Synergistic Global-space Camera and Human Reconstruction from Videos

Supplementary Material

In this document, we provide additional technical details, more ablation studies, and more discussions. We refer the readers to the accompanying webpage for video results.

7. SynCHMR Setting vs. Prior Work

We compare the setup of recent world-frame HMR methods that handle dynamic cameras in Tab. 5. Methods that estimate world-frame body parameters through learning-based approaches often ignore the camera at test time [33, 52, 64]. On the other hand, optimization approaches need to estimate the camera at test time to fit to the detected 2D joint key points [15, 30, 36, 46, 62, 65], and we have discussed the downsides of their camera estimation approaches in Sec. 2 of the main paper. It is still worth noting that none of these methods reconstruct dense scene point clouds, except Liu *et al*. [36], who adopt COLMAP [47] for this purpose. However, since COLMAP is not robust enough for in-the-wild videos, they demonstrate results only on sequences acquired in a controlled capture settings. In stark contrast, SynCHMR is designed to work on casual videos. It does not assume the scene is a ground plane as in [30, 62] or is scanned *a priori* as in [13, 17]. It has a light-weight setup but it reconstructs the most information – human meshes, camera trajectory, and dense scene, all in one coherent global space.

8. Training Objectives for SMPL Denoiser

We consider a simple linear layer for each prediction head and parameterize $\mathbf{\Phi}$ and $\boldsymbol{\theta}$ predictions as quaternions. Specifically, $\mathcal{P}_{\mathbf{\Phi}} : \mathbb{R}^D \rightarrow \mathbb{R}^4$, $\mathcal{P}_{\boldsymbol{\theta}} : \mathbb{R}^D \rightarrow \mathbb{R}^{J\times 4}$, $\mathcal{P}_{\boldsymbol{\beta}} : \mathbb{R}^D \rightarrow \mathbb{R}^{10}$, and $\mathcal{P}_{\mathbf{\Gamma}} : \mathbb{R}^D \rightarrow \mathbb{R}^3$, where J denotes the number of joints. Then we apply direct supervision of SMPL parameters to the predictions

$$\begin{aligned}
\mathcal{L}_{\mathbf{\Phi}} &= 1 - \|\mathrm{q}(\mathbf{\Phi})\mathrm{q}(\mathbf{\Phi}^*)^\top\|_1, \\
\mathcal{L}_{\boldsymbol{\theta}} &= 1 - \|\mathrm{q}(\boldsymbol{\theta})\mathrm{q}(\boldsymbol{\theta}^*)^\top\|_1, \\
\mathcal{L}_{\boldsymbol{\beta}} &= \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1, \\
\mathcal{L}_{\mathbf{\Gamma}} &= \|\mathbf{\Gamma} - \mathbf{\Gamma}^*\|_1,
\end{aligned}$$

where $\mathrm{q}(\cdot)$ stands for the quaternion representations and superscript * denotes the ground truth. Following [9], we also introduce a discriminator $\mathcal{C}$ to ensure the per-frame predictions are valid

$$\mathcal{L}_{\mathcal{C}} = \|\mathbf{1} - \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\beta})\|_2^2.$$

The parameters are first factorized into (i) body pose parameters, (ii) shape parameters, and (iii) per-part relative rotations and classified by the discriminator to be fake (0) or real (1). To account for human motion, we further supervise the velocities and accelerations of human joints

$$\begin{aligned}
\mathcal{L}_{\dot{\mathbf{J}}} &= \left\| \|\dot{\mathbf{J}}\|_2 - \|\dot{\mathbf{J}}^*\|_2 \right\|_1, \\
\mathcal{L}_{\ddot{\mathbf{J}}} &= \left\| \|\ddot{\mathbf{J}}\|_2 - \|\ddot{\mathbf{J}}^*\|_2 \right\|_1,
\end{aligned}$$

where $\mathbf{J}$ are SMPL regressed joint locations.

9. SLAM Evaluation

Qualitative ablation study. In Tab. 3 of the main paper we quantitatively analyze the contribution of each design choice in our human-aware SLAM; here we provide visual examples. In Fig. 6, we show the results where we gradually add each design choice as stronger priors to the native visual SLAM. Merely using RGB inputs in Fig. 6(a), naive DROID-SLAM [54] fails in capturing the geometry structure of the scene. This results in a back-folded corridor, which is far from reasonable. The dynamic human also confuses the SLAM model, leading to a messy human point cloud in the center and everything else surrounding it in a circular shape. Masking out the human in Fig. 6(b) only removes the messy human point cloud but still produces a broken geometry since the depth ambiguity remains. An extra estimated depth channel in Fig. 6(c)(d) helps to resolve the depth ambiguity and correct the scene geometry. However, as we filter out points with epipolar inconsistency, the resulting point cloud is rather sparse. This indicates depth estimation with ZoeDepth [2] does not guarantee each point has a consistent location across different frames, and SLAM fails to correct this error. Finally, our Human-aware Metric SLAM in Fig. 6(e) is able to output a dense point cloud. This reflects the success in finding more points with consistent 3D locations. As the scene reconstruction depends on camera pose estimation in SLAM, our pipeline potentially produces more accurate camera poses.

