This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DECO: Dense Estimation of 3D Human-Scene Contact In The Wild

Shashank Tripathi[1*] Agniv Chatterjee[1*] Jean-Claude Passy[1] Hongwei Yi[1]
Dimitrios Tzionas[2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Amsterdam, the Netherlands
{stripathi, achatterjee, jpassy, hyi, black}@tue.mpg.de d.tzionas@uva.nl

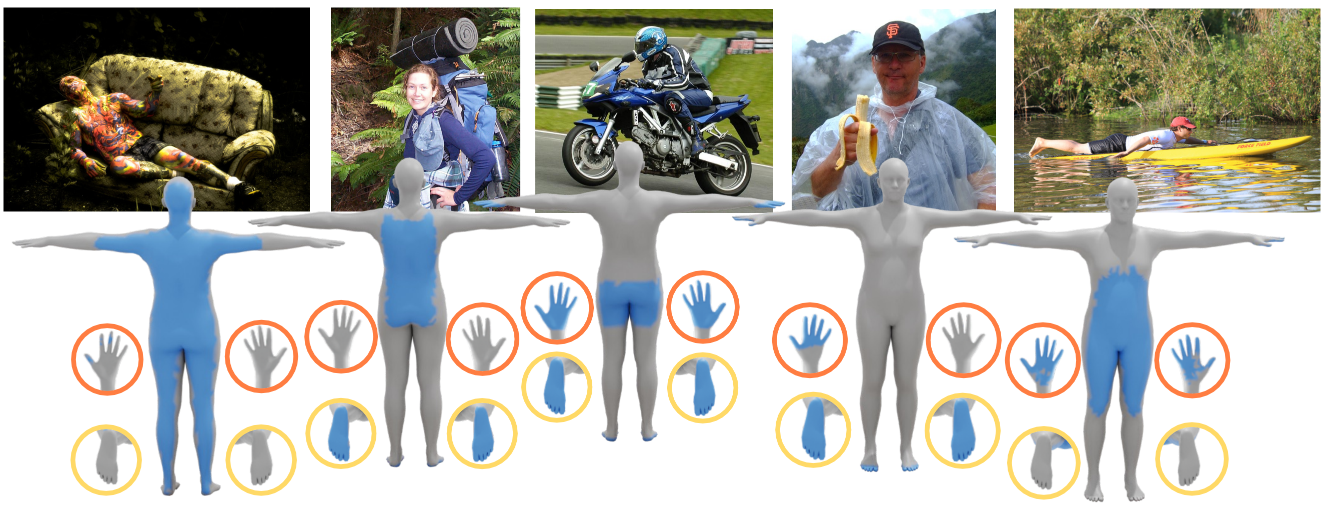

Figure 1: Given an RGB image, DECO infers dense vertex-level 3D contacts on the full human body. To this end, it reasons about the contacting body parts, human-object proximity, and the surrounding scene context to infer 3D contact for diverse human-object and human-scene interactions. **Blue areas** show the inferred contact on the body, hands, and feet for each image.

Abstract

Understanding how humans use physical contact to interact with the world is key to enabling human-centric artificial intelligence. While inferring 3D contact is crucial for modeling realistic and physically-plausible human-object interactions, existing methods either focus on 2D, consider body joints rather than the surface, use coarse 3D body regions, or do not generalize to in-the-wild images. In contrast, we focus on inferring dense, 3D contact between the full body surface and objects in arbitrary images. To achieve this, we first collect DAMON, a new dataset containing dense vertex-level contact annotations paired with RGB images containing complex human-object and human-scene contact. Second, we train DECO, a novel 3D contact detector that uses both body-part-driven and scene-context-driven attention to estimate vertex-level contact on the SMPL body. DECO builds on the insight that human observers recognize contact by reasoning about the contacting body parts, their proximity to scene objects, and the surrounding scene context. We perform extensive evaluations of our detector on DAMON as well as on the RICH and BEHAVE datasets. We significantly outperform existing SOTA methods across all benchmarks. We also show qualitatively that DECO generalizes well to diverse and challenging real-world human interactions in natural images. The code, data, and models are available at https://deco.is.tue.mpg.de.

1. Introduction

Humans rely on contact to interact with the world. While we use our hands and feet to support grasping and locomotion, we also leverage our entire body surface in our daily interactions with the world; see Fig. 1. We sit on our buttocks and thighs, lie on our backs, kneel on our knees, carry bags on our shoulders, and move heavy objects by holding them against our bodies. Executing everyday tasks involves diverse full-body and object contact. Thus, modeling and inferring contact from images or videos is essential for applications such as human activity understanding, robotics, biomechanics, and augmented or virtual reality.

* Equal contribution

Inferring contact from images has recently received attention. While some methods infer contact for hands [48], feet [51], self contact [15, 47], or person-person contact [14], others focus on human-scene or human-object contact for the full body [8, 28]. HOT [8] infers contact in 2D by training on in-the-wild images with crowd-sourced 2D contact areas, while BSTRO [28] infers 3D contact on a body mesh and is trained on images paired with 3D body and scene meshes reconstructed with a multi-camera system.

In contrast to prior work, we seek to represent detailed scene contacts across the full body and to infer these from in-the-wild images as illustrated in Fig. 1. To that end, we need both an appropriate training dataset and an inference method. Note that manipulating objects is fundamentally 3D. Thus, we must capture, model, and understand contact in 3D. Also note that some contacts support the body, while others do not. When sitting on a chair and drinking a cup of coffee, the body is supported by the buttocks on the chair and feet on the floor, while the coffee cup does not support the body. The former is critical for physical reasoning about human pose and motion, while the latter is important to understand how we interact with objects. The *type* of contact is therefore important to represent. For a method to robustly estimate contact for arbitrary images we need a rich dataset that combines in-the-wild images with precise 3D annotations; see Fig. 2. This is a huge challenge.

To address this challenge, we present a novel method and a new dataset. We first collect a dataset with 3D contact annotations for in-the-wild images using a novel interactive 3D labelling tool (Fig. 2). We then train a novel 3D contact detector that takes a single image as input and produces dense contact labels on a 3D body mesh (Fig. 1). Training on our new dataset means that the method generalizes well.

Contact data: To train a 3D contact detector that is both accurate and robust, we need appropriate training data. However, existing datasets for 3D contact [3, 24, 28] involve pre-scanning a 3D scene and estimating 3D human pose and shape (HPS) of people in the scene. These approaches are limited in the complexity of the human-scene interactions, the size of the dataset, and very few methods capture human-object interactions paired with image data [4, 29]. An alternative is to use synthetic data [59], but getting realistic synthetic data of complex human contacts is challenging, causing a domain gap between the dataset and real images.

