



UNIVERSITÄT
DES
SAARLANDES



MAX PLANCK INSTITUTE
FOR INFORMATICS

UNIVERSITY OF
MARYLAND

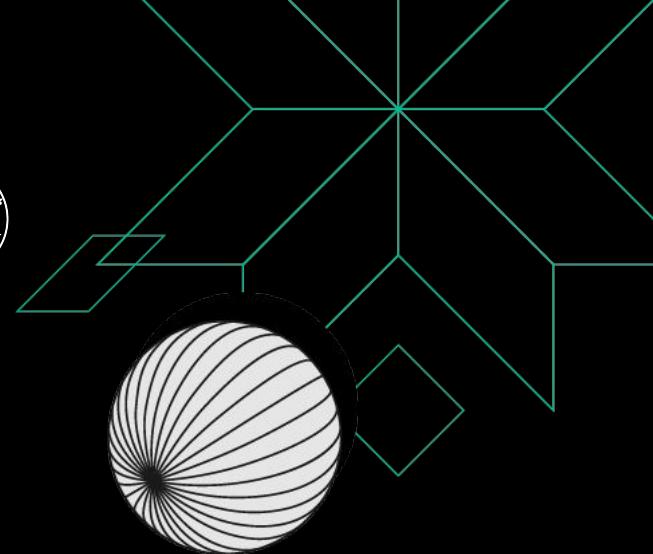
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



UNIVERSITY OF
OXFORD



Google



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

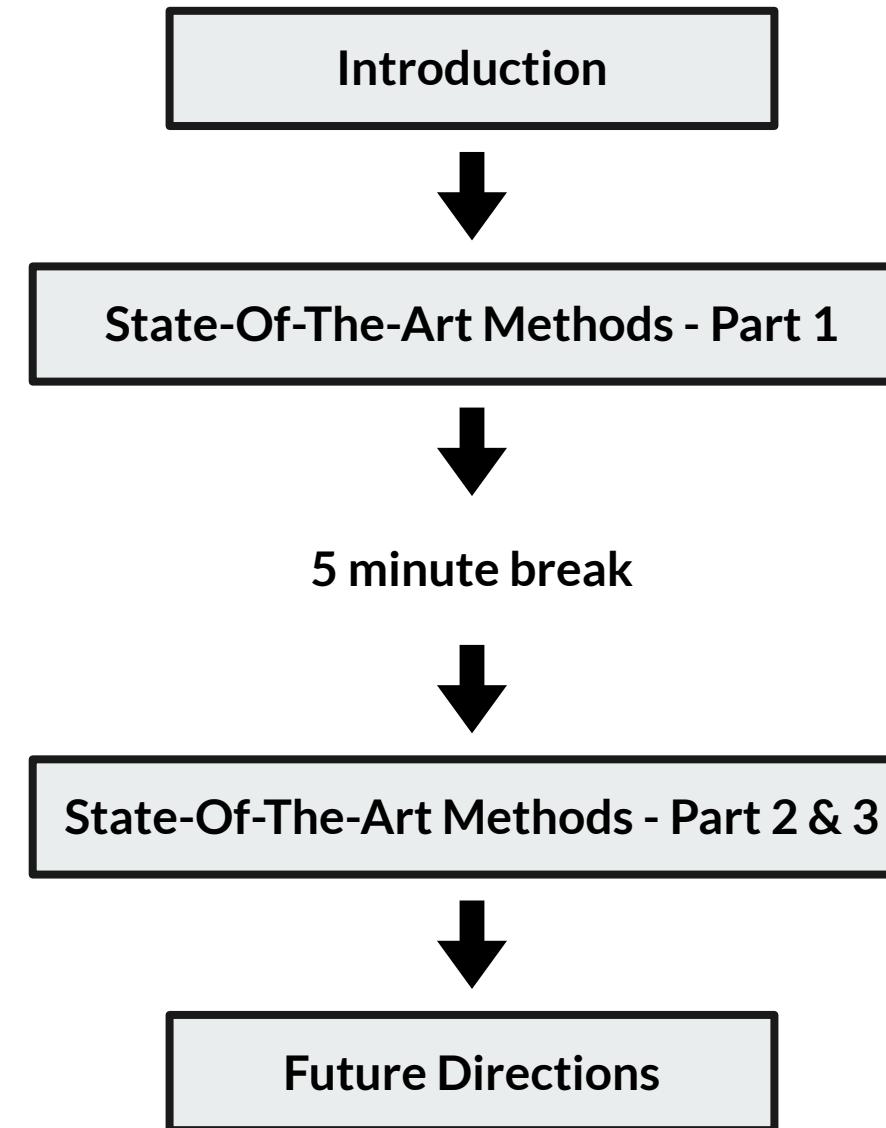
Presented by Raza Yunus

© Eurographics Conference 2024. All rights reserved.

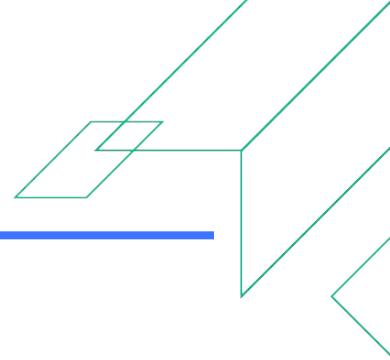


The 45th Annual Conference of the European Association for Computer Graphics is organized by CYENS Centre of Excellence in collaboration with the University of Cyprus and the Cyprus University of Technology.

Talk Schedule



Motivation & Applications



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



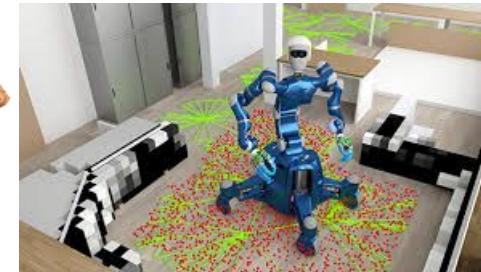
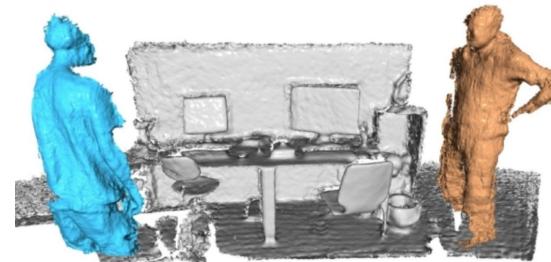
Motivation & Applications



Telepresence / VR



Movie & Gaming Industry



Robotics / AR

Motivation & Applications



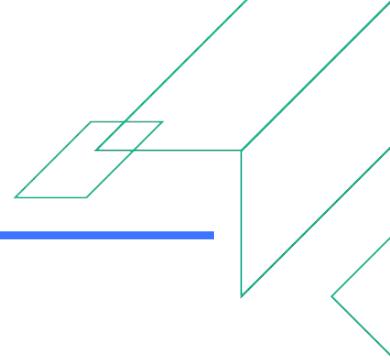
Recent advances are making non-rigid 3D reconstruction methods more and more powerful!



Robotics / AR

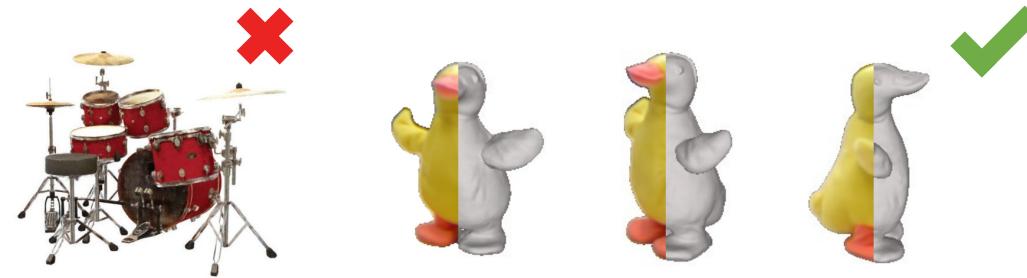
Scope

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



Scope

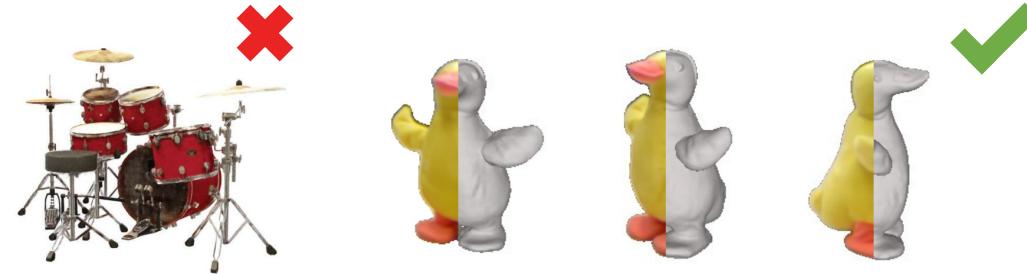
Recent Trends in 3D Reconstruction of General Non-Rigid Scenes



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

Scope

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



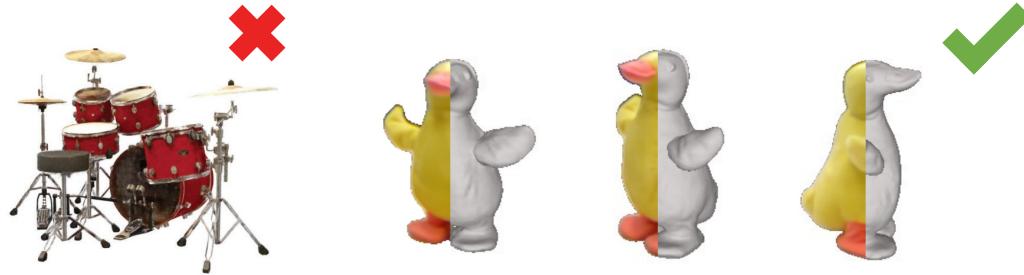
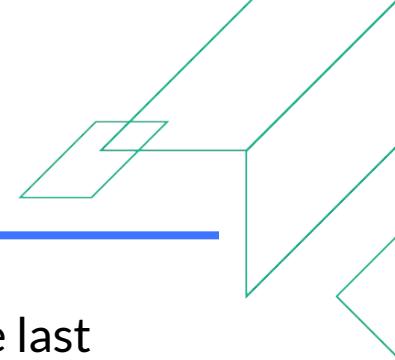
- Focus on methods that consider non-rigid deformations during reconstruction



- No domain-specific methods

Scope

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



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



- No domain-specific methods

- Covers methods mostly from the last three years
- We refer to older Eurographics STARs for a survey of earlier techniques:

DOI: 10.1111/cgf.14497
EUROGRAPHICS 2022
D. Mousavaei and G. Pascual
(Guest Editors)

Volume 41 (2022), Number 2
STAR – State of The Art Report

COMPUTER GRAPHICS Forum
Volume 41 (2022), Number 2
STAR – State of The Art Report

Advances in Neural Rendering



Figure 1: This state-of-the-art report discusses a large variety of neural rendering methods, from simple applications such as novel-view synthesis of static 3D meshes, to complex multi-view rendering and scene understanding. See Section 2 for more details on the various methods. Images adapted from [M3N3, NTU CMU 21, ZSPD 21, BBRP 21, LSS+21, PMP 21, JXZ 21, FPFM 21] ©2021 IEEE.

Abstract
Synthesizing photo-realistic images and videos is at the heart of computer graphics, and has been the focus of decades of research. Traditionally, rendering images or videos required using pre-defined deformable representations of geometry and material properties as input. Collectively, these appear as “rigid” models and what they can represent are limited to rigid objects and scenes. In contrast, recent advances have shown that non-rigid representations are also feasible and useful for various applications (e.g., created by an artist, point clouds (e.g., from a depth sensor), volumetric grids (e.g., from a CT scan), or implicit surface functions (e.g., learned via a neural network)). These representations are often referred to as “neural representations”. The main challenge in rendering is a long forward towards the goal of synthesizing photo-realistic images and video content. In recent years, we have seen many progress in this field through handling of applications that show different ways to interact with non-rigid representations via the neural network. This survey provides an overview of the methods that either directly or indirectly handle the classical rendering pipeline or learned 3D scene representation, often now referred to as neural scene representations. A brief discussion on the limitations of these methods is provided, followed by a summary of the latest developments in the synthesis of a captured scene. In addition to methods that handle static scenes, other new approaches to rendering these state-of-the-art methods, we provide an overview of fundamental concepts and definitions used in the current literature. We conclude with a discussion on open challenges and several implications.

1. Introduction
Synthesis of controllable 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 handle the rendering of static scenes, as well as the handling of the bending 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. To this end, all physical parameters of the scene have to be known for the rendering process. This is a major limitation of the rendering methods for non-rigid 3D reconstruction of various deformable objects and complete scenes from monocular videos or sets of monocular views. It restricts the fundamental types of deformations that can be handled by the methods. In this work, we will proceed toward techniques making stronger assumptions about the observed deformations (e.g., linear bending) and thus allowing the methods to handle more complex deformations. We also discuss the methods that generalize to non-rigid objects 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 the current literature. We conclude with a discussion on open challenges and several implications.

© 2022 The Author(s). Computer Graphics Forum published by DPG - The European Association for Computer Graphics on behalf of John Wiley & Sons Ltd. Published by John Wiley & Sons Ltd.

delivered by
EG EUROGRAPHICS
DIGITAL LIBRARY
www.eurographics.org

Tewari et al. (2022)

delivered by
EG EUROGRAPHICS
DIGITAL LIBRARY
www.eurographics.org

Tretschk et al. (2023)

DOI: 10.1111/cgf.14774
EUROGRAPHICS 2023
A. Borsigov and C. Theobalt
(Guest Editors)

Volume 41 (2023), Number 2
STAR – State of The Art Report

COMPUTER GRAPHICS Forum
Volume 41 (2023), Number 2
STAR – State of The Art Report

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



Figure 1: We review state-of-the-art methods for the dense 3D reconstruction of deformable objects from monocular images and videos, such as general objects, soft bodies in medical scenarios, animals, and human bodies and body parts. Images adapted from [JFMS 22, KRRK21, NTU 22, HZL 22, SITR 20, KTF 22, CXL 22, BZP 22, SNN 22].

Abstract
3D reconstruction of deformable 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—the problem admits many solutions leading to a variety of possible 3D reconstructions. Nevertheless, the advantage of using monocular cameras is their compactness and availability to the end user as well as their ease of use compared to multi-camera systems. This survey provides an overview of the state-of-the-art methods for dense 3D reconstruction of various deformable objects and complete scenes from monocular videos or sets of monocular views. It reviews the fundamental types of deformations that can be handled by the methods, the provided reconstruction pipelines, and the types of deformations (e.g., linear bending) that the methods can handle. We also discuss the methods that generalize to non-rigid objects and can be used for generative tasks. In addition to reviewing these state-of-the-art methods, we provide an overview of the datasets for training and evaluation of the discussed techniques. We conclude with a discussion on open challenges and several implications.

1. Introduction
Humans can close one eye, look around, and get a fair sense of the environment without the need for prior knowledge (e.g., 3D models, depth maps, etc.). This is due to the fact that the brain uses information about the scene geometry and material properties such as reflectivity or opacity. Given this information, modern ray tracing

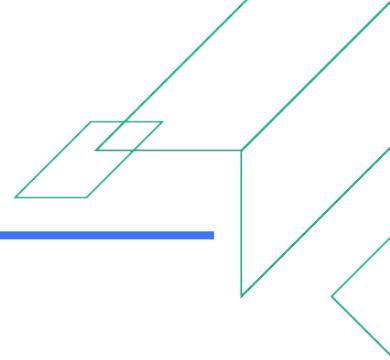
methods are able to synthesize the light transport from light sources to the virtual camera for synthesis.

To this end, all physical parameters of the scene have to be known for the rendering process. This is a major limitation of the rendering methods for non-rigid 3D reconstruction of various deformable objects and complete scenes from monocular videos or sets of monocular views. It restricts the fundamental types of deformations that can be handled by the methods. In this work, we will proceed toward techniques making stronger assumptions about the observed deformations (e.g., linear bending) and thus allowing the methods to handle more complex deformations. We also discuss the methods that generalize to non-rigid objects 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 the current literature. We conclude with a discussion on open challenges and several implications.

© 2023 The Author(s). Computer Graphics Forum published by DPG - The European Association for Computer Graphics on behalf of John Wiley & Sons Ltd. Published by John Wiley & Sons Ltd.

This is an open access article under the terms of the Creative Commons Attribution Non-Commercial License, available at <http://creativecommons.org/licenses/by-nc/4.0/>. You are free to copy, distribute, display, and perform the work. You must not change it in any way or sell it in a commercial context. You must credit the author(s) and respect any legal not

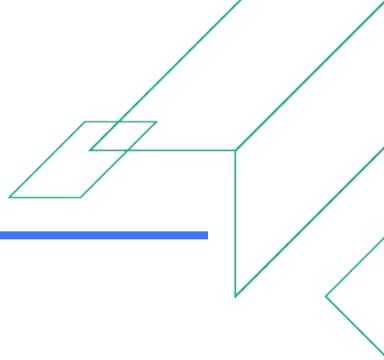
State-of-the-Art Report



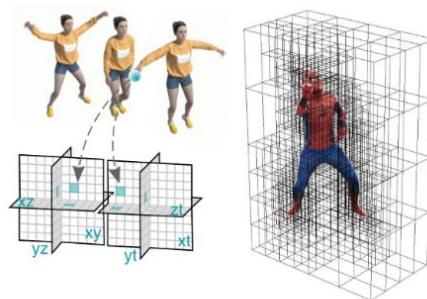
State-of-the-Art Report

Over 150 methods divided into four categories

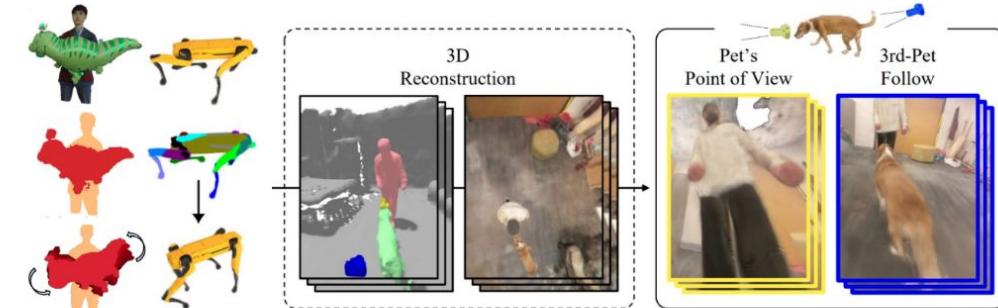
State-of-the-Art Report



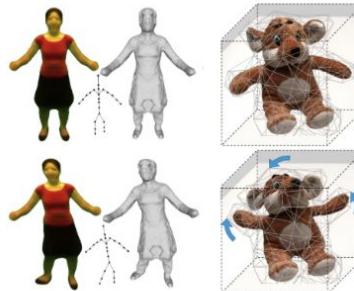
Over 150 methods divided into four categories



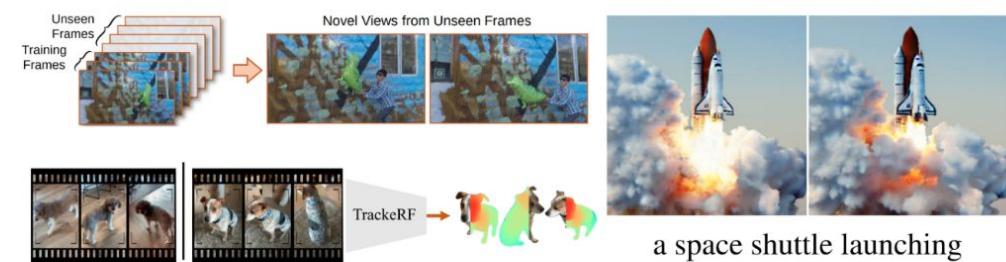
3D Non-Rigid Reconstruction and View Synthesis



Decompositional Scene Analysis

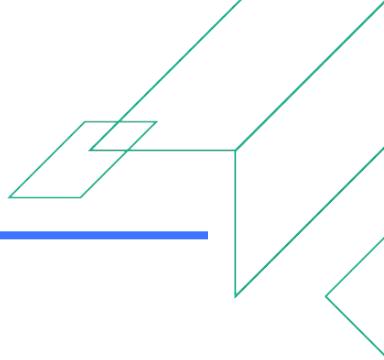


Editability and Control

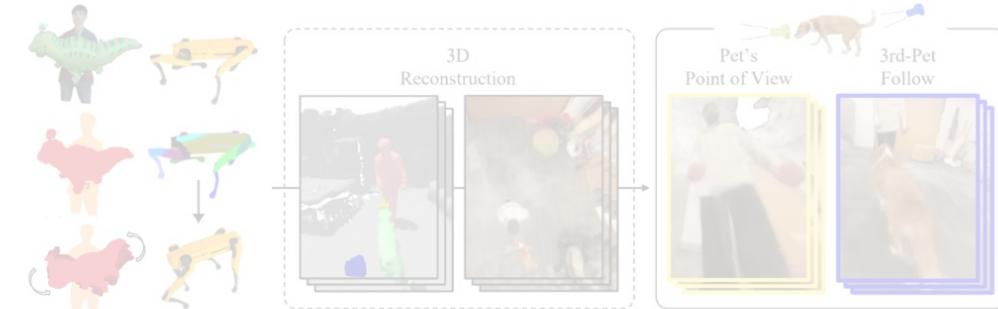
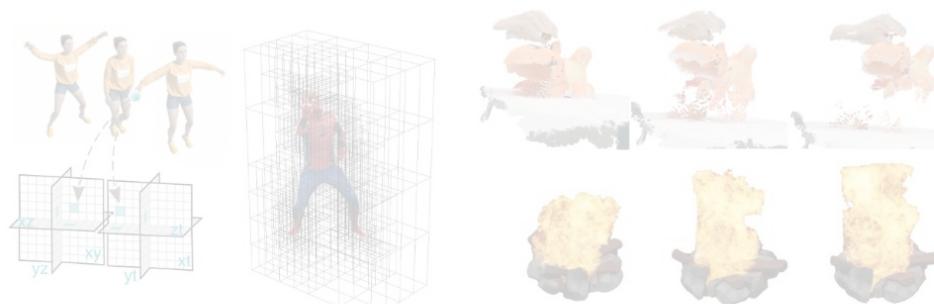


Generalizable and Generative Modeling

State-of-the-Art Report

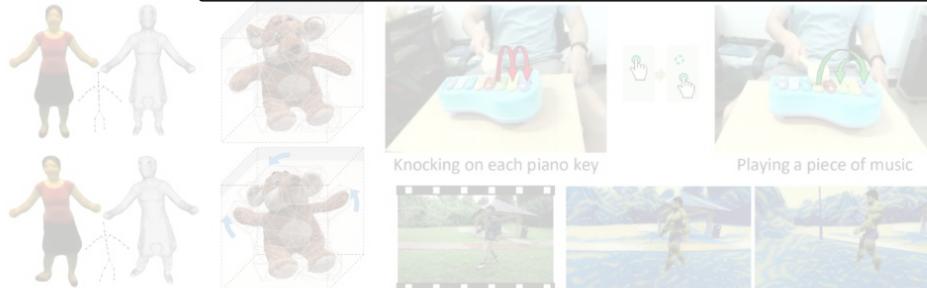


Over 150 methods divided into four categories



3D

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

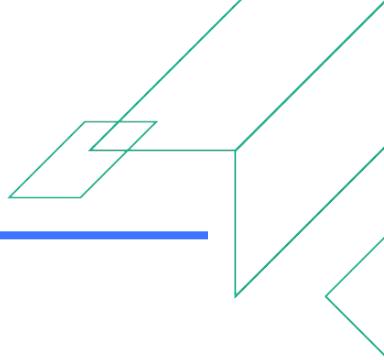


Editability and Control



Generalizable and Generative Modeling

Trends



Trends

1. Speed and Quality Advancements

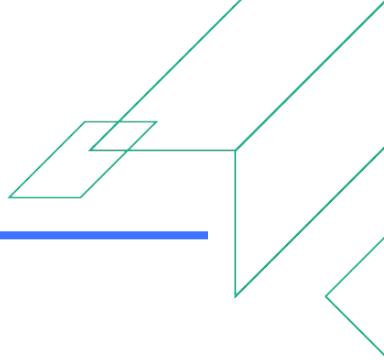
Trends

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

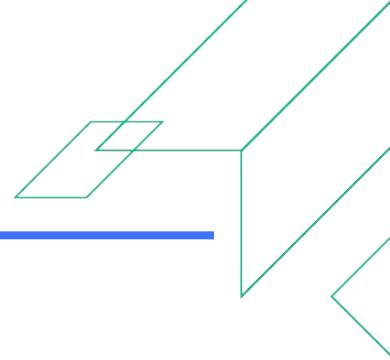


- Speed and Quality Advancements
- Hand Exemplars
- First, let's have a brief look at the different aspects of non-rigid 3D reconstruction
- Modelling Articulated Motion for General Objects



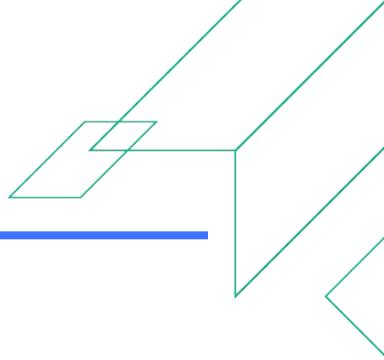
Task

Non-Rigid 3D Reconstruction and View Synthesis



Task

Non-Rigid 3D Reconstruction and View Synthesis



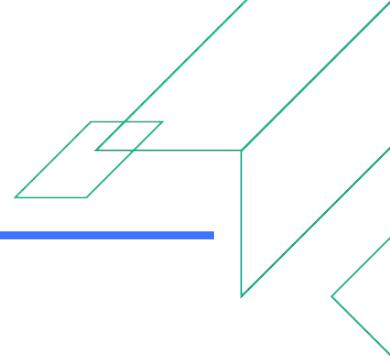
↓
Time



Observations



Sensors and Capture Settings



Sensors and Capture Settings



RGB

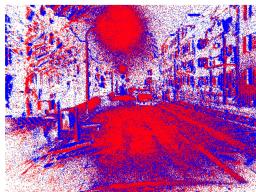


Passive Depth

Structured Depth

Time-of-Flight

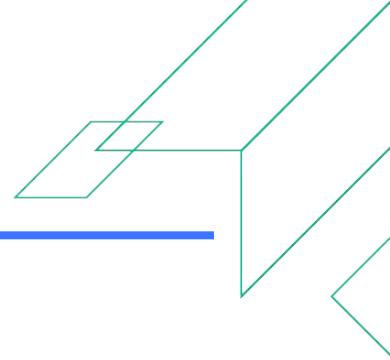
Depth



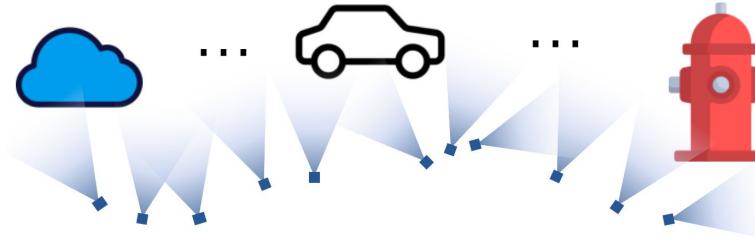
Event

Sensor Types

Sensors and Capture Settings



RGB



Monocular

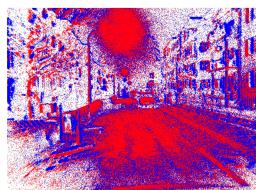


Depth

Passive Depth

Structured Depth

Time-of-Flight



Event



Multi-view

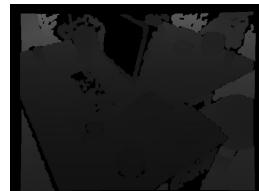
Sensor Types

Capture Settings

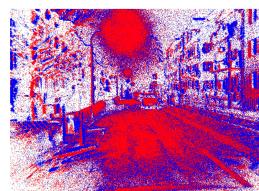
Sensors and Capture Settings



RGB

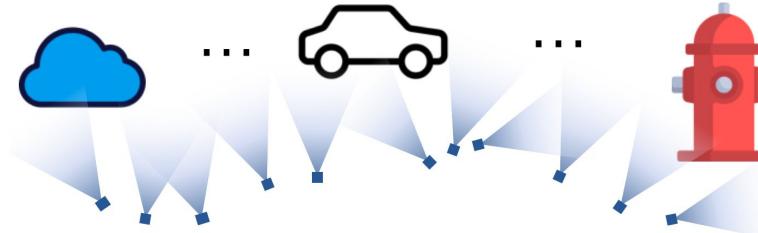


Depth



Event

Sensor Types

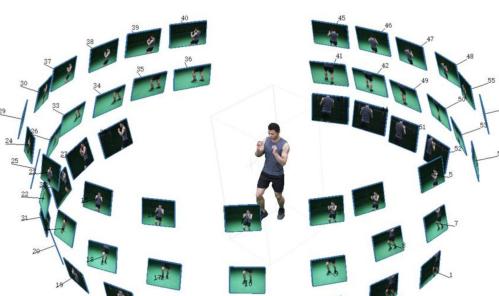


Monocular

Passive Depth

Structured Depth

Time-of-Flight



Multi-view

Capture Settings



Forward Facing



360 Degree



Freeform

Capture Trajectories

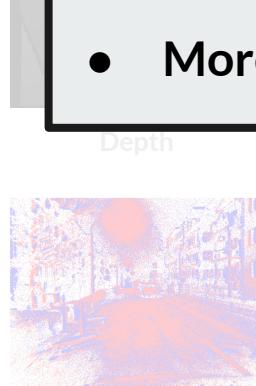
Sensors and Capture Settings



RGB



Forward Facing

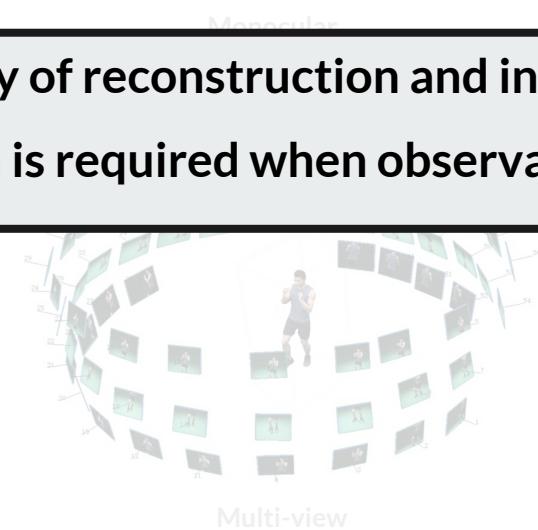


Depth



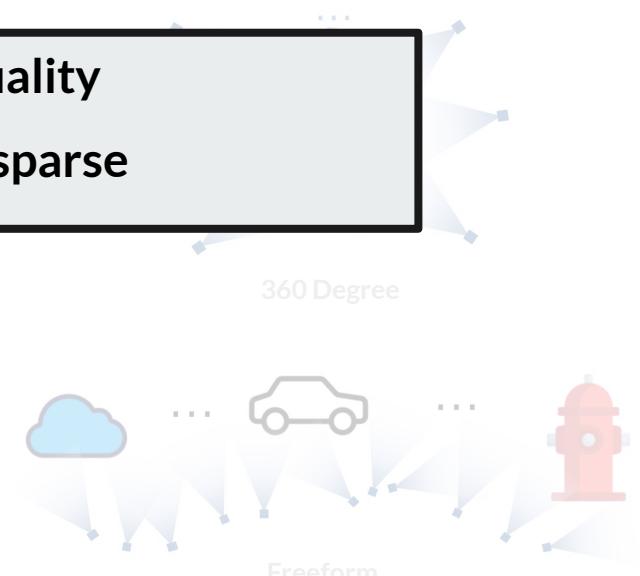
Event

Sensor Types



Multi-view

Capture Settings

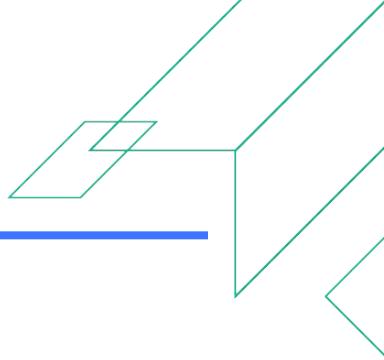


Freeform

Capture Trajectories

Task

Non-Rigid 3D Reconstruction and View Synthesis



↓
Time

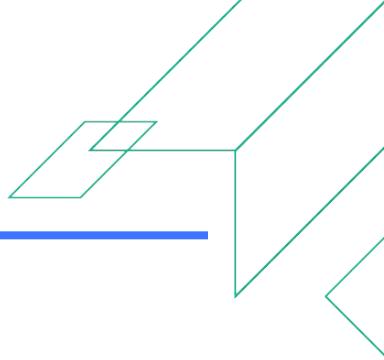


Observations

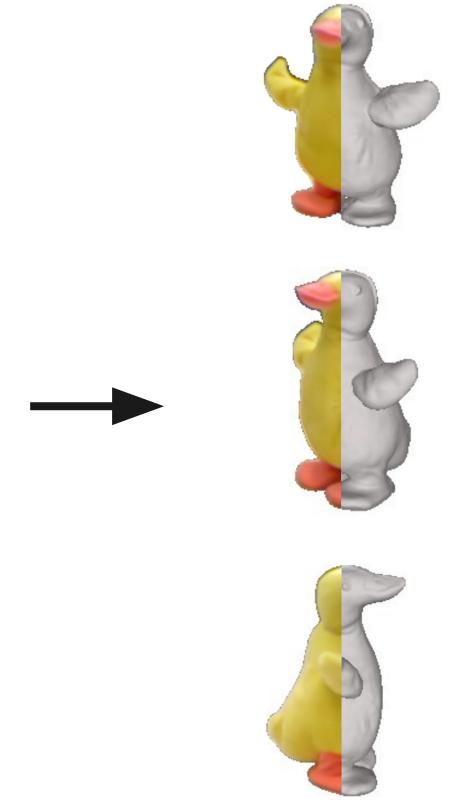


Task

Non-Rigid 3D Reconstruction and View Synthesis

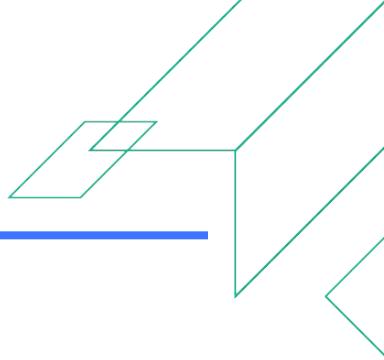


↓
Time

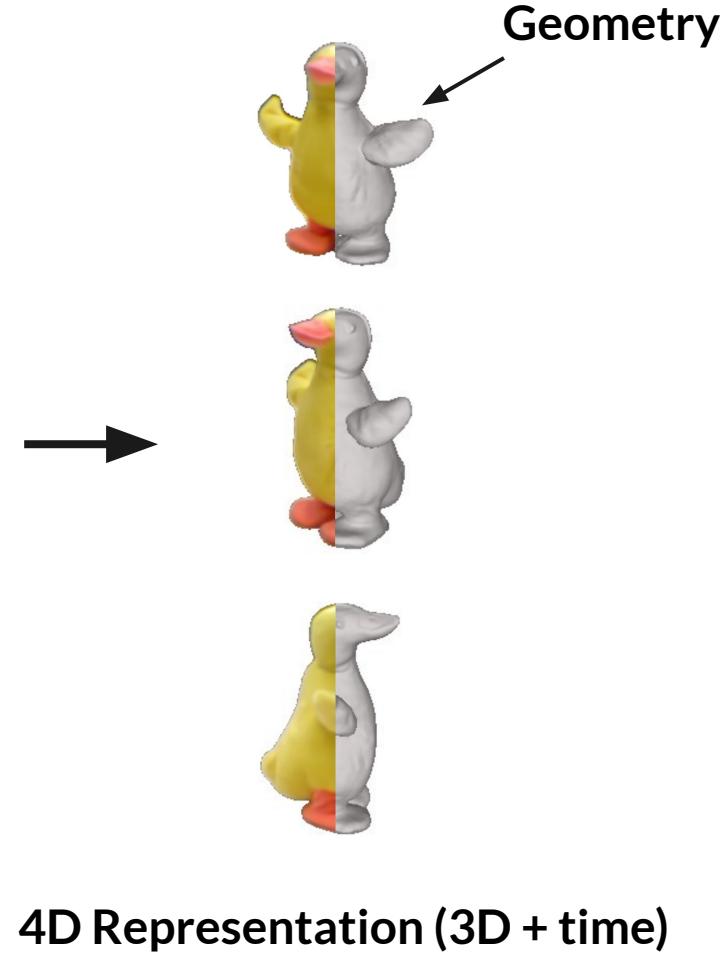


Task

Non-Rigid 3D Reconstruction and View Synthesis

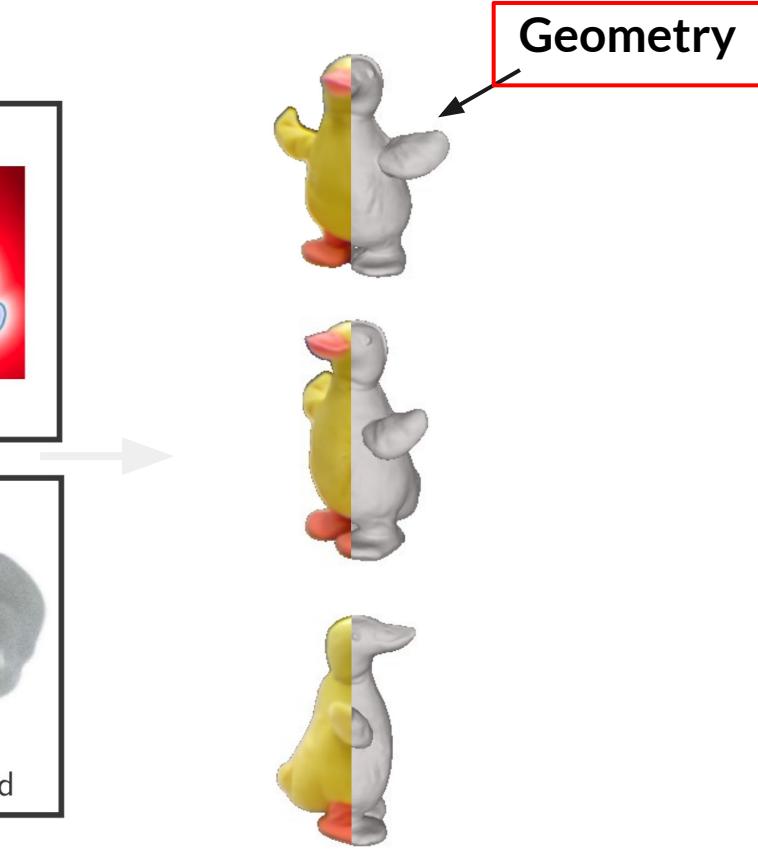
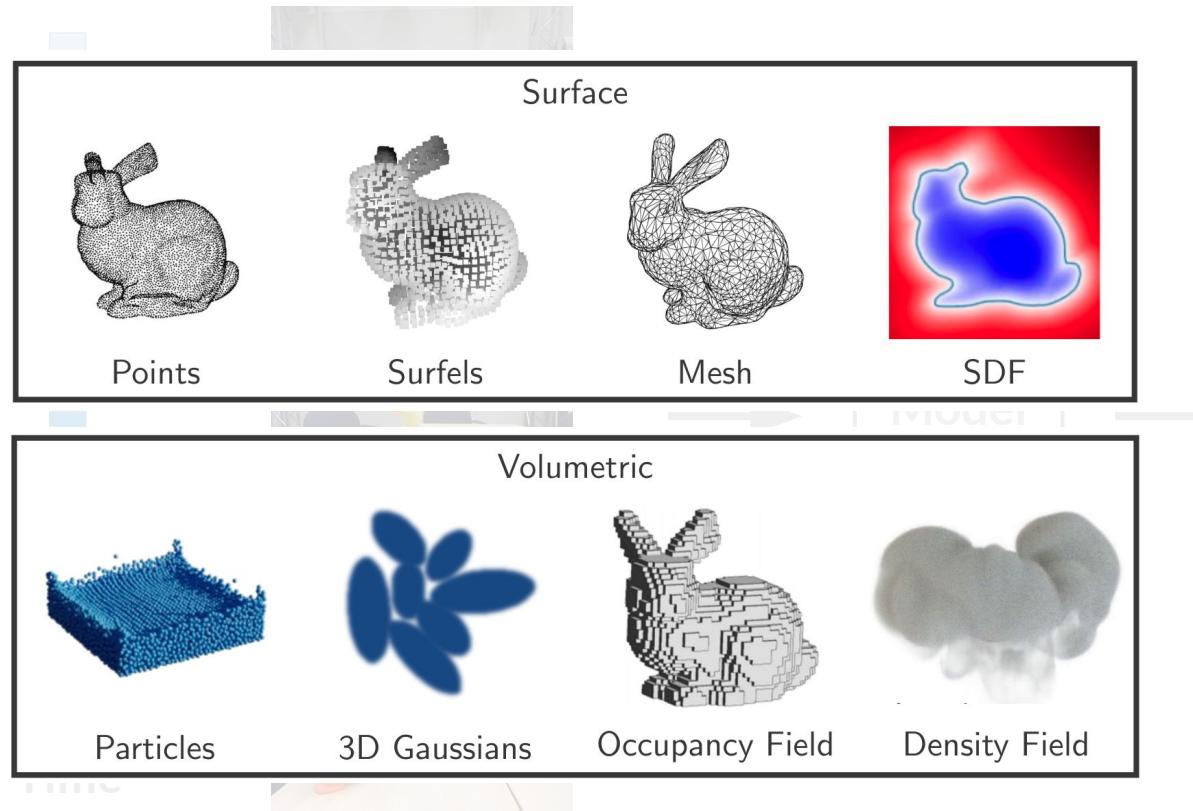


