CVPR
#4792

DS-Hybrid GS: auto-reconstruction and decomposition from monocular video

Anonymous CVPR submission

Paper ID 4792

Abstract

The dynamic nature of our environment, paired with the prevalence of monocular videos as a medium for capturing reality, necessitates efficient methods for reconstructing high-dimensional representations from low-dimensional video data while tackling dynamic and static elements separately for further editing. However, existing methods often rely heavily on costly and ambiguous priors, such as manually labeled masks, optical flow, and depth information, which can introduce inefficiencies and result in suboptimal reconstructions. In this work, we propose a novel reconstruction approach that not only facilitates the reconstruction of dynamic scenes from monocular videos but also effectively decouples dynamic and static components. To achieve this, we introduce an automatic segmentation pipeline that distinguishes between dynamic and static objects. Our method leverages large pre-trained models for generating high-quality masks and employs feature point registration to enhance generalization beyond traditional optical flow techniques. Additionally, we incorporate a semantic filter to further refine the segmentation results. The second stage of our process focuses on the reconstruction of dynamic scenes, relying solely on the dynamic masks obtained in the first step. This approach results in a hybrid dynamic scene representation that enables effective dynamic decomposition called ***DS-Hybrid GS****(Dynamic-Static-Hybrid Gaussian Splatting). By using only the dynamic masks as prompt inputs, our method becomes robust and applicable to a wider range of datasets. This work makes three key contributions: (1) an automatic monocular video reconstruction method that facilitates the decoupling of dynamic components, (2) an innovative dynamic element classifier based on point density matching, (3)In street scenes, it significantly improves the efficiency and reduces the cost of reconstruction, and (4) the insight that fewer constraints in the reconstruction process lead to greater robustness in monocular scene reconstruction.*

Figure 1. **Video with extreme camera motion** Previous methods, such as OmnimatteRF, address video with parallax effect well. However, when facing this kind of monocular video, preview methods often tend to treat the entire scene as part of the foreground due to the influence of optical flow. In contrast, our method, which uses only the mask as input, is not affected by the ambiguities of optical flow, enabling more accurate segmentation results.

1. Introduction

The world we inhabit can be viewed as a dynamic scene characterized by both temporal and spatial consistency. At the same time, video, particularly monocular video capturing everyday life in a causal manner, has become the primary medium for recording and representing the real world. Recovering high-dimensional representations of the real world from low-dimensional video data has long been a significant challenge and remains crucial for scaling up machine intelligence's ability to perceive and understand the world. Furthermore, the decoupling of dynamic and static components in the reconstructed scene is essential for subsequent applications and editing, playing a key role in the practical viability of reconstruction techniques.

Recent works [7, 12, 14, 15, 20, 22, 42] have demonstrated promising results in reconstructing dynamic scenes from monocular videos, with the majority treating dynamic and static scenes separately. To achieve robust outcomes, these studies often incorporate various priors and supervision, such as optical flow, depth information, and manual annotations to assist in reconstruction. However, methods relying on additional prior constraints can be costly and may encounter unavoidable ambiguities. Thus, obtaining economically efficient and robust information suitable for dynamic components has become a critical issue in this way of monocular video reconstruction. Meanwhile, some approaches [26, 39, 41] treat all scenes as dynamic for reconstruction. However, these methods are often limited by their rendering techniques and storage formats. Adding temporal information to static scenes can introduce significant redundancy, resulting in inefficient reconstruction processes unsuited for subsequent editing tasks. Therefore, optimizing these methods for cost reduction and efficiency, while ensuring their usability, remains an important challenge. In prior work on dynamic-static decomposition, methods [2] have demonstrated reasonable performance in simple cases, effectively separating dynamic and static components. Since the concept of Omnimatte [21] proposed as a general of co-effects such as shadows in dynamic scenes, related and following works [12, 16, 21, 33, 39] setting a higher standard for decomposition tasks. However, these methods struggle when applied to more complex dynamic scenarios. As illustrated in Fig. 1, significant camera motion can introduce optical flow ambiguities, which cause previous methods to fail and lack robustness. Addressing such cases, particularly more extreme dynamic situations remains an open problem.

Therefore, we first propose an efficient and rapid solution for obtaining dynamic scene information by designing a pipeline for the automatic segmentation of dynamic and static objects in videos. We leverage the priors from large pre-trained models to achieve high-quality masks and innovatively employ feature point registration to filter dynamic objects, offering greater generalization compared to similar optical flow methods. Additionally, we utilize a semantic filter to refine the results. Furthermore, addressing the demands for efficient storage and rendering while considering the characteristics of monocular videos, we introduce a hybrid dynamic scene representation that enables dynamism decomposition with only dynamic masks as additional input. This approach significantly reduces redundant information compared to previous dynamic methods. Furthermore, our research has substantiated that a reduction in the number of constraints leads to enhanced robustness in performance.

1. A fully automatic monocular video reconstruction method while enabling dynamic components decoupling.

2. Novel dynamic elements classifier base match points density.

3. In street scene reconstruction, we eliminated the reliance on bounding boxes, significantly reducing costs and enhancing reconstruction efficiency.

