EUROGRAPHICS 2024
A. Aristidou and R. McDonnell
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2
STAR – State of The Art Report

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

Raza Yunus[1,2] Jan Eric Lenssen[2] Michael Niemeyer[3] Yiyi Liao[4] Christian Rupprecht[5] Christian Theobalt[2]
Gerard Pons-Moll[6] Jia-Bin Huang[7] Vladislav Golyanik[2] Eddy Ilg[1]

[1]Saarland University, SIC [2]MPI for Informatics, SIC [3]Google [4]Zhejiang University [5]University of Oxford [6]University of Tübingen [7]UMD

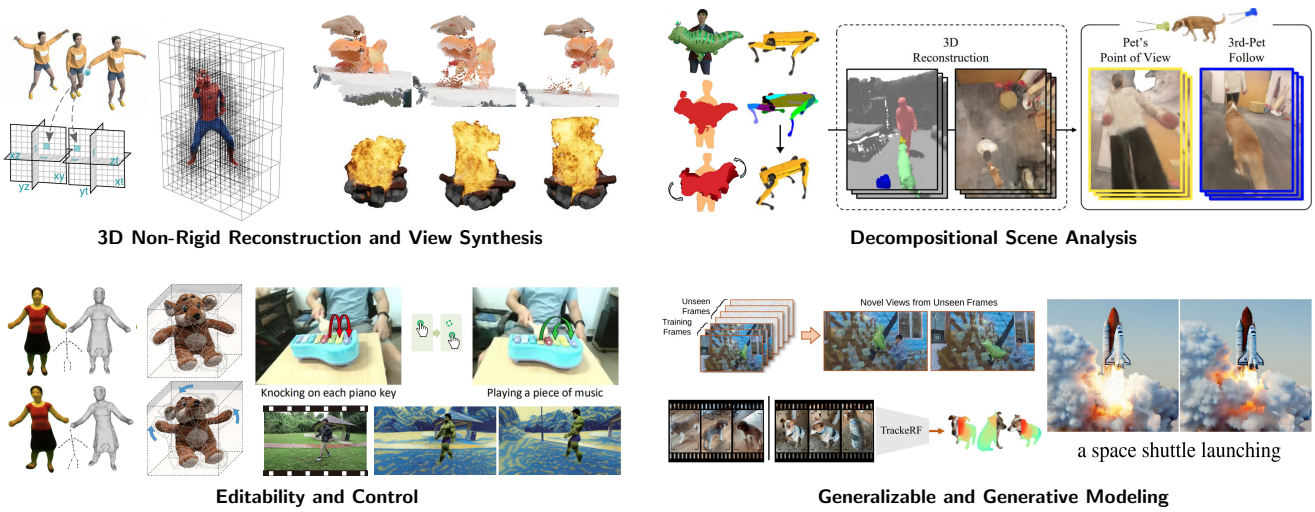

Figure 1: *This STAR covers fundamental concepts and recent trends in 3D reconstruction of general non-rigid scenes. We discuss techniques for reconstruction, decompositional scene analysis, editing and control, and generalizable and generative modeling. Image sources: [FKMW*23, WZL*22, CB22, MPCVG23, LLM*23, LCW*24, SYD*23, KKK*23, XH22, NIT*22, ZLX23, SSR*23, TDD23, BSR*24]*

Abstract

Reconstructing models of the real world, including 3D geometry, appearance, and motion of real scenes, is essential for computer graphics and computer vision. It enables the synthesizing of photorealistic novel views, useful for the movie industry and AR/VR applications. It also facilitates the content creation necessary in computer games and AR/VR by avoiding laborious manual design processes. Further, such models are fundamental for intelligent computing systems that need to interpret real-world scenes and actions to act and interact safely with the human world. Notably, the world surrounding us is dynamic, and reconstructing models of dynamic, non-rigidly moving scenes is a severely underconstrained and challenging problem. This state-of-the-art report (STAR) offers the reader a comprehensive summary of state-of-the-art techniques with monocular and multi-view inputs such as data from RGB and RGB-D sensors, among others, conveying an understanding of different approaches, their potential applications, and promising further research directions. The report covers 3D reconstruction of general non-rigid scenes and further addresses the techniques for scene decomposition, editing and controlling, and generalizable and generative modeling. More specifically, we first review the common and fundamental concepts necessary to understand and navigate the field and then discuss the state-of-the-art techniques by reviewing recent approaches that use traditional and machine-learning-based neural representations, including a discussion on the newly enabled applications. The STAR is concluded with a discussion of the remaining limitations and open challenges.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction;*** *Volumetric models; Point-based models; Mesh geometry models;* ***Motion capture; Shape representations; Appearance and texture representations;***

1. Introduction

3D reconstruction and rendering of non-rigidly deforming scenes are fundamental problems in computer vision and graphics. Depending on the sensor types, underlying scene assumptions, and types of motions and deformations, this problem is severely ill-posed and highly challenging. Applications of non-rigid 3D reconstruction pervade many domains of science, studying our world on different scales: tracking of celestial bodies and their agglomerations (cosmological scale); reconstruction and prediction of dynamics in the troposphere of Earth from satellite observations; reconstruction of melting glaciers over time (the scale of a planet and its ecosystems); reconstruction of humans and animals in interaction with their environments (level of ecosystems); non-rigid tracking of human faces, body parts, worn garments, and organ tissues (level of living organisms); non-rigid objects even on smaller scales (e.g. microorganisms). The intermediate scales involving humans and their environments—the focus of this report—have recently seen a lot of work due to their relevance to visual computing, while others remain difficult to study. Moreover, end-users utilize and enjoy technology involving non-rigid 3D reconstruction daily, such as movies, computer games, AR/VR headset applications, driver assistance systems, and mobile video editing applications, among others.

The emergence of neural scene representations marked a paradigm shift in non-rigid 3D reconstruction. Through the advances in differentiable rendering [TTM*22], these methods enable end-to-end optimization of the 3D scene representations directly from the available visual observations (images, videos, or other sensing modalities). For example, Neural Radiance Fields (NeRFs) [MST*20] represent a scene with a coordinate-based multi-layer perception that maps a 3D position in space and a 2D viewing direction into color and density. Using classic volume rendering techniques, NeRF achieved unprecedented quality of view synthesis. Compared to classical 3D reconstruction pipelines that involve multiple disconnected stages (e.g., structure from motion, multi-view stereo, surface reconstruction via depth fusion, and texturing), such neural scene representation approaches offer significantly more accurate geometry and appearance reconstruction.

Significant progress has been made in improving the accuracy of geometry reconstruction, appearance modeling, training, rendering speed, and supporting various input modalities for general 3D reconstruction tasks, from which the non-rigid setting [LNSW21, TTG*21, GSKH21] has also benefited. Scene representations such as hybrid neural representations [MESK22, XXP*22, CXG*22] and 3D Gaussian Splatting [KKLD23]—introduced in the static setting—have significantly reduced training times and enabled fast rendering (real-time in the case of the latter) of non-rigid scenes [CJ23, FYW*22, AHR*23, LKLR24, DWY*24]. Advances in machine learning techniques and especially in generative modeling [RBL*22] have enabled learning stronger priors for different aspects of the non-rigid reconstruction task [LZYX22, TDD23] and even generate new 4D sequences of scenes [SSP*23] and objects [EMS*23] from the learned distributions. The general trend of using self-supervised learning to discover underlying structures and concepts has also been seen in the context of static-dynamic scene decomposition [WZT*22, SCL*23] and joint/skeleton discovery [NIT*22, YZH*24]. Finally, the need to make non-rigid 3D reconstruction methods more accessible for increased applications has pushed development in challenging settings, such as reconstruction from a monocular video [WMJL23, CFF*22] and complex motion modeling without specialized templates [PMR*23, YVN*22, YWRR23]. These trends motivate the scope of this report, as shown in Fig. 1, which is described next.

1.1. Scope of the STAR

This state-of-the-art report (STAR) aims to provide researchers and practitioners with the background knowledge necessary for understanding the vibrantly evolving field, describe the core new techniques that recently (re-)shaped it, and discuss the recent progress in non-rigid 3D reconstruction.

Each observation level and application necessitates suitable sensor types ranging from LiDAR systems, specialized (multi-view) camera systems, event cameras, and endoscopic visual devices to becoming increasingly widespread RGB+depth (RGB-D) sensors and cameras in mobile devices. We consider all of them in the context of general non-rigid 3D reconstruction in this survey. Note that we do *not* cover methods that make strong assumptions or leverage domain-specific constraints about the observed scenes, such as parametric shape models (e.g., methods reconstructing humans or human faces), which are investigated in different active sub-fields by the community. In this report, we focus on aspects of general non-rigid scenes, i.e., their reconstruction, decompositional scene analysis, editability and control, and the emerging field of generalizable and generative modeling.

Existing surveys cover subsets of these aspects of general non-rigid scenes. The survey on Advances in Neural Rendering (2022) by Tewari et al. [TTM*22] is distantly related to ours. It focuses on Neural Radiance Field (NeRF)-based neural rendering methods for static and dynamic scenes and includes only a few (the first of their kind at the time) NeRF-based techniques for non-rigid 3D reconstruction of deformable scenes. Since then, the field substantially moved forward, and our STAR complements and covers the progress in non-rigid reconstruction. Another related survey published in 2018 focuses on 3D reconstruction with RGB-D cameras [ZSG*18] and also briefly addresses dynamic scenes. Ours provides a substantial update for non-rigid techniques and is significantly more comprehensive by discussing the state-of-the-art methods introduced in recent years. The most closely related survey to ours is Tretschk et al. [TKBR*23]. It also addresses non-rigid 3D reconstruction but focuses on the monocular setting exclusively. In contrast, this STAR addresses the 3D reconstruction of non-rigid scenes from various sensor types, including multi-view, RGB-D, LiDAR, and monocular data. Notably, these sensors enable a much wider variety of applications in different contexts than monocular cameras. For the monocular setting, we complement the discussion of monocular methods compared to Tretschk et al. [TKBR*23] with techniques that appeared in the last twelve months. In addition, we discuss decompositional scene analysis, editability and control, and the emerging field of generalizable and generative models for non-rigid scenes [PYG*23].

We predominantly include approaches published at top-tier computer vision, machine learning, and computer graphics venues from

late 2021 until late 2023. Since the field progresses rapidly, several recent technical reports on arXiv.org are also included.

1.2. Structure of the STAR

The following Sec. 2 reviews the fundamentals necessary to understand and navigate the field. Sec. 3 then presents a comprehensive discussion of the state-of-the-art techniques. It first addresses the general non-rigid 3D reconstruction and novel view synthesis methods in Sec. 3.1 and then proceeds with specific aspects, namely to decompose scenes into parts in Sec. 3.2, and to enable editing and controlling the scenes in Sec. 3.3. Finally, Sec. 3.4 discusses generalizable and generative modeling. The remaining two sections conclude the report by presenting the open challenges in Sec. 4 and giving a conclusion in Sec. 5.

2. Background

Non-rigid 3D reconstruction seeks to infer the time-varying geometry and appearance of a scene. Modeling a deforming sequence requires a scene representation that spatially captures a 3D scene in one of two ways: either with a deformation representation to parameterize how the scene changes from one time instant to another, or an extension of the scene representation into the temporal domain to model the evolution over time. This section overviews these fundamental building blocks of non-rigid 3D reconstruction. We assume knowledge about the fundamentals of computer vision, computer graphics, and machine learning on the reader's part.

We first look at capture settings that provide observations of scenes in Sec. 2.1. In Sec. 2.2, we introduce the data structures that form the basis for representing geometry and appearance. In Sec. 2.3, we then review the fundamental challenges and look at performing non-rigid 3D reconstruction with the given representations and observed data, where we discuss key aspects such as modeling deformations, optimizing the scene model, and incorporating data-driven priors.

2.1. Sensors and Capture Settings

Reconstructing real-world objects and scenes requires capturing the data first. The type and number of camera sensors used to observe the scene influence how ill-posed the problem is and, therefore, influence the required priors and the overall reconstruction quality. Every sensor set-up has advantages and disadvantages regarding consumer availability, cost, and whether it is technically feasible to use in a given setting. This section introduces the various aspects of capturing real-world scenes with a sensor: different sensor types and their parameterizations, and the sensor configurations for capturing a scene.

2.1.1. Camera Model

The *camera model* defines how a position in 3D space is projected to 2D image coordinates, given the *intrinsic* camera parameters. For simplicity, most methods use a pinhole camera model, and real data is often preprocessed to remove lens distortion in advance. *Extrinsic* camera parameters determine the *pose* of the camera in world coordinates using a rotation $R \in SO(3)$ and a translation $t \in \mathbb{R}^3$. Online reconstruction methods usually track camera poses simultaneously while performing the reconstruction. However, it is common practice for most offline methods to compute camera poses beforehand, using structure-from-motion (SfM) methods such as COLMAP [SF16, SZFP16]. The accuracy of the estimated camera poses is vital to the reconstruction quality; inaccuracies can lead to blurry reconstruction results or even failure.

2.1.2. Sensor Types

RGB. Today, practically every smartphone has an RGB camera, making it the most accessible sensor. Each pixel collects incoming photons and generates analog electrical signals transformed into a digital representation (i.e., an RGB image).

Passive Depth. Depth can be obtained *passively* by using two RGB cameras and estimating the disparity between every pixel using the principles of epipolar geometry, which constrains the observations of a 3D point in two cameras to lie on the corresponding epipolar lines [HZ03]. The baseline, focal length, and resolution of the stereo pair determine the depth range. Notably, recovering the depth from stereo RGB images is often ill-posed and results in poor performance for texture-less surfaces and under low-light conditions.

Structured Light. Structured light sensors add to the cost, but provide better results by *actively* sending patterns of visible or infrared light into the scene and observing and analyzing the distortion patterns. However, in direct sunlight, infrared light interferes with the projector's emitted light and prevents the depth measurements. Hence, these sensors are primarily suited for indoor applications.

Time-of-Flight. Analogous to echolocation in bats, these sensors measure depth based on the principle of time-of-flight, i.e., the time it takes for an emitted signal to reach the sender after reflection from the environment. The most common type in this class is the LIDAR sensor, which uses a laser projector to measure depth. While it allows for accurate measurements in sunlight and outdoor environments, the samples obtained are irregular and spatially sparse due to sequential scanning of the scene.

Event. Event cameras are relatively new types of sensors that output asynchronous per-pixel brightness changes over time instead of 2D scene snapshots at pre-defined time instants [LPD08] (like RGB sensors). Each sensor pixel stores a reference brightness and asynchronously fires an event when the brightness change exceeds a threshold with respect to the reference. This results in a sparse signal that only informs about the per-pixel brightness changes in the scene. The main advantage of event sensors is the high dynamic range, which enables accurate sensing even in low-light conditions, and the high temporal resolution (on the order of tens to hundreds of microseconds), which leads to significantly lower motion blur in fast-moving scenes compared to RGB cameras [RGW*21, RETG23, MLR*24].

2.1.3. Scene Capture

Capturing scene dynamics requires sensors to observe the environment across space and time. We define a *frame* as data captured by different sensors at a given time instant. A *single-view* frame contains the data from a single sensor, while a *multi-view* frame contains data from multiple synchronized sensors. A *video* is a

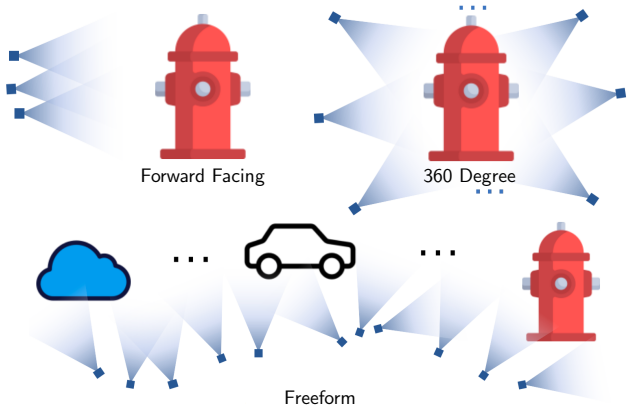

Figure 2: ***Capture Trajectories.*** *A frame can consist of images from a single or multiple cameras installed on a rig. The camera or rig can move along a forward-facing, a 360-circle, or a freeform trajectory. Image source: [WLC*23].*

collection of frames acquired over time along a certain *trajectory*, which may be forward-facing, a 360-degree circle, or freeform (see Fig.2). *Multiple videos* of an object or category can be utilized to build instance-level [YVN*22] or category-level [YWRR23] models. An *image collection* is a set of images of an object or category captured under different states and in different scenes (e.g. images of "dogs" found on the internet), which can be used to learn articulated, category-level models.

2.2. 3D Scene Representations

Capturing a non-rigidly deforming scene requires suitable representations to define various scene properties in space and time. *Geometry* informs us about where surfaces and occupied space are, and *appearance* informs us about the properties of outgoing light from a particular geometry point and how the occupied space looks when rendered into an image. *Deformation* captures how the geometry moves from one time instant to the next. Furthermore, a *compositional* scene representation provides a decomposition into its constituent static and dynamic parts.

In the following, we summarize the most common types of representations used for modeling 4D scenes. We describe a generic 3D representation as a function

$$\mathbf{y} = \rho(\mathbf{x}, \mathcal{H}; \theta), \tag{1}$$

where ρ is the model of the representation, $\mathbf{x} \in \mathbb{R}^3$ are 3D coordinates, $\mathcal{H}$ is a set of optional additional parameters (e.g., the view direction), and θ stores the scene information. The function outputs scene properties $\mathbf{y}$ for the given position $\mathbf{x}$, where some example properties that can be represented with $\mathbf{y}$ are color, irradiance, occupancy, signed distance, density, and BRDF parameters. Naively, all 3D representations can be extended to 4D (modelling time) by introducing a dependency on t for ρ.

An overview of common 3D representations is given in Fig. 3. We first look at discrete representations, which explicitly store scene information θ at discretely defined nodes, with ρ defining the interpolation of information from these nodes to any 3D point. Then, we describe continuous representations, usually implemented as neural networks which store scene information θ implicitly in their weights, before introducing hybrid variants which combine both forms of representations. Later in the section, we explore how scene properties y, i.e. geometry (Sec. 2.2.1), appearance (Sec. 2.2.2) and deformations (Sec. 2.2.3) are defined using these representations and how the scene can be decomposed into its constituent parts and modeled using separate representations (Sec. 2.2.4).

Point Clouds. Point clouds consist of irregularly sampled points in 3D space, which can host scene information θ. Point clouds are adaptive in their sampling and can easily be edited by moving points, unlike volumetric discretizations. A radius-based search is usually employed to interpolate scene information from the nearest points in the point cloud for an arbitrary 3D location $\mathbf{x}$. A point cloud is usually obtained by unprojecting and fusing depth maps if the scene is captured with a depth sensor or sampling a 3D mesh.

Meshes. The natural extension of point clouds in traditional 3D graphics—and the standard representation for any 3D graphics software—is *meshes*, which add connectivity between points, defining polygon primitives usually in the form of triangles, where any 3D point on the polygon can be accessed via barycentric interpolation from the respective vertices. Information can be stored not only on vertices but also on the primitives via u-v mapping, e.g., textures or normal maps. Differentiable renderers for mesh reconstruction allow optimizing elements in the representation [LLCL19, SGY*21]. In the context of deformations, a mesh can also be utilized as an *embedded deformation graph*, where explicit edge connectivity defines the neighborhood that moves together. The graph can be built on multiple levels of resolution and the underlying surface can be represented by TSDF voxel grids [NFS15], surfels [GT18, CRG*23] or points [RMT23].

Voxel Grids. *Voxel grids* are a traditional volumetric data structure and discretize a volume into regular, fixed-size voxels. The scene information is usually interpolated from the grid using trilinear interpolation. The main drawback is the cubic growth of memory requirement with the resolution. The efficiency is usually improved in one of the following ways (see bottom left part of Fig. 3):

- **Hashing:** Voxel hashing [NZIS13] introduces lookup hash tables to retrieve the values stored at voxels, improving the efficiency and scalability of voxel grids.
- **Octrees:** The 3D space is usually sparse and irregularly populated, so voxel grids can be made more memory efficient by using Octrees [Mea80] that hierarchically subdivide the utilized space and allow adaptive resolution.
- **Tensor Factorization:** TensoRF [CXG*22] introduced classical tensor decomposition techniques [CC70, DL08] in the context of scene representations, factorizing a voxel grid using low-rank vector and plane components, in turn allowing compact, memory-efficient representation, and efficient sampling for high resolutions. Interpolating scene information for an arbitrary 3D position $\mathbf{x}$ usually requires projecting the coordinates of $\mathbf{x}$ onto the respective low-rank components first. A special case is planar factorizations of a volume, also called *tri-*

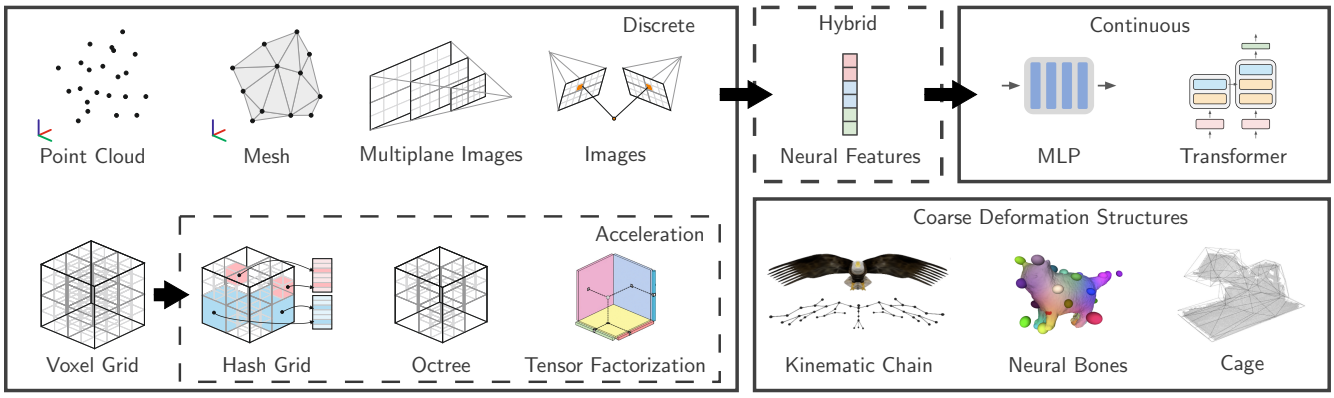

Figure 3: ***Common Scene Representations for Non-Rigid Reconstruction.*** *Scene representations can be discrete, such as point clouds, meshes, and grids, or continuous, such as MLPs or Transformers. Both paradigms can optionally be combined, where feature embeddings for continuous neural representations are stored in discrete structures. Some scene representations are only used for coarse deformation modeling, such as those shown in the bottom right. Image sources: [JLX*23a, XH22, YVN*22, KKK*23].*

planes [PNM*20, CLC*22a], which uses an axis-aligned plane for each spatial dimension. These representations have been recently extended to the temporal domain as well, by similarly factorizing the time dimension along with the spatial dimensions [CJ23, SZT*23, AHR*23, FKMW*23].

Multi-Plane Images (MPI). A static 3D scene can be represented by a discrete set of planes at varying distances to the camera in the camera frustum [ZTF*18]. Each plane encodes the RGB color and the alpha values. The MPI representation supports efficient novel view rendering, involving only homography warping and back-to-front alpha blending, but is limited to small viewpoint changes.

Image Sets. A set of captured images, along with the reprojection function and a feature aggregation strategy, can be used as a 3D scene representation in an image-based rendering setup. Scene properties at a 3D location are inferred by projecting it into all the images and retrieving the information from there. Usually, pixel-aligned features predicted by a CNN are used [YYTK21, WWG*21, LWC*23a], which lift the color information to a feature space with prior information about shape and appearance.

Multi-Layer-Perceptrons (MLPs) and Transformers. A few years ago, several works [PFS*19, MON*19, CZ19, SZW19] introduced using multi-layer perceptrons as a continuous representation for scenes. For such representations, called *neural fields*, ρ is a coordinate-based MLP with parameters θ that defines a continuous function over a volume, modeling properties of the scene at every possible point $x \in \mathbb{R}^3$. Neural fields hold the properties of being resolution-independent, continuous, and biased towards learning low-frequency functions [RBA*19]. To effectively allow modeling higher frequencies, positional encodings in the form of Fourier features are employed [TSM*20]. To consider relations between sample 3D points, usually spatial along a ray or temporal across time, a *transformer* architecture [VSP*17] can be utilized which adds a cross-attention mechanism to model these relations.

Hybrid. In the original continuous formulation, neural fields are globally parameterized, meaning that one set of MLP parameters encodes the representation of the whole scene. This entanglement slows optimization, as parameter updates for one location change the scene in many places, and many iterations are required to counteract this side effect. It also leads to scalability issues, as encoding larger scenes requires larger MLPs for more representation capacity. To mitigate these undesirable properties, MLPs are paired with discrete data structures to restrict their domain to a localized scene region [CLI*20]. Such pairing can be done by storing latent feature embeddings in a discrete data structure $\mathbf{z} = \rho(\mathbf{x}, \theta)$ (e.g., a voxel grid or point cloud) and conditioning an MLP Φ with weights ω on the spatially interpolated feature embedding $\mathbf{y} = \Phi(\mathbf{z}; \omega)$. During optimization, only the neural feature embeddings in the local neighborhood of the input point—determined by the interpolation used—are optimized (in auto-decoder fashion, see Sec. 2.3.4) together with the MLP weights. Hybrid methods usually gain efficiency and rendering quality in exchange for memory consumption and a locality assumption.

Coarse Deformation Structures. Certain representations are only utilized to define deformations on a coarse level. These representations use coarse, low-dimensional structures—which drive finer deformations—mainly serving two purposes: articulated control over object pose and motion regularization. Commonly used representations are:

- **Kinematic Chain:** These are defined by joints that are linked together with bones, forming a *skeleton*. Each bone defines the transformation relative to its parent bone in the skeleton up to the root joint. The transformation is commonly parameterized either as a screw [JLCN21, LDS*23], with two kinds of possible motion: a revolute rotation $\theta \in \mathbb{R}$ or a prismatic translation $t \in \mathbb{R}$, or with six degrees-of-freedom using a rotation and a translation (see Sec. 2.2.3 for more parameterizations). The skeleton is usually rigged to a geometry template. For category-specific methods, the topology and rest pose of the skeleton are pre-specified, and the joint locations are fitted to the observations. Category-agnostic methods use morphological techniques [Blu67] or data-driven prediction [XZK*20] to extract the skeleton.
- **Neural Bones:** These are defined by bones only, which are vol-

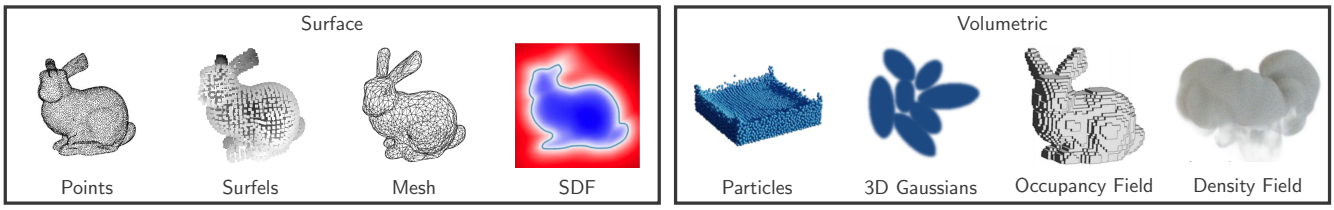

Figure 4: ***Representing Geometry.*** *With the different representations introduced in Fig. 3, geometry can be modeled either as surface or as volume. Image sources: [AP09, PFS*19, RBSC21, LCK*22, CLZ*22, GDWY22].*

umetrically predicted using an MLP, making them a part of *neural parametric models*—a class of methods where the articulation space of an object is parameterized by an auto-decoded MLP (see Sec. 2.3.4), thus bypassing the hand-crafted, object-specific constraints required by traditional parametric models like SMPL [LMR*15]. Each bone is associated with a rigid transform $T \in SE(3)$, which defines its location and orientation in space. Its influence on the surroundings can be defined by a 3D Gaussian ellipsoid that moves along with the bone [YVN*22].

- **Cage:** Rather than being inside the surface, like a skeleton, a *cage* [SP86] is an instance of a general free-form deformation scheme where the deformation of every point in space within an enclosing volume is defined by the deformation of the enclosing shape. The vertices of a cage are used as handles to deform the underlying surface volumetrically and can model certain deformations more naturally than a skeleton, e.g. a breathing character, torsions, or local scaling.

Based on the coarse deformation structure, blend skinning is used to deform geometry, modeling articulated, i.e. piece-wise rigid motion. Skinning weights are defined for each point in the scene representation with respect to each node, determining its influence on the point. Skinning then defines a deformation field over the scene representation, which interpolates the node deformations based on the defined skinning weights for each point, resulting in a non-rigid deformation based on piecewise rigid segments. Standard skinning functions are:

- **Linear Blend Skinning (LBS):** $\mathbf{x}' = \sum_i \mathbf{T}_i w_i \mathbf{x}$, where the $\mathbf{T}_i \in SE(3)$ and w_i defines the corresponding skinning weight for i-th bone respectively. Point $\mathbf{x}$ is deformed to $\mathbf{x}'$ through the corresponding weighted transformation. Being a basic linear interpolation, it suffers from several visual artifacts, most notably volume loss, self-intersection, and the *candy wrapper* artifact around the joints.
- **Dual Quaternion Blending (DQB):** $\mathbf{x}' = SE3(\frac{\sum_i w_i \hat{\mathbf{q}}_i}{||\sum_i w_i \hat{\mathbf{q}}_i||})\mathbf{x}$, where $\mathbf{T}_i$ is replaced by a unit dual-quaternion $\hat{\mathbf{q}} \in \mathbb{R}^8$ and $SE3(\cdot)$ converts a quaternion back to a rigid transform. It provides higher quality interpolation than LBS and resolves the candy wrapping artifacts, but can exhibit bulging effects in some areas.

