4D Dynamic Scene Reconstruction: A Comprehensive Survey

Ziren Gong, Guo Chen, Yongjia Li, Yihua Shao, Fabio Tosi, Stefano Mattoccia, Matteo Poggi, Hao Tang†, Fei Ma, Shuyan Li, Ziyang Yan†, Nicu Sebe, Ling Shao, Jianfei Cai, Qi Tian, Ming-Hsuan Yang

Abstract—4D scene reconstruction in dynamic environments is a fundamental problem in robotics and computer vision, as it aims to model scenes that evolve over time. Traditional geometric pipelines face inherent challenges in handling non-rigid motion, occlusions, appearance variations, and maintaining temporal consistency. Recent advances in neural radiance fields, including Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS), have significantly pushed the boundaries by enabling high-fidelity, continuous, and real-time 4D reconstruction beyond the capabilities of classical approaches. Despite this progress, a comprehensive survey dedicated to radiance field-based 4D reconstruction remains lacking. This paper presents a systematic review of recent developments, with a focus on NeRF- and 3DGS-based methods. We introduce a unified taxonomy and conceptual pipelines for existing 4D reconstruction systems, and provide a critical analysis of their strengths and limitations. In addition, we summarize representative datasets and evaluation metrics, and discuss key open challenges that warrant further investigation. This survey aims to offer a coherent and structured foundation for advancing 4D reconstruction in dynamic and complex real-world environments. A continuously updated version is available online at github webpage.

Index Terms—4D Scene Reconstruction, Neural Radiance Field, Gaussian Splatting, Literature Survey

1 Introduction

Scene reconstruction is a core problem in robotics and spatial AI, aiming to recover the three-dimensional structure and appearance of real-world environments from multi-view observations. It remains a fundamental challenge in computer vision and graphics, with applications spanning navigation, scene understanding, and novel view synthesis [1]–[5]. Considerable efforts have been devoted to developing methods for dense, accurate, and high-fidelity reconstruction.

The field has evolved substantially over the past three decades. Early approaches were based on classical geometric pipelines, with Structure from Motion (SfM) and Multi-View Stereo (MVS) forming the foundation of three-dimensional reconstruction. SfM estimates camera poses and sparse scene geometry via feature matching and bundle adjustment, while MVS densifies these reconstructions using photometric consistency across calibrated views, producing detailed surface models of static scenes [6], [7]. Despite their effectiveness, these methods are limited by computational efficiency and their reliance on hand-crafted features and geometric assumptions.

With the development of Simultaneous Localization and Mapping (SLAM) [8], scene reconstruction has progressed from offline batch processing to online frameworks that jointly perform camera tracking and environment mapping. However, traditional geometric pipelines remain sensitive to noise in incremental inputs, leading to cumulative trajectory drift and motion-induced artifacts that degrade reconstruction quality over time [5], [9].

The field has undergone a paradigm shift with the advent of deep learning and neural representations. Neural Radiance Fields (NeRF) [10] represent scenes as continuous implicit functions, enabling high-quality view synthesis. Subsequent extensions have improved training efficiency, rendering speed, and robustness [11]–[13]. More recently, 3D Gaussian Splatting (3DGS) [14] has emerged as an effective alternative, representing scenes with anisotropic 3D Gaussians and leveraging a differentiable tile-based rasterizer for real-time rendering while preserving fine details. These approaches depart from discrete geometric representations and instead learn continuous scene representations that capture complex geometry and appearance.

Despite advances in static scene modeling, extending reconstruction to dynamic environments introduces substantially greater complexity. High-fidelity 4D reconstruc-

- *Ziren Gong, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi are with the Department of Computer Science and Engineering, University of Bologna, Italy.*
  Guo Chen is with the Wangxuan Institute of Computer Technology, Peking University, China.
  Yongjia Li and Yihua Shao are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong.
  Hao Tang is with the School of Computer Science, Peking University, China.
  Fei Ma and Qi Tian are with Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China. Qi Tian is also with Huawei.
  Shuyan Li is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom.
  Ziyang Yan and Nicu Sebe are with the Department of Information Engineering and Computer Science, University of Trento, Italy.
  Ling Shao is with the University of the Chinese Academy of Sciences, China.
  Jianfei Cai is with the Faculty of IT, Monash University, Australia.
  Ming-Hsuan Yang is with Google DeepMind and the University of California, Merced, United States.
- *Corresponding author: Ziyang Yan, Hao Tang.*
- *E-mail: yanziyang199634@gmail.com; bjdxtanghao@gmail.com*

† denotes corresponding authors.

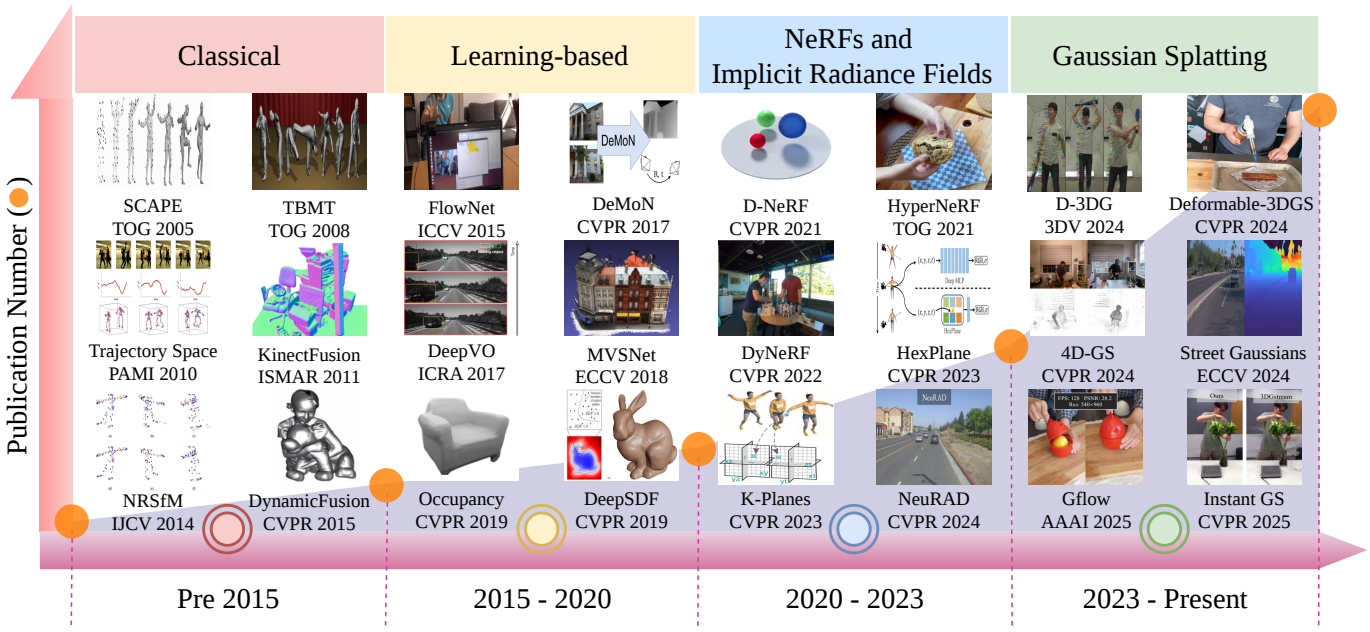

Fig. 1. **Trends in 4D reconstruction.** The increasing adoption of NeRF- and GS-based methods for dynamic scene modeling has led to a rapid growth in related publications in recent years.

tion must address the ambiguities of non-rigid motion, preserve long-range spatio-temporal coherence, and disentangle time-varying radiance from geometric deformation.

The rapid development of 4D dynamic reconstruction, driven by the transition from implicit neural representations to explicit Gaussian primitives, has led to an increasingly diverse and fragmented research landscape [15], [16]. As illustrated in Fig. 1, the field is undergoing a shift in design choices, with evolving trade-offs among rendering speed, temporal consistency, and memory efficiency. This survey is motivated by the need to consolidate recent advances into a coherent framework and provide a structured reference for both established methods and emerging approaches such as 4D Gaussian Splatting (4DGS).

Existing surveys [11], [17]–[22] primarily focus on static neural rendering and often treat dynamic reconstruction as a secondary topic. As a result, a systematic review of 4D dynamic reconstruction using scene-specific optimization remains lacking. In this survey, we focus on the two principal paradigms for scene-specific optimization in dynamic reconstruction: implicit 4D Neural Radiance Fields (4D NeRF) and explicit 4D Gaussian Splatting (4DGS). While classical Simultaneous Localization and Mapping (SLAM) and Structure from Motion (SfM) provide the geometric foundation for camera tracking and mapping, they are typically predicated on rigid-body assumptions or sparse representations. Emerging feed-forward, generalizable 4D reconstruction approaches enable rapid, offline-style inference across diverse scenes; however, they are fundamentally constrained by reliance on pre-trained dataset biases, often resulting in a *fidelity ceiling* in out-of-distribution scenarios. We emphasize per-scene optimization approaches, which remain the gold standard for achieving high-fidelity reconstruction and temporal consistency. Both 4D NeRF and 4DGS mitigate the generalization gap of feed-forward models by anchoring reconstruction to scene-specific observations, rather than the statistical priors of a training set. Furthermore, we analyze the performance of representative methods on benchmark datasets and discuss key open challenges for future research.

2 Preliminaries

2.1 History of Dynamic Scene Reconstruction

As shown in Fig. 1, dynamic scene reconstruction has evolved over the past two decades from geometry-based pipelines to learning-based methods and, more recently, to implicit neural representations. The introduction of NeRF established 4D reconstruction as a central research direction, further accelerated by 3DGS. This has led to rapid growth in recent work. This section reviews this progression and highlights key developments underlying modern radiance field and Gaussian-based approaches.

Classical Approaches: From Geometry to Volumetric Fusion (Pre 2015). Early work was dominated by geometry-based pipelines. Multi-view stereo (MVS) and structure-from-motion (SfM) were extended to dynamic settings, leading to non-rigid SfM (NRSfM) [23] and template-based mesh tracking [24]. These methods estimate per-frame geometry and track deformations relative to a canonical model. In parallel, depth sensors enabled volumetric fusion methods, such as KinectFusion [25] and DynamicFusion [26], which integrate depth maps into a canonical volume with non-rigid alignment.

These approaches have several limitations. Explicit geometry restricts modeling of complex deformations and topology changes. Per-frame optimization limits scalability for long sequences and high-resolution scenes. Appearance modeling is also limited, typically relying on texture maps and lacking view-dependent effects.

Transition to Learning-based Models (2015–2020). With the rise of deep learning, reconstruction methods began incorporating learned priors. Early works [27]–[32] improved components such as depth, scene flow, and pose estimation, while retaining classical pipelines. A key shift was the

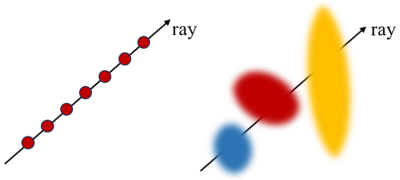

Fig. 2. **Comparison between NeRF and 3DGS**. NeRF (left) evaluates an MLP along each ray, whereas 3DGS (right) renders by rasterizing and blending Gaussians.

adoption of implicit representations, including occupancy networks [33] and neural SDFs [34], which model scenes as continuous fields. These representations are more expressive and compact, and enable dynamic extensions.

Dynamic NeRFs and Implicit Radiance Fields (2020–2023). Neural Radiance Fields (NeRF) [10] enable high-quality novel view synthesis by modeling scenes as continuous radiance fields. Dynamic variants, such as D-NeRF [35], Nerfies [36], and HyperNeRF [37], represent scenes using canonical fields with deformation models. These methods capture complex non-rigid motion. However, dynamic NeRFs are computationally expensive. Training often requires long runtimes, and inference is slow due to volumetric rendering. Temporal consistency remains difficult over long sequences. Later methods, including TiNeuVox [38] and NSFF [39], introduce explicit structures or motion priors to improve efficiency, but scalability remains limited.

Dynamic 3D Gaussian Splatting (2023–Present). Gaussian Splatting (GS) [14] represents scenes as sets of anisotropic 3D Gaussians with learnable parameters, including position, scale, opacity, rotation, and appearance. Dynamic extensions incorporate temporal modeling through deformation or per-frame transformations, leading to 4D Gaussian Splatting (4DGS). GS enables real-time rendering via rasterization. Its explicit structure facilitates handling occlusions and non-rigid motion. Each Gaussian jointly encodes geometry and appearance, yielding a compact representation. However, dynamic GS methods often require many primitives, leading to high memory usage [40], [41]. Improving efficiency and scalability remains an open problem.

2.2 Representing Radiance Fields

Recent advances in radiance field representations have improved 4D scene reconstruction, enabling high-fidelity modeling of geometry and appearance over time. Both NeRF and 3DGS represent scenes as radiance fields but differ in formulation. NeRF models the scene implicitly using a neural network that maps 3D coordinates and viewing directions to color and density. In contrast, 3DGS represents the scene explicitly with anisotropic 3D Gaussians optimized for efficient rendering. We briefly review these approaches and summarize their differences in Fig. 2.

Neural Radiance Fields. NeRF [10] models a 3D scene as a continuous function that maps a 3D point and viewing direction to color and density:

$$F_\theta(\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma). \tag{1}$$

Novel views are synthesized via differentiable volume rendering, where pixel color is computed by integrating radiance along a camera ray:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - e^{-\sigma_i \delta_i}\right) \mathbf{c}_i, \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right). \tag{2}$$

To capture high-frequency details, NeRF applies positional encoding to the input coordinates, enabling the MLP to represent complex geometry and appearance.

3D Gaussian Splatting. 3DGS [14] represents a scene explicitly as a set of anisotropic 3D Gaussians. Each Gaussian encodes position, density, and appearance, analogous to volumetric elements in NeRF but in an explicit form. Initialized from a sparse SfM point cloud, each Gaussian is parameterized by its mean μ and covariance Σ:

$$G(x) = \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \quad \Sigma = RSS^\top R^\top. \tag{3}$$

For rendering, Gaussians are projected to the image plane and composited via alpha blending:

$$C = \sum_i c_i \alpha_i \prod_{j<i} (1 - \alpha_j). \tag{4}$$

Each Gaussian defines a continuous density in space and contributes to pixel color along a camera ray, analogous to volumetric rendering in NeRF.

