Adaptive and Temporally Consistent Gaussian Surfels for Multi-view Dynamic Reconstruction

Decai Chen[1,2]   Brianne Oberson[1,3]   Ingo Feldmann[1]
Oliver Schreer[1]   Anna Hilsmann[1]   Peter Eisert[1,2]

[1]Fraunhofer HHI   [2]Humboldt University of Berlin   [3]Technical University of Berlin

{first}.{last}@hhi.fraunhofer.de

| Ground Truth | Ours
Train: 38s/frame
LPIPS: 0.040 | 4K4D
Train: 14min/frame
LPIPS: 0.047 | Ours
Train: 38s/frame | NeuS2
Train: 37s/frame |

Figure 1. Comparison of our proposed method on a scene from the DNA-Rendering dataset [10]. The training time and LPIPS scores (lower is better) are averaged across the sequence. Our approach not only achieves photorealistic novel view rendering with significantly reduced training time compared to the recent method [72], but also produces finer surface meshes, surpassing the state-of-the-art results [67].

Abstract

3D Gaussian Splatting has recently achieved notable success in novel view synthesis for dynamic scenes and geometry reconstruction in static scenes. Building on these advancements, early methods have been developed for dynamic surface reconstruction by globally optimizing entire sequences. However, reconstructing dynamic scenes with significant topology changes, emerging or disappearing objects, and rapid movements remains a substantial challenge, particularly for long sequences. To address these issues, we propose AT-GS, a novel method for reconstructing high-quality dynamic surfaces from multi-view videos through per-frame incremental optimization. To avoid local minima across frames, we introduce a unified and adaptive gradient-aware densification strategy that integrates the strengths of conventional cloning and splitting techniques. Additionally, we reduce temporal jittering in dynamic surfaces by ensuring consistency in curvature maps across consecutive frames. Our method achieves superior accuracy and temporal coherence in dynamic surface reconstruction, delivering high-fidelity space-time novel view synthesis, even in complex and challenging scenes. Ex-tensive experiments on diverse multi-view video datasets demonstrate the effectiveness of our approach, showing clear advantages over baseline methods. Project page: https://fraunhoferhhi.github.io/AT-GS

1. Introduction

Recovering dynamic scenes with high fidelity from multi-view videos presents a significant challenge in computer vision and graphics, with applications spanning virtual reality, cinematic effects, and interactive media. While many existing methods focus on creating visually appealing and immersive representations of dynamic environments, they often fall short when integrated in modern graphics engines, which require precise and temporally stable surface meshes for tasks such as geometry editing, physics-based simulations, animation, and texture mapping. Therefore, our goal is to develop a method that not only delivers photorealistic rendering of dynamic scenes but also ensures the reconstruction of geometrically accurate and temporally consistent surfaces.

In recent years, the rise of Neural Radiance Fields (NeRF) has gained considerable attention for their power-

ful ability to achieve photorealistic free-viewpoint rendering using compact volumetric representation and differentiable alpha composition [4, 47, 48]. Building on this foundation, numerous subsequent works [3, 17, 37, 42, 51, 53, 55] have further explored the synthesis of free-viewpoint videos for dynamic scenes. While NeRF-inspired approaches have driven significant progress, they often struggle with inefficiencies in training time and rendering speed. In contrast, the recent introduction of 3D Gaussian Splatting (3DGS) [31] marks a significant transition towards explicit point-based representations using differentiable rasterization, which offers more efficient training and high-fidelity real-time rendering. Recent advancements [14, 20, 25, 29, 35, 38, 43, 45, 59, 69, 73] have further demonstrated that Gaussian Splatting achieves superior performance in rendering complex, time-varying environments.

Existing surface reconstruction techniques, including those based on multi-view stereo [7, 11, 15, 75], neural implicit representations [8, 18, 39, 66], and more recently, 3D Gaussian Splatting [13, 22, 24, 77] have proven effective in static scenes. However, directly adapting them per-frame to time-varying real-world scenes presents challenges, such as significantly prolonged training times and temporal inconsistencies across frames. An alternative approach is to reconstruct the entire dynamic sequence within a single holistic model [2, 6, 12, 28, 44, 46, 68], such as by deforming a canonical space. However, globally representing dynamic scenes with significant topology changes, emerging or disappearing objects, and rapid movements remains a substantial challenge, particularly for long sequences.

To address these challenges, we propose Adaptive and Temporally Consistent Gaussian Surfels (AT-GS), a novel method for efficient and temporally consistent dynamic surface reconstruction from multi-view videos. Our approach utilizes a coarse-to-fine incremental optimization process based on a per-frame Gaussian surfels representation.

Initially, we train the first frame of the sequence using a standard static multi-view reconstruction technique representing the scene as Gaussian surfels [13]. For each subsequent frame, we learn the SE(3) transformation to coarsely align Gaussian surfels from the previous frame to the current one. We introduce a unified densification strategy combining the strengths of clone and split. Additionally, we design adaptive probability density function (PDF) sampling, guiding the splitting process using the magnitude of viewspace positional gradients.

Another challenge in dynamic reconstruction is maintaining temporal consistency. In dynamic reconstruction, slight temporal jittering between frames caused by the randomness of optimization can lead to visible artifacts, especially in regions with minimal textures, where different Gaussian configurations may produce visually identical results. To address this, we first predict the optical flow be-

tween neighboring frames of the same view and warp the rendered normal map from the previous frame to the current one. We then enforce consistency in the curvature maps derived from the normals of consecutive frames. This method indirectly ensures temporal coherence in the rendered depth maps and Gaussian orientations, resulting in more stable and accurate final surface geometry.

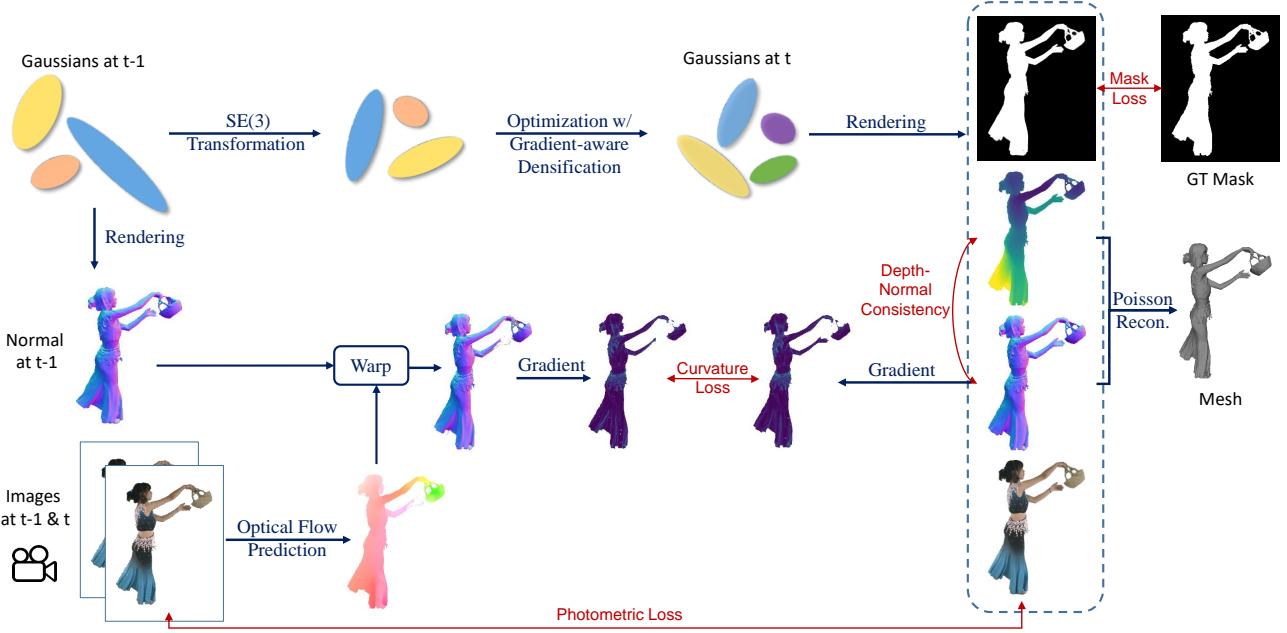In summary, our contributions include:

- A method for efficiently reconstructing dynamic surfaces from multi-view videos using Gaussian surfels.

