MonoFusion: Sparse-View 4D Reconstruction via Monocular Fusion

Zihan Wang Jeff Tan Tarasha Khurana[*] Neehar Peri[*] Deva Ramanan

Carnegie Mellon University

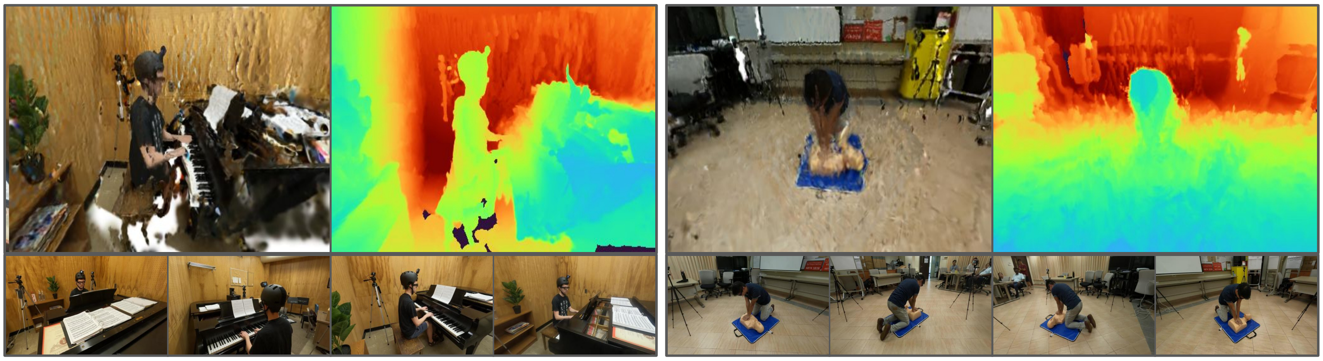

Figure 1. **Dynamic Scene Reconstruction from Sparse Views**. MonoFusion reconstructs dynamic human behaviors, such as playing the piano or performing CPR, from four equidistant inward-facing static cameras. We visualize the RGB and depth renderings of a 45° novel view between two training views. Training views are shown below for reference.

Abstract

We address the problem of dynamic scene reconstruction from sparse-view videos. Prior work often requires dense multi-view captures with hundreds of calibrated cameras (e.g. Panoptic Studio) - such multi-view setups are prohibitively expensive to build and cannot capture diverse scenes in-the-wild. In contrast, we aim to reconstruct dynamic human behaviors, such as repairing a bike or dancing, from a small set of sparse-view cameras with complete scene coverage (e.g. four equidistant inward-facing static cameras). We find that dense multi-view reconstruction methods struggle to adapt to this sparse-view setup due to limited overlap between viewpoints. To address these limitations, we carefully align independent monocular reconstructions of each camera to produce time- and view-consistent dynamic scene reconstructions. Extensive experiments on PanopticStudio and Ego-Exo4D demonstrate that our method achieves higher quality reconstructions than prior art, particularly when rendering novel views.

1. Introduction

Accurately reconstructing dynamic 3D scenes from multi-view videos is of great interest to the vision community, with applications in AR/VR [49] and robot manipulation [26]. Prior work often studies this problem in the context of dense multi-view videos, which require dedicated capture studios that are prohibitively expensive to build and are difficult to scale to diverse scenes in-the-wild. In this paper, we aim to strike a balance between the ease and informativeness of multi-view data collection by reconstructing skilled human behaviors such as repairing a bike and dancing from four equidistant inward-facing static cameras.

Problem setup. Despite recent advances in dynamic scene reconstruction [4, 15–17], current approaches often require dozens of calibrated cameras [23, 36], are category specific [61], or struggle to generate multi-view consistent geometry [34]. We study the problem of reconstructing dynamic human behaviors from an *in-the-wild capture studio*: a small set of (4) portable cameras with limited overlap but complete scene coverage. For example, such a capture mode is common in the large-scale Ego-Exo4D dataset [19]. We argue that sparse-view limited-overlap reconstruction presents unique challenges not found in dense multi-view setups and typical "sparse view" captures with large covisi-

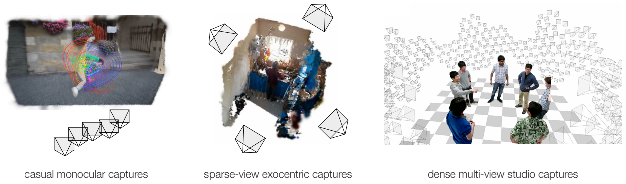
casual monocular captures
sparse-view exocentric captures
dense multi-view studio captures

Figure 2. **Problem Setup.** Our sparse-view setup (**middle**) strikes a balance between ill-posed reconstructions from casual monocular captures [17, 42] and well-constrained reconstructions from dense multi-view studio captures [23]. Unlike existing "sparse-view" datasets like DTU [22] and LLFF [37], our setup is more challenging because input views are 90° apart with limited cross-view correspondences.

blity (Fig. 2). For dense multi-view captures, it is often sufficient to rely solely on geometric and photometric cues for reconstruction, often making use of classic techniques from (non-rigid) structure from motion [12]. As a result, these methods fail in sparse-view settings with limited cross-view correspondences.

Key insights. We find that initializing sparse-view reconstructions with monocular geometry estimators like MoGe [55] produces higher quality results. However, naively merging independent monocular geometry estimates often yields inconsistent geometry across views (e.g. duplicate structures), resulting in local minima during 3D optimization. Instead, we carefully align independent monocular reconstructions to a global reference frame to ensure spatio-temporal consistency. Furthermore, many of the challenges in inferring view-consistent and time-consistent depth become dramatically simplified when working with *fixed cameras with known poses* (inherent to the in-the-wild capture setup that we target). For example, temporal consistent background geometry can be enforced by simply averaging predictions over time.

Contributions. We present three major contributions.

- We highlight the challenge of reconstructing skilled human behaviors in dynamic environments from sparse-view cameras in-the-wild.
- We demonstrate that monocular reconstruction methods can be extended to the sparse-view setting by carefully incorporating monocular depth and foundational priors.
- We extensively ablate our design choices and show that we achieve state-of-the-art performance on PanopticStudio and challenging sequences from Ego-Exo4D.

2. Related Work

Dynamics scene reconstruction. Dynamic scene reconstruction [4] has received significant interest in recent years. While classical work [9, 39] often relies on RGB-D sensors, or strong domain knowledge [2, 7], recent approaches [33, 34] based on neural radiance fields [38] have progressed towards reconstructing dynamic scenes in-the-wild from RGB video alone. However, such methods are computationally heavy, can only reconstruct short video clips with limited dynamic movement, and struggle with extreme novel view synthesis. Recently, 3D Gaussian Splatting [25, 36] has accelerated radiance field training and rendering via an efficient rasterization process. Follow-up works [35, 58, 65] repurpose 3DGS to reconstruct dynamic scenes, often by optimizing a fixed set of Gaussians in canonical space and modeling their motion with deformation fields. However, as Gao et al. [17] points out, such methods often struggle to reconstruct realistic videos. Many works address this shortcoming by relying on 2D point tracking priors [54], fusing Gaussians from many timesteps [30], modeling isotropic Gaussians [50], or exploiting domain knowledge such as human body priors [31, 52]. However, these approaches study the reconstruction problem in the monocular setting. As 4D reconstruction from a single viewpoint is under-constrained, practical robotics setups for manipulation [27] and hand-object interaction [11, 29, 53] adopt camera rigs where a sparse set of cameras capture the scene of interest. Similarly, datasets like Ego-Exo4D [19], DROID [27] and H2O [29] explore sparse-view capture for dynamic scenes in-the-wild.

