4D World Reconstruction of Humans, Scenes, and Camera Systems

by

Jerrin Bright

A research proposal
presented to the University of Waterloo
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2025

Abstract

4D world reconstruction, which aims to recover humans, scenes, and camera systems within a unified and metrically consistent world frame, is a foundational challenge in computer vision. Most existing pipelines reconstruct these components independently, which often leads to scale drift, depth ambiguity, and physically inconsistent human–scene interactions. These issues manifest as artifacts such as floating bodies, foot sliding, or interpenetration with the environment, and they are particularly severe in cluttered scenes involving occlusions and non-planar support surfaces.

Recent progress in global human mesh recovery has improved per-frame pose and shape estimation, but physical grounding is often weak because the scene is treated as a sparse tracking signal or a simplified support surface. Conversely, modern dense scene reconstruction methods can recover detailed geometry, yet they frequently struggle in the presence of dynamic humans and lack human-centric cues that would otherwise help resolve metric scale and occluded structure. As a result, human and scene reconstructions remain poorly coupled: humans do not reliably constrain scene geometry, and scenes do not consistently enforce physically plausible human placement.

This proposal is motivated by a **unified Human–Scene–Camera Reconstruction (HSR)** paradigm, in which humans, scenes, and cameras are estimated jointly within a single, metric coordinate system. The core hypothesis is that humans and scenes are mutually informative: human kinematics and anthropometric regularities provide strong cues for resolving scene scale and geometry, while reconstructed scene structure provides metric grounding and physical constraints for plausible human pose and placement.

To realize this paradigm, we propose an end-to-end world reconstruction model centered on a shared *world latent* representation that serves as the fusion space for bidirectional reasoning. The model initializes scene tokens using strong scene priors with state propagation across time, enabling temporally coherent scene features and incremental refinement. In parallel, a multi-human pose model prior extracts compact human embeddings and initial SMPL-X parameters. Rather than applying a single, isolated cross-attention layer, the world latent is updated recurrently by cross-attending to the world latent and propagating the state across frames, producing globally consistent world-frame estimates.

Physical and metric plausibility are further encouraged through *implicit regularization priors*, including gravity alignment, contact consistency, and anthropometric height constraints, which are enforced through dedicated loss terms. Together, these design choices aim to produce a coherent 4D world model with metrically grounded geometry, physically plausible human–scene interaction, and camera consistency, enabling downstream applications in robotics, AR/VR, sports analytics, autonomous navigation, and embodied AI.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Reconstructing a coherent, temporally consistent, and metrically accurate 4D world from images and videos remains one of the most fundamental and unsolved challenges in computer vision. Existing approaches typically address humans, scenes, and cameras in isolation or under restrictive assumptions. Human reconstruction methods often depend on canonical spaces or strong priors, limiting generality. Scene reconstruction methods struggle with dynamic elements, scale drift, and occlusions. Human–scene reconstruction methods improve interaction reasoning but still fall short of producing a unified metric world model that integrates humans, environment geometry, and camera motion.

The main problem addressed in this proposal is the development of a unified system capable of reconstructing the full 4D world, including dynamic humans, static and dynamic scenes, and camera systems, all within a consistent and metrically grounded coordinate frame. This requires resolving depth ambiguities, modeling human–scene interactions, and recovering physically plausible motion. A successful system would enable downstream applications such as controllable simulation, activity understanding, long-horizon prediction, and embodied AI.

1.2 Challenges

1. **Human and Scene Entanglement.** Humans and scenes are tightly coupled. Humans occlude the environment, interact with objects, and introduce complex non-

rigid motion. Scenes impose geometric and physical constraints on feasible human actions. Many existing methods model one component more accurately than the other, which leads to inconsistencies such as interpenetration, floating, or misaligned contact.

2. **Metric Scale Ambiguity.** Monocular and sparse view reconstruction methods often suffer from scale ambiguity or gradual scale drift. Without proper metric grounding, the reconstructed world does not have physical interpretability, limiting practical use in robotics, AR and VR, and world modeling.

3. **Dynamic and Non-Rigid Motion.** Recovering detailed human motion, clothing deformation, and non-rigid object behavior from limited viewpoints is highly underconstrained. Temporal modeling must capture long-term motion without overfitting or accumulating drift.

4. **Fragmented Literature.** Most existing research focuses on either human reconstruction, scene reconstruction, SLAM, neural radiance fields, or generative modeling. Very few approaches aim to unify these components within a single coherent system.

1.3 Motivation and Applications

A unified, world-grounded reconstruction of humans, scenes, and camera systems enables a wide range of impactful applications. Accurate 4D modeling in metric space supports robust perception, realistic simulation, and fine-grained human-environment understanding.

1. **Humanoid Robotics and Embodied AI.** Humanoid robots require accurate models of both their own motion and the surrounding environment to navigate, manipulate objects, and interact with humans. A world-grounded reconstruction system enables robots to recover unknown spaces from onboard cameras and operate with precise global control in unstructured homes, warehouses, and outdoor settings.

2. **Autonomous Driving in Unmapped Regions.** Current autonomous vehicles rely heavily on pre-mapped environments. Systems such as Waymo operate effectively within restricted regions but struggle in unfamiliar or unmapped domains. A world reconstruction method that simultaneously recovers scene geometry, dynamic agents, and camera trajectories allows autonomous vehicles to operate safely even without dense prior maps.

3. **Augmented and Virtual Reality Content Generation.** AR and VR experiences, including live overlays, wayfinding, and interactive replays, require accurate 3D models of both humans and the environment. A unified reconstruction system enables real-time model generation from monocular video, enabling virtual content to be anchored to the physical world with correct scale, occlusion, and physics.

4. **Sports Broadcasting and Immersive Viewing.** In sports, accurate reconstruction of athletes and the environment allows viewers to watch replays from any angle, inspect biomechanics, freeze time, and visualize the scene from an athlete's perspective. This enhances game analysis, training, and spectator engagement without the need for dense multi-camera motion capture setups.

This broad range of applications highlights the importance of developing world-grounded reconstruction techniques and motivates the contributions of this thesis.

1.4 Proposed Contributions

This work advances the state of the art in 4D world reconstruction by introducing a unified framework that jointly recovers humans, scenes, and cameras in metric space. The main contributions are as follows:

1. **Unified Human–Scene–Camera Reconstruction (HSR) formulation.** We formalize 4D world reconstruction as a single coupled inference problem in which humans, scenes, and cameras are estimated jointly, enabling bidirectional reasoning: humans provide constraints and cues for scene geometry and scale, while scenes and cameras provide metric grounding and physical constraints for human motion and placement.

2. **World-latent fusion for cross-modal reasoning.** We propose a shared *world latent* representation that acts as the central fusion space for human and scene embeddings. Instead of treating cross-attention as a single isolated module, the world latent is updated recurrently over time via bidirectional interaction with both (i) scene tokens and propagated scene state and (ii) human tokens, producing globally consistent world-frame estimates.

3. **Strong pretrained priors for scene and human initialization.** We leverage strong pretrained representations to initialize both scene and human estimates: a

stateful scene encoder provides dense scene features with temporal *state propagation* (predicting and passing forward the next state to condition subsequent frames), while a human encoder produces compact human embeddings and an initial parametric body estimate. These initializations are not treated as final outputs; instead, they are refined through interaction with the shared world latent to improve robustness under occlusions, fast motion, and view changes.

4. **Implicit human–scene priors as regularization.** We incorporate human-centric priors, including gravity alignment, probabilistic contact consistency, and anthropometric height constraints, as *implicit regularizers* in the end-to-end objective. These priors are not direct network inputs; instead, they shape the learned world representation and predictions through dedicated loss terms that encourage metric consistency and physically plausible human–scene interaction.

5. **Toward fully controllable, in-the-wild video-based 4D world reconstruction.** Building on the proposed HSR model, the long-term goal is a fully controllable 4D world representation learned from *in-the-wild* monocular videos (including large-scale Internet video), analogous to how modern foundation models in NLP leverage web-scale corpora. Such a controllable world model should support simulation and data generation, including novel-view rendering, editable camera trajectories, and physically consistent human–scene interaction, enabling scalable synthetic data generation, training embodied policies, and closed-loop evaluation in diverse human-centered environments.

1.5 Proposal Outline

The outline of this proposal is as follows. Chapter 2 reviews the existing literature in the field, exploring different human reconstruction, scene reconstruction, and human-scene reconstruction methods. Chapter 3 describes the proposed research framework, explaining in detail the novel backbones proposed to enable accurate world models. Chapter 4 discusses the progress to date, along with some baseline evaluations to show the gap in the literature. Finally, Chapter 5 explains the proposed research contributions, future directions, and the research timeline for the completion of the proposed research.

Chapter 2

Related Works

2.1 3D Human Reconstruction

3D human reconstruction aims to estimate the full 3D geometry of the human body from visual observations such as single or multi-view images, videos, or keypoints. Existing approaches can broadly be categorized into **parametric** and **non-parametric** methods. Parametric approaches rely on statistical body models to represent the human mesh as a function of low-dimensional pose and shape parameters. Non-parametric methods, on the other hand, attempt to directly infer the 3D mesh or vertex coordinates without assuming any predefined model structure.

2.1.1 Parametric Human Modeling

Parametric methods are typically built upon learned human body models such as the Skinned Multi-Person Linear Model (SMPL) [36]. SMPL provides a differentiable function $M(\theta, \beta) \in \mathbb{R}^{3\times 6980}$ that maps pose parameters θ and shape parameters β to a triangulated 3D body mesh. Here, θ represents the relative 3D rotations of 23 joints, and β denotes the 10 principal shape coefficients that capture inter-person shape variations. The model also learns pose-dependent deformations ϕ to account for non-rigid effects such as muscle bulging, enabling a compact yet expressive human representation. The SMPL-X model [45] extends SMPL by incorporating facial expressions and articulated hand motion, combining the SMPL body with the FLAME and MANO models to achieve a full-body expressive representation.

Building on SMPL, Kanazawa *et al.* introduced Human Mesh Recovery (HMR) [22], one of the first methods to regress the 3D human body parameters directly from a single RGB image. The model employs an image encoder to predict pose, shape, and camera parameters, from which the corresponding 3D mesh is reconstructed. A reprojection loss between the projected 3D joints and detected 2D keypoints enforces geometric consistency, while an adversarial discriminator encourages physically plausible body configurations. This work demonstrated that end-to-end learning from images can produce realistic 3D meshes without explicit 3D supervision.

To model temporal consistency, Human Mesh and Motion Recovery (HMMR) [23] extends HMR to sequential data. Features extracted using a ResNet-50 [18] backbone are fed into a temporal encoder that captures motion dynamics across frames. The encoded features are then regressed into pose and shape parameters, producing smooth mesh sequences. A hallucination network is further introduced to predict the meshes of adjacent frames from a single input frame, improving temporal coherence and robustness to missing data.

Recent transformer-based architectures have advanced parametric modeling. The 4D-Humans framework [14] replaces convolutional backbones with a Vision Transformer (ViT), where image patches are cross-attended with canonical SMPL query tokens before regressing to pose, shape, and camera parameters. Similarly, TokenHMR [10] proposes a discrete token-based representation of poses and introduces threshold-adaptive loss scaling to balance between accurate 3D reconstruction and image-aligned predictions. Perspective-aware extensions such as CameraHMR [44] further improve mesh realism by jointly estimating camera intrinsics, effectively mitigating foreshortening effects in monocular setups.

2.1.2 Non-Parametric Human Modeling

Non-parametric methods depart from fixed statistical body models and aim to directly estimate the 3D mesh or vertex positions from image-based cues. These approaches often rely on graph-based or transformer-based representations that allow flexible modeling of the body structure.

Pose2Mesh [6] introduced a graph convolutional framework that reconstructs the 3D human mesh directly from 2D keypoints, avoiding explicit use of parametric body models. The 2D keypoints are first detected using an integral pose estimation network [55], lifted to 3D using PoseNet, and subsequently refined by MeshNet, which employs spectral graph convolutions to regress vertex positions. Despite the absence of shape and pose parameters, Pose2Mesh effectively captures realistic body shapes directly from 2D skeletons.

Transformer-based models have shown strong performance in non-parametric settings. METRO [33] employs a transformer encoder with Progressive Dimensionality Reduction (PDR) to predict vertex-level coordinates. Features extracted from an HRNet backbone [54] are combined with positional embeddings for template joints and vertices, and random query masking encourages contextual learning between body parts. This adaptation of the BERT architecture allows joint and vertex representations to interact through attention mechanisms, enabling accurate and coherent mesh predictions.

More recent approaches explore probabilistic and depth-aware representations. ProHMR [27] reformulates mesh recovery as a probabilistic prediction problem, outputting a distribution over plausible 3D poses rather than a single deterministic estimate, thus explicitly modeling reconstruction ambiguity. I2LMeshNet [38] introduces a heatmap-based representation called *Lixels* to capture the nonlinear relationship between image pixels and 3D vertices, resulting in more stable mesh regression. Finally, depth-aware extensions such as D2A-HMR [3] integrate depth attention and distribution-aware loss functions to enhance alignment between reconstructed meshes and image evidence.