↓
Time



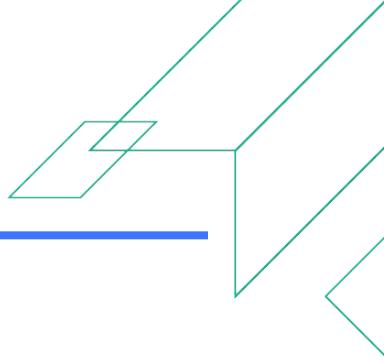
Task

Non-Rigid 3D Reconstruction and View Synthesis

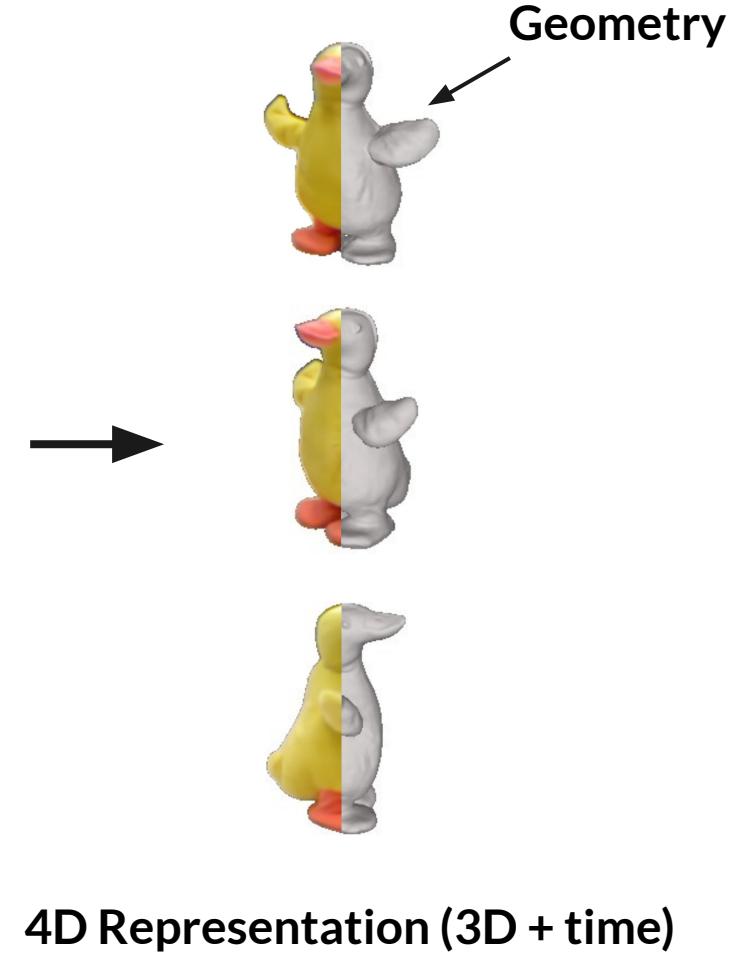


Task

Non-Rigid 3D Reconstruction and View Synthesis

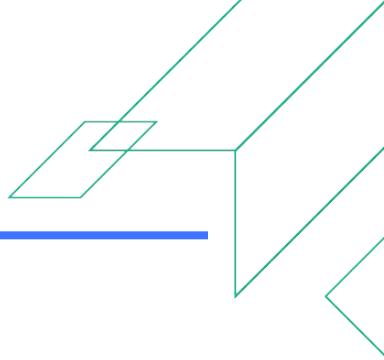


↓
Time



Task

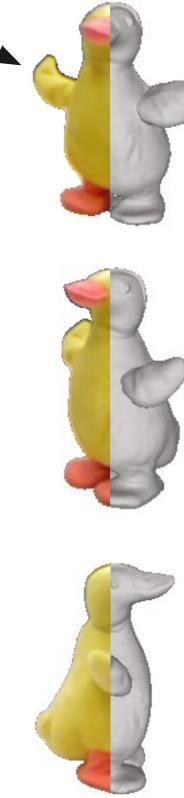
Non-Rigid 3D Reconstruction and View Synthesis



↓
Time



Appearance Geometry

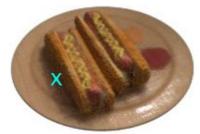


4D Representation (3D + time)



Task

Non-Rigid 3D Reconstruction and View Synthesis

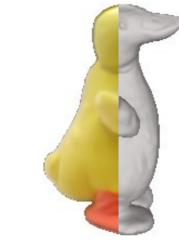
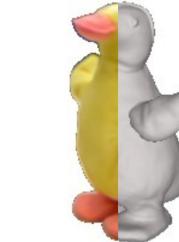


$$= \int_S ((b) \text{ Light Visibility} \times (c) \text{ Direct Illumination} + (d) \text{ Indirect Illumination}) \times (e) \text{ BRDF} d\omega_i$$

Interaction between environment lighting and surface materials!

Appearance

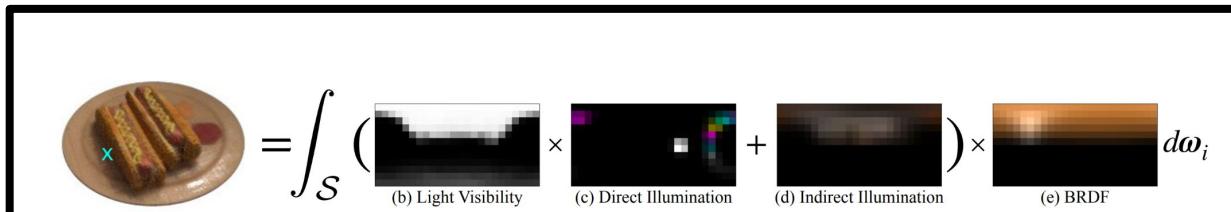
Geometry



4D Representation (3D + time)

Task

Non-Rigid 3D Reconstruction and View Synthesis

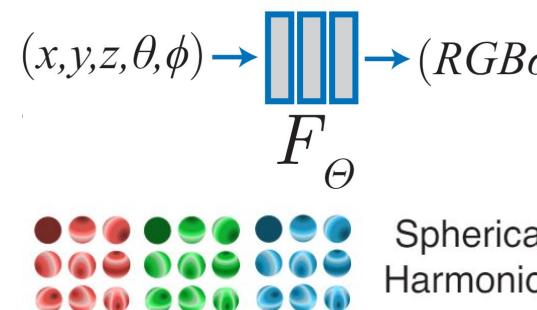


Interaction between environment lighting and surface materials!

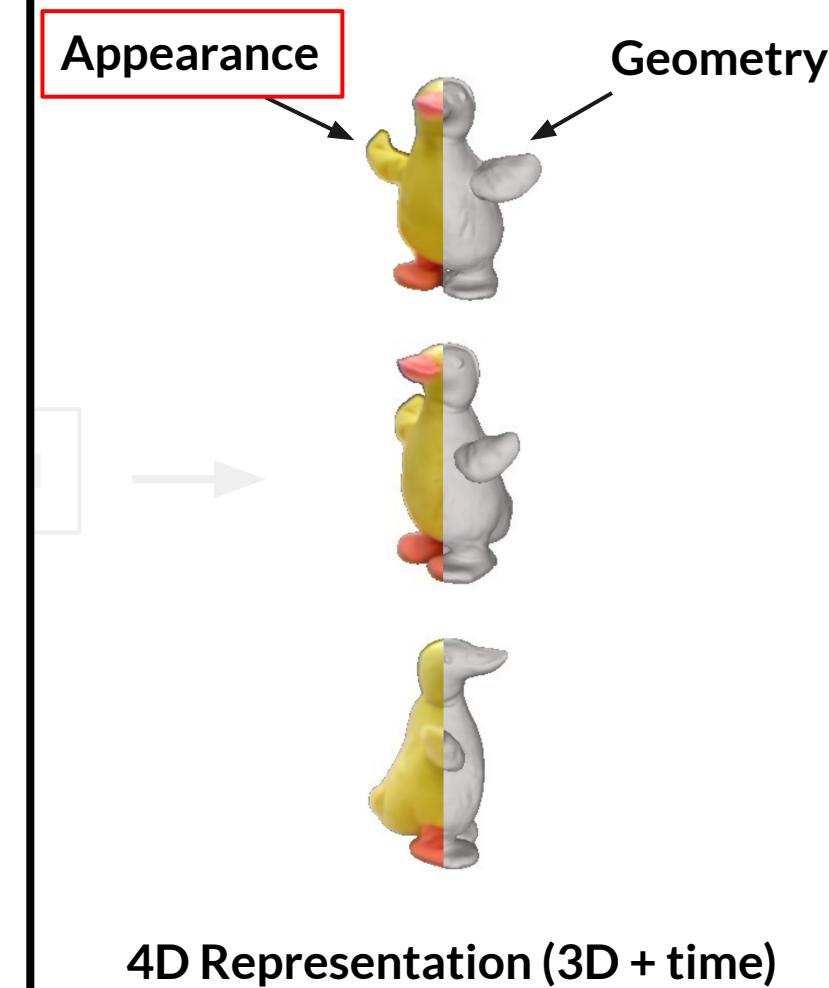
Simplifications:



Lambertian Surface Model

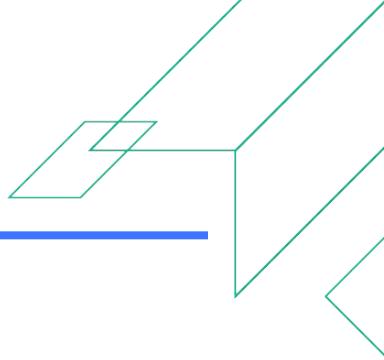


View-Dependent Outgoing Irradiance

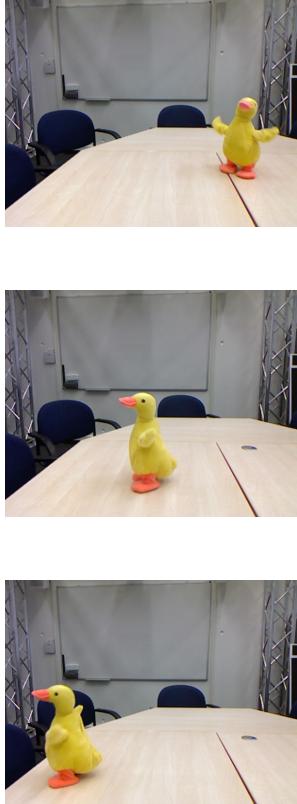


Task

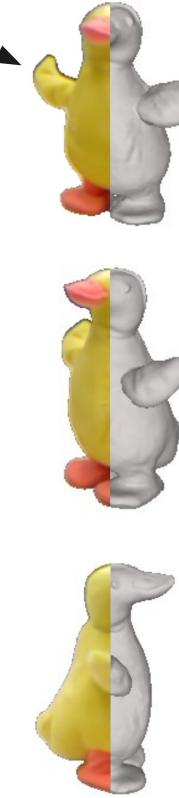
Non-Rigid 3D Reconstruction and View Synthesis



↓
Time



Appearance Geometry



4D Representation (3D + time)



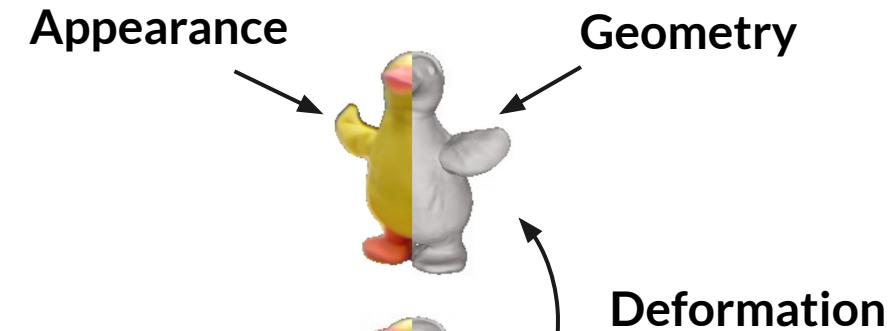
Task

Non-Rigid 3D Reconstruction and View Synthesis

↓
Time



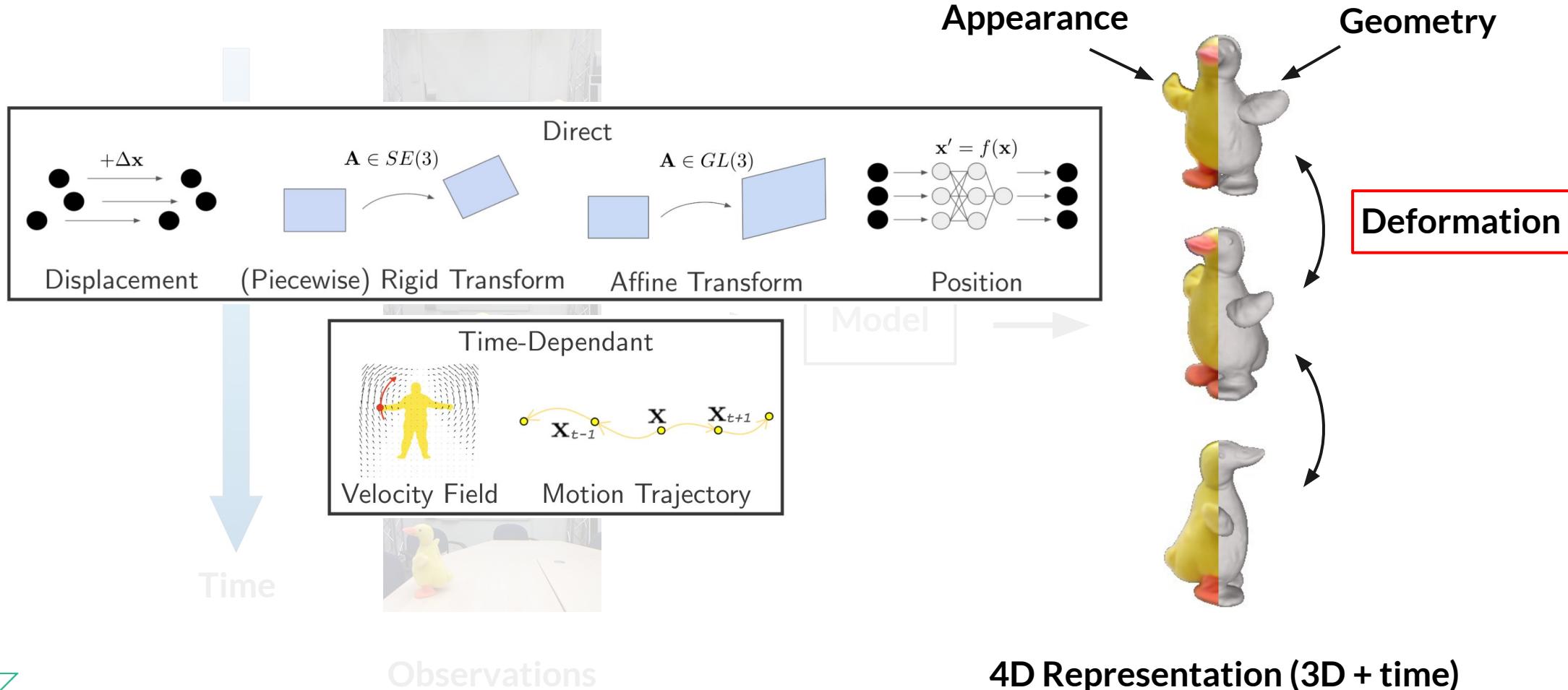
Observations



4D Representation (3D + time)

Task

Non-Rigid 3D Reconstruction and View Synthesis



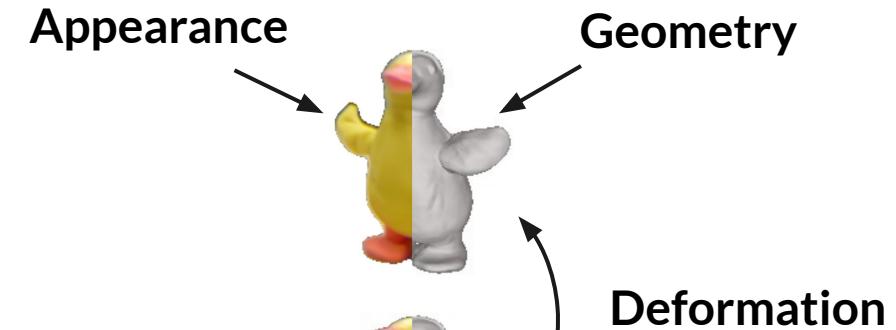
Task

Non-Rigid 3D Reconstruction and View Synthesis

↓
Time



Observations



4D Representation (3D + time)

Task

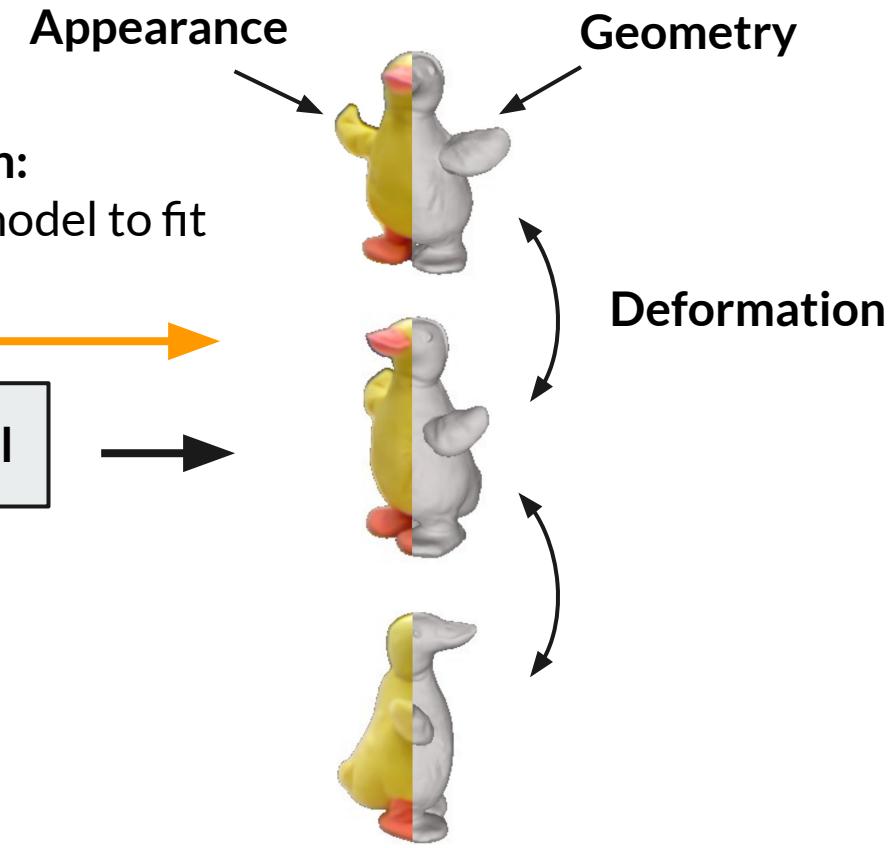
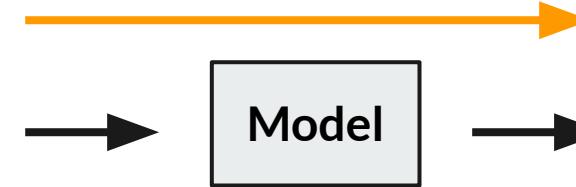
Non-Rigid 3D Reconstruction and View Synthesis

Time



Observations

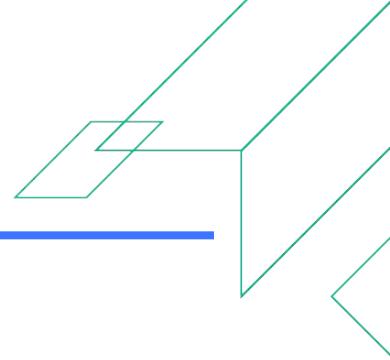
Reconstruction:
Optimize the model to fit
observations



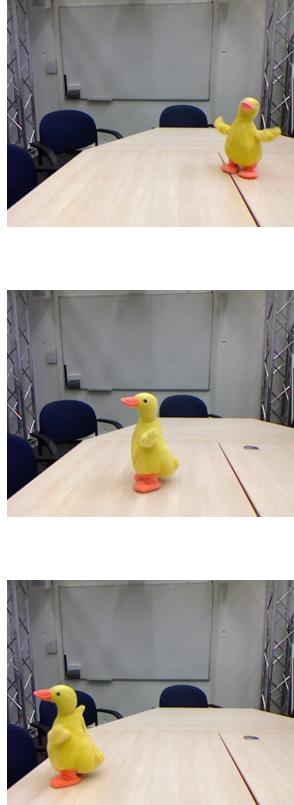
4D Representation (3D + time)

Task

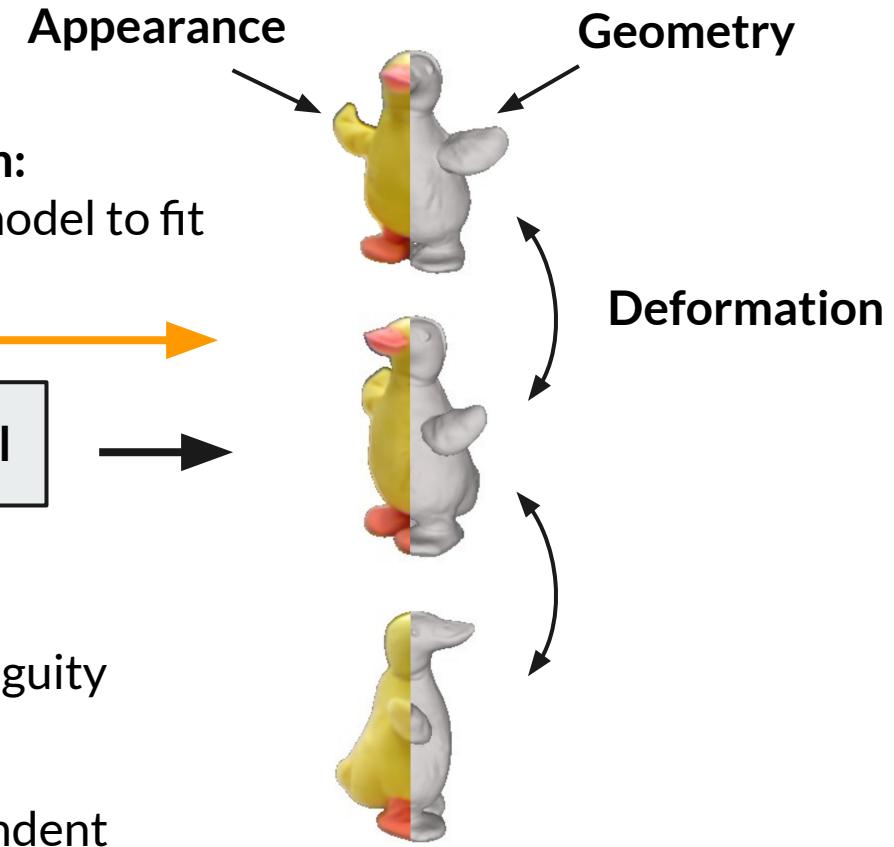
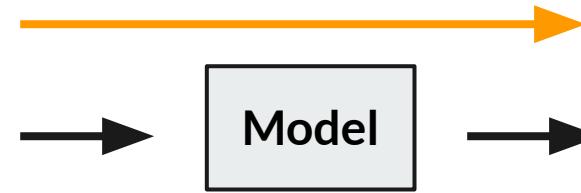
Non-Rigid 3D Reconstruction and View Synthesis



↓
Time



Reconstruction:
Optimize the model to fit
observations



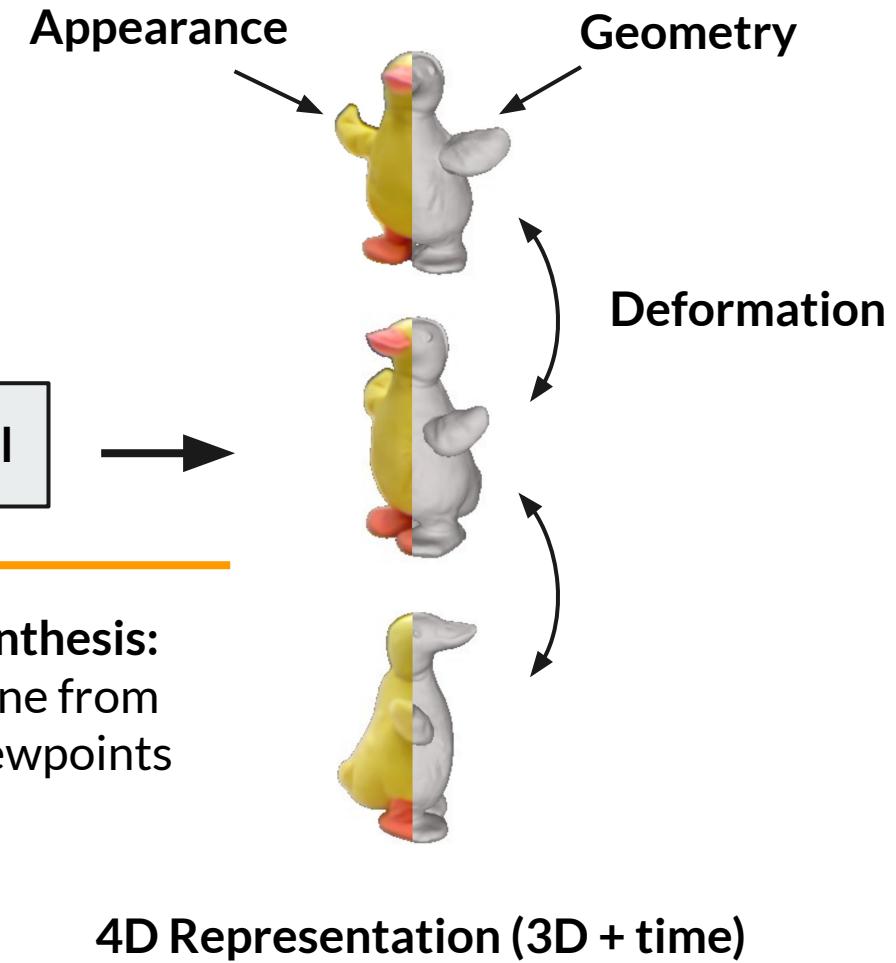
Task

Non-Rigid 3D Reconstruction and View Synthesis

↓
Time

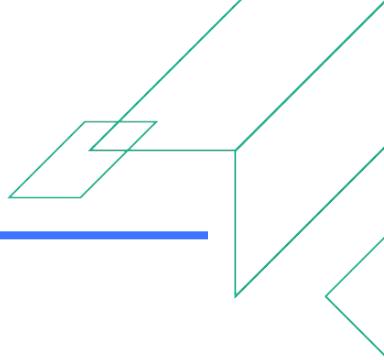


Novel View Synthesis:
Render the scene from
unobserved viewpoints



Task

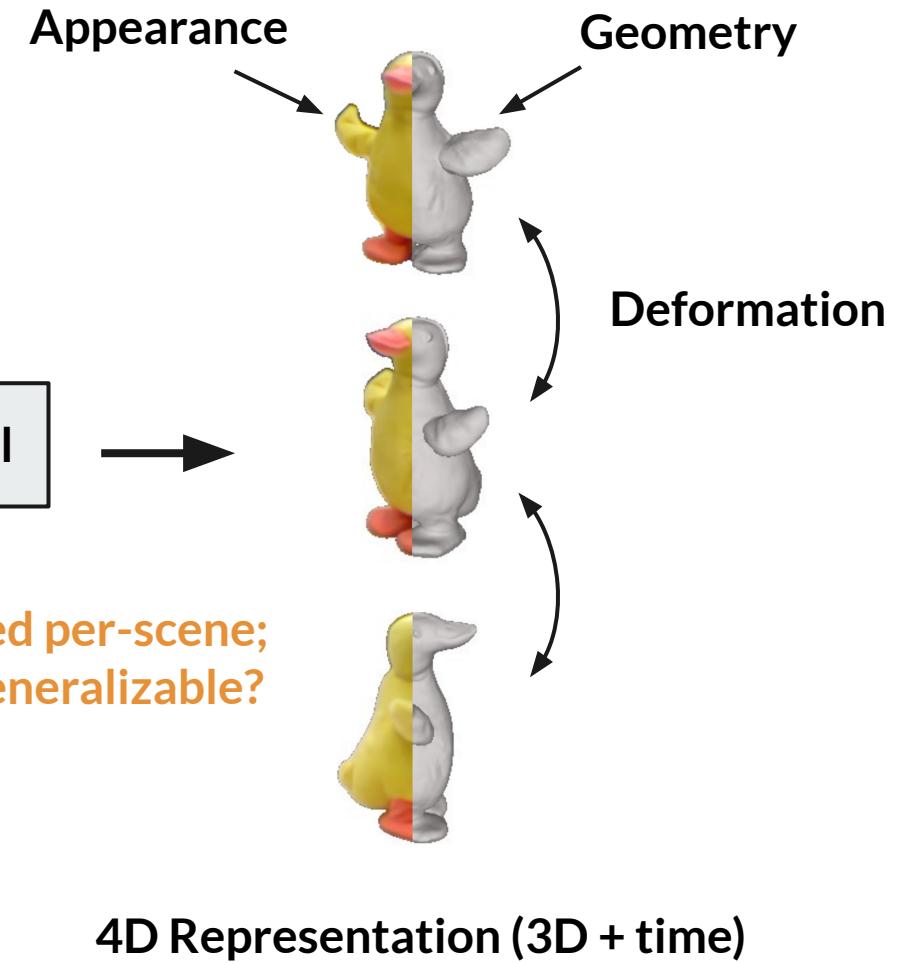
Non-Rigid 3D Reconstruction and View Synthesis



↓
Time

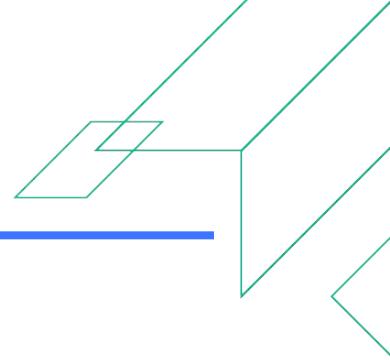


Model is optimized per-scene;
how to make it generalizable?

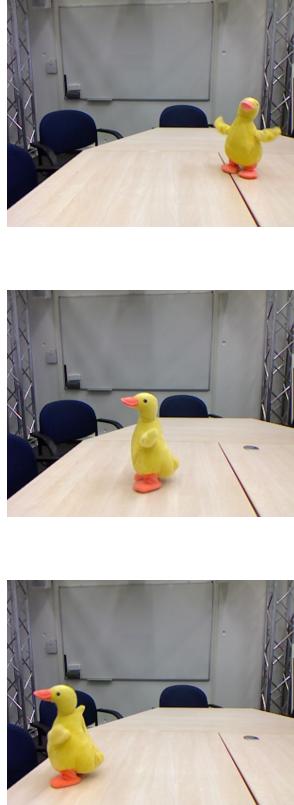


Task

Non-Rigid 3D Reconstruction and View Synthesis



↓
Time

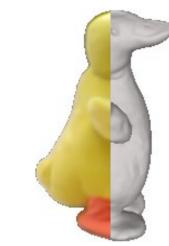
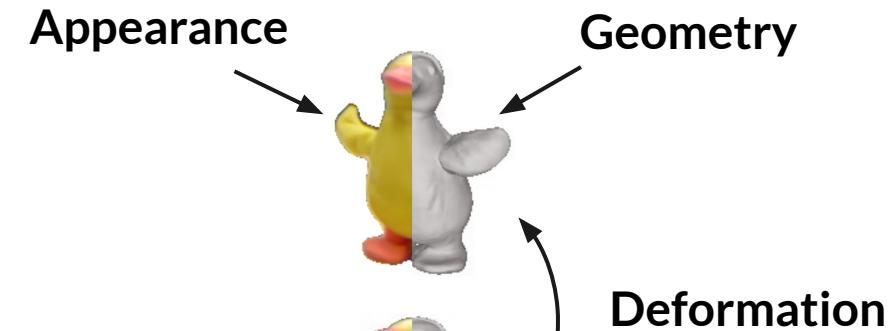


Observations

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



Model is optimized per-scene;
how to make it generalizable?