Novel-view synthesis from sparse views. Both NeRF and 3D Gaussian Splatting require dense input view coverage, which hinders their real-world applicability. Recent works aim to reduce the number of required input views by adding additional supervision and regularization, such as depth [8, 40] or semantics [21, 44, 66]. FSGS [70] builds on Gaussian splatting by producing faithful static geometry from as few as 3 views by unpooling existing Gaussians and adopting extra depth supervision. GaussianObject [60], on the other hand, adds noise to Gaussian attributes and relies on a pre-trained ControlNet [68] to repair low-quality rendered images. Other works such as MVSplat [5] build a cost volume representation and predict Gaussian attributes in a feed-forward manner. However, they only show success in novel view synthesis with small deviations from the nearest training view. For methods that rely on learned priors, high-quality novel view synthesis is often limited to images within the training distribution. Such methods cannot handle diverse real-world geometry. Diffusion-based reconstruction methods [18, 59] try to generate additional views consistent with the sparse input views, but often produce artifacts. In our case, four sparse view cameras are separated around 90° apart, posing unique challenges.

Feed-forward geometry estimation. Learning-based methods, such as monocular depth networks, are able to reconstruct 3D objects and scenes by learning strong priors from training data. While early works [10, 13]

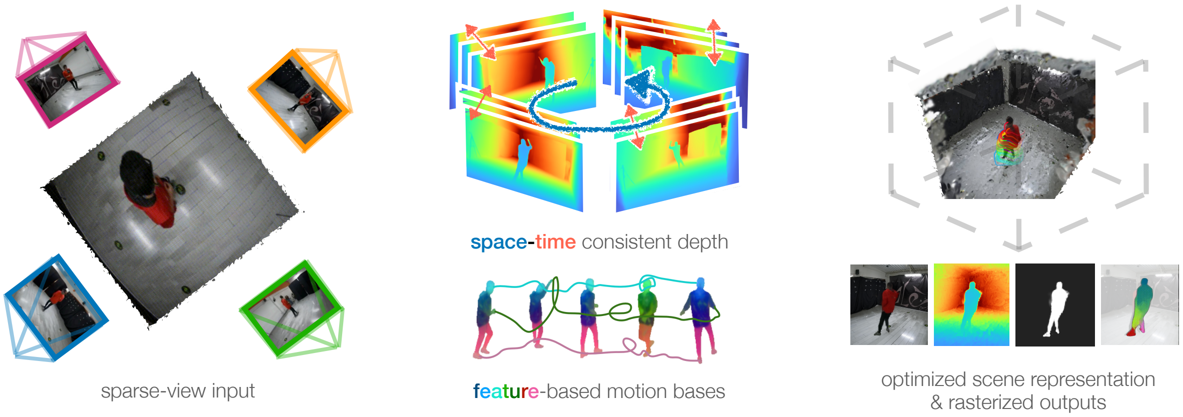
sparse-view input
space-time consistent depth
feature-based motion bases
optimized scene representation
& rasterized outputs

Figure 3. **Approach.** Given sparse-view video sequences of a scene (left), we aim to optimize a 3D gaussian representation over time. We begin by running DUSt3R [56], a *static* multi-view reconstruction method, on the sparse views of a given reference timestamp. This generates a global reference frame that connects all views. Next, we use MoGe [55] to independently predict depth maps for each camera. Since these depth predictions are only defined up to an *affine transformation*, we must estimate a scale and shift for each predicted depth map across all views and time instants. To achieve this, we leverage the fact that background pixels remain static over time. Specifically, for each time instant and each view, we align the background regions of each camera's depth map to the global reference frame by adjusting the scale and shift parameters accordingly (middle, top). This process requires a foreground-background mask for all input videos (which can be obtained using off-the-shelf tools like SAM [47]). To reduce occlusions and noisy depth predictions, we concatenate all aligned background depth points, and average corresponding background points (where correspondence across time is trivially given by the 2D pixel index of the unprojected pointmap) across time. Lastly, we find that motion bases constructed from feature-clustering form a more geometrically consistent set of bases (middle, bottom), than those initialized by noisy 3D tracks [54]. Our optimization yields a 4D scene representation from which we can rasterize RGB frames, depth maps, a foreground silhoutte, and object features from novel views (right).

focus on in-domain depth estimation, recent works build foundational depth models by scaling up the training data [45, 46, 55, 63, 64], resolving the metric ambiguity from various camera models [20, 43, 57], or relying on priors such as Stable Diffusion [14, 24, 48]. Unfortunately, monocular depth networks are not scale or view consistent, and often require extensive alignment against ground-truth to produce meaningful metric outputs. To address these shortcomings, DUSt3R [56] and MonST3R [67] propose the task of point map estimation, which aims to recover scene geometry as well as camera intrinsics and extrinsics given a pair of input images. These methods unify single-view and multi-view geometry estimation, and enable consistent depth estimation across either time or space.

3. Towards Sparse-View 4D Reconstruction

Given sparse-view (i.e. 3 – 4) videos from stationary cameras as input, our method recovers the geometry and motion of a dynamic 3D scene. We model the scene as a set of canonical 3D Gaussians (Sec. 3.1), which translate and rotate via a linear combination of motion bases. We initialize consistent scene geometry by carefully aligning multi-view geometry predictions (Sec. 3.2), and initialize motion trajectories by clustering per-point 3D semantic features distilled from 2D foundation models (Sec. 3.3). We formulate a joint optimization which simultaneously recovers geometry and motion (Sec. 3.4). Fig. 3 provides a summary of our method.

3.1. 3D Gaussian Scene Representation

We represent the geometry and appearance of dynamic 3D scenes using 3D Gaussian Splatting [25], due to its efficient optimization and rendering. Each Gaussian in the canonical frame t_0 is parameterized by $(\mathbf{x}_0, \mathbf{R}_0, \mathbf{s}, \alpha, \mathbf{c})$, where $\mathbf{x}_0 \in \mathbb{R}^3$ is the position of the Gaussian in canonical frame, $\mathbf{R}_0 \in \mathbb{SO}(3)$ is the orientation, $\mathbf{s} \in \mathbb{R}^3$ is the scale, $\alpha \in \mathbb{R}$ is the opacity, and $\mathbf{c} \in \mathbb{R}^3$ is the color. The position and orientation are time-dependent, while the scale, opacity, and color are persistent quantities shared over time. We additionally assign a semantic feature $\mathbf{f} \in \mathbb{R}^N$ to each Gaussian (Sec. 3.3), where N is an arbitrary number representing the embedding dimension of the feature. Empirically, we find that fixing the color and opacity of Gaussians results in a better performance. In summary, for the i-th 3D Gaussian, the optimizable attributes are given by $\Theta^{(i)} = \{\mathbf{x}_0^{(i)}, \mathbf{q}_0^{(i)}, \mathbf{s}^{(i)}, \mathbf{f}^{(i)}\}$. Following [69], the optimized Gaussians are rendered from a given camera into an RGB image and a feature map using a tile-based rasterization procedure.