In contrast, crowdsourced image annotations support many tasks in computer vision such as image classification [12], object detection [41, 72], semantic segmentation [27, 41], 2D human pose estimation [1, 6], and 3D body shape estimation [10, 61]. HOT [8] takes this approach for human-object contact, but the labels are all in 2D, while contact is fundamentally 3D. Consequently, we collect a large dataset with dense 3D contact annotations for in-the-wild images, called DAMON (**D**ense **A**nnotation of 3D hu**M**an **O**bject contact in **N**atural images). We enable this with a new interactive software tool that lets people "paint" contact areas on a 3D body mesh such that these reflect the observed contact in images. We use Amazon Mechanical Turk, train human annotators for our task, and collect a rich corpus of 3D contact annotations for standard datasets of in-the-wild images of diverse human-object interactions, i.e., V-COCO [22] and HAKE [37]; Fig. 2 shows samples of our dataset. Note how contact and support regions are distinguished as are the semantic labels related to object contact.

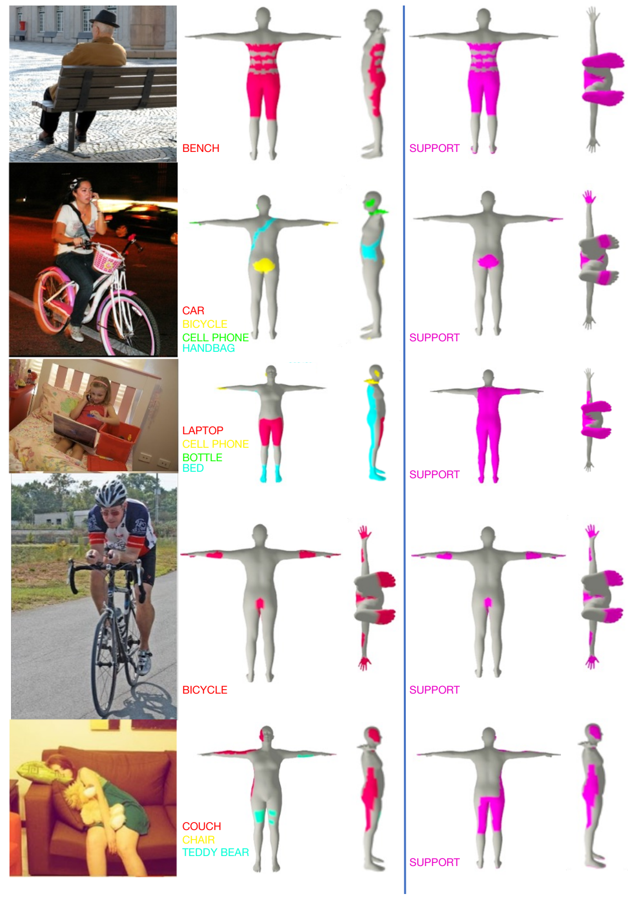

Figure 2: Sample contact annotations from the DAMON dataset. **Left to Right:** RGB image, two views showing human-supported contact (color-coded by object labels), and two views showing scene-supported contact.

Contact detection: As noted in the literature [8, 28], contact areas are ipso facto occluded in images, thus, detecting contact requires reasoning about the involved body-parts and scene elements. To this end, BSTRO [28] uses a transformer [39] with positional encoding based on body-vertex positions to implicitly learn the context around these, but has no explicit attention over body or scene parts. HOT [8, 28], on the other hand, focuses only on 2D, pulls image features, and processes them with two branches in parallel, a contact branch and a body-part attention branch; the latter helps the contact features attend areas on and around body parts.

We go beyond prior work to estimate detailed 3D contact on the body. Our method, DECO (**D**ense **E**stimation of 3D human-scene **CO**ntact in the wild), introduces two technical novelties: (1) DECO uses not only *body-part-driven attention*, but also adds *scene-context-driven attention*, as well as a *cross-attention* module; this explicitly encourages contact features computed from the image to attend to meaningful areas both on (and near) body parts and scene elements. (2) DECO uses a new 2D Pixel Anchoring Loss (PAL) that relates the inferred 3D contacts to the respective image pixels. For this, we infer a 3D body mesh with CLIFF [38] (SOTA for HPS), detect which vertices of this are in contact with DECO, project the 3D contact vertices onto the image, and encourage them to lie in HOT's corresponding 2D contact-area annotations. Note that this brings together both crowd-sourced 2D and 3D contact annotations.

Experiments: We perform detailed quantitative experiments and find that DECO outperforms BSTRO on the test sets of RICH and DAMON, when both are trained on the same data. Ablation studies show that our two-branch architecture effectively combines body part and scene information. We also provide ablation studies of the backbone and training data. We show that the inferred contact from DECO significantly outperforms methods that compute the geometric vertex distance between a reconstructed object and human mesh [73, 81]. Finally, we use DECO's estimated contact in the task of 3D human pose and shape estimation and find that exploiting estimated contact improves accuracy.

Contributions: In summary, our contributions are (1) We collect DAMON, a large-scale dataset with dense vertex-level 3D contact annotations for in-the-wild images of human-object interactions. (2) Using DAMON, we train DECO, a novel regressor that cross-attends to both body parts and scene elements to predict 3D contact on a body. DECO outperforms existing contact detectors, and all its components contribute to performance. This shows that learning 3D contact estimation from natural images is possible. (3) We integrate DECO's inferred 3D contacts into a 3D HPS method and show that this boosts accuracy. (4) Our data, models, and code are available at https://deco.is.tue.mpg.de.

2. Related Work

2.1. 2D contact in images

There exist multiple ways of representing human-object interactions (HOI) and human-scene interactions (HSI) in 2D. Several HOI methods [33, 49, 69, 75, 86] localize humans and objects as bounding boxes and assign a semantic label to indicate the *interactions* between them. However, the interaction labels focus on action and do not support contact inference. Chen et al. [8] output image-aligned contact heatmaps and body-part labels directly from the RGB image by training a regressor on approximate 2D polygon-level contact annotations. Some approaches learn part-specific contact regressors for hand [48, 57] and foot [52] contact but only detect rough bounding boxes around contacting regions or joint-level labels. Such coarse image-based contact annotations are ambiguous and not sufficient for many downstream tasks. We address these limitations by collecting a large-scale dataset of paired images and accurate vertex-level contact annotations directly on the 3D SMPL mesh.

Several methods estimate properties related to contact such as affordances [36, 54, 70], contact forces [60, 78, 85] and pressure [17, 20, 56]. However, collecting large datasets with ground-truth object affordances, forces, or pressure is challenging. Clever et al. [11] use simulation and a virtual pressure mat to generate synthetic pressure data for lying poses. Tripathi et al. [66] exploit interpenetration of the body mesh with the ground plane as a heuristic for pressure. Recent work [18, 60, 78] uses a physics simulator to infer contact forces. In contrast, we focus on annotating and estimating 3D contact, which is universal in HOI and is intuitively understood by annotators.

2.2. Joint- & patch-level 3D contact

Joint-level contact. 3D contact information is useful for 3D human pose estimation [52, 60, 73], 3D hand pose estimation [7, 21, 26], 3D body motion generation [51, 63, 82–84] and 3D scene layout estimation [77]. 3D pose estimation approaches use joint-level contact to *ground* the estimated 3D human mesh [16, 24, 76, 79, 81] or encourage realistic foot-ground contact to avoid foot-skating artefacts [30, 51, 58, 82, 87]. PhysCap [60] and others [51, 52, 79, 87] constrain the human pose by predicting skeleton joint-level foot-ground contact from video. Several approaches predict 3D contact states of 2D foot joints detected from RGB images by manually annotating contact labels [87] or computing contact labels from MoCap datasets [52, 60]. Rempe et al. [51] extend joint-level contact estimation to the toe, heel, knee and hands, but use heuristics such as a zero-velocity constraint to estimate contact from AMASS [45]. Zhang et al. [82] estimate contact between foot-ground vertices using alignment of normals between foot and scene surface points. Such joint-level annotations cannot represent the richness of how human bodies contact the world. In contrast DECO captures dense vertex-level contact across the full body.