Results on TUM-RGBD dataset. Tab. 3 of the main paper considers HMR datasets that provide ground truth camera trajectories. Here, we report the results on a common SLAM benchmark TUM-RGBD [51]. Since it does not contain humans in the scene, we can only apply our adapted video-consistent ZoeDepth [2], namely ZoeDepth$^+$, without calibrating the scales. In Tab. 6, we see that this depth-augmented version yields an average lower error than the original DROID-SLAM. This suggests that despite the unknown scale, estimated monocular depth still provides prior information to better reason about camera trajectories. One can see this as a byproduct of SynCHMR.

Methods	Test-time Camera Estimation	Test-time Scene Representation	World-frame SMPL Params. Estimation
Yu *et al*. [64]	no camera estimation	manually created shape primitives	RL-based
TRACE [52]	no camera estimation	no scene	feed-forward
D&D [33]	estimated acceleration and angular velocity	ground plane	feed-forward
Liu *et al*. [36]	COLMAP [47]	dense point cloud	optimization
GLAMR [65]	difference between the root transformations in the camera space and world space	ground plane	optimization
SmartMocap [46]	jointly solved with body params.; target: *only* body kpts.	no scene	optimization
BodySLAM [15]	jointly solved with body params. target: scene kpts and body kpts	no scene	optimization
PACE [30]		ground plane	optimization
SLAHMR [62]	DROID-SLAM [54] where humans are *not* excluded	ground plane	optimization
SynCHMR (ours)	human-aware metric SLAM (Sec. 3.2)	dense point cloud	scene-aware SMPL denoising (Sec. 3.3)

Table 5. **Comparison of methods that reconstruct humans in a global space from a video filmed by a dynamic camera.**