2.2.1. Representing Geometry

Using the representations introduced in the last section, a scene's geometry can be described either volumetrically or by surfaces. We provide an overview in Fig. 4. In the following, we first introduce how geometry can be described via discrete primitives such as points, meshes, surfels, and 3D Gaussians. Afterward, we look at signed distance functions, density fields, and occupancy fields, all of which can either be defined continuously or using regular discrete representations such as voxel grids.

Points and Meshes. The already introduced representations of point clouds and meshes can directly define a surface. Points—called *particles* in this context—can also be used to represent objects and fluids volumetrically, usually when their physical properties, such as elasticity and viscosity, and the deformation dynamics under applied forces need to be modeled based on the principles of continuum mechanics [SB12].

Surfels. Surfels [PZvBG00] are 2D disks hosted by point clouds, which locally approximate a surface. They are defined by their center, radius, and normal, determining the orientation. The density of the point samples determines the fidelity of the sampled surface.

3D Gaussians. Similarly, rendering via splatting Gaussians or ellipsoids has been a traditional technique for many years [BHZK05, RPZ02, ZPvBG01]. Very recently, 3D Gaussians are going through a renaissance, enabled by the introduction of a differentiable and efficient method for view synthesis [KKLD23]. It allows obtaining radiance fields of Gaussians via optimization from images. The Gaussians are hosted by a point cloud and represented by a 3D scale and a 3D orientation. In addition, view-dependent appearance parameterized through spherical harmonics, opacity, and other properties can be associated with the Gaussian. Rendering a scene composed of 3D Gaussians is done by splatting all the Gaussians to an image and performing alpha compositing based on depth.

Signed Distance Functions. A *signed distance function* (SDF) specifies the distance to the closest surface at each point, with the distance usually being positive outside and negative inside the object, therefore implicitly representing a surface. Therefore, the zero crossing of the SDF represents the surface and can be found by methods such as sphere tracing [Har96] or marching cubes [LC87]. Recent methods have combined SDFs with density fields and volumetric rendering [WLL*21, OPG21, YGKL21], utilizing the strength of both, i.e. obtaining accurate surfaces and utilizing the graceful reconstruction properties from density fields.

Density Field. A *density* measure $d \in [0, \infty)$ models how much light travels through a specific point in 3D space and how much is reflected; together with color, it is suitable for representing non-solid and solid volumetric objects.

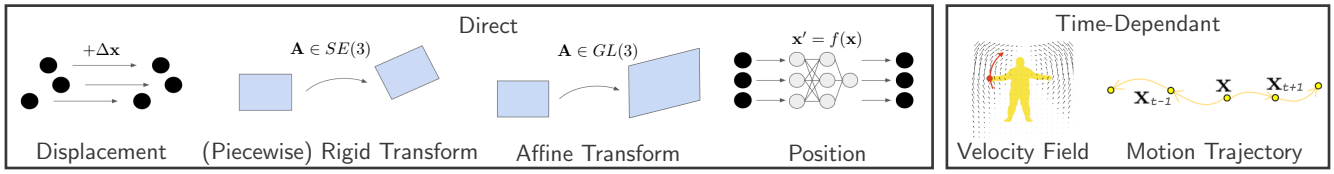

Figure 5: ***Representing deformations.*** *Parameterizations include transformations for directly deforming between two timesteps (left) or modeling deformations over multiple timesteps (right). Image sources: [NMOG19, LWC*23a].*

Occupancy Field. An *occupancy* measure $o(\mathbf{x}) \in [0,1]$ defines the probability that the space is occupied. It can be considered analogous to the *opacity* of a point. A common way of extracting surface S is by thresholding the occupancy function with τ: $S = \{\mathbf{x} \mid o(\mathbf{x}) = \tau\}$.

2.2.2. Representing Appearance

It is important to note that the observed color is not directly a property of an object, but the result of an interaction of environment light and the scene material properties. Let ω_o be the outgoing ray direction of a surface point $\mathbf{x}$ with normal $\mathbf{n}$ and $L_e(\omega_o, \mathbf{x})$ and $L_r(\mathbf{x}, \mathbf{n})$ be the emitted and reflected light respectively. The rendering equation describes the physical light transport in the scene:

$$L_o = L_e(\omega_o, \mathbf{x}) + L_r(\mathbf{x}, \mathbf{n}) \text{ and} \tag{2}$$

$$L_r(\mathbf{x}, \mathbf{n}) = \int_{\omega_i \in \Omega} f_r(\omega_i, \omega_o, \mathbf{x}) L_i(\omega_i, \mathbf{n}) d\omega_i \,, \tag{3}$$

with Ω being the hemisphere around the surface point $\mathbf{x}$ and ω_i the incoming ray direction. Here, f_r is the Bidirectional Reflection Distribution Function (BRDF) that models how the surface reflects light and $L_i(\omega_i, \mathbf{n})$ is the incoming light from direction ω_i, which may either directly come from the environment, i.e. *direct illumination*, or may *bounce* around the environment before reaching $\mathbf{x}$, i.e. *indirect illumination*.

In the context of this report, which focuses on general scenes, all methods either assume a *Lambertian* surface or use the simplification of baking in the incoming radiance. The outgoing radiance is either constant in all view directions or has a view-dependence, often modeled as a *radiance field*. These fields represent the scene as tuples of density σ and view-dependent irradiance $c(\mathbf{x}, \mathbf{d})$—parameterized either by a view direction-conditioned MLP [MST*20] or Spherical Harmonics [FTC*22] —in continuous 3D space. The representation can be translated into images by *volume rendering*, where the color $C(r)$ of a pixel is determined by weighted accumulation along a camera ray r [MST*20]:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), \mathbf{d}), \tag{4}$$

where $r(t)$ is the sample point along the ray at depth t, t_n and t_f are the near and far sample points, $\sigma(r(t))$ is the density at $r(t)$ and $T(t) = exp(-\int_{t_n}^{t} \sigma(r(s))ds)$ is the transmittance, specifying how much of the light emitted from $r(t)$ is visible in the rendered image. To accumulate radiance, different sampling strategies exist [LGTK23]. Some methods leverage ray transformers [WWG*21, LWC*23a] to capture contextual information using self-attention by modeling all ray samples simultaneously.

The simplification of baking in the incoming radiance does not hold when an object moves as the incoming direction changes, and modeling illumination accurately through the light transport process remains an open challenge (see Sec. 4). It is common to model the illumination changes due to motion by allowing unspecific changes of the surface color and capturing them through a per-time-step latent appearance code [PSH*21].

2.2.3. Representing Deformations

In the previous subsections, we examined how to represent static scene geometry and appearance. For dynamic, non-rigid scenes, the geometry deforms over time. Fig. 5 provides an overview of representations for such deformations. Formally, a 3D deformation $d : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}^3$ maps a point at coordinates $\mathbf{x} \in \mathbb{R}^3$ at time t to deformed coordinates $\mathbf{x}' \in \mathbb{R}^3$ at time t': $\mathbf{x}' = d(\mathbf{x}, t')$. Here, d performs a *forward warp*. The inverse $\mathbf{x} = d^{-1}(\mathbf{x}', t)$ is called a *backward warp*. We refer to models which have a well-defined forward warp and backward warp as *bidirectional*. In general, we distinguish between the following basic deformation models, where all parameters can either be defined explicitly or predicted by an MLP:

- **Position:** $\mathbf{x}' = \phi(\mathbf{x})$ where ϕ is usually a neural network. It can model arbitrary non-rigid deformations of point $\mathbf{x}$ to point $\mathbf{x}'$.
- **Displacements:** $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$ describes a displacement of the data point $\mathbf{x}$, where $\Delta\mathbf{x}$ is known as scene flow [VBR*99].
- **Rigid Transforms:** $\mathbf{x}' = \mathbf{R}\mathbf{x} + \Delta$, where $\mathbf{R} \in SO(3)$ is a 3D rotation and $\Delta \in \mathbb{R}^3$ is a translation.
- **Affine Transforms:** $\mathbf{x}' = \mathbf{A}\mathbf{x} + \Delta$, where $\mathbf{A} \in \mathbb{R}^{3\times3}$ and $\Delta \in \mathbb{R}^3$. $\mathbf{A}$ allows scaling and shearing in addition to rotation.

The above deformation types are independent of time and allow transforming only from one discrete time step to the next. In contrast, the following concepts allow us to model continuous transformations in time:

- A **velocity field** $v(\mathbf{x}_t, t)$ (c.f. [NMOG19]) can model motion of a particle $\mathbf{x}$ over a time interval τ via integration over time: $\mathbf{x}' = \mathbf{x}_0 + \int_0^\tau v(\mathbf{x}_t, t)dt$.
- A **motion trajectory** can be defined over a sequence of motion basis vectors $\{\mathbf{h}_k\}_{k=1}^K$ as $\mathbf{x}(t) = \sum_{k=1}^K \phi_{\mathbf{x},k}\mathbf{h}_k$, where $\phi_{\mathbf{x},k}$ are the coefficients used to obtain the trajectory $x(t)$ for point $\mathbf{x}$. Smaller K limits the expressiveness of the trajectory, thus introducing regularization. The displacement between two timesteps is given as $x(t_2) - x(t_1)$. As a basis, discrete cosines are a common choice [WELG21, LWC*23a].

When modeling deformation, there is a trade-off between data structure size and expressivity of deformation. Thus, a single affine

transform would fail to accurately represent motions of most deformable scenes. In contrast, modeling the deformation through a dense displacement field can describe arbitrary motion but requires a larger data structure. Moreover, the expressivity of deformation also acts as a regularization in reconstruction tasks. Note that these representations are not approximations of the underlying physics-based deformations. The latter can be modeled with physics simulation, most commonly based on the Finite Element Method [ZT00] or the Material Point Method [JST*16].

2.2.4. Compositional Representations

Real-world scenes are complex, consisting of multiple dynamic objects, each undergoing different motion, along with a background. Scene-level methods reconstruct all elements while object-level methods focus on reconstructing a single object. Furthermore, modeling the whole scene with a single representation might not be ideal. In this subsection, we introduce possible spatial and motion decompositions.

Spatial Decomposition. Decomposing the scene into its constituent parts allows modeling it more efficiently and enables certain useful properties. It can be represented either by one segmented model or by individual models for each decomposed region. For the latter, novel views can be compositionally rendered using blending weights (usually automatically arising from densities). Different types of spatial decompositions are:

- **Static/Dynamic Segmentation**: Separates the scene into a static background and dynamic foreground model. It is most commonly used as preprocessing to estimate the camera pose from the static background [LGM*23].
- **Instance Segmentation**: Provides masks for each object, which allows modeling individual object motion and inter-object dynamics [DHL*23]. Bounding boxes may further localize objects in the scene [TZFR23]. Identified scene elements can be edited by manipulating the individual object models, while it also makes it possible to introduce object-level priors.
- **Semantic Segmentation**: Additionally provides class labels for each object. Allows modeling class-level properties, e.g. object rigidity [CRG*23].
- **Motion Segmentation**: Groups parts with similar motion together. It is used commonly as a preprocessing step for methods that enable articulated control over the object [YZH*24].

Motion Decomposition. The motion of an object can also be decomposed according to different levels of expressivity. To model the *rigid* motion of individual objects based on the spatial decomposition, the *root pose* of each object can be estimated with respect to world coordinates (e.g. a trajectory of multiple humans in a room), with residual motion modeled in the object's own space. The simpler modeling allows handling larger deformations [SYD*23, WDSY23]. The motion of an individual object can be further decomposed into *articulated* and *non-rigid* (e.g. clothing deformation on top of human motion), where the articulated motion can be represented with a coarse deformation structure (see Sec.2.2).

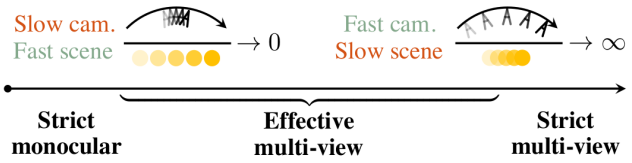

Figure 6: ***Effective Multi-View Setting.*** *A combination of fast scene motion and slow camera movement does not give enough observations of each point in the monocular setting, and depth ambiguities remain. Fast camera motion with slow scene movement provides plenty of observations for each scene point in a specific state, resolving depth ambiguities and effectively turning the monocular setting into multi-view. Image source: [GLT*22].*

2.3. Reconstruction

In the previous sections, we looked at the sensors and the capture settings that provide us with observations of a non-rigid scene and the models to represent it. Reconstructing a non-rigid scene requires optimizing one of these models to match the observed data. Once the reconstruction of geometry and appearance is obtained, *novel view synthesis* can be carried out by rendering the scene from unobserved viewpoints or times. In this section, we briefly outline the main challenges of non-rigid reconstruction in Sec. 2.3.1 and take a brief look at the key aspects of reconstruction: the design choices for modeling deformations of a non-rigid scene (Sec. 2.3.2), optimizing the scene model (Sec. 2.3.3) and leveraging data-driven priors (Sec. 2.3.4).

2.3.1. Challenges of Handling Non-rigid Scenes

Obtaining a 3D reconstruction of a non-rigid scene from a set of view-dependent observations is an inherently ill-posed problem, especially from a monocular sensor, as only one observation is available for each surface point undergoing deformation and multiple scene configurations can project into the same set of sensor readings. Depth sensors and multiview capture settings alleviate some issues by providing more constraints but suffer from inaccuracies and occlusions. We briefly discuss the main challenges in reconstructing non-rigid scenes in the context of different sensor and capture settings.

Depth Ambiguity. Depth cannot be recovered from a single monocular RGB observation, as all 3D points along a ray project to the same 2D observation. If the scene stays static and the camera moves, we can reconstruct depth where discriminative image features are present. However, if the object deforms, the relative motion between the camera and the scene can result in ambiguity and prevent reconstruction (see Fig. 6). Using a multiview capture setting reduces the ambiguity for reconstructing depth at each time step while using a depth sensor allows surface reconstruction, subject to the accuracy and noise of the sensor.

Occlusions and Non-Rigid Loop Closure. Even in dense multi-view capture setups, self-occlusions still occur. A significant challenge is recognizing when an occluded region becomes visible and whether it has been seen before. As the region may undergo non-rigid deformations while occluded, reintegrating new observations into the reconstruction, i.e. non-rigid loop closure, becomes hard.

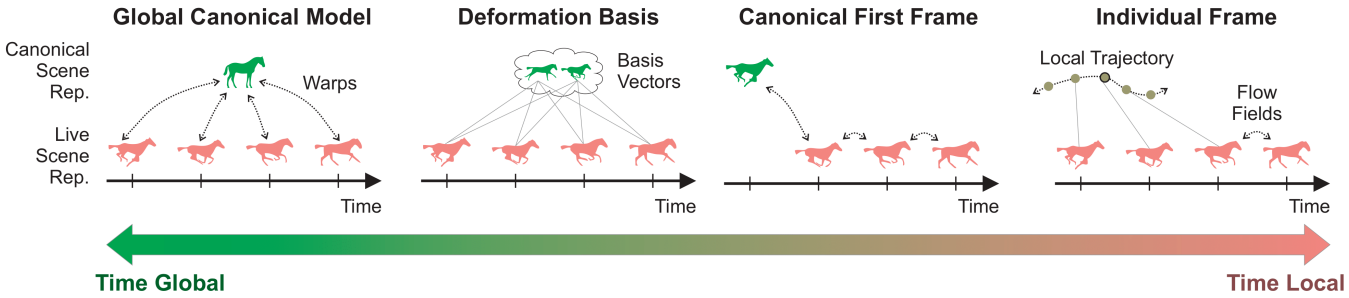

Figure 7: ***Spatio-temporal Modeling.*** *We categorize different spatio-temporal models according to the time consistency of their 3D representation. On one end of the spectrum, a representation with time-global consistency is assumed, which can be warped to all timeframes. The other end shows purely local models that allow the representation of each frame individually.*

View-Dependent Appearance. If observed surfaces have a view-dependant appearance, different colors of the same geometry point might be observed from different views and if the object deforms. In the inverse rendering setting, attributing the difference to geometry or view-dependent effects is ambiguous, as different 3D points with different view-dependent appearances can produce the same observations. A depth sensor can alleviate this ambiguity by providing geometry measurements.

2.3.2. Spatio-Temporal Modeling

When designing a model to capture a non-rigidly deforming scene, one important design choice that needs to be considered is *consistency over time*. A *globally consistent* model tracks a canonical geometry over all time steps. Such a property is helpful for tasks like virtual asset creation as well as motion analysis and editing, where a consistent geometry is required with correspondences to each deformed state. However, canonical space modeling limits the model's ability to handle large motions and topology changes, which deviate significantly from the canonical geometry. On the other extreme of the spectrum is modeling a separate geometry per time step, in which consistency is sacrificed, and individual reconstruction is performed. According to consistency, we structure existing models into four different categories: (1) *global canonical models*, (2) *deformation bases*, (3) *canonical first frames*, and (4) *individual frame modeling*. An overview is given in Fig. 7.

Global Canonical Model. In this paradigm, the scene representation is defined in a canonical space, i.e. a temporally global 3D representation. The motion representation is disentangled from this scene representation and modeled by a 4D deformation field. This automatically ensures long-term correspondences and explicitly defines a mapping between canonical points and their deformed states that correspond to the live frame, allowing consistent novel view synthesis and smooth interpolation in the time dimension. As the ability to handle topology changes is limited, this paradigm is more suitable for object-level modeling that has a temporally global representation. The generalizability of the deformation to new poses and novel views greatly depends on the direction of the modeled deformation field:

- **Backward Models:** For such models, the 4D deformation field represents the mapping of deformed states back to the canonical state at each time step. Backward models are based on the Eulerian view, i.e. how material flows through a fixed location over time. Hence, canonical frame tracking with these models is not temporally smooth, as over time different parts of the scene geometry will occupy a certain 3D location (see Fig. 8). This hampers generalization to new poses in the case of skinned models [CZB*21] and makes it hard to be represented by smooth models like MLPs [GSD*23]. Note that the geometry fused in canonical space can exhibit distortions due to imperfect backward warps [GSD*23].
- **Forward Models:** These models define the 4D deformation field from the canonical space to the deformed state of the frame at each time step (also referred to as *live frame*). The canonical geometry is either given as a template [LM18], fused from observations [LZYX22], or generated by a neural network [UEK23]. Tracking fixed points on a canonical geometry is temporally smooth (see Fig. 8). This allows better fitting of the deformation fields with MLPs [GSD*23] and better generalization to new poses in the case of skinning since time-invariant skinning weights can be learned in the canonical space [CZB*21]. However, to define warps to the live frame, the canonical space needs to be discretized. Forward modeling is based on the Lagrangian view, i.e. tracking motion over time associated with the same scene element. Thus, it allows defining classical spatial constraints on the deforming surface like ARAP [SA07] and isometry [PMR*23], and is easier to edit.
- **Bidirectional Models:** Non-rigid deformations are naturally bijective. Modeling backward and forward warps to and from the canonical space, and defining cycle consistency between them [CFF*22, YVN*22] enforces this natural property, improving the quality of the learned canonical representation.

Deformation Basis. Geometry in live frames can be represented as a combination of basis elements [CJ23, NRS*22]. The number of basis elements controls the amount of consistency over time and exposes control over the global vs. local trade-off. Intuitively, each basis vector can represent the scene in a certain state, and the combination of these states defines the extent of deformations that can be modeled, restricting the dimensionality of the model and thus providing regularization. Low-rank basis representation can also be introduced in the latent space, e.g. for auto-decoded latent codes of neural parametric models [YVN*22] or scene deformation MLPs [SGP*22].

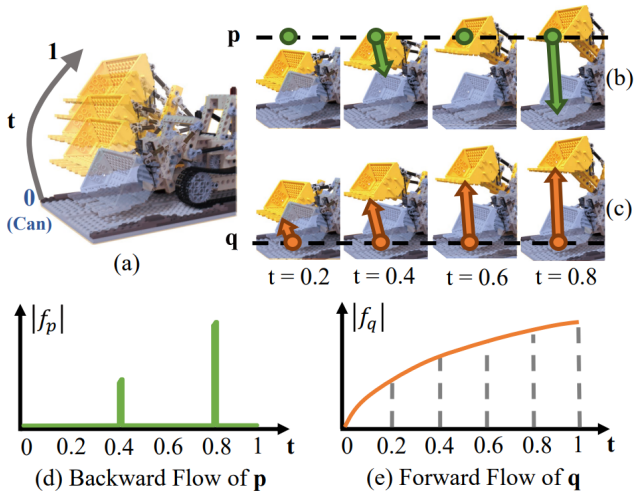

Figure 8: ***Backward vs. Forward Flow.*** *(a): A dynamic scene across time t; (b & d): Backward flow f_p that maps the live point **p** to the canonical frame with the corresponding norm, which is not smooth; (c & e): Forward flow f_q that maps the canonical point **q** to the live frame with the corresponding norm, which is smooth and continuous, benefiting motion model learning. Image source: [GSD*23].*

Canonical First Frame. These models propagate canonical geometry forward by defining deformation between time steps. In contrast to model-to-frame tracking of forward models, the trajectories obtained by frame-to-frame tracking are more susceptible to drift. Paradigms include tracking points across frames using rigid transforms [LKLR24] and using velocity fields as deformation representation [NMOG19], both of which are naturally invertible.

Individual Frame Modeling. Instead of tracking geometry from a canonical frame, a separate geometry can be reconstructed at each time step. These models directly add a time dimension to the scene representation. Due to a lack of geometric constraints between time steps, they can model a larger range of motion and arbitrary topology changes. However, because of the lack of time consistency, their novel view synthesis ability heavily depends on observation density. Common paradigms include extending the 3D neural field to 4D by additionally conditioning the MLP on time (or a per-time latent code, which provides a more compact representation [LSZ*22]), called *Space-Time Neural Fields* [LNSW21, XHKK21], and using a view-dependent scene representation at each time step, such as Mulitplane Images [TS20, ZW22]. Individual frame representations can be made *locally consistent* by modeling:

- **Flow Fields:** These define a warp to the next and previous state for each visible point.
- **Local Trajectories:** These allow extrapolating a point's motion across several time steps in the visible temporal neighborhood. A low-rank linear basis usually represents the trajectory, e.g. DCT [WELG21], which provides motion regularization.

Local consistency is then achieved by using the field or trajectory to do cross-time rendering [LWC*23a, LNSW21] or enforce cycle consistency [GSKH21, LNSW21, TZFR23].

2.3.3. Model Optimization

Optimization-based reconstruction involves finding model parameters θ, which are most likely to have produced the observed data. The optimization consists of a *data term*, which forces the solution to match the observations, and optional *prior terms*, which regularize the solution:

$$\theta^* = \arg\min_{\theta} L_{\text{data}}(\theta) + L_{\text{prior}}(\theta). \tag{5}$$

Next, we describe the most commonly used data terms, strategies for regularizing the solution and metrics used to evaluate the quality of reconstruction.

Data term. If the 3D capture data is available, e.g. from depth sensor or synthetic scenes [LTT*21], then the model can be directly supervised in 3D using a *geometric loss*. If only 2D image observations are used, we can render the scene into the observed views using a differentiable renderer and either apply a *photometric loss*, like l_1 and l_2, or a perceptual loss like LPIPS [ZIE*18].

Constraints, Regularization and Priors. The non-rigid reconstruction problem is highly underconstrained, as a potentially infinite number of solutions can explain an object's deformation between two timesteps (see Sec. 2.3.1). Many methods leverage off-the-shelf approaches to extract additional information from observations and utilize it to constrain the reconstruction. Common techniques are:

- **Object Masks:** Segmentation (see also Sec. 2.2.4) can be enforced either by applying the mask to the image as preprocessing or through a segmentation loss between the rendered mask from the model and the extracted mask from an off-the-shelf method.
- **Optical Flow:** To enforce 2D-3D motion consistency, predicted scene flow—which is the 3D variant of optical flow—can be constrained to match 2D optical flow after projection. RAFT [TD20] is a popular method for computing optical flow.
- **Feature Distillation:** Information from 2D features like DensePose embeddings [GNK18] can be distilled to 3D canonical feature embeddings for long-term registration [YVN*22]. 3D semantic features can also be distilled from 2D feature embeddings like DINO [CTM*21].
- **Pseudo Depth:** If a depth sensor is not available to measure the scene geometry, then a monocular [RLH*20] or video [ZCT*21] depth estimator could be used to get pseudo-depth estimates and impose constraints on the optimized geometry.

Further, techniques for soft regularization can be used to drive the optimization to a better local minimum. Representing deformations with an MLP for example inherently provides some regularization, as MLPs are smooth function approximators. Another way to provide regularization is to use skinning with a coarse deformation structure (see Sec. 2.2.3). For an explicit surface representation, spatial and temporal regularization can be introduced on the geometry points. Common geometric priors are:

- **Local Rigidity:** Specifies that neighboring geometry points should deform similarly, commonly known as the as-rigid-as-possible (ARAP) constraint [SA07].

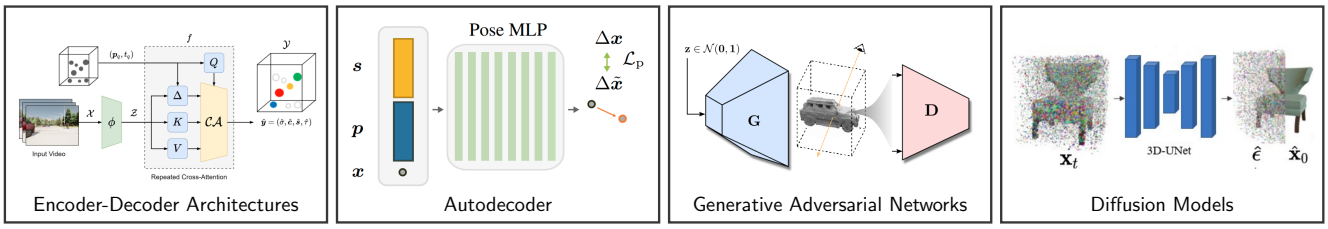

Figure 9: ***Methods for Data Priors in Non-Rigid Reconstruction.*** *The figure shows methods to capture dataset distribution in the context of non-rigid reconstruction. Prominent are different types of encoder-decoder architectures, autodecoders, GANs, and recently diffusion models. Typically, shape/appearance priors are learned from static datasets. Non-rigid aspects can be learned on top by modeling optical/scene flow, articulation spaces, or 2D/3D correspondences. Image sources: [PBT*21, VHTS*22, SSN*22, MSP*23].*

- **Isometry:** Preserves distance between two points on a manifold in some reference geometry, when they are deformed to an observed state [PMR*23, RMT23].
- **Small Motion:** Real motion, when captured at a high enough frame rate, is similar for deformations that are temporally close. This can be enforced through a similarity loss between the subsequent deformations [PPGT*23].

Other strategies include coarse-to-fine optimization [TGZ*24, MESK22], introducing a total variation loss which forces the feature embeddings for a representation to be spatially and temporally smooth [CJ23, WYF*24] and online tracking, where the model at the current time step is initialized from the previous time step [WHH*23b, TGZ*24, PMR*23]. The introduction of differentiable physics simulators [HMC*21, LLK19] has also allowed them to be used as a prior for the deformation modeling in reconstruction tasks [YYZ*23, KTE*22].

Evaluation Metrics. Evaluating the quality and faithfulness of the reconstructed scene requires comparing it to ground-truth data. Common ways to evaluate are:

- **Rendering Quality:** In the case of 2D observations, renders of the reconstructed scene can be compared with the observed views or ground-truth novel views using Peak Signal-to-Noise Ratio (PSNR), which measures faithfulness in absolute terms, or perceptual metrics like Structural Similarity (SSIM) [WBSS04] and LPIPS [ZIE*18].
- **Geometry Quality:** If 3D ground-truth data is available, then the reconstructed 3D geometry can be compared directly using chamfer distance in the case of point sets or mean absolute error (MAE) in the case of SDFs. Estimated normals can also be evaluated using mean angular error (MAE).
- **Trajectories:** The deformation of a scene can be evaluated using 3D point trajectories or their projection in 2D views. Absolute Trajectory Error (ATE) or Median Trajectory Error (MTE)—which is more robust to outliers—checks the global consistency with the ground-truth trajectory. Robustness can further be measured using average position accuracy, which measures the percentage of point tracks within a threshold distance to ground-truth, and survival rate, which is the average number of frames until tracking failure [ZHS*23]. Along with ATE, methods that also track camera poses [LGM*23, RMT23] usually evaluate the estimated camera trajectory using Relative Pose Error (RPE), which is well-suited to measure the drift [PZB*19].

2.3.4. Learning Data-Driven Priors

An alternative way to introduce additional constraints into the optimization framework of Eq. 5 is to model the distribution of a given dataset and use this distribution as a data prior for the reconstruction task. Four general techniques for modeling data priors with deep learning in non-rigid scenarios are (see Sec. 3.4 for advancements): *encoder-decoder architectures*, *autodecoders*, *generative adversarial networks*, and *diffusion/flow models*. The first two provide maximum likelihood estimates. The latter two naturally enable to sample from the learned distribution, which makes them ideal candidates for generative methods. While variational autoencoders, as part of the first category, also allow sampling from the posterior, they are rarely seen in the context of 4D modeling.

Naively applying these methods in a general 4D setting is intractable and requires quantities of data that do not exist today. Thus, the trend is to learn data priors in more creative ways, focusing on certain aspects instead of on the full data distribution. We categorize data priors of existing works into the following types:

- **3D Shape/Appearance Priors**: Learning common geometry and appearance distributions from a dataset of static objects and scenes. If the learned latent spaces are global, they often already enable deformation via interpolation and latent trajectories.
- **3D Deformation/Flow Priors**: Learning to predict where a set of entities will move in the immediate future, or how a canonical representation deforms over time. Learned from a dataset of videos or 4D scenes.
- **Articulation Priors**, learning how a category of objects can articulate from a dataset of moving/deforming objects. Usually controllable with a low-rank decomposition representation.
- **2D Correspondence Priors**: Learning to find correspondences between images or between images and 3D from a dataset with correspondence labels or via optical flow.