Discrete patch-level contact. Pre-defined contact regions or "patches" on the 3D body provide an intermediate representation for modeling surface-level contact. Müller et al. [47] and Fieraru et al. [15] crowdsource patch-level self-contact annotations between discrete body-parts patches on the same individual. Fieraru et al. [14] also collect patch-level contact between two interacting people.

While richer than joint-level contact, *patches* do not model fine-grained contact. In contrast, the DAMON dataset and DECO model contact on the vertex level, significantly increasing the contact resolution.

2.3. Dense vertex-level contact

Dense ground-truth contact can be computed if one has accurate 3D bodies in 3D scenes. For instance, PROX [24], InterCap [29], and BEHAVE [3] use RGB-D cameras to capture humans interacting with objects and scenes whereas HPS [23] uses a head-mounted camera and IMU data to localize a person in a pre-scanned 3D scene. RICH uses a laser scanner to capture high-quality 3D scenes and the bodies are reconstructed using multi-view cameras. GRAB [64] captures hand-object interactions using marker-based MoCap but lacks images paired with the ground-truth scene. Such datasets require a constrained capture setup and are difficult to scale. An alternative uses synthetic 3D data. HULC [59] generates contact by fitting SMPL to 3D joint trajectories in the GTA-IM [5] dataset. The contacts, however, lack detail and the domain gap between the video game and the real world limits generalization to natural images.

Several methods infer 3D bodies using dense 3D contact. PHOSA [81] jointly estimates 3D humans, objects and contacts for a limited set of objects for which there are predetermined, hand-crafted, contact pairs on the human and object. Other methods optimize the body and scene together using information about body-scene contact [55, 71, 73, 74, 77].

Some methods predict dense contact on the body mesh. POSA [25] learns a body-centric prior over contact. Given a posed 3D body, POSA predicts which vertices are likely to contact the world and what they are likely to contact. It assumes the pose is given. Closest to our work are BSTRO [28] and HULC [59], which infer dense contact on the body from an image. We go beyond these methods by providing a rich dataset of images in the wild with dense contact labels. Moreover we exploit contextual cues from body parts as well as the scene and objects using a novel attentional architecture.

3. DAMON Dataset

DAMON is a collection of *vertex-level* 3D contact labels on SMPL paired with color images of people in unconstrained environments with a wide diversity of human-scene and human-object interactions. We source our images from the HOT dataset [8] for the following reasons: (1) HOT curates valid human contact images from existing HOI datasets like V-COCO [22] and HAKE [37] by removing indirect human-object interactions, heavily cropped humans, motion blur, distortion or extreme lighting conditions; (2) HOT contains 15082 images containing 2D *image-level* contact annotations, which are complementary to the dense 3D contact annotations in our dataset. Example images and contact annotations from the DAMON dataset are shown in Fig. 2.

3.1. Types of contact

While existing HOI methods and datasets typically treat all contacts the same way, human contact is more nuanced. Physical contact can be classified into 3 categories: (1) *scene-supported contact*, i.e., humans supported by scene objects; (2) *human-supported contact*, i.e., objects supported by a human; and (3) *unsupported* contact, e.g., self-contact [15, 47] and human-human contact [14, 16]. Since datasets for the latter already exist, we focus on the first two categories, i.e., contact that involves support. Note that labeling contact in images is challenging. Focusing on support helps reduce ambiguous cases where humans are close to scene objects but not actually in contact. We use Amazon Mechanical Turk (AMT) to crowd-source annotations for DAMON; we ask people to annotate both *human-supported contact* for each individual object and *scene-supported contact*.

3.2. Annotation procedure

We create a novel user-friendly interface and tool that enables annotators to *"paint"* 3D vertex-level contact areas directly on the human mesh; see the interface in Sup. Mat. We show the original image with the type of contact to be annotated on the left and the human mesh to the right. We then ask annotators to "paint" contact labels on the $N_V = 6890$ vertices of the SMPL [43] template mesh, $\bar{\mathcal{M}} \in \mathbb{R}^{6890\times 3}$.

The tool has features such as mesh rotation, zoom in/out, paint-brush size selection, an eraser, and a reset button. Depending on the selected brush size, the tool "paints" contact annotations by selecting a *geodesic* neighborhood of vertices around the vertex currently under the mouse pointer. For a detailed description of the tool, see **video** in Sup. Mat.

The tool lets annotators label contact with multiple objects in addition to the scene-supported contact. For example annotations, see Fig. 2. For every image, to label human-supported contact, we cycle through object labels provided in the V-COCO and HAKE datasets. For scene-supported contact, we ask annotators to label contact with all supporting scene objects, including the ground. We automatically get body-part labels for contact vertices using SMPL's part segmentation. To support amodal contact estimation, we ask annotators to also label contact regions that may not be visible in the image but can be guessed confidently. We filter out ambiguous contact in images such as human-human contact, human-animal contact, and indirect human-object interactions, such as pointing; for details about data collection and how we limit ambiguity in the task, see Sup. Mat.

We ensure a high annotation quality with two quality checks: (1) We detect and filter out the inconsistent annotators; out of 100 annotators we keep only 14 good ones. (2) We have meta-annotators curate the collected annotations; images with noisy annotations are then pushed for a re-annotation. For details about quality control, see Sup. Mat.

We access DAMON's quality by computing two metrics:

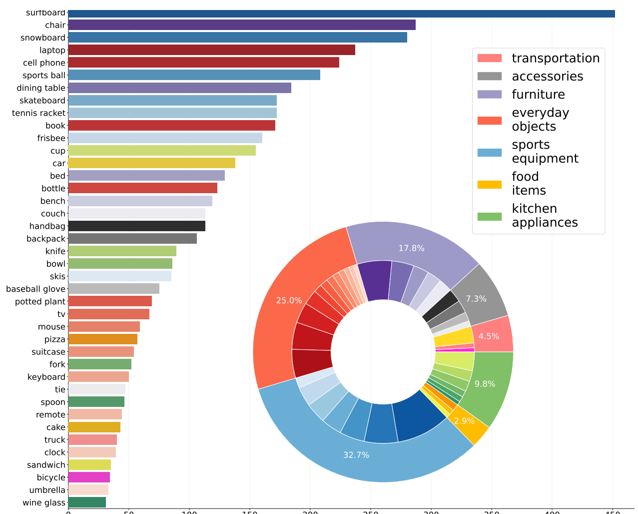

Figure 3: DAMON dataset statistics. **Histogram**: contact object labels (y-axis) and the number of images in which they are present (x-axis). We crop the plot in the interest of space; for the full long-tailed plot see Sup. Mat. **Pie chart**: object labels are grouped into 7 main categories; inner colors correspond to the colors in the histogram. 🔍 **Zoom in**.

