HAMSt3R: Human-Aware Multi-view Stereo 3D Reconstruction

Sara Rojas[1,2*] Matthieu Armando[2] Bernard Ghanem[1]
Philippe Weinzaepfel[2] Vincent Leroy[2] Grégory Rogez[2]

[1]KAUST [2]NAVER LABS Europe

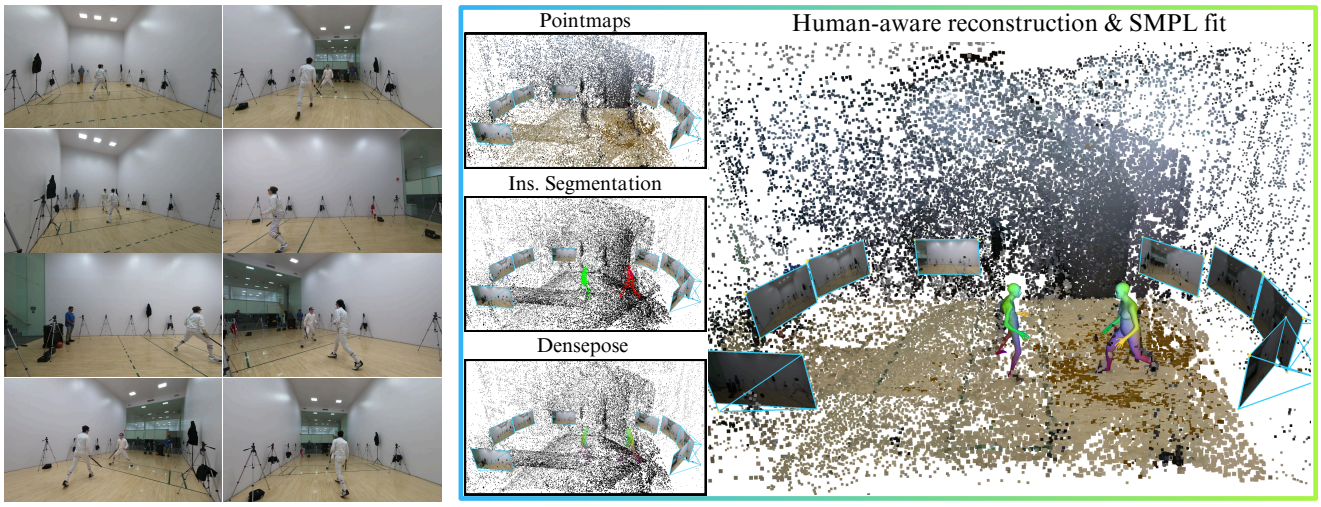

Figure 1. Given a set of unposed images, HAMSt3R reconstructs the 3D scene as a dense point map with human semantics, attaching instance segmentation and DensePose information directly to 3D points for human-aware modeling.

Abstract

Recovering the 3D geometry of a scene from a sparse set of uncalibrated images is a long-standing problem in computer vision. While recent learning-based approaches such as DUSt3R and MASt3R have demonstrated impressive results by directly predicting dense scene geometry, they are primarily trained on outdoor scenes with static environments and struggle to handle human-centric scenarios. In this work, we introduce HAMSt3R, an extension of MASt3R for joint human and scene 3D reconstruction from sparse, uncalibrated multi-view images. First, we exploit DUNE, a strong image encoder obtained by distilling, among others, the encoders from MASt3R and from a state-of-the-art Human Mesh Recovery (HMR) model, multi-HMR, for a better understanding of scene geometry and human bodies. Our method then incorporates additional network heads to segment people, estimate dense correspondences via DensePose, and predict depth in human-centric environments, enabling a more comprehensive 3D reconstruction. By leveraging the outputs of our different heads, HAMSt3R produces a dense point map enriched with human semantic information in 3D. Unlike existing methods that rely on complex optimization pipelines, our approach is fully feed-forward and efficient, making it suitable for real-world applications. We evaluate our model on EgoHumans and EgoExo4D, two challenging benchmarks containing diverse human-centric scenarios. Additionally, we validate its generalization to traditional multi-view stereo and multi-view pose regression tasks. Our results demonstrate that our method can reconstruct humans effectively while preserving strong performance in general 3D reconstruction tasks, bridging the gap between human and scene understanding in 3D vision.

1. Introduction

3D scene reconstruction from uncalibrated images is a fundamental problem in computer vision with a wide range of applications in robotics, augmented reality, and human-computer interaction. Traditionally, this task has long been approached by solving a succession of problems [53, 54]

*Work done while interning at NAVER LABS Europe.

using different algorithmic tools like image matching and bundle adjustments. However, recent learning-based methods such as DUSt3R [64] and MASt3R [33] introduced a new paradigm by directly regressing the 3D geometry of a scene, given a pair of images. These methods not only significantly improved reconstruction quality but also simplified the pipeline, inspiring numerous follow-up works [56, 63, 69], including some efforts focused on human-centric scene reconstruction [37, 41].

Despite these advances, estimating the geometry of scenes involving people remains a major challenge. Humans are highly articulated, exhibit complex deformations, and often appear in self-occluded poses, making their reconstruction significantly more difficult than static environments. While a complete 3D scene understanding should ideally facilitate Human Mesh Recovery (HMR)—the task of detecting and reconstructing people in 3D—humans themselves provide valuable cues for scene understanding, such as scale estimation. However, concurrent methods that jointly reconstruct humans and their surrounding environment [37, 41] rely on cumbersome optimization-based processes, limiting their scalability and practicality.

Furthermore, current learning-based reconstruction models such as MASt3R have been trained on buildings and outdoor scenes, with little focus on human subjects. As a result, they struggle when applied to images containing people, failing to capture articulated structures accurately and often producing incomplete or distorted reconstructions. Addressing this limitation requires integrating additional human-specific cues into the reconstruction pipeline.

