Capturing and Inferring Dense Full-Body Human-Scene Contact

Chun-Hao P. Huang[1]   Hongwei Yi[1]   Markus Höschle[1]   Matvey Safroshkin[1]   Tsvetelina Alexiadis[1]
Senya Polikovsky[1]   Daniel Scharstein[2]   Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany   [2]Middlebury College
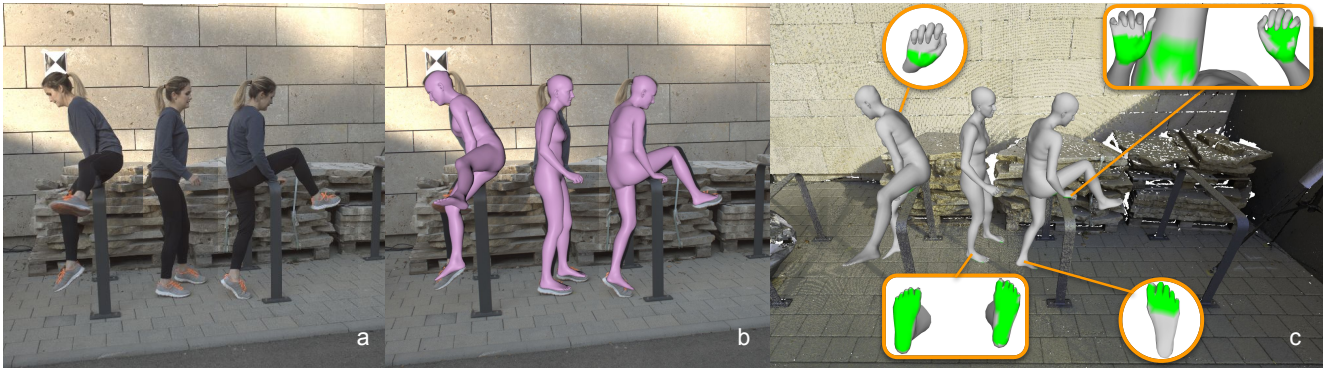{paul.huang, firstname.lastname, black}@tuebingen.mpg.de, schar@middlebury.edu

Figure 1. **RICH** is a new dataset containing videos of people in natural scenarios and standard clothing together with ground-truth 3D body pose and shape (a-b). A key novelty of RICH is that it also contains 3D scene scans, which enable dense and accurate labeling of human-scene contact (c, green). We exploit this to learn a regressor called BSTRO that takes an image and infers human-scene contact.

Abstract

Inferring human-scene contact (HSC) is the first step toward understanding how humans interact with their surroundings. While detecting 2D human-object interaction (HOI) and reconstructing 3D human pose and shape (HPS) have enjoyed significant progress, reasoning about 3D human-scene contact from a single image is still challenging. Existing HSC detection methods consider only a few types of predefined contact, often reduce the body and scene to a small number of primitives, and even overlook image evidence. To predict human-scene contact from a single image, we address the limitations above from both data and algorithmic perspectives. We capture a new dataset called RICH for "Real scenes, Interaction, Contact and Humans." RICH contains multiview outdoor/indoor video sequences at 4K resolution, ground-truth 3D human bodies captured using markerless motion capture, 3D body scans, and high resolution 3D scene scans. A key feature of RICH is that it also contains accurate vertex-level contact labels on the body. Using RICH, we train a network that predicts dense body-scene contacts from a single RGB image. Our key insight is that regions in contact are always occluded so the network needs the ability to explore the whole image for evidence. We use a transformer to learn such non-local relationships and propose a new Body-Scene contact TRansfOrmer (BSTRO). Very few methods explore 3D contact; those that do focus on the feet only, detect foot contact as a post-processing step, or infer contact from body pose without looking at the scene. To our knowledge, BSTRO is the first method to directly estimate 3D body-scene contact from a single image. We demonstrate that BSTRO significantly outperforms the prior art. Our code and dataset are available for research purposes at: https://rich.is.tue.mpg.de

1. Introduction

Understanding human actions and behaviors has long been studied in computer vision, with applications in robotics, healthcare, virtual try-on, AR/VR, and beyond. Remarkable progress has been made in both 2D human pose detection [8,30,34,45,70,85] and 3D human pose and shape estimation (HPS) from a single image [6, 38, 42, 43, 47, 59, 84, 99], thanks to realistic datasets annotated with 2D keypoints [2,35,49] and 3D data [32,36,52,69,80]. Despite this progress, something important is missing. Even the most basic human activities, such as walking, involve interaction

with the surrounding environment. Fundamentally, human-scene interaction (HSI) involves the contact relationships between a 3D human and a 3D scene, i.e., human-scene contact (HSC). Existing HPS methods, however, largely ignore the scene and estimate human poses and/or shapes in isolation, often leading to physically implausible results.

Since reconstructing the full 3D scene from a single image is challenging, recent HPS methods tackle this problem by making several simplifying assumptions about the scene and/or body. Many methods consider only the contact between feet and ground [64,67,86,92,93,97,105], or assume the ground is a even plane [63], which is often violated, e.g., walking up stairs. To infer contact, many state-of-the-art (SOTA) methods use MoCap datasets [51,53] to train a contact detector [64,97,105]. Others exploit physics simulation [67,92] or physics-inspired objectives [86] but reduce the body representation to a small set of primitives. Surprisingly, none of these methods use image evidence when predicting human-scene contact. This is primarily due to the lack of datasets with images and 3D contact ground truth.

Many methods do estimate human object interaction (HOI) from images but constrain the reasoning to 2D image regions [39,61,81,88,104]. That is, they estimate bounding boxes or heatmaps in the image corresponding to contact but do not relate these to the 3D body.

In this work, we address this problem with a framework that estimates 3D contact on the body directly from a single image. We make two main contributions. First, we create a new dataset that accurately captures human-scene contact by extending a markerless MoCap method to markerless HSC capture. Specifically, we capture multiview video sequences at 4K resolution in both indoor and outdoor environments. We also capture the precise 3D geometry of the scene using a laser scanner. Additionally, we capture high-resolution 3D scans of all subjects in minimal clothing and fit the SMPL-X body model [59] to the scans. Our markerless HSC approach allows us to compute accurate per-vertex scene contact, as visualized in Fig. 1c.

Compared to the PROX dataset [27], which captures HSC with monocular RGB-D input, multiview data has two advantages: (1) it effectively resolves occlusions, leading to better reconstructed bodies and consequently more accurate scene contact; (2) it works for outdoor environments, as shown in Fig. 1.

The resulting dataset, called RICH ("Real scenes, Interaction, Contact and Humans"), provides: (1) high-resolution multiview images of single or multiple subjects interacting with a scanned 3D scene, (2) dense full-body scene-contact labels, (3) high-quality outdoor/indoor scene scans, (4) high-quality 3D human shapes and poses, and (5) dynamic backgrounds and moving cameras.