2.3 Comparison with Existing Surveys

Recent advances in radiance field representations have driven the development of 4D reconstruction methods based on NeRF and 3DGS, with applications across diverse domains. Table 1 compares recent surveys in this area. Our survey provides broader coverage by including both NeRF- and GS-based methods, summarizing commonly used datasets and evaluation protocols, and incorporating a quantitative analysis for performance comparison.

Among existing works, Zhu et al. [42] is most closely related, but covers fewer methods (52 vs. 102) and datasets (10 vs. 22), and omits several scene types, such as autonomous driving. In addition, it does not provide a unified taxonomy or a comprehensive evaluation. This survey aims to offer a structured and comprehensive reference for 4D scene reconstruction.

3 Dynamic Neural Radiance Fields

3.1 Preliminaries of 4D NeRF

To extend static NeRF to dynamic scenes, temporal dynamics are incorporated by introducing a time dimension into the radiance field formulation. As shown in Fig. 3, existing approaches adopt different strategies to encode temporal information within neural radiance fields. Based on the taxonomy in Table 2, these methods can be categorized into four main classes.

Deformation-Field-Based Methods extend static NeRF by maintaining a canonical 3D representation and learning time-dependent spatial transformations. This approach is based on the observation that many dynamic scenes can be modeled as deformations of a canonical state, where geometric changes are captured by learnable displacement functions. The formulation decomposes the 4D problem

TABLE 1
Comparison of existing surveys on 4D scene reconstruction.

Survey	4D Scene Types	Evaluation Coverage	Datasets	Methods	Taxonomy
Fan et al. [43]	Human and animal motion	–	–	90	–
Zhu et al. [42]	General	NVS, Efficiency	10	52	✓
He et al. [44]	Autonomous driving	–	–	36	–
Cao et al. [45]	General	–	–	111	–
Zhao et al. [46]	Object, human, and animal motion	–	21	62	–
Ours	General	NVS, Geometry, Efficiency	22	102	✓

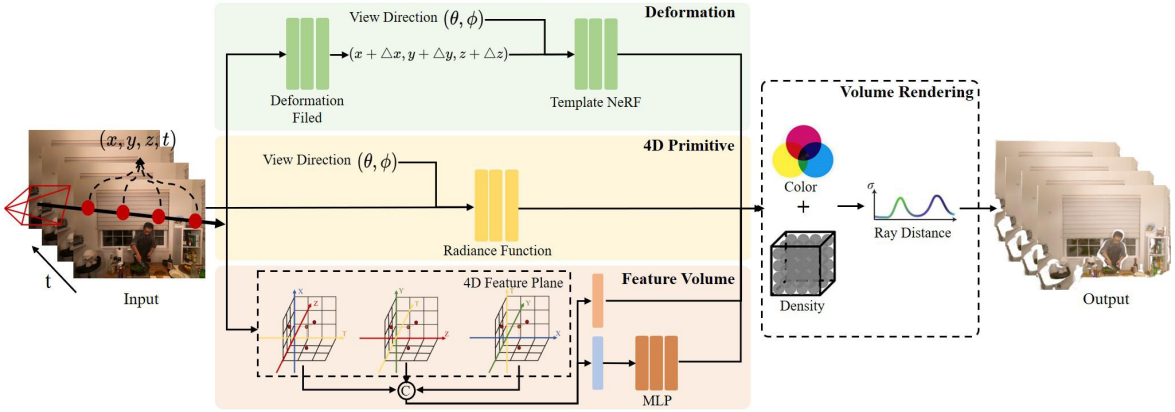

Fig. 3. General pipeline of **NeRF-based** 4D scene reconstruction methods. The pipeline illustrates representative strategies, including deformation-based, 4D primitive-based, and 4D feature volume-based frameworks. Temporal prior-based methods are not included due to their diversity.

into a canonical NeRF and a deformation network. The canonical NeRF F_θ models a reference frame, while the deformation network D_ϕ maps spatial coordinates at time t to the canonical space:

$$\begin{aligned} \mathbf{x}' &= D_\phi(\mathbf{x}, t), \\ F_\theta &: (\mathbf{x}', \mathbf{d}) \rightarrow (\mathbf{c}, \sigma), \end{aligned} \tag{5}$$

where $\mathbf{x}'$ denotes the canonical coordinate and t is the temporal variable. This formulation leverages static NeRF priors while introducing a compact mechanism to model temporal variations.

Implicit 4D Primitive-Based Methods model time as an additional input alongside spatial coordinates and viewing direction. This formulation extends NeRF to a unified 4D representation without explicit decomposition into canonical and deformation components. The radiance field directly maps space-time coordinates to color and density:

$$F_\theta : (\mathbf{x}, \mathbf{d}, t) \rightarrow (\mathbf{c}, \sigma), \tag{6}$$

where the temporal variable t is encoded jointly with spatial inputs, enabling the model to capture temporal variations. This formulation offers high flexibility for modeling complex dynamics but increases computational cost and requires careful temporal sampling.

4D Feature-Volume-Based Methods improve efficiency by decomposing the space-time domain into structured representations. The key idea is to approximate the 4D volume using lower-dimensional components, reducing memory and computation while preserving expressiveness. A common approach factorizes the 4D space using plane-based representations, such as tri-plane extensions:

$$\begin{aligned} \mathbf{f}(\mathbf{x}, t) = \mathbf{f}_{xy}(x, y) \odot \mathbf{f}_{xz}(x, z) \odot \mathbf{f}_{yz}(y, z) \\ \odot \mathbf{f}_{xt}(x, t) \odot \mathbf{f}_{yt}(y, t) \odot \mathbf{f}_{zt}(z, t), \end{aligned} \tag{7}$$

$$F_\theta : \mathbf{f}(\mathbf{x}, t) \rightarrow (\mathbf{c}, \sigma), \tag{8}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{f}_{xy}, \mathbf{f}_{xt}, \mathbf{f}_{yt}$ denote spatial and spatiotemporal feature planes. Alternative approaches use tensor decomposition (e.g., CP or Tucker) to represent the 4D volume with low-rank components, enabling efficient storage and rendering while maintaining temporal coherence.

Temporal-Prior-Based Methods incorporate external temporal cues as auxiliary supervision, rather than explicitly parameterizing time within the radiance field. The key idea is to enforce temporal coherence through additional constraints derived from complementary signals, without modifying the underlying NeRF formulation. These methods augment standard training with temporal consistency losses and motion priors:

$$\begin{aligned} F_\theta &: (\mathbf{x}, \mathbf{d}, t) \rightarrow (\mathbf{c}, \sigma), \\ \mathcal{L} &= \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{flow}}(t) + \lambda_2 \mathcal{L}_{\text{temp}}(t), \end{aligned} \tag{9}$$

where $\mathcal{L}_{\text{flow}}$ encodes motion cues (e.g., optical or scene flow) to enforce geometric consistency, and $\mathcal{L}_{\text{temp}}$ enforces temporal smoothness. These approaches improve coherence and can be combined with other designs.

3.2 Deformation-Field-Based Methods

These methods extend static NeRF by maintaining a canonical 3D representation and learning deformation functions $D(\mathbf{x}, t) \rightarrow \mathbf{x}'$ to model temporal variations. To address non-rigid reconstruction without explicit geometry, early methods employ deformation MLPs to map spatio-temporal

TABLE 2
Overview of NeRF-based 4D dynamic scene reconstruction methods. Methods are categorized into four types. For each method, we summarize its scene representation, key components, and additional priors.

Method	Venue	Inputs	Scenario	Target Domain	4D-style	Scene Encoding	Flow	Normal	Segment.	Extra Prior
D-NeRF [35]	CVPR2021	RGB	Indoor	Entity-Centric	Deformation Fields	MLP				
NR-NeRF [47]	CVPR2021	RGB	In-the-Wild	Scene-Centric	Deformation Fields	MLP				
STaR [48]	CVPR2021	RGB	Indoor	Scene-Centric	Deformation Fields	MLP				
Nerfies [36]	ICCV2021	RGB	Indoor	Entity-Centric	Deformation Fields	MLP				
HyperNeRF [37]	TOG2021	RGB	Indoor	Entity-Centric	Deformation Fields	MLP				
NDR [49]	NIPS2022	RGBD	Indoor	Entity-Centric	Deformation Fields	MLP				
DeVRF [50]	RGBNIPS2022	RGB	Indoor	Entity-Centric	Deformation Fields	Voxel Grid + MLP	✓			RAFT
TiNeuVox [38]	ACM SIG. 2022	RGB	Indoor	Entity-Centric	Deformation Fields	MLP				
NeRF-DS [51]	CVPR2023	RGB	Indoor	Entity-Centric	Deformation Fields	MLP		✓		
RoDynRF [52]	CVPR2023	RGB	In-the-Wild	Scene-Centric	Deformation Fields	Voxel Grid + MLP	✓			RAFT
Total-Recon [53]	ICCV2023	RGBD	Indoor	Scene-Centric	Deformation Fields	MLP	✓			VCN
DyBluRF [54]	CVPR2024	RGB	In-the-Wild	Scene-Centric	Deformation Fields	MLP	✓			RAFT
NSFF [39]	CVPR2021	RGB	In-the-Wild	Scene-Centric	4D Primitive	MLP	✓			RAFT
Video-NeRF [55]	CVPR2021	RGBD	In-the-Wild	Entity-Centric	4D Primitive	MLP				
DynNeRF [56]	ICCV2021	RGB	In-the-Wild	Scene-Centric	4D Primitive	MLP	✓			RAFT
NeRFlow [57]	ICCV2021	RGB	Indoor	Entity-Centric	4D Primitive	MLP	✓			Farneback G.
DyNeRF [58]	CVPR2022	RGB	Indoor	Scene-Centric	4D Primitive	MLP				
MonoNeRF [59]	ICCV2023	RGB	In-the-Wild	Scene-Centric	4D Primitive	MLP	✓			RAFT
Sync-NeRF [60]	AAAI2024	RGB	Indoor	Scene-Centric	4D Primitive	MLP				
4DNDF [61]	CVPR2024	LiDAR	Auto. Driving	Scene-Centric	4D Primitive	Hash Grid + MLP			✓	
Ml-nsg [62]	CVPR2024	RGB	Auto. Driving	Scene-Centric	4D Primitive	Hash Grid + MLP				
DecouplingNeRF [63]	TVCG2024	RGB	In-the-Wild	Scene-Centric	4D Primitive	MLP	✓			NSFF
DetNeRF [64]	AAAI2025	RGB	In-the-Wild	Scene-Centric	4D Primitive	MLP	✓			RAFT
HexPlane [65]	CVPR2023	RGB	Indoor	Entity-Centric	4D Feature Volumes	Feat. Plane + MLP				
K-Planes [66]	CVPR2023	RGB	Indoor	Entity-Centric	4D Feature Volumes	Feat. Plane + MLP				
SUDS [67]	CVPR2023	RGBD	Auto. Driving	Scene-Centric	4D Feature Volumes	Hash Grid + MLP	✓			RAFT & DINO
TIDNeRF [68]	CVPR2023	RGB	Indoor	Multi.-Centric	4D Feature Volumes	Hash Grid + MLP				
HyperReel [69]	CVPR2023	RGB	Indoor	Entity-Centric	4D Feature Volumes	Feat. Plane + MLP				
MixVoxels [70]	ICCV2023	RGB	In. & Wild	Scene-Centric	4D Feature Volumes	Voxel Grid + MLP				
MSTH [71]	NIPS2023	RGB	Indoor	Scene-Centric	4D Feature Volumes	Hash Grid + MLP				Kendall and Gal
NVFi [72]	NIPS2023	RGB	In. & wild	Multi.-Centric	4D Feature Volumes	Feat. Plane + MLP				HexPlane
NeRFPlayer [73]	TVCG2023	RGB	Indoor	Scene-Centric	4D Feature Volumes	Hybrid + MLP				
BLiRF [74]	AAAI2024	RGB	Indoor	Multi.-Centric	4D Feature Volumes	MLP				
Ced-NeRF [75]	AAAI2024	RGB	Indoor	Multi.-Centric	4D Feature Volumes	Hash Grid + MLP				
LiDAR4D [76]	CVPR2024	LiDAR	Auto. Driving	Scene-Centric	4D Feature Volumes	Hybrid + MLP	✓			flow MLP
DaReNeRF [77]	CVPR2024	RGB	Indoor	Scene-Centric	4D Feature Volumes	Feat. Plane + MLP				
Gear-NeRF [78]	CVPR2024	RGB	Indoor	Scene-Centric	4D Feature Volumes	Feat. Plane + MLP			✓	SAM
NeuRAD [79]	CVPR2024	RGBD	Auto. Driving	Scene-Centric	4D Feature Volumes	Hash Grid + MLP				
S-DyRF [80]	CVPR2024	RGB	Indoor	Multi.-Centric	4D Feature Volumes	Feat. Plane + MLP				HexPlane
RoDUS [81]	ECCV2024	RGB	Auto. Driving	Scene-Centric	4D Feature Volumes	Hash Grid + MLP	✓		✓	
EmerNeRF [82]	ICLR2024	RGBD	Auro. Driving	Scene-Centric	4D Feature Volumes	Hash Grid + MLP	✓			DINOv2
SLS4D [83]	TVCG2024	RGB	Indoor	Entity-Centric	4D Feature Volumes	Hybrid + MLP				
StreamRF [84]	NIPS2022	RGB	Indoor	Scene-Centric	Temporal Prior	Voxel Grid + MLP				
OTNeRF [85]	ICLR2024	RGB	Indoor	Entity-Centric	Temporal Prior	MLP				
STGC-NeRF [86]	AAAI2025	LiDAR	Auro. Driving	Scene-Centric	Temporal Prior	Hier. Repre. + MLP	✓	✓		GMSF

coordinates into a canonical space, e.g., **D-NeRF** [35]. On the other hand, **NR-NeRF** [47] introduces rigidity regularization to improve temporal correspondence. However, these methods remain less effective in the presence of significant topological changes or large displacements due to the limitations of a single continuous deformation field.

For complex scenes with articulated structures or multiple interacting objects, global deformation fields are often insufficient to capture localized motions. **STAR** [48], **Total-Recon** [53], and **NDR** [49] address this limitation by factorizing scenes into motion-aware canonical subspaces. With trajectory-based constraints, these methods jointly optimize geometry, camera poses, and non-rigid transformations. However, the optimization process remains computationally expensive and sensitive to the quality of the initial trajectory or pose estimates.

