BodyContact4D: A Multi-view Video Dataset for Understanding Human and Environment Interactions

Soyong Shin[1,2] Chaeeun Lee[1] Holly Chen[1] Jyun-Ting Song[1]
Eni Halilaj[1] Kris Kitani[1,2]

[1]Carnegie Mellon University [2]Meta

{soyongs, chaeeunl, hollyc, jyuntins, ehalilaj, kmkitani}@andrew.cmu.edu

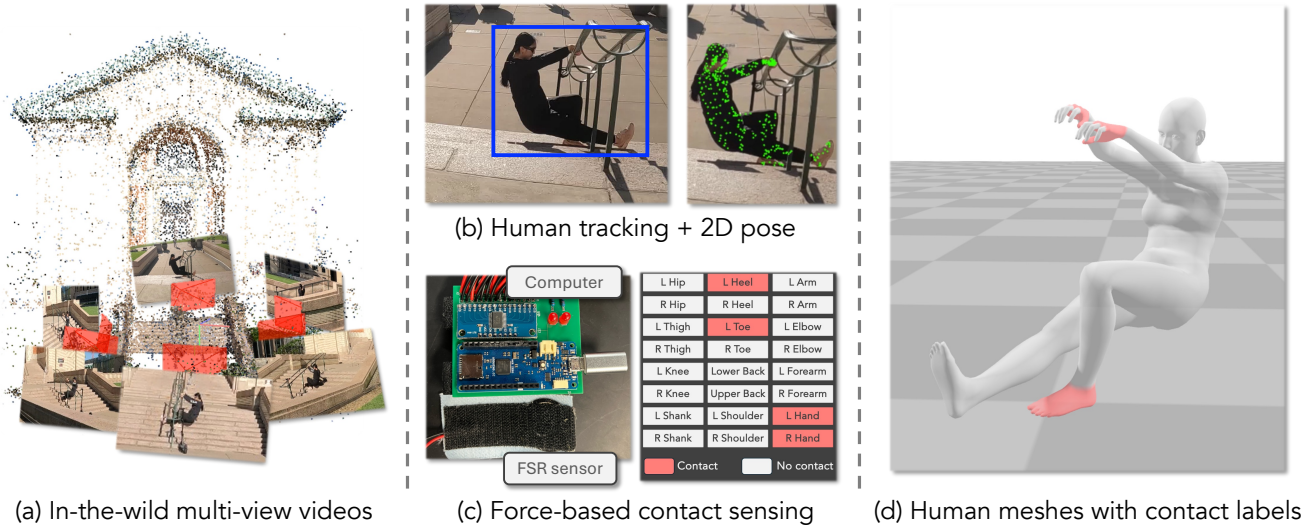

Figure 1. BodyContact4D Overview. (a) We capture synchronized multi-view videos in diverse in-the-wild environments. (b, c) We track human motion using dense surface keypoints and measure environmental contact through a customized force-sensing system. (d) As a result, accurate 3D human meshes with contact labels, highlighted in red, are obtained.

Abstract

To improve vision-based methods for understanding how people interact with their physical environment, we introduce a multi-view video and body-contact sensing dataset designed to capture dynamic human activities that involve interactions with the physical environment. The dataset includes activities such as parkour, physical training, and gym exercises, characterized by frequent body-environment contact. The proposed dataset includes 780K images across 120K pose sequences from 7 subjects. Each subject is captured by 6 synchronized third-person cameras, a single egocentric camera, and multiple contact sensors worn on the body. Using our proposed dataset, we benchmark state-of-the-art vision-based body contact models and show significant limitations in existing methods. Furthermore, we benchmark existing human pose estimation methods on our dataset and show that they fail under significant occlusion caused by close interactions with the environment, which indicates that our dataset can also be used to further develop pose estimation models to be more robust during interaction with the environment. To facilitate better human pose estimation from video, we introduce and evaluate a video-based human contact detection model that outperforms existing image-based methods, underscoring the potential improvements from integrating contact information into pose estimation models. See the project page at: `https://yohanshin.github.io/bodycontact4d.github.io/`

1. Introduction

Consider observing a person climbing a set of stairs. In order to fully understand the person's pose/motion (kinematics) and torques/forces (dynamics) being applied to the body,

Dataset	Type	# Scene	# Subject	# Images	# Cameras	3D Mesh	Contact Source
PROX [11]	Video	12	20	100K	1	Yes	Distance
RICH [12]	Video	5	22	577K	6-8	Yes	Distance
DAMON [33]	Image	–	–	5K	–	No	Human Annotation
3DIR [41]	Image	–	–	5K	–	No	Human Annotation
BodyContact4D (Ours)	Video	24	7	780K	6-7	Yes	Force Sensor

Table 1. Comparison of BodyContact4D with existing human-scene interaction datasets. Our dataset provides the largest number of image frames and scenes among video-based datasets, and uniquely offers accurate, scalable contact annotations through a customized force-sensing system rather than relying on distance-based heuristics or human annotations.

we must understand how the body interacts with the physical environment. When we consider the development of computer vision algorithms that can truly understand human motion, it is essential that those methods have the ability to detect physical contact between a person's body and their environment.

