 3D correspondences
3. Modelling Articulated Motion for General Objects

Modelling General Articulated Motion

Deformations of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton:



Modelling General Articulated Motion

Deformations of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton:

- Skeletons allow reposing of objects to unseen poses



Modelling General Articulated Motion

Deformations of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton:

- Skeletons allow reposing of objects to unseen poses
- If we know how an object category articulates, we can use that information as prior to estimate motion of new sequences of that object



Modelling General Articulated Motion

Deformations of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton:

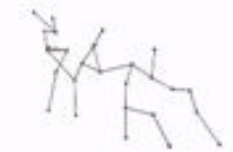
- Skeletons allow reposing of objects to unseen poses
- If we know how an object category articulates, we can use that information as prior to estimate motion of new sequences of that object
- Category-level object templates are available mostly for human categories (e.g. SMPL) which are obtained from expensive 3D data.



Modelling General Articulated Motion

Deformations of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton:

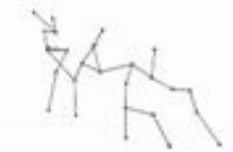
- Skeletons allow reposing of objects to unseen poses
- If we know how an object category articulates, we can use that information as prior to estimate motion of new sequences of that object
- Category-level object templates are available mostly for human categories (e.g. SMPL) which are obtained from expensive 3D data.
- How can we obtain them for general object categories where such data is not available?



Modelling General Articulated Motion

Deformations of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton:

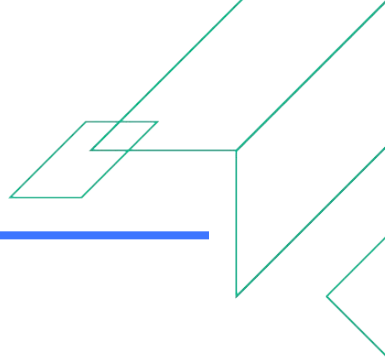
- Skeletons allow reposing of objects to unseen poses
- If we know how an object category articulates, we can use that information as prior to estimate motion of new sequences of that object
- Category-level object templates are available mostly for human categories (e.g. SMPL) which are obtained from expensive 3D data.
- How can we obtain them for general object categories where such data is not available?



From RGB videos!

Modelling General Articulated Motion

Self-Supervised Part Discovery for Reposing

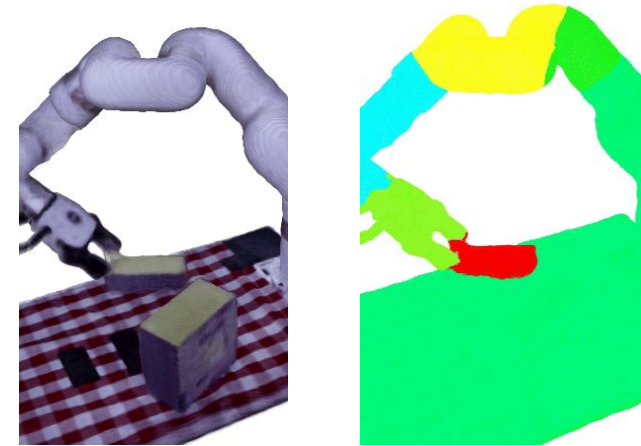
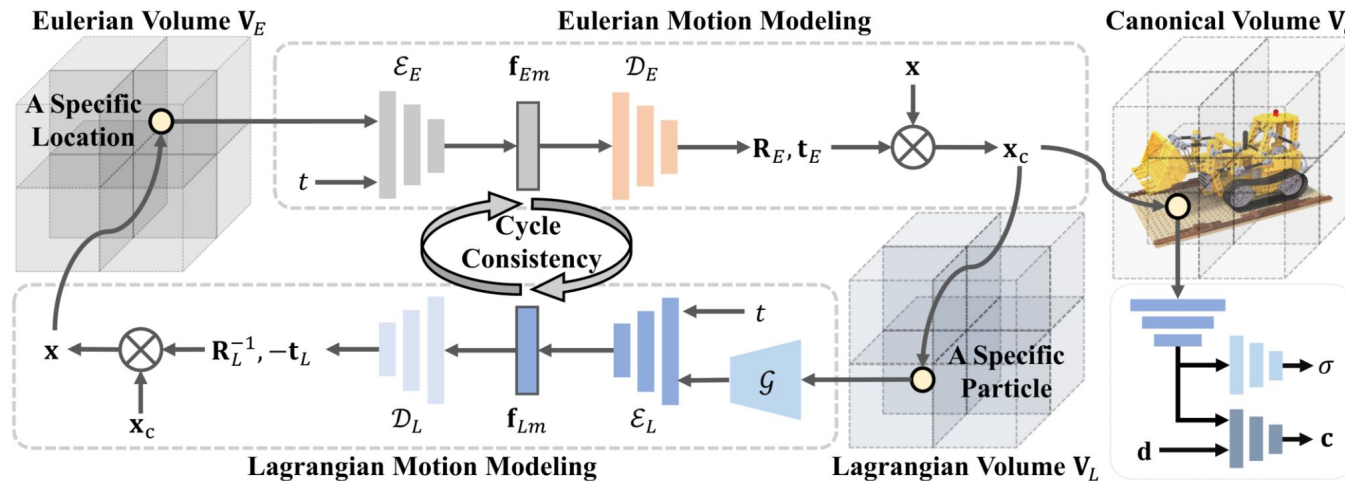


