Reconstructing People, Places, and Cameras

Supplementary Material

This is the supplementary material for our main paper "Reconstructing People, Places, and Cameras". We provide additional qualitative results (Sec. S.1), including in-the-wild examples, a discussion of our approach's limitations (Sec. S.2), ablation studies on the scale initialization and input views (Sec. S.3), evaluation details (Sec. S.4), implementation detail (Sec. S.5), and additional related work (Sec. S.6).

S.1. Additional Qualitative Results

In-the-wild demo. We first present HSfM's reconstruction results on images captured by two cell phones in Figures S.1 and S.2. The data was captured using a minimal setup consisting of two cell phones, two tripods, and the Riverside app[1] for straightforward time synchronization. Despite this simple setup, our multi-view optimization algorithm successfully handles challenging scenes, such as individuals jumping, without relying on any heuristic contact priors and small data-driven motion priors that previous works [44, 61, 69] use.

Benchmark evaluation. We provide additional qualitative results on EgoExo4D [17] in Figure S.4, showing challenging scenes such as kitchens, humans interacting with objects (e.g., playing the piano), and sports activities like soccer. In Figure S.7, we display further results on EgoHumans [24], demonstrating HSfM reconstructions of multiple people interacting, such as fencing. Figure S.3 show a scene from EgoHumans before and after HSfM optimization.

S.2. Discussion

Our goal is to study the mutual benefits of jointly reconstructing humans, scenes, and cameras. To this end, we assume that the re-identification of people across camera views is known, because misidentified individuals can introduce spurious effects, disrupting the optimization process. Please note that this limitation also applies to UnCaliPose [57]. To ensure a fair comparison and maintain focus on the core objectives of this study, we rely on ground-truth identities in both our approach and UnCaliPose in the main text.

Since re-identification may not be available at test time and manual identification is cumbersome, we tested the feasibility of automating the re-identification process using the re-identification module of UnCaliPose on the EgoHumans dataset [24]. For re-identification, UnCaliPose solves a constrained clustering optimization problem, assuming a known number of people in the scene and utilizing re-identification features extracted by an off-the-shelf re-identification network [31]. The re-identification process achieves an accuracy of 51.22% on EgoHumans. The main failure mode occurs with individuals wearing uniforms (e.g., tennis sequences: 12.04% accuracy, volleyball sequences: 25.71% accuracy), where appearance features are difficult to distinguish.

These findings indicate that manual re-identification remains necessary for accurate multi-view reconstruction of the world, including humans. Fortunately, modern tools like LabelMe[2] simplify this process. Looking ahead, we anticipate that ongoing advancements in large-scale data-model paradigms will significantly improve performance in multi-view re-identification. These advancements may include robust appearance feature matching [34] and the use of geometric similarities, such as human pose and location [16, 37].

While our method achieves good quantitative results, we observe a few failure cases stemming from preprocessing errors in reconstruction and detections. The most common issues arise from erroneous initial camera estimates generated by the scene reconstruction-based SfM[53], particularly in scenes with limited structure, insufficient overlap between images, or large areas affected by radial distortion. In instances where DUSt3R[53] fails to detect any cameras, we rely on human-centric camera poses to initialize the optimization. Another source of error involves missing or highly inaccurate keypoint detections, which can occur under conditions of heavy occlusion or poor lighting. In such cases, our method estimates frame-specific cameras solely based on pixel data, without incorporating human constraints. Despite these occasional errors, we find DUSt3R, ViTPose [59], and HMR2.0 [16] to exhibit remarkable robustness across a wide range of challenging scenarios.

S.3. Additional Ablation Studies

We analyze the effect of different scale initializations to validate the superiority of the human-centric scaling introduced in Section 4.1 in Table S.1. Without scale initialization ($\alpha = 1.0$), where we directly use the raw DUSt3R [53] scene and camera pose outputs as input to our optimization, the W-MPJPE is 11.89m, whereas ours is 1.04m. Additionally, the high metric-scale camera translation errors, such as 6.42m TE, and extremely low RRA values, demonstrate the necessity of proper initialization. This error occurs be-

[1] https://riverside.fm/

[2] https://github.com/wkentaro/labelme

	Human Metrics						Camera Metrics					
	W-MPJPE↓	GA-MPJPE↓	PA-MPJPE↓	TE↓	s-TE↓	AE↓	RRA@10↑	RRA@15↑	RTA@10↑	RTA@15↑	s-RTA@10↑	s-RTA@15↑
S1: $\alpha = 1.0$	11.89	0.85	0.09	6.42	5.50	121.88	0.01	0.01	0.01	0.02	0.01	0.05
S2: $\alpha = 100.0$	1.94	0.22	0.06	2.17	1.11	15.00	0.68	0.82	0.31	0.45	0.68	0.83
S3: HSfM (Ours)	**1.04**	**0.21**	**0.05**	**2.09**	**0.75**	**9.35**	**0.72**	**0.89**	**0.32**	**0.46**	**0.75**	**0.91**
2 Cam. HSfM (init)	3.73	0.42	0.06	1.53	-	9.81	0.41	0.87	0.08	0.12	-	-
2 Cam. HSfM (Ours)	2.63	0.26	**0.05**	**0.39**	-	10.37	0.41	**0.91**	**0.48**	**0.68**	-	-
4 Cam. HSfM (init)	4.26	0.51	0.06	2.36	1.14	10.96	0.52	0.79	0.26	0.38	0.49	0.74
4 Cam. HSfM (Ours)	1.15	0.27	0.06	2.00	**0.71**	8.92	0.68	0.88	0.35	0.50	**0.78**	**0.93**
8 Cam. HSfM (init)	5.06	0.53	0.06	2.36	0.96	7.61	0.71	0.87	0.25	0.40	0.65	0.88
8 Cam. HSfM (Ours)	**1.00**	**0.19**	**0.05**	1.97	0.90	**7.41**	**0.76**	0.90	0.41	0.54	0.72	0.88