2.1.3 3D Human Pose Estimation

Traditional 3D human pose estimation methods decompose the task by first extracting 2D poses from input images or videos, then reconstructing the 3D pose from these 2D pose sequences [70, 72, 74]. These approaches predict a single, most likely 3D pose for each 2D observation/ sequence. Recently, however, multihypothesis approaches have emerged, generating a set of plausible 3D poses for a given timestamp f from a given 2D pose sequence centered around f [50, 19, 32]. These approaches aim to better capture pose variability by combining multiple plausible 3D predictions using conditioning or averaging techniques.

Recent works on 3D HPE [50, 15, 19] have focused on using diffusion models, with the aim of naturally handling the indeterminacy and uncertainty in the observation. Diffusion models for 3D HPE offer various advantages: (1) Plausible human poses against noise and occlusions [73]; (2) Does not suffer from phenomena like posterior collapse, vanishing gradients, or training instabilities [19]; (3) Captures fine-grained dynamics with fidelity even during inherent ambiguities in representation [73, 11].

2.1.4 Global Human Mesh Reconstruction

Traditional human mesh recovery methods typically reconstruct human bodies in the *camera coordinate frame*, which limits their applicability in scenarios requiring real-world spatial understanding, such as motion capture, human-scene interaction, and augmented reality. To address this limitation, recent research has focused on **global human mesh reconstruction**, where the goal is to recover metrically scaled, world-aligned 3D human meshes with temporally consistent trajectories across moving cameras and dynamic scenes.

Simultaneous Localization and Human Mesh Recovery (SLAHMR) [68] combines SfM and human pose estimation within a unified optimization framework. The approach first estimates camera motion using DROID-SLAM [56] and identifies individual humans across the sequence using PHALP-based tracking [14]. A two-stage optimization process then aligns human pose parameters with 2D keypoints from ViTPose, followed by refinement of camera scale, shape, and pose to achieve world-consistent human meshes.

Building on this direction, World-grounded Humans with Accurate Motion (WHAM) [52] focuses on mitigating artifacts such as foot sliding and implausible global trajectories common in moving-camera setups. It first reconstructs pixel-aligned human meshes in camera space using 2D keypoints and image features, and then refines trajectories in world coordinates through a contact-aware optimization module.

TRAM [66] further extends this paradigm by leveraging DROID-SLAM for joint camera and scene depth estimation, followed by VIMO-based mesh recovery in camera space. Using metric scale alignment, TRAM converts these reconstructions to global coordinates, enabling realistic motion estimation even in unconstrained in-the-wild videos. World-grounded Humans and Cameras (WHAC) [69] uses

More recently, Gravity-View HMR (GVHMR) [51] introduces a gravity-aware framework that explicitly incorporates physical priors into mesh recovery. By conditioning the reconstructed human orientation on gravity vectors derived from video cues, GVHMR achieves more physically plausible and stable human poses.

Overall, global human mesh reconstruction represents an important shift from purely camera-centered mesh estimation to physically grounded, world-consistent reconstruction. By integrating visual SLAM, temporal tracking, and physical priors, these methods enable metrically accurate human motion understanding in complex real-world settings.

2.2 3D Scene Reconstruction

2.2.1 Sparse Feature-based Reconstruction

Classical sparse feature–based reconstruction is a foundational approach for 3D mapping and SLAM, relying on robust feature detection, matching, and bundle adjustment to recover camera poses and sparse scene points. ORB-SLAM [41], in both its v2 and v3 iterations, exemplifies this approach: ORB features are extracted and tracked across frames to establish correspondences, which are then optimized globally or locally through bundle adjustment to produce a consistent sparse 3D map. ORB-SLAM v3 extends the system to handle inertial measurements, supporting more robust and metric-scale tracking in challenging, dynamic, or large-scale environments.

Other approaches integrate visual-inertial fusion more explicitly. VINS [48] leverages tightly-coupled visual-inertial bundle adjustment to estimate both camera motion and sparse 3D landmarks, providing high accuracy in real-time for UAVs or mobile robots even under aggressive motion. Droid-SLAM [56] and DPVO [57] push this further by combining deep feature descriptors with traditional sparse BA pipelines, improving robustness under lighting changes, repetitive textures, or feature-poor scenes. These systems maintain the key advantage of sparse reconstruction: provable consistency through geometric constraints and bundle adjustment, while still achieving near real-time performance.

COLMAP [12] represents a versatile offline reconstruction framework that also builds on sparse feature correspondences. It combines feature extraction, matching, and incremental or global bundle adjustment to produce highly accurate camera poses and sparse point clouds, serving as a standard for structure-from-motion and multi-view stereo benchmarks. Although slower than real-time SLAM systems, COLMAP provides strong metric fidelity, dense reconstruction initialization, and a reliable baseline for evaluating learned or hybrid dense reconstruction methods.

2.2.2 Dense Reconstruction with Geometry Priors

Dense reconstruction has seen significant advances with recent learning-based methods that unify multi-view geometry, depth estimation, and camera pose inference into a single architecture. DUSt3R [63] introduces a novel paradigm that directly regresses dense per-pixel 3D pointmaps from image pairs or small image collections without requiring known intrinsics or camera poses. By relaxing traditional projective constraints, it demonstrates robustness in scenarios with few views, small motion, or non-Lambertian surfaces, though

its performance depends on the diversity of training data. Building on this, MonST3R [71] extends DUSt3R into dynamic scenes by predicting pointmaps and associating them across frames, enabling coherent reconstruction of moving or deforming regions while reducing error accumulation common in multi-stage pipelines. Similarly, CUT3R [62] reframes dense reconstruction as an online, stateful process; it maintains a latent scene representation that is incrementally updated as new frames arrive, producing refined pointmaps and camera parameters in a continuous fashion. This makes it particularly suitable for low-latency applications such as robotics or AR.

Other approaches focus on grounding geometric prediction across multiple tasks. VGGT [61] jointly predicts camera parameters, depth maps, point tracks, and per-pixel 3D points in a single transformer-based model, leveraging cross-task synergies to collapse what would traditionally be a multi-stage pipeline into one feed-forward system. π^3 [65] addresses the common reliance on a fixed reference view by using a permutation-equivariant architecture that treats input views as an unordered set, improving robustness to input ordering and enhancing pose, depth, and pointmap accuracy. MapAnything [25] takes a more generalist approach, unifying multiple reconstruction tasks into a single end-to-end model capable of multi-image SfM, MVS, monocular metric depth estimation, registration, and depth completion, with consistent metric outputs. Collectively, these methods highlight the power of learned geometry priors, temporal consistency, local surface priors, permutation equivariance, and persistent scene representations—to improve reconstruction quality while simplifying or even bypassing classical optimization-based pipelines.

2.2.3 4D Reconstruction from Monocular Video

Recent work pushes from static scene recovery toward 4D dynamic reconstruction in monocular video, often using Gaussian splatting or feed-forward video encoders to represent time-varying geometry. Mesh4D [21] targets compact, editable 4D mesh sequences by predicting per-frame deformation fields in a latent space, enabling mesh tracking and animation from single-view video. DreamScene4D [7] and Dynamic Gaussian Marbles [53] adopt Gaussian splatting with explicit scene decomposition or simplified Gaussian primitives to improve dynamic object modeling and novel view synthesis under occlusions. FreeTimeGS [64], MoDGS [34], and MoSca [29] further generalize dynamic Gaussian representations by allowing time-varying primitives, depth-prior initialization, or motion-scaffold constraints to stabilize reconstruction from casual monocular footage. Complementary system components such as RoMo [49] improve robustness in dynamic scenes by filtering moving regions during SfM, while UniK3D [46] targets camera-agnostic monocular geometry to

reduce sensitivity to intrinsics. Finally, 4RC [37] compresses an entire video into a spatio-temporal latent representation that can be queried at arbitrary timestamps, enabling flexible 4D geometry and motion retrieval. Together, these methods define a fast-growing class of monocular 4D reconstruction techniques that trade explicit multi-view constraints for learned priors, temporal regularization, and compact representations.

2.3 Human and Scene Reconstruction

Recent progress in human and scene reconstruction has increasingly emphasized the joint recovery of dynamic human motion, scene geometry, and camera parameters, motivated by the need for coherent and physically plausible understanding of human–environment interaction. Early work, such as Human Structure-from-Motion (HSfM) [39] extends classical SfM pipelines to articulated humans by incorporating parametric human body models into multi-view reconstruction. By leveraging human kinematic priors, HSfM resolves scale ambiguity and improves spatial consistency across humans, scene geometry, and cameras. However, its reliance on optimization-based pipelines and multi-view imagery limits scalability and applicability to casual monocular videos.

Complementary to video- and multi-view settings, PhySIC [42] targets the challenging *single-image* case, aiming to reconstruct metrically accurate SMPL-X humans together with dense scene surfaces and physically plausible human–scene interactions. By explicitly handling depth ambiguity, occlusions, and contact consistency, it produces metric-scale human meshes, dense scene geometry, and vertex-level contact maps from a single monocular RGB input. While this design enables physically grounded reconstructions without multi-view inputs, it is limited to static, per-image inference and does not model temporal motion.

More recent methods aim to unify human and scene reconstruction into a single framework. JOSH [35] represents an optimization-based approach that jointly reconstructs 4D human motion and dense scene geometry from monocular videos. Its key insight is to use explicit human–scene contact constraints as a coupling mechanism, allowing human motion, scene structure, and camera poses to be refined together. This joint optimization significantly improves reconstruction accuracy in in-the-wild videos, but its performance depends heavily on the quality of initialization from external human and scene reconstruction models, and contact visibility is critical for stable optimization.

To address the complexity and fragility of multi-stage optimization pipelines, Human3R [5] proposes a unified, single-stage feed-forward framework for joint 4D hu-

man–scene reconstruction. Built on top of a pretrained 4D reconstruction backbone, Human3R simultaneously predicts multi-person SMPL-X human meshes, dense scene geometry, and camera trajectories in a global world coordinate system from monocular videos. By eliminating explicit optimization loops and heavy preprocessing, it achieves real-time performance and improved robustness on casual, in-the-wild videos, albeit at the cost of limited appearance modeling and reliance on visible head position for human reconstruction.

UniSH [31] further advances this direction by explicitly targeting metric-scale, temporally consistent human–scene reconstruction across diverse input modalities, including monocular videos, multi-view images, and unordered image sets. UniSH combines strong pretrained human and scene priors, drawing from CameraHMR and Pi3, with a unified feed-forward architecture and a global alignment module to recover metric-scale reconstructions. By leveraging joint human–scene reasoning and large-scale in-the-wild training data, UniSH improves robustness to camera motion and scene complexity, though performance is usually inferior when compared to optimization-based techniques.

2.4 Human and Scene Interaction

Human–Scene Interaction (HSI) methods explicitly incorporate contact, support, and collision constraints to ensure reconstructed humans respect environmental structure. These approaches are particularly relevant for global human mesh recovery, where accurate grounding and physical consistency are critical. Early work in this direction includes PROX [16], which introduced one of the first frameworks for scene-aware human mesh optimization. Given a static scene mesh, PROX optimizes SMPL-X parameters to enforce contact between the body and the environment while penalizing interpenetration via signed distance fields. This approach demonstrated that leveraging scene geometry significantly improves global plausibility, particularly for lower-body joints and grounding. CHORE [67] further extends this idea to dynamic human–object interactions, using implicit neural fields to reconstruct both humans and objects in 3D from monocular videos, while enforcing multi-frame consistency and minimizing contact violations.

Other lines of research focus on enriching contact modeling. Methods such as CAPE and SCAPE+ capture clothing and soft-tissue deformation during body–scene contact, while BEHAVE datasets [1] introduce real-world recordings of humans interacting with everyday objects to encourage models that learn grounded, object-aware priors. These datasets have enabled approaches that jointly estimate human pose, object pose, and support relations, improving reconstructions in cluttered or occluded environments.

2.5 Summary

Across human, scene, and joint reconstruction, prior work has made significant advances in modeling 3D geometry, motion, and physical plausibility. Parametric mesh recovery approaches, such as SMPL and its derivatives, enable robust single-view and multi-view reconstruction of human pose and shape. Non-parametric mesh estimators relax these constraints by directly predicting vertex-level geometry, while recent diffusion-based frameworks address inherent ambiguity by modeling distributions of plausible 3D poses. Global human mesh recovery methods further attempt to lift humans into world coordinates using SLAM or multi-view alignment. However, these approaches generally rely on accurate camera poses and often struggle to ground humans in metrically consistent environments, especially in static-camera or monocular scenarios.

Scene reconstruction methods provide complementary advances through feature-based SLAM, deep geometry priors, and feed-forward models that predict depth, point maps, and camera parameters. Despite these advances, most systems reconstruct environments independently of humans, and rarely consider dynamic agents, physical interactions, or contact relations. Human–Scene Interaction (HSI) methods partially address this gap by enforcing scene-aware constraints such as contact, support, and collision avoidance. Yet, these approaches typically assume that the scene geometry is known or sufficiently accurate and do not consider treating humans as active sources of information for improving scene reconstruction.

Despite recent progress toward joint human–scene reconstruction, most existing approaches exhibit limited bidirectional interaction between human and scene representations. In many feed-forward architectures, human and scene modules are pretrained and optimized largely independently, with coupling restricted to global alignment variables such as scale or translation. As a result, human motion does not meaningfully influence the reconstruction of scene geometry, and scene structure does not impose strong geometric or physical constraints on human pose, articulation, or motion. This decoupled design leads to joint inference pipelines that are end-to-end in execution but not in representation or reasoning, limiting their ability to exploit human–scene priors such as contact, support, collision, and affordance cues.