4. An insight that fewer constraints provide more robustness in monocular reconstruction.

2. Related work

3D reconstruction Throughout the development of computer vision, recovering spatial information from images has remained a challenging problem. Traditional methods [1, 29, 30, 32] have primarily focused on reconstructing geometric information. In recent years, however, novel view synthesis approaches have emerged, such as Neural Radiance Fields (NeRF) [24] and its subsequent extensions [3, 4, 9, 35, 48] which are capable of capturing view-dependent effects. However, vanilla NeRF requires querying the MLP for hundreds of points each ray, significantly constraining its training and rendering speed. Although some works [10, 17, 19, 25, 28, 44] have attempted to improve the training or rendering speed, these methods remain confined to the nuances of differentiable volume rendering until 3D Gaussian Splatting(3DGS) [13] proposed. 3DGS utilizes rasterization to achieve real-time rendering of high-quality results in complex scenes. While numerous subsequent works have made advancements in geometry reconstruction [11], large-scale representation [31], and anti-aliasing [45], we argue that the inherent effectiveness of Gaussian primitives, coupled with the theoretical foundation of Gaussian Mixture Models (GMM) for fitting arbitrary shape probability distributions, renders 3DGS a more robust representation for static scenes in our work.

Monocular video reconstruction While the input of reconstruction of a scene is diverse, monocular video is the most common and challenging set. With the emergence of NeRF and 3DGS, various works [8, 14, 15, 20, 22, 26, 36, 37, 40, 41, 46, 47] have attempted to address this issue. Many of these approaches [14, 15, 37, 47] utilize prior information, such as optical flow and depth information, to guide the reconstruction process. However, as dycheck [8] points out, most works focus on quasi-static scenes and are not generalizable for most videos in our lives. In our work, our primary focus is on utilizing robust prior knowledge to reconstruct causal monocular videos.

Video dynamics decomposition Video dynamic decomposition plays a fundamental and vital role in diverse video editing. Traditional methods have largely depended on green screens, multi-view observations, or rotoscoping. However, these approaches do not apply to typical monocular videos encountered in everyday situations. Thus, several methods [2, 18] have attempted to address the decoupling of dynamic components in monocular videos, successfully isolating the RGBA representation of both the foreground and background. However, prior methods primarily focused on the main dynamic components or considered shadows in isolation [38], neglecting the overall associated effects of dynamic elements, such as shadows and lighting. Omnimatte [21], was the first to propose a generic framework capable of learning all associated effects. In recent years, highly relevant improvements have emerged, whether by incorporating 3D information [16, 33], employing self-supervised techniques to obtain foregrounds [39, 42], or utilizing UV mapping to facilitate follow editing [12], all of which have shown promising results. However, similar to NeRF [24], these methods are limited by their rendering techniques or over-reliance on priors like optical flow and depth estimation, resulting in insufficient robustness. Recent work on 4D Gaussian Splatting (4DGS) has also demonstrated some capabilities for dynamic reconstruction and static-dynamic decoupling. Building on this, we aim to enhance the robustness of the video dynamic Omnimatte decomposition framework by refining 4DGS, thereby broadening its applicability to causal videos.

3. Method

Given a monocular video, our task is to generate a high-quality reconstructed scene while decoupling dynamic and static elements. Monocular video reconstruction is ill-posed since the observation of dynamic objects is limited under one view of one frame and is always insufficient. Although the static scene usually has richer views and information to achieve a stable and reliable result, preview methods strongly rely on various priors to help reconstruct the dynamic part. Our method can reconstruct a dynamic scene from a general monocular video taken freely and achieve the decomposition of dynamic and static scenes. It is fully self-supervised and does not require additional training data, no manual labeling, and no optical flow, but only input videos.

The overview of our method is divided into two stages as shown in Figure 2, stage 1 is to obtain a high-quality mask for the dynamic part including (b)Automask and (c)Dynamic classifier, and stage 2 is to decouple the static part and the dynamic part and reconstruct the dynamic scene. In stage 1 we will first generate enough numeral temporal consistent masks, then we will use a robust matching model [5] to gain matched points cross frames and use epipolar geometry to classify whether the masked object is dynamic or not. After that, we propose a semantic filter to avoid potential ambiguity in poor-feature regions. More details will be explained in Section 3.1.

As for stage 2 in Section 3.2, we will introduce the static dynamic hybrid representation based Gaussian Splatting which largely refers to 4d-gaussian splatting [41] and the training and optimization details will be discussed in Section 3.3.

3.1. Dynamic Mask Estimation

Previous methods for obtaining dynamic masks typically rely on manual labeling or optical flow techniques [23]. However, manual labeling is labor-intensive and costly, while optical flow often fails in areas with limited features or when camera motion exceeds object movement, as seen in street scenes. To address these limitations, we leverage large pre-trained models in the preprocessing phase, providing a more robust and automated pipeline for dynamic mask generation. Specifically, we use SAM2 [27] to initialize masks with general semantic priors and RoMa [5] to classify dynamism based on these priors.