Modelling General Articulated Motion

Self-Supervised Part Discovery for Reposing

Motion-based grouping:

- Models both backward and forward motion with feature grids
- Features from forward motion are grouped into slots using an attention mechanism
- Similar motion \Rightarrow same slot \Rightarrow same part
- Discovered parts can be skeletonized and reposed



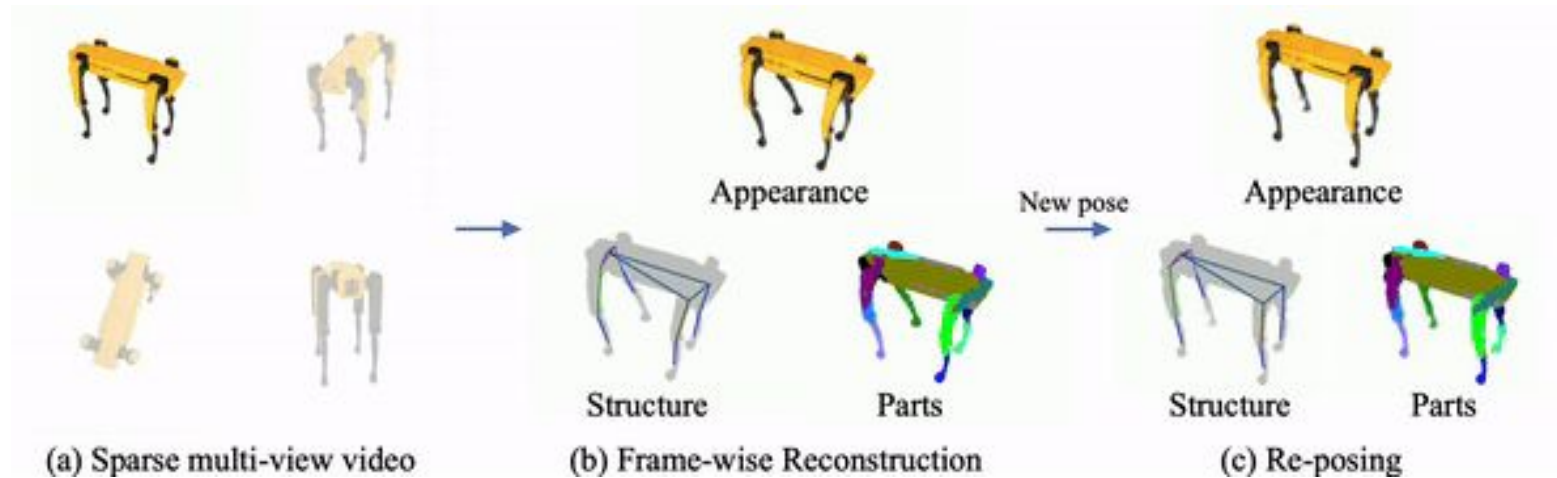
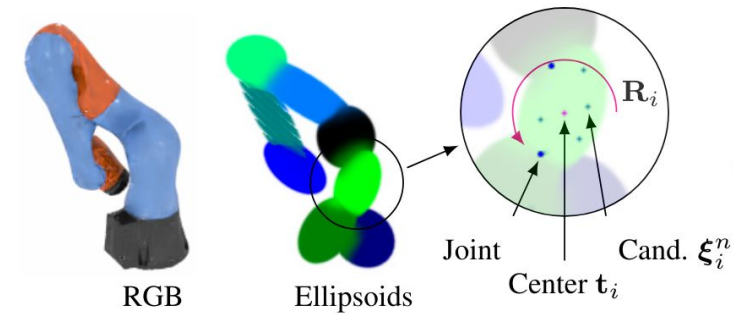
MovingParts

Modelling General Articulated Motion

Self-Supervised Part Discovery for Reposing

Unsupervised Part Prediction:

- Represent parts by ellipsoids in 3D
- Each ellipsoid has a rotation and a position
- Optimize the per-frame ellipsoids prediction MLP from multi-view videos
- Repose using discovered ellipsoids



