Capturing and Inferring Dense Full-Body Human-Scene Contact

Chun-Hao P. Huang1  Hongwei Yi1  Markus Höschle1  Matvey Safroshkin1  Tsvetelina Alexiadis1
Senya Polikovsky1  Daniel Scharstein2  Michael J. Black1
1Max Planck Institute for Intelligent Systems, Tübingen, Germany 2Middlebury College
{paul.huang, firstname.lastname, black}@tuebingen.mpg.de, schar@middlebury.edu

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 [7, 28, 32, 42, 67, 82] and 3D human pose and shape estimation (HPS) from a single image [5, 36, 39, 40, 44, 56, 81, 95], thanks to realistic datasets annotated with 2D keypoints [1, 33, 46] and 3D data [30, 34, 49, 66, 77]. 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 [61, 64, 83, 89, 90, 93, 101], or assume the ground is a even plane [60], which is often violated, e.g., walking up stairs. To infer contact, many state-of-the-art (SOTA) methods use MoCap datasets [48, 50] to train a contact detector [61, 93, 101]. Others exploit physics simulation [64, 89] or physics-inspired objectives [83] 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 [37, 58, 78, 85, 100]. 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 [56] 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 [25], 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. 1c.

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 [75], which has been successfully employed for natural-language processing [11] and HPS estimation with occlusion [44].

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 [24, 25], 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 [26], 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 [30, 55, 77]. We confirm that the performance of a SOTA HPS method [17] 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 [2, 35, 47, 56, 86]. SMPLify [5] fits the SMPL model to the output of a 2D keypoint detector [57] and we build on it here.

In contrast, deep neural networks regress body-model parameters directly from pixels [10, 16, 17, 23, 36, 38–40, 62, 68, 69]. To deal with the lack of in-the-wild 3D ground truth, some methods use 2D keypoints [36, 71, 74] or linguistic attributes [9] as weak supervision, while some directly fine-tune the network w.r.t. an input image at test time [34]. Kolotouros et al. [40] combine HMR [36] and SMPLify [5] 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 [8, 12, 41, 44, 45, 52, 91]. We refer readers to [72, 99] for a comprehensive review. None of the above methods estimate HSC.
Table 1. Comparison of contact-related methods and datasets.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Scene Contact</th>
<th>Body</th>
<th>Contact Cues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Body / Scene</td>
<td></td>
<td>Train / Test</td>
</tr>
<tr>
<td>Zantir et al. [90]</td>
<td>foot joints / ground</td>
<td>mesh</td>
<td>- / dist.</td>
</tr>
<tr>
<td>Zou et al. [101]</td>
<td>joint</td>
<td>joint</td>
<td>both 2D vel.</td>
</tr>
<tr>
<td>Rempke et al. [61]</td>
<td>part</td>
<td>part</td>
<td>vel. &amp; dist. / vel. &amp; dist.</td>
</tr>
<tr>
<td>Physit-cap [65]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HoMoR [60]</td>
<td>8 joints / ground</td>
<td>mesh</td>
<td></td>
</tr>
<tr>
<td>LEMO [93]</td>
<td>foot vert. / ground</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimPoE [89]</td>
<td>foot parts / ground</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xie et al. [83]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HolisticMesh [80]</td>
<td>dense body mesh / scene mesh</td>
<td>mesh</td>
<td>- / dist.</td>
</tr>
<tr>
<td>PHOSA [92]</td>
<td></td>
<td></td>
<td>- / dist.</td>
</tr>
<tr>
<td>Zhang et al. [97]</td>
<td></td>
<td></td>
<td>- / dist.</td>
</tr>
<tr>
<td>PLACE [74]</td>
<td></td>
<td></td>
<td>dist. / -</td>
</tr>
<tr>
<td>POSA [26]</td>
<td></td>
<td></td>
<td>dist. / pose</td>
</tr>
<tr>
<td>BSTRO (ours)</td>
<td>as above</td>
<td>mesh</td>
<td>dist. / image</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Contact Label</th>
<th>Img</th>
<th>Scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTP [53]</td>
<td>self-contact</td>
<td>✓</td>
<td>N/A</td>
</tr>
<tr>
<td>GRAB [70]</td>
<td>hand-object</td>
<td>✓</td>
<td>N/A</td>
</tr>
<tr>
<td>ContactHands [54]</td>
<td>hand-X</td>
<td>✓</td>
<td>N/A</td>
</tr>
<tr>
<td>Fieraru et al. [18]</td>
<td>person-person</td>
<td>✓</td>
<td>N/A</td>
</tr>
<tr>
<td>Fieraru et al. [19]</td>
<td>self-contact</td>
<td>✓</td>
<td>N/A</td>
</tr>
<tr>
<td>PiGraph [63]</td>
<td>joint-scene</td>
<td>✓</td>
<td>RGBD scans</td>
</tr>
<tr>
<td>i3DB [51]</td>
<td>N/A</td>
<td>✓</td>
<td>CAD</td>
</tr>
<tr>
<td>GPA [79]</td>
<td>N/A</td>
<td>✓</td>
<td>Cubes</td>
</tr>
<tr>
<td>Guzov et al. [24]</td>
<td>foot-ground</td>
<td>✓</td>
<td>laser scans</td>
</tr>
<tr>
<td>PROX [25]</td>
<td>body-scene</td>
<td>✓</td>
<td>RGBD scans</td>
</tr>
<tr>
<td>RICH (ours)</td>
<td>body-scene</td>
<td>✓</td>
<td>laser scans</td>
</tr>
</tbody>
</table>