This highlights a critical limitation in the literature: *humans are rarely treated as active contributors to reconstructing the environment, and scenes are rarely used to metrically ground human motion in a unified world frame.* In real-world scenarios, human kinematics, contact events, and motion patterns provide strong cues about scene geometry, and conversely, the scene provides essential metric context for human motion, particularly when visual data is sparse, noisy, or ambiguous.

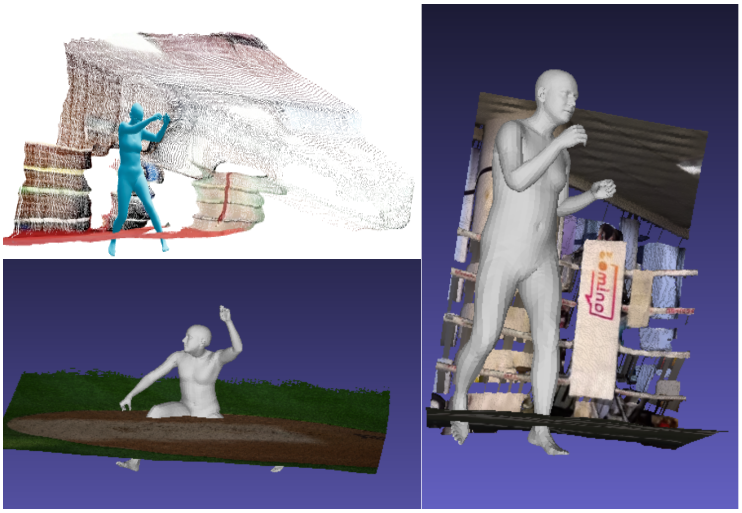

Figure 2.1: Ground penetration artifacts in existing reconstructions (three samples). This highlights the lack of reliable contact and support constraints between humans and scenes.

This motivates a **unified Human–Scene–Camera Reconstruction (HSR)** paradigm, in which humans, scenes, and cameras are estimated jointly within a single, metrically consistent world frame. By leveraging human motion to infer uncertain aspects of the environment and simultaneously using scene and camera geometry to ground humans in metric coordinates, the proposed framework enables bidirectional reasoning and full 4D world reconstruction. Such integration is critical for robust, physically plausible reconstruction in everyday, human-centered environments and forms the foundation for applications in robotics, AR/VR, autonomous navigation, and immersive media. Figure 2.1–2.3 illustrates representative failure cases that motivate this direction.

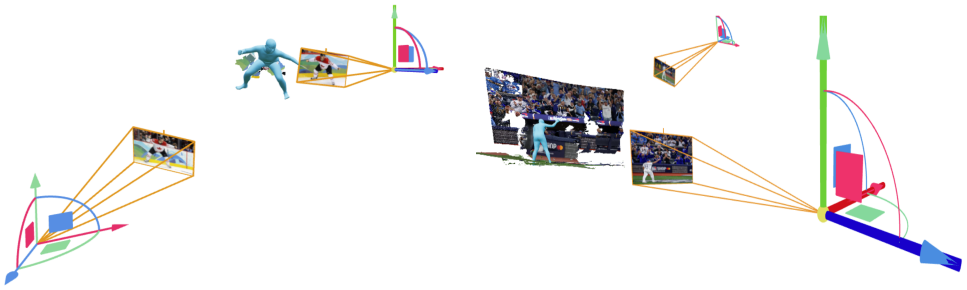

Figure 2.2: Sparse scene reconstruction even when multiple input views are available, indicating weak geometric consistency and incomplete scene recovery.

Figure 2.3: Floating humans and incorrect scale/translation relative to the scene, underscoring the need for metric grounding and joint optimization.

Chapter 3

Proposed Research

3.1 Overview

This chapter details the proposed research framework for unified 4D world reconstruction, with a focus on jointly reconstructing the global scene, human, and the camera systems in a shared grounded world coordinate frame. Unlike prior methods that treat these components independently or in a loosely coupled manner, the proposed approach formulates world reconstruction as a single integrated problem, where each component both constrains and informs the others.

Figure 3.1 illustrates the overarching architecture of the proposed unified human-scene-camera reconstruction system. The core hypothesis underlying this research is that humans, scenes, and cameras are mutually informative. Human kinematics and anthropometric regularities provide strong metric and geometric cues for resolving ambiguities in scene reconstruction and camera alignment. Conversely, reconstructed scene geometry and physical structure impose essential constraints on plausible human pose, motion, and placement. Camera parameters and trajectories serve as the connective tissue that binds these components into a consistent global model.

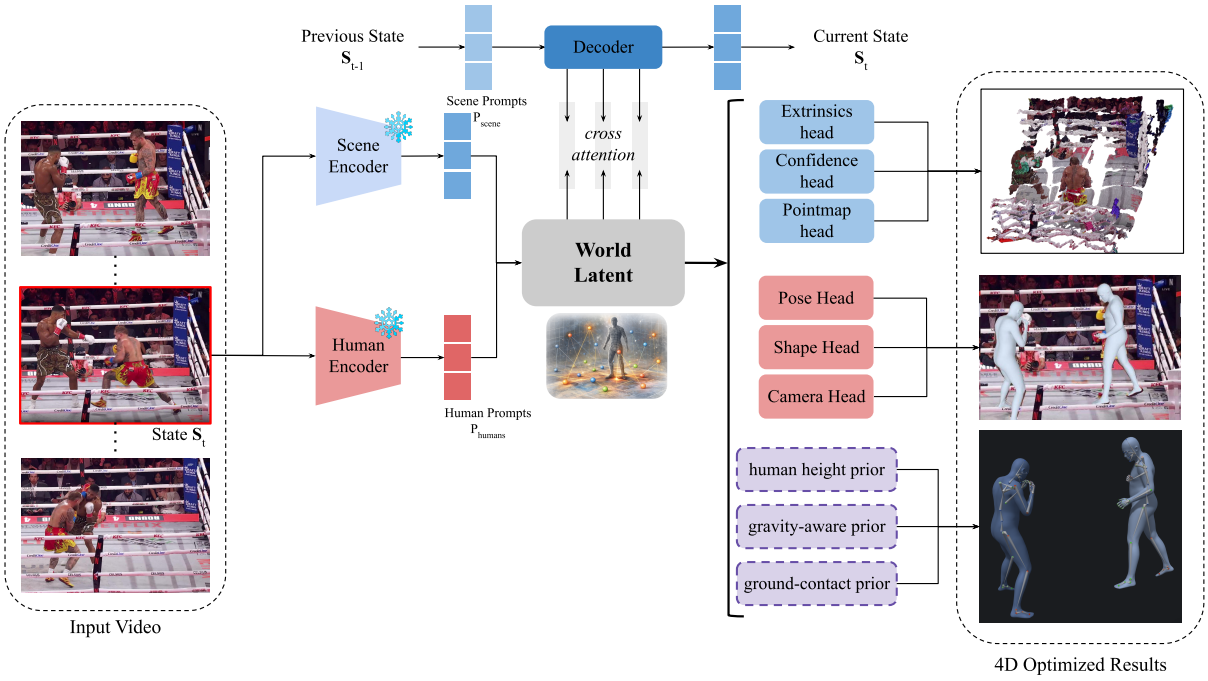

Figure 3.1: Overview of the human-scene-camera reconstruction model

3.2 Unified Human-Scene-Camera Reconstruction

3.2.1 Human Reconstruction

Human reconstruction focuses on recovering consistent, metrically meaningful 3D human pose and shape of all the humans in the frame, given a single video. The proposed approach builds on top of a very strong pretrained encoder [44] as the backbone representation.

Specifically, the human reconstruction module leverages a vision transformer encoder trained to extract pose-aware, perspective-sensitive image features. Then the decoder utilizes three MLP heads to output the root-relative parametric human body parameters and camera-relative global translation.

This module is not aimed at producing the final world-aligned human trajectories; it provides a strong but local initialization that is subsequently refined using scene context and global constraints.

3.2.2 Scene Reconstruction

Scene reconstruction aims to provide a strong initialization of scene geometry and camera motion cues that can be refined jointly with humans. We plan to use the CUT3R encoder [62] as the primary source of scene features, and to maintain a model *state* across time, rather than treating each frame independently. Intuitively, the state acts as a persistent memory of the scene that can be updated as new frames arrive.

Concretely, at timestep t, the CUT3R encoder produces (i) per-frame scene tokens/features and (ii) a state representation s_t. A lightweight transition module then predicts the next state s_{t+1}, which is passed forward to the subsequent timestep and used to condition the next encoding pass. This design provides temporally coherent scene features and supports incremental refinement under occlusions, motion blur, or rapid camera motion.

3.2.3 World Latent Fusion and State Propagation

To couple humans, scene, and camera in a unified coordinate frame, we introduce a shared *world latent* representation that serves as the central fusion space for cross-modal reasoning. At each timestep, scene tokens (from the CUT3R encoder and its propagated state) and human tokens (from the MultiHMR-style human encoder) cross-attend with the world latent, enabling bidirectional information exchange between modalities.

The world latent is intended to: (1) aggregate temporally consistent scene context via the propagated CUT3R state; (2) integrate human-centric cues that help resolve depth and scale ambiguities; and (3) support downstream prediction heads for globally consistent quantities such as world-frame human translations and camera alignment. While the exact parameterization of the world latent remains an open design choice, the key requirement is that it provides a shared space where human and scene embeddings are fused and updated recurrently over time.

3.2.4 Human-Scene Prior Modeling

A key novelty of the proposed research lies in modeling human–scene priors as a mechanism for metric grounding and physical plausibility. In our formulation, these priors are *not* provided as direct inputs to the encoder; instead, they act as *implicit regularizers* that shape the fused world latent and the final predictions through dedicated loss terms and consistency constraints. Humans are treated as informative agents whose kinematics and

anthropometry provide strong cues about the surrounding environment. The framework incorporates human-centric priors in several forms.

Anthropometric Attributes. First, n random images are used to extract the human crops and feed them into a gender and age prediction model [28] to get the demographic attributes. These attributes are then mapped to statistically grounded height distributions derived from population-level anthropometric data. The resulting height estimates serve as soft metric constraints that anchor both human scale and scene scale in the world frame.

Gravity-aligned body orientation. Global body orientation is used to infer a gravity-aligned reference direction. This prior provides a consistent notion of verticality across time and views, reducing ambiguity in both human pose orientation and scene alignment. Rather than being enforced as a hard constraint, gravity alignment is introduced as a regularization term that encourages physically plausible orientations while allowing flexibility under noisy observations.

Ground contact relationships. To further improve physical plausibility and metric consistency, the framework incorporates weak ground contact priors between reconstructed humans and the scene geometry. The mesh vertices are encouraged to remain close to locally supporting surfaces. To modulate these constraints, we train a learned body-centric contact prior inspired by POSA [17], which operates on human body vertices expressed in a gravity-aligned coordinate frame. This prior produces probabilistic per-vertex contact likelihoods and coarse semantic contact types. Importantly, these contact relationships are modeled probabilistically and do not assume explicit force estimation, rigid ground planes, or known scene geometry, allowing contact to emerge naturally during joint human-scene-camera alignment.

3.3 Training Objective

During training, we provide the model with a sequence of N observations (e.g., consecutive video frames or a small image collection). When metric-scale 3D supervision is available for the scene, we optionally replace some image inputs with raymap-style geometric queries (excluding the first view) so that the scene branch learns to predict metric pointmaps; when the supervision scale is unknown, raymap querying is disabled to avoid introducing inconsistent scale.

Scene geometry losses. Following confidence-aware pointmap regression used in recent dense reconstruction methods, we supervise predicted pointmaps with a confidence-

weighted regression objective [30]:

$$\mathcal{L}_{\text{conf}} = \sum_{t=1}^{N} \sum_{i} \Big(c_{t,i} \left\| \hat{\boldsymbol{x}}_{t,i} - \boldsymbol{x}_{t,i} \right\|_2 - \alpha \log c_{t,i} \Big), \tag{3.1}$$

where $\hat{\boldsymbol{x}}_{t,i}$ and $\boldsymbol{x}_{t,i}$ denote predicted and target 3D points (after the appropriate scale normalization), and $c_{t,i}$ is the predicted confidence. When metric ground truth is available, we tie the normalization factors so the network can learn absolute scale. We additionally supervise camera pose (e.g., rotation and translation) with an ℓ_2 loss,

$$\mathcal{L}_{\text{pose}} = \sum_{t=1}^{N} \Big(\|\hat{\boldsymbol{q}}_t - \boldsymbol{q}_t\|_2 + \|\hat{\boldsymbol{\tau}}_t - \boldsymbol{\tau}_t\|_2 \Big), \tag{3.2}$$

and, when raymap inputs provide RGB targets, we apply an MSE color loss $\mathcal{L}_{\text{rgb}} = \|\hat{\boldsymbol{I}} - \boldsymbol{I}\|_2^2$.

Human losses. The human branch is trained end-to-end with (i) a detection/classification loss for person presence on the feature grid (binary cross-entropy), and (ii) regression losses for predicted body model parameters and depth/translation (typically ℓ_1). When mesh supervision is available, we also include a mesh-space loss and a reprojection loss that penalizes disagreement between projected predicted meshes and image evidence.

Joint world losses and regularization. The fused world latent is trained with cross-component consistency terms that encourage agreement between humans, scenes, and cameras. Concretely, we add regularizers for (i) metric consistency (e.g., scale/height alignment), (ii) gravity alignment of global orientation, and (iii) physically plausible human–scene interaction via contact and non-penetration. These priors are implemented as loss terms acting on the world-latent outputs, rather than as direct network inputs.