Reconstructing high-fidelity dynamic radiance fields from sparse observations or uncalibrated cameras remains challenging. To improve stability, **DeVRF** [50] and **RoDynRF** [52] employ voxel-based representations for efficient canonicalization. These methods further incorporate auxiliary priors, such as monocular depth, disparity, and reprojection constraints, to jointly estimate camera motion and dynamic scene evolution. However, their performance depends heavily on the quality of the external priors, and inaccurate disparity estimates can introduce artifacts into the 4D representation.

Typical deformation models often struggle with view-dependent specularities and motion blur, leading to entangled geometry and appearance. To alleviate this issue, **DyBluRF** [54] and **NeRF-DS** [51] incorporate surface-normal conditioning and mask-guided deformation to separate transient appearance effects from scene geometry. However, modeling complex radiance variations remains challenging in regions with rapid motion or strong reflections, often resulting in blurred textures or residual artifacts.

3.3 Implicit 4D Primitive-Based Methods

These methods extend NeRF by directly modeling dynamics through an explicit temporal dimension, where the radiance field is defined as $F(\mathbf{x}, t, \mathbf{d}) \rightarrow (\sigma, \mathbf{c})$. This formulation avoids canonical decomposition and provides a unified space-time representation. To improve temporal coherence in dynamic scenes, several methods incorporate explicit motion fields into volumetric representations. **NSFF** [39] introduced neural scene flow fields to jointly optimize geometry, radiance, and dense 3D motion, while **NeRFlow** [57] coupled radiance fields with continuous flow for con-

sistent monocular view synthesis. However, these methods remain sensitive to flow estimation errors, often producing artifacts or blurred geometry under rapid motion.

To address the under-constrained nature of monocular reconstruction, several works adopt static–dynamic decomposition. **DynNeRF** [56] and **Video-NeRF** [55] incorporate monocular depth priors to regularize geometry and appearance, enabling stable free-viewpoint rendering of dynamic content. **DetNeRF** [64] further extends this by employing occlusion-aware modeling to explicitly separate static backgrounds from moving components. The efficacy of these methods is strictly bounded by the quality of external priors

For long sequences and complex scenes with multiple moving agents, **DecouplingNeRF** [63] and **ML-NSG** [62] use hierarchical neural scene graphs to decompose scenes into object-centric components for scalable reconstruction. However, the hierarchical design introduces additional complexity, and these methods may struggle with objects exhibiting unpredictable motion or frequent occlusions.

Recent advances increasingly focus on modeling motion dynamics and incorporating non-RGB sensors. **Sync-NeRF** [60] and **MonoNeRF** [59] learn implicit velocity fields and feature correspondences to improve temporal alignment and robustness. Beyond RGB inputs, **4D-NDF** [61] models LiDAR sequences using time-dependent signed distance functions (SDFs) to jointly reconstruct static structures and dynamic objects. However, velocity-based methods remain sensitive to temporal aliasing, while multimodal approaches are affected by modality-specific noise.

3.4 4D Feature-Volume-Based Methods

To improve efficiency, these methods replace implicit neural fields with structured representations that factorize the 4D space-time domain into lower-dimensional components. Common strategies include tensor decomposition, planar factorization, and multi-resolution grids, enabling faster training and rendering with reduced memory.

To reduce the computational cost of coordinate-based MLPs, several methods decompose 4D space into lower-dimensional representations. **HexPlane** [65] and **K-Planes** [66] project 4D volumes onto orthogonal 2D planes, while **HyperReel** [69] and **MixVoxels** [70] combine voxel grids with ray-conditioned sampling and deformation modeling. These factorizations enable efficient high-quality rendering through structured memory layouts, but often involve a trade-off between memory usage and spatial resolution.

To ensure smooth transitions between temporal snapshots, several methods incorporate temporal interpolation and frequency-aware modeling. **TID-NeRF** [68] integrates temporal modeling into explicit representations, while **BLiRF** [74] models radiance fields as band-limited signals using neural trajectory bases and low-rank spatial decomposition. **NeRFPlayer** [73] decomposes scenes into static and dynamic components with sliding-window temporal encoding. **Ced-NeRF** [75] improves generalization with hybrid grid-based representations, and **DaReNeRF** [77] encodes temporal information using direction-aware wavelet representations. However, high-frequency motion and abrupt topological changes are often blurred by the underlying spectral constraints or interpolation schemes.

For large-scale scenes, hash-based encodings improve scalability. **SUDS** [67] and **MSTH** [71] utilize multiresolution hash grids with uncertainty-aware masking to handle complex urban dynamics, while **NeuRAD** [79] and **RoDUS** [81] incorporate rolling shutter compensation and semantic signals for robust performance in autonomous driving scenarios. **EmerNeRF** [82] and **LiDAR4D** [76] further extend these ideas to multimodal settings by using learned flow and temporal feature slots to aggregate information across sparse frames. However, hash-based representations may suffer from feature collisions in complex scenes, potentially introducing aliasing artifacts or geometric noise.

Hybrid representations support specialized tasks by incorporating domain-specific inductive biases. **NVFi** [72] introduces physics-informed velocity fields for motion transfer and future-state prediction, while **Gear-NeRF** [78] leverages semantic priors for object-level tracking and motion-aware sampling. Methods such as **S-DyRF** [80] further enable temporally consistent stylization in dynamic scenes. However, the reliance on specialized priors can limit generalization across diverse scenarios.

3.5 Temporal-Prior-Based Methods

While previous approaches explicitly parameterize time within the radiance field, practical reconstruction often benefits from external temporal cues used as constraints or guidance. These methods leverage signals such as optical flow, physical constraints, or sequential modeling to regularize the under-constrained dynamic reconstruction problem. To support real-time and streaming applications, several frameworks replace global optimization with temporally incremental updates. **StreamRF** [84], for example, employs a grid-based representation with sequential updates, modeling temporal evolution through incremental changes to a base model and enabling efficient streaming via difference-based compression. However, these methods are prone to error accumulation, where small update misalignments propagate over long sequences.

For scenes with complex stochastic motion, **OTNeRF** [85] enforces temporal consistency by modeling scene dynamics as low-frequency shifts in pixel distributions. However, such statistical regularization often suppresses high-frequency geometric details. In scenarios with sparse observations or severe occlusions, incorporating physics-based or geometric priors from non-RGB sensors becomes important. **STGC-NeRF** [86] introduces spatio-temporal geometric constraints for LiDAR-based reconstruction, using scene flow to establish cross-frame correspondences and improve robustness under sparse inputs. However, the effectiveness of these constraints depends heavily on motion estimation quality, and large inter-frame displacements can introduce temporal aliasing in high-speed scenes.

4 DYNAMIC 3D GAUSSIAN SPLATTING

The core challenge in extending 3DGS to the temporal domain lies in parameterizing dynamic scene evolution [87]. Existing 4D reconstruction frameworks differ in how they model time, ranging from explicit geometric embedding to implicit motion modeling. As shown in Fig. 4 and Table 3,

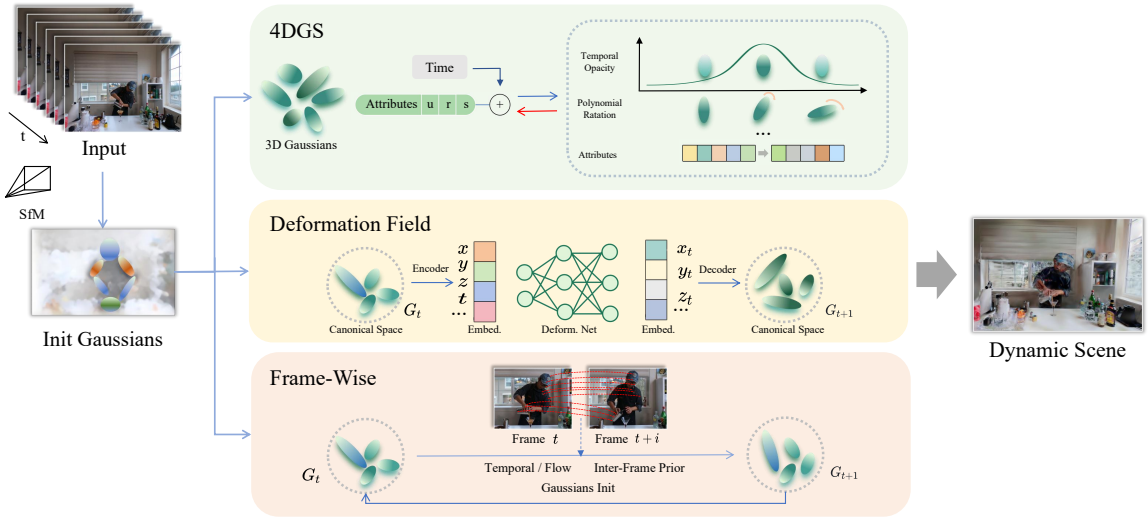

Fig. 4. General pipeline of **3DGS-style** 4D scene reconstruction methods. The pipeline presents the representative 4D strategies in explicit 4D primitive-based, deformation-field-based, and frame-wise-training frameworks.

these approaches can be grouped into three paradigms: (i) **Explicit 4D Primitive-Based Methods**, which treat time as a geometric dimension; (ii) **Deformation Field-Based Methods**, which separate geometry and motion via learned transformations; and (iii) **Frame-wise Training Methods**, which optimize discrete states for temporal consistency. These paradigms involve trade-offs in flexibility, efficiency, and temporal coherence.

4.1 Fundamentals of 4DGS

The standard 3DGS framework represents scenes using static primitives, limiting its applicability to stationary environments [88]. Extending to the 4D spatio-temporal domain requires modeling the evolution of scene attributes—position, rotation, and appearance—over time. Existing methods differ in how this temporal evolution is formulated.

Explicit 4D Primitive-Based Methods treat time as an intrinsic dimension and represent scenes using 4D Gaussian primitives. Each primitive is parameterized by a mean $\boldsymbol{\mu} \in \mathbb{R}^4$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{4\times4}$, decomposed into rotation and scaling:

$$\boldsymbol{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^\top\mathbf{R}^\top, \quad \mathbf{S} = \mathrm{diag}(s_x, s_y, s_z, s_t), \tag{10}$$

where the rotation $\mathbf{R}$ couples spatial and temporal dimensions. To render a frame at time t, the 4D Gaussian is sliced into a 3D Gaussian via the conditional distribution $p(\mathbf{x} \mid t)$. By partitioning $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ into spatial $(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$, temporal (μ_t, Σ_{tt}), and cross terms $(\boldsymbol{\Sigma}_{xt})$, the resulting 3D Gaussian parameters are:

$$\boldsymbol{\mu}' = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xt}\Sigma_{tt}^{-1}(t-\mu_t), \quad \boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xt}\Sigma_{tt}^{-1}\boldsymbol{\Sigma}_{xt}^\top. \tag{11}$$

This formulation enables integration with standard splatting while maintaining temporal consistency.

Deformation Field-Based Methods decouple temporal dynamics from scene geometry by maintaining static 3D Gaussians in a canonical space $\mathcal{G}_{\text{can}}$. A learnable deformation network $\mathcal{F}_\theta$ maps canonical coordinates and time to attribute offsets. Given a query point $\mathbf{p}$ and time t, the network predicts:

$$(\Delta\boldsymbol{\mu}, \Delta q, \Delta s) = \mathcal{F}_\theta(\mathbf{p}, t), \tag{12}$$

where $\mathbf{p}$ is typically the Gaussian center $\boldsymbol{\mu}$ [89]. The deformed Gaussian set at time t is:

$$\mathcal{G}(t) = \{\boldsymbol{\mu} + \Delta\boldsymbol{\mu},\ q \otimes \Delta q,\ s + \Delta s,\ \alpha,\ c\}, \tag{13}$$

where $\otimes$ denotes quaternion multiplication. Rendering is performed using standard differentiable splatting.

Frame-Wise Training Methods optimize Gaussian parameters independently at each timestamp, often guided by priors such as rigidity or optical flow. The state of Gaussian i at time t is updated from the previous frame:

$$\boldsymbol{\mu}_{i,t} = \mathbf{R}_{i,t}\boldsymbol{\mu}_{i,t-1} + \mathbf{T}_{i,t}, \quad q_{i,t} = q(\mathbf{R}_{i,t}) \otimes q_{i,t-1}, \tag{14}$$

where $\mathbf{R}_{i,t}$ and $\mathbf{T}_{i,t}$ denote local rotation and translation. To handle topology changes, the Gaussian set is dynamically updated:

$$\mathcal{G}_t = \hat{\mathcal{G}}_t \cup \mathcal{G}_{\text{new}}, \tag{15}$$

where $\hat{\mathcal{G}}_t$ denotes propagated Gaussians and $\mathcal{G}_{\text{new}}$ newly initialized ones.

4.2 Explicit 4D Primitive-Based Methods

These methods extend 3D Gaussian Splatting by incorporating time directly into Gaussian primitives. Combined with temporal slicing and tile-based rasterization, they achieve high reconstruction quality and temporal coherence.

4DGS [90] generalizes 3DGS to 4D by treating time as an additional dimension. It uses 4D scaling and dual quaternions for rotation, and renders by slicing 4D Gaussians into conditional 3D Gaussians.

Subsequent works refine spatial–temporal disentanglement. **4D-RotorGS** [91] adopts rotor-based rotation with entropy regularization, while **SpaceTimeGS** [92] models topology changes using temporal basis functions and compact feature decoding.

To better capture motion, **CD-3DGS** [93] and **FreeTimeGS** [94] parameterize motion using Fourier or linear functions, with additional supervision such as optical flow. **SplatFlow** [95] further incorporates learned motion flow fields. In order to improve robustness, **DriveDreamer4D** [96] and **DeSiRe-GS** [97] introduce external supervision,

including synthetic data and motion-aware decomposition, to handle occlusions and sparse observations.

Explicit 4D primitive-based methods directly embed time into Gaussian parameterization, achieving strong temporal consistency through conditional slicing of 4D Gaussians into 3D counterparts [90]. Combined with tile-based rasterization, this formulation enables efficient real-time rendering. Extensions such as temporal basis functions [92] and rotor-based rotation with entropy regularization [91] further improve spatio-temporal disentanglement and topology modeling. However, the higher per-primitive parameter dimensionality introduces substantial memory overhead, as each Gaussian requires a 4D mean and a 4×4 covariance matrix. These methods are particularly suitable for indoor dynamic scenes with dense multi-view capture and real-time rendering requirements.

4.3 Deformation-Field-Based Methods

Deformation field-based methods model dynamics by applying a learnable deformation field to a canonical set of 3D Gaussians, typically implemented as an MLP. This approach decouples appearance from motion by predicting only the temporal changes in position, rotation, and scale, while assuming other Gaussian attributes remain fixed.

