



Feed-Forward Multi-view Multi-person Reconstruction with Contrastive Human-Aware 3D Representation

Yuanwang Yang¹ · Buzhen Huang¹ · Zongxuan Ren¹ · Jing Huang¹ · Kun Li¹

Received: 15 January 2026 / Accepted: 11 August 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Multi-view human reconstruction has been extensively studied under simplified settings, yet scaling these methods to robust and efficient multi-person reconstruction in unconstrained environments requires a more general and scalable modeling paradigm. Existing bottom-up methods often rely on accurate camera calibration and explicit cross-view matching, and therefore struggle in multi-person scenarios with severe occlusions and ambiguities. We attribute these limitations to the lack of an explicit and view-agnostic 3D representation for jointly reasoning about human structure and multi-view consistency. Based on this observation, we propose a new top-down paradigm that maintains a unified, instance-centric human-aware 3D space, enabling simultaneous camera calibration, cross-view association, and human reconstruction via cross-modal contrastive learning. Specifically, observations from multiple views are lifted and fused directly into this shared 3D space, where geometric structure, visual appearance, and human-centric semantic cues are jointly encoded at the instance level. To enhance the discriminability and consistency of the 3D representation, we introduce a spatial contrastive learning strategy that aligns 3D features corresponding to the same human instance across different views and modalities, while separating features from different instances. This design allows correspondence reasoning, semantic aggregation, and instance discrimination to be performed natively in the 3D space, which enforces cross-view consistency and improves robustness under severe occlusions. Based on the learned instance-aware 3D representation, we recover structured human body models in a feed-forward manner by regressing SMPL parameters from instance-level 3D human tokens. Extensive experiments demonstrate that adopting a unified human-aware 3D feature space as the core representation leads to robust, accurate, and efficient multi-view human reconstruction in challenging real-world scenarios. The code and data will be made publicly available.

Keywords Unconstrained 3D reconstruction · Multi-person pose and shape estimation · Human-scene interaction

Communicated by Rama Chellappa.

Yuanwang Yang and Buzhen Huang contributed equally to this work.

✉ Kun Li
lik@tju.edu.cn

Yuanwang Yang
yyw@tju.edu.cn

Buzhen Huang
hbz@tju.edu.cn

Zongxuan Ren
xuan591847902@163.com

Jing Huang
hj00@tju.edu.cn

¹ College of Intelligence and Computing, Tianjin University,
135 Yaguan Road, Tianjin 300350, Tianjin, China

1 Introduction

Reconstructing 3D humans from multiple camera views is a long-standing problem with wide applications in VR/AR, sports analysis, and human-computer interaction. Although extensive research has been devoted to this topic, achieving robust and efficient reconstruction in complex real-world scenarios remains challenging. Traditional multi-view systems (Dong et al., 2021; Ionescu et al., 2013; Joo et al., 2015; Tu et al., 2020) rely heavily on accurate camera calibration and well-controlled environments, making them fragile under practical conditions such as imperfect setups, heavy occlusions, or large-scale scenes (Fig. 1).

The majority of existing multi-view human reconstruction approaches (Dong et al., 2021; Huang et al., 2023; Zhang et al., 2021) follow a bottom-up “2D-to-3D” paradigm. They first extract 2D cues (e.g., keypoints, silhouettes,

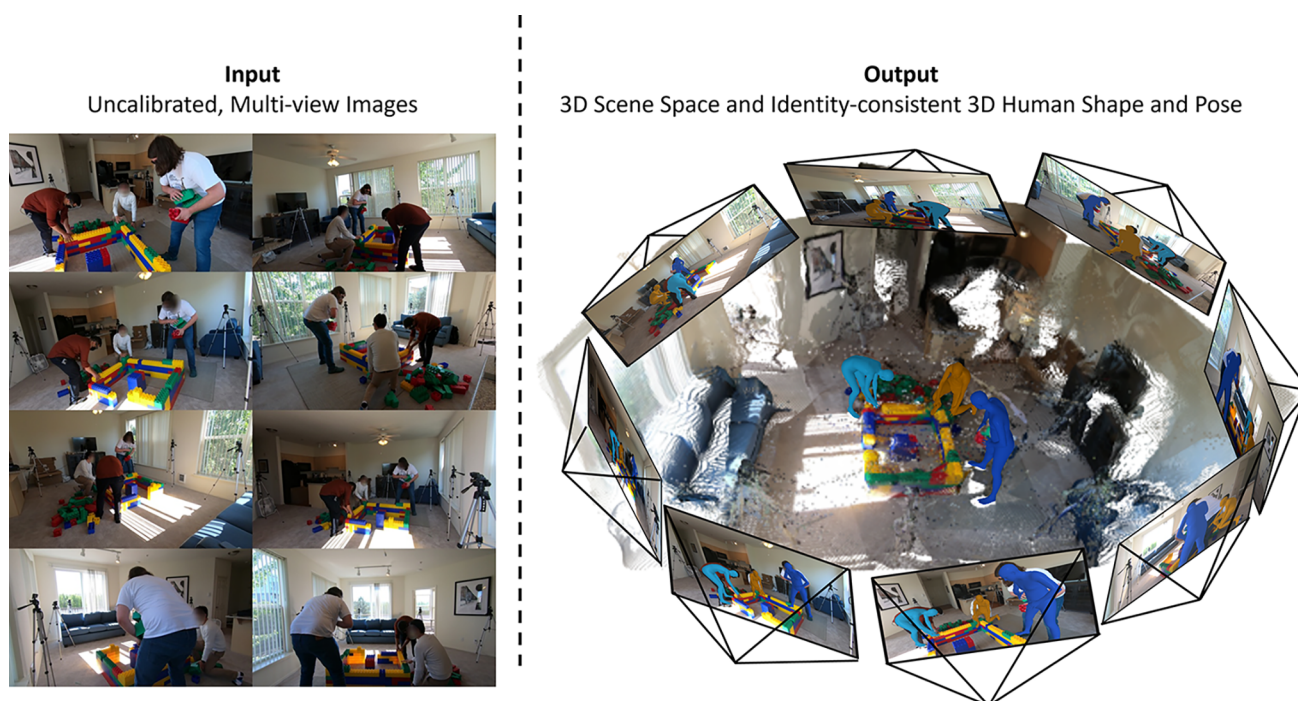


Fig. 1 We propose a unified, human-aware 3D representation learned via cross-modal contrastive learning. This representation provides a persistent and view-agnostic 3D space in which camera calibration, cross-view association, and human reconstruction can be jointly performed

or per-view body parameters) from each image, establish cross-view correspondences, and then fuse these observations into a 3D representation. As one of the dominant approaches, triangulation-based methods directly associate and fuse multi-view 2D observations with camera geometry, but minor inaccuracies in camera parameters can propagate and result in global geometric inconsistencies (Dong et al., 2021; Huang et al., 2020; Tu et al., 2020; Zhang et al., 2021). Although some works (Huang et al., 2023, 2024; Müller et al., 2025; Yu et al., 2022) relax the reliance on precise camera geometry with semantic association, they remain inherently vulnerable to occlusion, as missing or corrupted 2D observations directly undermine correspondence estimation and subsequent feature fusion. In addition, these methods often depend on iterative optimization or matching procedures, which are computationally expensive and difficult to scale to complex scenes or real-time applications.

These challenges are not merely implementation issues, but are inherent to the bottom-up paradigm, which reasons about 3D structure through fragmented and view-dependent 2D observations. In contrast, humans and objects exist and interact in a shared 3D space, and a top-down approach enforces multi-view consistency by reasoning directly within a unified 3D representation rather than across independent image views. This motivates a shift in perspective: instead of treating 3D reconstruction as the result of fusing 2D predictions, we advocate for maintaining an explicit, shared 3D

representation as the central reasoning space throughout the reconstruction process.

In this work, we propose a new top-down paradigm that maintains a unified human-aware 3D space to enable simultaneous camera calibration, cross-view association, and human reconstruction. With this 3D space, multi-view information is lifted and fused at the feature level, where geometric structure, appearance cues, and human-centric semantics are jointly encoded in a shared representation. To enhance the discriminability and consistency of the 3D representation, a spatial contrastive learning strategy is further introduced. Consequently, this 3D feature space serves as a persistent, view-agnostic set of spatial tokens, and thus cross-view association, occlusion handling, and instance reasoning can be performed natively within the 3D space.

Specifically, given uncalibrated multi-view images, we initialize the 3D space with a geometry-aware prior (Wang et al., 2025) and subsequently finetune the representation through differentiable 3D Gaussian Splatting (3DGS) rendering (Kerbl et al., 2023). After initialization, the 3D representation encodes only limited human prior knowledge and lacks explicit instance discriminability. To this end, we further introduce a spatial contrastive learning strategy to improve the consistency and discriminability of the 3D feature space. In particular, human points in the 3D space are re-projected into pose, appearance, and geometry fields to sample expressive human semantics. We then apply contrastive learning across sampled features from different views

and instances to enforce semantic consistency. Meanwhile, the sampled features are used to regress identity embeddings, which are further contrasted across the 3D and 2D spaces. Spatial contrastive learning not only enriches the human prior knowledge within the 3D space, but also improves geometric alignment (*e.g.*, camera parameters) by enforcing semantic consistency. With the learned 3D representation, structured SMPL parameters (Loper et al., 2023) are finally regressed in a feed-forward manner from instance-aware 3D human tokens.

In summary, our main contributions are as follows:

- We propose a novel multi-view human reconstruction framework that explicitly maintains a unified 3D feature space as the central representation, which enables feed-forward and simultaneous calibration, association and reconstruction.
- We design a human-aware 3D representation and refinement strategy that integrates multi-view geometry, appearance, and human semantics at the representation level, thereby substantially improving robustness to occlusion and camera calibration errors.
- We introduce a spatial contrastive learning strategy that leverages semantic consistency to enhance geometric alignment and reconstruction accuracy in multi-view, multi-person scenarios.

