# UniFusion: Sparse-View 4D Reconstruction via Unified Spatio-temporal Depth Alignment

Yongzhe Lyu*, Shaofei Wang*, Yixin Chen, and Siyuan Huang

State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China

**Abstract.** In this paper, we address the challenging problem of 4D reconstruction from sparse-view videos. This setup usually relies on monocular depth estimation to provide priors for the reconstruction model. A key challenge arises from limited cross-view overlap and temporal variation, making monocular depth predictions inconsistent across views and time. Existing methods align spatial and temporal dimensions in separate stages, requiring foreground segmentation masks while failing to leverage temporal cues for cross-view alignment. Contrary to these methods, we propose a unified spatial-temporal depth alignment framework that jointly resolves cross-view and cross-time inconsistencies without distinguishing foreground/background. Our method represents depth maps across views and time as a set of spatio-temporal neural fields. This representation not only yields fast convergence, but also captures spatio-temporal correlation among depth maps implicitly, without dependence on external segmentation/tracking models. We also propose a multi-view depth-order loss while leveraging the classic scale-and-shift-invariant loss to further improve the final depth quality. The aligned depths initialize and supervise Gaussian splatting models for 4D reconstruction. Experiments on Ego-Exo4D and EgoHuman demonstrate that our improved depth alignment substantially benefits dynamic Gaussian-splatting-based reconstruction methods for novel-time/view synthesis and geometry accuracy/consistency.

**Keywords:** 4D reconstruction · sparse views · multi-view depth alignment · Gaussian splatting · dynamic scenes

## 1 Introduction

Reconstructing dynamic 3D scenes from multi-view videos is of broad interest for augmented/virtual reality, robotics, and content creation. Dense multi-view capture systems [7, 18, 24, 56] employ hundreds of synchronized cameras, achieving impressive reconstruction quality, but require expensive, lab-confined infrastructure. This has motivated a strong practical push toward in-the-wild capture with only a few portable cameras, as exemplified by the Ego-Exo4D dataset [16], which uses three to four cameras to record diverse human activities. The central tension is clear: practitioners want the reconstruction quality of dense-camera studios

* These authors contributed equally.

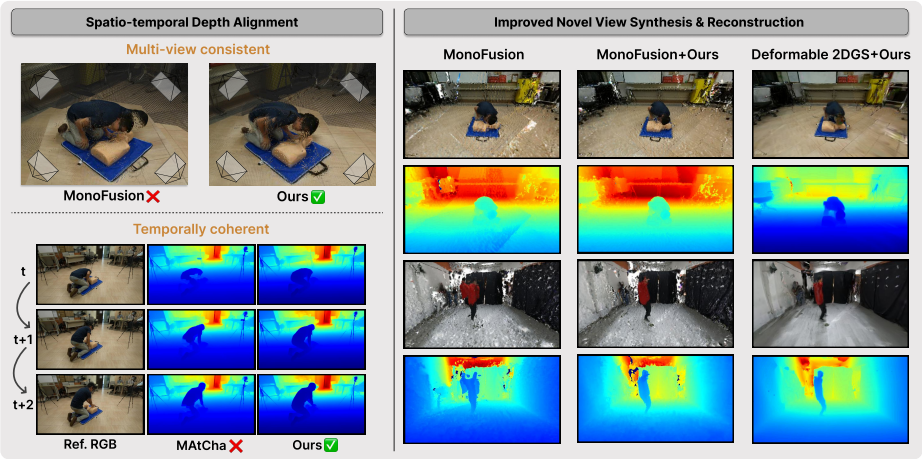


**Fig. 1: Sparse-view 4D Reconstruction via spatially and temporally aligned monocular depths.** *Left:* given 4-view video streams with minimal overlap, we leverage a unified spatio-temporal depth alignment framework to produce multi-view consistent (top) and temporally coherent (bottom) depth maps. *Right*: the aligned depth maps serve as both initialization and supervision for 4D reconstruction, substantially improving the quality over state-of-the-art methods, achieving reasonable novel view synthesis even when shifted $\sim 45^\circ$ from the training view.

from the convenience of sparse-camera setups. Crucially, this sparse-view setting differs from typical "sparse-view" benchmarks where a set of forward-facing cameras shares a large overlap; in our setting, cameras are separated by roughly 90°, providing complete scene coverage but limited cross-view correspondences.

Recent monocular depth and geometry estimators [21, 27, 40, 66, 67] predict remarkably detailed per-view geometry from single images. These foundation models provide strong geometric priors for sparse-view reconstruction. However, monocular depth predictions carry unknown per-view scale/shift and are inconsistent across both views and time. Naively merging them produces chaotic, duplicated geometry with unresolvable artifacts. The central challenge is therefore *alignment*: bringing independently predicted monocular depth maps into a consistent spatio-temporal domain. This is fundamentally harder than the static case, because genuine scene motion is entangled with depth prediction errors, making inconsistencies across views and time difficult to disentangle.

MonoFusion [70] addresses sparse-view dynamic reconstruction by carefully fusing monocular predictions with sparse SfM points [68], achieving promising results. However, it handles spatial alignment and temporal alignment as separate stages. This separation introduces three drawbacks. First, it uses only background for spatial alignment; this requires foreground masks [5, 32, 55] to distinguish static from dynamic regions, adding a fragile dependency on an external segmentation model. Second, temporal alignment heavily depends on tracking models [8–10, 33, 61, 78], which not only result in prolonged preprocessing time

but also introduce additional error accumulation. Third, the independent stages fail to leverage temporal cues for improving cross-view alignment and vice versa. Meanwhile, MAtCha [17] has also demonstrated effective monocular depth map alignment (which they term as "chart alignment") for static sparse-view reconstruction, but does not address temporal variation.

In this work, we propose *unified spatio-temporal depth alignment*: a single framework that jointly optimizes monocular depth alignment across all views and time steps. Our key idea is to represent depth maps across views and time as a set of spatio-temporal neural fields, where all temporal frames within each view share the same spatial representation augmented with a temporal encoding for time-dependent variation. This shared representation provides compactness and smoothness priors that yield fast convergence, while naturally capturing spatio-temporal correlation among depth maps without explicit motion modeling. Crucially, this formulation treats all regions uniformly: the representation learns what varies temporally without being told, eliminating the need for foreground segmentation and pixel tracking, as well as making the method more robust and simpler to deploy. We further propose a multi-view depth-order loss and leverage the classic scale-and-shift-invariant loss to improve the final depth quality.