Table S.1. **Ablation on the number of input view cameras.** We evaluate the performance of HSfM by varying the number of input view cameras and assessing human reconstruction and camera pose estimation in the world coordinate frame. The experiments are conducted on EgoHumans, excluding samples without ground truth camera poses for all views in the specified combinations (2, 4, and 8). Compared to the initialization, our joint optimization improves all human pose and camera pose metrics, regardless of the number of input cameras. We do not report the scaled version of camera translation errors for the 2-camera cases, as the predictions become identical to the ground truth camera translations after scale alignment.

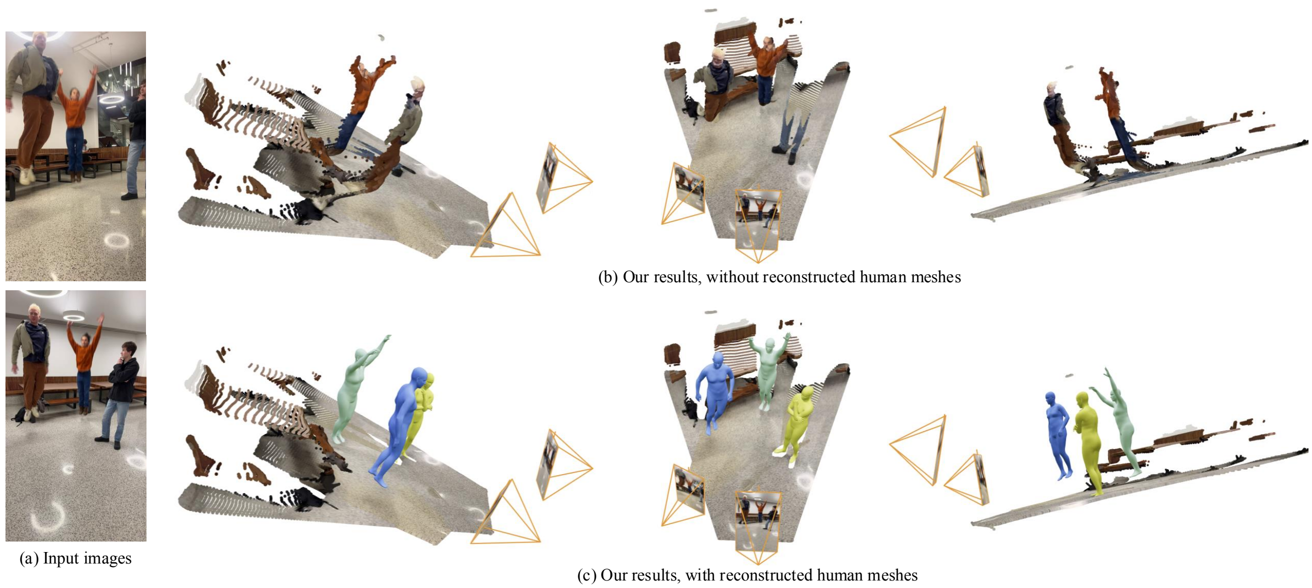


Figure S.1. **Qualitative results *in the wild*.** We show reconstructions on *in-the-wild* images taken with two smartphones (a), demonstrating the reconstruction of humans and scenes. Unlike previous works [61, 71], which adopt human-scene contact priors that hinder generalization to scenarios without ground foot contact, HSfM recovers accurate world locations of the human meshes that are coherent with the static scene structure. The use of humans in our framework (c) not only serves as a reliable initialization for 3D structure in the SfM formulation but also provides more faithful and complete information about people in the world, which a noisy human point cloud (b) cannot offer. For visualization purposes, the human point cloud is removed using SAM2 [38].

cause the raw camera and scene outputs have a significantly smaller scale than the real world due to their scale normalization during training.

Choosing a large scale value ($\alpha = 100$) generally covers the real-world capture scene sufficiently but does not perform as well as our human-centric scaling approach (W-MPJPE: 1.94m vs. 1.04m). The camera metrics are also worse than ours (e.g., RRA values are 5–7% lower than ours). This implies that, without proper scaling, the optimization is prone to failure due to poor initialization and local minima problems.

One common local minimum observed was humans being placed behind the camera while still reprojecting to the correct pixel locations. To address this, we increased the scale α until all humans were placed in front of all cameras, ensuring positive depth values in all camera coordinate systems. While this initialization produced similar quantitative results in successful cases, it completely failed for 2% of

Figure S.2. **Qualitative results *in the wild*.** We show reconstructions on *in-the-wild* images taken with two cell phones and the reconstruction of humans and scene. Our method places people in the world and reconstructs accurate human-scene contact, *e.g*. between the person's right foot and box.