(1) *Label accuracy:* We manually curate from RICH [28] and PROX [24] 100 images that have highly-accurate 3D poses and contact labels. We treat these as ground-truth contact, and compute the IoU of our collected annotations. (2) *Level of annotators' agreement:* We ask annotators to label the same set of 100 images, and compute *Fleiss' Kappa (κ)* metric. For a detailed analysis of results, see Sup. Mat.

3.3. Dataset statistics

Out of HOT's 15082 images we annotate 5522 images via our annotation tool (Sec. 3.2); we "paint" contact vertices, and assign to each vertex an appropriate label out of 84 object (Fig. 3) and 24 body-part labels. An image has on average 3D contacts for 1.5 object labels. We use HOT's train/test/val data splits.

We also show aggregate vertex-level contact probabilities on the SMPL mesh across the whole DAMON dataset in Fig. 4. The individual body-part close-ups in Fig. 4 show normalized contact probabilities for that body part. It is evident that, while we typically use our hands and feet for contact, we also frequently use the rest of our body, especially the buttocks, back of the head, chest, lips, and ears to interact with everyday objects. To our knowledge, no such analysis of full-body contact for in-the-wild images has previously been reported. This motivates the need for modeling dense full-body contact.

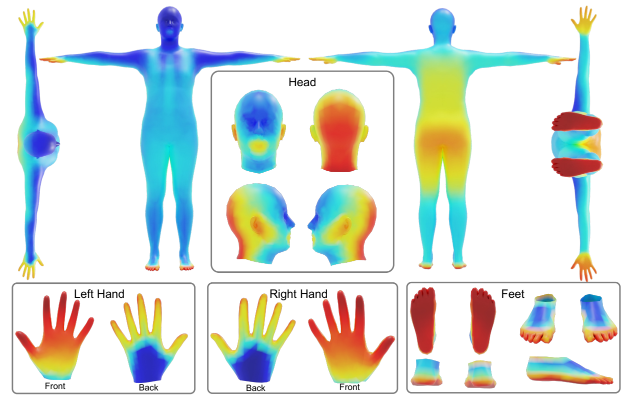

Figure 4: Aggregate statistics showing contact probabilities across the body vertices in the DAMON dataset. The body part closeups show the contact probabilities normalized for that body part. Red implies higher probability of contact while blue implies lower probability. 🔍 **Zoom in**.

4. Method: DECO

Contact regions in images are ipso facto occluded. This makes human-object contact estimation from in-the-wild images a challenging and ill-posed problem. We tackle this with a new **DE**nse **CO**ntact estimator, DECO, which uses scene and part context.

Our contributions are two fold: (1) To reason about the contacting body parts, human-object proximity, and the surrounding scene context, we use a novel architecture with three branches, i.e., a scene-context, a part-context, and a per-vertex contact-classification branch. (2) We use a novel 2D pixel-anchoring loss that constrains the solution space by grounding the inferred 3D contact to the 2D image space.

4.1. Model architecture

Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, DECO predicts contact probabilities on the SMPL [43] mesh. We use SMPL as it is widely used for HPS estimation [31, 32, 34, 35, 38, 80]. SMPL parameterizes the human body with pose and shape parameters, $\mathbf{\Theta} = [\boldsymbol{\theta} \in \mathbb{R}^{72}, \boldsymbol{\beta} \in \mathbb{R}^{10}]$ and outputs a 3D mesh $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$. SMPL's template mesh $\bar{\mathcal{M}}$ is segmented into $J = 24$ parts, $P_k \in \mathcal{P}$, which allows part-labeling of contact vertices. Moreover, SMPL's mesh topology is consistent with the SMPL-H [53] model and has the same vertices below the neck as the SMPL-X model [28], making our contact representation widely applicable.

Figure 5 shows DECO's architecture. Intuitively, contact estimation relies on both part and scene features as they are complementary. We use two separate encoders $\mathcal{E}_s$ and $\mathcal{E}_p$ to extract scene features $\boldsymbol{F_s}$ and body-part features $\boldsymbol{F_p}$. For the encoder backbone, we use both the transformer-based SWIN [42] and the CNN-based HRNET [68]. We integrate scene features $\boldsymbol{F_s}$ and body-part features $\boldsymbol{F_p}$ via a cross-

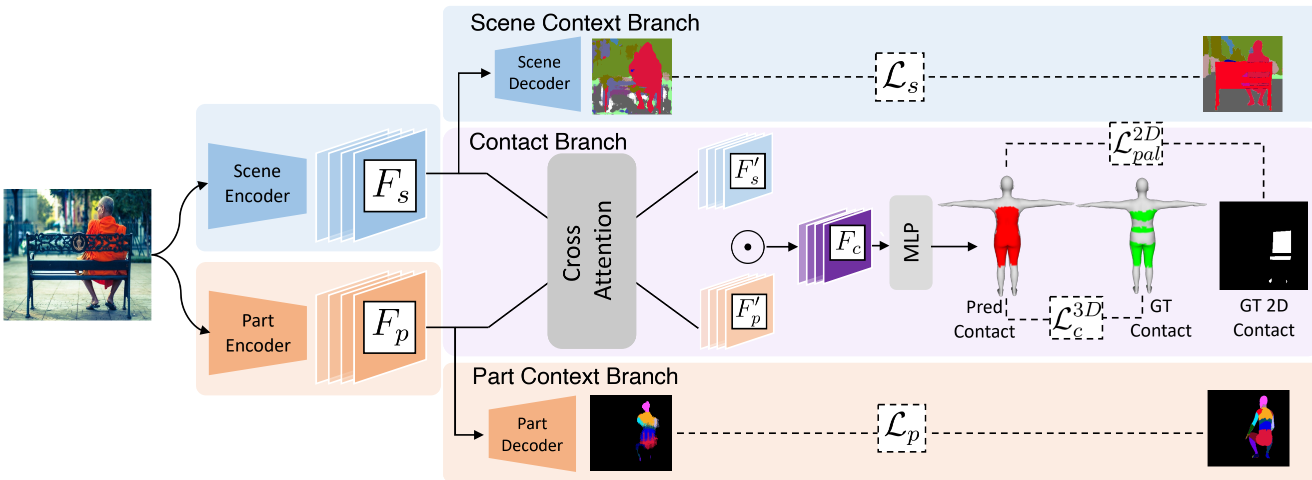

Figure 5: DECO architecture (Sec. 4.1). DECO reasons about body parts, human-object proximity, and the surrounding scene context. To this end, it uses three branches, i.e., a scene-context, a part-context, and a per-vertex contact-classification branch. Cross attention guides the features to focus attention on (and around) body parts and scene elements that are relevant for contact.

attention module inspired by [44, 67]. Previous methods either concatenate multi-modal features [46], use channel-wise multiplication [34], adopt trainable fusion [65] or use bilinear interpolation between multi-modal features [62]. However, such methods simply combine the multi-modal features without explicitly exploiting their interactions. In contrast, DECO's cross-attention guides the network to "attend" to relevant regions in $\boldsymbol{F_s}$ and $\boldsymbol{F_p}$ to reason about contact.