2 Related Work

2.1 Human Reconstruction from Unconstrained Cameras

Real-world deployments often involve unknown, unsynchronized, or distorted cameras, as well as sparse and wide-baseline viewpoints. The key challenge is that uncalibrated cameras prevent direct enforcement of multi-view geometric consistency, so both camera parameters and 3D humans must be inferred from weak and noisy cues; this coupling is particularly brittle under occlusion, wide baselines, distortion, and unsynchronized viewpoints. To deal with these unconstrained settings, several methods (Javerliat et al., 2025; Li et al., 2024; Müller et al., 2025; Xu & Kitani, 2022; Yu et al., 2022) jointly estimate camera parameters and 3D human poses. HSfM (Müller et al., 2025) uses SMPL-derived anthropometric constraints as metric references and jointly optimizes scene structure, human meshes, and camera parameters through bundle adjustment, addressing both scale ambiguity and missing calibration. Yu et al. (2022) map per-view image features to a canonical human surface that is invariant to viewpoint, pose and fuse them via confidence-weighted attention, enabling reconstruction without camera calibration. Additional approaches (Li et al., 2024) incor-

porate depth or 3D re-identification cues for cross-view association but require extra sensors. EasyRet3D (Yin et al., 2025) addresses uncalibrated multi-view multi-human reconstruction by explicitly modeling cross-view identity association through a tracking-based formulation. However, it introduces a dependency on explicit tracking quality and can be sensitive to association errors in crowded or heavily occluded scenes. Javerliat et al. (2025) focuses on self-calibration and geometric recovery by jointly estimating camera intrinsics/extrinsics and 3D structure from weak observations, leveraging robust sampling, temporal consistency, and anthropometric priors to resolve scale and depth ambiguities. Although these methods demonstrate the feasibility of calibration-free reconstruction, they still rely on accurate 2D detections, robust cross-view matching, or reliable dense semantic correspondences.

2.2 Multi-view 3D Human Pose and Shape Estimation

As a foundational component of 3D human reconstruction, estimating pose from multiple views has been extensively studied (Chen et al., 2024). Early approaches (Chen et al., 2024) follow a bottom-up pipeline: 2D poses are detected independently per view and subsequently triangulated or fitted by kinematic models to recover 3D joint locations. Such pipelines are highly sensitive to 2D detection noise and often fail under heavy occlusion or multi-person interaction. To mitigate errors introduced during 2D association, voxel-based methods (Luo et al., 2025; Tu et al., 2020) aggregate multi-view features into a unified 3D space and perform inference directly on volumetric representations, achieving improved robustness at the cost of substantial computation and memory. Transformer-based methods further enable end-to-end multi-view reasoning. For instance, MvP (Zhang et al., 2021) leverages learnable joint queries and projective attention to fuse cross-view geometry, while MV-SSM (Chharia et al., 2025) introduces state-space modeling to enhance generalization to unseen camera layouts. Meanwhile, multi-view human mesh reconstruction has been widely studied, particularly in controlled environments where intrinsic and extrinsic camera parameters are known (Furukawa & Ponce, 2009; Hu et al., 2023; Starck & Hilton, 2007). Classical optimization-based methods extend single-view body models (*e.g.*, SMPL/SMPL-X (Loper et al., 2023)) by jointly refining camera poses, keypoints, silhouettes, and mesh parameters, achieving metrically consistent meshes when calibration and initialization are reliable (Bogo et al., 2016; Li et al., 2021). More recent learning-based approaches aggregate multi-view features to directly regress a coherent mesh (Baradel et al., 2024).

2.3 Feed-Forward 3D Reconstruction for Scene and Humans

Feed-forward 3D reconstruction (Cabon et al., 2025; Elflein et al., 2025; Jang et al., 2025; Li et al., 2025; Liu et al., 2025; Wang et al., 2025; Yang et al., 2025) has recently emerged as an efficient method for predicting dense 3D structures from images directly, without relying on iterative optimization. VGGT introduces a feed-forward multi-view geometry framework that directly predicts camera parameters together with dense geometric outputs (e.g., depth and point-based 3D representations) from one to many images in a single forward pass, avoiding expensive test-time optimization. Building on this capability, recent works increasingly treat VGGT-style predictors as a front-end to bootstrap geometry and camera estimates, and then plug these initialization signals into downstream tasks such as rendering, tracking, interaction understanding, or task-specific refinement. Meanwhile, recent work has begun to reconstruct humans, cameras, and the surrounding scene in a shared metric coordinate system, rather than treating human mesh recovery and scene reconstruction as separate problems. HSfM (Müller et al., 2025) tackles this problem from sparse, uncalibrated multi-view images by jointly reconstructing human meshes, the surrounding scene, and cameras, using human observations to stabilize camera estimation and scene geometry. SynCHMR (Zhao et al., 2024) addresses monocular videos by combining human motion recovery with a SLAM-style formulation to estimate metric camera motion, scene geometry, and a consistent human trajectory in a unified coordinate system. Liu et al. (2025) performs optimization-based 4D human–scene reconstruction from monocular videos by leveraging human–scene contact constraints to jointly refine human motion, camera poses, and dense scene geometry, while its distilled variant JOSH3R (Liu et al., 2025) enables efficient feed-forward inference trained from JOSH pseudo-labels. Chen et al. (2025) builds on CUT3R (Wang et al., 2025) and performs feed-forward human–scene reconstruction from monocular videos, jointly predicting multi-person SMPL-X bodies, camera motion, and dense scene geometry in a shared world coordinate system. Recently, Rojas et al. (2025) extends Leroy et al. (2024) with human-aware heads and performs fully feed-forward reconstruction, and it predicts dense 3D point maps directly from sparse views, enabling efficient joint recovery of the scene and people. Successfully combines feed-forward 3D reconstruction with joint optimization of the scene and human bodies. In contrast, we construct a unified implicit 3D embedding space that jointly models pose, appearance, and spatial relationships in an end-to-end manner. This enables direct cross-view correspondence learning and feature fusion in 3D space, substantially improving the robustness of multi-human SMPL reconstruction under unconstrained settings.

3 Method

As shown in Fig. 2, we present a human-aware multi-view 3D reconstruction framework that explicitly maintains and optimizes a unified 3D feature space as an intermediate representation. Instead of reasoning directly in the 2D image domain or regressing parametric body models from per-view features, our approach encodes geometric structure, semantic information, and human identity within this shared 3D space. This design allows camera calibration, cross-view association, and human reconstruction to be addressed jointly in a feed-forward pipeline. Details of the proposed framework are described in the following sections.

3.1 Human-Aware Unified 3D Feature Space Initialization

Directly operating on 2D features is inherently fragile under occlusions, viewpoint changes, and inter-view misalignment. While multi-view aggregation alleviates some of these issues, ambiguity remains when reasoning purely in image space. We therefore lift image features into 3D and maintain an explicit spatial representation, allowing information to be aggregated, filtered, and refined in a geometry-aware manner, which naturally mitigates these issues.

3.1.1 Construction of 3D Feature Space

As shown in Fig. 2 (a), given a set of uncalibrated images $\{I_v\}_{v=1}^V$ from V views, we first construct a view-consistent 3D feature tokens:

$$\mathcal{T} = \{(\mathbf{p}_i, \mathbf{f}_i)\}_{i=1}^N, \quad (1)$$

where $\mathbf{p}_i \in \mathbb{R}^3$ denotes a 3D location and $\mathbf{f}_i \in \mathbb{R}^C$ represents a learned 3D feature embedding. N is the number of points and varies with the complexity of the scene.

To initialize the 3D feature space, we employ a pretrained VGGT model (Wang et al., 2025) as a geometry-aware prior. Given each input image I_v , VGGT predicts a depth map D_v , a per-pixel confidence map C_v , a dense feature map F_v and per-view camera parameters p_v :

$$(D_v, C_v, F_v, p_v) = \Phi_{\text{VGGT}}(I_v). \quad (2)$$

Using the predicted depth and camera parameters, each pixel $\mathbf{u} \in \mathbb{R}^2$ is unprojected into 3D space:

$$\mathbf{x}_v(\mathbf{u}) = \pi^{-1}(\mathbf{u}, D_v(\mathbf{u}); p_v). \quad (3)$$