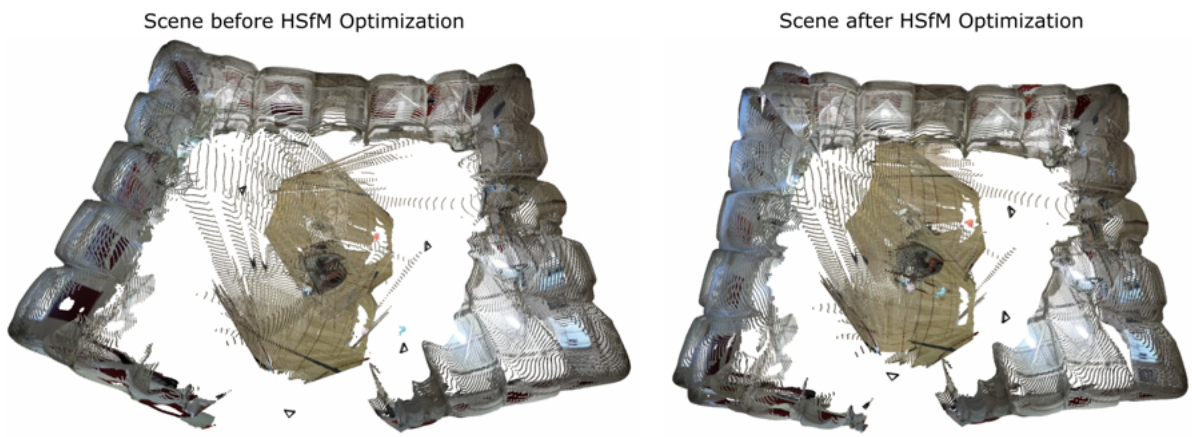


Figure S.3. **Qualitative result of HSfM reconstruction.** Top view of scene from EgoHumans before and after HSfM optimization.

samples, demonstrating that naive initialization approaches are not reliable.

Next, we vary the number of cameras to evaluate the robustness of our method. We tested 2, 4, and 8 input view cameras. As indicated in the table, our joint optimization consistently improves all metrics, regardless of the number of input view cameras. With only 2 cameras, W-MPJPE is 2.63m and GA-MPJPE is 0.26m, indicating accurate human placement in the world. The consistently better camera results compared to the initialization further validate the benefit of incorporating humans into the traditional SfM formulation. The robustness of our method is further demonstrated qualitatively in Figure S.2, where the data is captured by two cameras in in-the-wild scenes.

S.4. Evaluation Details

S.4.1. Evaluation Metrics

In this section, we provide additional details about our human pose and camera metrics.

W-MPJPE describes the mean per-joint position error, measured in the world frame. To bring predicted human meshes into the ground-truth's world coordinate system, we use an SE(3) rigid alignment from the estimated camera positions to the ground-truth camera positions.

PA-MPJPE describes the Procrustes-aligned variant of MPJPE, which measures position errors after Sim(3) alignment of joints for each human. This metric evaluates local pose accuracy in a way that is not dependent on camera position estimates or human body scale.

GA-MPJPE evaluates group-aligned joint position errors, computed after Sim(3) alignment for all humans in a scene. This measures people relative to each other, without considering the scene or camera positions.

TE measures the mean Euclidean distance between predicted and ground truth camera positions, after SE(3) alignment. TE evaluates metric accuracy of camera positions.

s-TE is the scale-aligned version of TE, where we preprocess positions with Sim(3) instead of SE(3) alignment. This measures scale-invariant errors for estimated cameras.

AE measures the average Angle Error between camera pairs. We compute relative orientations for each pair of cameras in a scene. We then measure the difference between ground-truth and predicted pairwise orientations, convert to degrees, and average.

CCA [27] measures the Camera Center Accuracy, after the SE(3) alignment process used for TE. CCA@τ is the proportion of camera positions with absolute error within τ% of the overall scene scale. Following existing work, we compute the scene scale as the furthest distance between a ground-truth camera and the centroid.

s-RTA measures the the scale-aligned version of RTA, after the Sim(3) alignment process used for s-TE.

RRA [52] measures the Relative Rotation Accuracy of camera estimates, computed using the same camera pairs as AE. RRA@τ is the proportion of pairwise camera orientations with angular error of τ degrees or lower.

S.4.2. Evaluation Datasets

EgoHumans: In the main text's tables and Table S.1's 4 view case, we used the following camera configurations for each sequence:

- For `01_tagging` sequences: `camera 1`, `camera 4`, `camera 6`, and `camera 8`.
- For `02_lego` sequences: `camera 2`, `camera 3`, `camera 4`, and `camera 6`.
- For `03_fencing` sequences: `camera 4`, `camera 5`, `camera 10`, and `camera 13`.
- For `04_basketball` sequences: `camera 1`, `camera 3`, `camera 4`, and `camera 8`.
- For `05_volleyball` sequences: `camera 2`, `camera 4`, `camera 8`, and `camera 11`.

- For 06_badminton sequences: camera 1, camera 2, camera 5, and camera 7.
- For 07_tennis sequences: camera 4, camera 9, camera 12, and camera 20.

In Table S.1's 2 view case, we used the following camera configurations for each sequence:

- For 01_tagging sequences: camera 1 and camera 2.
- For 02_lego sequences: camera 3 and camera 5.
- For 03_fencing sequences: camera 5 and camera 13.
- For 04_basketball sequences: camera 2 and camera 7.
- For 05_volleyball sequences: camera 6 and camera 12.
- For 06_badminton sequences: camera 5 and camera 7.
- For 07_tennis sequences: camera 9 and camera 12.

