Multi-THuMBS: Multi-person Tracking of 3D Human Meshes Beyond Video Shots

Jeongwan On[1], Muhammad Salman Ali[1], Muneeb A. Khan[1], Sunwoo Park[1], Inwoong Moon[1], Hyung Jin Chang[2], Jaekwang Kim[3], Seong Jong Ha[3], and Seungryul Baek[1]

[1] UNIST, South Korea
[2] University of Birmingham, UK
[3] AI/DT Division, CJ Corporation, South Korea

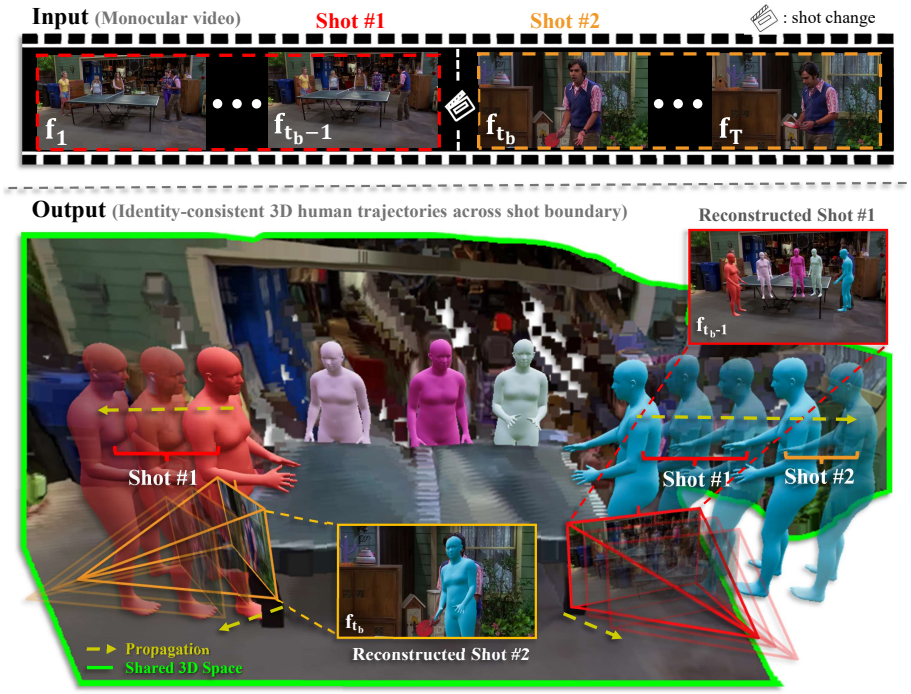


Fig. 1: Our method reconstructs globally aligned 3D human motions from multi-shot videos with multiple people, effectively handling abrupt shot changes. We apply a state-of-the-art 3D scene reconstruction technique [39] to two frames before and after the shot change ($\mathbf{f}_{t_b-1}$, $\mathbf{f}_{t_b}$) to construct a Shared 3D Space. We further register and align the 3D human meshes of shot boundaries to the Shared 3D Space and propagate the alignment throughout Shot 1 and Shot 2. Through this propagation, our method maintains consistent identity assignment and motion continuity across shot changes.

Abstract. Tracking multi-person 3D human meshes from in-the-wild videos is a highly challenging problem due to complex interactions, frequent occlusions, and severe truncation inherent in unconstrained environments. While recent approaches have improved robustness against

these issues, they largely overlook the critical challenge prevalent in real-world footage: frequent shot changes. These abrupt transitions in camera viewpoints often cause existing methods to lose track of human identities and fail in reconstructing temporally coherent trajectories. Although several recent works have explored 3D human mesh tracking under shot changes, they are still limited to single-person scenarios, making them inadequate for real-world videos where multiple people interact and appear simultaneously. To address this limitation, we propose Multi-THuMBS (Multi-person Tracking of 3D Human Meshes Beyond Video Shots) that leverages a state-of-the-art 3D scene prior to reconstruct the two boundary frames in a single shared 3D space. Human meshes are then registered within the shared 3D space, maintaining per-person identity and motion consistency across shot changes. Extensive experiments demonstrate that our approach yields significant improvements in 3D human mesh recovery, camera pose estimation, and identity tracking, thereby ensuring high-fidelity motion reconstruction with consistent identity preservation across shots compared to previous state-of-the-art methods.

Keywords: Multi-person 3D human mesh tracking · Multi-shot video · Re-identification

1 Introduction

Tracking and reconstruction of multi-person 3D human meshes from videos is a fundamental goal in computer vision with applications spanning animation, sports analytics, AR/VR, and human-computer interaction. Recent methods for multi-person 3D human mesh tracking [33, 34, 41] have achieved remarkable accuracy in real-world videos, effectively handling challenges such as occlusion and depth ambiguity. However, these advancements heavily rely on the assumption of continuous camera motion within a single shot. Real-world videos ranging from movies and broadcast sports to edited online content are typically composed of multiple shots with abrupt transitions (as illustrated in Figures 1 and 2). Despite the ubiquity of such footage, the problem of tracking 3D human motion across shot changes has been largely overlooked.

To address discontinuous camera motion, multi-view reconstruction frameworks [3, 23, 31, 43] have been proposed to integrate observations from disjoint viewpoints into a shared 3D space. However, these methods are inherently limited to simultaneous captures at single time instances, lacking the ability to model temporal human motion. Building upon these spatial alignment principles, a few pioneering works [26, 47] have attempted to extend reconstruction across shot boundaries in multi-shot videos. However, these methods are fundamentally limited to single-person scenarios and do not address the complex identity associations required when multiple individuals appear simultaneously across different shots. As a result, although prior studies partially tackle either spatial misalignment or cross-shot reconstruction, the problem of identity-consistent multi-person 3D reconstruction across shot boundaries remains largely

unsolved. This gap highlights the need for a unified framework that can jointly handle abrupt viewpoint changes, global spatial alignment, and robust multi-person identity tracking across multi-shot videos.

However, extending multi-person tracking to multi-shot scenarios introduces severe challenges that existing methods struggle to address (as shown in Fig. 2): (i) Abrupt shot changes introduce sudden variations in camera viewpoints, leading to drastic appearance changes in the image space. As a result, existing appearance-based re-identification methods often fail to maintain consistent identities across shots. (ii) Shot changes also cause abrupt changes in camera positions, making it difficult to obtain globally aligned human meshes. (iii) The number of visible individuals may change across shot boundaries, making identity association ambiguous and temporal correspondence difficult to establish. In such settings, existing methods fail to recognize shot changes, leading to erroneous global pose estimates at the boundary. Moreover, by enforcing temporal continuity across these discontinuities, they introduce unrealistic motion artifacts such as foot sliding, which severely degrades reconstruction performance.