To enable better reasoning about human-environment interaction, recent methods for human pose estimation have incorporated human contact reasoning by providing datasets with contact annotations (*e.g.,* foot is in contact with the ground). Some datasets use human annotators to provide contact labels by analyzing static images [33, 41]. This process can be error-prone as physical contact is not always easy to observe visually. For example, hands can be very close to an object without making any contact immediately prior to grasping the object. Additionally, occlusion can make it difficult to visually detect when contact actually occurs. Other methods attempt to automate contact annotations by pre-scanning static scenes and computing distances between a reconstructed 3D human mesh and scene geometry [11, 12]. Pre-scanning an environment can be restrictive as it may be difficult to capture a diverse set of natural environments, and it can be especially challenging if the environment has articulated objects that move (e.g., gym equipment).

To overcome these limitations, we introduce *BodyContact4D*, a multi-modal dataset designed to capture dynamic human-environment interactions. BodyContact4D combines multi-view video recordings with full-body contact sensing to provide accurate annotations of human-scene interactions. The overview of our data collection setup and the processing framework are shown in Figure 1. The dataset features humans naturally interacting with surrounding structures across diverse and realistic environments. Inspired by previous datasets [16, 17, 19], we capture synchronized multi-view video data using multiple third-person cameras combined with an egocentric view obtained using Aria glasses [2]. To enable data collection in diverse real-world locations, we use a compact setup consisting of six third-person cameras, maintaining comprehensive multi-perspective coverage while enhancing portability and versatility. In addition to the full-body contact annotations obtained through force sensors, BodyContact4D provides 3D mesh reconstructions of subjects performing dynamic interactive motions in challenging settings. Overall, BodyContact4D includes over 780K synchronized multi-view images of 7 subjects performing diverse movements across 24 different real-world locations. As summarized in Table 1, BodyContact4D significantly expands dataset scale and uniquely supports dynamic scenes through customized contact sensing.

To ensure accurate 3D human body mesh recovery under challenging scenarios involving severe occlusions from human-environment contacts, we implement a multi-stage optimization pipeline utilizing a dense human landmark detection network trained on synthetic data [3, 25]. To obtain ground-truth human-scene contacts, we design a customized full-body contact sensing system. This system employs multiple force sensing resistors (FSR) [42] integrated into an inner suit, allowing subjects to wear their regular daily outfits over it, thus minimizing visual intrusiveness. Given the numerous potential contact points on the human body, our system is specifically engineered for easy attachment and detachment of sensors. For each data collection session, we pre-assess which body regions are most likely to contact the environment and accordingly adjust the sensor configuration. We measure full-body contact and label the contacts according to 24 distinct body segments, ensuring detailed and structured annotations of human-scene interactions.

We benchmark existing 3D human pose and shape estimation models on the BodyContact4D test set. Due to the presence of challenging poses resulting from dynamic interactions with the environment and frequent occlusions, existing models exhibit limited accuracy compared to widely-used 3D human benchmark datasets [13, 15, 36]. Additionally, to test the utility of our dataset, we train a video-based human contact detection network built upon DECO [33], the state-of-the-art per-frame contact detection model. Experimental results demonstrate that our proposed approach achieves superior accuracy in human-scene contact detection.

Our contributions are summarized as follows. (1) We introduce BodyContact4D, a multi-modal dataset capturing

dynamic human-environment interactions. Our integration of a novel full-body contact sensing system with robust multi-view body fitting provides comprehensive 3D motion data and detailed contact annotations in diverse real-world environments. (2) We collected human motion data involving rich interactions with structural environments, which naturally produce challenging poses and severe occlusions rarely represented in existing datasets. Our evaluation of existing 3D human pose estimation models on our data provides clear evidence of these challenges and highlights the dataset's value as a benchmark. (3) We provide a baseline for video-based human contact detection by fine-tuning an existing image-based model on our dataset, demonstrating that it outperforms the existing image-based approaches.

2. Related Work

2.1. 3D Human Pose Datasets

Early research in human pose estimation primarily utilized large-scale 2D datasets such as MPII [1] and COCO [20], advancing our understanding of human pose representations. The initial 3D datasets relied on marker-based motion capture systems to obtain accurate ground-truth poses [13, 32, 34], but these methods were limited by visually intrusive markers and constrained laboratory settings. Later methods employed multi-view camera setups to triangulate 3D joint positions from 2D keypoints detected by neural networks pretrained on 2D datasets [4], thereby avoiding visual intrusion but still remaining restricted to controlled studio environments [14, 23]. Recent approaches have targeted capturing poses in more natural settings by minimal sensors such as handheld cameras combined with wearable sensors [15, 36]. However, accurately obtaining 3D poses from sparse setups remains challenging, these datasets captured human motion without severe occlusions. To address these limitations, BodyContact4D employs a portable multi-view camera system, enabling diverse indoor and outdoor data collection. We specifically capture sequences involving significant human-environment interactions, providing accurate 3D human motions paired with videos featuring substantial occlusions and challenging poses.