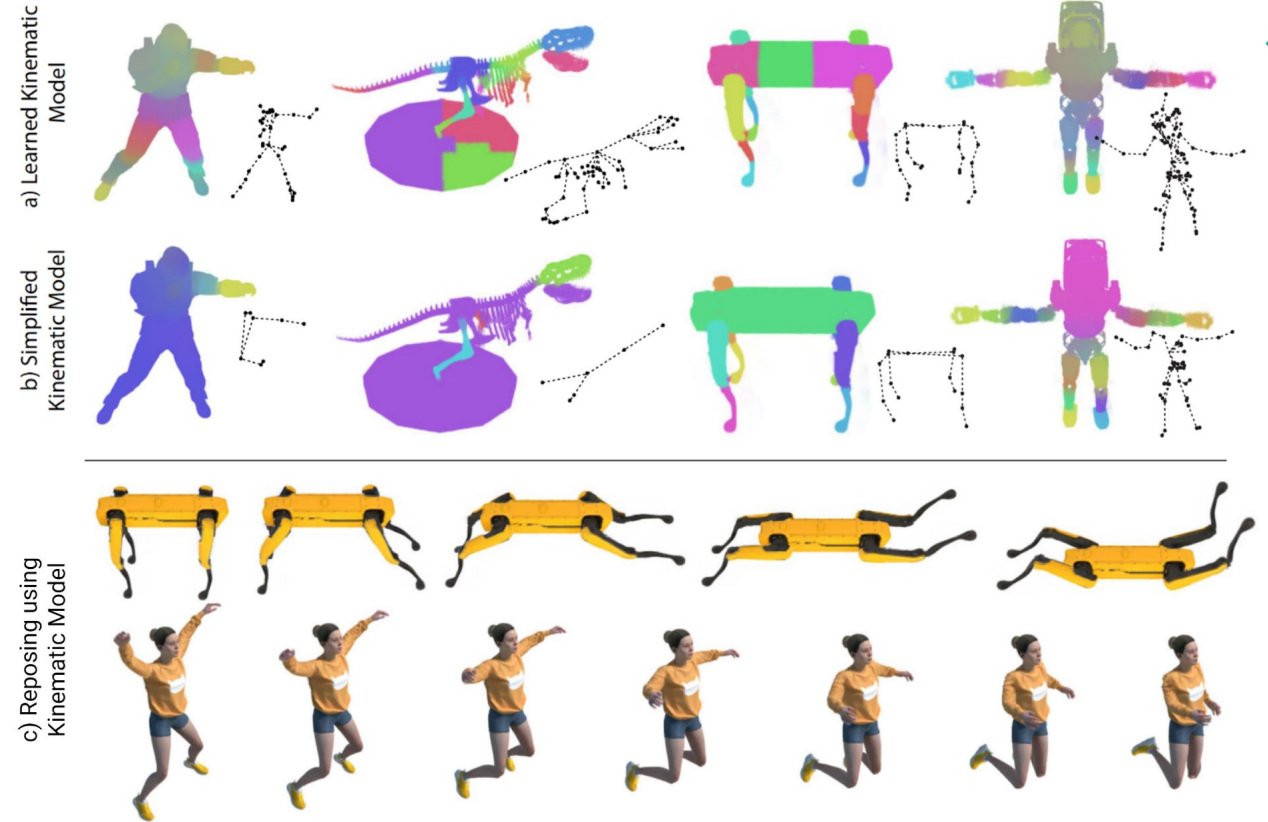
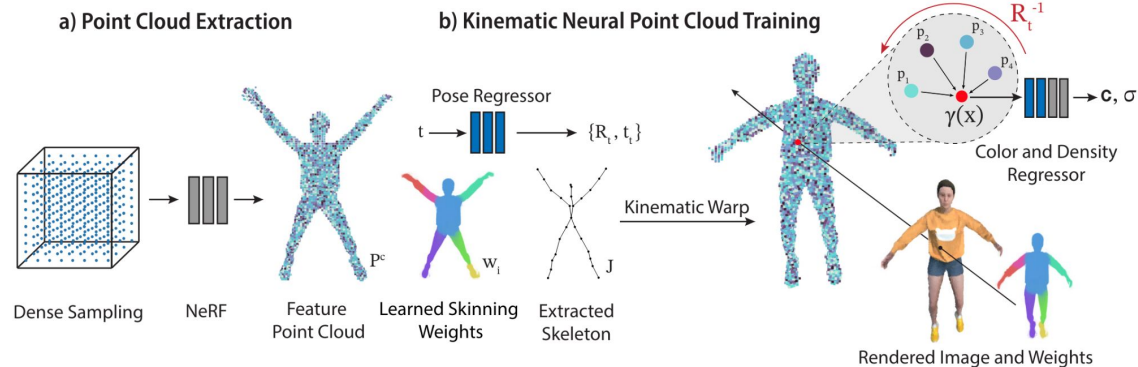
Watch-it-Move

Modelling General Articulated Motion

Skeleton Discovery for Reposing

Morphological Operations:

- Point-based canonical representation extracted from a dynamic NeRF backbone
- Medial Axis Transform used to extract skeleton from canonical points
- Linear blend skinning-based model to learn forward dynamics from observations
- Repose using the learnt template
- Also fast because of the point-based hybrid representation



Uzolas et al.

Modelling General Articulated Motion

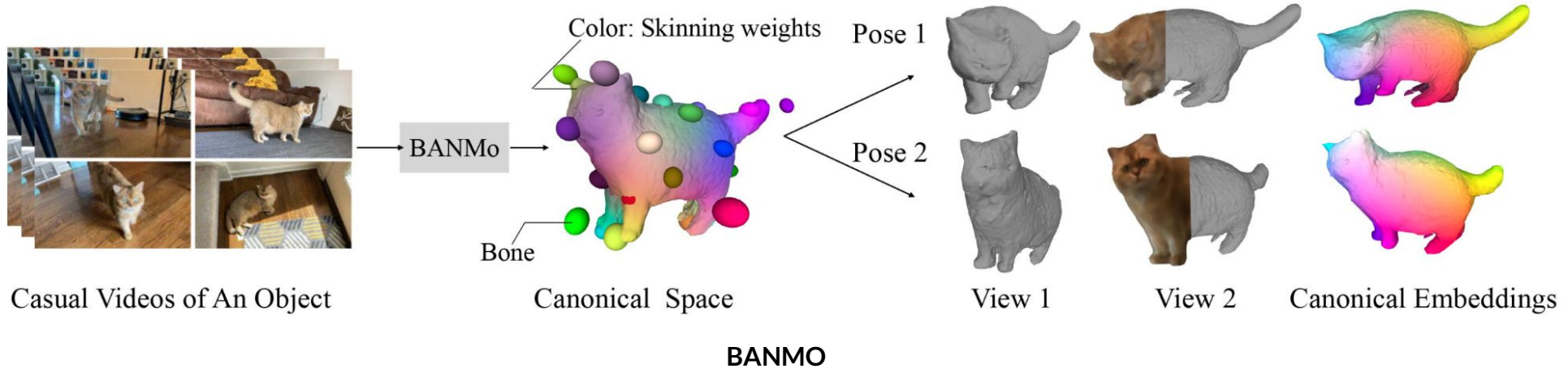
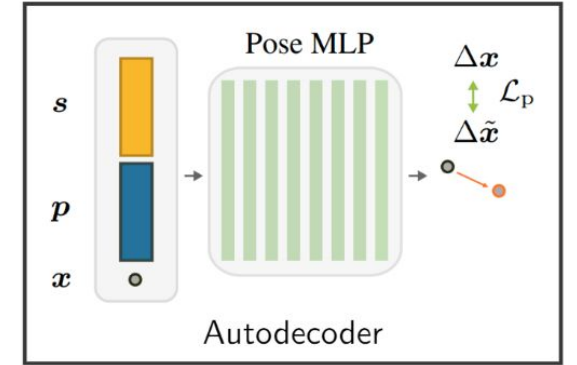
- Previous methods are trained on a single video sequence
- Can we utilize multiple videos of the same object to build an instance-level model?