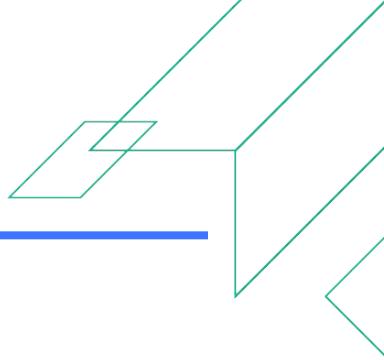


4D Representation (3D + time)

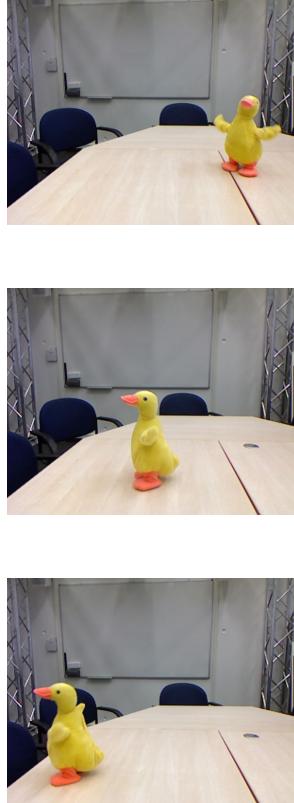


Task

Non-Rigid 3D Reconstruction and View Synthesis

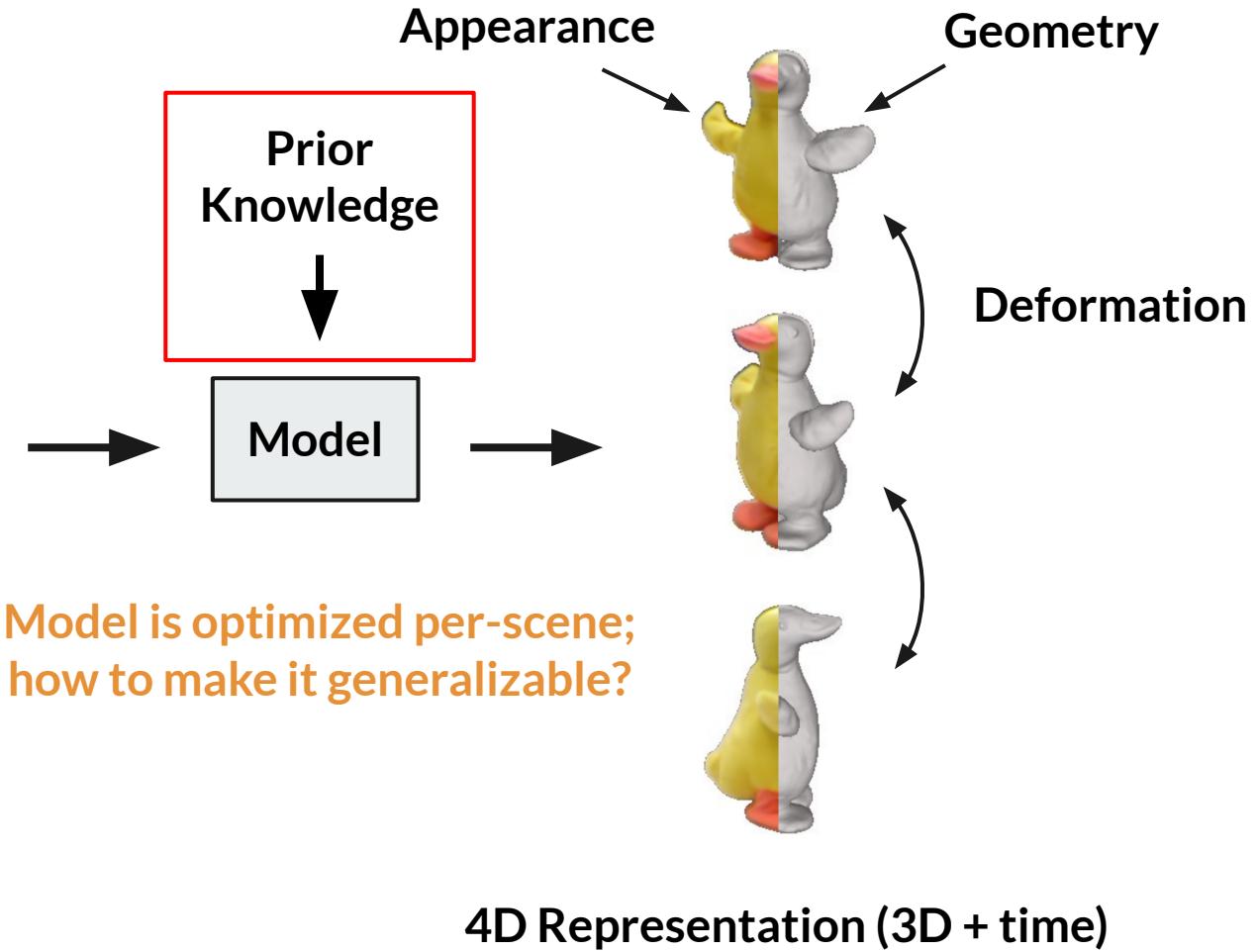


↓
Time



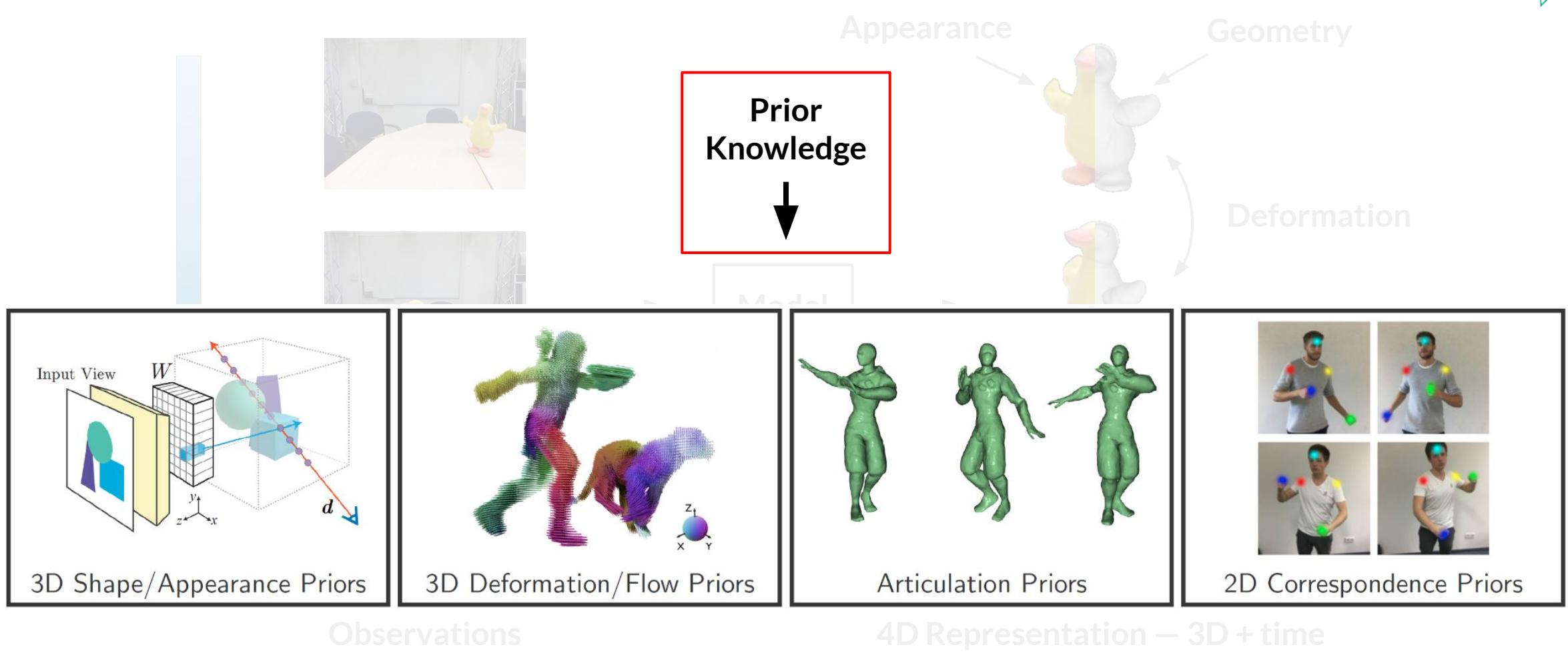
Observations

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



Task

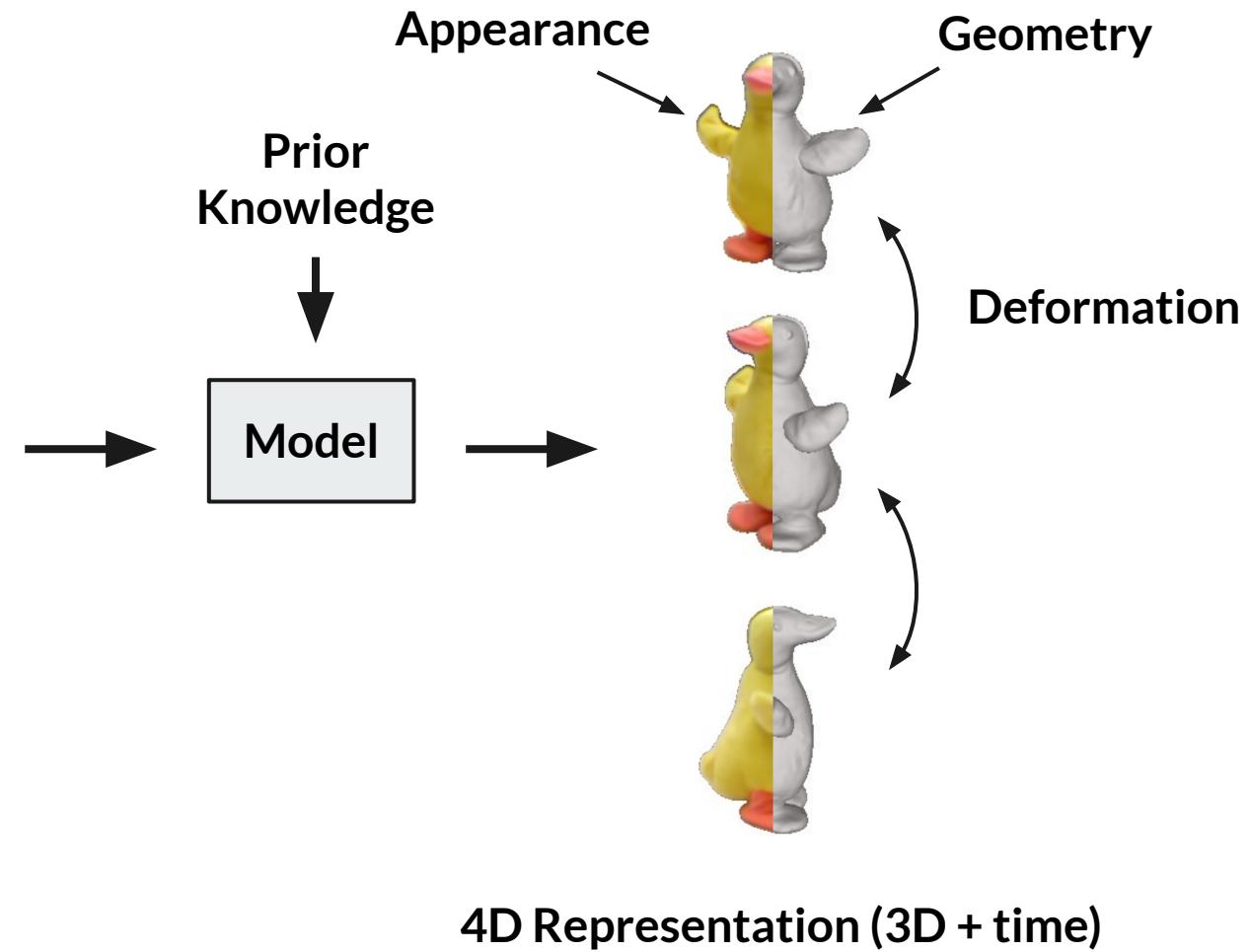
Non-Rigid 3D Reconstruction and View Synthesis



Task

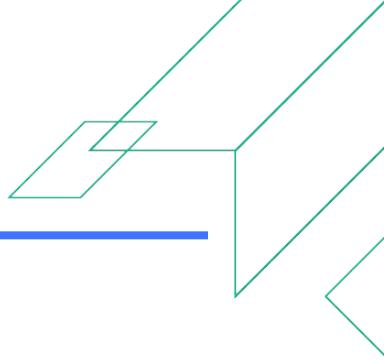
Non-Rigid 3D Reconstruction and View Synthesis

↓
Time

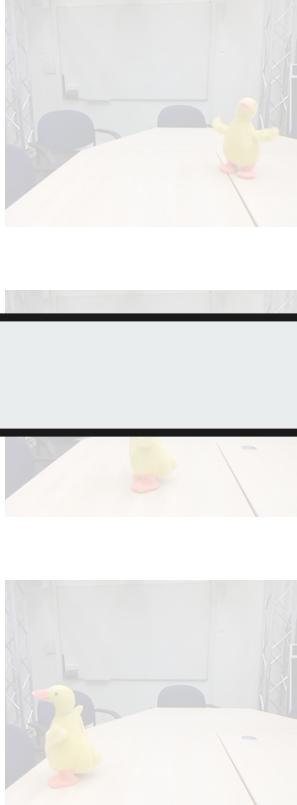


Task

Non-Rigid 3D Reconstruction and View Synthesis



↓
Time



Observations



Let's look at the trends

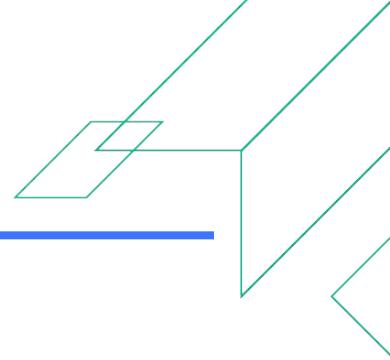
4D Representation – 3D + time



Trends

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

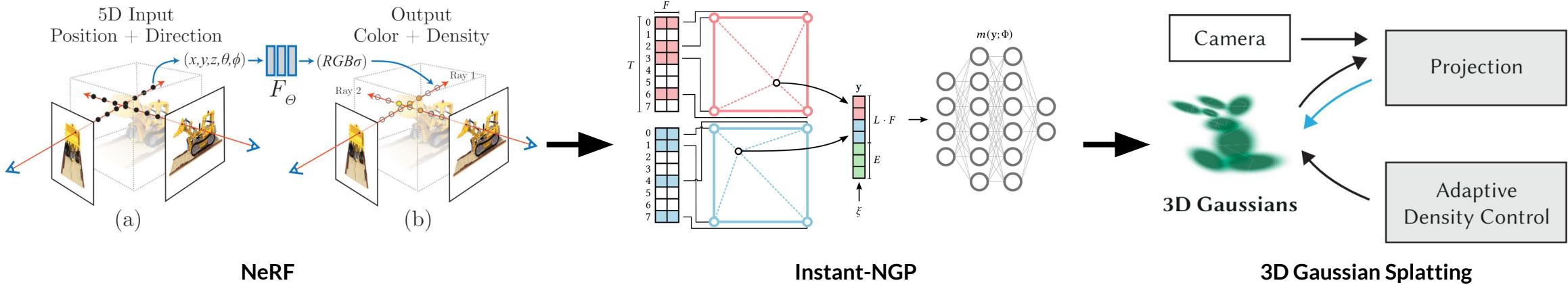
Trends

- 
- 
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

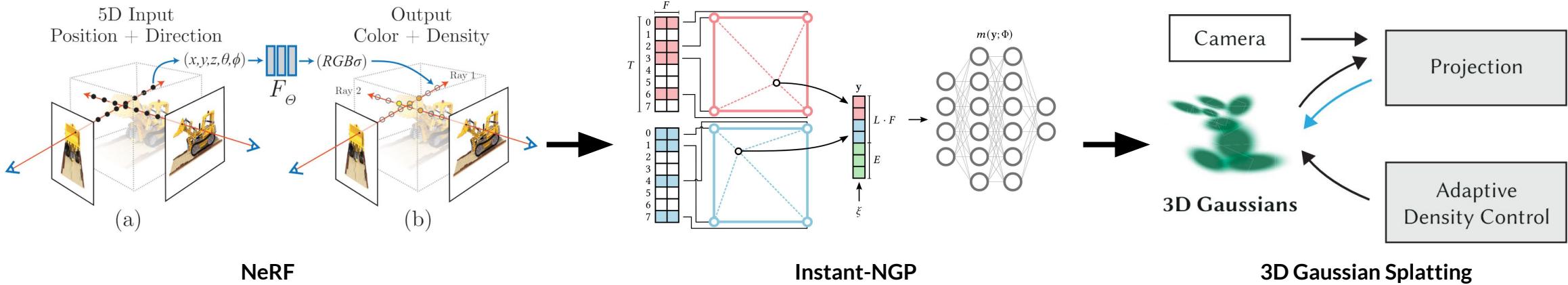
Quality or speed advancements in non-rigid setting follows the advancements in rigid setting:



Speed and Quality Advancements

Seminal Works in 3D Rigid Reconstruction and View Synthesis

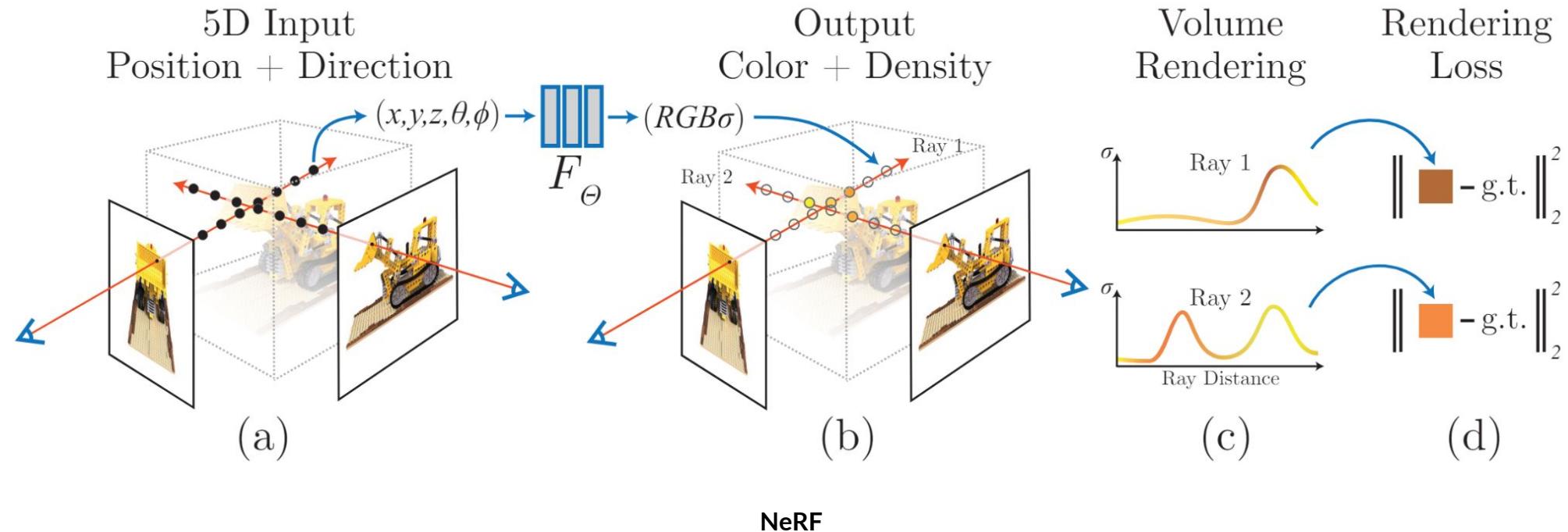
Quality or speed advancements in non-rigid setting follows the advancements in rigid setting:



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

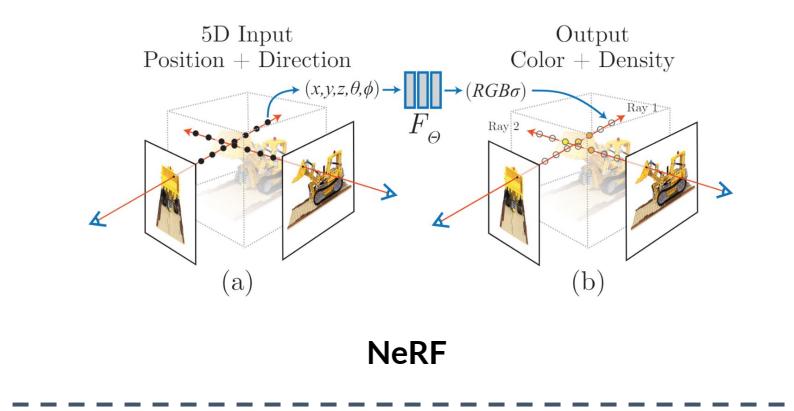
Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



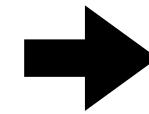
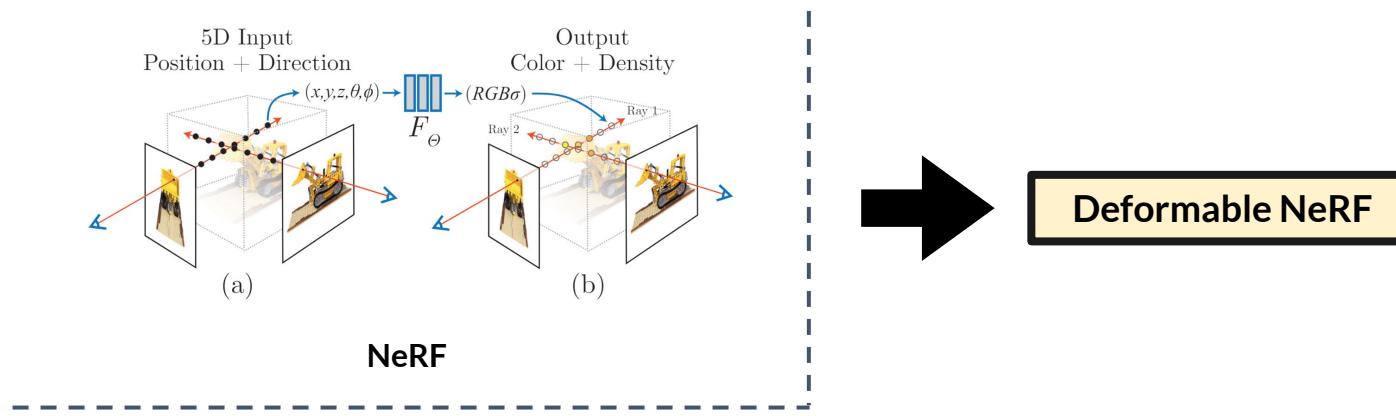
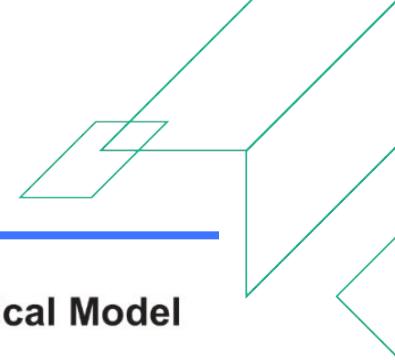
Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



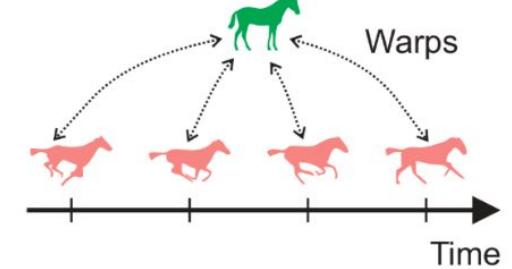
Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



Deformable NeRF

Global Canonical Model

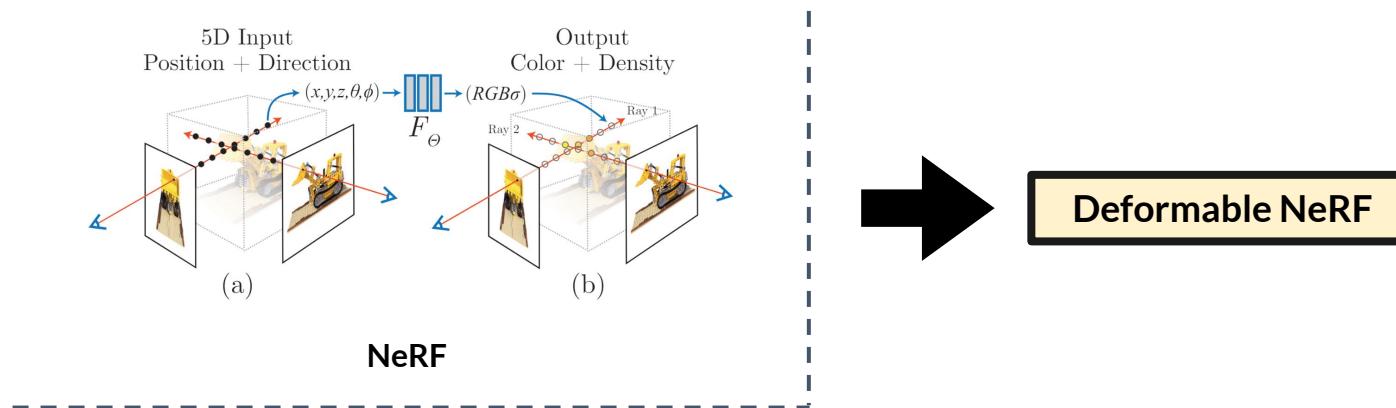
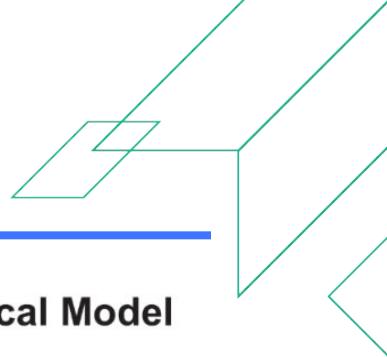


Nerfies



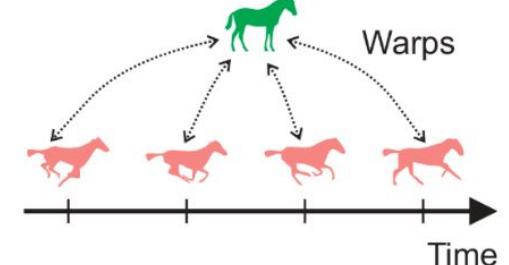
Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



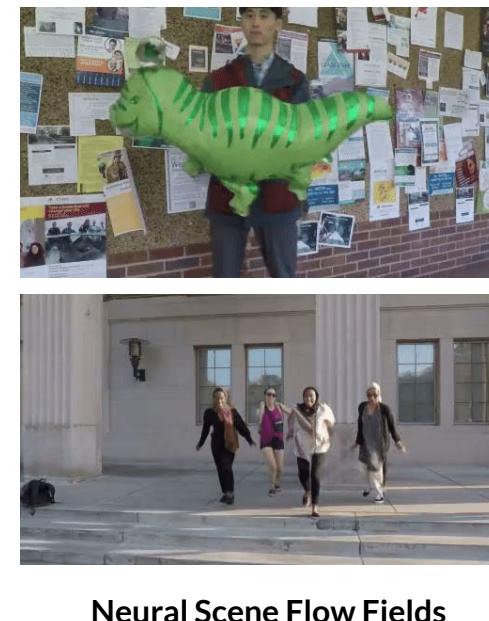
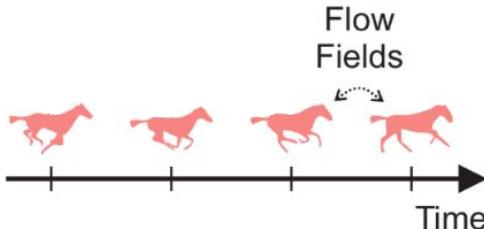
Deformable NeRF

Global Canonical Model



Space-Time NeRF

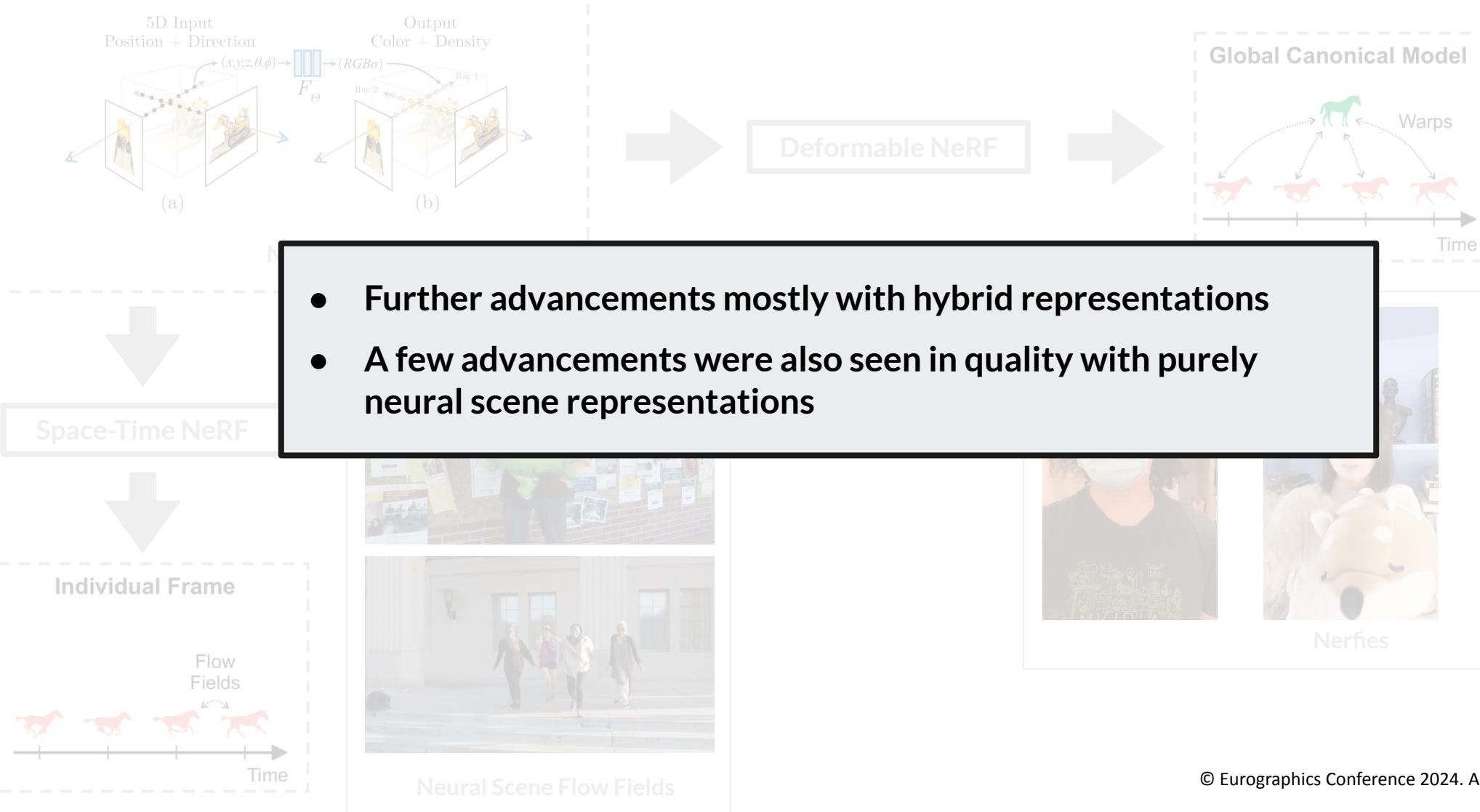
Individual Frame



Nerfies

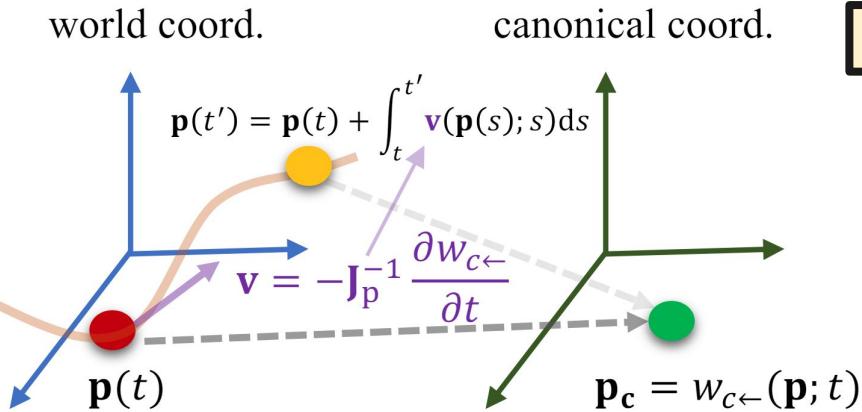
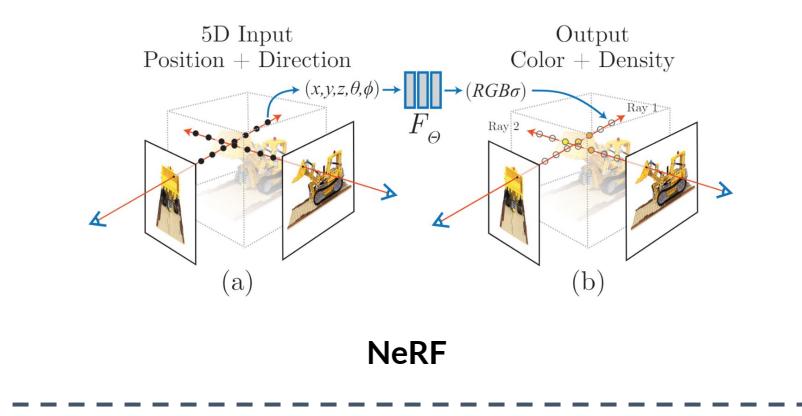
Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations

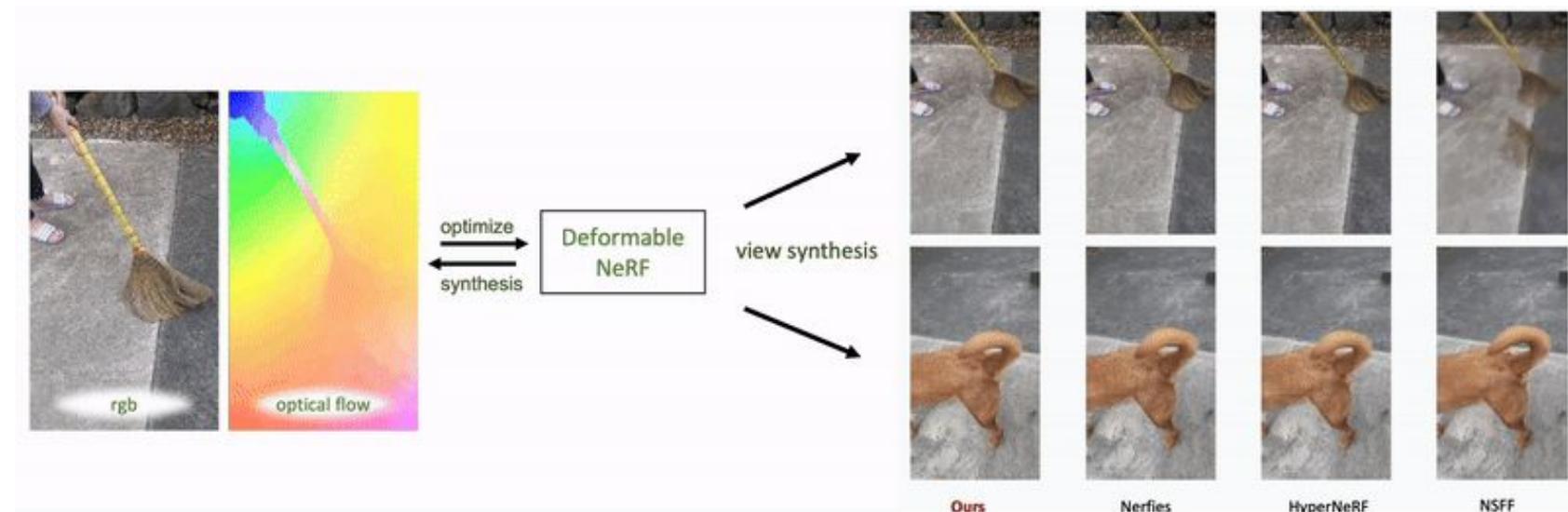


Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



Optical Flow Supervision:

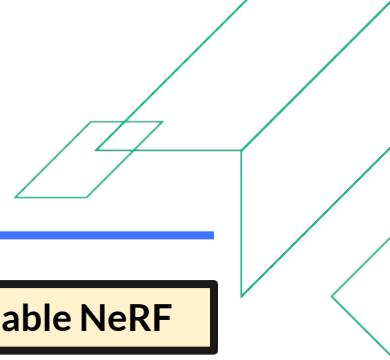


FSD-NeRF

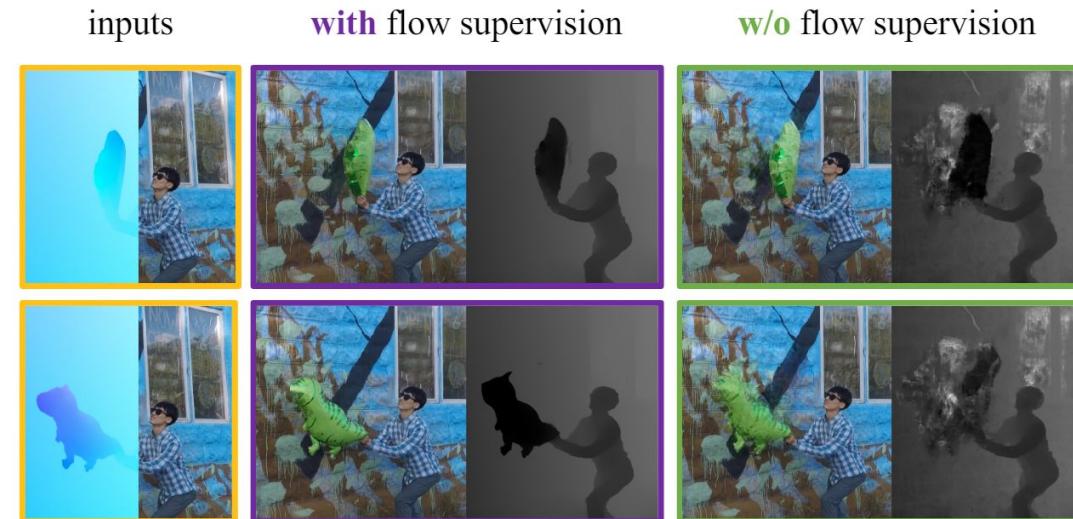
© Eurographics Conference 2024. All rights reserved.

