DeGauss: Dynamic-Static Decomposition with Gaussian Splatting for Distractor-free 3D Reconstruction

Rui Wang Quentin Lohmeyer Mirko Meboldt Siyu Tang

ETH Zürich

{ruiwang46, qlohmeye, meboldt}@ethz.ch siyu.tang@inf.ethz.ch

https://batfacewayne.github.io/DeGauss.io/

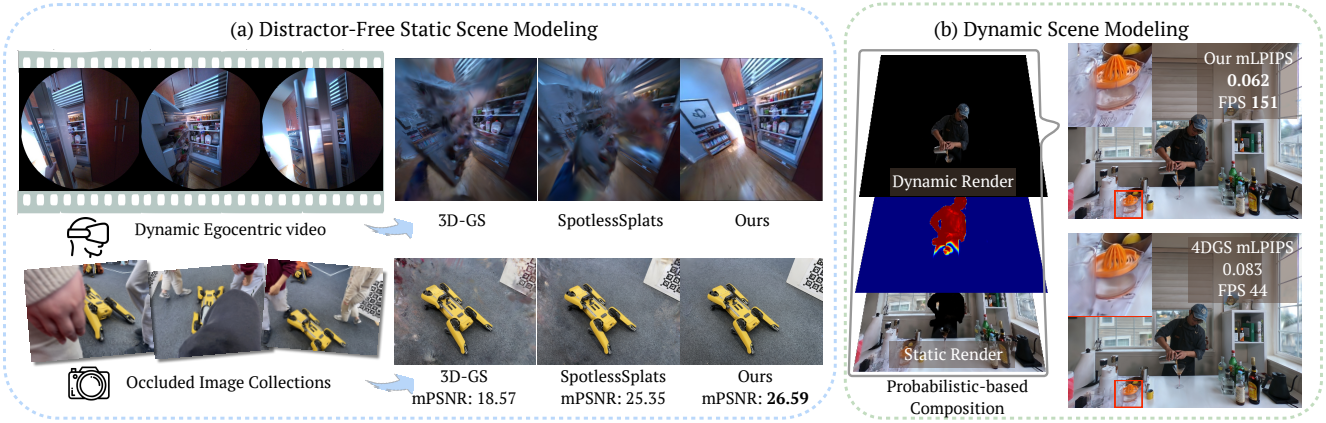

Figure 1. With self-supervised foreground-background gaussian splats modeling and accurate decomposition, **DeGauss** simultaneously enables **(a)**: SOTA distractor-free static scene reconstruction for casual captures (no dynamic modeling in the static background) and **(b)**: efficient, high-quality dynamic-static representation for dynamic scenes (no static modeling in dynamic foreground).

Abstract

Reconstructing clean, distractor-free 3D scenes from real-world captures remains a significant challenge, particularly in highly dynamic and cluttered settings such as egocentric videos. To tackle this problem, we introduce DeGauss, a simple and robust self-supervised framework for dynamic scene reconstruction based on a decoupled dynamic-static Gaussian Splatting design. DeGauss models dynamic elements with foreground Gaussians and static content with background Gaussians, using a probabilistic mask to coordinate their composition and enable independent yet complementary optimization. DeGauss generalizes robustly across a wide range of real-world scenarios, from casual image collections to long, dynamic egocentric videos, without relying on complex heuristics or extensive supervision. Experiments on benchmarks including NeRF-on-the-go, ADT, AEA, Hot3D, and EPIC-Fields demonstrate that DeGauss consistently outperforms existing methods, establishing a strong baseline for generalizable, distractor-free 3D reconstruction in highly dynamic, interaction-rich environments.

1. Introduction

Recent advances in Neural Radiance Fields (NeRF) [18] and 3D Gaussian Splatting [10] have enabled scalable 3D scene reconstruction and high-quality novel view synthesis from image collections. However, these methods perform well primarily on datasets captured under controlled conditions, where scenes remain mostly static and consistent across views. They struggle to generalize to casual captures containing dynamic elements, such as moving objects and humans. In such cases, dynamic content is often modeled as view-dependent artifacts, resulting in numerous ”floaters” in the reconstructed scene.

This limitation is further amplified in egocentric videos, a rapidly growing data source that introduces unique challenges for 3D scene reconstruction[7, 16, 29, 32, 41]. Egocentric videos, typically recorded with head-mounted, forward-facing cameras, are characterized by rapid, embodied motion. Besides substantial camera movement and motion blur, these videos frequently capture dynamic objects that the camera wearer interacts with, as well as the wearer's own body. These factors introduce significant challenges for

standard scene reconstruction methods.

The key question we aim to address in this work is how to reconstruct clean, distractor-free 3D scenes from real-world, in-the-wild videos. We focus on developing a robust and generalizable framework capable of handling a wide range of everyday capture scenarios, from casual, uncontrolled image collections to long-duration, highly dynamic egocentric recordings. By explicitly tackling the presence of dynamic elements, we aim to push 3D scene reconstruction beyond static environments toward realistic, interaction-rich settings.

To model dynamics in 3D reconstruction, recent methods such as NeRF-on-the-go, WildGaussians, and SpotlessSplats [12, 22, 24] propose to suppress transient regions during training, achieving state-of-the-art distractor-free scene reconstruction on casual image collections. These approaches leverage reconstruction loss residuals and semantic features [19, 30] to identify and mask dynamic content, as transient regions often exhibit higher reconstruction errors. However, these methods typically rely on careful initialization and stable optimization, which limits their ability to handle the complex dynamics of egocentric videos, where continuous human-scene interactions, severe motion blur, and rapid illumination changes make static-dynamic separation particularly challenging.

Meanwhile, several self-supervised NeRF-based methods aim to jointly model dynamic and static elements through explicit dynamic branches and masking strategies [17, 31, 38]. While these methods improve generalization across diverse inputs, they suffer from long training times and struggle to balance dynamic and static representations. For 3D scenes captured with temporally sparse image inputs, the dynamic branch may fail to fully segment dynamic elements, leaving floaters in the static reconstruction [23]. In contrast, for highly dynamic egocentric videos, the dynamic branch often over-segments dynamic regions, dominating the reconstruction and leaving the static scene under-represented [20].