Modelling General Articulated Motion

Modelling Articulations with Neural Bones

- Canonical space is shared between videos
- Bone positions and transforms are estimated per-frame using an auto-decoded MLP
- Articulated using volumetric skinning

Model captures the articulations across videos, providing better regularization



Modelling General Articulated Motion

Modelling Articulations with Neural Bones

- Use optimized pose embeddings from a driving video for another structurally similar geometry model for motion retargeting!



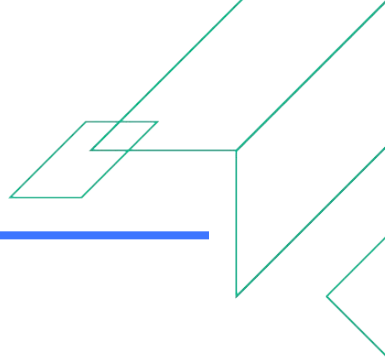
Driving Sequence



Target Geometry

BANMO

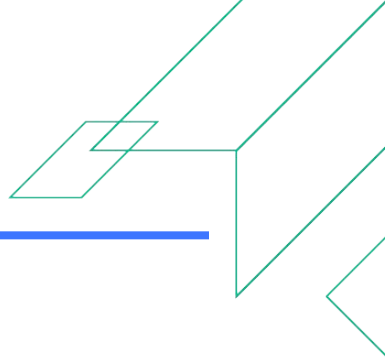
Modelling General Articulated Motion



- A category of objects articulates in the same way (e.g. different breeds of cats)
- Can we learn category-level templates from videos to regularize motion even further and use it as a prior for instances?



Modelling General Articulated Motion



- A category of objects articulates in the same way (e.g. different breeds of cats)
- Can we learn category-level templates from videos to regularize motion even further and use it as a prior for instances?

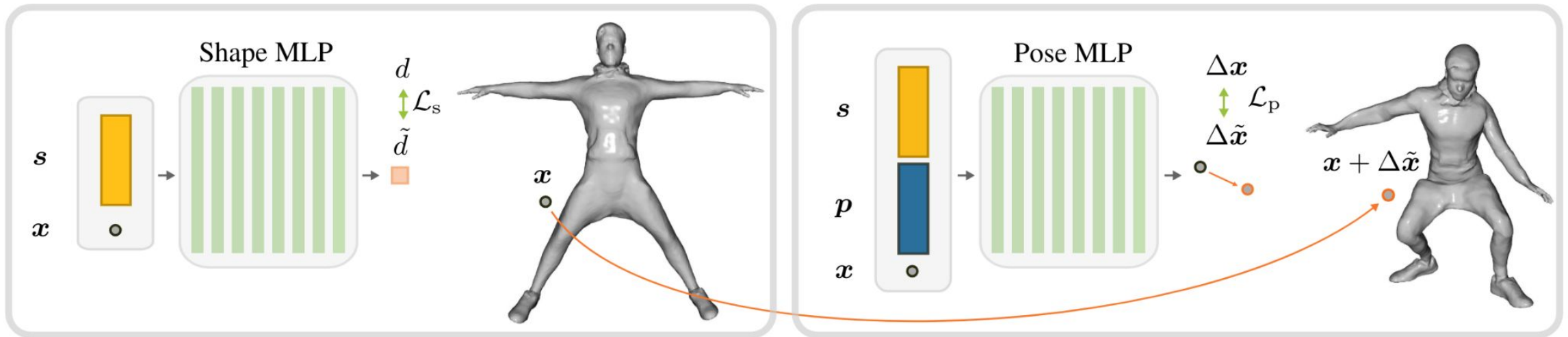
Yes, but we need to capture shape variations between category instances as well!