2.2. Human Interaction Datasets

While the aforementioned datasets primarily focus on capturing human poses, recent efforts have increasingly aimed to capture explicit interaction signals. For example, a body of work [8, 14, 17, 24, 38, 43] has focused on capturing human–human social interactions, emphasizing scenarios with close proximity, dynamic motions, and complex inter-person occlusions. Beyond human–human interactions, another significant research direction explores interactions between humans and surrounding objects. One active subset within this area emphasizes fine-grained hand-object interactions and dexterous manipulations [5, 7, 10]. In parallel, another line of research broadly focuses on interactions between humans and structural environments, capturing explicit contact signals beyond fine-grained hand-object manipulations. Recent datasets such as DAMON [33] and 3DIR [41] provide detailed annotations of human contact points, including interactions with handheld objects as well as structural scenes, relying primarily on human annotations and thus resulting in limited dataset sizes. Other datasets like PROX [11] and RICH [12] instead automate contact annotation by pre-scanning static scenes and computing human-scene interactions using distance-based heuristics between the human mesh and the scanned environment geometry. However, such an approach inherently limits the diversity of captured scenes due to the extensive time required for scene scanning. It also restricts applicability to static environments, potentially diminishing the realism of captured interactions. MMVP [44] partly addresses this by incorporating pressure insoles with videos, but the dataset is limited to annotation of only foot contact. In contrast, we measure full-body human-scene contact using the customized wearable force sensors rather than relying on heuristic distance computations. This allows us to capture true physical interactions—contact events characterized by actual pressure exerted between humans and their environments, in diverse and dynamic real-world scenarios.

3. Method

3.1. BodyContact4D Dataset

Data Collection. Our objective is to capture dynamic human–scene interactions naturally occurring in diverse real-world environments. Following recent multimodal human motion datasets such as Harmony4D [17], EgoHumans [16], and EgoExo [19], we employ synchronized multi-view GoPro cameras with 4K resolution (3840×2160) and 30 frames per second (fps) to comprehensively record these activities from multiple third-person perspectives. However, instead of using a large-scale setup (e.g., 20 cameras in Harmony4D), we simplify our system to include only 6 cameras. This streamlined arrangement enhances portability, enabling efficient data collection across varied in-the-wild locations. Additionally, similar to EgoHumans and EgoExo, we incorporate egocentric views captured with Meta's Aria glasses [2], offering complementary first-person perspectives. Figure 2 illustrates the dynamic human–environment interactions in the diverse in-the-wild environments. In total, our dataset currently consists of approximately 800K synchronized images from 7 subjects, captured across 24 unique and diverse real-world scenes involving a broad range of interactive human motions.

Multi-view Camera Calibration. For each data collection session, we first pre-scan the environment using one external camera and Meta's Aria glasses [2]. Using these recorded

Figure 2. BodyContact4D Dataset Samples. The dataset includes a wide range of indoor and outdoor scenes, diverse dynamic interactions, and their corresponding reconstructed 3D human meshes with contact labels.

video streams, we employ COLMAP [29], a widely-used Structure-from-Motion (SfM) framework, to estimate intrinsic parameters, extrinsic parameters, and lens distortion coefficients for all cameras. However, since COLMAP's reconstruction is scale-ambiguous, we leverage the metric-scale trajectory provided by the Aria glasses' onboard tracking system. Specifically, we perform Procrustes analysis [22] between the scale-ambiguous Aria camera trajectory from COLMAP and the metric-aware trajectory from the Aria tracking system, thereby transforming our calibrated camera system into a consistent metric-scale and gravity-aligned coordinate system.

3.1.1 3D Human Motion Reconstruction

In our dataset, accurately reconstructing 3D human motion is challenging due to the relatively sparse camera setup and significant occlusions arising from human-environment interactions. To begin addressing these challenges, we first track human bounding boxes across video sequences, even under partial or complete occlusions, using the recently proposed SAMURAI [40]. Subsequently, we integrate dense human keypoint detection with an optimization-based fitting algorithm, enabling the recovery of accurate 3D human body meshes with minimal manual intervention.