- A unified and gradient-aware densification strategy for optimizing dynamic 3D Gaussians with fine details.

- A temporal consistency approach that ensures stable and coherent surface reconstructions across frames by enforcing consistency on curvature maps.

- Extensive experiments that demonstrate our method's advantages including fast training, high-fidelity novel view synthesis, and accurate surface geometry.

2. Related Works

2.1. Dynamic Novel View Synthesis

Recent advancements in dynamic view synthesis have been significantly driven by the development of Neural Radiance Fields (NeRF) [47]. Building on this foundation, various methods extend NeRF by conditioning it on temporal states such as frame indices, time coordinates, body poses, or per-frame embeddings, to handle dynamic scenes [19, 34, 37, 40, 61, 64]. To enhance training and rendering efficiency, recent works factorize the 4D space into lower-dimensional components, such as planes, thereby reducing computational complexity [3, 17, 26, 42, 55]. Another line of research focuses on explicitly modeling deformation fields that map the deformed space to a canonical space where the NeRF is embedded [33, 50, 51, 53, 62]. Alternatively, some approaches employ incremental training, learning the per-frame differences to achieve efficient dynamic scene representation [36, 58, 65]. Nevertheless, like NeRF, these methods face challenges with slow training and rendering times.

The recently introduced 3D Gaussian Splatting technique [31] dramatically enhances training and rendering speeds through differentiable rasterization. Similar to extending NeRF to dynamic scenes, various approaches predict a deformation field that maps canonical Gaussians to their positions at each observed timestep [1, 20, 23, 25, 27, 41, 56, 63, 69, 71, 73]. One alternative direction is to learn temporally continuous motion trajectories of 3D Gaussians via basis functions [29, 35, 38, 43]. Besides, other approaches lift 3D Gaussians into 4D by directly incorporating the time dimension [16, 74]. Instead of using Gaussians, 4K4D [72]

Figure 2. **Pipeline of Our Method.** Starting with the Gaussian surfels from the previous frame $(t-1)$, we first estimate their coarse translation and rotation to align with the current frame (t). Subsequently, we optimize all Gaussian attributes, incorporating our gradient-guided densification strategy. For each training view, we render opacity, depth, normal, and color maps (from top to bottom in the dashed box) using differentiable tile-based rasterization. Additionally, we predict optical flow between consecutive frames, which warps the rendered normal map from frame $t-1$ to frame t. We then ensure temporal consistency of the underlying surface by comparing curvature maps derived from the warped and rendered normal maps. Furthermore, we apply photometric loss, depth-normal consistency loss, and mask loss for supervision. Finally, Poisson reconstruction is employed to generate a mesh from the unprojected depth and normal maps.

represents dynamic scenes through point clouds encoded with 4D features, enabling photorealistic and real-time rendering. These methods model entire dynamic scenes using a holistic and temporally smooth representation, which is particularly suitable for monocular video inputs. However, the assumption of accurate cross-frame correspondence often breaks down in complex dynamic scenes with significant topological changes and transient objects. Incrementally optimizing dynamic scenes frame by frame from multi-view videos can mitigate these limitations. Specifically, Luiten *et al.* [45] maintain fixed opacity, color, and size for the optimized 3DGS model from the first frame, and learn per-frame 6-DoF motion of each Gaussian for dense tracking. More recently, 3DGStream [59] introduces a two-stage training scheme: the first stage trains a Neural Transformation Cache [48] for translating and rotating Gaussians with attributes inherited from the first frame, while the second stage spawns and optimizes additional Gaussians using densification and pruning. While relying on the initial Gaussians reduces storage overhead, these methods struggle when subsequent frames deviate significantly from the first frame, such as with emerging objects or topological changes, especially for long sequences. In contrast, our approach allows for the full optimization of per-frame Gaussians including densification and pruning, which enables rapid adaptation to complex dynamic scenes.

Notably, several works guide the movement of Gaussians by supervising the projected Gaussian scene flow with the estimated optical flow of input images [20, 23, 29]. However, these approaches maintain a fixed number of Gaussians over time to ensure consistent tracking. In contrast, our method allows flexible densification and pruning, eliminating the need for per-Gaussian correspondence across frames, thereby allowing rapid adaptation to new scenes.

2.2. Gaussians-based Surface Reconstruction

Recent advancements in 3D Gaussian Splatting have demonstrated significant progress in surface reconstruction. An earlier work, SuGaR [22], introduces regularization terms to better align Gaussians with scene surfaces, leveraging Poisson reconstruction to generate meshes from sampled point clouds. NeuSG [9] and GSDF [76] integrate 3D Gaussian Splatting with SDF to jointly optimize these representations. Despite these improvements, challenges such as irregular Gaussian shapes and the presence of artifacts remain. To address these issues, methods like 2DGS

[24] and Gaussian-Surfels [13] flatten 3D volumes into 2D planar disks (i.e., surface elements or surfels) to achieve more precise reconstructions. Additionally, GOF [77] employs a Gaussian opacity field to facilitate direct geometry extraction. More recently, PGSR [5] introduced unbiased depth rendering alongside various regularization techniques, while RaDe-GS incorporated a rasterized approach to render depths and surface normals from 3DGS.

Concurrently, several approaches have extended static 3DGS-based surface reconstruction to dynamic scenes from monocular videos. DG-Mesh [44] proposes cycle-consistent deformation between canonical and deformed Gaussians, mapping these to tracked mesh facets and optimizing Gaussians across all time frames. Vidu4D [68] optimizes a bone-based deformation field to transform Gaussian surfels from a canonical state to a warped state, refining rotation and scaling parameters. MaGS [46] introduces a mutually adsorbed mesh-Gaussian representation, allowing for relative displacement between the mesh and Gaussians, with joint optimization of these elements. In contrast to these holistic optimization approaches, our method incrementally trains per-frame Gaussian surfels without requiring the entire video sequence during optimization. By employing unified gradient-aware densification and a curvature-based temporal consistency strategy, our method achieves robust dynamic reconstruction, even in complex scenes with significant topological changes.

3. Method

In this section, we first introduce our incremental training pipeline for Gaussian surfels tailored to dynamic scenes, as detailed in Sec. 3.1. Next, we elaborate on our unified, gradient-guided densification strategy in Sec. 3.2, which refines dynamic 3D Gaussians with fine-grained detail. In Sec. 3.3, we present our curvature-based temporal consistency approach, ensuring stable and coherent surface reconstructions over time. Finally, we provide details on the model training process in Sec. 3.4.

3.1. Incremental Gaussian Surfels

The overview of AT-GS is demonstrated in Fig. 2. Our goal is to accurately reconstruct both the appearance and the geometry of dynamic scenes from multi-view video sequences. The input consists of multi-view RGB image sequences, denoted as $I_{i,j} : i \in [1, N], j \in [1, M]$, where N represents the number of frames and M the number of views. Additionally, camera calibration parameters, including intrinsics and poses, are provided.

