A. Video Examples

Please refer to https://romosfm.github.io/ to view videos of our results. We show video motion segmentation results on FBMS59, DAVIS16, and TrackSegv2 compared to OCLR-adap [18]. We further show masked video results on the Casual Motion dataset, and some in-the-wild video samples.

B. Optical flow limitations – Figure 1

Despite recent advancements that have made optical flow prediction networks a powerful and versatile tool, there are inherent limitations to optical flow. One is the ambiguity of flow predictions for shadows [13]. This can lead to an inability to detect moving shadows as distinct moving entities in our segmentation masks (top of Fig. 1).

Another key limitation are objects that appear and disappear almost instantly, such as the arm in our 'Table Objects' scene. These abrupt changes behave similar to occluded areas where the flow is ambiguous and fail the cycle consistency check, rendering nearly all pixels from such objects unusable for our weak inlier/outlier annotations (bottom of Fig. 1).

C. Scene optimization with distractors – Fig. 2

Videos, as a collection of images of a scene, can be used to reconstruct the 3D scene using methods like Neural Radiance Fields (NeRFs) [8] or 3D Gaussian Splatting (3DGS) [4]. However, transient inconsistencies, such as passing pedestrians, often violate the static scene assumption of these techniques, appearing as noise in the reconstruction. These inconsistencies, referred to as *distractors*, can be filtered out through robust 3D optimization methods, such as those proposed in [10–12].

RoMo can similarly be be applied to the problem of 3D optimization from such videos by incorporating its motion masks into a standard 3DGS model. We filter out dynamic pixels from the photometric loss, following the approach in Sabour et al. [11]. Similar to Sabour et al. [12], the structural similarity loss is not utilized in training the 3DGS model. Qualitative results for this application are presented in Fig. 2 for the "patio" scene from the NeRF On-the-go dataset [10], which has the temporal order of frames preserved, allowing us to compute optical flow.

Observe that RoMo effectively masks moving human distractors in this scene. We compare against results from SpotLessSplats (SLS) [12], a robust 3D optimization method for 3DGS. The results show that SLS masks more effectively capture shadows and secondary effects, which RoMo misses due to optical flow limitations as discussed earlier. However, the results for SLS show leaked distractors in areas of the scene which are sparsely sampled in the training set. This is due to the imbalance of learning rates between the mask predictor of SLS and its 3D model, i.e. the 3D model overfits to the distractor faster than the mask predictor learns its mask. Adjusting the training schedule to better balance the learning of the mask predictor and the 3DGS model can help mitigate this issue. This highlights the inherent challenge of finding an optimal learning rate balance between the two modules in SLS. In contrast, our approach avoids this problem entirely, as RoMo masks are computed as *preprocessing* on the video and provided as input to the 3DGS model optimization. Because RoMo masks operate independently of the 3D optimization pipeline, they can more seamlessly integrate with various 3D reconstruction methods, such as NeRF and 3DGS. We believe that while our motion masks might not fully capture all inconsistencies for robust 3D optimization, they can serve as a strong initialization for robust masks, which can then be further refined using methods such as SLS. Furthermore, since RoMo does not require camera poses, as many robust 3D optimizations [10–12] do, it can help in cases were SfM pipelines like COLMAP [14, 15] fail due to high distractor rates.

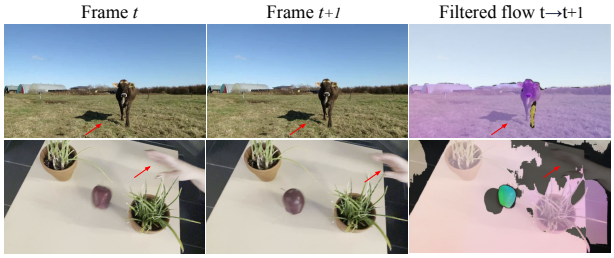


Figure 1. **Optical flow ambiguities in the presence of shadows and occlusion. Top:** Optical flow of the cow's shadow follows the ground beneath it, although it has a similar movement to the cow. **Bottom:** The fleetingly appearing arm does not pass optical flow cycle consistency and is completely filtered akin to occluded areas.