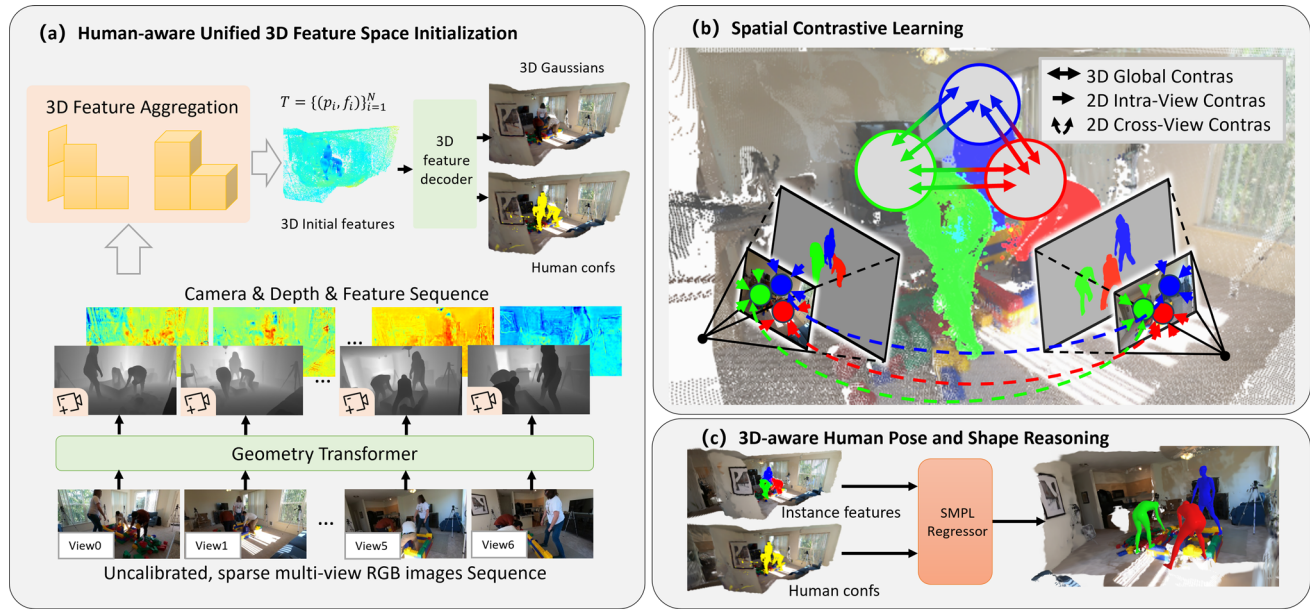


Fig. 2 Overview. Given a set of multi-view unconstrained images, our method first initializes a 3D space using a geometry-aware prior and 3D Gaussian Splatting (a). To further enhance human prior knowledge and instance-level discriminability for the 3D representation, spatial

contrastive learning is applied by sampling across different semantic fields (b). Finally, camera and human representations are jointly optimized within the 3D space and regressed in a feed-forward fashion (c)

Thus, the image features $F_v(\mathbf{u})$ and corresponding confidence weights $C_v(\mathbf{u})$ can be mapped into 3D space.

$$\mathbf{f}_v(\mathbf{u}) = F_v(\mathbf{u}), \quad (4)$$

$$c_v(\mathbf{u}) = C_v(\mathbf{u}). \quad (5)$$

This process yields a dense, view-dependent 3D representation, but directly operating on it is computationally expensive and prone to errors from noisy depth estimation and redundant observations. To obtain a more stable and compact representation, we aggregate neighboring 3D points from multiple views using a confidence-weighted soft aggregation scheme, where aggregation weights are computed as

$$w_{ij} = c_i \cdot c_j \cdot \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \quad (6)$$

where σ is a fixed hyper-parameter controlling the spatial aggregation range, and the fused point center $\mathbf{p}_i$ and feature $\mathbf{f}_i$ are then defined as

$$\mathbf{p}_i = \frac{\sum_j \tilde{w}_{ij} \mathbf{x}_j}{\sum_j \tilde{w}_{ij}}, \quad \mathbf{f}_i = \frac{\sum_j \tilde{w}_{ij} \mathbf{f}_j}{\sum_j \tilde{w}_{ij}}. \quad (7)$$

In contrast to voxelization (Luo et al., 2025; Zhang et al., 2022), our representation operates directly on continuous 3D

points using soft spatial aggregation, avoiding explicit discretization into fixed voxel grids. This formulation is more flexible with respect to varying scene scales and irregular point distributions. With the aggregation, the 3D feature space is view-agnostic and spatially anchored, forming a robust 3D feature tokens that integrate multi-view evidence while suppressing noisy or inconsistent observations.

3.1.2 Human-Aware 3D Feature Decoding

Since the 3D features from VGGT only describe geometry cues, we then explicitly encode human-related semantics into the 3D feature space. Each fused 3D feature token $(\mathbf{p}_i, \mathbf{f}_i)$ serves as a shared latent representation from which multiple attributes are decoded. Specifically, a lightweight MLP g_θ maps $\mathbf{f}_i$ to Gaussian attributes (Kerbl et al., 2023):

$$(\rho_i, \mathbf{s}_i, \mathbf{q}_i, \mathbf{h}_i^{\text{SH}}, z_i) = g_\theta(\mathbf{f}_i), \quad (8)$$

where ρ_i denotes density, $\mathbf{s}_i$ scale, $\mathbf{q}_i$ rotation quaternion, $\mathbf{h}_i^{\text{SH}}$ spherical harmonic coefficients, and z_i is a human-related logit.

The Gaussian opacity is obtained via a bounded monotonic mapping:

$$\alpha_i = \psi(\rho_i), \quad (9)$$

while the human confidence is defined as

$$h_i = \sigma(z_i). \quad (10)$$

The human confidence is defined at the 3D level rather than per image. This enables consistent human reasoning across views and prevents conflicting predictions caused by occlusion or partial visibility in individual images. Moreover, both geometric, appearance, and semantic attributes are decoded from the same 3D feature space. This allows the representation to jointly encode scene structure and human-related semantics, rather than treating human segmentation as a separate 2D task.

3.1.3 Initial 3D Space Training

We train the initial 3D space by learning Gaussian parameters and the human confidence field, without introducing instance-level supervision or SMPL regression.

Specifically, each 3D token is associated with a Gaussian primitive parameterized by its position $\mathbf{p}_i$, scale $\mathbf{s}_i$, rotation $\mathbf{q}_i$, opacity α_i , and spherical harmonic coefficients $\mathbf{h}_i^{\text{SH}}$. Given a camera view v , we render color images by splatting these Gaussians onto the image plane. For a pixel $\mathbf{x}$, the rendered color is computed as

$$\hat{\mathbf{I}}_v(\mathbf{x}) = \sum_i \mathbf{c}_i(\mathbf{x}) \alpha_i \mathcal{G}_i(\mathbf{x}) \prod_{j < i} (1 - \alpha_j \mathcal{G}_j(\mathbf{x})), \quad (11)$$

where $\mathcal{G}_i(\mathbf{x})$ denotes the projected Gaussian footprint and $\mathbf{c}_i(\mathbf{x})$ is the view-dependent color obtained from spherical harmonics. The photometric reconstruction loss $\mathcal{L}_{\text{rgb}}$ supervises Gaussian appearance by comparing rendered images with the input views:

$$\mathcal{L}_{\text{rgb}} = \sum_v \sum_{\mathbf{x}} \|\hat{\mathbf{I}}_v(\mathbf{x}) - \mathbf{I}_v(\mathbf{x})\|_1, \quad (12)$$

where $\hat{\mathbf{I}}_v$ denotes the image rendered from Gaussians in view v .

To supervise human confidence in 3D, we render a soft human mask by modulating Gaussian opacity with the predicted confidence:

$$\tilde{\alpha}_i = \alpha_i \cdot h_i. \quad (13)$$

Rasterizing $\tilde{\alpha}_i$ yields a per-view human mask $\hat{M}_v$, which is supervised by available 2D masks M_v using a binary cross-entropy loss:

$$\mathcal{L}_{\text{mask}} = \sum_v \text{BCE}(\hat{M}_v, M_v). \quad (14)$$

We further introduce a regularization term $\mathcal{L}_{\text{reg}}$ to stabilize Gaussian optimization and encourage spatially coherent human confidence. Specifically, we penalize excessively opaque Gaussians, regularize Gaussian scales to avoid degenerate footprints, and enforce local smoothness of the 3D human confidence field:

$$\mathcal{L}_{\text{reg}} = \lambda_\alpha \sum_i \alpha_i^{2+\lambda_s} \sum_i \|\log \mathbf{s}_i\|_1 + \lambda_h \sum_i \sum_{j \in \mathcal{N}(i)} (h_i - h_j)^2, \quad (15)$$

where $\mathcal{N}(i)$ denotes the spatial neighbors of token i (e.g., within a fixed radius in 3D).

3.2 3D Feature Enhancement with Spatial Contrastive Learning

The 3D feature space in Sect. 3.1 provides stable anchors but may lose fine-grained details and lacks explicit instance discriminability. We therefore *enhance* the 3D token features and *learn* instance-discriminative fields with spatial contrastive learning by selectively sampling from more human semantics. Importantly, both enhancement and instance reasoning are defined on the shared 3D token field, which mitigates view-dependent ambiguities from occlusion and truncation.

3.2.1 Human Semantics Sampling

After initializing the 3D space, the resulting 3D features contain only limited human-related information; therefore, we further exploit richer semantic cues from the image domain. Specifically, we adopt ROMP (Sun et al., 2021) features P_v to provide additional pose information, since they encode pixel-wise human pose and body structure cues. The appearance cues from DINOv3 (Siméoni et al., 2025) A_v are also utilized. In addition, since the aggregation adjust the point positions, we further sampling from the VGGT feature maps F_v .

However, not all 3D tokens require semantic enhancement. We only focus on the tokens with high human confidence h_i , which are more likely to lie on human surfaces:

$$\mathcal{H} = \{i \mid h_i > \tau_h\}, \quad (16)$$

where τ_h is a confidence threshold. Restricting the sampling points to $\mathcal{H}$ significantly reduces computational cost and prevents background clutter from contaminating the refinement process. Tokens outside $\mathcal{H}$ retain their original geometric features and continue to contribute to scene reconstruction.