To implement cross-attention, we exchange the key-value pairs in the multi-head attention block between the two branches. Specifically, we initialize the query, key, and value matrices for each branch i.e. $\{\mathcal{Q}_s, \mathcal{K}_s, \mathcal{V}_s\} = \{\boldsymbol{F_s}, \boldsymbol{F_s}, \boldsymbol{F_s}\}$ for the scene branch and $\{\mathcal{Q}_p, \mathcal{K}_p, \mathcal{V}_p\} = \{\boldsymbol{F_p}, \boldsymbol{F_p}, \boldsymbol{F_p}\}$ for the part branch. Then we obtain the contact features $\boldsymbol{F_c}$ after multi-head attention as

$$\boldsymbol{F'_s} = \text{softmax}(\mathcal{Q}_p \mathcal{K}_s^T / \sqrt{C_t}) \mathcal{V}_s, \tag{1}$$

$$\boldsymbol{F'_p} = \text{softmax}(\mathcal{Q}_s \mathcal{K}_p^T / \sqrt{C_t}) \mathcal{V}_p, \tag{2}$$

$$\boldsymbol{F_c} = LN(\boldsymbol{F'_s} \odot \boldsymbol{F'_p}), \tag{3}$$

where C_t is a scaling factor [67], $\odot$ is the Hadamard operator and LN represents layer-normalization [2]. We obtain final contact predictions $\bar{y}_c \in \mathbb{R}^{6890 \times 1}$ after filtering $\boldsymbol{F_c}$ via a shallow MLP followed by sigmoid activation.

The DECO architecture encourages the scene and part encoders, $\mathcal{E}_s$ and $\mathcal{E}_p$, to focus on relevant features by upsampling $\boldsymbol{F_s}$ and $\boldsymbol{F_p}$ using scene decoder $\mathcal{D}_s$ and part decoder $\mathcal{D}_p$ respectively. The output of $\mathcal{D}_s$ is a predicted scene segmentation map, $\bar{\boldsymbol{X}}_s \in \mathbb{R}^{H \times W \times N_o}$, where N_o are the number of objects in MS COCO [40]. Similarly, we obtain the part features $\bar{\boldsymbol{X}}_p \in \mathbb{R}^{H \times W \times (J+1)}$ from $\mathcal{D}_p$, where J are the number of body parts and the extra channel is for the background class.

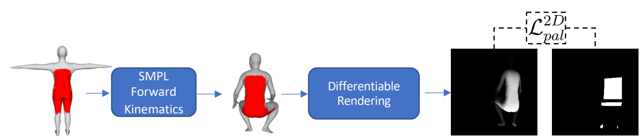

Figure 6: The Pixel Anchoring Loss (PAL) grounds 3D contact predictions to image pixels by rendering the contact-colored posed mesh on the image plane. The rendered contact mask is compared with 2D contact ground truth contact from HOT [8]

.

We train DECO end-to-end (Fig. 5) with the loss:

$$\mathcal{L} = w_c \mathcal{L}^{3D}_c + w_{pal} \mathcal{L}^{2D}_{pal} + w_s \mathcal{L}^{2D}_s + w_p \mathcal{L}^{2D}_p, \tag{4}$$

where $\mathcal{L}^{3D}_c$ is the binary-cross entropy loss between per-vertex predicted contact $\bar{y}_c$ and ground-truth contact labels y^{gt}_c. $\mathcal{L}^{2D}_s$ and $\mathcal{L}^{2D}_p$ are segmentation losses between the predicted and the ground-truth masks. We describe $\mathcal{L}^{2D}_{pal}$ in the following section. Steering weights w are set empirically.

4.2. 2D Pixel Anchoring Loss (PAL)

To relate contact on the 3D mesh with image pixels, we propose a novel pixel anchoring loss (PAL); see Fig. 6. We run the SOTA HPS network CLIFF [38] on input image I to infer the camera scale s, camera translation, $\mathbf{t}^c$, and SMPL parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, in the camera coordinates assuming camera rotation, $\mathbf{R}^c = \boldsymbol{I}_3$ and body translation, $\mathbf{t}^b = \mathbf{0}$. Using the estimated SMPL parameters, we obtain the *posed* mesh $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t}^b)$, which is colored using DECO-predicted per-vertex contact probability, $\bar{y}_c$, in a continuous and differentiable manner. We denote the posed mesh colored with contact probability by $\mathcal{M}_c$. We use the PyTorch3D [50] differentiable renderer to render $\mathcal{M}_c$ on the image under weak perspective, resulting in the 2D contact probability map, $\bar{\boldsymbol{X}}_c^{2D}$. $\mathcal{L}_{\text{pal}}^{2D}$ is computed as the binary-cross entropy loss between $\bar{\boldsymbol{X}}_c^{2D}$ and the ground-truth 2D contact mask from HOT [8], $\boldsymbol{X}_c^{2D}$.

5. Experiments

Implementation Details. We experiment with both Swin Transformer [42] and HRNET [68] as backbone architectures for $\mathcal{E}_s$ and $\mathcal{E}_p$. We initialize the two encoder configurations with ImageNet and HRNET pretrained weights respectively. We obtain pseudo ground-truth scene segmentation masks, $\boldsymbol{X}_s \in \mathbb{R}^{H \times W \times N_o}$, containing semantic labels for $N_o = 133$ categories, by running inference using the SOTA image segmentation network, Mask2Former [9]. To get ground-truth part segmentations, $\bar{\boldsymbol{X}}_p \in \mathbb{R}^{H \times W \times (J+1)}$, we follow [34] to use the SMPL part segmentation and segment the posed ground-truth mesh when available (e.g. in RICH and PROX) into $J = 24$ parts, rendering each part mask as a separate channel. Since there are no ground-truth 3D meshes in DAMON, we obtain pseudo ground-truth meshes by running the SOTA human pose and shape network, CLIFF [38]. This strategy works better in practice than using a human-parsing network (e.g. Graphonomy [19]). It has the advantage of *left-right sided* part labels, which helps in circumventing left-right ambiguity. It also retains full-visibility under occlusion, which allows reasoning about parts not visible in the original image.

Training and Evaluation. To train DECO, we use the DAMON dataset along with existing datasets with 3D contact labels: RICH [28] and PROX [24]. We evaluate our method on the test splits of DAMON and RICH. To evaluate out-of-domain generalization performance, we also show evaluation on the test split of BEHAVE [3], which is not used in training. We follow [28] and report both count-based evaluation metrics: precision, recall and F1 score and geodesic error (in cm, see [28] for details). For additional implementation and training details, please refer to Sup. Mat.

5.1. 3D Contact Estimation

We compare DECO with BSTRO [28] and POSA [25], both of which give dense vertex-level contact on the body mesh. Since POSA needs a posed body mesh as input, we show POSA results when given ground-truth meshes, called POSA^{GT} and meshes reconstructed by PIXIE [13], called $\text{POSA}^{\text{PIXIE}}$. For a fair comparison, we make sure to use the same training data splits in all our evaluations.