In this work, we present HAMSt3R which extends MASt3R to explicitly handle human-centric scenes by jointly reconstructing both humans and their surrounding environments (see Figure 1). To achieve this, we first leverage DUNE [51], a strong image encoder which is pre-trained by distilling those from several teacher models, including MASt3R and Multi-HMR [6], a state-of-the-art HMR model, to help the network gain human understanding capabilities. We then introduce additional processing heads for instance segmentation, dense pose estimation, and binary mask generation. These components allow our model to distinguish human regions from the background, estimate dense correspondences based on the SMPL model [38] (*e.g.* DensePose predictions [26]), and integrate human-specific priors into the reconstruction process. By leveraging the outputs of our different heads, HAMSt3R produces a dense point map enriched with human semantic information in 3D. Specifically, each predicted 3D point is classified as human or non-human, with human points mapped to precise body locations of specific individuals. Predictions across multiple images pairs can be aggregated with global alignment, enabling dense, structured human semantics in 3D. By adapting a state-of-the-art stereo-based reconstruction pipeline to the complexities of human shape recovery, we offer an efficient and scalable alternative to existing optimization-based approaches. Our model effectively bridges the gap between general scene reconstruction and articulated human modeling, enabling high-fidelity 3D reconstructions from sparse and unstructured image collections. To train the new human-centric heads, we introduce a large-scale, multi-view, synthetic dataset of humans in indoor environments, created by combining the procedural scene generation of Infinigen [44] with *HumGen3D* [1] human generator.

Following [41], we evaluate our approach on two challenging human-centric benchmarks, namely EgoHumans [31] and EgoExo4D [34], which feature a variety of indoor and outdoor scenarios with one or several individuals across diverse environments. To ensure that our model maintains strong performance in traditional reconstruction setting - *i.e.*, scenes without humans- we also evaluate it for the task of multi-view stereo depth estimation across several benchmarks following [64]. Additionally, we assess its ability to perform multi-view pose regression on the CO3Dv2 [46] and RealEstate10K [72] datasets following [33]. Our thorough evaluation shows that our method remains robust across both human-centric and general reconstruction tasks.

The remaining of the paper is organized as follows: after reviewing the related work, we present our methodology, followed by a description of our experiments. Finally, we draw conclusions and discuss potential future directions for improving human-centric 3D scene reconstruction.

2. Related Work

We review past work on structure-from-motion, multi-view human reconstruction and both of them jointly.

Structure-from-Motion (SfM) [16, 17, 27] consists in reconstructing 3D scene geometry and camera poses given a set of images. The most popular approach is COLMAP [53, 54] that relies on traditional feature matching to perform incremental bundle adjustments. For many years, most work has focused on improving various parts of this pipeline such as keypoint detection and description [7, 19, 39, 48, 49], feature matching [22, 52, 60], initialization strategies [4, 58] or optimization techniques [36, 65]. Recently, there has been a significant paradigm shift towards fully-learnable approaches [10, 21, 57, 62, 64]. In particular, DUSt3R [64] has shown outstanding performance in unconstrained 3D reconstruction from as few as 2 images. Their core idea is to regress pointmaps for each image, expressed in the coordinate system of the first image. Several extensions have been since proposed including MASt3R [33] which also regresses pixel-aligned dense descriptors, Splatt3r [56] which outputs pixel-aligned parameters for 3D Gaussian splat-

ting [30], MONSt3R [69] which enables handling dynamic objects or MUSt3R [11] and CUT3R [63] which focus on improving efficiency when processing large image sets. In this paper, we build upon MASt3R to enable joint 3D reconstruction of humans and scenes from sparse uncalibrated views. While prior methods focus on rigid scene reconstruction, our approach explicitly incorporates human understanding while maintaining strong performance on structures, such as buildings and other man-made elements.

Multi-view Human Reconstruction has been extensively studied, particularly in controlled environments where camera parameters are known [23, 28, 29, 59]. In such settings, multi-view geometry can be leveraged for accurate 3D shape estimation, effectively transforming single- and multi-person reconstruction to a triangulation task [27]. When intrinsic and extrinsic camera parameters are available, single-view reconstruction techniques can also be extended to multi-view settings by enforcing geometric consistency. For instance, SMPLify [9] has been adapted to estimate 3D human body geometry in a shared coordinate frame, where accuracy is assessed by minimizing the 2D reprojection error of keypoints and silhouettes across multiple views, ensuring a geometrically coherent model [35]. Recent approaches have explored setups with unknown camera poses, employing end-to-end learning to jointly estimate camera parameters and 3D human poses [66, 68]. However, these methods are often limited to single-person scenarios [68] or lack scene context integration [66].

Joint Reconstruction of Scene and Humans has been studied in some very recent concurrent methods. JOSH [37] begins with an off-the-shelf scene reconstruction model [33, 64]; the resulting geometry offers valuable contact cues that guide human fitting. HSfM [41] instead assumes accurate 2D human keypoints across views to refine camera poses and, in turn, the surrounding scene. SynCHMR [71], following the SLAHMR [67] insight that human meshes can disambiguate SLAM, stitches together camera-frame HMR, monocular depth, and a human-aware SLAM pipeline before a global optimisation fuses the scene, cameras, and a single actor. All three pipelines therefore depend on pre-computed modules and iterative refinement to reconcile them. Our method removes these dependencies: in a single forward pass, we jointly predict metric 3D point-maps, dense human semantics, and camera parameters, providing a fully integrated and markedly more efficient solution.