First, we perform full training of Gaussian surfels [13] for frame 0 from a sparse point cloud generated by Structure-from-Motion (SfM) [54] or random initialization, following standard static reconstruction practices. For each subsequent frame t, we begin with the Gaussians from the previous frame $t-1$ and efficiently adapt them to the current frame using a coarse-to-fine strategy.

In the coarse stage, only the centers and rotations of the Gaussians are updated. Inspired by 3DGStream [59], we train a per-frame Neural Transformation Cache (NTC) [48, 49], consisting of multi-resolution hash encoding and a shallow MLP, which maps spatial positions to SE(3) transformations. Leveraging the spatial smoothness of the voxel-grid representation and linear interpolation, NTC facilitates faster convergence compared to directly optimizing the center and rotation of each Gaussian.

In the fine stage, we refine all learnable parameters (center, rotation, scale, view-dependent color, and opacity) of the Gaussians while allowing for pruning and adaptive gradient-guided densification (Sec. 3.2) to capture fine details and accommodate new objects. Furthermore, we leverage curvature-based temporal consistency (Sec. 3.3) to ensure stable and coherent surface reconstructions over time.
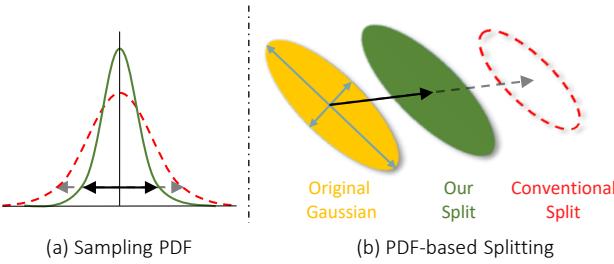
After optimization, we render color, depth, normal, and opacity maps using alpha blending. Similar to [13], the rendered depth maps are back-projected and merged in the global 3D space to form a point cloud, with normals derived from the rendered normal maps. Lastly, screened Poisson reconstruction [30] is applied to generate a surface mesh.

3.2. Unified and Gradient-aware Densification

In conventional 3DGS methods [13, 31], Gaussians with large positional gradients are densified through either cloning or splitting, depending on their sizes. Cloning produces a new Gaussian that retains the size and position of the original, while the original Gaussian shifts in the direction of the positional gradient. On the other hand, splitting generates two smaller Gaussians with positions sampled from the original Gaussian, and then removes the original. Although effective in static reconstruction, where training occurs from scratch, this conventional densification approach can lead to sub-optimal performance in incremental training of dynamic scenes.

After initializing from the previous frame and applying the SE(3) transformation, our Gaussian representation is already near optimal convergence, and our objective is to optimize it within a limited number of iterations. This presents two challenges. First, due to small gradients and momentum during optimization, the centers of the Gaussians are prone to becoming trapped in suboptimal local minima. Therefore, it is preferable to move the Gaussians rather than preserve them in their original positions, as it is done in conventional cloning. Second, excessive updates to the Gaussians can degrade the results, particularly during splitting, which can displace Gaussians far from their original positions.

To address these challenges, we propose a new adaptive densification strategy for the fine stage of per-frame

(a) Sampling PDF   (b) PDF-based Splitting

Figure 3. **Gradient-aware splitting.** (a) 1D illustration of sampling PDFs (normal distributions), which determine the positions of new split Gaussians. (b) Conventional splitting (red dashed ellipse) samples a new, smaller Gaussian from a multivariate normal distribution centered at the original Gaussian, with standard deviations equal to its scales. In contrast, our approach (green solid ellipse) adaptively guides the sampling using view-space positional gradients, while preserving the size of the original Gaussian.

optimization. This approach unifies the strengths of conventional cloning and splitting into a single step to densify Gaussians with large positional gradients, regardless of their size. Specifically, the original Gaussian moves in the direction of positional gradient, while a new Gaussian of the same size is added by gradient-guided sampling.

As illustrated in Fig. 3, splitting creates a new Gaussian centered at a point sampled from a multivariate normal distribution. Conventional split uses the scales of the original Gaussians as the standard deviations $\boldsymbol{\sigma}$ of the sampling PDFs, which can cause excessive displacement, especially for Gaussians with moderate gradients that require only minor adjustments. This issue is more pronounced in incremental dynamic training scenarios, where the Gaussian representation is already close to convergence and only a few iterations are available to correct positional inaccuracies.

To overcome this, we utilize the magnitude of the loss function gradients $\|\nabla \mathcal{L}\|$ with respect to view-space positions of the Gaussians to adaptively guide the sampling PDFs. We first clamp $\|\nabla \mathcal{L}\|$ to a maximum of twice their average of all Gaussians being split $\overline{\|\nabla \mathcal{L}\|}$, in order to eliminate outliers. We then normalize these clamped gradients $\|\nabla \mathcal{L}\|$ and use them to scale the standard deviations of their corresponding sampling PDFs:

$$\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma} \times \frac{\min\left(\|\nabla \mathcal{L}\|, 2\overline{\|\nabla \mathcal{L}\|}\right)}{2\overline{\|\nabla \mathcal{L}\|}}. \tag{1}$$

Finally, we initialize the centers of the split Gaussian by sampling from a multivariate normal distribution with standard deviations $\hat{\boldsymbol{\sigma}}$. In this manner, Gaussians with larger gradients are more likely to move further away, allowing significant updates, while those with smaller gradients undergo finer corrections. Additionally, instead of shrink-

ing the split Gaussians, we maintain their original size to achieve faster convergence. Our unified, adaptive densification strategy effectively balances movement and stability, enabling efficient and precise optimization during incremental training for dynamic scenes.

3.3. Curvature-based Temporal Consistency

Another key challenge in reconstructing dynamic surfaces is ensuring temporal consistency. We observe that concatenating per-frame meshes leads to noticeable jittering artifacts during temporal visualization, as shown in the supplementary video. This temporal inconsistency arises for two main reasons. First, the training process for 3D Gaussian Splatting is susceptible to randomness stemming from GPU scheduling, even with identical input data [21]. Similarly, the inherent randomness in optimization can yield varying results in dynamic reconstruction, even in regions with minimal changes in the input images. Second, low-textured areas lack sufficient multi-view photometric constraints, allowing different Gaussian representations to fit the same input images. This jittering degrades visual quality, undermining downstream applications like virtual reality, where temporal coherence is crucial.

To overcome this issue, we leverage the local rigidity prior to enhance the temporal consistency of dynamic surfaces. As illustrated in Fig. 2, for each training view of frame t, we estimate both forward (from $t-1$ to t) and backward (from t to $t-1$) optical flow between consecutive timesteps using an off-the-shelf technique [60]. We then compute a confidence mask for the backward flow using cycle consistency. With this masked backward flow, the rendered normal map $\mathbf{N}_{t-1}$ at frame $t-1$ is warped to $\hat{\mathbf{N}}_t$ for frame t.

Next, we approximate the curvature map $\mathbf{C}$ derived from the normal map $\mathbf{N}$ by:

$$\mathbf{C} = \| \left(|\nabla_x \mathbf{N}| + |\nabla_y \mathbf{N}|\right) \|_2. \tag{2}$$

where $\nabla_x \mathbf{N}$ and $\nabla_y \mathbf{N}$ represent the partial derivatives of $\mathbf{N}$ with respect to the pixel space coordinates x and y, while $\|\cdot\|_2$ denotes the Euclidean norm calculated along the spatial (i.e., xyz) dimensions. Using Eq. (2), we compute the curvature maps for both $\hat{\mathbf{N}}_t$ and $\mathbf{N}_t$, denoted as $\hat{\mathbf{C}}_t$ and $\mathbf{C}_t$, respectively.