In Table S.1's 8 view case, we used the following camera configurations for each sequence:

- For 01_tagging sequences: all 8 available cameras.
- For 02_lego sequences: all 8 available cameras.
- For 03_fencing sequences: camera 1, camera 3, camera 5, camera 7, camera 9, camera 11, camera 13, camera 15.
- For 04_basketball sequences: all 8 available cameras.
- For 05_volleyball sequences: camera 1, camera 3, camera 5, camera 7, camera 9, camera 11, camera 13, camera 15.
- For 06_badminton sequences: camera 1, camera 3, camera 5, camera 7, camera 9, camera 11, camera 13, camera 15.
- For 07_tennis sequences: camera 1, camera 3, camera 5, camera 7, camera 9, camera 11, camera 13, camera 15.

EgoExo4D: EgoExo4D scenes are typically captured using four to six RGB cameras and an egocentric device (Aria glasses). For our experiments and the baselines, we use only the RGB images from sequences with correct re-identification. Sequences containing ego-centric RGB views, such as helmet-mounted cameras, are excluded. We evaluate 182 videos from the validation set, sampling one random frame per video. The videos include ground-truth annotations for human poses, locations, and camera parameters. We evaluate on the following takes/frames:
cmu_soccer06_3/1426
cmu_soccer12_2/6807
cmu_soccer16_2/6373
georgiatech_bike_06_12/170
georgiatech_bike_06_2/97
georgiatech_bike_06_6/74
georgiatech_bike_06_8/15
georgiatech_bike_07_10/28
georgiatech_bike_07_12/38
georgiatech_bike_07_2/97
georgiatech_bike_07_4/46
georgiatech_bike_07_6/67
georgiatech_bike_07_8/138
georgiatech_bike_14_12/593
georgiatech_bike_14_2/1214
georgiatech_bike_14_6/575
georgiatech_bike_14_8/97
georgiatech_bike_15_2/1508
georgiatech_bike_15_4/844
georgiatech_bike_15_6/1103
georgiatech_bike_15_8/3153
georgiatech_bike_16_2/882
georgiatech_bike_16_6/3031
georgiatech_bike_16_8/1274
georgiatech_covid_02_10/2227
georgiatech_covid_02_12/6974
georgiatech_covid_02_14/2926
georgiatech_covid_02_2/67
georgiatech_covid_02_4/67
georgiatech_covid_04_10/999
georgiatech_covid_04_12/6160
georgiatech_covid_04_4/2996
georgiatech_covid_04_6/4528
georgiatech_covid_06_2/47
georgiatech_covid_06_4/64
georgiatech_covid_18_10/5524
georgiatech_covid_18_12/3457
georgiatech_covid_18_2/2413
georgiatech_covid_18_4/3534
georgiatech_covid_18_6/4389
georgiatech_covid_18_8/458
iiith_cooking_59_2/7795
iiith_cooking_64_2/298
iiith_cooking_89_6/1177
iiith_cooking_90_4/1383
iiith_soccer_015_2/1610
nus_cpr_12_1/1338
nus_cpr_12_2/76
sfu_basketball012_10/774
sfu_basketball012_12/399
sfu_basketball012_2/945
sfu_basketball012_3/1506
sfu_basketball012_4/66
sfu_basketball012_6/526
sfu_basketball012_7/1581
sfu_basketball012_8/329
sfu_basketball016_2/247
sfu_basketball_04_8/209
sfu_basketball_05_22/1902