To estimate vertex-level HSC from a single color image, we develop BSTRO (Body-Scene contact TRansfOrmer),

and train it with RICH. Our key insight in building BSTRO is that contact is not directly observable in images due to occlusion; thus, to infer contact, the network architecture must be able to explore the whole image for evidence. The transformer architecture enables BSTRO to learn non-local relationships and use scene information to "hallucinate" unobserved contact. We employ a multi-layer transformer [78], which has been successfully employed for natural-language processing [13] and HPS estimation with occlusion [47].

In summary, our key contributions are: (1) We present RICH, a novel dataset that captures people interacting with complex scenes. It is the first dataset that provides *both scans of outdoor scenes and images* for monocular HSC estimation, unlike existing methods [26,27], which lack one or the other. (2) We propose BSTRO, a monocular HSC detector. It is *body-centric* so it does not require 3D scene reconstructions to infer contact. Unlike POSA [28], which is also body-centric, BSTRO directly estimates dense scene contact from the input image without reconstructing bodies. (3) We evaluate recent HSC methods and show that BSTRO gives SOTA results. (4) Since RICH has pseudo-ground-truth body fits, we also evaluate SOTA HPS methods and analyze their performance with respect to scene-contact, which is not supported by existing HPS datasets [32,58,80]. We confirm that the performance of a SOTA HPS method [19] degrades in the presence of scene contact.

2. Related Work

We review existing methods that consider contact between humans and scenes. Since many of them employ a 3D body reconstruction method as a backbone in the pipeline, we first briefly discuss recent HPS trends and then focus on how the prior art incorporates scene contact.

2.1. Human Pose and Shape Estimation (HPS)

Monocular HPS methods reconstruct 3D human bodies from a single color image. Many methods output the parameters of statistical 3D body models [3,37,50,59,89]. SMPLify [6] fits the SMPL model to the output of a 2D keypoint detector [60] and we build on it here.

In contrast, deep neural networks regress body-model parameters directly from pixels [11,18,19,25,38,41–43, 65,71,72]. To deal with the lack of in-the-wild 3D ground truth, some methods use 2D keypoints [38,74,77] or linguistic attributes [10] as weak supervision, while some directly fine-tune the network w.r.t. an input image at test time [36]. Kolotouros et al. [43] combine HMR [38] and SMPLify [6] in a training loop for better 3D supervision. On the other hand, non-parametric or model-free approaches directly estimate 3D vertex locations without body parameters [9,14,44,47,48,55,94]. We refer readers to [75,103] for a comprehensive review. None of the above methods estimate HSC.

Table 1 (upper part):

Methods	Scene Contact Body / Scene	Body	Contact Cues Train / Test
Zanfir et al. [93]	foot joints / ground	mesh	- / dist.
Zou et al. [105]		joint	both 2D vel.
Rempe et al. [64]			vel. & dist. / vel. & dist.
PhysCap [67]		part	
HuMoR [63]	8 joints / ground	mesh	
LEMO [97]	foot vert. / ground		
SimPoE [92]	foot parts / ground	part	physics simulation
Xie et al. [86]			
HolisticMesh [83]	dense body mesh / scene mesh	mesh	- / dist.
PROX [27]			- / dist.
PHOSA [96]			- / dist.
Zhang et al. [101]			dist. / -
PLACE [98]			dist.
POSA [28]			dist. / pose
BSTRO (ours)	as above	mesh	dist. / image

Datasets	Contact Label	Img	Scene
MTP [56]	self-contact	✓	N/A
GRAB [73]	hand-object	✗	N/A
ContactHands [57]	hand-X‡	✓	N/A
Fieraru et al. [20]	person-person	✓	N/A
Fieraru et al. [21]	self-contact	✓	N/A
PiGraph [66]	joint-scene	✓	RGBD scans
i3DB [54]	N/A	✓	CAD
GPA [82]	N/A	✓	Cubes
Guzov et al. [26]	foot-ground	✓¶	laser scans
PROX [27]	body-scene	✓	RGBD scans
RICH (ours)	body-scene	✓	laser scans

Table 1. **Comparison of contact-related methods and datasets.**
‡: X can be self, person and object. ¶: egocentric images. Vert.: vertex; vel.: velocity; dist.: distance.

Markerless MoCap exploits synchronized videos from multiple calibrated cameras and has a long history with commercial solutions, but these focus on estimating a 3D skeleton. To model HSC, we need to extract a full 3D body shape and, therefore, focus on such methods here. Early methods, either bottom-up [5,24,68] or top-down [4,22,79], are fragile, need subject-specific templates and manual input, and do not generalize well to in-the-wild images.

Powered by CNNs, recent methods leverage multiview consistency to improve keypoint detection [29, 33, 62, 76], to re-identify subjects across views [16] or across view and time [15, 100], but they estimate only joints, not body meshes. Dong et al. [17] reconstruct SMPL bodies for multiple subjects and Zhang et al. [102] additionally estimate hands and facial expressions. They demonstrate results for lab scenarios, while our HSC capture method in Sec. 3.1 works in less constrained outdoor scenes.

All methods above reconstruct human bodies in isolation without taking into account the interaction with scenes. Consequently, the results often contain physically implausible artifacts such as foot skating and ground penetration.

2.2. Human Scene Interaction (HSI)

2D Human-Object Interaction (HOI) methods localize 2D image regions with HOI and recognize the semantic interactions in them. Most methods represent humans and objects very roughly as bounding boxes [39, 104]; only a few use body meshes for humans and spheres for objects [46].

3D Contact. Knowing which part of the body and scene are in contact provides compact yet rich information that enables many applications, such as HSI recognition [7] or placing virtual humans into a scene [28]. The upper part of Table 1 summarizes how body-scene contact gets incorporated in methods of different goals and tasks.

Early work uses scene contact as part of the HSI feature [54,66] but represents a human body roughly as a stick figure. Recent HPS methods [27, 63, 64, 105] use contact to improve the estimated body poses. Ideally, when both the body and scene are "perfectly reconstructed," applying a threshold to the 3D Euclidean distances between them is sufficient to infer accurate contact. Prior work takes this thresholding approach to annotate contact [26, 27, 56, 73]. At test time, PROX [27] assumes scene scans to be known a priori; PHOSA [96] estimates 3D objects, 3D people, and the contacts between them but only for a limited class of objects. Since reconstructing a 3D scene in high quality with correct layout and spatial arrangement is still an open challenge [90], monocular HSC detection methods resort to other heuristics. The most common one is a zero-velocity assumption; i.e., surfaces in contact should not slide relative to each other. This assumption is widely employed to reduce foot-skating artifacts [63, 64, 67, 105]. Some of these detect contact with a separate neural network at test time, taking the 2D/3D joints in a temporal window as input [64, 67, 105], while others integrate it in a body motion prior [63, 97]. These approaches use MoCap datasets such as AMASS [51] and Mixamo [53] to build training data, where contact is automatically labelled via thresholding the distance to the ground and/or the velocity.