Mask initialization The process of obtaining the mask is always under-considered. The cost of annotation and robustness challenge is often magnified or prioritized for resolution in practical applications. To address this issue, we initialize masks through SAM2(Segment Anything model 2) [27]. Leveraging the capabilities of SAM2, we can automatically perform a comprehensive segmentation on an initial frame, generate the corresponding mask, and propagate through the whole video. This self-generated mask approach is more labor-efficient and scalable compared to interactive methods. As a large model, it exhibits greater generalization capability compared to fine-tuned, task-specific segmentation models. Additionally, it provides more refined edges and retains higher-frequency details than previous methods.

Matching-based classifier After gaining the masks, we need to classify their dynamism. Previous methods for determining physical dynamism predominantly rely on optical flow, with some approaches even deriving dynamic masks directly from optical flow. However, optical flow operates under a strong assumption of content consistency between frames, which conflicts with the incomplete observation of dynamic objects in monocular video sequences. Additionally, optical-flow-based methods [23] tend to fail when encountering non-rigid dynamic objects or objects with less prominent image features. In short, optical flow represents the correspondence instead of motion itself, which results in misalignment.

We propose a feature-matching-based approach to identify dynamic objects and design a semantic filter to incorporate commonsense knowledge, thereby mitigating unavoidable errors arising from ambiguities in image features. Given two paired frames at t_i, $t_{i+\Delta t}$, where Δt is fixed frame intervals for equal frame rate video to ensure there will be significant dynamics. We use RoMa [5] to estimate a dense warp $W^{t \rightarrow t_{i+\Delta t}}$ and a matchability score $p(P_{t_i})$, where P_{t_i} means the matched key points in frame t_i pair. We sample key points paired according to the matchabil-