To overcome these limitations, we propose the first geometry-driven framework Multi-THuMBS (**Multi**-person **T**racking of 3D **Hu**man **M**eshes **B**eyond Video **S**hots) designed for identity-consistent, multi-person 3D human mesh reconstruction and tracking across shot boundaries, as illustrated in Fig. 1. Our key insight is to leverage 3D scene geometry as a robust spatial anchor to bridge human trajectories across abrupt viewpoint changes. Given a multi-shot video, our framework first detects shot change and reconstructs 3D human meshes for each shot, along with visual cues such as human poses and UV texture maps [7]. To establish a shared 3D space, we reconstruct dense scene point clouds and camera parameters at boundary frames by leveraging the 3D scene priors provided by VGGT [39]. A multi-stage optimization then aligns human meshes from adjacent shots to this shared space, maintaining spatial consistency across shot boundaries. This geometric alignment enables a re-identification (Re-ID) strategy that integrates appearance, geometry, and pose cues, rather than relying solely on appearance cues, which become unreliable due to drastic variations in camera viewpoints caused by abrupt shot changes. During Re-ID, we compute pairwise 3D distances between individuals and apply thresholding to account for the varying number of visible individuals across shot boundaries, thereby mitigating false matches. The resulting correspondences are temporally propagated throughout both shots, followed by global trajectory smoothing. At shot boundaries, a cross-camera consistency loss enforces coherent reprojections, producing globally aligned and identity-consistent 3D human trajectories across the entire video.

In summary, our main contributions are as follows:

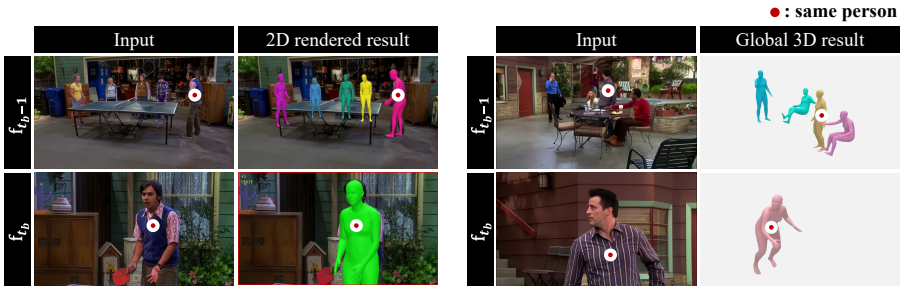
- We propose the Multi-THuMBS. To the best of our knowledge, this is the first work to extend multi-person 3D human mesh tracking to multi-shot settings. Our method maintains both identity and motion consistency across shot boundaries by leveraging a learned 3D prior to construct a shared 3D space for adjacent shots, within which individual trajectories are seamlessly connected.

- We propose a re-identification (Re-ID) strategy that integrates appearance, pose, and geometric (3D distance) cues to establish robust identity association across shot boundaries. Unlike appearance-only methods, our approach maintains reliable identity association even under drastic viewpoint shifts and severe appearance changes caused by abrupt shot changes.
- We introduce a global optimization framework that enforces spatiotemporal consistency through joint trajectory smoothing and cross-camera reprojection loss. This joint optimization stabilizes pose estimates, corrects geometric misalignment, and ensures globally consistent 3D human motion across multi-shot videos.
- Extensive experiments across multiple benchmarks encompassing multi-camera environments [11, 12, 46] and dynamic in-the-wild videos [8, 28] demonstrate consistent improvements in pose estimation, camera tracking, and identity association. Detailed ablation studies further support the contribution of each component of our framework.

2 Related Work

3D Human Mesh Reconstruction and Tracking. Early research in 3D human mesh reconstruction [10, 22] focused on single-person and single-image scenarios. While these methods established a baseline for pose estimation, they fundamentally lacked the ability to understand human motion dynamics. To overcome this limitation, subsequent studies [14, 21] extended frameworks to the video domain, incorporating temporal modeling to capture motion cues. However, these approaches remained confined to single-person settings, failing to generalize to complex, in-the-wild multi-person scenes. To address this, several works [7, 29] expanded reconstruction to multi-person tracking. Despite their success in tracking identities, they were restricted to local image-space estimation, lacking global spatial reasoning in world coordinates. Addressing this, optimization-based frameworks such as SLAHMR [44] emerged, jointly optimizing human and camera trajectories to achieve globally consistent reconstruction. However, their reliance on iterative optimization resulted in prohibitive inference times, making them unsuitable for scalable applications. Consequently, the field shifted towards feed-forward approaches to enable more efficient global 3D human motion reconstruction. Recent methods like WHAM [34] and PromptHMR [41] estimate global human motion using purely learning-based architectures, avoiding the high computational cost of iterative optimization.

Nevertheless, a critical limitation persists: they assume continuous camera motion and fail to handle shot changes, leading to tracking drifts or re-initialization errors at shot boundaries. Although Pavlakos *et al.* [27] partially addressed tracking across shot changes, their scope was limited to single-person sequences. To address these challenges, HumanMM [47] explicitly models shot changes and occlusions for multi-shot motion reconstruction, yet it remains limited to single-person scenarios. Similarly, ShowMak3r [13] jointly optimizes human pose and neural scene geometry for cross-shot alignment in multi-person settings; however, it primarily focuses on visual quality rather than motion accuracy.



(a) **Case 1.** Ambiguity in identity association (b) **Case 2.** Inconsistent global pose estimation

Fig. 2: The existing state-of-the-art human mesh reconstruction method [41] suffers from two distinct limitations: (**Case 1**) First, the method incorrectly assigns different identities across the shot change, as indicated by the different mesh colors; (**Case 2**) Second, the same persons' meshes are located in the completely different 3D locations after the shot change. In both cases, the number of visible individuals varies across the shot boundary, further complicating both re-identification and global alignment.

To date, no prior work jointly addresses multi-person tracking, and shot discontinuity (see the Supplementary Material for a detailed comparison).

Person Re-Identification. Early person re-identification (Re-ID) approaches relied on handcrafted features [5, 6], which were highly sensitive to viewpoint and illumination changes. Deep CNN-based methods [1, 38] improved robustness by learning discriminative representations, while weakly supervised frameworks [17] and architecture refinements [20] reduced dependence on large annotated datasets. Transformer-based architectures have also advanced Re-ID performance through stronger global reasoning. ViTs [4] enable long-range feature modeling, and TransReID [9] adapted them for identity tracking in crowded, multi-person scenes. Pose2ID [45] proposed a training-free feature centralization strategy that enhances identity representation without additional Re-ID training, while KPR [35] introduced keypoint-promptable matching to resolve occlusion ambiguities. Despite their strong identity robustness, existing Re-ID methods are primarily appearance-based and struggle at shot boundaries, where viewpoint, lighting, and appearance shifts cause unreliable identity association. Moreover, they also rely on long temporal sequences requiring high computational complexity, limiting their generalization to real-world, multi-shot scenarios.