3.2. Space-Time Consistent Depth Initialization

Similar to recent methods [51, 54], we rely on data-driven monocular depth priors to initialize the position and appearance of 3D Gaussians over time. Given the success of initializing 3DGS with monocular depth estimates in single-view settings [54], one might think to naturally extend this to multi-view settings by repeating monocular depth initialization for each view. However, this naive initialization yields conflicting geometry signals, as monocular depth estimators commonly predict up to an unknown scale and shift factor. Thus, the unprojected monocular depths from separate views are often inconsistent, resulting in duplicated object parts.

Multi-view pointmap prediction. DUSt3R [56] predicts multi-view consistent pointmaps across K input images by first performing pairwise pointmap inference, followed by a global 3D optimization that searches for per-image pointmaps and pairwise similarity transforms (rotation, translation, and scale) that best aligns all pointmaps with each other.

We run DUST3R on the multiview images at time t, but constrain the global optimization to be consistent with the K known stationary camera extrinsics $\{\mathbf{P}_k\}$ and intrinsics $\{\mathbf{K}_k\}$. This produces per-image global pointmaps $\{\chi_k^t\}$ in metric coordinates. One can then compute a depth map by simply projecting each pointmap back to each image with the known cameras

$$d_k^t(u,v)\begin{bmatrix}u & v & 1\end{bmatrix}^T = \mathbf{K}_k \mathbf{P}_k \chi_k^t(u,v) \tag{1}$$

This produces metric-scale multi-view consistent depth maps $d_k^t(u,v)$. However, such depth maps will not be consistent over time.

Spatio-temporal alignment of monocular depth with multi-view consistent pointmaps. In fact, even beyond temporally inconsistency, such multiview predictors tend to underperform on humans since they are trained on multi-view data where dynamic humans are treated as outliers. Instead, we find monocular depth estimators such as MoGe [55] to be far more accurate, but such predictions are not metric (since they are accurate only up to an affine transformation) and are not guaranteed to be consistent across views *or* times. Instead, our strategy is to use the multi-view depth maps from DUST3R as a metric target to *align* monocular depth predictions, which we write as $m_k^t(u,v)$. Specifically, we search for scale and shift factors a_k^t and b_k^t that minimize the following error:

$$\underset{\{a_k^t, b_k^t\}}{\arg\min} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{u,v \in \mathrm{BG}_k^t} \left\| (a_k^t m_k^t(u,v) + b_k^t) - d_k^t(u,v) \right\|^2 \tag{2}$$

where BG_k^t refers to a pixelwise background mask for camera k at frame t. The above uses metric background points as a target for aligning all monodepth predictions. The above optimization can be solved quite efficiently since each time t and view k can be optimized independently with a simple least-squares solver (implying our approach will easily scale to long videos). However, the above optimization will still produce scale factors that are not temporally consistent since the targets are temporally inconsistent as well. But we can exploit the constraint that background points should be *static* across time for stationary cameras. To do so, we replace $d_k^t(u,v)$ with a static target $d_k(u,v)$ obtained by averaging depth maps over time or selecting a canonical reference timestamp. The final set of scaled time- and view-consistent depthmaps are then unprojected back to 3D pointmaps. Note that this tends to produce accurate predictions for static background points, but the dynamic foreground may remain noisy because they cannot be naively denoised by simple temporal averaging. Rather, we rely on motion-based 3DGS optimization to enforce smoothness of the foreground, described next.

3.3. Grouping-based Motion Initialization

Beyond initializing time- and view-consistent geometry in the canonical frame, we also aim to initialize reasonable estimates of the scene motion. We model a dynamic 3D scene as a set of $\mathcal{N}$ canonical 3D Gaussians, along with time-varying rigid transformations $\mathbf{T}_{0\to t} = [\mathbf{R}_{0\to t}\mathbf{t}_{0\to t}] \in \mathbb{SE}(3)$ that warp from canonical space to time t:

$$\mathbf{x}_t = \mathbf{R}_{0\to t}\mathbf{x}_0 + \mathbf{t}_{0\to t} \quad \mathbf{R}_t = \mathbf{R}_{0\to t}\mathbf{R}_0 \tag{3}$$

Motion bases. Similar to Shape of Motion [54], we make the observation that in most dynamic scenes, the underlying 3D motion is often low-dimensional, and composed of simpler units of rigid motion. For example, the forearms tend to move together as one rigid unit, despite being composed of thousands of distinct 3D Gaussians. Rather than storing independent 3D motion trajectories for each 3D Gaussian (i), we define a set of B learnable basis trajectories $\{\mathbf{T}_{0\to t}^{(i,b)}\}_{b=1}^{B}$. The time-varying rigid transforms are written as a weighted combination of basis trajectories, using fixed per-point basis coefficients $\{w^{(i,b)}\}_{b=1}^{B}$:

$$\mathbf{T}_{0\to t}^{(i)} = \sum_{b=1}^{B} \mathbf{w}^{(i,b)} \mathbf{T}_{0\to t}^{(i,b)} \tag{4}$$

Motion bases via feature clustering. Unlike Shape of Motion which initializes motion bases by clustering 3D tracks, our key insight is that semantically grouping similar scene parts together can help regularize dynamic scene motion, without ever initializing trajectories from noisy 3D track predictions. Inspired by the success of robust and universal feature descriptors [41], we obtain pixel-level features for each input image by evaluating DINOv2 on an image pyramid. We average features across pyramid levels and reduce the dimension to 32 via PCA [1]. We choose the small DINOv2 model with registers, as it produces fewer peaky feature artifacts [6].

Given the consistent pixel-aligned pointmaps $\chi_{t,k}^{\text{(time+view)}}$, we associate each pointmap with the 32-dim feature map $\mathbf{f}_{t,k}$ computed from the corresponding image. We perform k-means clustering on per-point features $\mathbf{f}$ to produce b initial clusters of 3D points. After initializing 3D Gaussians from pointmaps, we set the motion basis weight $\mathbf{w}^{(i,b)}$ to be the L2 distance between the cluster center and 3D Gaussian center. We initialize the basis trajectories $\mathbf{T}_{0\to t}^{(b)}$ to be identity, and optimize them via differentiable rendering.

3.4. Optimization

As observed in prior work [16, 32], using photometric supervision alone is insufficient to avoid bad local minima in a sparse-view setting. Our final optimization procedure is a combination of photometric losses, data-driven priors, and regularizations on the learned geometry and motions.

