

Supplementary Material

A Additional experimental details

In this section, we present detailed experimental setups for Section 4.

Baselines. We consider 7 different baseline point mapping models, DUST3R [2], MonST3R [3], MegaSaM [19], Align3R [16], Fast3R [22], Spann3R [20], and CUT3R [4]. We experiment with the checkpoint provided in the official open-source repository hosted by their authors, following the default image processing in each model, *e.g.*, the input dimensions are, the longer side length of 512 in DUST3R [2], MonST3R [3], Align3R [16], and Fast3R [22], the longer side length of 672, in MegaSaM [19], and, the square 256×256 in Spann3R [20].

Multi-frame Processing. Unless otherwise specified, we always choose the temporal window size of the inference $W = 6$ for evaluating our method and the baselines. We note that the pair-wise processing baselines are iteratively executed to match the required window size. To evaluate the feed-forward camera pose estimation in Section 4.3, we employ the weighted Procrustes solver to derive the relative rotation and translation between the frames, and the weighted least squares solver to estimate the camera intrinsic parameters, similar to the experimental configuration in CUT3R [4].

B Additional discussion

Although the pair-wise architecture [2, 3] can produce pointmaps for more than 2 frames by executing multiple pair-wise inferences, its design inevitably enforces the assumption that the distributions of consecutive pointmaps are independent. For example, given $\{\mathbf{I}^i, \mathbf{I}^j, \mathbf{I}^k\}$, a pair-wise model assumes that a joint density $\Pr(\mathbf{Y}^{i|j}, \mathbf{Y}^{i|k}, \mathbf{Y}^{j|k})$ is proportional to $\Pr(\mathbf{Y}^{i|j}) \cdot \Pr(\mathbf{Y}^{i|k}) \cdot \Pr(\mathbf{Y}^{j|k})$. However, in practice, including the scenarios represented by our evaluation, there exists an extreme case where \mathbf{I}^i and \mathbf{I}^k are completely non-overlapping, so that the pair-wise model assigns an erroneous estimate of $\Pr(\mathbf{Y}^{i|k})$, which can induce significant failure modes of estimating the joint density. Since Track3R can relax this constraint for multiple frames, it can learn the joint point mapping and trajectory prior that is more close to the true nature of the dynamic scenes.

C Raw point map examples

Within the supplementary material, we provide the raw point map predicted by Track3R, corresponding to the following results collected from the DAVIS dataset [5], depicted in Figures 4 to 7.

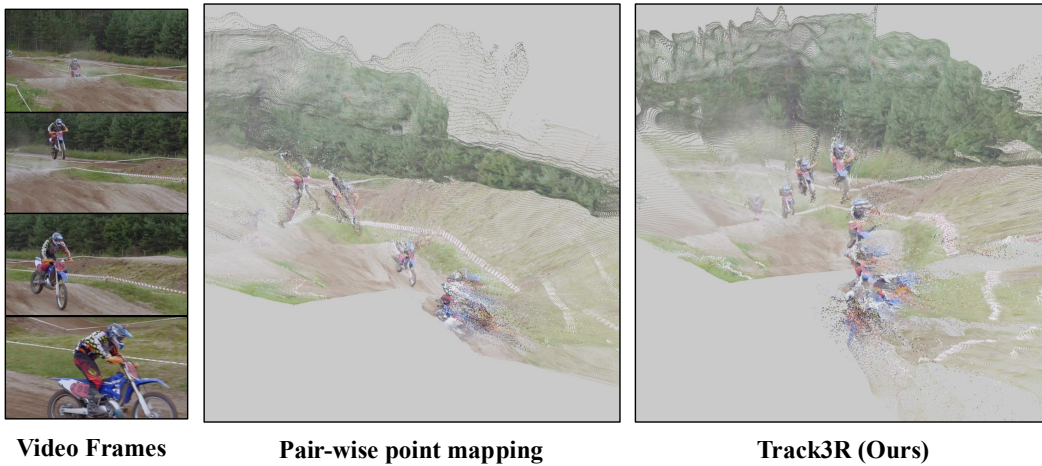


Figure 4: **Visualization of point map.** The point map predicted by the pair-wise point mapping baseline [3] and Track3R are depicted.