To model individual distributions of these kinds, data priors are captured individually and used as constraints in the full reconstruction setting. In the following, we describe the individual building blocks and their application in more detail (see Fig.9).

Encoder-Decoder Architectures. A standard technique in modeling data distributions is autoencoders or more general, dense prediction architectures in which input and output modalities can differ. In the scope of this work, they appear in their 3D and 2D variants: some 3D variants take voxelized point clouds as input and

decode them into dense occupancy, SDF, or 3D flow fields (see Sec. 2.2.3). Other 3D variants directly work on sparse point clouds and produce point-level predictions, such as scene flow [LQG19]. Encoder-decoder architectures on 2D data also come in many different flavors. A dominant category is 2D-to-3D models, which take images as input and produce (temporal) feature volumes, which can be rendered from arbitrary views [RZS21]. Also, several 2D-to-2D models find their application as an additional constraint on 3D modeling, such as optical flow predictors (e.g. RAFT [TD20]), surface embeddings (e.g. CSE [NNS*20], DensePose [GNK18]), or semantic features (e.g. DINO [CTM*21]). An advantage is that these data priors often come off-the-shelf as pre-trained (foundation) models and do not need to be trained on large-scale 3D or 4D data.

Autodecoders. The second category is autodecoders [PFS*19]. In comparison to general encoder-decoder architectures, they omit the encoder and find the latent representation via optimization instead. During training, the decoder learns a function space, which is fixed during inference, where the representation is found via test-time optimization, selecting a specific instance in the function space. They can be used in settings where designing an encoder is difficult or where the encoder fails to capture fine-grained details in the observation. As a downside, test-time optimization is typically slower than forward encoding during inference. Autodecoders are typically used to learn spaces of deformable or articulated SDFs, or deformable point templates [PBT*21, WISL23].

Generative Adversarial Networks. GANs as traditional generative models for 2D data find their application for non-rigid generation as well. Since generating dense 3D/4D data with GANs is computationally challenging, existing approaches generate some intermediate representation, such as triplanes for shape [CLC*22b] or neural field MLP weights [BPP*23].

Diffusion Models. The last approach that appears, and arguably the most important for the immediate future, is diffusion models. Given a high-dimensional random variable $\mathbf{X} \in \mathbb{R}^d$ (which can model the representation or the data itself), diffusion models [HJA20], and related methods, such as score-based modeling [SSK*21] and the recent flow matching [LCBH*23], model the distribution $p(\mathbf{x})$ implicitly by modeling the *score function* $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ instead. This is done by training a score estimating neural network $F : \mathbb{R}^d \to \mathbb{R}^d$ that allows sampling from the distribution, i.e. $\mathbf{x} \sim P(\mathbf{X})$, through iterative application. The network is trained to follow a path from corrupted to clean states, where the corruption is usually Gaussian noise. The trained network can be used in different ways:

- **Sampling** $\mathbf{x} \sim P(\mathbf{X})$: Sample Gaussian noise $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$ and iteratively apply F.
- **Evaluating** $p(\mathbf{x})$: We can obtain a value that correlates with $p(\mathbf{x})$ via *score distillation sampling* [PJBM22], allowing the model to be used as data prior for 3D or 4D reconstruction tasks.
- **Editing and In-painting**: We can conditionally re-sample part of the representation via Repaint [LDR*22], allowing to replace certain parts of the representation with other content.

3. State-of-the-Art Methods

Our discussion starts with reviewing state-of-the-art methods for non-rigid reconstruction and view synthesis in Sec 3.1. Afterwards, we categorize the methods based on the added functionality incorporated into the non-rigid reconstruction. In Sec. 3.2, we provide an overview of methods that further decompose the scene into its static and dynamic parts. Sec. 3.3 discusses editable methods that provide control over the reconstruction. Finally, we turn to generalizable and generative methods in Sec. 3.4.

3.1. 3D Non-Rigid Reconstruction and View Synthesis

In this section, we discuss recent scene representations and deformation models for general non-rigid reconstruction and view synthesis. These are evaluated on the quality of the reconstructed geometry and renderings from observed and novel viewpoints, which should look as realistic as possible. One essential criterion for capturing scene dynamics is the ability to capture large, unconstrained motion while providing temporal correspondences across frames (see Sec. 2.3.2 for the trade-off between the two). Another important aspect for evaluation is the efficiency of the method in terms of memory consumption, training time, and rendering speed. In Sec. 3.1.1, we first look at recent advances in neural scene representations, the introduction of which significantly advanced the state-of-the-art for high-fidelity view synthesis. Then, we discuss the hybrid neural representations in Sec. 3.1.2, which significantly improve the efficiency of the neural scene representations and have become the de facto standard. Lastly, Sec. 3.1.3 looks at the non-neural scene representations, which have traditionally been used by classical methods and are often capable of real-time performance. We provide an overview of selected methods in Tab. 1, comparing across the introduced representations.

3.1.1. Neural Scene Representations

After the introduction of neural fields in 3D reconstruction [MON*19, PFS*19, CZ19] and subsequently, in neural rendering [MST*20, TFT*20], the use of neural fields as the underlying scene representation has also gained popularity for non-rigid 3D reconstruction and view synthesis problems. The significant increase in neural field-based methods can be attributed to its simple, flexible architecture and state-of-the-art performance. Most works use a deformable neural scene representation combined with a differentiable rendering technique (e.g., volume rendering). The flexible formulation and continuous nature allow for optimization with various types of inputs, ranging from monocular to multi-view videos to per-point space-time trajectories. In the following, we discuss state-of-the-art methods structured by how they represent the deformation (see Sec. 2.3.2 for background information). More specifically, we categorize methods into a.) *Space-Time Neural Fields*, the class of 4D extensions of neural fields, b.) *Deformable Neural Fields*, variants of neural fields that are equipped with a deformation operation, and c.) *Velocity Fields*, neural fields combined with motion-describing vector fields.

Space-Time Neural Fields. The majority of space-time neural fields build on the NeRF [MST*20] formulation (see Sec. 2.2) and provide time as an additional input to the neural field. Given the

Method	S	SR	DR	STM	Sp
Neural Scene Representations					
DyNeRF [LSZ*22]			—		
PREF [SGP*22]					
HyperNeRF [PSH*21]					
Unbiased4D [JHS*23]					
4DRegSDF [CCP*23]					
NDR [CFF*22]					
DySurf [CLF*23]					
DE-NeRF [MPCVG23]					
FSDNeRF [WMJL23]					
RFNet-4D [VNH*22]					
Hybrid Neural Scene Representations					
Hexplane [CJ23]					
K-Planes [FKMW*23]			—		
Tensor4D [SZT*23]			—		
Park et al. [PSJ*23]			—		
Wang et al. [WZL*22]			—		
HyperReel [AHR*23]					
Guo et al. [GSD*23]					
MixVoxels [WTL*23]			—		
DeVRF [LCM*22]					
SceNeRFlow [TGZ*24]					
ReRF [WHH*23a]			—		
StreamRF [LSW*22]			—		
TiNeuVox [FYW*22]					
DynIBaR [LWC*23a]					
4K4D [XPL*24]			—		
Im4D [LPX*23]			—		
PAC-NeRF [LQC*22]					
Non-Neural Scene Representations					
NR-SLAM [RMT23]					
OccFusion [LZYX22]					
Temp-MPI [XC22]					
Prokudin et al. [PMR*23]					
Luiten et al. [LKLR24]					
Yang et al. [YGZ*24]					
Wu et al. [WYF*24]					
Yang et al. [YYP*24]					
NPGs [DWY*24]					
3DGStream [SJL*24]					

Table 1: ***Selected Non-Rigid Reconstruction and View Synthesis methods. Supervision (S)****: Video, Multi-view Video, Depth, Event, Mask, Optical Flow, Pseudo Depth;* ***Scene Representation (SR)****: Density, Occupancy, SDF, Radiance;* ***Deformation Representation (DR)****: Position, Displacement, Rigid Transform, Affine Transform, Velocity Field, Motion Trajectory, DQB;* ***Spatio-Temporal Modeling (STM)****: Canonical, Deformation Basis, Canonical First Frame, Individual, Forward, Backward, Bidirectional;* ***Speed (Sp)****: Offline, Online, Real-time Rendering, Real-time (more than 20 FPS).* ***Data structures*** *used for SR and DR are (appears first): MLP, Transformer, Voxel Grid, Octree, Tensor Factorization, Triplane, MPI, Images, Point Cloud, 3D Gaussians, EDG, Physical Simulation.*

simplicity of this approach, space-time approaches have also been investigated with other parameterizations, e.g. with hybrid methods, as we will see in Sec. 3.1.2. While early works use monocular videos as input [LNSW21, XHKK21, DZY*21, GSKH21] and often rely on pre-computed optical flow and depth maps from off-the-shelf methods, You et al. [YH23] extend [GSKH21] to circumvent the need for pre-computed input by employing surface consistency and patch-based multi-view constraints. This leads to improvements, especially when off-the-shelf predictions are inaccurate. DCTNeRF [WELG21] is a space-time neural field that also predicts coefficients for discrete cosine transforms (DCT). This way, they model deformations to a canonical field for the entire trajectory, leading to sharper reconstructions in dynamic parts. In Neural Scene Chronology [LWC*23b], the space-time neural field is optimized from time-stamped internet photo collections of landmarks. While the geometry is assumed to be static, the neural field is optimized with per-image illumination embeddings, and learned step functions are employed for temporally varying scene changes to enable spatial, illumination, and time view synthesis.

Space-time neural fields have also been used to perform 4D reconstruction from multi-view video captures (see also Tab. 1). DyNeRF [LSZ*22] optimizes a space-time neural radiance field in a coarse-to-fine manner with importance sampling to speed up optimization. They, among others, show the compression property of neural fields: a 28-camera multi-view video clip can be compressed from over 1GB of storage to 28MB in MLP weights. PREF [SGP*22] optimizes a space-time neural field along with a time-embedding predictor and a time-embedding-conditioned motion field. They train the space-time neural field directly and with the motion field's predictions, resulting in time- and space-interpolating view synthesis and correspondences over time. Zhang et. al. [ZLC22] train a space-time neural radiance field from multi-scopic capture recordings. Optical flow-based predictions [JSJ*18] are used to also supervise between time frames, and the camera parameters are optimized jointly next to the scene representation.

Deformable Neural Fields. Another popular class of approaches combines a canonical static neural field and a deformation field (also called a "ray bending field") to perform 4D non-rigid reconstruction [PCPM21, PSB*21, TTG*21]. A key property of this approach is that any arbitrary 3D point can be evaluated for its deformation, rendering this approach very flexible and this idea has hence been explored also in the context of e.g. particle-based methods (see Sec. 3.1.2) as well as editing approaches discussed in Sec. 3.3. Recently, in HyperNeRF [PSH*21], Park *et al.* extends Nerfies [PSB*21] to model the canonical neural field in a higher-dimensional hyperspace to better handle topological variations. In DyLiN [YJM*23], this representation is distilled into a neural deformable light field, leading to a significant inference speed-up.

Another line of work focuses on enabling mesh extractions after optimization, directly enabling multiple applications such as the usage in traditional graphics software (see Sec. 2.2 for an overview). To this end, the density-based representation in the canonical neural field is exchanged with the NeuS [WLL*21] signed distance formulation in Unbiased4D [JHS*23] to enable mesh extractions via marching cubes. Notably, their approach requires a proxy geometry as input to stabilize the deformation field optimization.

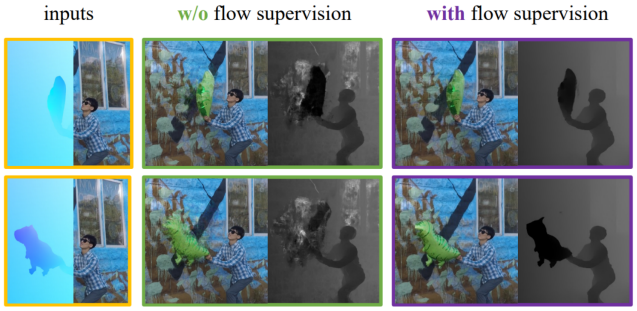

Figure 10: ***Optical Flow Supervision for Monocular Deformable NeRF.*** *Novel view images and depth maps for a monocular video with rapid object motion. Results from FSDNeRF [WMJL23].*

4DRegSDF [CCP*23] further proposes additional surface regularization strategies based on local rigidity leading to improved surface reconstructions. NDR [CFF*22] uses monocular RBG-D input and the canonical hyperspace is combined with a similar SDF representation, and the deformations are modeled via bijective mappings implemented as invertible neural networks.

Due to its efficient and simple design, the deformable field representation has recently been used for other downstream tasks or data modalities including 4D reconstruction from sinograms [RKA*21], point cloud interpolation [LCQ*21, ZQZ*22, ZWL*23], and deformable tissue reconstruction from a stereo video [ZCL*23, WLFD22]. Multi-view video input has also been investigated. In DySurf [CLF*23], an $SE(3)$ deformation field is combined with a canonical field in hyperspace, and they employ the SDF formulation from [YGKL21] to enable marching cube-based mesh extraction next to view synthesis. To model fast-deforming scenes, DE-NeRF [MPCVG23] combines sparse RGB input with a dense asynchronous capture from an event camera (see Sec. 2.1.2 for context), and the radiance field is optimized jointly with the unknown event camera poses. In EvDNeRF [BMC*24], the batching of events is performed progressively, leading to an increase in time resolution for each update. This way, fine temporal details can be reconstructed. TöRF [ALG*21] directly uses phasor images obtained from raw ToF sensor measurements to regularize reconstruction from a monocular video, reducing the number of required views and achieving sharp details. Using raw sensor measurements helps the methods bypass issues with processed depth maps, such as limited depth range and difficulty with low-reflectance objects and regions affected by multi-path interference.

Velocity Fields. Occupancy Flow [NMOG19] introduced combining a static canonical occupancy field with a neural time-dependent velocity field (see Sec. 2.2.3 for background information) for 4D reconstruction from point cloud input. Forward and backward transformations are solved using neural ODEs [CRBD18], thereby naturally providing dense correspondences. CaSPR [RBZ*20] extends this representation with a spatio-temporally-canonicalized object space and solves the neural ODE in a latent spatio-temporal space. In RFNet-4D [VNH*22], the objective is split into reconstruction and correspondence optimization with a joint point cloud encoder leading to improved results. Chu et. al. [CLZ*22] propose a combination of a space-time neural radiance field and a velocity field enabling dynamic fluid reconstruction from monocular input. They optimize the velocity field with a reconstruction loss considering adjacent frames and solve the resulting ODE with the Euler method. In FSDNeRF [WMJL23], Wang et. al. show that the inverse Jacobian of an ordinary, non-invertible deformable radiance field can be used to construct a local velocity field, which they solve with the fourth-order Runge-Kutta. This enables direct flow supervision leading to improved reconstructions of dynamic parts of the scene (see Fig. 10).

3.1.2. Hybrid Neural Scene Representations

This section focuses on hybrid representations for non-rigid 3D reconstruction that recently became popular. They combine neural components with explicit structures such as voxel grids, feature planes, images, and particles, bringing advantages of better feature localization, reduced training and inference time, and reduced memory requirements.

Planar Factorization. HexPlane [CJ23] extends the factorized tensor representation—introduced in TensoRF [CXG*22] for static scenes—to dynamic scenes, representing them by six planes of learned features and achieves 100× speedup in rendering non-rigid 4D volumes compared to previous methods with a single MLP. It computes features for points of a non-rigid scene by fusing the pre-plane feature vectors (using various feature fusion policies). A concurrent and related approach, K-Planes [FKMW*23] decomposes a D-dimensional space into $\binom{D}{2}$ feature planes (e.g. dynamic 4D volumes are factorized into six planes: three for space and three for spatio-temporal variations) and compresses the full 4D grid of data by three orders of magnitude, without requiring any GPU kernels. Both HexPlanes and K-Planes decode features with a small MLP to regress the output scene colors in the hybrid verison and utilize spherical harmonics in the explicit version. Inspired by priors from the non-rigid-structure-from-motion literature [STG*20, PPB21], BaLi-RF [RSAH23] represents a scene as a low-rank tensor decomposition similar to TensoRF [CXG*22] and the time dimension as a finite linear combination of neural time-basis functions which they empirically observe to be more expressive than DCT, Fourier, and Bernstein basis functions. They show that this representation of scenes as a composition of band-limited, high-dimensional signals can better reconstruct long-range dynamics. Peng *et al.* [PYS*23] predict per-scene MLP maps by a 2D CNN decoder. Instead of directly storing features on the six spatio-temporal planes, Guo *et al.* [GPY*23] introduces a dynamic codebook to store such features, with the planes just storing the index of features in the codebook, thus achieving further compression.

Other methods use factorized 4D spatio-temporal tensors to represent scenes in more specialized settings. Tensor4D [SZT*23] assumes four sparse input RGB views of a dynamic scene (e.g. from the corners of a holographic display); motion and detail changes are learned from coarse to fine with the hierarchical tri-projection decomposition policy, which results in nine decomposed planes. HyperReel [AHR*23] adopts the triplane representation for novel view synthesis from multi-view camera rigs (small baselines) in the context of 6-DoF videos. It combines a ray-conditioned sampling network with a keyframe-based dynamic volume represen-

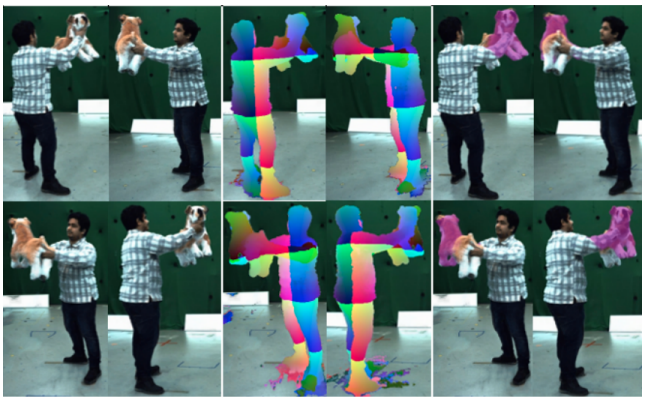

Figure 11: ***Results of SceNeRFlow [TGZ*24] for an example multi-view video sequence.*** *(Left column): Two rendered RGB novel views for the same time instant; (middle column): Their correspondence visualizations; (right column): Their temporally consistent recoloring. Image source: [TGZ*24] ©2024 IEEE.*

tation and allows immersive AR/VR experiences. Feng and colleagues [FAR*23] propose a new method for synthesizing novel views of a motion-magnified non-rigid scene with subtle deformations; it is a combination of non-rigid NeRF and the Eulerian principle for motion magnification. 4D volumes factorized into explicit static and dynamic fields were also used for non-rigid NeRF reconstruction of endoscopic scenes [YWW*23], resulting in a compact memory footprint and significantly accelerated optimization.

Voxel Grids. Voxel grids have the same goal as planar factorization approaches, namely accelerated training and inference. NeuS2 [WHH*23b] is a general multi-view 3D reconstruction approach on individual frames which is accelerated by InstantNGP [MESK22], with incremental learning for consecutive frames. Similarly, OD-NeRF [YLL23b] utilizes the occupancy grid estimated from the previous frames by tracking correspondences between the neighboring frames. It achieves 6 fps during training and on-the-fly rendering speed of dynamic scenes. Residual radiance fields (ReRF) [WHH*23a] is a compact representation for free-viewpoint rendering of long dynamic scenes relying on compact motion- and residual feature grids to exploit inter-frame feature similarities. Next, Li *et al.*'s NeRF streaming approach [LSW*22] uses an explicit grid with incremental learning and difference-based NeRF compression. This allows on-the-fly handling of new observations.

A few other recent methods using voxel grids push the frontiers of novel view synthesis and reconstruction through various alternative formulations (problem settings) and further extensions. Božič and colleagues [BPZ*21] introduced an encoder that works on a voxel grid filled with SDF values and outputs a deformation graph with local SDF fields. Park *et al.* [PSJ*23] propose a non-rigid NeRF method which is based on temporal interpolation of feature vectors, either on neural networks or voxel (hash) grids. Their method is remarkable due to its simplicity, as it avoids deformation or flow estimation modules. The MixVoxels approach [WTL*23] represents a dynamic scene as a mixture of static and dynamic voxels and processes them with separate networks, which results in fast training and rendering. The method can efficiently perform multiple queries simultaneously, thereby improving the rendering quality of dynamic objects. Another advantage of MixVoxels is that static voxels can be processed with a lightweight model. Fourier PlenOctrees [WZL*22] combine generalized NeRFs, plen-octrees, volumetric fusion, and Fourier transforms for real-time rendering of general dynamic scenes. It maintains an implicit network to model the Fourier coefficients of time-varying density and color attributes. ENeRF [LPX*22] builds a cascade cost volume to predict a coarse scene geometry, which allows for sampling of fewer points near the surface (thereby achieving substantial acceleration). MoNeRF [KGC*24] performs non-rigid NeRF reconstruction from multi-view data; only a single RGB frame is provided to the method per each timestamp, which the authors introduce as *monocularized*. The deformation module of MoNeRF decouples the processing of spatial and temporal information for the acceleration of training and inference. The authors emphasize monocularization as a way to accelerate the training. Moreover, MoNeRF can also reconstruct multi-view sequences.

TiNeuVox [FYW*22] uses a very light MLP for the deformation network to model coarse trajectories and achieve fast training while enhancing the temporal information in the radiance network through multi-resolution temporal embeddings. DeVRF [LCM*22] uses a hybrid, voxel-based 4D deformation field to model motion. To relax optimization, a canonical representation of the static scene is first learned through multi-view images, which is used as the basis for dynamic scene reconstruction from a few multi-view videos. SceNeRFlow [TGZ*24] addresses the joint novel view synthesis and long-term correspondence estimation for general scenes from multi-view RGB videos. Even though SceNeRFlow achieves improved results compared to previous techniques, the new setting with correspondences remains an open challenge in the field, see Fig. 11 for the visualizations. Instead of modeling the backward deformation from time steps to a canonical field (see Fig. 8 for its downside), ForwardFlowDNeRF [GSD*23] models the forward deformation and uses splatting and inpainting modules to deal with many-to-one and one-to-many mappings. The explicit voxel grid representation allows efficiently registering the canonical frame to the live frame, leading to smoother object motions. DynamicSurf [MA24] utilizes a canonical feature grid, trained in a coarse-to-fine manner, and a topology-aware deformation field to speed up neural surface reconstruction.

Image-based. Recently, a few works aim to combine image-based rendering techniques with neural fields. Inspired by static hybrid methods [LPL*22, SESM22, WWG*21], these approaches propose to use image-based features obtained via backprojection in combination with the neural representation enabling reconstructions of non-object-centric scenes with long-range trajectories. Based on PIFu [SHN*19]—one of the earliest approaches utilizing image-based features for single-view or multi-view static reconstruction—Function4D [YZG*21] extends the representation to real-time dynamic scene reconstruction from a few synchronized multi-view RGB-D inputs. Using the SDF-based surface reconstructed by a DynamciFusion-based [NFS15] pipeline, a transformer-based image-feature aggregation scheme from multi-view SDF projections is proposed which learns a more detailed and complete occupancy and appearance representation. In DynIBaR [LWC*23a],

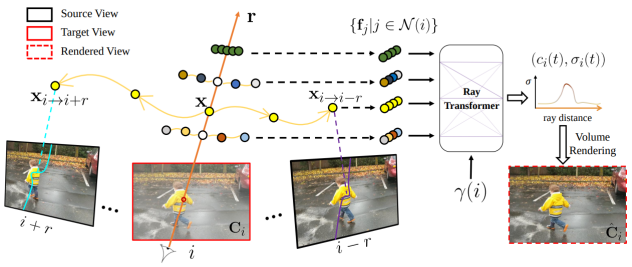

Figure 12: ***Image-based rendering via motion-adjusted feature aggregation.*** *DynIBaR [LWC*23a] uses motion trajectories to aggregate features from temporally neighboring frames, improving time consistency and increasing the range of novel view synthesis. Image source: [LWC*23a].*

learned trajectory basis functions (initialized with DCT coefficients) are used to aggregate features from nearby views efficiently leading to improved space-time view synthesis results particularly for long-range captures with complex camera motion (see Fig. 12). With a focus on real-time 4D rendering, Im4D [LPX*23] combines a grid-based representation for a space-time neural field with an image-based appearance prediction module utilizing nearby views. While the grid-based backbone enables fast training and inference, the image-based rendering module allows for reconstructing fine texture details. The very recently proposed method 4K4D [XPL*24] uses a related hybrid representation with a similar focus on real-time rendering. A 4D point cloud representation is combined with 4D feature planes to represent the time-varying geometry. A piece-wise constant IBR term and a continuous spherical harmonics module are combined to achieve high-quality appearance reconstruction with 80 FPS at 4k resolution.

Particle-based. Point-NeRF [XXP*22] inspired several reconstruction approaches, which adopt the idea of using point clouds—or, generally, particles—as geometric proxies in neural rendering and reconstruction of non-rigid scenes. For instance, DAP-NeRF [LL23] quantizes the dynamic motion field as a collection of appearance particles, each of which carries the semantically meaningful visual information of a small dynamic scene element.

Some particle-based methods target long-term novel view synthesis scenarios (with the input monocular videos following dynamic objects) [PK24] or reconstruction and editing of dynamically vibrating scenes [PPGT*23], while other ones estimate physical parameters of non-rigid objects (such as stiffness and volume preservation coefficients) from multi-view videos [CTS*22, LQC*22] or infer fluid dynamics from sequential image observations [GDWY22]. Point-DynRF [PK24] analyses the entire long sequence and distinguishes the background from the moving objects; its joint optimization scheme alleviates degenerate solutions. ModalNeRF [PPGT*23] uses a particle-based deformation field to bend the observation-space rays into the canonical space. The trained deformation field can be used for modal analysis and synthesis (e.g. to magnify the observed motions and increase oscillation amplitudes). The VEOs approach [CTS*22] estimates a point cloud of an object non-rigidly deforming under the influence of external forces observed in a multi-view video. The material parameters of the object—estimated with a differentiable particle-based simulator—ensure that the reconstructions match the observations. They also can be used for finding new non-rigid states (in response to new force fields and collision constraints), which can be re-rendered as 2D images, thanks to integrating the new simulations with NeRF. Another approach, PAC-NeRF [LQC*22], uses a hybrid Eulerian-Lagrangian NeRF representation with Lagrangian particles and applies the conservation laws of continuum mechanics for the object states to be physically plausible. Finally, NeuroFluid [GDWY22] is a recent intuitive physics approach to inferring the 3D dynamics of a fluid from its 2D surface observations. At its core is the particle-driven neural renderer which helps to optimize the particle transition model and minimize the discrepancy between the rendered and observed images.

3.1.3. Non-Neural Scene Representations

Before the recent popularity of neural methods for non-rigid scenes (see Secs. 3.1.1 and 3.1.2), the main approaches for monocular non-rigid reconstruction for general scenes were Shape-from-Template (SfT) [SGTS19, CPPFJ*21, FJPCP*21, KTE*22, SWK24] and Non-Rigid Structure-from-Motion (NRSfM) [GJST20, STG*20, KVG22, WLPL22, ZYMY22, GB22, PPB21, ZDY*21]. SfT works with a pre-acquired template, deforming it to fit the observations over time. NRSfM does not require a template but relies on the information provided by deformation cues in the form of point tracks across 2D images [GJST20, STG*20]. In comparison to these approaches, differentiable neural rendering-based methods are more flexible and achieve reconstruction with higher fidelity. Because of the significant training time required by the neural approaches, real-time online non-rigid reconstruction methods still predominantly use classical scene representations and data structures like meshes, voxels, and point clouds. Moreover, point-based representations have been living through a renaissance recently, as they are often very efficient and can profit from recent advances in optimization algorithms. We next categorize and discuss methods that use these classical representations exclusively; non-rigid SLAM and dense RGB-D reconstruction for the online setting, and point-based reconstruction and multiplane images-based view synthesis for the offline setting.

Monocular Non-Rigid SLAM. Non-Rigid Simultaneous Localization and Mapping (SLAM) is a challenging problem, especially in the underconstrained setting of monocular RGB observations [SGM*19]. Most SLAM methods operating in dynamic environments only segment the scene to use the static region for mapping and subsequent camera tracking [BFCN18, YLL*18, BBLT18, NDS*20]. However, in specific environments, discarding dynamic regions of the scene is not feasible since it can either consist of most of the observed environment, as in incorporeal scenes, or the reconstruction of deforming objects is essential to understand and interact with the scene, as in AR/VR applications.