In summary, existing approaches lack a unified framework for multi-person 3D human mesh reconstruction with motion and identity-consistent tracking across multi-shot videos. Existing state-of-the-art methods address only partial aspects of this problem: PromptHMR [41] handles multi-person tracking but is limited to single shots; HumanMM [47] processes multi-shot videos but only for single-person; ShowMak3r [13] handles multi-person multi-shot scenarios but prioritizes rendering quality over motion and identity consistency. In contrast,

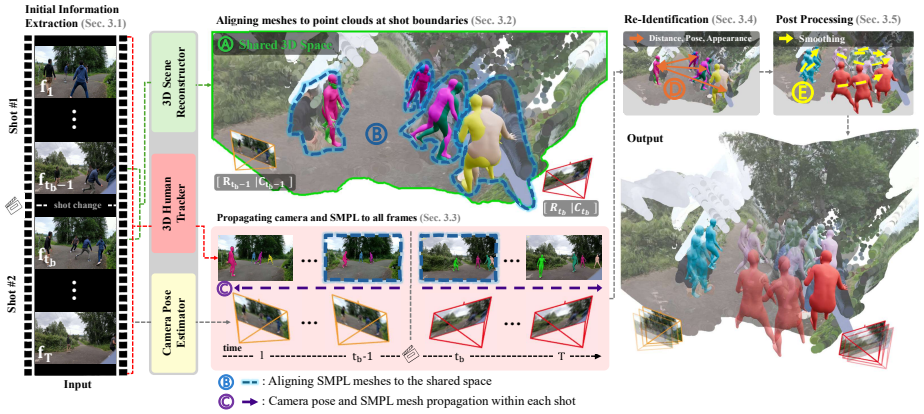


Fig. 3: Overview of the Multi-THuMBS. The input video is split into two shots at the detected boundary. We estimate 3D human meshes and camera poses for all frames, reconstruct a Ⓐ shared 3D space from the boundary frames $\{\mathbf{f}_{t_b-1}, \mathbf{f}_{t_b}\}$, and Ⓑ align both shots to this shared space. We then Ⓒ propagate camera poses and global positions within each shot, Ⓓlink identities across the boundary using spatial proximity integrated with pose and appearance cues, and apply Ⓔ temporal smoothing to obtain stable, motion- and identity-consistent 3D trajectories.

our framework is the first to achieve globally consistent multi-person 3D reconstruction with motion and identity tracking across shot boundaries.

3 Methodology

Given a multi-shot video $\mathcal{V} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$ comprising n shots, we aim to reconstruct a multi-person motion and identity-consistent 3D human meshes $\mathbf{m}$ across shot boundaries. Each shot $\mathcal{S}_j$ consists of consecutive frames $\{\mathbf{f}_t\}$ captured at a specific time t. We focus on the two-shot case ($n = 2$) for clarity; extensions to $n > 2$ are provided in supplementary material. Let f_{t_b} represent the shot boundary frame, dividing $\mathcal{V}$ into $\mathcal{S}_1 = \{\mathbf{f}_t\}_{t=1}^{t_b-1}$ (pre-boundary) and $\mathcal{S}_2 = \{\mathbf{f}_t\}_{t=t_b}^{T}$ (post-boundary), where T is the total number of frames.

We first employ PySceneDetect [2] to identify t_b, then independently reconstruct 3D human meshes $\{\mathbf{m}_t^i\}_{t=1}^{T}$ for each detected person i using 4DHumans [7]. To align these meshes across the boundary, we reconstruct scene point clouds $\mathbf{P}_{t_b-1}, \mathbf{P}_{t_b}$ at the frames $\mathbf{f}_{t_b-1}$, $\mathbf{f}_{t_b}$ using VGGT [39], which provides one-to-one pixel-to-point correspondence. Per-person point clouds $\mathbf{P}_{t_b-1}^i, \mathbf{P}_{t_b}^i$ are extracted by selecting 3D points within segmentation masks $\mathbf{M}_t^i$ obtained from Grounded SAM [30]. Aligning the boundary meshes $\{\mathbf{m}_{t_b-1}^i, \mathbf{m}_{t_b}^i\}$ to these point clouds synchronizes the coordinate systems of $\mathcal{S}_1$ and $\mathcal{S}_2$.

Finally, we perform identity matching using spatial proximity fused with pose and appearance similarities and apply temporal smoothing. Each component of

our framework is described in detail below, and the overall workflow is visualized in Figure 3.

3.1 Initial Information Extraction

For each shot $\mathcal{S}_1$ and $\mathcal{S}_2$, we extract 3D human meshes $\{\mathbf{m}_t^i\}_{t=1}^T$ using 4DHumans [7], which provides stable intra-shot identity tracking. Each mesh $\mathbf{m}_t^i$ represents the i-th person at time t and is parameterized by SMPL [18] with global orientation $\mathbf{\Phi}_t^i \in \mathbb{R}^{3\times3}$, root translation $\mathbf{\Gamma}_t^i \in \mathbb{R}^3$, pose parameters $\theta_t^i \in \mathbb{R}^{23\times3}$, and shape parameters $\beta^i \in \mathbb{R}^{10}$ (time-invariant). Additionally, we extract 2D keypoints $\mathbf{J}_t^i$ using ViTPose [42] within the bounding boxes rendered from the 3D meshes $\mathbf{m}_t^i$, which serve as supervision for the subsequent alignment optimization.