Modelling General Articulated Motion

Category-level Modelling from Depth Videos

- Use auto-decoders to model both shape and pose variations
- Shape embeddings can capture category-level variations while pose embeddings capture instance articulations
- Optimize shape and then pose at test-time



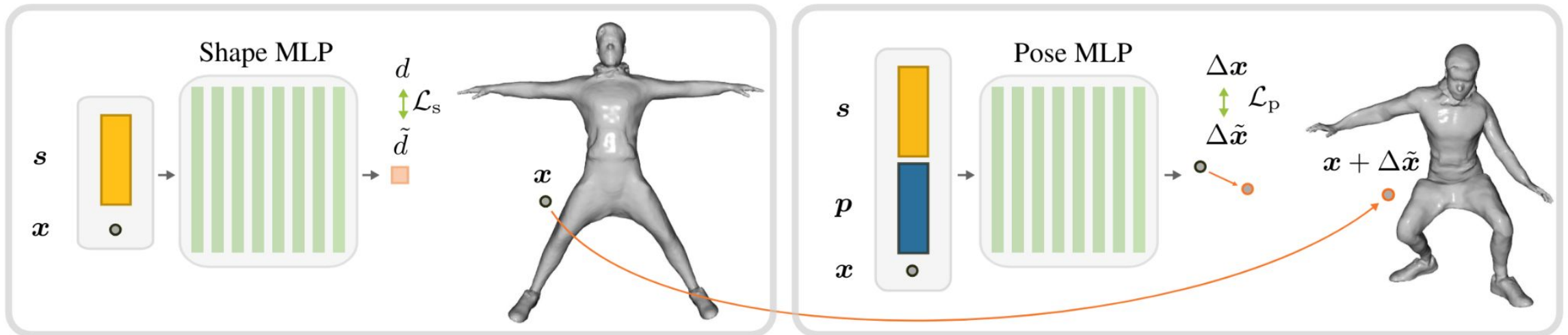
NPMs

Modelling General Articulated Motion

Category-level Modelling from Depth Videos

- Use auto-decoders to model both shape and pose variations
- Shape embeddings can capture category-level variations while pose embeddings capture instance articulations
- Optimize shape and then pose at test-time

Learned from depth sequences.
Can we do it from RGB videos?

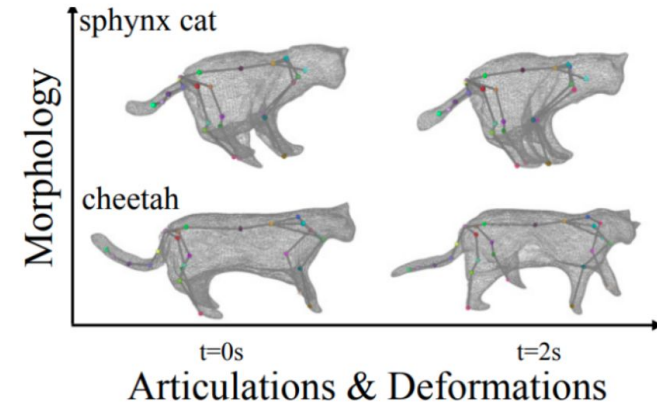
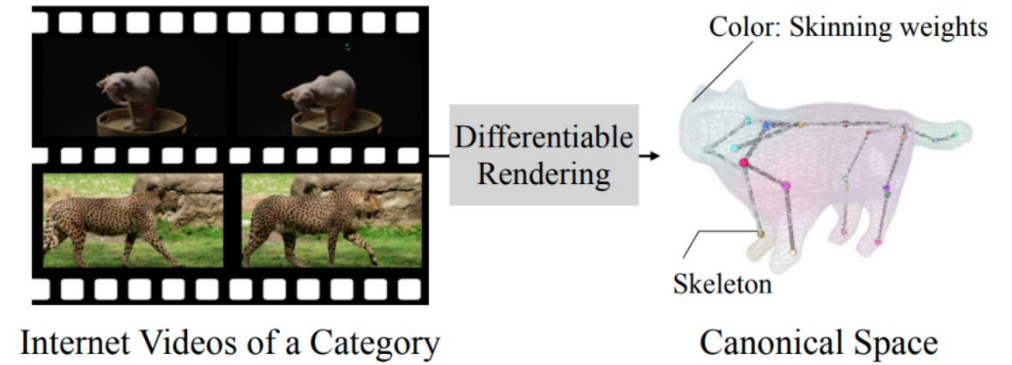


NPMs

Modelling General Articulated Motion

Category-level Modelling from RGB Videos

- Learn category-level shape and skeleton model from internet videos of a category
- Predict the instance-level bone locations for category skeleton using an auto-decoded MLP, similar to BANMO
- Capture instance-level articulations using BANMO