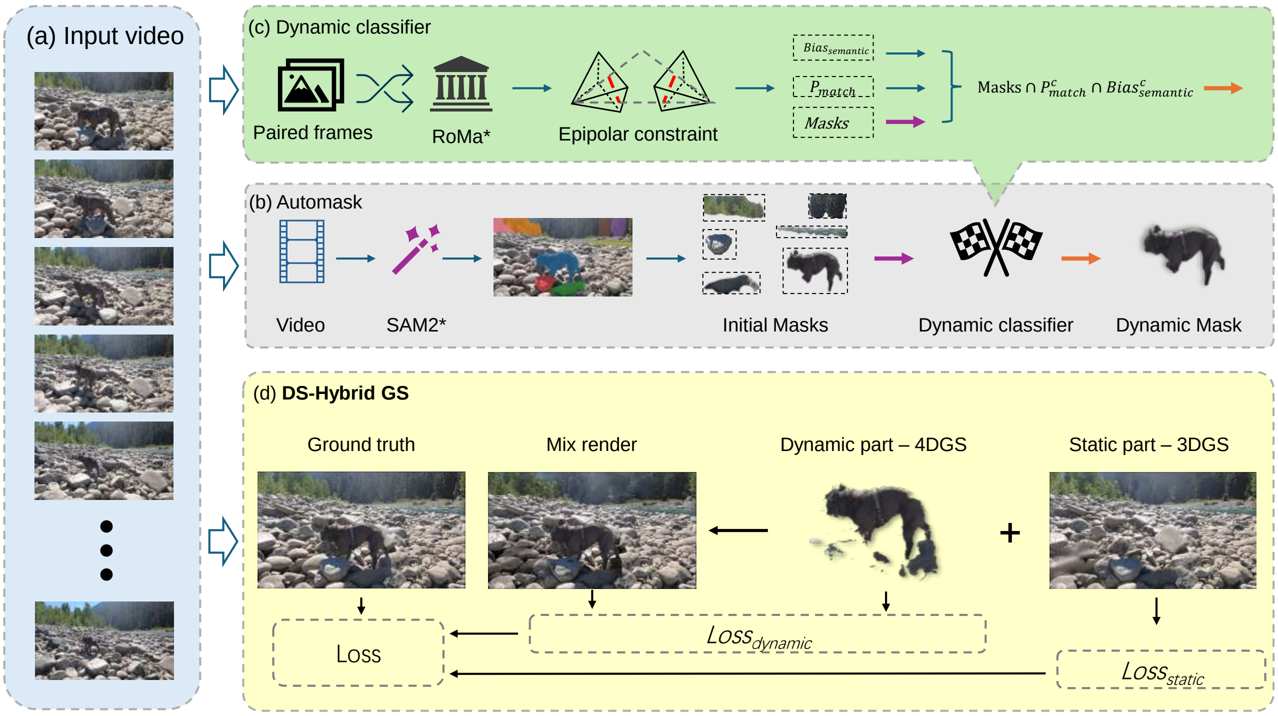

Figure 2. **Pipeline** (a) Our pipeline takes a causal video as input, enabling automatic segmentation and reconstruction with decomposition of dynamic components. (b) First, we preprocess the video using SAM2 [27] to generate a set of high-quality masks for the entire video. Based on these initial masks, we design a dynamic classifier to determine whether the masked object is dynamic. (c) With consistent masks across frames, we select paired frames at fixed intervals and use RoMa [5], a dense feature matching model, for coarse key points matching. Since dynamic objects often violate epipolar constraints, we apply epipolar constraint filtering to obtain refined match key points P_{match}, and the superscript ^{c} means complementary set. Additionally, we introduce a semantic bias to avoid ambiguity and mismatches in low-feature areas, such as the sky or road. Consequently, the final criterion for a dynamic mask is that it falls within sparsely matched key points and is not within the semantic bias. (d) Finally, given the dynamic part masks obtained from the monocular video, we can initialize a more accurate point cloud via structure-from-motion [29], enabling scene reconstruction with the dynamic parts decoupled.

ity score. Since dynamic objects violate the epipolar constraint, we can conclude that the registration points within the dynamic mask are likely to be sparser as shown in Fig 3. Once we obtain accurately matched key points between two frames, we can employ the RANSAC [6] method to estimate the essential matrix to exclude points that do not conform to epipolar constraint. This condition can be utilized to determine whether the objects within the mask are dynamic. Moreover, ambiguity always exists in some low-texture parts, thus we propose a semantic bias set to avoid some usual textures that may result in sparse match points, like sky or road, which significantly work in street scenes. Above all, we classify an object as dynamic or not by following the equation:

$$\mathrm{D} = \left\{ x \in \mathrm{I} \;\middle|\; \frac{\Delta t}{T} \sum_{i=1} \mathrm{Density}(P_{t_i}) < \tau \text{ and } x \notin \mathrm{Bias} \right\} \tag{1}$$

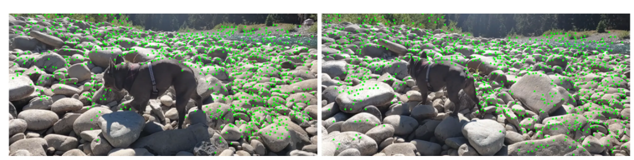

Figure 3. **Match points Example**

where $x \in$ I represents points in the initial mask, $\frac{\Delta t}{T}\sum_{i=1}\mathrm{Density}(P_{t_i})$ is the average key point density across frames with interval Δt, and τ is the dynamic threshold for identifying sparse matches. $x \notin$ Bias excludes points within the semantic bias set (e.g., sky or road).

3.2. Dynamic Scene Representation

As a scene representation method, 3D Gaussian Splatting (3DGS) [13] benefits from a well-optimized rasterization system on GPUs, achieving high-quality real-time novel view synthesis. And 4D Gaussian splatting(4DGS)

[41], build on 3DGS, achieve real-time photorealistic dynamic novel view synthesis. Thus, we aim to combine the strengths of both approaches by proposing a hybrid scene representation method that effectively integrates dynamic and static components.

Preliminary: 4D Gaussian Splatting 3DGS represent the whole scene as a cloud of 3D Gaussians while each Gaussian has a theoretically infinite scope. Compare to normalized Gaussian function in origin 3DGS, 4DGS prove the the unnormalized Gaussian function of a multivariate Gaussian can be factorized as the production of the unnormalized Gaussian functions of its condition and margin distributions and hold the critical properties. Thus the influence of a Gaussian on a given spatial position $x \in \mathbb{R}^3$ defined by an unnormalized Gaussian function:

$$p(x|\mu, \Sigma) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \tag{2}$$

where $\mu \in \mathbb{R}^3$ is its mean vector, and $\Sigma \in \mathbb{R}^{3\times 3}$ is an anisotropic covariance matrix. For the Mean vector μ of a 3D Gaussian is parameterized as $\mu = (\mu_x, \mu_y, \mu_z)$ in static scene and $\mu = (\mu_x, \mu_y, \mu_z, \mu_t)$ for a 4D Gaussian dynamic scene. And the covariance matrix Σ both Gaussian is factorized same into a scaling matrix S and a rotation matrix R as:

$$\Sigma = RSS^T R^T, \tag{3}$$

where S is summarized by its diagonal elements $S = \text{diag}(s_x, s_y, s_z)$, whilst R is constructed from a unit quaternion q for static Gaussian. In contrast, 4D Gaussian extent time dimension equally to a space dimension in scale matrix and rotaion matrix.

Moreover, a 3D Gaussian also includes a set of coefficients of spherical harmonics (SH) for representing view-dependent color$c_i(d)$, where c_i denotes the color of the i-th visible Gaussian from the viewing direction d_i, along with an opacity α. 4DGS proposes to represent $c_i(d, t)$ as the combination of a series of 4D spherindrical harmonics (4DSH) which are constructed by merging SH with different 1D-basis functions.