Finally, we define the loss function for temporal consistency as the mean squared error (MSE) between both curvature maps:

$$\mathcal{L}_t = \mathrm{MSE}\left(\hat{\mathbf{C}}_t, \mathbf{C}_t\right). \tag{3}$$

This loss function allows us to maintain local rigidity by ensuring that the orientation of Gaussian surfels is temporally coherent. By supervising the curvature map of the current frame with the warped curvature map from the previous

frame, we achieve smoother temporal transitions and more accurate reconstruction of the dynamic scenes.

3.4. Optimization

Our reconstruction process is optimized in an end-to-end manner. Following static reconstruction with Gaussian surfels [13], we incorporate a comprehensive set of loss functions to guide the optimization process. In addition to our temporal consistency loss $\mathcal{L}_t$, the per-frame optimization is supervised by several other loss functions: photometric loss $\mathcal{L}_p$, depth-normal consistency loss $\mathcal{L}_c$, opacity loss $\mathcal{L}_o$, and mask loss $\mathcal{L}_m$. The total loss function is thus formulated as:

$$\mathcal{L} = \mathcal{L}_p + \lambda_t \mathcal{L}_t + \lambda_c \mathcal{L}_c + \lambda_o \mathcal{L}_o + \lambda_m \mathcal{L}_m, \quad (4)$$

where $\lambda_t, \lambda_c, \lambda_o, \lambda_m$ are weighting factors that balance the contribution of each respective term.

In contrast to [13], our method does not require additional monocular normal prior or its associated loss term, as our framework is capable of supervising the normal maps more effectively through the temporal context. During the coarse stage of per-frame training, we simplify the optimization by focusing solely on the photometric loss $\mathcal{L}p$ and mask loss $\mathcal{L}m$, which are sufficient for learning the coarse transformation of the Gaussian representation. As the training progresses to the fine stage, we apply the full set of loss terms described in Eq. (4) to optimize the fine-grained details of the dynamic scene reconstruction. For all datasets, we train the coarse stage for 200 iterations and the fine stage for 800 iterations. Thanks to our adaptive incremental optimization strategy, we achieve efficient on-the-fly training at approximately 30 seconds per frame.

4. Experiments

4.1. Datasets

We evaluate our method on the DNA-Rendering [10] and NHR [70] datasets for dynamic scene reconstruction. The DNA-Rendering dataset features dynamic scenes of humans with complex clothing, rapid movement, and challenging objects, such as reflective surfaces. It is captured using 48 cameras at a resolution of 2448×2048 and 12 cameras at 4096×3000. In contrast, the NHR dataset includes 56 cameras capturing three sport scenes and 72 cameras for the basketball scene, with resolutions of 1024×768 and 1224×1024. Our experiments are conducted on five commonly used sequences from DNA-Rendering dataset and all four sequences from the NHR dataset, each containing 150 frames. Following 4K4D [72], four views are reserved for testing, while the remaining views are used for training.

4.2. Comparison

Free-viewpoint Rendering. We compare AT-GS with state-of-the-art methods for novel view synthesis in dy-

Type	Method	PSNR↑	SSIM↑	LPIPS↓
Holistic	4K4D	34.52	0.985	0.025
	STG	28.49	0.966	0.041
Incremental	NeuS2	33.80	0.987	0.032
	3DGStream	30.78	0.974	0.047
	Ours	35.44	0.988	0.024

Table 1. Quantitative results on the DNA-Rendering dataset [10]. The best values are highlighted in red, and the second-best values in yellow. Metrics are averaged across all scenes.

Type	Method	PSNR↑	SSIM↑	LPIPS↓
Holistic	4K4D	33.65	0.972	0.039
	STG	28.05	0.949	0.074
Incremental	NeuS2	33.04	0.972	0.047
	3DGStream	30.70	0.955	0.083
	Ours	33.55	0.973	0.054

Table 2. Quantitative results on the NHR dataset [70].

namic scenes, following the official implementations of these methods. These methods are categorized into two groups: (1) holistic approaches, such as 4K4D [72] and SpacetimeGaussians (STG) [38], which optimize entire video sequences as a whole; and (2) incremental methods, including 3DGStream [59] and NeuS2 [67], which train each frame sequentially.

Qualitative and quantitative comparisons on the DNA-Rendering dataset [10] are presented in Fig. 4 and Tab. 1, respectively. To quantitatively evaluate the rendering quality, we report the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [78] based on the VGG network [57]. These metrics are calculated and averaged across all testing views and frames. As shown in Tab. 1, our method outperforms existing approaches across all three metrics. Specifically, STG and 3DGStream struggle to recover areas with fast motion, such as the hands, as illustrated in both scenes of Fig. 4. Additionally, NeuS2 fails to reconstruct fine-grained details like the basket frame in the second row of Fig. 4. 4K4D, despite achieving photo-realistic appearances in dynamic regions, suffers from artifacts around object boundaries and struggles with non-Lambertian surfaces, such as the camera lens. In contrast, our method synthesizes novel views with higher visual fidelity, even in complex dynamic scenes.