Table 1. Comparison of contact-related methods and datasets.

xiv: 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 [4,22,65] or top-down [3,20,76], 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 [27,31,59,73], to re-identify subjects across views [14] or across view and time [13,96], but they estimate only joints, not body meshes. Dong et al. [15] reconstruct SMPL bodies for multiple subjects and Zhang et al. [98] 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 interaction in them. Most methods represent humans and objects very roughly as bounding boxes [37,100]; only a few use body meshes for humans and spheres for objects [43].

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 [6] or placing virtual humans into a scene [26]. 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 [51,63] but represents a human body roughly as a stick figure. Recent HPS methods [25,60,61,101] 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 [24,25,53,70]. At test time, PROX [25] assumes scene scans to be known a priori; PHOSA [92] 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 [87], 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 [60,61,64,101]. Some of these detect contact with a separate neural network at test time, taking the 2D/3D joints in a temporal window as input [61,64,101], while others integrate it in a body motion prior [60,93]. These approaches use MoCap datasets such as AMASS [48] and Mixamo [50] to build training data, where contact is automatically labelled via thresholding the distance to the ground and/or the velocity.

POSA [26] observes that scene contact is correlated with body poses and introduces a generative model to sample contact given a posed mesh. Some methods [64,83,89] 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 [87] 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 [19,53] or person-person contact [18], but not HSC. The most relevant datasets for HSC are [24] and PROX [25]. The former provides egocentric images for localization, which are not suitable for HSC detection from images. PROX [25] 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 [61, 64, 83, 88], which represent a body as a set of coarse geometry primitives, we follow [25, 26] to capture realistic human-scene contact with a parametric SMPL-X body model [56]. The vertex locations on a SMPL-X mesh $M(\theta, \beta, \psi) \subset \mathbb{R}^3$ are controlled by parameters for pose $\theta$, shape $\beta$, and facial expression $\psi$. $\theta$ consists of body pose $\theta_b$ and hand pose $\theta_h$. Hand pose $\theta_h$ is a function $\theta_h(Z_b)$ of a PCA latent vector $Z_b \in \mathbb{R}^{12}$.

Given videos captured by $C$ synchronized cameras, we first identify each subject across views and across time with [14, 84]. 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 540K 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 [84], followed by MvPose [14] to match the tracklets across views. Other methods that build such 4D associations [13, 96] 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 [56] to accommodate multiview data. SMPLify-X optimizes the pose $\theta$, shape $\beta$ and facial expression $\psi$ of SMPL-X to match the observed 2D keypoints [7] by minimizing the following objective:

$$
E(\theta, \beta, \psi) = E_J + E_{\text{reg}}
$$

$$
E_{\text{reg}} = \lambda_{\theta_b} E_{\theta_b} + \lambda_{\alpha} E_{\alpha} + \lambda_{\beta} E_{\beta} + \lambda_{C} E_{C} + \lambda_{\psi} E_{\psi},
$$

where $E_J$ is the data term, and $E_{\text{reg}}$ includes several regularization terms: $\theta_b$ is the pose vector for the body, which is a function $\theta_b(Z_b)$, where $Z_b \in \mathbb{R}^{12}$ is a VAE latent representation and $E_{\theta_b}$ 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. $\lambda$’s denote weights for each respective term. Interested readers are referred to [56] for details.