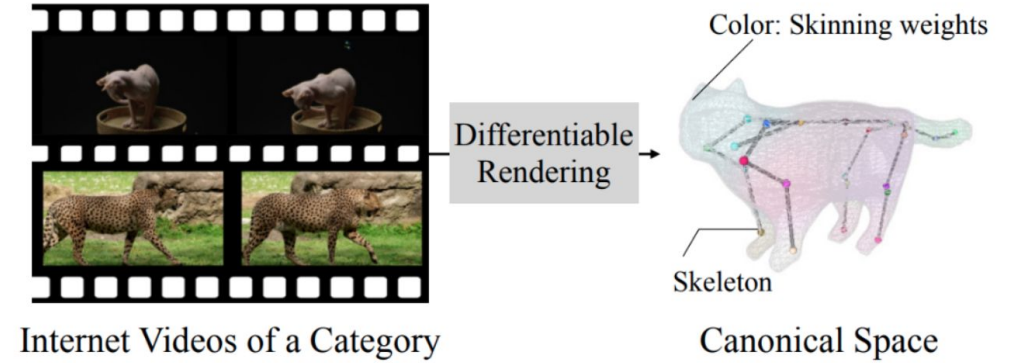


RAC

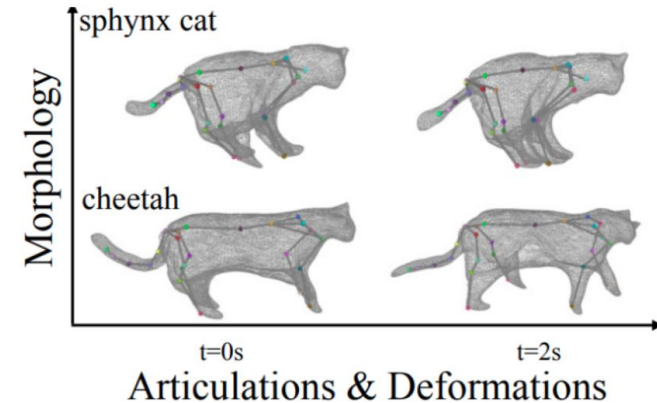
Modelling General Articulated Motion

Category-level Modelling from RGB Videos

- Learn category-level shape and skeleton model from internet videos of a category
- Predict the instance-level bone locations for category skeleton using an auto-decoded MLP, similar to BANMO
- Capture instance-level articulations using BANMO



Can we do it from image collections, which are more commonly available for general categories than videos?

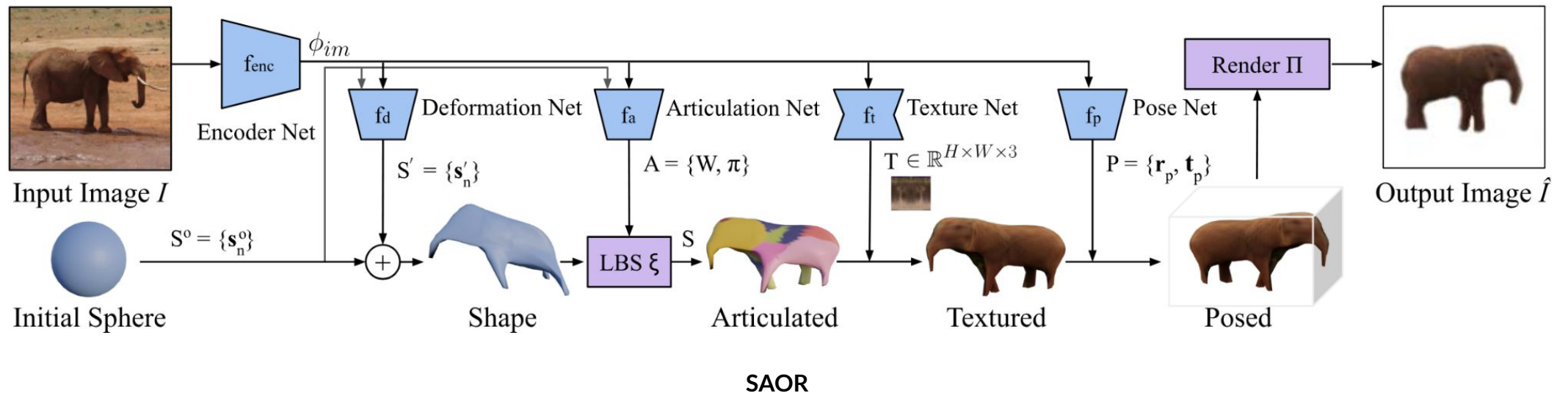


RAC

Modelling General Articulated Motion

Category-level Modelling from Image Collections

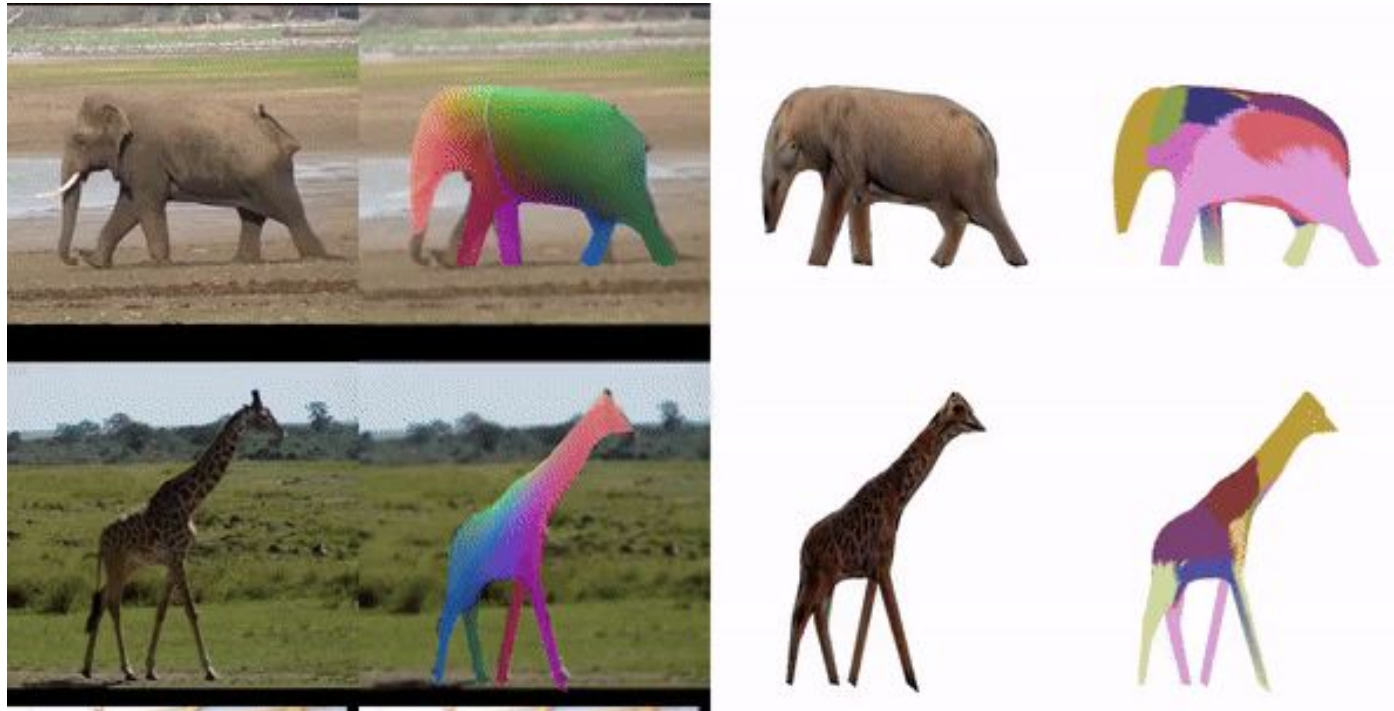
- Shape, articulation, pose and texture are directly predicted with separate decoders from an encoded image
- Category-level prior learned by shape and articulation decoders
- Enables prediction from single image at test-time