During each training step, we sample a random timestep t and camera k. We render the image $\hat{\mathbf{I}}_{t,k}$, mask $\hat{\mathbf{M}}_{t,k}$, features $\hat{\mathbf{F}}_{t,k}$, and depth $\hat{\mathbf{D}}_{t,k}$. We compute reconstruction loss by comparing to off-the-shelf estimates:

$$\mathcal{L}_{\text{recon}} = \left\|\hat{\mathbf{I}} - \mathbf{I}\right\|_1 + \lambda_{\text{m}} \left\|\hat{\mathbf{M}} - \mathbf{M}\right\|_1 + \lambda_{\text{f}} \left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1 + \lambda_{\text{d}} \left\|\hat{\mathbf{D}} - \mathbf{D}\right\|_1 \tag{5}$$

We additionally enforce a rigidity loss between randomly sampled dynamic Gaussians and their k nearest neighbors. Let $\hat{\mathbf{X}}_t$ denote the location of a 3D Gaussian at time t, and let $\hat{\mathbf{X}}_{t'}$ denote its location at time t'. Over neighboring 3D Gaussians i, we define:

$$\mathcal{L}_{\text{rigid}} = \sum_{\text{neighbors } i} \left\|\hat{\mathbf{X}}_t - \hat{\mathbf{X}}_t^{(i)}\right\|_2^2 - \left\|\hat{\mathbf{X}}_{t'} - \hat{\mathbf{X}}_{t'}^{(i)}\right\|_2^2 \tag{6}$$

4. Experimental Results

Implementation details. We optimize our representation with Adam [28]. We use 18k gaussians for the foreground and 1.2M for the background. We fix the number of $\mathbb{SE}(3)$ motion bases to 28 and obtain these from feature clustering (Sec. 3.3). For the depth alignment, we use points above the confidence threshold of 95%. We show results on 7 10-sec long sequences at 30fps with a resolution of 512 × 288. Training takes about 30 minutes on a single NVIDIA A6000 GPU. Our rendering speed is about 30fps.

Datasets. We conduct qualitative and numerical evaluation on Panoptic Studio [23] and a subset of Ego-Exo4D [19] which we call ExoRecon.

Panoptic Studio is a massively multi-view capture system which consists of 480 video streams of humans performing skilled activities. Out of these 480 views, we manually select 4 camera views, 90° apart to simulate the same exocentric camera setup as Ego-exo4D (see below). Given these 4 training view cameras, we find 4 other intermediate cameras that lie 45° apart from the training views and use these for evaluating novel view synthesis from 45° camera views.

For in-the-wild evaluation of sparse-view reconstruction, we repurpose Ego-Exo4D [19], which includes sparse-view videos of skilled human activities. While many Ego-Exo4D scenarios are out of scope for dynamic reconstruction with existing methods (due to fine-grained object motion, specular surfaces, or excessive scene clutter), we find one scene each from the 6 different scenarios in Ego-Exo4D with considerable object motion: *dance*, *sports*, *bike repair*, *cooking*, *music*, *healthcare*. For each scene, we extract 300 frames of synchronized RGB video streams, captured from 4 different cameras with known parameters. We remove fisheye distortions from all RGB videos and assume a simple pinhole camera model after undistortion. We call this subset ExoRecon, and show results on these sequences. Please see supplement for more visuals.

Metrics. We follow prior work [36, 62] in evaluating the perceptual and geometric quality of our reconstructions using PSNR, SSIM, LPIPS and absolute relative (AbsRel) error in depth. We compute these metrics on the entire image, and also on only the foreground region of interest. We additionally evaluate the quality of the dynamic foreground silhouette by reporting mask IoU, computed as $(\hat{\mathbf{M}}\&\mathbf{M})/(\hat{\mathbf{M}}||\mathbf{M})$. Similar to prior work [62], our evaluation views are a set of held-out frames, subsampled from the input videos from 4 exocentric cameras, in both Panoptic Studio and ExoRecon.

Note that since the cameras in our setup are stationary, above evaluation only analyses the *interpolation* quality of different methods. More explicitly, we also benchmark novel-view synthesis on Panoptic Studio with an evaluation camera placed 45° away from the training view cameras. Since such a ground-truth evaluation camera is not available in ExoRecon, we only show qualitative results.

Baselines. We compare our method with prior work on dynamic scene reconstruction from single or multiple views. Among methods that operate on monocular videos, we run Shape of Motion [54] on 8 scenes from Panoptic Studio following the setup of Dynamic 3D Gaussians [36] and our curated dataset ExoRecon that covers 6 diverse scenes. Finally, we consider two multi-view dynamic reconstruction baselines, Dynamic 3D Gaussians [36], and a naive multi-view extension of Shape of Motion (**MV-SOM**). To construct the latter baseline, we simply concatenate the Gaussians, motion bases, and optimization objectives as four separate instances of single-view SOM. We verify that all baselines reconstruct reasonable training views in the supplement.

Dataset	Method	Full Frame				Dynamic Only			
		PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	PSNR ↑	SSIM ↑	LPIPS ↓	IOU ↑
Panoptic Studio	SOM [54]	17.86	0.687	0.460	0.491	18.75	0.701	0.236	0.358
	Dyn3D-GS [36]	25.37	0.831	0.266	0.207	26.11	0.862	0.129	—
	MV-SOM [54]	26.28	0.858	0.241	0.331	26.80	0.883	0.161	0.886
	MonoFusion	**28.01**	**0.899**	**0.117**	**0.149**	**27.52**	**0.944**	**0.022**	**0.965**
ExoRecon	SOM [54]	14.73	0.535	0.482	0.843	15.63	0.559	0.450	0.294
	Dyn3D-GS [36]	24.28	0.692	0.539	0.612	24.61	0.673	0.384	—
	MV-SOM [54]	26.91	0.890	0.138	0.474	27.31	0.919	0.078	0.845
	MonoFusion	**30.03**	**0.921**	**0.067**	**0.290**	**29.41**	**0.946**	**0.016**	**0.963**

Table 1. **Quantitative analysis of held-out view synthesis.** We benchmark our method against state-of-the-art approaches by evaluating the novel-view rendering and geometric quality on both the dynamic foreground region and the entire scene, across the held-out frames from input videos. MV-SOM is a multi-view version of Shape-of-Motion [54] that we construct by instantiating four different instances of single-view shape of motion, and optimize them together. On Panoptic Studio, groundtruth depth for computing the AbsRel metric is obtained from 27-view optimization of the original Dynamic 3DGS, and for ExoRecon, we project the released point clouds obtained via SLAM from Aria glasses. When evaluating single-view baselines, SOM [54], we naively aggregate their predictions from the four views and evaluate this aggregated prediction against the evaluation cameras.

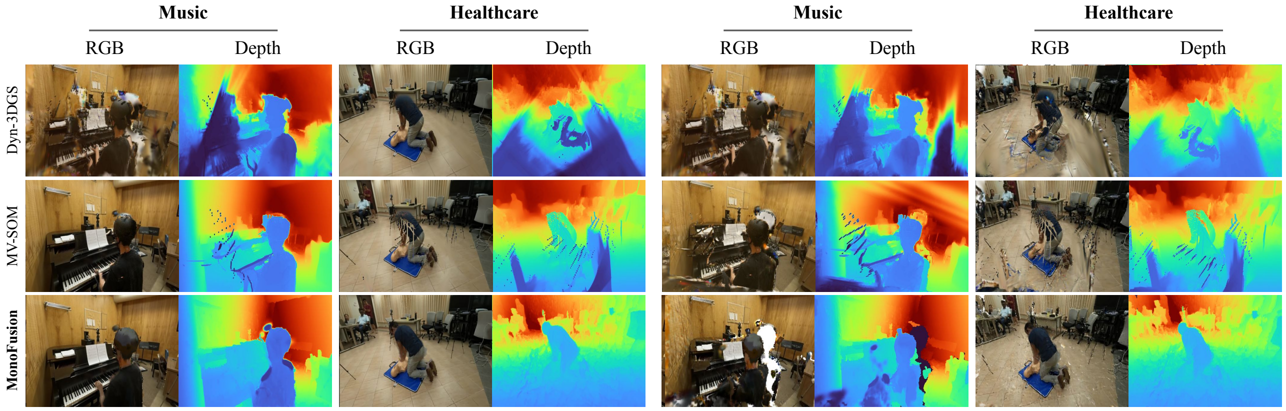

Figure 4. **Qualitative analysis of held-out view synthesis on ExoRecon.** We show qualitative results of held-out view synthesis (**left**) and a 5° deviation from the static camera position at the held-out timestamp (**right**). As compared to other multi-view baselines, our method does dramatically better at interpolating the motion of dynamic foreground (left), even from new camera views (right). We posit that Dynamic 3DGS suffers because of lack of geometric constraints and MV-SOM has duplicate foreground artifacts because of conflicting depth initialization from the four views.