We report results on RICH-test, BEHAVE-test, and DAMON-test in Tab. 1. For evaluation on RICH-test, we train both BSTRO and DECO on the RICH training split only. This ablates the effect of the DAMON dataset, allowing us to isolate the contribution of the DECO architecture. As shown in Tab. 1, we outperform all baselines across all metrics. Specifically, we report a significant $\sim$11% improvement in F1 score and 7.93 cm improvement in the geodesic error over the closest baseline, BSTRO. Further, we observe that adding $\mathcal{L}_{pal}^{2D}$ improves the geodesic error considerably with only a slight trade-off in F1 score. Here, we reiterate the observation in [28] that, while POSA matches DECO in recall, it comes at the cost of precision, resulting in worse F1 scores. Since POSA does not rely on image evidence and only takes the body pose as input, it tends to predict false positives. For qualitative results, see Fig. 7 and Sup. Mat.

Next, we retrain both BSTRO and DECO on all available training datasets, RICH, PROX and DAMON, and evaluate on the DAMON test split. POSA training needs a GT body which is not available in DAMON. This evaluation tests generalization to unconstrained Internet images. Note that to train with $\mathcal{L}_{pal}^{2D}$, we include HOT images with 2D contact annotations even if they do not have 3D contact labels from DAMON. For these images, we simply turn off $\mathcal{L}_c^{3D}$. This is because DECO, unlike BSTRO, is compatible with both 3D and 2D contact labels. DECO significantly outperforms all baselines and results in an F1 score of 0.55 vs 0.46 for BSTRO with a 16.18 cm improvement in geodesic error. Notably, the improvement over baselines when including PROX and DAMON in training is higher compared with training only on RICH, which indicates that DECO scales better with more training images compared to BSTRO.

Finally, we evaluate out-of-domain generalization on the unseen BEHAVE [3] dataset. BEHAVE focuses on a single human-object contact per image, even if multiple contacting objects may be present. The focus on single object-contact in the GT contact annotations partly explains why most methods struggle with this dataset. Further, since BEHAVE does not label contact with the ground, for the purpose of evaluation, we mask out contact predictions on the feet. As reported in Tab. 1, we outperform all baselines on both F1 and geodesic error, which indicates that DECO has a better generalization ability.

5.2. Ablation Study

In Tab. 2 we evaluate the impact of our design choices. First, we analyze the effect of using a shared encoder for the scene and the part branch vs separate encoders for both. Compared to having separate encoders without branch-specific losses, a single encoder performs better, which can be attributed to having fewer training parameters. However, any configuration using $\mathcal{L}_s^{2D}$ or $\mathcal{L}_p^{2D}$ outperforms the shared encoder. While $\mathcal{L}_p^{2D}$ contributes improvements to precision, $\mathcal{L}_s^{2D}$ contributes to better recall. This is expected since, intuitively, attending to body parts helps with inferring fine-grained contact, whereas scene context helps to reason

Methods	RICH [28]				DAMON				BEHAVE [3]			
	Precision ↑	Recall ↑	F1 ↑	geo. (cm)↓	Precision ↑	Recall ↑	F1 ↑	geo. (cm) ↓	Precision ↑	Recall ↑	F1 ↑	geo. (cm) ↓
BSTRO [28]	0.65	0.66	0.63	18.39	0.51	0.53	0.46	38.06	0.13	0.03	0.04	50.45
POSA$^{\text{PIXIE}}$ [13, 25]	0.31	0.69	0.39	21.16	0.42	0.34	0.31	33.00	0.11	0.07	0.06	54.29
POSA$^{\text{GT}}$ [13, 25]	0.37	**0.76**	0.46	19.96	-	-	-	-	0.10	0.09	0.06	55.43
DECO	**0.71**	**0.76**	**0.70**	17.92	0.64	**0.57**	**0.55**	**21.32**	0.25	**0.21**	**0.18**	46.33
DECO $+\mathcal{L}^{2D}_{pal}$	**0.71**	0.74	0.69	**10.46**	**0.65**	**0.57**	**0.55**	21.88	**0.27**	0.18	**0.18**	**44.51**

Table 1: Comparison of DECO with SOTA models on RICH [28], DAMON, and BEHAVE [3]. See discussion in Sec. 5.1.