Overall, the full objective is a weighted sum

$$\mathcal{L} = \lambda_{\text{scene}} \mathcal{L}_{\text{conf}} + \lambda_{\text{pose}} \mathcal{L}_{\text{pose}} + \lambda_{\text{rgb}} \mathcal{L}_{\text{rgb}} + \lambda_{\text{human}} \mathcal{L}_{\text{human}} + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}, \tag{3.3}$$

where $\mathcal{L}_{\text{human}}$ aggregates detection, parameter/mesh, and reprojection losses, and $\mathcal{L}_{\text{prior}}$ aggregates the implicit regularization priors.

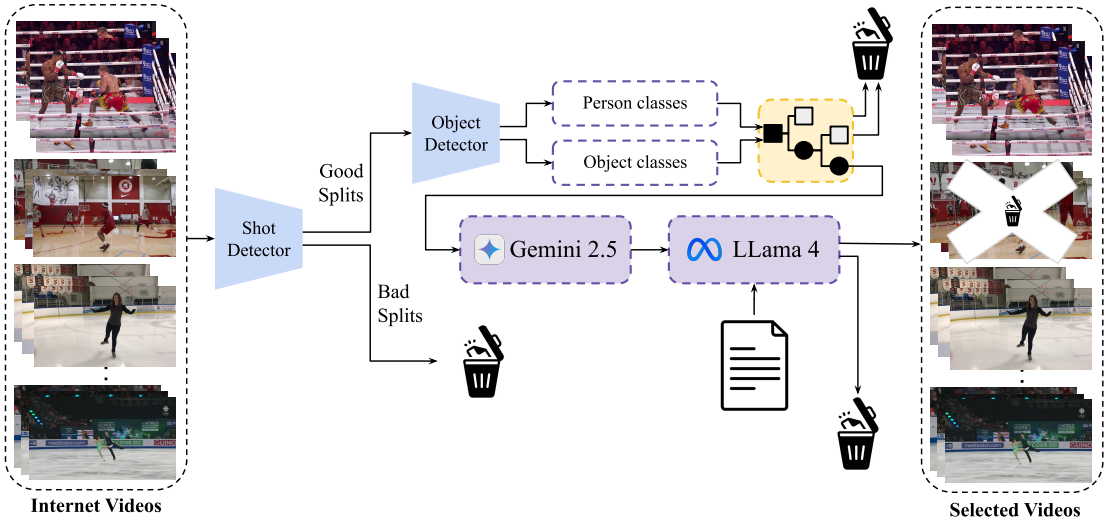

Figure 3.2: **Dataset curation pipeline overview:** starting from raw in-the-wild web videos, we apply automated shot-boundary detection and geometric filtering to produce candidate clips, then perform semantic verification with a VLM to validate split points and identify subatomic actions, and accept high-quality segments for training

3.4 Internet-Scale Dataset Curation

Scalability and data requirements. A central goal of this thesis is to move toward world reconstruction models that can learn from large-scale, in-the-wild Internet video, with minimal manual annotation. While early experiments may focus on domains with visually clear motion (e.g., dance videos where the subject remains the primary focus), this choice serves as a proof-of-concept rather than a restriction. The proposed framework requires only monocular videos with a visible human subject and sufficient temporal continuity; it does not assume laboratory capture, calibrated cameras, or motion-capture systems, enabling direct scaling to broader web video sources.

Automated filtering pipeline. To obtain training clips with reliable temporal and geometric structure, we adopt a multi-stage automated pipeline as illustrated in Figure 3.4. First, to mitigate scene transitions and montages common in raw web videos, we enforce strict temporal continuity by detecting shot boundaries (PySceneDetect [47]) and discarding sequences with discontinuities. Second, we isolate sequences to a maximum 5 person scene using a person detector [13], keeping clips less cluttered and consistently detected. Third, we enforce a spatial prominence constraint (e.g., average bounding-box

height exceeding a fixed fraction of the image height) to ensure sufficient resolution for human reconstruction. Finally, we remove sequences with truncation at image borders and discard clips with significant overlap between the primary subject and other detected boxes, reducing the prevalence of severe occlusion.

Semantic verification with video-language models. In practice, purely geometric filtering can be noisy at scale: shot-boundary detectors may over-segment fast camera motion or lighting changes, and person detection does not guarantee that the retained clip depicts a coherent action segment useful for learning world dynamics. To improve dataset quality, we introduce an additional semantic validation stage using a video-language model (Gemini [8]). Given each candidate clip and its proposed split points, the model is used to (i) verify that a detected boundary corresponds to a true semantic transition (rather than a false positive), and (ii) identify the *subatomic action* occurring within each retained segment (e.g., "pitching windup", "jump landing", "turn", "sit-to-stand").

Action cross-verification and acceptance rules. The predicted subatomic action labels are then cross-verified by an LLM [59] against a curated action dictionary/database (with canonical names and synonyms). Only segments whose action label maps cleanly to an allowed action class are admitted into the training set; segments that are ambiguous, off-topic, or inconsistent with the action taxonomy are excluded. This two-stage semantic gating (video-language verification + LLM dictionary matching) provides an explicit mechanism for scaling dataset curation while controlling noise, and aligns the training distribution with the motion and interaction primitives needed for learning a controllable, in-the-wild 4D world model.

Chapter 4

Progress to date

This chapter summarizes preliminary experiments and system components developed in support of the proposed unified world reconstruction framework. While the full joint optimization is ongoing, the results presented here demonstrate feasibility, highlight key challenges, and motivate the design choices described in Chapter 3.

4.1 Domain-Robust Human Reconstruction

To address severe motion blur, self-occlusion, and visual degradation prevalent in broadcast sports footage, we developed a domain-robust 3D human pose estimation system that serves as a strong initialization for downstream world reconstruction. The system is designed to prioritize robustness under adverse visual conditions rather than general-purpose pose accuracy.

Transformer Backbone. The core visual representation is obtained using a transformer-based image backbone that produces a compact, global feature vector for each input frame. Given an input image I_t at time t, the backbone extracts a feature embedding

$$\mathbf{f}_t = \Phi(I_t), \quad \mathbf{f}_t \in \mathbb{R}^D, \tag{4.1}$$

where $\Phi(\cdot)$ denotes the pretrained vision transformer and D is the feature dimensionality (1280 in our implementation). The backbone is initialized from a large-scale pretrained human motion and reconstruction model [14], providing strong priors over human appearance and articulation.

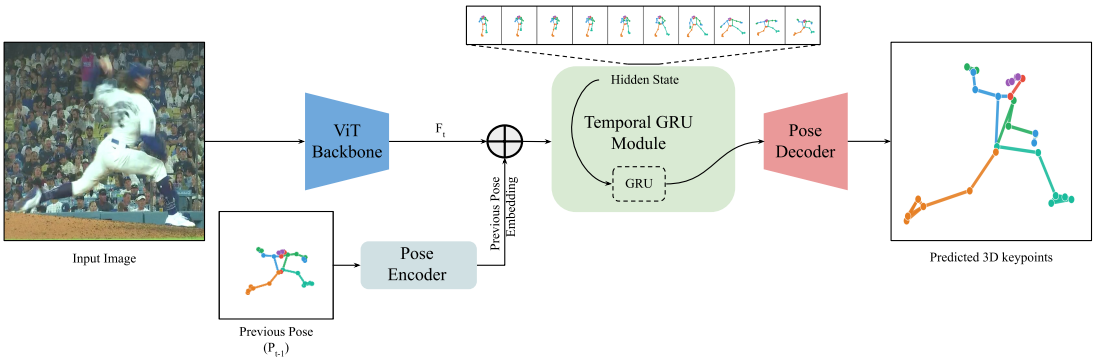

Figure 4.1: Overview of the proposed 3D pose estimation architecture

The extracted feature $\mathbf{f}_t$ captures global context and pose-relevant visual cues and serves as the sole visual input to the temporal pose head. No temporal information is modeled at this stage; each frame is processed independently by the backbone.

Previous Pose Encoding. At time step t, the model receives the predicted 3D pose from the previous frame,

$$\mathbf{P}_{t-1} \in \mathbb{R}^{J\times 3}, \tag{4.2}$$

where J is the number of joints. The pose is flattened and embedded using a lightweight multilayer perceptron:

$$\mathbf{e}_{t-1} = \psi(\mathrm{vec}(\mathbf{P}_{t-1})), \quad \mathbf{e}_{t-1} \in \mathbb{R}^{256}, \tag{4.3}$$

where $\psi(\cdot)$ denotes the pose encoder MLP. This embedding provides an explicit geometric prior, allowing the network to condition on the most recent pose configuration without directly exposing raw joint coordinates to the temporal module. For the first frame of a sequence, $\mathbf{P}_{t-1}$ is initialized to zeros.

Feature Fusion. The visual feature $\mathbf{f}_t$ and the pose embedding $\mathbf{e}_{t-1}$ are concatenated and projected into a unified latent space via a fusion MLP:

$$\mathbf{x}_t = \phi\left([\mathbf{f}_t \,\|\, \mathbf{e}_{t-1}]\right), \quad \mathbf{x}_t \in \mathbb{R}^H, \tag{4.4}$$

where $\phi(\cdot)$ denotes the feature fusion network and H is the hidden dimensionality. This fused representation forms the sole explicit input to the recurrent module.

Recurrent Temporal Modeling. Temporal dynamics are modeled using a gated recurrent unit (GRU) operating in a single-step, online fashion. At each frame, the GRU updates its hidden state according to

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}), \tag{4.5}$$

where $\mathbf{h}_{t-1}$ is the hidden state carried over from the previous timestep. The hidden state implicitly encodes motion trends, velocity, and longer-term temporal context accumulated over past frames. No explicit pose sequence is provided to the GRU; all temporal information is conveyed through the recurrent hidden state.

Pose Decoder. The GRU output $\mathbf{h}_t$ is decoded into the current 3D pose prediction using a final MLP:

$$\mathbf{P}_t = \delta(\mathbf{h}_t), \quad \mathbf{P}_t \in \mathbb{R}^{J\times 3}, \tag{4.6}$$

where $\delta(\cdot)$ denotes the pose decoder. The predicted pose is fed back as input to the next timestep, forming a recurrent feedback loop across frames.

While the current system is trained on baseball footage, it serves as a proof-of-concept for domain-robust human reconstruction under extreme visual conditions. This architecture will be extended toward broader domains through data diversification and integration with scene-level constraints, allowing the human reconstruction module to generalize beyond sports-specific motion patterns.

4.2 Human-Scene Context Priors

4.2.1 Anthropometric Attributes

As an initial step toward explicit human-scene prior modeling, we implemented and validated a demographic-based anthropometric prior that provides coarse metric grounding for reconstructed humans. This component corresponds to the first module outlined in the proposed human-scene prior framework.

Demographic Estimation. Given an input image sequence, we first extract human-centric crops using the detected bounding boxes. For each subject, we apply a pretrained age and gender estimation model (MiVOLO [28]) to predict demographic attributes. Given a crop I^{human}, the model produces estimates

$$(\hat{a}, \hat{g}) = \text{MiVOLO}(I^{\text{human}}), \tag{4.7}$$

where $\hat{a}$ denotes the predicted age and $\hat{g}$ denotes the predicted gender, accompanied by a confidence score. When multiple frames are available for a given video, predictions are aggregated across a random subset of n frames to reduce per-frame noise and obtain a stable demographic estimate.

Anthropometric Height Prior. The estimated age and gender are then mapped to a statistically grounded height distribution derived from population-level anthropometric data. Rather than predicting a single deterministic height, we model human stature as a Gaussian random variable

$$H \sim \mathcal{N}(\mu_c, \sigma_c^2), \tag{4.8}$$

where the mean μ_c and standard deviation σ_c are selected based on the demographic category c inferred from $(\hat{a}, \hat{g})$. Specifically, we use the following categories:

- **Baby** ($\hat{a} < 3$): $\mu = 0.801$ m, $\sigma = 0.126$ m
- **Kid** ($3 \leq \hat{a} < 8$): $\mu = 1.122$ m, $\sigma = 0.12$ m
- **Teen** ($8 \leq \hat{a} < 15$): $\mu = 1.477$ m, $\sigma = 0.156$ m
- **Adult Female** ($\hat{a} \geq 15$, confident female): $\mu = 1.647$ m, $\sigma = 0.0707$ m
- **Adult Male** ($\hat{a} \geq 15$, confident male): $\mu = 1.784$ m, $\sigma = 0.0759$ m
- **Neutral Adult** ($\hat{a} \geq 15$, uncertain gender): $\mu = 1.715$ m, $\sigma = 0.10$ m

When gender confidence is low, we default to the neutral adult distribution to avoid overconfident or biased metric assumptions. The resulting height distribution is not enforced as a hard constraint. Instead, it is intended to serve as a soft metric prior that anchors human scale and, by extension, scene scale in the world frame. In practice, this prior will be incorporated as a regularization term that penalizes large deviations between the reconstructed human height $\hat{H}$ and the demographic prior:

$$\mathcal{L}_{\text{height}} = \frac{(\hat{H} - \mu_c)^2}{\sigma_c^2}. \tag{4.9}$$

4.2.2 Analysis of Demographic-Based Height Priors

To assess whether the proposed demographic-based anthropometric prior provides meaningful and non-degenerate metric information, we analyze the height distributions inferred from a real-world sports video. Specifically, we apply the age and gender estimation pipeline to 557 human instances extracted from a boxing tournament video and assign each subject a demographic-conditioned height distribution as described in Section 4.2.1.