D. Ground-truth in Casual Motion

We investigate the validity of our groundtruth camera poses by evaluating the accuracy of COLMAP poses and the robotic arm's pose reproducibility. To validate COLMAP's poses on static captures we evaluate it's photometric consistency by training 44 3DGS models on the static capture of the "billiard" scene, each time leaving one image out. The average PSNR of the held-out images is 38.7dB, indicating excellent photometric consistency for which camera pose must be precise. To test the reproducibility of camera trajectories by the robotic arm, we performed multiple runs of the same trajectory on the robotic arm with the scene designed to have many visual cues to ensure COLMAP's success and ran COLMAP on all captured videos. Across 6 runs of the arm with a static scene, COLMAP yields an average ATE↓ of 0.04 ± 0.02, an order of magnitude smaller than the errors reported in Fig. 9 of the main paper.This clearly shows that COLMAP on a static scene, is more than satisfactory as a

baseline for SfM methods on the same scene but with one or more dynamic objects.

E. RoMo on Static Scenes

To ensure RoMo does not provide non-zero masks on static SfM benchmarks which could degrade SfM performance we test RoMo on the first 100 frames of the 27 static fully-lit scenes in the ETH3D [16] test set. It correctly identified static content (producing zero dynamic masks) in 24 scenes. On average, over all scenes it produces only 0.4% nonzero pixels per video, but with no impact on COLMAP's estimates. This significantly outperforms the synthetically supervised SOTA motion segmentation method OCLR [18], which produced zero masks on only 2 scenes, with an average of 23.5% nonzero pixels per video.

F. Results on "Casual Motion" – Figure 3

Figure 3 presents a more detailed breakdown of results from our "Casual Motion" dataset (main paper Figure 9). It illustrates that supervised baselines, which rely heavily on synthetic data, have less reliable estimates of camera pose compared to classic camera estimation methods like `COLMAP` [14, 15]. The 'Money Leaf' scene exemplifies significant challenges for `ParticleSFM` [21], `LEAP-VO` [2], and `MonST3R` [20], all of which produce notably inferior results compared to `COLMAP`. In contrast, our method leverages `COLMAP`'s strength as a robust camera pose estimator while addressing its limitations. This enhancement is evident both quantitatively and qualitatively, particularly at the beginnings and ends of trajectories. In these regions, where slower camera movements with smaller translation are overshadowed by the larger motions of dynamic objects, `COLMAP`'s estimates often falter. Our approach corrects these errors effectively by incorporating dynamic masks.

G. Failure scenes of ParticleSFM – Figure 4

Figure 4 presents a detailed comparison of camera pose estimation baselines on scenes where `ParticleSFM` struggles. The 'Table Objects' scene is particularly challenging due to rapid camera and rapid object movements, which result in motion blur and sparse dynamic objects. These factors make masking difficult for all methods, including ours. `COLMAP` is generally robust to this scene because the movements, though rapid, are temporally sparse. Poor masking however, can lead to failures in the robust baselines. Qualitative results show that `ParticleSFM` [21] focuses its detected tracks (blue and green) and filtered dynamic tracks (green) on the static flowerpots, which provide texture and reliable cues for bundle adjustment in an otherwise plain-textured scene. This incorrect masking causes `ParticleSFM` to completely fail at camera estimation. `MonST3R` [20] produces good masks in some frames but fails with empty masks

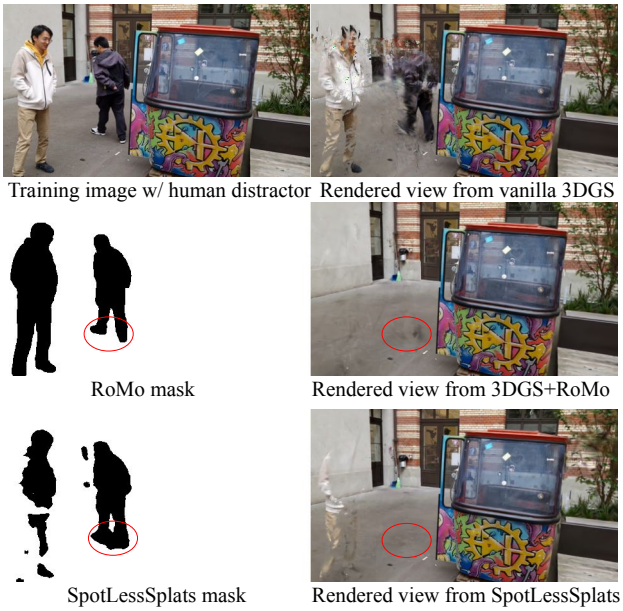


Figure 2. **Application of RoMo in 3D optimization** – with in-the-wild videos, shows that RoMo can completely mask distractor humans in the scenes but fails to capture shadows due to optical flow limitations as described in Appendix B.

in others. `LEAP-VO` [2] shows no evidence of filtering tracks associated with dynamic objects (green arrows). Our method partially fails to detect the fleetingly appearing arm but successfully masks out the moving fruits even under heavy blur.