We instantiate 4D reconstruction with our spatio-temporally aligned depths and demonstrate consistent improvements over different Gaussian splatting variants. Notably, we demonstrate that, when initialized and supervised properly, straightforward Gaussian splatting frameworks [23, 71] substantially outperform existing complex multi-stage processing pipelines [70].

In summary, our contributions are:

- **Unified alignment framework.** We propose to use spatio-temporal neural fields for depth alignment, jointly optimizing monocular depth alignment across all views and time steps in a single model, replacing the separate spatial and temporal alignment stages of prior work.
- **Minimal external dependencies.** Our method handles static and dynamic regions uniformly without distinguishing foreground and background, while capturing motion implicitly without relying on external tracking models. These are natural consequences of the unified formulation rather than adding architectural complexity.
- **Empirical validation.** The aligned depths serve as initialization and supervision for 4D Gaussian splatting, providing consistent improvement for the novel-view/time synthesis (PSNR $+4 \sim 6$dB) and geometric quality (absolute relative error $-14\% \sim 40\%$) on Ego-Exo4D and EgoHuman.

Code is available at https://yongzhelyu.github.io/UniFusion/.

## 2 Related Work

**Monocular depth estimation:** Monocular depth estimation has advanced rapidly, progressing from early CNN-based approaches [13] to modern foundation models with strong zero-shot generalization. MiDaS [54] and Omnidata [12]

established the paradigm of training on diverse, large-scale datasets—the former by mixing existing depth datasets for zero-shot cross-dataset transfer, the latter by generating consistent annotations from 3D scans at scale. Subsequent discriminative methods further improved accuracy: ZoeDepth [3] bridges relative and metric depth through a two-stage architecture, and Metric3D v2 [21] incorporates camera intrinsics to resolve metric ambiguity. In parallel, diffusion-based approaches such as Marigold [27] fine-tune latent diffusion models on synthetic depth data, achieving strong generalization without real-world depth labels. The Depth Anything series [74, 75] scales monocular depth training, first through large-scale self-training on unlabeled images, then by distilling a synthetic-data teacher via pseudo-labeled real images. MoGe [66] takes a complementary approach, predicting affine-invariant 3D point maps rather than depth to eliminate focal-distance ambiguity, with its successor [67] further recovering metric scale depth. Despite these progresses, a fundamental limitation persists: each image is processed independently, yielding per-view predictions with unknown scale and shift. While video depth methods such as DepthCrafter [22] and Geo4D [25] repurpose video diffusion models for temporally consistent depth sequences, they do not address multi-view geometric consistency. Aligning these independently predicted depths across views and over time remains the central challenge that motivates our work.

**Multi-view geometry and depth alignment:** Classical structure from motion (SfM) [20, 47, 49], exemplified by COLMAP [57], produces sparse but globally consistent 3D reconstructions. Recent learning-based methods relax these requirements: DUSt3R [68] and MASt3R [36] predict dense pointmaps from image pairs without explicit camera calibration, VGGT [64] predicts cameras, point maps, and depth from unposed images in a single forward pass. Although these methods produce multi-view consistent point maps, the point maps are still relatively coarse and lack fine details, compared to dedicated monocular depth prediction models. MAtCha [17] introduces chart alignment—fitting per-view neural deformation fields to jointly align monocular depths [74] and point maps [36] toward multi-view consistency, achieving high-quality static geometry from sparse views. However, it does not address dynamic reconstruction. No existing method jointly aligns monocular depths across both views and time; we extend chart alignment to the spatio-temporal domain to fill this gap.

**Dynamic scene reconstruction:** Neural radiance fields [46] enabled photo-realistic novel-view synthesis, and subsequent extensions model dynamic scenes through learned deformation fields or direct temporal conditioning [1, 24, 31, 37, 52, 60, 62, 63]. D-NeRF [53], Nerfies [50], and HyperNeRF [51] pioneered this direction but typically require dense multi-view input or moving cameras and are slow to train. Efficient spatio-temporal representations such as HexPlane [4], K-Planes [14], and Tensor4D [58] factorize 4D volumes into planar or voxel components for fast training, while ResFields [44] adds temporal conditioning through weight residuals. The advent of 3D Gaussian splatting [28] has brought real-time rendering to dynamic scenes. Deformation-based methods [2, 19, 26, 30, 34, 39, 41,

42,59,65,72,77,79] employ explicit motion models to adjust Gaussian parameters over time, while 4D-primitive approaches [11, 15, 35, 38, 43, 69, 71, 73, 76] instead directly model dynamics with temporally extended Gaussian primitives. Sparse-view dynamic reconstruction remains severely under-explored: SplatFields [45] addresses sparse-view synthesis but focuses on foreground objects without background modeling. MonoFusion [70] achieves full dynamic scene reconstruction by fusing monocular estimates, but relies on separate spatial and temporal alignment stages with additional foundational model dependencies [33, 48, 55] aside from monocular depth and SfM points. We demonstrate that unified spatio-temporal depth alignment, combined with a straightforward deformable Gaussian splatting pipeline, outperforms more complex multi-stage approaches, underscoring that effectively leveraging depth priors matters more than architectural complexity.

## 3 Method

Given $N$ synchronized cameras with either known or estimated [36, 64, 68] intrinsics and extrinsics observing a dynamic scene over $T$ time steps, our goal is to reconstruct the 4D scene for novel-view/time synthesis.

Specifically, we leverage a unified spatio-temporal depth alignment framework to provide strong geometric initialization and supervision for 4D reconstruction. It jointly corrects all monocular depth maps across views and time through a neural-field-based deformation model (Sec. 3.1), optimized with multi-view geometric losses (Sec. 3.2). The aligned depths then initialize and supervise subsequent Gaussian splatting reconstruction approaches (Sec. 3.3). An overview is shown in Fig. 2.

### 3.1 Spatio-Temporal Depth Alignment

Our key idea is to represent all depth maps across views and time as a set of spatio-temporal neural fields. Within these fields, all temporal frames of a given view share the same spatial representation, while frame-specific variation is captured by temporal conditioning in the network weights. This formulation allows joint optimization across space and time without separating static and dynamic regions, and hence without requiring segmentation masks or point trajectories.

We build on the per-view depth parameterization of MAtCha [17], which was designed for *static* multi-view alignment, and extend it to the spatio-temporal setting through residual temporal conditioning.

