# CRISP: Contact-guided Real2Sim from Monocular Video with Planar Scene Primitives

Anonymous ICCV submission

Paper ID 42

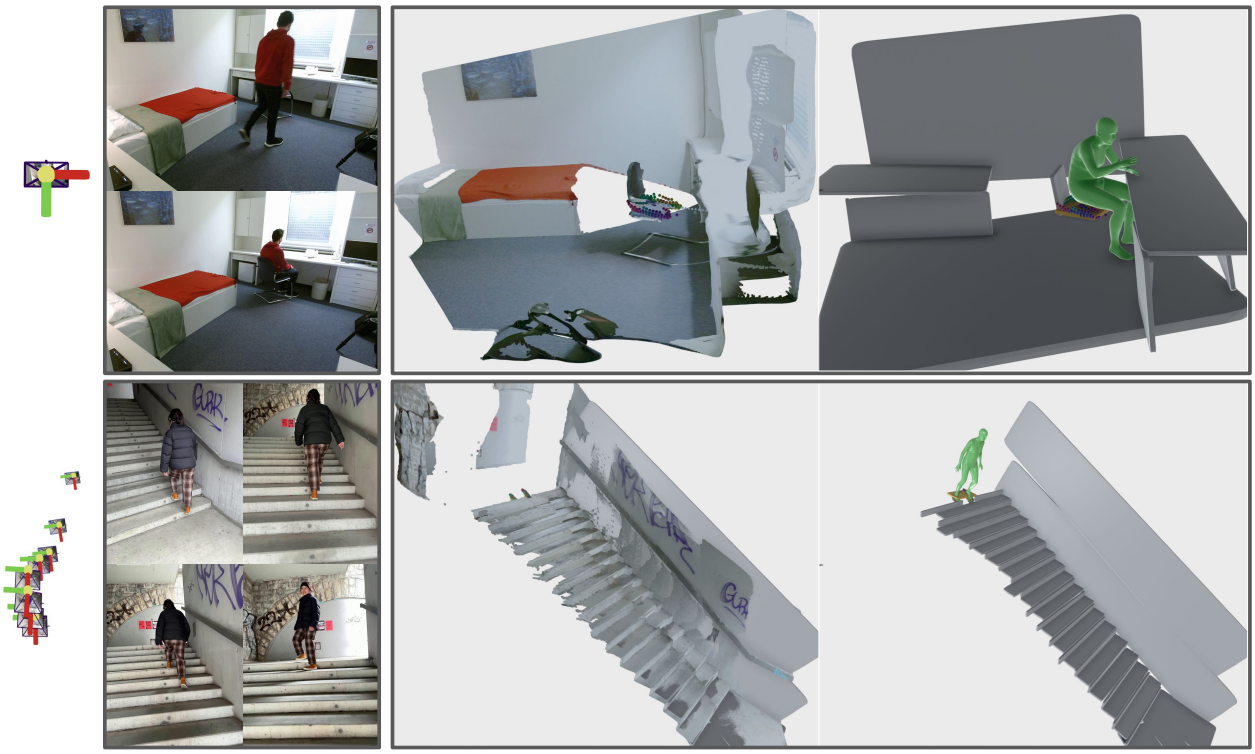

Figure 1. We present CRISP, a framework for seamless real-to-sim human–scene interaction from monocular video. From unconstrained video of a human interacting with a static environment (left), our pipeline reconstructs the camera, 4D human motion, and a scene point cloud in world coordinates. Estimated contact points (colored dots) serve as geometric anchors to recover occluded regions as a complementary point cloud (middle). Finally, compact planar primitives are fitted to support contact-faithful physics in simulation (right). Gray primitives are derived from the scene point cloud, while yellow primitives are derived from the contact point cloud.

## Abstract

*Modeling contact-accurate human–scene interaction from monocular video is a crucial step toward real-to-sim transfer in computer vision and robotics. However, this task remains highly challenging due to the inherent ambiguities of monocular perception and the limitations of Human Mesh Recovery (HMR) and single-view 3D geometry estimation. Existing methods often fail to capture reliable contact and scene structure, making them unsuitable for converting in-the-wild videos into simulation-ready assets. In this work, we introduce CRISP, a framework that integrates HMR, 4D reconstruction, and contact prediction into a unified front-end for recovering human motion, scene structure, and contact cues. These signals jointly guide the completion of occluded geometry, after which we fit compact planar primitives that merge the scene point cloud and the contact point cloud into a unified, simulation-friendly representation. Finally, we integrate the reconstructed assets into a physics-based simulator and use reinforcement learning to enforce*