sfu_basketball_05_26/29
sfu_basketball_09_11/32
sfu_basketball_09_12/1114
sfu_cooking028_12/1049
sfu_cooking_007_7/77
sfu_cooking_008_3/4164
sfu_cooking_008_5/3559
sfu_covid_004_2/2828
sfu_covid_004_4/5360
sfu_covid_008_16/1595
unc_basketball_02-24-23_01_3/84
unc_basketball_02-24-23_02_10/466
unc_basketball_02-24-23_02_11/927
unc_basketball_03-30-23_02_10/45
unc_basketball_03-30-23_02_14/7
unc_basketball_03-30-23_02_15/40
unc_basketball_03-30-23_02_17/9
unc_basketball_03-30-23_02_18/20
unc_basketball_03-30-23_02_19/7
unc_basketball_03-30-23_02_4/107
unc_basketball_03-30-23_02_5/25
unc_basketball_03-30-23_02_7/1141
uniandes_basketball_001_23/768
uniandes_basketball_001_24/1386
uniandes_basketball_001_26/146
uniandes_basketball_001_27/439
uniandes_basketball_003_38/32
uniandes_basketball_004_23/369
uniandes_basketball_004_44/261
uniandes_basketball_004_45/667
uniandes_dance_002_11/201
uniandes_dance_002_2/439
uniandes_dance_008_29/276
uniandes_dance_008_30/166
uniandes_dance_008_31/31
uniandes_dance_008_32/11
uniandes_dance_008_33/1105
uniandes_dance_008_34/753
uniandes_dance_008_35/607
uniandes_dance_008_36/1045
uniandes_dance_008_37/913
uniandes_dance_008_38/706
uniandes_dance_016_10/841
uniandes_dance_016_11/279
uniandes_dance_016_12/932
uniandes_dance_016_13/453
uniandes_dance_016_14/951
uniandes_dance_016_30/577
uniandes_dance_016_31/1709
uniandes_dance_016_32/377
uniandes_dance_016_33/1158
uniandes_dance_016_36/1247
uniandes_dance_016_37/145
uniandes_dance_016_38/1416
uniandes_dance_016_39/399
uniandes_dance_016_3/1239
uniandes_dance_016_42/1406
uniandes_dance_016_43/1271
uniandes_dance_016_44/1268
uniandes_dance_016_45/838
uniandes_dance_016_6/1361
uniandes_dance_016_7/1040
uniandes_dance_016_8/1488
uniandes_dance_017_6/1592
uniandes_dance_019_17/1003
uniandes_dance_019_18/509
uniandes_dance_019_19/1537
uniandes_dance_019_20/1089
uniandes_dance_019_22/81
uniandes_dance_019_24/484
uniandes_dance_019_25/183
uniandes_dance_019_26/1814
uniandes_dance_019_27/283
uniandes_dance_019_28/1411
uniandes_dance_019_46/412
uniandes_dance_019_47/790
uniandes_dance_019_49/1617
uniandes_dance_019_51/481
uniandes_dance_019_52/875
uniandes_dance_019_54/766
uniandes_dance_019_55/679
uniandes_dance_019_56/561
uniandes_dance_019_57/1073
uniandes_dance_019_58/192
uniandes_dance_024_11/1619
uniandes_dance_024_12/104
uniandes_dance_024_13/1419
uniandes_dance_024_14/1180
uniandes_dance_024_15/378
uniandes_dance_024_16/1569
uniandes_dance_024_17/1317
uniandes_dance_024_45/844
uniandes_dance_024_47/732
uniandes_dance_024_48/261
uniandes_dance_024_49/325
upenn_0706_Dance_4_2/2512
upenn_0706_Dance_4_3/1277
upenn_0706_Dance_4_4/1670
upenn_0706_Dance_4_5/1904
upenn_0713_Dance_3_2/164
upenn_0713_Dance_3_3/586
upenn_0713_Dance_3_4/0
upenn_0713_Dance_3_5/243
upenn_0713_Dance_4_2/125
upenn_0713_Dance_4_3/1280
upenn_0713_Dance_4_4/308

```
upenn_0713_Dance_4_5/262
upenn_0713_Dance_5_4/238
upenn_0713_Dance_5_6/2534
upenn_0721_Piano_1_2/140
upenn_0721_Piano_1_3/648
upenn_0722_Piano_1_2/83
upenn_0727_Partner_Dance_3_1_2/62
utokyo_pcr_2001_29_2/5799
utokyo_pcr_2001_29_4/3491
utokyo_pcr_2001_29_6/550
utokyo_pcr_2001_30_2/2121
utokyo_pcr_2001_30_4/1696
utokyo_pcr_2001_32_2/6641
utokyo_pcr_2001_32_4/6048
utokyo_soccer_8000_43_2/3262
utokyo_soccer_8000_43_4/3472
utokyo_soccer_8000_43_6/2781
```

S.5. Implementation Details

Given sparse-view images, HSfM jointly estimates SMPL-X [35] parameters for humans, scene pointmaps, and camera poses (rotation and translation), in the world coordinate frame. The SMPL-X parameters for humans are initialized using predictions from HMR2 [16] converted to SMPL-X following the conversion procedure in [33]. Scene pointmaps and camera parameters are initialized with estimates from DUSt3R [53]. We use Adam [25] optimizer and set the number of optimization steps proportional to the scene scale with a minimum of 500 steps. This allows sufficient time to accurately determine scene scale and camera poses and people's location. The learning rate is set to 0.015 with a linear reduction schedule. To tune hyperparameters, we use the first frame (4 cameras) of sequences `01_tagging` and `04_basketball` of EgoHumans as these scene encompass a good range of scene scales.

S.6. Additional related work

Monocular Human Mesh Reconstruction. Most methods estimate 3D humans in the camera coordinate system for a single person [14, 16, 23, 35] or for multiple people with depth estimation [2, 46, 67]. Recent works jointly estimate human and camera motion in the world coordinate frame [29, 43, 44, 47, 54, 61, 65, 69]. They leverage temporal dynamics from video sequences to improve reconstruction quality over time. While single-view reconstruction methods are valuable for their minimal input requirements, they often suffer from ambiguity, especially due to occlusion. Our approach leverages multi-view data to enhance reconstruction accuracy and integrates scene context, providing a more detailed and reliable reconstruction of multi-person interactions within their environment.