To construct a shared 3D space across different shots, we employ VGGT [39] over traditional structure-from-motion (SfM) methods [16, 36, 40]. This is because VGGT remains robust under the sparse viewpoints and extreme viewpoint changes typically observed at shot boundaries. However, VGGT is inherently designed to process multi-view images captured at a single time step, making it infeasible to apply directly to videos. In particular, VGGT struggles with performance degradation in dynamic scenes involving moving people or backgrounds, and its high computational cost makes it unsuitable for long-term videos. Therefore, adopting the assumption from [47] that the temporal gap at a shot boundary is negligible, we apply VGGT only to the boundary frames, $\mathbf{f}_{t_b-1}$ and $\mathbf{f}_{t_b}$, treating them as a set of multi-view captures. Specifically, we construct the 3D point clouds $\{\mathbf{P}_{t_b-1}, \mathbf{P}_{t_b}\}$ along with their corresponding intrinsic parameters $\mathbf{K}$ and camera poses $\{[\mathbf{R}_{t_b-1}|\mathbf{C}_{t_b-1}], [\mathbf{R}_{t_b}|\mathbf{C}_{t_b}]\}$, where $\mathbf{R}_t \in \mathbb{R}^{3\times3}$ and $\mathbf{C}_t \in \mathbb{R}^3$ denote the rotation and center translation, respectively. Since VGGT assigns a unique 3D point to each pixel, we then extract per-person point clouds $\{\mathbf{P}_{t_b-1}^i, \mathbf{P}_{t_b}^i\}$ by selecting points within the segmentation masks $\mathbf{M}_t^i$ obtained via Grounded SAM [30].

Finally, to address the scale ambiguity between the meshes and point clouds, we compute the maximum centroid-to-vertex distance for both $\{\mathbf{m}_{t_b}^i, \mathbf{m}_{t_b-1}^i\}$ and $\{\mathbf{P}_{t_b}^i, \mathbf{P}_{t_b-1}^i\}$, determine their ratio, average it across all persons, and apply this factor to uniformly rescale $\{\mathbf{P}_{t_b}, \mathbf{P}_{t_b-1}\}$.

3.2 Aligning Meshes to Point Clouds at Shot Boundary

With the scale uniformly aligned, our next step is to spatially anchor the human meshes into VGGT's shared 3D space. However, the per-person 3D meshes $\{\mathbf{m}_{t_b-1}^i, \mathbf{m}_{t_b}^i\}$ and point clouds $\{\mathbf{P}_{t_b-1}^i, \mathbf{P}_{t_b}^i\}$ are initially defined in different coordinate systems. To align them, we optimize the human root translation $\{\mathbf{\Gamma}_{t_b-1}^i, \mathbf{\Gamma}_{t_b}^i\}$, global orientation $\{\mathbf{\Phi}_{t_b-1}^i, \mathbf{\Phi}_{t_b}^i\}$, and the camera pose $\{[\mathbf{R}_{t_b-1}|\mathbf{C}_{t_b-1}]$, $[\mathbf{R}_{t_b}|\mathbf{C}_{t_b}]\}$. Direct joint optimization of these spatial parameters, however, is highly non-convex and often becomes unstable due to severe parameter coupling. To address this, we propose a three-stage progressive optimization (Algorithm 1) that sequentially refines them. The full optimization objective combines

2D reprojection, silhouette alignment, and depth consistency:

$$\min_{\mathbf{\Gamma}_t^i, \mathbf{\Phi}_t^i, [\mathbf{R}_t|\mathbf{C}_t]} \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}, \tag{1}$$

where $t \in \{t_b - 1, t_b\}$, $\mathcal{L}_{\text{2D}}$ enforces 2D reprojection consistency, $\mathcal{L}_{\text{sil}}$ ensures silhouette consistency, and $\mathcal{L}_{\text{depth}}$ enforces depth alignment with scene geometry. These losses and parameters are optimized across three stages to avoid local minima, as detailed below.

In the first stage, we initialize the root translation $\{\mathbf{\Gamma}_{t_b-1}^i, \mathbf{\Gamma}_{t_b}^i\}$ by leveraging the one-to-one mapping between the image pixels and the VGGT point cloud. Specifically, we extract the 2D root joint from the detected keypoints $\{\mathbf{J}_{t_b-1}^i, \mathbf{J}_{t_b}^i\}$ and assign the 3D coordinate of its corresponding VGGT point as the initial translation.

Algorithm 1 Aligning meshes to point clouds at shot boundary

Input: $\{\mathbf{m}_{t_b-1}^i, \mathbf{m}_{t_b}^i\}$: Per-person local meshes
Input: $[\mathbf{R}_{t_b-1}|\mathbf{C}_{t_b-1}], [\mathbf{R}_{t_b}|\mathbf{C}_{t_b}]$: Camera poses
Output: $\{{}^w\mathbf{m}_{t_b-1}^i, {}^w\mathbf{m}_{t_b}^i\}$: Global meshes
Output: $[\hat{\mathbf{R}}_{t_b-1}|\hat{\mathbf{C}}_{t_b-1}], [\hat{\mathbf{R}}_{t_b}|\hat{\mathbf{C}}_{t_b}]$: Refined poses

```
max_iter ← [500, 500, 1500];
for stage s = 1 to 3 do
    iter ← 1;
    repeat
        for t ∈ {t_b − 1, t_b} do
            for each person i do
                if s = 1 then
                    Initialize Γ_t^i;
                else if s = 2 then
                    Optimize Γ_t^i, Φ_t^i;
                    Minimize λ_2D^(2) L_2D;
                else
                    Optimize Γ_t^i, Φ_t^i, [R_t|C_t];
                    Minimize λ_2D^(3) L_2D + λ_sil L_sil +
                    λ_depth L_depth;
            end
        end
        iter ← iter + 1;
    until iter > max_iter[s];
end
```

In the second stage, we jointly refine both $\{\mathbf{\Gamma}_{t_b-1}^i, \mathbf{\Gamma}_{t_b}^i\}$ and $\{\mathbf{\Phi}_{t_b-1}^i, \mathbf{\Phi}_{t_b}^i\}$ for 500 iterations to better align with 2D observations with $\lambda_{\text{2D}}^{(2)} = 10.0$:

$$\min_{\mathbf{\Gamma}_t^i, \mathbf{\Phi}_t^i} \lambda_{\text{2D}}^{(2)}\mathcal{L}_{\text{2D}}. \tag{2}$$

The 2D reprojection loss $\mathcal{L}_{\text{2D}}$ enforces consistency between projected 3D joints and detected 2D keypoints:

$$\mathcal{L}_{\text{2D}} = \sum_i \| \mathit{\Pi}_f(\mathbf{R}_t \cdot J(\phi_t^i) + \mathbf{C}_t) - \mathbf{J}_t^i \|_2^2, \tag{3}$$

where ϕ_t^i denotes SMPL parameters (i.e., $\mathbf{\Gamma}_t^i$, $\mathbf{\Phi}_t^i$, θ_t^i, β_t^i) and $\mathbf{J}_t^i$ are 2D keypoints extracted using ViTPose [42], $J(\cdot)$ denotes the SMPL joint regressor [18], and $\mathit{\Pi}_f(\cdot)$ represents perspective projection using camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times3}$.