3. Methodology

An overview of our model is shown in Figure 2. Building on the DUSt3R/MASt3R architecture, our approach takes two input images I_0 and I_1 which are encoded using a siamese ViT encoder. Each image is then processed by a separate ViT decoder where cross-attention is applied with the tokens from the other image. In addition to the 3D head of MASt3R (which predicts pointmaps and descriptor features for matching), our model is also trained to produce additional human-specific outputs, including human instance segmentation and DensePose predictions [26]. This helps the model gain a more meaningful understanding of the human geometry and its relation to the scene, adding rich semantic information to the reconstruction process. These outputs can be further leveraged for tasks like SMPL predictions through an optimization procedure. We first provide background on MASt3R (Section 3.1), then describe the image encoder adopted from [51] (Section 3.2) and the additional heads and training strategy (Section 3.3).

3.1. Background on MASt3R

Given an image pair $I_0, I_1 \in \mathbb{R}^{H\times W\times 3}$ (3 channels for RGB), MASt3R [33] jointly performs local 3D reconstruction and pixelwise matching. To achieve this, the network predicts a pixel-aligned pointmap $X \in \mathbb{R}^{H\times W\times 3}$ for each image, *i.e.*, containing the predicted 3D coordinates of the scene point corresponding to each pixel, expressed in the coordinate system of the first image's camera. A confidence map C is also produced. Additionally, another head predicts a small descriptor for each pixel, enabling efficient matching via approximate but fast nearest neighbor search.

For training, MASt3R uses a confidence-aware regression loss for pointmaps and a InfoNCE loss for the local descriptor learning. We denote its total loss as $\mathcal{L}_{\text{MASt3R}}$.

In terms of architecture, each image is processed by a ViT encoder Enc [20] to obtain a feature map $F = Enc(I) \in \mathbb{R}^{h\times w\times d}$ for an image I. A dual ViT decoder Dec, incorporating cross-attention blocks, then processes both feature maps while attending to the tokens from the other image: $F_0', F_1' = Dec\big(Enc(I_0), Enc(I_1)\big)$. Finally, the prediction heads operate on F_0' or F_1', producing pixel-wise outputs via either a linear head or a DPT head [45]. In this paper, we exclusively use linear heads.

3.2. A Strong Image Encoder

We replace the original MASt3R encoder with a stronger image encoder Enc: a distilled ViT-B/14 backbone obtained through the multi-teacher strategy of DUNE [51]. Distillation fuses complementary competencies from three powerful teachers—(1) a generalist image encoder (DINOv2 [42]), (2) the encoder of a state-of-the-art multi-person human-mesh-recovery model (Multi-HMR [6]), and (3) the MASt3R encoder itself [33]. Their representations are aligned by the UNIC projection mechanism [50], yielding visual features that are simultaneously robust for scenes and humans. Unlike DUNE, which attaches separate decoders per task and therefore reconstructs humans only from single images, our method couples this encoder with a unified, end-to-end architecture that jointly performs

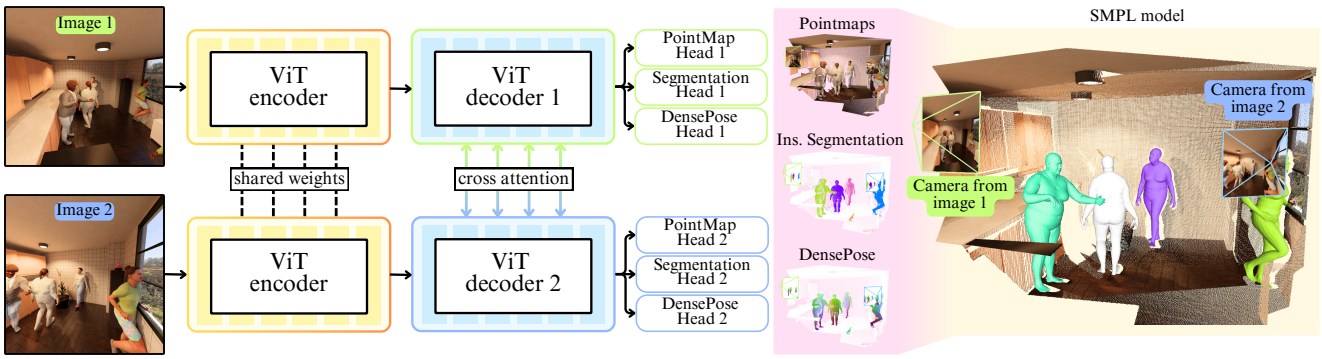

Figure 2. **Overview of HAMSt3R.** From left to right: (1) Input stereo images are processed via a Siamese ViT encoder, (2) extracted features are passed to dual decoders with cross-attention, (3) separate heads generate 3D pointmaps and dense human semantic information, in the form of instance segmentation, DensePose, and binary mask predictions. (4) These outputs can be lifted to 3D using the Pointmaps and can be used, for example, to fit a SMPL body model for each human.

3D scene reconstruction, instance segmentation, and cross-view human reconstruction.

3.3. HAMSt3R

Unlike MASt3R, which primarily focuses on objects and scenes (*e.g.* indoor environments or buildings), our method — as shown in Figure 2 — is trained on scenes containing humans. HAMSt3R takes two images, I_0 and I_1, as input and encodes them using a shared Vision Transformer (ViT) and a cross-attention decoder, generating global feature maps F_0 and F_1. These feature embeddings are then utilized by various heads. In addition to the original point and matching heads, we introduce an object segmentation head that produces segmentation masks for each individual in both images and a DensePose head that predicts DensePose maps for each person, mapping human pixels to the 3D surface of the human body, represented by the SMPL mesh [38]. The following paragraphs describe these additional heads in detail.

Instance Segmentation Head. An instance segmentation head is added to MASt3R, extending the original backbone with a transformer-based design inspired by Mask2Former [14]. This head is specifically designed to segment human instances from the background, generating masks that capture the full appearance of people, including hair and clothing. The segmentation branch is supervised by classification and mask losses following the strategy proposed in [13, 14], but with an extension to account for the two input views. In particular, the classification loss distinguishes between human and background, while the mask loss combines binary cross-entropy and dice loss. The key idea is that the model's understanding of 3D geometry allows it to assign consistent instance labels to each person across different viewpoints.