POSA [28] observes that scene contact is correlated with body poses and introduces a generative model to sample contact given a posed mesh. Some methods [67, 86, 92] apply physics to encourage foot-ground contact and ensure physically plausible motions. However, they have to approximate the body as a set of boxes, cylinders or spheres. MOVER [90] uses human scene contact to improve monocular estimation of 3D scene layout.

All these approaches first reconstruct bodies (2D or 3D), and then reason about contact, effectively ignoring valuable image information. To go further, we need a dataset consisting of natural images and 3D body-scene contact labels. As summarized in the lower part of Table 1, many existing contact-related datasets consider self contact [21, 56] or person-person contact [20], but not HSC. The most relevant datasets for HSC are [26] and PROX [27]. The former provides egocentric images for localization, which are not suitable for HSC detection from images. PROX [27] can be used for our task but it consists of only indoor scenes

and is of lower quality. The ground-truth bodies in PROX are computed by fitting to RGBD data, which is sensitive to occlusions. This not only limits the type of HSI in the dataset (mostly walking, sitting, lying) but also influences the quality of body fits.

3. Methods: RICH Dataset

Overview and preliminaries. Unlike [64,67,86,91], which represent a body as a set of coarse geometry primitives, we follow [27,28] to capture realistic human-scene contact with a parametric SMPL-X body model [59]. The vertex locations on a SMPL-X mesh $M(\theta, \beta, \psi) \subset \mathbb{R}^3$ are controlled by parameters for pose θ, shape β, and facial expression ψ. θ consists of body pose θ_b and hand pose θ_h. Hand pose θ_h is a function $\theta_h(Z_h)$ of a PCA latent vector $Z_h \in \mathbb{R}^{12}$.

Given videos captured by C synchronized cameras, we first identify each subject across views and across time with [16, 87]. For each identified subject, we reconstruct a SMPL-X body by a multiview fitting method that is robust to noisy 2D keypoint detections, and we place it in a pre-scanned scene to compute body-scene contact (Sec. 3.1). With this approach, we build a monocular body-scene interaction dataset (RICH) comprising 577K images paired with SMPL-X parameters and scene contact labels (Sec. 5).

3.1. Capturing Dense Body-Scene Contact

We first track subjects temporally in each video with AlphaPose [87], followed by MvPose [16] to match the tracklets across views. Other methods that build such *4D associations* [15,100] could also be applied here.

At time t, we now have at most C bounding boxes of the same person and we aim to reconstruct the body. To this end, we adapt SMPLify-X [59] to accommodate multi-view data. SMPLify-X optimizes the pose θ, shape β and facial expression ψ of SMPL-X to match the observed 2D keypoints [8] by minimizing the following objective:

$$
\begin{aligned}
E(\beta, \theta, \psi) &= E_J + E_{\text{reg}} \\
E_{\text{reg}} &= \lambda_{\theta_b} E_{\theta_b} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C},
\end{aligned} \tag{1}
$$

where E_J is the data term, and E_{reg} includes several regularization terms: θ_b is the pose vector for the body, which is a function $\theta_b(Z_b)$, where $Z_b \in \mathbb{R}^{32}$ is a VAE latent representation and E_{θ_b} is an L_2 prior defined on Z_b. $E_\alpha(\theta_b)$ penalizes strong bending of elbows and knees. $E_\beta(\beta)$ is an L_2 prior on the body shape and E_C is a term penalizing mesh-intersections. λ's denote weights for each respective term. Interested readers are referred to [59] for details.

Multiview per-person reconstruction. For each person, we compute 2D keypoints [8] in each camera c. Instead of fitting them using SMPLify-X in each view, we combine all 2D landmarks in a multiview energy term: $\sum_c E_J^c$. Unlike in [59], where one needs to estimate camera translation first,
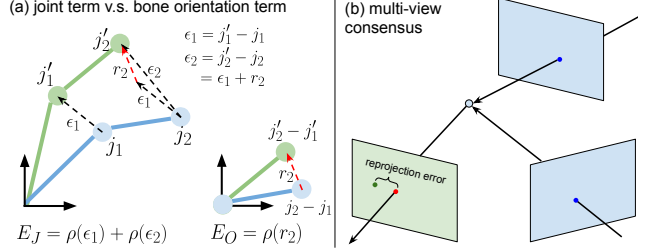

Figure 2. Illustration of bone orientation term and multiview consensus. ρ is a Geman-McClure robust estimator [23]. See text and Sup. Mat. for more discussion.

the perspective projection here is well defined by the pre-calibrated intrinsics and extrinsics. To pursue high-quality fits, body shape β is estimated in advance by registering a SMPL-X template to minimally-clothed 3D scans following [31]. β is hence no longer a free variable in Eq. 1 and we set $\lambda_\beta = 0$. In addition to E_J, which measures joint errors, we also use E_O that measures errors in "bone orientations." Figure 2(a) illustrates the intuition behind this term. Since posing human bodies requires traversing a kinematic chain, with the joint term E_J, the error of parent joints ϵ_1 is accumulated in the error of child joints ϵ_2. When $\|\epsilon_2\|$ gets too large, the influence is downweighted because our robust loss treats it as an outlier. Instead, E_O factors out the errors of ancestors and focuses on the error of the joint per se. Our final objective is $E_{\text{mv}}(\theta, \psi) = \sum_c E_J^c + \sum_c E_O^c + E_{\text{reg}}$.

Due to noisy 2D detections, keypoints in each view often disagree with each other. One may count on the robustifier to identify outliers and reduce their contribution. This depends, however, on the current estimated body in the optimization, so it assumes good initialization. Instead, we check the multiview consistency of landmarks as illustrated in Fig. 2(b). For each joint, we take the detections in two views (blue), triangulate a 3D point and project it to the third view (green). If the distance between the projected point (red) and the detection (green) in the third view is large, that means the three detections do not agree with each other and at least one of them is wrong. Instead of making hard decision separating outliers from inliers, we exhaustively compute all triplets of views, accumulate the reprojection error and downweight the contribution in $\sum_c E_J^c$ for views with high errors. We term this *multiview consensus*, as it behaves like a soft majority voting mechanism. As long as there are more correct detections than wrong ones, it can reduce the influence of noisy landmarks, independent of the current body estimate.