We further evaluate our method on the NHR dataset [70]. As demonstrated in Tab. 2 and Fig. 5, our method achieves rendering quality comparable to 4K4D [72] and NeuS2 [67], while significantly outperforming other methods. Notably, training 4K4D takes more than a day, whereas
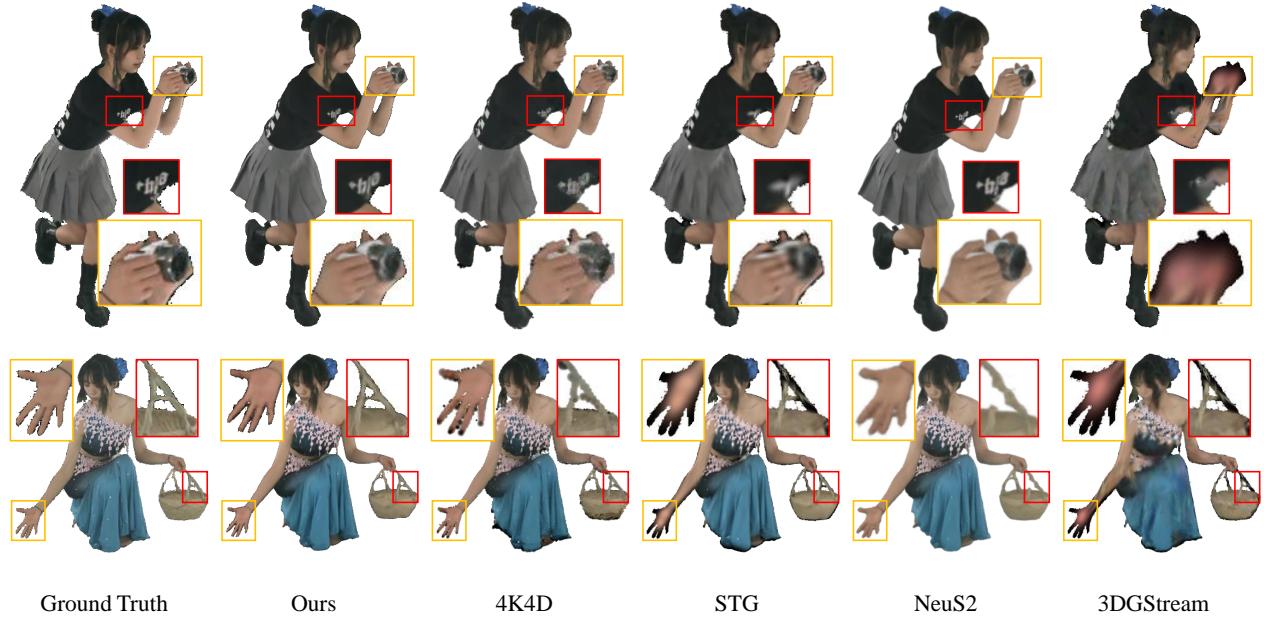
Figure 4. Qualitative comparison of novel view synthesis on the DNA-Rendering dataset [10].
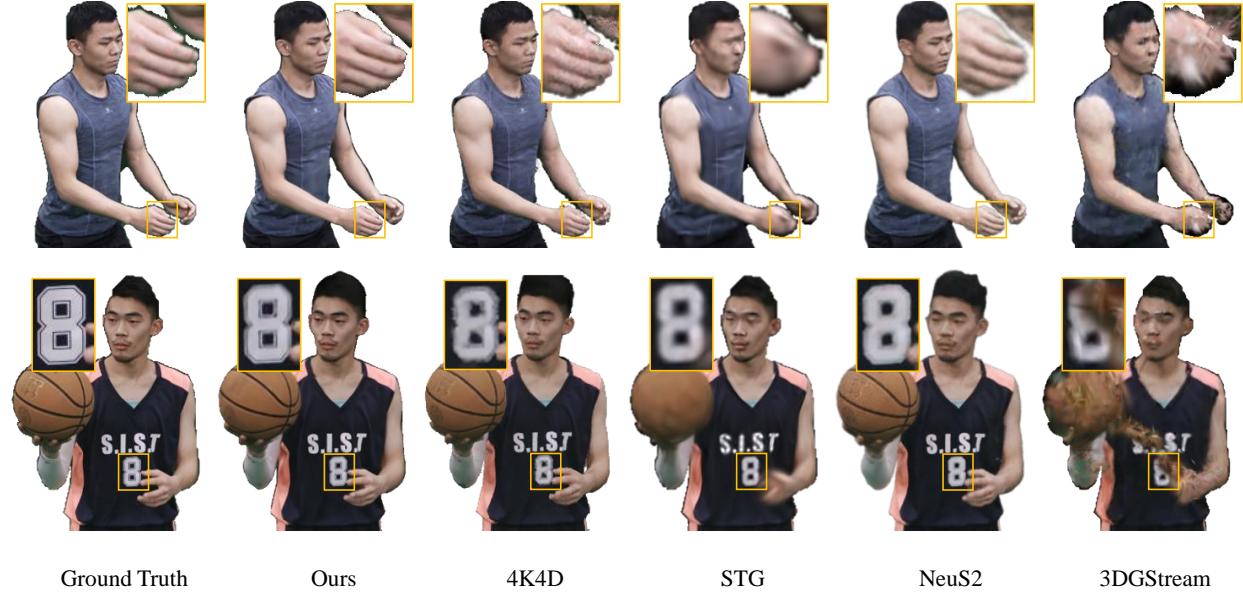

Figure 5. Qualitative comparison of novel view synthesis on the NHR dataset [70].

our method requires only about 1.5 hours.

Surface reconstruction. We compare our method with NeuS2 [67], a state-of-the-art neural scene reconstruction method, to evaluate the geometric quality of dynamic reconstructions. Since ground truth geometry is not available for either dataset, we provide a qualitative comparison in Fig. 6. Specifically, NeuS2 tends to produce blurry surfaces on fine details, such as fingers, and is prone to noisy artifacts. In contrast, our method achieves more accurate ge-

ometry reconstruction with enhanced detail, even for thin objects like the phone in the first row and in challenging areas like the occluded upper body in the third row.

4.3. Ablation Study

To evaluate the contribution of our proposed components, we conduct both quantitative and qualitative analyses across all sequences from the NHR dataset. The "w/o GD" variant replaces the proposed gradient-aware densification
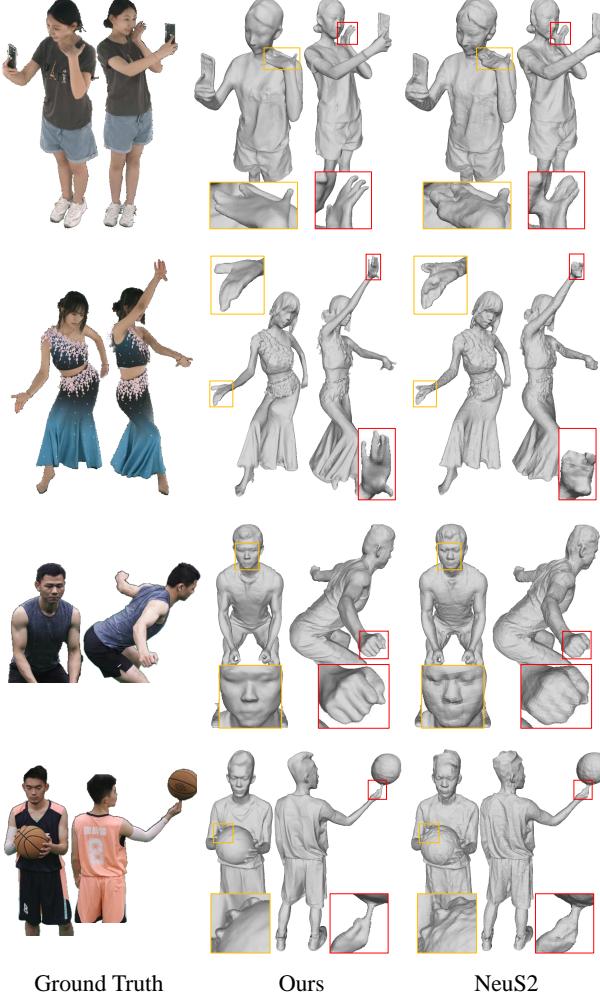
Figure 6. Qualitative comparison of surface reconstruction on the DNA-Rendering and NHR datasets. Compared to the state-of-the-art method, our approach produces superior results with finer and more accurate geometric details.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o GD + w/o TC	33.468	0.9729	0.0549
w/o GD	33.505	0.9731	0.0545
w/o TC	33.499	0.9732	0.0541
Ours Full	33.547	0.9733	0.0539

Table 3. Ablation study on the NHR dataset. Both Gradient-aware Densification (GD) and curvature-based Temporal Consistency (TC) contribute to improved overall results. Metrics are averaged across all scenes.
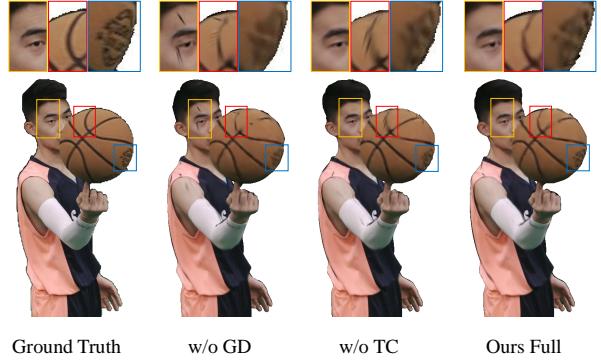

Ground Truth · w/o GD · w/o TC · Ours Full

Figure 7. Qualitative ablation study on the NHR dataset.

strategy with conventional densification. As shown in Tab. 3 and Fig. 7, this modification introduces visual artifacts and results in inferior rendering, particularly in terms of perceptual quality (LPIPS). Similarly, the "w/o TC" variant omits the proposed curvature-based temporal consistency loss, leading to degradation in the fidelity of view synthesis. Additionally, an ablation comparison of this variant on dynamic meshes is provided in the supplementary video, highlighting the effectiveness of temporal consistency in maintaining visual coherence over time.