DensePose Head. The DensePose head is introduced as an additional branch to predict DensePose maps using SMPL projection templates. It consists of a linear layer that generates a four-channel prediction: an RGB dense pose map $P_{\text{dp}} \in \mathbb{R}^{H\times W\times 3}$, where each pixel is assigned an RGB color that encodes its corresponding 3D location on the SMPL template mesh, and a binary mask indicating the regions where the DensePose mapping is valid. The DensePose representation, when integrated with the 3D reconstruction, is expected to enhance the model's ability to reason about human poses, body parts, and their interaction with the environment, which adds semantics beyond just the raw geometric points. The predicted DensePose map is supervised by an L2 loss computed as $\mathcal{L}_{\text{dp}} = \|P_{\text{dp}} - P_{\text{gt}}\|_2^2$, where P_{gt} represents the ground truth DensePose map. Unlike the original DensePose approach that uses discrete body-part classes, our method employs a continuous mapping that directly formulates the problem as a 3D regression task, as done in [5].

In addition to the RGB DensePose map, the head produces a binary mask $M \in \mathbb{R}^{H\times W}$ that specifically distinguishes human regions corresponding to the SMPL projection — this excludes areas such as hair and clothing — from the background. This is in contrast to the instance segmentation mask, which covers the entire human silhouette, including hair and clothing. The binary mask is optimized using a cross entropy-loss $\mathcal{L}_{\text{mask}}$, which ensures that the SMPL model can be accurately fitted to the relevant human regions.

Semantic 3D Human Representation. By combining the outputs of our different heads, we obtain dense point maps with human semantic information in 3D. Specifically, every 3D point predicted by our method can be classified as either human or non-human, with human points mapped to specific locations on the body of the corresponding human instance. An advantage of our method is that we can readily fit the SMPL body model to every detected human, *e.g.* for numerical evaluation (see Section 4.1). This kind of seman-

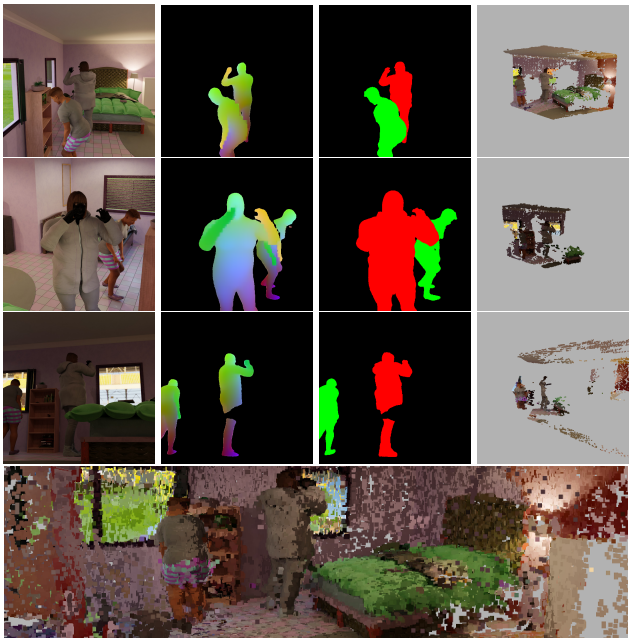

Figure 3. **Results on HumGen3D data** (on a scene not seen during training), using global alignment: Given a set of images of a scene (three out of eight of them are shown in the first column), we run our model on all possible image pairs, and aggregate predictions from the human heads in 2D, for each view (second and third columns). We apply the alignment method of MASt3R [33] to align the individual pointmaps in 3D (fourth column) and they can be combined into a unified reconstruction (bottom).

tic understanding is crucial for applications like tracking, behavior analysis, or interaction with the environment.

Dealing with more than 2 views. To handle scenes with an arbitrary number of images, we run our network independently on each possible image pair and align the output pointmaps in 3D using the procedure from MASt3R [33]. Since instance segmentation is performed independently for each pair, person IDs can vary across pairs—for instance, a person labeled as 'human 1' in one pair might be labeled as 'human 2' in another. To maintain consistent IDs across the scene, we resolve ID correspondences using 2D overlap before integrating all pairs into a consistent 3D representation. For DensePose, we aggregate the predictions for a same image (produced by each pair) by performing a weighted average of the DensePose outputs, using confidence scores as weights. This prioritizes higher-confidence predictions, resulting in a more accurate and stable human surface representation in 3D. Predictions from multiple image pairs are subsequently combined after a global alignment step, ensuring a coherent 3D representation of human semantics across the entire scene. This is illustrated in Figure 3. Note that in the binocular case, our approach is fully feed-forward.

Training. The DUNE image encoder is frozen to preserve the distilled features while the decoders and the new heads are fine-tuned. The overall training loss $\mathcal{L}$ is a weighted sum of the MASt3R loss $\mathcal{L}_{\text{MASt3R}}$, the segmentation loss $\mathcal{L}_{\text{seg}}$, the dense pose loss $\mathcal{L}_{\text{dp}}$, and the binary mask loss $\mathcal{L}_{\text{mask}}$:

$$\mathcal{L} = \mathcal{L}_{\text{MASt3R}} + \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{dp}} + \lambda_3 \mathcal{L}_{\text{mask}}, \qquad (1)$$

where λ_1, λ_2 and λ_3 are loss weights; details on their selection can be found in the Supp. Mat. Training is performed by mixing 50% of the original MASt3R dataset with 50% human-specific data in each epoch. For the original MASt3R dataset, supervision is exclusively applied to the point maps and the matching head. In contrast, the human-specific data is used to supervise all heads. This tailored supervision strategy ensures that each network component is optimally trained based on the available data. Training samples consist of image pairs from multi-camera setups or closely spaced frames from videos, ensuring diverse viewpoints while maintaining spatial coherence. All images are downscaled to a maximum dimension of 518 pixels.