Lamarca *et al.* [LM18] provide the first real-time camera tracking method that operates in deformable scenes, based on an SfT method working with a pre-acquired template. It is extended by DefSLAM [LPBM20], which proposes the first SLAM approach to build and extend a deformable map. They use an isometric NRSfM

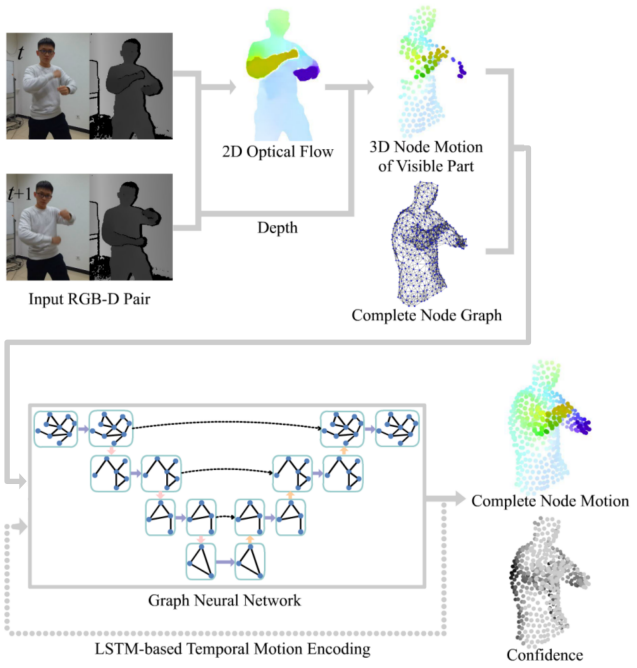

Figure 13: ***Occlusion-aware Reconstruction.*** *OcclusionFusion [LZYX22] pretrains a GNN on DeformingThings4D [LTT*21] to estimate the motion of occluded regions. Image source: [LZYX22].*

approach [PPB17] to acquire surface templates at keyframe rate, while the SfT method of Lamarca *et al.* is employed to track those templates at frame rate. SD-DefSLAM [GRLM*21] adds photometric tracking to perform data association on top of DefSLAM, improving robustness and accuracy in varying illumination and strongly deforming environments. Two significant limitations of these approaches are the use of a triangular mesh, which cannot model discontinuous changes to the surface, and the isometric deformation assumption, which limits the use cases significantly. The recently proposed NR-SLAM [RMT23] solves these shortcomings by proposing a time-varying point cloud representation for the deformable map, coupled with a Dynamic Deformation Graph and a Visco-Elastic Deformation model for physically-inspired regularization, allowing more expressive deformations than the isometric case. However, it requires the camera-over-deformation assumption, which states that most image change comes from camera motion. This assumption is used to initialize and extend the map with rigid Structure-from-Motion and results in a limitation of use cases. A common setting for these methods is incorporeal scenarios such as endoscopic videos, where most of the results are demonstrated [ASF*23, MSY10]. A few methods [LPBM20, GRLM*21] also show results on small-scale non-rigidly deforming objects such as a kerchief.

Online Dense RGB-D Non-Rigid Reconstruction. These object-level methods focus on accurate surface reconstruction and deformation tracking by fusing RGB-D data, with camera pose estimates emerging from such tracking. The seminal DynamicFusion [NFS15] and its follow-up VolumeDeform [IZN*16] introduced the paradigm of using an Embedded Deformation Graph [SSP07] (EDG) with the extracted mesh from a TSDF voxel grid or volumetric control lattices for non-rigid deformation modeling. However, due to fixed connectivity, meshes cannot handle topology changes well when the deformation is discontinuous over 3D space. SobolevFusion [SBI18] tackles this issue by removing the deformation graph and directly optimizing for the deformation field, i.e. scene flow, which registers the live TSDF volume with the canonical one. It performs gradient updates in Sobolev space instead of L^2 space [SBCI17] first, which favors coarse-scale changes over fine-scale ones in the beginning, making the optimization less susceptible to local minima. AcceleratedFusion [SBI20] further incorporates a fast numerical optimization scheme based on Nestrov accelerated gradient descent [Nes83]. Li *et al.* [LG20], on the other hand, handle topology changes by allowing non-manifold volumetric grids for both TSDF and EDG, where node connectivity can be updated through cell splitting and replication.

SurfelWarp [GT18] replaces the performance- and memory-intensive TSDF voxel grid with surfels. It leads to a 50% decrease in frame processing time and a 98% decrease in memory consumption compared to DynamicFusion, albeit at a slight cost of quality. Explicit surfel representation also supports handling erroneous observations and topology change issues more efficiently. In Chang *et al.* [CB22] and MonoSTAR [CRG*23], a more sophisticated topology-change handling approach, based on the historical distance between the nodes of the deformation graph, and additionally on object rigidity from semantic information in the case of [CRG*23], results in the surfel-based deformation graph being split up to represent separate objects (see Sec. 3.2.2).

In recent years, data-driven learned modules have also been introduced for improved non-rigid reconstruction. DeepDeform [BZTN20] and Bozic *et al.* [BPZ*20] learn 2D feature correspondences to guide the reconstruction in case of little geometric variation, although the exhaustive correspondence computation hinders real-time performance (see Sec.3.4.2 - Correspondence priors). OcclusionFusion [LZYX22] learns a data-driven motion prior to estimate complete object motion, including occluded regions (see Sec.3.4.2 - Scene flow priors). Such motion estimation and 2D optical flow constraints allow it to achieve state-of-the-art results for real-time non-rigid reconstruction. DeepDeform provides a dataset for benchmarking results, though most methods also conduct experiments on self-acquired sequences. Moreover, other than SobolevFusion, none of the methods reconstruct appearance.

Efficient Point Representations for Offline Reconstruction. The main drawback of neural scene representations (see Sec. 3.1.1) is the prohibitively long training and rendering times, as thousands of MLP evaluations are required to render the scene at high resolutions. While hybrid approaches (see Sec. 3.1.2) have drastically reduced training and rendering times, no neural approach has produced scalable reconstructions of dynamic scenes that produce high-quality novel views while training quickly and rendering in real-time. Recently, such results have been achieved by explicit point-based representations, which are fast to process, scalable, and allow adding spatial deformation constraints. While still requiring

offline training, these approaches achieve state-of-the-art results regarding reconstruction and view synthesis quality.

Prokudin *et al.* [PMR*23] introduce a compact and efficient point-based canonical surface representation with a flexible and accurate neural deformation field for each frame to model fine-grained surface deformations from ground truth meshes, outperforming alternative neural scene representations in terms of reconstructed quality, training time, runtime and model size. The method can capture intricate non-rigid deformations, e.g. of a skirt, where traditional linear blend skinning-based models fail by incorporating regularization techniques such as as-isometric-as-possible [KMP07, HAWG08] and additionally supervising for highly dynamic and complex deformations with keypoint correspondences. They also show the results of their surface modeling method for dynamic reconstruction from multi-view videos. Zhang *et al.* [ZBRH22] proposes a dynamic view synthesis method from multi-view videos by differentiably splatting point primitives, with learnable spherical harmonics to represent color. The method requires an object mask to generate the point cloud and learns the model per frame.

Luiten *et al.* [LKLR24] is one of the first extensions of 3D Gaussian splatting [KKLD23] for general dynamic scenes. They parameterize the Gaussians with 6-DOF rotation and translation, each equipped with a color and opacity, and impose multiple physical regularizers (e.g. local rigidity and local isometry). Without needing an MLP to regress the radiance, they can render novel views of dynamic scenes with state-of-the-art quality and efficiency from multi-view videos, see Fig. 14. Dense, consistent trajectories for the Gaussians also emerge naturally from the dynamic view synthesis. [LKLR24] allows the Gaussians only to move and rotate over time. Concurrent works [YGZ*24, WYF*24] model the deformations more faithfully by learning static Gaussians in the canonical space, and a deformation field, which allows the scale of the Gaussian to change, along with the position and rotation, from monocular observations. NPGs [DWY*24] utilize 3D Gaussians initialized on a coarse point model to tackle the issue of high-quality novel view synthesis (from camera poses that are significantly different from the training views) for casual monocular video captures. A coarse point model—constrained by a low-rank deformation basis—is first obtained for the observed object at each timestep, which is then used as an anchor for 3D Gaussians, providing regularization for the reconstruction of the sparsely observed dynamic object.

Instead of canonical modeling, [YYP*24] directly extends 3D Gaussians into the time domain as 4D Gaussians, each with a 4D rotation, 4D scale, and 4D spherical harmonics to represent the view-dependent color that can change over time, achieving a rendering speed of 114 FPS. SpacetimeGaussians [LCLX24] also extend 3D Gaussians with an extra 1D Gaussian to represent the temporal opacity at any time. Instead of spherical harmonics, they store feature embeddings on the Gaussians for appearance, which are splatted and then decoded by an MLP to achieve high-fidelity renderings of up to 8K resolution at 60 FPS. While all the methods discussed in this subsection so far require offline training, 3DGStream [SJL*24] achieves high-speed, compact, and on-the-fly per-frame reconstruction within 12 seconds and with up to 200 FPS rendering speed by utilizing a multi-resolution hash grid as a cache for the transformations of the 3D Gaussians.

Figure 14: ***Dynamic View Synthesis by Tracking 3D Gaussians.*** *(Left:) Gaussian centers; (right:) Rendered image and depth from unseen view with 3D trajectories (including occlusions). Luiten et al. [LKLR24] parameterizes each scene with 200 to 300k Gaussians, tracked across frames with an accuracy 10x better than previous state-of-the-art while also rendering at 850 FPS. Image source: [LKLR24] ©2024 IEEE.*

Multi-plane Images for Novel View Synthesis. Multi-plane Image (MPI) representations have been successfully applied for view synthesis from multiple posed images [FBD*19], a narrow baseline stereo pair [ZTF*18, STB*19], a single image [TS20] or semantic maps [HTLH20]. They can also be augmented to model view-dependent appearance [WPYS21] and time-dependent variations [LXDS20]. Several recent works have extended the MPI-based representations for modeling dynamic video [LXL*21, ZW22, XC22, MLLS23]. Using a stereo video, Lin et al. [LXL*21] decompose the dynamic scene into a static and a dynamic MPI representation and blend them with a predicted 3D mask volume. The decomposition helps to enable temporally stable view extrapolation. Instead of explicit dynamic/static decomposition, Temporal-MPI [XC22] proposes to learn a low- and high-frequency temporal basis to model the dynamic scene. The plane representation in MPI can also be extended to a sphere to support a larger field of view. For example, MatryODShka [ALG*20] trains multi-sphere image (MSI) representations for real-time view synthesis of dynamic video from 360° omnidirectional stereo video.

3.2. Decompositional Scene Analysis

Real-world complex scenes usually comprise static regions and dynamic objects undergoing different motions. Decomposing the scene into multiple parts based on their motion is a natural choice that enables many downstream applications, including interaction analysis, scene editing, and future prediction. It is a challenging task, especially when no direct decomposition supervision or category-level prior is available. Scene-level methods discussed in Sec. 3.1 mostly capture the scene using a single geometry and deformation representation. In this section, we focus on decompositional approaches (see Sec. 2.2.4 for types of decompositions) applicable to non-rigid dynamic scenes, excluding methods related to static decomposition [NG21, YZX*21, YGW21, YSL*22]

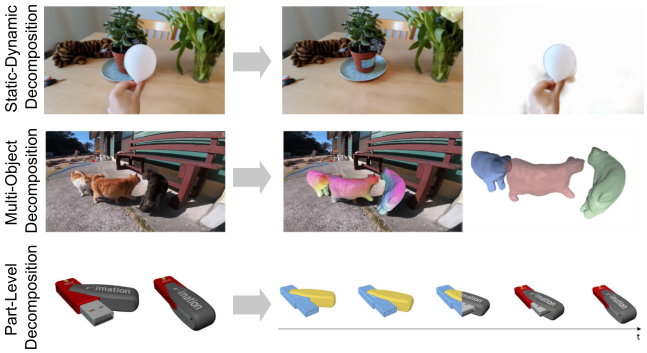

Figure 15: ***Decompositional Scene Analysis****. We show representative methods for static-dynamic decomposition [WZT*22], multi-object decomposition [WDSY23], and part-level decomposition [LMAS23], respectively. The left illustrates the input whereas the right shows the decomposition results.*

Method	S	SR	DR	STM	T/Sp
Static/Dynamic Decomposition					
NeRF-DS [YLL23a]					
PPR [YYZ*23]					
RoDynRF [LGM*23]					
Chen et al. [CT22]					
D^{2}NeRF [WZT*22]					
NeuralDiff [TLV21]					
NeRFPlayer [SCL*23]					
Multi-Object Decomposition					
RPD [WDSY23]					
Total-Recon [SYD*23]					
FactoredNeRF [WM23]					
Driess et al. [DHL*23]					
SUDS [TZFR23]					
SAFF [LLM*23]					
Bujanca et al. [BLL19]					
Li et al. [LZNH20]					
Bujanca et al. [BLL22]					
STAR-no-prior [CB22]					
MonoSTAR [CRG*23]					
Part-level Decomposition					
Art.Fusion [LZG18]					
PARIS [LMAS23]					
MovingParts [YZH*24]					
WIM [NIT*22]					

Table 2: ***Selected Decompositional Scene Analysis methods. Supervision (S)****: Video, Multi-view Video, Depth, LIDAR, Mask, Semantic Segmentation, Optical Flow, Image Features, Pseudo Depth, Bounding Box;* ***Scene Representation (SR)****: Density, Occupancy, SDF, Radiance, Semantics;* ***Deformation Representation (DR)****: Position, Displacement, Rigid Transform, Screw Transform, LBS, DQB;* ***Spatio-Temporal Modeling (STM)****: Canonical, Canonical First Frame, Individual, Forward, Backward, Bidirectional;* ***Tracking Type/Speed (T/Sp)****: Camera, Object Root Pose, Offline, Online, Real-time (everything above 20 FPS).* ***Data structures*** *used for SR and DR are (appears first): MLP, Voxel Grid, Octree, Point Cloud, Surfels, EDG, Joint, Forward Dynamics Model.*

or rigid-only decomposition [OMT*21, KGY*22]. Tab. 2 provides an overview. As illustrated in Fig. 15, we find three categories of granularity: static-vs-dynamic decomposition (Sec. 3.2.1) divides the scene into moving and stationary components, multi-object decomposition (Sec. 3.2.2) goes one step further by separating individual objects, and finally part-based decomposition (Sec. 3.2.3) operates at the level of parts. In the following sections, we will discuss these types of decomposition and all relevant works in detail.

3.2.1. Static-Dynamic Decomposition

Methods in this category decompose the scene into static and dynamic regions, yet do not consider further decomposition of different dynamic objects. We introduce these methods based on the supervision signals that are leveraged for their decomposition.

Mask-based. DynNeRF [GSKH21] is one of the first approaches to combine a dynamic scene representation with a static one through a learned blending weight field, leveraging a coarse binary motion segmentation mask to help decomposition. Similarly, NeRF-DS [YLL23a] leverages masks of moving objects to guide the training of the deformation field, providing a strong cue to decompose the moving foreground from the static background. It allows them to handle the specular effects of moving objects separately from the background. Assuming binary masks of articulated objects for decomposition, PPR [YYZ*23] proposes a first end-to-end framework for jointly optimizing dynamic 3D reconstruction and physical systems. While all aforementioned methods utilize decomposition to improve the reconstruction quality, RoDynRF [LGM*23] is the first method for jointly estimating static and dynamic radiance fields and the camera parameters. To facilitate robust camera parameter estimation based on the static regions, RoDynRF constructs a binary motion mask based on Mask R-CNN and optical flow estimation to enable static-dynamic decomposition. Another interesting application enabled by static-dynamic decomposition is creating endless looping 3D video [MLLS23].

Self-supervised. In contrast to the aforementioned methods that rely on motion masks for decomposition, another line of methods studies self-supervised decomposition. NSFF [LNSW21] is one of the first approaches that decompose the static and dynamic regions based on an unsupervised 3D blending weight field, with the intuition that static regions can be rendered with higher fidelity with static representation. Following NSFF, [CT22] demonstrates that proxy 2D optical flow can help decomposition in driving scenes by encouraging the static branch to model regions of low scene-flow magnitude. Without relying on proxy flow supervision, D^{2}NeRF [WZT*22] achieves decomposition with a novel skewed-entropy loss to regularize a position being either occupied by the static scene or a dynamic object, but not both. NeuralDiff [TLV21] encourages the foreground occupancy to be sparse and learns to decompose an egocentric video into static parts, dynamic objects, and the actor. A few hybrid methods, discussed in Sec. 3.1.2, also decompose the scene into static and dynamic parts, achieving computational efficiency by using a more lightweight model for the

static part [WTL*23, PSJ*23, LL23]. NeRFPlayer [SCL*23] further segments the newly observed areas of the scene, along with static and deforming parts, using a decomposition field regularized with a global parsimony loss. A streaming-friendly hybrid representation is further proposed based on the scene decomposition.

3.2.2. Multi-Object Decomposition

We now introduce methods that aim to decompose the scene into multiple dynamic objects, optionally along with a static part. In addition to offline methods that decompose the scene for motion decomposition or semantic understanding, many online dynamic reconstruction methods also fall into this category.

Mask-based Motion Decomposition. Building on the single dynamic object reconstruction method BANMo [YVN*22], both RPD [WDSY23] and Total-Recon [SYD*23] learn a decompositional scene representation for multiple objects, assuming that segmentation masks of objects are given. Both methods decompose object motion into root and articulated motion to model large deformations. Enabled by root pose decomposition and RGB-D data, Total-Recon shows the ability to synthesize views from extreme viewpoints while also enabling embodied view synthesis, i.e., point-of-view of moving actors and 3rd-person follow cameras (see Fig. 1, top row, on the right). RPD reconstructs the dynamic objects given only a monocular video and is the first method to estimate object root poses without any category-specific priors. BANMO further decomposes the object motion into articulated and non-rigid motion, from which both RPD and TotalRecon benefit. FactoredNeRF [WM23] assumes the root motion and segmentation masks of objects are provided at keyframes, enabling the decomposition of the scene into the background and multiple moving objects. This factored representation allows for object manipulations including removing an object or changing an object's trajectory. Instead of focusing on reconstruction, Driess *et al.* [DHL*23] proposes a novel approach to combine implicit object representations with graph-based neural dynamics models to enable future prediction, achieved by learning a compositional NeRF auto-encoder.

Data Prior-based. Several recent methods omit the requirement of segmentation masks or bounding by distilling 2D self-supervised features into the 3D space. While this feature distillation idea is firstly applied to static scenes [KMS22,TLLV22], SUDS [TZFR23] and SAFF [LLM*23] extend it to dynamic scene decomposition. SUDS distills 2D DIVO-ViT features into dynamic urban scenes, where unsupervised instance segmentation masks and 3D bounding boxes can be obtained by geometric clustering in the feature space. Similarly, SAFF distills 2D DINO-ViT semantic and attention saliency features, allowing for extracting segmentation masks based on their clustering.

Online Decomposition. Decomposing the rigid and non-rigid dynamic content during real-time reconstruction, e.g., in SLAM, is also common for scene understanding tasks in robotic vision. As shown in Tab. 2, these methods are distinguished from offline methods as they usually take unposed RGB-D sequences as input, and leverage surfels or voxel grids as the scene representation to enable real-time reconstruction and tracking. One line of work [BLL19, LZNH20, BLL22] exploits semantic instance segmentation models to decompose each observed RGB-D frame into several dynamic parts along with a static background, performing tracking and fusion on each segmented surface independently. In contrast to the aforementioned methods, STAR-no-prior [CB22] reverses the order of segmentation and reconstruction. By segmenting dynamic objects based on detecting topology changes, STAR-no-prior does not assume category-level priors. Mono-STAR [CRG*23] modifies the STAR-no-prior framework for the monocular RGB-D case. It uses the semantic class information obtained by segmentation to constrain the embedded deformation graph, modeling the rigidity usual for objects in that class.

3.2.3. Part-level Decomposition

A few methods take a further step to decompose objects into multiple parts. As part-level supervision is hard to obtain, most works focus on self-supervised part discovery. PARIS [LMAS23] decomposes articulated objects into a static and a dynamic NeRF in a self-supervised manner given two articulation states of a single object. ArticulatedFusion [LZG18] is an online method that clusters the nodes of the deformation graph with similar trajectories, thus regularizing the motion while segmenting the model into parts. Similarly, MovingParts [YZH*24] studies self-supervised parts discovery by a motion-based grouping mechanism that uses slot-based attention, assuming each group follows a rigid motion. Watch-it-move [NIT*22] is the first approach that learns re-poseable part decomposition from multi-view videos and foreground masks without any prior knowledge of the structure. Skeletonization of these part-level decomposition methods enables object re-posing, as discussed in Sec. 3.3.3.

3.3. Editability and Control

Many applications require editing properties of the scene, for example, to control the location, appearance, and pose of objects in the scene while keeping photo-realism. A key challenge when building such models is how to make models coherent under the considered edits: If an object is moved, the model should be able to fill in the dis-occluded background, and the object needs to be rendered in a new pose which potentially has not been seen during training. Deformable objects will non-rigidly deform when they are articulated. Editing the texture/appearance of scenes requires reasoning about lighting, shadows, occlusions, and temporal coherency.

In contrast to non-neural representations in Sec. 3.1.3 which can be edited by manipulating the explicit primitives, one limitation of the pure neural field representations of Sec. 3.1.1 is that they lack controllability. We refer to [YLW*21] for editing mesh-based representations and focus on editing of reconstructed neural scenes in this section. We consider methods that either allow editing of dynamic scene reconstructions or allow deformable editing of static scene reconstructions. An overview of selected methods is provided in Tab. 3.

We classify methods according to the type of control they achieve. Editability requires disentangling the scene representation and conditioning it on a set of control parameters, which can then be manipulated. Sec. 3.3.1 introduces methods that allow changing the scene's contents. In Sec. 3.3.2, we discuss methods that allow scene dynamics manipulation, while Sec. 3.3.3 focuses on methods that provide control over the object's pose.

Method	S/P	SR	DR	STM	CP
Scene Editing					
FactoredNeRF [WM23]					
Control-NERF [LGO*23]			—	—	
NeuVV [ZWL*22]					
Dyn-E [ZPS*23]					
DynVideo-E [LCW*24]					
Scene Dynamics Control					
CoNeRF [KYK*22]					
EditableNeRF [ZLX23]					
Wang et al. [WMDH22]					
VIRDO++ [WZFF22]					
Li et al. [LLS*22]					
ACID [SJC*22]					
Object Pose Control					
NeRF-Editing [YSL*22]					
NeuMesh [CJH*22]			—		
NeuralEditor [CLW23]			—		
NeuPhysics [QGL22]					
FAST-SNARF [CJS*23]					
TAVA [LTV*22]					
DANBO [SBR22]					
NPC [SBR23]					
HOSNeRF [LCY*23]					
CAMM [KKK*23]					
Uzolas et al. [UEK23]					
Transfer4D [MNH23]					
Liu et al. [LGW23]					
CageNeRF [PYL*22]					
Xu et al. [XH22]					
BANMo [YVN*22]					
MoDA [SCC*23]					
RAC [YWRR23]					

Table 3: ***Selected Editable and Controllable methods. Supervision/Priors (S/P)****: Video, Multiple Training Scenes, Multi-view, Multi-view Video, Depth, External Forces/Contact Point, Optical Flow, Mask, Pseudo Depth, Diffusion Prior;* ***Scene Representation (SR)****: Density, Occupancy, SDF, Radiance, Canonical Features;* ***Deformation Representation (DR)****: Position, Displacement, Screw Transform, LBS, DQB, Forward Dynamics;* ***Spatio-Temporal Modeling (STM)****: Canonical, Deformation Basis, Canonical First Frame, Individual, Forward, Backward, Bidirectional;* ***Control Parameters (CP)****: Neural Features, Appearance, Edited Image, Style Reference, Keypoints, Attributes, Action, Mesh, Cage, Driving Video.* ***Data structures*** *for SR and DR are (appears first): MLP, Voxel Grid, Octree, Point Cloud, Skeleton, Neural Bones, Cage, EDG.*

3.3.1. Scene Editing

This section discusses methods for scene editing, e.g. moving objects around, removing them, changing their appearance, or composing a new scene with different objects and backgrounds. We categorize and discuss these approaches next.

Compositional NeRFs. The original NeRF formulation does not directly allow editing scene content, as it encodes the full scene within a neural network. One way to gain control is to decompose the scene into its constituent objects by learning per-object models and composing them at rendering time, as described in Sec. 3.2. Once the per-object model is learned, individual objects can be moved, removed, duplicated, or their trajectories can be changed [WM23]. For scenes with many objects, this might be impractical as one NeRF per object needs to be learned, and controlling general properties of the scene, such as global illumination, is more challenging. Controlling a compositional dynamic scene is significantly more complicated than static scenes, as it requires tracking individual objects.

Decoupled scene representation and rendering. For neural representations, another way to edit the scene is to decompose the rendering network from the scene representation, as done in Neural Sparse Voxel Fields [LGZL*20] and Control-NERF [LGO*23]. These methods store neural features at a coarse voxelization of the 3D scene, and the radiance is predicted as a function of learnable scene-specific neural features. The rendering network is trained on multiple scenes to make it generalizable, which allows it to render a new scene after edits not seen during training. The scene can then be composed literally by merging different geometries represented as voxel-based neural features, and they can also be deformed directly with non-rigid geometric deformations. The main limitation of such methods is that storing high-dimensional features in a 3D grid is memory-consuming. While [LGZL*20, LGO*23] model static scenes, [TDD23] achieves similar editing capabilities for dynamic scenes as a by-product of generalizable modeling (see Sec. 3.4.2) by learning 3D point features for multiple dynamic scenes, which are aggregated from 2D image feature grids in an image-based rendering setup. Non-neural representations are already decoupled from the rendering pipeline. The approach proposed by Luiten *et al.* [LKLR24] (discussed in Sec. 3.1.3) allows adding or removing subsets of 3D Gaussians, or combining Gaussians from multiple scenes together. Edits to one frame automatically propagate to other frames over time due to their consistent Gaussian tracking.

Appearance Editing. A sub-class of editing methods leaves the scene's geometry intact and solely edits the appearance. NeuVV [ZWL*22] extends the PlenOctrees approach [YLT*21] introduced for static scenes to dynamic NeRFs, enabling real-time rendering. PlenOctrees bakes the color and density attributes from the underlying NeRF into an octree representation. As editing these color attributes directly (i.e. spherical harmonics coefficients) does not produce meaningful edits, NeuVV stores additional color values for the voxels edited by the user, which are blended in during rendering, thus enabling appearance editing. In Dyn-E [ZPS*23], the appearance of a local region on the surface representation is edited using a reference image, which can then be propagated to the rest of the dynamic scene consistently through invertible networks. Recently, style transfer was also introduced for dynamic scenes. DynVideo-E [LCW*24] transfers style separately to the background [BMV*22] and foreground [LCY*23] model—learned from monocular video—from corresponding reference images. The foreground style transfer is made consistent under animation and large-scale view and motion changes by additionally supervising with Score Distillation Sampling [PJBM22] from (1) a 3D diffusion prior [LWVH*23] to distill the inherent geometric information from the 2D reference image, and (2) a 2D text-based diffusion

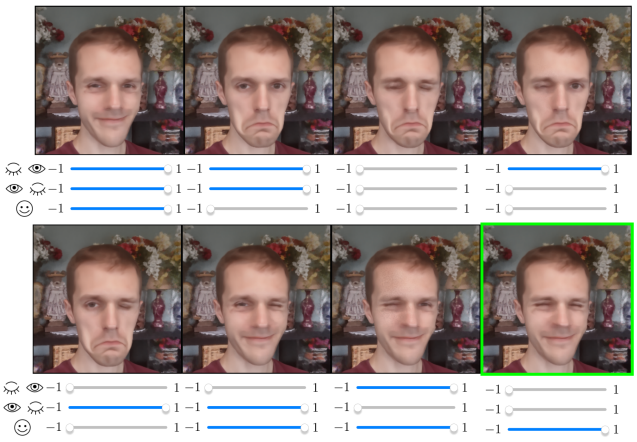

Figure 16: ***Novel Attribute Rendering.*** *Interpolation of user-annotated attributes learned by CoNeRF [KYK*22]. The expression in green border is not seen during training. Image source: [KYK*22].*

prior [RBL*22] to guide the rendered views, which is finetuned to the reference image using DreamBooth [RLJ*23]. A few methods [XLSL23, ZLY*23a] allow deformable style transfer for static scenes by incorporating deformation fields into their approach to cater to the geometric differences between the source observation and target style image.

3.3.2. Scene Dynamics Control

Next, we look at methods that enable users to manipulate scene dynamics through different types of control parameters. We structure the method based on the type of control variable, i.e. control based on attributes, keypoints, or actions.

Attribute-Based. Attributes are specific properties of an object that can exhibit different states, e.g. an eye being open, closed or in an intermediate state. CoNeRF [KYK*22] is the first method to extend neural radiance fields, specifically HyperNeRF [PSH*21], to allow user control over the scene dynamics, by treating these attributes as latent variables that a neural network can regress. It requires a few annotation masks from the user, as little as two throughout the video, for each region that needs to be controlled. The changes observed throughout the video for a particular annotated region are captured into a latent attribute, which the user can toggle at test-time to interpolate between the states and even generalize to combinations not seen in the video (see Fig. 16). DyLin [YJM*23] extends the method introduced in [KYK*22] to dynamic light field networks.

Keypoint-Based. CoNeRF provides control over the scene regions only using one-dimensional attributes and requires manual annotations from the user. In contrast, EditableNeRF [ZLX23], also building on HyperNeRF [PSH*21], proposes an approach that conditions the radiance of a sample point in the canonical space on a weighted combination of automatically detected keypoints. This allows deforming the overall scene into a new state by just deforming the keypoints, providing multi-dimensional control over the scene, which is consistent under novel views. However, the ability of the method to handle edits with large and complex motion is limited.

Action-Based. Recently, 3D neural scene representations have also been introduced for goal-driven deformable object manipulation and visuomotor control in robotics. For these tasks, typically, the future state of the scene is predicted by a forward dynamics model based on an action (e.g. parameterized control commands for a robotic arm like grasp, move, and release). Such a model learns the dynamics of the scene, taking non-rigid deformations into account, and is driven by the visual observations obtained from the scene, auto-decoded using the given state. The learned predictor can then be used to find the optimal set of actions to achieve the goal state in the target visual observation. In Wang *et al.* [WMDH22], the NeRF is conditioned on a set of 3D keypoints encoded from camera observation. The forward dynamics model predicts keypoints for the next time step given the current keypoints, gaining control over the representation. Li *et al.* [LLS*22] achieves similar control from multi-view observations using a state embedding rather than keypoints and an additional time contrastive loss to distinguish different views in one frame from views at another time step. ACID [SJC*22] encodes an RGB-D observation into canonical triplane features, which are decoded to learn an occupancy field for the current scene state and a correspondence field to the target observation. Control is achieved by predicting a flow field from action, which utilizes the correspondence field to drive to the target state. VIRDO [WFZF22] pretrains an object module to learn the nominal shape of the object, i.e. the shape before deformation and utilizes an encoder to incorporate measured external forces and contact points in learning the deformation field. It serves as a foundation for VIRDO++ [WZFF22], which uses an action module to predict the force and contact points for the next state, thus controlling the deformation dynamics. Finally, Driess et al. [DHL*23] use a compositional NeRF representation to model inter-object dynamics and learn a collision-aware dynamics prediction module from multi-view observations.