To further avoid local minima, we apply a state-of-the-art in-the-wild body regressor (PARE [42]) to initialize θ. We run PARE on the bounding box from each view, fuse the results by averaging the poses, and covert the fused body from SMPL to SMPL-X. The SMPL-X body pose gives the

initial value of θ for minimizing E_{mv}. We first solve E_{mv} in a frame-wise manner and then refine a batch of T frames jointly with two additional terms on body and hand motions: $E_{\mathrm{batch}}(\theta_1, \cdots, \theta_T) = \sum_{t=1}^{T} E_{\mathrm{mv}}^t + \lambda_{\mathrm{sm}}^b E_{\mathrm{sm}}^b + \lambda_{\mathrm{sm}}^h E_{\mathrm{sm}}^h$. E_{sm}^b is the smoothness term in [97] and E_{sm}^h encourages neighboring frames to have similar hand-pose PCA vectors Z_h.

We place the reconstructed bodies into pre-scanned 3D scenes to estimate the body-scene contact. The scene mesh and HDR textures were acquired using an industrial laser scanner, Leica RTC360. To put the bodies in the scene, we solve the rigid transformation between camera coordinates and scan coordinates with manually identified correspondences. To annotate human-scene contact automatically, our approach is similar to POSA [28]. Specifically, for each vertex on the body mesh, we compute the point-to-surface distance to the scene scan. If the distance is lower than a threshold and the normal is compatible, we accept the hypothesis that it is *in contact*. Considering the thickness of shoe soles, the threshold is 5cm for the vertices at the bottom of feet and 2.5cm for the rest of body. This is different from POSA, which uses 5cm for the whole body to collect training data from PROX [27]. Furthermore, the pseudo-ground-truth body poses in PROX are obtained by fitting the SMPL-X template to monocular RGBD data. As shown in the bottom row of Fig. 5, PROX accuracy suffers from occlusion, sometimes resulting in severe penetration with the scene. The errors in body fits are carried over to the ground-truth HSC data for POSA. In contrast, in RICH, bodies are recovered from multiview data, which reduces the issues caused by occlusion and depth ambiguity.

4. Methods: BSTRO

Here we introduce BSTRO for dense HSC estimation from a single image. This relies on RICH, described in Sec. 5 in detail. Existing HSC methods usually take a multi-stage approach. Given an input image, they first reconstruct the body mesh and use it as a proxy to infer contact. Formally, let f denote the function recovering a body mesh M from the input image I, $M = f(I)$. f can be an energy-minimization process such as [59] or a neural network as in [38, 42]. To estimate contact, SOTA methods differ from each other in two ways: (1) the features extracted from M, e.g., Euclidean distance to the 3D scene, velocity and body poses (cf. Table 1); (2) the prediction functions, e.g., simple thresholding, neural network, or physics engine. With a slight abuse of notation, we denote these feature extraction and contact estimation processes collectively as g, which takes the body M as input and predicts a contact vector $\mathbf{c} = g(M)$. Each element in $\mathbf{c}$ is 1 if the corresponding part of the body (vertex, joint or body part) is in contact with the scene, and 0 otherwise. For example, g represents the decoder of a conditional VAE in POSA [28], taking the vertex locations of M as input, while in [63, 64, 67], g is a
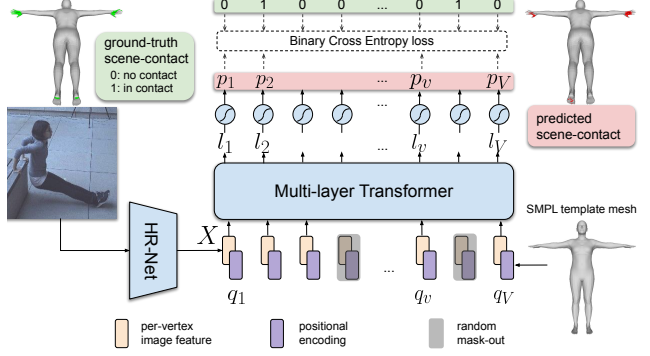

Figure 3. **BSTRO model architecture.** Given an input image, BSTRO predicts dense per-vertex contact labels by exploiting image information, without reconstructing 3D poses or 3D bodies.

MLP operating on the motion of M.

With this formulation, the body-scene contact $\mathbf{c}$, whether defined on a dense mesh or on a set of sparse joints/parts, is a composite function of g and f: $\mathbf{c} = g \circ f(I)$, where g is agnostic to the input image. In contrast, our goal is to detect dense body-scene contact directly from the input I: $\mathbf{c} = g(I)$. To our knowledge, this was explored only for self-contact [21] and person-person contact [20] and only at a coarse region level, not the vertex level.

We use SMPL as the body representation for BSTRO, hence $\mathbf{c} \in \{0, 1\}^V$, where V=6890 is the number of vertices on a SMPL mesh, as opposed to V=10475 on a SMPL-X mesh. The reason for this choice is that a SMPL-X mesh has nearly 50% of the vertices on the head, which rarely participates in natural body-scene contact, so we would like to reduce the dimensionality of the output space. See Sup. Mat. for more discussion of this design choice.

We model g as a neural network and train it end-to-end in a supervised way with the $(I, \mathbf{c})$ pairs sampled from RICH. The network architecture is designed based on our key observation. That is, regions in contact are not directly observable due to occlusion. However, there is rich information in the image to tell which parts of the body are in contact with the scene. Estimating HSC from images is therefore inherently a "hallucination" task. Without really "seeing" the regions in contact, the network needs to explore the image freely and attend to regions it finds informative.

We use a multi-layer transformer [13] to learn such a non-local relationship from data and propose the *Body-Scene contact TRansfOrmer* (BSTRO). Figure 3 visualizes the architecture of BSTRO. It takes an image of a person as input, extracts features $X \in \mathbb{R}^{2048}$ with a CNN backbone, and appends vertex locations of the SMPL template as positional encoding. The feature after concatenation is denoted as $q \in \mathbb{R}^{2051}$. The input query of the transformer is a set of q: $Q = \{q_v\}_{v=1}^V$. The transformer outputs an array of logits l_v, which, after applying sigmoid functions, result in elements $p_v \in [0, 1]$ encoding the probability of vertex v

being in contact. Finally, the dense scene-contact vector **c** is obtained by thresholding p_v at 0.5. Note that BSTRO is a non-parametric method, in spirit similar to [47] that makes prediction for each vertex directly without passing through a parametric model.

Training. We apply the binary cross entropy loss between the ground truth contact and the predicted contact probability p_v. One can think of this as a multi-label classification problem, where each category (vertex) has its own probability of being true (in contact) or not.

To gain robustness to occlusion, we employ Masked Vertex Modeling (MVM) [47]. Specifically, at each iteration, we randomly mask out some queries in Q and still ask the transformer to estimate contact for all vertices. In order to predict the output of a missing query, the model has to explore other relevant queries. This simulates occlusions where bodies are only partially visible and also encourages the network to hallucinate contact.

5. RICH Dataset

We capture 22 subjects performing various human-scene interactions in 5 static 3D scenes with 6-8 static cameras and, in some scenes, with an additional (untracked) moving camera (Fig. 4 rightmost scene). Subjects gave prior written informed consent for the capture, use, and distribution of their data for research purposes. The experimental methodology has been reviewed by the University of Tübingen Ethics Committee with no objections.