Dense Keypoints Detection. Accurately reconstructing detailed 3D human pose and shape benefits significantly from densely distributed surface keypoints, rather than sparse keypoints defined at joint centers. Following prior work [25], which introduced dense keypoint detection models trained on synthetic data [3], we adopt a similar approach with several notable differences. Specifically, we employ an even denser keypoint configuration to enhance fine-grained pose estimation. Additionally, unlike the Transformer decoder and uncertainty-based regression used in CameraHMR, we utilize a conventional heatmap-based keypoint detection method. Empirically, we observe that the heatmap-based method yields more accurate pose estimation on our collected data. While dense surface keypoints provide detailed visual cues for pose estimation, the model relies exclusively on synthetic training data due to the lack of densely annotated real-world datasets, occasionally limiting its performance in challenging scenarios involving occlusions or complex interactions. To address this limitation, we additionally incorporate ViTPose [39], which is trained extensively on large-scale real-world datasets, to estimate 17 conventional sparse keypoints. As a result, at each camera $c \in (1, ..., C)$ and time frame $t \in (1, ..., T)$, we obtain the set of 2D body keypoints and confidence $x_{c,t}^{2D} \in \mathbf{R}^{K\times 3}$ where K is the number of keypoints ($K = 454$).

3D mesh optimization. Given the detected 2D keypoints $\mathbf{x}^{2D} = \{x_{c,t}^{2D}\}$ and triangulated 3D keypoints $\mathbf{x}^{3D} = \{x_t^{3D}\}$, we employ a parametric body model to reconstruct 3D human body meshes. The goal of this stage is to find the optimal set of parameters Θ_t at each time t, where $\Theta_t = \{\theta_t, \tau_t, \beta\}$. The pose parameter $\theta_t \in \mathbf{R}^{21\times 3}$ represents the 3D joint rotations of 21 body joints, and $\tau_t \in \mathbf{R}^6$ denotes the global orientation and translation of the body with respect to the global coordinate system. The shape parameter $\beta \in \mathbf{R}^{11}$ captures the 11 principal directions of human shape variability derived from PCA. Although we do not fit hand poses or facial expressions, we utilize the SMPL-X model [27] to leverage the learned body pose prior, VPoser, defined within the SMPL-X body pose parameter space. Using VPoser, the pose parameter can be expressed as $\theta_t = V(z_t)$, where V is the decoder and $z_t \in \mathbf{R}^{32}$ is the latent vector of human body poses.

Our optimization framework consists of multiple sequential stages. First, we initialize the global translation and orientation parameter τ_t by computing the transformation between unposed template keypoints and the triangulated 3D target keypoints, specifically using shoulder and hip keypoints. Next, we jointly optimize pose, shape, and global

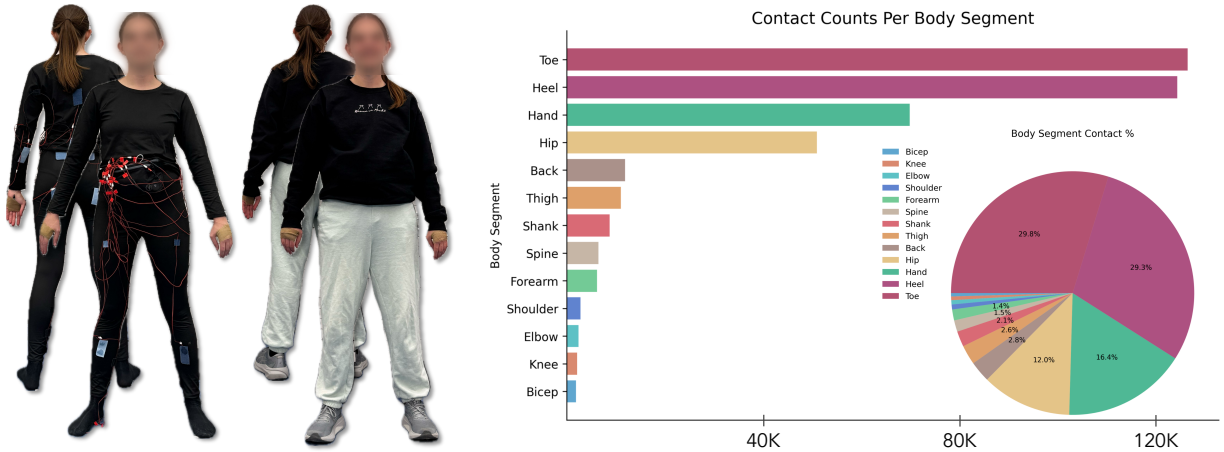

Figure 3. Contact Sensing System and Data Distribution. (a) Visualization of the contact sensing setup, illustrating subjects wearing the sensor-integrated inner suit and regular outfits, highlighting its minimal visual intrusiveness. (b) Distribution of contact annotations, showing the number of frames in which contact occurs for each body segment.

parameters using a short initial segment from each sequence, where subjects maintain relatively simple poses (*e.g.*, T or A poses). After obtaining robust shape estimates from this initial stage, we fix the β parameters and continue optimizing only z_t and τ_t. Our objective function is defined as

$$\mathcal{L} = \lambda_{3D}\mathcal{L}_{3D}+\lambda_{2D}\mathcal{L}_{2D}+\lambda_z\mathcal{L}_z+\lambda_\beta\mathcal{L}_\beta+\lambda_{smooth}\mathcal{L}_{smooth},$$

where λ denotes the weights on each loss term. Here, $\mathcal{L}_{2D}$ and $\mathcal{L}_{3D}$ are 2D keypoints reprojection and 3D keypoints losses, respectively, $\mathcal{L}_z$ and $\mathcal{L}_\beta$ denote the L2 regularization of the latent pose vector and shape parameter, and $\mathcal{L}_{smooth}$ represents the temporal smoothness term. For a more detailed description of each loss term, please refer to the supplementary materials.