In the final stage, we incorporate 3D geometric scene constraints to resolve depth ambiguity by optimizing all parameters in Eq. 1 for 1500 iterations with $\lambda_{\text{2D}}^{(3)} = 50.0$, $\lambda_{\text{sil}} = 0.01$, and $\lambda_{\text{depth}} = 1.0$.

For each boundary frame $t \in \{t_b - 1, t_b\}$, the depth consistency loss $\mathcal{L}_{\text{depth}}$ penalizes the 3D distance between SMPL vertices $\mathbf{v}_k$ whose projections lie within $\mathbf{M}_t^i$ and corresponding scene points $\mathbf{p}_k \in \mathbf{P}_t^i$:

$$\mathcal{L}_{\text{depth}} = \sum_k \|\mathbf{v}_k - \mathbf{p}_k\|_2^2. \tag{4}$$

Similarly, the silhouette loss $\mathcal{L}_{\text{sil}}$ penalizes vertices projecting outside $\mathbf{M}_t^i$:

$$\mathcal{L}_{\text{sil}} = \sum_{\mathbf{v}_k} \text{dist}\left(\Pi_f(\mathbf{v}_k)\right), \tag{5}$$

where $\text{dist}(u) \in \mathbb{R}^{H \times W}$ is a precomputed distance transform assigning each pixel u its distance to the nearest mask boundary [32].

This process produces globally aligned meshes $\{{}^w\mathbf{m}_{t_b-1}^i, {}^w\mathbf{m}_{t_b}^i\}$, derived from the aligned SMPL parameters $\{{}^w\mathbf{\Phi}_t, {}^w\mathbf{\Gamma}_t, \theta_t, \beta\}_{t \in \{t_b-1, t_b\}}$, along with refined camera poses $\{[\hat{\mathbf{R}}_{t_b-1}|\hat{\mathbf{C}}_{t_b-1}], [\hat{\mathbf{R}}_{t_b}|\hat{\mathbf{C}}_{t_b}]\}$. These outputs serve as anchors for subsequent temporal propagation.

3.3 Propagating Camera Poses and SMPL Parameters

As discussed in Sec. 3.1, while VGGT [39] successfully constructs a shared 3D space at the boundary frames $\mathbf{f}_{t_b-1}$ and $\mathbf{f}_{t_b}$, applying it to all video frames is computationally prohibitive and prone to errors in dynamic scenes with significant human and background motion. To efficiently track cameras across non-boundary frames, we adopt DROID-SLAM [37], which provides lightweight, robust, and temporally smooth camera tracking within a single continuous shot. Therefore, to achieve global consistency across multiple shots, we align these intra-shot trajectories to our shared global space by computing the relative transformation between the DROID-SLAM and VGGT camera poses at the shot boundary.

Specifically, for each boundary frame $t \in \{t_b - 1, t_b\}$, the relative transformation matrix $\mathcal{R}_t \in SE(3)$ is computed as:

$$\mathcal{R}_t = [\hat{\mathbf{R}}_t|\hat{\mathbf{C}}_t] \cdot [\mathbf{R}_t^*|\mathbf{C}_t^*]^{-1}, \tag{6}$$

where $[\mathbf{R}_t^*|\mathbf{C}_t^*]$ denote the camera poses estimated by DROID-SLAM. Here, $\mathcal{R}_t$ serves as the transformation aligning the local DROID-SLAM coordinate system to our shared global space. Once computed at the boundary, these transformations are propagated to all non-boundary frames within their respective shots. Specifically, $\mathcal{R}_{t_b-1}$ is applied to the DROID-SLAM poses $[\mathbf{R}_t^*|\mathbf{C}_t^*]$ for all frames where $t \in \mathcal{S}_1$, and similarly, $\mathcal{R}_{t_b}$ is applied to the cameras for all frames where $t \in \mathcal{S}_2$. Through this propagation, the locally consistent intra-shot trajectories are seamlessly integrated into a unified global coordinate system.

Similarly, we propagate the globally aligned SMPL parameters, specifically the global orientation ${}^w\mathbf{\Phi}_t^i$ and root translation ${}^w\mathbf{\Gamma}_t^i$, from the boundary frames to all non-boundary frames. This step is required because only the human meshes

at the boundary frames are aligned to the shared global space. By applying corresponding relative transformations to the SMPL parameters, we ensure that the entire sequence of human meshes is temporally coherent and accurately situated within the shared global space. The detailed algorithm for propagation is provided in the supplementary material.

3.4 Re-Identification

With aligned meshes in a shared 3D space, identity matching fundamentally reduces to spatial proximity. For boundary frames $\mathbf{f}_{t_b-1}$ and $\mathbf{f}_{t_b}$, we establish identity correspondences by computing pairwise Euclidean distances between root translations ${}^{w}\mathbf{\Gamma}^{i}_{t_b-1}$ and ${}^{w}\mathbf{\Gamma}^{j}_{t_b}$ of globally aligned meshes ${}^{w}\mathbf{m}^{i}_{t_b-1}$ and ${}^{w}\mathbf{m}^{j}_{t_b}$:

$$\mathbb{D}_{ij} = \left\| {}^{w}\mathbf{\Gamma}^{i}_{t_b-1} - {}^{w}\mathbf{\Gamma}^{j}_{t_b} \right\|_2, \tag{7}$$

where i and j denote index of each mesh in the pre- and post-boundary frames.

To further strengthen identity association under challenging viewpoints and lighting variations, we incorporate complementary appearance and pose cues. Appearance dissimilarity $\mathbb{A}_{ij}$ compares UV texture maps extracted using 4DHumans [7], computing pixel-wise color differences over valid non-occluded regions. Pose dissimilarity $\mathbb{P}_{ij}$ measures axis-angle differences between SMPL joint rotations, considering only joints visible in both frames based on projected 2D keypoint visibility. These cues are integrated into a unified cost matrix:

$$\mathbb{U}_{ij} = \lambda_{geo}\mathbb{D}_{ij} + \lambda_{app}\mathbb{A}_{ij} + \lambda_{pos}\mathbb{P}_{ij}, \tag{8}$$

where $\lambda_{geo} = 1.0$, $\lambda_{app} = 0.2$, and $\lambda_{pos} = 0.35$ denote the weights for the geometric, appearance, and pose terms, respectively. An ablation study analyzing the impact of these hyperparameters is provided in the Supplementary Material. We then apply the Hungarian algorithm [15] on $\mathbb{U}_{ij}$ to obtain the optimal one-to-one matching. To prevent false associations when individuals enter or exit the scene, we discard any matched pairs with a spatial distance $\mathbb{D}_{ij} > \tau$, where $\tau = 1.0$ meter. This hybrid formulation improves the geometry-driven matching and identity disambiguation in cases where spatial proximity alone is insufficient. Ultimately, this process yields the final set of matched human pairs $\mathcal{M}$ across the shot boundary.