Deformable 3D-GS [89] introduces deformation fields into 3DGS by mapping Gaussians to a canonical space and using an MLP to predict position, scale, and rotation offsets, with annealing to reduce rendering jitter. Subsequent works improve efficiency and structural consistency through sparse control. **SC-GS** [104] and **SP-GS** [118] drive dense Gaussians using sparse control points and superpoints, respectively, while **GaussianPrediction** [109] employs graph convolutional networks on clustered key points for motion prediction. Feature encoding is further optimized by **4D-GS** [102], which uses K-Planes with voxel encoding, and **GaGS** [105], which combines point-based MLPs with voxel U-Nets for geometry-aware features.

To reduce artifacts from global deformation, several methods separate static and dynamic components. **GauFRe** [143] and **SWinGS** [108] restrict deformation to dynamic regions. **HUGS** [101] models static backgrounds and dynamic objects with separate parameterizations. **Gflow** [119] and **EfficientGS** [144] further incorporate priors such as depth and optical flow to localize deformable regions.

Hierarchical methods address multi-scale dynamics. **Grid4D** [112] decomposes spatiotemporal encoding using hash grids and attention. **MoDec-GS** [123] and **Hicom** [115] adopt global-to-local cascades, while **ADC-GS** [132] uses anchor-driven deformation to combine transformations with local refinement.

In addition, external priors and structured models are introduced to regularize deformation. **MotionGS** [116] and **MoDGS** [131] use optical flow decomposition. **BARD-GS** [122] and **4D-GS Wild** [114] address challenging scenarios with pose interpolation and diffusion-based regularization. **TaylorGaussian** [129], **Gaussian-Flow** [100], and **SplineGS** [124] model motion using analytic representations, while **FreeGave** [125] enforces physical constraints.

Specialized designs target specific scenarios. **SpectroMotion** [130] handles specular materials, while **GIFStream** [126] and **MoSca** [128] enable efficient streaming. **Marbles** [111] and **MonoFusion** [145] address monocular settings with simplified primitives and motion models. **OmniRe** [133] adopts neural scene graphs to separate static and dynamic components under a shared deformation field.

Deformation-field-based methods decouple temporal dynamics from canonical geometry by predicting per-Gaussian attribute offsets through a learnable deformation network [89], offering improved parameter efficiency over explicit 4D primitives. This formulation is particularly effective for modeling non-rigid motion and view-dependent specularities. However, MLP-based deformation prediction may introduce rendering jitter, while the additional network inference can limit rendering speed. As one of the most widely explored paradigms, deformation-field-based methods are well-suited for scenes with non-rigid motion and complex appearance changes, and naturally support downstream tasks such as scene editing [104], [127] and neural scene graph decomposition [133].

4.4 Frame-Wise Training Methods

Frame-wise methods optimize 3D Gaussians independently at each timestamp via per-frame reconstruction, optionally incorporating inter-frame constraints for temporal consistency. These approaches are simple and flexible but often incur high storage cost and limited long-term coherence.
D-3DG [134] models Gaussian centers and orientations as time-varying, while keeping color, opacity, and scale fixed. Motion is guided by rigidity and rotation priors. However, independent optimization can lead to weak temporal consistency and high storage overhead.

To reduce redundancy in complex scenes, many methods adopt dynamic–static decomposition. **Street Gaussians** [137] and **DrivingGaussian** [138] use graph-based representations to separate static backgrounds and dynamic objects [146]. **Casual-FVS** [139] decomposes scenes into static planes and dynamic points with flow-based blending, while **LiveSplats** [147] employs hierarchical optimization for real-time processing.

For online scenarios with topology changes, **3DGStream** [135] introduces a Neural Transformation Cache to transform existing Gaussians and incrementally add new ones. **4D-Fly** [141] propagates Gaussians across frames using anchor-based updates, expanding the representation only when needed.

In order to reduce storage, interpolation-based methods use keyframes. **Ex4DGS** [140] and **4DGC** [40] reconstruct intermediate frames via interpolation or motion prediction. **Ex4DGS** applies CHip and Slerp for trajectory smoothing, while **4DGC** uses multi-resolution motion grids for efficient transformation estimation. On the other hand, several methods incorporate strong supervision to enforce temporal consistency. **GaussianFlow** [148] enforces consistency between 3D motion and 2D observations via optical flow. **MAGS** [142] further improve supervision using dense correspondences and uncertainty-aware flow modeling.

Frame-wise methods provide architectural simplicity and flexibility by optimizing 3D Gaussians independently at each timestamp, with dynamic set updates ($\mathcal{G}_t = \hat{\mathcal{G}} * t \cup \mathcal{G} * \text{new}$) that naturally handle topology changes [135],

TABLE 3
Overview of 3DGS-based 4D dynamic scene reconstruction methods. Methods are categorized into three types. For each method, we summarize its key components and additional priors.

Method	Venue	Input	Scenario	Target Domain	4D-style	Text	Flow	Normal	Segment.	Extra Prior
4DGS [90]	ICLR2024	RGB	Indoor	Multi.-Centric	Explicit 4D Primitive					
CD-3DGS [93]	ECCV2024	RGB	Indoor	Entity-Centric	Explicit 4D Primitive		✓			RAFT
SpacetimeGS [92]	CVPR2024	RGB	Indoor	Entity-Centric	Explicit 4D Primitive					
4DRotorGS [91]	ACM SIG. 2024	RGB	Indoor	Entity-Centric	Explicit 4D Primitive		✓			
FreeTimeGS [94]	CVPR2025	RGB	Indoor	Multi.-Centric	Explicit 4D Primitive					ROMA
DeSiRe-GS [97]	CVPR2025	RGBD	Auto. Driving	Scene-Centric	Explicit 4D Primitive			✓		
SplatFlow [95]	CVPR2025	RGBD	Auto. Driving	Scene-Centric	Explicit 4D Primitive		✓		✓	RAFT
DriveDreamer4D [96]	CVPR2025	RGBD	Auto. Driving	Scene-Centric	Explicit 4D Primitive					
ST-4DGS [98]	ACM SIG. 2024	RGB	In. & Wild	Entity-Centric	Deformation Fields		✓			RAFT
SaRO-GS [99]	ACM MM2024	RGB	Indoor	Entity-Centric	Deformation Fields					
GaussianFlow [100]	CVPR2024	RGB	Indoor	Entity-Centric	Deformation Fields		✓			Videoflow
HUGS [101]	CVPR2024	RGB	Auto. Driving	Scene-Centric	Deformation Fields		✓			Unimatch
4D-GS [102]	CVPR2024	RGB	Indoor	Entity-Centric	Deformation Fields					
DeformGS [103]	CVPR2024	RGB	Indoor	Entity-Centric	Deformation Fields				✓	
SC-GS [104]	CVPR2024	RGB	In. & Wild	Entity-Centric	Deformation Fields					
Deformable-3DGS [89]	CVPR2024	RGB	Indoor	Entity-Centric	Deformation Fields					
GaGS [105]	CVPR2024	RGB	In. & Wild	Entity-Centric	Deformation Fields					
DynMF [106]	ECCV2024	RGB	Indoor	Entity-Centric	Deformation Fields					
ED-3DGS [107]	ECCV2024	RGB	In. & Wild	Multi.-Centric	Deformation Fields					
SwinGS [108]	ECCV2024	RGB	Indoor	Entity-Centric	Deformation Fields		✓			RAFT
GSPrediction [109]	ACM SIG.2024	RGB	Indoor	Entity-Centric	Deformation Fields					
AmbientGaussian [110]	ACM SIG. 2024	RGB	In-the-Wild	Scene-Centric	Deformation Fields					
Marbles [111]	SIG-ASIA 2024	RGB	Indoor	Entity-Centric	Deformation Fields				✓	Trackanything
Grid4D [112]	NIPS2024	RGB	Indoor	Entity-Centric	Deformation Fields				✓	
Vidu4D [113]	NIPS2024	RGB	Indoor	Entity-Centric	Deformation Fields	✓				
4D-GS Wild [114]	NIPS2024	RGB	In. & Wild	Multi.-Centric	Deformation Fields	✓	✓			RAFT & BLIP & Stable Diffusion
HiCoM [115]	NIPS2024	RGB	Indoor	Entity-Centric	Deformation Fields					
MotionGS [116]	NIPS2024	RGB	Indoor	Multi.-Centric	Deformation Fields		✓			Gaussianflow
DN-4DGS [117]	NIPS2024	RGB	In. & Wild	Entity-Centric	Deformation Fields					
SP-GS [118]	ICML2024	RGB	Indoor	Entity-Centric	Deformation Fields					SuperPoint
Gflow [119]	AAAI2025	RGB	In-the-Wild	Scene-Centric	Deformation Fields		✓			DUSt3R & UniMatch
EfficientGS [120]	AAAI2025	RGB	Indoor	Multi.-Centric	Deformation Fields		✓			COLMAP
Instant GS [121]	CVPR2025	RGB	Indoor	Entity-Centric	Deformation Fields		✓			GM-Flow
BARD-GS [122]	CVPR2025	RGB	Indoor	Entity-Centric	Deformation Fields					
MoDec-GS [123]	CVPR2025	RGB	In. & Wild	Multi.-Centric	Deformation Fields					
SplineGS [124]	CVPR2025	RGB	In-the-Wild	Scene-Centric	Deformation Fields					
FreeGave [125]	CVPR2025	RGB	Indoor	Entity-Centric	Deformation Fields				✓	SAM
GIFStream [126]	CVPR2025	RGB	Indoor	Entity-Centric	Deformation Fields					
Instruct-4DGS [127]	CVPR2025	RGB	Indoor	Entity-Centric	Deformation Fields	✓				InstructPix2Pix
MoSca [128]	CVPR2025	RGB	In-the-Wild	Entity-Centric	Deformation Fields		✓			RAFT
TaylorGaussian [129]	CVPR2025	RGB	Indoor	Entity-Centric	Deformation Fields					
SpectroMotion [130]	CVPR2025	RGB	Indoor	Entity-Centric	Deformation Fields			✓		
MoDGS [131]	ICLR2025	RGBD	Indoor	Entity-Centric	Deformation Fields		✓		✓	RAFT & SAM2
ADC-GS [132]	IJCAI2025	RGB	In. & Wild	Entity-Centric	Deformation Fields					
Omnire [133]	ICLR2025	RGBD	Auto. Driving	Scene-Centric	Deformation Fields					
D-3DG [134]	3DV2024	RGB	Indoor	Entity-Centric	Frame-wise training					
3DGStream [135]	CVPR2024	RGB	Indoor	Entity-Centric	Frame-wise training					
NPGs [136]	CVPR2024	RGB	Indoor	Entity-Centric	Frame-wise training					
StreetGaussian [137]	ECCV2024	RGBD	Auto. Driving	Scene-Centric	Frame-wise training				✓	Video K-Net
DrivingGaussian [138]	CVPR2024	RGBD	Auto. Driving	Scene-Centric	Frame-wise training					
Casual-FVS [139]	ECCV2024	RGB	In-the-Wild	Scene-Centric	Frame-wise training		✓		✓	RAFT & SAM
Ex4DGS [140]	NIPS2024	RGB	Indoor	Entity-Centric	Frame-wise training					
4DGC [40]	CVPR2025	RGB	Indoor	Entity-Centric	Frame-wise training					
4D-Fly [141]	CVPR2025	RGB	Indoor	Entity-Centric	Frame-wise training					
MAGS [142]	TCSVT2025	RGB	Indoor	Entity-Centric	Frame-wise training		✓			RAFT

[141]. This design is particularly suitable for online and streaming scenarios. However, the independent per-frame parameterization results in storage costs that grow linearly with sequence length, while long-range temporal coherence remains limited without explicit inter-frame coupling. Keyframe interpolation [40], [140] and flow-supervised consistency losses [142] partially alleviate these issues, but introduce approximation errors or reliance on external priors. Consequently, this paradigm is best suited for short sequences, casual monocular videos [139], and streaming reconstruction tasks where real-time adaptability is prioritized over long-term temporal consistency [149], [150].

5 Performance Evaluation

In this section, we summarize representative datasets for dynamic scene synthesis, categorizing them according to key properties and research objectives (Table 4). In addition, we explore novel view synthesis and geometric reconstruction in representative benchmarks, highlighting the best results as **first**, second, and third. We organize quantitative data from papers with a common evaluation protocol and cross-verified results. Since some works do not release codes or specific configurations, our priority is to include papers with consistent benchmarks, ensuring a reliable basis for verifiable comparison with a shared evaluation framework across multiple sources.

5.1 Benchmark Datasets

5.1.1 *Synthetic vs. Real-world Data*

Synthetic: D-NeRF [35] and **ParticleNeRF** [151] provide clean articulated and deformable scenes, making them suitable for evaluating deformation modeling and motion tracking. In autonomous driving, **SS3DM** [152] offers synchronized RGB, LiDAR, and semantic annotations, enabling controlled evaluation of multimodal fusion methods. However,

TABLE 4
Taxonomy of dynamic scene datasets based on benchmark properties. Datasets are grouped by their primary research focus and capture characteristics.

Dataset	Scene Type	Sensor Setup	Resolution	Frame Rate	Scene/Seq	Temporal Scale
Synthetic Datasets (Ground Truth Geometry/Motion)						
D-NeRF [35]	Indoor	1 Cam	800×800	–	8	50–200 frames
ParticleNeRF [151]	Indoor	40 Cams	–	–	6	–
SS3DM [152]	Autonomous Driving	6 Cams + 5 LiDAR	–	10 FPS	28	13K frames
Real-world: Monocular & Sparse View						
DAVIS [153]	Outdoor	1 Cam	–	–	150	10k frames total
HyperNeRF [37]	Indoor/Outdoor	1–2 Cams	540×960	15 FPS	17	8–15s/seq
DyCheck [154]	Indoor	1 iPhone + 7 Static	–	–	14	200–500 frames
Stereo4D [155]	Indoor/Outdoor	2 Cams	Diverse	–	200K clips	–
NeRF-DS [51]	Outdoor	2 Cams	–	–	8	–
Real-world: Dense Multi-view & Human-Centric						
Panoptic Studio [156]	Indoor	480 Cams	–	–	5	–
ENeRF-Outdoor [157]	Outdoor	18 Cams	–	–	3	1200 frames
Neu3DV [58]	Indoor	18–21 Cams	2704×2028	30 FPS	6	10s/seq
Technicolor [158]	Indoor (RGB-only)	16 Cams	2048×1088	25 FPS	12	–
NVIDIA Dynamic [159]	Outdoor	12 Cams	–	–	7	90–200 frames
Meeting Room [160]	Indoor	13 Cams	1280×720	30 FPS	3	300 frames
Google Immersive [161]	Indoor/Outdoor	≤46 Cams	–	–	15	–
Real-world: Long Horizon & Multimodal						
Waymo Open [162]	Autonomous Driving	5 Cams + 5 LiDAR	1920×1280	10 FPS	1150	∼12M frames
nuScenes [163]	Autonomous Driving	6 Cams + LiDAR	1600×900	12 FPS	1000	∼5.5h
Argoverse 2 [164]	Autonomous Driving	7 Cams + LiDAR	1920×1200	30 FPS	1000	∼1000h
PandaSet [165]	Autonomous Driving	6 Cams + LiDAR	1920×1080	20 FPS	103	∼1h
OmniHD-Scenes [166]	Autonomous Driving	6C + LiDAR + 6R	Diverse	15 FPS	1501	∼30s/seq
KITTI [167]	Autonomous Driving	2 Stereo + LiDAR	1242×375	10 FPS	22	∼6h
WayveScenes101 [168]	Autonomous Driving (RGB-only)	5 Cams	–	10 FPS	101	20s/seq

limited domain diversity and simplified rendering reduce their ability to assess real-world robustness.