Speed and Quality Advancements

Photo-realistic View Synthesis: Neural Scene Representations



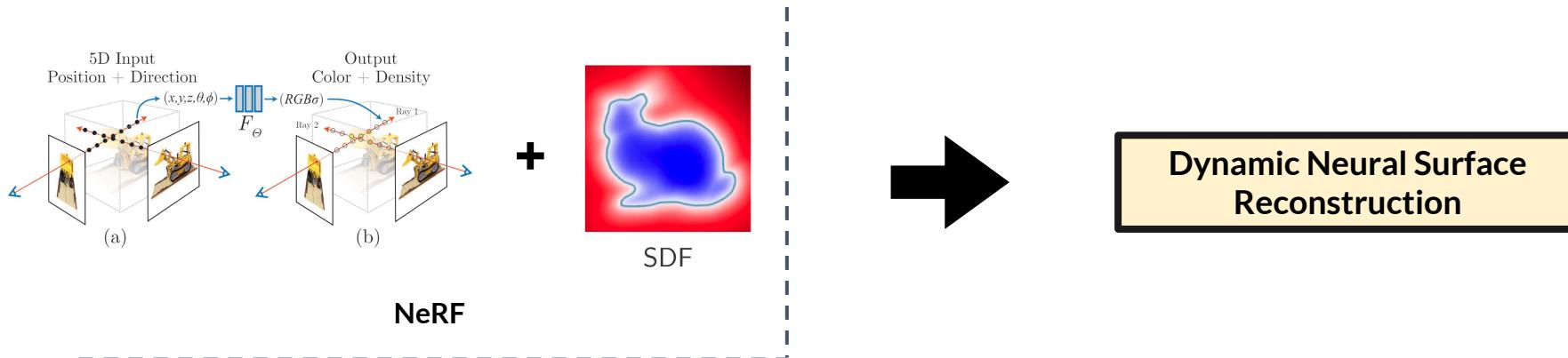
Optical Flow Supervision:



FSD-NeRF

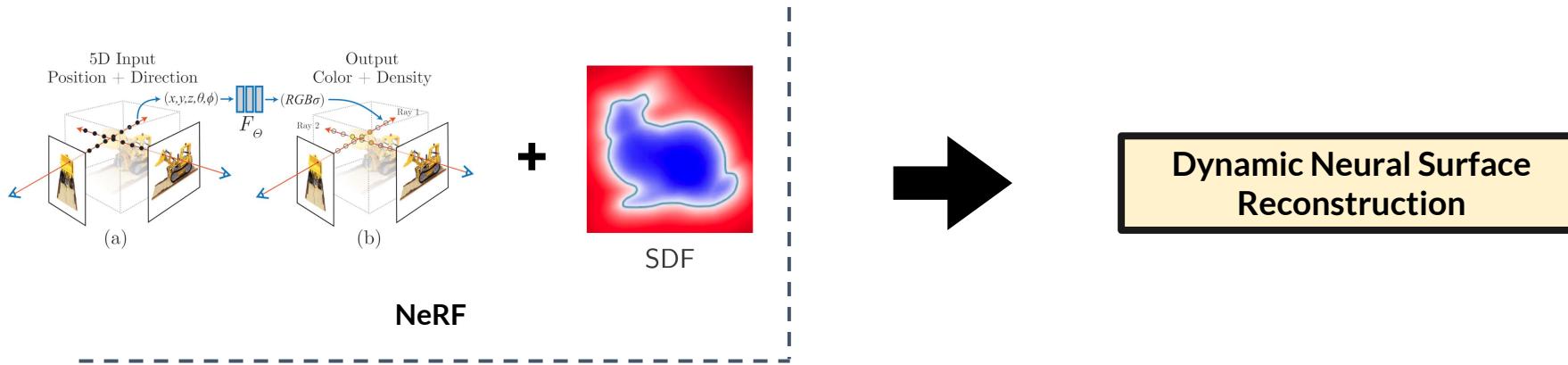
Speed and Quality Advancements

High-fidelity Geometry: Neural Scene Representations

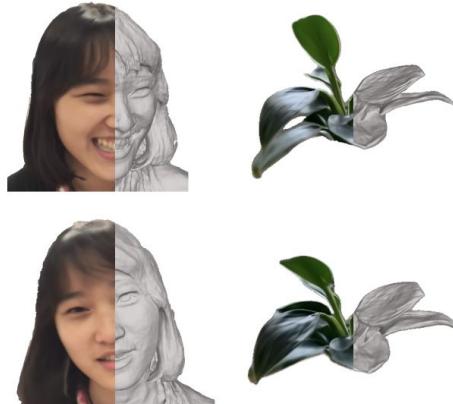


Speed and Quality Advancements

High-fidelity Geometry: Neural Scene Representations

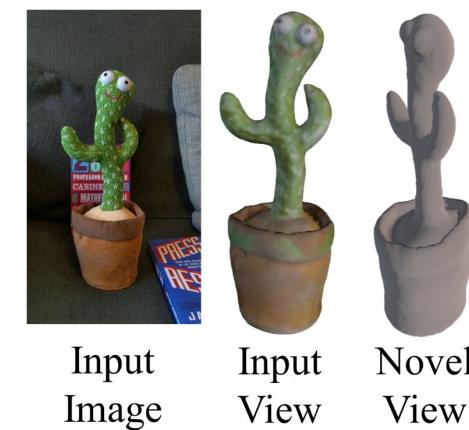


- RGB-D with mask



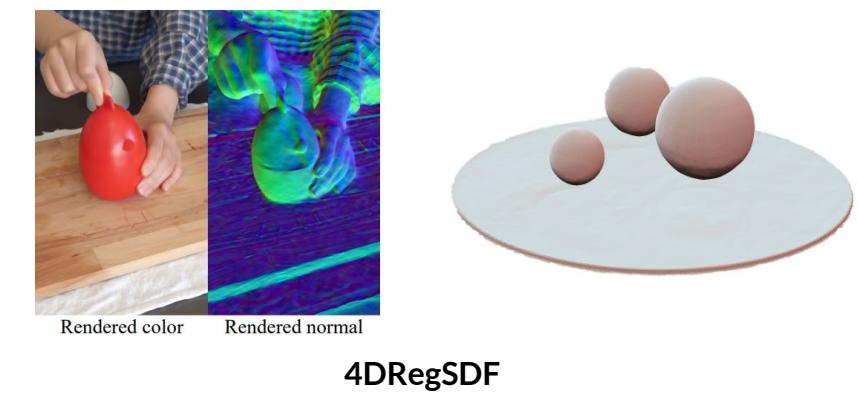
NDR

- RGB with mask and mesh proxy



Unbiased4D

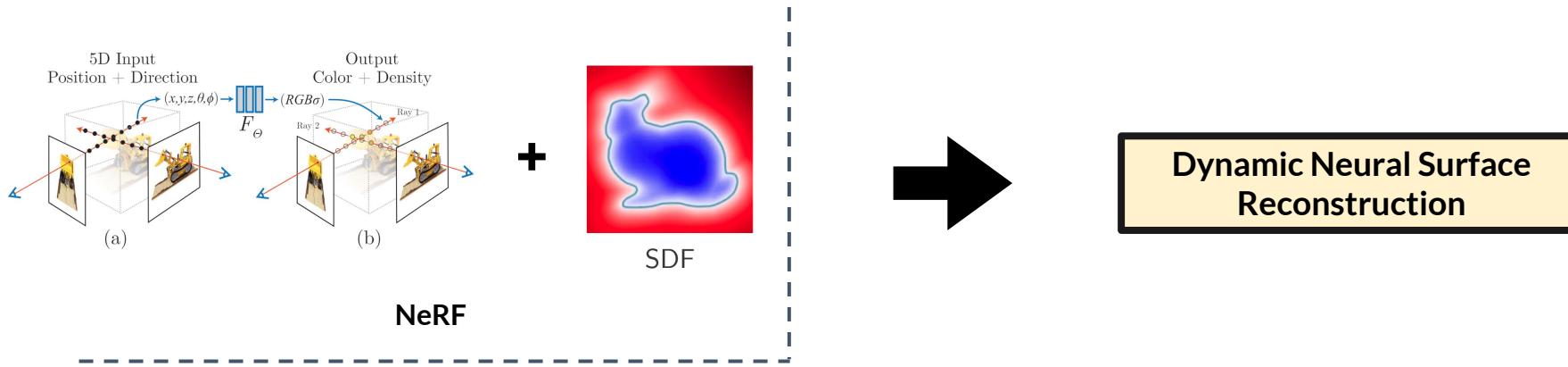
- RGB only



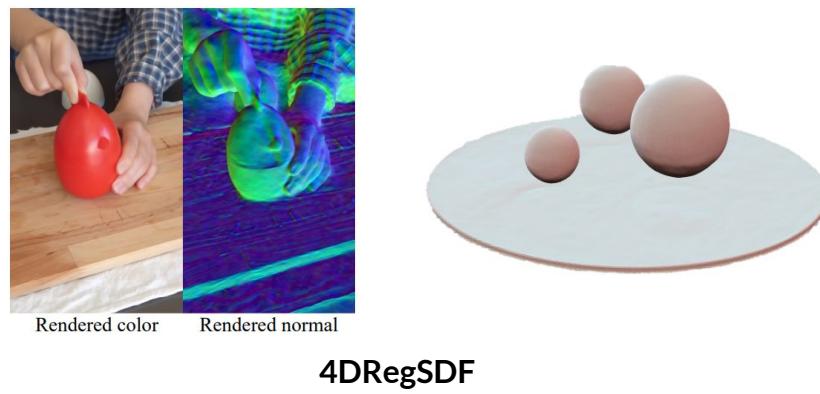
4DRegSDF

Speed and Quality Advancements

High-fidelity Geometry: Neural Scene Representations



- RGB only



$$\sum \left(\text{Local rigidity constraint} \right) : \text{enforce local rigidity}$$

Total variation of curvature

$$\sum \left(\text{Smoothness constraint} \right) : \text{limit unnecessary kinks}$$

Absolute curvature of SDF

$$\sum \left(\text{Eikonal loss} \right) : \text{make gradient of SDF valid}$$

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

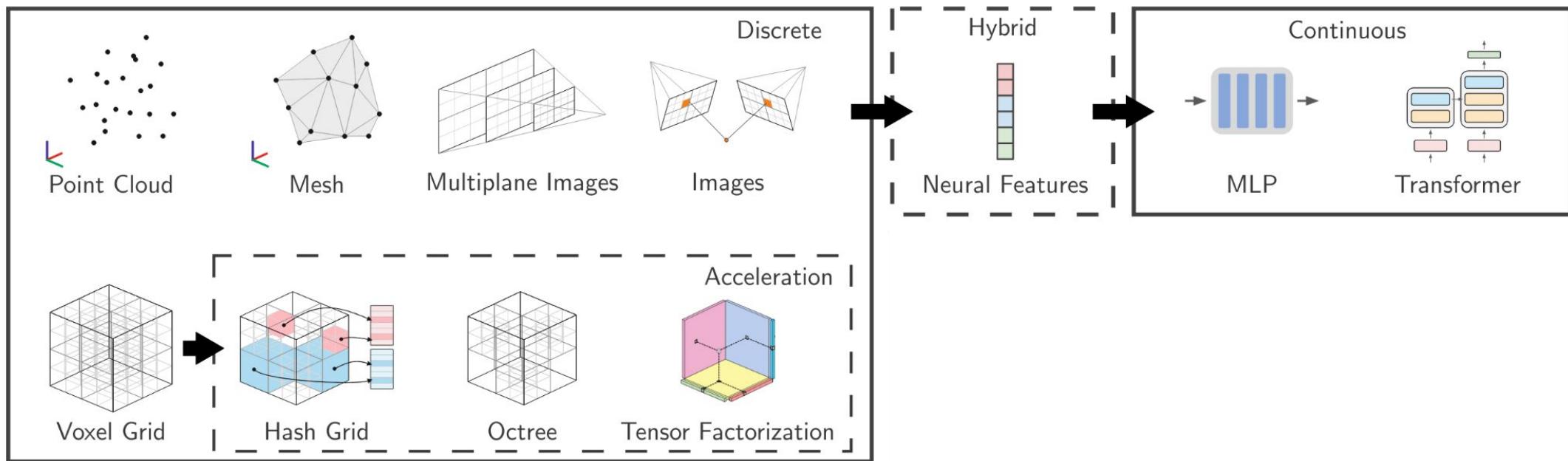
• RGB with mask and
mesh proxy

• RGB only

Surface Reconstruction

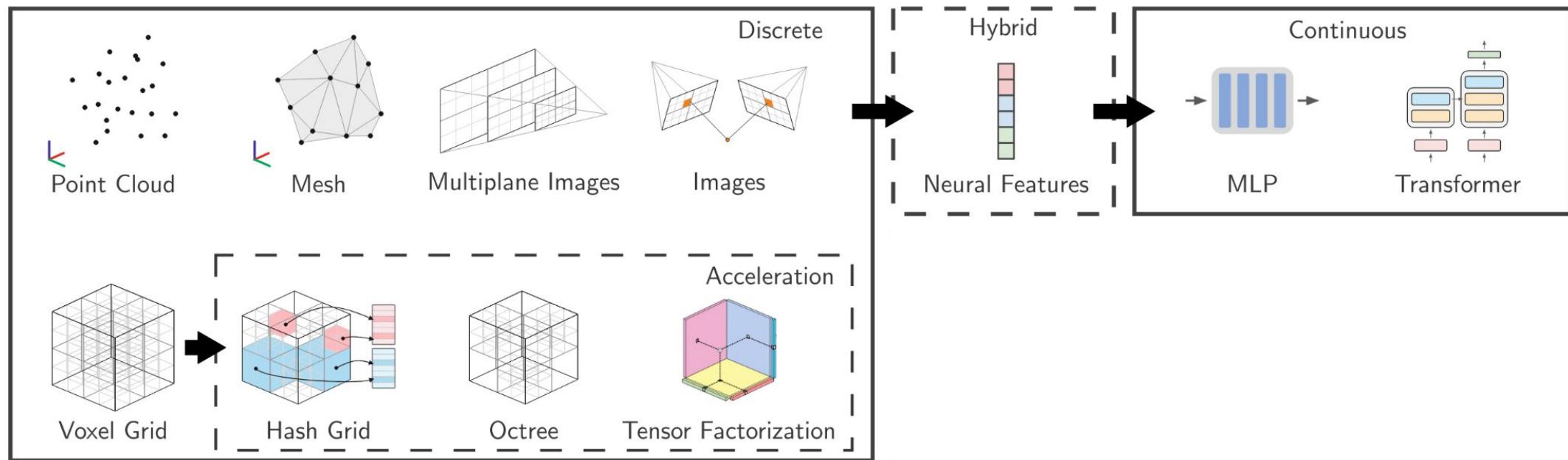
Speed and Quality Advancements

3D Scene Representations



Speed and Quality Advancements

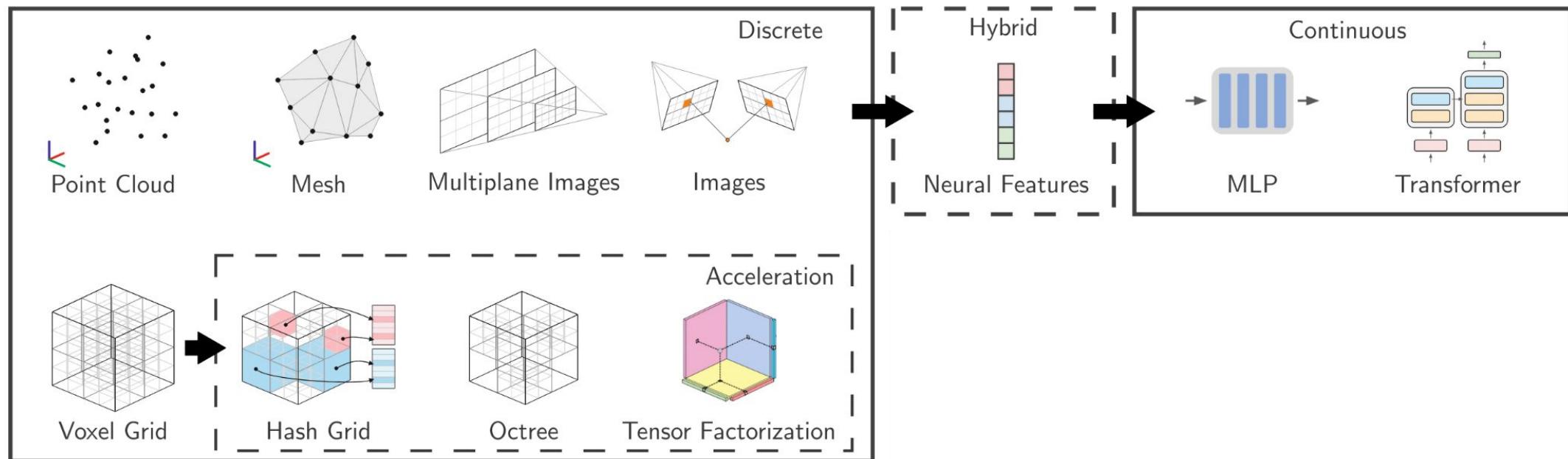
3D Scene Representations



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

Speed and Quality Advancements

3D Scene Representations

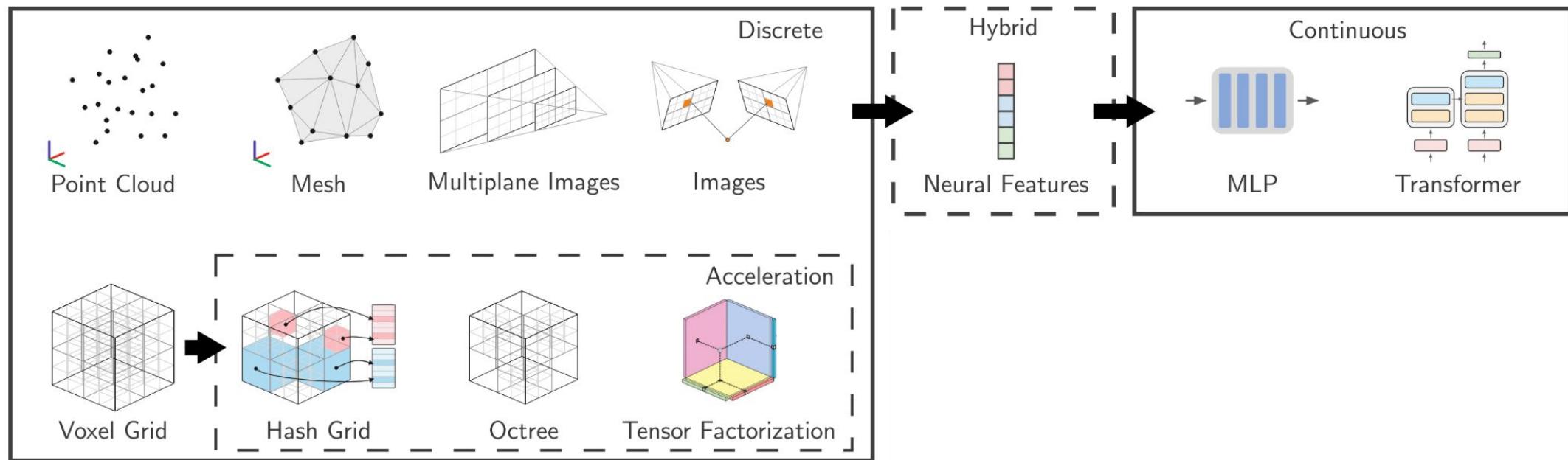


$\mathbf{x} \in \mathbb{R}^3$ are the 3D coordinates

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

Speed and Quality Advancements

3D Scene Representations

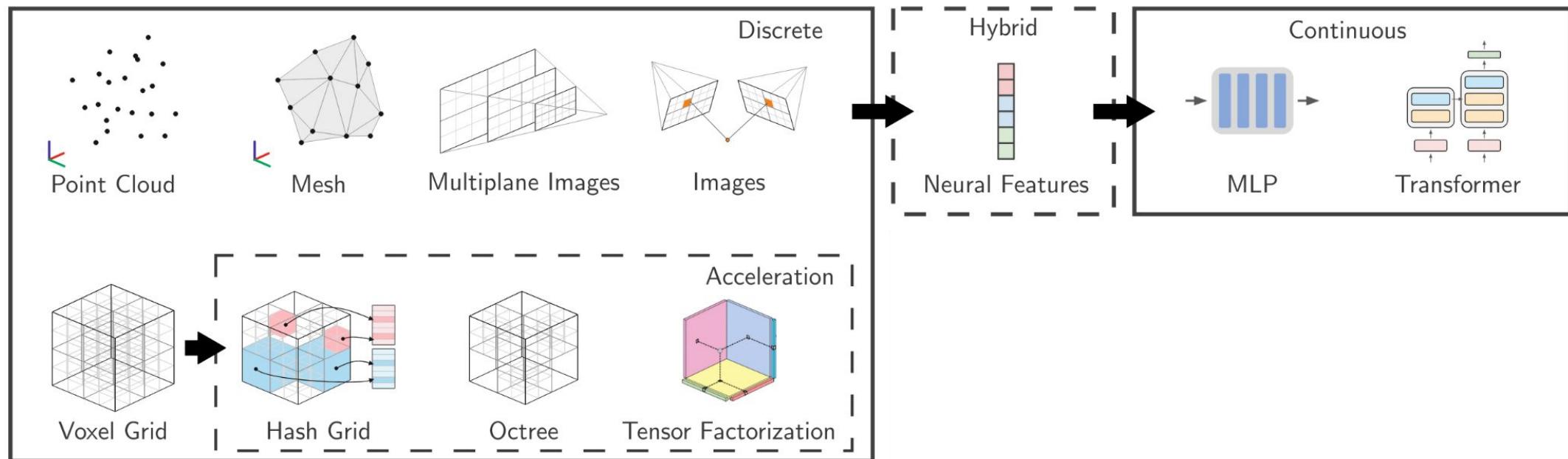


$\mathbf{x} \in \mathbb{R}^3$ are the 3D coordinates

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

Speed and Quality Advancements

3D Scene Representations

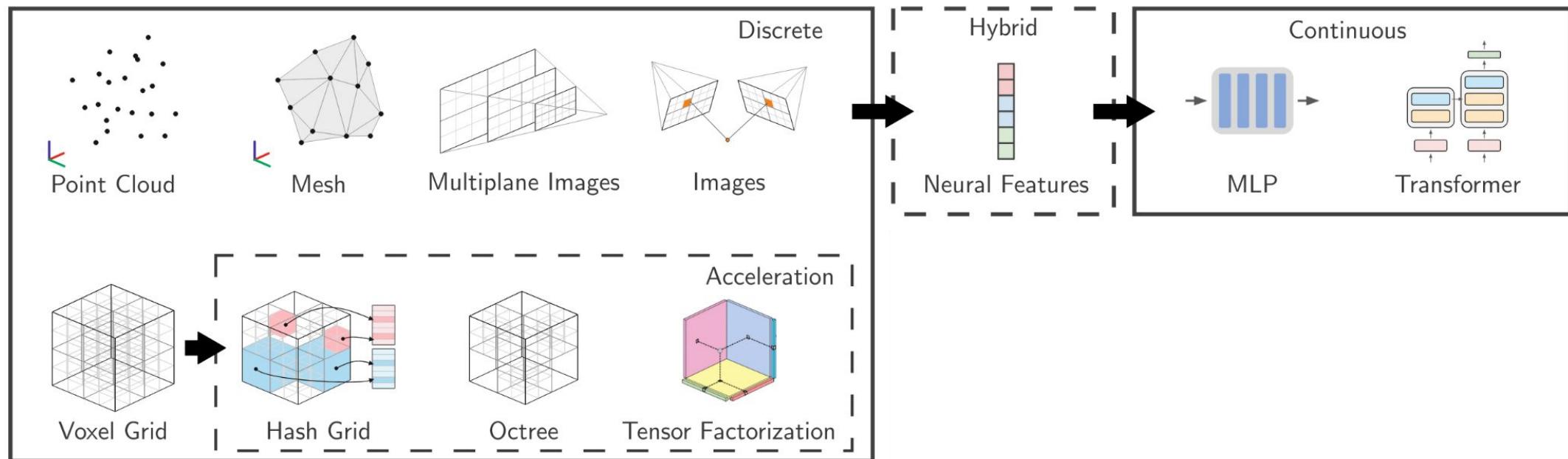


$\mathbf{x} \in \mathbb{R}^3$ are the 3D coordinates

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

Speed and Quality Advancements

3D Scene Representations



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

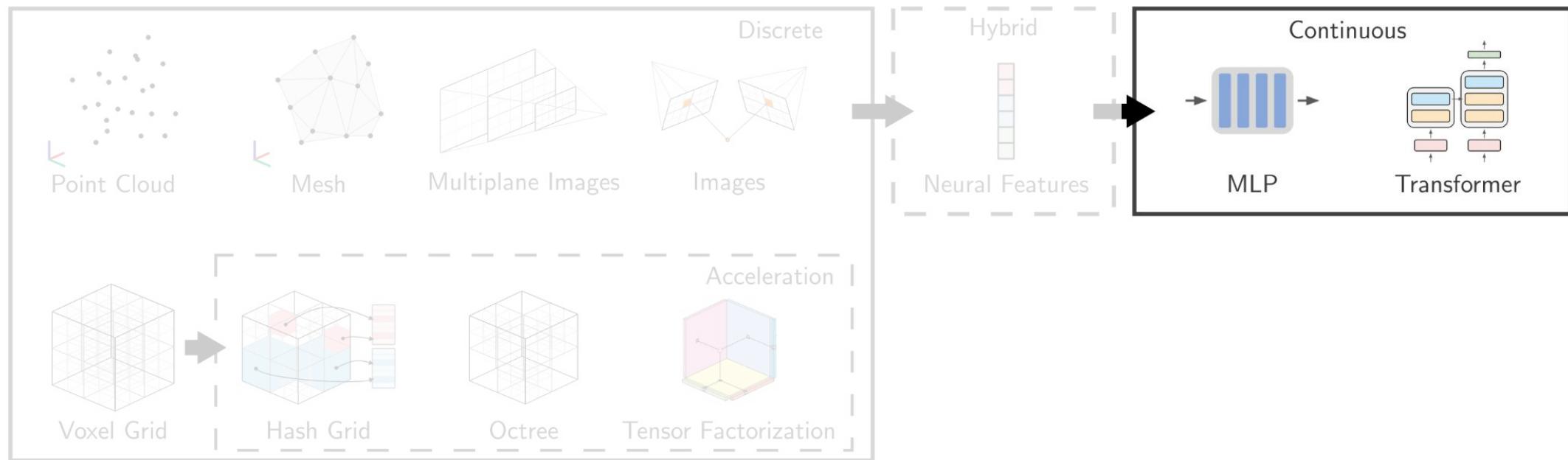
$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta), \quad \mathcal{H}$ are optional additional inputs (e.g. view direction)

θ stores the scene information

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

Speed and Quality Advancements

3D Scene Representations

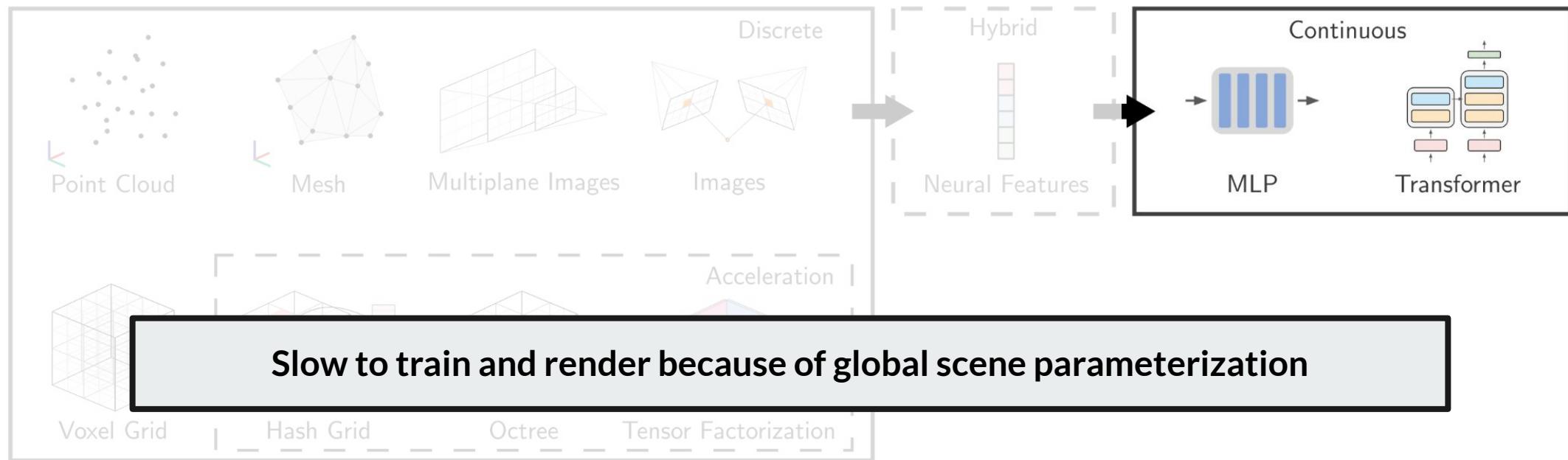


$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta), \quad \begin{matrix} \theta \\ \rho \end{matrix} \text{ is stored in network parameters}$$

θ is stored in network parameters
 ρ is an MLP or Transformer

Speed and Quality Advancements

3D Scene Representations

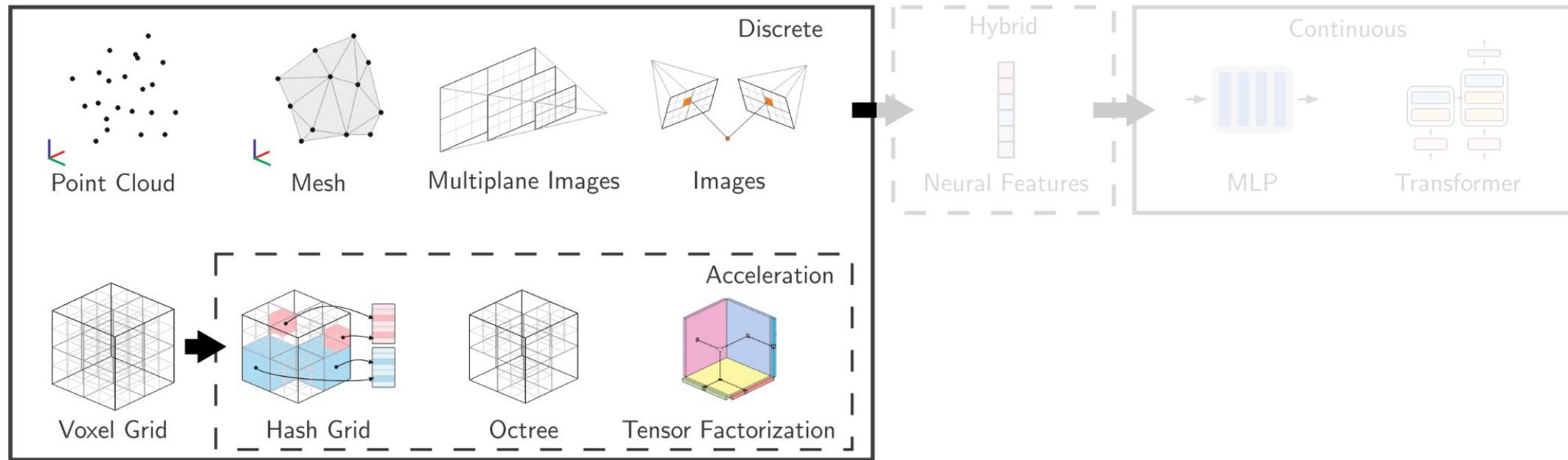


$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta), \quad \begin{matrix} \theta \\ \rho \end{matrix} \text{ is stored in network parameters}$$

θ is stored in network parameters
 ρ is an MLP or Transformer

Speed and Quality Advancements

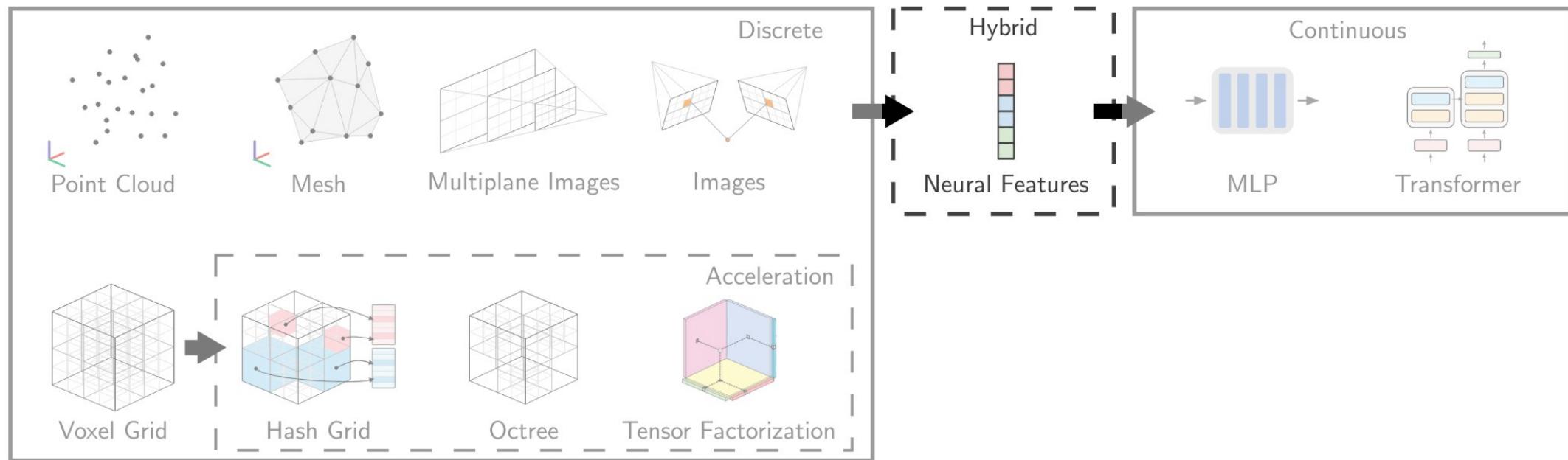
3D Scene Representations



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

Speed and Quality Advancements

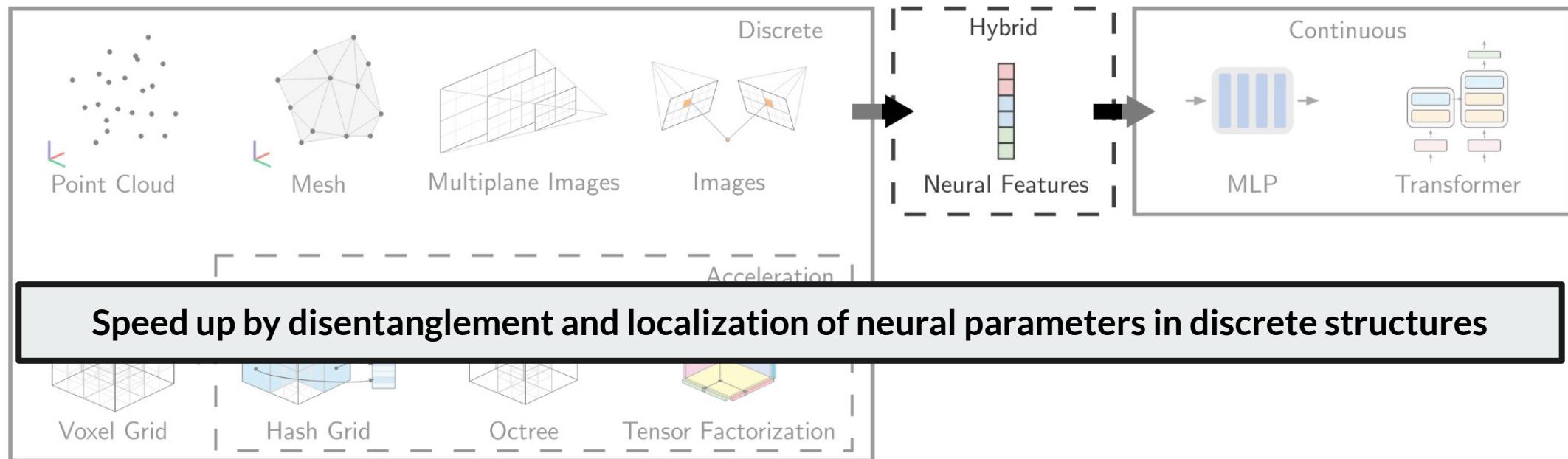
3D Scene Representations



$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta)$, θ are neural features stored in a discrete structure
 ρ defines interpolation of discrete information followed by network query