RICH has in total 142 single or multi-person multiview videos, with a total of 90K posed 3D body meshes, together with 90K dense full-body contact labels in both SMPL-X and SMPL mesh topology, and 577K high resolution (4K) images. Compared to PROX, RICH consists of mostly outdoor environments with areas of roughly $60m^2$. The images in RICH are real, not limited to a single subject, have dynamic backgrounds and varied viewpoints. All these features make it suitable for training and evaluating monocular HSC methods. Figure 4 shows several examples of RICH.

In addition, since RICH provides SMPL-X fits, i.e., pseudo-ground-truth human poses and shapes, it can also serve as a monocular or multiview HPS benchmark. It contains more subjects than 3DPW [80], more accurate body shapes than AGORA [58], and real human-scene interaction unlike Human3.6M [32]. In our experiments we analyze the performance of SOTA HPS methods with respect to body-scene contact. Such analyses are not feasible with existing HPS datasets.

6. Experiments

6.1. Dataset Split

We split 142 multiview videos in RICH into 62, 28, 52 for training, validation, and testing purposes, respectively.

The test set consists of several subsets designed for varied evaluation protocols. Each subset is defined by whether or not each of three attributes has been observed in training: scene, human-scene interaction, and subject. The most challenging subset is when they are all unseen in RICH-train. The split ensures there is one completely withheld scene and 7 unseen subjects in the test set. See Sup. Mat. for more breakdowns in terms of 3D bodies and images.

6.2. Evaluation Metrics and Baselines

We apply standard detection metrics (precision, recall, and F1 score) to evaluate the estimated dense HSC. Since vertex density varies over the SMPL template, the same number of false positives, say, on the palm and on the thigh correspond to different areas on the body surface, but this is not reflected in the scores above. To better understand how well an HSC method estimates contact, we additionally consider a measure that translates the count-based scores to errors in metric space. Specifically, for each vertex predicted in contact, we compute its shortest *geodesic distance* to a ground-truth vertex in contact. If it is a true positive, this distance is zero; if not, this distances indicates the amount of prediction error along the body.

We evaluate three HSC baselines on the RICH-test. Zou et al. [105] use the velocity of 4 2D keypoints on the feet to predict contact; HuMoR [63] estimates contact for 8 joints while reconstructing human motions. These two methods estimate contact for sparse joints, not dense vertices, so we mark all vertices that correspond to a joint as contact when the method predicts the joint is in contact. POSA [28] requires a 3D body mesh in the canonical space as input to sample dense body contact. We consider two choices of 3D bodies for POSA: (1) using the results from a SOTA body regressor PIXIE [19], or (2) using ground-truth bodies to evaluate the impact of errors in estimated body pose.

6.3. Main Results

The results on RICH-test are reported in Table 2. We see that HuMoR yields overall lowest detection scores and highest geodesic errors. This is partially due to the fact that it only considers contact with an even ground plane, while RICH-test contains more varied real scene interactions.

POSA, in general, has higher recall compared to other methods. This, however, comes with a cost of precision, meaning that there are many false positives. Comparing rows (c) and (d) we see that recall is significantly better when using ground-truth bodies. BSTRO yields significantly better precision but with lower recall than POSA. Still, it has the highest F1 score and lowest geodesic error, which shows that it strikes a good balance between precision and recall. Figure 6 shows some visual examples. RICH has accurately fitted SMPL-X bodies and body-scene contact. Given an input image, BSTRO estimates
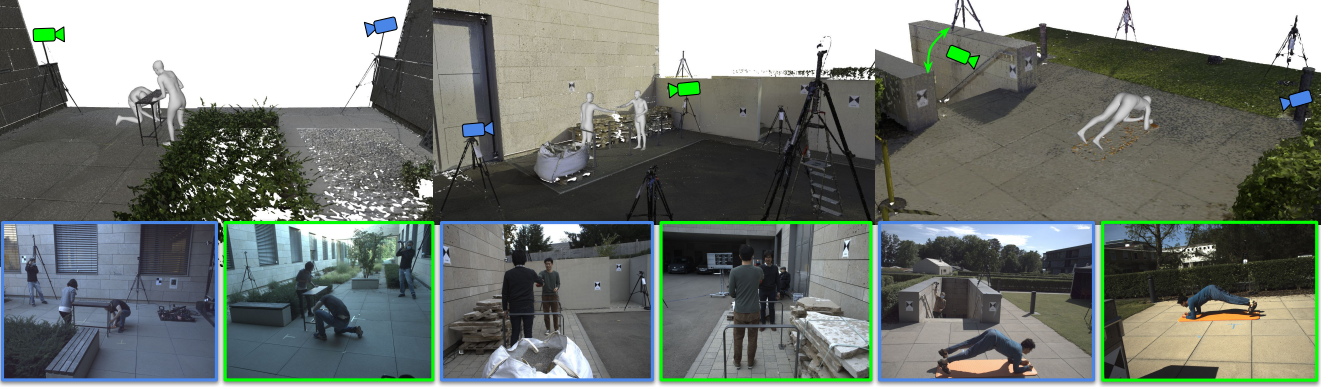
Figure 4. **RICH dataset.** In each scene we capture subjects' motions with 6-8 static cameras and, for some scenes, with 1 additional moving camera. Top row: scans of three example outdoor scenes with example 3D body meshes. Bottom row: RGB images from these scenes. The color border matches the camera icon of the same color.
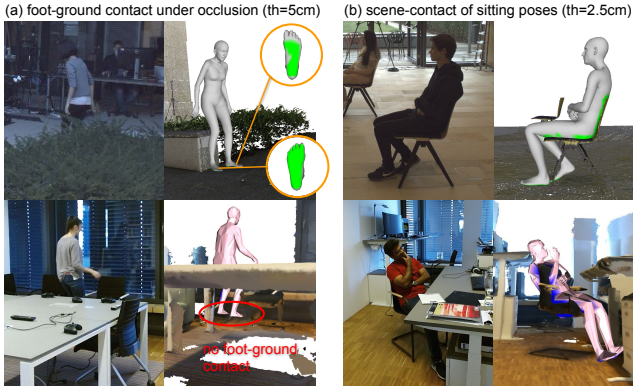

Figure 5. Comparison of HSC annotations in RICH (top) and POSA [28] (bottom). The noisy body fits in PROX [27] result in undesirable HSC labels in POSA: (a) no foot-ground contact under occlusion; (b) severe penetration with chairs.

scene contact that is closer to the ground truth, whereas POSA$^{\text{PIXIE}}$ yields false positives frequently (red circles) and sometimes misses the contact on the hands. While the training dataset is limited, BSTRO also works on in-the-wild images, as shown in the right part of Fig. 6.