3.1.2 Full-body Contact Sensing

To accurately measure full-body contact between humans and their surrounding environments, we developed a wearable sensing system based on FSRs. An FSR is a thin and flexible sensor that measures pressure based on a decrease in electrical resistance as force is applied perpendicularly to its surface [42]. Our system is specifically designed to be visually non-intrusive, by placing the sensors to an inner suit over which subjects wear their regular clothing, thus avoiding the introduction of distinct sensor-related visual cues into the dataset (see Figure 3 for illustration). Given the numerous potential contact points on the human body, we rehearsed each sequence to identify and mark the points anticipated to experience contact. The sensors were then precisely positioned at these locations. To facilitate convenient and quick repositioning, velcro strips were used to attach the sensors. Continuous pressure measurements were recorded at each sensor and converted into binary contact masks by applying a heuristic threshold. This approach reduces false-positive detections arising from pressure exerted by the subject's clothing. For hand locations, we used elastic bandages to securely position and insulate the sensors, ensuring accurate and stable contact measurements. At the beginning and end of each sequence, the subjects performed hand clapping to temporally synchronize the FSR and camera data. While our sensing system is capable of measuring force over 80 fps, we downsampled the data to 30 fps to align it with the video frames.

3.2. Video-based Contact Estimation

To demonstrate the utility of our dataset, we introduce and train the first video-based human contact detection model named *DECO-VID*, which builds upon DECO [33], the current state-of-the-art image-based contact estimation method. Figure 4 illustrates the architecture of our proposed video-based human contact detection model. Given a sequence of images within a temporal window of size W, $\mathbf{I} = \{I_1, ..., I_W\}$, we first extract context tokens f_t^{DECO} from each frame using DECO:

$$\mathbf{f}_t^{DECO} = E(I_t), \quad t = 1, ..., W$$

where E is the pretrained DECO feature extractor. We then expand these tokens to a combined representation $f_t = \{f_t^{DECO}; \hat{\theta}_t\}$ by incorporating 3D human pose estimation, where $\hat{\theta}_t$ denotes the SMPL pose parameters predicted by CameraHMR [25]. The combined features are temporally encoded using Transformer encoders [35] to effectively model temporal dependencies across frames. Subsequently,

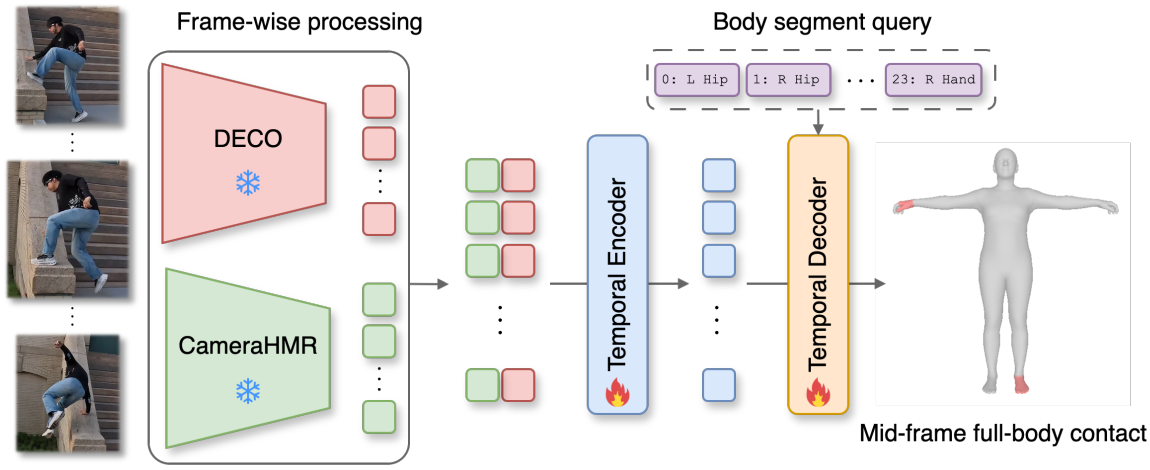

Figure 4. Network Architecture of DECO-VID. Context tokens extracted by DECO and 3D pose predictions from CameraHMR are temporally encoded by Transformer encoders and subsequently decoded by Transformer decoders using segment-specific queries to estimate human–scene contact probabilities.

temporally encoded features $\tilde{\mathbf{F}}$ are processed by Transformer decoders, where each targeted body segment acts as a learnable query. This decoding step produces segment-specific contact probabilities for the middle frame in the sequence:

$$\mathbf{P} = \text{TransformerDecoder}(\mathbf{Q}, \tilde{\mathbf{F}})$$

$$\mathbf{P} = \{p_1, ..., p_S\}, \quad \mathbf{Q} = \{q_1, ..., q_S\}$$

where S is the number of the candidate body segments. Throughout training, we freeze the weights of both the DECO feature extractor and CameraHMR to prevent overfitting to visual patterns specific to our dataset. This ensures the generalizability of learned representations. We train the model on BodyContact4D train set and evaluate on test set.