Speed and Quality Advancements

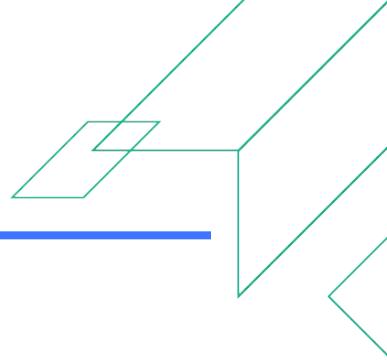
3D Scene Representations


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

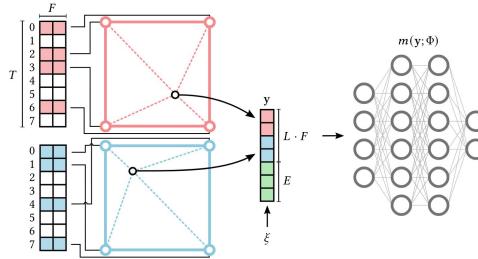
are neural features stored in a discrete structure
defines interpolation of discrete information followed by network query

Speed and Quality Advancements

Seminal Hybrid Scene Representations for Rigid Setting

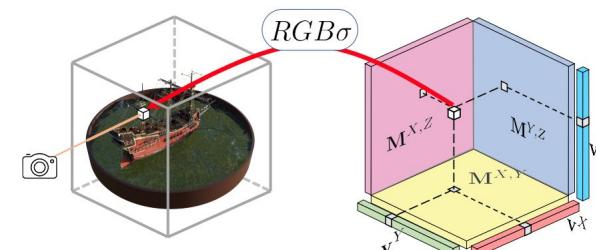


Voxel Grid



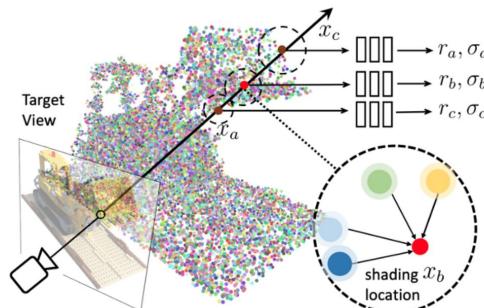
Instant-NGP

Planar Factorization



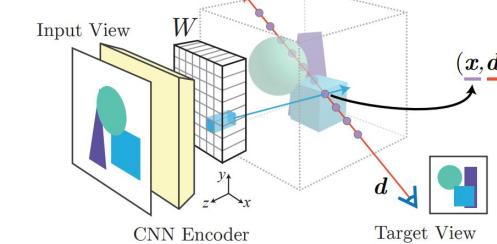
TensoRF

Points



Point-NeRF

Image-based

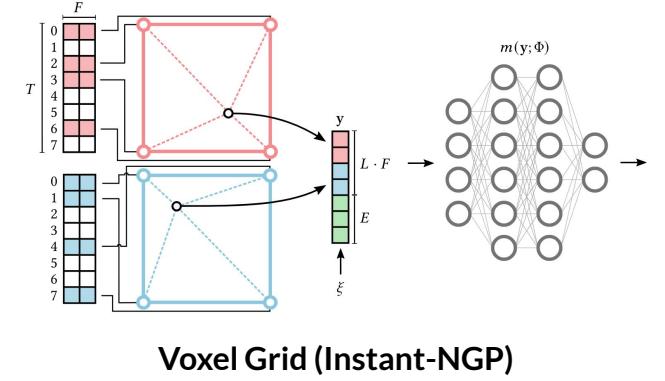


pixelNeRF



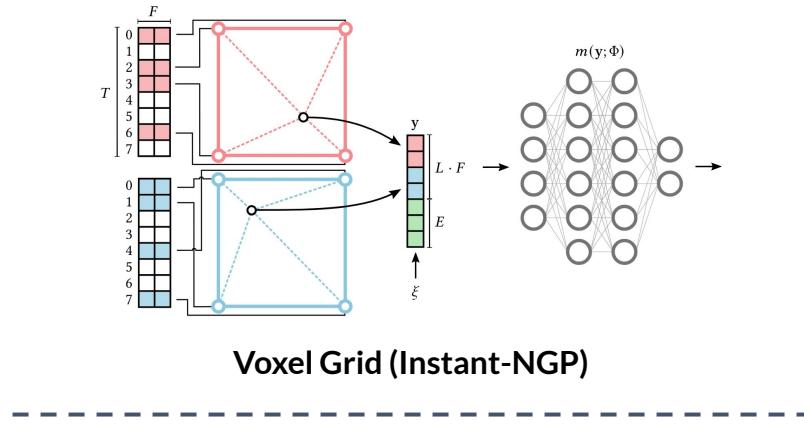
Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations



Speed and Quality Advancements

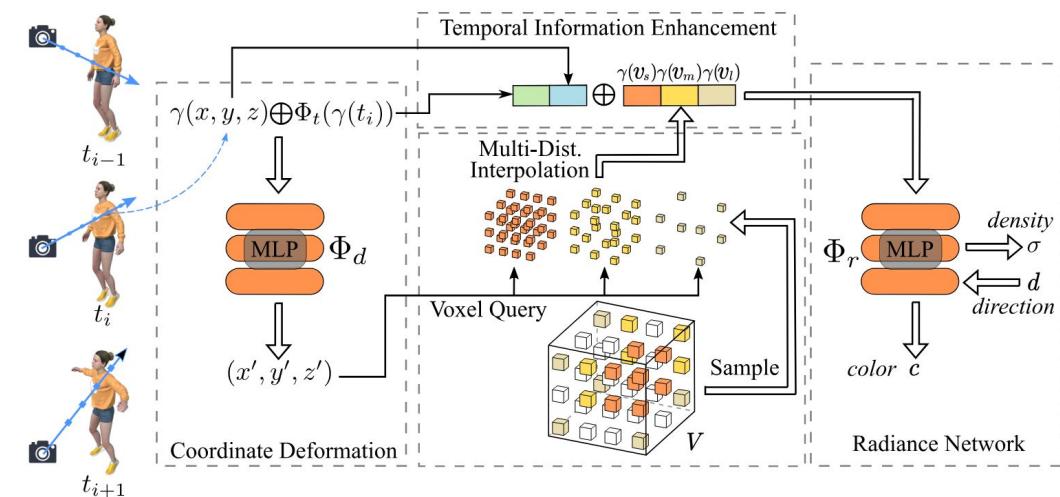
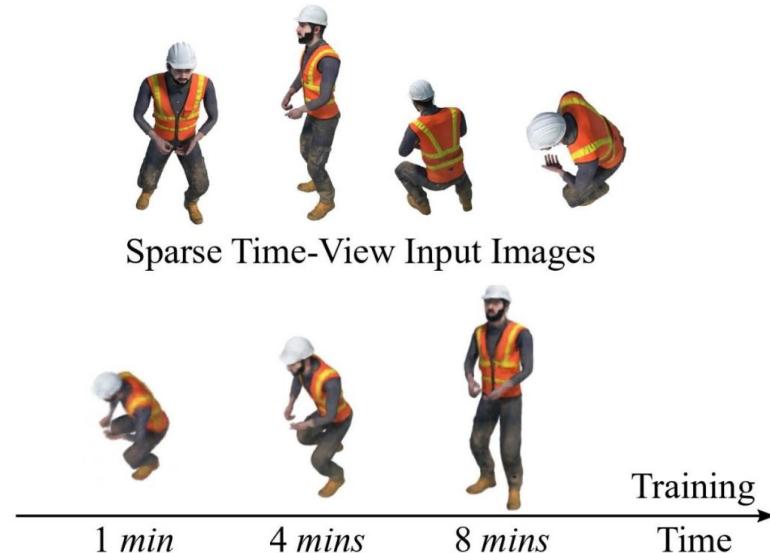
Faster Training and Rendering: Hybrid Neural Scene Representations



Voxel Grid (Instant-NGP)

Dynamic Voxel NeRF

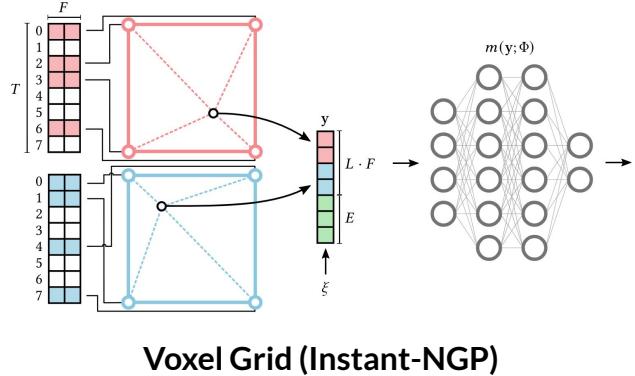
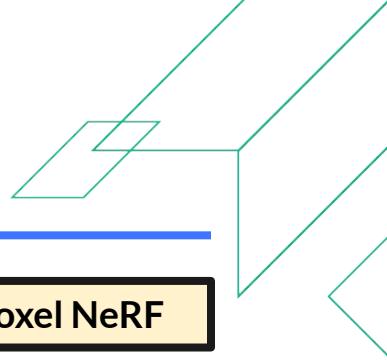
- Both the deformation field and canonical space are parameterized by MLPs with voxel grids
- Very light deformation MLP for fast training
- Canonical radiance field enhanced through temporal embeddings



TiNeuVox

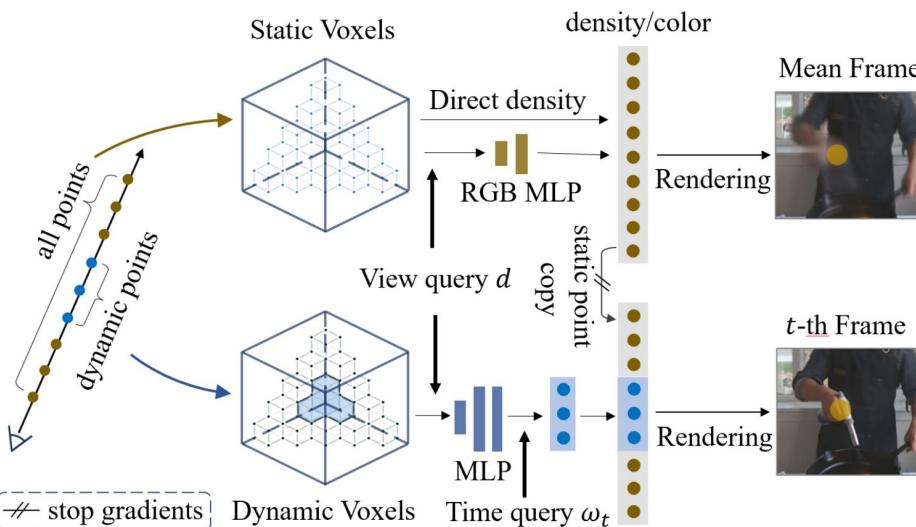
Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations



Dynamic Voxel NeRF

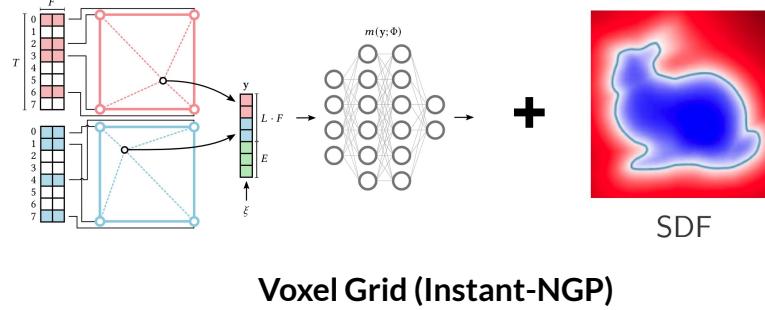
- Uses static-dynamic decomposition for better scalability and motion modelling capacity
- Lightweight static model for fast training from multi-view videos and near real-time rendering



MixVoxels

Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations



Voxel Grid (Instant-NGP)

Dynamic Neural Surface Reconstruction

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

Novel view synthesis Geometry reconstruction

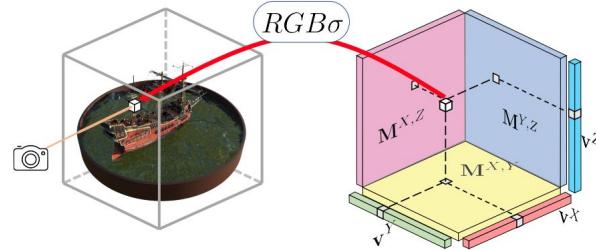
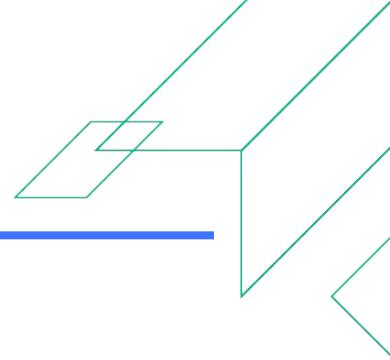


Reference images

NeuS2

Speed and Quality Advancements

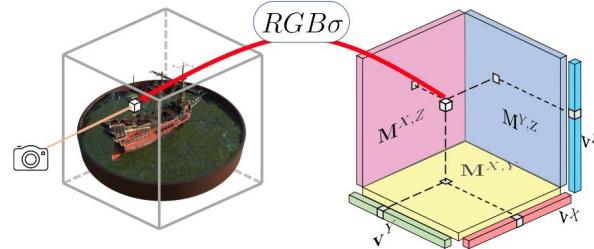
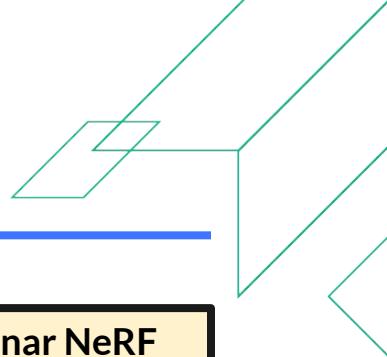
Faster Training and Rendering: Hybrid Neural Scene Representations



Planar Factorization (TensoRF)

Speed and Quality Advancements

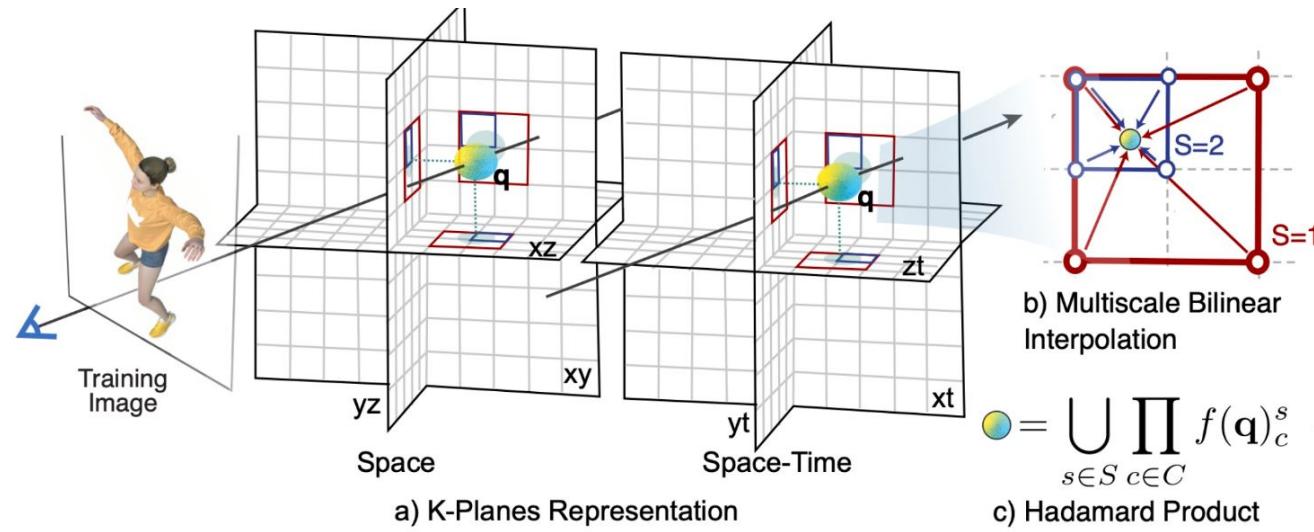
Faster Training and Rendering: Hybrid Neural Scene Representations



Planar Factorization (TensoRF)

Space-Time Planar NeRF

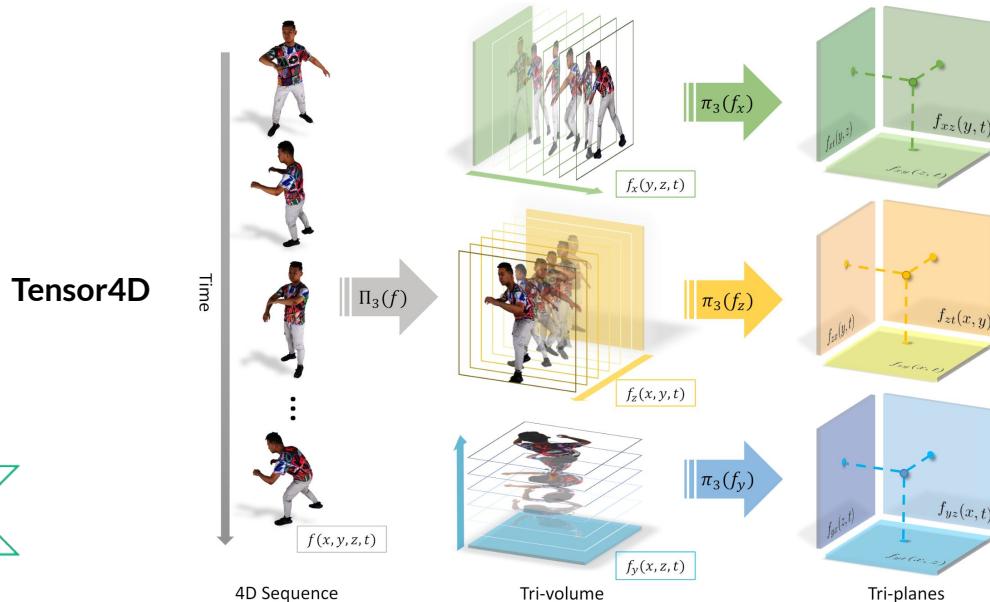
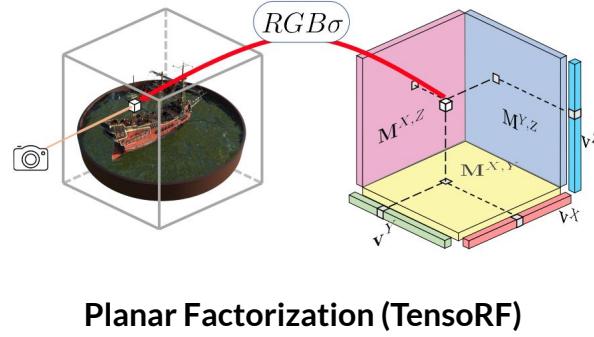
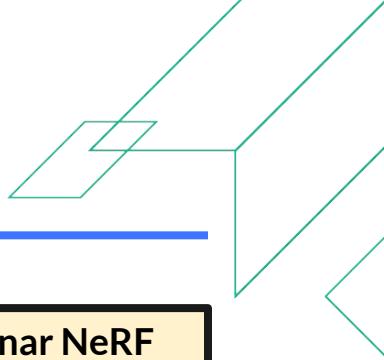
Compaction of voxel grid
for memory efficiency



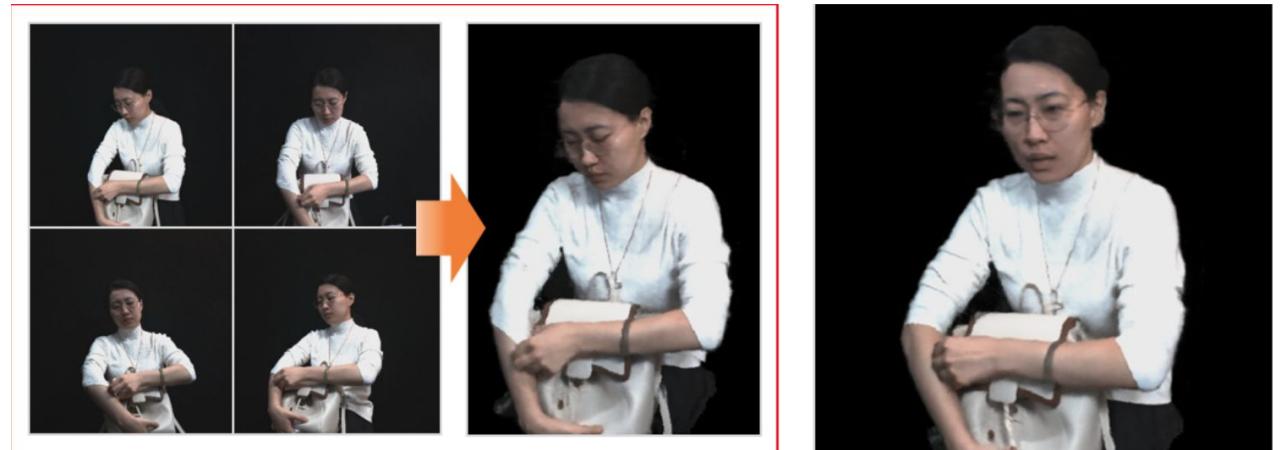
K-Planes

Speed and Quality Advancements

Faster Training and Rendering: Hybrid Neural Scene Representations



- 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

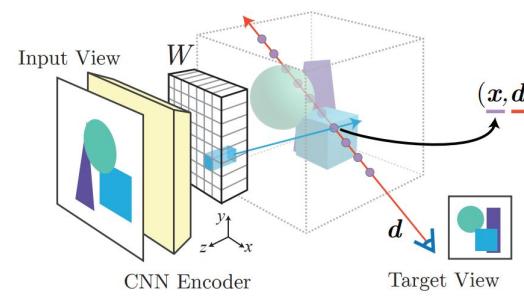
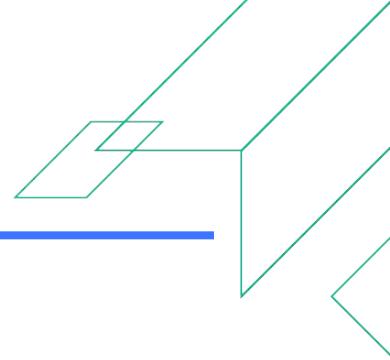


Image-based (pixelNeRF)

Speed and Quality Advancements

High-quality Rendering: Hybrid Neural Scene Representations

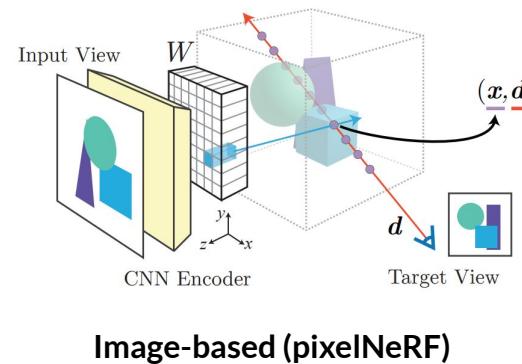
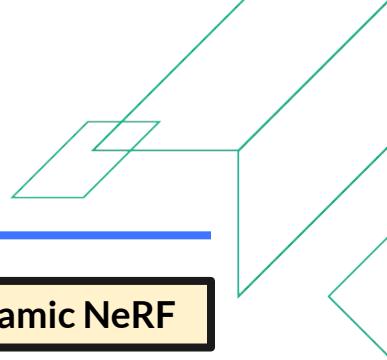
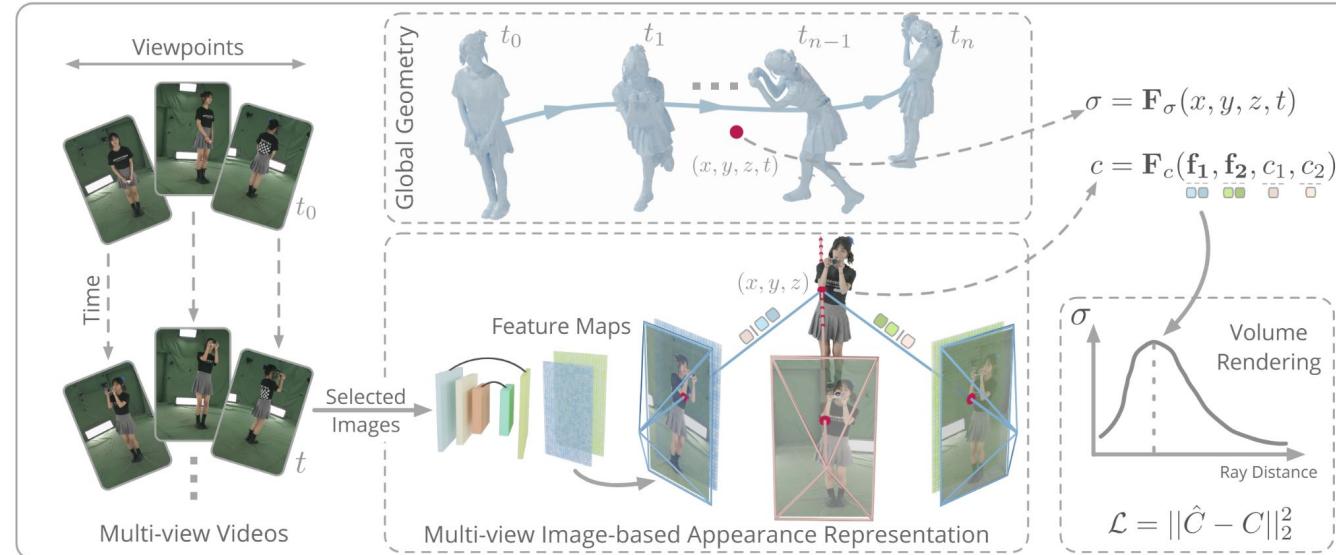


Image-based (pixelNeRF)

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!



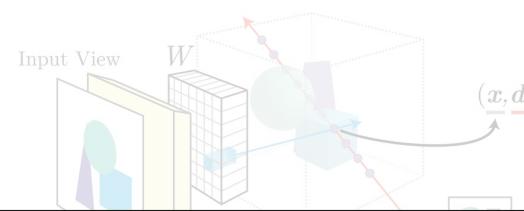
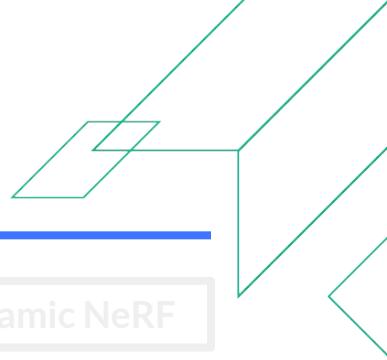
Im4D



4K4D

Speed and Quality Advancements

High-quality Rendering: Hybrid Neural Scene Representations

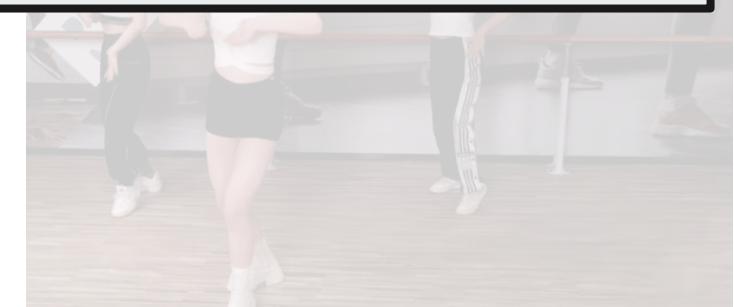
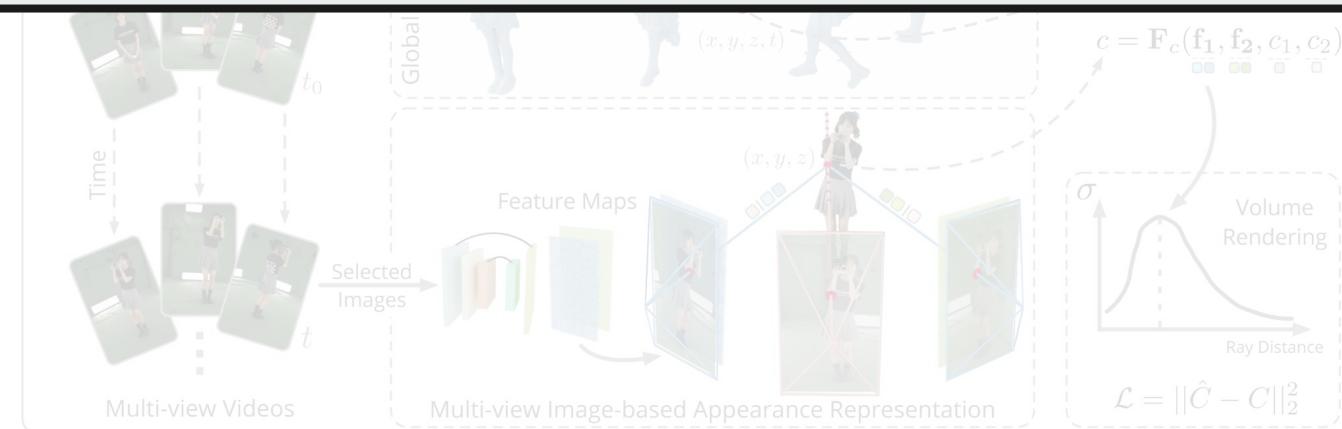


Discriminative image
features for high-quality

Image-based Dynamic NeRF

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



Speed and Quality Advancements

High-quality Rendering: Hybrid Neural Scene Representations

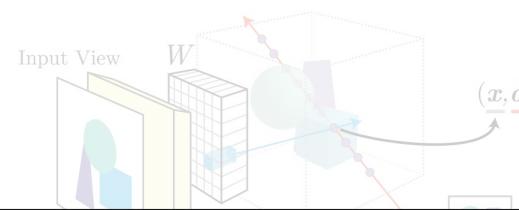
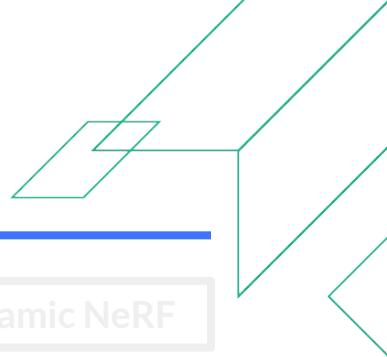


Image-based Dynamic NeRF

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!

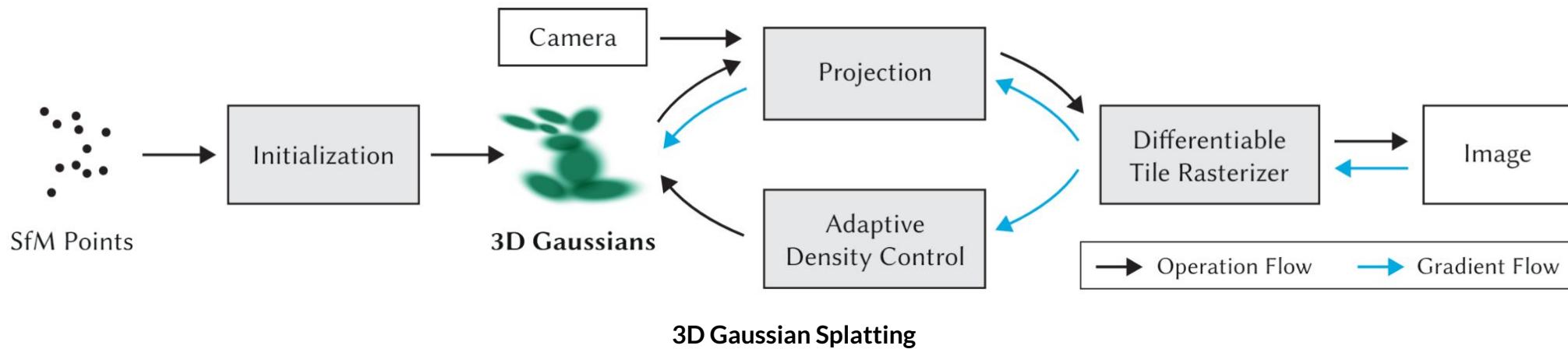
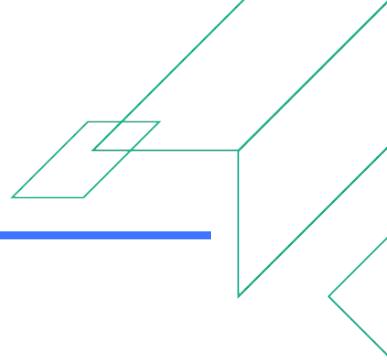


Im4D

© Eurographics Conference 2024. All rights reserved.