6.4. Generalization

To analyze how well BSTRO generalizes, we split RICH-test into several subsets. Each subset represents whether BSTRO has observed similar images of the three attributes: scene, human-scene interaction (HSI), and subject. This allows us to inspect the importance of each attribute, and to know which aspect future methods should focus on. Note that this is a unique feature of RICH, as existing HSC datasets from MoCap [51,53] and HPS datasets [32,36,80] do not support such an analysis.

In Table 3, ✓ means BSTRO has seen similar images of that attribute during training, while ✗ means it has not. For example, images in row (a) share the same subjects and

similar HSI with training data but the scenes are new. Intuitively, this is an easy subset and indeed the scores are high in this scenario. Once HSI is withheld, the performance drops (row (e)). This drop is more pronounced than the drop caused by withholding a subject (row (c)). Comparing each of the rows (c,d,e) to row (f), we observe that seeing similar HSI at training helps the most. Seeing the same scenes or same subjects does not guarantee gains in performance. Finally, row (f) represents the most challenging subset, where scene, HSI, and subjects are all unseen during training. We see that BSTRO still yields results that are comparable to other subsets. Subset (c) contains many images with person-person occlusion, e.g., Fig. 6 bottom left, which partially explains why it is the most challenging.

6.5. HPS Evaluation on RICH-test

Besides evaluating human-scene contact, RICH can also serve as a benchmark for monocular HPS methods. Unlike existing HPS benchmarks with real images such as 3DPW [80] or Human3.6M [32], the *real* scene contact in RICH enables a new way of analyzing the performance of an HPS method. In particular, we use PIXIE [19], a recent monocular HPS method, to regress SMPL-X bodies from RICH-test. We compare the estimated SMPL-X bodies with the pseudo-ground-truth SMPL-X fits from Sec. 3.1, and compare the error when body-scene contact is present or absent.

We consider Mean Per-Joint Position Error (MPJPE) and Vertex-to-Vertex Error (V2V) to measure the discrepancies in joints and body meshes respectively. For freely moving cameras, we apply Procrustes alignment (PA) before calculating the two errors, hence PA-MPJPE and PA-V2V. Procrustes alignment factors out differences in rotation, scale and translation, focusing on measuring the difference in "pure body poses." PA hides many sources of errors so we use it only when ground-truth camera extrinsic parameters are not available. For calibrated cameras, on the other hand, we factor out only translation by aligning the estimated and

input   SMPL-X fits   GT HSC   BSTRO   POSA^{PIXIE}   BSTRO results on in-the-wild images
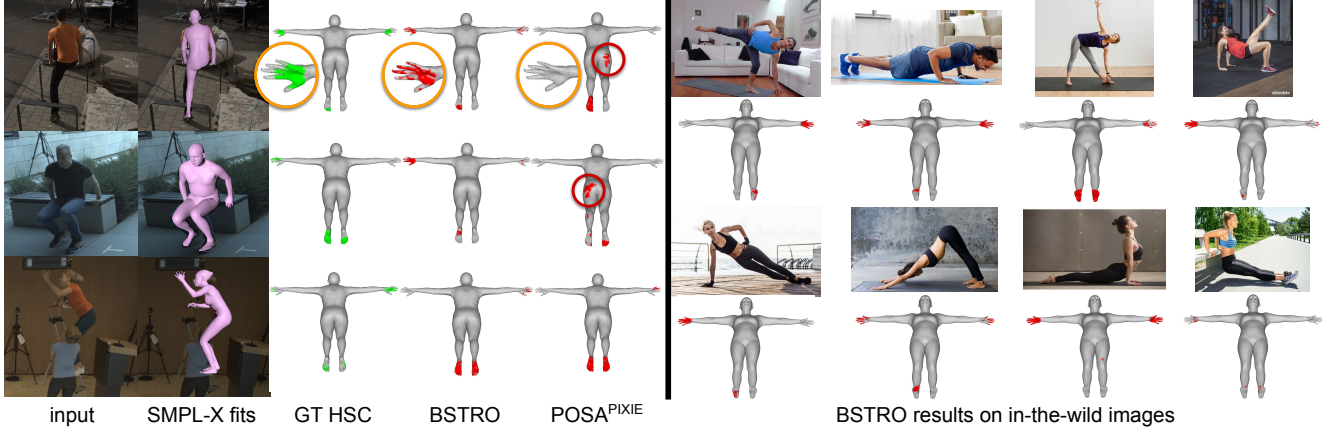
Figure 6. Left: qualitative results on RICH-test. GT HSC stands for ground-truth human-scene contact computed from the SMPL-X fits and scene scans. BSTRO estimates more accurate scene contact than POSA^{PIXIE}. Right: qualitative results on in-the-wild images.

Methods	precision ↑	recall ↑	F1 ↑	geo. error ↓
a. Zou et al. [105]	0.277	0.609	0.359	17.48cm
b. HuMoR [63]	0.248	0.527	0.314	25.35cm
c. POSA [28]^{GT}	0.375	0.768	0.464	19.96cm
d. POSA [28]^{PIXIE}	0.312	0.699	0.399	21.16cm
e. BSTRO	**0.699**	**0.774**	**0.708**	**10.98cm**

Table 2. **Evaluation on RICH-test.** POSA^{GT} means taking ground-truth bodies as input, while POSA^{PIXIE} takes the estimated bodies from PIXIE [19].

	scene	HSI	subject	p. ↑	r. ↑	F1 ↑	geo. err. ↓
a.	✗	✓	✓	**0.766**	0.819	*0.769*	**4.77cm**
b.	✓	✓	✗	0.734	0.756	0.718	9.64cm
c.	✓	✗	✗	0.556	0.593	0.520	25.11cm
d.	✗	✓	✗	*0.753*	**0.870**	**0.790**	*6.61cm*
e.	✗	✗	✓	0.644	*0.820*	0.705	16.64cm
f.	✗	✗	✗	0.682	0.814	0.721	9.00cm
g.	full RICH-test set			0.699	0.774	0.708	10.98cm

Table 3. The performance of BSTRO on each subset of RICH-test. p.: precision; r.: recall. ✓/✗: observed attribute at training. Bold/italic: the best and the 2nd best results in each metric.

ground-truth bodies to their pelvis locations, denoted with a prefix "TR." We ignore foot-ground contact, which is ubiquitous, and compare the results when there is meaningful scene contact vs. no scene contact.

On average, images containing meaningful scene contact yield 214.0mm/172.81mm TR-MPJPE/TR-V2V, higher than 161.81mm/121.71mm for images with no contact other than foot-ground contact. This is partially due to the fact that scene contact usually comes with scene occlusion, and this shows a direction where monocular HPS methods can improve. The corresponding errors in moving cameras are 84.15mm/83.16mm PA-MPJPE/PA-V2V for images with meaningful contact and 63.67mm/64.37mm for those without. We again observe that the presence of scene contact makes HPS more challenging, yielding higher errors. This shows that scene contact impacts all aspects of the problem: from pure body poses to global orientation and translation.