Real-world: HyperNeRF [37] and **Technicolor** [158] introduce complex lighting, calibration errors, and dynamic backgrounds. These datasets are more suitable for evaluating generalization and robustness, but often lack accurate ground truth, making quantitative evaluation more challenging.

5.1.2 Monocular vs. Multi-view Capture

Monocular: D-NeRF [35], **HyperNeRF** [37], and **DAVIS** [153] contain sequences captured from a single moving camera. These benchmarks are well-suited for evaluating methods that rely on strong priors, such as generative models or deformation-aware representations. However, monocular setups often suffer from scale ambiguity and limited spatial coverage.

Multi-view: Panoptic Studio [156] and **Google Immersive** [161] provide dense spatial coverage, making them suitable for evaluating reconstruction fidelity and view consistency. Intermediate-scale datasets such as **NeRF-DS** [51] and **KITTI** [167] instead offer sparse multi-view setups that better reflect real-world constraints and are useful for evaluating view-sparse reconstruction methods.

5.1.3 Short-term vs. Long-term Horizons

Short Horizon: Neu3DV [58] and **Meeting Room** [160] typically contain short clips with localized motions. These datasets are well-suited for evaluating deformation modeling and short-term motion consistency, but may not capture long-range dynamics.

Long Horizon: Argoverse 2 [164] and **nuScenes** [163] provide large-scale temporal data across diverse driving environments. These benchmarks are useful for evaluating temporal consistency and long-term prediction, although sparse viewpoints and noisy annotations introduce additional challenges.

5.1.4 Human-Centric vs. General Dynamic Scenes

Human-Centric: Panoptic Studio [156] and **ENeRF-Outdoor** [157] focus on articulated human motion. These datasets are particularly suitable for evaluating deformation modeling, skeletal motion tracking, and fine-grained geometry reconstruction.

General Dynamics: NVIDIA Dynamic Scene [159] and **Stereo4D** [155] include diverse dynamic elements such as animals, fluids, and object interactions. These datasets better reflect real-world complexity, but introduce greater challenges for motion decomposition and scene understanding.

5.1.5 RGB-only vs. Multimodal Datasets

RGB-only: Technicolor [158] and **WayveScenes101** [168] rely solely on visual inputs. These benchmarks are suitable for evaluating appearance modeling, but suffer from depth ambiguity and sensitivity to lighting variations.

Multimodal: Waymo Open Dataset [162], **PandaSet** [165], and **OmniHD-Scenes** [166] integrate LiDAR, radar, and IMU signals, making them suitable for evaluating geometric accuracy and robust reconstruction under challenging conditions. **DyCheck** [154] further incorporates smartphone LiDAR, enabling evaluation in lightweight capture settings.

5.2 Evaluation Metrics

Evaluation of 4D reconstruction involves three primary aspects: rendering quality, geometric accuracy, and computational efficiency. We adopt widely used metrics in the

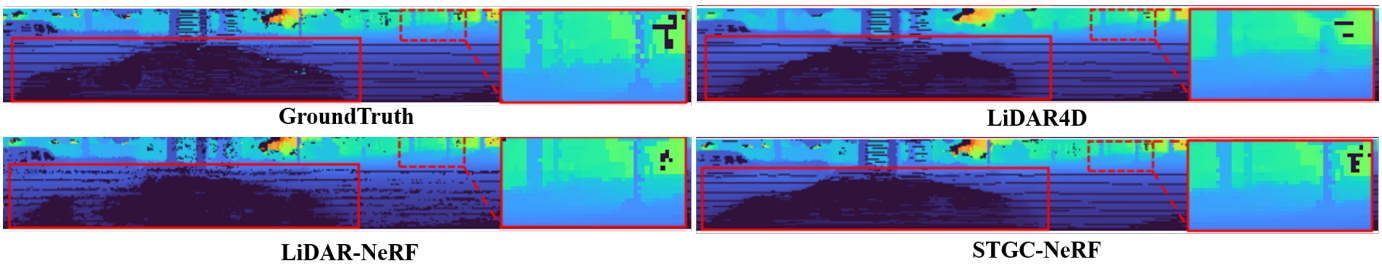

Fig. 5. **Qualitative reconstruction point map and depth of NeRF-style methods on the NuScenes [163] dataset.** Image from [86].

literature and further analyze their strengths and limitations for dynamic 4D reconstruction.

5.2.1 View Synthesis Metrics

PSNR measures reconstruction fidelity through pixel-wise error. Although widely used for novel view synthesis, it primarily evaluates per-frame appearance quality and does not fully capture temporal consistency in dynamic scenes.
SSIM evaluates perceptual similarity in terms of structure and contrast. Compared to PSNR, it better reflects structural preservation, but remains a frame-wise metric without explicit temporal modeling.
LPIPS measures perceptual similarity using deep feature representations. While it correlates better with human perception, it mainly evaluates appearance quality rather than dynamic geometric accuracy.

5.2.2 Geometric and Spatiotemporal Metrics

Chamfer Distance (CD) measures geometric similarity between predicted and ground-truth point sets. It is widely used to evaluate geometric fidelity, but remains sensitive to point density and may not adequately capture topology changes in dynamic scenes.
F-Score evaluates reconstruction quality through precision and recall under a distance threshold. It provides a more balanced assessment of geometric accuracy, although the results can vary with threshold selection.
RMSE measures the error between predicted and ground-truth geometry. While it offers a direct measure of geometric accuracy, it is sensitive to outliers and may not fully reflect perceptual quality.

5.2.3 Efficiency and Practicality

FPS measures inference speed and computational efficiency. However, FPS alone does not capture other practical factors such as memory consumption and scalability, which are also critical in 4D reconstruction.

Despite these limitations, these metrics remain widely adopted in existing 4D reconstruction works and provide a common basis for comparison. However, the lack of standardized protocols for evaluating temporal coherence, topology changes, and long-term consistency remains an open challenge in dynamic 4D reconstruction.

5.3 Novel View Synthesis

We evaluate rendering fidelity using PSNR, D-SSIM, SSIM, and LPIPS. Benchmarks are conducted on widely used datasets, including Neu3D [58], D-NeRF [35], NeRF-DS [51], NVIDIA Dynamic Scene [159], and Waymo [162].

TABLE 5
Neu3D [58] NeRF-style 4D reconstruction results. PSNR (↑), SSIM (↑), and LPIPS (↓) are used as metrics.

Methods	PSNR (↑)	SSIM (↑)	LPIPS (↓)
DyNeRF	29.6	0.961	0.083
StreamRF	28.3	-	-
HexPlane	29.5	-	0.097
K-Planes	31.6	**0.964**	-
TIDNeRF	29.9	-	0.096
HyperReel	31.1	0.927	0.096
MixVoxels	31.7	0.944	0.064
MSTH	**32.4**	-	**0.056**
NeRFPlayer	30.7	0.931	0.111
Sync-NeRF	31.9	0.916	0.146
DecouplingNeRF	28.6	0.917	0.123
Ced-NeRF	30.6	0.919	-
Gear-NeRF	31.8	0.936	0.058
DaReNeRF	32.3	-	0.084

TABLE 6
Neu3D [58] 3DGS-style 4D reconstruction results. PSNR (↑), SSIM (↑), and LPIPS (↓) are used as the evaluation metrics.

	PSNR (↑)	SSIM (↑)	LPIPS (↓)
4DGS	32.01	-	0.055
4DRotorGS	31.62	0.940	0.140
FreeTimeGS	**33.19**	-	**0.036**
SpacetimeGS	32.05	-	0.044
CD-3DGS	30.46	0.955	0.150
SaRO-GS	32.15	-	0.044
ST-4DGS	32.67	0.946	0.166
4DGC	31.58	0.943	-
ADC-GS	31.67	**0.981**	0.061
GIFStream	31.75	0.938	0.051
TaylorGaussian	33.02	0.970	0.053
GaGS	31.31	0.950	0.140
DynMF	31.70	0.946	0.180
DN-4DGS	32.02	0.944	0.043
ED-3DGS	31.31	0.945	0.037
Ex4DGS	32.11	0.970	0.048
MAGS	31.30	0.943	0.053

Neu3D. Table 5 reports NeRF-style results under the protocol of [58]. Performance steadily improves from early implicit models to recent hybrid approaches. DyNeRF establishes a strong baseline, while MSTH and DaReNeRF further advance the state of the art. A notable trend is the strong performance of 4D feature-volume-based architectures, which consistently achieve leading results across multiple metrics. In addition, specialized designs highlight the benefits of structured priors; for example, the uncertainty-aware modeling in MSTH and the semantic segmentation constraints in Gear-NeRF demonstrate the effectiveness of incorporating probabilistic cues and geometric semantics into dynamic scene optimization.

Table 6 presents results for 3DGS-based methods. Recent approaches, such as FreeTimeGS and TaylorGaussian,

TABLE 7
D-NeRF [35] NeRF-style 4D reconstruction results. PSNR (↑), SSIM (↑), and LPIPS (↓) are used as metrics.

Methods	PSNR (↑)	SSIM (↑)	LPIPS (↓)
D-NeRF	30.50	0.95	0.070
TiNeuVox	32.67	0.97	0.041
HexPlane	31.04	0.97	0.040
K-Planes	31.61	0.97	0.049
TIDNeRF	32.73	0.97	0.033
Ced-NeRF	34.21	**0.99**	0.037
DaReNeRF	31.95	0.97	0.030
SLS4D	**34.84**	0.98	**0.025**

TABLE 8
NeRF-DS [51] 3DGS-style 4D reconstruction results. PSNR (↑), SSIM (↑), and LPIPS (↓) are used as metrics.

Methods	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Deformable-3DGS	24.10	0.85	0.18
SC-GS	24.10	0.89	**0.14**
4D-GS	24.18	0.88	**0.14**
SP-GS	23.33	0.84	0.21
MotionGS	24.54	0.87	0.17
DN-4DGS	24.36	0.87	0.17
EfficientGS	**24.65**	**0.90**	**0.14**

outperform earlier methods in PSNR while maintaining strong perceptual quality. Deformation-based methods also perform competitively: ADC-GS achieves the best SSIM, and ED-3DGS yields strong LPIPS scores. A key observation from the evaluation is the performance improvement enabled by integrating external priors. In addition, specialized frameworks highlight the effectiveness of structured constraints; for example, the feature matching priors in FreeTimeGS, which achieves high rendering fidelity, and the optical flow constraints used in ST-4DGS and CD-3DGS demonstrate the benefits of external guidance for dynamic scene optimization. These priors effectively regularize Gaussian primitives in highly dynamic regions, reducing floaters and multi-view inconsistencies commonly observed in purely photometric optimization. Figure 8 shows qualitative comparisons on Neu3D.

D-NeRF. Table 7 reports reconstruction quality under the protocol of [35], showing that 4D feature-volume-based methods consistently achieve state-of-the-art performance across diverse dynamic benchmarks. While earlier approaches such as TiNeuVox establish strong baselines, Ced-NeRF and SLS4D leverage semi-explicit representations and high-dimensional feature volumes to better disentangle static geometry from temporal dynamics. This trend reflects a broader shift in modeling small-scale dynamic indoor scenes, where grid-based feature structures outperform purely coordinate-based MLPs.

NeRF-DS. Table 8 reports rendering performance under the protocol of [51]. EfficientGS achieves the best results, while motion-aware methods such as MotionGS and DN-4DGS perform strongly on dynamic regions. This trend highlights the effectiveness of deformation-field-based representations for explicit 4D modeling. By decoupling temporal motion from canonical geometry, these methods avoid the parameter growth associated with unified 4D primitives. Moreover, this formulation is well-suited for modeling complex non-rigid dynamics and view-dependent specularities, which are particularly challenging in the NeRF-DS dataset.

NVIDIA Dynamic Scene Dataset. Qualitative evaluations in unstructured, in-the-wild environments (Fig. 7) highlight the strong performance of frameworks incorporating geometric priors. DynNeRF maintains superior free-view consistency and temporal stability, largely due to its use of multi-view constraints and 3D scene flow for regularization. This demonstrates the effectiveness of geometric priors in dynamic scene reconstruction.

TABLE 9
NuScenes [163] 3D geometric reconstruction results. * denotes methods with LiDAR supervision; † uses protocols from [133].

methods	CD (↓)	F-Score (↑)	RMSE (↓)
NeRF-style			
D-NeRF	0.33	0.85	7.11
TiNeuVox-B	0.39	0.86	7.21
K-Planes	0.30	0.89	6.80
LiDAR4D*	0.24	0.89	6.78
STGC-NeRF*	**0.22**	**0.91**	**6.54**
3DGS-style			
Deformable-3DGS†	0.38	-	2.97
StreetGaussian†	0.27	-	2.19
OmniRE†	**0.24**	-	**1.89**

Waymo. As illustrated in Figure 6, general-purpose 4D reconstruction methods often struggle with distant or fast-moving objects in large-scale driving scenes, leading to ghosting artifacts or geometric collapse. In contrast, EmerNeRF achieves more robust results by integrating 3D scene flow with DINOv2 features, providing a more stable optimization signal and improved robustness to lighting variations. OmniRE further delivers high-fidelity reconstruction by incorporating category-specific semantic priors for human modeling, effectively constraining the solution space to physically plausible structures. These results highlight the importance of semantic guidance and structural priors for 4D reconstruction under limited viewpoint overlap.