Modelling General Articulated Motion

Category-level Modelling from Image Collections

- Per-frame video reconstruction



SAOR

Trends

1. Speed and Quality Advancements
2. Handling of Large Deformations / Long-term 3D correspondences
3. **Modelling Articulated Motion for General Objects**



Trends

1. Speed and Quality Advancements

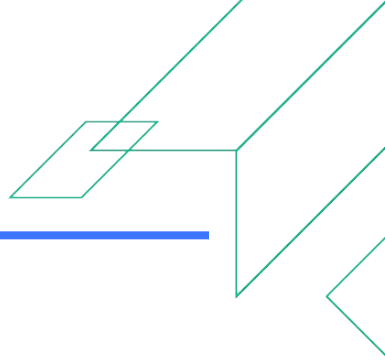
2. Ha

Non-rigid 3D reconstruction is far from solved!

ndences

3. Modelling Articulated Motion for General Objects

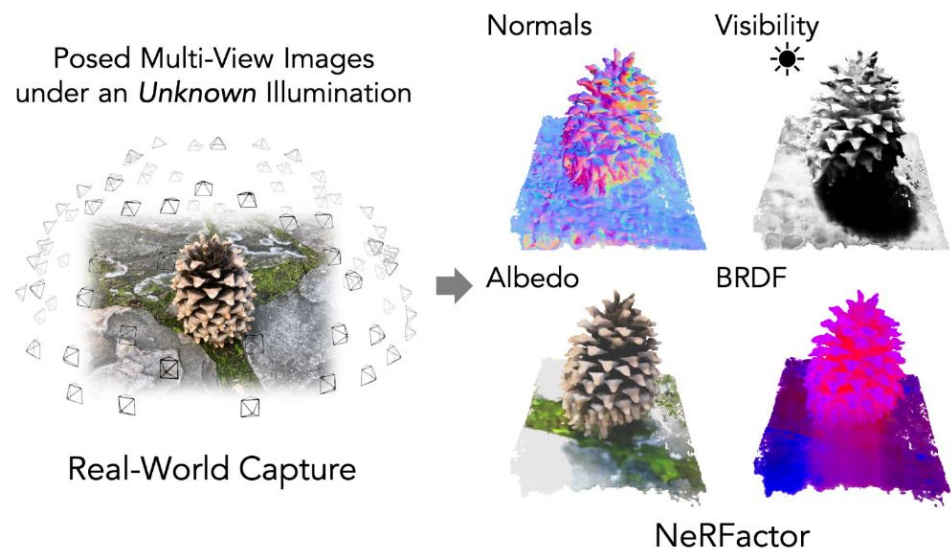
Remaining Challenges and Future Directions



Remaining Challenges and Future Directions

Intrinsic Decomposition and Relighting

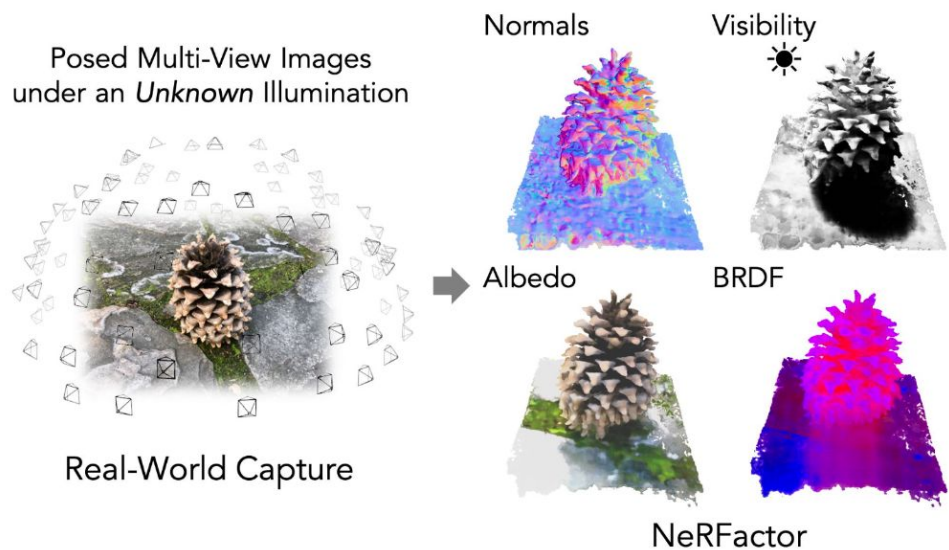
- Current methods for general scenes do not estimate materials and lighting
- Required to correctly relight objects in new environments