3.5 Post-processing

Finally, we apply a post-processing step to refine temporal and geometric consistency across the entire video. This step smooths joint trajectories, maintains plausible body poses, and enforces cross-camera geometric consistency at the boundaries, yielding stable multi-person motions. Specifically, we formulate an optimization objective that integrates a joint-level temporal smoothing term $\mathcal{L}_{\text{smooth}}$ across all frames with a cross-camera reprojection constraint $\mathcal{L}_{\text{cross}}$ at the shot boundaries:

$$\min_{{}^{w}\boldsymbol{\phi}^{i}_{t}} \lambda_{\text{smooth}}\,\mathcal{L}_{\text{smooth}} + \lambda_{\text{cross}}\,\mathcal{L}_{\text{cross}}, \tag{9}$$

where ${}^{w}\phi_t^i$ denotes the set of globally aligned SMPL parameters (*i.e.*, ${}^{w}\mathbf{\Phi}_t^i$, ${}^{w}\mathbf{\Gamma}_t^i$, θ_t^i, β_t^i), and the loss weights are set to $\lambda_{\text{smooth}} = 10.0$ and $\lambda_{\text{cross}} = 1.0$.

The temporal smoothing loss $\mathcal{L}_{\text{smooth}}$ enforces temporal continuity of joint trajectories while regularizing body shape and pose parameters to maintain plausible human configurations:

$$\mathcal{L}_{\text{smooth}} = \sum_t \sum_i \left\| {}^{w}\mathbf{J}_t^i - {}^{w}\mathbf{J}_{t+1}^i \right\|_2^2 + \sum_i \left\| \beta^i \right\|_2^2 + \sum_{i,t} \left\| \zeta_t^i \right\|_2^2, \tag{10}$$

where ${}^{w}\mathbf{J}_t^i$ denotes 3D joints of globally aligned mesh, β^i denotes the SMPL shape parameter, and ζ_t^i denotes the VPoser [25] latent vector representing pose ${}^{w}\theta_t^i$. The cross-camera loss $\mathcal{L}_{\text{cross}}$ extends the 2D reprojection loss (Eq. 3) to enforce geometric consistency across two different camera viewpoints at the shot boundary. For a matched person pair $i \in \mathcal{M}$, the mesh from one shot is projected using the camera of the adjacent shot, and vice versa:

$$\begin{aligned} \mathcal{L}_{\text{cross}} = \sum_{i \in \mathcal{M}} \Big[& \left\| \Pi_f \left(\mathbf{R}_{t_b} J({}^{w}\phi_{t_b-1}^i) + \mathbf{C}_{t_b} \right) - \mathbf{J}_{t_b}^i \right\|_2^2 \\ & + \left\| \Pi_f \left(\mathbf{R}_{t_b-1} J({}^{w}\phi_{t_b}^i) + \mathbf{C}_{t_b-1} \right) - \mathbf{J}_{t_b-1}^i \right\|_2^2 \Big], \end{aligned} \tag{11}$$

By refining the pose trajectories and ensuring coherent reprojections across adjacent shots, the post-processing stage effectively produces stable, identity-consistent 3D human motions throughout the entire multi-shot sequence.

4 Experiments

4.1 Implementation Details

Our framework is implemented in PyTorch [24] and optimized using the AdamW [19] optimizer. For a video containing 150 frames at 1920×1080 resolution, the complete optimization takes approximately 10 minutes on a single NVIDIA RTX 3090 GPU. All detailed dataset constructions, implementation settings, and evaluation protocols are provided in the supplementary material.

Datasets. To comprehensively evaluate our method, we construct a multi-shot benchmark by modifying multi-view sequences from EgoHumans [11], EgoBody [46], and Harmony4D [12] to introduce shot boundaries at specific frames. For qualitative and quantitative motion evaluation, we use AVA [8] and sitcom videos from *Friends* (1994) and *The Big Bang Theory* (2007), which naturally contain shot changes. All videos are segmented into clips based on shot boundaries to align with our evaluation protocol.

Metrics. We evaluate pose with MPJPE and MPVPE, temporal smoothness with Accel, camera localization with ATE, and identity consistency via identity switches (IDs). We report three MPJPE variants: W-MPJPE (initial-frame alignment) for trajectory consistency, and WA-MPJPE (trajectory-level alignment) for shape accuracy. For real videos lacking ground-truth annotations, we additionally evaluate cross-shot motion quality using *cross-shot* PCK (PCK*) [26], Jitter, and Foot Sliding (FS).

Table 1: Quantitative comparison with state-of-the-art methods on EgoHumans, EgoBody, and Harmony4D. We report human pose metrics (W-MPJPE, WA-MPJPE, MPJPE, MPVPE, Acceleration).

	Method	Human Metrics				
		W-MPJPE↓	WA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
EgoHumans	Multishot [26]	474.1	287.4	347.1	408.2	63.15
	GVHMR [33]	404.8	204.7	287.4	371.4	59.0
	PromptHMR [41]	1778.2	440.9	285.3	364.1	74.3
	HSfM† [23]	544.2	187.7	263.1	294.3	48.7
	Ours	**279.0**	**166.0**	**228.3**	**262.2**	**27.3**
EgoBody	Multishot [26]	185.1	144.1	147.1	166.5	17.9
	GVHMR [33]	174.0	133.3	108.0	147.1	14.7
	PromptHMR [41]	1228.1	395.3	99.0	133.9	17.4
	HSfM† [23]	113.1	96.3	113.3	123.2	10.9
	Ours	**99.2**	**72.8**	**72.0**	**94.9**	**6.0**
Harmony4D	Multishot [26]	248.0	231.6	511.2	609.5	37.1
	GVHMR [33]	244.9	166.7	244.7	334.1	29.6
	PromptHMR [41]	1746.3	399.8	675.7	746.0	66.8
	HSfM† [23]	372.0	178.4	225.6	257.6	28.3
	Ours	**221.0**	**116.9**	**215.9**	**278.3**	**17.4**

Table 2: Comparison of cross-shot Re-ID (IDs) and camera pose estimation (ATE) on EgoHumans, EgoBody, and Harmony4D. Ours (*dist.*) denotes the baseline that uses only 3D distance term (Eq. 7) for the Re-ID.