In 4DGS rendering, given a pixel with spatial coordinates (u, v) and timestamp t in view $\mathcal{I}$, its color $\mathcal{I}(u, v, t)$, after being further factorized as a product of a conditional probability $p_i(u, v|t)$ and a marginal probability $p_i(t)$ at time t, can be computed by blending visible 4D Gaussians$p_i(x, y, z, t)$, that have been sorted according to their depth. Whereas, a 4D Gaussian can also be factorized into $p_i(x, y, z|t)$ and t, where $p_i(x, y, z|t)$ is a 3D Gaussian whose projection in view plane can be approximated by 2D Gaussian $p_i(u, v|t)$. Same as the linearize the perspective transformations in [13, 41, 49], mean of the derived 2D Gaussian is obtained as:

$$\mu_i^{2d} = \text{Proj}\left(\mu_i | E, K\right)_{1:2}, \tag{4}$$

where $\text{Proj}\left(\cdot | E, K\right)$ denotes the transformation from the world space to the image space given the intrinsic K and extrinsic E. The covariance matrix is given by

$$\Sigma_i^{2d} = (JE\Sigma E^T J^T)_{1:2,1:2}, \tag{5}$$

where J is the Jacobian matrix of the perspective projection. After get the 2D Gaussian $p_i(u, v|t)$ for alpha blending, the rendering equation can be described as below:

$$\begin{aligned} I(u, v, t) = &\sum_{i=1}^{N} p_i(t) p_i(u, v|t) \alpha_i c_i(d, t) \\ &\times \prod_{j=1}^{i-1} (1 - p_j(t) p_j(u, v|t) \alpha_j). \end{aligned} \tag{6}$$

where $c_i(d, t)$ denotes the color of the i-th visible Gaussian from the viewing direction d_i at timestamp t, α_i represents its opacity.

3D-4D hybrid representation Although 4DGS has demonstrated remarkable results in reconstructing dynamic scenes, its performance in novel view synthesis is suboptimal. This is primarily due to the tendency of 4DGS to overfit static scenes, resulting in issues with spatial consistency for static elements and significantly increasing both rendering and storage overhead. In other words, 4DGS cannot distinguish the view effect or time effect with only a monocular video as input. Inspired by techniques in video matting, we propose a hybrid 3D-4D hybrid scene representation. In this approach, dynamic regions are first masked out, allowing separate optimization of the static scene before refining the dynamic regions. It is noteworthy that certain "Omnimatte" elements, like shadows and reflections within the static scene, may also be learned as part of the dynamic representation. To obtain a cleaner decouped scene, a retraining strategy with background constraints can further enhance the decoupling of static and dynamic components. And since we majorly focus on monocular reconstruction, we simplify the general mix rendering formula into alpha blending which is enough for most monocular video cases shown below:

$$I_{\text{blend}}(u, v, t) = \sum_{i=1}^{N} \alpha_i I_i(u, v, t) + \left(1 - \sum_{i=1}^{N} \alpha_i\right) I_b(u, v) \tag{7}$$

where $I_i(u, v, t)$ and α_i means the i-th dynamic foreground color and its alpha, and $I_b(u, v)$ is the time-invariant static background color. Our representation enables accurate scene reconstruction with improved spatial consistency, making it efficient for subsequent editing tasks. Furthermore, the internal representation is flexible and can be updated to incorporate any advanced image-based reconstruction methods as they become available.

3.3. Optimization

Our training process is divided into two main stages: first, training the static components, followed by the optimization of the dynamic elements. Supervision is provided by a combination of three distinct loss functions. During the static training phase, dynamic regions are masked out, and only the real images of the static components are used to guide the reconstruction through a static loss term. Once the static scene converges, the static model is frozen, and we proceed with the reconstruction of the dynamic scene. In this stage, a background loss is introduced to preserve the clarity of the background in the dynamic regions after initializing the dynamic part, while the overall reconstruction is supervised by comparing the rendered scene with the full real image using a reconstruction loss.

The total loss function used for training is as follows:

$$Loss = \lambda_{dssim} L_1 + (1 - \lambda_{dssim}) L_{ssim} + \lambda_{stage} L_{bg} \quad (8)$$

Here, L_1 is the L1 loss, which measures the absolute difference between the predicted image and the ground truth image. L_{ssim} is the SSIM loss, which is derived from the Structural Similarity Index Measure and evaluates perceptual similarity. L_{bg} represents the background loss, which contains the same helps to reduce artifacts in the dynamic layers. The hyperparameter λ_{dssim} determines the relative contribution of the L1 and SSIM losses, while λ_{stage} is a binary parameter that ensures the background loss is applied only during the dynamic stage of training.

Initialization We first obtain the dynamic mask using our proposed method and preprocess the images with this mask. These preprocessed images are then used to derive the camera poses and initial point cloud required for Gaussian Splatting through structure-from-motion [29, 30]. Both the static and dynamic scene representations are initialized based on this point cloud.

