HumanBA: Human-Aware Bundle Adjustment via Global Human-Camera Decoupling

Supplementary Material

A. More Details on the Evaluation Datasets

Subset of SLOPER4D [3]. The publicly released SLOPER4D evaluation set contains six sequences. Among these, we select three for our experiments based on whether they trigger the bundle adjustment keyframe limit in DROID-SLAM [6] and DEVA tracking limit [2]. The selected sequences are:

> *seq008_running_001, seq007_garden_001, seq005_library_002.*

As reported in Tab. 5, we evaluate on this same subset with TRAM or fair comparison.

Subset of EMDB2 [4] with pronounced inter-frame motion spikes. To highlight the effectiveness of our method in rectifying SLAM failures (*e.g.*, jumping errors and motion spikes), we identify and evaluate on EMDB2 sequences that suffer from pronounced inter-frame motion spikes, as summarized in Tab. 2. A sequence is considered to contain a motion spike if the initial SLAM trajectory exhibits instantaneous camera velocity exceeding 10 m/s, which we treat as jitter-induced outlier motion. The sequences filtered under this criterion are:

> *P2/19_indoor_walk_off_mvs, P2/20_outdoor_walk, P3/27_indoor_walk_off_mvs, P3/30_outdoor_stairs_down, P4/35_indoor_walk, P4/37_outdoor_run_circle, P6/48_outdoor_walk_downhill, P6/49_outdoor_big_stairs_down, P9/78_outdoor_stairs_up_down.*

B. Additional Visualization on SLOPER4D

We provide additional qualitative visualizations of global human motion on the SLOPER4D dataset [3] to further illustrate the effectiveness of HumanBA. The examples below show that our approach produces more accurate and stable world-space human trajectories compared to TRAM [7].

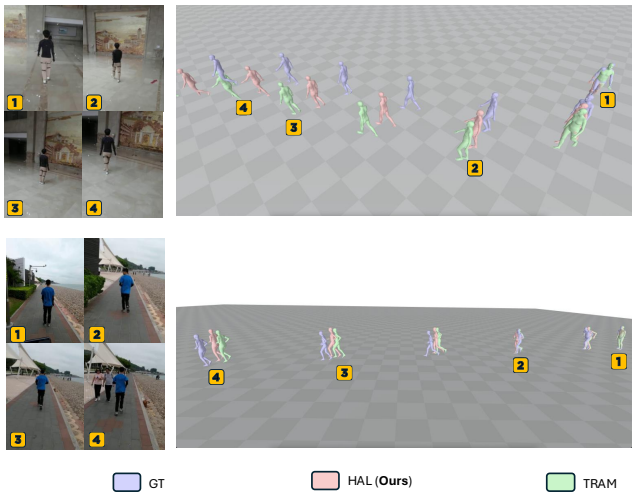

Figure 8. Additional qualitative results of global human motion on SLOPER4D dataset [3]. HumanBA recovers more accurate and stable world-space human trajectories compared to TRAM [7].

C. Discussion

In this section, we provide additional discussion about our method as follows:

1) Why not use the full SMPL mesh as BA landmarks? The SMPL [5] model contains 6890 vertices, whereas DROID-SLAM [6] operates on a 64×48 internal feature map. Using all mesh vertices as BA landmarks is therefore problematic: many vertices—including points on the front and back of the human body that lie at very different depths—would project onto the same feature-map cell, leading to ambiguity and degraded optimization. As shown in Sec. 5.2 and Tab. 1, more landmarks do not necessarily lead to better BA performance; what matters is the reliability of the constraints. Hence, we use only the 24 body joints as landmarks and apply our adaptive, motion-aware confidence weighting (Sec. 4.3), which proves effective in experiments (Sec. 5).

2) Differences from BA-Track [1]. BA-Track and HumanBA share the intuition that dynamic foreground motion is not inherently harmful to SLAM. However, BA-Track is object-agnostic and relies on generic dynamic 3D point tracking, which is computationally heavy and less suitable for long human-centric videos. HumanBA instead exploits human-specific priors through a parametric body model, making it substantially lighter and more scalable. Moreover, HumanBA is designed to let global human motion and global camera motion reinforce each other inside BA.

Since BA-Track is not human-centric and does not report human metrics, we compare camera accuracy only. As shown in Tab. 7, BA-Track performs worse on highly dynamic human videos, while HumanBA benefits from structured kinematic priors that provide more stable constraints.