Height Distribution with Gaussian Priors. Figure 4.2.2(a) shows a histogram of the assigned mean heights across all subjects, with overlaid Gaussian distributions corresponding to each demographic category. The histogram is weighted by the number of samples

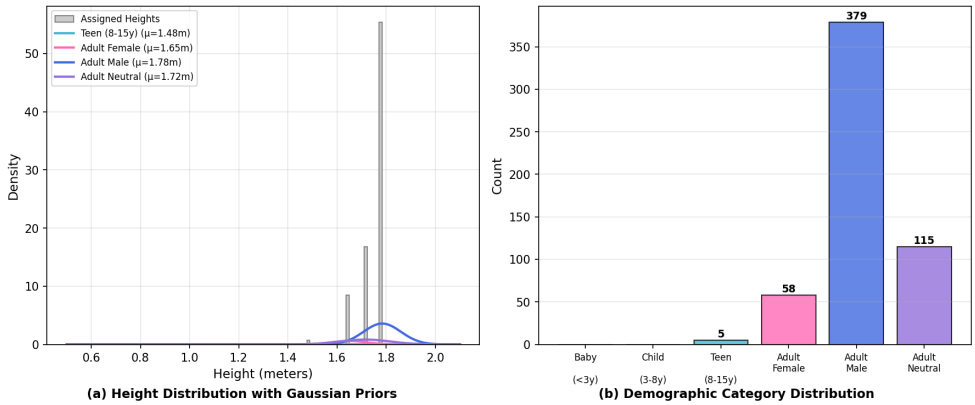

Figure 4.2: Height prior validation- Demographic Distribution Analysis

per group. The resulting distribution exhibits clear clustering around biologically plausible values. Notably, adult male subjects peak around 1.78 m and adult female subjects around 1.65 m, closely matching established population-level anthropometric statistics [40]. Dashed vertical lines indicate the mean height μ_c for each demographic group, highlighting the separation between categories.

Category-Wise Prior Comparison. Figure 4.2.2(b) presents a category-wise comparison of the height priors, where bar heights correspond to the demographic means μ_c and error bars denote one standard deviation ($\pm\sigma_c$). Sample counts (n) are reported for each category to ensure transparency and reproducibility. Adult categories exhibit limited overlap in their uncertainty ranges, with a statistically meaningful separation between adult male and adult female priors of approximately 13 cm, aligning with global anthropometric measurements [4, 40].

Observations and Limitations. Empirically, this demographic-based prior provides reasonable coarse-scale estimates that are often sufficient to disambiguate global scale in monocular or weakly constrained settings. However, its accuracy is limited by errors in age and gender prediction, intra-class variance in human stature, and dataset bias. As such, this prior is not intended to fully resolve metric ambiguity on its own, but rather to provide an informative initialization and regularization signal.

4.2.3 Ground Contact Relationships: Progress

As a step toward modeling human–scene contact priors, we integrate a contact prediction module based on POSA, which operates on SMPL-X body meshes expressed in a gravity-aligned coordinate frame. Since the upstream human reconstruction model (SAM3D)

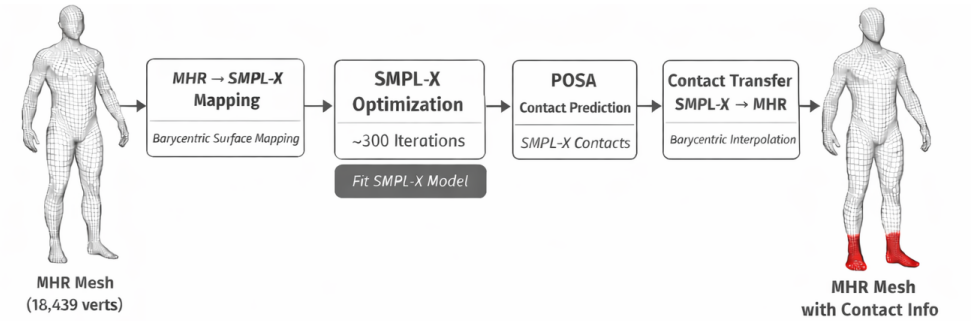

Figure 4.3: Overview of the ground contact mechanism

produces body meshes in the MHR representation, an intermediate conversion from MHR to SMPL-X is required to interface with POSA.

MHR to SMPL-X Conversion. We employ the official MHR→SMPL-X conversion pipeline, which performs a geometry-consistent transfer of pose and shape using precomputed barycentric surface correspondences. Given MHR vertices $\mathbf{V}^{\text{MHR}} \in \mathbb{R}^{18439\times 3}$ and camera translation, the conversion proceeds as follows: (i) MHR vertices are mapped to the SMPL-X surface topology via barycentric interpolation, producing target SMPL-X vertices, (ii) SMPL-X parameters (global orientation, body pose, shape coefficients, and translation) are optimized to reproduce these target vertices.

This conversion is solved via iterative optimization rather than a feedforward network. The full procedure consists of a coarse fitting stage followed by fine-grained refinement, totaling approximately 300 optimization iterations. Each iteration involves a SMPL-X forward pass, vertex-level loss computation, backpropagation, and parameter updates.

Accuracy of the Conversion. The reported reconstruction error of approximately 0.008 cm measures the discrepancy between the target SMPL-X vertices obtained via barycentric surface mapping and the vertices produced by the optimized SMPL-X parameters. Importantly, this metric reflects the quality of the SMPL-X parameter fitting and does not measure direct MHR–to–SMPL-X vertex error, which would be ill-defined due to the differing mesh topologies.

A regional error analysis indicates that the barycentric surface mapping itself is highly accurate (mean error ∼0.32 cm), while most residual error arises from the limited expressiveness of the SMPL-X parametric model. The largest discrepancies occur in the head and facial regions, where MHR contains higher-resolution geometry. In contrast, the torso, limbs, and hands exhibit mean errors of approximately 1.3 cm, which is well within acceptable bounds for contact reasoning.

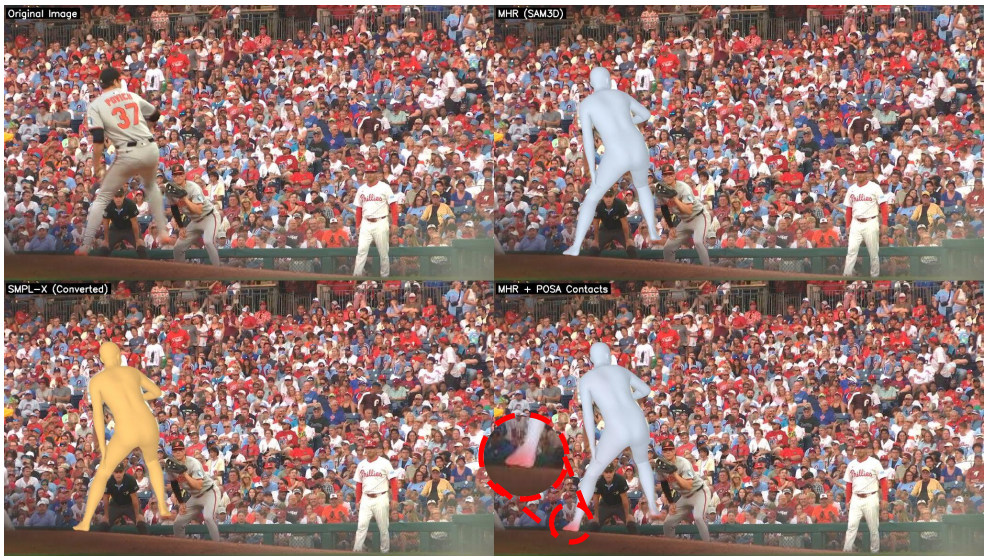

Figure 4.4: Overview of ground contact prediction and mapping. Contact probabilities are predicted on the SMPL-X mesh (highlighted in gold) using POSA and subsequently transferred back to the original MHR mesh via barycentric surface correspondence, preserving the original pose (contact highlighted in red).

Contact Prediction and Mapping Back to MHR. POSA predicts per-vertex contact probabilities on the SMPL-X mesh (10,475 vertices). To preserve the original MHR pose produced by SAM3D and avoid propagating SMPL-X conversion artifacts downstream, we map the predicted contact probabilities back to the MHR mesh using the same precomputed barycentric surface correspondence.

Specifically, for each MHR vertex, the correspondence provides the associated SMPL-X triangle and barycentric weights. Contact values are transferred via barycentric interpolation of the contact probabilities at the triangle's vertices. This produces a dense per-vertex contact field on the original MHR mesh without modifying its geometry or pose.

As a result, the SMPL-X conversion only influences how POSA infers contact likelihoods, while all subsequent processing operates directly on the original MHR representation. This design ensures that conversion inaccuracies do not propagate into downstream pose, geometry, or scene alignment components.

Computational Cost and Limitations. The primary limitation of the current approach is computational efficiency. The MHR→SMPL-X conversion requires approximately 4.2 seconds per image due to iterative optimization, dominating the overall runtime. Additionally, while the conversion preserves pose faithfully at a coarse level, small discrepancies are

introduced due to the restricted expressiveness of the SMPL-X parameterization. These limitations motivate learning an end-to-end, feedforward contact prediction model that operates directly on MHR representations, eliminating the need for SMPL-X conversion. Such a model would enable faster inference and tighter geometric consistency while retaining the benefits of learned human–scene contact priors.

4.3 Evaluation on Benchmarks

Camera-Space Human Reconstruction

We first examine camera-space human reconstruction methods, which recover articulated human pose and shape relative to the camera coordinate system without explicit world grounding. Representative state-of-the-art approaches include HMR2.0 [14], TokenHMR [10], and CameraHMR [44]. These methods are evaluated using standard reconstruction metrics such as PA-MPJPE, MPJPE, and PVE on datasets including 3DPW [60], EMDB [24], and SPEC-SYN [26].

As shown in prior work (Table 4.3), modern camera-space models achieve strong reconstruction accuracy across datasets, with TokenHMR and HMR2.0 consistently improving over earlier baselines through better temporal modeling and transformer-based representations. These results demonstrate that camera-space human reconstruction is largely a solved problem under moderate visual conditions.

However, despite strong quantitative performance, camera-space methods remain fundamentally ambiguous with respect to global scale, orientation, and placement in the world. Reconstructions are only defined up to a similarity transform, and errors in camera estimation or depth ambiguity directly propagate to downstream tasks. In addition, performance degrades in challenging real-world scenarios such as broadcast footage with severe motion blur, self-occlusion, and limited viewpoint diversity. These limitations motivate the need for additional temporal, contextual, and scene-level constraints beyond camera-space reconstruction alone.

Global-Space Human Reconstruction

We next consider world-space human reconstruction methods, which aim to recover metrically grounded human motion trajectories in a shared global coordinate frame. Representative approaches include GVHMR [51], TRAM [66], and WHAM [52]. These methods

Table 4.1: **Camera-space human reconstruction performance on standard benchmarks.** We report PA-MPJPE, MPJPE, and PVE on 3DPW [60], EMDB [24], and SPEC-SYN [26], as reported in prior work. These methods reconstruct human pose and shape relative to the camera coordinate frame without explicit world grounding.

	3DPW			EMDB			SPEC-SYN		
Method	PA-MPJPE↓	MPJPE↓	PVE↓	PA-MPJPE↓	MPJPE↓	PVE↓	PA-MPJPE↓	MPJPE↓	PVE↓
HMR2.0 [14]	44.4	69.8	82.2	61.5	97.8	120.0	55.8	133.3	153.0
TokenHMR [10]	43.8	70.5	86.0	49.8	88.1	104.2	51.8	110.5	127.6
CameraHMR [44]	40.0	62.3	74.8	45.4	82.7	97.0	**31.8**	**58.9**	**70.0**

Table 4.2: **World-space human reconstruction performance on global motion benchmarks.** We report world-grounded metrics on RICH [20] and EMDB-2 [24], including WA-MPJPE$_{100}$, W-MPJPE$_{100}$, relative trajectory error (RTE), temporal jitter, and foot-sliding. Lower values indicate better global consistency and physical plausibility.

	RICH					EMDB-2				
Method	WA-MPJPE$_{100}$ ↓	W-MPJPE$_{100}$ ↓	RTE↓	Jitter↓	Foot-Sliding↓	WA-MPJPE$_{100}$ ↓	W-MPJPE$_{100}$ ↓	RTE↓	Jitter↓	Foot-Sliding↓
GVHMR [51]	129.4	236.2	**3.8**	49.7	18.1	280.8	726.6	11.4	46.3	20.7
TRAM [66]	238.1	925.4	610.4	1578.6	230.7	529.0	1702.3	17.7	2987.6	370.7
WHAM [52]	**109.9**	**184.6**	4.1	**19.7**	**3.3**	**135.6**	**354.8**	**6.0**	**22.5**	**4.4**

integrate human pose estimation with camera motion estimation and temporal reasoning to produce globally consistent reconstructions.

Global evaluation metrics, such as WA-MPJPE$_{100}$, W-MPJPE$_{100}$, relative trajectory error (RTE), temporal jitter, and foot-sliding, are reported on datasets including RICH [20] and EMDB-2 [24]. As shown in Table 4.2, recent methods substantially reduce global pose error and temporal artifacts compared to naive combinations of camera tracking and per-frame pose estimation.