In this work, we propose **DeGauss**: Dynamic-Static Decomposition with Gaussian Splatting for Distractor-free 3D Reconstruction. It is a simple and robust self-supervised framework that leverages dynamic-static Gaussian Splatting to effectively model and separate dynamic elements from input scenes. **DeGauss** generalizes across a wide range of scenarios, from casual image collections such as the NeRF-on-the-go dataset [22] to highly dynamic egocentric video sequences like ADT [20], AEA [16], Hot3D [20], and EPIC-Fields [32], consistently delivering superior performance without complex heuristics or elaborate designs.

Our key insight is to leverage the complementary strengths of dynamic and static Gaussians for coordinated optimization for dynamic scene reconstruction. Specifically, dynamic Gaussian methods [36, 39] learn deformation fields for temporal modeling but tend to overfit to training views and generalize poorly to novel viewpoints [6, 28]. In contrast, static Gaussians, while limited in handling motion, offer more stable representations across views, modeling dynamic elements as view-dependent artifacts (e.g., floaters). To combine their advantages, we propose a decoupled foreground-background Gaussian representation, where dynamic elements are modeled with foreground Gaussians and static content with background Gaussians. A probabilistic mask, rasterized from the foreground Gaussians, controls the composition of the two branches and enables coordinated yet independent optimization. During training, ambiguous regions are updated jointly, while floaters in the static branch are progressively suppressed through partial opacity resets and pruning. To further improve robustness under varying illumination, we introduce a brightness control mask to enhance non-Lambertian effects modeling capability of the background branch during training and mitigate dynamic-static ambiguities in those regions. Beyond producing clean, distractor-free 3D reconstructions, our formulation offers an efficient, hybrid representation of dynamic scenes through this decoupled dynamic-static design. We show that our method achieves superior results compared to baseline dynamic scene modeling approaches, with notable advantages across diverse datasets [13, 21]. In summary, our contributions are:

- We propose **DeGauss**, a decoupled foreground-background design which leverages dynamic-static Gaussian splatting for robust and generalizable dynamic-static decomposition.
- Our proposed method achieves state of the art distractor-free reconstruction results for both highly challenging egocentric videos and image collections.
- We demonstrate that **DeGauss** enables high-quality and efficient dynamic scene modeling through the decoupled dynamic-static representation.

2. Related Work

Distractor-Free Neural Reconstruction based on loss residual of input images and renders during reconstruction was investigated in [4, 23]. In [7], it is additionally combined with open-world 3D segmentation task with Segment Anything masks [11]. NeRF-on-the-Go [22] leverages DINOV2 features [19], color residuals, and an MLP predictor for dynamic elements mask. This approach was later extended to gaussian splatting [10] in WildGaussians [12]. SpotlessSplats [24] utilizes clustered diffusion-based features [30] and SOTA distractor-free scene modeling for image collections. However, these methods are sensitive to initialization and fail to leverage semantic information when within-class dynamic-static ambiguities or scene deformations arise, which limits their generalizability in more challenging settings.

Self-Supervised Scene Decomposition for neural fields was first introduced in Nerf-W [17], which decomposes and models the whole scene with dynamic and static neural fields.

This approach is further generalized to egocentric videos in NeuralDiff [31], decomposing the entire scene into dynamic, static, and actor branches. D^{2}NeRF [38] enhances decomposition results for small scenes and short clips by incorporating assignment regularization and a shadow field. However, in general, these methods face balancing issues between static-dynamic reconstruction and do not generalize well to long video inputs.

Dynamic Gaussian Splatting modeling via explicit trajectory modeling to track gaussian dynamic was investigated in [8, 15]. Deformable-GS [39] employs a deformation network to encode Gaussian deformations. 4DGS [36] leverages a Hex-plane[3] encoder and MLP-based decoders to model time-dependent Gaussian attribute parameters. However, these methods struggle to predict different deformations for gaussians with proximity, leading to over-smoothed dynamic motion. A Recent method [37] tackles this with dynamic-static separation by pre-computing static-dynamic decomposition masks based on video pixel intensity variation. However, this method only works for fixed-view camera inputs with simple motion.

Concurrent work: Recent methods [14, 34] fit separate per-camera-space gaussians for every training view to model and segment out dynamic elements with self-supervised modeling for image collections. However, the lack of shared distractors modeling across images makes it sensitive to initialization and hard to generalize. With foreground dynamic gaussians, our method achieves SOTA distractor-free scene reconstruction results for both challenging egocentric videos [2, 16, 20, 32] and casual image collections [22].

3. Method

3.1. 3D Gaussian Splatting

3D Gaussian Splatting [10] provides an explicit representation of a 3D scene using Gaussian primitives. Each primitive is defined by a mean vector $\mathbf{x} \in \mathbb{R}^3$ and a covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{3\times3}$, where

$$\mathcal{G}(x) = \exp\Big(-\frac{1}{2}\,(x-\mathbf{x})^T\,\Sigma^{-1}\,(x-\mathbf{x})\Big), \quad (1)$$
$$\text{s.t.} \quad \Sigma = \mathbf{R}\text{diag}(\mathbf{s})\text{diag}(\mathbf{s})^T\mathbf{R}^T$$

with $\mathbf{R}$ being the are rotation matrix that could be represented by quaternion $\mathbf{r}$ and $\mathbf{s}$ being the scale vector.

To render these Gaussians onto the image plane, we use differentiable splatting [40], which applies a projection transformation $\mathcal{P}(\mathcal{G})$. The final color $\mathbf{C}$ at each pixel is then computed by blending the contribution of all Gaussians, sorted by their depth:

$$\mathbf{C} = \sum_{i=1}^{N} \mathbf{c}_i\,\sigma_i\,\mathcal{P}_i(\mathcal{G}_i)\prod_{j=1}^{i-1}(1-\sigma_j\,\mathcal{P}_j(\mathcal{G}_j)). \quad (2)$$

Here, $\mathbf{c}_i \in \mathbb{R}^k$ are spherical harmonic (SH) coefficients (for an SH basis of degree k), and $\sigma_i \in \mathbb{R}$ denotes the opacity of the ith Gaussian.