3.3.3. Object Pose Control

The methods discussed in this section either deform objects through explicit geometry manipulation or model object articulations using one of the coarse deformation structures introduced in Sec. 2.2.3. We categorize the works into those using explicit geometry deformation, skeletons, or cages.

Geometry-based. Unlike neural fields, explicit geometry representations, like meshes and point clouds, are easier to edit as they provide direct access to the surface. Non-rigid transforms can be directly applied by deforming the vertices. Recent methods gain control over the neural field-based geometry by linking it with explicit representations. The neural model is then trained under deformations of the explicit geometry to produce the correct renders. NeRF-Editing [YSL*22] uses the extracted mesh from NeuS [WLL*21] to gain control over its geometry. After the user edits the mesh, the offset from the original state is calculated, and rays are deformed accordingly, with the help of tetrahedralization of the mesh [HSW*20]. VolTeMorph [GKE*22] proposes a method to directly deform a tetrahedral mesh for editing the geometry rather than editing a surface mesh and transferring the motion to the volumetric one, like in NeRF-Editing. NeuPhysics [QGL22] extends

the concept of NeRF-Editing to dynamic scenes by adding an additional bending layer on top of the motion field to account for deformations due to edits, which is further regularized by strong physics priors using physical simulation. NeuMesh [CJH*22] proposes a hybrid representation by directly storing a separate latent code for appearance and geometry on mesh vertices extracted from a pre-trained NeuS. The deformed mesh, after editing, can be rendered by interpolating vertex codes along the ray. It also allows texture editing by replacing the appearance latent codes with ones from a different scene. Bypassing the expensive extraction step and limited topology change handling of meshes, NeuralEditor [CLW23] directly stores neural features on a point cloud and renders via interpolating the point features, similar to Point-NeRF [XXP*22]. However, they show that directly deforming the points in Point-NeRF does not lead to high-quality renders of the deformed object. They incorporate more geometric information, like point normals, to guide the editing process, which also enables modeling the color with the Phong reflection model [Pho75], obtaining a much more precise point cloud overall. The method achieves high-fidelity rendering results on deformed scenes, even in a zero-shot inference manner, without additional training.

Skeleton Prior. The deformation of humans, animals, and many other articulated objects can be represented and controlled by an underlying skeleton. Such skeleton-based rigging is common for meshes, and it has recently been used to control neural scene representations as well, enabling articulated deformations to novel, unseen poses. When using a neural field representation, the task is to infer the geometry or radiance for every 3D point in the live frame (or posed frame) as a function of the pose - while also inferring the influence of each bone transform on every point via blend skinning - and learn the deformation from a canonical pose or vice versa. Now we look at methods where the skeleton is fitted to the observations a priori using off-the-shelf methods, and only the shape or appearance of the object is learned on top, along with the skinning weights. Note that we do not focus on human-specific [JYS*22, PZX*21, XAS21, LHR*21] or human body part-specific [MBW*23, BZH*23, SWW*20] methods which are based on surface templates and skinning weights from parametric models, such as SMPL [LMR*15]. We cover methods that can conceptually generalize to generic objects beyond humans and do not require massive amounts of human-specific data.

The first work to demonstrate *articulated* pose control of a neural field is NASA [DLJ*20] via a *backward model*. The main idea is to consider objects as a collection of rigid part-based models and transform a posed point back (hence backward model) to the canonical space according to the inverse rigid pose of the part. Since the part association for the point is unknown, the method transforms the point according to each part separately (most basic form of *inverse skinning*). It obtains occupancy as the maximum of the part-based model occupancies (a point is occupied if any of the parts occupies it). The part-based occupancy models are learned as a function of pose to deal with non-rigid deformations. A common limitation of pure inverse skinning methods is that they do not consider non-rigid deformation beyond articulated effects. For this, inverse skinning has been generalized further to represent *deformable* shapes in Neural-GIF [TSTPM21]. Inspired by the idea of deforming the input points before the evaluation of the SDF or occupancy [SP91], it learns both articulated and non-rigid deformation (e.g., clothing and soft tissue) by first un-posing the points with inverse skinning and then further deforming them with a learned deformation field to the canonical space. Inverse skinning is made smooth by linearly blending body-part deformations weighted by skinning weights predicted by the network. Like dynamic NeRFs [PCPM21], learning the deformation field instead of an occupancy/distance field as a function of deformation parameters leads to more detail with smaller models.

While models like Neural-GIF [TSTPM21] produce highly detailed surfaces, the inability of inverse skinning to generalize to highly articulated poses remains (see Fig. 8 for the same issue of backward modeling in dynamic NeRFs). To address this issue, SNARF [CZB*21] was the first method to propose *forward skinning* (forward model), directly finding the point in canonical space that will deform to the target posed point by numerically solving a non-linear equation. Time-invariant skinning weights are predicted in canonical space, removing the pose dependency of inverse skinning. Forward skinning is shown to generalize better to new, unseen poses, albeit with some loss of surface detail. FAST-SNARF [CJS*23] achieves a 150x speed-up over SNARF by parameterizing the skinning weights field using a low-resolution voxel-grid, which works, as they demonstrate that the field does not contain high-frequency details.

The use of a canonical representation coupled with either forward skinning [LTV*22] or inverse skinning [NSLH22] has been exploited to represent *articulated neural radiance fields* as well. The main difference with neural implicit surface models is that supervision is done via rendering loss instead of direct 3D ground truth. TAVA [LTV*22] builds on SNARF and learns the deformation through multi-view images. It further demonstrates generalization to articulated animals. Since learning the backward mapping from the live frame to the canonical spaces is, in general, hard for novel poses, DANBO [SBR22] leverages the skeleton structure by predicting a localized volumetric representation of each body part with a graph neural network and biasing the model to attend to nearby body parts, reducing spurious correlations. Recently, NPC [SBR23] replaced the purely volumetric NeRF representation of the scene with a collection of canonical neural points, which can be posed and volume rendered by interpolating the neural features. While all the previously discussed methods reconstruct only the object, HOSNeRF [LCY*23] learns to reconstruct the background as well from a single monocular in-the-wild video, along with a foreground model which is based on a skeleton (human-based in the paper [WCS*22]) extended with bones to represent objects, modeling human-object interaction. The smoothness and consistency of human-object deformations is further improved by defining cycle consistency between a forward and backward deformation module. These modules add a non-rigid deformation field on top of linear blend skinning to model fine deformations. GART [LWP*24] and ASH [PZK*24] learn a Gaussian Splatting representation on top of skeletons, achieving fast reconstruction and real-time rendering of avatars.

Skeleton Discovery. The aforementioned methods achieve a higher level of control by relying on an underlying skeleton model. While it is reasonable to assume such prior knowledge in the case of hu-

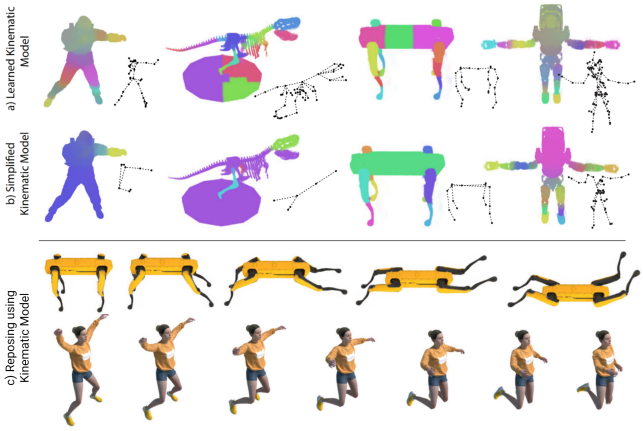

Figure 17: ***Skeleton Discovery and Reposing.*** *Shown are LBS weights learned by Uzolas et al. [UEK23] from multi-view image data before (a) and after (b) post-processing. Using the learned LBS weights, the full models can be reposed by skeletal animation (c). Image source: [UEK23].*

mans, it might be hard to obtain such skeletons for general shapes. Hence, several methods learn the skeleton structure directly from the observed data. CAMM [KKK*23] learns the skeleton structure via inverse rendering from monocular videos of a deforming object. First, the canonical shape is estimated with learnable *anchor* points, representing surface deformation via blend skinning and learned weights. Then, a skeleton initialized with a pre-trained RigNet [XZK*20] is hooked to these anchor points using the proposed association mechanism, which can then be used to drive surface deformations. To avoid the aforementioned challenges of backward modeling (generalization issues, non-invertible mappings), the recent work by Uzolas *et al.* [UEK23] uses a point-based representation, coupled with forward skinning to learn the canonical space, time-varying body-part deformations, and the skeleton. The skeleton is initialized with the medial-axis transform instead of RigNet, which is further refined by pruning or merging bones, as shown in Fig. 17. Transfer4D [MNH23] proposes a pipeline to automatically track non-rigid objects and extract a skeleton as a post-process, which is used to re-target motion to semantically similar shapes. Liu *et al.* [LGW23] optimize part segmentation, skeleton, and kinematics from point cloud videos. As the problem is under-constrained, a two-stage approach is proposed which first optimizes the more tractable 6-DOF rigid model without kinematic constraints, later projecting it to a 1-DOF model with screw-parameterized joints and a valid kinematic tree.

Neural Parametric Models. Another way to avoid the object-specific constraints of pre-defined skeletons is to model the articulation space of an object using an auto-decoded MLP, a paradigm introduced by NPMs [PBT*21]. Once the latent codes are optimized for poses at observed timesteps, they can be interpolated in the latent space to get novel poses. They also allow motion retargeting by using pose embeddings from a driving video sequence, given that the geometry model has been learned and fixed for another sequence. Building on previous works on articulated shape reconstruction from a few (or single) monocular videos [YSJ*21a, YSJ*21b], BANMo [YVN*22] represents the articulation space of an object using neural bones instead of displacements like NPMs [PBT*21]. Bone positions and orientations are predicted for each timestep and volumetric skinning is used to articulate the object space using these bones. Both forward and backward skinning weights are learned to induce cycle consistency. Follow-up work MoDA [SCC*23] replaces the linear blend skinning of BANMo with dual quaternion blending, enabling better surface reconstruction and fewer skin-collapsing artifacts than the former. Both methods demonstrate the ability to transfer motion between structurally similar avatars. RAC [YWRR23] extends BANMO to reconstruct a category-level model and skeleton, where the auto-decoded MLP predicts the instance-level joint locations for the category-level skeleton, allowing additional pose control. NPGs [DWY*24] use a deformation basis to constrain the articulation space, optimizing the basis coefficients for the poses observed at each timestep, which can then be interpolated to get novel intermediate poses.

Cage-based. Concurrent works [PYL*22,XH22] introduced cages, well-known in computer graphics for mesh editing, to neural fields. CageNeRF [PYL*22] optimizes a cage with respect to the mean-value coordinates (MVC) [JSW23] of the vertices of the underlying mesh, which is extracted from the SDF-based radiance field using marching cubes. The cage is edited using the module deforming it based on a given target mesh of the deformed state. Since MVCs are invariant under cage deformations, an MVC field is built using the deformed cage, which is then used to access the canonical radiance field while rendering views from the deformed state. Taking a different approach, Xu *et al.* [XH22] compute the reverse deformation of the deformed cage to the canonical cage to query the radiance and density while rendering. They also propose a faster cage coordinates computation method based on harmonic coordinates [JMD*07]. Overall, cages enable category-agnostic control over object geometry and are easy to generalize under various settings; however, the deformation quality is significantly dependent on the cage generation process, requiring manual refinement for complex shapes.

3.4. Generalizable and Generative Modelling

Previously described methods in Sections 3.1, 3.2, and 3.3 follow the approach of pure optimization, which finds an optimal 4D representation that respects a given set of observations and manually defined constraints. As a natural extension of this principle, one can try to learn constraints from data. This principle can serve two goals: to be applicable in much more under-constrained situations with few observations by generalizing from training data (*generalizable*), or without any observations at all to sample from the learned model itself (*generative*). To follow this principle, a model—sometimes called a *data prior*—has to be learned from a large amount of data that optimally captures the scale and variety of natural non-rigid scenes. In the setting of general non-rigid reconstruction, the amount of datasets that allow us to learn large-scale models is very limited, which is why we are still in the early stages of this research area and only a few approaches exist. Please refer to Sec. 2.3.4 for an overview of fundamental techniques.

In this section, we will discuss the early steps in the direction

Dataset	#V	M	R	A	Size
Fauna Dataset [LLL*24]	1		✔		78,168
EPIC Fields [TDZ*23]	1		✔		18.8k
Common Pets in 3D [SSR*23]	1		✔	—	4.2k
BEHAVE [BXP*22]	4		✔		321
SAIL-VOS 3D [HWYS21]	1	-	✘		237.6k
DeformingThings4D [LTT*21]	—		✘		1.9k
DeepDeform [BZTN20]	1	-	✔		390k
Dynamic Scene [YKG*20]	12		✔		8
3DPW [vMHB*18]	1		✔		51k
D-FAUST [BRPMB17]	—		✔		129

Table 4: ***Large Datasets.*** *A list of datasets that are or have the potential to be used for generalizable non-rigid 3D modeling. We exclude datasets that cover only one domain (such as humans) while making an exception for D-FAUST, as it has been used in many general methods.* ***# Views (#V). Modality (M):*** *RGB, with Depth, 3D Objects.* ***Real Data (R):*** *Yes ✔, No ✘.* ***Annotations (A):*** *Camera Poses, Registered Templates, 3D Meshes, Segmentations, Animated Models, Optical Flow, IMU Poses, Matchings.* ***Size:*** *# Frames, # Sequences.*

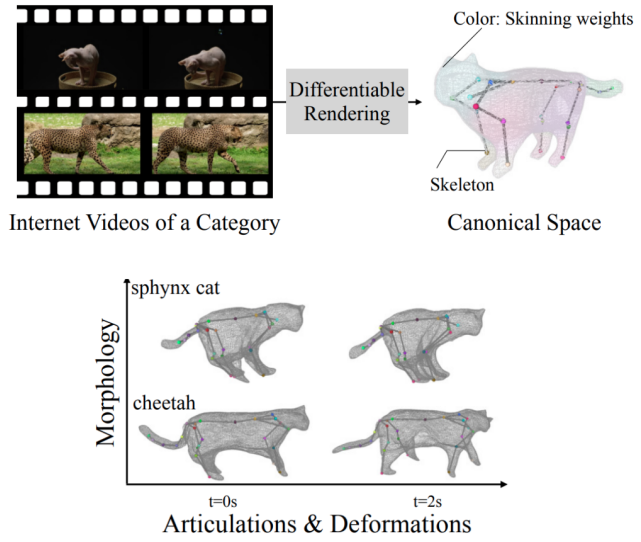

Figure 18: ***Category-level Shape and Articulation Prior.*** *RAC [YWRR23] learns a category-level 3D model and skeleton from a collection of internet videos, which can then be fitted to specific instances. Image source: [YWRR23].*

of generalizable and generative non-rigid reconstruction. We begin by introducing existing large-scale datasets in Sec. 3.4.1, which are required to learn effective models from data. We summarize generalizable models in Sec. 3.4.2 and generative models in Sec. 3.4.3.

3.4.1. Datasets

Datasets to learn generalizable or generative models for non-rigid reconstructions are still rare. An overview is given in Tab. 4. We restrict ourselves to datasets that have been or can probably be used to learn data priors, requiring scale and generality. We exclude datasets that capture only a single domain, such as humans or hands, except when they are used for general methods. Existing generalizable methods in the scope of this work learn models from two types of datasets: (multi-view) video datasets and datasets containing animated 3D models.

3.4.2. Generalizable 4D modeling

The bulk of generalizable methods for non-rigid reconstruction work on the level of objects. Thus, we start by first introducing these methods. Then, we turn to more general, scene-level methods, which include learning priors for scene flow, temporal correspondences, or scene radiance fields.

Object-level models in 3D space. We start with OccupancyFlow [NMOG19], which models forward and backward flow between occupancy fields as continuous neural fields while solving a neural ODE for deformation. Tang et al. [TXJZ21] learn to autoencode object scan sequences with a spatio-temporal autoencoder that predicts occupancy and correspondences. Similarly, Jiang et al. [JZW*21] learn to disentangle object identity, object pose, and motion by formulating a neural ODE in the latent space. All these methods are trained on D-FAUST. A slightly different trend for object-level reconstruction is using neural parametric models [PBT*21]. They learn a deformable template as an MLP autodecoder from a dataset of 3D objects. Neural templates have several useful properties. They can disentangle identity from deformation [XMY*21, MA22], automatically perform part decomposition [PSTD22], and provide correspondences between instances [MA22].

Object-level Models in Image Space. Previously described approaches take 3D input to predict deformation. In contrast, a parallel line of work, started by REDO [RZS21], makes predictions about deformation from images. REDO learns to recover shapes from videos by training CNNs to predict flow fields in pixel space for each time step. The CNN is trained on SAIL-VOS 3D, DeformingThings4D or 3DPW. In the last two years, this principle gained more traction. TrackeRF [SSR*23] learns object-level non-rigid deformations by predicting trajectories of 3D points, given image-aligned feature information. Tan et al. [TYR23] use a dataset of pretrained dynamic NeRFs obtained by BANMo [YVN*22] to train a video encoder to predict viewpoint, appearance, and articulation from a video of a novel object.

Articulation Priors. A special case of object-level non-rigid reconstruction concerns articulated objects with multiple rigid parts. Works discussed in Sec. 3.3.3 - Skeleton Priors assume that a skeleton is already available, predicted using an off-the-shelf method. Such methods learn category-level articulation priors using a large amount of 3D data, which is usually only available for characters [XZK*20]. Rather than 3D data, RAC [YWRR23]—following the neural parametric model paradigm—learns category-level 3D models and skeletons from monocular internet videos of a category. Using a 3D background model to separate objects from the background, it learns a structured latent space within a category, optimizing over multiple instances by specializing the skeleton to instance morphology, as shown in Fig.18. In contrast, BANMO [YVN*22] learns articulations on instance-level only.

Another line of work with significant recent progress learns de-

formable templates for generalizable, category-level few/single-view reconstruction from image collections instead of temporal data, which can then be animated or optimized for per-frame video reconstruction [YRH*23, YHL*23, WLJ*23, AMA24, LLL*24]. Hi-LASSIE [YHL*23] extracts a class-level part-based skeleton from around 20-30 images of an articulated animal class using semantic cues from DINO features [CTM*21]. The skeleton is then specialized to instances and textured during test-time optimization. ARCTIC3D [YRH*23] improves the robustness and quality of Hi-LASSIE by introducing 2D diffusion-based priors, which are utilized during input preprocessing, shape and texture optimization, and animation.

A few methods [WLJ*23, AMA24, LLL*24] enable direct test-time prediction of shape, articulation, texture, and viewpoint from a single view using embeddings from an image encoder. MagicPony [WLJ*23], which removes the requirement of DOVE [WJRV23] for explicit video data during training, initializes a heuristic-based skeleton from the learned category shape which is then specialized to instances while SAOR [AMA24] directly predicts the part-based segmentation and the corresponding transformations to get the instance shape, which is then articulated using skinning weights. Building on MagicPony, 3D-Fauna [LLL*24] exploits the 3D shape similarities across different animal categories by predicting the category shape as a combination of latent shape bases, retrieved based on similarity to the image embeddings.

Considerable effort has also been made recently to learn articulation states of object-categories [MQK*21, HJZ23, WCM*22, TLYCS22], to allow articulated reconstruction from image or scan inputs. The learned articulated states (e.g., joint angles) can be used to repose the object, enabling control similar to the methods discussed in Sec. 3.3.3. Recently, CARTO [HIZ*23] brought this concept to the scene level by learning to jointly predict object locations, their 6D pose, and their articulation in a scene of multiple objects.

Method	Type	Prior	Data
Generalizable - Object-level			Input
GNPM [MA22]	◀	🧊 🚶	📉
SPAM [PSTD22]	◀	🧊 🚶	📉
TrackeRF [SSR*23]	⋈	🧊 ▦ ↗ ⁘	🎥
Tan et al. [TYR23]	⋈	🧊 ▦ 🚶	🎥
RAC [YWRR23]	◀	🧊 🚶 ⁘	🎥
SAOR [HIZ*23]	⋈	🧊 🚶	🖼
Generalizable - Scene-level			Input
MonoNeRF [TDD23]	⋈	🧊 ▦ ↗ ⁘	🎥 ▦
FlowIBR [BBNB23]	⋈	🧊 ▦ ↗ ⁘	🎥 ▦
Van Hoorick et al. [VHTS*22]	⋈	🧊 ▦ ↗	🎥 ▦ 📉
Zhou et al. [ZFB23]	⋈	🧊 ↗	📉
OcclusionFusion [LZYX22]	⋈	🧊 ↗	🎥 📉
CARTO [HIZ*23]	⋈	🧊 🚶	🖼 ▦
Generative Approaches			Train
Kim et al. [KY22]	⇆	🧊 ▦ ↗	🎥
HyperDiffusion [EMS*23]	⇆	🧊 🚶	🧊
Bahmani et al. [BPP*23]	◆	🧊 ▦ ↗	🎥
GenCorres [YHS*24]	⋈	🧊 🚶	🧊
Tang et al. [TMW*22]	⋈	🧊 🚶	🧊
NAP [LDS*23]	⇆	🧊 🚶	🧊
PGM [MSL*24]	⇆	▦	🎥
MAV3D [SSP*23]	⇆	▦	🎥

Table 5: ***Generalizable and generative methods.*** *Recent works on generalizable or generative non-rigid reconstruction.* ***Method type****: Encoder-Decoder ⋈, Autodecoder ◀, GAN ◆, Diffusion ⇆.* ***Learned priors****: Shape 🧊, Appearance ▦, 3D Deformation/flow ↗, Articulation 🚶, 2D Correspondences ⁘.* ***Input or training data****: Image 🖼, Video 🎥, Multi-View ▦, Depth 📉, Shapes 🧊.*

Scene Flow Priors. Another line of work combines traditional non-rigid reconstruction techniques with a scene flow prior, e.g. by utilizing FlowNet3D [LQG19, WLHJ*20]. 4DComplete [LTT*21] utilizes FlowNet3D and a 4D voxel encoder to learn scene motion over time. OcclusionFusion [LZYX22] trains a motion estimator based on GNNs and LSTMs to predict the motion of points for an embedded deformation graph on the canonical model (see Fig. 13). Both learn to propagate motion from visible to occluded regions and are trained on DeformingThings4D [LTT*21]. 4DComplete additionally predicts the occluded geometry as well. Trained on D-FAUST, Zhou et al. [ZFB23] use MLPs to decode flow and SDF from 3D feature volumes unprojected from depth frames, which are enhanced by self- and cross-attention between the source and the target frame.

Correspondence Priors. Priors can be learned from data to recover correspondences between different frames. As an example, DeepDeform [BZTN20] trains a Siamese CNN architecture to perform non-rigid matching between RGB-D frames, trained in a semi-supervised setting on their own DeepDeform dataset with sparse correspondence annotations. The matches are then used as a constraint in the reconstruction pipeline. Neural Non-Rigid Tracking [BPZ*20] follows a similar principle but predicts dense correspondences instead in an end-to-end learnable framework. Another way to find dense correspondence matches is to distill pre-trained image features (e.g. 2D DensePose CSE embeddings [NNS*20]) from observations into the canonical 3D model. Such 3D features are constrained to match the corresponding 2D features across observations, providing long-term registration [YVN*22, SCC*23, YWRR23, KKK*23].

4D Scene Generalization. The first data-driven methods for general non-rigid scenes appeared in the past two years. MonoNeRF [TDD23] combines off-the-shelf optical flow with a model that provides pixel-aligned volume features for generalization to novel videos. It jointly learns to predict dense motion, shape, and appearance and is trained on the Dynamic Scene dataset. Van Hoorick et al. [VHTS*22] train a point transformer on multi-view RGB-D videos to learn to track occluded objects from a single view. The transformer attends to other timeframes, which enables it to look up currently occluded information and fill in missing information. Recently, FlowIBR [BBNB23] combines a learned static scene prior with a 2D optical flow prior for fitting a single dynamic scene. During test-time optimization, the method optimizes ray bending in the inferred static representation to minimize optical flow and cycle consistency losses. Zhao *et al.* [ZCM*24] shows that given monocular depth estimation and optical flow priors for geometry

and motion learning, their pseudo-generalized method is able to avoid costly scene-specific appearance optimization and still give results on par with scene-specific methods.

3.4.3. Generative Models

The field of generative models on general non-rigid objects and scenes is in the early stages of development. We include them in this survey because recent work in rigid 3D reconstruction [CNC*23, ZT23] indicates that generative models will also be used as data priors for non-rigid reconstruction in the near future. Generating large, general scenes is already very challenging in a rigid setting, due to the lack of large-scale datasets and scalable generative models. Thus, most existing non-rigid generative methods either focus on single objects or lift 2D priors into 3D.

Generation of Articulated or Deformable Objects. Generative models for deformable objects have seen progress in recent years. ShapeFlow [JHTG20] learns the deformation space of an object category by training an SDF neural field and a flow field on ShapeNet. Using an explicit representation of meshes, ARAPReg [HHS*21] designs a spectral ARAP regularization for graph shape generators, constraining them to produce a series of shapes that move as rigidly as possible [SA07]. Recent advances include GenCorres [YHS*24], which brings the ARAPReg-like principle to implicit neural fields, enforcing cycle consistency and rigidity constraints, and the work of Tang et al. [TMW*22], which further extends the modeling of deformation spaces by allowing generation conditioned on user-given constraints via dragging handles, exposing more control over deformation. HyperDiffusion [EMS*23] can sample from a 4D shape distribution by training a diffusion network to generate MLP weights of an SDF neural field. Further, for articulated objects with non-rigid parts, NAP [LDS*23] trains a diffusion model on articulation trees of objects to generate partially rigid shapes in different articulation states. Another line of work is 3D video generation. Bahmani et al. [BPP*23] train a GAN that modulates neural field weights to generate 3D videos of faces. Kim et al. [KY22] generate a deformation sequence of medical image volumes with a diffusion model synthesizing intermediate 3D frames. Similar to object-level approaches, their method focuses on a very narrow data domain.

General 4D Generation. The previously described methods generate articulated or deformable objects but are not able to generate full scenes. Non-rigid scene generation was largely out of reach, until the appearance of foundational diffusion models. These powerful open-world generators also made their impact in the field of 4D generation. PGM [MSL*24] generates frames with a 2D diffusion model, conditioned on a 4D game engine state consisting of several explicit parts. MAV3D [SSP*23] uses two types of diffusion models, text-to-image and text-to-video generators, and fuses the generated images/videos into 3D with a two-step procedure. First, a static scene is generated by fusing the outputs of the text-to-image generator into a HexPlane representation via 3D-aware score distillation sampling (see Sec. 2). Then, the text-to-video generator is used to animate the generated scene. The generation results depend on those given by the diffusion generators. The method can generate multiple objects at once but is still limited to very small scenes. 4D-fy [BSR*24] introduces a 3D-aware text-to-image model to deal with the Janus problem—a scene that has repeated 3D structure when seen from multiple viewpoints—and optimizes the scene using text-to-image, 3D-aware text-to-image, and text-to-video models in an alternative fashion to retain the desirable qualities from each model (see Fig. 1, bottom row, on the right).

4. Remaining Challenges and Discussion

Despite the remarkable and rapid progress in non-rigid 3D reconstruction of general scenes, multiple challenges remain. This section highlights the key open challenges and future directions.

Intrinsic Decomposition and Relighting. In static reconstruction methods, it is common today to model view-dependent appearance [MST*20, WLL*21]. However, in dynamic scenes, not only the view direction changes but also the incoming light direction to moving elements in the scene. This implies that accurately reconstructing a non-rigid scene requires modeling physical light transport and environment maps. Doing so is also necessary in order to be able to relight objects with new illumination conditions if the objects from a reconstructed scene are to be edited or extracted and placed into a new scene. However, current non-rigid view synthesis approaches for general scenes are based on the simple emission-absorption model, which cannot model such changes. They either ignore illumination changes entirely, implicitly capture them using per-time step latent codes, or model appearance individually for different time steps. Modeling surface materials and environment maps is already well-explored for static scenes [ZXY*23, JLX*23b, ZSD*21, SDZ*21, BBJ*21], albeit the use of data priors is minimal, and most of the methods rely on a sufficiently large number of views to perform a per-scene reconstruction. For non-rigid scenes, methods that do intrinsic decomposition are based on human-specific templates [ICN*23, CL22, ZYW*23, BLS*21, LMM*22, TAL*07, WSVT13, LWS*13]. A few early approaches exist for general non-rigid scenes [WZN*14, GLD*19] but investigation with the new neural implicit representation paradigms is still missing. Transparent materials are, even for static methods, rarely addressed [WZS23, CLZ*23].

Faster Scene Representations. 3D Gaussian splatting [KKLD23] is a recent method for novel-view synthesis that has enabled real-time rendering for static scenes, resulting in quick adoption of the representation across all applications such as editing [CCZ*23, FWZ*24], surface reconstruction [GL24, CLL23], generative modeling [CWL23, TRZ*24] and SLAM [KKJ*23, YLGO23], among others. As 4D scenes take longer to optimize because of the extra time dimension, a faster representation will be even more useful and will improve scalability. A few initial works have already been proposed for the general non-rigid setting [LKLR24, WYF*24, DWY*24] and extension to non-rigid applications discussed in this report are likely to follow. However, even the 3D Gaussian Splatting representation requires offline, scene-specific optimization. Real-time, online geometry reconstruction and tracking is possible using classical representations and RGB-D input [LZYX22]. Doing the same for high-fidelity appearance reconstruction would open up significantly more AR/VR applications.