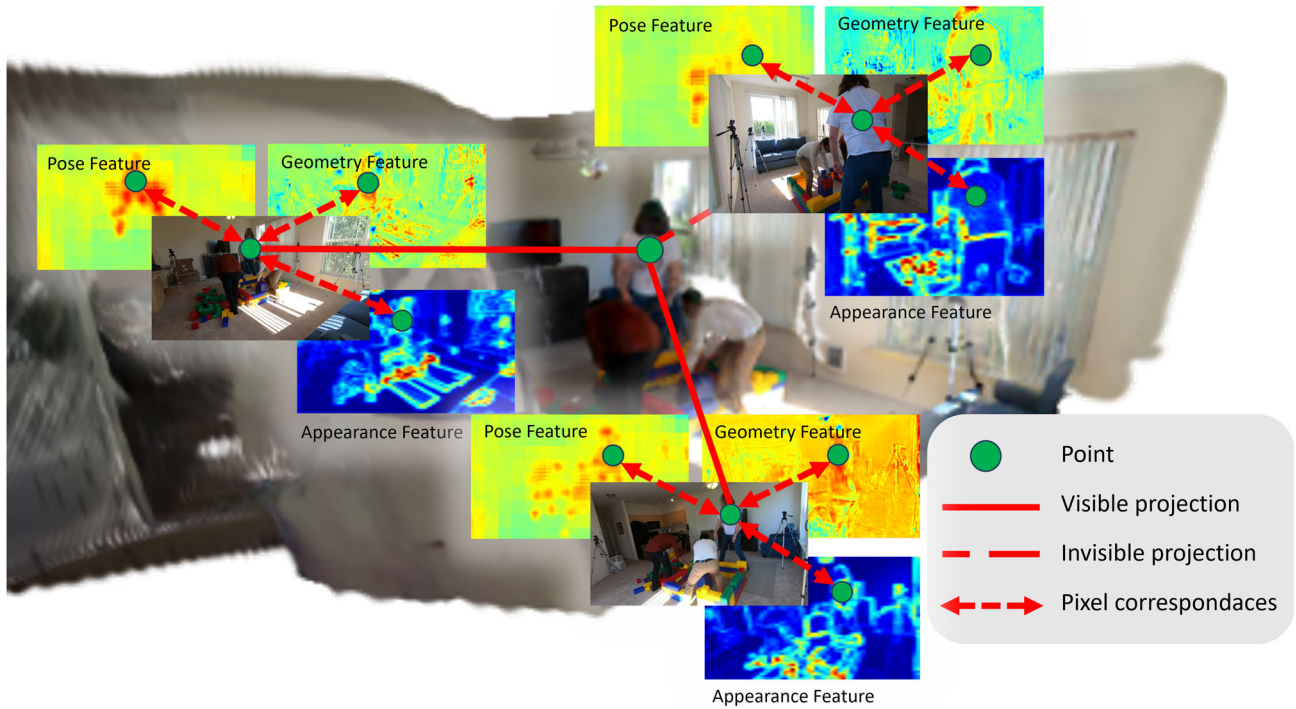


Fig. 3 Human points in the 3D space are projected into different semantic fields to obtain features for spatial contrastive learning

For each selected 3D token ($\mathbf{p}_i, \mathbf{f}_i$) and each camera view v , we project the 3D position $\mathbf{p}_i$ onto the image plane:

$$\mathbf{u}_i^{(v)} = \pi(\mathbf{p}_i; p_v), \quad (17)$$

where $\pi(\cdot)$ denotes the camera projection function defined by the estimated camera parameters.

To ensure reliable feature sampling, we perform visibility-aware filtering. A reprojection is considered valid if it satisfies two conditions: (i) the projected location lies within image boundaries, and (ii) the projected depth is consistent with the predicted depth map.

Formally, the visibility indicator $m_i^{(v)}$ is defined as

$$m_i^{(v)} = \mathbb{I}(\mathbf{u}_i^{(v)} \in \Omega) \cdot \mathbb{I}(|z_i^{(v)} - D_v(\mathbf{u}_i^{(v)})| < \epsilon), \quad (18)$$

where Ω denotes the image domain, $z_i^{(v)}$ is the projected depth of $\mathbf{p}_i$ in view v , $D_v(\cdot)$ is the predicted depth map, and ϵ is a tolerance threshold.

As shown in Fig. 3, by sampling different semantic maps P_v , A_v , and F_v with each valid reprojection point $\mathbf{u}_i$, we then obtain pose, appearance and geometry features $P_v(\mathbf{u}_i)$, $A_v(\mathbf{u}_i)$, and $F_v(\mathbf{u}_i)$.

3.2.2 Cross-Modal Feature Fusion

The sampled features $P_v(\mathbf{u}_i)$, $A_v(\mathbf{u}_i)$, and $F_v(\mathbf{u}_i)$ are concatenated as $\mathbf{g}_i^{(v)}$ and then aggregated across multiple views to refine the original 3D feature. Each view contributes with a confidence weight proportional to its visibility:

$$w_i^{(v)} = \frac{m_i^{(v)}}{\sum_{v'} m_i^{(v')} + \delta}, \quad (19)$$

where δ is a small constant for numerical stability.

The refined 3D feature is computed via a residual fusion formulation:

$$\tilde{\mathbf{f}}_i = \mathbf{f}_i + \sum_v w_i^{(v)} \cdot \phi([\mathbf{f}_i, \mathbf{g}_i^{(v)}]), \quad (20)$$

where $\phi(\cdot)$ is a lightweight fusion network implemented as an MLP, and $[\cdot]$ denotes feature concatenation.

This residual design preserves the stability of the original 3D representation while selectively injecting high-level semantic information. Importantly, fusion is performed independently for each 3D token, enabling scalable and parallel refinement.

By selectively sampling and fusing 2D features through reprojection, we compensate for the insufficient human information and enable the refined 3D feature space to better

align with articulated human structure. Moreover, repeated reprojection and semantic supervision implicitly enforce cross-view consistency, allowing the method to naturally scale to higher-resolution features and additional modalities without modifying the core architecture.

3.2.3 3D Instance Embedding and Spatial Contrastive Learning

Based on the refined 3D features, we further learn instance-aware representations to distinguish multiple humans in the shared 3D space. For each 3D token, we predict a compact instance embedding using a lightweight projection head:

$$\mathbf{e}_i = \frac{h_\phi(\tilde{\mathbf{f}}_i)}{\|h_\phi(\tilde{\mathbf{f}}_i)\|_2}, \quad (21)$$

where h_ϕ denotes an MLP and ℓ_2 normalization enforces a bounded embedding space. The projection head $h_\phi(\cdot)$ is optimized during contrastive learning, while the backbone features $\tilde{\mathbf{f}}_i$ remain fixed. Instance embeddings are only computed for tokens with high human confidence, which suppresses background interference and reduces unnecessary computation.

To learn discriminative embeddings, we adopt a spatial contrastive learning objective that links 3D tokens with 2D instance supervision. Each 3D token is associated with a Gaussian primitive parameterized by its fixed geometric attributes (position, scale, rotation, and opacity) and its instance embedding. During rendering, the instance embedding is treated as the feature carried by the Gaussian and projected onto the image plane via Gaussian splatting. Geometric parameters are detached during this process to prevent degenerate solutions.

Formally, the rendered instance embedding map at pixel $\mathbf{x}$ is given by

$$\mathbf{E}(\mathbf{x}) = \sum_i \mathbf{e}_i \alpha_i \mathcal{G}_i(\mathbf{x}) \prod_{j < i} (1 - \alpha_j \mathcal{G}_j(\mathbf{x})), \quad (22)$$

which softly aggregates contributions from visible 3D tokens along each viewing ray. This differentiable rendering establishes a correspondence between 3D tokens and 2D pixels, allowing instance supervision to be propagated from image space back to the shared 3D representation.

We apply the same contrastive objective under three complementary supervision domains: intra-view, cross-view, and 3D-space contrastive learning. Intra-view contrastive learning is performed within each image to improve local instance discrimination and suppress ambiguity caused by overlapping humans in image space. Cross-view contrastive learning aligns embeddings belonging to the same person across different camera views, improving view consistency

and cross-view association robustness. Finally, 3D-space contrastive learning directly regularizes the shared 3D representation after multi-view feature aggregation, enforcing identity consistency in the unified 3D space. These three forms of supervision are complementary: intra-view supervision improves local separability, cross-view supervision improves identity consistency across viewpoints, while 3D-space supervision stabilizes the globally aggregated instance-aware representation. In all cases, embeddings belonging to the same person are encouraged to be close, while embeddings from different persons are separated. The contrastive loss is defined as

$$\mathcal{L}_{sc} = - \sum_i \log \frac{\exp(\mathbf{e}_i^\top \mathbf{u}_{y_i})}{\sum_k \exp(\mathbf{e}_i^\top \mathbf{u}_{y_k})}, \quad (23)$$

where y_i denotes the instance label and $\mathbf{u}_{y_i}$ is the prototype embedding of instance y , computed as the mean of embeddings assigned to that instance. The same formulation in Eq. (23) is used in all three settings, differing only in how instance labels are obtained.

The final spatial contrastive learning loss is defined as a weighted combination of the three terms:

$$\mathcal{L}_{SCL} = \lambda_{iv} \mathcal{L}_{sc}^{iv} + \lambda_{cv} \mathcal{L}_{sc}^{cv} + \lambda_{3D} \mathcal{L}_{sc}^{3D}, \quad (24)$$

where λ_{iv} , λ_{cv} , and λ_{3D} balance the contributions of intra-view, cross-view, and 3D-space supervision.

By enforcing instance consistency directly in the shared 3D feature space, spatial contrastive learning enables robust multi-person separation across views and provides instance-aware 3D tokens for subsequent SMPL regression.

In addition to instance-level contrastive learning, we further introduce a semantic contrastive objective on $P_v(\mathbf{u}_i)$, $A_v(\mathbf{u}_i)$, and $F_v(\mathbf{u}_i)$ to improve the consistency of semantic representations across views. The key observation is that, for a given person, semantic features extracted from different viewpoints should remain consistent, even under occlusion or partial visibility. Such consistency is essential for learning a coherent 3D feature space that captures both spatial structure and semantic meaning.

Concretely, we treat the refined 3D features as semantic descriptors of the underlying 3D tokens. For each human-related token, its 3D location is projected into multiple camera views, and only valid projections are selected based on visibility. Semantic features sampled from different views but corresponding to the same 3D token and the same identity are encouraged to be similar, while features associated with different identities are pushed apart. The contrastive loss is the same as Eq. (24). This identity-aware contrastive supervision is applied directly to semantic features, rather than instance embeddings.