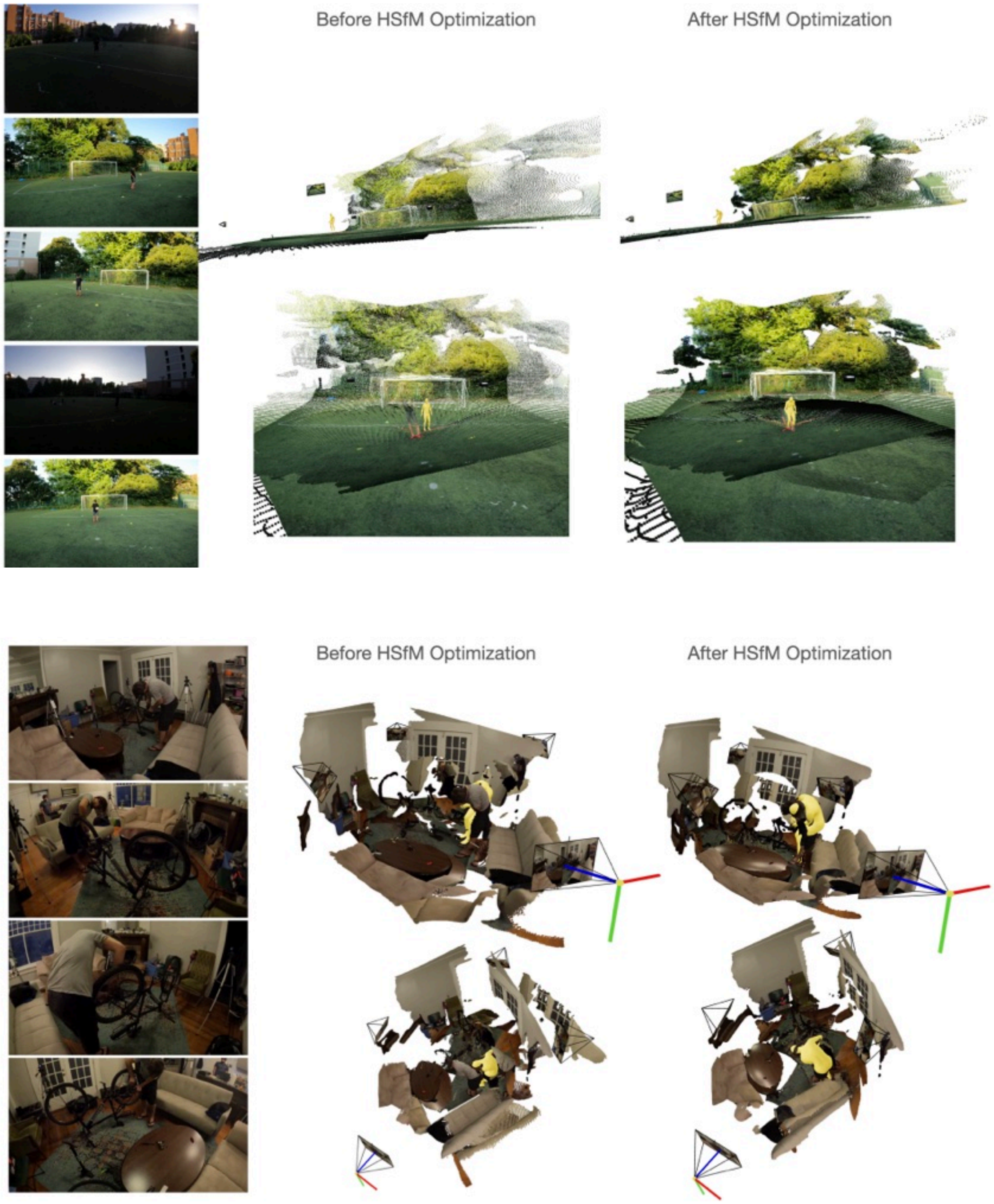
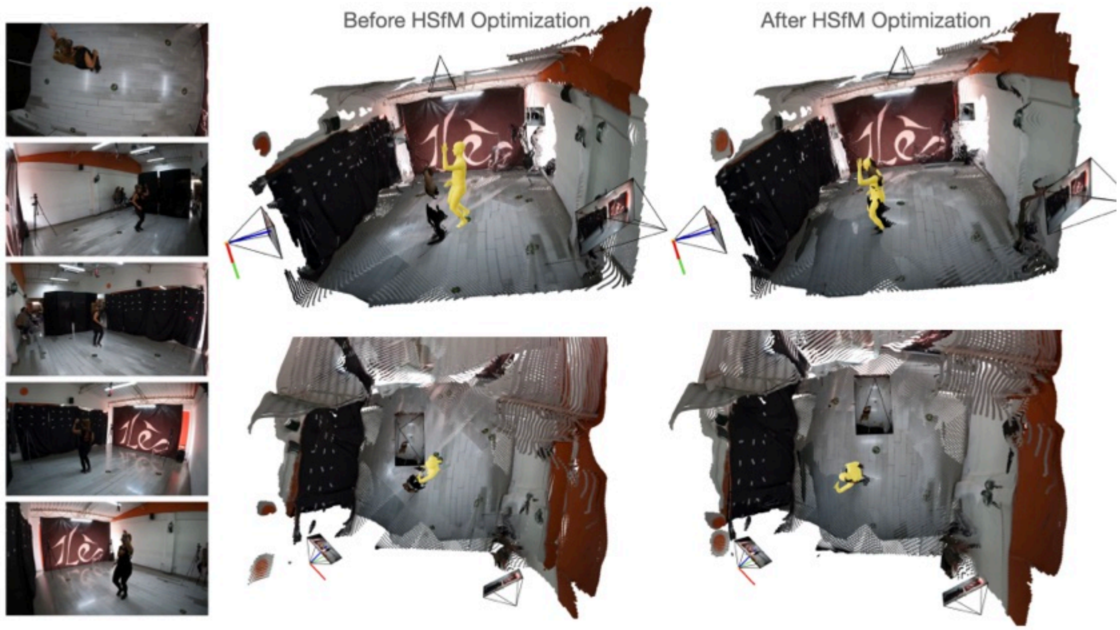


Figure S.4. **Qualitative results.** We show reconstruction on EgoExo4D. On the left, the input images to our method, the scene, humans, and cameras before optimization (HSfM (init.)) in the center, and the reconstruction of our method after joint optimization on the right.