3.2. Foreground deformable gaussian

We extend the set of foreground Gaussians $\mathcal{G}_f$ to embed customized mask elements for dynamic scene decomposition, and the complete features could be defined as $\mathcal{G}_f = \{\mathbf{x}, \mathbf{s}, \mathbf{r}, \sigma, \mathbf{c}, m_f, m_b, b\}$. Here, the standard attributes $\{m_f, m_b, b\}$ are the foreground probabilistic attributes, background probabilistic attributes, and brightness control attributes, respectively.

The deformed foreground Gaussians are obtained as: $\mathcal{G}'_f = \Delta\mathcal{G}_f + \mathcal{G}_f$. The spatial-temporal module comprises an encoder $\mathcal{H}$ and a decoder $\mathcal{D}$. The encoder, based on Hexplane [3], extracts spatio-temporal features based on reference time t with $\mathbf{f_d} = \mathcal{H}(\mathcal{G}_f, t)$, and the multi-head decoder $\mathcal{D}$ predicts the deformation of each gaussian features with $\Delta\mathcal{G}_f = \mathcal{D}(\mathbf{f_d})$. Separate MLPs are employed to predict the deformation of each gaussian attribute. The decoder $\mathcal{D}$ comprises: $\mathcal{D} = \{\phi_\mathbf{x}, \phi_\mathbf{r}, \phi_\mathbf{s}\,\phi_\sigma, \phi_\mathbf{c}, \phi_{m_f}, \phi_{m_b}, \phi_b\}$. With this, the deformed feature could be addressed as:

$$\begin{aligned}(\mathbf{x}', \mathbf{r}', \mathbf{s}', \sigma', \mathbf{c}', m'_f, m'_b, b') = \Big(&\mathbf{x} + \phi_\mathbf{x}(\mathbf{f}_d),\\ &\mathbf{r} + \phi_\mathbf{r}(\mathbf{f}_d),\ \mathbf{s} + \phi_\mathbf{s}(\mathbf{f}_d), \sigma + \phi_\sigma(\mathbf{f}_d),\ \mathbf{c} + \phi_\mathbf{c}(\mathbf{f}_d),\\ &m_f + \phi_{m_f}(\mathbf{f}_d), m_b + \phi_{m_b}(\mathbf{f}_d),\ b + \phi_b(\mathbf{f}_d)\Big). \quad (3)\end{aligned}$$

3.3. Probabilistic Composition Mask Rasterization

Given the predicted mask elements $\{m'_f, m'_b\}$ and the deformed attributes $\{\mathbf{x}', \mathbf{r}', \mathbf{s}', \sigma'\}$, we can directly use differentiable rendering to compute the raw foreground probability $\mathbf{M}_f$ and $\mathbf{M}_b$ via based on Eq. (2):

$$\mathbf{M}_f = \sum_{i=1}^{N} m'_{f_i}\sigma'_i\mathcal{P}_i(\mathcal{G}'_{f_i})\prod_{j=1}^{i-1}(1-\sigma'_j\mathcal{P}_j(\mathcal{G}'_{f_j})), \quad (4)$$

$$\mathbf{M}_b = \sum_{i=1}^{N} m'_{bi}\sigma'_i\mathcal{P}_i(\mathcal{G}'_{f_i})\prod_{j=1}^{i-1}(1-\sigma'_j\mathcal{P}_j(\mathcal{G}'_{f_j})); \quad (5)$$

With $\mathbf{P} = M_f + M_b + \epsilon$, where ϵ is a small constant to avoid division by zero, the foreground and background probabilistic masks could be given by:

$$\mathbf{P}_f = (1/\mathbf{P}) * \mathbf{M}_f, P_b = (1/\mathbf{P}) * \mathbf{M}_b. \quad (6)$$

This probabilistic formulation naturally discourages mid-range values (near 0.5), pushing the prediction toward 0 or 1 and yielding a clean dynamic-static decomposition.

3.4. Background Brightness Control

Casual captures often exhibit significant illumination variations, creating ambiguities in geometry and view-dependent

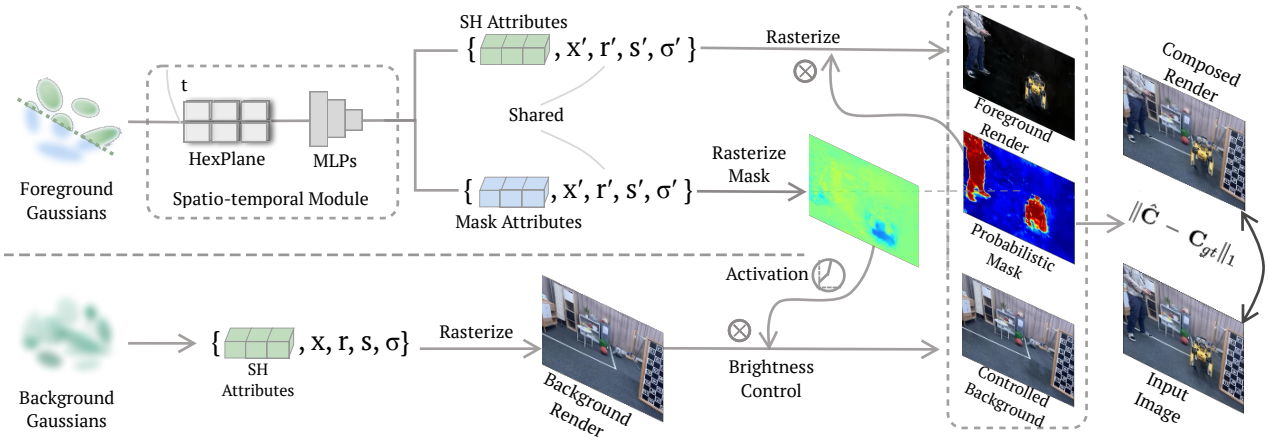

Figure 2. Our method simultaneously reconstructs the 3D scene and learns an unsupervised decomposition into decoupled static background and dynamic foreground branches, where the update is loosely controlled by the mask rasterization process. This decoupled formulation guarantee flexible yet accurate scene decomposition result.

appearance modeling. While non-Lambertian effects can be progressively captured through the spherical harmonic (SH) coefficients of Gaussian Splatting, the high expressiveness of dynamic Gaussians in the foreground branch often leads to over-segmentation of dynamic elements in regions with large illumination variations. To address this, we introduce a brightness control mask that enhances the background branch's capacity to model non-Lambertian effects. The raw brightness control mask could be obtained via rasterizing foreground gaussian with brightness control element b:

$$\mathbf{B} = \sum_{i=1}^{N} b'_i \sigma'_i \mathcal{P}_i(\mathcal{G}'_{f_i}) \prod_{j=1}^{i-1} (1 - \sigma'_j \mathcal{P}_j(\mathcal{G}'_{f_j})) \tag{7}$$