Reliable Camera Pose Estimation. Even when using offline reconstruction, estimating camera poses in a strongly deforming

scene and removing the dependency on reliable pose estimation from rigid Structure-from-Motion remains challenging. RoDynRF [LGM*23] is a step towards this goal but requires additional optical flow and monocular depth supervision.

Long-Term Dense Correspondences. While many works propose systems that allow for establishing 3D correspondences over time on synthetic or lab-captured data [TGZ*24, LKLR24], most of these methods do not yield satisfactory results for general real-world scenes with long-range and complex camera trajectories. Methods that model changes over long timespans [LWC*23a] do so at the expense of 3D correspondences. One step in the direction of improving these correspondences is provided by TotalRecon [SYD*23], which uses depth information and motion decomposition into each object's root body and articulated motion to scale to minute-long videos containing challenging motion. OmniMotion [WCC*23] can also establish dense, occlusion-aware motion trajectories. However, they use a quasi-3D representation that does not explicitly disentangle the camera and scene motion; thus, the resulting representation is not a physically accurate 3D scene reconstruction.

Reconstruction from Sparse Casual Captures. Most works require at least a dense video stream as input, while the majority of videos captured in the real world cover the scene relatively sparsely, with small view baselines, occlusions, and fast object motion relative to the camera. Most current approaches are evaluated on datasets that contain multi-view signals in [GLT*22] and they perform poorly in the sparse capture setting. How to enable dense 4D reconstruction from such sparse coverage is one of the key future challenges.

Specialized Sensors. As we have seen in this report, utilizing multi-view capture systems or RGB-D cameras is common but only few works utilize specialized sensors such as LiDAR [ALG*21, TZFR23] and event cameras [MPCVG23, BMC*24, MLR*24] for non-rigid 3D reconstruction. LIDARs are commonly used in autonomous vehicles to get reliable depth estimates in outdoor urban scenes, while they are also becoming common in mobile phones, with the potential to enable a whole host of AR applications. Event sensors provide high temporal resolution, especially for fast-moving scenes. Incorporation of these sensors in a multimodal setting could be highly beneficial for non-rigid 3D reconstruction and view synthesis in these varied and challenging scenarios.

Compositionality and Multi-Object Interaction. Reconstruction of two or more general deformable objects interacting with each other remains an open challenge. Most discussed methods capture dynamics for single objects or on a scene level, disregarding the interaction between different scene parts. A few category-specific methods explicitly model human-object interactions [SXZ*22, ZLY*23b, JYS*23, JJS*22], hand-object interactions [QCZ*23, XYZ*23, HYZ*22, HJH*22, ZBYX19, TA18] and interaction between different human parts [SGPT23, MDB*19], but only a few approaches exist for modeling interactions between general dynamic objects [LSS*22]. Initial steps have been taken in this direction by the compositional methods discussed in Sec. 3.2, which incorporate physically plausible background-foreground interactions [YYZ*23], occlusions [WDSY23] and collision dynamics [DHL*23] between multiple objects. However, more mature handling remains elusive. Such interaction handling is also important for geometry and pose editing, where objects come into contact and separate again over time.

Vision-Language Models for Non-Rigid Scenes. While text-driven generative models have become popular for static 3D content [PJBM22, HTE*23], the relationship between text and general non-rigid scenes has not been explored until very recently. First examples of Text-to-X tasks relevant for non-rigid scenes, namely 4D [SSP*23] and Articulated 3D [KAZ*23, LDS*23] have appeared recently. Text-driven editing [MPE*23] and motion synthesis [DMGT23] has also been proposed for non-rigid scenes but they remain human-specific for now. Other than generative use cases, imbuing non-rigid scenes with language embeddings—already achieved for static scenes [KKG*23, QLZ*24]—will result in a richer representation and enable more interactive use cases.

Fakes and Authenticity. Many recent methods for general non-rigid 3D reconstruction can provide photo-realistic novel views of recorded scenes with humans, as this report evidences. Such synthetic views usually do not deviate much from what has been observed, though they hallucinate small details (e.g. in the areas occluded in the input views). If, however, the methods allow editing of any kind (e.g. appearance editing or adding or removing the entire scene elements; automatically or with manual assistance), the generated scenes pose a risk of being perceived as real and should be declared as edited. Moreover, in this case, the consent of the participants regarding possible scene edits and their intended usage is required. At the same time, the concern of imagery falsification does not only apply to scenes with humans. Detection of synthesized imagery and forensics of visual data, including general non-rigid scenes, is a sub-field of research with its own body of work [ZGLA23], and approaches discussed in this STAR could assist in detecting synthesized content more effectively.

Physics-based Methods. While physics-based priors have been successfully applied in sparse 3D reconstruction (e.g. 3D human motion capture) [SGXT20, IYOK20, SGX*21, DSJ*21, GAXS22, XWI*21, LBX*22], only a few methods adopt them in the non-rigid 3D reconstruction of general and dense scenes or objects [CLZ*22, QGL22, YYZ*23, LQC*22].All physics simulators—as sophisticated and accurate as they can be—make assumptions and simplifications, where it is impossible to model all effects that influence 2D observations and the underlying 3D reconstructed states. Hence, hard physics-based constraints have upper bounds on the accuracy they can provide. Consider the differentiable physics-based simulator used in ϕ-SfT [KTE*22], imposing hard physics constraints. It remains open if the relaxation of such constraints could improve the 3D surface reconstruction accuracy of such methods. Furthermore, methods jointly estimating physical parameters and the geometry [KTE*22, QGL22] allow editing of the estimated dynamic 3D scenes in a physically meaningful way, and more works exploring this are foreseeable in the future.

Generalizable Modeling and Generative Priors. Existing research on 3D non-rigid generalizable and generative models focuses on learning data priors on an object level, such as for humans, faces, or simple object categories. Models for general non-rigid scenes remain a challenge since large-scale datasets with dynamic scenes are rare and existing generative models do not scale

well to general 4D data. Probabilistic diffusion models [HJA20] or flow matching [LCBH*23], for example, are promising candidates for both generalizable and generative models, and are challenging to scale to 3D data. They only have been applied in rigid object-level settings [KVNM23, MSP*23, MKRV23, CGC*23] or to generate intermediate representations so far, such as the weights of an MLP [EMS*23]. The recent trend to distill knowledge from 2D diffusion models, such as Stable Diffusion [RBL*22], ControlNet [ZRA23], or Instruct-Pix2pix [BHE23], might be a promising way forward, also for the non-rigid setting. However, the challenge to solve is how to efficiently achieve spatial and temporal consistency when different views or frames are generated independently. MAV3D [SSP*23], for example (see Sec. 3.4.3), takes multiple hours to generate a very simple scene in low resolution. For a comprehensive review of diffusion models in visual computing, we refer to the recent survey of Po *et al.* [PYG*23].

5. Conclusion

This state-of-the-art report focused on the recent trends in the fast-growing research field of non-rigid 3D reconstruction of general scenes. Its central aspects are deformation modeling, different ways to learn data-driven priors, and leveraging inductive biases of neural methods. While the reviewed approaches allow reconstructing deformable geometry from different sensor types and producing different 3D output representations, the latest methods have been strongly influenced by computer graphics and versatile implicit 3D representations. The vast majority of the reviewed methods are neural, although remarkably, not all of them. Two major upcoming trends that we observe are real-time capable Gaussian Splatting and diffusion-based priors for generalizable models. Furthermore, we see models leveraging pre-trained features or segmentation approaches to move towards compositionality, as this simplifies modeling geometry and motion priors and enables editing capabilities. Another trend is moving towards self-supervised learning for scene decomposition, e.g. for structure and skeleton discovery, instead of relying on masks or templates. We conclude the report by presenting an overview of the open challenges and promising future research areas. It is expected that ideas developed first for the static setting will be quickly adopted for the non-rigid case, e.g. leveraging priors from diffusion models or introducing surface materials, and other rapidly evolving research directions such as generative AI will continue influencing it likewise.

Acknowledgments

We thank Cameron Braunstein and Navami Kairanda for their feedback on the draft. R. Yunus was supported by the Max Planck & Amazon Science Hub. V. Golyanik and C. Theobalt were supported by the ERC Consolidator Grant 4DReply (770784). G. Pons-Moll is supported by the funding from the Carl Zeiss Foundation, DFG-409792180 (Emmy Noether Programme, project: Real Virtual Humans), German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. Open Access funding enabled and organized by Projekt DEAL.

References

[AHR*23] ATTAL B., HUANG J.-B., RICHARDT C., ZOLLHOEFER M., KOPF J., O'TOOLE M., KIM C.: Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16610–16620. 2, 5, 13, 14

[ALG*20] ATTAL B., LING S., GOKASLAN A., RICHARDT C., TOMPKIN J.: Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *European Conference on Computer Vision (ECCV)* (2020), pp. 441–459. 18

[ALG*21] ATTAL B., LAIDLAW E., GOKASLAN A., KIM C., RICHARDT C., TOMPKIN J., O'TOOLE M.: Törf: Time-of-flight radiance fields for dynamic scene view synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021), vol. 34, pp. 26289–26301. 14, 28

[AMA24] AYGÜN M., MAC AODHA O.: Saor: Single-view articulated object reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 26

[AP09] AGARWAL P., PRABHAKARAN B.: Robust blind watermarking of point-sampled geometry. *IEEE Transactions on Information Forensics and Security 4*, 1 (2009), 36–48. 6

[ASF*23] AZAGRA P., SOSTRES C., FERRÁNDEZ Á., RIAZUELO L., TOMASINI C., BARBED O. L., MORLANA J., RECASENS D., BATLLE V. M., GÓMEZ-RODRÍGUEZ J. J., ET AL.: Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data 10*, 1 (2023), 671. 17

[BBJ*21] BOSS M., BRAUN R., JAMPANI V., BARRON J. T., LIU C., LENSCH H.: Nerd: Neural reflectance decomposition from image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 12684–12694. 27

[BBLT18] BRASCH N., BOZIC A., LALLEMAND J., TOMBARI F.: Semantic monocular slam for highly dynamic environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), IEEE, pp. 393–400. 16

[BBNB23] BÜSCHING M., BENGTSON J., NILSSON D., BJÖRKMAN M.: Flowibr: Leveraging pre-training for efficient neural image-based rendering of dynamic scenes, 2023. arXiv:2309.05418. 26

[BFCN18] BESCOS B., FÁCIL J. M., CIVERA J., NEIRA J.: Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters (RA-L) 3*, 4 (2018), 4076–4083. 16

[BHE23] BROOKS T., HOLYNSKI A., EFROS A. A.: Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 29

[BHZK05] BOTSCH M., HORNUNG A., ZWICKER M., KOBBELT L.: High-quality surface splatting on today's gpus. In *Eurographics/IEEE VGTC Symposium on Point-Based Graphics (PBG)* (2005), pp. 17–141. doi:10.1109/PBG.2005.194059. 6

[BLL19] BUJANCA M., LUJÁN M., LENNOX B.: Fullfusion: A framework for semantic reconstruction of dynamic scenes. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019). 19, 20

[BLL22] BUJANCA M., LENNOX B., LUJÁN M.: Acefusion-accelerated and energy-efficient semantic 3d reconstruction of dynamic scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2022), IEEE, pp. 11063–11070. 19, 20

[BLS*21] BI S., LOMBARDI S., SAITO S., SIMON T., WEI S.-E., MCPHAIL K., RAMAMOORTHI R., SHEIKH Y., SARAGIH J.: Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (ToG) 40*, 4 (2021), 1–15. 27

[Blu67] BLUM H.: A transformation for extracting new descriptors of shape. In *Models for Perception of Speech and Visual Form*, Wathen-Dunn W., (Ed.). MIT Press, Cambridge, MA, 1967. 5

[BMC*24] BHATTACHARYA A., MADAAN R., CLADERA F., VEMPRALA S., BONATTI R., DANIILIDIS K., KAPOOR A., KUMAR V., MATNI N., GUPTA J. K.: Evdnerf: Reconstructing event data with dynamic neural radiance fields. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (January 2024), pp. 5846–5855. 14, 28

[BMV*22] BARRON J. T., MILDENHALL B., VERBIN D., SRINIVASAN P. P., HEDMAN P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5470–5479. 21

[BPP*23] BAHMANI S., PARK J. J., PASCHALIDOU D., TANG H., WETZSTEIN G., GUIBAS L., GOOL L. V., TIMOFTE R.: 3d-aware video generation. *Transactions on Machine Learning Research (TMLR)* (2023). URL: https://openreview.net/forum?id=SwlfyDq6B3. 12, 26, 27

[BPZ*20] BOZIC A., PALAFOX P., ZOLLHÖFER M., DAI A., THIES J., NIESSNER M.: Neural non-rigid tracking. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020), vol. 33, pp. 18727–18737. 17, 26

[BPZ*21] BOZIC A., PALAFOX P., ZOLLHOFER M., THIES J., DAI A., NIESSNER M.: Neural deformation graphs for globally-consistent non-rigid reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 1450–1459. 15

[BRPMB17] BOGO F., ROMERO J., PONS-MOLL G., BLACK M. J.: Dynamic FAUST: Registering human bodies in motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). 25

[BSR*24] BAHMANI S., SKOROKHODOV I., RONG V., WETZSTEIN G., GUIBAS L., WONKA P., TULYAKOV S., PARK J. J., TAGLIASACCHI A., LINDELL D. B.: 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 1, 27

[BXP*22] BHATNAGAR B. L., XIE X., PETROV I., SMINCHISESCU C., THEOBALT C., PONS-MOLL G.: Behave: Dataset and method for tracking human object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (jun 2022). 25

[BZH*23] BHARADWAJ S., ZHENG Y., HILLIGES O., BLACK M. J., ABREVAYA V. F.: Flare: Fast learning of animatable and relightable mesh avatars. *ACM Transactions on Graphics (ToG) 42* (Dec. 2023), 15. doi:https://doi.org/10.1145/3618401. 23

[BZTN20] BOZIC A., ZOLLHOFER M., THEOBALT C., NIESSNER M.: Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 7002–7012. 17, 25, 26

[CB22] CHANG H., BOULARIAS A.: Scene-level tracking and reconstruction without object priors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2022), IEEE, pp. 3785–3792. 1, 17, 19, 20

[CC70] CARROLL J. D., CHANG J.-J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika 35*, 3 (1970), 283–319. 4

[CCP*23] CHOE J., CHOY C., PARK J., KWEON I. S., ANANDKUMAR A.: Spacetime surface regularization for neural dynamic scene reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 17871–17881. 13, 14

[CCZ*23] CHEN Y., CHEN Z., ZHANG C., WANG F., YANG X., WANG Y., CAI Z., YANG L., LIU H., LIN G.: Gaussianeditor: Swift and controllable 3d editing with gaussian splatting, 2023. arXiv:2311.14521. 27

[CFF*22] CAI H., FENG W., FENG X., WANG Y., ZHANG J.: Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022), vol. 35, pp. 967–981. 2, 9, 13, 14

[CGC*23] CHEN H., GU J., CHEN A., TIAN W., TU Z., LIU L., SU H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 29

[CJ23] CAO A., JOHNSON J.: Hexplane: A fast representation for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 130–141. 2, 5, 9, 11, 13, 14

[CJH*22] CHONG BAO AND BANGBANG YANG, JUNYI Z., HUJUN B., YINDA Z., ZHAOPENG C., GUOFENG Z.: Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision (ECCV)* (2022). 21, 23

[CJS*23] CHEN X., JIANG T., SONG J., RIETMANN M., GEIGER A., BLACK M. J., HILLIGES O.: Fast-snarf: A fast deformer for articulated neural fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 45*, 10 (2023), 11796–11809. 21, 23

[CL22] CHEN Z., LIU Z.: Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 606–623. 27

[CLC*22a] CHAN E. R., LIN C. Z., CHAN M. A., NAGANO K., PAN B., DE MELLO S., GALLO O., GUIBAS L. J., TREMBLAY J., KHAMIS S., ET AL.: Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 16123–16133. 5

[CLC*22b] CHAN E. R., LIN C. Z., CHAN M. A., NAGANO K., PAN B., MELLO S. D., GALLO O., GUIBAS L., TREMBLAY J., KHAMIS S., KARRAS T., WETZSTEIN G.: Efficient geometry-aware 3D generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 12

[CLF*23] CHEN D., LU H., FELDMANN I., SCHREER O., EISERT P.: Dynamic multi-view scene reconstruction using neural implicit surface. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), IEEE, pp. 1–5. 13, 14

[CLI*20] CHABRA R., LENSSEN J. E., ILG E., SCHMIDT T., STRAUB J., LOVEGROVE S., NEWCOMBE R.: Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision (ECCV)* (2020). 5

[CLL23] CHEN H., LI C., LEE G. H.: Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance, 2023. arXiv:2312.00846. 27

[CLW23] CHEN J.-K., LYU J., WANG Y.-X.: NeuralEditor: Editing neural radiance fields via manipulating point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 21, 23

[CLZ*22] CHU M., LIU L., ZHENG Q., FRANZ E., SEIDEL H.-P., THEOBALT C., ZAYER R.: Physics informed neural fields for smoke reconstruction with sparse data. *ACM Transactions on Graphics (ToG) 41*, 4 (2022), 1–14. 6, 14, 28

[CLZ*23] CHEN X., LIU J., ZHAO H., ZHOU G., ZHANG Y.-Q.: Nerrf: 3d reconstruction and view synthesis for transparent and specular objects with neural refractive-reflective fields, 2023. arXiv:2309.13039. 27

[CNC*23] CHAN E. R., NAGANO K., CHAN M. A., BERGMAN A. W., PARK J. J., LEVY A., AITTALA M., MELLO S. D., KARRAS T., WETZSTEIN G.: GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 27

[CPPFJ*21] CASILLAS-PEREZ D., PIZARRO D., FUENTES-JIMENEZ D., MAZO M., BARTOLI A.: The isowarp: the template-based visual geometry of isometric surfaces. *International Journal of Computer Vision (IJCV) 129*, 7 (2021), 2194–2222. 16

[CRBD18] CHEN T. Q., RUBANOVA Y., BETTENCOURT J., DUVENAUD D.: Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)* (2018), Bengio S., Wallach H. M., Larochelle H., Grauman K., Cesa-Bianchi N., Garnett R., (Eds.), pp. 6572–6583. 14

[CRG*23] CHANG H., RAMESH D. M., GENG S., GAN Y., BOULARIAS A.: Mono-star: Mono-camera scene-level tracking and reconstruction. In *IEEE International Conference on Robotics and Automation (ICRA)* (2023), pp. 820–826. doi:10.1109/ICRA48891.2023.10160778. 4, 8, 17, 19, 20

[CT22] CHEN Q.-A., TSUKADA A.: Flow supervised neural radiance fields for static-dynamic decomposition. In *IEEE International Conference on Robotics and Automation (ICRA)* (2022), IEEE, pp. 10641–10647. 19

[CTM*21] CARON M., TOUVRON H., MISRA I., JÉGOU H., MAIRAL J., BOJANOWSKI P., JOULIN A.: Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021). 10, 12, 26

[CTS*22] CHEN H.-Y., TRETSCHK E., STUYCK T., KADLECEK P., KAVAN L., VOUGA E., LASSNER C.: Virtual elastic objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 15827–15837. 16

[CWL23] CHEN Z., WANG F., LIU H.: Text-to-3d using gaussian splatting, 2023. arXiv:2309.16585. 27

[CXG*22] CHEN A., XU Z., GEIGER A., YU J., SU H.: Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)* (2022). 2, 4, 14

[CZ19] CHEN Z., ZHANG H.: Learning implicit fields for generative shape modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019). 5, 12

[CZB*21] CHEN X., ZHENG Y., BLACK M. J., HILLIGES O., GEIGER A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 11594–11604. 9, 23

[DHL*23] DRIESS D., HUANG Z., LI Y., TEDRAKE R., TOUSSAINT M.: Learning multi-object dynamics with compositional neural radiance fields. In *Conference on Robot Learning (CoRL)* (2023), PMLR, pp. 1755–1768. 8, 19, 20, 22, 28

[DL08] DE LATHAUWER L.: Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications 30*, 3 (2008), 1033–1066. 4

[DLJ*20] DENG B., LEWIS J. P., JERUZALSKI T., PONS-MOLL G., HINTON G., NOROUZI M., TAGLIASACCHI A.: Nasa neural articulated shape approximation. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 612–628. 23

[DMGT23] DABRAL R., MUGHAL M. H., GOLYANIK V., THEOBALT C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 9760–9770. 28

[DSJ*21] DABRAL R., SHIMADA S., JAIN A., THEOBALT C., GOLYANIK V.: Gravity-aware monocular 3d human-object reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021). 28

[DWY*24] DAS D., WEWER C., YUNUS R., ILG E., LENSSEN J. E.: Neural parametric gaussians for monocular non-rigid object reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 2, 13, 18, 24, 27

[DZY*21] DU Y., ZHANG Y., YU H.-X., TENENBAUM J. B., WU J.: Neural radiance flow for 4d view synthesis and video processing. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), IEEE Computer Society, pp. 14304–14314. 13

[EMS*23] ERKOÇ Z., MA F., SHAN Q., NIESSNER M., DAI A.: Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 2, 26, 27, 29

[FAR*23] FENG B. Y., ALZAYER H., RUBINSTEIN M., FREEMAN W. T., HUANG J.-B.: 3d motion magnification: Visualizing subtle motions from time-varying radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 9837–9846. 15

[FBD*19] FLYNN J., BROXTON M., DEBEVEC P., DUVALL M., FYFFE G., OVERBECK R., SNAVELY N., TUCKER R.: Deepview: View synthesis with learned gradient descent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019). 18

[FJPCP*21] FUENTES-JIMENEZ D., PIZARRO D., CASILLAS-PEREZ D., COLLINS T., BARTOLI A.: Texture-generic deep shape-from-template. *IEEE Access 9* (2021), 75211–75230. 16

[FKMW*23] FRIDOVICH-KEIL S., MEANTI G., WARBURG F. R., RECHT B., KANAZAWA A.: K-planes: Explicit radiance fields in space, time, and appearance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12479–12488. 1, 5, 13, 14

[FTC*22] FRIDOVICH-KEIL AND YU, TANCIK M., CHEN Q., RECHT B., KANAZAWA A.: Plenoxels: Radiance fields without neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 7

[FWZ*24] FANG J., WANG J., ZHANG X., XIE L., TIAN Q.: Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 27

[FYW*22] FANG J., YI T., WANG X., XIE L., ZHANG X., LIU W., NIESSNER M., TIAN Q.: Fast dynamic radiance fields with time-aware neural voxels. In *ACM SIGGRAPH Asia* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. doi:10.1145/3550469.3555383. 2, 13, 15

[GAXS22] GÄRTNER E., ANDRILUKA M., XU H., SMINCHISESCU C.: Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 13106–13115. 28

[GB22] GRASSHOF S., BRANDT S. S.: Tensor-based non-rigid structure from motion. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), pp. 3011–3020. 16

[GDWY22] GUAN S., DENG H., WANG Y., YANG X.: Neurofluid: Fluid dynamics grounding with particle-driven neural radiance fields. In *International Conference on Machine Learning (ICML)* (2022), PMLR, pp. 7919–7929. 6, 16

[GJST20] GOLYANIK V., JONAS A., STRICKER D., THEOBALT C.: Intrinsic dynamic shape prior for dense non-rigid structure from motion. In *International Conference on 3D Vision (3DV)* (2020). 16

[GKE*22] GARBIN S. J., KOWALSKI M., ESTELLERS V., SZYMANOWICZ S., REZAEIFAR S., SHEN J., JOHNSON M., VALENTIN J.: Volumorph: Realtime, controllable and generalisable animation of volumetric representations, 2022. arXiv:2208.00949. 22

[GL24] GUÉDON A., LEPETIT V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 27

[GLD*19] GUO K., LINCOLN P., DAVIDSON P., BUSCH J., YU X., WHALEN M., HARVEY G., ORTS-ESCOLANO S., PANDEY R., DOURGARIAN J., ET AL.: The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG) 38*, 6 (2019), 1–19. 27

[GLT*22] GAO H., LI R., TULSIANI S., RUSSELL B., KANAZAWA A.: Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 8, 28

[GNK18] GÜLER R. A., NEVEROVA N., KOKKINOS I.: Densepose: Dense human pose estimation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018). 10, 12

[GPY*23] GUO H., PENG S., YAN Y., MOU L., SHEN Y., BAO H., ZHOU X.: Compact neural volumetric video representations with dynamic codebooks. In *Advances in Neural Information Processing Systems (NeurIPS)* (2023), Oh A., Neumann T., Globerson A., Saenko K., Hardt M., Levine S., (Eds.), vol. 36, Curran Associates, Inc., pp. 75884–75895. URL: https://proceedings.

neurips.cc/paper_files/paper/2023/file/ef63b00ad8475605b2eaf520747f61d4-Paper-Conference.pdf. 14

[GRLM*21] GÓMEZ-RODRÍGUEZ J. J., LAMARCA J., MORLANA J., TARDÓS J. D., MONTIEL J. M.: Sd-defslam: Semi-direct monocular slam for deformable and intracorporeal scenes. In *IEEE International Conference on Robotics and Automation (ICRA)* (2021), IEEE, pp. 5170–5177. 17

[GSD*23] GUO X., SUN J., DAI Y., CHEN G., YE X., TAN X., DING E., ZHANG Y., WANG J.: Forward flow for novel view synthesis of dynamic scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 16022–16033. 9, 10, 13, 15

[GSKH21] GAO C., SARAF A., KOPF J., HUANG J.-B.: Dynamic view synthesis from dynamic monocular video. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021). 2, 10, 13, 19

[GT18] GAO W., TEDRAKE R.: Surfelwarp: Efficient non-volumetric single view dynamic reconstruction. In *Robotics: Science and Systems (RSS)* (2018). 4, 17

[Har96] HART J. C.: Sphere tracing: a geometric method for the antialiased ray tracing of implicit surfaces. *The Visual Computer 12*, 10 (Dec. 1996), 527–545. 6

[HAWG08] HUANG Q.-X., ADAMS B., WICKE M., GUIBAS L. J.: Non-rigid registration under isometric deformations. *Computer Graphics Forum 27*, 5 (2008), 1449–1457. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2008.01285.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2008.01285.x, doi:https://doi.org/10.1111/j.1467-8659.2008.01285.x. 18

[HHS*21] HUANG Q., HUANG X., SUN B., ZHANG Z., JIANG J., BAJAJ C.: Arapreg: An as-rigid-as possible regularization loss for learning deformable shape generators. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5815–5825. 27

[HIZ*23] HEPPERT N., IRSHAD M. Z., ZAKHAROV S., LIU K., AMBRUS R. A., BOHG J., VALADA A., KOLLAR T.: Carto: Category and joint agnostic reconstruction of articulated objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 21201–21210. 26

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020), Larochelle H., Ranzato M., Hadsell R., Balcan M., Lin H., (Eds.). 12, 29

[HJH*22] HUANG D., JI X., HE X., SUN J., HE T., SHUAI Q., OUYANG W., ZHOU X.: Reconstructing hand-held objects from monocular video. In *ACM SIGGRAPH Asia* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. doi:10.1145/3550469.3555401. 28

[HJZ23] HSU C.-C., JIANG Z., ZHU Y.: Ditto in the house: Building articulation models of indoor scenes through interactive perception. In *IEEE International Conference on Robotics and Automation (ICRA)* (2023). 26

[HMC*21] HEIDEN E., MILLARD D., COUMANS E., SHENG Y., SUKHATME G. S.: Neuralsim: Augmenting differentiable simulators with neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)* (2021), IEEE, pp. 9474–9481. 11

[HSW*20] HU Y., SCHNEIDER T., WANG B., ZORIN D., PANOZZO D.: Fast tetrahedral meshing in the wild. *ACM Transactions on Graphics (ToG) 39*, 4 (2020), 117–1. 22

[HTE*23] HAQUE A., TANCIK M., EFROS A., HOLYNSKI A., KANAZAWA A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 28

[HTLH20] HUANG H.-P., TSENG H.-Y., LEE H.-Y., HUANG J.-B.: Semantic view synthesis. In *European Conference on Computer Vision (ECCV)* (2020), pp. 592–608. 18

[HWYS21] HU Y.-T., WANG J., YEH R. A., SCHWING A. G.: SAIL-VOS 3D: A Synthetic Dataset and Baselines for Object Detection and 3D Mesh Reconstruction from Video Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 25

[HYZ*22] HU H., YI X., ZHANG H., YONG J.-H., XU F.: Physical interaction: Reconstructing hand-object interactions with physics. In *ACM SIGGRAPH Asia* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. doi:10.1145/3550469.3555421. 28

[HZ03] HARTLEY R., ZISSERMAN A.: *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[ICN*23] IQBAL U., CALISKAN A., NAGANO K., KHAMIS S., MOLCHANOV P., KAUTZ J.: Rana: Relightable articulated neural avatars. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 23142–23153. 27

[IYOK20] ISOGAWA M., YUAN Y., O'TOOLE M., KITANI K. M.: Optical non-line-of-sight physics-based 3d human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). 28

[IZN*16] INNMANN M., ZOLLHÖFER M., NIESSNER M., THEOBALT C., STAMMINGER M.: VolumeDeform: Real-time Volumetric Non-rigid Reconstruction. In *European Conference on Computer Vision (ECCV)* (2016). 17

[JHS*23] JOHNSON E., HABERMANN M., SHIMADA S., GOLYANIK V., THEOBALT C.: Unbiased 4d: Monocular 4d reconstruction with a neural deformation model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)* (2023), pp. 6597–6606. 13