*Time-independent Encoding.* For each view $k$ at time $t$, we first fit an affine transformation to coarsely align the monocular depth obtained from Depth Anything V2 [75] to sparse SfM points from MASt3R [36] via weighted least squares, then backproject the aligned depth $D_k^t$ to 3D:

$$\psi_k^{t,(0)}(\mathbf{u}) = \mathbf{o}_k + D_k^t(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}), \tag{1}$$

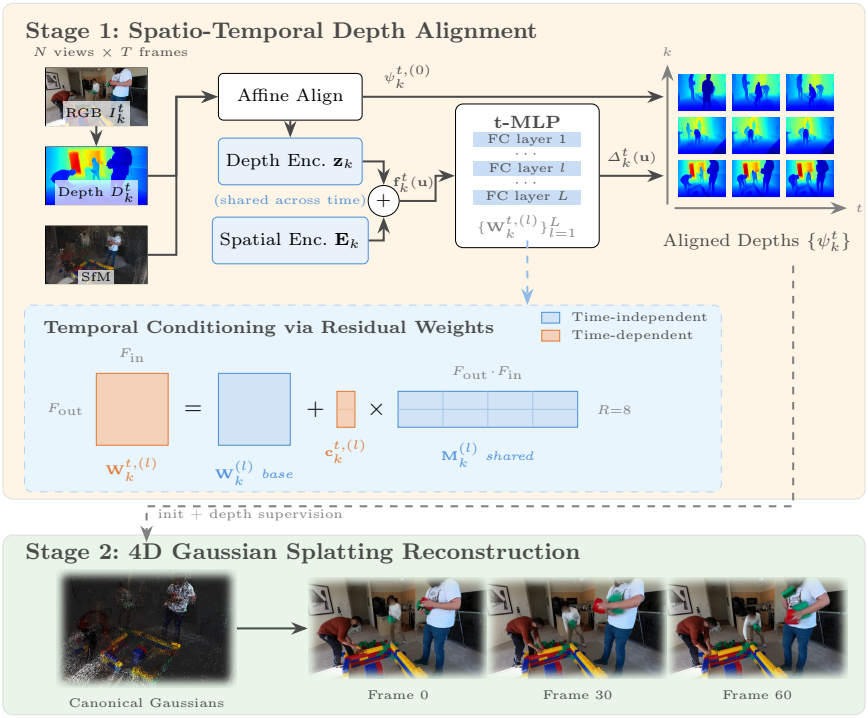


**Fig. 2: Overview of our method.** Given monocular depths (estimated from RGB images) and sparse SfM points from sparse synchronized cameras (*left*), we align them jointly across views and time via per-view spatio-temporal neural fields (*Stage 1*). Each view maintains spatial and depth encodings shared across time, while a per-view deformation MLP predicts depth offsets along ray directions. Temporal variation is captured through low-rank residual weight conditioning (*inset*): each layer's weights decompose into a shared base plus a compact residual parameterized by per-frame temporal coefficients. The aligned depths initialize and supervise 4D Gaussian splatting for novel-view/time synthesis (*Stage 2*).

where $\mathbf{u} \in [0,1]^2$ are normalized pixel coordinates. $\mathbf{o}_k$ and $\mathbf{r}_k(\mathbf{u})$ are the camera center and ray directions at the $k$th view, respectively. Following MAtCha [17], for each view $k$ we maintain two learnable encodings *shared* across all $T$ time steps: a *spatial encoding* $\mathbf{E}_k \in \mathbb{R}^{r_h \times r_w \times d}$, a sparse 2D feature grids interpolated bilinearly as $\mathbf{E}_k(\mathbf{u})$; and a *depth encoding* $\mathbf{z}_k$, a 1D grid along the depth axis queried by quantile-normalized depth $\mathbf{z}_k(D_k^t(\mathbf{u}))$. The two are combined by addition to form the time-independent encoding:

$$\mathbf{f}_k^t(\mathbf{u}) = \mathbf{E}_k(\mathbf{u}) + \mathbf{z}_k(D_k^t(\mathbf{u})). \tag{2}$$

*Deformation MLP* The combined encoding Eq. (2) is fed to per-view deformation networks $\{f_{\theta_k}\}_{k=1}^N$ to predict refined/aligned depth maps of a *static* scene. With a slight abuse of notations, we temporarily remove the $t$ superscript for

Eq. (2). The deformation network $f_{\theta_k} : \mathbb{R}^d \to \mathbb{R}$ takes the input encoding $\mathbf{f}_k(\mathbf{u})$ and predicts a scalar deformation $\Delta_k(\mathbf{u})$ along the ray direction. Concretely, each deformation network consists of three fully-connected hidden layers with 64-dimensional features and ReLU activations. Layer $l$ of view $k$'s network computes:

$$\mathbf{h}_k^{(l)} = \mathrm{ReLU}\big(\mathbf{W}_k^{(l)}\,\mathbf{h}_k^{(l-1)} + \mathbf{b}_k^{(l)}\big), \tag{3}$$

where $\mathbf{W}_k^{(l)} \in \mathbb{R}^{F_{\text{out}}^{(l)} \times F_{\text{in}}^{(l)}}$ are learnable weights and $\mathbf{h}_k^{(0)} = \mathbf{f}_k(\mathbf{u})$. Each deformation MLP is essentially a *neural field* that takes an input encoding derived from 2.5D query coordinates and predicts depth offsets $\Delta_k(\mathbf{u})$. Note that at this point, the MLP does not take temporal conditioning as inputs, which we will describe next.

**Temporal conditioning via residual weights.** To capture temporal variation while sharing spatial representations across time, we replace the static weights $\mathbf{W}_k^{(l)}$ in each hidden layer of the deformation MLP with time-dependent versions via per-timestep weight residuals [44]. This essentially extends the spatial neural field to a *spatio-temporal* neural field, enabling joint alignment of depth maps across views and time simultaneously. The smoothness bias of neural fields also significantly reduces the temporal jittering of aligned depth maps.