Moreover, to prevent modeling dark dynamic objects with the brightness control mask and enable the modeling of over-brightness, we further introduce a piece-wise linear activation function for the brightness control mask, and the transformed brightness control mask $\hat{\mathbf{B}}$ is given by:

$$\hat{\mathbf{B}} = \begin{cases} \mathbf{B} + 0.5, & 0 \leq \mathbf{B} \leq 0.75, \\ k\,(\mathbf{B} - 0.75) + 1.25, & 0.75 < \mathbf{B} \leq 1, \end{cases} \tag{8}$$

, where k is an over-brightness modeling coefficient, we choose $k = 35$ in practice. The raw background render $\mathbf{C}_b$ is rasterized by background gaussian $\mathcal{G}_b$ with equation (2). The controlled background is then given with $\hat{\mathbf{C}}_b = \hat{\mathbf{B}} * \mathbf{C}_b$.

3.5. Dynamic Foreground Representation

With deformed gaussian $\mathcal{G}'_f$, the raw foreground render could be given by:

$$\mathbf{C}_f(u, v) = \sum_{i=1}^{N} \mathbf{c}'_{f_i} \sigma'_i \mathcal{P}_i(\mathcal{G}'_{f_i}) \prod_{j=1}^{i-1} (1 - \sigma'_j \mathcal{P}_j(\mathcal{G}'_{f_j})), \tag{9}$$

And the final foreground render $\hat{\mathbf{C}}_f$ is obtained by applying the foreground probabilistic mask to the raw foreground render, $\mathbf{C}_f = \mathbf{P}_f \mathbf{C}_f$. This formulation comes with several advantages. On one hand, we could efficiently allow the presence of utility gaussians that are important for probabilistic composition mask $\mathbf{P}_f$, $\mathbf{P}_b$ and brightness control mask $\hat{\mathbf{B}}$ but do not contribute to foreground render. Moreover, such a design could efficiently reduce the presence of unregulated gaussian movement for dynamic scene modeling with this added degree of freedom and avoid artifacts caused by unconstrained gaussian movement.

3.6. Unsupervised scene decomposition

With the established composition mask $\mathbf{P}_f$, $\mathbf{P}_b$ and brightness control mask $\hat{\mathbf{B}}$, the composed render $\hat{C}$ is defined as:

$$\hat{\mathbf{C}} = \mathbf{P}_f * \mathbf{C}_f + \mathbf{P}_b * \hat{\mathbf{B}} * \mathbf{C}_b \tag{10}$$

Compositional rendering with color mixing in NeRF-based methods [13, 17, 31] sorts and integrates static and dynamic density and radiance along each ray(compose during rendering), leading to early ray termination during training on local minima and reconstructs static scene without fine details[23, 31]. In our decoupled design, the dynamic/static gaussians rasterize the foreground/background renders $\mathbf{C}_f$ and $\mathbf{C}_b$ independently and compose (after rending) with the probabilistic mask $\mathbf{P}_f$. This design enables full gradient flow and allow gradually formulated composition mask during training, as shown in Fig. 3. The wrongly modeled elements are gradually pruned during gaussian density control process, yielding accurate, clean yet flexible dynamic-static separation results that is much more robust to local minima.

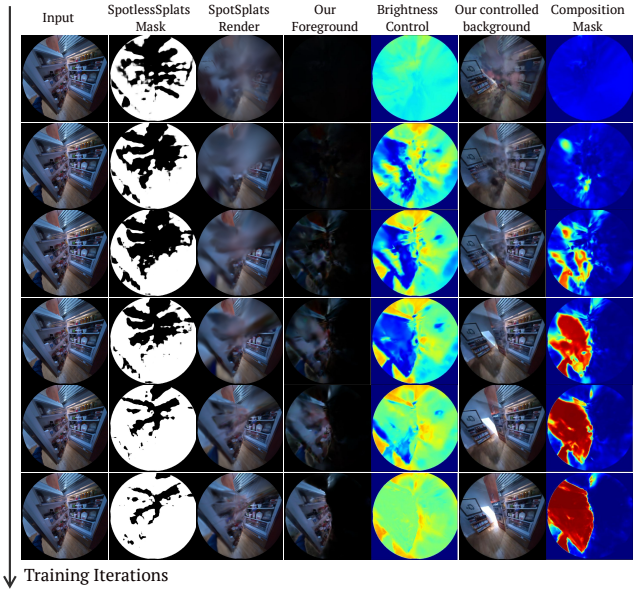

Figure 3. Compared to SpotlessSplats [24], which is constrained by initialization and overfit to floaters. Our method offers significantly greater robustness in handling local minimas. The brightness control mask effectively resolves the static-dynamic ambiguity due to strong illumination variations and promote the decomposition process during training.

3.7. Loss function

Loss function design is crucial to balance the expressiveness of foreground and background branches while reconstructing the scene with high-quality details. As the loss gradient magnitude controls the densification process of gaussians [10], we design the loss function $\mathcal{L}$, which comprises two parts $\mathcal{L}_{\text{main}}$ and $\mathcal{L}_{\text{uti}}$, separating loss gradients for adaptive densification process, to effectively suppress the spawning of floaters and controlling the number of utility gaussians in foreground branch:

$$\begin{aligned}\mathcal{L} = &\underbrace{\mathcal{L}_1 + \mathcal{L}_{\text{diversity}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_f + \mathcal{L}_b}_{\mathcal{L}_{\text{main}}} \\ &+ \underbrace{\mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{entropy}} + \mathcal{L}_{\text{brightness}} + \mathcal{L}_{\text{scale}}}_{\mathcal{L}_{\text{uti}}}. \end{aligned} \tag{11}$$

While both main loss $\mathcal{L}_{\text{main}}$ and utility loss $\mathcal{L}_{\text{uti}}$ are used for optimizable parameters' update, only the gradient magnitude of $\mathcal{L}_{\text{main}}$ are used to densify foreground and background gaussians. We refer readers to Appendix **A.** for a detailed definition of each loss term.