The 'Stairs' scene presents a highly occluded environment. `ParticleSFM` fails to estimate camera poses for the final frames with the most occlusions, likely due to the sparsity of remaining tracks (blue region in Figure 4). `MonST3R` occasionally misses moving people, and `LEAP-VO` does not filter tracks of dynamic objects. In contrast, RoMo fully masks the dynamic people in this scene.

Finally, in the 'Umbrella Garden' scene, `ParticleSFM` fails to find sufficient tracks due to the high occlusion rate during its initial stage, leading to a complete failure.

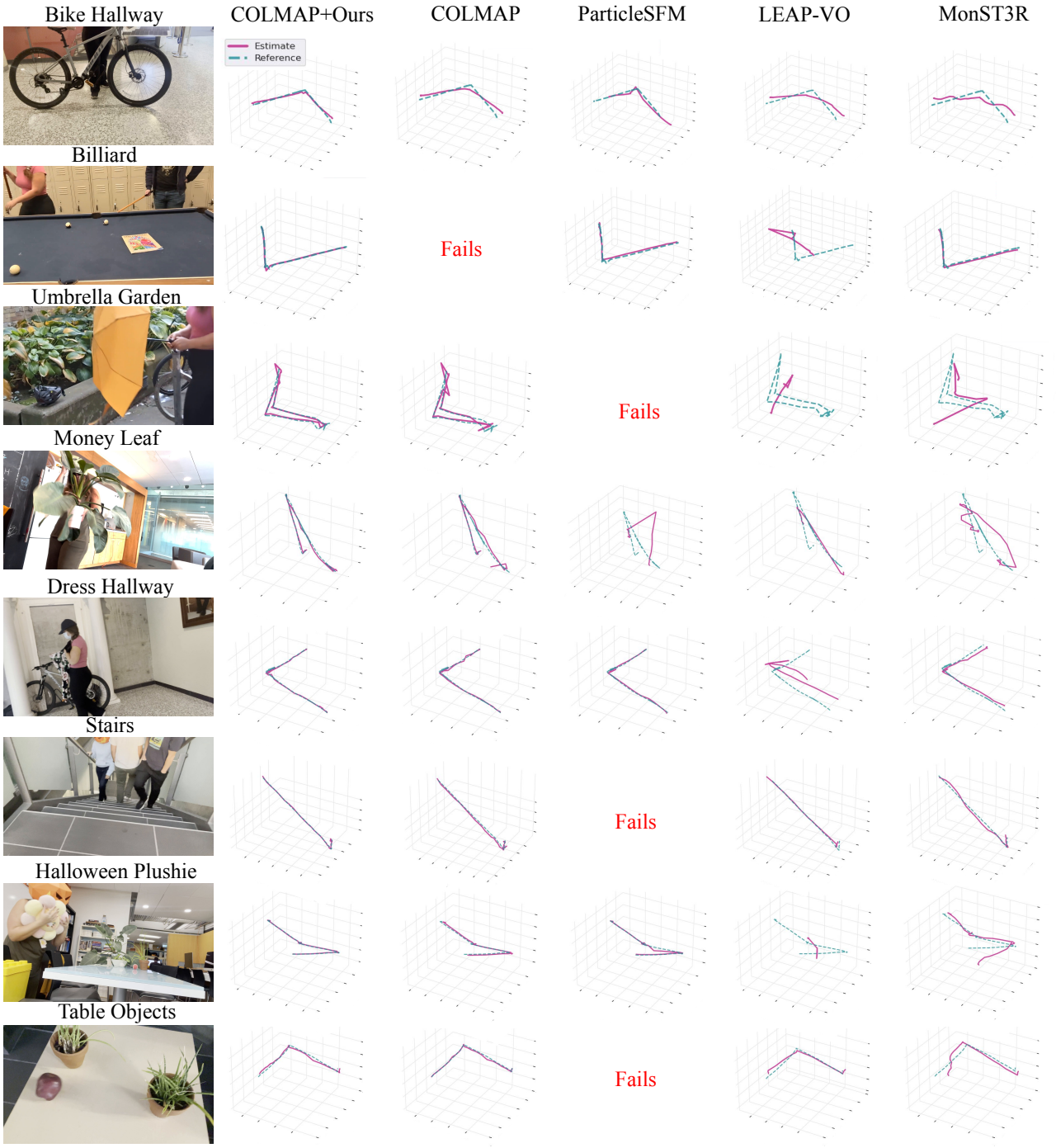
H. Detailed results on MPI Sintel – Table 1

Table 1 presents a per-scene breakdown of results on MPI Sintel for both our method and unmasked TAPIR [3] tracks used with TheiaSfM [17].