Specifically, the weight matrix of the hidden layer $l$ at view $k$ decomposes into a base $\mathbf{W}_k^{(l)}$ shared across all frames plus a time-dependent residual:

$$\mathbf{W}_k^{t,(l)} = \mathbf{W}_k^{(l)} + \delta\mathbf{W}_k^{t,(l)}. \tag{4}$$

To keep the parameter count compact, we factor the residual via vector-matrix (VM) decomposition [6, 44]:

$$\delta\mathbf{W}_k^{t,(l)} = (\mathbf{c}_k^{t,(l)})^\top \mathbf{M}_k^{(l)}, \tag{5}$$

where $\mathbf{c}_k^{t,(l)} \in \mathbb{R}^R$ are per-frame temporal coefficients and $\mathbf{M}_k^{(l)} \in \mathbb{R}^{R \times (F_{\text{out}}^{(l)} \cdot F_{\text{in}}^{(l)})}$ is a shared matrix ($R$=8 in practice), adding only $T{\cdot}R + R{\cdot}F_{\text{out}}{\cdot}F_{\text{in}}$ parameters per layer. Eq. (3) can then be extended to handle temporal conditioning by simply querying the weight matrix with $t$:

$$\mathbf{h}_k^{t,(l)} = \mathrm{ReLU}\big(\mathbf{W}_k^{t,(l)}\,\mathbf{h}_k^{t,(l-1)} + \mathbf{b}_k^{(l)}\big). \tag{6}$$

The final aligned depth is as follows:

$$\psi_k^t(\mathbf{u}) = \psi_k^{t,(0)}(\mathbf{u}) + \Delta_k^t(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}). \tag{7}$$

This design is what enables unified spatio-temporal alignment. For static regions, the temporal residuals are free to vanish ($\delta\mathbf{W}^t \approx \mathbf{0}$), producing consistent outputs across frames automatically. For dynamic regions, the residuals activate to produce frame-specific adjustments. Crucially, the network learns this distinction without being told—no foreground/background segmentation is needed. It also learns motion implicitly. This stands in contrast to prior work [70] that requires explicit masks to extract static regions for alignment, and needs point tracking to initialize dynamic regions.

### 3.2 Optimization Objectives

All views and time steps are optimized jointly. We first describe our proposed multi-view depth-order loss and the scale-and-shift-invariant loss, then summarize the alignment losses inherited from MAtCha [17].

*Multi-view depth-order loss.* When multiple views observe the same scene at a given time step, their depth maps must be mutually consistent: a 3D point visible in one view, when reprojected into another, should not lie in front of that view's recorded surface. We enforce this geometric constraint with a cross-view depth-order loss. For each view pair $(k, j)$ at the same time step, we unproject view $k$'s deformed depth $\psi_k^t$ to 3D and reproject it into view $j$. Let $z_{j\leftarrow k}(\mathbf{u})$ denote the reprojected depth in view $j$, and $\hat{d}_j(\mathbf{u}')$ the depth read from view $j$'s depth map at the corresponding pixel $\mathbf{u}' = \pi_j(\Pi_k^{-1}(\mathbf{u}))$. The loss penalizes violations where the reprojected point is closer than the recorded surface:

$$\mathcal{L}_{\text{order}} = \frac{1}{|\mathcal{N}|^2\,|\mathcal{U}|} \sum_{(k,j)\in\mathcal{N}} \sum_{\mathbf{u}\in\mathcal{U}} \max\left(0,\ \frac{\hat{d}_j(\mathbf{u}') - z_{j\leftarrow k}(\mathbf{u})}{\hat{d}_j(\mathbf{u}')}\right), \tag{8}$$

where $\mathcal{U}$ denotes the set of all pixels, $\mathcal{N}$ denotes the set of all view pairs. This loss provides a purely geometric, multi-view consistency signal that requires no monocular depth prior and no explicit cross-view correspondences—the depth maps themselves supply the supervisory signal through reprojection.

*Scale-and-shift-invariant loss.* To provide a complementary dense metric signal, we employ a scale-and-shift-invariant (SSI) loss [54]. Each depth map is normalized by subtracting its median and dividing by its mean absolute deviation:

$$\widetilde{d} = \frac{d - \text{median}(d)}{\text{MAD}(d)}, \qquad \text{MAD}(d) = \text{mean}\big(|d - \text{median}(d)|\big), \tag{9}$$

and the loss is the L1 distance between normalized maps:

$$\mathcal{L}_{\text{ssi}} = \big\|\widetilde{d}_{\text{pred}} - \widetilde{d}_{\text{ref}}\big\|_1. \tag{10}$$

While the depth-order loss enforces multi-view geometric consistency, the SSI loss captures per-view shape similarity after factoring out scale and shift, providing complementary supervision.

*Alignment losses.* We adopt three losses from MAtCha [17] for aligning the deformed depths to sparse observations and preserving monocular geometry. (i) An *SfM fitting loss* $\mathcal{L}_{\text{fit}}$ aligns the deformed depths to sparse 3D points from MASt3R [36, 68], using learnable per-pixel confidence to downweight noisy correspondences. (ii) *Structure preservation losses* $\mathcal{L}_{\text{struct}}$ penalize changes in surface normals (cosine distance) and depth gradients (L1) between the initial and deformed depth maps, ensuring that global alignment corrections do not destroy fine-grained monocular geometry. (iii) A *cross-view matching loss* $\mathcal{L}_{\text{match}}$ projects deformed 3D points into other views and minimizes reprojection error for multi-view consistency.

*Regularization and total loss.* We regularize the spatial encoding magnitudes (L2) and depth encoding smoothness (total variation). The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{fit}} + \lambda_{\text{s}}\mathcal{L}_{\text{struct}} + \lambda_{\text{m}}\mathcal{L}_{\text{match}} + \lambda_{\text{o}}\mathcal{L}_{\text{order}} + \lambda_{\text{ssi}}\mathcal{L}_{\text{ssi}} + \mathcal{L}_{\text{reg}}, \tag{11}$$

where $\lambda_{\text{s}}$, $\lambda_{\text{m}}$, $\lambda_{\text{o}}$, and $\lambda_{\text{ssi}}$ are scalar weights.

### 3.3 4D Gaussian Splatting Reconstruction

We reconstruct 4D scenes using Gaussian splatting initialized from the aligned depth maps. As shown in Sec. 4, our spatio-temporally aligned depth maps consistently improve the quality of a variety of baselines [70, 71].

During training of Gaussian splatting, we follow depth supervision schedules that come with the method [70] if it provides one. For the method [71] that do not come with a depth supervision schedule, we follow that of [17] unless specified otherwise. Please find details of each Gaussian splatting approach used in the Supp. Mat.

## 4 Experiments

We evaluate our approach in this section. Sec. 4.1 describes the experimental setup; Sec. 4.2 compares against state-of-the-art methods; and Sec. 4.3 ablates our design choices.

### 4.1 Experimental Setup

*Datasets.* We evaluate on two real-world, in-the-wild datasets. **Ego-Exo4D** [16] provides in-the-wild sparse-view captures of diverse human activities. Following MonoFusion [70], we use the ExoRecon subset: six scenarios (dance, sports, bike repair, cooking, music, healthcare), each with 300 frames from four exocentric cameras separated by $\sim$90°. Since most of these sequences contain only 4 cameras, we use every 3rd frame for training and every other two frames for validation. **EgoHuman** [29] captures multi-human activities with egocentric and exocentric views. We chose the legoassemble and fencing sequences to conduct our experiments. These two sequences contain 8 and 20 cameras, respectively, which are ideal setups for testing the model's in-the-wild novel-view synthesis capability. We select four exocentric cameras also with $\sim$90° separation and use 300 frames per sequence, while using cameras in between the training cameras for evaluation. We use ground-truth camera poses for both datasets.