Remaining Challenges and Future Directions

Intrinsic Decomposition and Relighting

- Current methods for general scenes do not estimate materials and lighting
- Required to correctly relight objects in new environments



Faster Scene Representations

- Gaussian Splatting has introduced real-time rendering with photorealistic appearance
- Photorealistic reconstruction still requires offline training



Remaining Challenges and Future Directions

Reliable Camera Pose Estimation

- Current view synthesis methods rely on static Structure-from-Motion for camera poses
- Noisy when large and complex motions are present



Remaining Challenges and Future Directions

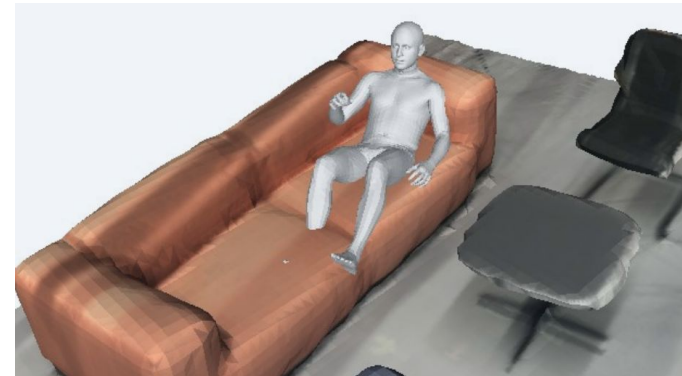
Reliable Camera Pose Estimation

- Current view synthesis methods rely on static Structure-from-Motion for camera poses
- Noisy when large and complex motions are present



Multi-Object Interaction

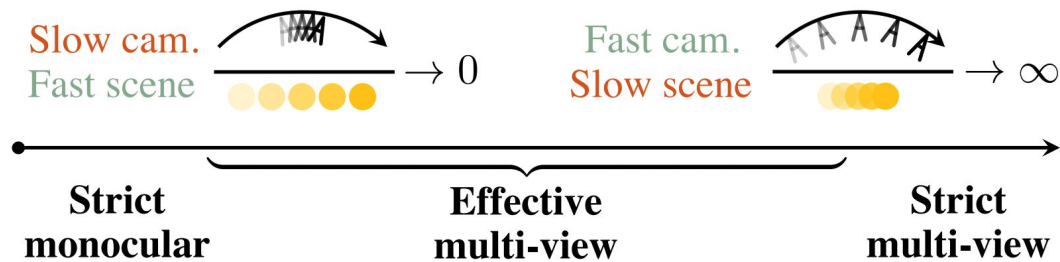
- Interaction between objects is not explicitly modelled by current methods for general objects
- Useful to enforce the correct dynamics and constraints



Remaining Challenges and Future Directions

Reconstruction from Sparse Casual Captures

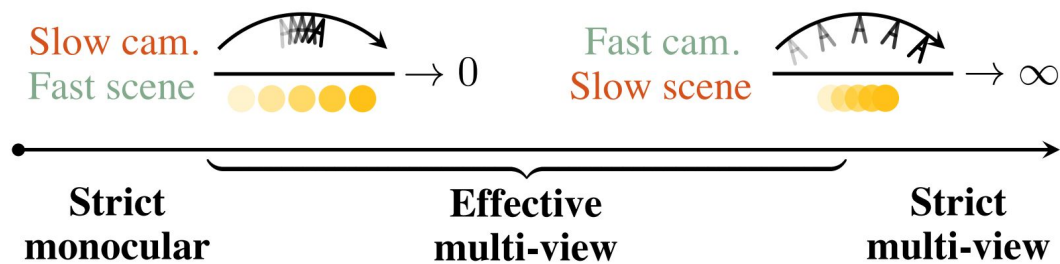
- Most methods evaluate on data with multi-view cues
- Reconstruction quality degrades for sparse, realistic monocular captures



Remaining Challenges and Future Directions

Reconstruction from Sparse Casual Captures

- Most methods evaluate on data with multi-view cues
- Reconstruction quality degrades for sparse, realistic monocular captures



Long-Term Dense Correspondences

- Recent works allow establishing 3D correspondences over time on lab-captured data
- Results not satisfactory for general real scenes with large and complex motion



Remaining Challenges and Future Directions

Generalizable Modeling and Generative Priors

- Text-to-image and text-to-video 2D diffusion models have been used as priors for 3D non-rigid scene generators
- We can see these powerful generative models being utilized for the non-rigid reconstruction task as well



Input

Synthesized Novel Views

Thank you.

More Information:

<https://razayunus.github.io/non-rigid-star>

Contact Information:

<https://razayunus.github.io>

Thanks to all authors for their contribution to the STAR!

© Eurographics Conference 2024. All rights preserved.