5. Conclusion

In this work, we present AT-GS, a novel approach for efficient dynamic surface reconstruction from multi-view videos. By introducing a unified and gradient-aware densification strategy, we optimize dynamic 3D Gaussians with fine-grained details, overcoming the challenges of local minima and ensuring high-fidelity reconstruction. Additionally, our temporal consistency approach, which enforces curvature map consistency across frames, addresses the issue of temporal jittering, leading to stable and coherent surface reconstructions. Through extensive experiments on diverse datasets, we demonstrate that our method outperforms existing approaches in terms of rendering quality and geometric accuracy.

Limitations. First, because we focus on fast dynamic reconstruction, the limited number of training iterations per frame may hinder performance in handling extremely challenging objects, such as the small sparkling sequins on skirts, as seen in Fig. 4. Additionally, since the Gaussian representation for each frame is stored separately, the storage overhead scales linearly with the video length, reducing storage efficiency for very long sequences. Addressing these limitations will be an avenue for future work.

Acknowledgements

Adaptive and Temporally Consistent Gaussian Surfels for Multi-view Dynamic Reconstruction
–Supplementary Material–

6. Implementation Details

In all our experiments, training is conducted on a GPU server equipped with an AMD EPYC 9654 CPU and an NVIDIA RTX 6000 Ada GPU, utilizing the Adam optimizer [32], PyTorch 2.3.1 [52], and CUDA 11.8. For each dynamic scene, we begin with static reconstruction using Gaussian surfels [13] for the first frame, obtaining a surfel-based Gaussian representation from a sparse point cloud generated by COLMAP [54]. For each subsequent frame, we initialize the scene from the previous frame and apply our coarse-to-fine training approach, with 200 iterations for the coarse stage and 800 iterations for the fine stage. Training takes 31.7 seconds per frame on the NHR dataset [70] and 37.5 seconds per frame on the DNA-Rendering dataset [10].

In the coarse stage, the learning rate for the Neural Transformation Cache is set to 0.002. In the fine stage, our unified, adaptive densification of Gaussians starts at iteration 230 and ends at iteration 600, with a densification interval of 30 iterations. Additionally, the Gaussian opacity reset interval is set to 200 iterations. We set the spherical harmonics degree to 1 for the NHR dataset and 2 for the DNA-Rendering dataset, as the latter contains more non-Lambertian objects. All other hyperparameters are kept consistent with 3DGS [31].

For the loss function, we set λ_o to 0.01 and λ_m to 0.1. Additionally, we gradually increase λ_m from 0.01 to 0.11, while linearly decaying λ_t from 0.04 to 0.02.

7. Additional Dataset Details

For the DNA-Rendering dataset [10], we evaluate our method on five widely used sequences: 0008_01, 0012_11, 0013_01, 0013_03, and 0013_09, with images downsampled by a factor of 2 and cropped to focus on the foreground region. Following 4K4D [72], we select views 11, 25, 37, and 57 as testing views, with the remaining views used for training. For all scenes in the NHR dataset [70], we reserve views 18, 28, 37, and 46 for evaluation, while the rest serve as the training set.

8. Additional Ablation Study

In this section, we quantitatively evaluate the effectiveness of our method in enhancing temporal consistency. Specifically, we render dynamic mesh sequences from a fixed testing view and calculate SSIM, PSNR, and LPIPS between consecutive frames. Temporal consistency is then measured by averaging these metrics across the entire sequence, with higher scores indicating greater similarity between consecutive frames. Since the scene movement remains consistent for the same rendering view, more similar images across frames suggest higher temporal consistency. As shown in Tab. 4, our curvature-based temporal consistency (TC) module significantly improves smoothness across frames. Additionally, a qualitative evaluation of temporal consistency is provided in the supplementary video.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o GD + w/o TC	29.268	0.946	0.0145
w/o GD	29.569	0.9507	0.0129
w/o TC	29.271	0.9469	0.0145
Ours Full	29.589	0.9514	0.0129

Table 4. Ablation study on the temporal consistency of rendered mesh videos on the NHR dataset.

9. More Results

Free-Viewpoint Rendering. In Tab. 5 and Tab. 6, we provide a detailed per-scene quantitative comparison of our rendering results against various baselines on both the DNA-Rendering and NHR datasets. Additionally, as shown in Fig. 8, our method consistently achieves photo-realistic rendering with fine-grained details.

Surface Reconstruction. We include further qualitative comparisons of dynamic surface geometry on the DNA-Rendering and NHR datasets in Fig. 9. Our method reconstructs high-quality surface meshes across various complex dynamic scenes.

10. Supplementary Video

The supplementary video includes the following:

- Additional ablation study on the impact of temporal consistency loss on dynamic surface meshes.

- A comparison between our method and NeuS2 [67] on dynamic surface meshes.

- Additional results showcasing free-viewpoint renderings of both color images and surface meshes.

11. Potential Societal Impact

While AT-GS advances dynamic surface reconstruction and novel view synthesis, its deployment carries potential

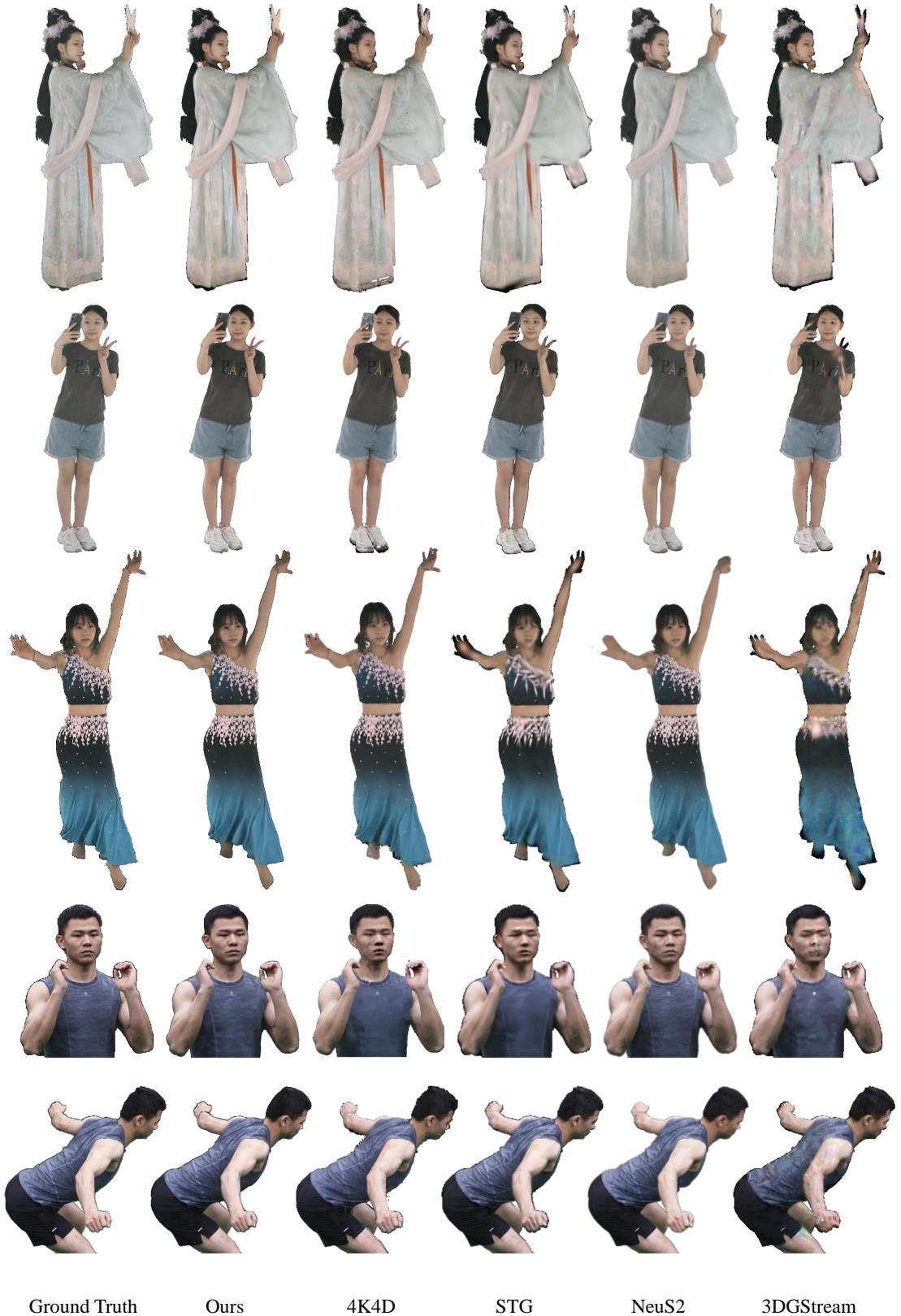| Ground Truth | Ours | 4K4D | STG | NeuS2 | 3DGStream |