*Metrics.* We report PSNR (↑), SSIM (↑), and LPIPS (↓) to measure photometric quality of rendered images. For Ego-Exo4D, we report both full-image metrics and dynamic-region metrics, following MonoFusion's evaluation protocol [70]; we additionally report Absolute Relative Error (AbsRel, ↓) for depth quality. For EgoHuman, we evaluate novel view synthesis on held-out cameras. We also report direct evaluation metrics on depth alignment quality (reprojection consistency, flow-warped depth error, depth-order violation rate) in the Supp. Mat.

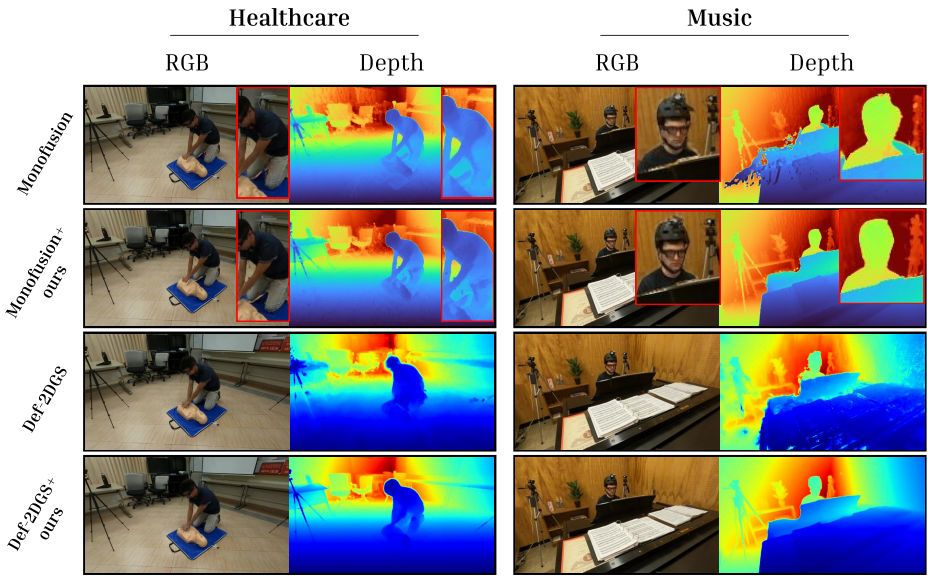


**Fig. 3: Qualitative analysis of held-out frames of training view on EgoExo4D.** Our aligned depth enhances the geometry quality of both MonoFusion and deformable 2DGS, removing ghosting artifacts that are clearly visible in the held-out frames (top two rows), while substantially improving the overall depth quality of deformable 2DGS(bottom two rows).

*Baselines.* We compare against methods reported in MonoFusion [70], copying their published numbers where available (marked with †): SOM, Dynamic 3D Gaussians (Dyn3D-GS) [43], MV-SOM, MV-SOM-SD [65], and MonoFusion [70]. We note that MonoFusion's official code does not contain any evaluation script to reproduce their reported metrics. Some evaluation details, especially on the dynamic regions, are only vaguely described in their paper. We thus re-implement the evaluation script and re-evaluate MonoFusion (no † superscript), which may cause differences in certain metrics compared to those reported in the original MonoFusion paper. We discuss details in the Supp. Mat.

To demonstrate the effectiveness of our proposed depth alignment, we apply the aligned depth to both MonoFusion and a simple deformable 2DGS baseline.

### 4.2 Comparisons with State-of-the-Art

Table 1 presents quantitative results on Ego-Exo4D for both full-image and dynamic-region metrics. Our method consistently improves different Gaussian splatting backends. Notably, our depth alignment substantially improves the geometry quality (AbsRel $-14\%$ for MonoFusion, $-40\%$ for deformable 2DGS). We also note that Def-2DGS + Ours achieves the best overall performance, significantly outperforming the state-of-the-art method (PSNR $+4.25$dB, AbsRel

**Table 1: Quantitative comparison on Ego-Exo4D.** Baseline results (marked †) are copied from MonoFusion [70]. "X + Ours" denotes backend X initialized and supervised with our aligned depths.

| Method | Full image | | | | Dynamic region | | |
|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | AbsRel ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| SOM† | 14.73 | 0.535 | 0.482 | 0.843 | 15.63 | 0.559 | 0.450 |
| Dyn3D-GS† | 24.28 | 0.692 | 0.539 | 0.612 | 24.61 | 0.673 | 0.384 |
| MV-SOM-DS† | 28.37 | 0.906 | 0.079 | 0.398 | 28.23 | 0.931 | 0.063 |
| MV-SOM† | 26.91 | 0.890 | 0.138 | 0.474 | 27.31 | 0.919 | 0.078 |
| MonoFusion† | 30.43 | 0.927 | 0.061 | 0.290 | 29.71 | 0.947 | 0.017 |
| MonoFusion | 30.41 | 0.944 | 0.079 | 0.277 | 29.48 | 0.958 | 0.040 |
| MonoFusion + Ours | 31.63 | 0.955 | 0.064 | 0.238 | 30.44 | 0.972 | 0.032 |
| | (+1.22) | (+.011) | (-.015) | (-.039) | (+.96) | (+.014) | (-.008) |
| Def-3DGS | 34.15 | 0.952 | 0.062 | 0.330 | 37.24 | 0.982 | 0.028 |
| Def-2DGS | 34.15 | **0.964** | 0.058 | 0.376 | 38.64 | 0.984 | 0.026 |
| Def-2DGS + Ours | **34.66** | 0.963 | **0.049** | **0.222** | **38.70** | **0.984** | **0.025** |
| | (+.51) | (-.001) | (-.009) | (-.154) | (+.06) | | (-.001) |

−0.055). The relatively smaller rendering gain over Def-2DGS is mainly because Ego-Exo4D evaluates held-out frames from training cameras, where Def-2DGS can saturate photometric metrics by overfitting on training views. Nevertheless, the large AbsRel reduction shows that our alignment still substantially improves geometry, and the stronger gains under held-out-camera evaluation on the EgoHuman dataset (shown later in this section) further confirm its benefit for challenging novel-view settings.