4. Experiments

4.1. 3D Human Mesh Recovery Benchmark

We evaluate several state-of-the-art 3D human pose and shape estimation models on the BodyContact4D `test set`. Specifically, we include per-frame methods such as BEDLAM-CLIFF [3], HMR2.0 [9], TokenHMR [6], NLF [28], and CameraHMR [25], as well as video-based approaches including WHAM [31] and TRAM [37]. During evaluation, we provide ground-truth bounding boxes for all methods. For models conditioned on camera intrinsics, we additionally test them by providing the ground-truth intrinsic parameters to evaluate performance when exact camera calibration is available.

Evaluation metrics. We follow conventional evaluation protocols from existing benchmarks [3, 12, 13, 15, 26, 36]. Specifically, we report mean-per-joint-position-error (MPJPE) and per-vertex error (PVE), measuring Euclidean distances between predicted and ground-truth 3D joints and mesh vertices, respectively. We also provide these metrics after rigid alignment (PA-MPJPE, PA-PVE) via Procrustes Analysis (PA). Additionally, temporal coherence is evaluated using Acceleration Error (Accel), computed as the difference in joint accelerations from ground truth, and Jitter, computed as the norm of the third-order temporal derivatives of predicted joint positions. Since lower jitter indicates smoother motion but does not necessarily imply higher accuray, we also report the ground truth jitter α as $Jitter_\alpha$. Predictions closer to the ground truth indicate better motion estimation quality. Lastly, we report 3D Percentage of Correct Keypoints (3DPCK) and Area Under Curve (AUC) [17, 26]. We use 24 body joints and 6,890 vertices configurations from SMPL [21] for body-joints and vertex metrics, respectively.

Evaluation results. Table 2 summarizes our extensive evaluation of state-of-the-art models on the BodyContact4D test set. Notably, existing models significantly underperform on our dataset compared to their previously reported results on widely-used benchmarks. This demonstrates the expensive nature of our dataset, which includes dynamic human–scene interactions that inherently induce frequent occlusions—conditions less prominently represented in existing datasets. For instance, CameraHMR [25] achieved an average MPJPE of 62.7 mm and a PVE of 73.5 mm across 3DPW [36], EMDB [15], and SPEC-SYN [18]. In contrast, on our dataset, CameraHMR obtained MPJPE and PVE values of 79.5 mm and 91.0 mm, respectively, representing increases of 26% and 24%. Similar trends were consistently observed across all evaluated methods. NLF [28] consistently shows the best geometric accuracy (MPJPE, PVE,

Models	MPJPE ↓	PA-MPJPE ↓	PVE ↓	PA-PVE ↓	Accel ↓	Jitter$_{0.9}$ →	3DPCK ↑	AUC ↑
BEDLAM-CLIFF [3]	133.8	76.8	150.9	86.2	50.6	27.0	65.2	41.6
BEDLAM-CLIFF† [3]	125.7	77.1	139.2	86.8	50.7	27.3	69.2	45.1
HMR2.0 [9]	129.3	59.3	148.9	67.3	22.0	11.6	66.5	42.7
TokenHMR [6]	129.5	57.7	149.2	65.5	17.1	8.9	66.0	42.8
WHAM‡ [31]	122.6	68.7	138.7	77.4	**5.4**	**2.0**	70.6	46.4
TRAM‡ [37]	104.5	55.8	116.0	64.8	5.7	3.5	75.8	53.2
NLF [28]	94.8	**53.8**	106.8	59.7	27.9	14.9	82.1	55.7
NLF† [28]	**82.6**	55.4	**92.0**	**61.5**	28.4	15.1	**87.0**	**61.5**
CameraHMR [25]	88.6	56.4	99.8	62.5	20.3	10.6	85.3	58.5
CameraHMR† [25]	86.8	56.5	97.0	62.7	19.8	10.3	86.0	59.2

Table 2. Evaluations of state-of-the-art methods on the BodyContact4D `test set` for 3D human pose and shape estimation. Best results are in **bold**. † denotes when the ground truth camera intrinsic was used during the evaluation and ‡ indicates the temporal models.