*realistic, contact-faithful human–scene interaction. Our approach achieves over 97% success rate on human-centric video benchmarks (EMDB, PROX) and delivers 1.9× faster throughput for Reinforcement Learning training compared to prior pipelines. This demonstrates the ability of CRISP to generate paired human motion and interacting environments at scale, greatly advancing real-to-sim applications in robotics and embodied AI.*

## 1. Introduction

Humans constantly interact with their environments—sitting on chairs, lying on sofas, climbing stairs. Accurately recovering such human–scene interactions (HSI) from monocular video would unlock large-scale applications in embodied AI and robotics [1]. However, this problem remains highly challenging due to the inherent ambiguities of monocular perception and the limitations of Human Mesh Recovery (HMR) and single-view 3D geometry estimation.

Most prior work on HMR [14, 15, 18] focuses on human motion and neglects scene context. The resulting motions lack physical grounding and fail to provide paired scene geometry that can be faithfully simulated in physics engines. Recent methods [6, 15] that attempt joint human–scene recovery often rely on predicted contact cues to align humans with reconstructed scenes, but these cues are unreliable for complex, in-the-wild videos with frequent occlusions. The most related concurrent work, VideoMimic [1], directly imports dense reconstructed meshes into simulators. We find this strategy suboptimal: dense triangle meshes yield unstable contacts, slow training, and cannot robustly handle occluded regions. These limitations highlight a core challenge: reliably bridging monocular video to simulation requires representations that capture human–scene contact while remaining lightweight and stable for physics engines, which is a gap not addressed by existing methods.

We introduce CRISP, a contact-aware real-to-sim pipeline that converts monocular human-centric videos into contact-accurate, simulation-ready assets. CRISP integrates HMR, 4D reconstruction, and contact prediction into a unified front-end to jointly recover human motion, scene structure, and contact cues. These signals guide the completion of occluded geometry, which we represent using lightweight planar primitives that merge scene and contact point clouds into a compact, simulation-friendly form. Finally, we integrate these assets into a physics-based simulator, where reinforcement learning refines interaction dynamics to enforce realistic, contact-faithful human–scene interaction.

Our contributions are summarized: (1) We propose CRISP, a contact-aware real-to-sim pipeline. It converts unconstrained monocular videos into simulation-ready human motion and scene primitives explicitly designed for stable contact in physics engines. (2) Contact-guided scene completion. We leverage predicted body–scene contacts as geometric priors to complete occluded scene regions, followed by normal-guided planar fitting that produces compact primitives suitable for simulation. (3) Robust performance and efficiency. On human–scene benchmarks, CRISP attains a 96.9% real-to-sim success rate and achieves 1.9× faster reinforcement learning training throughput compared to baseline.

## 2. Related work

3D human motion recovery is most widely formulated as recovering the parameters of a parametric human model, such as SMPL [7] and its variants [10]. Recently, feed-forward HMR methods directly regress SMPL parameters in camera coordinates from images or videos [3] - however this does not recover the global motion trajectory, scene geometry, or human-scene contact. To recover world-grounded global motion trajectories, several recent works estimate cameras via visual odometry [14] or SLAM [18], but do not recover scene geometry or contact. To recover human-scene contact, WHAM [15] is trained to predict the likelihood of foot-ground contact using estimated contact labels from both AMASS and 3D video datasets. WHAM then trains a trajectory refinement network that updates the root orientation and velocity based on foot contact information. However, despite recent progress, recovering global motion trajectories together with scene geometry and contact remains a challenging problem. The final result is often inconsistent over time (involving jittering), or fails to recover accurate and physically-plausible motion within the context of the scene. In light of this, we propose to unify all these different threads: we recover world-grounded human motion jointly with scene geometry, and leverage human-scene contact signals within a physics-based simulator to provide contact-rich feedback and refine human motion via reinforcement learning.