By enforcing semantic consistency across views at the 3D level, the proposed objective improves the model's ability to extract stable semantic representations over the entire 3D space. As a result, the learned features exhibit stronger view invariance and spatial coherence, which benefits subsequent SMPL regression.

3.2.4 Training with Spatial Contrastive Learning

During the contrastive learning stage, Gaussian parameters and the 3D feature tokens continue to be optimized jointly.

The spatial contrastive learning loss $\mathcal{L}_{\text{SCL}}$ is applied to both instance embeddings and semantic features sampled via multi-view reprojection. Instance-level contrastive supervision encourages 3D tokens belonging to the same person to share similar embeddings, while separating tokens from different individuals. In parallel, semantic contrastive supervision enforces view-consistent semantic representations for the same identity across different viewpoints.

Photometric reconstruction loss $\mathcal{L}_{\text{rgb}}$ is retained throughout this stage to preserve geometric consistency and prevent drift in the underlying 3D structure. By jointly optimizing geometry, appearance, and contrastive objectives, the model learns a discriminative and spatially coherent 3D feature space that supports robust multi-person reasoning.

3.3 SMPL Parameter Regression from Human Tokens

With the refined 3D representation, we then recover structured human body models for each person in the scene. Instead of iterative fitting, we adopt a feed-forward regression formulation that predicts SMPL parameters directly from the learned 3D representation.

3.3.1 Person-Token Pooling with Human-Guided Weighting

Let $\{(\mathbf{p}_i, \tilde{\mathbf{f}}_i)\}_{i=1}^N$ denote the fused 3D tokens, where $\mathbf{p}_i \in \mathbb{R}^3$ is the 3D location and $\tilde{\mathbf{f}}_i \in \mathbb{R}^C$ is the refined feature (Sect. 3.2.2). We also obtain a human confidence score $h_i \in [0, 1]$ for each token and a pseudo instance label $\ell_i \in \{0, 1, \dots, P\}$, where $\ell_i = 0$ indicates background and $\ell_i = p$ indicates the p -th person instance.

We first construct a point descriptor by concatenating 3D features and coordinates, followed by a projection MLP:

$$\mathbf{d}_i = \psi([\tilde{\mathbf{f}}_i; \mathbf{p}_i]) \in \mathbb{R}^T, \quad (25)$$

where $\psi(\cdot)$ denotes a lightweight MLP and T is the token dimension. We then pool a per-person token by confidence-weighted averaging over all tokens associated with the same person:

$$\mathbf{t}_p = \frac{\sum_{i=1}^N \mathbb{I}[\ell_i = p] w_i \mathbf{d}_i}{\sum_{i=1}^N \mathbb{I}[\ell_i = p] w_i + \epsilon}, \quad w_i = h_i^\gamma, \quad (26)$$

where γ is a hyper-parameter controlling how strongly we emphasize high-confidence human tokens, and ϵ is a small constant. This design effectively suppresses noisy or floating points by down-weighting low-confidence tokens.

We additionally keep the accumulated weight

$$s_p = \sum_{i=1}^N \mathbb{I}[\ell_i = p] w_i, \quad (27)$$

which serves as a validity indicator for whether person p is reliably observed and can be used to mask invalid instances during training.

3.3.2 Feed-Forward SMPL Parameter Regression

Given the pooled person token $\mathbf{t}_p$, we regress SMPL parameters with a small MLP head:

$$\mathbf{z}_p = \rho(\mathbf{t}_p), \quad (28)$$

followed by linear predictors for pose, shape, and translation:

$$\hat{\boldsymbol{\theta}}_p = \mathbf{W}_\theta \mathbf{z}_p, \quad \hat{\boldsymbol{\beta}}_p = \mathbf{W}_\beta \mathbf{z}_p, \quad \hat{\mathbf{t}}_p = \mathbf{W}_t \mathbf{z}_p, \quad (29)$$

where $\hat{\boldsymbol{\theta}}_p \in \mathbb{R}^{144}$ denotes the 6D pose parameters, $\hat{\boldsymbol{\beta}}_p \in \mathbb{R}^{10}$ denotes the shape coefficients, and $\hat{\mathbf{t}}_p \in \mathbb{R}^3$ denotes the global translation. Optionally, we also predict a global scale parameter $\hat{s}_p$ (implemented as a positive scalar via *softplus*) to compensate for scale ambiguity in monocular or normalized coordinate systems:

$$\hat{s}_p = \text{softplus}(\mathbf{W}_s \mathbf{z}_p) + \epsilon. \quad (30)$$

The regression outputs define an SMPL mesh for each person instance, enabling downstream tasks such as multi-view rendering, silhouette supervision, and instance-level evaluation without iterative optimization.

3.3.3 Training with SMPL Regression

In the final stage, we train the SMPL regression head while continuing to optimize the Gaussian representation with photometric supervision. Given the pooled person tokens (Sect. 3.3), the network predicts SMPL pose, shape, and translation parameters for each person instance.

The SMPL regression is supervised using a combination of 2D reprojection, 3D joint alignment, and silhouette consistency:

$$\mathcal{L}_{\text{smpl}} = \lambda_{\text{j2d}} \mathcal{L}_{\text{j2d}} + \lambda_{\text{j3d}} \mathcal{L}_{\text{j3d}} + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}}. \quad (31)$$

The 2D reprojection loss $\mathcal{L}_{j2d}$ penalizes the discrepancy between projected SMPL joints and ground-truth 2D key-points across views. When available, the 3D joint loss $\mathcal{L}_{j3d}$ directly supervises predicted SMPL joints in 3D space. In addition, a silhouette loss $\mathcal{L}_{seg}$ is applied by rendering the predicted SMPL meshes into each view and matching them with the corresponding human masks, which helps constrain body extent and resolve depth ambiguities under occlusions.

During this stage, photometric supervision $\mathcal{L}_{rgb}$ is retained to preserve geometric and appearance consistency of the Gaussian representation. This joint optimization enables accurate recovery of structured SMPL parameters from the instance-aware 3D feature space.

4 Experiments

We evaluate the proposed framework on challenging multi-view human reconstruction benchmarks under unconstrained conditions. Our experiments are designed to validate the effectiveness of maintaining a unified human-aware 3D feature space, as well as the contributions of reprojection-based enhancement, spatial contrastive learning, and 3D token-based SMPL regression.

4.1 Datasets

EgoHumans (Khrodar et al., 2023) is a multi-camera, multi-person dataset designed for evaluating human pose estimation and reconstruction in real-world environments. It consists of video sequences captured by multiple synchronized cameras, where two to four individuals perform everyday activities. The dataset covers a wide range of scene scales and layouts, including both indoor and outdoor environments, and exhibits frequent inter-person occlusions, viewpoint variations, and camera placement irregularities. These characteristics make EgoHumans particularly suitable for evaluating the robustness of multi-view reconstruction methods under realistic conditions. Following common practice, we construct train/test splits at the scene level: for each scene, one sequence is randomly selected for testing, while the remaining sequences are used for training.

OcMotion (Huang et al., 2022) is a multi-camera dataset focusing on single-person motion under severe object-induced occlusions. It contains diverse everyday activities where the human body is partially or heavily occluded by surrounding objects, resulting in incomplete and view-dependent observations. OcMotion provides a complementary evaluation setting that emphasizes robustness to dynamic occlusion and missing visual evidence. For OcMotion, we follow the official dataset split as defined in the original benchmark.

4.2 Metrics

We evaluate both human reconstruction accuracy and camera estimation quality using standard metrics.

For human reconstruction, we report **CA-MPJPE** (Camera-aligned MPJPE, meters) first aligns the estimated camera parameters with the ground-truth camera parameters, and then computes MPJPE between the predicted and ground-truth 3D joints. We additionally report **GA-MPJPE** (Procrustes-aligned MPJPE, meters), which aligns the predicted multi-person configuration with the ground-truth multi-person layout as a whole, and evaluates the relative spatial relationships among all individuals. We further report **PA-MPJPE** (Group-aligned MPJPE, meters), which independently applies Procrustes alignment to each predicted person and evaluates per-person pose accuracy for articulated human reconstruction. These metrics together provide a more comprehensive and interpretable evaluation of both local pose accuracy and global multi-person spatial structure.

For camera evaluation, we report **AE** (angular error) and **s-TE** (scale-normalized translation error) to measure rotation and translation accuracy, respectively. We additionally report **s-CCA@10** and **AUC@10**, which evaluate camera consistency within a fixed angular threshold. Lower AE and s-TE, and higher s-CCA@10 and AUC@10 indicate more accurate and consistent camera estimation.

4.3 Results

We compare our method with representative state-of-the-art multi-view reconstruction approaches on the EgoHumans (Khrodar et al., 2023) and OcMotion (Huang et al., 2022) dataset. In particular, we include comparison with HSfM (Müller et al., 2025), a recent framework for jointly reconstructing people, scene structure, and camera parameters from multi-view images. HSfM represents a strong baseline that integrates geometry and human modeling in a unified optimization framework.

Tables 1 and 2 report quantitative results on OcMotion and EgoHumans under different occlusion settings, respectively, while Table 3 presents camera pose estimation accuracy.