4.3.1 Observations and Failure Modes

Despite these advances, world-space methods typically rely on strong assumptions about camera motion or pre-segmented static scenes. Performance degrades in settings with static cameras, dominant dynamic humans, or unreliable camera motion estimates—conditions commonly encountered in broadcast video. Moreover, most approaches treat scene geometry implicitly or ignore it altogether, limiting their ability to enforce physically plausible human–scene interactions.

These observations highlight that while world-space human reconstruction has made

Table 4.3: **Global human-scene reconstruction performance on EMDB-2 [24] and RICH [20].** We report world-grounded motion metrics including WA-MPJPE$_{100}$, W-MPJPE$_{100}$, and relative trajectory error (RTE). **Opt. Free** indicates feed-forward inference, while **Scene** denotes explicit scene reconstruction.

Method	Opt. Free	Scene	EMDB-2			RICH		
			WA-MPJPE ↓	W-MPJPE ↓	RTE(%) ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE(%) ↓
TRAM [66]	✗	✗	76.4	222.4	1.4	127.8	238.0	6.0
WHAM [52]	✓	✗	135.6	334.8	6.0	108.4	190.1	4.5
GVHMR [51]	✓	✗	111.0	276.5	2.0	**78.8**	**126.3**	**2.4**
JOSH [35]	✗	✓	**68.9**	**174.7**	**1.3**	89.0	132.5	3.0
JOSH3R [35]	✗	✓	220.0	661.7	13.1	-	-	-
UniSH [31]	✓	✓	118.5	270.1	5.8	118.1	183.2	4.8
Human3R [31]	✓	✓	112.2	267.9	2.2	110.0	184.9	3.3

significant progress, robust global alignment in unconstrained environments remains challenging without explicit modeling of scene geometry and human–scene priors.

4.4 Human-Scene-Camera Reconstruction

4.4.1 Evaluation on Existing Benchmarks

We evaluate recent human–scene reconstruction approaches on global motion estimation with scene awareness. Table 4.3 evaluates global human-scene reconstruction methods on EMDB-2 [24] and RICH [20] using world-grounded motion metrics. Optimization-based approaches such as JOSH [35], which jointly refine human motion and scene structure through iterative inference, achieve the strongest global accuracy and temporal consistency. This highlights the effectiveness of explicit optimization for enforcing long-range constraints and resolving metric ambiguities.

4.4.2 Computational Efficiency Benchmark

We report a lightweight computational benchmark comparing representative feed-forward human-scene reconstruction models. Inference was performed on a monocular video consisting of 60 frames at a resolution of 512×288. We measure model throughput in frames per second (FPS), approximate floating-point operations (FLOPs) per frame, and parameter counts to highlight trade-offs between reconstruction fidelity and computational efficiency.

Table 4.4: Computational efficiency comparison for human–scene reconstruction models. Inference was performed on a 60-frame monocular video at 512×288 resolution.

Method	**FPS**	**FLOPs / Frame**	**Total Params**	**Trainable Params**
PROX	-	∼2.67 TFLOPs	15.73B	3.83B
HSfM	-	∼0.84 TFLOPs	2.98B	2.98B
Human3R	8.36	∼1.20 TFLOPs	1.17B	531M
UniSH	29.94	∼1.51 TFLOPs	1.64B	–

4.4.3 Observations and Failure Modes

Feed-forward methods such as GVHMR [51], TRAM [66], WHAM [52], and UniSH [31] trade some global accuracy for improved efficiency, robustness, and scalability. Notably, UniSH represents a unified feed-forward formulation that explicitly incorporates scene reconstruction alongside human motion, narrowing the performance gap to optimization-based methods while maintaining fast inference. These results suggest that while iterative optimization remains advantageous for global alignment, a unified feed-forward formulation provides a strong and practical foundation. *Motivated by this observation, the proposed research builds on a unified feed-forward framework and aims to further close the gap to optimization-based approaches through joint human-scene-camera alignment and physically grounded priors.*

Chapter 5

Conclusion

5.1 Proposed Contributions

This work advances the state of the art in 4D world reconstruction by introducing a unified framework that jointly recovers humans, scenes, and cameras in metric space. The main contributions are as follows:

1. **Unified Human–Scene–Camera Reconstruction formulation.** We formalize 4D world reconstruction as a single coupled inference problem in which humans, scenes, and cameras are estimated jointly, enabling bidirectional reasoning: humans provide constraints and cues for scene geometry and scale, while scenes and cameras provide metric grounding and physical constraints for human motion.

2. **World-latent fusion for cross-modal reasoning.** We propose a shared *world latent* representation that acts as the central fusion space for human and scene embeddings. Instead of treating cross-attention as a single isolated module, the world latent is updated recurrently over time via bidirectional interaction with both (i) scene tokens and propagated scene state and (ii) human tokens, producing globally consistent world-frame estimates.

3. **Strong pretrained priors for scene and human initialization.** We leverage strong pretrained representations to initialize both scene and human estimates: a stateful scene encoder provides dense scene features with temporal *state propagation* (predicting and passing forward the next state to condition subsequent frames), while a human encoder produces compact human embeddings and an initial parametric

body estimate. These initializations are not treated as final outputs; instead, they are refined through interaction with the shared world latent to improve robustness under occlusions, fast motion, and view changes.

4. **Implicit human–scene priors as regularization.** We incorporate human-centric priors, including gravity alignment, probabilistic contact consistency, and anthropometric height constraints, as *implicit regularizers* in the end-to-end objective. These priors are not direct network inputs; instead, they shape the learned world representation and predictions through dedicated loss terms that encourage metric consistency and physically plausible human–scene interaction.

5. **Toward fully controllable, in-the-wild video-based 4D world reconstruction.** Building on the proposed HSR model, the long-term goal is a fully controllable 4D world representation learned from *in-the-wild* monocular videos (including large-scale Internet video), analogous to how modern foundation models in NLP leverage web-scale corpora. Such a controllable world model should support simulation and data generation, including novel-view rendering, editable camera trajectories, and physically consistent human–scene interaction, enabling scalable synthetic data generation, training embodied policies, and closed-loop evaluation.

5.2 Research Timeline

Table 5.1: Proposed PhD Research Timeline

Term	**Planned Progress and Milestones**
Fall 2024	• Completed two graduate courses on (i) Generative AI and LLMs; (ii) Deep Reinforcement Learning
Winter 2025	• Completed coursework in Modern Computer Vision • Published three research papers on (i) Synthetic in-the-wild 3D data generation; (ii) Multi-player detection and tracking; and (iii) Puck detection and trajectory estimation

Continued on next page

Term	Planned Progress and Milestones
Summer 2025	• In-person research with collaborators (Baltimore Orioles)
Fall 2025	• Resolved data synchronization issues in high-resolution baseball videos with broadcast ground truth • Enhanced biomechanical modeling (e.g., extreme shoulder external rotation, bone kinematics inconsistencies)
Winter 2026	• Take Comprehensive Examination • 3 papers submitted and accepted into CVPRW on (i) Injury-Risk Screening from monocular videos; (ii) Pre-Release Baseball Pitch Type Anticipation; and (iii) Synthetic domain-specific 4D human generation
Summer 2026	• Prepare WACV submission on human–scene reconstruction
Fall 2026	• Prepare CVPR submission on unified human–scene–camera reconstruction with context priors (core PhD contribution)
Winter 2027	• Develop a video-aware 4D world reconstruction framework extending bidirectional human–scene reconstruction
Summer 2027	• Build a full navigable 4D world simulation from monocular video enabling novel-view synthesis, free virtual camera control, dynamic relighting, and basic physical interaction
Fall 2027	• Complete thesis writing and revisions • Defend PhD and submitted final dissertation

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *CVPR*, pages 15935–15946, 2022.

[2] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, June 2023.

[3] Jerrin Bright, Bavesh Balaji, Harish Prakash, Yuhao Chen, David A Clausi, and John Zelek. Distribution and depth-aware transformers for 3d human mesh recovery. *arXiv:2403.09063*, 2024.

[4] CDC. Anthropometric reference data for children and adults. https://www.cdc.gov/nchs/fastats/body-measurements.htm, 2025.

[5] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. *arXiv:2510.06219*, 2025.

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. *CoRR*, abs/2008.09047, 2020.

[7] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *NEURIPS*, 2024. Decompose-recompose approach with dynamic Gaussian Splatting for multi-object 4D scenes.

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*, 2025.

[9] Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *CVPR*, pages 682–692, June 2023.

[10] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, pages 1323–1333, 2024.

[11] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *ICCV*, pages 14861–14872, 2023.

[12] Alex Fisher, Ricardo Cannizzaro, Madeleine Cochrane, Chatura Nagahawatte, and Jennifer L Palmer. Colmap: A memory-efficient occupancy grid mapping framework. *Robotics and Autonomous Systems*, 142:103755, 2021.

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv:2107.08430*, 2021.

[14] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, pages 14783–14794, 2023.

[15] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, June 2023.

[16] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019.

[17] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016.

[19] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *ICCV*, pages 15977–15987, 2023.

[20] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, pages 13274–13285, 2022.

[21] Zeren Jiang, Chenyang Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Mesh4d: 4d mesh reconstruction and tracking from monocular video, 2026. Feed-forward 4D mesh reconstruction with deformation fields from monocular RGB videos.

[22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *CoRR*, abs/1712.06584, 2017.

[23] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, pages 5614–5623, 2019.

[24] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *ICCV*, pages 14632–14643, 2023.

[25] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv:2509.13414*, 2025.

[26] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *ICCV*, pages 11035–11045, 2021.

[27] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, pages 11605–11614, 2021.

[28] Maksim Kuprashevich and Irina Tolstykh. Mivolo: Multi-input transformer for age and gender estimation. In *International conference on analysis of images, social networks and texts*, pages 212–226. Springer, 2023.

[29] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds, 2024. Sparse SE(3) motion scaffold graph for smooth deformation in casual monocular videos.

[30] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.

[31] Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, et al. Unish: Unifying scene and human reconstruction in a feed-forward pass. *arXiv:2601.01222*, 2026.

[32] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13147–13156, 2022.

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. *CoRR*, abs/2012.09760, 2020.

[34] Qingming Liu, Yuan Liu, et al. Modgs: Dynamic gaussian splatting from casually-captured monocular videos with depth priors, 2025. Accepted at ICLR 2025. Depth-guided dynamic Gaussian optimization from monocular videos.

[35] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *arXiv:2501.02158*, 2025.

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), nov 2015.

[37] Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4rc: 4d reconstruction via conditional querying anytime and anywhere, 2026. Encode-once, query-anytime 4D latent space with conditional transformer decoder.

[38] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, pages 752–768. Springer, 2020.

[39] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21948–21958, 2025.

[40] Sarah-Maria Müller, Joël Floris, Sabine Rohrmann, Kaspar Staub, and Katarina L Matthes. Body height among adult male and female swiss health survey participants in 2017: Trends by birth years and associations with self-reported health status and life satisfaction. *Preventive Medicine Reports*, 29:101980, 2022.

[41] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[42] Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. Physic: Physically plausible 3d human-scene interaction and contact from a single image. *arXiv:2510.11649*, 2025.

[43] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. 2019.

[44] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571. IEEE, 2025.

[45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[46] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *CVPR*, 2025. Camera-agnostic monocular 3D estimation via spherical harmonics representation.

[47] PySceneDetect Developers. Pyscenedetect documentation. https://pyscenedetect.readthedocs.io/. Online; accessed January 31, 2026.

[48] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE transactions on robotics*, 34(4):1004–1020, 2018.

[49] Leila Goli Sabour et al. Romo: Robust motion segmentation improves structure from motion, 2024. Zero-shot motion segmentation to enhance SfM in dynamic scenes.

[50] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *ICCV*, pages 14761–14771, 2023.

[51] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024.

[52] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, pages 2070–2080, 2024.

[53] Colton Stearns, Adam W. Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *ACM SIGGRAPH Asia 2024 Conference Papers*, 2024.

[54] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[55] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. *CoRR*, abs/1711.08229, 2017.

[56] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *NEURIPS*, 34:16558–16569, 2021.

[57] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *NEURIPS*, 36:39033–39051, 2023.

[58] Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

[60] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617, 2018.

[61] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[62] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[63] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024.

[64] Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou. Freetimegs: Free gaussian primitives at anytime and anywhere for dynamic scene reconstruction, 2025. Flexible 4D Gaussians with adaptive appearance/disappearance for complex motions.

[65] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable Permutation-Equivariant Visual Geometry Learning. *arXiv e-prints*, pages arXiv–2507, 2025.

[66] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*, pages 467–487. Springer, 2024.

[67] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *ECCV*, pages 125–145. Springer, 2022.

[68] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild, 2023.

[69] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *ECCV*, pages 20–37. Springer, 2024.

[70] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *CVPR*, pages 13232–13242, June 2022.

[71] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv:2410.03825*, 2024.

[72] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *CVPR*, pages 8877–8886, June 2023.

[73] Jieming Zhou, Tong Zhang, Zeeshan Hayder, Lars Petersson, and Mehrtash Harandi. Diff3dhpe: A diffusion model for 3d human pose estimation. In *ICCV*, pages 2092–2102, October 2023.

[74] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023.