## 3. Preliminary

**Human Mesh Recovery in camera coordinate.** We represent the per-frame 3D human body with SMPL [8], a parametric triangular mesh model. At time $t$, the posed mesh is $\mathbf{M}_t = \mathcal{M}(\boldsymbol{\theta}_t, \boldsymbol{\beta}, \mathbf{r}_t, \boldsymbol{\pi}_t) \in \mathbb{R}^{6890\times 3}$, where $\boldsymbol{\theta}_t \in \mathbb{R}^{23\times 3}$ are the relative joint rotations (axis–angle) of the 23 body joints, $\boldsymbol{\beta} \in \mathbb{R}^{10}$ encodes body shape, and $(\mathbf{r}_t, \boldsymbol{\pi}_t) \in \mathbb{R}^3 \times \mathbb{R}^3$ denote the root (global) orientation and translation with respect to the camera. For convenience we group the parameters as $\boldsymbol{\Theta}_t = (\boldsymbol{\theta}_t, \boldsymbol{\beta})$ and $\mathbf{T}_t = (\mathbf{r}_t, \boldsymbol{\pi}_t)$, so that $\mathbf{T}_t$ captures global rigid motion while $\boldsymbol{\Theta}_t$ captures local articulated pose and identity. The scale of the SMPL mesh is in metric units, represent-

ing the real size of the human in meters. In this work we adopt GVHMR [14], a transformer-based model trained in regression manner to predict SMPL pose $\boldsymbol{\theta}_t$ together with camera-aligned root motion $(\mathbf{r}_t, \boldsymbol{\pi}_t)$ from video. At inference, it outputs per-frame SMPL mesh in the camera coordinate frame $\mathbf{M}_t^{(c)} = \mathcal{M}(\boldsymbol{\theta}_t, \boldsymbol{\beta}, \mathbf{r}_t, \boldsymbol{\pi}_t)$.

# 4. Method

## 4.1. Human, Scene, and Camera Initialization

Given an unconstrained monocular video $\mathcal{V} = \{ I_i \in \mathbb{R}^{H\times W} \}_{i=1}^{N}$ depicting a human interacting with either a *static* scene $\mathcal{S}$ (e.g. parkour, stair climbing, sitting on a sofa) or *dynamic* hinged objects $\mathcal{O} = \{ O_i \}_{i=1}^{N}$ (e.g. a door or a scooter), our goal is to recover *paired* 3-D human motion and scene geometry in a world coordinate. A minimal solution must therefore infer camera poses $\mathcal{T}_i = [\mathcal{R}_i \mid t_i] \in \mathrm{SE}(3)$; camera intrinsics $\mathcal{K} \in \mathbb{R}^{3\times3}$; a per-frame dense depth map $\mathcal{D} = \{(d_i)\}_{i=1}^{N}$. We replace the depth estimator in optimization stage of MegaSAM [5] with MoGe[17], producing a *scale-invariant* dense point cloud $\mathcal{P}$ together with calibrated camera parameters $\{\mathcal{K}, \mathcal{T}_i\}_{i=1}^{N}$. The intrinsics $\mathcal{K}$ are then forwarded to GVHMR [14] to obtain a human-mesh reconstruction in camera space and lifted to the world frame ensuring that the human, scene, and camera share a single coordinate system. As $\mathcal{P}$ is scale–ambiguous, we recover metric scale by taking the median of per-frame scales calculated by dividing the depth of the human center from the SMPL mesh by the human depth from the predicted dense depth map. The scaled cloud $\tilde{\mathcal{P}}$ is therefore metrically consistent with both the camera calibration and the human mesh.

## 4.2. Normal–based planar primitive fitting

We adopt a planar-world assumption: we assume the scene structure can be modeled as a compact set of planar primitive. Despite its simplicity, this representation is sufficient for common human–scene interactions (sitting, lying down, walking, climbing stairs, etc.). Unlike methods that first learn a globally consistent neural field from 2D segmentation priors to extract planar primitives [**?** ], our approach directly estimates planes from 3D cues, yielding a lightweight and simulation-ready reconstruction. We estimate a normal field $\mathcal{F}$ from the point set $\mathcal{P}$, then run KMeans followed by DBSCAN to cluster points into planar patches. For each primitive, we align the local axis $\mathbf{z}$ with the scene normal vector $\mathbf{n}$, estimate the in-plane axes $(\mathbf{x}, \mathbf{y})$ via PCA, fit a plane using RANSAC, and expand along $(\mathbf{x}, \mathbf{y})$ to cover the inlier points; the thickness is a hyperparameter fixed as $0.05\,\mathrm{m}$. For dynamic camera setup, we also adopt off-the-shelf flow estimator [20] to infer plane correspondence across frames.

## 4.3. Goal-conditioned reinforcement learning.

Following [11], we train a fully-constrained motion-tracking policy $\pi^{\mathrm{FC}}$ that maps the current state of the humanoid to actuator torques. Because the dataset provides only *kinematic* reference clips, $\pi^{\mathrm{FC}}$ must infer the required motor commands. Our fully-constrained controller is trained end-to-end to imitate target motions by conditioning on the full-body motion sequence. The training objective is formulated as a motion-tracking reward and optimized using reinforcement learning. The objective of motion tracking is to predict the next actions based on the current character state, and a sequence of future target poses.