Table 1 reports results on OcMotion, which involves challenging single-person scenarios with severe object-induced occlusions. Our method achieves the best CA-MPJPE and competitive PA-MPJPE, outperforming Multi-HMR, DMMR and HSfM in global reconstruction accuracy. Table 2 further evaluates performance on the EgoHumans dataset under non-severely and severely occluded scenarios. In particular, we define severe occlusion as the presence of a subject occluded in more than 50% of the views. All remaining samples are grouped into the non-severely occluded subset. Across all settings, our method consistently achieves

Table 1 Quantitative comparison of multi-view human reconstruction accuracy on OcMotion

Method	CA-MPJPE ↓	PA-MPJPE ↓
Multi-HMR Baradel et al. (2024)	–	0.06
DMMR Huang et al. (2024)	–	0.04
HSfM Müller et al. (2025)	0.35	0.05
Ours	0.19	0.04

Bold values indicate the best result for each metric

Table 2 Quantitative comparison of multi-view human reconstruction accuracy on EgoHumans under non-severely and severely occluded settings

Method	Non-severely occluded			Severely occluded		
	CA-MPJPE ↓	GA-MPJPE ↓	PA-MPJPE ↓	CA-MPJPE ↓	GA-MPJPE ↓	PA-MPJPE ↓
Multi-HMR Baradel et al. (2024)	–	0.85	0.10	–	1.23	0.12
DMMR Huang et al. (2024)	–	0.54	0.23	–	0.81	0.20
HSfM Müller et al. (2025)	0.84	0.34	0.08	1.18	0.53	0.13
Ours	0.80	0.21	0.07	0.89	0.28	0.14

Bold values indicate the best result for each metric

strong CA-MPJPE and GA-MPJPE, indicating robust multi-person spatial reconstruction under complex occlusions. Compared to Multi-HMR and DMMR, our method better preserves global spatial consistency, especially under heavy occlusion where cross-view correspondence becomes highly ambiguous. HSfM shows competitive inter-person structure modeling but suffers from larger absolute reconstruction errors. Overall, our method achieves consistently strong global reconstruction accuracy while maintaining competitive or superior local pose accuracy across both datasets and occlusion levels.

As shown in Table 3, our method consistently achieves substantially lower camera pose errors across all metrics on both EgoHumans and OcMotion. In particular, the improvements in AE and s-TE indicate accurate estimation of absolute camera orientation and translation, while the high s-CCA@10 and AUC@10 scores demonstrate strong cross-view consistency. These results suggest that maintaining a unified 3D feature space enables reliable joint reasoning over camera geometry and scene structure without explicit calibration.

Figure 4 presents qualitative reprojection comparisons. HSfM tends to prioritize relative spatial relationships between individuals, which often leads to inaccurate reprojection results and visible misalignments in the image plane. Multi-HMR produces accurate reprojections in the input views but lacks explicit 3D reasoning, resulting in inconsistent reconstructions across different viewpoints. In comparison, our method produces accurate and consistent reprojections across views, reflecting improved multi-view coherence and pose estimation.

Finally, we visualize the reconstructed 3D scene together with the recovered SMPL bodies in Fig. 5. The results show that our framework is able to reconstruct both scene geome-

try and human bodies in a coherent and physically plausible manner, further validating the effectiveness of maintaining a unified human-aware 3D feature space.

4.4 Runtime and Memory Analysis

We further analyze the computational efficiency of the proposed framework under different numbers of input views, and compare it with a representative optimization-based baseline HSfM.

Since all components in our framework are implemented in a feed-forward regression manner, the overall computational cost is primarily determined by the number of input frames rather than iterative optimization or test-time refinement. This makes the proposed method fundamentally different from optimization-based reconstruction pipelines.

Table 4 reports the runtime breakdown of different stages in our pipeline, including 3D scene reconstruction, instance-level human segmentation, and SMPL regression. We observe that the 3D scene reconstruction stage scales with the number of input views due to additional depth estimation and feature lifting from VGGT. The instance-level human segmentation stage depends on the number of reconstructed human-related 3D points and gradually stabilizes as spatial density converges. The SMPL regression module remains lightweight and relatively stable across different input sizes.

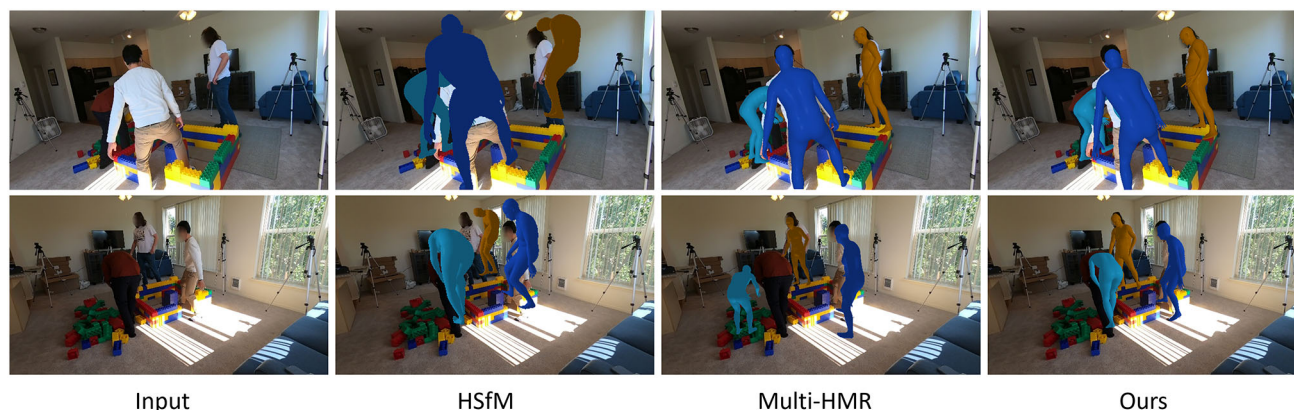
Overall, although runtime increases with more input views, the growth is sub-linear rather than strictly proportional, indicating good scalability in multi-view settings. In contrast, optimization-based methods such as HSfM require iterative refinement, leading to significantly higher computational cost as shown in Table 5.

We further report a comparison of runtime and GPU memory consumption between our method and HSfM. As shown

Table 3 Quantitative comparison of camera pose accuracy on EgoHumans and OcMotion

Method	EgoHumans				OcMotion			
	AE↓	s-TE↓	s-CCA@10↑	AUC@10↑	AE↓	s-TE↓	s-CCA@10↑	AUC@10↑
UnCaliPose Xu and Kitani (2022)	130.68	4.38	0.02	0.23	154.43	4.31	0.01	0.19
HSfM Müller et al. (2025)	29.58	0.77	0.17	0.62	0.97	0.13	0.50	0.80
Ours	0.60	0.03	0.95	0.94	0.36	0.04	0.99	0.94

Bold values indicate the best result for each metric

**Fig. 4** Qualitative comparison of multi-view human reconstruction**Table 4** Runtime breakdown (s) of the proposed method under different numbers of input frames

Input frames	1	2	4	8	16	32
3D scene reconstruction	0.099	0.131	0.219	0.382	0.812	1.886
Instance-level human segmentation	0.064	0.184	0.471	0.501	0.501	0.488
SMPL regression	0.046	0.046	0.048	0.052	0.084	0.180
Total	0.209	0.361	0.729	0.936	1.397	2.554

Table 5 Runtime and peak GPU memory comparison with HSfM under different numbers of input frames

Method	Metric	1	2	4	8
HSfM	Time (s)	170	177	199	220
HSfM	Mem. (GB)	4.62	5.13	6.29	10.79
Ours	Time (s)	0.209	0.361	0.729	0.936
Ours	Mem. (GB)	8.26	8.72	11.20	15.82

in Table 5, our approach achieves substantially lower runtime while maintaining comparable or higher memory efficiency under increasing numbers of input views. This demonstrates that the proposed framework provides a favorable trade-off between reconstruction accuracy and computational efficiency in multi-view scenarios.

4.5 Cross-View Association Evaluation

To further evaluate the quality of cross-view instance association, we further apply the AIDP (Gan et al., 2021) metrics

for cross-view association evaluation. AIDP evaluates the precision of subject association across views by computing pairwise identity matching accuracy over all view pairs and averaging the results. This metric directly reflects the consistency of instance-level correspondence in multi-view settings.

We first compare our method with HSfM under this evaluation protocol. In addition, to further examine whether geometric cues alone are sufficient for reliable association, we construct a VGGT-based Pose-Aware ReID baseline. Specifically, given detected human bounding boxes, we extract keypoint predictions and confidence scores from VGGT, and perform confidence-weighted aggregation to obtain compact pose-centric descriptors. Cross-view identity association is then performed via nearest-neighbor matching based on Euclidean distance.