Method	Re-ID Metrics (IDs↓)		
	EgoHumans	EgoBody	Harmony4D
PromptHMR [41]	10.40	1.29	8.00
HSfM† [23]	3.87	0.20	1.58
KPR [35]	2.54	0.05	1.19
Pose2ID [45]	4.62	0.72	1.32
Ours (*dist.*)	1.66	0.00	0.54
Ours	**0.97**	**0.00**	**0.46**

Method	Camera Metrics (ATE↓)		
	EgoHumans	EgoBody	Harmony4D
VGGT [39]	1.4	1.3	1.4
PromptHMR [41]	2.3	1.6	2.3
HSfM† [23]	2.8	2.9	3.2
Ours	**0.7**	**0.1**	**0.7**

Table 3: Ablation study on EgoHumans. (1) '*w/o cam*' denotes a baseline that removes camera pose optimization. (2) '*w/o post*' removes post-processing in Sec. 3.5. (3) '*w/o stage*' disables the hierarchical optimization in Sec. 3.2. (4) '*w/o align*' skips the alignment in Sec. 3.2 entirely.

Method	Human Metrics				Camera Metrics
	W-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓	ATE↓
(1) w/o *cam*	389.7	230.1	264.7	33.7	1.40
(2) w/o *post*	311.2	241.6	298.3	47.9	0.77
(3) w/o *stage*	491.9	368.1	359.4	33.9	2.75
(4) w/o *align*	882.7	422.4	392.1	34.9	1.40
Ours	**278.8**	**228.3**	**262.1**	**27.3**	**0.77**

Table 4: Quantitative comparison with the state-of-the-art method on edited videos (*i.e.*, AVA, *Friends*, and *The Big Bang Theory*). We report motion quality metrics (PCK*, Jitter, FS). PCK* denotes *cross-shot* PCK [26]

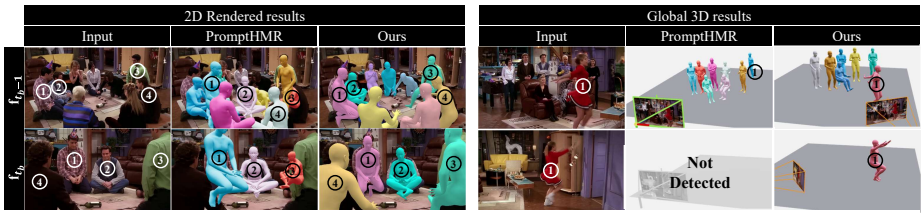
Method	Motion quality		
	PCK* ↑	Jitter↓	FS↓
PromptHMR [41]	62.7	162.44	23.11
Ours	**90.7**	**31.5**	**10.7**

4.2 Results

Baselines. We compare our approach against state-of-the-art methods for multi-human 3D reconstruction and tracking including Multishot [26], GVHMR [33], PromptHMR [41], and HSfM†. where HSfM† denotes a modified variant of HSfM [23] adapted to our multi-shot, multi-person setting (implementation details in supplementary material).

For camera pose estimation, we additionally evaluate against PromptHMR [41], VGGT [39], and HSfM†, which provide comparisons for our boundary-aware camera optimization approach. For re-identification across shot boundaries, we include appearance-based methods KPR [35] and Pose2ID [45], which rely on visual features and are sensitive to viewpoint and illumination changes.

Quantitative results. As shown in Table 1, our method achieves state-of-the-art spatial pose accuracy and temporal smoothness on EgoHumans [11], EgoBody [46], and Harmony4D [12]. For camera estimation and Re-ID (Table 2),



(a) **Case 1.** Hard crowding

(b) **Case 2.** Severe body truncation driven by camera close-up

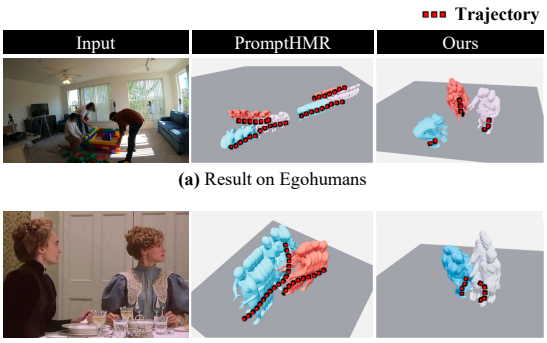
Fig. 4: Qualitative results for challenging cases: (**Case 1**) When the number of visible persons changes drastically across a shot boundary, PromptHMR loses track of specific individuals, whereas our method successfully maintains consistent tracking. (**Case 2**) Under severe body truncation driven by camera close-up, PromptHMR predicts completely incorrect camera poses, while our method robustly estimates accurate camera locations and global trajectories.

we achieve the lowest ATE, demonstrating robust alignment under abrupt viewpoint changes. For Re-ID, both our spatial-only and full variants substantially outperform appearance-based baselines [35, 45], validating the efficacy of our shared 3D space. Finally, on in-the-wild datasets (AVA, sitcoms), our method significantly outperforms PromptHMR [41] in PCK*, Jitter, and FS (Table 4), proving our ability to generate smooth, physically plausible transitions across abrupt shot changes.

Qualitative results. Figures 4–5 qualitatively compare our pipeline with PromptHMR [41]. Figure 4 demonstrates our superior camera pose estimation and person tracking at shot boundaries. Under challenging conditions like hard crowding and severe occlusion, PromptHMR frequently loses track of specific individuals or predicts completely incorrect camera poses. In contrast, our method reliably re-identifies persons across transitions and recovers accurate camera parameters, ensuring correct mesh-image alignment.

Furthermore, our method effectively suppresses temporal drift in global 3D trajectories (Fig. 5). PromptHMR exhibits noticeable sliding artifacts and abrupt positional shifts at shot boundaries, particularly during rapid viewpoint changes or zoom-ins in datasets like EgoHumans [11] and AVA [8]. Conversely, our approach maintains stable and coherent multi-person motion trajectories across these complex transitions.

Overall, these improvements underline the robustness of our pipeline in handling challenging shot transitions, achieving temporally consistent and spatially accurate reconstructions where prior methods struggle.



(a) Result on Egohumans

(b) Result on AVA

Fig. 5: Qualitative comparison of global trajectories focusing on foot sliding artifacts. Compared to PromptHMR, our method produces smoother and more physically plausible motion with reduced foot sliding, as shown by the aligned trajectories (red dotted lines).

Ablation study. Table 3 presents an ablation study of each component of our method on EgoHumans [11] using pose accuracy (W-MPJPE, MPJPE, MPVPE), temporal consistency (Accel), and camera accuracy (ATE). Removing the *cam* module significantly increases both ATE and pose errors, emphasizing the importance of camera refinement. Excluding *post* degrades temporal smoothness (Accel), validating its role in trajectory refinement. Disabling *stage* causes performance drops across all metrics, underscoring the necessity of progressive optimization. The absence of *align* causes the most significant performance decline, highlighting the critical importance of accurate human-scene alignment for cross-shot reconstruction. The full model (**Ours**) achieves the best performance across all metrics, confirming that each component contributes to the overall framework.