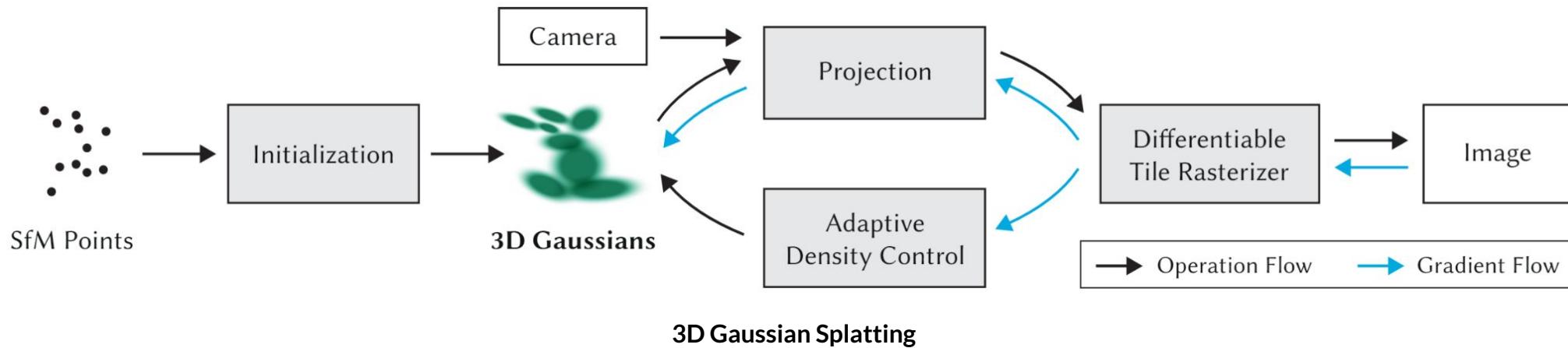
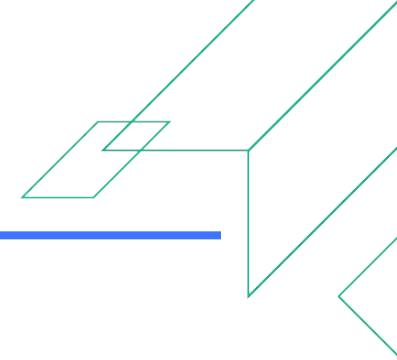
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



Speed and Quality Advancements

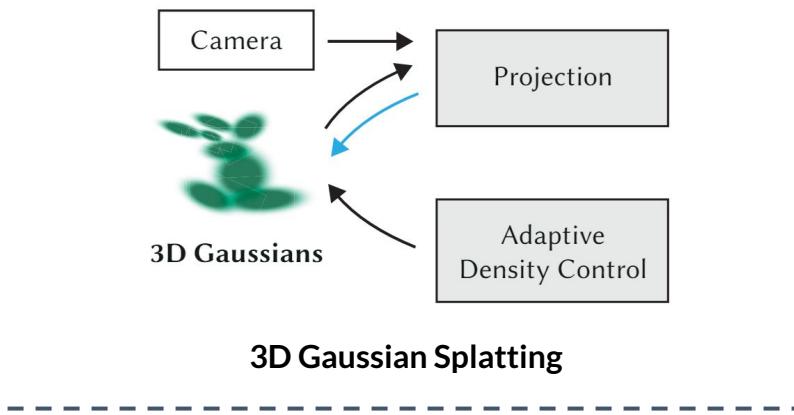
Real-time Rendering: 3D Gaussian Splatting



- 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

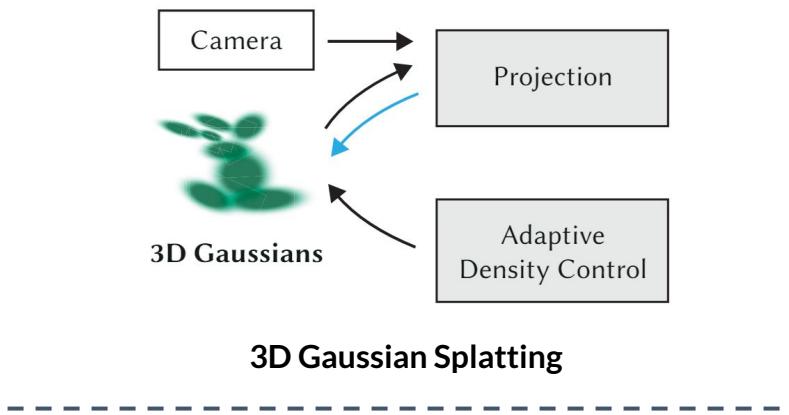
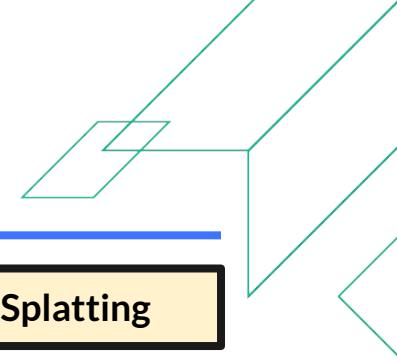
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting

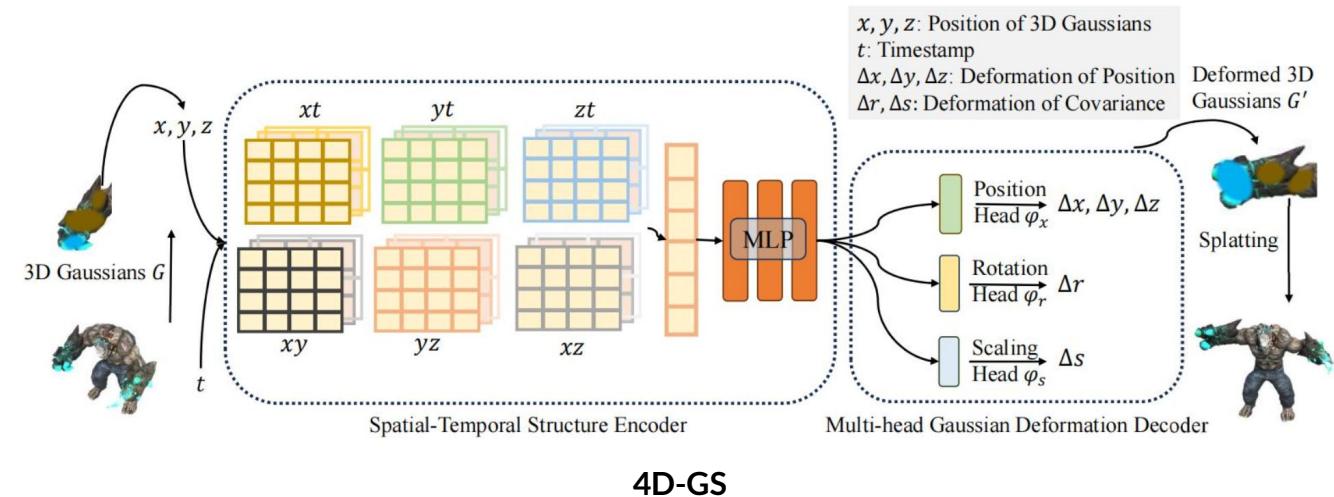


Deformable 3D Gaussian Splatting

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

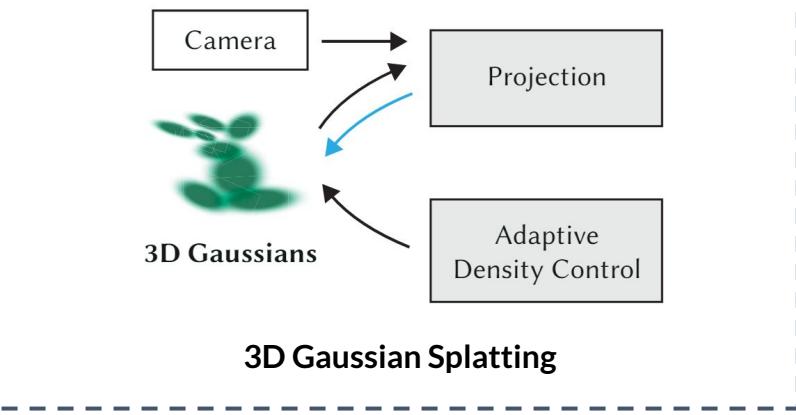


Deformable3DGS



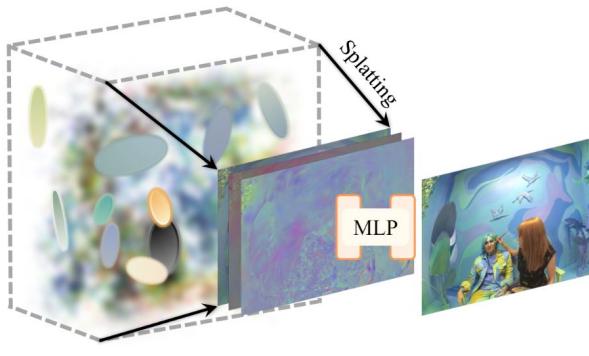
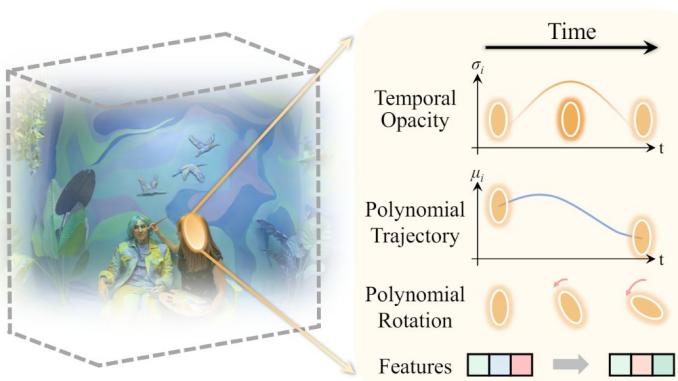
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



Space-time 3D Gaussian Splatting

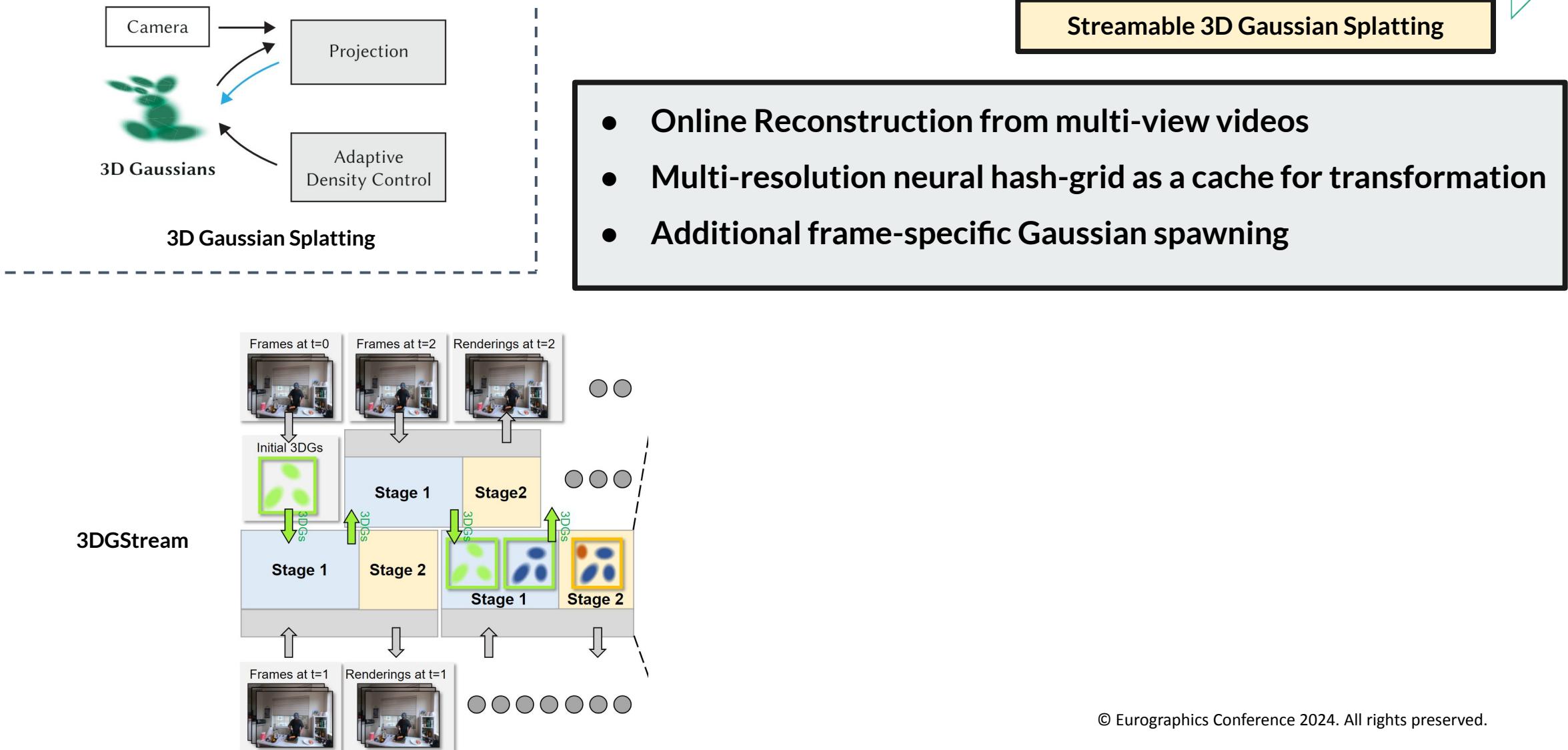
- Extra 1D Gaussian added to 3D Gaussians
- Features instead of SHs, with an MLP to convert them into an RGB image after splatting
- 8K video rendering at 66 FPS!



SpacetimeGaussians

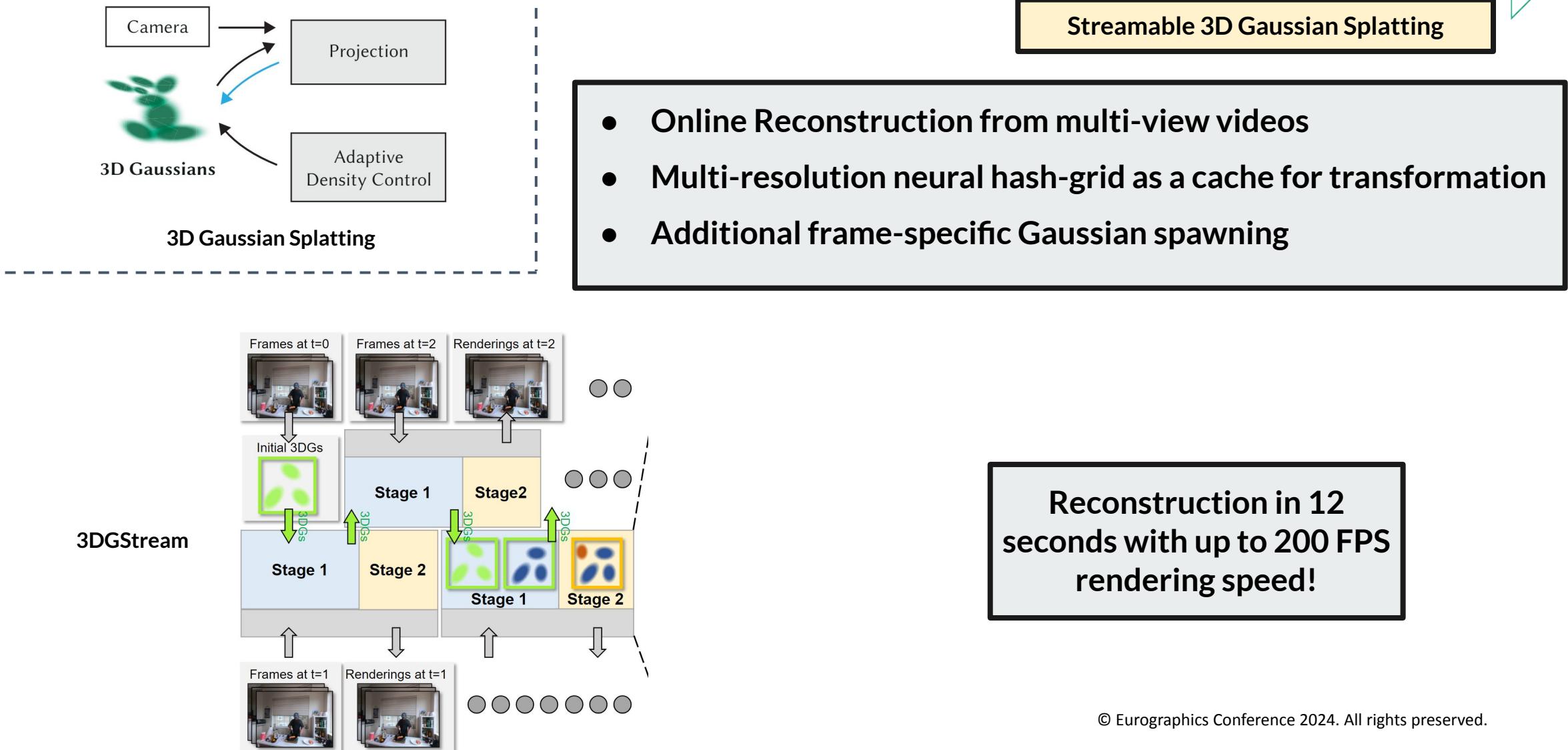
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



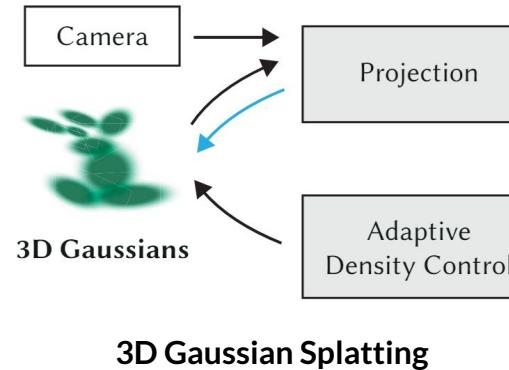
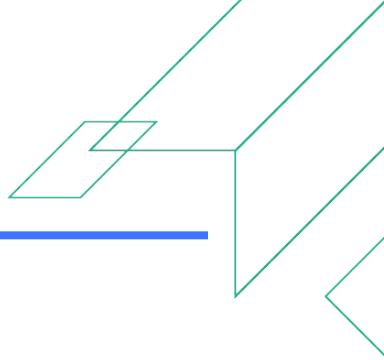
Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting



Speed and Quality Advancements

Real-time Rendering: 3D Gaussian Splatting

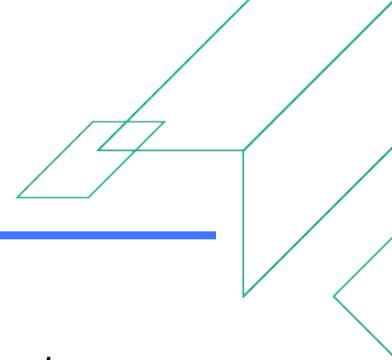


Even though it is getting close, real-time reconstruction
is still only possible using classical representations

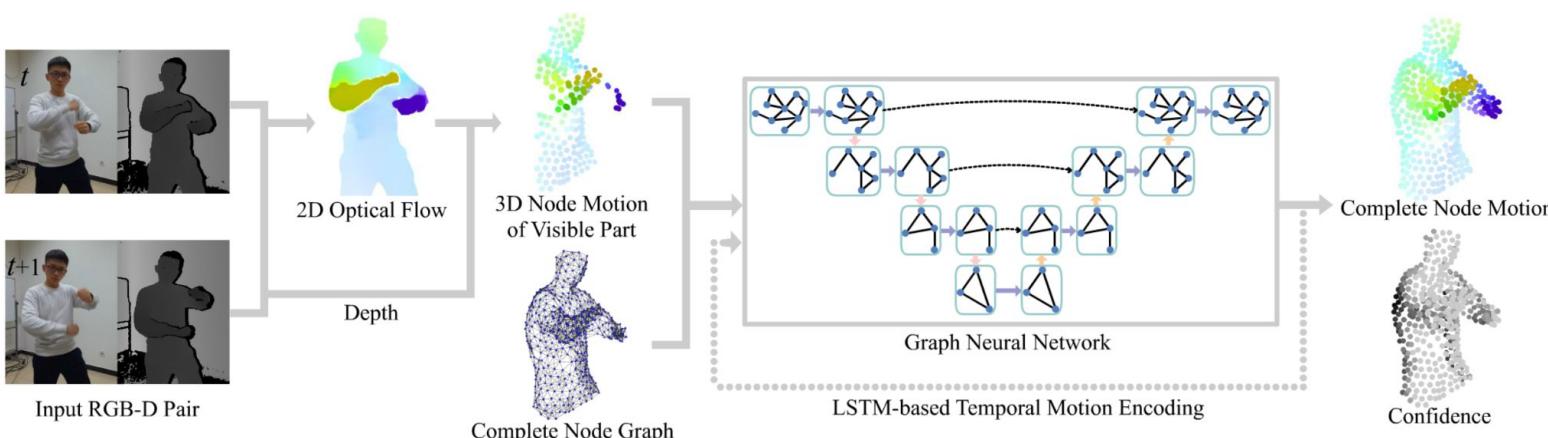


Speed and Quality Advancements

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 Deformations
 3. Modelling Articulated Motion for General Objects
- 5 minute break!**

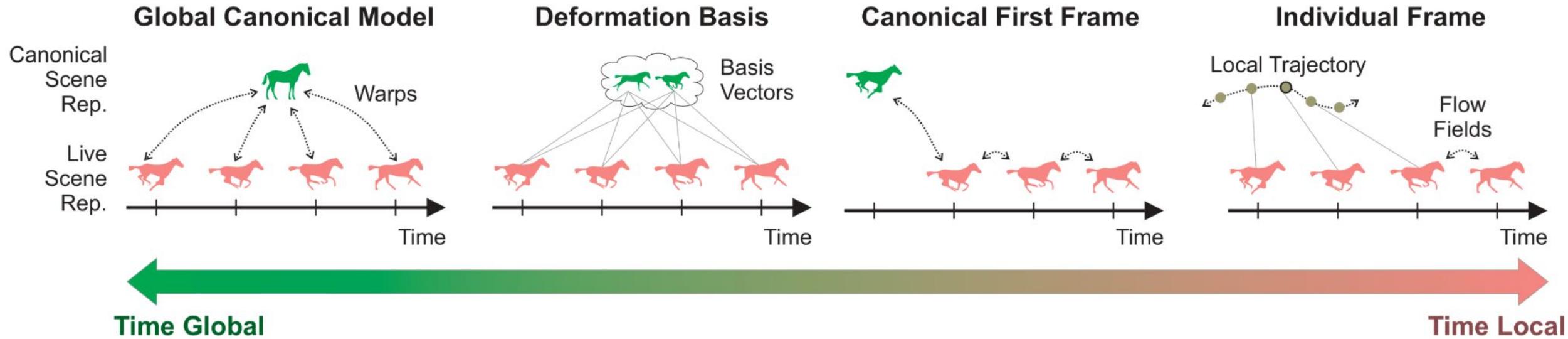
Trends

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

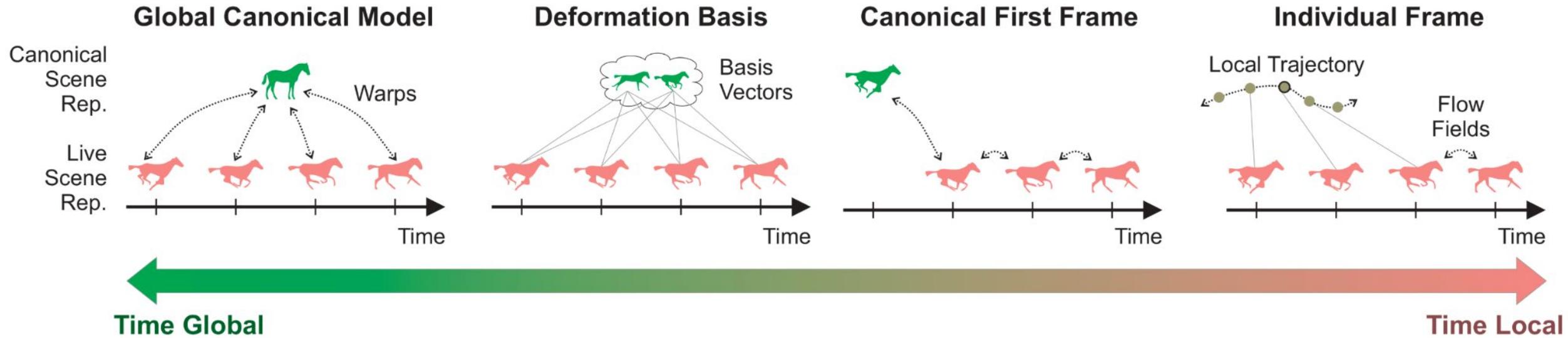
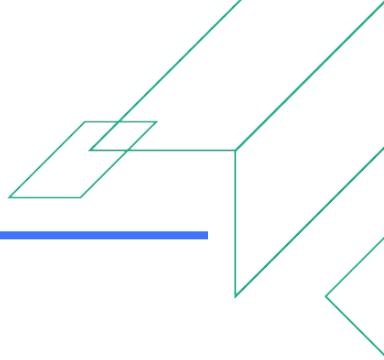
Spatio-Temporal Modelling



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

Large Motion vs. Long-term 3D Correspondences

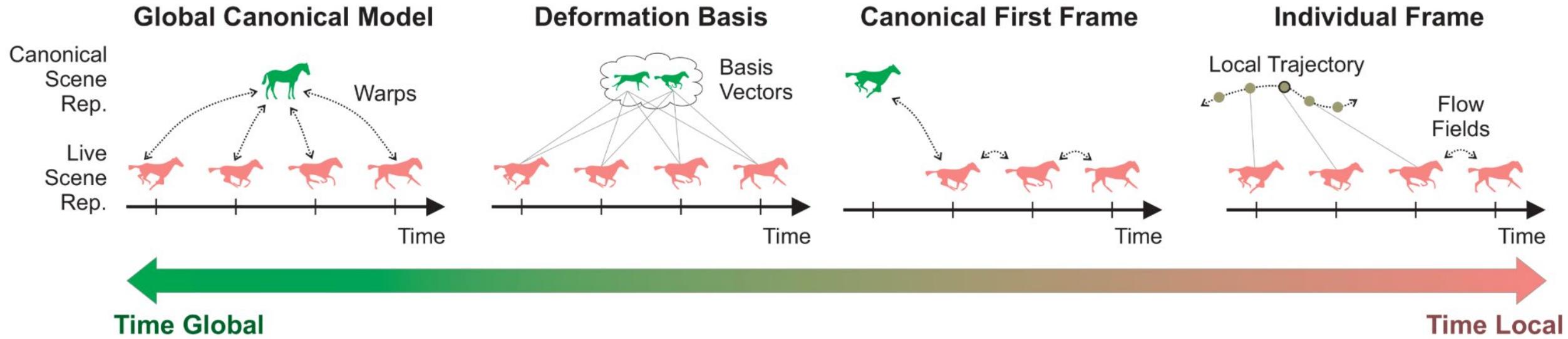
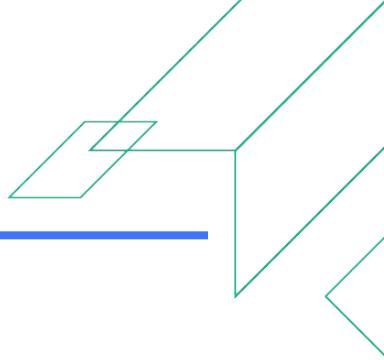
Spatio-Temporal Modelling



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

Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling

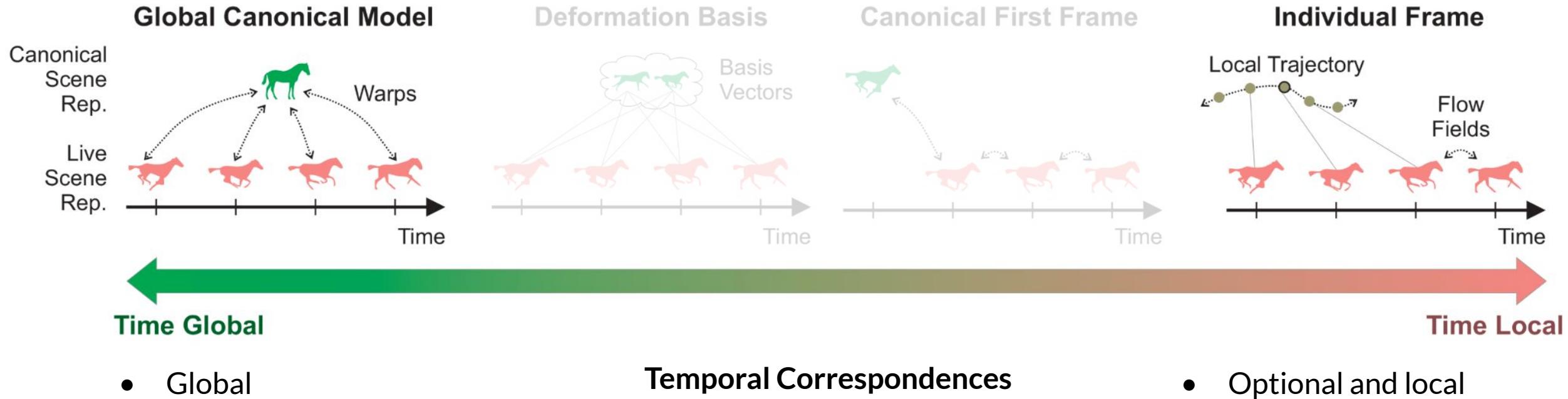
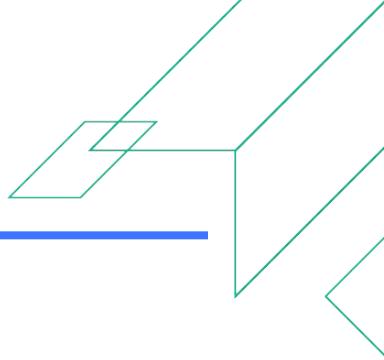


Time consistency enables applications like 3D editing and virtual asset creation

But we don't want to compromise on motion modelling

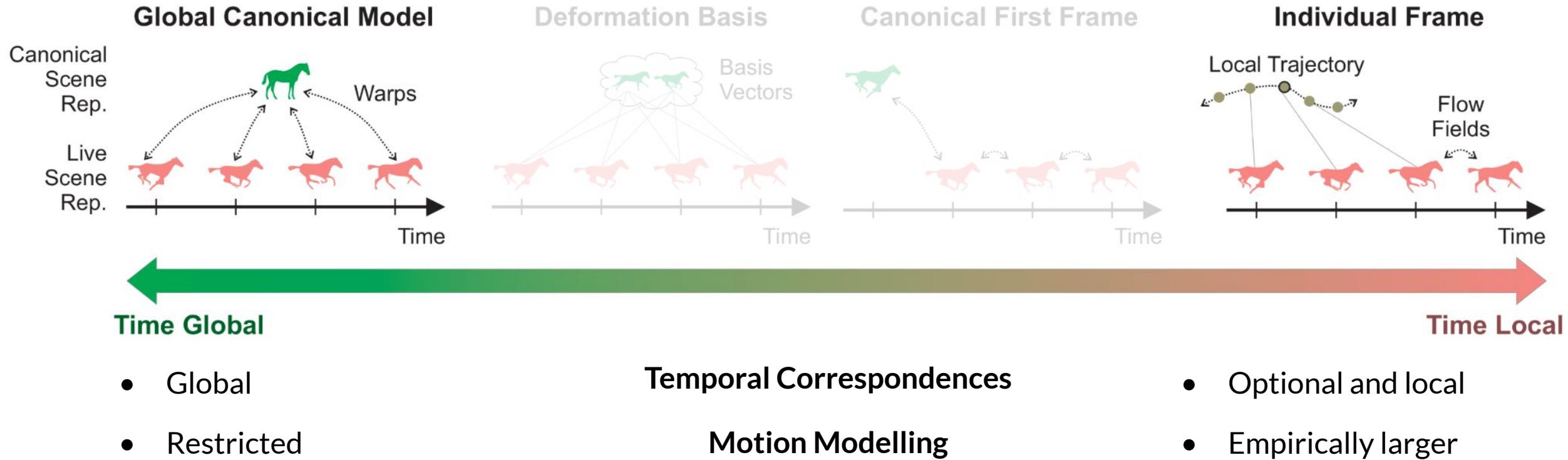
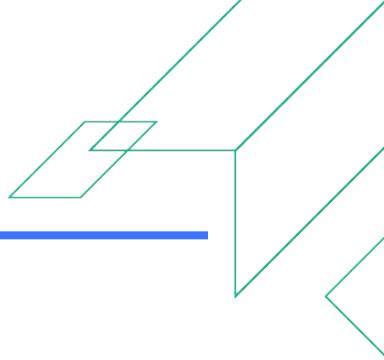
Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



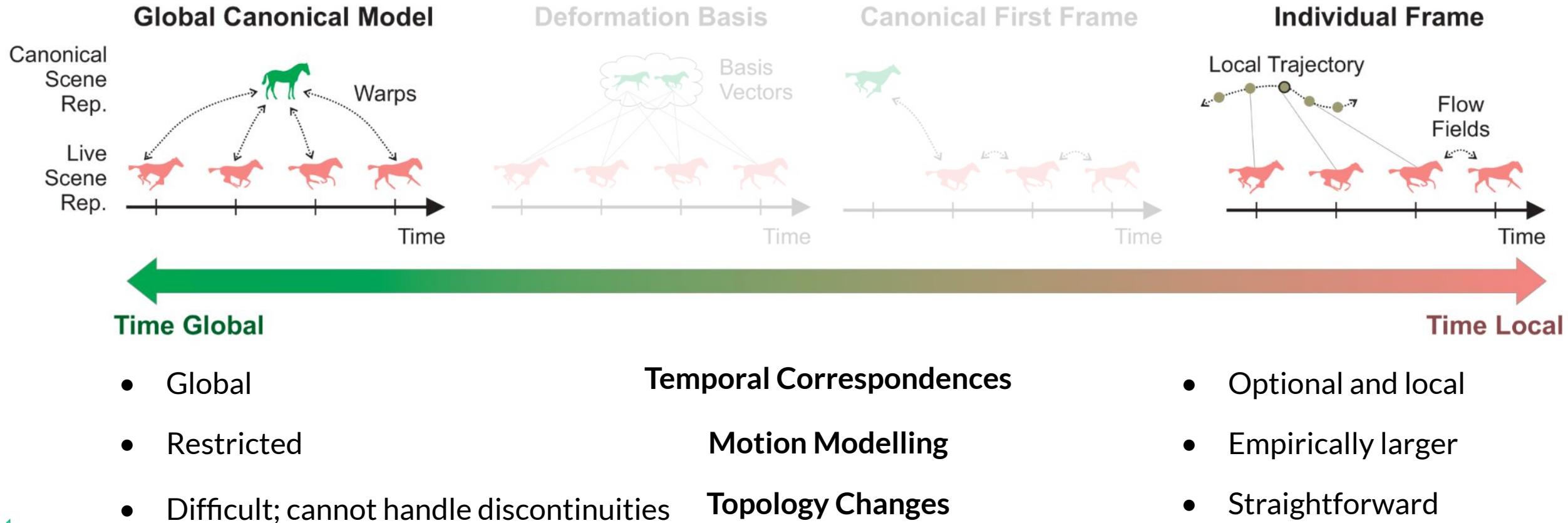
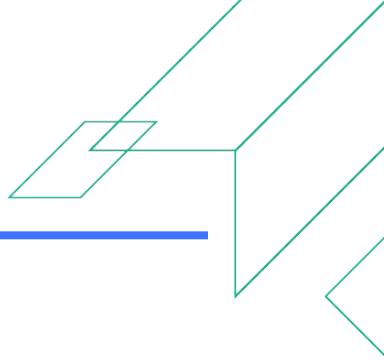
Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



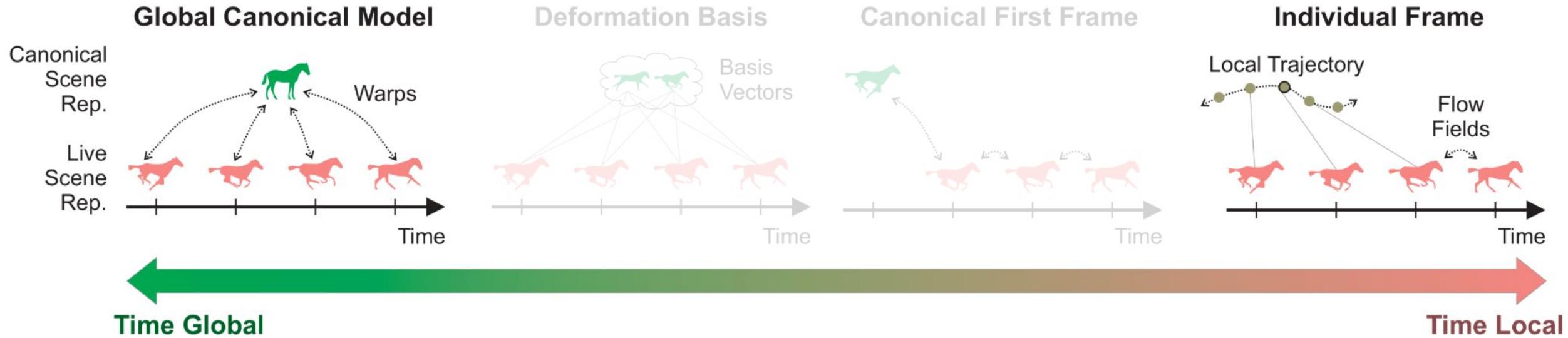
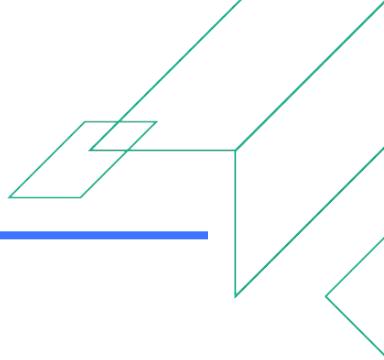
Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



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

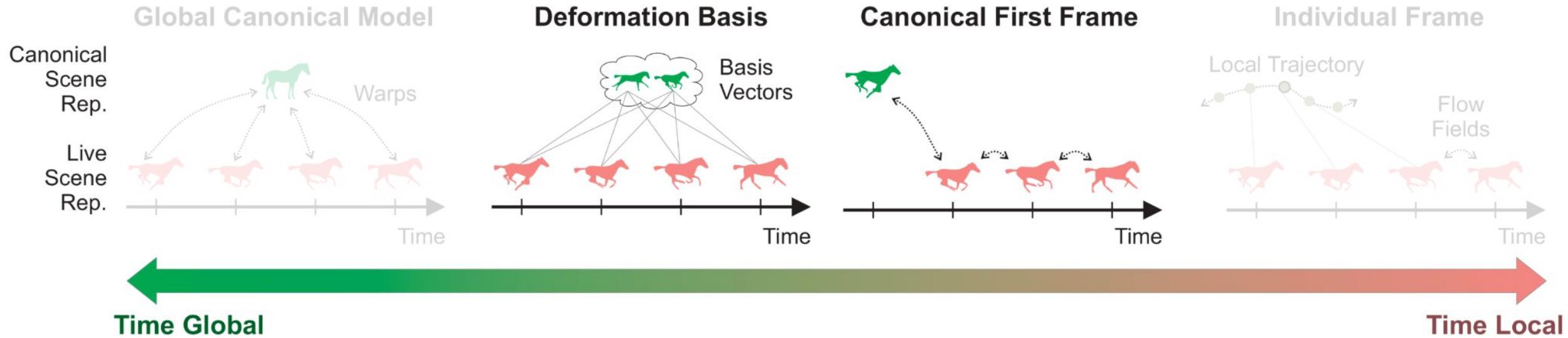
Topology Changes

Appearance Changes

- Optional and local
- Empirically larger
- Straightforward
- Straightforward

Large Motion vs. Long-term 3D Correspondences

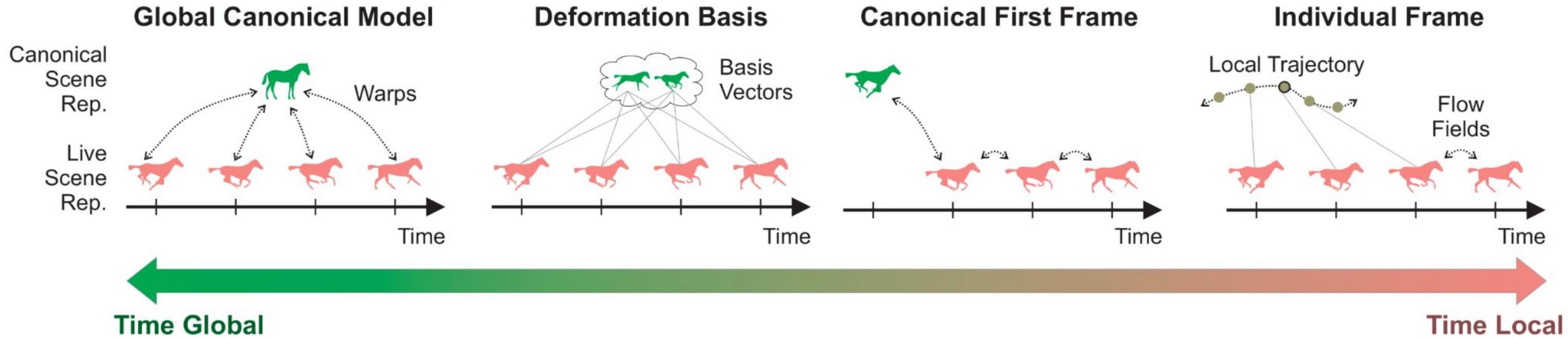
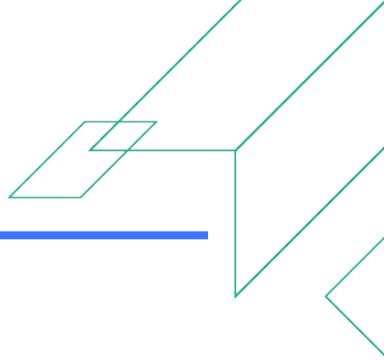
Spatio-Temporal Modelling



Trade-offs to balance the best of both worlds!