Table 1. **Human-centric training datasets** used for finetuning on human-centric scenes alongside the original MASt3R training datasets. All datasets provide camera pose, depth, and instance segmentation. BEDLAM's 9.2k scenes are motion clips from 8 3D environments and 95 HDRIs. For EgoBody, we use an off-the-shelf segmentation tool [47] to obtain the instance segmentations.

Dataset	Domain	Type	Scenes
HumGen3D	Synthetic	Indoor	10k
BEDLAM [8]	Synthetic	Indoor & Outdoor	9.2k
HuMMan [12]	Real	Studio	339
EgoBody [70]	Real	Indoor	125

Training Datasets. Obtaining large-scale, multi-view data with accurate camera poses, depth maps, and parametric body annotations is particularly challenging for human-centric scenes. Real-world captures often require specialized equipment or extensive manual post-processing, making them both costly and error-prone. Consequently, we rely primarily on synthetic data where camera intrinsics, poses, and depth can be automatically recorded during rendering. We generate our own training data with the following pipeline: For each person in the data, we first sample a random body shape and pose from the AMASS dataset [40], then map a human model to it, from the *HumGen3D* [1] human generator plugin for Blender. These Humgen3D humans are then placed in detailed indoor 3D scenes that are procedurally generated with Infinigen Indoors [44]. Finally, these scenes are rendered with Blender, along with the necessary annotations (depth, instance masks, and DensePose). We generate 524k images, rendered from 10k scenes and 1000 unique 3D environments, with approximately 5 persons per scene. To increase diversity in environment types and the number of subjects, we also incorporate BEDLAM [8], which includes both indoor and outdoor settings,

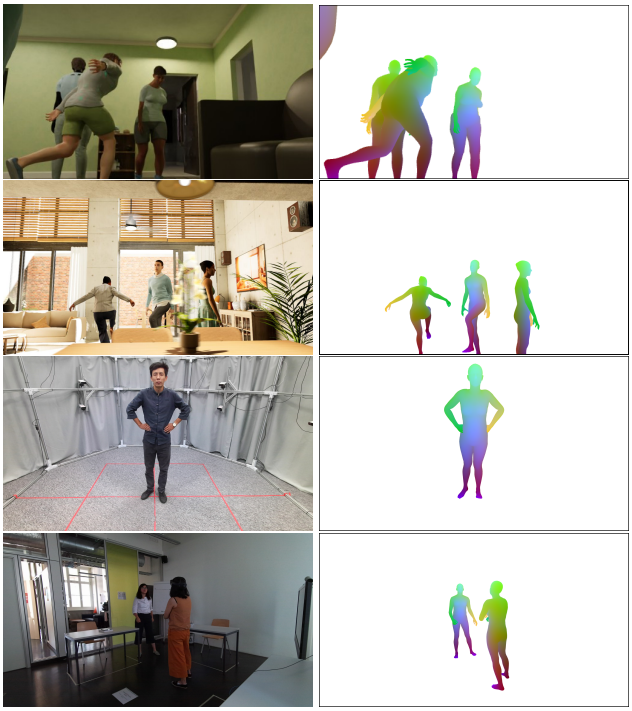

Figure 4. **Illustration of the human-centric datasets** used in this paper and listed in Table 1, namely HumGen3D, BEDLAM, HuMMan and EgoBody. For each dataset, we show an input image (left), along with its corresponding DensePose annotations (right).

and HuMMAN [12], which features single individuals performing complex poses. Lastly, we include EgoBody [70], a real dataset captured with multiple Kinect sensors that contains up to two individuals per scene and provides accurate depth maps. A summary of these datasets is provided in Table 1 and some examples are shown in Figure 4.

4. Experimental results

We evaluate our approach on both human-centric and traditional 3D tasks. We first present the evaluation protocol in Section 4.1 and then discuss results in Section 4.2.

4.1. Datasets and Metrics

Human-centric experiments. We evaluate the effectiveness of HAMSt3R across diverse indoor and outdoor environments, covering various activities involving one or multiple individuals. Following [41], we evaluate on EgoHumans [31] and EgoExo4D [34], and report different Mean Per-Joint Position Error (MPJPE) metrics (expressed in meters): W-MPJPE, when measured in the world coordinate system, PA-MPJPE, its Procrustes-Aligned version, and Group-Aligned MPJPE, (GA-MPJPE), after alignment between people. To obtain 3D joint predictions with our method, we fit SMPL to our predictions using an optimization procedure that minimizes the distance between all predicted 3D points of the person, and the corresponding vertices on the SMPL model. This is accompanied by an additional loss that serves as a prior on the pose and shape, using VPoser [43] for pose regularization. We build upon the MvSMPLFitting framework [15], which extends SMPLify-X [43] to multi-view settings. For evaluating camera poses, we also report the average camera translation error TE, and its Sim(3) aligned version (s-TE), the camera Angle Error (AE), the Relative Rotation Accuracy (RRA), the Camera Center Accuracy (CCA) and its version computed after Sim(3) alignment (s-CCA).

Traditional 3D tasks. To ensure that our model still performs well in classical 3D vision tasks, we also evaluate it for the task of multi-view stereo depth estimation on KITTI [24], ScanNet [18], ETH3D [55], DTU [3], Tanks and Temples [32], following [64]. We report the Absolute Relative Error (rel) and Inlier Ratio (τ) with a threshold of 1.03 on each test set and the averages across all test sets. Additionally, we assess the ability of HAMSt3R to perform multi-view pose estimation on the CO3Dv2 [46] and RealEstate10K [72] datasets following [33] and report the Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) for each image pair to evaluate the relative pose error and we select a threshold of 15° to report $RTA@15$ and $RRA@15$. Additionally, we calculate the mean Average Accuracy (mAA30).

