POMATO: Marrying Pointmap Matching with Temporal Motions for Dynamic 3D Reconstruction

Supplementary Material

A. Pointmap Matching for Global Alignment.

Given a sequence of video frames, the target of global alignment is to project all pairwise estimated pointmaps to the same global world coordinates. DUSt3R constructs a connectivity pairwise graph and aims to minimize the re-projection error for each image pair globally where the dynamic regions are supposed to be separated from the static regions. To this end, MonST3R [52] further introduces an assistant optical flow network [46] to help mask the dynamic regions and provide a pseudo label of 2D matching for minimizing the re-projection error in static regions. However, the introduced assistant model will introduce inevitable domain gaps and additional computation costs. Besides, the optical flow model is tailored for matching within two adjacent frames, suffering an obvious degeneration with the large view displacement. In POMATO, for an image pair $\{\mathbf{I}^i, \mathbf{I}^j\}$, the dynamic mask $\mathbf{D}^{j,i}$ is calculated by comparing the difference between $\mathbf{X}^{j,i}$ and $\mathbf{X}_m^{j,i}$:

$$\mathbf{D}^{j,i} = ||\mathbf{X}_m^{j,i} - \mathbf{X}^{j,i}|| > \alpha, \tag{9}$$

where α is a dynamic threshold defined as $3 \times \text{median}(||\mathbf{X}_m^{j,i} - \mathbf{X}^{j,i}||)$.

Given the updated camera intrinsic $\tilde{K}$ after an iteration of optimization, the target matching 2D coordinates $\mathbf{F}_m^{j,i} \in \mathbb{R}^{H \times W \times 2}$ can be calculated as $\mathbf{F}_m^{j,i} = p(\tilde{\mathbf{K}}\mathbf{X}_m^{j,i})$ where p is a mapping from 3D camera coordinates to 2D pixel coordinates. The optical flow loss proposed in MonST3R can thus be modified with our dynamic mask and 2D matching coordinates. Details about the optical flow loss are referred to MonST3R [52].

B. Fast 3D Reconstruction with video POMATO

Given a sequence of images less than the temporal window length of 12 frames, dynamic 3D reconstruction can be obtained by directly estimating the pointmaps of all reference images to the coordinate of the key frame as discussed in the Sec.3.4. Here, we provide more visualization results of this feed-forward manner and demonstrate the effectiveness of introducing the temporal motion module. As shown in Fig.8, directly applying the pairwise reconstruction will suffer from an obvious scale shift among different frames. After the temporal motion module, the consistency within the video sequence obtains an obvious enhancement.

C. Training Data Details

The details about the training datasets can be found in Tab.6. The finetuning procedure of POMATO was conducted exclusively using synthetic training datasets.

D. More Visualizations on Dynamic Scenes

We provide more visualizations in Fig. 9 and Fig. 10. MonST3R suffers obvious degeneration when the view displacement is large as reflected by the erroneous pose estimation while POMATO can still provide a consistent camera trajectory.

Dataset	Domain	Scene Type	# of Frames	# of Scenes	Dynamics	Ratio
PointOdyssey [54]	Synthetic	Indoors & Outdoors	200k	131	Realistic	57.1%
TartanAir [45]	Synthetic	Indoors & Outdoors	100k	163	None	14.3%
DynamicReplica [21]	Synthetic	Indoors	145k	524	Realistic	14.3%
ParallelDomain4D [40]	Synthetic	Outdoors	750k	15015	Driving	8.6%
Carla [10]	Synthetic	Outdoors	7k	5	Driving	5.7%

Table 6. **An overview of all training datasets and sample ratio.** All datasets provide both camera pose, depth, and most of them include dynamic objects.