	Full body			Hand			Foot		
Models	Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑
WHAM [31]	–	–	–	–	–	–	0.54	0.72	0.67
GVHMR [30]	–	–	–	0.38	0.80	0.52	0.52	0.94	0.67
BSTRO [12]	0.51	0.64	0.57	0.52	0.48	0.50	0.54	0.86	0.66
DECO [33]	0.48	0.70	0.57	0.44	0.28	0.34	0.50	**0.99**	0.66
DECO-VID	**0.64**	**0.74**	**0.69**	**0.64**	**0.83**	**0.73**	**0.66**	0.80	**0.72**

Table 3. Evaluations of state-of-the-art methods on the BodyContact4D `test set` for human contact estimation. Best results are in **bold**.

3DPCK). In general image-based models [25, 28] provide higher geometric accuracy, while temporal models [31, 37] tend to yield smoother reconstructions as reflected by lower Accel and Jitter metrics.

4.2. Human–Scene Contact Estimation Benchmark

We benchmark state-of-the-art methods on the human-scene contact estimation task using our BodyContact4D test set, as summarized in Table 3. Specifically, we evaluate two recent image-based human contact detection models, DECO [33] and BSTRO [12], alongside two video-based methods, WHAM [31] and GVHMR [30], which leverage detected hand and foot contacts to refine global human trajectories.

Training setup. To rigorously assess how existing state-of-the-art models generalize to the challenging, dynamic interactions in BodyContact4D , we evaluate all baseline methods (DECO, BSTRO, WHAM, GVHMR) in a zero-shot setting using their official public weights. In contrast, our proposed **DECO-VID** serves as a reference for the performance achievable when leveraging the BodyContact4D training set. Specifically, DECO-VID utilizes the pre-trained DECO feature extractor (which remains frozen to prevent overfitting to visual patterns) and trains only the newly added temporal encoder and regressor modules on the BodyContact4D training split.

Evaluation metrics. We evaluate human–scene contact estimation performance using standard classification metrics: precision, recall, and F1 score. In order to encompass models only detect part of the body contact, we report results separately for three categories; full-body, hand, and foot. While BodyContact4D provides contact labels based on predefined 24 body segments, image-based models (DECO [33], BSTRO [12]) predict per-vertex contacts. To address this discrepancy, we convert their predictions to our segment-based format. Specifically, if a model predicts contact for more than 30% of the vertices corresponding to a particular body segment, we classify that segment as "in contact."

Evaluation results. Table 3 demonstrates that our proposed DECO-VID model consistently outperforms baseline methods, achieving over 10% improvement across all 24 body segments. Notably, DECO-VID shows particularly strong performance in detecting hand contacts, significantly surpassing existing approaches, while also matching their performance on foot contacts. Figure 5 qualitatively compares DECO-VID with DECO and BSTRO, highlighting our method's effectiveness in challenging scenarios.