**Observations.** At each timestep, the policy observes the character's joint orientations and positions together with linear and angular velocities, all expressed in the root frame to remove global motion. We also provide a short look-ahead of the next N target poses: for each future step, the input contains joint-wise orientation and position deltas relative to the current pose plus a scalar time-to-target. The policy is trained end-to-end with asynchronous PPO [13] to track the reference motion while avoiding interpenetration and respecting joint limits, yielding stable contact-aware control in the reconstructed scene.

**Action space.** Following prior work [12], actions are parametized as desired joint targets for a Proportional–Derivative (PD) controller. The stochastic policy $\pi\big(a_t \mid s_t,\, g_t\big)$ is modelled as a multivariate Gaussian with fixed diagonal covariance $\boldsymbol{\Sigma} = \sigma_\pi^2 \mathcal{I}$.

**Reward Function.** The reward function $r_t$ encourages the agent to closely follow a reference motion by minimizing discrepancies between the simulated character's state and the target trajectory:

$$r_t = w^{\mathrm{gp}} r_t^{\mathrm{gp}} + w^{\mathrm{gr}} r_t^{\mathrm{gr}} + w^{\mathrm{rh}} r_t^{\mathrm{rh}} + w^{\mathrm{jv}} r_t^{\mathrm{jv}} + w^{\mathrm{jav}} r_t^{\mathrm{jav}} + w^{\mathrm{eg}} r_t^{\mathrm{eg}}, \tag{1}$$

where each $r_t^{\{\cdot\}}$ represents an individual reward component, and $w^{\{\cdot\}}$ denotes its corresponding weight. Specifically, the reward encourages accurate imitation of global joint positions (gp), global joint rotations (gr), root height (rh), joint velocities (jv), and joint angular velocities (jav).

# 5. Experiments

**Datasets.** We evaluate on the following datasets: EMDB [4] and PROX [2]. The EMDB dataset provides labels for global human motion without the scene model, and weuse the EMDB-2 subset with 21 sequences (4 indoor + 17 ourdoor) for evaluation. The PROX dataset provides labels for and paired 3D scene scan of 12 indoor setting