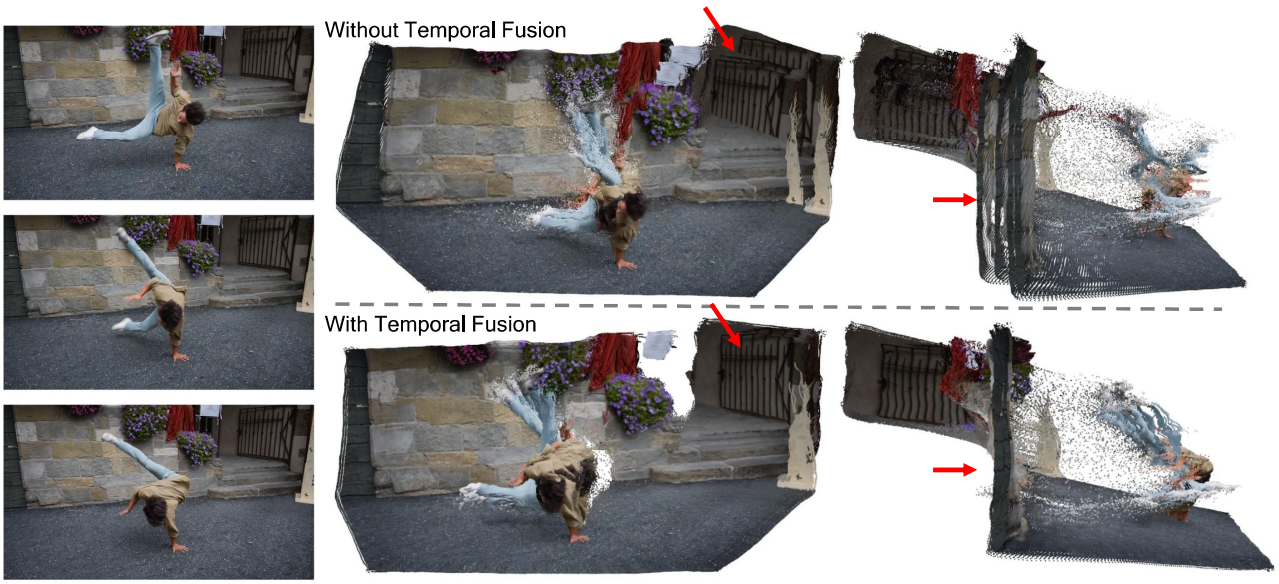

Figure 8. **Fast 3D reconstruction with our temporal motion module**. Given a sequence of images less than temporal window length, our POMATO can directly obtain a global pointmap under the key frame coordinate.

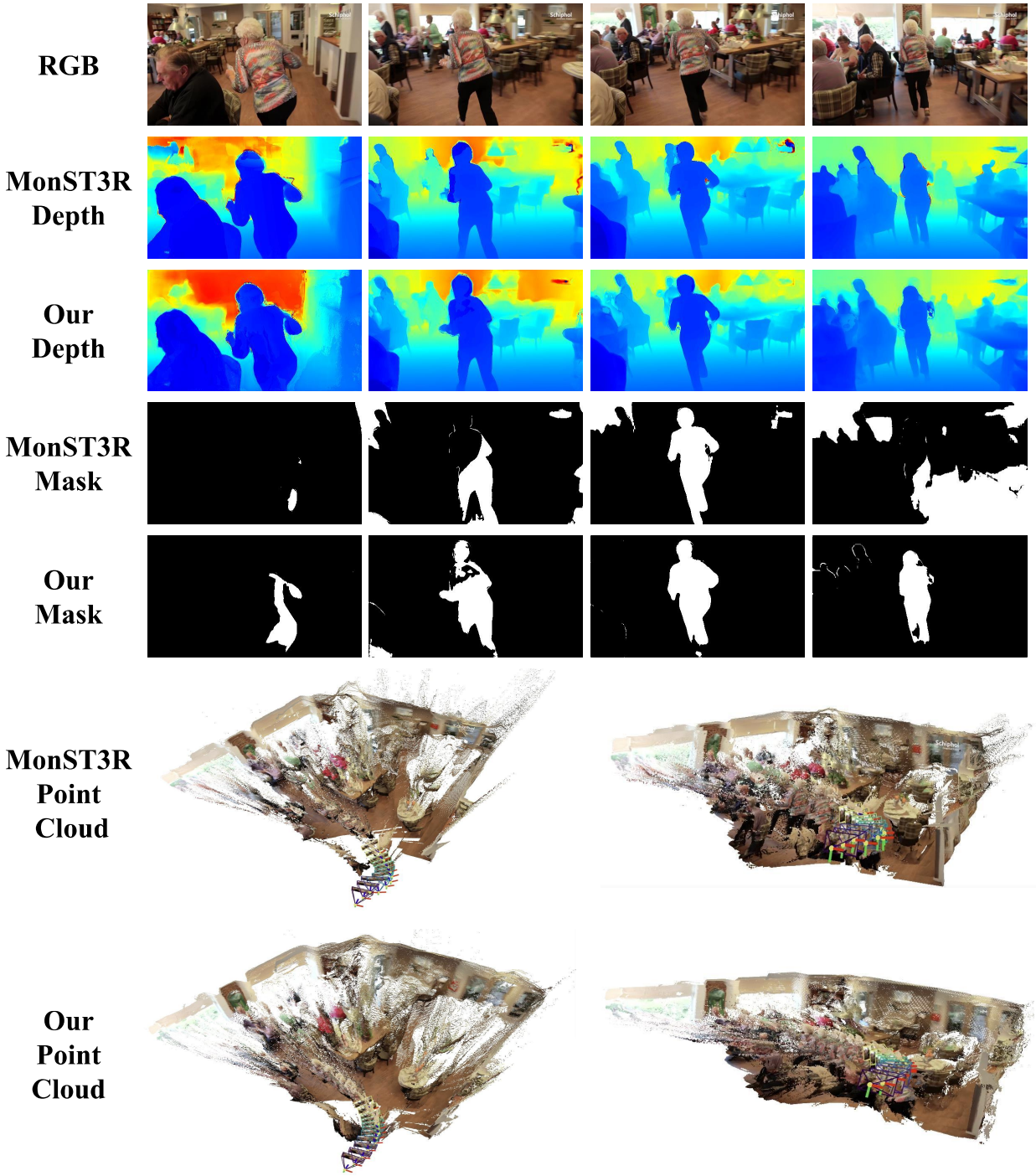

Figure 9. Compared with MonST3R, our POMATO can provide more complete dynamic masks and consistent geometry.