4.2. Results

Human-centric experiments. We report human metrics in Table 2, comparing various human pose estimation baselines on EgoHumans and EgoExo4D. Specifically, we evaluate against UnCaliPose [66] and the concurrent work HSfM [41], as well as a monocular baseline Multi-HMR [6]. For the latter, we simply select a random view among the set, to be used as input. For HSfM, we also provide numerical evaluation before their optimization step (init). While both HSfM and UnCaliPose jointly reconstruct humans and cameras, HSfM is more comparable to our approach as it also reconstructs the environment and leverages DUSt3R to estimate the cameras. HAMSt3R outperforms the other baselines on EgoExo4D in World coordinate metric (W-MPJPE = 0.51 m), in particular HSfM (0.56 m) that uses a global scene optimization (and bundle adjustment guided by 2D human keypoint predictions) to optimize the humans, depth maps, and cameras. However, after Procrustes alignment, our performance (PA-MPJPE= 0.09m) is slightly below HSfM's performance (PA-MPJPE= 0.06m). The similar results of HSfM (init) for PA-MPJPE (0.07m) indicates that most of its human pose estimation accuracy comes from its strong initialization using the off-the-shelf HMR2 [25]. A possible reason for our slightly lower performance in PA-MPJPE is that we discard RGB information before fitting the SMPL model, making it challenging to

Table 2. **Human-centric evaluation metrics** on EgoHumans and EgoExo4D.

Method	EgoHumans			EgoExo4D	
	W-MPJPE ↓	GA-MPJPE ↓	PA-MPJPE ↓	W-MPJPE ↓	PA-MPJPE ↓
Multi-HMR [6]	7.66	0.99	0.12	2.88	0.07
UnCaliPose [66]	3.51	0.67	0.13	2.90	0.13
HSfM (init) [41]	4.28	0.51	0.06	5.29	0.07
HSfM [41]	**1.04**	**0.21**	**0.05**	0.56	**0.06**
HAMSt3R (Ours)	3.80	0.42	0.14	**0.51**	0.09

Table 3. **Camera pose evaluation** on EgoHumans and EgoExo4D.

Method	EgoHumans						EgoExo4D					
	TE ↓	s-TE ↓	AE ↓	RRA@10 ↑	CCA@10 ↑	s-CCA@10 ↑	TE ↓	s-TE ↓	AE ↓	RRA@10 ↑	CCA@10 ↑	s-CCA@10 ↑
UnCaliPose [66]	2.63	2.63	60.90	0.28	-	0.33	2.43	1.16	65.61	0.19	-	0.24
DUSt3R [64]	-	1.15	11.00	0.61	-	0.49	-	0.33	9.92	0.81	-	0.64
MASt3R [33]	4.97	0.92	10.42	0.61	0.06	0.65	0.96	0.35	11.70	0.79	0.06	0.68
HSfM (init) [41]	2.37	1.15	11.00	0.52	0.26	0.49	1.27	0.33	9.92	0.81	0.05	0.64
HSfM [41]	2.09	0.75	9.35	0.72	**0.32**	0.75	0.95	0.36	11.57	0.78	0.07	0.67
DUNE	**1.43**	**0.17**	**3.51**	**0.96**	0.27	**0.97**	1.03	0.26	6.45	0.94	0.25	0.85
HAMSt3R (Ours)	2.33	0.40	10.24	0.77	0.06	0.75	**0.60**	**0.15**	**2.85**	**0.99**	**0.42**	**0.87**

Table 4. **Camera pose evaluation on EgoHumans according to scene scale.** Average results over the entire dataset (All), compared with a finer analysis based on the split between large-scale/open scenes (Badminton, Tennis, Volleyball) and smaller environments (Basketball, Fencing, Lego, Tagging).

Env.	Method	TE ↓	s-TE ↓	AE ↓	RRA@10 ↑	CCA@10 ↑	s-CCA@10 ↑
All	DUNE	**1.43**	**0.17**	**3.51**	**0.96**	**0.27**	**0.97**
	HAMSt3R (Ours)	2.33	0.40	10.24	0.77	0.06	0.75
Large	DUNE	**0.98**	**0.24**	**4.78**	**0.90**	**0.34**	**0.93**
	HAMSt3R (Ours)	3.30	0.66	16.303	0.58	0.00	0.51
Small	DUNE	2.04	**0.11**	2.36	**1.00**	**0.14**	0.99
	HAMSt3R (Ours)	**1.31**	0.12	**2.32**	**1.00**	**0.14**	**1.00**

match neural HMR methods that leverage richer appearance cues. The same trend is observed on EgoHumans, where we outperform the state-of-the-art method UnCalibPose on W-MPJPE and GA-MPJPE but not on PA-MPJPE. Notably, on this dataset, the concurrent work HSfM achieves significantly better results, particularly on the PA-MPJPE metric (0.17m for us *vs*. 0.05m for HSfM), where their global optimization provides only marginal improvements over their strong initialization (0.06m). Our lower performance on EgoHumans compared to EgoExo4D is likely due to the nature of scenes, which often features large, open, outdoor environments. Our method to estimate SMPL parameters appears more sensitive to lower-resolution inputs as obtaining accurate SMPL fits becomes difficult when a person occupies only a small portion of the image.