Figure 8. Additional qualitative comparison of novel view synthesis on the DNA-Rendering and NHR datasets.

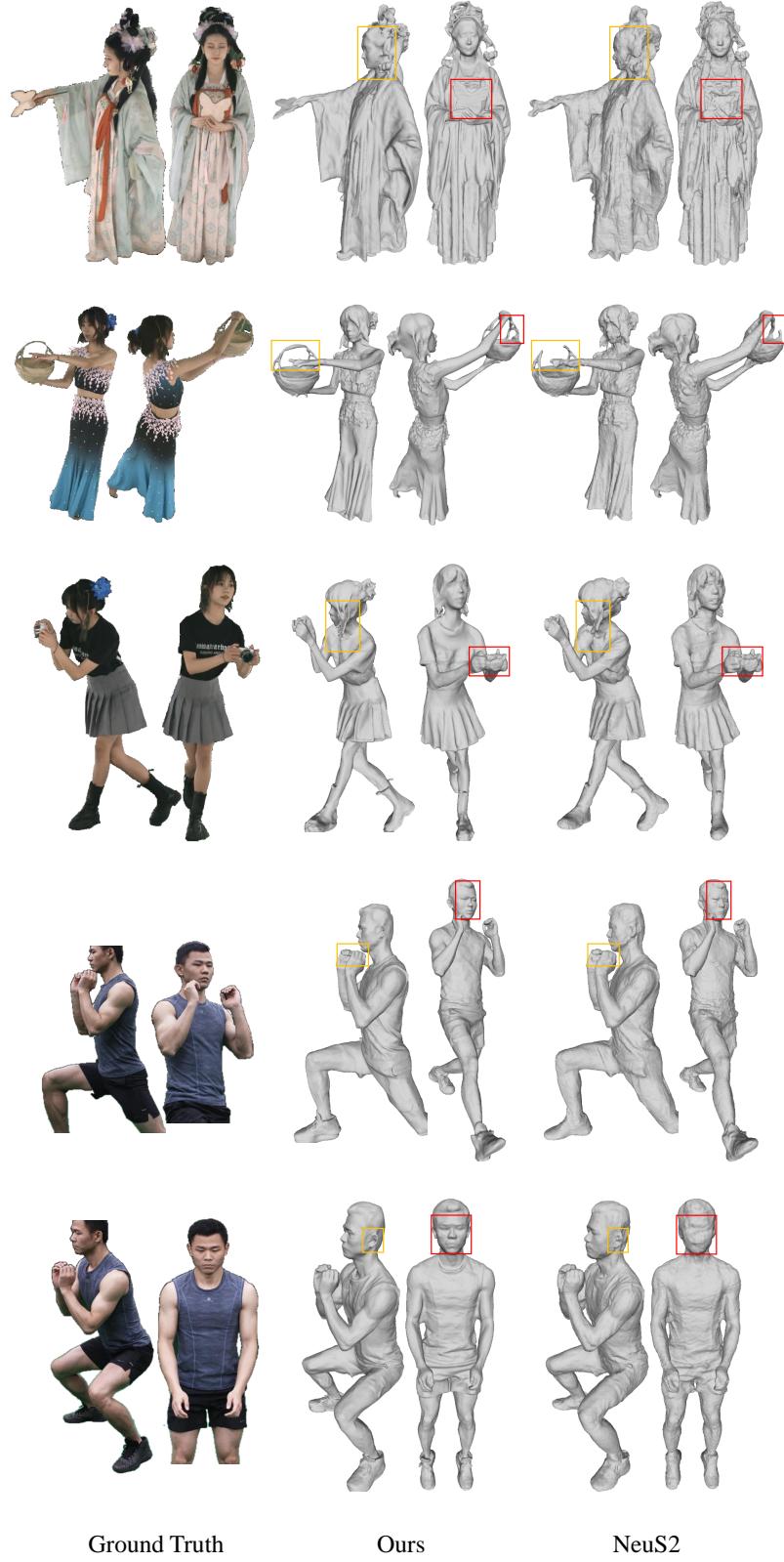Ground Truth   Ours   NeuS2

Figure 9. Additional comparison of surface reconstruction on the DNA-Rendering and NHR datasets.

Type	Method	0008_01			0012_11			0013_01		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Holistic	4K4D	31.36	0.974	0.047	35.81	0.990	0.018	34.52	0.987	0.021
	STG	24.08	0.944	0.068	33.55	0.986	0.023	25.47	0.957	0.047
Incremental	NeuS2	30.24	0.980	0.054	35.54	0.992	0.023	33.33	0.987	0.030
	3DGStream	27.46	0.960	0.075	33.88	0.986	0.033	29.14	0.969	0.047
	Ours	32.07	0.980	0.039	37.03	0.992	0.018	35.46	0.988	0.022

Type	Method	0013_03			0013_09			**Average**		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Holistic	4K4D	34.41	0.986	0.022	36.48	0.989	0.020	34.52	0.985	0.025
	STG	27.49	0.965	0.037	31.84	0.977	0.031	28.49	0.966	0.041
Incremental	NeuS2	33.60	0.987	0.029	36.27	0.990	0.025	33.80	0.987	0.032
	3DGStream	29.78	0.972	0.045	33.63	0.982	0.037	30.78	0.974	0.047
	Ours	35.43	0.988	0.020	37.19	0.990	0.020	35.44	0.988	0.024

Table 5. Per-scene quantitative results on the DNA-Rendering dataset [10]. The best values are highlighted in red, and the second-best values in yellow. Our method achieves the highest rendering quality compared to all other baselines.