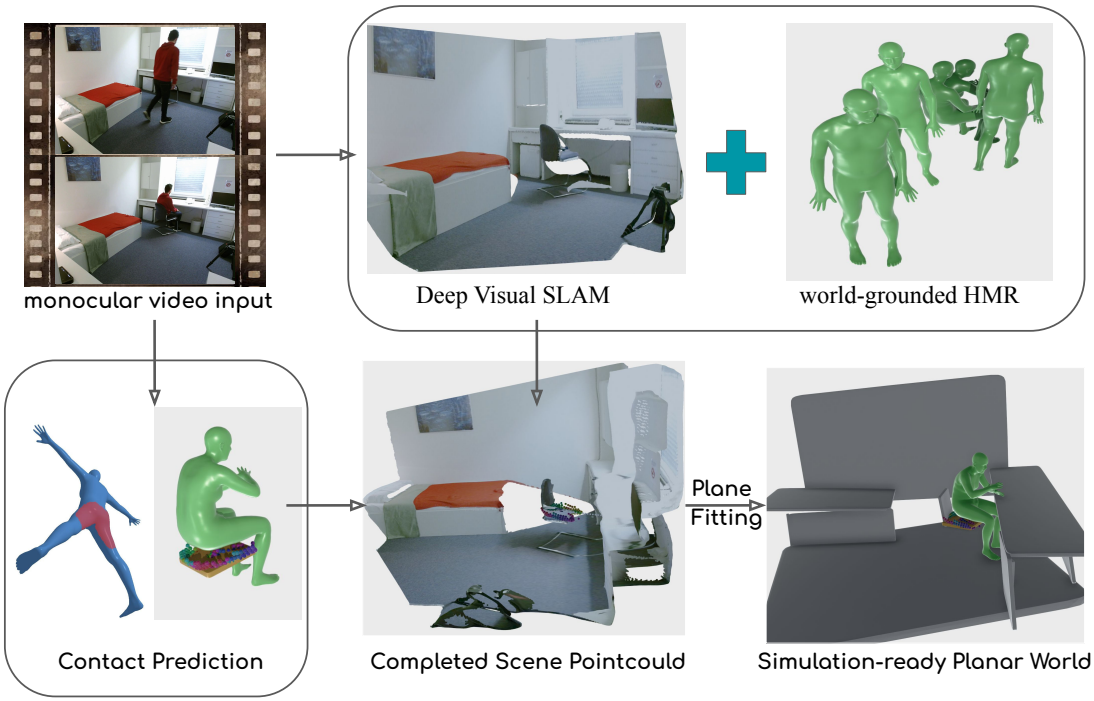


Figure 2. **Framework overview.** From a monocular video, we recover a common world frame by coupling deep visual SLAM with world-grounded HMR. Contact prediction provides human–scene anchors to complete occluded regions as a point cloud. We then fit a compact set of planar primitives to obtain a simulation-ready scene for RL training.

**Baselines.** We ablate three types of geometry representations and run best convexification strategy [19] that enables real2sim, and compare to prior work on world-grounded HMR. *Geometry.* (i) VDBFusion [16] (average weighted TSDF followed by Marching Cubes [9]; and (ii) NKSR. We omit JOSH [6] and VideoMimic [1] as code are not publicly available.

Table 1. **EMDB human motion estimation results.**

| Method | EMDB | | |
|---|---|---|---|
| | WA-MPJPE$_{100}$↓ | W-MPJPE$_{100}$↓ | RTE↓ |
| WHAM[15] | 98.4489 | 267.5326 | 3.299 |
| TRAM[18] | 83.6142 | 249.5018 | 1.927 |
| GVHMR[14] | **74.8000** | 200.700 | 1.900 |
| **Physis(Ours)** | **76.5195** | **178.0635** | **1.620** |

**Metrics.** We evaluate global human motion estimation following prior works [15, 18] to use the world-grounded human pose estimation metric W-MPJPE$_{100}$ or WA-MPJPE$_{100}$ for evaluation. We also evaluate the global trajectory errors normalized by the total distance after aligning the whole trajectory and measure Root Translation Error (RTE%). For RL training, we ablate several design choices regrading success rate. For preliminary result, we define success rate as whether the agent successfully follows the entire motion sequences reconstructed from monocular video. We also compare the throughput during RL training, measured in frames per second (FPS). This metric reflects the efficiency of different reconstruction backends when used in our RL loop.

Table 2. **Success rates (%) and overall FPS.**

| Method | EMDB (21) | PROX (12) | Average↑ | FPS↑ |
|---|---|---|---|---|
| VDBfusion | 90.40 | 81.82 | 87.50 | 8K |
| NKSR | 52.38 | 72.73 | 65.40 | 8K |
| **Physis (Ours)** | **100.0** | **91.67** | **96.97** | **15K** |

**Discussion and Conclusion.** Our proposed framework attains the highest success rates on both datasets—100% on EMDB-2 and 91.67% on PROX—yielding a 96.97% average, which is +9.47 percentage points over the strongest baseline (VDBfusion at 87.50%). Moreover, our planar, simulation-ready representation delivers 15K FPS training throughput which brought 2X speedup. The gains are consistent with our design: contact-guided completion plus compact plane primitives reduce collision pairs and contact resolution cost, improving both stability (higher success) and efficiency (higher FPS). The results highlight that explicitly leveraging predicted contact points to drive planar reconstruction is an effective route to reliable, simulation-ready real-to-sim from in-the-wild monocular video.

## References


[1] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025. 2, 4

[2] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 3

[3] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[4] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14632–14643, 2023. 3

[5] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. *arXiv preprint arXiv:2412.04463*, 2024. 3

[6] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *arXiv preprint arXiv:2501.02158*, 2025. 2, 4

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 2

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[9] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, page 163–169, New York, NY, USA, 1987. Association for Computing Machinery. 4

[10] 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 *CVPR*, 2019. 2

[11] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic. *ACM Transactions on Graphics*, 37(4):1–14, 2018. 3

[12] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 3

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. 3

[14] 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 2024 Conference Papers*, 2024. 2, 3, 4

[15] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4

[16] Ignacio Vizzo, Tiziano Guadagnino, Jens Behley, and Cyrill Stachniss. Vdbfusion: Flexible and efficient tsdf integration of range sensor data. *Sensors*, 22(3), 2022. 4

[17] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. 3

[18] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. *arXiv preprint arXiv:2403.17346*, 2024. 2, 4

[19] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (TOG)*, 41(4):1–18, 2022. 4

[20] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, et al. Ufm: A simple path towards unified dense correspondence with flow. *arXiv preprint arXiv:2506.09278*, 2025. 3