Implementation details For the SAM2 hyperparameters, we set 64 points per side, 128 points per batch and only one crop number of layers. As for the RoMa matching model [5], we did not finetune or modify the model. The dynamic threshold in most cases is 0.01. As for the optimization part, we use Adam optimizer and we perform both 10000 iterations in static and hybrid training stages for general scenes like in Omnimatte-wild datasets [21] and both 30000 iterations for street scenes since it is harder to coverage with the same learning rate 0.00016. The λ_{dssim} is 0.2. Training a general scene usually takes 3.5 hours on a single RTX4090 graphic card. And all the preprocessing is also done on the same device. Our code and dataset will be made public and available.

4. Experiments

In this section, we present a comprehensive comparison with state-of-the-art methods via both qualitative and quantitative evaluations. For qualitative results, we assess our method across various datasets, unlike previous methods that primarily focus on video reconstruction with limited camera motion dynamics, we also evaluate our approach on `Waymo` [34] datasets, which involve more complex and dynamic scenarios. Additionally, our automatic mask generation method is rigorously tested on these practical datasets to validate its applicability and robustness in real-world conditions. Meanwhile, quantitative results are presented on `Waymo` dataset for the reconstruction performance and `Movies` dataset to examine decomposition capability and visual fidelity. The `Movies` dataset, proposed by OmnimatteRF [16], includes ground truth backgrounds specifically designed to evaluate model decoupling performance. The baseline setting and preprocessing are the same as the instruction.

4.1. Qualitative Evaluation

We present a qualitative comparison of methods in Fig. 4 and Fig. 5. In Fig. 4, our method demonstrates strong performance on this dataset. By using only the dynamic car masks, our approach successfully incorporates the associated shadows into the dynamic region. In contrast, the baseline method struggles with accurately decomposing the static background, as the ambiguity in optical flow often causes the network to mistakenly treat all scenes as dynamic. In Fig. 5, we also evaluate our method in `iPhone` dataset [26], `dynamic scenes`dataset [43], and `Movies` dataset with separation results.

4.2. Quantitative Evaluation

We select PSNR, SSIM, and LPIPS as the evaluation index in both the reconstruction metric and decomposition metric. We first quantitatively evaluated our method on `Waymo` dataset, comparing the reconstruction performance against baseline methods as shown in Table 1. The results, along with several sampled visualizations in Fig 4, demonstrate that our approach outperforms current state-of-the-art methods of reconstruction from monocular video with decomposition ability. For further analysis, we also evaluated our method on the Movie dataset by comparing metrics between our segmented backgrounds and the ground truth backgrounds, as presented in Table 2. (Some results cite from OmnimatteRF [16]) Our methods perform second best in this metric and have more details in some regions of

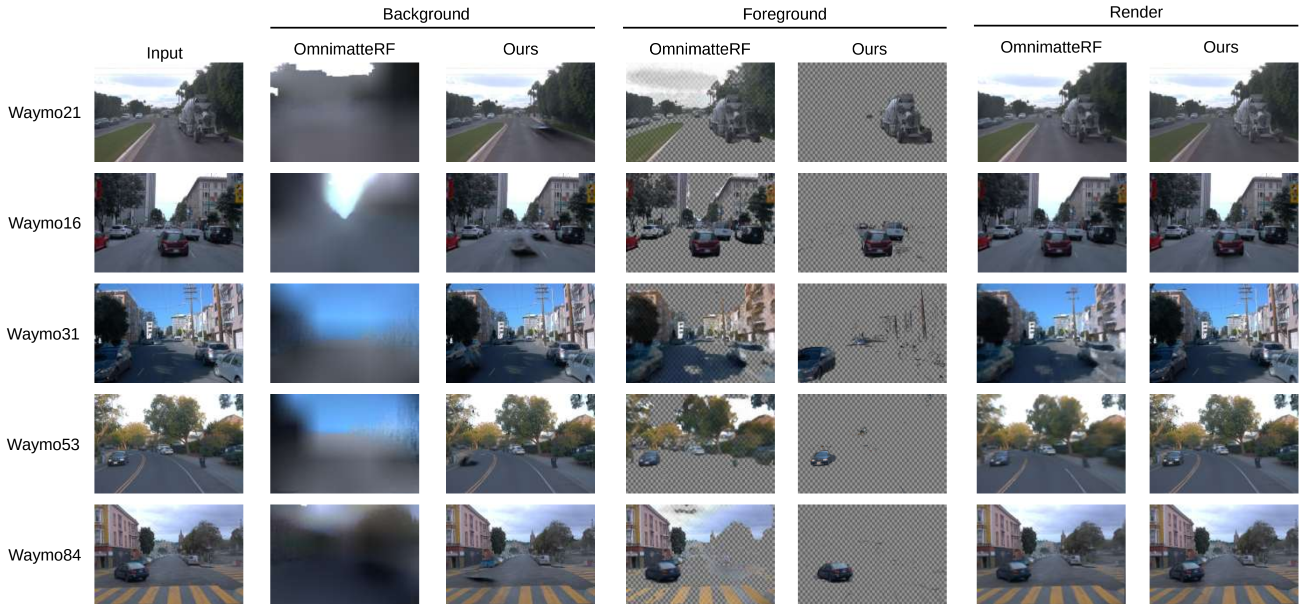

Figure 4. **`Waymo` qualitative evaluation** We evaluated our method and OmnimatteRF on several `Waymo` dataset cases. The baseline method fails to handle these cases with extreme camera motion. The input mask is generated by our automatic method instead of the dynamic masks projected from the labeled bounding box in the dataset. For the foreground render we apply an alpha threshold of 0.5 which is the same as in the baseline code.