I. Additional details on baseline experiments

In Fig. 6 of the main paper, we present baseline results on the `Casual Motion` dataset. For `OCLR`[18], we follow the authors experiment settings, where the number of detected moving objects is set to a fixed number of three. We perform test-time adaptation of DINO features to the test video for `OCLR-adap`. The unsupervised networks for

STM [6] and EM [7] were trained on the Flying Things 3D dataset [5] and then applied directly to the Casual Motion dataset, consistent with the authors' evaluation protocol of testing their networks across different datasets. For STM, we adopt the original paper's approach of selecting the mask that best matches the ground truth at test time. Since the Casual Motion dataset contains only a test set, we could not train a separate model for the Motion Grouping baseline [19] and instead utilized weights from a network trained on the FBMS59 dataset [9]. After experimentation, we found that a video gap of 1 yielded optimal results for this baseline, which we report in our evaluation.



Scene	ATE						RPE-T						RPE-R					
	COLMAP+Ours	COLMAP	COLMAP+OCLR	MonST3R	LEAP-VO	ParticleSFM	COLMAP+Ours	COLMAP	COLMAP+OCLR	MonST3R	LEAP-VO	ParticleSFM	COLMAP+Ours	COLMAP	COLMAP+OCLR	MonST3R	LEAP-VO	ParticleSFM
Bike Hallway	**2.77**	13.38	-	6.10	5.74	10.36	3.64	5.92	-	**2.22**	3.95	4.13	1.53	4.46	-	**0.57**	0.72	0.84
Billiard	1.90	-	-	1.54	13.94	**1.40**	3.08	-	-	3.89	5.34	**1.99**	1.42	-	-	1.30	8.19	**0.99**
Umbrella Garden	**7.84**	10.80	-	25.83	26.26	-	**6.99**	14.08	-	11.96	12.18	-	**2.30**	**2.30**	-	3.82	6.39	
Money Leaf	3.87	4.27	**3.48**	14.64	16.74	12.08	**6.37**	6.43	**6.37**	6.63	7.16	8.92	**4.53**	4.56	4.56	4.67	6.65	5.9
Dress Hallway	**1.96**	2.61	2.10	4.50	17.56	2.10	**2.68**	4.06	2.76	4.21	4.64	2.90	**1.97**	2.56	**1.97**	2.83	1.98	1.79
Stairs	**0.51**	0.72	-	0.72	1.35	-	**0.53**	0.65	-	1.16	1.04	-	**0.12**	0.15	-	0.74	0.30	-
Halloween Plushie	**1.16**	**1.16**	-	5.78	19.27	1.38	**0.99**	1.05	-	4.18	4.51	1.10	**0.43**	0.45	-	1.64	5.93	0.46
Table Objects	2.31	**1.06**	2.64	4.90	4.09	-	1.12	**0.73**	1.19	3.97	1.45	-	**0.45**	0.72	0.56	2.43	0.8	-

Figure 3. **Detailed results on "Casual Motion"** – show that our method can be paired with a bundle adjustment technique (COLMAP [14]) to make it more robust to dynamic scenes, often outperforming SoTA methods for camera estimation on such scenes.