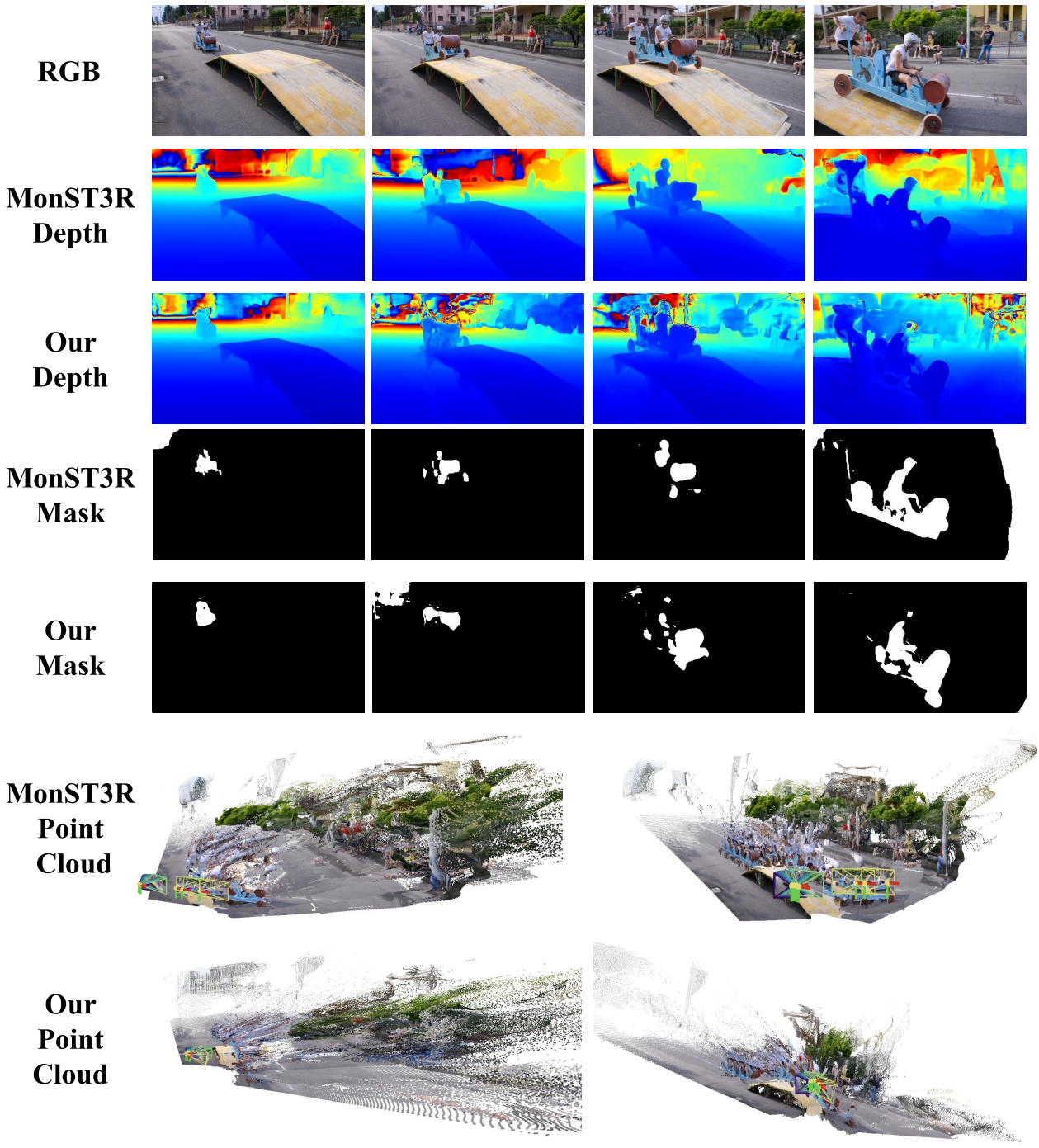

Figure 10. MonST3R suffers obvious degeneration when the view displacement is large as reflected by the erroneous pose estimation while POMATO can still provide a consistent camera trajectory.