4.1. Comparison to State-of-the-Art

Evaluation on held-out views. In Tab. 1, we compare our method to recent dynamic scene reconstruction baselines [36, 54, 67], following evaluation protocols from prior work [54, 62]. Our method beats prior art on both Panoptic Studio and ExoRecon datasets, when evaluated on held-out views across photometric (PSNR, SSIM, LPIPS) and geometric error (AbsRel) metrics. Note that when initializing Dynamic 3DGS [36] with 4 views we find that COLMAP fails, and so the point cloud initialization for this baseline is from a 27-view COLMAP optimization.

Interestingly, we find that though the monocular 4D reconstruction method Shape of Motion (SOM) [54] always fails to output accurate metric depth, it shows incredibly robustness to a limited camera shift. We hypothesize that the foundational priors of Shape of Motion allow it to produce reasonable results in under-constrained scenarios, while test-time optimization methods, especially ones that do not always rely on data-driven priors like [36], can more easily fall into local optima (e.g. those caused by poor initialization) which are difficult to optimize out of via rendering losses alone.

Evaluation on a 45° novel-view On Panoptic Studio, we use the four evaluation cameras to evaluate the predictions from our method with photometric errors. We also evaluate the rendered depth against a ‘pseudo-groundtruth’ depth ob-

Method	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**IOU** ↑	**AbsRel** (↓)
SOM	16.73	0.554	0.491	0.287	0.578
Dyn3D-GS	23.31	0.776	0.316	—	0.273
MV-SOM	21.56	0.541	0.433	0.482	0.413
MonoFusion	**25.73**	**0.847**	**0.158**	**0.943**	**0.188**

Table 2. **Quantitative analysis of 45° novel-view synthesis on Panoptic Studio.** We benchmark our method against state-of-the-art approaches by evaluating both the dynamic foreground region and the entire scene. Notably, the evaluation is conducted on novel views where the cameras are at least 45 degree shifted from all training views. We additionally evaluate the geometric reconstruction quality with absolute relative (AbsRel) error in rendered depth.

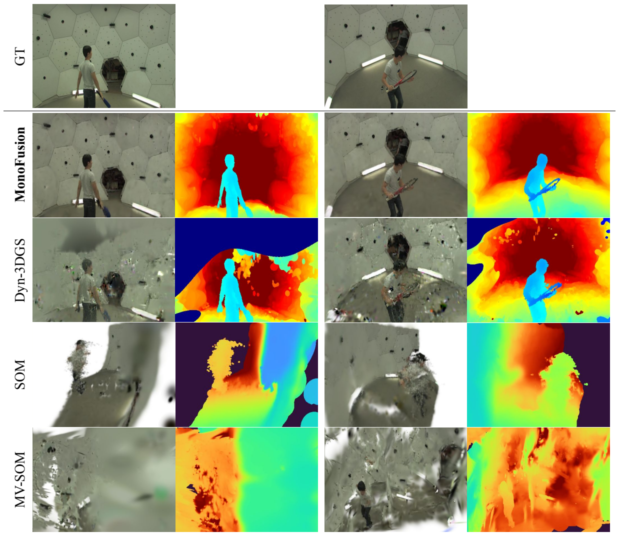

Figure 5. **Qualitative results of** 45° **novel-view synthesis results on Panoptic Studio.** We show qualitative novel-view synthesis results of our method compared to baselines on the softball (left) and tennis (right) sequences. We visualize the groundtruth RGB image for the 45° at the top. Our rendered extreme novel-view RGB image closely matches ground truth. We find that all other baselines struggle to generalize to extreme novel views.

tained from the optimization of Dynamic 3DGS [36] from their 24 training views. We find that all methods achieve low photometric errors, highlighting the difficulty of learning plausible dynamic reconstructions from limited viewpoints. Despite this, our method outperforms all baselines, achieving state-of-the-art results on 45° novel-view synthesis.

4.2. Ablation Study

We ablate the design decisions in our pipeline in Tab. 3. Our proposed space-time consistent depth plays a crucial role in learning accurate scene geometry and appearance (yielding a 3.4 PSNR improvement, Row 1 vs 3). Next, we find that the feature-metric loss $\mathcal{L}_{\text{feat}} = \left\|\hat{\mathbf{F}} - \mathbf{F}\right\|$ provides a trade-off between learning photometric properties vs.learning foreground motion and silhouette. Although the PSNR decreases, we see an increase in mask IoU (Row 1 vs 2 and Row 3 vs 4). Similarly, freezing the color of all Gaussians across frames aids learning the motion mask, as measured by mask IoU. Finally, our motion bases constructed from feature-clustering improve overall scene optimization (final row).

Method	$\mathcal{L}_{\text{feat}}$	$\mathbf{d}_n$	$\mathbf{T}^{(b)}_{0\to t}$	↑PSNR	↑SSIM	↓LPIPS	↑IoU
Baseline	✗	✗	✗	26.19	0.915	0.077	0.60
+ $\mathcal{L}_{\text{feat}}$	✓	✗	✗	25.39	0.933	0.087	0.63
+ Our depth / no $\mathcal{L}_{\text{feat}}$	✗	✓	✗	29.55	0.944	0.037	0.73
+ Our depth / $\mathcal{L}_{\text{feat}}$	✓	✓	✗	29.31	0.941	0.041	0.75
+ Motion bases (**Ours**)	✓	✓	✓	**30.40**	**0.947**	**0.037**	**0.81**

Table 3. **Ablation study of pipeline components.** We ablate our choice of feature-metric loss, spacetime consistent depth, and feature-based motion bases. While the proposed depth and feature-based motion bases considerably improve 4D reconstruction (evaluated by photometric errors), we find that our feature loss helps learn better motion masks (evaluated by IoU).

Velocity-based vs. feature-based motion bases In the monocular setting, we empirically found that both designs performed equally well. However, in our 4 camera sparse view setting, we found that feature-based motion bases perform much better than velocity-based motion bases. The reason is that for velocity-based motion bases, we infer 3D velocity by querying the 2D tracking results plus depth per frame following Shape-of-Motion[54]. Thus, noisy foreground depth estimates where the estimated depth of the person flickers between foreground and backward will negatively influence the quality of velocity-based motion bases, causing rigid body parts to move erratically. In contrast, feature-based motion bases, where features are initialized from more reliable image-level observations, are more robust to noisy 3D initialization and force semantically-similar parts to move in similar ways. To validate our points, in Fig. 11 we use PCA analysis to visualize the inferred features and find that they are consistent not only on temporal axis but also across cameras.