Method	indoor_walk_off_mvs		outdoor_stairs_up	
	ATE-S↓	ATE↓	ATE-S↓	ATE↓
BA-Track	1.72	0.89	3.52	0.27
Masked DROID + HumanBA **(ours)**	**0.36**	**0.18**	**0.60**	**0.13**

Table 7. Comparison with BA-Track on camera metrics.

3) Robustness to noisy HMR estimates. HumanBA is designed to be robust to imperfect HMR predictions rather than assuming accurate human pose input. We treat human-derived anchors as soft constraints with adaptive confidence, instead of trusting all human observations equally. As shown in Fig. 9, HumanBA still provides gains under varying HMR quality, while removing the adaptive weighting leads to more noticeable degradation. This shows that the benefit of HumanBA comes from selectively using reliable human cues, not from assuming perfect HMR.

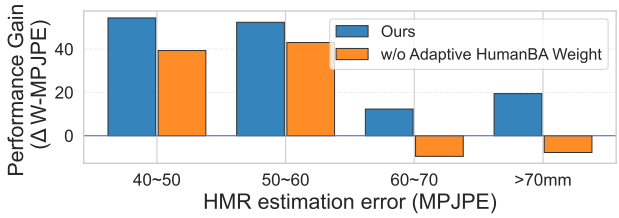

Figure 9. HumanBA remains effective under different HMR accuracy levels.

4) Why are the improvements on camera poses sometimes limited? HumanBA is not intended to replace dense SLAM correspondences, but to complement them in challenging cases. When the background is sufficiently textured and masking is accurate, the SLAM backend already has enough static evidence, leaving less room for improvement. In contrast, HumanBA is more helpful when masking removes many observations or when the human occupies a large fraction of the image. Fig. 10 shows that the gain becomes more pronounced as human foreground occupancy increases.

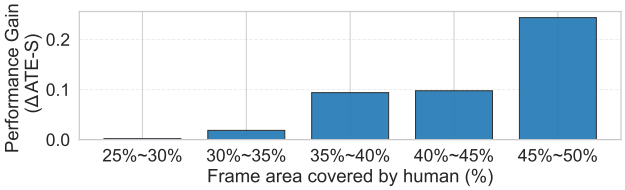

Figure 10. HumanBA helps more when the human occupies a larger image area.

5) Detailed runtime. HumanBA is lightweight in practice. As shown in Tab. 8, our method takes 9.7 minutes per 1000 frames, including 1.5 minutes for initialization and 8.2 minutes for HumanBA itself. This is substantially faster than SLAHMR [8] and also much more efficient than BA-Track.

	SLAHMR	BA-Track	HumanBA **(ours)**
Runtime / 1000 frames	402 min	>33 min	9.7 min
Breakdown	–	BA: 33 min	Init: 1.5 min; HumanBA: 8.2 min

Table 8. Runtime comparison per 1000 frames.

6) Behavior on long and ultra-long sequences. HumanBA successfully processes all EMDB videos in our experiments, which contain up to roughly 2000 frames. The few failures we observed only arise on ultra-long SLOPER4D sequences exceeding 6000 frames, and are caused by the keyframe-budget limit of the SLAM backend rather than HumanBA itself. In principle, HumanBA can be combined with sliding-window BA, hierarchical pose graphs, or more lightweight backends without changing the core idea.

7) Why not jointly optimize human pose (or human anchors) inside BA? We also tried jointly optimizing the human anchor, but it slightly degrades both camera and human metrics, as shown in Tab. 9. This is likely because camera and human variables begin to compensate for each other, weakening the constraints for camera refinement. To avoid this, HumanBA keeps the human estimates decoupled and instead uses adaptive weighting to suppress unreliable human constraints.

Method	ATE-S↓	ATE↓	W-MPJPE↓	WA-MPJPE↓
Jointly optimize human anchor	0.739	0.384	199.05	70.64
Masked DROID + HumanBA **(ours)**	**0.682**	**0.358**	**195.97**	**70.10**

Table 9. Effect of jointly optimizing human anchors.

References

[1] Weirong Chen, Ganlin Zhang, Felix Wimbauer, Rui Wang, Nikita Araslanov, Andrea Vedaldi, and Daniel Cremers. Back on track: Bundle adjustment for dynamic scene reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 1

[2] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[3] Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 682–692, 2023. 1

[4] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 1

[6] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021. 1

[7] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision (ECCV)*, pages 467–487. Springer, 2024. 1

[8] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21222–21232, 2023. 2