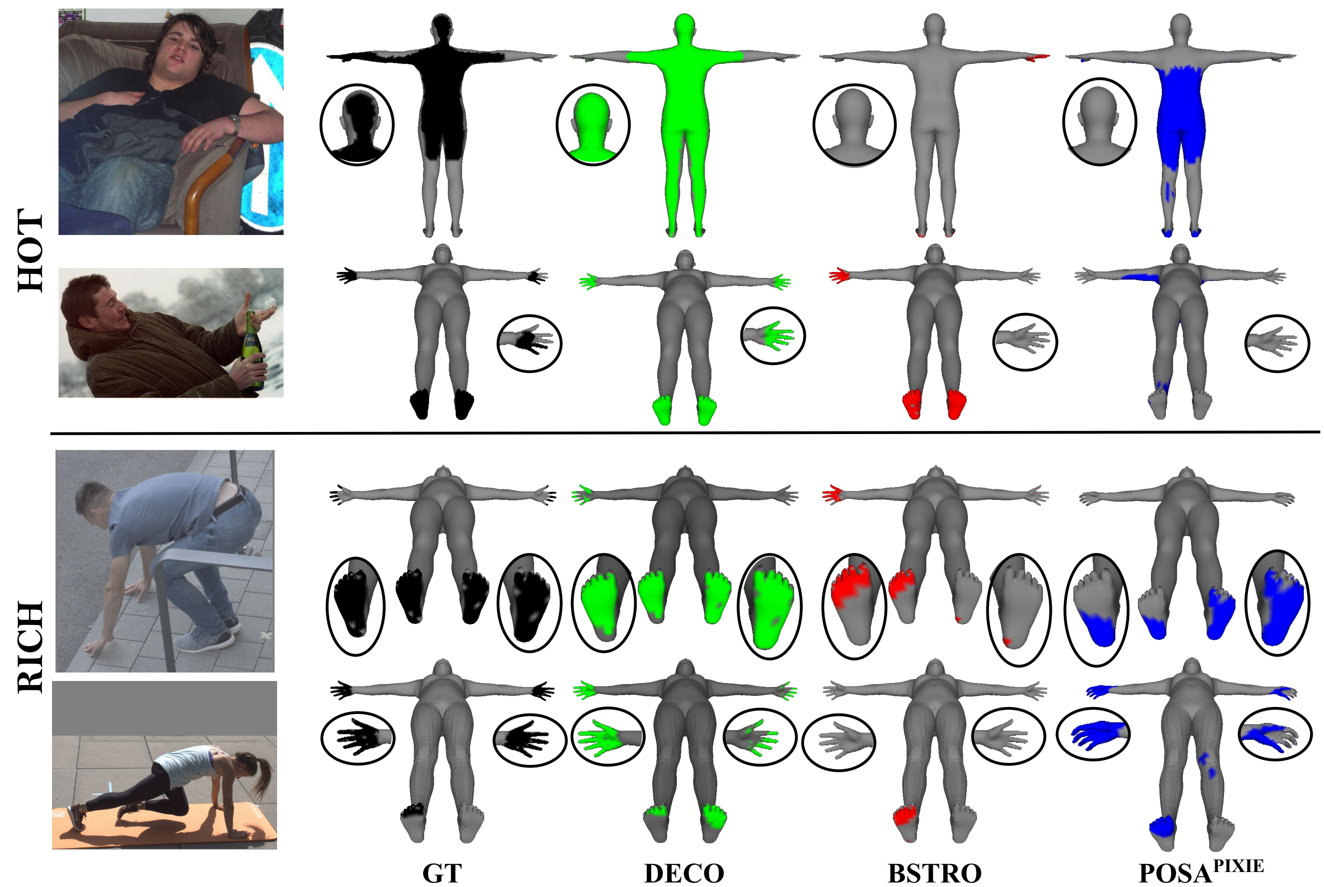

Figure 7: Qualitative evaluation of DECO (green), BSTRO (red) and POSA$^{\text{PIXIE}}$ (blue), alongside Ground Truth (**black**).

$\mathcal{E}_s$	$\mathcal{E}_p$	$\mathcal{L}^{2D}_s$	$\mathcal{L}^{2D}_p$	Back.	Pre. ↑	Rec. ↑	F1 ↑	geo. (cm) ↓
shared		✗	✗	HR	0.68	**0.76**	0.68	20.85
✓	✓	✗	✗	HR	0.67	0.76	0.67	23.54
✓	✓	✓	✗	HR	0.68	0.75	0.68	18.44
✓	✓	✗	✓	HR	0.70	0.74	0.68	18.37
✓	✓	✓	✓	SW	0.68	0.71	0.66	18.54
✓	✓	✓	✓	HR	**0.71**	**0.76**	**0.70**	**17.92**

Table 2: Ablation study for DECO design choices (Sec. 5.2). We ablate: (1) using separate or joint encoders for the scene and body parts, (2) using branch-specific losses, (3) using an HRNET (HR) or Swin (SW) backbone. Bold denotes best performance.

about the existence of contact regions. Each one separately helps with geodesic error, but the best performance comes when used together, in terms of both F1 score and geodesic error. Finally, we see that the HRNET backbone outperforms the Swin backbone. This is likely because HRNET is pre-trained on human-centric tasks (like our task), whereas Swin in pretrained on ImageNet image classification.

5.3. Inferred versus geometric contact

An alternative to directly inferring contact, as DECO does, is to first recover the 3D body and scene and then compute contact geometrically using the distance between the body and scene [73, 81]. If 3D human and scene recovery were accurate, this could be a viable alternative to DECO's inferred contact. To test this hypothesis we perform an experiment using the two SOTA techniques for 3D human and object estimation, PHOSA [81] and CHORE [73]. PHOSA works only on 8 objects, and CHORE works on 13. In contrast, DECO supports all 80 object classes in MS-COCO. Because they are optimization based, PHOSA and CHORE are slow, taking 4 mins and 66 secs per image respectively. DECO is real-time and takes 0.012 secs for inference. For fair comparison, we split the DAMON dataset and evaluate using test sets that include only objects supported by either PHOSA or CHORE. We reconstruct the human and object and then recover contact using thresholded distance. CHORE achieves an F1 score of 0.08 as opposed to DECO's score of 0.48. Similarly, PHOSA achieves an F1 score of 0.18 as opposed to DECO's score of 0.60. Given the current state of 3D human pose and scene estimation, DECO significantly outperforms geometry-based contact estimation.

6. HPS using DECO contacts

Next we evaluate whether contact information inferred by DECO can be used to improve human pose and shape (HPS) regression; we do so using the PROX "quantitative" dataset [24]. PROX uses an optimization method to fit SMPL-X bodies to images. It further assumes a-priori known 3D scenes and uses manually-annotated contact regions on the body to encourage these body vertices to be in contact with the scene if they are sufficiently close, while penalizing body-scene penetration.

Specifically, we replace the manually-annotated contact vertices with the inferred SMPL-X body-part contact vertices from baseline methods as well as the detailed contact estimated by DECO. For a fair comparison, we follow the same experimental setup as HOT [8] and evaluate all methods using the Vertex-to-Vertex (V2V) error. For the "No contact" setup, we turn off all contact constraints in the optimization process. PROX uses the contact regions on the body from the original method [24]. HOT uses the body-part vertices from the body-part labels predicted by the HOT detector. We also report V2V errors when using the ground-truth (GT) contact vertices. The results in Tab. 3 illustrate the value of inferring detailed contact on the body.

All baselines in Tab. 3 use PROX's [24] hyperparameters for a fair comparison. PROX uses a Geman-McClure robust error function (GMoF) for the contact term (see Eq.4 in [24]), so that the manually-defined contact areas that lie "close enough" to the scene are snapped onto it. The robust scale term, $\rho_C = 5e-02$, is tuned for PROX's naive contact prediction; this is relatively conservative as PROX uses no image contact for this prediction. Since DECO takes into account the image features, and makes a much more informed contact prediction, we we can "relax" this robustness term, and trust the output of DECO regressor more. In Tab. 4 we report a sensitivity analysis by varying ρ_C with DECO's contact predictions. The results verify that we can trust DECO's

Method	No Contact	PROX [24]	HOT [8]	DECO Contact	GT Contact
V2V ↓	183.3	174.0	172.3	**171.6**	163.0

Table 3: HPS estimation performance using contact derived from different sources.

GMoF ρ_C	1e-03	5e-02	1e-01	1.0	2.0	3.0	5.0
V2V ↓	180.07	171.6	170.0	**169.0**	176.5	179.6	183.5

Table 4: Sensitivity analysis for the ρ_C value in the Geman-McClure error function (GMoF) of the contact term.

contacts more, and that there is a sweet spot for $\rho_C = 1.0$. This suggests that exploiting inferred contact is a promising direction for improving HPS estimates.

7. Conclusion

We focus on detecting 3D human-object contact from a single image taken in the wild; existing methods perform poorly for such images. To this end, we use crowd-sourcing to collect DAMON, a rich dataset of in-the-wild images paired with pseudo ground-truth 3D contacts on the vertex level, as well as labels for the involved objects and body parts. Using DAMON, we train DECO, a novel model that detects contact on a 3D body from a single color image. DECO's novelty lies in cross-attending to both the relevant body parts and scene elements, while it also anchors the inferred 3D contacts to the relevant 2D pixels. Experiments show that DECO outperforms existing work by a good margin, and generalizes reasonably well in the wild. To enable further research, we release our data, models and code.

Future work: DECO currently reasons about contact between a single person, the scene, and multiple objects. Our labelling tool and DECO could be extended to fine-grained human-human, human-animal and self-contact. Another promising, but challenging, direction would be to leverage captions in existing datasets, or methods that infer captions for unlabeled images, via large language models (LLM).