Effect of different number of motion bases. When the number of motion bases is not expressive enough (in our experience when the number of motion bases < 20), there are often obvious flaws in the reconstruction, such as missing arms or the two legs joining together into a single leg. In reality, we do not observe that increasing the number of motion bases further hurts the performance. Empirically, the capacity of our design (which is **28** motion bases) can effectively handle different scene dynamics.

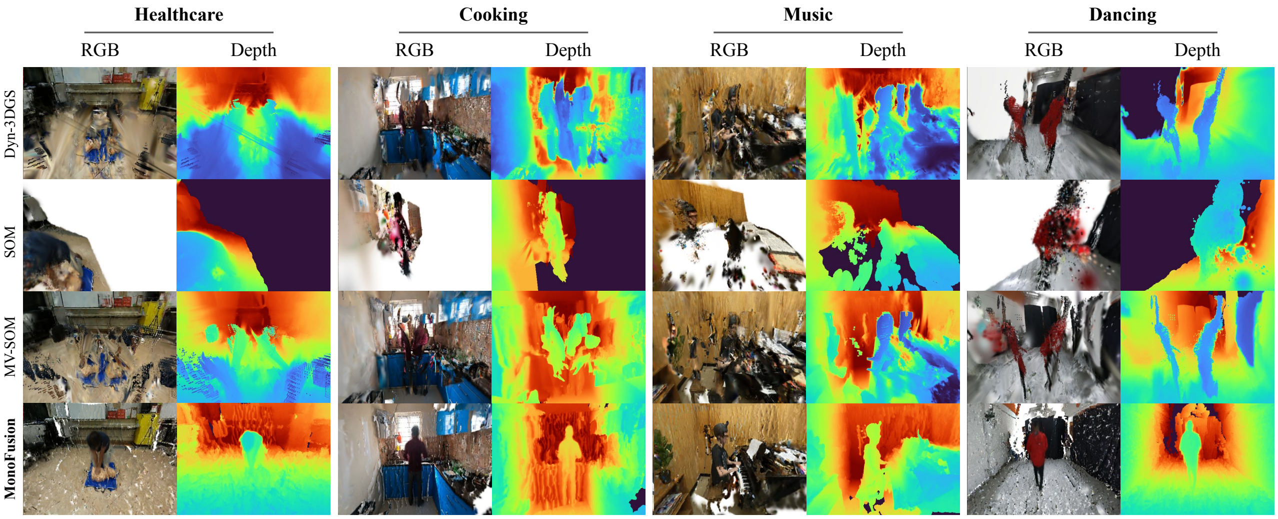

Figure 6. **Qualitative results of 45° extreme novel view synthesis results on ExoRecon (1/2).** We visualize the rasterized RGB image and depth map from each method for 4 diverse EgoExo sequences (see supplement for more scenes). Existing monocular methods and their extension to multi-view produce poor results rendered from a drastically different novel view. MV-SOM improves upon SOM by optimizing a 4D scene representation with four view constraints, but it still suffers from duplication artifacts. Our method's careful point cloud initialization and feature-based motion bases further improve on MV-SOM.

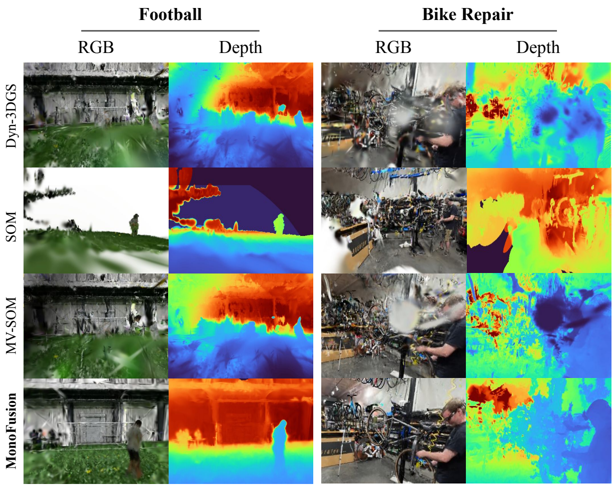

Figure 7. **Qualitative results of 45° extreme novel view synthesis results on ExoRecon (2/2).** We show qualitative novel-view synthesis results of our method compared to baselines on challenging sequence on ExoRecon: highly-dynamic, large scene with small foreground *football* (left) and complex, highly-occluded scene *bike repair* (right). Notably MonoFusion significantly beats other baselines in terms of quality.

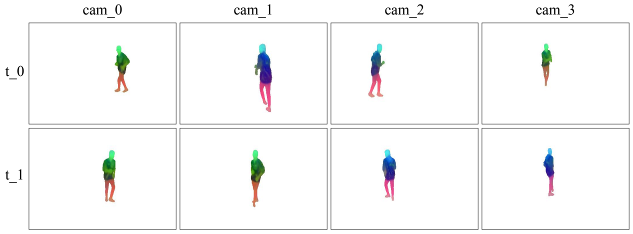

Figure 8. **Spatial-Temporal Visualization of feature PCA.** We perform PCA analysis and transform the 32-dim features from Sec. 3.3 down to 3 dimensions for visualization purposes. We find that the features are consistent across views and across time. Notably, when the person turns around between t_0 and t_1 in observations from cam_1 and cam_2, the feature remains robust and consistent. The semantic consistency of features aids explainability, provides a strong visual clue for tracking, and gives confidence in our feature-guided motion bases.

5. Conclusion

In this work, we address the problem of sparse-view 4D reconstruction of dynamic scenes. Existing multi-view 4D reconstruction methods are designed for dense multi-view setups (e.g. Panoptic Studio). In contrast to prior work, we aim to strike a balance between the ease and informativeness of multi-view data capture by reconstructing skilled human behaviors from four equidistant inward-facing static cameras. Our key insight is that carefully incorporating *priors*, in the form of monocular depth and feature-based motion clustering, are important to enable plausible and photorealistic 4D reconstructions of dynamic scenes. Our empirical analysis shows that we achieve state-of-the-art performance on novel space-time synthesis as compared to prior art for 4D reconstruction, on challenging scenes and object dynamics.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 4

[2] Joel Carranza, Christian Theobalt, Marcus Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22:569–577, 2003. 2

[3] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. *arXiv preprint arXiv:2304.12308*, 2023. 1

[4] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint*, 2024. 1, 2

[5] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images, 2024. 2

[6] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 5

[7] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *SIGGRAPH*, 2008. 2

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021. 2

[9] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2016*, 35, 2016. 2

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, pages 2366–2374, 2014. 2

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2

[12] David A Forsyth and Jean Ponce. A modern approach. *Computer vision: a modern approach*, 17:21–48, 2003. 2

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. *CVPR*, pages 2002–2011, 2018. 2

[14] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024. 3

[15] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[16] 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. 5

[17] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[18] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint 2405.10314*, 2024. 2

[19] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2023. 1, 2, 5

[20] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 3

[21] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[22] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 2

[23] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview

system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(1): 190–204, 2017. 1, 2, 5

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3

[26] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. *arXiv preprint arXiv:2409.18121*, 2024. 1

[27] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021. 2

[30] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2

[31] Jiahui Lei, Yufu Weng, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, 2024. 2

[32] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[33] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[34] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[35] 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*, 2023. 2