The camera pose estimation metrics are reported in Table 3, where we evaluate our method against HSfM and UnCaliPose, as well as the original DUSt3R and MASt3R. To better analyze our performance, we also report camera metrics when estimating the cameras using DUNE. On EgoExo4D, HAMSt3R clearly outperforms all the other baselines for all the considered metrics. On the EgoHumans dataset, DUNE consistently achieves the best performance overall. In contrast, training our method seems to reduce

Table 5. **Multi-view pose regression evaluation** on the CO3Dv2 [46] and RealEstate10K [72] with 10 random frames.

Method	Co3Dv2 ↑			RealEstate10K ↑
	RRA@15	RTA@15	mAA(30)	mAA(30)
DUSt3R [64]	93.3	88.4	77.2	61.2
MASt3R [33]	**94.6**	**91.9**	**81.8**	76.4
DUNE	92.2	90.7	78.7	**80.1**
HAMSt3R (Ours)	90.7	90.2	76.3	77.4

the accuracy of camera estimates. We hypothesize that this is due to the large scale of some environments in the EgoHumans dataset, combined with our use of MASt3R log-scaling of the 3D regression loss. This non-linear scaling does not penalize distant points as heavily, which may negatively affect performance on larger scenes. This hypothesis is further supported by separate evaluations of performance on large and small scenes, where our method performs significantly better on smaller environments; See experiments in Table 4.

Traditional 3D tasks. Table 6 presents our multi-view stereo depth estimation results. As expected, overall performance decreases compared to the original DUSt3R and MASt3R models. This is due to two factors: (1) we use DUNE's encoder that distills MASt3R alongside an HMR model, enhancing human understanding but reducing depth accuracy for structures (see the drop in performance between MASt3R and DUNE) and buildings, and (2) we introduce additional human-centric heads and tasks during training while using 50% human-centric data, further impacting depth estimation in non-human regions. Nevertheless, HAMSt3R remains competitive, performing on par with or even surpassing recent deep learning architectures such as Deepv2D [61]. A similar trend is observed in the multi-view pose regression results reported in Table 5, where performance drops on CO3Dv2 but unexpectedly improves

Table 6. **Multi-view depth evaluation** (ScanNet) denote training on data from the same domain.

Method	KITTI		ScanNet		ETH3D		DTU		T&T		Average	
	rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑	rel. ↓	τ ↑
DeepV2D (ScanNet) [61]	10.00	36.20	**4.40**	54.80	11.80	29.30	7.70	33.00	8.90	46.40	8.60	39.90
DUSt3R [64]	5.88	47.67	**3.01**	**72.54**	3.04	75.17	2.92	73.94	2.93	78.51	3.56	69.56
MASt3R [33]	**3.54**	**65.68**	4.17	65.22	**2.44**	**82.77**	3.46	66.89	**2.04**	**87.88**	**3.13**	**73.69**
DUNE [51]	4.88	50.76	4.24	59.68	2.48	77.97	**2.69**	**75.63**	2.60	79.19	3.38	68.65
HAMSt3R (Ours)	5.60	45.66	4.43	56.50	2.96	71.68	5.31	57.62	3.01	73.53	4.26	61.00

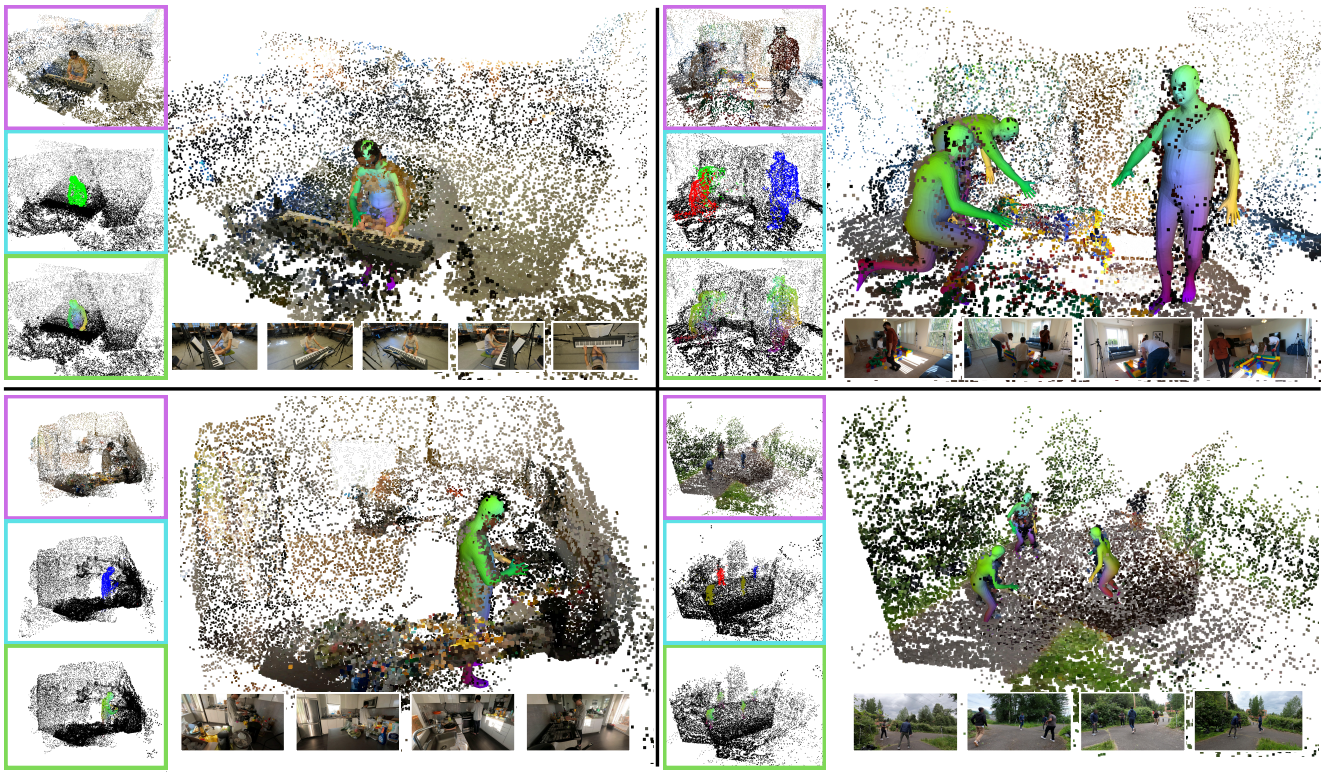

Figure 5. **Qualitative results of HAMSt3R.** Results from EgoExo4D (left) and EgoHumans (right). Each example includes point clouds (pink), instance segmentation (blue), and dense pose (green), with corresponding input images below (Best viewed when zoomed in).