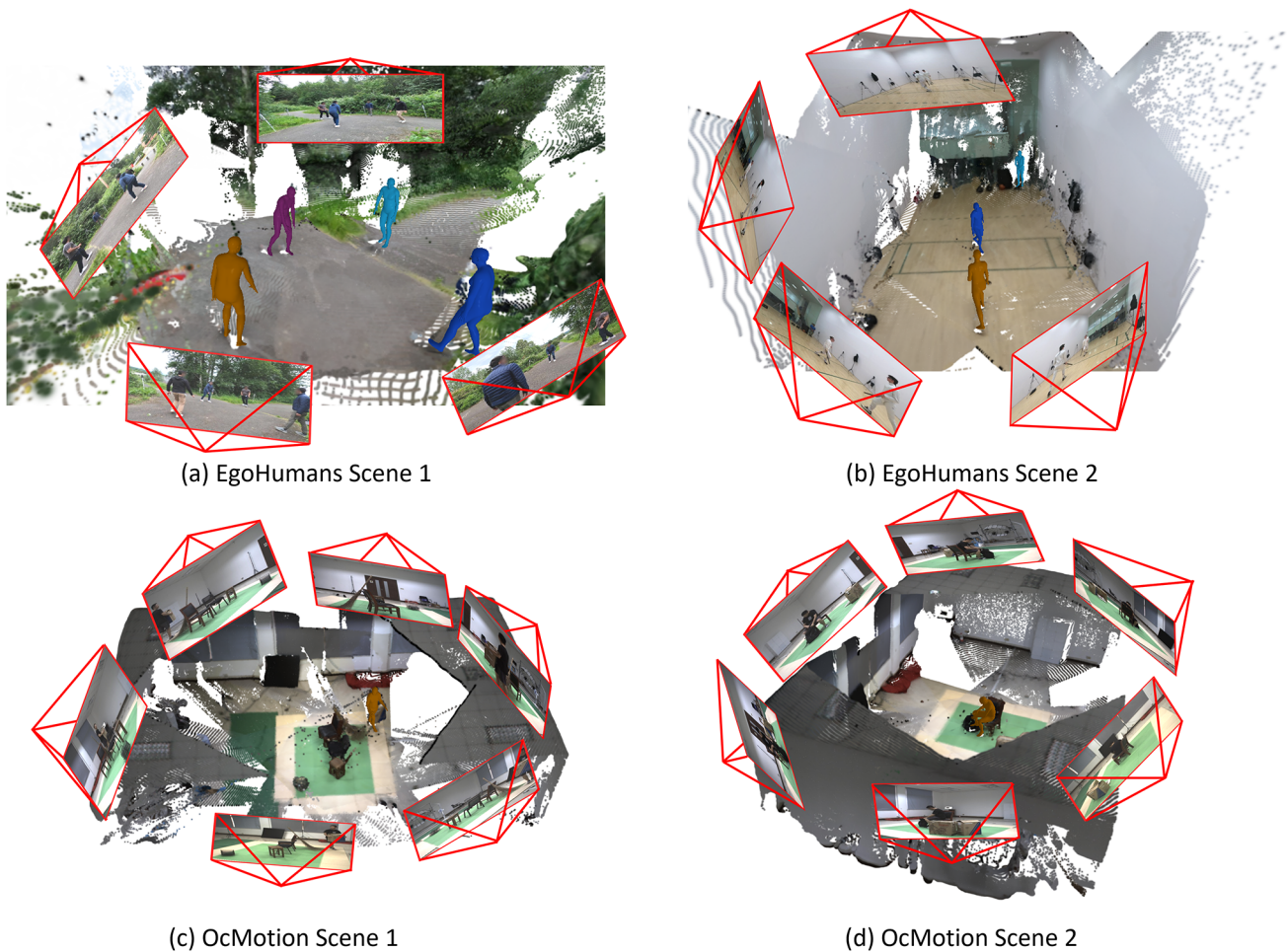


Fig. 5 Qualitative results of the proposed method on EgoHumans and OcMotion. We visualize the reconstructed 3D scene together with the recovered SMPL bodies from multiple views. The results demonstrate that our method produces spatially coherent human reconstructions

Table 6 Cross-view instance association evaluation using AIDP metric. Higher is better

Method	AIDP $\uparrow$
HSfM Müller et al. (2025)	75.68
VGGT-based Pose-Aware ReID	82.02
Ours	98.54

Bold values indicate the best result for each metric

As shown in Table 6, our method achieves the highest AIDP score, outperforming both HSfM and the VGGT-based baseline. This indicates that geometry- or pose-based heuristics can provide reasonable cues for association, but remain limited under challenging multi-view conditions. In contrast, our spatial contrastive learning explicitly enforces instance-level consistency in the unified 3D feature space, leading to more robust cross-view identity alignment.

4.6 Ablation Study

4.6.1 Architecture Ablation

We conduct ablation studies on EgoHumans to analyze the contributions of key components in the proposed framework. In particular, we investigate how different design choices affect multi-view human perception, instance consistency, and SMPL regression from the unified 3D feature space.

Specifically, we consider the following variants. **w/o unified 3D feature space** removes explicit 3D perception and directly regresses human bodies from per-view 2D feature maps. **w/o reprojection enhancement** disables reprojection-based pose feature sampling and regresses SMPL parameters solely from the initial 3D feature space containing geometric cues. **w/o SCL** removes spatial contrastive learning and aggregates features based on 2D instance segmentation without enforcing instance consistency in 3D.

Table 7 summarizes the ablation results. Removing the unified 3D feature space leads to a significant performance

drop. Without explicit 3D perception, the model relies solely on view-dependent 2D features, which are highly sensitive to occlusion and viewpoint variation. As a result, it becomes difficult to associate observations of the same individual across views, leading to degraded reconstruction accuracy.

Disabling reprojection-based enhancement also degrades performance. In this setting, SMPL parameters are regressed only from the initial 3D feature space that mainly encodes geometric information. Without injecting pose-related image features via reprojection, optimization becomes more difficult and convergence is less stable, resulting in inferior reconstruction quality.

Removing spatial contrastive learning further impacts performance. When instance consistency is enforced only through 2D instance segmentation, feature aggregation becomes vulnerable to segmentation noise and view-dependent errors. This leads to contaminated 3D features and reduced robustness, especially in the presence of occlusion and clutter.

4.6.2 Robustness to Geometry Initialization

To further evaluate the robustness of our framework to geometry initialization, we replace the VGGT backbone with AnySplat (Jiang et al., 2025), a representative geometry-aware reconstruction method that provides depth and camera priors with Gaussian Splatting-based refinement.

Both VGGT and AnySplat are geometry-aware reconstruction pipelines that produce compatible initializations of depth, features, and camera parameters, making them suitable for evaluating the sensitivity of our framework to different geometry backbones. Unlike VGGT, which directly predicts geometry and feature representations, AnySplat introduces an additional rendering-based refinement stage, resulting in slightly different camera and point cloud quality. Nevertheless, both methods share a similar underlying geometry foundation, allowing a fair comparison of backbone robustness.

As shown in Table 8, the performance remains highly consistent under both initialization strategies, with only marginal differences across all evaluation metrics. This indicates that our method is largely robust to the choice of geometry initialization. The results further suggest that our method is robust to the choice of geometry initialization, and benefits primarily from the unified 3D feature representation and spatial contrastive learning.

4.6.3 Analysis of 3D Human Awareness

The limitations of relying solely on 2D instance segmentation further motivate a deeper analysis of human perception in 3D. Although 2D segmentation-based aggregation can achieve high coverage and include most human-related features, it

exhibits structural weaknesses when applied to multi-view 3D reconstruction.

In practice, segmentation errors near object boundaries and occlusion regions are common. When such masks are lifted into 3D, these inaccuracies lead to enlarged selection regions and ray-like leakage into the background, contaminating the aggregated 3D features. Moreover, 2D instance segmentation does not naturally provide consistent instance identities across views. Associating masks belonging to the same person from different viewpoints requires additional cross-view matching, which is highly sensitive to occlusion, truncation, and segmentation noise. Inconsistent instance assignments across views further degrade feature aggregation and undermine multi-view coherence.

These observations indicate that accurate human perception in multi-view settings cannot be reliably achieved by inheriting instance information from 2D segmentation alone. Instead, human awareness and instance consistency should be learned directly in the shared 3D feature space, where multi-view observations can be aggregated and regularized in a geometry-aware manner. In the following, we further validate this design choice.

On a single-person setting with per-view masks and ground-truth SMPL models, we compare the proposed learned 3D human confidence with a non-learned baseline that fuses 2D masks into 3D via visibility-aware averaging. Ground-truth human labels for 3D tokens are derived based on their distance to the SMPL surface.

Table 9 reports token-level Precision, Recall, F1 score, and PR-AUC. The 2D mask fusion baseline achieves very high recall but suffers from extremely low precision, indicating severe background leakage. In contrast, the learned 3D confidence substantially improves precision and F1 score while maintaining competitive recall, and achieves a significantly higher PR-AUC, demonstrating better global separability and calibration.

Figure 6 provides qualitative evidence. Directly lifting 2D masks into 3D introduces ray-like artifacts and background noise around occlusion boundaries, whereas the learned 3D confidence yields a spatially compact and coherent human region. These results indicate that human awareness is not merely inherited from 2D segmentation, but emerges as a spatially consistent property learned in the unified 3D feature space, forming a reliable foundation for subsequent human-guided modules.

5 Conclusion

In this paper, we presented a novel feed-forward framework for multi-view multi-person 3D reconstruction. Departing from conventional pipelines that rely on view-dependent 2D reasoning or tightly coupled optimization, our method

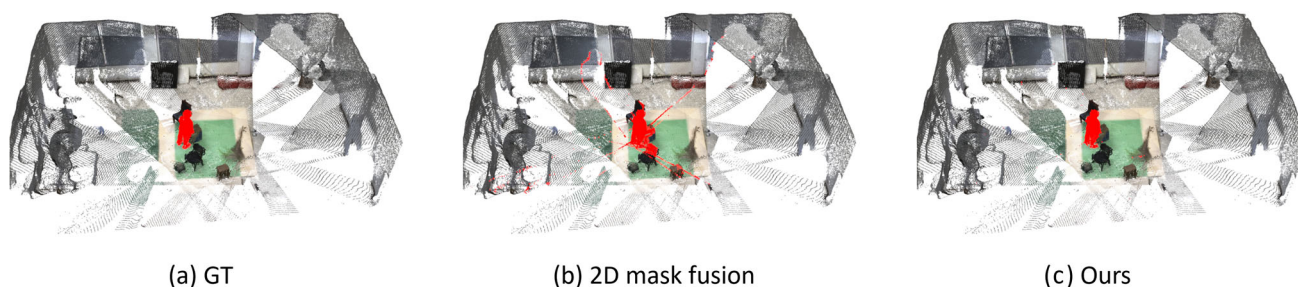
Table 7 Ablation study on key components of the proposed framework on EgoHumans

Method	CA-MPJPE↓	GA-MPJPE↓	PA-MPJPE↓
w/o unified 3D feature space	3.12	3.28	2.53
w/o reprojection enhancement	1.28	1.42	1.74
w/o SCL	1.94	0.49	0.37
Ours	0.95	0.30	0.18