7. Conclusion

While there is rapid progress on estimating 3D human pose and shape from images, much of this work ignores the scene and the interaction of the body with that scene. Capture and analysis of body-scene contact, however, is critical to understanding human action in detail. To address this, and to help the research community study this problem, we created RICH, a new dataset with challenging natural video sequences, high-resolution 3D scene scans, ground-truth body shapes, high-quality reference poses, and detailed 3D contact labels. We use the contact information to train a new method (BSTRO) that takes a single image of a person interacting with a scene and infers the 3D contacts on their body. We also use the dataset to evaluate human pose estimation and find that scenes with significant contact cause problems for the state of the art. The dataset and code are available for research purposes.

Limitations and future work. RICH considers only contact with static scenes so does not account for the body contact with dynamic scenes, e.g., with hand-held objects, or human-human interaction. One extension would estimate the rigid-body pose of an object given its 3D model and simultaneously reconstruct the hand/body that interacts with it. Another interesting direction would jointly estimate the body pose, shape, and scene contact in one single network.

References

[1] https://github.com/vchoutas/smplx/tree/master/transfer_model. 13

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

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005. 2

[4] Alexandru Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 3

[5] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3D pictorial structures for multiple human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, 2016. 1, 2

[7] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 1, 4

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *European Conference on Computer Vision*, 2020. 2

[10] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression using metric and semantic attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[11] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, 2020. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 13

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2, 5

[14] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J. Cashman, and Jamie Shotton. Full-body motion from a single head-mounted device: Generating SMPL poses from partial observations. In *International Conference on Computer Vision*, 2021. 2

[15] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3D pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3, 4

[16] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3D pose estimation from multiple views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 4

[17] Zijian Dong, Jie Song, Xu Chen, Chen Guo, and Otmar Hilliges. Shape-aware multi-person pose estimation from multi-view images. In *International Conference on Computer Vision*, 2021. 3

[18] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to regress bodies from images using differentiable semantic rendering. In *International Conference on Computer Vision*, 2021. 2

[19] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision*, 2021. 2, 6, 7, 8

[20] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3, 5

[21] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3D human self-contact. In *AAAI Conference on Artificial Intelligence*, 2021. 3, 5

[22] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 3

[23] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, volume 52, 1987. 4

[24] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3D structure with a statistical image-based shape model. In *International Conference on Computer Vision*, 2003. 3

[25] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[26] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[27] 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*, 2019. 2, 3, 4, 5, 7

[28] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 4, 5, 6, 7, 8

[29] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[30] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *International Conference on Computer Vision*, 2019. 1

[31] David Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conference on Computer Vision*, 2012. 4, 13

[32] 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. 1, 2, 6, 7, 13

[33] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision*, 2019. 3

[34] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, 2020. 1

[35] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[36] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision*, 2021. 1, 2, 7

[37] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[38] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5

[39] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-end human-object interaction detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 13

[41] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[42] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision*, 2021. 1, 2, 4, 5

[43] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, 2019. 1, 2

[44] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[45] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *International Conference on Computer Vision*, 2021. 1

[46] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2D-3D joint representation for human-object interaction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[47] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 6, 13

[48] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *International Conference on Computer Vision*, 2021. 2

[49] 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 *European Conference on Computer Vision*, 2014. 1

[50] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[51] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, 2019. 2, 3, 7

[52] 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 *International Conference on 3D Vision*, 2017. 1

[53] *Mixamo*, 2021. https://www.mixamo.com. 2, 3, 7

[54] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J. Mitra. iMapper: interaction-guided scene mapping from monocular videos. *ACM Transactions on Graphics*, 38(4):92:1–92:15, 2019. 3

[55] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *European Conference on Computer Vision*, 2020. 2

[56] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[57] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Nguyen. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020. 3

[58] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6, 13

[59] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 4, 5

[60] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[61] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *European Conference on Computer Vision*, 2018. 2

[62] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3D human pose estimation. In *International Conference on Computer Vision*, 2019. 3

[63] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *International Conference on Computer Vision*, 2021. 2, 3, 5, 6, 8

[64] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *European Conference on Computer Vision*, 2020. 2, 3, 4, 5

[65] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *International Conference on Computer Vision Workshops*, 2021. 2

[66] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning interaction snapshots from observations. *ACM Transactions on Graphics*, 35(4):139:1–139:12, 2016. 3

[67] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 39(6):235:1–235:16, 2020. 2, 3, 4, 5

[68] Leonid Sigal, Alexandru Balan, and Michael J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing Systems*, 2008. 3

[69] Leonid Sigal, Alexandru Balan, and Michael 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. 1

[70] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[71] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *International Conference on Computer Vision*, 2021. 2

[72] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[73] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, 2020. 3

[74] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human shape and pose prediction. In *British Machine Vision Conference*, 2017. 2

[75] Yating Tian, Hongwen Zhang, Yebin Liu, and limin Wang. Recovering 3D human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 2

[76] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. VoxelPose: Towards multi-camera 3D human pose estimation in wild environment. In *European Conference on Computer Vision*, 2020. 3

[77] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, 2017. 2

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[79] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):1–9, 2008. 3

[80] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision*, 2018. 1, 2, 6, 7, 13

[81] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *International Conference on Computer Vision*, 2019. 2

[82] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3D human pose with scene constraints. *arXiv preprint arXiv:1905.07718*, 2019. 3

[83] Zhenzhen Weng and Serena Yeung. Holistic 3D human and scene mesh estimation from single view images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[84] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[85] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*, 2018. 1

[86] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision*, 2021. 2, 3, 4

[87] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *British Machine Vision Conference*, 2018. 4

[88] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[89] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[90] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2022. 3

[91] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 40(6):208, 2021. 4

[92] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[93] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes – the importance of multiple scene constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[94] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3D human mesh regression with dense correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[95] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 13

[96] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision*, 2020. 3

[97] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4D human body capture in 3D scenes. In *International Conference on Computer Vision*, 2021. 2, 3, 5

[98] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision*, 2020. 3

[99] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[100] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D association graph for realtime multi-person motion capture using multiple video cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3, 4

[101] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[102] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Light-weight multi-person total capture using sparse multi-view cameras. In *International Conference on Computer Vision*, 2021. 3

[103] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *arXiv:2012.13392*, 2022. 2

[104] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[105] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *Winter Conference on Applications of Computer Vision*, 2020. 2, 3, 6, 8

Supplementary Material

A. SMPL-X vs. SMPL HSC labels

We build RICH by fitting a SMPL-X template to multi-view data and compute the human-scene contact (HSC) as explained in the Sec. 3 and Sec. 5 of the main paper (Fig. 7). The contact labels are defined in SMPL-X format and we map them to SMPL format for training BSTRO. This is feasible since there is an 1-to-1 correspondence between SMPL-X and SMPL vertices below the neck, as shown in Fig. 8.