[JHTG20] JIANG C., HUANG J., TAGLIASACCHI A., GUIBAS L. J.: Shapeflow: Learnable deformation flows among 3d shapes. *Advances in Neural Information Processing Systems (NeurIPS) 33* (2020), 9745–9757. 27

[JJS*22] JIANG Y., JIANG S., SUN G., SU Z., GUO K., WU M., YU J., XU L.: Neuralhofusion: Neural volumetric rendering under human-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 6155–6165. 28

[JLCN21] JAIN A., LIOUTIKOV R., CHUCK C., NIEKUM S.: Screwnet: Category-independent articulation model estimation from depth images using screw theory. In *IEEE International Conference on Robotics and Automation (ICRA)* (2021), IEEE, pp. 13670–13677. 5

[JLX*23a] JIN H., LIU I., XU P., ZHANG X., HAN S., BI S., ZHOU X., XU Z., SU H.: Tensoir: Tensorial inverse rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 165–174. 5

[JLX*23b] JIN H., LIU I., XU P., ZHANG X., HAN S., BI S., ZHOU X., XU Z., SU H.: Tensoir: Tensorial inverse rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 165–174. 27

[JMD*07] JOSHI P., MEYER M., DEROSE T., GREEN B., SANOCKI T.: Harmonic coordinates for character articulation. *ACM Transactions on Graphics (ToG) 26*, 3 (2007), 71–es. 24

[JSJ*18] JIANG H., SUN D., JAMPANI V., YANG M., LEARNED-MILLER E. G., KAUTZ J.: Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), Computer Vision Foundation / IEEE Computer Society, pp. 9000–9008. 13

[JST*16] JIANG C., SCHROEDER C., TERAN J., STOMAKHIN A., SELLE A.: The material point method for simulating continuum materials. In *ACM SIGGRAPH 2016 Courses*. 2016, pp. 1–52. 8

[JSW23] JU T., SCHAEFER S., WARREN J.: Mean value coordinates for closed triangular meshes. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 223–228. 24

[JYS*22] JIANG W., YI K. M., SAMEI G., TUZEL O., RANJAN A.: Neuman: Neural human radiance field from a single video. In *European*

Conference on Computer Vision (ECCV) (2022), Springer, pp. 402–418. 23

[JYS*23] JIANG Y., YAO K., SU Z., SHEN Z., LUO H., XU L.: Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 595–605. 28

[JZW*21] JIANG B., ZHANG Y., WEI X., XUE X., FU Y.: Learning compositional representation for 4d captures with neural ode. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5340–5350. 25

[KAZ*23] KOLOTOUROS N., ALLDIECK T., ZANFIR A., BAZAVAN E. G., FIERARU M., SMINCHISESCU C.: Dreamhuman: Animatable 3d avatars from text, 2023. arXiv:2306.09329. 28

[KGC*24] KAPPEL M., GOLYANIK V., CASTILLO S., THEOBALT C., MAGNOR M.: Fast non-rigid radiance fields from monocularized data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2024). 15

[KGY*22] KUNDU A., GENOVA K., YIN X., FATHI A., PANTOFARU C., GUIBAS L. J., TAGLIASACCHI A., DELLAERT F., FUNKHOUSER T.: Panoptic neural fields: A semantic object-aware neural scene representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 12871–12881. 19

[KKG*23] KERR J., KIM C. M., GOLDBERG K., KANAZAWA A., TANCIK M.: Lerf: Language embedded radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 19729–19739. 28

[KKJ*23] KEETHA N., KARHADE J., JATAVALLABHULA K. M., YANG G., SCHERER S., RAMANAN D., LUITEN J.: Splatam: Splat, track & map 3d gaussians for dense rgb-d slam, 2023. arXiv:2312.02126. 27

[KKK*23] KUAI T., KARTHIKEYAN A., KANT Y., MIRZAEI A., GILITSCHENSKI I.: Camm: Building category-agnostic and animatable 3d models from monocular videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 6586–6596. 1, 5, 21, 24, 26

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG) 42*, 4 (2023), 1–14. 2, 6, 18, 27

[KMP07] KILIAN M., MITRA N. J., POTTMANN H.: Geometric modeling in shape space. *ACM Transactions on Graphics (ToG) 26*, 3 (jul 2007), 64–es. URL: https://doi.org/10.1145/1276377.1276457, doi:10.1145/1276377.1276457. 18

[KMS22] KOBAYASHI S., MATSUMOTO E., SITZMANN V.: Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 20

[KTE*22] KAIRANDA N., TRETSCHK E., ELGHARIB M., THEOBALT C., GOLYANIK V.: f-sft: Shape-from-template with a physics-based deformation model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 3948–3958. 11, 16, 28

[KVG22] KUMAR S., VAN GOOL L.: Organic priors in non-rigid structure from motion. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 71–88. 16

[KVNM23] KARNEWAR A., VEDALDI A., NOVOTNY D., MITRA N.: Holodiffusion: Training a 3D diffusion model using 2D images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 29

[KY22] KIM B., YE J. C.: Diffusion deformable model for 4d temporal medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2022), Springer, pp. 539–548. 26, 27

[KYK*22] KANIA K., YI K. M., KOWALSKI M., TRZCIŃSKI T., TAGLIASACCHI A.: Conerf: Controllable neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 18623–18632. 21, 22

[LBX*22] LI J., BIAN S., XU C., LIU G., YU G., LU C.: D &d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision (ECCV)* (2022). 28

[LC87] LORENSEN W. E., CLINE H. E.: Marching cubes: A high resolution 3d surface construction algorithm. In *ACM SIGGRAPH* (New York, NY, USA, 1987), SIGGRAPH '87, Association for Computing Machinery, p. 163–169. URL: https://doi.org/10.1145/37401.37422, doi:10.1145/37401.37422. 6

[LCBH*23] LIPMAN Y., CHEN R. T. Q., BEN-HAMU H., NICKEL M., LE M.: Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)* (2023). 12, 29

[LCK*22] LAHOUD J., CAO J., KHAN F. S., CHOLAKKAL H., ANWER R. M., KHAN S., YANG M.-H.: 3d vision with transformers: A survey, 2022. arXiv:2208.04309. 6

[LCLX24] LI Z., CHEN Z., LI Z., XU Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 18

[LCM*22] LIU J.-W., CAO Y.-P., MAO W., ZHANG W., ZHANG D. J., KEPPO J., SHAN Y., QIE X., SHOU M. Z.: Devrf: Fast deformable voxel radiance fields for dynamic scenes. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022), vol. 35, pp. 36762–36775. 13, 15

[LCQ*21] LU F., CHEN G., QU S., LI Z., LIU Y., KNOLL A.: Pointinet: Point cloud frame interpolation network. In *AAAI Conference on Artificial Intelligence* (2021), vol. 35, pp. 2251–2259. 14

[LCW*24] LIU J.-W., CAO Y.-P., WU J. Z., MAO W., GU Y., ZHAO R., KEPPO J., SHAN Y., SHOU M. Z.: Dynvideo-e: Harnessing dynamic nerf for large-scale motion- and view-change human-centric video editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 1, 21

[LCY*23] LIU J.-W., CAO Y.-P., YANG T., XU Z., KEPPO J., SHAN Y., QIE X., SHOU M. Z.: Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 18483–18494. 21, 23

[LDR*22] LUGMAYR A., DANELLJAN M., ROMERO A., YU F., TIMOFTE R., GOOL L. V.: Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 12

[LDS*23] LEI J., DENG C., SHEN B., GUIBAS L., DANIILIDIS K.: Nap: Neural 3d articulated object prior. In *Advances in Neural Information Processing Systems (NeurIPS)* (2023). 5, 26, 27, 28

[LG20] LI C., GUO X.: Topology-change-aware volumetric fusion for dynamic scene reconstruction. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 258–274. 17

[LGM*23] LIU Y.-L., GAO C., MEULEMAN A., TSENG H.-Y., SARAF A., KIM C., CHUANG Y.-Y., KOPF J., HUANG J.-B.: Robust dynamic radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 13–23. 8, 11, 19, 28

[LGO*23] LAZOVA V., GUZOV V., OLSZEWSKI K., TULYAKOV S., PONS-MOLL G.: Control-nerf: Editable feature volumes for scene rendering and manipulation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2023), pp. 4340–4350. 21

[LGTK23] LI R., GAO H., TANCIK M., KANAZAWA A.: Nerfacc: Efficient sampling accelerates nerfs. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (Los Alamitos, CA, USA, oct 2023), IEEE Computer Society, pp. 18491–18500. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01699, doi:10.1109/ICCV51070.2023.01699. 7

[LGW23] LIU S., GUPTA S., WANG S.: Building rearticulable models for arbitrary 3d objects from 4d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 21138–21147. 21, 24

[LGZL*20] LIU L., GU J., ZAW LIN K., CHUA T.-S., THEOBALT C.: Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020), vol. 33, pp. 15651–15663. 21

[LHR*21] LIU L., HABERMANN M., RUDNEV V., SARKAR K., GU J., THEOBALT C.: Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (ToG) 40*, 6 (2021), 1–16. 23

[LKLR24] LUITEN J., KOPANAS G., LEIBE B., RAMANAN D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *International Conference on 3D Vision (3DV)* (2024). 2, 10, 13, 18, 21, 27, 28

[LL23] LIN A., LI J.: Dynamic appearance particle neural radiance field, 2023. arXiv:2310.07916. 16, 20

[LLCL19] LIU S., LI T., CHEN W., LI H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019). 4

[LLK19] LIANG J., LIN M., KOLTUN V.: Differentiable cloth simulation for inverse problems. *Advances in Neural Information Processing Systems (NeurIPS) 32* (2019). 11

[LLL*24] LI Z., LITVAK D., LI R., ZHANG Y., JAKAB T., RUPPRECHT C., WU S., VEDALDI A., WU J.: Learning the 3d fauna of the web. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 25, 26

[LLM*23] LIANG Y., LAIDLAW E., MEYEROWITZ A., SRIDHAR S., TOMPKIN J.: Semantic attention flow fields for dynamic scene decomposition. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 1, 19, 20

[LLS*22] LI Y., LI S., SITZMANN V., AGRAWAL P., TORRALBA A.: 3d neural scene representations for visuomotor control. In *Conference on Robot Learning (CoRL)* (2022), PMLR, pp. 112–123. 21, 22

[LM18] LAMARCA J., MONTIEL J. M. M.: Camera tracking for slam in deformable maps. In *European Conference on Computer Vision Workshop (ECCVW)* (2018). 9, 16

[LMAS23] LIU J., MAHDAVI-AMIRI A., SAVVA M.: PARIS: Part-level reconstruction and motion analysis for articulated objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 19, 20

[LMM*22] LI G., MEKA A., MUELLER F., BUEHLER M. C., HILLIGES O., BEELER T.: Eyenerf: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (ToG) 41*, 4 (2022), 1–16. 27

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (ToG) 34*, 6 (oct 2015). URL: https://doi.org/10.1145/2816795.2818013, doi:10.1145/2816795.2818013. 6, 23

[LNSW21] LI Z., NIKLAUS S., SNAVELY N., WANG O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 6498–6508. 2, 10, 13, 19

[LPBM20] LAMARCA J., PARASHAR S., BARTOLI A., MONTIEL J.: Defslam: Tracking and mapping of deforming scenes from monocular sequences. *IEEE Transactions on Robotics (T-RO) 37*, 1 (2020), 291–303. 16, 17

[LPD08] LICHTSTEINER P., POSCH C., DELBRUCK T.: A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits 43*, 2 (2008), 566–576. 3

[LPL*22] LIU Y., PENG S., LIU L., WANG Q., WANG P., THEOBALT C., ZHOU X., WANG W.: Neural rays for occlusion-aware image-based rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 15

[LPX*22] LIN H., PENG S., XU Z., YAN Y., SHUAI Q., BAO H., ZHOU X.: Efficient neural radiance fields for interactive free-viewpoint video. In *ACM SIGGRAPH Asia* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. doi:10.1145/3550469.3555376. 15

[LPX*23] LIN H., PENG S., XU Z., XIE T., HE X., BAO H., ZHOU X.: High-fidelity and real-time novel view synthesis for dynamic scenes. In *ACM SIGGRAPH Asia* (New York, NY, USA, 2023), SA '23, Association for Computing Machinery. doi:10.1145/3610548.3618142. 13, 16

[LQC*22] LI X., QIAO Y.-L., CHEN P. Y., JATAVALLABHULA K. M., LIN M., JIANG C., GAN C.: Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *International Conference on Learning Representations (ICLR)* (2022). 13, 16, 28

[LQG19] LIU X., QI C. R., GUIBAS L. J.: Flownet3d: Learning scene flow in 3d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019). 12, 26

[LSS*22] LI Z., SHIMADA S., SCHIELE B., THEOBALT C., GOLYANIK V.: Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *International Conference on 3D Vision (3DV)* (2022). 28

[LSW*22] LI L., SHEN Z., WANG Z., SHEN L., TAN P.: Streaming radiance fields for 3d video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022), vol. 35, pp. 13485–13498. 13, 15

[LSZ*22] LI T., SLAVCHEVA M., ZOLLHOEFER M., GREEN S., LASSNER C., KIM C., SCHMIDT T., LOVEGROVE S., GOESELE M., NEWCOMBE R., ET AL.: Neural 3d video synthesis from multi-view video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5521–5531. 10, 13

[LTT*21] LI Y., TAKEHARA H., TAKETOMI T., ZHENG B., NIESSNER M.: 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 12706–12716. 10, 17, 25, 26

[LTV*22] LI R., TANKE J., VO M., ZOLLHÖFER M., GALL J., KANAZAWA A., LASSNER C.: Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 419–436. 21, 23

[LWC*23a] LI Z., WANG Q., COLE F., TUCKER R., SNAVELY N.: Dynibar: Neural dynamic image-based rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4273–4284. 5, 7, 10, 13, 15, 16, 28

[LWC*23b] LIN H., WANG Q., CAI R., PENG S., AVERBUCH-ELOR H., ZHOU X., SNAVELY N.: Neural scene chronology. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 20752–20761. 13

[LWP*24] LEI J., WANG Y., PAVLAKOS G., LIU L., DANIILIDIS K.: Gart: Gaussian articulated template models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 23

[LWS*13] LI G., WU C., STOLL C., LIU Y., VARANASI K., DAI Q., THEOBALT C.: Capturing relightable human performances under general uncontrolled illumination. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 275–284. 27

[LWVH*23] LIU R., WU R., VAN HOORICK B., TOKMAKOV P., ZAKHAROV S., VONDRICK C.: Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 9298–9309. 21

[LXDS20] LI Z., XIAN W., DAVIS A., SNAVELY N.: Crowdsampling the plenoptic function. In *European Conference on Computer Vision (ECCV)* (2020). 18

[LXL*21] LIN K.-E., XIAO L., LIU F., YANG G., RAMAMOORTHI R.: Deep 3d mask volume for view synthesis of dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 1749–1758. 18

[LZG18] LI C., ZHAO Z., GUO X.: Articulatedfusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera. In *European Conference on Computer Vision (ECCV)* (2018), pp. 317–332. 19, 20

[LZNH20] LI Y., ZHANG T., NAKAMURA Y., HARADA T.: Split-fusion: Simultaneous tracking and mapping for non-rigid scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), IEEE, pp. 5128–5134. 19, 20

[LZYX22] LIN W., ZHENG C., YONG J.-H., XU F.: Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 1736–1745. 2, 9, 13, 17, 26, 27

[MA22] MOHAMED M., AGAPITO L.: Gnpm: Geometric-aware neural parametric models. In *International Conference on 3D Vision (3DV)* (2022), IEEE, pp. 166–175. 25, 26

[MA24] MOHAMED M., AGAPITO L.: Dynamicsurf: Dynamic neural rgb-d surface reconstruction with an optimizable feature grid. In *International Conference on 3D Vision (3DV)* (2024). 15

[MBW*23] MUNDRA A., B R M., WANG J., HABERMANN M., THEOBALT C., ELGHARIB M.: Livehand: Real-time and photorealistic neural hand rendering. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023). 23

[MDB*19] MUELLER F., DAVIS M., BERNARD F., SOTNYCHENKO O., VERSCHOOR M., OTADUY M. A., CASAS D., THEOBALT C.: Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG) 38*, 4 (2019), 1–13. 28

[Mea80] MEAGHER D.: *Octree Encoding: A New Technique for the Representation, Manipulation and Display of Arbitrary 3-D Objects by Computer*. Tech. rep., 10 1980. 4

[MESK22] MÜLLER T., EVANS A., SCHIED C., KELLER A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG) 41*, 4 (2022), 102:1–102:15. 2, 11, 15

[MKRV23] MELAS-KYRIAZI L., RUPPRECHT C., VEDALDI A.: PC^2 projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 29

[MLLS23] MA L., LI X., LIAO J., SANDER P. V.: 3d video loops from asynchronous input. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 310–320. 18, 19

[MLR*24] MILLERDURAI C., LUVIZON D., RUDNEV V., JONAS A., WANG J., THEOBALT C., GOLYANIK V.: 3d pose estimation of two interacting hands from a monocular event camera. In *International Conference on 3D Vision (3DV)* (2024). 3, 28

[MNH23] MAHESHWARI S., NARAIN R., HEBBALAGUPPE R.: Transfer4d: A framework for frugal motion capture and deformation transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12836–12846. 21, 24

[MON*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3d reconstruction in function space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019). 5, 12

[MPCVG23] MA Q., PAUDEL D. P., CHHATKULI A., VAN GOOL L.: Deformable neural radiance fields using rgb and event cameras. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 3590–3600. 1, 13, 14, 28

[MPE*23] MENDIRATTA M., PAN X., ELGHARIB M., TEOTIA K., R M. B., TEWARI A., GOLYANIK V., KORTYLEWSKI A., THEOBALT C.: Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Transactions on Graphics (ToG) 42*, 6 (dec 2023). URL: https://doi.org/10.1145/3618368, doi:10.1145/3618368. 28

[MQK*21] MU J., QIU W., KORTYLEWSKI A., YUILLE A., VASCONCELOS N., WANG X.: A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021). 26

[MSL*24] MENAPACE W., SIAROHIN A., LATHUILIÈRE S., ACHLIOPTAS P., GOLYANIK V., TULYAKOV S., RICCI E.: Promptable game models: Text-guided game simulation via masked diffusion models. *ACM Transactions on Graphics (ToG) 43*, 2 (jan 2024). URL: https://doi.org/10.1145/3635705, doi:10.1145/3635705. 26, 27

[MSP*23] MÜLLER N., SIDDIQUI Y., PORZI L., BULO S. R., KONTSCHIEDER P., NIESSNER M.: Diffrf: Rendering-guided 3d radiance field diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 11, 29

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)* (2020). 2, 7, 12, 27

[MSY10] MOUNTNEY P., STOYANOV D., YANG G.-Z.: Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine 27*, 4 (2010), 14–24. doi:10.1109/MSP.2010.936728. 17

[NDS*20] NAIR G. B., DAGA S., SAJNANI R., RAMESH A., ANSARI J. A., JATAVALLABHULA K. M., KRISHNA K. M.: Multi-object monocular slam for dynamic environments. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (2020), IEEE, pp. 651–657. 16

[Nes83] NESTEROV Y. E.: A method for solving the convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$. In *Doklady Akademii Nauk SSSR* (1983), vol. 269, pp. 543–547. 17

[NFS15] NEWCOMBE R. A., FOX D., SEITZ S. M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 343–352. 4, 15, 17

[NG21] NIEMEYER M., GEIGER A.: Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 11453–11464. 18

[NIT*22] NOGUCHI A., IQBAL U., TREMBLAY J., HARADA T., GALLO O.: Watch it move: Unsupervised discovery of 3d joints for re-posing of articulated objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 3677–3687. 1, 2, 19, 20

[NMOG19] NIEMEYER M., MESCHEDER L., OECHSLE M., GEIGER A.: Occupancy flow: 4d reconstruction by learning particle dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5379–5389. 7, 10, 14, 25

[NNS*20] NEVEROVA N., NOVOTNY D., SZAFRANIEC M., KHALIDOV V., LABATUT P., VEDALDI A.: Continuous surface embeddings. *Advances in Neural Information Processing Systems (NeurIPS) 33* (2020), 17258–17270. 12, 26

[NRS*22] NOVOTNY D., ROCCO I., SINHA S., CARLIER A., KERCHENBAUM G., SHAPOVALOV R., SMETANIN N., NEVEROVA N., GRAHAM B., VEDALDI A.: Keytr: keypoint transporter for 3d reconstruction of deformable objects in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5595–5604. 9

[NSLH22] NOGUCHI A., SUN X., LIN S., HARADA T.: Unsupervised learning of efficient geometry-aware neural articulated representations. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 597–614. 23

[NZIS13] NIESSNER M., ZOLLHÖFER M., IZADI S., STAMMINGER M.: Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG) 32*, 6 (2013), 1–11. 4

[OMT*21] OST J., MANNAN F., THUEREY N., KNODT J., HEIDE F.: Neural scene graphs for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 2856–2865. 19

[OPG21] OECHSLE M., PENG S., GEIGER A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021). 6

[PBT*21] PALAFOX P., BOŽIČ A., THIES J., NIESSNER M., DAI A.: Npms: Neural parametric models for 3d deformable shapes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 12695–12705. 11, 12, 24, 25

[PCPM21] PUMAROLA A., CORONA E., PONS-MOLL G., MORENO-NOGUER F.: D-nerf: Neural radiance fields for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), Computer Vision Foundation / IEEE, pp. 10318–10327. 13, 23

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019). 5, 6, 12

[Pho75] PHONG B. T.: Illumination for computer generated pictures. *Communications of the ACM 18*, 6 (jun 1975), 311–317. doi:10.1145/360825.360839. 23

[PJBM22] POOLE B., JAIN A., BARRON J. T., MILDENHALL B.: Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)* (2022). 12, 21, 28

[PK24] PARK B., KIM C.: Point-dynrf: Point-based dynamic radiance fields from a monocular video. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (January 2024), pp. 3171–3181. 16

[PMR*23] PROKUDIN S., MA Q., RAAFAT M., VALENTIN J., TANG S.: Dynamic point fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 2, 9, 11, 13, 18

[PNM*20] PENG S., NIEMEYER M., MESCHEDER L., POLLEFEYS M., GEIGER A.: Convolutional occupancy networks. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 523–540. 5

[PPB17] PARASHAR S., PIZARRO D., BARTOLI A.: Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 40*, 10 (2017), 2442–2454. 17

[PPB21] PARASHAR S., PIZARRO D., BARTOLI A.: Robust isometric non-rigid structure-from-motion. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 44*, 10 (2021), 6409–6423. 14, 16

[PPGT*23] PETITJEAN A., POIRIER-GINTER Y., TEWARI A., CORDONNIER G., DRETTAKIS G.: Modalnerf: Neural modal analysis and synthesis for free-viewpoint navigation in dynamically vibrating scenes. *Computer Graphics Forum 42*, 4 (2023), e14888. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14888, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14888, doi:https://doi.org/10.1111/cgf.14888. 11, 16

[PSB*21] PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., MARTIN-BRUALLA R.: Nerfies: Deformable neural radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), IEEE, pp. 5845–5854. 13

[PSH*21] PARK K., SINHA U., HEDMAN P., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., MARTIN-BRUALLA R., SEITZ S. M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics (ToG) 40*, 6 (dec 2021). 7, 13, 22

[PSJ*23] PARK S., SON M., JANG S., AHN Y. C., KIM J.-Y., KANG N.: Temporal interpolation is all you need for dynamic neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4212–4221. 13, 15, 20

[PSTD22] PALAFOX P., SARAFIANOS N., TUNG T., DAI A.: Spams: Structured implicit parametric models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 12851–12860. 25, 26

[PYG*23] PO R., YIFAN W., GOLYANIK V., ABERMAN K., BARRON J. T., BERMANO A. H., CHAN E. R., DEKEL T., HOLYNSKI A., KANAZAWA A., LIU C. K., LIU L., MILDENHALL B., NIESSNER M., OMMER B., THEOBALT C., WONKA P., WETZSTEIN G.: State of the art on diffusion models for visual computing, 2023. arXiv:2310.07204. 2, 29

[PYL*22] PENG Y., YAN Y., LIU S., CHENG Y., GUAN S., PAN B., ZHAI G., YANG X.: Cagenerf: Cage-based neural radiance field for generalized 3d deformation and animation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022), vol. 35, pp. 31402–31415. 21, 24

[PYS*23] PENG S., YAN Y., SHUAI Q., BAO H., ZHOU X.: Representing volumetric videos as dynamic mlp maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4252–4262. 14

[PZB*19] PROKHOROV D., ZHUKOV D., BARINOVA O., ANTON K., VORONTSOVA A.: Measuring robustness of visual slam. In *International Conference on Machine Vision Applications (MVA)* (2019), IEEE, pp. 1–6. 11

[PZK*24] PANG H., ZHU H., KORTYLEWSKI A., THEOBALT C., HABERMANN M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 23

[PZvBG00] PFISTER H., ZWICKER M., VAN BAAR J., GROSS M.: Surfels: surface elements as rendering primitives. In *ACM SIGGRAPH* (USA, 2000), SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., p. 335–342. URL: https://doi.org/10.1145/344779.344936, doi:10.1145/344779.344936. 6

[PZX*21] PENG S., ZHANG Y., XU Y., WANG Q., SHUAI Q., BAO H., ZHOU X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 9054–9063. 23

[QCZ*23] QU W., CUI Z., ZHANG Y., MENG C., MA C., DENG X., WANG H.: Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 15100–15111. 28

[QGL22] QIAO Y.-L., GAO A., LIN M.: Neuphysics: Editable neural geometry and physics from monocular videos. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022), vol. 35, pp. 12841–12854. 21, 22, 28

[QLZ*24] QIN M., LI W., ZHOU J., WANG H., PFISTER H.: Langsplat: 3d language gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 28

[RBA*19] RAHAMAN N., BARATIN A., ARPIT D., DRAXLER F., LIN M., HAMPRECHT F., BENGIO Y., COURVILLE A.: On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)* (2019), PMLR, pp. 5301–5310. 5

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 10684–10695. 2, 22, 29

[RBSC21] ROSSI A., BARBIERO M., SCREMIN P., CARLI R.: Robust visibility surface determination in object space via plücker coordinates. *Journal of Imaging 7*, 6 (2021). URL: https://www.mdpi.com/2313-433X/7/6/96, doi:10.3390/jimaging7060096. 6

[RBZ*20] REMPE D., BIRDAL T., ZHAO Y., GOJCIC Z., SRIDHAR S., GUIBAS L. J.: Caspr: Learning canonical spatiotemporal point cloud representations. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020), vol. 33, pp. 13688–13701. 14

[RETG23] RUDNEV V., ELGHARIB M., THEOBALT C., GOLYANIK V.: Eventnerf: Neural radiance fields from a single colour event camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 3

[RGW*21] RUDNEV V., GOLYANIK V., WANG J., SEIDEL H.-P., MUELLER F., ELGHARIB M., THEOBALT C.: Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021). 3

[RKA*21] REED A. W., KIM H., ANIRUDH R., MOHAN K. A., CHAMPLEY K., KANG J., JAYASURIYA S.: Dynamic ct reconstruction from limited views with implicit neural representations and parametric motion fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 2258–2268. 14

[RLH*20] RANFTL R., LASINGER K., HAFNER D., SCHINDLER K., KOLTUN V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 44*, 3 (2020), 1623–1637. 10

[RLJ*23] RUIZ N., LI Y., JAMPANI V., PRITCH Y., RUBINSTEIN M., ABERMAN K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 22500–22510. 22

[RMT23] RODRIGUEZ J. J. G., MONTIEL J. M. M., TARDOS J. D.: Nr-slam: Non-rigid monocular slam, 2023. arXiv:2308.04036. 4, 11, 13, 17

[RPZ02] REN L., PFISTER H., ZWICKER M.: Object space ewa surface splatting: A hardware accelerated approach to high quality point rendering. In *Eurographics* (September 2002), vol. 21, pp. 461 – 470. 6

[RSAH23] RAMASINGHE S., SHEVCHENKO V., AVRAHAM G., HENGEL A. V. D.: Blirf: Bandlimited radiance fields for dynamic scene modeling, 2023. arXiv:2302.13543. 14

[RZS21] REN Z., ZHAO X., SCHWING A.: Class-agnostic reconstruction of dynamic objects from videos. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021), vol. 34, pp. 509–522. 12, 25

[SA07] SORKINE O., ALEXA M.: As-rigid-as-possible surface modeling. In *Symposium on Geometry Processing (SGP)* (Goslar, DEU, 2007), SGP '07, Eurographics Association, p. 109–116. 9, 10, 27

[SB12] SIFAKIS E., BARBIC J.: Fem simulation of 3d deformable solids: a practitioner's guide to theory, discretization and model reduction. In *ACM SIGGRAPH 2012 Courses*. 2012, pp. 1–50. 6

[SBCI17] SLAVCHEVA M., BAUST M., CREMERS D., ILIC S.: Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1386–1395. 17

[SBI18] SLAVCHEVA M., BAUST M., ILIC S.: Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 2646–2655. 17

[SBI20] SLAVCHEVA M., BAUST M., ILIC S.: Variational level set evolution for non-rigid 3d reconstruction from a single depth camera. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 43*, 8 (2020), 2838–2850. 17

[SBR22] SU S.-Y., BAGAUTDINOV T., RHODIN H.: Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 107–124. 21, 23

[SBR23] SU S.-Y., BAGAUTDINOV T., RHODIN H.: Npc: Neural point characters from video. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 21, 23

[SCC*23] SONG C., CHEN T., CHEN Y., WEI J., FOO C. S., LIU F., LIN G.: Moda: Modeling deformable 3d objects from casual videos, 2023. arXiv:2304.08279. 21, 24, 26