3.8. Partial Opacity Reset

In methods as [12, 24], directly employing periodic opacity reset [10] is not feasible, as it induces instability during training. Owing to the added stability with the foreground-background formulation, we perform periodic partial opacity reset for 50% for background-foreground gaussians. This guarantees stable training, effectively controls gaussian density, and handles local minima.

4. Experiments

4.1. Implementation Details

Initialization The scene boundary and the background gaussians are initialized from point clouds generated using COLMAP [25, 25] or sensor perception [9] for Aria Sequences[2, 16, 20]. The foreground Gaussians are initialized from randomly generated points distributed within this scene boundary.
Coarse Training Stage: During the coarse training stage, we disable the deformation module in the foreground branch and train both the foreground and background models for 1,000 iterations with short video clips and image collections or for iterations equal to sequence length for long captures.
Fine Training Stage: In the fine training stage, we jointly optimize the foreground and background branches end-to-end. For short video clips and image collections of less than 500 images, training iterations are set to 20k. For input long video clips of a few thousand frames, the training iteration is set to 120k.

4.2. Datasets

Egocentric video sequences are with intensive camera wearer activities and varying illumination conditions, which pose challenges to scene modeling methods. We take one sequence from ADT [20], AEA [16], Hot3D [2], and Epic-Field [32] dataset, respectively, ranging from 2800-5000 frames, to evaluate our method against baseline methods [10, 24, 31] in diverse scenarios. For each sequence, every 1 out of 5 frames is held out during training.
NerF On-the-Go Dataset [22] comprises several hundred input images featuring moving distractors alongside a smaller set of clean images reserved for testing. We train our methods on the noisy occluded images and assess the quality of novel view synthesis on the clean hold-out set.
Neu3D Dataset [13] was captured using 15 to 20 static cameras recording relatively simple activities over 300 frames. Camera view 0 is the testing set, with the remaining views used for training.
HyperNeRF Dataset [21] features real-world activities captured with smooth trajectories. However, as noted in [8], the camera poses are considerably inaccurate. Therefore, we focus primarily on qualitative visualizations for this dataset.

4.3. Results

To assess the performance of our method for the distractor-free scene reconstruction task in the presence of noisy inputs, we conduct evaluations on both egocentric videos and image collections. For egocentric video data [2, 16, 20, 32]—which

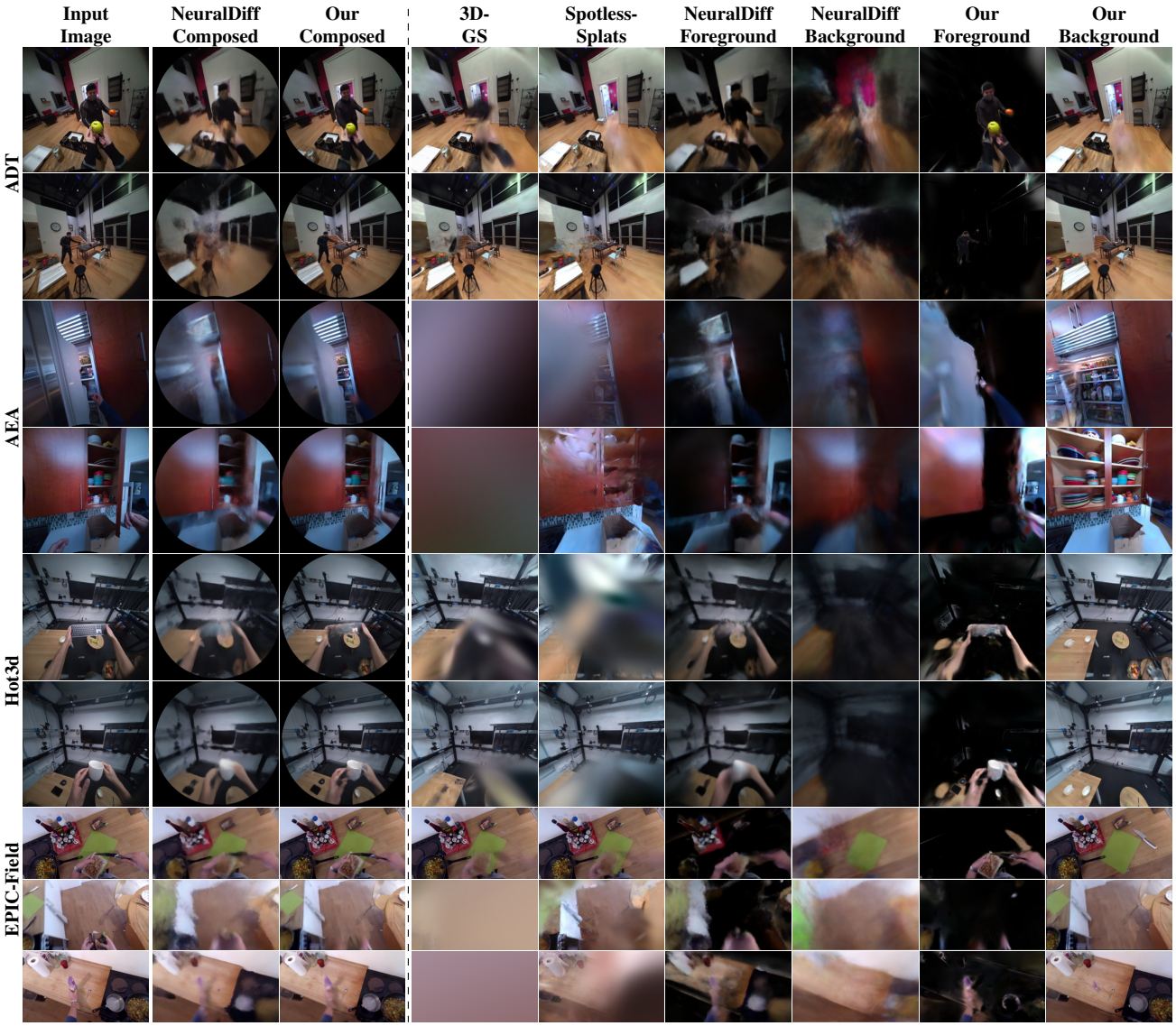

Figure 4. Qualitative comparison of baseline methods [10, 24, 31] for distractor-free scene reconstruction on the Aria and EPIC-Field sequences. Left of the dashed line: composed render comparisons; right: static reconstruction comparison(without camera masks).