Appendix A

Supplementary Materials

A.1 Existing Benchmarked Datasets

A.2 EMBD

EMBD [24], standing for Electromagnetic DataBase of Global 3D Human Pose and Shape in the Wild. It comprises 105,000 frames, equivalent to 58 minutes of challenging 3D human motion recorded in diverse scenes. The dataset includes high-quality pose annotations using SMPL parameters, shape annotations, global body root trajectories, and camera trajectories. It encompasses 81 sequences distributed across 10 participants (5 males and 5 females), with videos captured using a hand-held mobile phone.

In terms of scene geometry and 3D scene representation, EMBD includes camera motion but does not offer explicit scene depths or groundtruth scene geometry. This limitation implies that it cannot directly support scene geometry learning without additional processing or augmentation from external sources. The groundtruth quality is achieved through a capture method employing body-worn, wireless electromagnetic (EM) sensors combined with a hand-held iPhone, resulting in an expected accuracy of 2.3 cm positional error and 10.6 degrees angular error, which surpasses previous in-the-wild datasets.

Primarily, EMBD is designed to support accurate 3D human pose and shape estimation in unconstrained environments, including global body and camera tracking. However, it is not intended for scene reconstruction, as it lacks 3D geometry as part of the dataset. This dataset is utilized in models such as JOSH [35] for enhancing global pose estimation capabilities.

A.2.1 EMDB-2

EMDB-2 serves as a specialized 25-sequence subset of the original EMBD dataset, specifically curated for global human motion evaluation. It inherits the core attributes of EMBD, focusing on high-quality 3D SMPL pose and shape parameters with global trajectories. This subset is employed in models like UniSH [31] and Human3R [5] to assess the accuracy of global human motion estimation, providing a targeted benchmark for evaluating how well algorithms handle extended motion sequences in diverse settings.

A.2.2 EMDB-1

EMDB-1 is another subset of EMBD, with a primary emphasis on local motion and camera-relative reconstruction. It complements EMDB-2 by shifting the focus to per-frame predictions of human body pose and shape. This subset is particularly used in Human3R [5] to evaluate algorithm performance in reconstructing human poses and shapes on a frame-by-frame basis, offering a balanced counterpart to the global-oriented EMDB-2.

A.3 SLOPER4D

SLOPER4D [9] is a novel scene-aware dataset collected in large urban environments, aimed at facilitating research on global human pose estimation with human-scene interactions in the wild. The dataset records activities from 12 human subjects across 10 diverse urban scenes, consisting of 15 sequences with trajectory lengths exceeding 200 meters and up to 1,300 meters. It includes over 100,000 LiDAR frames, 300,000 video frames, and 500,000 IMU-based motion frames.

The modality and content involve data retrieved via a head-mounted device integrated with LiDAR and a camera. Frame-wise annotations encompass 2D keypoints, 3D pose parameters, and global translations, accompanied by reconstructed scene point clouds. SLOPER4D provides rich annotations and reconstructed 3D scene meshes, offering multi-modal spatial context through LiDAR point clouds, RGB videos with 2D/3D annotations, accurate global human pose annotations, and the reconstructed scene.

Regarding scene geometry and 3D scene representation, SLOPER4D features explicit scene geometry as an integral part, making it highly useful for training scene-aware models. The groundtruth quality is enhanced by a joint optimization method that fits local SMPL

meshes to the scene and fine-tunes camera calibration, resulting in plausible and scene-natural 3D human poses. This suggests a sophisticated processing pipeline beyond raw capture, optimizing pose and camera calibration frame by frame for consistency and quality.

Primarily, SLOPER4D focuses on human pose and scene interaction, designed for global human pose estimation with scene context rather than solely human reconstruction. It can support human-scene reconstruction in terms of globally consistent 3D humans within real 3D scenes (with proper interactions), but it is not optimized for full scene reconstruction, as scene geometry is provided as contextual LiDAR reconstruction rather than a learning target. This dataset is used in JOSH [35] for advancing scene-aware pose estimation.

A.4 RICH

RICH [20], which stands for Real Scenes, Interaction, Contact, and Humans. It includes multiview outdoor and indoor video sequences at 4K resolution, groundtruth 3D human bodies captured using markerless motion capture, 3D body scans, and high-resolution 3D scene scans. In total, it comprises 142 single or multi-person multiview videos, with 90,000 posed 3D body meshes, 90,000 dense full-body contact labels in both SMPL-X and SMPL mesh topologies, and 577,000 4K images.

The modality and content feature high-resolution multiview images of single or multiple subjects interacting with scanned 3D scenes, dense full-body scene-contact labels, high-quality outdoor/indoor scene scans, high-quality 3D human shapes and poses, and dynamic backgrounds with moving cameras. It also provides accurate vertex-level contact labels on the body.

For scene geometry and 3D scene representation, RICH offers explicit scene geometry through high-resolution 3D scene scans that enable dense body-scene contact annotation, including vertex-level contact data that relies on accurate 3D scene geometry and human mesh information. The groundtruth quality is derived from markerless motion capture in multiview setups, effectively resolving occlusions to produce better-reconstructed bodies and scene contacts. Scene contact labels are generated via precise geometric alignment.

Primary uses of RICH include supporting reasoning about 3D human-scene contact from images, serving as a monocular or multiview human pose and shape benchmark, and potentially aiding end-to-end scene reconstruction (though it targets contact and interactions rather than full scene rebuild). It is also employed for human 3D reconstruction and 3D scene geometry tasks. In specific models, RICH is used in JOSH, UniSH, Human3R,

and PhySIC; for instance, in UniSH, it evaluates the model's accuracy in estimating global human motion.

A.5 Bonn Dataset

The Bonn Dataset [43], is typically employed in human-scene reconstruction and dynamic SLAM research. It is designed to challenge RGB-D SLAM and reconstruction systems with highly dynamic indoor scenes where dynamic elements, such as people and moving objects, interact within otherwise static environments. The dataset contains 24 dynamic sequences and 2 static sequences as baseline references, capturing scenarios like playing with balloons that cause significant motion and occlusions.

Modality and content include RGB and depth sequences, with registered RGB frames, depth maps, and intrinsic camera calibration, suitable for 3D mapping and pose estimation. Dynamic content ranges from person tracking to moving obstacles. For scene geometry and 3D scene representation, it provides a high-resolution groundtruth 3D point cloud of the static environment, captured using a Leica BLK360 terrestrial laser scanner, comprising approximately 394 million points. Static scene models are available in PLY format, allowing evaluation of reconstructed geometry against high-precision scans.

Groundtruth quality features per-frame pose labels from an OptiTrack Prime 13 motion capture system for camera trajectories, with data acquired using an ASUS Xtion Pro LIVE RGB-D sensor. Primary uses encompass benchmarking dynamic SLAM and mapping algorithms for robustness to moving objects, evaluating camera tracking accuracy under dynamic conditions, and testing reconstructed scene geometry against the groundtruth model. It is used in UniSH [31] for related evaluations.

A.6 BEDLAM

BEDLAM [2] is a synthetic dataset that addresses limitations in real datasets by offering large-scale variations, though it may be constrained in scene diversity compared to real captures. It contains monocular RGB videos with groundtruth 3D bodies in SMPL-X format, enabling the training of 3D human pose and shape regressors. The dataset emphasizes diversity in body shapes, motions, skin tones, hair, and clothing to enhance generalization. Clothing is realistically simulated on moving bodies using physics simulations, with varied lighting and camera motions for added realism. It includes around 10,450 sequences at 30

fps, comprising RGB images, segmentation masks, depth maps, and groundtruth SMPL-X parameters.

Scene geometry and 3D scene representation are not the primary focus; while environments are rendered, they serve as backgrounds for human motion synthesis rather than independent benchmarks with dense scans. Depth maps provide per-frame pixel-aligned depth information, useful for depth-aware models, but these are synthetic. Groundtruth quality is exact by design, generated from the SMPL-X model and rendering pipeline, with no physical capture hardware involved; instead, motions are animated and rendered with physics-based clothing and diverse lighting. The dataset supplies exact per-frame 3D parameters and CSV files for direct supervision.

Primary uses include training and evaluating 3D human pose and shape estimation from monocular video, benchmarking generalization to real datasets via synthetic training, and supporting research into human appearance and motion diversity. It is also applicable for tasks like depth estimation and segmentation. BEDLAM is utilized in UniSH [31] for coarse-grained alignment to learn initial localization and in Human3R [5] for related evaluations.

A.7 BEDLAM2

BEDLAM2 [58] is a large-scale synthetic video dataset which provides millions of frames specifically designed for training and evaluating 3D human pose and shape estimation methods, with a strong emphasis on estimating humans in world coordinates. The dataset includes accurate ground-truth 3D body parameters using the SMPL-X model and comprehensive camera data, enabling robust training and benchmarking of approaches that handle both human motion and dynamic camera movements.

The dataset consists of monocular synthetic video sequences at 1280×720 resolution and 30 fps, totaling 27,480 sequences and more than 8 million rendered PNG images. It provides SMPL-X body parameters (3D pose and shape) for each frame in world coordinates, along with camera extrinsics and intrinsics for all frames. These sequences feature multiple 3D environments with diverse lighting conditions (HDRI-based from Poly Haven), realistic rendering of clothing, strand-based hair, shoes (as displacement maps), and body textures. Depth maps are available for approximately 44% of the images (in 16-bit EXR format).

Regarding scene geometry and 3D scene representation, BEDLAM2 generates full synthetic 3D environments, including diverse indoor and outdoor scenes with realistic lighting, occlusions, and geometry. It also produces ground-truth depth maps for the rendered

scenes, though these are synthetic and primarily support human-focused tasks rather than full scene reconstruction. The dataset incorporates realistic motion blur to enhance visual fidelity. Rendering is performed using Unreal Engine 5.3.2, with no physical capture involved; everything is generated through a synthetic pipeline.

The groundtruth quality is exact by construction, derived directly from the SMPL-X parameterization, animation pipeline, and rendering process. This ensures precise per-frame 3D body parameters, camera poses, and associated metadata (provided in CSV and JSON formats).

Primary uses of BEDLAM2 include 3D human pose and shape estimation from monocular video, camera motion estimation, human motion analysis in world coordinates, and related tasks such as 3D/4D point tracking, structure from motion with non-rigid elements, depth estimation, optical flow, and dynamic scene understanding. While it can support aspects of scene reconstruction through its rendered environments and depth maps, the dataset is more tailored toward human pose and shape estimation due to the synthetic nature of the scene rendering (less rich and varied than real-world 3D scans).

Compared to the original BEDLAM dataset, BEDLAM2 offers several significant improvements and benefits:

- Much larger scale and richer content, with over 8 million frames providing substantially greater diversity for training.
- Enhanced camera diversity and realism, introducing a wide range of realistic camera motions (e.g., zoom, orbit, panning, tracking) and varying focal lengths to better simulate dynamic, real-world camera behaviors—addressing limitations in the original BEDLAM's more restricted camera variation.
- Increased realism and variation in human appearance, including more diverse body shapes, strand-based realistic hair, simulated clothing, shoes, and improved skin textures.
- Broader scene variation with more 3D environments and enhanced lighting diversity.
- More comprehensive ground-truth data, leading to state-of-the-art improvements when training methods (particularly those estimating humans in world coordinates) compared to the original BEDLAM.

BEDLAM2 builds on the foundation of BEDLAM by significantly advancing synthetic data quality for 3D human understanding tasks, making it a powerful resource for models requiring robust generalization to real-world videos with complex motions and viewpoints.

A.8 3DPW

3DPW [60], referring to 3D Poses in the Wild, is designed for recovering accurate 3D human pose in unconstrained environments using IMUs and a moving camera, distinguishing itself from datasets limited to small recording volumes by incorporating video from a moving phone camera. The dataset includes 60 monocular RGB video sequences captured in real-world indoor and outdoor settings, with synchronized RGB imagery, 2D pose annotations, 3D human poses, and camera poses for every frame. Auxiliary assets comprise 3D body scans and 18 re-poseable human models with clothing variations.

Scene geometry and 3D scene representation do not include dense 3D scene geometry like meshes or LiDAR scans; the emphasis is on human body and camera motion, with provided camera poses enabling trajectory analysis. Environments are part of the video but not captured as explicit 3D data. groundtruth quality fuses video and IMU data for accurate 3D human pose annotations and camera poses. The dataset is split into train/validation/test subsets with evaluation protocols for benchmarking.

Primary uses focus on 3D human pose estimation in unconstrained environments and evaluation of pose and camera motion approaches. In Human3R [5], it is specifically used to assess human pose and shape reconstruction.

A.9 PROX

PROX [16], standing for Proximal Relationships with Object eXclusion, is a dataset used in human-scene interaction and pose estimation research. It comprises RGB-D video recordings of humans interacting with static 3D scenes. The dataset has two parts: Qualitative PROX with around 100,000 RGB-D frames of 20 subjects in 12 indoor scenes, and Quantitative PROX with 180 static RGB-D frames of a single subject captured with motion capture. It includes RGB images, depth maps, camera data from Kinect-One sensors, and SMPL-X body model fittings for many frames.

Scene geometry and 3D scene representation feature high-resolution scanned 3D static scene meshes for each environment, along with signed distance fields for geometric reasoning. Static scene models are aligned with RGB-D data via camera-to-world transformations, accompanied by calibration files and background geometry. Groundtruth quality in Qualitative PROX involves 30 fps RGB-D recordings with Kinect-One sensors for dynamic interactions, while Quantitative PROX uses synchronized motion capture for precise pose

and shape. Body meshes are fitted and refined, with provided camera intrinsics and extrinsics for alignment. Primary uses include human-scene interaction evaluation, physics and contact benchmarks, pose and shape estimation, and scene reconstruction. It is employed in PhySIC [42] for related tasks.