Large Motion vs. Long-term 3D Correspondences

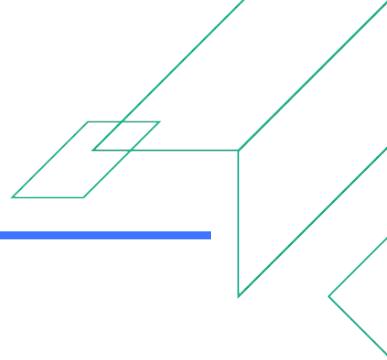
Spatio-Temporal Modelling



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

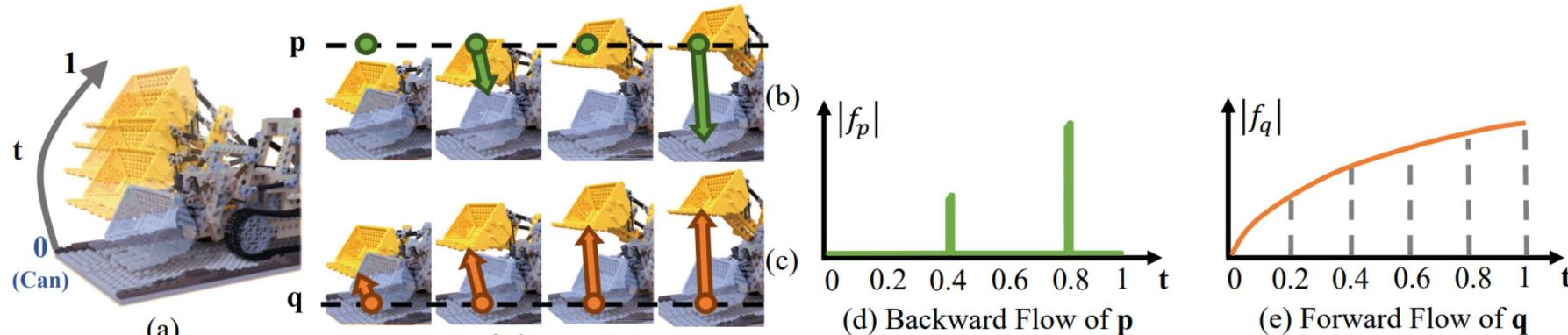
Large Motion vs. Long-term 3D Correspondences

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

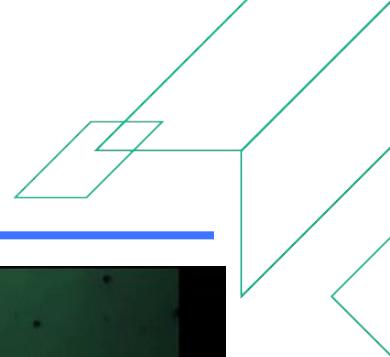
Time Global

Global Canonical Model

Time Local

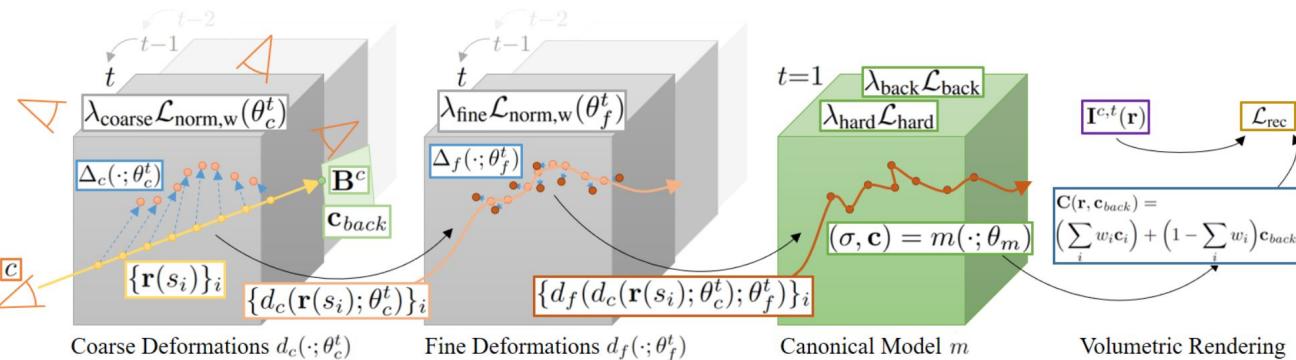
Large Motion vs. Long-term 3D Correspondences

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

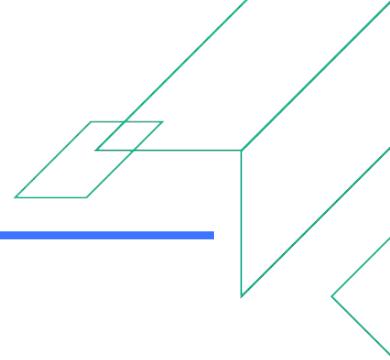
Time Global

Global Canonical Model

Time Local

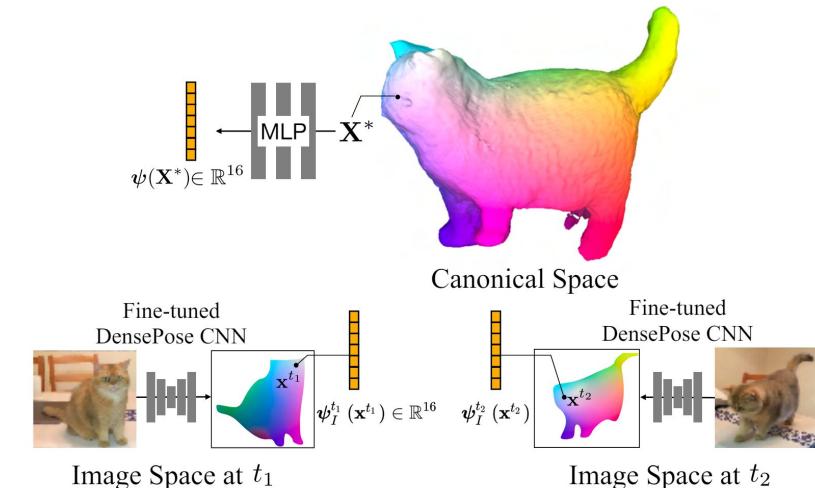
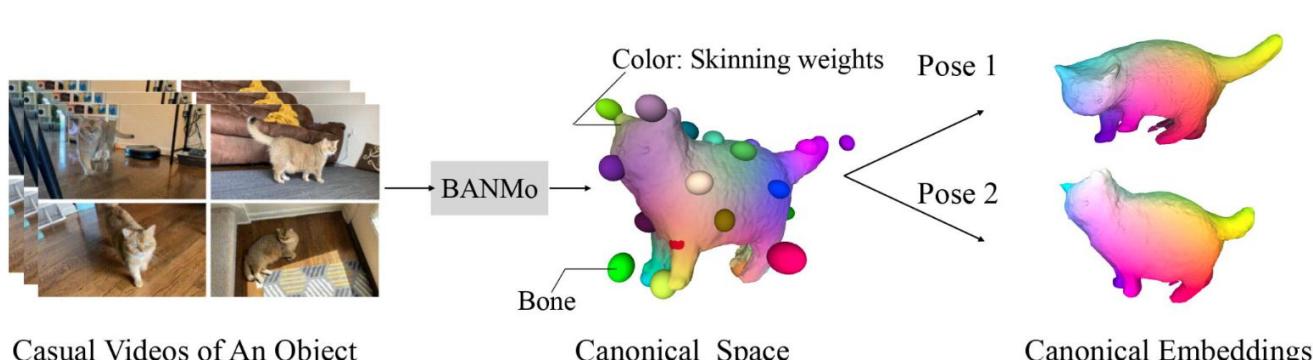
Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



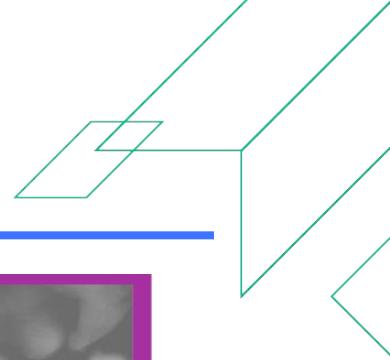
Canonical Feature Embeddings:

- 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



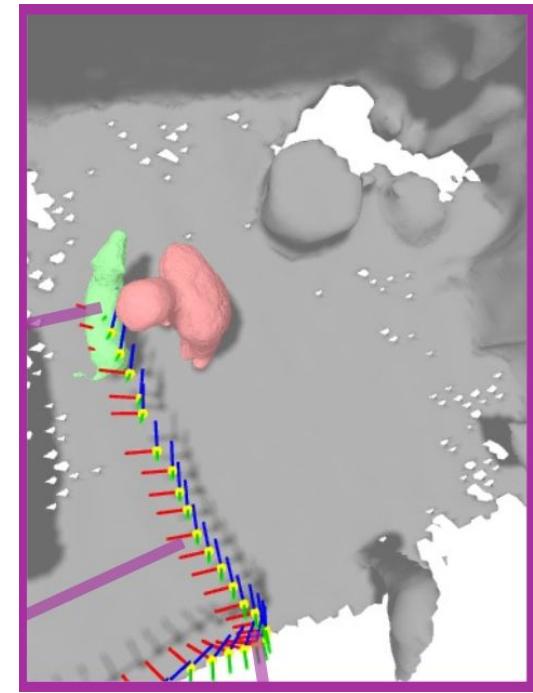
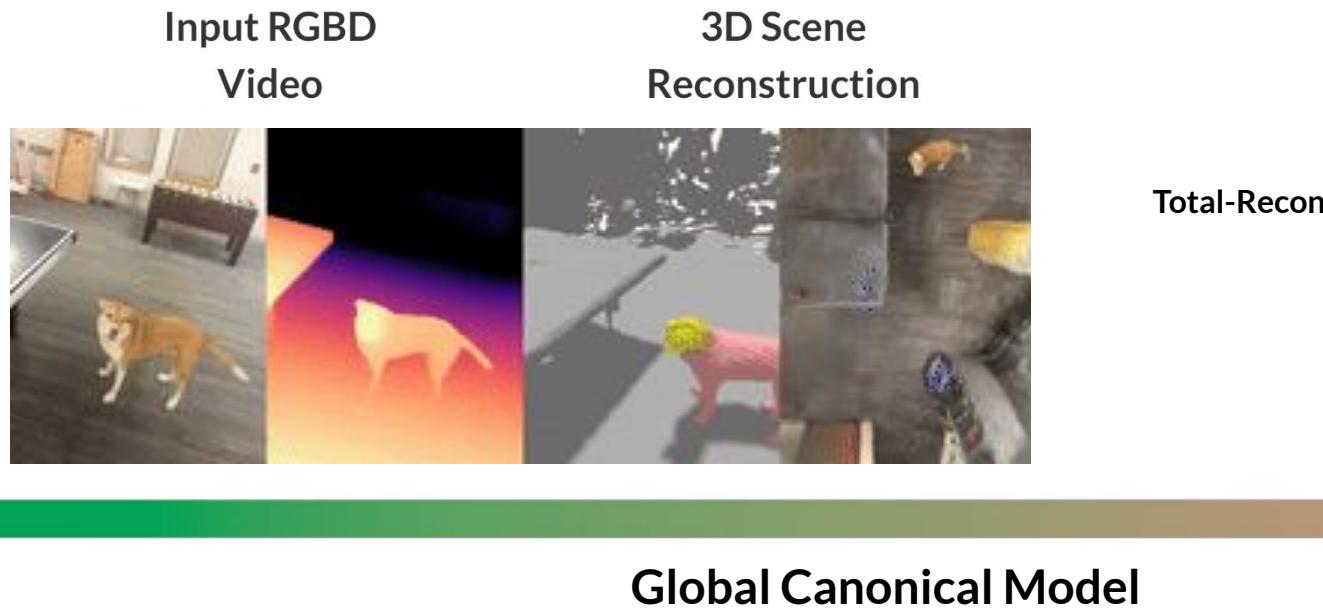
Large Motion vs. Long-term 3D Correspondences

Spatio-Temporal Modelling



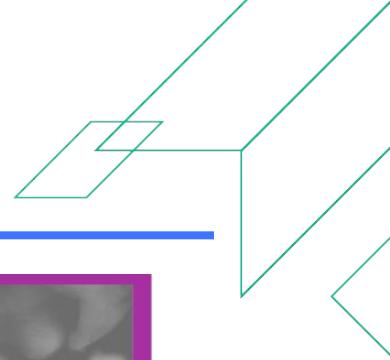
Decomposed Motion Modelling:

- 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



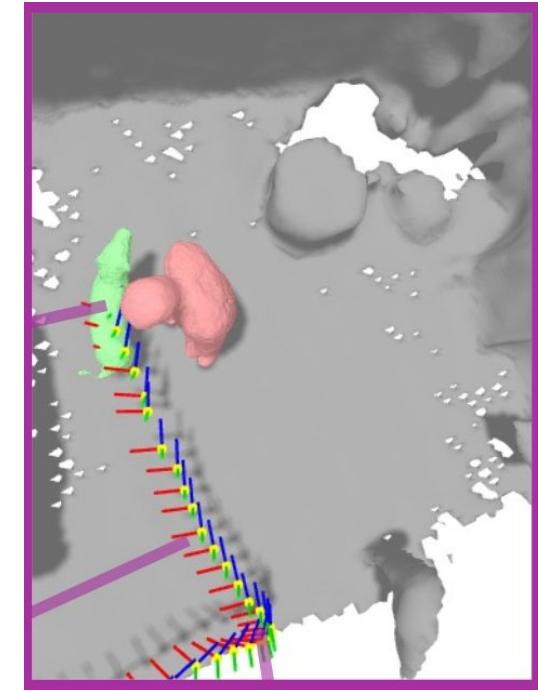
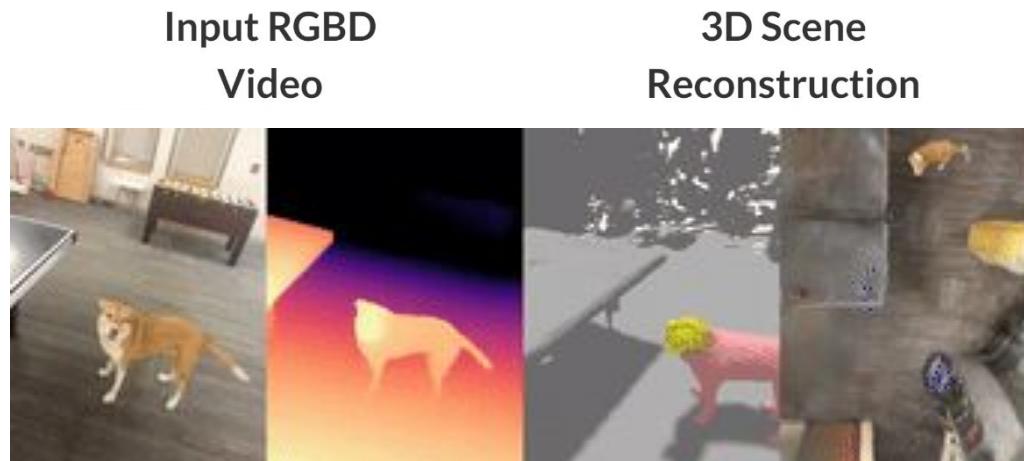
Large Motion vs. Long-term 3D Correspondences

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



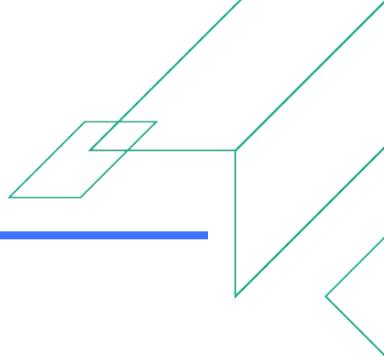
Scales up to minute-long RGB-D videos with large motion!

Time Global

Time Local

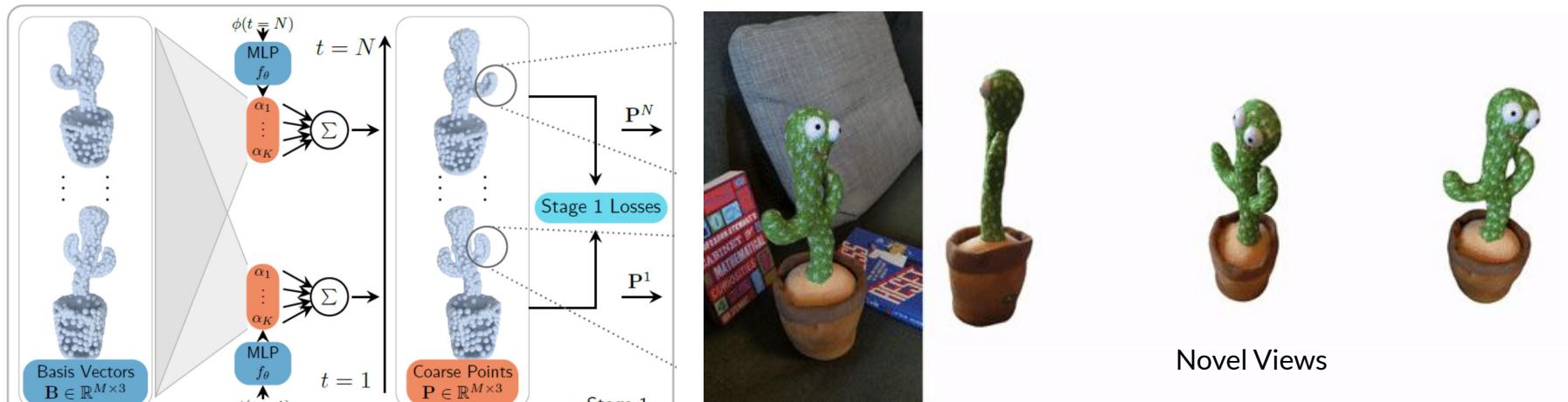
Large Motion vs. Long-term 3D Correspondences

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



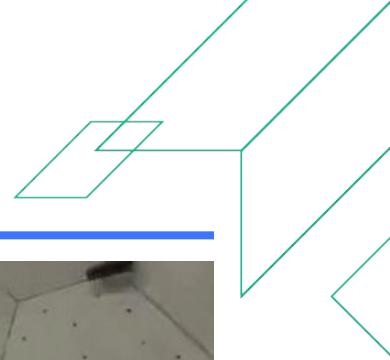
Time Global

Deformation Basis

Time Local

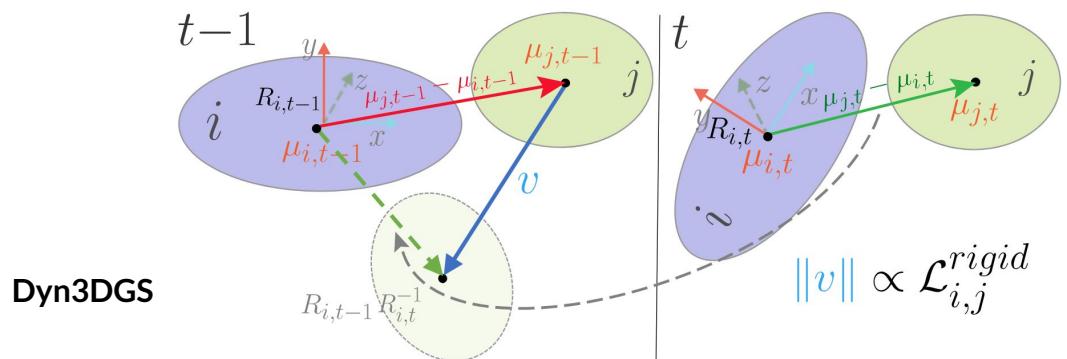
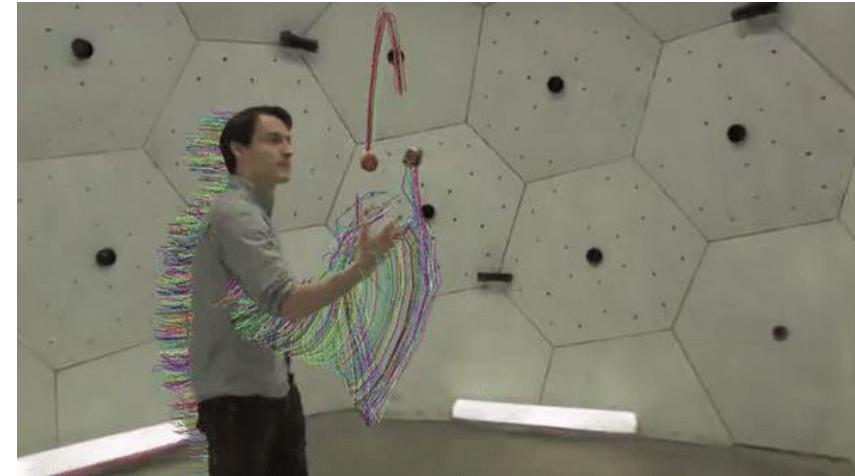
Large Motion vs. Long-term 3D Correspondences

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



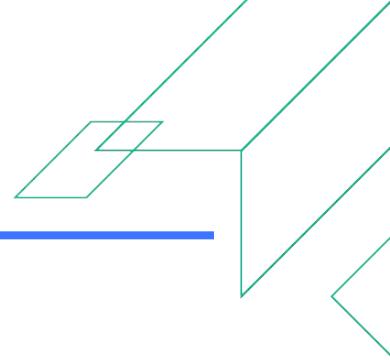
Time Global

Canonical First Frame

Time Local

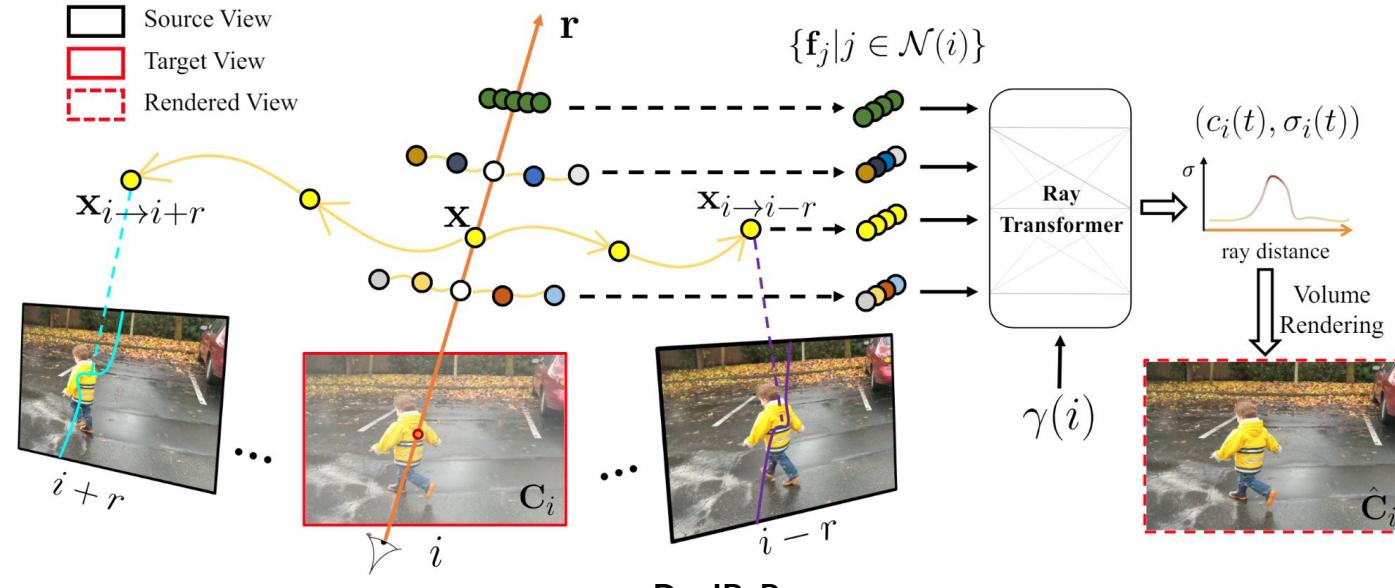
Large Motion vs. Long-term 3D Correspondences

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



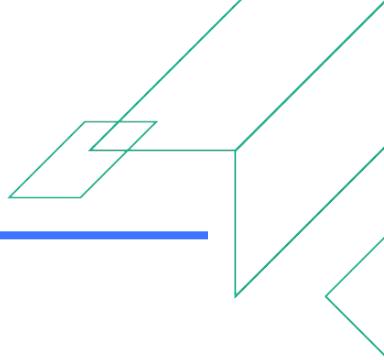
Time Global

Individual Frame

Time Local

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



Time Global

Individual Frame

Time Local

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

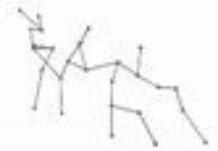
Deformations of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton:



Modelling General Articulated Motion

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

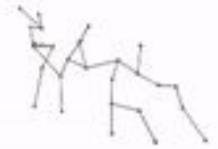
- Skeletons allow reposing of objects to unseen poses



Modelling General Articulated Motion

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

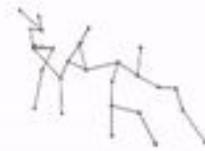
- Skeletons allow reposing of objects to unseen poses
- If we know how an object category articulates, we can use that information as prior to estimate motion of new sequences of that object



Modelling General Articulated Motion

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

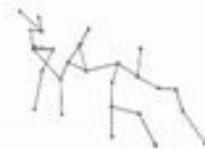
- Skeletons allow reposing of objects to unseen poses
- If we know how an object category articulates, we can use that information as prior to estimate motion of new sequences of that object
- Category-level object templates are available mostly for human categories (e.g. SMPL) which are obtained from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

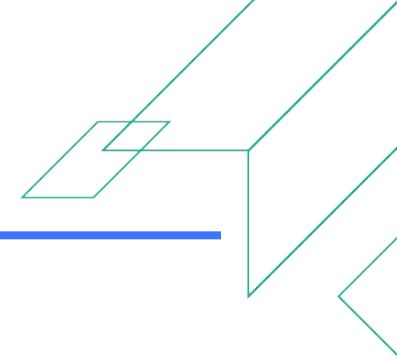
- Skeletons allow reposing of objects to unseen poses
- If we know how an object category articulates, we can use that information as prior to estimate motion of new sequences of that object
- Category-level object templates are available mostly for human categories (e.g. SMPL) which are obtained from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



From RGB videos!

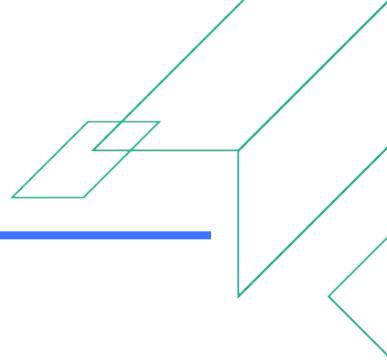
Modelling General Articulated Motion

Self-Supervised Part Discovery for Reposing



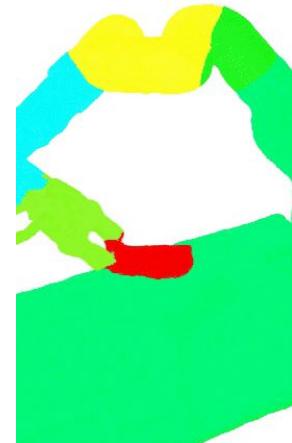
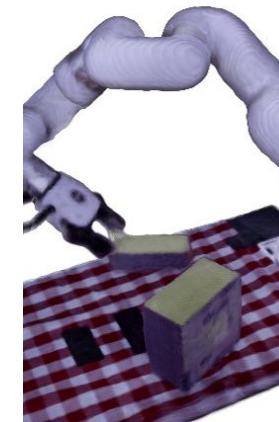
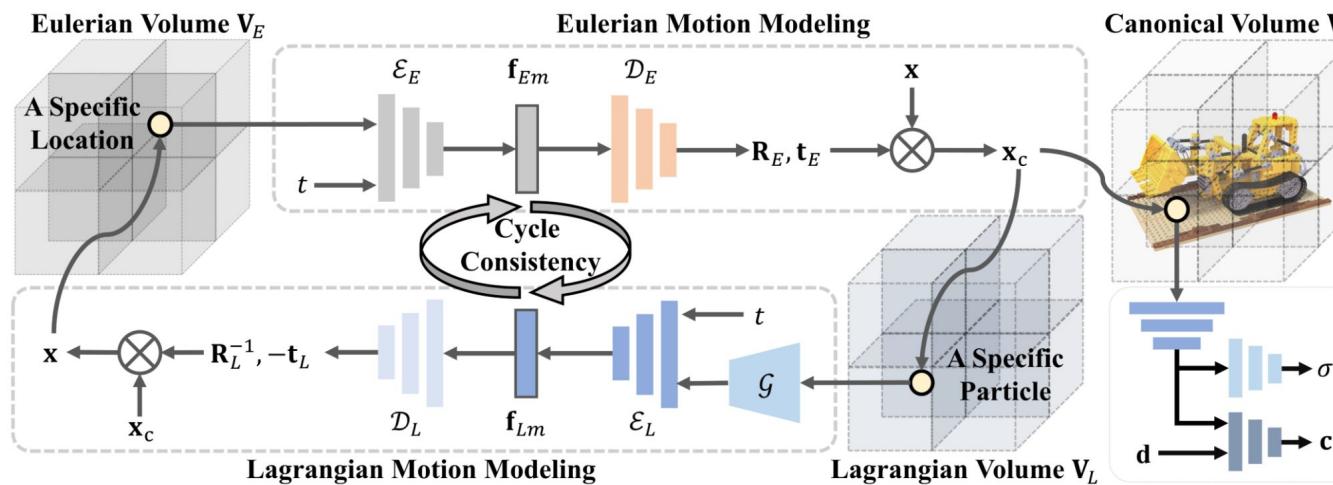
Modelling General Articulated Motion

Self-Supervised Part Discovery for Reposing



Motion-based grouping:

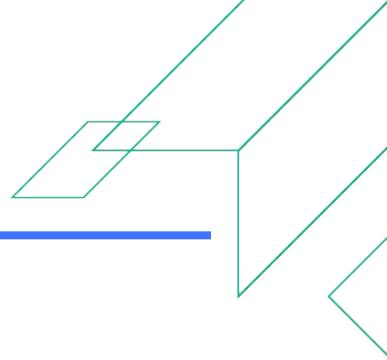
- Models both backward and forward motion with feature grids
- Features from forward motion are grouped into slots using an attention mechanism
- Similar motion \rightarrow same slot \rightarrow same part
- Discovered parts can be skeletonized and reposed



MovingParts

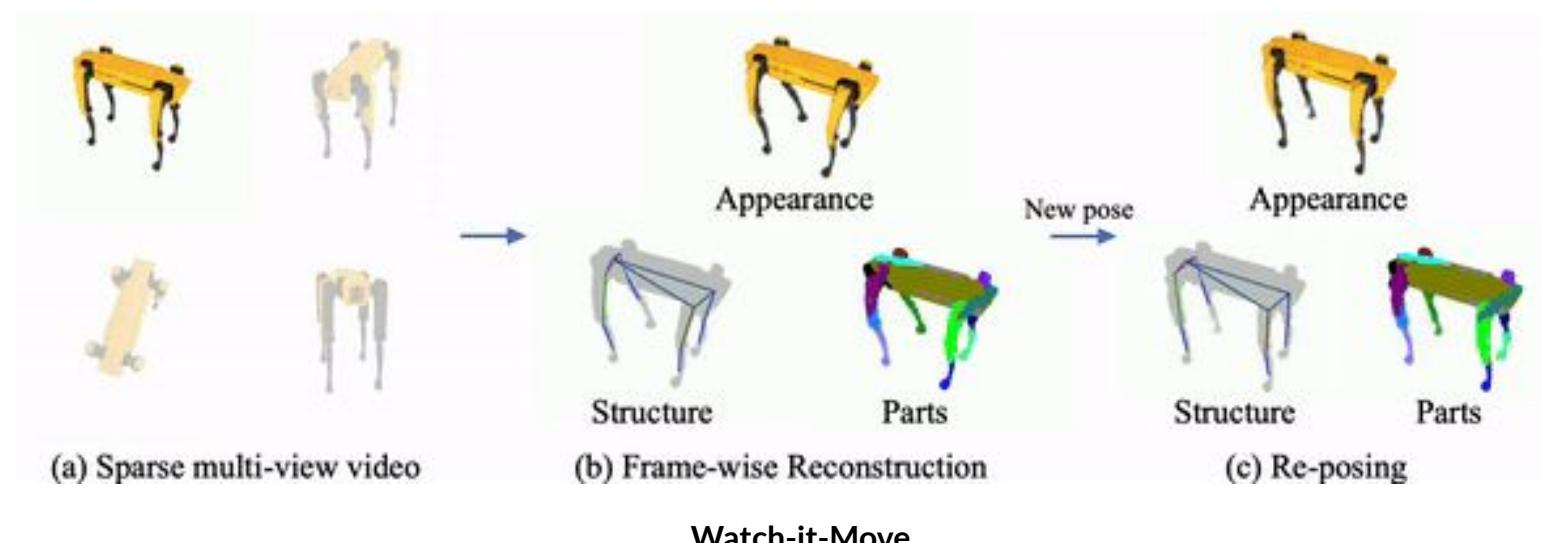
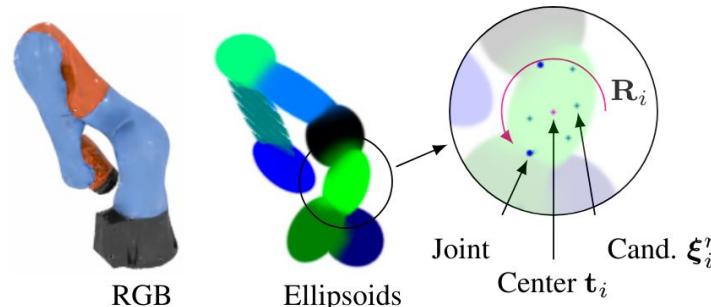
Modelling General Articulated Motion