Table 1. Distractor free scene reconstruction on NeRF On-the-go Dataset[22].The best , second best , and third best are highlighted. ‡: ±0.005 SSIM and LPIPS due to rounding uncertainty of originally reported result. Our method shows generally superior performance over state-of-the-art methods.

	Mountain			Fountain			Corner			Patio			Spot			Patio-High			Mean		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
RobustNeRF [23]	17.54	0.496	0.383	15.65	0.318	0.576	23.04	0.764	0.244	20.39	0.718	0.251	20.65	0.625	0.391	20.54	0.578	0.366	19.64	0.583	0.369
NeRF On-the-go [22]	20.15	0.644	0.259	20.11	0.609	0.314	24.22	0.806	0.190	20.78	0.754	0.219	23.33	0.787	0.189	21.41	0.718	0.235	21.67	0.720	0.234
3DGS [10]	19.40	0.638	0.213	19.96	0.659	0.185	20.90	0.713	0.241	17.48	0.704	0.199	20.77	0.693	0.316	17.29	0.604	0.363	19.30	0.668	0.253
WildGaussian [12]	20.43	0.653	0.255	20.81	0.662	0.215	24.16	0.822	0.139	21.44	0.800	0.138	23.82	0.816	0.138	22.23	0.725	0.206	22.16	0.746	0.182
DeSplat‡ [34]	19.59	0.715	0.175	20.27	0.685	0.175	26.05	0.885	0.095	20.89	0.815	0.115	26.07	0.905	0.095	22.59	0.845	0.125	22.58	0.813	0.130
Spotlesssplats [24]	21.64	0.725	0.195	22.38	0.768	0.166	25.77	0.877	0.117	22.40	0.833	0.108	25.35	0.866	0.127	22.98	0.808	0.155	23.42	0.813	0.145
Ours	22.31	0.746	0.163	22.40	0.764	0.139	25.94	0.869	0.078	22.88	0.850	0.087	26.59	0.886	0.089	23.35	0.799	0.124	23.91	0.819	0.113

lack clean view references—we present qualitative comparisons with baseline methods [10, 24, 31] in Fig. 4. Compared to baseline methods [10, 24, 31], our method models high-quality distractor-free static background with accurate foreground separation. We additionally report video comparisons in our supplementary materials. For image collections

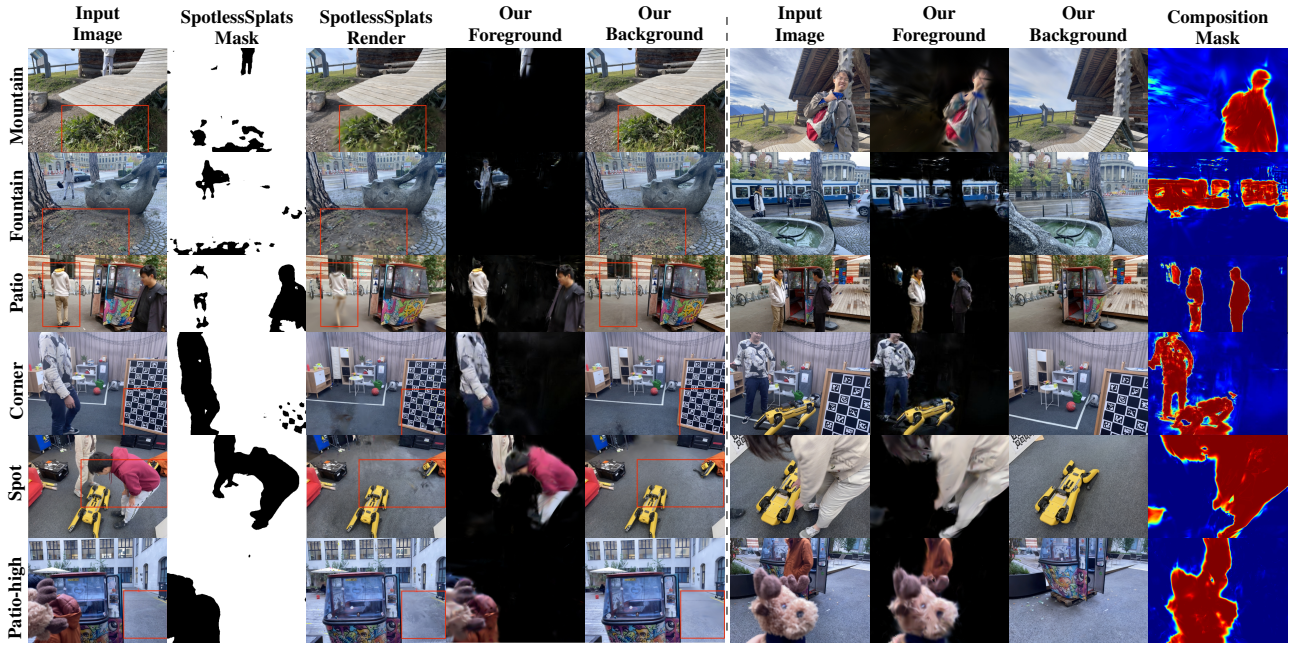

Figure 5. Occlusion handling on the NeRF-on-the-Go dataset [22]. Compared to SpotlessSplats [24], our method better preserves fine details in the training views (please consider zooming in for a clearer view) and reduces misclassification of dynamic regions, leading to consistently better LPIPS on testing images. Right of dashed line: more results.

Table 2. Comparison dynamic modeling on Neu3D Dataset [13]. The best , second best , and third best are highlighted. Noticeably, our method shows a consistently better LPIPS score compared to baseline methods.