Figures 3 and 4 shows qualitative comparisons. Our method renders sharper details and more geometrically coherent results, especially in dynamic regions where MonoFusion's separate alignment stages produce temporal artifacts. On the other hand, deformable 2DGS can overfit to training views quite well, but do not produce reasonable depth/geometry due to the lack of proper depth supervision. This is evidenced in novel view rendering, where severe artifacts manifest. In comparison, with our depth initialization and supervision,

**Table 2: Quantitative comparison on EgoHuman.** We evaluate novel view synthesis on two sequences from the EgoHuman dataset, substantially outperforming all baselines.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| SOM | 13.97 | 0.652 | 0.645 |
| MV-SOM | 13.36 | 0.588 | 0.564 |
| SplatFields | 17.74 | 0.675 | 0.571 |
| MonoFusion | 14.15 | 0.436 | 0.628 |
| MonoFusion + Ours | 20.41 | 0.612 | 0.492 |
| | (+6.26) | (+.176) | (-.136) |
| Def-2DGS | 18.88 | 0.657 | 0.431 |
| Def-2DGS + Ours | **22.76** | **0.856** | **0.277** |
| | (+3.88) | (+.199) | (-.154) |

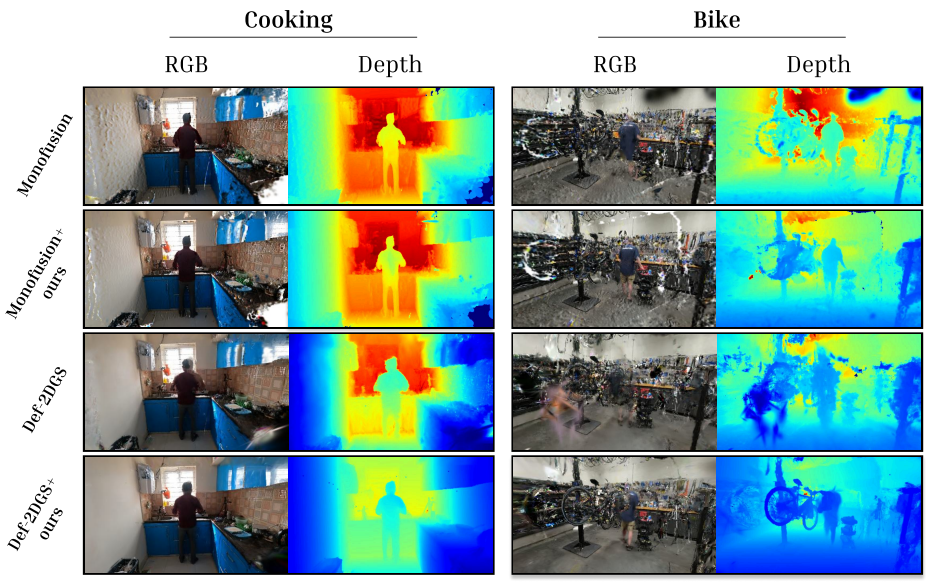


**Fig. 4: Qualitative analysis of novel view synthesis on EgoExo4D.** We show extreme novel view synthesis results ($\sim 45°$ from training view), further confirming that our aligned depth consistently benefits the baseline models.

we achieve reasonable results
on extreme novel views ($30° \sim 45°$).

The novel view synthesis is further demonstrated quantitatively on the Ego-Humans dataset Table 2, where significant improvements of novel view synthesis are achieved over both MonoFusion (+6.26dB) and deformable 2DGS (+3.88dB). Qualitative comparisons are shown in Figure 5.

### 4.3 Ablation Studies

Table 3 progressively ablates our design choices on Ego-Exo4D across different backends.

*(a) Per-frame alignment* applies MAtCha's depth/chart alignment [17] independently per frame, providing cross-view consistency without temporal modeling.

*(b) Time embedding* ablates a naive, temporal-embedding-conditioned MLP extension to (a). This naive implementation results in severe quality degradation.

*(c) Temporal residual weights* incorporates our proposed residual temporal weight ( Sec. 3.1). This gives the single largest improvement. We also note that the chart/depth alignment step of MAtCha over 100 training frames from 4 views takes about half a day, due to the time-consuming per-frame optimization. In

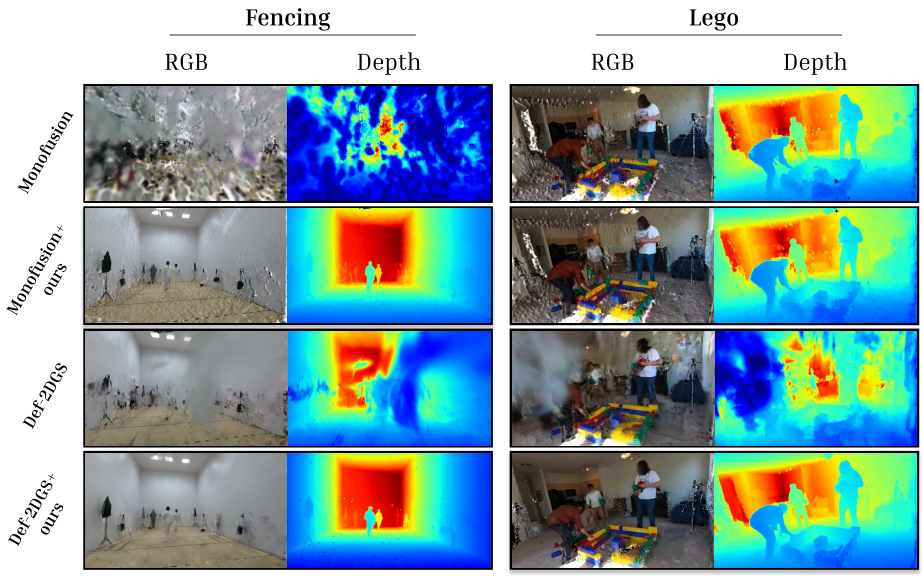


**Fig. 5: Qualitative Results on the EgoHumans dataset.** On the more challenging EgoHumans dataset, our aligned depth helps MonoFusion to recover from catastrophic failure, while significantly improving the novel view rendering quality of deformable 2DGS.

contrast, our spatial-temporal alignment takes only one hour to align the same amount of data, while also achieving improved quality.

*(d) Depth order loss* adds $\mathcal{L}_{\text{order}}$ (Eq. (8)), which regularizes relative depth orderings and yields further improvement.

*(e) Scale-and-shift invariant loss* completes the model with $\mathcal{L}_{\text{ssi}}$ (Eq. (10)), bringing additional gains.