Self-Supervised Part Discovery for Reposing



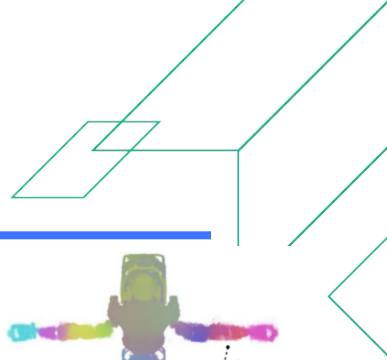
Unsupervised Part Prediction:

- 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



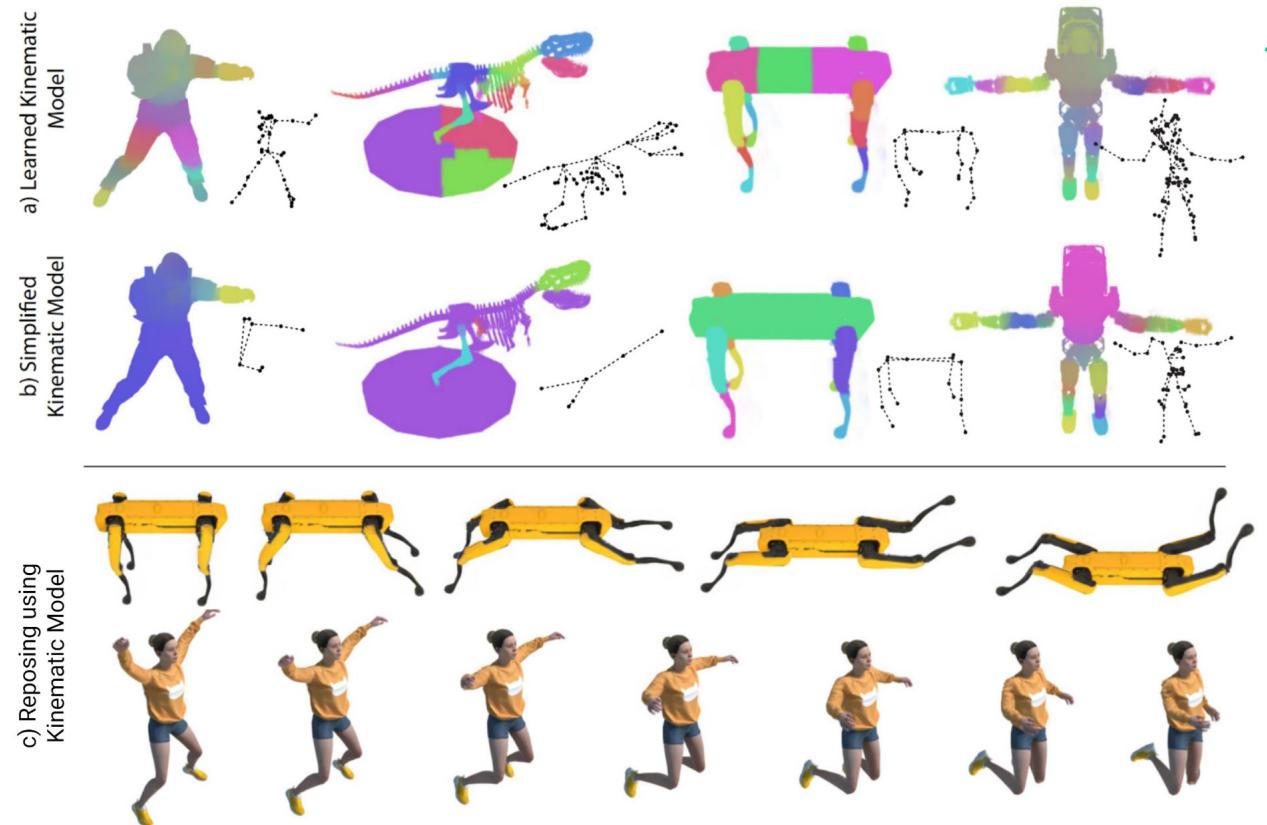
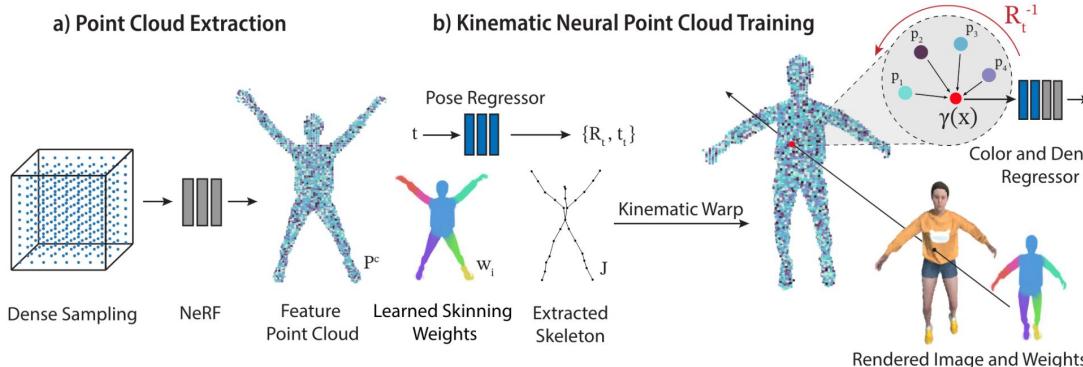
Modelling General Articulated Motion

Skeleton Discovery for Reposing



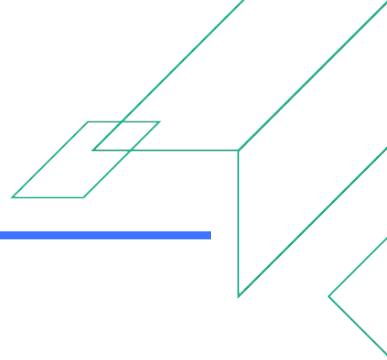
Morphological Operations:

- 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?

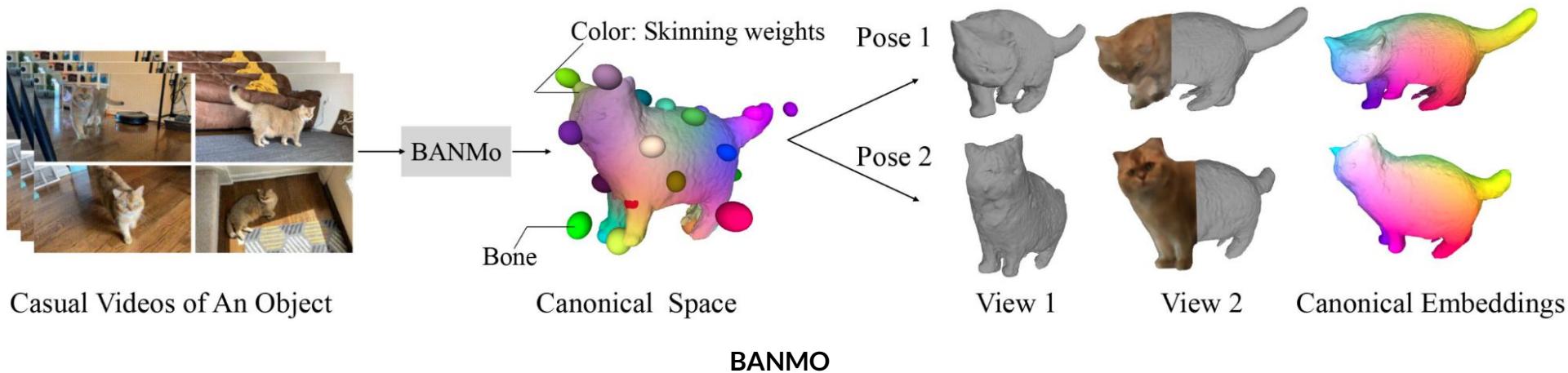
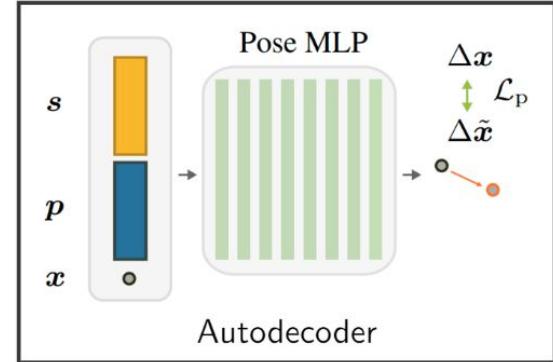


Modelling General Articulated Motion

Modelling Articulations with Neural Bones

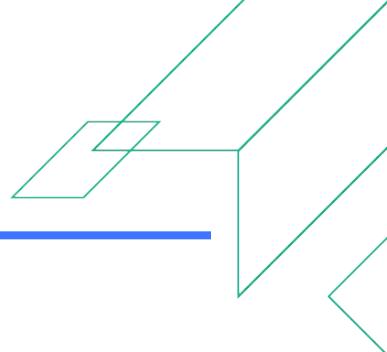
- Canonical space is shared between videos
- Bone positions and transforms are estimated per-frame using an auto-decoded MLP
- Articulated using volumetric skinning

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



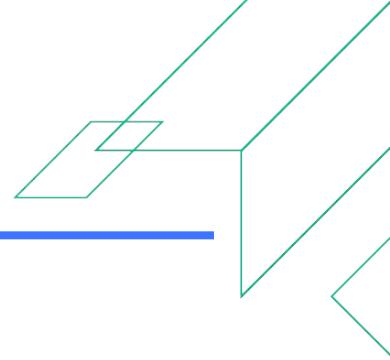
BANMO



Target Geometry



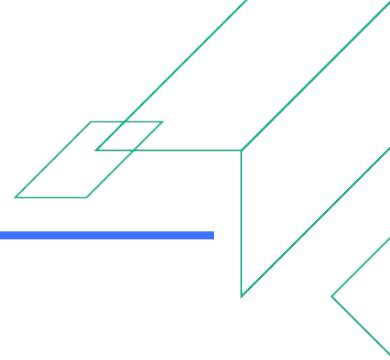
Modelling General Articulated Motion



- 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



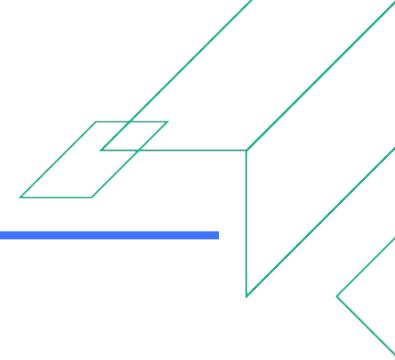
- A category of objects articulates in the same way (e.g. different breeds of cats)
- Can we learn category-level templates from videos to regularize motion even further and use it as a prior for instances?

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

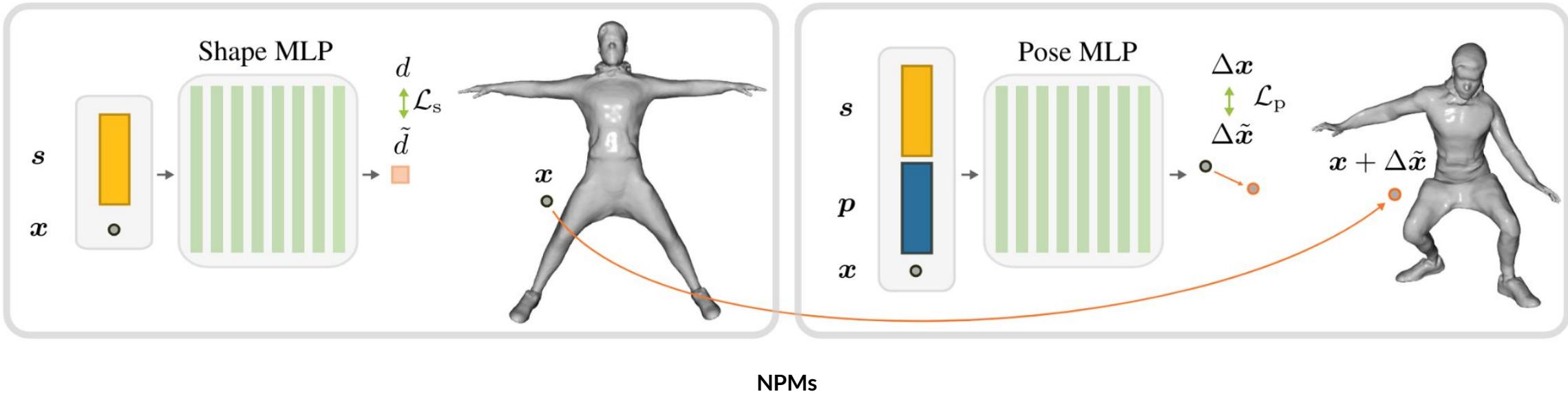


Modelling General Articulated Motion

Category-level Modelling from Depth Videos

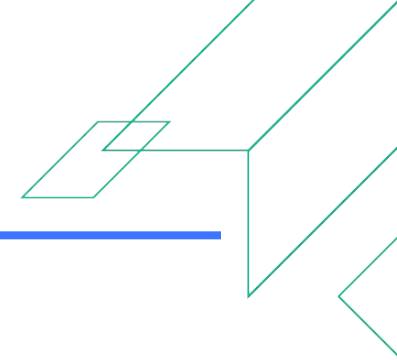


- Use auto-decoders to model both shape and pose variations
- Shape embeddings can capture category-level variations while pose embeddings capture instance articulations
- Optimize shape and then pose at test-time



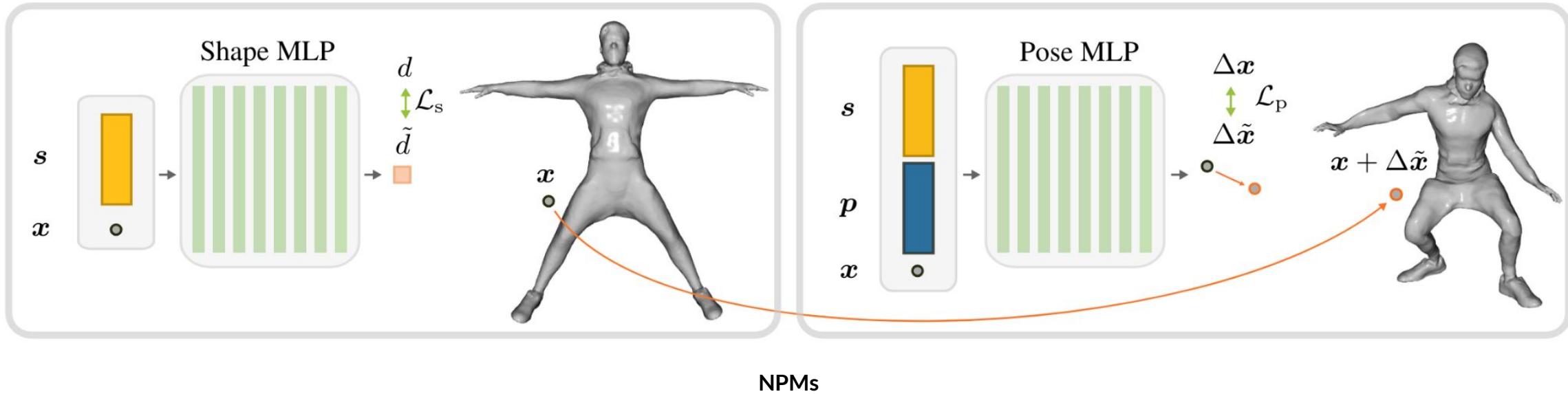
Modelling General Articulated Motion

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

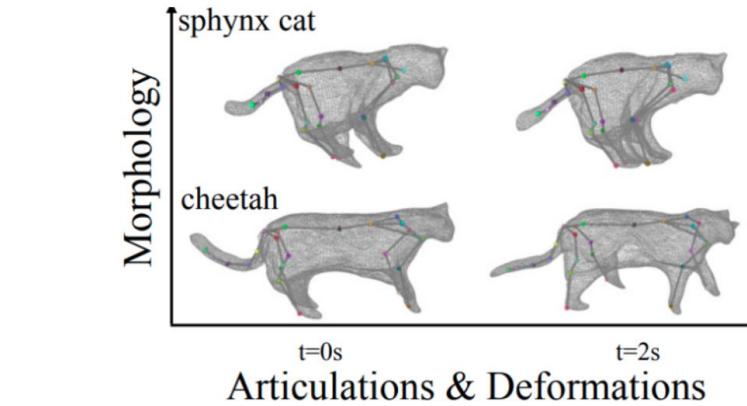
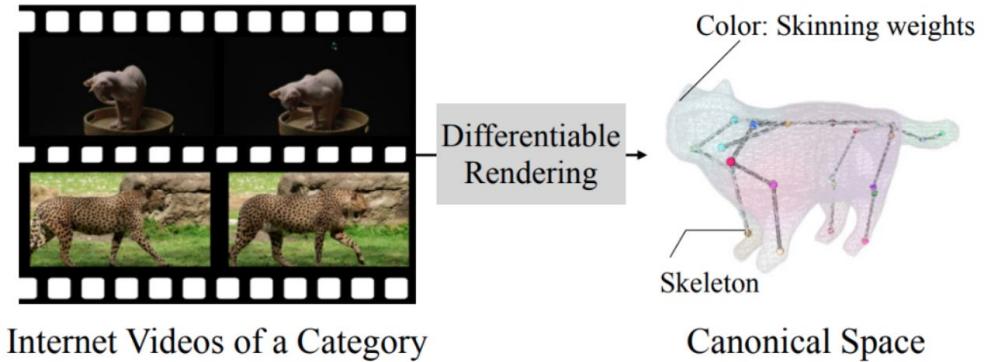
Learned from depth sequences.
Can we do it from RGB videos?



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



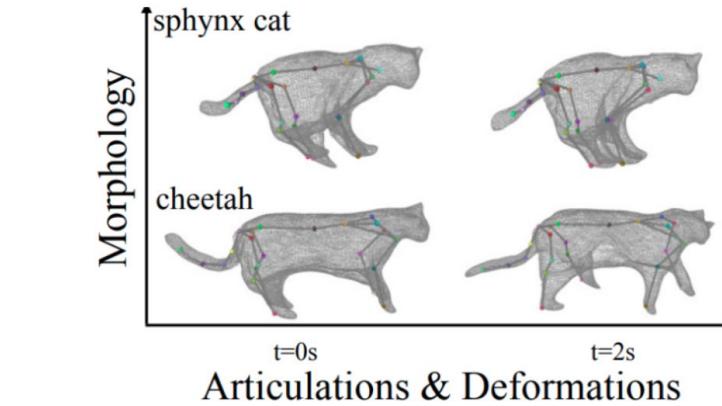
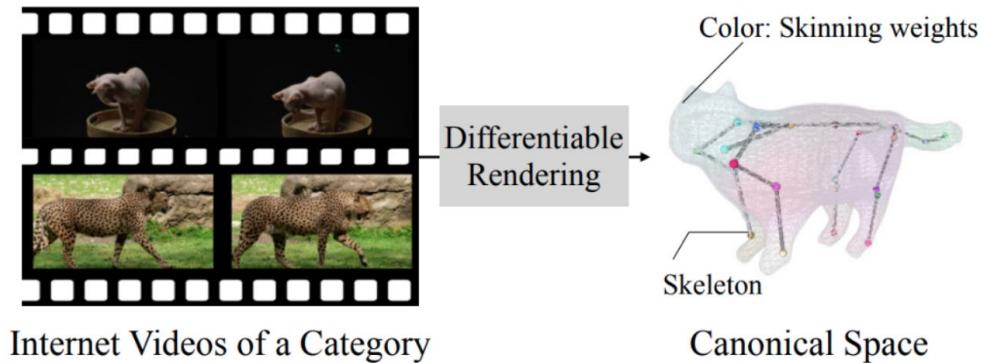
RAC

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

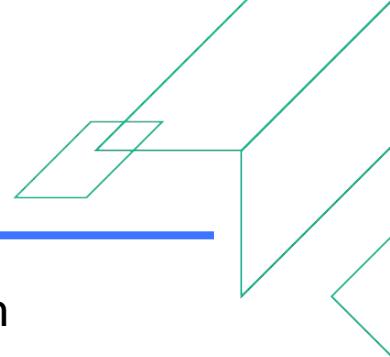
Can we do it from image collections,
which are more commonly available
for general categories than videos?



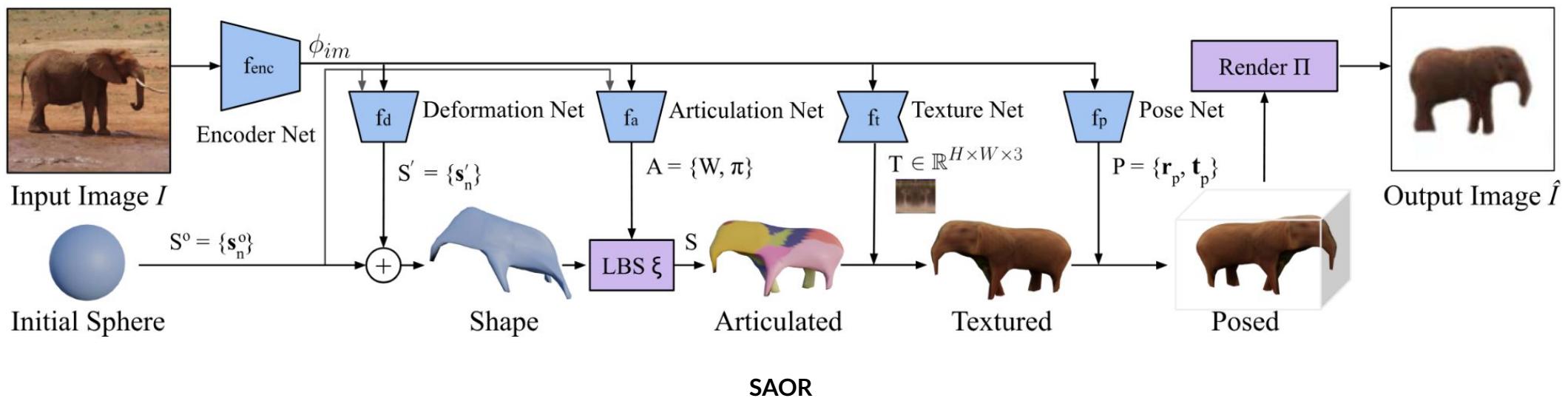
RAC

Modelling General Articulated Motion

Category-level Modelling from Image Collections



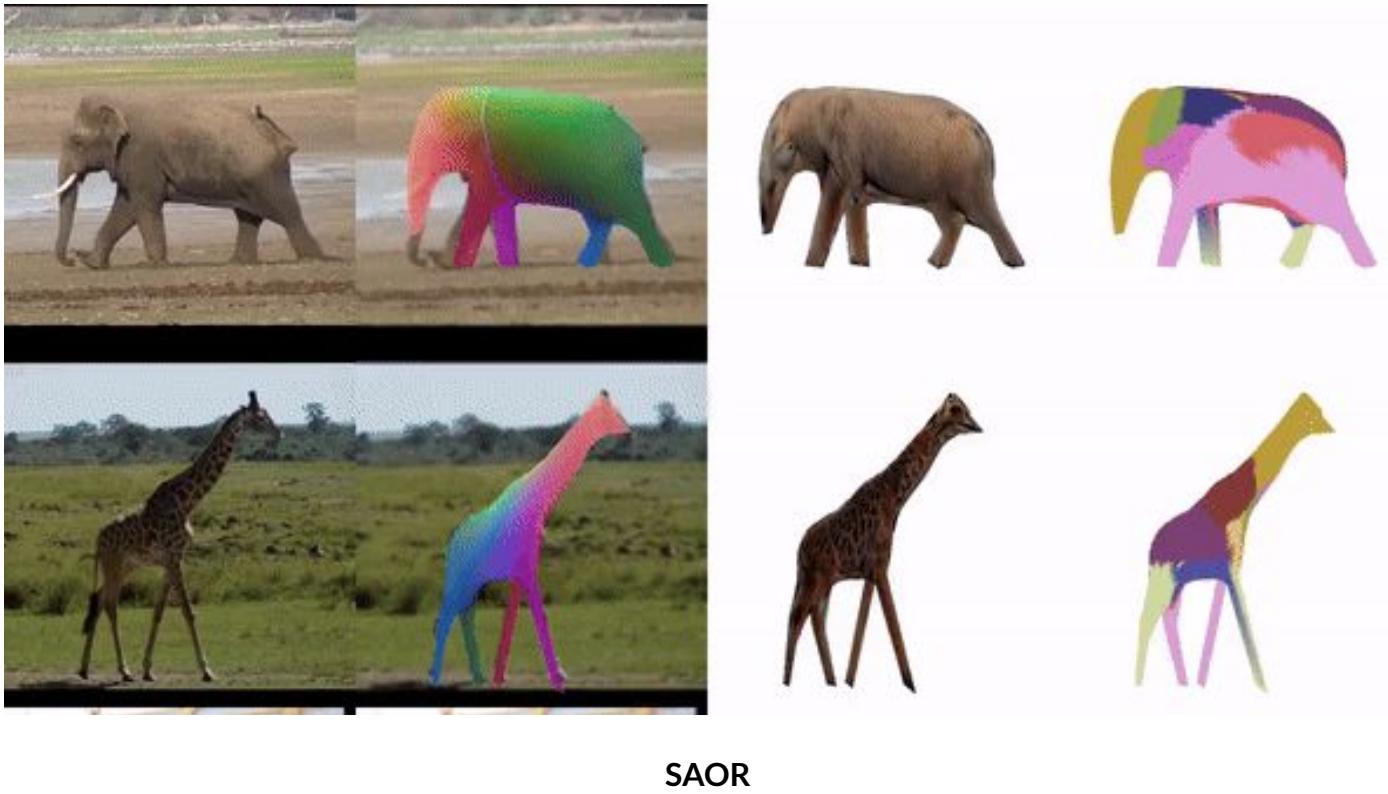
- Shape, articulation, pose and texture are directly predicted with separate decoders from an encoded image
- Category-level prior learned by shape and articulation decoders
- Enables prediction from single image at test-time



Modelling General Articulated Motion

Category-level Modelling from Image Collections

- Per-frame video reconstruction



Trends

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

Trends

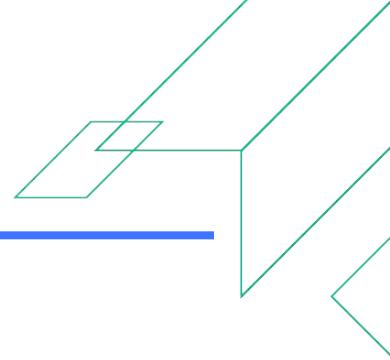
1. Speed and Quality Advancements

2. Hand Gestures

Non-rigid 3D reconstruction is far from solved!

3. Modelling Articulated Motion for General Objects

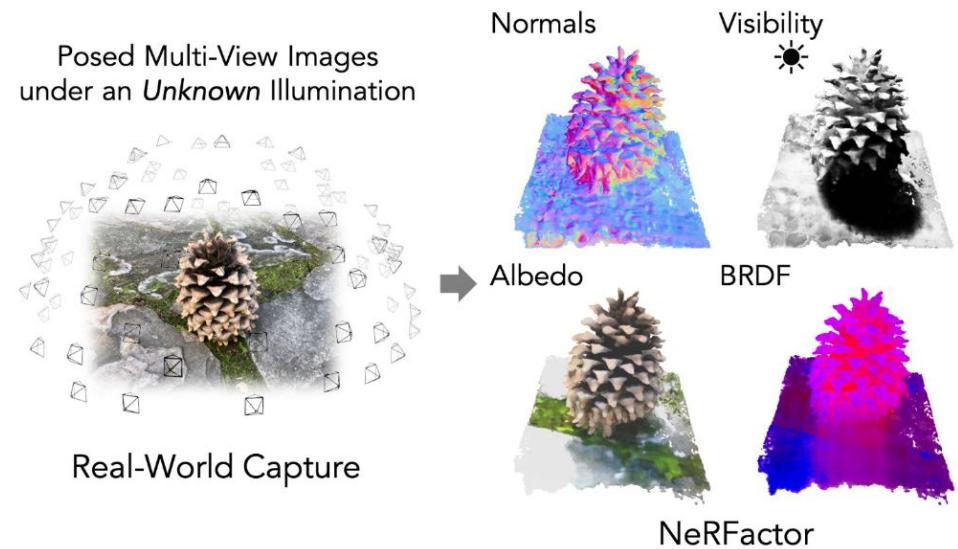
Remaining Challenges and Future Directions



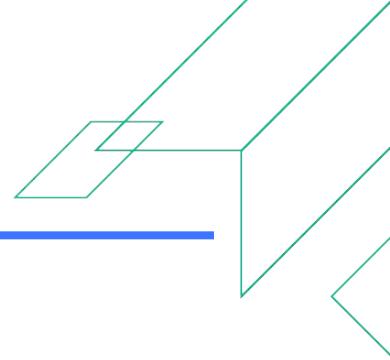
Remaining Challenges and Future Directions

Intrinsic Decomposition and Relighting

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

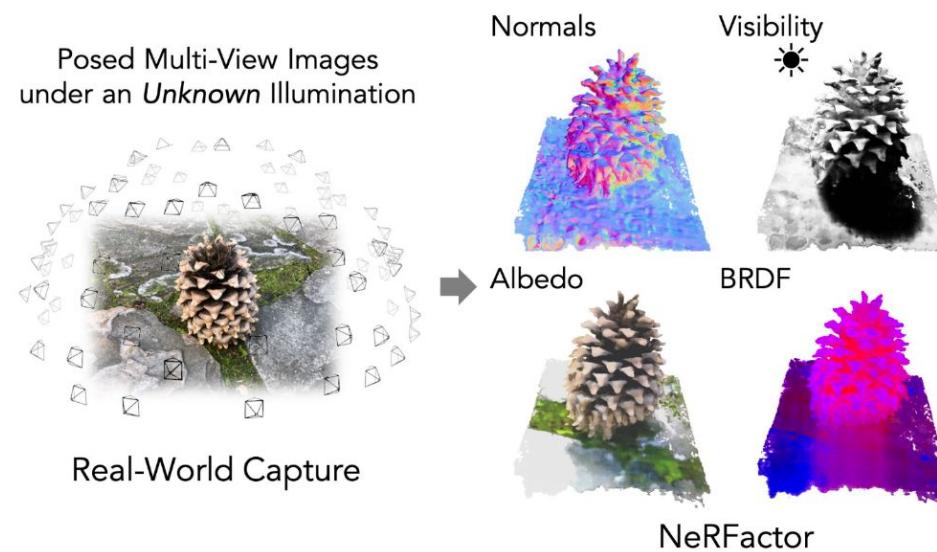


Remaining Challenges and Future Directions



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



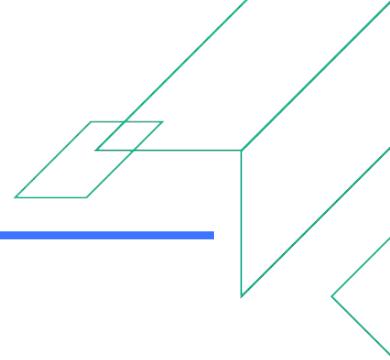
Remaining Challenges and Future Directions

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



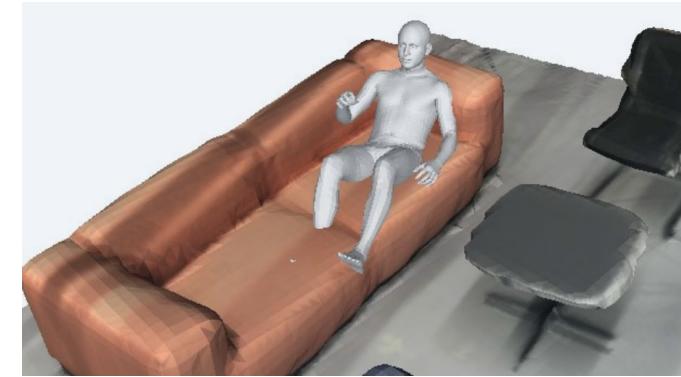
Reliable Camera Pose Estimation

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



Multi-Object Interaction

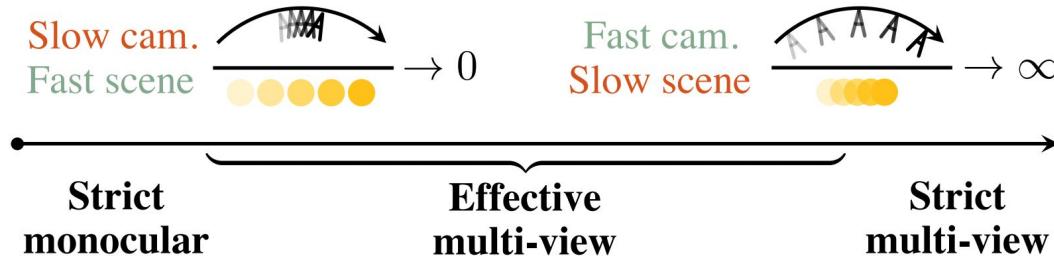
- Interaction between objects is not explicitly modelled by current methods for general objects
- Useful to enforce the correct dynamics and constraints



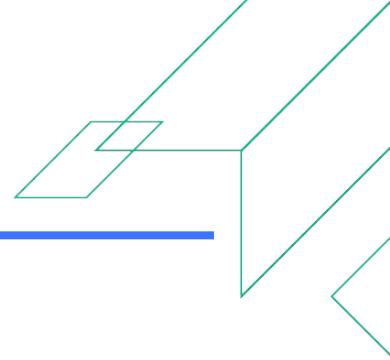
Remaining Challenges and Future Directions

Reconstruction from Sparse Casual Captures

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

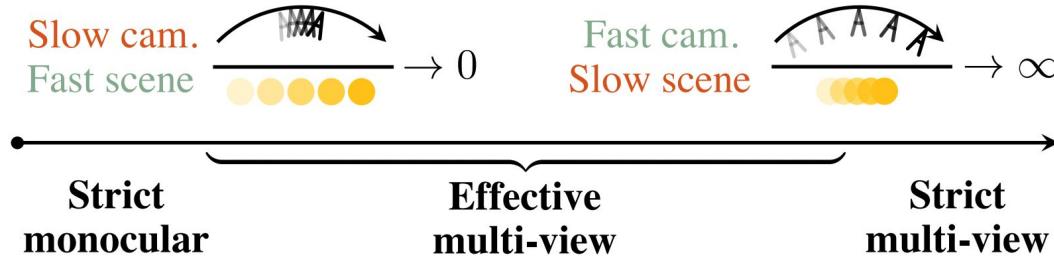


Remaining Challenges and Future Directions



Reconstruction from Sparse Casual Captures

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

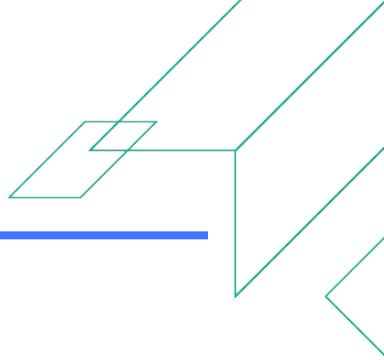


Long-Term Dense Correspondences

- 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

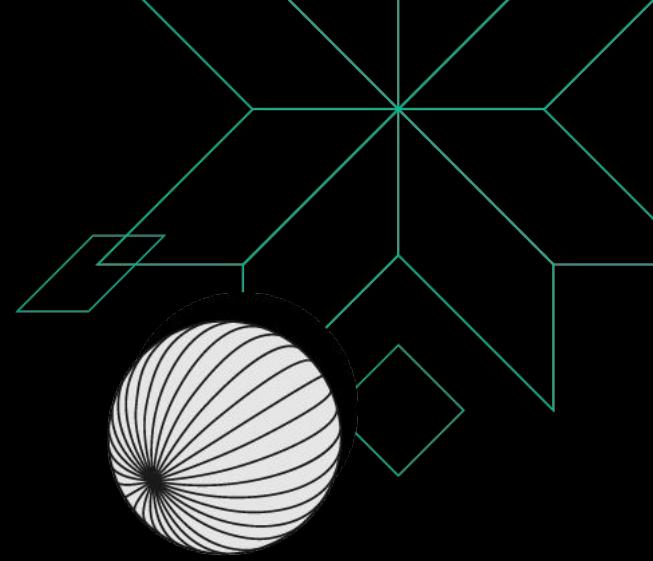




Thank you.

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 reserved.



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.