Acknowledgements: We sincerely thank Alpar Cseke for his contributions to DAMON data collection and PHOSA evaluations, Sai K. Dwivedi for facilitating PROX downstream experiments, Xianghui Xie for help with CHORE evaluations, Lea Müller for her help in initiating the contact annotation tool, Chun-Hao P. Huang for RICH discussions and Yixin Chen for details about the HOT paper. We are grateful to Mengqin Xue and Zhenyu Lou for their collaboration in BEHAVE evaluations and Tsvetelina Alexiadis for valuable data collection guidance. Their invaluable contributions enriched this research significantly. This work was funded by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). **Disclosure:** https://files.is.tue.mpg.de/black/CoI_ICCV_2023.txt

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5167–5176, 2018. 2

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. 6

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 15935–15946, 2022. 2, 4, 7, 8

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, volume 12358, pages 361–378. Springer, 2020. 2

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, volume 12346, pages 387–404, 2020. 4

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

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *International Conference on Computer Vision (ICCV)*, pages 12397–12406, 2021. 3

[8] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 6, 7, 9

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 7

[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 *Computer Vision and Pattern Recognition (CVPR)*, pages 2718–2728, 2022. 2

[11] Henry M. Clever, Zackory M. Erickson, Ariel Kapusta, Greg Turk, C. Karen Liu, and Charles C. Kemp. Bodies at Rest: 3D human pose and shape estimation from a pressure image using synthetic data. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6214–6223, 2020. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2

[13] 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 (3DV)*, 2021. 7, 8

[14] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7221, 2020. 2, 3, 4

[15] 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. 2, 3, 4

[16] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. REMIPS: Physically consistent 3D reconstruction of multiple interacting people under weak supervision. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 19385–19397. Curran Associates, Inc., 2021. 3, 4

[17] Christopher Funk, Savinay Nagendra, Jesse Scott, Bharadwaj Ravichandran, John H Challis, Robert T Collins, and Yanxi Liu. Learning dynamics from kinematics: Estimating 2D foot pressure maps from video frames. *arXiv:1811.12607*, 2018. 3

[18] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3D human motion reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13180–13190, 2022. 3

[19] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7450–7459, 2019. 7

[20] Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. PressureVision: Estimating hand pressure from a single RGB image. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[21] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021. 3

[22] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv:1505.04474*, 2015. 2, 4

[23] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 4318–4329, 2021. 4

[24] 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 (ICCV)*, pages 2282–2292, 2019. 2, 3, 4, 5, 7, 9

[25] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, 2021. 4, 7, 8

[26] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 3

[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B.

Girshick. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[28] Chun-Hao Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4, 5, 7, 8

[29] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, volume 13485, pages 281–299, 2022. 2, 4

[30] Leslie Ikemoto, Okan Arikan, and David Forsyth. Knowing when to put your foot down. In *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games*, page 49–53, 2006. 3

[31] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 5

[32] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1705–1715, 2022. 5

[33] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-end human-object interaction detection with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[34] 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 (ICCV)*, pages 11127–11137, 2021. 5, 6, 7

[35] 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 (ICCV)*, pages 2252–2261, 2019. 5

[36] Hema S Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *European Conference on Computer Vision (ECCV)*, pages 831–847. Springer, 2014. 3

[37] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Computer Vision and Pattern Recognition (CVPR)*, pages 382–391, 2020. 2, 4

[38] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, volume 13665, pages 590–606, 2022. 3, 5, 6, 7

[39] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 2

[40] Tsung-Yi Lin, Michael Maire, Serge J. 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 (ECCV)*, volume 8693, pages 740–755, 2014. 6

[41] Tsung-Yi Lin, Michael Maire, Serge J. 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 (ECCV)*, 2014. 2

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 5, 7

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 4, 5

[44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019. 6

[45] 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 (ICCV)*, pages 5441–5450, 2019. 3

[46] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. *Transactions on Graphics (TOG)*, 36(4):44:1–44:14, 2017. 6

[47] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, 2021. 2, 3, 4

[48] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. *Conference on Neural Information Processing Systems (NeurIPS)*, 33:7841–7851, 2020. 2, 3

[49] 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 (ECCV)*, pages 401–417, 2018. 3

[50] Nikhila Ravi, Jeremy Reizenstein, David Novotný, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *CoRR*, abs/2007.08501, 2020. 7

[51] 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 (ICCV)*, pages 11488–11499, 2021. 2, 3

[52] 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 (ECCV)*, pages 71–87. Springer, 2020. 3

[53] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 5

[54] Anirban Roy and Sinisa Todorovic. A multi-scale CNN for affordance segmentation in RGB images. In *European Conference on Computer Vision (ECCV)*, pages 186–201, 2016.

3
[55] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, and Silvia Zuffi. BITE: Beyond priors for improved three-D dog pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8867–8876, June 2023. 4
[56] Jesse Scott, Bharadwaj Ravichandran, Christopher Funk, Robert T Collins, and Yanxi Liu. From image to stability: learning dynamics from human pose. In *European Conference on Computer Vision (ECCV)*, pages 536–554, 2020. 3
[57] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9869–9878, 2020. 3
[58] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3D human motion reconstruction from monocular video with skeleton consistency. *Transactions on Graphics (TOG)*, 40(1):1–15, 2020. 3
[59] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. HULC: 3D human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision (ECCV)*, pages 516–533, 2022. 2, 4
[60] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 3
[61] Stephan Streuber, M. Alejandra Quiros-Ramirez, Matthew Q. Hill, Carina A. Hahn, Silvia Zuffi, Alice O'Toole, and Michael J. Black. Body Talk: Crowdshaping realistic 3D avatars with words. *Transactions on Graphics (TOG)*, 35(4), 2016. 2
[62] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *International Conference on Computer Vision (ICCV)*, pages 5348–5357, 2019. 6
[63] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3
[64] 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 (ECCV)*, volume 12349, pages 581–600, 2020. 4
[65] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 3961–3970, 2017. 6
[66] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. 3
[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 6
[68] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(10):3349–3364, 2021. 5, 7
[69] 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 (ICCV)*, pages 5694–5702, 2019. 3
[70] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2596–2605, 2017. 3
[71] Zhenzhen Weng and Serena Yeung. Holistic 3D human and scene mesh estimation from single view images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 334–343, 2020. 4
[72] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2
[73] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*, pages 125–145. Springer, 2022. 3, 4, 8, 9
[74] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4757–4768, June 2023. 4
[75] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
[76] M. Yamamoto and K. Yagishita. Scene constraints-aided tracking of human body. In *Computer Vision and Pattern Recognition (CVPR)*, pages 151–156 vol.1, 2000. 3
[77] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12965–12976, June 2023. 3, 4
[78] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7159–7169, 2021. 3
[79] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 3
[80] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11426–11436, 2021. 5
[81] 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 (ECCV)*, pages

34–51. Springer, 2020. 3, 4, 8, 9

[82] 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 (ICCV)*, pages 11343–11353, 2021. 3

[83] 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 (3DV)*, pages 642–651, 2020.

[84] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6193–6203, 2020. 3

[85] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3823–3833, 2016. 3

[86] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 11825–11834, 2021. 3

[87] 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 (WACV)*, pages 459–468, 2020. 3