Waymo	016			021			031			053			084		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
OmnimatteRF	0.271	0.854	30.60	0.288	0.877	30.82	0.435	0.715	22.47	0.393	0.742	24.04	0.198	0.902	33.31
Ours	**0.097**	**0.968**	**32.67**	**0.103**	**0.962**	**32.19**	**0.096**	**0.956**	**30.45**	**0.129**	**0.938**	**28.15**	**0.066**	**0.975**	**33.85**

Table 1. **Reconstruction quantitative evaluations.** We present the reconstruction comparison of our method and baselines on the waymo datasets. The better results are in **bold**.

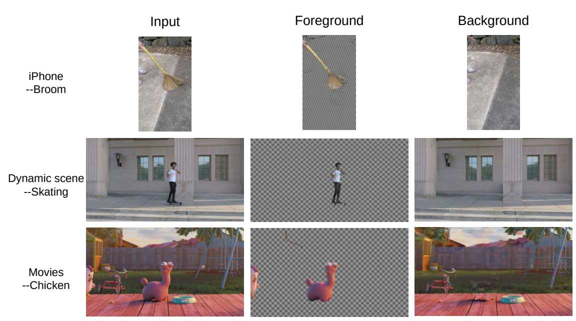

Figure 5. **Various datasets qualitative evaluation**

the background compared to the best method as shown in Fig. 6.

4.3. Ablation study

We demonstrate the effect of our background constraint. When directly reconstructing the scene, the static part's reconstruction is limited by the constrained observation from a monocular video, leading to unavoidable artifacts in some high-frequency regions. Meanwhile, in the dynamic part, 4DGS tends to overfit the static scene components. To mitigate this, we introduce a background loss that encourages the regions outside the dynamic masks to remain cleaner, helping to refine the reconstruction of the dynamic parts as we shown in Fig 7

5. Conclusion

We introduce an automated method for the complete reconstruction of scenes from monocular videos, with the capability to automatically decouple dynamic and static components. Extensive experiments have demonstrated that our approach is not only comparable to existing methods on simpler datasets but also exhibits superior performance and robustness in more complex and rapidly changing scenarios. Additionally, our method for automatically obtaining dynamic masks is readily transferable to other techniques.

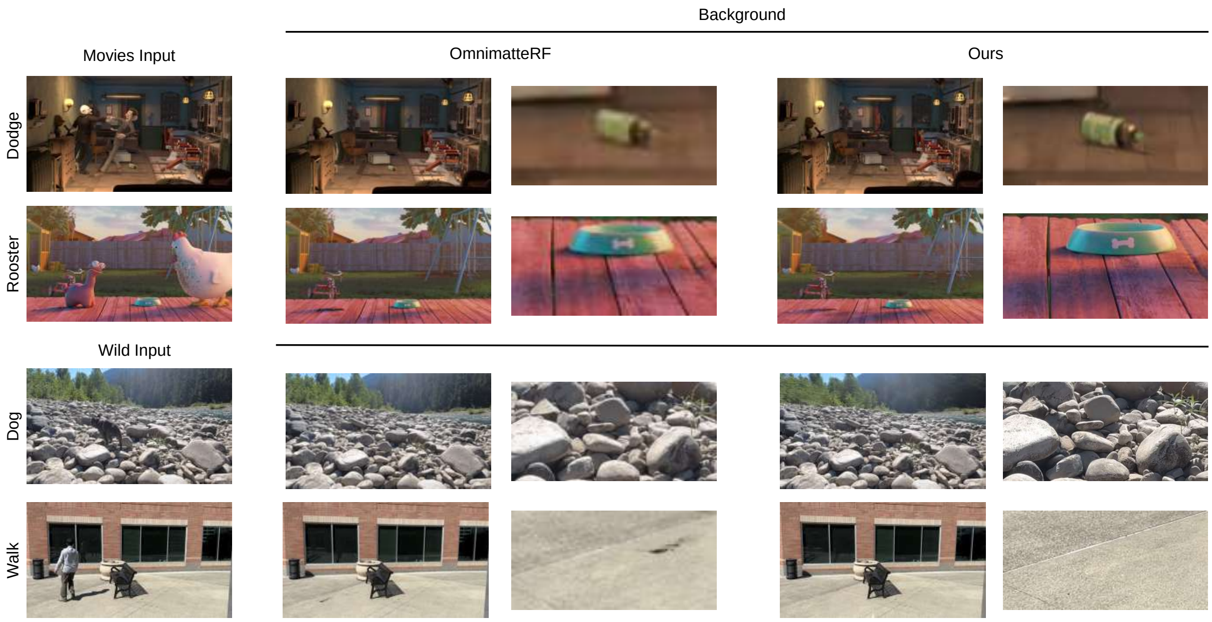

Figure 6. **Background visualizations** We also evaluate our background decomposition capabilities in comparison to OmnimatteRF on the datasets they proposed. Additionally, due to the advantages of Gaussian splatting, our method produces more precise results for the background in the near-field, particularly under similar training conditions.