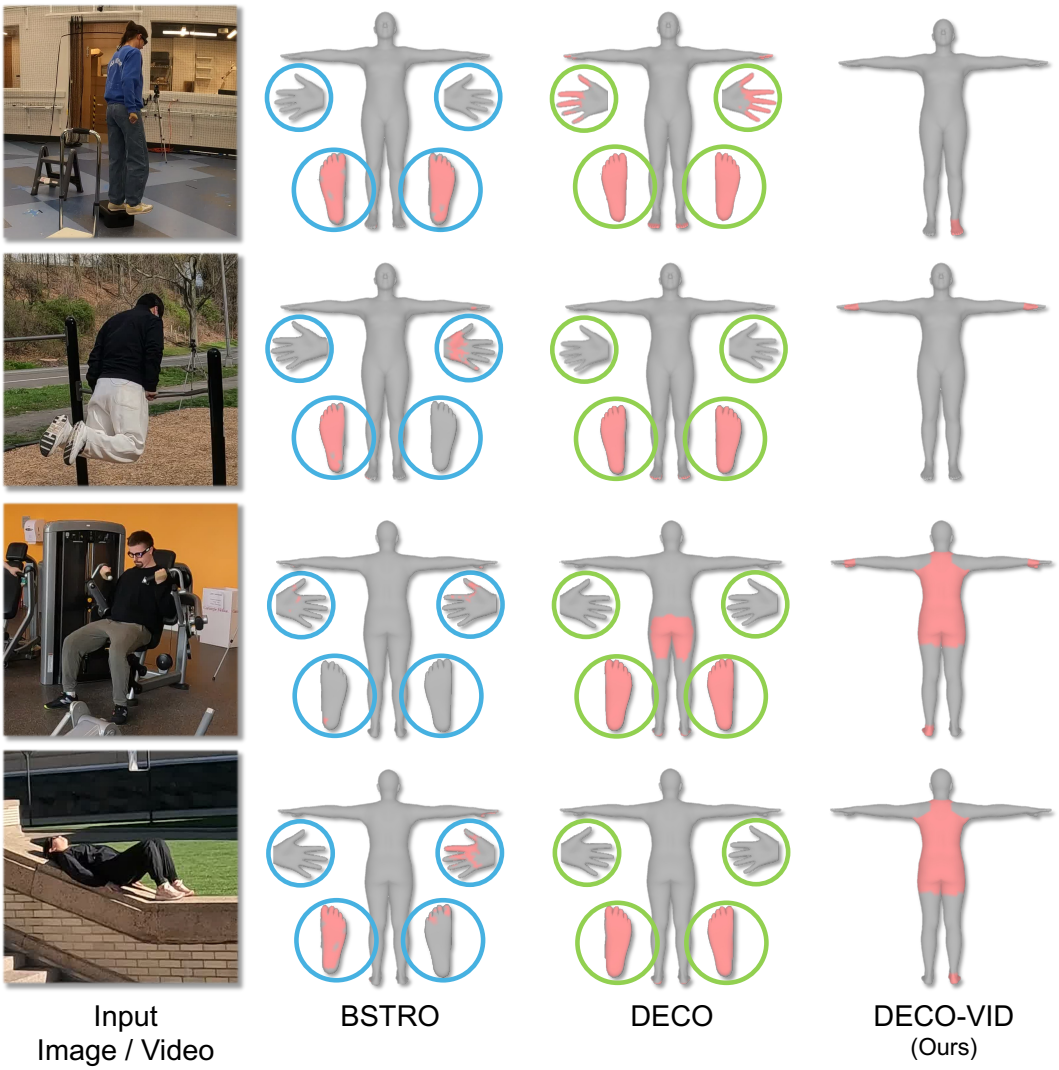

Figure 5. Qualitative Comparison of Contact Detection. We compare our DECO-VID against baseline methods, DECO [33] and BSTRO [12]. Since DECO and BSTRO predict per-vertex contacts, we provide additional close-up views for hands and feet to clarify their predictions.

5. Conclusion

We introduce *BodyContact4D*, a large-scale multi-view video dataset specifically designed to capture dynamic human–environment interactions in diverse real-world settings. Integrating synchronized multi-view video with a customized full-body force-sensing system, our dataset provides comprehensive and accurate annotations of human-scene contacts alongside accurate 3D human motion reconstructions. Benchmark evaluations on 3D human pose and shape estimation models demonstrated that our dataset significantly challenges existing methods due to frequent occlusions and the dynamic nature of human–scene interactions. We further presented *DECO-VID*, the first baseline video-based human contact detection model, which outperforms existing approaches by leveraging temporal context.

Limitations. Our point-based sensors may yield false negatives in uninstrumented areas due to specific setup constraints, or false positives resulting from clothing pressure and body bracing. Additionally, we abstract absolute force measurements into binary labels, prioritizing vision-based classification over continuous dynamics. Finally, the current lack of 3D scene reconstructions necessitates scene-agnostic mesh recovery, which can occasionally result in mesh-scene penetrations. Future work could integrate physics-based simulations with joint human-scene reconstruction to better address these challenges and advance the robustness of interaction modeling.

Acknowledgement Soyong Shin is supported by the Meta AI Mentorship (AIM) program.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Sally A. Applin and Catherine Flick. Facebook's project aria indicates problems for responsible innovation when broadly deploying ar and other pervasive technology in the commons. *Journal of Responsible Technology*, 5:100010, 2021.

[3] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[7] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[8] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023.

[10] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

[11] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, 2019.

[12] 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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[14] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. *ICCV*, 2015.

[15] 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 *International Conference on Computer Vision (ICCV)*, 2023.

[16] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. *arXiv preprint arXiv:2305.16487*, 2023.

[17] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions, 2024.

[18] 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 *Proc. International Conference on Computer Vision (ICCV)*, pages 11015–11025, Piscataway, NJ, 2021. IEEE.

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

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, 2014.

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015.

[22] B. Luo and E.R. Hancock. Iterative procrustes alignment with the em algorithm. *Image and Vision Computing*, 20(5): 377–396, 2002.

[23] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[24] Lea Muller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3D social interaction from images. 2024.

[25] Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *International Conference on 3D Vision (3DV)*, 2025.

[26] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[27] 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)*, pages 10975–10985, 2019.

[28] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. 2024.

[29] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] 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.

[31] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

[32] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, 2010.

[33] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8001–8013, 2023.

[34] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[36] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.

[37] 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.

[38] Guo Wen, Bie Xiaoyu, Alameda-Pineda Xavier, and Moreno-Noguer Francesc. Multi-person extreme motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[39] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.

[40] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory, 2024.

[41] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. *arXiv preprint arXiv:2312.08963*, 2023.

[42] S.I. Yaniger. Force sensing resistors: A review of the technology. In *Electro International, 1991*, pages 666–668, 1991.

[43] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[44] He Zhang, Shenghao Ren, Haolei Yuan, Jianhui Zhao, Fan Li, Shuangpeng Sun, Zhenghao Liang, Tao Yu, Qiu Shen, and Xun Cao. Mmvp: A multimodal mocap dataset with vision and pressure sensors. *CVPR*, 2024.