With this mapping, we convert the ground-truth HSC labels from SMPL-X to SMPL without losing information. As a result, we benefit from realistic hand articulation in SMPL-X and still keep the dimension of the output space small (SMPL). Such a mapping also makes RICH a suitable HSC benchmark for both body models. Since the two models share the set of vertices of interest, choosing either of them does not influence the detection scores or errors.

On the other hand, the human pose and shape (HPS) parameters of the two models differ. Converting HPS parameters between SMPL-X and SMPL requires extra processing [1] and one always loses the hand articulation when converting from SMPL-X to SMPL. Therefore, RICH provides only SMPL-X as pseudo ground truth. To evaluate methods that regress SMPL parameters using RICH, users should convert SMPL to SMPL-X, which does not result in a loss of information.

B. RICH Dataset

The 142 multi-view videos in RICH are recorded at a rate of 30 frames per second. We separate them into subsets of 62, 28, 52 for training, validation, and testing purposes, respectively. This amounts to 303K, 149K, 125K images of 4K resolutions (in total 577K), and 40K, 18K, 32K 3D SMPL-X bodies along with dense scene-contact labels (in total 577K) in each subset. By "body" here, we mean any SMPL-X mesh. Note that the number of unique "people" in the dataset is much smaller than the number of bodies because every posed mesh constitutes a separate body.

Compared to the recent HPS dataset AGORA [58], RICH has more 3D bodies (90K vs. 4K), more images (577K vs. 19K) and more accurate body shapes (registrations to minimally-closed scans [31] vs. clothed scans [95]). It has more subjects in varied body shapes than 3DPW [80] (22 vs. 18) and subjects are in natural clothing as opposed to those in Human3.6M [32]. Last but not least, RICH provides high-quality scene scans and scene contact labels that none of the above datasets provides.

C. Bone-orientation Term E_O

Following the illustration in Fig. 2(a) of the main paper, the bone-orientation term E_O factors out the residual of the parent joint ϵ_1 from the residual of the child joint ϵ_2:

$$\begin{aligned} r_2 &= \epsilon_2 - \epsilon_1, \\ &= (j_2' - j_2) - (j_1' - j_1), \\ &= (j_2' - j_1') - (j_2 - j_1), \\ &= b_2' - b_2, \end{aligned}$$

where $b_2' = j_2' - j_1'$ and $b_2 = j_2 - j_1$ denote the "bone vector" of target points (detected landmarks) and estimated SMPL-X joints respectively. It follows that

$$\|r_2\|_2^2 = \|b_2'\|_2^2 + \|b_2\|_2^2 - b_2^\top b_2'. \tag{2}$$

Since b_2' involves only the detected landmarks and $\|b_2\|$ is fixed given a constant body shape β, the first two terms are constant when optimizing the multi-view objective E_{mv}. $\|r_2\|_2^2$ is therefore minimized when $b_2^\top b_2'$ is maximized, i.e., when b_2 has the same orientation as b_2'.

D. BSTRO Implementation Details

We sample RICH-train to build the image-HSC pairs $(I, \mathbf{c})$ for training BSTRO. For each sequence, we consider only every other frame, and for each frame, we use the dynamic view and one randomly selected static view, or two static views if no moving camera is available. This sampling strategy ensures sufficient variations in viewpoints and background, while keeping the total number of the training pairs tractable.

We train with in total 24K $(I, \mathbf{c})$ pairs from RICH-train and use Adam [40] optimizer with an initial learning rate of 1e-4 for 100 epochs. The HR-Net backbone is initialized with the weights pre-trained on ImageNet [12], Human3.6M [32] or 3DPW [80]. The best checkpoint is selected by the best performance on RICH-validation with RICH-test completely withheld. We refer interested readers to [47] for the architecture of the multi-layer transformer.
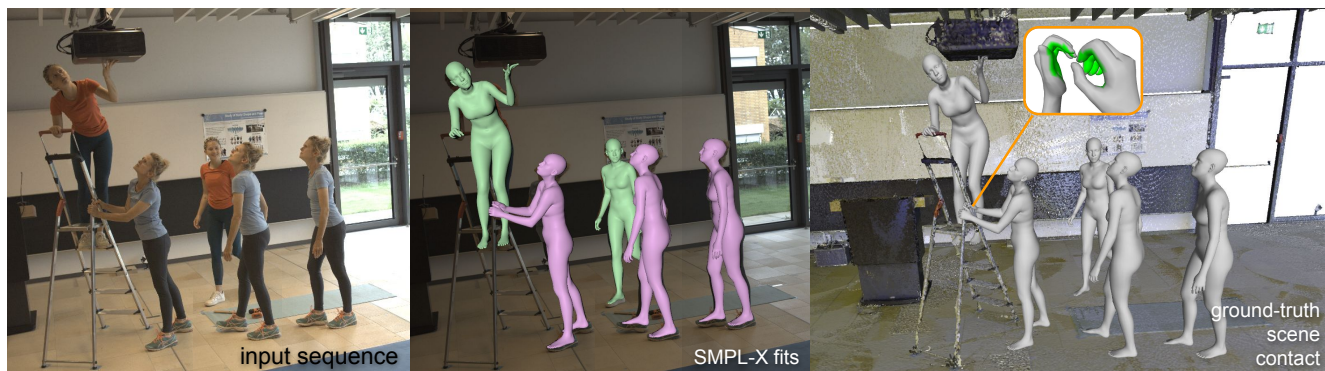
Figure 7. The **RICH dataset** contains multiple people interacting with a real scene. It provides complex natural images, precise 3D scene scans, pseudo ground-truth SMPL-X bodies, and dense body contact labels.
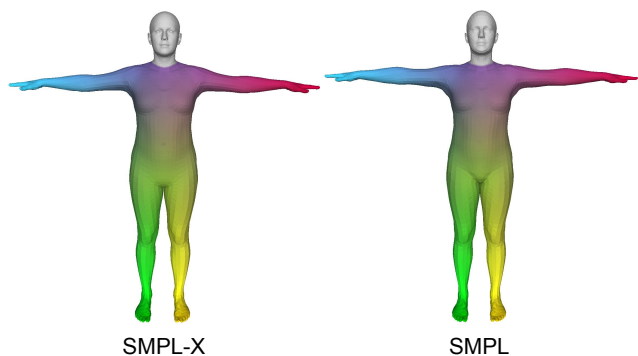

SMPL-X   SMPL

Figure 8. SMPL-X and SMPL bodies share the same set of vertices for regions below the neck. The same vertices are visualized in the same colors.