Multiview per-person reconstruction. For each person, we compute 2D keypoints [7] in each camera $c$. Instead of fitting them using SMPLify-X in each view, we combine all 2D landmarks in a multiview energy term: $\sum_c E^c J$. Unlike in [56], where one needs to estimate camera translation first, the perspective projection here is well defined by the pre-calibrated intrinsics and extrinsics. To pursue high-quality fits, body shape $\beta$ is estimated in advance by registering a SMPL-X template to minimally-clothed 3D scans following [29]. $\beta$ 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 $E_{\psi}$ 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 $\epsilon_1$ is accumulated in the error of child joints $\epsilon_2$. When $\lVert \epsilon_2 \rVert$ gets too large, the influence is downweighted because our robust loss treats it as an outlier. Instead, $E_{\psi}$ factors out the errors of ancestors and focuses on the error of the joint per se. Our final objective is $E_{\text{mv}}(\theta, \psi) = \sum_c E^c J + \sum_c E^c \psi + E_{\text{reg}}$.

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 $\sum_c E^c J$ 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 [39]) to initialize $\theta$. 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 \( \theta \) for minimizing \( E_{mv} \). We first solve \( E_{mv} \) for each time step \( t \) independently and then refine a batch of \( T \) frames jointly with a motion smoothness term \( E_{smooth} \) [93]:
\[
E_{batch}(\theta_1, \ldots, \theta_T) = \sum_{t=1}^{T} E_{mv}^{t} + \lambda_{smooth} E_{smooth}^{t}.
\]

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 [26]. 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 [25]. 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: Bestro

Here we introduce Bestro 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 [56] or a neural network as in [36, 39]. 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 \( c = g(M) \). Each element in \( c \) 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 [26], taking the vertex locations of \( M \) as input, while in [60, 61, 64], \( g \) is a MLP operating on the motion of \( M \).

With this formulation, the body-scene contact \( c \), whether defined on a dense mesh or on a set of sparse joints/parts, is a composite function of \( g \) and \( f \): \( c = g \circ f(I) \), where \( g \) is agnostic to the input image. In contrast, our goal is to detect dense body-scene contact directly from the input \( I \): \( c = g(I) \). To our knowledge, this was explored only for self-contact [19] and person-person contact [18] and only at a coarse region level, not the vertex level.

We use SMPL as the body representation for Bestro, hence \( c \in \{0, 1\}^V \), 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 \((I, c)\) 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 [11] 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 \( \theta \): \( Q = \{q_v\}_{v=1}^{V} \). 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
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 134 single or multi-person multiview videos, with a total of 85K posed 3D body meshes, together with 85K dense full-body contact labels in both SMPL-X and SMPL mesh topology, and 540K 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 [77], more accurate body shapes than AGORA [55], and real human-scene interaction unlike Human3.6M [30]. 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 134 multiview videos in RICH into 57, 27, 50 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. [101] use the velocity of 4 2D keypoints on the feet to predict contact; HuMoR [60] 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 [26] 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 [17], 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 scene contact that is closer to the ground truth, whereas POSA$^\text{PIXIE}$ yields false positives frequently (red circles) and
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 [26] (bottom). The noisy body fits in PROX [25] result in undesirable HSC labels in POSA: (a) no foot-ground contact under occlusion; (b) severe penetration with chairs.

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 [48, 50] and HPS datasets [30, 34, 77] 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 scenes and similar HSI with training data but the subjects are new. Intuitively, this is an easy subset and indeed the scores are best in this scenario. Once HSI is withheld, the performance drops (row (b)). This drop is more pronounced than the drop caused by withholding a scene (row (c)). Comparing each of the rows (b,c,d) to row (e), 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 (e) 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 (b) 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 [77] or Human3.6M [30], the real scene contact in RICH enables a new way of analyzing the performance of an HPS method. In particular, we use PIXIE [17], 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
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.

Acknowledgments. We thank Taylor McConnell, Claudia Gallatz, Mustafa Ekinci, Camilo Mendoza, Galina Henz, Tobias Bauch, and Mason Landry for the data collection and cleaning. We thank all participants who contributed to the dataset and Benjamin Pellkofer for IT support. Daniel Scharstein was supported by NSF grant IIS-1718376.

Disclosure: https://files.is.tue.mpg.de/black/Col_CVPR_2022.txt
References


Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J.


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


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

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


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


13284