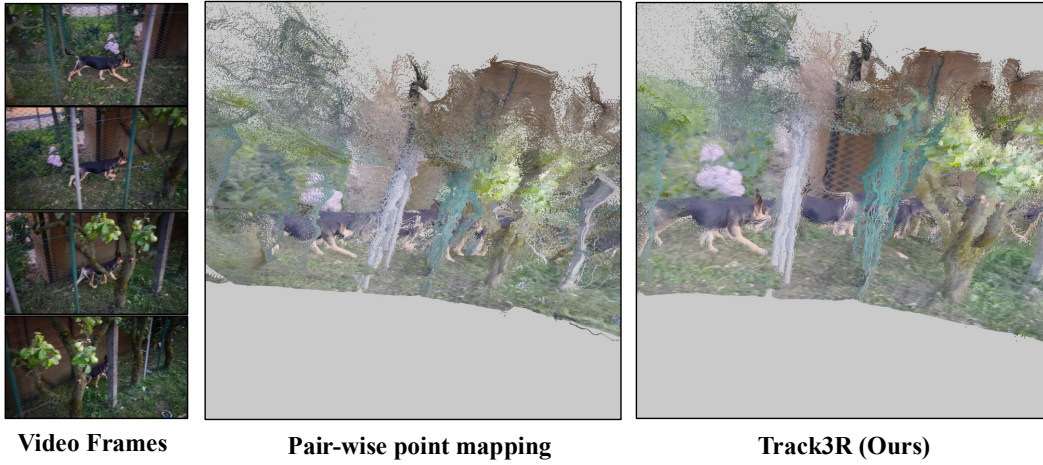


Figure 5: **Visualization of point map.** The point map predicted by the pair-wise point mapping baseline [3] and Track3R are depicted.

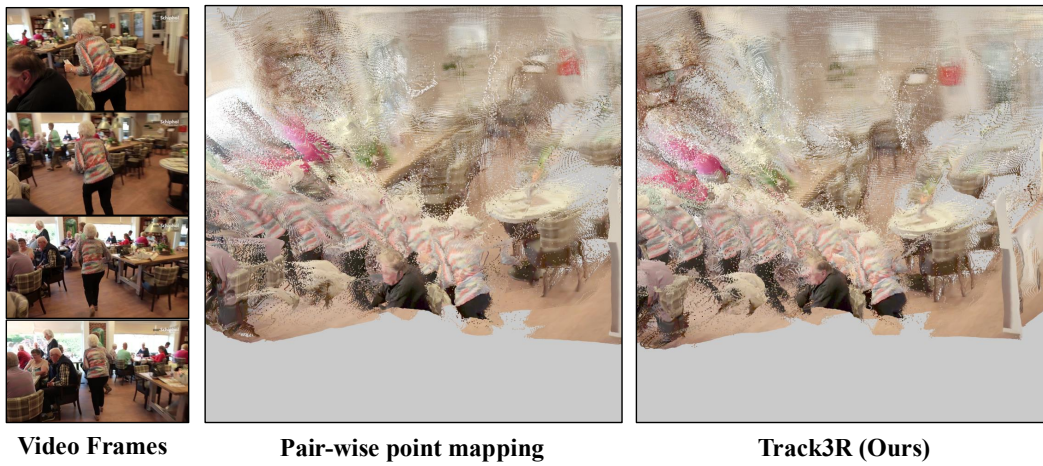


Figure 6: **Visualization of point map.** The point map predicted by the pair-wise point mapping baseline [3] and Track3R are depicted.

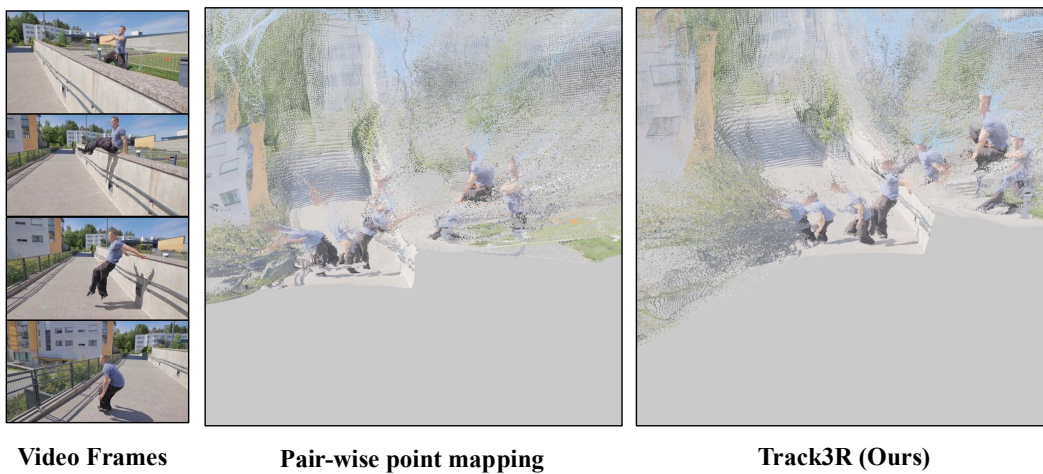


Figure 7: **Visualization of point map.** The point map predicted by the pair-wise point mapping baseline [3] and Track3R are depicted.