[36] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 1, 2, 5, 6, 7

[37] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 2

[38] 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 *European Conference on Computer Vision (ECCV)*, 2020. 2

[39] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[40] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. *arXiv preprint arXiv:2112.00724*, 2021. 2

[41] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4

[42] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[43] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[45] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[46] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 3

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[49] Chonghyuk Song, Gengshan Yang, Kangle Deng, Jun-Yan Zhu, and Deva Ramanan. Total-recon: Deformable scene reconstruction for embodied view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17671–17682, 2023. 1

[50] Colton Stearns, Adam W. Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. *arXiv preprint arXiv:2406.18717*, 2024. 2

[51] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. 2024. 3

[52] Jeff Tan, Donglai Xiang, Shubham Tulsiani, Deva Ramanan, and Gengshan Yang. Dressrecon: Freeform 4d human reconstruction from monocular video. *arXiv preprint arXiv:2409.20563*, 2024. 2

[53] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. *arXiv preprint arXiv:2406.06843*, 2024. 2

[54] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 2, 3, 4, 5, 6, 7

[55] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. 2, 3, 4

[56] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 3, 4

[57] Hao Chen Zhipeng Cai Gang Yu Kaixuan Wang Xiaozhi Chen Chunhua Shen Wei Yin, Chi Zhang. Metric3d: Towards zero-shot metric 3d prediction from a single image. *ICCV*, 2023. 3

[58] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint*, 2023. 2

[59] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024. 2

[60] Chen Yang, Sikuang Li, Jiemin Fang, Ruofan Liang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. *SIGGRAPH Asia*, 2024. 2

[61] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[62] Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Apoorva Sharma, Maximilian Igl, Peter Karkus, Danfei Xu, et al. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. *arXiv preprint arXiv:2501.00602*, 2024. 5, 6

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 3

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 3

[65] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint*, 2023. 2

[66] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[67] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 3, 6, 1

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 3813–3824, 2023. 2

[69] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. *CVPR*, 2024. 3

[70] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. FSGS: real-time few-shot view synthesis using gaussian splatting. *CoRR*, abs/2312.00451, 2023. 2

MonoFusion: Sparse-View 4D Reconstruction via Monocular Fusion

Supplementary Material

A. Results on Additional Sequences

We provide novel-view synthesis results from additional sequences to show the generalization ability among all categories of EgoExo4D. Our additional results cover basic life scenarios, such as healthcare, dancing, cooking, music and bike repair. Specifically, we provide qualitative novel view rendering results from 5° and 45° novel views results. We also include 4D visualizations on the attached website.

B. Training View Renderings

To build confidence in our implementations, we validate every baseline we run by verifying that each method looks reasonable at training views. It is worth noticing that in each iteration of optimization, we sample a batch of frames out of the video to optimize the overall loss. As the loss is optimized as a global minimum averaged over all frames, it is possible that some artifacts remain for certain frames.

C. Training Details

In this section, we report the learning rate and loss weights of Gaussians in our optimization process. These hyperparameters are shared across every scene that we evaluated on. Specifically, $\mathcal{L}_{\text{smooth\_bases}}$ enforces smooth motion bases by penalizing high accelerations in rotations and translations. $\mathcal{L}_{\text{smooth\_tracks}}$ promotes smooth object tracks by penalizing large accelerations in object positions across frames. $\mathcal{L}_{\text{depth\_grad}}$ aligns the gradients of the predicted and ground truth depth maps to preserve structural details. $\mathcal{L}_{\text{z\_accel}}$ penalizes high accelerations along the depth axis to reduce jitter in depth estimation. $\mathcal{L}_{\text{scale\_val}}$ constrains the variance of scale parameters of Gaussians to achieve consistent representations.

Table 4. Learning Rates for Foreground (FG), Background (BG), and Motion Parameters

Parameter	FG LR	BG LR	Motion LR
means	1.6×10^{-4}	1.6×10^{-4}	–
opacities	1×10^{-2}	1×10^{-2}	–
scales	5×10^{-3}	1×10^{-3}	–
quats	1×10^{-3}	1×10^{-3}	–
colors	0	1×10^{-2}	–
feats	1×10^{-3}	1×10^{-3}	–
motion_coefs	1×10^{-3}	–	–
rots	–	–	1.6×10^{-4}
transls	–	–	1.6×10^{-4}

Table 5. Loss Weights Configuration

Loss Parameter	Weight	Loss Parameter	Weight
w_{rgb}	7.0	w_{mask}	5.0
w_{feat}	7.0	$w_{\text{smooth\_bases}}$	0.1
$w_{\text{depth\_reg}}$	1.0	$w_{\text{smooth\_tracks}}$	2.0
$w_{\text{depth\_const}}$	0.1	$w_{\text{scale\_var}}$	0.01
$w_{\text{depth\_grad}}$	0.1	$w_{\text{z\_accel}}$	1.0
w_{track}	2.0		

D. Alternative design choice

E. Limitations and Future Work

We address two key limitations of our work. First, like previous methods, we rely heavily on 2D foundation models to estimate priors (e.g. depth and dynamic masks) for gradient-based differentiable rendering optimization. Thus, imprecise priors can harm the downstream rendering process. In addition, the current pipeline requires a user prompt to specify dynamic masks for each moving object [3], which can be labor-intensive for complex scenes. To solve this, distilling dynamic masks from foundation models or inferring dynamic masks from image level priors (as in [67]) could be beneficial.

Second, most off-the-shelf feed-forward depth estimation networks are trained on simple scene-level datasets, with few dynamic movers (e.g. people) in the foreground. In practice, we observe that the depth of humans in dynamic scenes is often incorrect when observed from other views. For example, DUSt3R often estimates the depth of a human to be the same as the depth of surrounding walls, causing the human to blend into the background. We believe that these fundamental problems with the depth predictions *cannot* be solved by any alignment in the output space. To mitigate this issue, we plan to further fine-tune DUSt3R or MonST3R on existing dynamic human datasets.