5.4 Geometric Reconstruction

We evaluate geometric fidelity using surface extraction and point-based metrics, considering both vision-only and vision-LiDAR methods. For image-based methods, we follow the D-NeRF [35] surface extraction pipeline. For vision–LiDAR methods, we directly compute point cloud error via nearest-neighbor distance. Meshes are extracted from SDF zero-crossings using marching cubes [169]. We report Chamfer Distance (CD), RMSE, and F1-score (F1) with a 5 cm threshold.

NuScenes. Evaluations on the NuScenes dataset (Table 9) show that frameworks incorporating LiDAR supervision significantly outperform vision-only approaches, highlighting the importance of active depth sensing for resolving scale ambiguities in large-scale environments. STGC-NeRF and LiDAR4D achieve the lowest global geometric error (Fig. 5) by leveraging spatio-temporal flow and surface normal priors for reconstruction regularization. Furthermore, the 3DGS-based OmniRE yields superior local depth accuracy and robustness to outliers, suggesting that deformation-field-based representations are effective for capturing fine geometric details and preserving structural integrity in dynamic driving scenes.

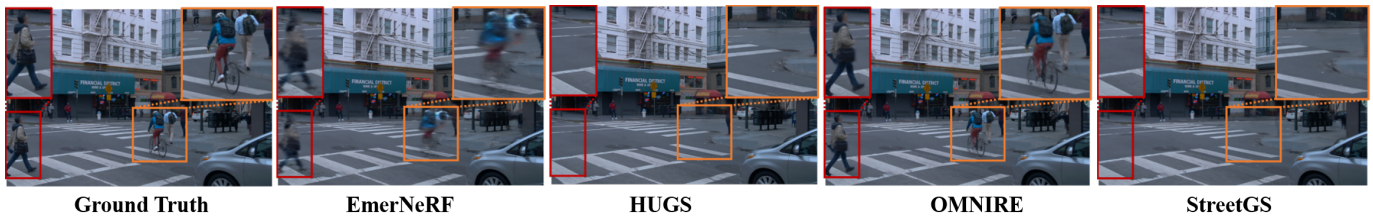

Fig. 6. **Qualitative novel view synthesis results of 4D reconstruction methods on the Waymo [162] dataset.** Image from [133].

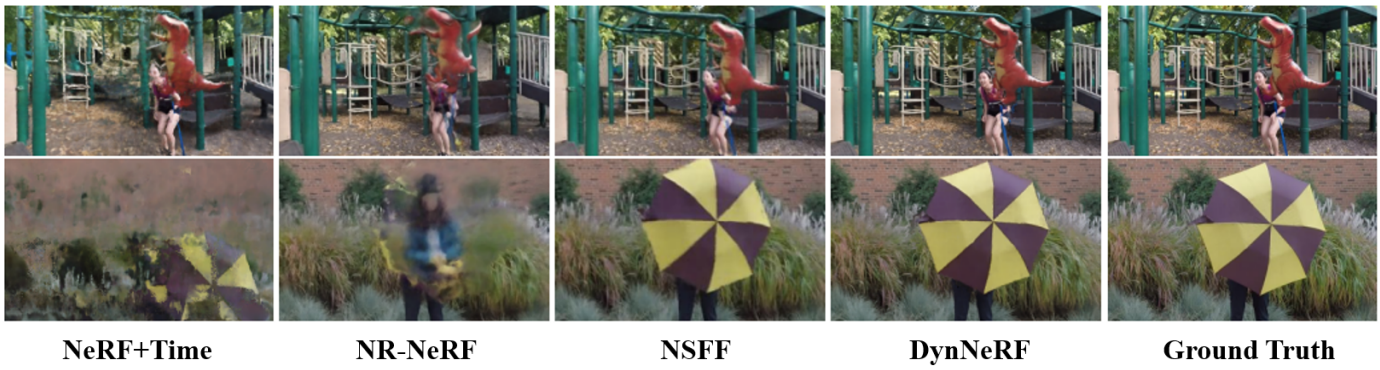

Fig. 7. **Qualitative novel view synthesis results of NeRF-style methods on the NVIDIA Dynamic Scene [159] dataset.** Image from [56].

Fig. 8. **Qualitative Novel View Synthesis results of 3DGS-style framework on the Neu3D [58] dataset.** Image sourced from [121].

5.5 Model Efficiency

We evaluate efficiency using GPU memory (peak GB), FPS, and training time on four NVIDIA A100 GPUs. Table 10 summarizes results for NeRF-style and 3DGS-style methods on Neu3D. Among NeRF-style methods, DevRF and StreamRF are the most efficient due to voxel-based representations and compact temporal encoding. In contrast, 3DGS-based methods achieve higher rendering speed owing to efficient rasterization, with 4DRotorGS attaining the highest FPS. Overall, 3DGS-based methods offer superior speed but require larger model capacity, while NeRF-style methods are more memory-efficient at the cost of slower rendering.

6 Future Prospects

6.1 Technical Prospects

Feed-Forward 4D Representations. Existing NeRF and Gaussian Splatting methods largely rely on per-scene optimization, resulting in high computational cost and limited scalability. A major trend is the transition toward generalizable feed-forward models. By leveraging large reconstruction models and transformer-based architectures [170]–[173], the field is gradually shifting from optimizing individual scenes to directly inferring scene representations. This paradigm enables near real-time 4D reconstruction from sparse inputs and helps bridge low-level reconstruction with higher-level spatio-temporal understanding.

Hybrid Explicit–Implicit Representations. The distinction between explicit and implicit representations is increasingly converging toward hybrid paradigms. Future 4D models are expected to combine explicit structures for efficient rasterization with implicit latent fields for modeling non-Lambertian effects and temporal topology changes [130], [174]. Such hybrid designs are important for scaling 4D representations to open-world dynamic scenes, where purely explicit methods face memory limitations and purely implicit methods struggle with real-time interaction.

Integration with Generative Priors. Another emerging direction is the integration of generative models [175],

TABLE 10
Performance analysis of 4D reconstruction methods. GPU memory, frame per second (FPS), and training time are evaluated.

Methods	4D-style	FPS	training time (h)	Params (Mb)
NeRF-style				
D-NeRF	Deformation fields	<1	22.3	**3**
DyNeRF	4D Primitive	<1	1344	7
NeRFPlayer	4D feature volumes	<1	6	-
HyperReel	4D feature volumes	6.1	2.2	360
MixVoxel	4D feature volumes	4.3	1.3	500
K-Planes	4D feature volumes	-	3.7	51
HexPlanes	4D feature volumes	-	12	200
MSTH	4D feature volumes	**15**	0.3	135
Ced-NeRF	4D feature volumes	6.3	**0.2**	-
StreamRF	Temporal prior	10.9	0.3	31
3DGS-style				
4DGS	Explicit 4D Primitive	30	5.0	1183
4DRotorGS	Explicit 4D Primitive	**277**	1.0	-
SpacetimeGS	Explicit 4D Primitive	140	0.31	200
CD-3DGS	Explicit 4D Primitive	118	1.0	338
4D-GS	Deformation Fields	30	0.67	90
ST-4DGS	Deformation Fields	37	2.7	339
Instant Gaussian Stream	Deformation Fields	204	**0.23**	2370
GaGS	Deformation Fields	12	2.0	48
DynMF	Deformation Fields	135	0.67	-
HiCoM	Deformation Fields	274	1.7	270
DN-4DGS	Deformation Fields	15	0.83	112
ED-3DGS	Deformation Fields	74.5	1.87	**35**
Ex4DGS	Frame-wise training	121	0.6	115
3DGStream	Frame-wise training	215	1.0	2340
4DGC	Frame-wise training	168	1.2	150

[176] into 4D reconstruction pipelines. Unlike traditional optimization-based methods that rely heavily on observations, generative priors enable plausible completion, motion prediction, and view synthesis under sparse or degraded inputs [177], [178]. This integration may shift reconstruction from purely observation-driven modeling toward predictive and generative frameworks, leading to more robust 4D reconstruction and synthesis.

Interactive and Controllable 4D Scenes. Beyond passive reconstruction, future 4D systems are expected to support interaction and controllability [104]. These capabilities enable users to manipulate dynamic scenes, edit object behaviors, and simulate alternative scenarios. Such developments may transform 4D representations into interactive world models for applications in simulation, digital twins, and content creation.

6.2 Application Prospects

Embodied AI Simulation. 4D reconstruction enables temporally consistent environments for embodied AI [135], [179], [180]. Compared with static representations, dynamic 4D scenes provide richer interaction signals and more realistic training environments, supporting robust perception, planning, and interaction in complex settings [181]–[183].

Dynamic Scene Understanding and Editing. 4D representations facilitate scene understanding, navigation, and editing in dynamic environments [184]–[186]. By modeling temporal evolution, these methods support motion prediction, scene simulation, and controllable editing for both analysis and content generation.

Human-Centered Applications. 4D reconstruction further enables accurate modeling of human motion and interaction [187]–[189]. These capabilities support applications in AR/VR, telepresence, healthcare, and digital twins. Future systems may further improve realism and interactivity, enabling more immersive user experiences.

7 Conclusion

4D scene reconstruction advances spatio-temporal 3D vision by modeling dynamic environments. This survey reviews NeRF- and 3DGS-based approaches, analyzing their efficiency, scalability, and temporal coherence. We further summarize representative datasets, evaluation metrics, and existing methods. Finally, we discuss open challenges and future directions, including feed-forward modeling, hybrid representations, and generative priors, as well as applications in embodied AI and dynamic scene understanding. This survey serves as both a reference and a roadmap for future research in 4D scene reconstruction.

References

[1] Jiahui Zhang, Yuelei Li, Anpei Chen, Muyu Xu, Kunhao Liu, Jianyuan Wang, Xiao-Xiao Long, Hanxue Liang, Zexiang Xu, Hao Su, et al., "Advances in feed-forward 3d reconstruction and view synthesis: A survey," *arXiv preprint arXiv:2507.14501*, 2025. 1

[2] Shuting He, Peilin Ji, Yitong Yang, Changshuo Wang, Jiayi Ji, Yinglin Wang, and Henghui Ding, "A survey on 3d gaussian splatting applications: Segmentation, editing, and generation," *arXiv preprint arXiv:2508.09977*, 2025. 1

[3] Ziyang Yan, Yihua Shao, Minwen Liao, Siyu Chen, Nan Wang, Muyuan Lin, Jenq-Neng Hwang, Hao Zhao, Fabio Remondino, and Lei Li, "3dsceneeditor: Controllable 3d scene editing with gaussian splatting," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2026, pp. 1852–1863. 1

[4] Hong Li, Chongjie Ye, Houyuan Chen, Weiqing Xiao, Ziyang Yan, Lixing Xiao, Zhaoxi Chen, Jianfeng Xiang, Shaocong Xu, Xuhui Liu, et al., "Near: Coupled neural asset-renderer stack," *arXiv preprint arXiv:2511.18600*, 2025. 1

[5] Ziyang Yan, Mengrui Yin, Yihua Shao, and Fabio Remondino, "Evaluating 3d gaussian splatting for urban scene reconstruction," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 251–258, 2025. 1

[6] Johannes L Schonberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113. 1

[7] Nazanin Padkan, Ziyan Yan, and Fabio Remondino, "Evaluating monocular depth estimation methods on industrial objects," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 175–181, 2025. 1

[8] Peter Cheeseman, Robert Smith, and Michael Self, "A stochastic map for uncertain spatial relationships," in *4th international symposium on robotic research*. MIT Press Cambridge, 1987, pp. 467–474. 1

[9] Fernando Nobre, Michael Kasper, and Christoffer Heckman, "Drift-correcting self-calibration for visual-inertial slam," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 6525–6532. 1

[10] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. 1, 3

[11] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li, "Nerf: Neural radiance field in 3d vision, a comprehensive review," *arXiv preprint arXiv:2210.00379*, 2022. 1, 2

[12] Fabio Remondino, Ali Karami, Ziyang Yan, Gabriele Mazzacca, Simone Rigon, and Rongjun Qin, "A critical analysis of nerf-based 3d reconstruction," *Remote Sensing*, vol. 15, no. 14, pp. 3585, 2023. 1

[13] Ziyang Yan, Gabriele Mazzacca, Simone Rigon, Elisa Mariarosaria Farella, Pawel Trybala, Fabio Remondino, et al., "Nerfbk: a holistic dataset for benchmarking nerf-based 3d reconstruction," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, no. 1, pp. 219–226, 2023. 1

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023. 1, 3

[15] Minwen Liao, Haobo Dong, Xinyi Wang, Kurban Ubul, Yihua Shao, and Ziyang Yan, "Gm-moe: Low-light enhancement with gated-mechanism mixture-of-experts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8766–8776. 2

[16] Yihua Shao, Deyang Lin, Minxi Yan, Siyu Chen, Fanhu Zeng, Minwen Liao, Ao Ma, Ziyang Yan, Haozhe Wang, Yan Wang, et al., "Tr-dq: Time-rotation diffusion quantization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, vol. 40, pp. 8869–8877. 2

[17] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar, "Neural fields in visual computing and beyond," in *Computer graphics forum*. Wiley Online Library, 2022, vol. 41, pp. 641–676. 2

[18] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He, "3d gaussian splatting as new era: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, pp. 4429–4449, 2024. 2

[19] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao, "Recent advances in 3d gaussian splatting," *Computational Visual Media*, vol. 10, no. 4, pp. 613–642, 2024. 2

[20] Yanqi Bao, Tianyu Ding, Jing Huo, Yaoli Liu, Yuxin Li, Wenbin Li, Yang Gao, and Jiebo Luo, "3d gaussian splatting: Survey, technologies, challenges, and opportunities," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2

[21] Guikun Chen and Wenguan Wang, "A survey on 3d gaussian splatting," *arXiv preprint arXiv:2401.03890*, 2024. 2

[22] Fabio Tosi, Youmin Zhang, Ziren Gong, Erik Sandström, Stefano Mattoccia, Martin R Oswald, and Matteo Poggi, "How nerfs and 3d gaussian splatting are reshaping slam: a survey," *arXiv preprint arXiv:2402.13255*, vol. 4, pp. 1, 2024. 2

[23] Yuchao Dai, Hongdong Li, and Mingyi He, "A simple prior-free method for non-rigid structure-from-motion factorization," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 101–122, 2014. 2