on RealEstate10K. The distillation process is beneficial on this benchmark (see DUNE's performance) and the performance drop is less important when training HAMSt3R.

Qualitative results are presented in Figure 5, illustrating human reconstructions from EgoExo4D and EgoHumans. Our method successfully estimates dense human poses, instance segmentation, and point clouds across a variety of indoor and outdoor settings. The reconstructions capture realistic human shapes and spatial configurations, demonstrating the robustness of our approach even in challenging scenes. While some minor artifacts are visible in complex environments, our results align well with the quantitative findings, reinforcing our strong performance in world-coordinate metrics.

5. Conclusion

We have introduced HAMSt3R, the first feed-forward method for jointly reconstructing people and their surroundings from sparse stereo views. Given multiple images of a scene involving one or several persons, it produces dense point maps with human semantic information in 3D. Unlike optimization-based approaches, our method avoids common drawbacks such as computational slowness and sensitivity to hyperparameters. Through extensive evaluation, we demonstrate that our approach, combined with SMPL fitting, outperforms prior methods in estimating human poses and achieves competitive results to concurrent work in small environments, all while maintaining strong performance in general scene reconstruction—even in the absence of people. For future work, we aim to extend our method to videos and dynamic scenes, further enhancing its applicability to real-world scenarios.

Acknowledgements. The research reported in this publication was partially supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940.

References

[1] Humgen3d. https://www.humgen3d.com/. 2, 5

[2] Pexels. https://www.pexels.com/. Accessed: 2025-07-29. 12

[3] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016. 6

[4] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *ECCV*, 2010. 2

[5] Matthieu Armando, Salma Galaaoui, Fabien Baradel, Thomas Lucas, Vincent Leroy, Romain Brégier, Philippe Weinzaepfel, and Grégory Rogez. Cross-view and cross-pose completion for 3d human understanding. In *CVPR*, 2024. 4

[6] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 2, 3, 6, 7

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 2

[8] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 5

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 3

[10] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2

[11] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *CVPR*, 2025. 3

[12] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *ECCV*, 2022. 5, 6

[13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 4

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 4

[15] V. Choutas. Mvsmplfitting. https://github.com/vchoutas/smplify-x. 6

[16] David J Crandall, Andrew Owens, Noah Snavely, and Daniel P Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. *IEEE trans. PAMI*, 2012. 2

[17] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In *CVPR*, 2017. 2

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 6

[19] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised Interest Point Detection and Description. In *CVPR*, 2018. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[21] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 2

[22] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, 2024. 2

[23] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. PAMI*, 2010. 3

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 6

[25] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 6

[26] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 3

[27] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 3

[28] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *ICCV*, 2023. 3

[29] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, 2022. 3

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 3

[31] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *ICCV*, 2023. 2, 6

[32] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graphics*, 2017. 6

[33] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 2, 3, 5, 6, 7, 8

[34] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021. 2, 6

[35] Zhongguo Li, Magnus Oskarsson, and Anders Heyden. 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In *WACV*, 2021. 3

[36] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 2

[37] 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, 3

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graphics*, 2015. 2, 4

[39] G Lowe. Sift-the scale invariant feature transform. *IJCV*, 2004. 2

[40] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 5

[41] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *CVPR*, 2025. 2, 3, 6, 7, 12

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 3

[43] 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. 6

[44] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *CVPR*, 2024. 2, 5

[45] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 3

[46] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 2, 6, 7

[47] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 5, 12

[48] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *NeurIPS*, 2019. 2

[49] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011. 2

[50] Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Unic: Universal classification models via multi-teacher distillation. In *ECCV*, 2024. 3

[51] Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Pau de Jorge, Diane Larlus, and Yannis Kalantidis. Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In *CVPR*, 2025. 2, 3, 8

[52] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[53] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2

[54] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 1, 2

[55] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 6

[56] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 2

[57] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. In *3DV*, 2025. 2

[58] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH*, 2006. 2

[59] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 2007. 3

[60] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 2

[61] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 7, 8

[62] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *CVPR*, 2024. 2

[63] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2, 3

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 3, 6, 7, 8, 12

[65] Yuxi Xiao, Nan Xue, Tianfu Wu, and Gui-Song Xia. Level-s^2 fm: Structure from motion on neural level set of implicit surfaces. In *CVPR*, 2023. 2

[66] Yan Xu and Kris Kitani. Multi-view multi-person 3d pose estimation with uncalibrated camera networks. In *BMVC*, 2022. 3, 6, 7

[67] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 3

[68] Zhixuan Yu, Linguang Zhang, Yuanlu Xu, Chengcheng Tang, Luan Tran, Cem Keskin, and Hyun Soo Park. Multi-view human body reconstruction from uncalibrated cameras. In *NeurIPS*, 2022. 3

[69] 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. In *ICLR*, 2025. 2, 3

[70] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 5, 6

[71] Yizhou Zhao, Tuanfeng Y. Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao Paul Huang. Synergistic global-space camera and human reconstruction from videos. *CVPR*, 2024. 3

[72] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 2, 6, 7