The trends are consistent for both MonoFusion and deformable 2DGS, confirming that our design choices are complementary to the downstream 4D reconstruction method. Additional ablations, including those on 1) the EgoHuman dataset, 2) residual rank and depth order penalty designs, and 3) different camera baselines (wider/narrower), can be found in the Supp. Mat.

## 5 Conclusion

We presented UniFusion, a unified spatio-temporal depth alignment framework for sparse-view 4D reconstruction. By representing per-view depth corrections as spatio-temporal neural fields with temporal residual weight conditioning, our method jointly resolves cross-view and cross-time depth inconsistencies in a single optimization, eliminating the need for separate alignment stages, foreground

**Table 3: Progressive ablation of alignment components** on Ego-Exo4D on two Gaussian splatting backends.

| Configuration | MonoFusion PSNR ↑ | MonoFusion SSIM ↑ | MonoFusion LPIPS ↓ | Def-GS PSNR ↑ | Def-GS SSIM ↑ | Def-GS LPIPS ↓ |
|---|---|---|---|---|---|---|
| (a) Per-frame MAtCha | 30.14 | 0.943 | 0.070 | 32.60 | 0.956 | 0.071 |
| (b) + time embedding | 28.53 | 0.935 | 0.088 | 32.60 | 0.947 | 0.071 |
| (c) + residual weights | 30.90 | 0.948 | 0.067 | 33.61 | 0.960 | 0.055 |
| (d) + $\mathcal{L}_{\text{order}}$ | 31.52 | 0.953 | 0.065 | 33.90 | 0.961 | 0.055 |
| (e) + $\mathcal{L}_{\text{ssi}}$ (full) | 31.63 | 0.955 | 0.064 | 34.66 | 0.963 | 0.049 |

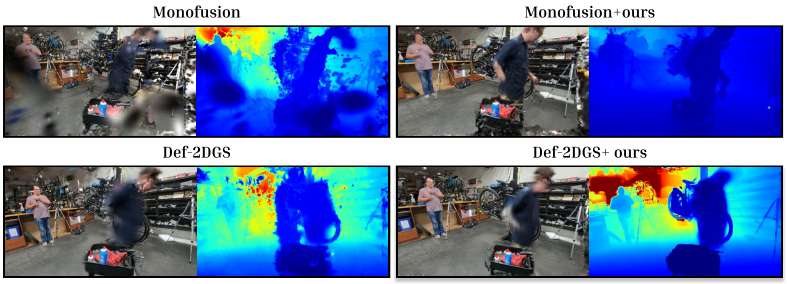


**Fig. 6: Limitations.** While our improved depth consistently enhanced the geometry quality of the different baselines, we note that the naive deformable 2DGS can sometimes become the bottleneck and fail to capture the fast/complex motion. In such a case, MonoFusion, when combined with proper depth initialization/supervision from our method, can produce more reasonable foreground geometry and rendering.

segmentation masks, or pixel tracking models. The aligned depths serve as both initialization and supervision for downstream Gaussian splatting reconstruction.

Experiments on Ego-Exo4D and EgoHuman demonstrate that our alignment consistently improves multiple Gaussian splatting backends, reducing geometric error (AbsRel) by 14–40% and improving novel-view synthesis by 4–6 dB PSNR, while being about 10× faster than naive per-frame spatial alignment. These results highlight that accurately aligned geometric priors are more impactful than downstream architectural complexity for sparse-view 4D reconstruction.

*Limitations.* Our framework relies on monocular depth foundation models and sparse SfM points for initialization; reconstruction quality is therefore bounded by the accuracy of these priors. We also note that the basic deformable 2DGS can fail due to fast and complex motion; in such a case, point tracking information could still help (Figure 6). Promising future directions include further accelerating the alignment process for scalable 4D reconstruction, as well as integrating our alignment module with feed-forward multi-view foundational methods for real-time applications.

## References


1. Attal, B., Huang, J.B., Richardt, C., Zollhoefer, M., Kopf, J., O'Toole, M., Kim, C.: HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
2. Bae, J., Kim, S., Yun, Y., Lee, H., Bang, G., Uh, Y.: Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In: Eur. Conf. Comput. Vis. (2024)
3. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
4. Cao, A., Johnson, J.: HexPlane: A fast representation for dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts. In: Int. Conf. Learn. Represent. (2026)
6. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: Eur. Conf. Comput. Vis. (2022)
7. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. **34**(4), 69:1–69:13 (2015)
8. Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: TAP-vid: A benchmark for tracking any point in a video. In: Adv. Neural Inform. Process. Syst. (2022)
9. Doersch, C., Luc, P., Yang, Y., Gokay, D., Koppula, S., Gupta, A., Heyward, J., Rocco, I., Goroshin, R., Carreira, J., Zisserman, A.: BootsTAP: Bootstrapped training for tracking-any-point. In: Asian Conf. Comput. Vis. (2024)
10. Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: TAPIR: Tracking any point with per-frame initialization and temporal refinement. In: Int. Conf. Comput. Vis. (2023)
11. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4D-Rotor Gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. In: Proc. of ACM SIGGRAPH (2024)
12. Eftekhar, A., Sax, A., Bachmann, R., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans. In: Int. Conf. Comput. Vis. (2021)
13. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Adv. Neural Inform. Process. Syst. (2014)
14. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-Planes: Explicit radiance fields in space, time, and appearance. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
15. Gao, Z., Planche, B., Zheng, M., Choudhuri, A., Chen, T., Wu, Z.: 7DGS: Unified spatial-temporal-angular Gaussian splatting. arXiv preprint arXiv:2503.07946 (2025)
16. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua,

C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
17. Guédon, A., Ichikawa, T., Yamashita, K., Nishino, K.: Matcha gaussians: Atlas of charts for high-quality geometry and photorealism from sparse views. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
18. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P., Izadi, S.: The relightables: Volumetric performance capture of humans with realistic relighting. ACM Trans. Graph. **38**(6), 217:1–217:19 (2019)
19. Guo, Z., Zhou, W., Li, L., Wang, M., Li, H.: Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction. IEEE Transactions on Circuits and Systems for Video Technology **35**(4), 3119–3133 (2025)
20. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
21. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10579–10596 (2024)
22. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: DepthCrafter: Generating consistent long depth sequences for open-world videos. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
23. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: Proc. of ACM SIGGRAPH (2024)
24. Işık, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: HumanRF: High-fidelity neural radiance fields for humans in motion. ACM Trans. Graph. **42**(4), 1–12 (2023)
25. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In: Int. Conf. Comput. Vis. (2025)
26. Jiawei, X., Zexin, F., Jian, Y., Jin, X.: Grid4D: 4D decomposed hash encoding for high-fidelity dynamic gaussian splatting. In: Adv. Neural Inform. Process. Syst. (2024)
27. Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 1–14 (2023)

29. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.: Egohumans: An egocentric 3d multi-human benchmark. In: Int. Conf. Comput. Vis. (2023)
30. Kim, M., Lim, J., Han, B.: 4d gaussian splatting in the wild with uncertainty-aware regularization. In: Adv. Neural Inform. Process. Syst. (2024)
31. Kim, S., Bae, J., Yun, Y., Lee, H., Bang, G., Uh, Y.: Sync-nerf: Generalizing dynamic nerfs to unsynchronized videos. In: AAAI (2024)
32. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Int. Conf. Comput. Vis. (2023)
33. Koppula, S., Rocco, I., Yang, Y., Heyward, J., Carreira, J., Zisserman, A., Brostow, G., Doersch, C.: TAPVid-3D: A benchmark for tracking any point in 3D. In: Adv. Neural Inform. Process. Syst. (2024)
34. Labe, I., Issachar, N., Lang, I., Benaim, S.: Dgd: Dynamic 3d gaussians distillation. In: Eur. Conf. Comput. Vis. (2024)
35. Lee, J., Won, C., Jung, H., Bae, I., Jeon, H.G.: Fully explicit dynamic gaussian splatting. In: Adv. Neural Inform. Process. Syst. (2024)
36. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: Eur. Conf. Comput. Vis. (2024)
37. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., Lv, Z.: Neural 3d video synthesis from multi-view video. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
38. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
39. Liang, Y., Khan, N., Li, Z., Nguyen-Phuoc, T., Lanman, D., Tompkin, J., Xiao, L.: Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis. In: IEEE Winter Conf. Appl. Comput. Vis. (2025)
40. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: recovering the visual space from any views. In: Int. Conf. Learn. Represent. (2025)
41. LIU, Q., Liu, Y., Wang, J., Lyu, X., Wang, P., Wang, W., Hou, J.: MoDGS: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In: Int. Conf. Learn. Represent. (2025)
42. Lu, Z., Guo, X., Hui, L., Chen, T., Yang, M., Tang, X., Zhu, F., Dai, Y.: 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
43. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: Int. Conf. 3D Vis. (2024)
44. Mihajlovic, M., Prokudin, S., Pollefeys, M., Tang, S.: ResFields: Residual neural fields for spatiotemporal signals. In: Int. Conf. Learn. Represent. (2024)
45. Mihajlovic, M., Prokudin, S., Tang, S., Maier, R., Bogo, F., Tung, T., Boyer, E.: SplatFields: Neural gaussian splats for sparse 3d and 4d reconstruction. In: Eur. Conf. Comput. Vis. (2024)
46. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: Eur. Conf. Comput. Vis. (2020)
47. Oliensis, J.: A critique of structure-from-motion algorithms. Computer Vision and Image Understanding **80**(2), 172–214 (2000)
48. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V.,

Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
49. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion. Acta Numerica **26**, 305–364 (2017)
50. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Int. Conf. Comput. Vis. (2021)
51. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (2021)
52. Peng, S., Yan, Y., Shuai, Q., Bao, H., Zhou, X.: Representing volumetric videos as dynamic mlp maps. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
53. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
54. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1623–1637 (2022)
55. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: Int. Conf. Learn. Represent. (2025)
56. Remelli, E., Bagautdinov, T.M., Saito, S., Wu, C., Simon, T., Wei, S.E., Guo, K., Cao, Z., Prada, F., Saragih, J.M., Sheikh, Y.: Drivable volumetric avatars using texel-aligned features. In: Proc. of ACM SIGGRAPH (2022)
57. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
58. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
59. Shaw, R., Nazarczuk, M., Song, J., Moreau, A., Catley-Chandar, S., Dhamo, H., Pérez-Pellitero, E.: Swings: Sliding windows for dynamic 3d gaussian splatting. In: Eur. Conf. Comput. Vis. (2024)
60. Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., Geiger, A.: Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. IEEE Trans. Vis. Comput. Graph. **29**(5), 2732–2742 (2023)
61. Vecerik, M., Doersch, C., Yang, Y., Davchev, T., Aytar, Y., Zhou, G., Hadsell, R., Agapito, L., Scholz, J.: RoboTAP: Tracking arbitrary points for few-shot visual imitation. In: Int. Conf. on Robotics and Automation (2024)
62. Wang, C., Kang, D., Cao, Y., Bao, L., Shan, Y., Zhang, S.H.: Neural point-based volumetric avatar: Surface-guided neural points for efficient and photorealistic volumetric head avatar. In: Proc. of ACM SIGGRAPH Asia (2023)
63. Wang, F., Tan, S., Li, X., Tian, Z., Liu, H.: Mixed neural voxels for fast multi-view video synthesis. In: Int. Conf. Comput. Vis. (2023)
64. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
65. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: Int. Conf. Comput. Vis. (2025)

66. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
67. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. In: Adv. Neural Inform. Process. Syst. (2025)
68. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
69. Wang, Y., Yang, P., Xu, Z., Sun, J., Zhang, Z., Chen, Y., Bao, H., Peng, S., Zhou, X.: Freetimegs: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. (2025)
70. Wang, Z., Tan, J., Khurana, T., Peri, N., Ramanan, D.: Monofusion: Sparse-view 4d reconstruction via monocular fusion. In: Int. Conf. Comput. Vis. (2025)
71. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
72. Xu, Z., Peng, S., Lin, H., He, G., Sun, J., Shen, Y., Bao, H., Zhou, X.: 4k4d: Real-time 4d view synthesis at 4k resolution. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
73. Xu, Z., Xu, Y., Yu, Z., Peng, S., Sun, J., Bao, H., Zhou, X.: Representing long volumetric video with temporal gaussian hierarchy. ACM Trans. Graph. **43**(6), 1–18 (2024)
74. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
75. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Adv. Neural Inform. Process. Syst. (2024)
76. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In: Int. Conf. Learn. Represent. (2024)
77. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
78. Zholus, A., Doersch, C., Yang, Y., Koppula, S., Patraucean, V., He, X.O., Rocco, I., Sajjadi, M.S.M., Chandar, S., Goroshin, R.: Tapnext: Tracking any point (tap) as next token prediction. In: Int. Conf. Comput. Vis. (2025)
79. Zhu, R., Liang, Y., Chang, H., Deng, J., Lu, J., Yang, W., Zhang, T., Zhang, Y.: Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. In: Adv. Neural Inform. Process. Syst. (2024)