[24] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun, "Performance capture from sparse multi-view video," in *ACM SIGGRAPH 2008 papers*, pp. 1–10. ACM New York, NY, USA, 2008. 2

[25] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011, pp. 127–136. 2

[26] Richard A Newcombe, Dieter Fox, and Steven M Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352. 2

[27] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 785–801. 2

[28] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox, "Demon: Depth and motion network for learning monocular stereo," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5622–5631. 2

[29] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766. 2

[30] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1647–1655. 2

[31] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *arXiv preprint arXiv:1704.07804*, 2017. 2

[32] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 2043–2050. 2

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4455–4465. 3

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174. 3

[35] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10313–10322. 3, 5, 9, 10, 11, 12

[36] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5865–5874. 3, 5

[37] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz, "Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields," *arXiv preprint arXiv:2106.13228*, 2021. 3, 5, 10

[38] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian, "Fast dynamic radiance fields with time-aware neural voxels," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9. 3, 5

[39] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6494–6504. 3, 5

[40] Qiang Hu, Zihan Zheng, Houqiang Zhong, Sihua Fu, Li Song, Xiaoyun Zhang, Guangtao Zhai, and Yanfeng Wang, "4dgc: Rate-aware 4d gaussian compression for efficient streamable free-viewpoint video," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 875–885. 3, 8, 9

[41] Yuheng Yuan, Qiuhong Shen, Xingyi Yang, and Xinchao Wang, "1000+ fps 4d gaussian splatting for dynamic scene rendering," *arXiv preprint arXiv:2503.16422*, 2025. 3

[42] Jiaxuan Zhu and Hao Tang, "Dynamic scene reconstruction: Recent advance in real-time rendering and streaming," *arXiv preprint arXiv:2503.08166*, 2025. 3, 4

[43] Jinlong Fan, Xuepu Zeng, Jing Zhang, Mingming Gong, Yuxiang Yang, and Dacheng Tao, "Advances in radiance field for dynamic scene: From neural field to gaussian field," *arXiv preprint arXiv:2505.10049*, 2025. 4

[44] Lei He, Leheng Li, Wenchao Sun, Zeyu Han, Yichen Liu, Sifa Zheng, Jianqiang Wang, and Keqiang Li, "Neural radiance field in autonomous driving: A survey," *arXiv preprint arXiv:2404.13816*, 2024. 4

[45] Yukang Cao, Jiahao Lu, Zhisheng Huang, Zhuowen Shen, Chengfeng Zhao, Fangzhou Hong, Zhaoxi Chen, Xin Li, Wenping Wang, Yuan Liu, et al., "Reconstructing 4d spatial intelligence: A survey," *arXiv preprint arXiv:2507.21045*, 2025. 4

[46] Mingrui Zhao, Sauradip Nag, Kai Wang, Aditya Vora, Guangda Ji, Peter Chun, Ali Mahdavi-Amiri, and Hao Zhang, "Advances in 4d representation: Geometry, motion, and interaction," *arXiv preprint arXiv:2510.19255*, 2025. 4

[47] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt, "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12939–12950. 5

[48] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove, "Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13139–13147. 5

[49] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang, "Neural surface reconstruction of dynamic scenes

with monocular rgb-d camera," *Advances in Neural Information Processing Systems*, vol. 35, pp. 967–981, 2022. 5

[50] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou, "Devrf: Fast deformable voxel radiance fields for dynamic scenes," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36762–36775, 2022. 5

[51] Zhiwen Yan, Chen Li, and Gim Hee Lee, "Nerf-ds: Neural radiance fields for dynamic specular objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8285–8295. 5, 10, 11, 12

[52] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang, "Robust dynamic radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13–23. 5

[53] Chen Gao et al., "Total-recon: Deformable scene reconstruction for embodied view synthesis," *arXiv preprint arXiv:2304.12317*, 2023. 5

[54] Huiqiang Sun, Xingyi Li, Liao Shen, Xinyi Ye, Ke Xian, and Zhiguo Cao, "Dyblurf: Dynamic neural radiance fields from blurry monocular video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7517–7527. 5

[55] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim, "Space-time neural irradiance fields for free-viewpoint video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9416–9426. 5, 6

[56] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang, "Dynamic view synthesis from dynamic monocular video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5692–5701. 5, 6, 13

[57] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu, "Neural radiance flow for 4d view synthesis and video processing," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 14304–14314. 5

[58] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al., "Neural 3d video synthesis from multi-view video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5511–5521. 5, 10, 11, 13

[59] Fengrui Tian, Shaoyi Du, and Yueqi Duan, "Mononerf: Learning a generalizable dynamic radiance field from monocular videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17857–17867. 5, 6

[60] Seoha Kim, Jeongmin Bae, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh, "Sync-nerf: Generalizing dynamic nerfs to unsynchronized videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 2777–2785. 5, 6

[61] Xingguang Zhong, Yue Pan, Cyrill Stachniss, and Jens Behley, "3d lidar mapping in dynamic environments using a 4d implicit neural representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15417–15427. 5, 6

[62] Tobias Fischer, Lorenzo Porzi, Samuel Rota Bulo, Marc Pollefeys, and Peter Kontschieder, "Multi-level neural scene graphs for dynamic urban environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21125–21135. 5, 6

[63] Meng You and Junhui Hou, "Decoupling dynamic monocular videos for dynamic view synthesis," *IEEE Transactions on Visualization and Computer Graphics*, 2024. 5, 6

[64] Boyu Zhang, Wenbo Xu, Zheng Zhu, and Guan Huang, "Detachable novel views synthesis of dynamic scenes using distribution-driven neural radiance fields," *arXiv preprint arXiv:2301.00411*, 2023. 5, 6

[65] Ang Cao and Justin Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141. 5, 6

[66] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12479–12488. 5, 6

[67] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan, "Suds: Scalable urban dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12375–12385. 5, 6

[68] Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, and Nahyup Kang, "Temporal interpolation is all you need for dynamic neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 4212–4221. 5, 6

[69] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim, "Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16610–16620. 5, 6

[70] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu, "Mixed neural voxels for fast multi-view video synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19649–19659. 5, 6

[71] Feng Wang, Zilong Chen, Guokang Wang, Yafei Song, and Huaping Liu, "Masked space-time hash encoding for efficient dynamic scene reconstruction," *Advances in neural information processing systems*, vol. 36, pp. 70497–70510, 2023. 5, 6

[72] Jinxi Li, Ziyang Song, and Bo Yang, "Nvfi: Neural velocity fields for 3d physics learning from dynamic videos," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34723–34751, 2023. 5, 6

[73] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger, "Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2732–2742, 2023. 5, 6

[74] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Anton Van Den Hengel, "Blirf: Bandlimited radiance fields for dynamic scene modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 4641–4649. 5, 6

[75] Youtian Lin, "Ced-nerf: A compact and efficient method for dynamic neural radiance fields," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 3504–3512. 5, 6

[76] Zehan Zheng, Fan Lu, Weiyi Xue, Guang Chen, and Changjun Jiang, "Lidar4d: Dynamic neural fields for novel space-time view lidar synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5145–5154. 5, 6

[77] Ange Lou, Benjamin Planche, Zhongpai Gao, Yamin Li, Tianyu Luan, Hao Ding, Terrence Chen, Jack Noble, and Ziyan Wu, "Darenerf: Direction-aware representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5031–5042. 5, 6

[78] Xinhang Liu, Yu-Wing Tai, Chi-Keung Tang, Pedro Miraldo, Suhas Lohit, and Moitreya Chatterjee, "Gear-nerf: free-viewpoint rendering and tracking with motion-aware spatio-temporal sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19667–19679. 5, 6

[79] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson, "Neurad: Neural rendering for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14895–14904. 5, 6

[80] Xingyi Li, Zhiguo Cao, Yizheng Wu, Kewei Wang, Ke Xian, Zhe Wang, and Guosheng Lin, "S-dyrf: Reference-based stylized radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20102–20112. 5, 6

[81] Thang-Anh-Quan Nguyen, Luis Roldão, Nathan Piasco, Moussab Bennehar, and Dzmitry Tsishkou, "Rodus: Robust decomposition of static and dynamic elements in urban scenes," in *European Conference on Computer Vision*. Springer, 2024, pp. 112–130. 5, 6

[82] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al., "Emernerf: Emergent spatial-temporal scene decomposition via self-supervision," *arXiv preprint arXiv:2311.02077*, 2023. 5, 6

[83] Qi-Yuan Feng, Hao-Xiang Chen, Qun-Ce Xu, and Tai-Jiang Mu, "Sls4d: sparse latent space for 4d novel view synthesis," *IEEE Transactions on Visualization and Computer Graphics*, 2024. 5

[84] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan, "Streaming radiance fields for 3d video synthesis," *Advances in*

Neural Information Processing Systems, vol. 35, pp. 13485–13498, 2022. 5, 6

[85] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, Hisham Husain, and Anton Hengel, "Improving the convergence of dynamic nerfs via optimal transport," in *International Conference on Representation Learning*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., 2024, vol. 2024, pp. 19823–19840. 5, 6

[86] Shangshu Yu, Xiaotian Sun, Wen Li, Qingshan Xu, Zhimin Yuan, Sijie Wang, Rui She, and Cheng Wang, "Stgc-nerf: Spatial-temporal geometric consistency for lidar neural radiance fields in dynamic scenes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 9644–9652. 5, 6, 11

[87] Hong Li, Chongjie Ye, Houyuan Chen, Weiqing Xiao, Ziyang Yan, Lixing Xiao, Zhaoxi Chen, Jianfeng Xiang, Shaocong Xu, Xuhui Liu, et al., "Near: Coupled neural asset-renderer stack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 29834–29844. 6

[88] Wei Yao, Shuzhao Xie, Letian Li, Weixiang Zhang, Zhixin Lai, Shiqi Dai, Ke Zhang, and Zhi Wang, "Sd-gs: Structured deformable 3d gaussians for efficient dynamic scene reconstruction," *arXiv preprint arXiv:2507.07465*, 2025. 7

[89] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin, "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20331–20341. 7, 8, 9

[90] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang, "Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting," in *International Conference on Learning Representations (ICLR)*, 2024. 7, 8, 9

[91] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen, "4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11. 7, 8, 9

[92] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu, "Spacetime gaussian feature splatting for real-time dynamic view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8508–8520. 7, 8, 9

[93] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama, "A compact dynamic 3d gaussian representation for real-time dynamic view synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 394–412. 7, 9

[94] Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou, "Freetimegs: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21750–21760. 7, 9

[95] Su Sun, Cheng Zhao, Zhuoyang Sun, Yingjie Victor Chen, and Mei Chen, "Splatflow: Self-supervised dynamic gaussian splatting in neural motion flow field for autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27487–27496. 7, 9

[96] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, et al., "Drivedreamer4d: World models are effective data machines for 4d driving scene representation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12015–12026. 7, 9

[97] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan, "Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6782–6791. 7, 9

[98] Deqi Li, Shi-Sheng Huang, Zhiyuan Lu, Xinran Duan, and Hua Huang, "St-4dgs: Spatial-temporally consistent 4d gaussian splatting for efficient dynamic scene rendering," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11. 9

[99] Jinbo Yan, Rui Peng, Luyang Tang, and Ronggang Wang, "4d gaussian splatting with scale-aware residual field and adaptive optimization for real-time rendering of temporally complex dynamic scenes," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7871–7880. 9

[100] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao, "Gaussianflow: 4d reconstruction with dynamic 3d gaussian particle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21136–21145. 8, 9

[101] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao, "Hugs: Holistic urban 3d scene understanding via gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21336–21345. 8, 9

[102] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang, "4d gaussian splatting for real-time dynamic scene rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20310–20320. 8, 9

[103] Bardienus P Duisterhof, Zhao Mandi, Yunchao Yao, Jia-Wei Liu, Jenny Seidenschwarz, Mike Zheng Shou, Deva Ramanan, Shuran Song, Stan Birchfield, Bowen Wen, et al., "Deformgs: Scene flow in highly deformable scenes for deformable object manipulation," *arXiv preprint arXiv:2312.00583*, 2023. 9

[104] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi, "Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 4220–4230. 8, 9, 14

[105] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Min Yang, Xiao Tang, Feng Zhu, and Yuchao Dai, "3d geometry-aware deformable gaussian splatting for dynamic view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8900–8910. 8, 9

[106] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis, "Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 252–269. 9

[107] Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh, "Per-gaussian embedding-based deformation for deformable 3d gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 321–335. 9

[108] Richard Shaw, Michal Nazarczuk, Jifei Song, Arthur Moreau, Sibi Catley-Chandar, Helisa Dhamo, and Eduardo Pérez-Pellitero, "Swings: sliding windows for dynamic 3d gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 37–54. 8, 9

[109] Boming Zhao, Yuan Li, Ziyu Sun, Lin Zeng, Yujun Shen, Rui Ma, Yinda Zhang, Hujun Bao, and Zhaopeng Cui, "Gaussianprediction: Dynamic 3d gaussian prediction for motion extrapolation and free view synthesis," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12. 8, 9

[110] Meng-Li Shih, Jia-Bin Huang, Changil Kim, Rajvi Shah, Johannes Kopf, and Chen Gao, "Modeling ambient scene dynamics for free-view synthesis," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11. 9

[111] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas, "Dynamic gaussian marbles for novel view synthesis of casual monocular videos," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11. 8, 9

[112] Jiawei Xu, Zexin Fan, Jian Yang, and Jin Xie, "Grid4d: 4d decomposed hash encoding for high-fidelity dynamic gaussian splatting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 123787–123811, 2024. 8, 9

[113] Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu, "Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels," *Advances in Neural Information Processing Systems*, vol. 37, pp. 131316–131343, 2024. 9

[114] Mijeong Kim, Jongwoo Lim, and Bohyung Han, "4d gaussian splatting in the wild with uncertainty-aware regularization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 129209–129226, 2024. 8, 9

[115] Qiankun Gao, Jiarui Meng, Chengxiang Wen, Jie Chen, and Jian Zhang, "Hicom: Hierarchical coherent motion for dynamic streamable scenes with 3d gaussian splatting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 80609–80633, 2024. 8, 9

[116] Ruijie Zhu, Yanzhe Liang, Hanzhi Chang, Jiacheng Deng, Jiahao Lu, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang, "Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 101790–101817, 2024. 8, 9