	Cut Beef			Cook Spinach			Sear Steak			Flame Steak			Flame Salmon			Coffee Martini			Mean		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRFPlayer[27]	31.83	0.928	0.119	32.06	0.930	0.116	32.31	0.940	0.111	27.36	0.867	0.215	26.14	0.849	0.233	32.05	0.938	0.111	30.29	0.909	0.151
HyperReel [1]	32.25	0.936	0.086	31.77	0.932	0.090	31.88	0.942	0.080	31.48	0.939	0.083	28.26	0.941	0.136	28.65	0.897	0.129	30.72	0.931	0.101
HexPlane [3]	30.83	0.927	0.115	31.05	0.928	0.114	30.00	0.939	0.105	30.42	0.939	0.104	29.23	0.905	0.088	28.45	0.891	0.149	30.00	0.922	0.113
KPlanes [5]	31.82	0.966	0.114	32.60	0.966	0.114	32.52	0.974	0.104	32.39	0.970	0.102	30.44	0.953	0.132	29.99	0.953	0.134	31.63	0.964	0.117
MixVoxels [33]	31.30	0.965	0.111	31.65	0.965	0.113	31.43	0.971	0.103	31.21	0.970	0.108	29.92	0.945	0.163	29.36	0.946	0.147	30.81	0.960	0.124
SWinGS [26]	31.84	0.945	0.099	31.96	0.946	0.094	32.21	0.950	0.092	32.18	0.953	0.087	29.25	0.925	0.100	29.25	0.925	0.100	31.12	0.941	0.095
4DGS [36]	32.66	0.946	0.053	32.46	0.949	0.052	32.49	0.957	0.041	32.75	0.954	0.040	29.00	0.912	0.081	27.34	0.905	0.083	31.12	0.937	0.058
Ours	32.56	0.957	0.042	32.61	0.950	0.041	33.20	0.956	0.035	32.75	0.955	0.034	29.23	0.916	0.068	28.80	0.916	0.062	31.52	0.942	0.047

Figure 6. Our method robustly handles various challenges, preserving clean and high quality static background.

dataset Nerf-on-the-go[22] with clean reference test views, we report detailed per-scene metrics including peak signal-to-noise ratio (PSNR), perceptual quality (LPIPS) [42], and structural similarity index (SSIM) [35] against baseline methods[10, 12, 22–24, 34] on the hold-out test set in Tab. 1. Our methods generalize to image collections and achieve state-of-the-art results. Notably, our method consistently achieves significantly better LPIPS scores over the previous SOTA method SpotlessSplats [24]. We show our method robustly handles occlusion and reconstructs fine static details compared to SpotlessSplats [24]in Fig. 5. Additionally, our methods could naturally handle various input challenges, such as camera motion blur and lens flare, as shown in Fig. 6.

Moreover, we compare our method's composed render quality with various baseline methods [1, 3, 5, 26, 27, 33, 36] in Tab. 2, where our methods achieve consistently better LPIPS scores. We qualitatively show the dynamic reconstruction comparison and the rendering FPS of [36] and our method in Fig. 7(on RTX4090), where our methods show better reconstructed fine details and better test-time rendering efficiency. Moreover, we compare our method with 4DGS [38] on HyperNeRF [21] dataset in Fig. 8, showing that our method effectively regularizes gaussian movements

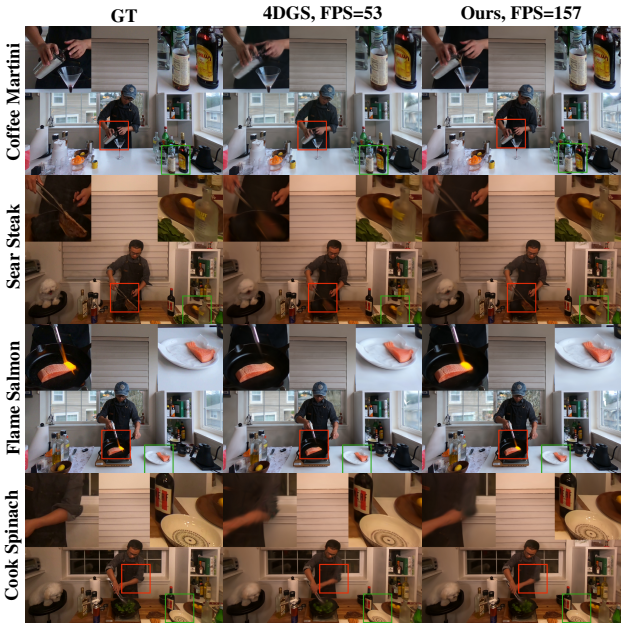

Figure 7. Qualitative comparison with 4DGS [36] on the Neu3D [13] dataset. FPS is tested with fix-view rendering as [36].

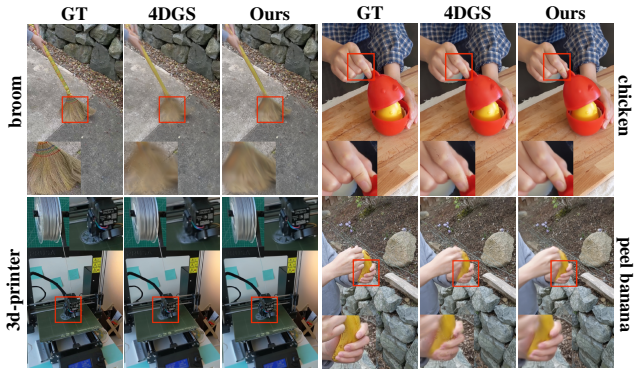

Figure 8. Qualitative comparison of our method with 4DGS [36] on HyperNerf Vrig dataset [21]. Please consider zooming in for a clearer view.

with probabilistic controlled dynamic foreground representation and reduces unregularized moving artifacts.

5. Ablation study

Brightness Control(BC) is introduced to enhance the background branch's capacity to model non-Lambertian effects and mitigate dynamic-static ambiguities caused by varying illuminations, as shown in Fig. 9. w/o BC leads to downgraded performance in Tab. 3.

Partial Opacity Reset(POR) controls the gaussian density, facilitates floaters pruning, and mitigates local minima assignment. We show in Fig. 9 and Tab. 3 that this design leads to cleaner separation.

Background Mask Element(m_b') is introduced to promote

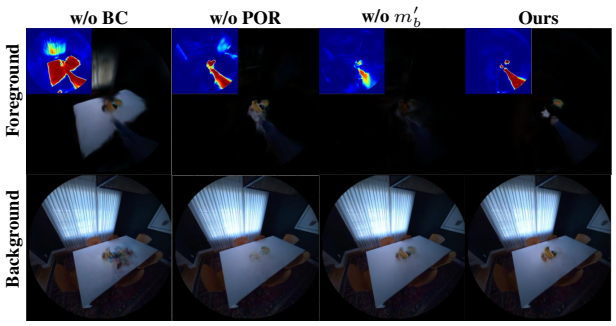

Figure 9. Ablation Study on AEA [16] dataset.