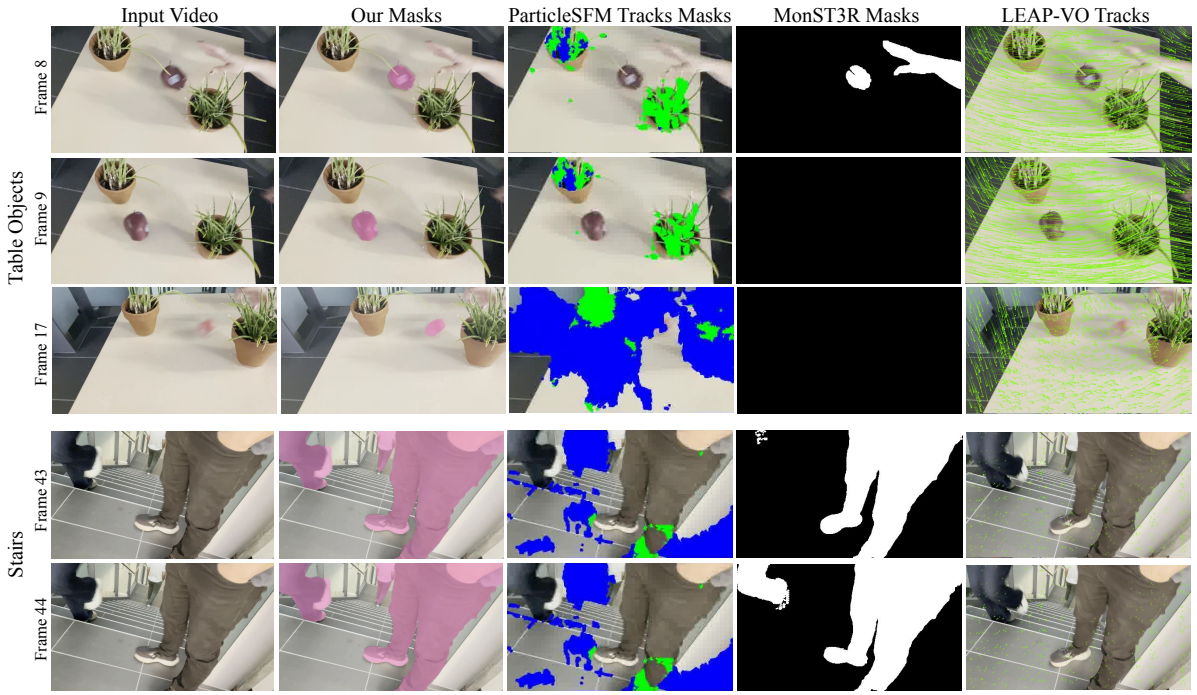


Figure 4. **Detailed results on ParticleSFM [21] failing scenes** – shows that over masking static regions can lead to bundle adjustment failure. Moreover, sparse tracks on highly occluded frames can lead to failure.

Scene	Our Masks + TAPIR tracks + TheiaSFM			TAPIR tracks + TheiaSFM		
	ATE	RPE (T)	RPE (R)	ATE	RPE (T)	RPE (R)
alley_2	0.001	0.001	0.018	0.001	0.001	0.020
ambush_4	0.014	0.015	0.188	0.017	0.014	0.159
ambush_5	0.004	0.004	0.068	0.037	0.027	0.750
ambush_6	0.003	0.002	0.047	0.150	0.090	1.802
cave_2	0.773	0.176	0.626	0.782	0.170	0.683
cave_4	0.005	0.003	0.019	0.078	0.046	0.283
market_2	0.014	0.012	0.112	0.068	0.028	8.483
market_5	0.010	0.003	0.027	0.012	0.004	0.029
market_6	0.006	0.005	0.037	0.051	0.022	0.800
shaman_3	0.001	0.001	0.213	0.005	0.003	0.680
sleeping_1	0.009	0.009	0.898	0.011	0.013	1.267
sleeping_2	0.001	0.001	0.026	0.001	0.001	0.026
temple_2	0.002	0.002	0.009	0.002	0.002	0.008
temple_3	0.456	0.128	0.743	0.626	0.204	1.452
Avg	**0.093**	**0.026**	**0.217**	0.132	0.045	1.175

Table 1. Per scene breakdown of MPI Sintel [1] results.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 5

[2] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. LEAP-VO: Long-term Effective Any Point Tracking for Visual Odometry. In *CVPR*, 2024. 2

[3] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking Any Point with Per-Frame Initialization and Temporal Refinement. In *ICCV*, 2023. 2

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM Transactions on Graphics*, 2023. 1

[5] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[6] Etienne Meunier and Patrick Bouthemy. Unsupervised Space-Time Network for Temporally-Consistent Segmentation of Multiple Motions. In *CVPR*, 2023. 3

[7] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. EM-Driven Unsupervised Learning for Efficient Motion Segmentation. *PAMI*, 2023. 3

[8] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[9] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of Moving Objects by Long Term Video Analysis. *PAMI*, 2014. 3

[10] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[11] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *CVPR*, 2023. 1

[12] 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. 1

[13] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In *NeurIPS*, 2023. 1

[14] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2, 4

[15] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 1, 2

[16] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[17] Chris Sweeney. Theia multiview geometry library: Tutorial & reference. http://theia-sfm.org, 2015. 2

[18] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting Moving Objects via an Object-Centric Layered Representation. In *NeurIPS*, 2022. 1, 2

[19] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-Supervised Video Object Segmentation by Motion Grouping. In *ICCV*, 2021. 3

[20] 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

[21] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. ParticleSfM: Exploiting Dense Point Trajectories for Localizing Moving Cameras in the Wild. In *ECCV*, 2022. 2, 5