Method	sport_1			sport_2			sport_3			basketball			**Average**		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
4K4D	33.37	0.975	0.026	34.57	0.968	0.052	34.19	0.968	0.051	32.49	0.977	0.027	33.65	0.972	0.039
STG	28.65	0.952	0.068	29.88	0.958	0.065	26.34	0.940	0.084	27.35	0.949	0.080	28.05	0.949	0.074
NeuS2	33.53	0.975	0.038	33.62	0.971	0.047	33.35	0.972	0.044	31.66	0.970	0.057	33.04	0.972	0.047
3DGStream	31.73	0.960	0.070	31.12	0.955	0.082	30.86	0.954	0.083	29.08	0.951	0.096	30.70	0.955	0.083
Ours	33.64	0.974	0.046	34.42	0.973	0.056	34.14	0.974	0.052	31.99	0.972	0.060	33.55	0.973	0.054

Table 6. Per-scene quantitative results on the NHR dataset [70].

negative societal impacts. When combined with generative technology, it could be misused to create hyper-realistic deepfakes or synthetic media, leading to disinformation, privacy breaches, and security risks. The high-fidelity reconstruction capabilities may also be exploited for intrusive surveillance, further raising privacy concerns. Additionally, although more efficient than some methods, the computational demands of AT-GS could contribute to environmental impact due to energy consumption, especially at scale. It is essential for researchers to remain vigilant and prioritize ethical use, alongside exploring safeguards to mitigate these risks.

References

[1] Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. *arXiv preprint arXiv:2404.03613*, 2024. 2

[2] 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*, 35:967–981, 2022. 2

[3] 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*, pages 130–141, 2023. 2

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 2

[5] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *arXiv preprint arXiv:2406.06521*, 2024. 4

[6] Decai Chen, Haofei Lu, Ingo Feldmann, Oliver Schreer, and Peter Eisert. Dynamic multi-view scene reconstruction using neural implicit surface. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[7] Decai Chen, Markus Worchel, Ingo Feldmann, Oliver Schreer, and Peter Eisert. Accurate human body reconstruction for volumetric video. In *2021 International Conference on 3D Immersion (IC3D)*, pages 1–8. IEEE, 2021. 2

[8] Decai Chen, Peng Zhang, Ingo Feldmann, Oliver Schreer, and Peter Eisert. Recovering fine details for neural implicit

surface reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4330–4339, 2023. 2

[9] Hanlin Chen, Chen Li, and Gim Hee Lee. Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance. *arXiv preprint arXiv:2312.00846*, 2023. 3

[10] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. *arXiv preprint*, arXiv:2307.10173, 2023. 1, 6, 7, 9, 12

[11] Xinlin Ren Chenjie Cao and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[12] Jaesung Choe, Christopher Choy, Jaesik Park, In So Kweon, and Anima Anandkumar. Spacetime surface regularization for neural dynamic scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17871–17881, 2023. 2

[13] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 4, 6, 9

[14] 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*, pages 10715–10725, 2024. 2

[15] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8585–8594, 2022. 2

[16] 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*, pages 1–11, 2024. 2

[17] 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*, pages 12479–12488, 2023. 2

[18] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[19] 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*, pages 5712–5721, 2021. 2

[20] 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. 2, 3

[21] graphdeco inria. Issue #89: Question and answers about randomness in the optimization progress of the gaussian splatting. https://github.com/graphdeco-inria/gaussian-splatting/issues/89, 2024. 5

[22] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 2, 3

[23] Zhiyang Guo, Wengang Zhou, Li Li, Min Wang, and Houqiang Li. Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction. *arXiv preprint arXiv:2403.11447*, 2024. 2, 3

[24] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 4

[25] 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*, pages 4220–4230, 2024. 2

[26] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2

[27] Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19734–19745, 2024. 2

[28] Erik Johnson, Marc Habermann, Soshi Shimada, Vladislav Golyanik, and Christian Theobalt. Unbiased 4d: Monocular 4d reconstruction with a neural deformation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6607, 2023. 2

[29] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama. An efficient 3d gaussian representation for monocular/multi-view dynamic scenes. *arXiv preprint arXiv:2311.12897*, 2023. 2, 3

[30] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 4

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 4, 9

[32] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 9

[33] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2

[34] P. Knoll, W. Morgenstern, A. Hilsmann, and P. Eisert. Animating nerfs from texture space: A framework for pose-dependent rendering of human performances. In *Proc. Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, pages 404–413, Rome, Italy, 2024. 2

[35] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. *arXiv preprint arXiv:2312.00112*, 2023. 2

[36] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming radiance fields for 3d video synthesis. *Advances in Neural Information Processing Systems*, 35:13485–13498, 2022. 2

[37] 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*, pages 5521–5531, 2022. 2

[38] 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*, pages 8508–8520, 2024. 2, 6

[39] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 2

[40] 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 (CVPR)*, 2021. 2

[41] 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. *arXiv preprint arXiv:2312.11458*, 2023. 2

[42] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. High-fidelity and real-time novel view synthesis for dynamic scenes. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–9, 2023. 2

[43] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 2

[44] Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from monocular videos. *arXiv preprint arXiv:2404.12379*, 2024. 2, 4

[45] 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)*, pages 800–809. IEEE, 2024. 2, 3

[46] Shaojie Ma, Yawei Luo, and Yi Yang. Reconstructing and simulating dynamic 3d objects with mesh-adsorbed gaussian splatting. *arXiv preprint arXiv:2406.01593*, 2024. 2, 4

[47] 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. In *ECCV*, 2020. 2

[48] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2, 3, 4

[49] Thomas Müller, Fabrice Rousselle, Jan Novák, and Alexander Keller. Real-time neural radiance caching for path tracing. *arXiv preprint arXiv:2106.12372*, 2021. 4

[50] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2

[51] 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. *ACM Trans. Graph.*, 40(6), dec 2021. 2

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 9

[53] 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*, 2020. 2

[54] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 9

[55] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 2

[56] Richard Shaw, Jifei Song, Arthur Moreau, Michal Nazarczuk, Sibi Catley-Chandar, Helisa Dhamo, and Eduardo Perez-Pellitero. Swags: Sampling windows adaptively for dynamic 3d gaussian splatting. *arXiv preprint arXiv:2312.13308*, 2023. 2

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[58] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with de-

composed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 2

[59] 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*, pages 20675–20685, 2024. 2, 3, 4, 6

[60] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5

[61] 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*, pages 17903–17913, 2023. 2

[62] 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 *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 2

[63] Diwen Wan, Ruijie Lu, and Gang Zeng. Superpoint gaussian splatting for real-time high-fidelity dynamic scene reconstruction. *arXiv preprint arXiv:2406.03697*, 2024. 2

[64] 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*, pages 19706–19716, 2023. 2

[65] Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, and Minye Wu. Neural residual radiance fields for streamably free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 76–87, 2023. 2

[66] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2

[67] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 1, 6, 7, 9

[68] 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. *arXiv preprint arXiv:2405.16822*, 2024. 2, 4

[69] 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*, pages 20310–20320, 2024. 2

[70] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 6, 7, 9, 12

[71] Yuting Xiao, Xuan Wang, Jiafei Li, Hongrui Cai, Yanbo Fan, Nan Xue, Minghui Yang, Yujun Shen, and Shenghua Gao. Bridging 3d gaussian and mesh for freeview video rendering. *arXiv preprint arXiv:2403.11453*, 2024. 2

[72] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20040, 2024. 1, 2, 6, 9

[73] 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*, pages 20331–20341, 2024. 2

[74] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 2

[75] 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)*, pages 767–783, 2018. 2

[76] Mulin Yu, Tao Lu, Linning Xu, Lihan Jiang, Yuanbo Xiangli, and Bo Dai. Gsdf: 3dgs meets sdf for improved rendering and reconstruction. *arXiv preprint arXiv:2403.16964*, 2024. 3

[77] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient high-quality compact surface reconstruction in unbounded scenes. *arXiv:2404.10772*, 2024. 2, 4

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6