Table 8 Robustness to geometry initialization

Geometry initialization	CA-MPJPE↓	GA-MPJPE↓	PA-MPJPE↓
w/ AnySplat Jiang et al. (2025)	1.02	0.37	0.19
VGGT (default)	0.95	0.30	0.18

We compare VGGT-based initialization with an alternative AnySplat-based initialization

**Fig. 6** Qualitative comparison of 3D human confidence fields. While 2D mask fusion covers most human regions, it introduces substantial false positives in background and occluded areas. In contrast, the learned

3D confidence produces a spatially compact and clean human region with significantly reduced background noise

Table 9 Comparison between learned 3D human confidence and 2D mask fusion on the single-person dataset

Method	Precision ↑	Recall ↑	F1 ↑	PR-AUC ↑
2D mask fusion	0.420	0.989	0.590	0.006
Ours (3D learned)	0.886	0.825	0.855	0.863

Evaluation is performed at the token level using ground-truth SMPL supervision

Bold values indicate the best result for each metric

explicitly maintains a shared 3D representation where geometry, semantics, and human identity are jointly encoded and progressively refined. By initializing the 3D space with a geometry-aware prior and 3D Gaussian representation, the framework establishes stable spatial anchors under unconstrained camera settings. Building upon this foundation, spatial contrastive learning enhances the discriminability and consistency of the 3D features by leveraging multi-view semantic cues, enabling robust multi-person separation and cross-view association in the presence of occlusion and partial visibility. Structured human body models are then recovered in a feed-forward manner directly from instance-aware 3D tokens, avoiding iterative fitting and allowing camera parameters and human representations to be jointly optimized. Extensive experiments on challenging multi-

view benchmarks demonstrate that the proposed approach achieves competitive or superior human reconstruction accuracy while significantly improving camera estimation quality.

While the present work focuses on static multi-view reconstruction, the proposed framework can be naturally extended to 4D dynamic settings by incorporating temporal association of 3D tokens across sequential frames. In particular, the instance-aware 3D representation provides a structured and identity-consistent basis that can naturally support temporal correspondence, enabling potential extensions to articulated motion and moving-camera scenarios. We further note that the current formulation does not explicitly model temporal evolution, and therefore future work may explore temporal update mechanisms such as token propagation, merging, and splitting within the same 3D Gaussian representation.

We also acknowledge that our framework relies on VGGT as a geometry-aware foundation model for initializing depth, features, and camera parameters. While VGGT provides strong generalization for constructing the initial 3D feature space, its performance is not perfect and may be limited in cases of large viewpoint gaps, low-texture regions, or challenging depth discontinuities, which can affect downstream reconstruction quality. Nevertheless, our method is not tightly coupled to VGGT itself and operates on the constructed 3D feature representation. Therefore, it can be

naturally integrated with alternative or improved geometry foundation models in future work.

We believe that extending the framework toward full 4D reconstruction and improving robustness under more challenging geometry initialization remain important and promising directions for future research.

Funding This work was supported in part by National Key R&D Program of China (2023YFC3082100), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2026C04), and Science Fund for Distinguished Young Scholars of Tianjin (No.22JCJCJC00040).

Data availability The datasets used in this study are publicly available. Detailed descriptions and corresponding references are provided in Sect. 4.1.

References

- Baradel, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., & Lucas, T. (2024). Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision* (pp. 202–218). Springer.
- Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., & Black, M. J. (2016). Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision* (pp. 561–578). Springer.
- Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., & Leroy, V. (2025). Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 1050–1060).
- Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., & Pons-Moll, G. (2025). Human3r: Everyone everywhere all at once. arXiv preprint [arXiv:2510.06219](https://arxiv.org/abs/2510.06219).
- Chen, X., Wang, L., Wang, Q., & Tang, J. (2024). 3d human pose estimation and tracking based on feedback mechanism. In *2024 7th International Conference on Intelligent Robotics and Control Engineering (IRCE)* (pp. 292–296). IEEE.
- Chharia, A., Gou, W., & Dong, H. (2025). Mv-ssm: Multi-view state space modeling for 3d human pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 11590–11599).
- Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., & Zhou, X. (2021). Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 6981–6992.
- Elflein, S., Zhou, Q., & Leal-Taixé, L. (2025). Light3r-sfm: Towards feed-forward structure-from-motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 16774–16784).
- Furukawa, Y., & Ponce, J. (2009). Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8), 1362–1376.
- Gan, Y., Han, R., Yin, L., Feng, W., & Wang, S. (2021). Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 282–290).
- Hu, S., Hong, F., Pan, L., Mei, H., Yang, L., & Liu, Z. (2023). Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9352–9364).
- Huang, B., Ju, J., Shu, Y., & Wang, Y. (2023). Simultaneously recovering multi-person meshes and multi-view cameras with human semantics. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6), 4229–4242.
- Huang, C., Jiang, S., Li, Y., Zhang, Z., Traish, J., Deng, C., Ferguson, S., & Da Xu, R. Y. (2020). End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. *European Conference on Computer Vision* (pp. 477–493). Springer.
- Huang, B., Ju, J., Shu, Y., & Wang, Y. (2024). Simultaneously recovering multi-person meshes and multi-view cameras with human semantics. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6), 4229–4242.
- Huang, B., Shu, Y., Ju, J., & Wang, Y. (2022). Occluded human body capture with self-supervised spatial-temporal motion prior [arXiv:2207.05375](https://arxiv.org/abs/2207.05375) arXiv preprint.
- Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1325–1339.
- Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., & Revaud, J. (2025). Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 1071–1081).
- Javerliat, C., Raimbaud, P., & Lavoué, G. (2025). Kineo: Calibration-free metric motion capture from sparse rgb cameras [arXiv:2510.24464](https://arxiv.org/abs/2510.24464) arXiv preprint.
- Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., & Zhao, F. (2025). Anysplat: Feed-forward 3d Gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6), 1–16.
- Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., & Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3334–3342).
- Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 1–4.
- Khrodgar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., & Kitani, K. (2023). *Ego-humans: An ego-centric 3d multi-human benchmark* (p. ICCV).
- Leroy, V., Cabon, Y., & Revaud, J. (2024). Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)* (pp. 71–91).
- Li, W., Liu, S., Qiao, P., & Dou, Y. (2025). Mono3r: Exploiting monocular cues for geometric 3d reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 11081–11090).
- Li, Z., Oskarsson, M., & Heyden, A. (2021). 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1888–1897).
- Liu, S., Li, W., Qiao, P., & Dou, Y. (2025). Regist3r: Incremental registration with stereo foundation model. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 4484–4493).
- Liu, Z., Lin, J., Wu, W., & Zhou, B. (2025). Joint optimization for 4d human-scene reconstruction in the wild [arXiv:2501.02158](https://arxiv.org/abs/2501.02158) arXiv preprint.
- Li, Y.-J., Xu, Y., Khrodgar, R., Park, J., & Kitani, K. (2024). Multi-person 3d pose estimation from multi-view uncalibrated depth cameras [arXiv:2401.15616](https://arxiv.org/abs/2401.15616) arXiv preprint.
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2023). Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries* (vol. 2, pp. 851–866).

- Luo, J., Xie, S., Quan, T., Ren, X., & Miao, Y. (2025). Voxel-based multi-person multi-view 3d pose estimation in operating room. *Applied Sciences*, 15(16), 9007.
- Müller, L., Choi, H., Zhang, A., Yi, B., Malik, J., & Kanazawa, A. (2025). Reconstructing people, places, and cameras. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 21948–21958).
- Rojas, S., Armando, M., Ghanem, B., Weinzaepfel, P., Leroy, V., & Rogez, G. (2025). Hamst3r: Human-aware multi-view stereo 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 5027–5037). Authors: Sara Rojas; Matthieu Armando; Bernard Ghanem; Philippe Weinzaepfel; Vincent Leroy; Grégory Rogez.
- Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., & Massa, F. (2025). Dinov3. arXiv preprint [arXiv:2508.10104](https://arxiv.org/abs/2508.10104).
- Starck, J., & Hilton, A. (2007). Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3), 21–31.
- Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M. J., & Mei, T. (2021). Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 11179–11188).
- Tu, H., Wang, C., & Zeng, W. (2020). Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. *European Conference on Computer Vision* (pp. 197–212). Springer.
- Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., & Novotny, D. (2025). Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 5294–5306).
- Wang, Q., Zhang, Y., Holynski, A., Efros, A. A., & Kanazawa, A. (2025). Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 10510–10522).
- Xu, Y., & Kitani, K. (2022). Multi-view multi-person 3d pose estimation with uncalibrated camera networks. In *Proceedings of the 33rd British Machine Vision Conference (BMVC)*, 132.
- Yang, J., Sax, A., Liang, K. J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., & Feiszli, M. (2025). Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 21924–21935).
- Yin, J. O., Li, T., Wang, J., Zhang, Y., & Yuille, A. (2025). Easyret3d: Uncalibrated multi-view multi-human 3d reconstruction and tracking. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 3128–3137). IEEE.
- Yu, Z., Zhang, L., Xu, Y., Tang, C., Tran, L., Keskin, C., & Park, H. S. (2022). Multiview human body reconstruction from uncalibrated cameras. *Advances in Neural Information Processing Systems*, 35, 7879–7891.
- Zhang, J., Cai, Y., Yan, S., & Feng, J. (2021). Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34, 13153–13164.
- Zhang, Y., Wang, C., Wang, X., Liu, W., & Zeng, W. (2022). Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), 2613–2626.
- Zhao, Y., Wang, T. Y., Raj, B., Xu, M., Yang, J., & Huang, C.-H. P. (2024). Synergistic global-space camera and human reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1216–1226).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.