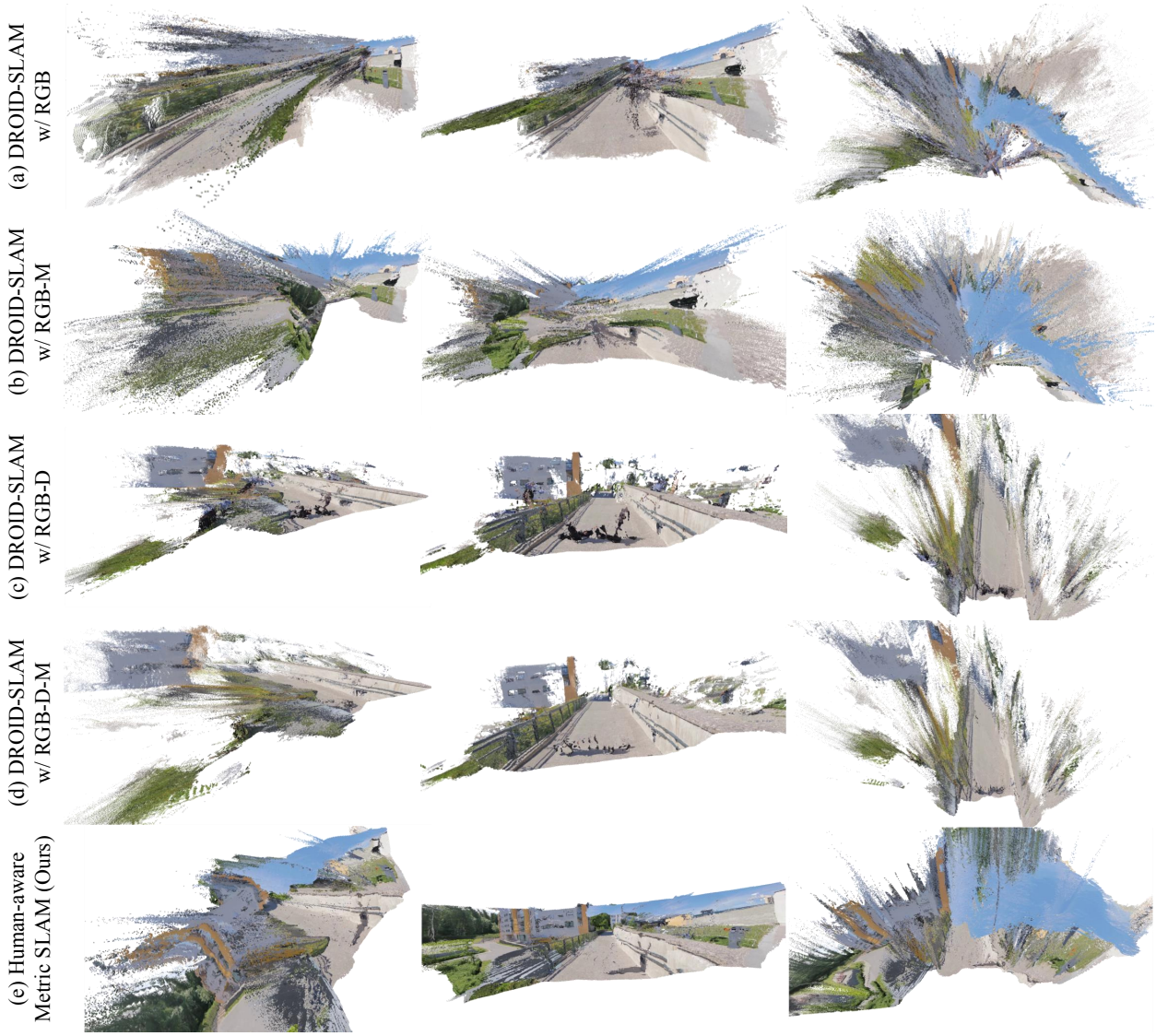

Figure 6. **Qualitative comparisons of the parkour sequence from DAVIS [42].** (a) naive DROID-SLAM [54] reconstructed point cloud with RGB input; (b) DROID-SLAM reconstructed point cloud with RGB input, where the foreground humans are masked out by an instance segmentation method Mask2Former [4]; (c) DROID-SLAM reconstructed point cloud with RGB-D input, where the depth channel is from ZoeDepth [2] estimations, the same below; (d) DROID-SLAM reconstructed point cloud with RGB-D and instance segmentation mask inputs (e) our proposed Human-aware Metric SLAM reconstructed point cloud. Please see the webpage for video results.

RGB	Depth	Mask	360	desk	desk2	floor	plant	room	rpy	teddy	xyz	avg
✓	✗	✗	162.3	75.1	682.8	54.2	257.7	930.5	40.4	480.0	16.4	340.2
✓	ZoeDepth$^+$	✗	101.3	153.9	75.6	817.4	219.4	96.3	32.6	201.2	21.8	223.4

Table 6. **Comparison between native DROID-SLAM (top) and our depth-augmented version (bottom) on TUM-RGBD [51].**

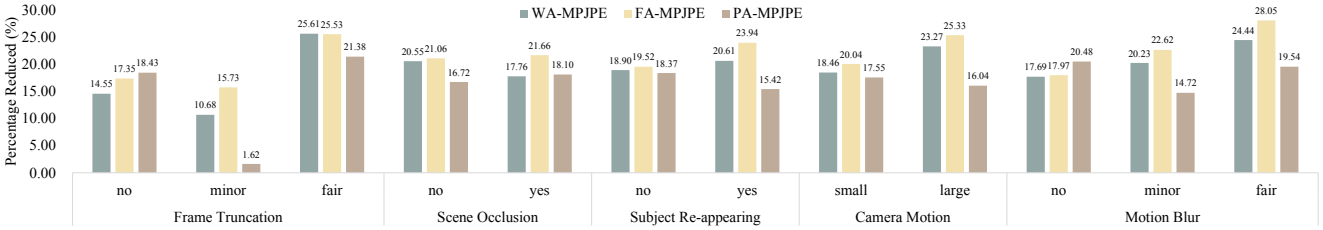

Figure 7. **Percentage of MPJPEs reduced by Scene-aware SMPL Denoiser in videos with varied attributes.** The larger the better (↑).

Figure 8. **SLAHMR and SynCHMR human meshes integrated with static scenes.** Video visualizations are included in the webpage.

10. HMR Evaluation

Qualitative comparison. In Fig. 8, we compare the estimated human body meshes and scene point clouds of (a) SLAHMR [62] and (b) our SynCHMR. We observe incompatible scales and structures in SLAHMR visualizations. This can be the reason why SLAHMR uses a ground plane instead of point clouds in the global refinement stage.

SMPL denoiser analysis. To better understand the impact of our scene-aware SMPL denoiser, we annotate the test set of EgoBody [69] with 5 attributes: frame truncation, scene occlusion, subject reappearing, camera motion, and motion blur. In Fig. 7, we plot the amount of error *reduced* by SMPL denoiser in these attributes. First, it confirms that the denoiser always brings improvement as there are no negative numbers. Second, we identify truncation, large camera motion, and motion blur as three primary scenarios where the denoiser helps greatly, as we see noticeable upward trends for them. The underlying mechanism might be our SMPL denoiser captures more comprehensive scene information with dynamic scene modeling, which is beneficial in these situations where single-frame observations are bad and one needs to rely on cross-frame clues.

Runtime analysis. We report the runtime of our SynCHMR along with state-of-the-art models in Tab. 2. Note that the runtime for PACE [30] does not include camera-frame initialization with HybrIK [32]. To integrate per-frame human bodies into a smooth motion, SLAHMR [62] employs a HuMoR-like motion prior, which is slow due to its auto-regressive nature. PACE [30] improves this by proposing a parallel motion prior. Similarly, while adding in scene awareness, our feed-forward SMPL Denoiser also benefits from the parallel inference of the Transformer architecture.