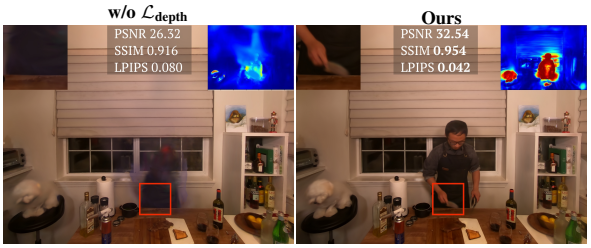

Figure 10. Ablation Study on Neu3D dataet [13] cut beef scene.

Table 3. Ablation study on Nerf-on-the-go dataset[22]

Sequence from	PSNR↑	SSIM↑	LPIPS↓
w.o BC	23.54	0.814	0.118
w.o POR	23.56	0.814	0.117
w.o $\mathcal{L}_{depth}$	23.68	0.816	0.113
w.o. m_b'	23.83	0.817	0.115
Ours	**23.91**	**0.819**	**0.113**

cleaner separation and discourage mid-range probabilities. Though the improvements are not significant for image collections with good initializations, it leads to better dynamic-static modeling and separation results as shown in Fig. 9.

Loss $\mathcal{L}_{depth}$ is introduced to promote reconstruction with smooth background geometry and loosely regularize foreground and background depth prediction. As shown in Fig. 10, this component efficiently prevents unconstrained gaussians from occluding test-time render for sparse, fixed camera input. $\mathcal{L}_{depth}$ also leads to better rendering quality as shown in Tab. 3.

6. Conclusion

This paper proposes DeGauss to robust decompose dynamic-static elements in the scene with gaussian splatting. With decoupled dynamic-static gaussian branches controlled by mask attributes rasterized by foreground gaussians, our method achieves flexible yet accurate dynamic-static decomposition that widely generalizes to various scenarios, leading to clean distractor-free static scene modeling and high-quality and efficient dynamic scene modeling.

Acknowledgements. The authors would like to sincerely thank Siwei Zhang for the rigorous proof-reading and Hui Zhang for the discussion.

References

[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 3, 5

[3] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 3, 7

[4] Jiahao Chen, Yipeng Qin, Lingjie Liu, Jiangbo Lu, and Guanbin Li. Nerf-hugs: Improved neural radiance fields in non-static scenes using heuristics-guided segmentation. *CVPR*, 2024. 2

[5] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 7

[6] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. 2

[7] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *European Conference on Computer Vision*, pages 382–400. Springer, 2025. 1, 2

[8] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. *arXiv preprint arXiv:2312.14937*, 2023. 3, 5

[9] Selcuk Karakas, Pierre Moulon, Wenqi Zhang, Nan Yang, Julian Straub, Lingni Ma, Zhaoyang Lv, Elizabeth Argall, Georges Berenger, Tanner Schmidt, Kiran Somasundaram, Vijay Baiyya, Philippe Bouttefroy, Geof Sawaya, Yang Lou, Eric Huang, Tianwei Shen, David Caruso, Bilal Souti, Chris Sweeney, Jeff Meissner, Edward Miller, and Richard Newcombe. Aria data tools. https://github.com/facebookresearch/aria_data_tools, 2022. 5

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1, 2, 3, 5, 6, 7

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[12] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *NeurIPS*, 2024. 2, 5, 6, 7

[13] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 2, 4, 5, 7, 8

[14] Jingyu Lin, Jiaqi Gu, Lubin Fan, Bojian Wu, Yujing Lou, Renjie Chen, Ligang Liu, and Jieping Ye. Hybridgs: Decoupling transients and statics with 2d and 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 788–797, 2025. 3

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

[16] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 1, 2, 3, 5, 8

[17] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 2, 4

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[20] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 2, 3, 5

[21] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2, 5, 7, 8

[22] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8931–8940, 2024. 2, 3, 5, 6, 7, 8

[23] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20626–20636, 2023. 2, 4, 6

[24] Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David J Fleet, and Andrea Tagliasacchi. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *arXiv preprint arXiv:2406.20055*, 2024. 2, 5, 6, 7

[25] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[26] Richard Shaw, Michal Nazarczuk, Jifei Song, Arthur Moreau, Sibi Catley-Chandar, Helisa Dhamo, and Eduardo Perez-Pellitero. Swings: Sliding windows for dynamic 3d gaussian splatting. *arXiv preprint arXiv:2312.13308*, 2023. 7

[27] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 7

[28] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[29] Jiankai Sun, Jianing Qiu, Chuanyang Zheng, John Tucker, Javier Yu, and Mac Schwager. Aria-nerf: Multimodal egocentric view synthesis. *arXiv preprint arXiv:2311.06455*, 2023. 1

[30] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2

[31] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. Neuraldiff: Segmenting 3d objects that move in egocentric videos. In *2021 International Conference on 3D Vision (3DV)*, pages 910–919. IEEE, 2021. 2, 3, 4, 5, 6

[32] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 5

[33] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19706–19716, 2023. 7

[34] Yihao Wang, Marcus Klasson, Matias Turkulainen, Shuzhe Wang, Juho Kannala, and Arno Solin. DeSplat: Decomposed Gaussian splatting for distractor-free rendering. *arXiv preprint arxiv:2411.19756*, 2024. 3, 6, 7

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[36] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2, 3, 7, 8

[37] Jiahao Wu, Rui Peng, Zhiyan Wang, Lu Xiao, Luyang Tang, Jinbo Yan, Kaiqiang Xiong, and Ronggang Wang. Swift4d: Adaptive divide-and-conquer gaussian splatting for compact and efficient reconstruction of dynamic scene. In *The Thirteenth International Conference on Learning Representations*. 3

[38] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *Advances in neural information processing systems*, 35:32653–32666, 2022. 2, 3, 7

[39] 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 arXiv:2309.13101*, 2023. 2, 3

[40] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38(6):1–14, 2019. 3

[41] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. *arXiv preprint arXiv:2406.19811*, 2024. 1

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7