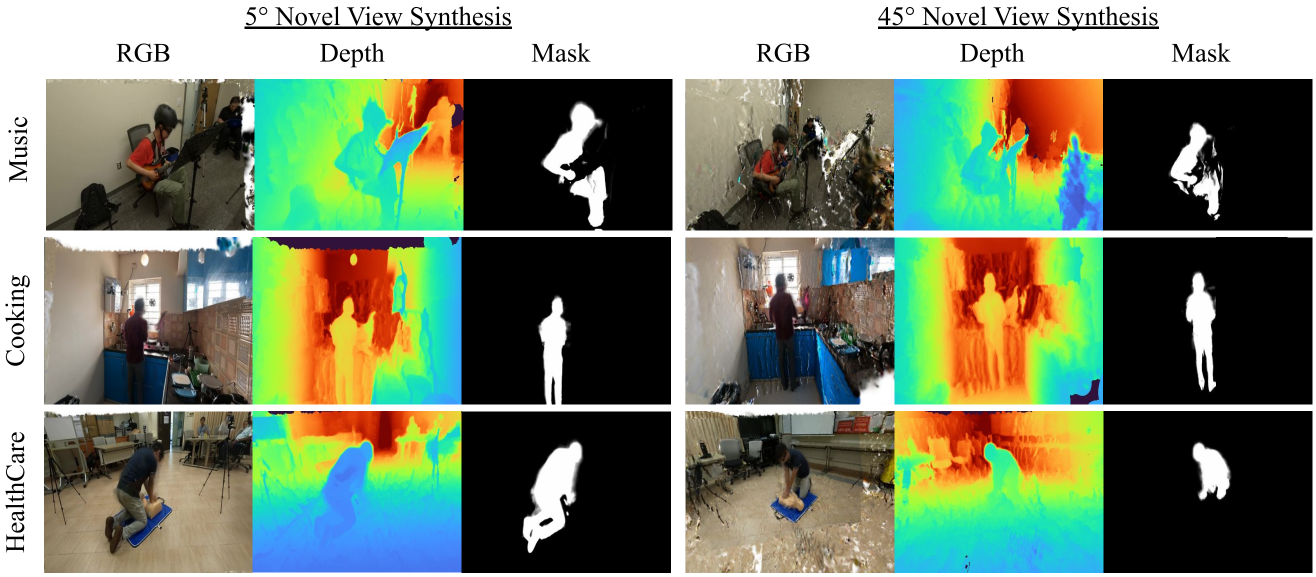

Figure 9. **Novel view synthesis results from more video sequences.** In each row, we visualize the rasterized RGB image, depth map, and foreground mask from our method for various diverse scene including music (top), cooking (middle), and healthcare (bottom). We include results for 5° (left) and 45° (right) novel view synthesis results. Notably, the rendered RGB and depth maps produce consistent reconstructions and plausible geometry.

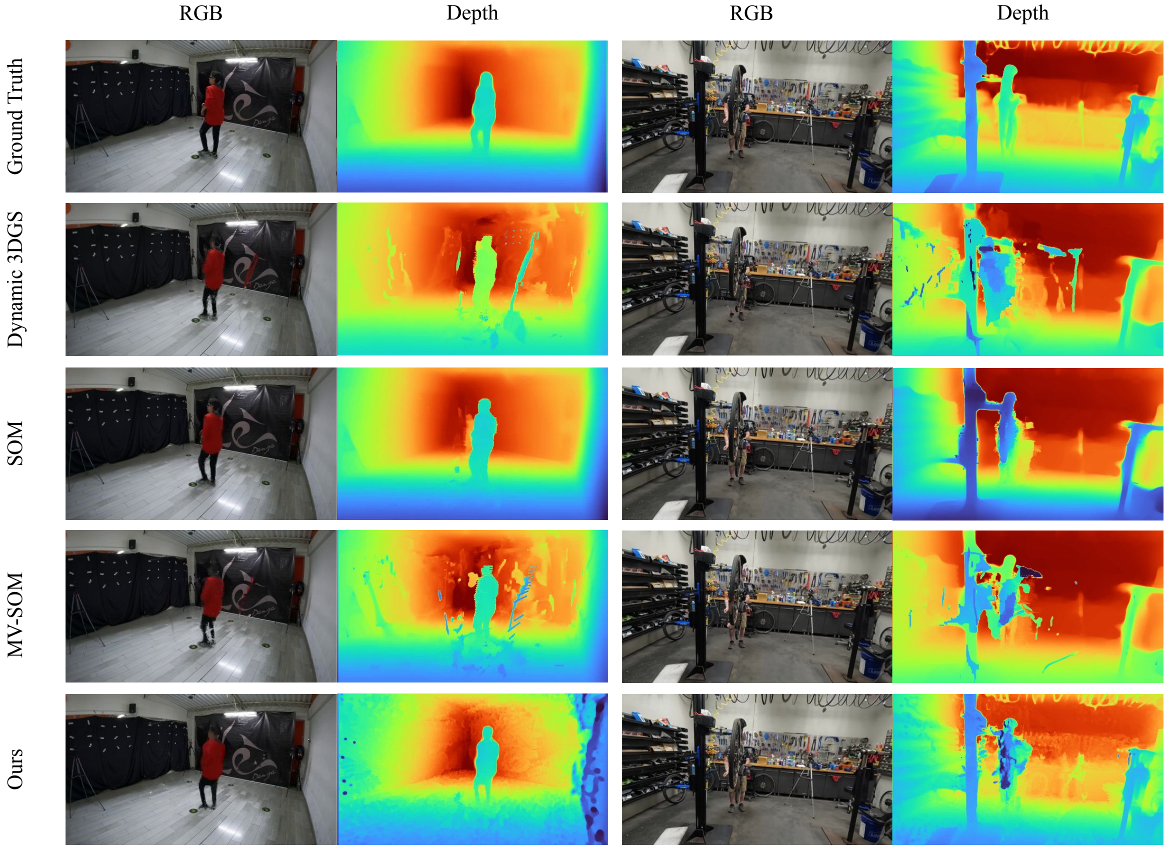

Figure 10. **Training view results.** We visualize the rasterized RGB image and depth map from each method for the dancing (left) and bike repair (right) sequences. All methods are capable of producing reasonable training views and depth maps. It is worth noticing that in each iteration of optimization, we sample a batch of frames out of the video to optimize the overall loss. As the loss is optimized as a global minimum averaged over all frames, it is possible that some artifacts remain for certain frames.

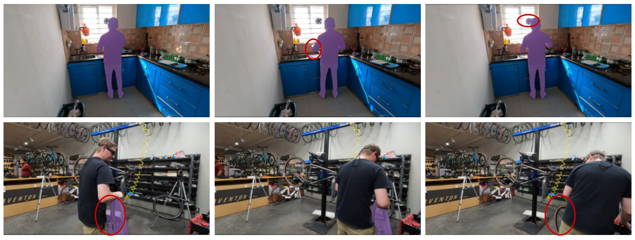

Figure 11. **Failure example of SAM-V2.** We qualitatively inspect the SAM-V2 dynamic foreground masks on the kitchen (top) and bike repair (bottom) scenes. The dynamic mask is highlighted in *purple*, and failures in dynamic mask estimation are highlighted in *red circle*. We observe that SAM-V2 can miss important body parts (e.g. the person's hands) or get confused by the background (as shown in top row). Long-term occlusion will also lead to tracking failure (as shown in bottom row). These failure cases suggest that dynamic mask tracking in complex scenes remains an open challenge.

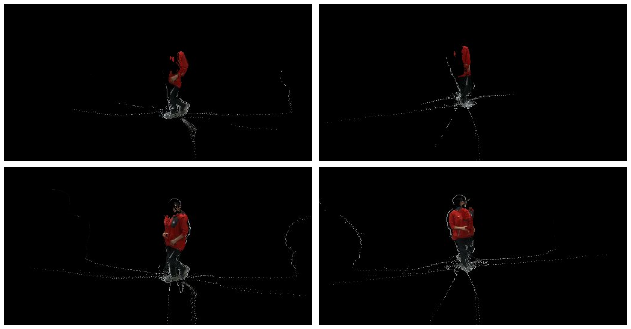

Figure 12. **Visualization of foreground projection for different checkpoints.** Here we show the projection by known cameras and ground-truth foreground masks, using the point cloud from DUSt3R (*top row*) and MonST3R (*bottom row*) for two selected cameras (each column represents one camera). Notably, although MonST3R is fine-tuned on temporal frame sequences instead of multi-view information, MonST3R benefits from the presence of dynamic foreground movers in its fine-tuning dataset and thus gives a better foreground result.