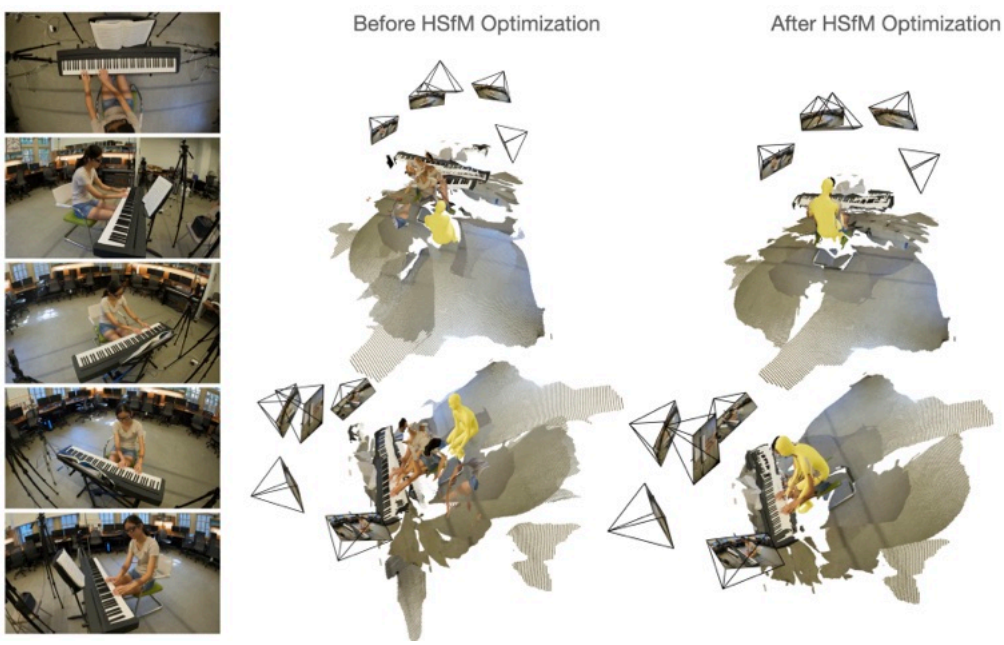


Figure S.5. **Continuation of Fig. S.4**

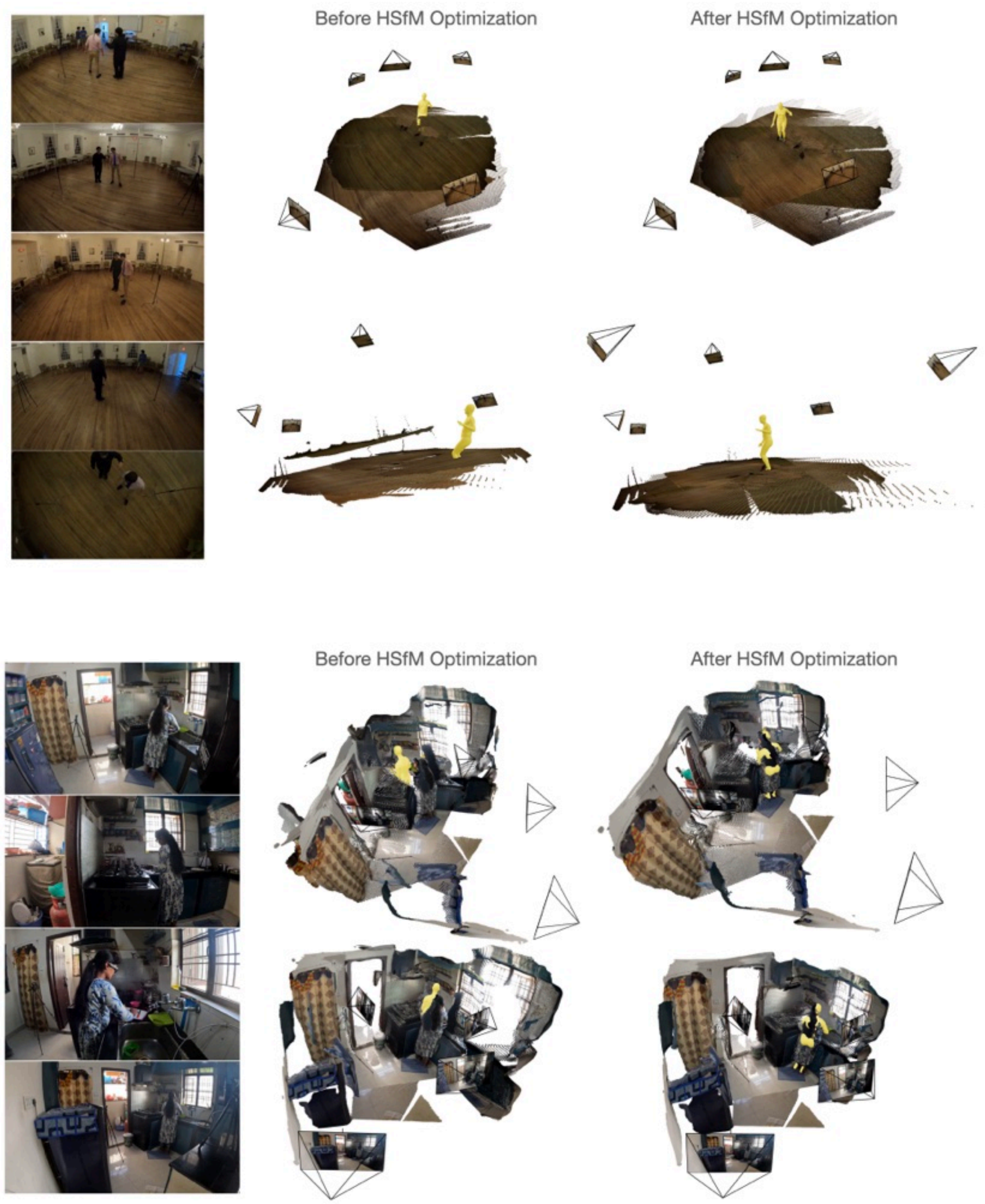


Figure S.6. **Continuation of Fig. S.4**

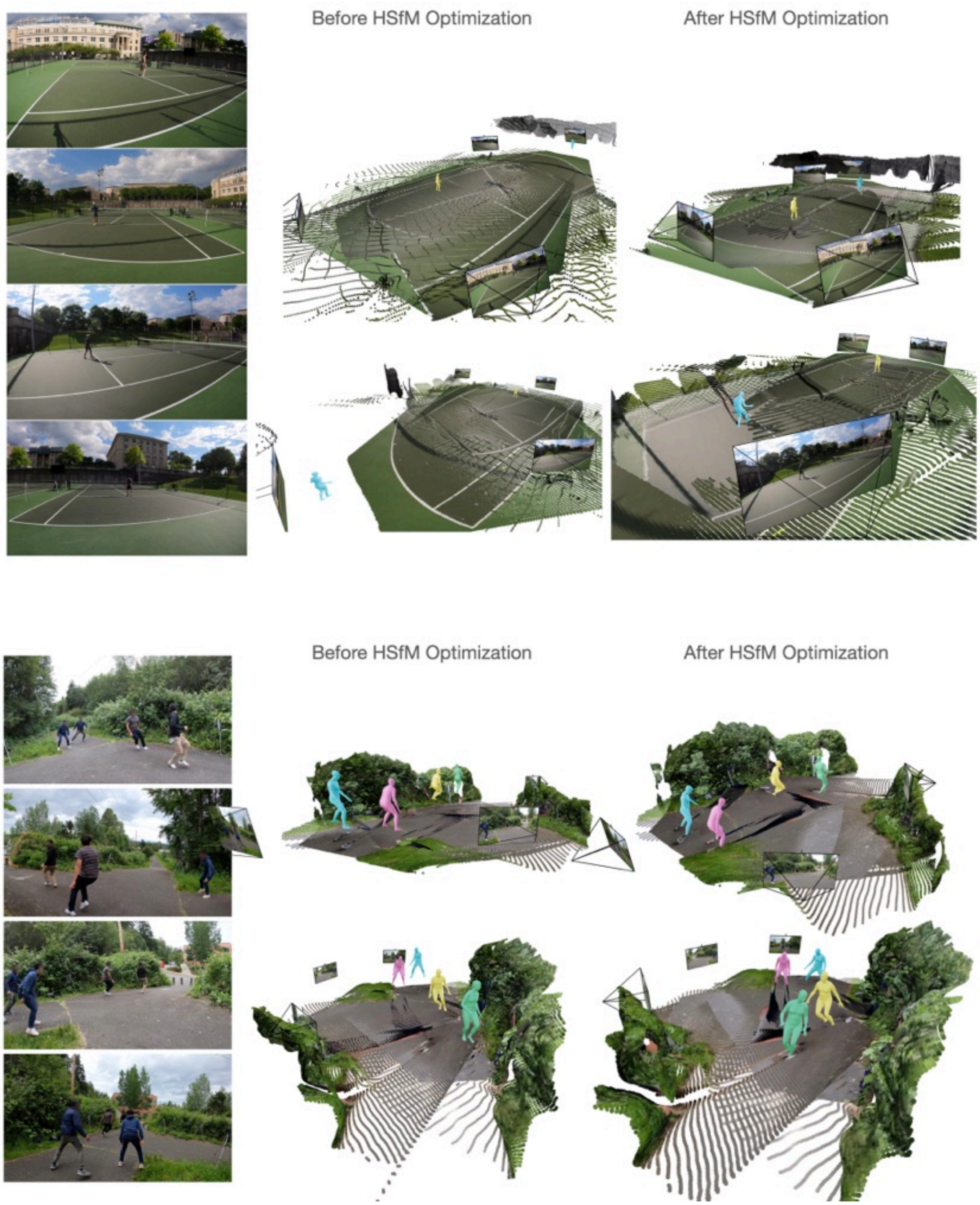


Figure S.7. **Qualitative results.** We show reconstructions on EgoHumans. On the left, the input images to our method, the scene, humans, and cameras before optimization (HSfM (init.)) in the center, and the reconstruction of our method after joint optimization on the right.

Figure S.8. **Continuation of Fig. S.7**