[117] Jiahao Lu, Jiacheng Deng, Ruijie Zhu, Yanzhe Liang, Wenfei Yang, Xu Zhou, and Tianzhu Zhang, "Dn-4dgs: Denoised deformable network with temporal-spatial aggregation for dynamic scene rendering," *Advances in Neural Information Processing Systems*, vol. 37, pp. 84114–84138, 2024. 9

[118] Diwen Wan, Ruijie Lu, and Gang Zeng, "Superpoint gaussian splatting for real-time high-fidelity dynamic scene reconstruction," *arXiv preprint arXiv:2406.03697*, 2024. 8, 9

[119] Shizun Wang, Xingyi Yang, Qiuhong Shen, Zhenxiang Jiang, and Xinchao Wang, "Gflow: Recovering 4d world from monocular video," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 7862–7870. 8, 9

[120] Hanyang Kong, Xingyi Yang, and Xinchao Wang, "Efficient gaussian splatting for monocular dynamic scene rendering via sparse time-variant attribute modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 4374–4382. 9

[121] Jinbo Yan, Rui Peng, Zhiyan Wang, Luyang Tang, Jiayu Yang, Jie Liang, Jiahao Wu, and Ronggang Wang, "Instant gaussian stream: Fast and generalizable streaming of dynamic scene reconstruction via gaussian splatting," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16520–16531. 9, 13

[122] Yiren Lu, Yunlai Zhou, Disheng Liu, Tuo Liang, and Yu Yin, "Bard-gs: Blur-aware reconstruction of dynamic scenes via gaussian splatting," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16532–16542. 8, 9

[123] Sangwoon Kwak, Joonsoo Kim, Jun Young Jeong, Won-Sik Cheong, Jihyong Oh, and Munchurl Kim, "Modec-gs: Global-to-local motion decomposition and temporal interval adjustment for compact dynamic 3d gaussian splatting," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11338–11348. 8, 9

[124] Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim, "Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26866–26875. 8, 9

[125] Jinxi Li, Ziyang Song, Siyuan Zhou, and Bo Yang, "Freegave: 3d physics learning from dynamic videos by gaussian velocity," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12433–12443. 8, 9

[126] Hao Li, Sicheng Li, Xiang Gao, Abudouaihati Batuer, Lu Yu, and Yiyi Liao, "Gifstream: 4d gaussian-based immersive video with feature stream," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21761–21770. 8, 9

[127] Joohyun Kwon, Hanbyel Cho, and Junmo Kim, "Instruct-4dgs: Efficient dynamic scene editing via 4d gaussian-based static-dynamic separation," *arXiv preprint arXiv:2502.02091*, 2025. 8, 9

[128] Jiahui Lei, Yijia Weng, Adam W Harley, Leonidas Guibas, and Kostas Daniilidis, "Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6165–6177. 8, 9

[129] Bingbing Hu, Yanyan Li, Rui Xie, Bo Xu, Haoye Dong, Junfeng Yao, and Gim Hee Lee, "Learnable infinite taylor gaussian for dynamic view rendering," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26844–26854. 8, 9

[130] Cheng-De Fan, Chen-Wei Chang, Yi-Ruei Liu, Jie-Ying Lee, Jiun-Long Huang, Yu-Chee Tseng, and Yu-Lun Liu, "Spectromotion: Dynamic 3d reconstruction of specular scenes," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21328–21338. 8, 9, 13

[131] LIU Qingming, Yuan Liu, Jiepeng Wang, Xianqiang Lyu, Peng Wang, Wenping Wang, and Junhui Hou, "Modgs: Dynamic gaussian splatting from casually-captured monocular videos with depth priors," in *The Thirteenth International Conference on Learning Representations*, 2025. 8, 9

[132] He Huang, Qi Yang, Mufan Liu, Yiling Xu, and Zhu Li, "Adc-gs: Anchor-driven deformable and compressed gaussian splatting for dynamic scene reconstruction," *arXiv preprint arXiv:2505.08196*, 2025. 8, 9

[133] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, et al., "Omnire: Omni urban scene reconstruction," *arXiv preprint arXiv:2408.16760*, 2024. 8, 9, 12, 13

[134] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan, "Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 800–809. 8, 9

[135] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing, "3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20675–20685. 8, 9, 14

[136] Devikalyan Das, Christopher Wewer, Raza Yunus, Eddy Ilg, and Jan Eric Lenssen, "Neural parametric gaussians for monocular non-rigid object reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10715–10725. 9

[137] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng, "Street gaussians: Modeling dynamic urban scenes with gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 156–173. 8, 9

[138] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21634–21643. 8, 9

[139] Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu, "Fast view synthesis of casual videos with soup-of-planes," in *European Conference on Computer Vision*. Springer, 2024, pp. 278–296. 8, 9

[140] Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon, "Fully explicit dynamic gaussian splatting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 5384–5409, 2024. 8, 9

[141] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, Yue Qian, Xiaohang Zhan, and Yueqi Duan, "4d-fly: Fast 4d reconstruction from a single monocular video," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16663–16673. 8, 9

[142] Zhiyang Guo, Wengang Zhou, Li Li, Min Wang, and Houqiang Li, "Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 8, 9

[143] Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao, "Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 2642–2652. 8

[144] Wenkai Liu, Tao Guan, Bin Zhu, Luoyuan Xu, Zikai Song, Dan Li, Yuesong Wang, and Wei Yang, "Efficientgs: Streamlining gaussian splatting for large-scale high-resolution scene representation," *IEEE MultiMedia*, 2025. 8

[145] Zihan Wang, Jeff Tan, Tarasha Khurana, Neehar Peri, and Deva Ramanan, "Monofusion: Sparse-view 4d reconstruction via monocular fusion," *arXiv preprint arXiv:2507.23782*, 2025. 8

[146] Guo Chen, Jiarun Liu, Sicong Du, Chenming Wu, Deqi Li, Shi-Sheng Huang, Guofeng Zhang, and Sheng Yang, "Gs-roadpatching: Inpainting gaussians via 3d searching and placing for driving scenes," in *ACM SIGGRAPH 2025 Conference Papers*, 2025, pp. 1–11. 8

[147] Junkai Huang, Saswat Subhajyoti Mallick, Alejandro Amat, Marc Ruiz Olle, Albert Mosella-Montoro, Bernhard Kerbl, Francisco Vicente Carrasco, and Fernando De la Torre, "Echoes of the coliseum: Towards 3d live streaming of sports events," *ACM Transactions on Graphics (TOG)*, vol. 44, no. 4, pp. 1–17, 2025. 8

[148] Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann, "Gaussianflow: Splatting gaussian dynamics for 4d content creation," *arXiv preprint arXiv:2403.12365*, 2024. 8

[149] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al., "Eventvad: Training-free event-aware video anomaly detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 2586–2595. 9

[150] Yihua Shao, Yeling Xu, Xinwei Long, Siyu Chen, Ziyang Yan, Yang Yang, Haoting Liu, Yan Wang, Hao Tang, and Zhen Lei, "Accidentblip: Agent of accident warning based on ma-former," *arXiv preprint arXiv:2404.12149*, 2024. 9

[151] Jad Abou-Chakra, Feras Dayoub, and Niko Sünderhauf, "Particlenerf: A particle-based encoding for online neural radiance

fields," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5963–5972. 9, 10

[152] Yubin Hu, Kairui Wen, Heng Zhou, Xiaoyang Guo, and Yong-jin Liu, "Ss3dm: Benchmarking street-view surface reconstruction with a synthetic 3d mesh dataset," *Advances in Neural Information Processing Systems*, vol. 37, pp. 106649–106666, 2024. 9, 10

[153] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017. 10

[154] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa, "Monocular dynamic view synthesis: A reality check," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33768–33780, 2022. 10

[155] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski, "Stereo4d: Learning how things move in 3d from internet stereo videos," *arXiv preprint arXiv:2412.09621*, 2024. 10

[156] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3334–3342. 10

[157] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou, "Efficient neural radiance fields for interactive free-viewpoint video," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9. 10

[158] Facebook Research, "Technicolor dataset for hyperreel," https://github.com/facebookresearch/hyperreel. 10

[159] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz, "Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5336–5345. 10, 11, 13

[160] "Meeting room image classification dataset," https://images.cv/dataset/meeting-room-image-classification-dataset. 10

[161] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec, "Immersive light field video with a layered mesh representation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 86–1, 2020. 10

[162] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2443–2451. 10, 11, 13

[163] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631. 10, 11, 12

[164] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al., "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023. 10

[165] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al., "Pandaset: Advanced sensor suite dataset for autonomous driving," in *2021 IEEE international intelligent transportation systems conference (ITSC)*. IEEE, 2021, pp. 3095–3101. 10

[166] Lianqing Zheng, Long Yang, Qunshu Lin, Wenjin Ai, Minghao Liu, Shouyi Lu, Jianan Liu, Hongze Ren, Jingyue Mo, Xiaokai Bai, et al., "Omnihd-scenes: A next-generation multimodal dataset for autonomous driving," *arXiv preprint arXiv:2412.10734*, 2024. 10

[167] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013. 10

[168] Jannik Zürn, Paul Gladkov, Sofia Dudas, Fergal Cotter, Sofi Toteva, Jamie Shotton, Vasiliki Simaiaki, and Nikhil Mohan, "Wayvescenes101: A dataset and benchmark for novel view synthesis in autonomous driving," *arXiv preprint arXiv:2407.08280*, 2024. 10

[169] William E Lorensen and Harvey E Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, pp. 347–353. ACM New York, NY, USA, 1998. 12

[170] Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, et al., "Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos," *arXiv preprint arXiv:2412.03526*, 2024. 13

[171] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu, "Gs-lrm: Large reconstruction model for 3d gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–19. 13

[172] Chieh Hubert Lin, Zhaoyang Lv, Songyin Wu, Zhen Xu, Thu Nguyen-Phuoc, Hung-Yu Tseng, Julian Straub, Numair Khan, Lei Xiao, Ming-Hsuan Yang, et al., "Dgs-lrm: Real-time deformable 3d gaussian reconstruction from monocular videos," *arXiv preprint arXiv:2506.09997*, 2025. 13

[173] Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv, "4dgt: Learning a 4d gaussian transformer using real-world monocular videos," *arXiv preprint arXiv:2506.08015*, 2025. 13

[174] Shuangkang Fang, I Shen, Takeo Igarashi, Yufeng Wang, ZeSheng Wang, Yi Yang, Wenrui Ding, Shuchang Zhou, et al., "Nerf is a valuable assistant for 3d gaussian splatting," *arXiv preprint arXiv:2507.23374*, 2025. 13

[175] Jiawei Ren, Cheng Xie, Ashkan Mirzaei, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al., "L4gm: Large 4d gaussian reconstruction model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 56828–56858, 2024. 13

[176] Ziqiao Ma, Xuweiyi Chen, Shoubin Yu, Sai Bi, Kai Zhang, Chen Ziwen, Sihan Xu, Jianing Yang, Zexiang Xu, Kalyan Sunkavalli, et al., "4d-lrm: Large space-time reconstruction model from and to any view at any time," *arXiv preprint arXiv:2506.18890*, 2025. 13

[177] Yinghao Chen, Yeying Jin, Xiang Chen, Yanyan Wei, Ziyang Yan, and Yaowen Fu, "Unpaired image deraining using reward-guided self-reinforcement strategy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 1342–1354. 14

[178] Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian, "Medical sam3: A foundation model for universal prompt-driven medical image segmentation," *arXiv preprint arXiv:2601.10880*, 2026. 14

[179] Haozhe Lou, Yurong Liu, Yike Pan, Yiran Geng, Jianteng Chen, Wenlong Ma, Chenglong Li, Lin Wang, Hengzhen Feng, Lu Shi, et al., "Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 15379–15386. 14

[180] Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager, "Splat-nav: Safe real-time robot navigation in gaussian splatting maps," *IEEE Transactions on Robotics*, 2025. 14

[181] Yihua Shao, Siyu Liang, Zijian Ling, Minxi Yan, Haiyang Liu, Siyu Chen, Ziyang Yan, Chenyu Zhang, Haotong Qin, Michele Magno, et al., "Gwq: Gradient-aware weight quantization for large language models," *arXiv preprint arXiv:2411.00850*, 2024. 14

[182] Yihua Shao, Minxi Yan, Yang Liu, Siyu Chen, Wenjie Chen, Xinwei Long, Ziyang Yan, Lei Li, Chenyu Zhang, Nicu Sebe, et al., "In-context meta lora generation," *arXiv preprint arXiv:2501.17635*, 2025. 14

[183] Yihua Shao, Xiaofeng Lin, Xinwei Long, Siyu Chen, Minxi Yan, Yang Liu, Ziyang Yan, Ao Ma, Hao Tang, and Jingcai Guo, "Icm-fusion: in-context meta-optimized lora fusion for multi-task adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, vol. 40, pp. 8860–8868. 14

[184] Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Haiyang Liu, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al., "Renderworld: World model with self-supervised 3d label," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 6063–6070. 14

[185] Nan Wang, Yuantao Chen, Lixing Xiao, Weiqing Xiao, Bohan Li, Zhaoxi Chen, Chongjie Ye, Shaocong Xu, Saining Zhang, Ziyang Yan, et al., "Unifying appearance codes and bilateral grids for driving scene gaussian splatting," *arXiv preprint arXiv:2506.05280*, 2025. 14

[186] Shijie Zhou, Hui Ren, Yijia Weng, Shuwang Zhang, Zhen Wang, Dejia Xu, Zhiwen Fan, Suya You, Zhangyang Wang, Leonidas Guibas, et al., "Feature4x: Bridging any monocular video to 4d agentic ai with versatile gaussian feature fields," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14179–14190. 14

[187] Zhe Fan, Shi-Sheng Huang, Yichi Zhang, Dachao Shang, Juyong Zhang, Yudong Guo, and Hua Huang, "Rgavatar: Relightable 4d gaussian avatar from monocular videos," *IEEE Transactions on Visualization and Computer Graphics*, 2025. 14

[188] Zhixia Zhao, Qiyue Li, Jie Li, Richang Hong, and Zhi Liu, "Viewgauss: A head movement dataset for 6dof gaussian splatting video viewing," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 13016–13022. 14

[189] Pinxuan Dai, Peiquan Zhang, Zheng Dong, Ke Xu, Yifan Peng, Dandan Ding, Yujun Shen, Yin Yang, Xinguo Liu, Rynson WH Lau, et al., "4d gaussian videos with motion layering," *ACM Transactions on Graphics (TOG)*, vol. 44, no. 4, pp. 1–14, 2025. 14