5 Conclusion

In this paper, we propose a novel pipeline that performs human mesh reconstruction for multi-shot videos involving multiple-persons. A key contribution of our method is its ability to maintain identity consistency across shots successfully reconstructing meshes without losing person-specific identity information and achieving reliable re-identification. This is achieved by explicitly involving 3D scene reconstruction method and aligning 3D meshes to the shared 3D space. Experimental results demonstrated the effectiveness of our method and ablation studies confirmed the usefulness of each component.

Limitation. The primary objective of Multi-THuMBS is to reconstruct continuous 3D trajectories, which inherently requires shot transitions to occur within the same physical environment (intra-scene). When a video undergoes an inter-scene transition with no spatial overlap, the concept of a continuous 3D motion becomes physically meaningless. While our geometry-driven Re-ID effectively handles intra-scene transitions, exploring identity tracking across entirely disjoint spaces remains an interesting future direction.

Acknowledgements

This work is supported by NRF grants (No. RS-2025-00521013 20%, No. RS-2025-02216916 10%) and IITP grants (No. RS-2020-II201336 Artificial intelligence graduate school program(UNIST) 10%; No. RS-2025-25442824 AI Star Fellowship Program(UNIST) 10%), funded by the Korean government (MSIT). This work is supported by CJ Corporation 50%.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An Improved Deep Learning Architecture for Person Re-Identification. In: CVPR (2015)
2. Brandon, C.: PySceneDetect: Automatic Scene Cut Detection Tool (2014)
3. Choi, C., Kim, J., Cha, G., Kim, M., Wee, D., Kim, Y.M.: Humans as a calibration pattern: Dynamic 3d scene reconstruction from unsynchronized and uncalibrated videos. In: ICCV (2025)
4. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person Re-Identification by Symmetry-Driven Accumulation of Local Features. In: CVPR (2010)
6. Gheissari, N., Sebastian, T.B., Hartley, R.: Person Reidentification Using Spatiotemporal Appearance. In: CVPR (2006)
7. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and Tracking Humans with Transformers. In: ICCV (2023)
8. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In: CVPR (2018)
9. He, S., Yu, H.X., Shi, X., Zhang, X., Zheng, W.S., Sun, J.: TransReID: Transformer-Based Object Re-Identification. In: ICCV (2021)
10. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end Recovery of Human Shape and Pose. In: CVPR (2018)
11. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.: EgoHumans: An Egocentric 3D Multi-Human Benchmark. In: ICCV (2023)
12. Khirodkar, R., Song, J.T., Cao, J., Luo, Z., Kitani, K.: Harmony4D: A Video Dataset for In-The-Wild Close Human Interactions. In: NeurIPS (2024)
13. Kim, S., Do, S., Park, J.: Showmak3r: Compositional tv show reconstruction. In: CVPR (2025)
14. Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: Video Inference for Body Pose and Shape Estimation. In: CVPR (2020)
15. Kuhn, H.W.: The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly (1955)
16. Leroy, V., Cabon, Y., Revaud, J.: Grounding Image Matching in 3D with MASt3R. In: ECCV (2024)
17. Liu, H., Yu, Y., Zhou, Z., Shao, M., Fu, Y.: End-to-End Comparative Attention Network for Person Re-Identification. In: TIP (2017)
18. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A Skinned Multi-Person Linear Model. In: ACM ToG (2015)

19. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)
20. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPR (2019)
21. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 3DV (2018)
22. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: VNect: real-time 3D human pose estimation with a single RGB camera. In: ACM ToG (2017)
23. Müller, L., Choi, H., Zhang, A., Yi, B., Malik, J., Kanazawa, A.: Reconstructing people, places, and cameras. In: CVPR (2025)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeuRIPS (2019)
25. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In: CVPR (2019)
26. Pavlakos, G., Kanazawa, A., Malik, J., Black, M.J., Daniilidis, K.: Multi-Shot Person Tracking and Re-Identification in TV Series. In: CVPR (2022)
27. Pavlakos, G., Malik, J., Kanazawa, A.: Human mesh recovery from multiple shots. In: CVPR (2022)
28. Pavlakos, G., Weber, E., Tancik, M., Kanazawa, A.: The one where they reconstructed 3d humans and environments in tv shows. In: ECCV (2022)
29. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3D appearance, location & pose. In: CVPR (2022)
30. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
31. Rojas, S., Armando, M., Ghanem, B., Weinzaepfel, P., Leroy, V., Rogez, G.: Hamst3r: Human-aware multi-view stereo 3d reconstruction. In: ICCV (2025)
32. SciPy: Exact euclidean distance transform, available at: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.distance_transform_edt.html`
33. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-Grounded Human Motion Recovery via Gravity-View Coordinates. In: ACM ToG (2024)
34. Shin, S., Kim, J., Halilaj, E., Black, M.J.: WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion. In: CVPR (2024)
35. Somers, V., Alahi, A., Vleeschouwer, C.D.: Keypoint promptable re-identification. In: ECCV (2024)
36. Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z.: MV-DUSt3R+: Single-Stage Scene Reconstruction from Sparse Views In 2 Seconds. In: CVPR (2025)
37. Teed, Z., Deng, J.: DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In: NeurIPS (2021)
38. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification. In: ECCV (2016)
39. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual Geometry Grounded Transformer. In: CVPR (2025)

40. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D Vision Made Easy. In: CVPR (2024)
41. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: PromptHMR: Promptable Human Mesh Recovery. In: CVPR (2025)
42. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: NeurIPS (2022)
43. Yang, F., Gu, K., Nguyen, H.L., Tse, T.H.E., Yao, A.: Humans as Checkerboards: Calibrating Camera Motion Scale for World-Coordinate Human Mesh Recovery. In: ICCV (2025)
44. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling Human and Camera Motion from Videos in the Wild. In: CVPR (2023)
45. Yuan, C., Zhang, G., Ma, C., Zhang, T., Niu, G.: From poses to identity: Training-free person re-identification via feature centralization. In: CVPR (2025)
46. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: ECCV (2022)
47. Zhang, Y., Wu, G., Chen, L.H., Zhao, Z., Lin, J., Jiang, X., Wu, J., Li, Z., Yang, H.F., Wang, H., Zhang, L.: HumanMM: Global Human Motion Recovery from Multi-shot Videos. In: CVPR (2025)