[SCL*23] SONG L., CHEN A., LI Z., CHEN Z., CHEN L., YUAN J., XU Y., GEIGER A.: Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 29*, 5 (2023), 2732–2742. 2, 19, 20

[SDZ*21] SRINIVASAN P. P., DENG B., ZHANG X., TANCIK M., MILDENHALL B., BARRON J. T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 7495–7504. 27

[SESM22] SUHAIL M., ESTEVES C., SIGAL L., MAKADIA A.: Light field neural rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 15

[SF16] SCHONBERGER J. L., FRAHM J.-M.: Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4104–4113. 3

[SGM*19] SU Y., GOLYANIK V., MINASKAN N., ALI S. A., STRICKER D.: A shape completion component for monocular non-rigid slam. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2019), IEEE, pp. 332–337. 16

[SGP*22] SONG L., GONG X., PLANCHE B., ZHENG M., DOERMANN D., YUAN J., CHEN T., WU Z.: Pref: Predictability regularized neural motion fields. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 664–681. 9, 13

[SGPT23] SHIMADA S., GOLYANIK V., PÉREZ P., THEOBALT C.: Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (ToG) 42*, 6 (2023). 28

[SGTS19] SHIMADA S., GOLYANIK V., THEOBALT C., STRICKER D.: IsMo-GAN: Adversarial learning for monocular non-rigid 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)* (2019). 16

[SGX*21] SHIMADA S., GOLYANIK V., XU W., PÉREZ P., THEOBALT C.: Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (ToG) 40*, 4 (2021). 28

[SGXT20] SHIMADA S., GOLYANIK V., XU W., THEOBALT C.: Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG) 39*, 6 (2020). 28

[SGY*21] SHEN T., GAO J., YIN K., LIU M.-Y., FIDLER S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 4

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 2304–2314. 15

[SJC*22] SHEN B., JIANG Z., CHOY C., SAVARESE S., GUIBAS L. J., ANANDKUMAR A., ZHU Y.: Acid: Action-conditional implicit visual dynamics for deformable object manipulation. In *Robotics: Science and Systems (RSS)* (2022). 21, 22

[SJL*24] SUN J., JIAO H., LI G., ZHANG Z., ZHAO L., XING W.: 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 13, 18

[SP86] SEDERBERG T. W., PARRY S. R.: Free-form deformation of solid geometric models. In *ACM SIGGRAPH* (New York, NY, USA, 1986), SIGGRAPH '86, Association for Computing Machinery, p. 151–160. doi:10.1145/15922.15903. 6

[SP91] SCLAROFF S., PENTLAND A.: Generalized implicit functions for computer graphics. In *ACM SIGGRAPH* (New York, NY, USA, 1991), SIGGRAPH '91, Association for Computing Machinery, p. 247–250. URL: https://doi.org/10.1145/122718.122745, doi:10.1145/122718.122745. 23

[SSK*21] SONG Y., SOHL-DICKSTEIN J., KINGMA D. P., KUMAR A., ERMON S., POOLE B.: Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)* (2021). 12

[SSN*22] SCHWARZ K., SAUER A., NIEMEYER M., LIAO Y., GEIGER A.: Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022). 11

[SSP07] SUMNER R. W., SCHMID J., PAULY M.: Embedded deformation for shape manipulation. *ACM Transactions on Graphics (ToG) 26*, 3 (jul 2007), 80–es. URL: https://doi.org/10.1145/1276377.1276478, doi:10.1145/1276377.1276478. 17

[SSP*23] SINGER U., SHEYNIN S., POLYAK A., ASHUAL O., MAKAROV I., KOKKINOS F., GOYAL N., VEDALDI A., PARIKH D., JOHNSON J., TAIGMAN Y.: Text-to-4d dynamic scene generation. In *International Conference on Machine Learning (ICML)* (2023), ICML'23, JMLR.org. 2, 26, 27, 28, 29

[SSR*23] SINHA S., SHAPOVALOV R., REIZENSTEIN J., ROCCO I., NEVEROVA N., VEDALDI A., NOVOTNY D.: Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4881–4891. 1, 25, 26

[STB*19] SRINIVASAN P. P., TUCKER R., BARRON J. T., RAMAMOORTHI R., NG R., SNAVELY N.: Pushing the boundaries of view extrapolation with multiplane images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019). 18

[STG*20] SIDHU V., TRETSCHK E., GOLYANIK V., AGUDO A., THEOBALT C.: Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision (ECCV)* (2020), Springer, pp. 204–222. 14, 16

[SWK24] STOTKO D., WANDEL N., KLEIN R.: Physics-guided shape-from-template: Monocular video perception through neural surrogate models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 16

[SWW*20] SMITH B., WU C., WEN H., PELUSE P., SHEIKH Y., HODGINS J. K., SHIRATORI T.: Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (ToG) 39*, 6 (2020), 1–14. 23

[SXZ*22] SU Z., XU L., ZHONG D., LI Z., DENG F., QUAN S., FANG L.: Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022). 28

[SYD*23] SONG C., YANG G., DENG K., ZHU J.-Y., RAMANAN D.: Total-recon: Deformable scene reconstruction for embodied view synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 1, 8, 19, 20, 28

[SZFP16] SCHÖNBERGER J. L., ZHENG E., FRAHM J.-M., POLLEFEYS M.: Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)* (2016), Springer, pp. 501–518. 3

[SZT*23] SHAO R., ZHENG Z., TU H., LIU B., ZHANG H., LIU Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16632–16642. 5, 13, 14

[SZW19] SITZMANN V., ZOLLHÖFER M., WETZSTEIN G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)* (2019), vol. 32. 5

[TA18] TSOLI A., ARGYROS A. A.: Joint 3d tracking of a deformable object in interaction with a hand. In *European Conference on Computer Vision (ECCV)* (2018). 28

[TAL*07] THEOBALT C., AHMED N., LENSCH H., MAGNOR M., SEIDEL H.-P.: Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 13*, 4 (2007), 663–674. 27

[TD20] TEED Z., DENG J.: Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)* (2020). 10, 12

[TDD23] TIAN F., DU S., DUAN Y.: MonoNeRF: Learning a generalizable dynamic radiance field from monocular videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 1, 2, 21, 26

[TDZ*23] TSCHERNEZKI V., DARKHALIL A., ZHU Z., FOUHEY D., LARINA I., LARLUS D., DAMEN D., VEDALDI A.: EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)* (2023). 25

[TFT*20] TEWARI A., FRIED O., THIES J., SITZMANN V., LOMBARDI S., SUNKAVALLI K., MARTIN-BRUALLA R., SIMON T., SARAGIH J., NIESSNER M., PANDEY R., FANELLO S., WETZSTEIN G., ZHU J.-Y., THEOBALT C., AGRAWALA M., SHECHTMAN E., GOLDMAN D. B., ZOLLHÖFER M.: State of the art on neural rendering. *Computer Graphics Forum 39*, 2 (2020), 701–727. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14022, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14022, doi:https://doi.org/10.1111/cgf.14022. 12

[TGZ*24] TRETSCHK E., GOLYANIK V., ZOLLHÖFER M., BOZIC A., LASSNER C., THEOBALT C.: Scenerflow: Time-consistent reconstruction of general dynamic scenes. In *International Conference on 3D Vision (3DV)* (2024). 11, 13, 15, 28

[TKBR*23] TRETSCHK E., KAIRANDA N., B R M., DABRAL R., KORTYLEWSKI A., EGGER B., HABERMANN M., FUA P., THEOBALT C., GOLYANIK V.: State of the art in dense monocular non-rigid 3d reconstruction. *Computer Graphics Forum 42*, 2 (2023), 485–520. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14774, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14774, doi:https://doi.org/10.1111/cgf.14774. 2

[TLLV22] TSCHERNEZKI V., LAINA I., LARLUS D., VEDALDI A.: Neural feature fusion fields: 3d distillation of selfsupervised 2d image representations. In *International Conference on 3D Vision (3DV)* (2022). 20

[TLV21] TSCHERNEZKI V., LARLUS D., VEDALDI A.: Neuraldiff: Segmenting 3d objects that move in egocentric videos. In *International Conference on 3D Vision (3DV)* (2021), IEEE, pp. 910–919. 19

[TLYCS22] TSENG W.-C., LIAO H.-J., YEN-CHEN L., SUN M.: Clanerf: Category-level articulated neural radiance field. In *IEEE International Conference on Robotics and Automation (ICRA)* (2022), IEEE, pp. 8454–8460. 26

[TMW*22] TANG J., MARKHASIN L., WANG B., THIES J., NIESSNER M.: Neural shape deformation priors. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022), vol. 35, pp. 17117–17132. 26, 27

[TRZ*24] TANG J., REN J., ZHOU H., LIU Z., ZENG G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations (ICLR)* (2024). 27

[TS20] TUCKER R., SNAVELY N.: Single-view view synthesis with multiplane images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). 10, 18

[TSM*20] TANCIK M., SRINIVASAN P., MILDENHALL B., FRIDOVICH-KEIL S., RAGHAVAN N., SINGHAL U., RAMAMOORTHI R., BARRON J., NG R.: Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020), vol. 33, pp. 7537–7547. 5

[TSTPM21] TIWARI G., SARAFIANOS N., TUNG T., PONS-MOLL G.: Neural-gif: Neural generalized implicit functions for animating people in clothing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 11708–11718. 23

[TTG*21] TRETSCHK E., TEWARI A., GOLYANIK V., ZOLLHÖFER M., LASSNER C., THEOBALT C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), IEEE, pp. 12939–12950. 2, 13

[TTM*22] TEWARI A., THIES J., MILDENHALL B., SRINIVASAN P., TRETSCHK E., YIFAN W., LASSNER C., SITZMANN V., MARTIN-BRUALLA R., LOMBARDI S., SIMON T., THEOBALT C., NIESSNER M., BARRON J. T., WETZSTEIN G., ZOLLHÖFER M., GOLYANIK V.: Advances in neural rendering. *Computer Graphics Forum 41*, 2 (2022), 703–735. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14507, arXiv:https:

//onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14507, doi:https://doi.org/10.1111/cgf.14507. 2

[TXJZ21] TANG J., XU D., JIA K., ZHANG L.: Learning parallel dense correspondence from spatio-temporal descriptors for efficient and robust 4d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 6022–6031. 25

[TYR23] TAN J., YANG G., RAMANAN D.: Distilling neural fields for real-time articulated shape reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4692–4701. 25, 26

[TZFR23] TURKI H., ZHANG J. Y., FERRONI F., RAMANAN D.: Suds: Scalable urban dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12375–12385. 8, 10, 19, 20, 28

[UEK23] UZOLAS L., EISEMANN E., KELLNHOFER P.: Template-free articulated neural point clouds for reposable view synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2023). URL: https://openreview.net/forum?id=fyfmHi8ay3. 9, 21, 24

[VBR*99] VEDULA S., BAKER S., RANDER P., COLLINS R., KANADE T.: Three-dimensional scene flow. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (1999), vol. 2, IEEE, pp. 722–729. 7

[VHTS*22] VAN HOORICK B., TENDULKAR P., SURÍS D., PARK D., STENT S., VONDRICK C.: Revealing occlusions with 4d neural fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 3011–3021. 11, 26

[vMHB*18] VON MARCARD T., HENSCHEL R., BLACK M. J., ROSENHAHN B., PONS-MOLL G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)* (September 2018). 25

[VNH*22] VU T.-A., NGUYEN D. T., HUA B.-S., PHAM Q.-H., YEUNG S.-K.: Rfnet-4d: Joint object reconstruction and flow estimation from 4d point clouds. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 36–52. 13, 14

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U., POLOSUKHIN I.: Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017). 5

[WBSS04] WANG Z., BOVIK A., SHEIKH H., SIMONCELLI E.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861. 11

[WCC*23] WANG Q., CHANG Y.-Y., CAI R., LI Z., HARIHARAN B., HOLYNSKI A., SNAVELY N.: Tracking everything everywhere all at once. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 28

[WCM*22] WEI F., CHABRA R., MA L., LASSNER C., ZOLLHOEFER M., RUSINKIEWICZ S., SWEENEY C., NEWCOMBE R., SLAVCHEVA M.: Self-supervised neural articulated shape and appearance models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 15816–15826. 26

[WCS*22] WENG C.-Y., CURLESS B., SRINIVASAN P. P., BARRON J. T., KEMELMACHER-SHLIZERMAN I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 16210–16220. 23

[WDSY23] WANG Y., DONG Y., SUN F., YANG X.: Root pose decomposition towards generic non-rigid 3d reconstruction with monocular videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 8, 19, 20, 28

[WELG21] WANG C., ECKART B., LUCEY S., GALLO O.: Neural trajectory fields for dynamic novel view synthesis, 2021. arXiv:2105.05994. 7, 10, 13

[WFZF22] WI Y., FLORENCE P., ZENG A., FAZELI N.: Virdo: Visio-tactile implicit representations of deformable objects. In *IEEE International Conference on Robotics and Automation (ICRA)* (2022), IEEE, pp. 3583–3590. 22

[WHH*23a] WANG L., HU Q., HE Q., WANG Z., YU J., TUYTELAARS T., XU L., WU M.: Neural residual radiance fields for streamably free-viewpoint videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 76–87. 13, 15

[WHH*23b] WANG Y., HAN Q., HABERMANN M., DANIILIDIS K., THEOBALT C., LIU L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 3295–3306. 11, 15

[WISL23] WEWER C., ILG E., SCHIELE B., LENSSEN J. E.: Simnp: Learning self-similarity priors between neural points. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 12

[WJRV23] WU S., JAKAB T., RUPPRECHT C., VEDALDI A.: Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision (IJCV) 131*, 10 (2023), 2623–2634. 26

[WLC*23] WANG P., LIU Y., CHEN Z., LIU L., LIU Z., KOMURA T., THEOBALT C., WANG W.: F2-nerf: Fast neural radiance field training with free camera trajectories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 4

[WLFD22] WANG Y., LONG Y., FAN S. H., DOU Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2022), Springer, pp. 431–441. 14

[WLHJ*20] WANG Z., LI S., HOWARD-JENKINS H., PRISACARIU V., CHEN M.: Flownet3d++: Geometric losses for deep scene flow estimation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2020), pp. 91–98. 26

[WLJ*23] WU S., LI R., JAKAB T., RUPPRECHT C., VEDALDI A.: Magicpony: Learning articulated 3d animals in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 8792–8802. 26

[WLL*21] WANG P., LIU L., LIU Y., THEOBALT C., KOMURA T., WANG W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021), Ranzato M., Beygelzimer A., Dauphin Y. N., Liang P., Vaughan J. W., (Eds.), pp. 27171–27183. 6, 13, 22, 27

[WLPL22] WANG C., LI X., PONTES J. K., LUCEY S.: Neural prior for trajectory estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 6532–6542. 16

[WM23] WONG Y.-S., MITRA N. J.: Factored neural representation for scene understanding. *Computer Graphics Forum 42*, 5 (2023), e14911. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14911, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14911, doi:https://doi.org/10.1111/cgf.14911. 19, 20, 21

[WMDH22] WANG W., MORGAN A. S., DOLLAR A. M., HAGER G. D.: Dynamical scene representation and control with keypoint-conditioned neural radiance field. In *IEEE International Conference on Automation Science and Engineering (CASE)* (2022), IEEE, pp. 1138–1143. 21, 22

[WMJL23] WANG C., MACDONALD L. E., JENI L. A., LUCEY S.: Flow supervision for deformable nerf. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 21128–21137. 2, 13, 14

[WPYS21] WIZADWONGSA S., PHONGTHAWEE P., YENPHRAPHAI J., SUWAJANAKORN S.: Nex: Real-time view synthesis with neural basis expansion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 18

[WSVT13] WU C., STOLL C., VALGAERTS L., THEOBALT C.: On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics (ToG) 32*, 6 (2013), 1–11. 27

[WTL*23] WANG F., TAN S., LI X., TIAN Z., SONG Y., LIU H.: Mixed neural voxels for fast multi-view video synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 19706–19716. 13, 15, 20

[WWG*21] WANG Q., WANG Z., GENOVA K., SRINIVASAN P. P., ZHOU H., BARRON J. T., MARTIN-BRUALLA R., SNAVELY N., FUNKHOUSER T.: Ibrnet: Learning multi-view image-based rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 4690–4699. 5, 7, 15

[WYF*24] WU G., YI T., FANG J., XIE L., ZHANG X., WEI W., LIU W., TIAN Q., WANG X.: 4d gaussian splatting for real-time dynamic scene rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 11, 13, 18, 27

[WZFF22] WI Y., ZENG A., FLORENCE P., FAZELI N.: Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects. In *Conference on Robot Learning (CoRL)* (2022). 21, 22

[WZL*22] WANG L., ZHANG J., LIU X., ZHAO F., ZHANG Y., ZHANG Y., WU M., YU J., XU L.: Fourier plenoctrees for dynamic radiance field rendering in real-time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 13524–13534. 1, 13, 15

[WZN*14] WU C., ZOLLHÖFER M., NIESSNER M., STAMMINGER M., IZADI S., THEOBALT C.: Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics (ToG) 33*, 6 (2014), 1–10. 27

[WZS23] WANG D., ZHANG T., SÜSSTRUNK S.: NEMTO: Neural Environment Matting for Novel View and Relighting Synthesis of Transparent Objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 27

[WZT*22] WU T., ZHONG F., TAGLIASACCHI A., COLE F., OZTIRELI C.: Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022), vol. 35, pp. 32653–32666. 2, 19

[XAS21] XU H., ALLDIECK T., SMINCHISESCU C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021), vol. 34, pp. 14955–14966. 23

[XC22] XING W., CHEN J.: Temporal-mpi: Enabling multi-plane images for dynamic scene modelling via temporal basis learning. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 323–338. 13, 18

[XH22] XU T., HARADA T.: Deforming radiance fields with cages. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 159–175. 1, 5, 21, 24

[XHKK21] XIAN W., HUANG J.-B., KOPF J., KIM C.: Space-time neural irradiance fields for free-viewpoint video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 9421–9431. 10, 13

[XLSL23] XU S., LI L., SHEN L., LIAN Z.: Desrf: Deformable stylized radiance field. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 709–718. 22

[XMY*21] XU B., MA L., YE Y., SCHMIDT T., TWIGG C. D., LOVEGROVE S.: Identity-disentangled neural deformation model for dynamic meshes, 2021. arXiv:2109.15299. 25

[XPL*24] XU Z., PENG S., LIN H., HE G., SUN J., SHEN Y., BAO H., ZHOU X.: 4k4d: Real-time 4d view synthesis at 4k resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 13, 16

[XWI*21] XIE K., WANG T., IQBAL U., GUO Y., FIDLER S., SHKURTI F.: Physics-based human motion estimation and synthesis from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 28

[XXP*22] XU Q., XU Z., PHILIP J., BI S., SHU Z., SUNKAVALLI K., NEUMANN U.: Point-nerf: Point-based neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 2, 16, 23

[XYZ*23] XIE W., YU Z., ZHAO Z., ZUO B., WANG Y.: Hmdo: Markerless multi-view hand manipulation capture with deformable objects. *Graphical Models 127* (2023), 101178. 28

[XZK*20] XU Z., ZHOU Y., KALOGERAKIS E., LANDRETH C., SINGH K.: Rignet: Neural rigging for articulated characters. *ACM Transactions on Graphics (ToG) 39* (2020). 5, 24, 25

[YGKL21] YARIV L., GU J., KASTEN Y., LIPMAN Y.: Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021), Ranzato M., Beygelzimer A., Dauphin Y. N., Liang P., Vaughan J. W., (Eds.), pp. 4805–4815. 6, 14

[YGW21] YU H.-X., GUIBAS L., WU J.: Unsupervised discovery of object radiance fields. In *International Conference on Learning Representations (ICLR)* (2021). 18

[YGZ*24] YANG Z., GAO X., ZHOU W., JIAO S., ZHANG Y., JIN X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 13, 18

[YH23] YOU M., HOU J.: Decoupling dynamic monocular videos for dynamic view synthesis, 2023. arXiv:2304.01716. 13

[YHL*23] YAO C.-H., HUNG W.-C., LI Y., RUBINSTEIN M., YANG M.-H., JAMPANI V.: Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 4853–4862. 26

[YHS*24] YANG H., HUANG X., SUN B., BAJAJ C., HUANG Q.: Gencorres: Consistent shape matching via coupled implicit-explicit shape generative models. In *International Conference on Learning Representations (ICLR)* (May 2024). 26, 27

[YJM*23] YU H., JULIN J., MILACSKI Z. A., NIINUMA K., JENI L. A.: Dylin: Making light field networks dynamic. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 12397–12406. 13, 22

[YKG*20] YOON J. S., KIM K., GALLO O., PARK H. S., KAUTZ J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 5336–5345. 25

[YLGO23] YUGAY V., LI Y., GEVERS T., OSWALD M. R.: Gaussian-slam: Photo-realistic dense slam with gaussian splatting, 2023. arXiv:2312.10070. 27

[YLL*18] YU C., LIU Z., LIU X.-J., XIE F., YANG Y., WEI Q., FEI Q.: Ds-slam: A semantic visual slam towards dynamic environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), IEEE, pp. 1168–1174. 16

[YLL23a] YAN Z., LI C., LEE G. H.: Nerf-ds: Neural radiance fields for dynamic specular objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 19

[YLL23b] YAN Z., LI C., LEE G. H.: Od-nerf: Efficient training of on-the-fly dynamic neural radiance fields, 2023. arXiv:2305.14831. 15

[YLT*21] YU A., LI R., TANCIK M., LI H., NG R., KANAZAWA A.: Plenoctrees for real-time rendering of neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5752–5761. 21

[YLW*21] YUAN Y.-J., LAI Y.-K., WU T., GAO L., LIU L.: A revisit of shape editing techniques: From the geometric to the neural viewpoint. *Journal of Computer Science and Technology 36*, 3 (jun 2021), 520–554. doi:10.1007/s11390-021-1414-9. 20

[YRH*23] YAO C.-H., RAJ A., HUNG W.-C., LI Y., RUBINSTEIN M., YANG M.-H., JAMPANI V.: Artic3d: Learning robust articulated 3d shapes from noisy web image collections. In *Advances in Neural Information Processing Systems (NeurIPS)* (2023). 26

[YSJ*21a] YANG G., SUN D., JAMPANI V., VLASIC D., COLE F., CHANG H., RAMANAN D., FREEMAN W. T., LIU C.: Lasr: Learning articulated shape reconstruction from a monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 15980–15989. 24

[YSJ*21b] YANG G., SUN D., JAMPANI V., VLASIC D., COLE F., LIU C., RAMANAN D.: Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. *Advances in Neural Information Processing Systems (NeurIPS) 34* (2021), 19326–19338. 24

[YSL*22] YUAN Y.-J., SUN Y.-T., LAI Y.-K., MA Y., JIA R., GAO L.: Nerf-editing: geometry editing of neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 18353–18364. 18, 21, 22

[YVN*22] YANG G., VO M., NEVEROVA N., RAMANAN D., VEDALDI A., JOO H.: Banmo: Building animatable 3d neural models from many casual videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). 2, 4, 5, 6, 9, 10, 20, 21, 24, 25, 26

[YWRR23] YANG G., WANG C., REDDY N. D., RAMANAN D.: Reconstructing animatable categories from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16995–17005. 2, 4, 21, 24, 25, 26

[YWW*23] YANG C., WANG K., WANG Y., YANG X., SHEN W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2023). 15

[YYP*24] YANG Z., YANG H., PAN Z., ZHU X., ZHANG L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)* (2024). 13, 18

[YYTK21] YU A., YE V., TANCIK M., KANAZAWA A.: pixelNeRF: Neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). 5

[YYZ*23] YANG G., YANG S., ZHANG Z., MANCHESTER Z., RAMANAN D.: Ppr: Physically plausible reconstruction from monocular videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 11, 19, 28

[YZG*21] YU T., ZHENG Z., GUO K., LIU P., DAI Q., LIU Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5746–5756. 15

[YZH*24] YANG K., ZHANG X., HUANG Z., CHEN X., XU Z., SU H.: Movingparts: Motion-based 3d part discovery in dynamic radiance field. In *International Conference on Learning Representations (ICLR)* (May 2024). 2, 8, 19, 20

[YZX*21] YANG B., ZHANG Y., XU Y., LI Y., ZHOU H., BAO H., ZHANG G., CUI Z.: Learning object-compositional neural radiance field for editable scene rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 13779–13788. 18

[ZBRH22] ZHANG Q., BAEK S.-H., RUSINKIEWICZ S., HEIDE F.: Differentiable point-based radiance fields for efficient view synthesis. In *ACM SIGGRAPH Asia* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. doi:10.1145/3550469.3555413. 18

[ZBYX19] ZHANG H., BO Z.-H., YONG J.-H., XU F.: Interactionfusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. *ACM Transactions on Graphics (ToG) 38*, 4 (2019), 1–11. 28

[ZCL*23] ZHA R., CHENG X., LI H., HARANDI M., GE Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2023), Springer, pp. 13–23. 14

[ZCM*24] ZHAO X., COLBURN A., MA F., ÁNGEL BAUTISTA M., SUSSKIND J. M., SCHWING A. G.: Pseudo-Generalized Dynamic View Synthesis from a Video. In *International Conference on Learning Representations (ICLR)* (2024). 26

[ZCT*21] ZHANG Z., COLE F., TUCKER R., FREEMAN W. T., DEKEL T.: Consistent depth of moving objects in video. *ACM Transactions on Graphics (ToG) 40*, 4 (2021), 1–12. 10

[ZDY*21] ZENG H., DAI Y., YU X., WANG X., YANG Y.: Pr-rrn: pairwise-regularized residual-recursive networks for non-rigid structure-from-motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 5600–5609. 16

[ZFB23] ZHOU B., FRANCO J.-S., BOYER E.: Human body shape completion with implicit shape and flow learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 26

[ZGLA23] ZANARDELLI M., GUERRINI F., LEONARDI R., ADAMI N.: Image forgery detection: a survey of recent deep-learning approaches. *Multimedia Tools and Applications* (2023). 28

[ZHS*23] ZHENG Y., HARLEY A. W., SHEN B., WETZSTEIN G., GUIBAS L. J.: Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 19855–19865. 11

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018). 10, 11

[ZLC22] ZHANG T., LAU Y.-F., CHEN Q.: A portable multiscopic camera for novel view and time synthesis in dynamic scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2022), IEEE, pp. 2409–2416. 13

[ZLX23] ZHENG C., LIN W., XU F.: Editablenerf: Editing topologically varying neural radiance fields by key points. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 8317–8327. 1, 21, 22

[ZLY*23a] ZHANG J., LAN Y., YANG S., HONG F., WANG Q., YEO C. K., LIU Z., LOY C. C.: Deformtoon3d: Deformable neural radiance fields for 3d toonification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 9144–9154. 22

[ZLY*23b] ZHANG J., LUO H., YANG H., XU X., WU Q., SHI Y., YU J., XU L., WANG J.: Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 8834–8845. 28

[ZPS*23] ZHANG S., PENG S., SHENTU Y., SHUAI Q., CHEN T., YU K., BAO H., ZHOU X.: Dyn-e: Local appearance editing of dynamic neural radiance fields, 2023. arXiv:2307.12909. 21

[ZPvBG01] ZWICKER M., PFISTER H., VAN BAAR J., GROSS M.: Surface splatting. In *ACM SIGGRAPH* (New York, NY, USA, 2001), SIGGRAPH '01, Association for Computing Machinery, p. 371–378. URL: https://doi.org/10.1145/383259.383300, doi:10.1145/383259.383300. 6

[ZQZ*22] ZENG Y., QIAN Y., ZHANG Q., HOU J., YUAN Y., HE Y.: Idea-net: Dynamic 3d point cloud interpolation via deep embedding alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 6338–6347. 14

[ZRA23] ZHANG L., RAO A., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 29

[ZSD*21] ZHANG X., SRINIVASAN P. P., DENG B., DEBEVEC P., FREEMAN W. T., BARRON J. T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG) 40*, 6 (2021), 1–18. 27

[ZSG*18] ZOLLHÖFER M., STOTKO P., GÖRLITZ A., THEOBALT C., NIESSNER M., KLEIN R., KOLB A.: State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum 37*, 2 (2018), 625–652. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13386, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13386, doi:https://doi.org/10.1111/cgf.13386. 2

[ZT00] ZIENKIEWICZ O. C., TAYLOR R. L.: *The finite element method: solid mechanics*, vol. 2. Butterworth-heinemann, 2000. 8

[ZT23] ZHOU Z., TULSIANI S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 27

[ZTF*18] ZHOU T., TUCKER R., FLYNN J., FYFFE G., SNAVELY N.: Stereo magnification: Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics (ToG) 37*, 4 (2018), 1–12. 5, 18

[ZW22] ZHANG Y., WU J.: Video extrapolation in space and time. In *European Conference on Computer Vision (ECCV)* (2022), pp. 313–333. 10, 18

[ZWL*22] ZHANG J., WANG L., LIU X., ZHAO F., LI M., DAI H., ZHANG B., YANG W., XU L., YU J.: Neuvv: Neural volumetric videos with immersive rendering and editing, 2022. arXiv:2202.06088. 21

[ZWL*23] ZHENG Z., WU D., LU R., LU F., CHEN G., JIANG C.: Neuralpci: Spatio-temporal neural field for 3d point cloud multi-frame non-linear interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 909–918. 14

[ZXY*23] ZHANG Y., XU T., YU J., YE Y., JING Y., WANG J., YU J., YANG W.: Nemf: Inverse volume rendering with neural microflake field. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 22919–22929. 27

[ZYMY22] ZENG H., YU X., MIAO J., YANG Y.: Mhr-net: Multiple-hypothesis reconstruction of non-rigid shapes from 2d views. In *European Conference on Computer Vision (ECCV)* (2022), Springer, pp. 1–17. 16

[ZYW*23] ZHENG Y., YIFAN W., WETZSTEIN G., BLACK M. J., HILLIGES O.: Pointavatar: Deformable point-based head avatars from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 21057–21067. 27