A.10 Benchmarking Literature Review

Tables A.1–A.3 summarize and contrast these approaches from complementary perspectives. Table A.1 provides a high-level comparison of objectives, inputs, outputs, and practical trade-offs, while Table A.2 highlights differences in human and scene representations, metric scale handling, and dynamic modeling. Finally, Table A.3 details architectural and optimization strategies, revealing a clear shift from optimization-heavy pipelines toward unified, feed-forward models with shared metric coordinate systems. Collectively, these methods illustrate the evolving design space of human–scene reconstruction and motivate the adoption of unified, metric, and temporally consistent models for challenging real-world settings such as sports videos.

Taken together, these trends are summarized quantitatively and structurally in Tables A.5–A.7. The tables distill how recent methods differ in their goals, representations, and architectural choices, ranging from relative-scale dense pointmap regression to unified, metric-scale, feed-forward formulations that jointly predict geometry, camera parameters, and, in some cases, motion. In particular, they highlight a shift away from iterative, optimization-heavy pipelines toward single-stage transformer architectures with structured decoders and multi-view attention, as well as the growing importance of metric consistency and temporal coherence. This perspective provides a compact reference for comparing scene priors and design trade-offs, and directly motivates their reuse and extension in human–scene reconstruction settings.

A.11 Future Research Directions

1. **Benchmarking controllable 4D world models for embodied AI.** Once a video-based, fully controllable 4D world reconstruction model is established, an important next step is to evaluate it as a *training substrate* for embodied agents. For humanoid robotics, the reconstructed world can be queried from an ego-centric perspective by synthesizing camera trajectories aligned to a person's viewpoint and generating consistent observations (geometry, contacts, and motion) for policy learning. This

enables benchmarking whether policies trained with world-model-generated data improve robustness and generalization relative to policies trained with conventional simulation or limited real data. A similar evaluation can be performed for autonomous navigation by extracting controllable, view-consistent trajectories through reconstructed environments and measuring improvements in downstream tasks such as localization, obstacle avoidance, and long-horizon planning.

2. **Scaling evaluation and task diversity.** Beyond reconstruction accuracy, future work should develop standardized benchmarks that measure controllability, physical consistency, and downstream utility across diverse in-the-wild Internet videos. This includes stress-testing under occlusion, fast motion, challenging illumination, and crowded scenes, and quantifying how well the model supports counterfactual camera control (e.g., novel viewpoints) and interaction-consistent rollouts.

3. **Unified scene–human encoding.** A further architectural direction is to replace separate human and scene encoders with a single, jointly trained encoder that produces a shared tokenization of both people and environment. Such a unified representation may improve bidirectional coupling, reduce failure modes caused by module boundaries, and enable cleaner end-to-end learning of a world latent that jointly captures dynamic humans and static scene structure.

Table A.1: General comparison of human–scene reconstruction methods.

Method	**Primary Goal**	**Input**	**Output**	**Strengths**	**Potential Limitations**
JOSH	Joint 4D human–scene reconstruction via optimization; jointly refines human motion, scene geometry, and camera poses	Monocular videos	4D human motion and dense 3D scene geometry with optimized camera poses	Joint optimization improves overall reconstruction quality; works on in-the-wild videos	Highly dependent on initialization quality; performance degrades when human–scene contacts are sparse or invisible
Human3R	Unified, single-stage 4D human–scene reconstruction from casual videos	Monocular RGB video	Global multi-person SMPL-X meshes, dense 3D scene point clouds, and camera trajectories	True one-stage feed-forward inference; real-time performance with low memory usage	Relies on head visibility for human detection; does not model clothing or appearance
HSfM	Joint reconstruction of humans, scenes, and cameras with metric scale	Multi-view images	Human meshes in shared world frame, dense metric scene point clouds, and camera parameters	Improved accuracy through joint reasoning; resolves scale ambiguity using human priors	Designed for multi-view inputs; applicability to monocular videos is not explicitly demonstrated
UniSH	Metric-scale 4D human–scene reconstruction with unified reasoning	Monocular videos, multi-view images, unordered image sets	Metric human meshes, dense scene surfaces, camera poses, temporally consistent 4D outputs	Efficient feed-forward inference; strong generalization via pre-trained human and scene models	Performance degrades under extreme motion or occlusion; sensitive to noisy initial estimates
PhySIC	Physically plausible joint human–scene reconstruction from a single image	Single monocular RGB image (one or more humans and a scene)	Metric SMPL-X human meshes, dense scene surfaces, and vertex-level contact maps	Physically plausible interactions; occlusion-aware reasoning; metric-scale outputs	Single-image only (no temporal motion); depends on monocular depth and robust initialization for occluded regions

Table A.2: Comparison of representations, geometry, and modelling strategies in human–scene reconstruction.

Method	**Human Representation**	**Scene Representation**	**Architectural Framework**	**Metric Scale**	**Dynamic Modelling**
JOSH	SMPL human meshes with full 4D motion	Dense scene point clouds and depth maps	Optimization-based system refining pre-trained human and scene initializations	Yes	Explicit temporal modeling via joint optimization and contact constraints
Human3R	Multi-person SMPL-X meshes in world frame	Dense 3D scene point clouds reconstructed per frame	Unified feed-forward framework built on CUT3R with prompt tuning	Yes	Online 4D reconstruction via world-frame multi-person tracking
HSfM	Multiple SMPL-X human meshes in a shared world frame	Dense metric scene point clouds	Optimization-based integration of SfM and human reconstruction	Yes	No explicit temporal tracking of human motion
UniSH	Full 3D human meshes with temporally smooth motion	Dense metric scene surfaces	Unified feed-forward network leveraging Pi3 and CameraHMR priors	Yes	Captures 4D human motion and joint human–scene dynamics
PhySIC	SMPL-X parameterized humans (pose, shape, articulation)	Dense scene surface reconstructed from fused monocular depth with occlusion-aware inpainting	Multi-stage monocular prediction + depth fusion + joint refinement/optimization	Yes	None (single-image static reconstruction)

Table A.3: Detailed architectural comparison of human–scene reconstruction methods.

Method	**Architectural Framework**	**Scene Modelling**	**Interaction Modelling**	**Training Strategy**
JOSH	Pure optimization-based framework refining pretrained initializations	Dense scene geometry jointly optimized with human motion and cameras	Explicit human–scene contact constraints drive coupling	No end-to-end learning; relies entirely on optimization
Human3R	Unified single-stage feed-forward network built on CUT3R	Dense scene geometry reconstructed in global/world coordinates	Implicit joint reconstruction without explicit contact modelling	End-to-end training on BEDLAM using parameter-efficient prompt tuning
HSfM	Joint optimization combining data-driven models with SfM	Metric scene point clouds refined via joint optimization	Implicit coupling through joint camera and human optimization	No end-to-end training; pretrained models used for initialization
UniSH	Unified feed-forward architecture with scene branch, human branch, and AlignNet	Dense pointmap-based scene reconstruction using Pi3 backbone	Unified metric coordinate system with AlignNet-based coupling	Surface distillation from expert depth models; coarse-to-fine supervision on in-the-wild data
PhySIC	Multi-stage pipeline: monocular scene estimation, human reconstruction/alignment, joint refinement	Metric-scale scene depth fusion with occlusion-aware inpainting to complete unseen surfaces	Vertex-level contact maps and physics-inspired constraints to avoid interpenetration and enforce support	Hybrid use of pretrained models with confidence-weighted joint optimization (no end-to-end training)

Aspect	**Optimization-based Methods** (e.g., JOSH [35])	**Feed-forward Methods** (e.g., Human3R [5], UniSH [31])
Human–Scene Coupling	Strong, explicit bidirectional coupling via shared optimization objectives and constraints	Weak or indirect coupling; interaction typically limited to global alignment (e.g., scale, translation)
Use of Human–Scene Priors	Explicit modeling of contact, support, and collision constraints	Implicit or absent; priors not enforced at the geometric level
Influence of Humans on Scene	Human motion and contacts directly influence scene geometry refinement	Scene geometry is largely invariant to human motion
Influence of Scene on Humans	Scene geometry constrains human pose and motion during optimization	Human reconstruction weakly conditioned on scene context
Metric Consistency	Naturally enforced through joint optimization in a shared world frame	Often achieved via post-hoc alignment or scale prediction
Initialization Sensitivity	Highly sensitive to quality of initial human, scene, and camera estimates	Less sensitive due to amortized inference
Computational Cost	High; requires iterative optimization per sequence	Low; single forward pass enables real-time or near real-time inference
Scalability	Poor scalability to long videos or large datasets	Highly scalable and suitable for large-scale deployment
Robustness to Occlusion and Noise	Fragile under occlusion, fast motion, or missing contacts	More robust to noisy or incomplete observations
Generalization	Limited generalization; optimization must be tuned per scenario	Strong generalization when pretrained on large-scale in-the-wild data

Table A.4: Comparison between optimization-based and feed-forward human–scene reconstruction approaches.

Table A.5: General comparison of recent scene reconstruction methods.

Method	**Primary Goal**	**Input**	**Output**	**Strengths**	**Potential Limitations**
DUSt3R	Dense 3D reconstruction without camera calibration or known extrinsics	Unconstrained RGB images	Dense 3D scene geometry (point maps, depth, correspondences)	Unified geometry without calibration; works for single or multiple images	Limited handling of dynamic scenes; scalability issues for large image sets; primarily designed for image pairs
MASt3R	Improve dense correspondence accuracy by treating matching as a 3D task	Image pairs	Dense correspondences and point maps	Significantly improved matching accuracy; strong performance on map-free localization	High computational cost; not suitable for real-time applications; slow for single image pairs
VGGT	Unified prediction of multiple 3D scene attributes from images	One or more images	3D point tracks, depth maps, camera parameters	Single model for multiple 3D tasks; efficient feed-forward inference	Performance degrades with large image sets or complex geometry; no support for fisheye or panoramic cameras
Pi3	Reference-free visual geometry reconstruction	Images, unordered image sets, or videos	Camera poses and 3D scene structure	Removes dependence on a fixed reference view; robust to input ordering	Scale-invariant (no metric scale); limited support for dynamic scenes
MapAnything	Metric-scaled 3D geometry and camera estimation in a single pass	Images with optional geometric inputs	Metric 3D scene geometry and camera poses	Produces metric-scale outputs; supports many geometric inference tasks	Requires complex optional inputs for best performance; does not model uncertainty in inputs
Any4D	Dense metric 4D reconstruction with geometry and motion	Videos or multi-view images with optional multi-modal inputs	Dense 4D geometry with per-pixel motion	Metric-scale outputs; joint geometry and motion prediction	Relies on simulated or clean multi-modal data; assumes reference object appears in the first frame

Table A.6: Comparison of scene representations, geometry modeling, and architectural design choices.

Method	**Scene Representation**	**Backbone**	**Inference Method**	**Metric Scale**	**Stage Design**	**Dynamic Scene Handling**
DUSt3R	Dense per-pixel 3D point maps	Transformer-based	Feed-forward with optional optimization	Relative	Multi-stage	Assumes static rigid scenes
MASt3R	Dense 3D point maps	Transformer-based	Feed-forward with matching post-processing	Relative	Multi-stage	Assumes static rigid scenes
VGGT	Dense point maps and point tracks	Transformer-based	Feed-forward	Relative	Single-stage	Static geometry only
Pi3	Local point maps	DINOv2-based ViT	Feed-forward	Scale-invariant	Single-stage	Static multi-view geometry
MapAnything	Factored multi-view geometry (depth, rays, poses)	Transformer-based	Feed-forward	Metric	Single-stage	Static geometry only
Any4D	Dense per-pixel 4D geometry and motion	Multi-view transformer	Feed-forward	Metric	Single-stage	Explicit per-pixel scene flow

Table A.7: Detailed architectural comparison of recent scene reconstruction methods.

Method	**Backbone**	**Post-Backbone Design**	**Attention Structure**	**Prediction Heads**	**Training Details**
DUSt3R	Transformer encoder-decoder (CroCo-inspired)	Pairwise point map regression with global alignment	Self-attention followed by cross-attention between views	Per-view dense point map and confidence heads	Trained on 8 diverse datasets using 3D regression and confidence-aware losses
MASt3R	DUSt3R transformer backbone	Adds dense matching head for feature correspondence	Shared multi-head attention across geometry and matching tasks	Point map regression head and dense descriptor head	Trained on 14 datasets with matching and InfoNCE losses; coarse-to-fine correspondence sampling
VGGT	Large transformer without explicit geometric priors	Unified transformer outputs for multiple 3D tasks	Alternating frame-wise and global self-attention	Camera, depth, and point map heads	Multi-task training on ScanNet, MegaDepth, CO3Dv2, and others
MapAnything	Multi-view transformer backbone	Factored decoders for depth, pose, and scale	Alternating multi-view attention	DPT geometry head, pose head, scale MLP	End-to-end training with pose, ray, and scale losses across 13 datasets
Any4D	Multi-view transformer backbone	Dual DPT heads for geometry and motion	Alternating multi-view attention	Geometry DPT, motion DPT, pose and scale decoders	Trained with mixed static and dynamic datasets using geometry and scene-flow supervision