Movies	Donkey			Dog			Chicken			Rooster			Dodge		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
D^{2}NeRF	-	-	-	0.370	0.694	22.73	-	-	-	0.340	0.708	25.13	0.408	0.729	20.95
Omnimatte	0.315	0.653	19.11	0.279	0.706	21.74	0.312	0.704	20.95	0.220	0.741	23.14	0.067	0.879	23.88
LNA	0.104	0.849	18.79	0.154	0.828	26.08	0.190	0.818	19.22	0.131	0.804	26.46	0.068	0.937	24.94
4DGS	0.357	0.614	16.48	0.427	0.628	19.64	0.390	0.653	19.42	0.558	0.490	15.41	0.333	0.701	19.41
OmnimatteRF	**0.005**	**0.990**	**38.24**	**0.030**	**0.976**	**31.44**	**0.021**	**0.978**	**32.86**	**0.024**	**0.969**	**27.65**	**0.006**	**0.991**	**39.11**
Ours	0.242	0.783	20.60	0.154	0.941	27.51	0.171	0.891	25.67	0.189	0.906	23.95	0.066	0.970	31.86

Table 2. **Decomposition quantitative evaluations.** We present the background reconstruction comparison of our method and baselines on the Movies datasets. The best results are in **bold**. The second best results are in underline. Results marked - are the ones where the method failed to give good separations.

Figure 7. **Ablation** The input consists of the image and dynamic mask, while the output includes the dynamic rendering results and the alpha visualization. The alpha visualization provides a more intuitive way to assess the effectiveness of the background loss, as it clearly illustrates the separation between dynamic and static regions in the scene.

Moreover, our method has several possible further works. First, our method can also be conveniently expanded to multi-view reconstruction in street scenes, significantly reducing annotation costs while providing high-quality reconstruction results. Second, we could incorporate additional regularization and constraints, such as depth, to improve the method. But, in this work, we believe that fewer constraints and regularization lead to greater robustness in the approach.

However, our method is not without limitations. The automatic mask method requires threshold adjustments tailored to specific kinds of scenes, and it may fail with relatively stationary objects. To address this, we propose to extend the video range in autonomous driving scenarios.

In summary, our method holds broad potential for application across various domains.

CVPR #4792

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Rick Szeliski. Building rome in a day. *Communications of the ACM*, 54: 105–112, 2011. 2

[2] Xue Bai, Jue Wang, David Simons, and Guillermo Sapiro. Video snapcut: robust video object cutout using localized classifiers. *ACM Transactions on Graphics (ToG)*, 28(3):1–11, 2009. 2

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *arXiv e-prints*, 2021. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *IEEE*, 2023. 2

[5] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3, 4, 6

[6] Martin A Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[7] 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. 1

[8] 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

[9] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5511–5520, 2022. 2

[10] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5875–5884, 2021. 2

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

[12] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 1, 2, 3

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 4, 5

[14] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[15] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4273–4284, 2023. 1, 2

[16] Geng Lin, Chen Gao, Jia-Bin Huang, Changil Kim, Yipeng Wang, Matthias Zwicker, and Ayush Saraf. Omnimatterf: Robust omnimatte with 3d background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23471–23480, 2023. 2, 3, 6

[17] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[18] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 2

[19] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2

[20] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[21] Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T Freeman, and Michael Rubinstein. Omnimatte: Associating objects and their effects in video. In *CVPR*, 2021. 2, 3, 6

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

[23] Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Self-supervised segmentation by grouping optical-flow. In *Computer Vision – ECCV 2018 Workshops*, pages 528–534, Cham, 2019. Springer International Publishing. 3

[24] 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. 2, 3

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[26] 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, 6

[27] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-

Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 4

[28] Christian Reiser, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron, and Peter Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2

[29] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 6

[30] 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. 2, 6

[31] Qing Shuai, Haoyu Guo, Zhen Xu, Haotong Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Real-time view synthesis for large scenes with millions of square meters. 2024. 2

[32] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2

[33] Mohammed Suhail, Erika Lu, Zhengqi Li, Noah Snavely, Leonid Sigal, and Forrester Cole. Omnimatte3d: Associating objects and their effects in unconstrained monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 630–639, 2023. 2, 3

[34] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[35] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2

[36] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2

[37] 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

[38] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1880–1889, 2020. 3

[39] 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

[40] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9421–9431, 2021. 2

[41] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 5

[42] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2666, 2022. 1, 3

[43] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020. 6

[44] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[45] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024. 2

[46] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13144–13152, 2021. 2

[47] 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. 2

[48] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv e-prints*, 2020. 2

[49] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa splatting. *IEEE TVCG*, 2002. 5