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

1. DAMON Data Collection and Quality

We select images for annotation from the HOT [1] curated subset of V-COCO [2] and HAKE [6] by filtering out images containing multiple people or images with a single person but fewer than 10 visible keypoints. For keypoint estimation, we use the transformer-based SOTA 2D keypoint estimator ViTPose [7].

We take several steps to limit ambiguity in the contact annotation task. Here, we focus on *scene-* and *human-supported* contact. The requirement for support resolves ambiguous cases, e.g. humans close to scene objects but not in contact. We use the object labels in V-COCO and HAKE to filter out images containing unsupported human-human and human-animal contact. V-COCO and HAKE also contain action labels that we leverage to filter out ambiguous indirect contact which does not involve physical touch, such as direct, greet, herd, hose, point, teach, etc. The training video (in Sup. Mat.) advises workers to orient the 3D mesh and to visualize themselves in the same posture as the person in the image. This helps infer contact while avoiding left-right ambiguity. Our Fleiss' Kappa score indicates significant agreement between annotators (see Sec. 1.3), suggesting that our protocol effectively minimizes task ambiguities.

To facilitate crowd-sourced 3D contact annotation using Amazon Mechanical Turk (AMT), we build a new annotation tool which we describe in detail in the following section. Please see the **Supplemental Video**.

1.1. Dense Contact Annotation Tool

We built a dense contact annotation tool to collect annotations from the DAMON dataset images. The code for the tool is written using Dash, a popular Python framework for building web applications. This application is deployed inside a Docker container under an uWSGI application server, eventually served by a NGINX web server acting as a reverse proxy. The annotation tool is accessible under a public URL used to create the Human Intelligent Tasks on AMT.

Interface and use. As seen in Fig. 1, the application is made of four parts. The top part contains a title and general instructions about how to use the annotation tool. The left part is made of the image and a label describing which contact should be annotated (object or supporting contact). The right part contains the mesh to be annotated by hovering over it. The mesh can be translated, rotated, and zoomed-in/out. A slider allows the user to select the size of the brush, and buttons are available for switching modes (draw/erase), erasing the full selection, and resetting the camera. Finally, a confirmation button is located at the bottom of the window to submit an annotation to the server. The user must provide one annotation for several human-object contacts and for the supporting contact. Once the last annotation has been submitted, a dialog box appears to ask for optional feedback about the annotation task for the current image. This helps workers report ambiguous contact scenarios.

Callbacks. Dash applications work with *callbacks*. Callbacks are functions that are fired when an input component is updated (e.g., a button is clicked) and that update output components. *Regular* callbacks are executed on the server-side: they are simpler to implement, but slower to execute. On the other hand, *client-side* callbacks are faster but require a more complex implementation. The user will spend most of their time annotating the high-resolution mesh. It should therefore be smooth and fast. As such, we implemented this logic in JavaScript as a client-side callback. Other callbacks, for instance when the camera is reset or the brush size is updated, rarely happen and do not require a fast response. Therefore, they have been implemented as server-side callbacks. During their execution, a spinner appears to let the worker know that the application is updating.

Caching. When a vertex is annotated, vertices belonging to a neighboring region are also annotated. The extent of this neighboring region is correlated with the brush size that the user chooses. When we start the application, we compute, for each vertex and for each brush size, all of its neighboring vertices. As the mesh is static, this has to be done only once. Therefore, we cache this result and use it for all annotations.

Video. Please watch the **Supplementary Video** for an in-depth tour of our tool, its features and the annotation protocol. Note that this is the same video we showed AMT workers for training purposes during qualification.

1.2. DAMON Additional Statistics

Figure 3 shows the full version of Fig. 3 in the main paper. The DAMON dataset is long-tailed and it covers contact

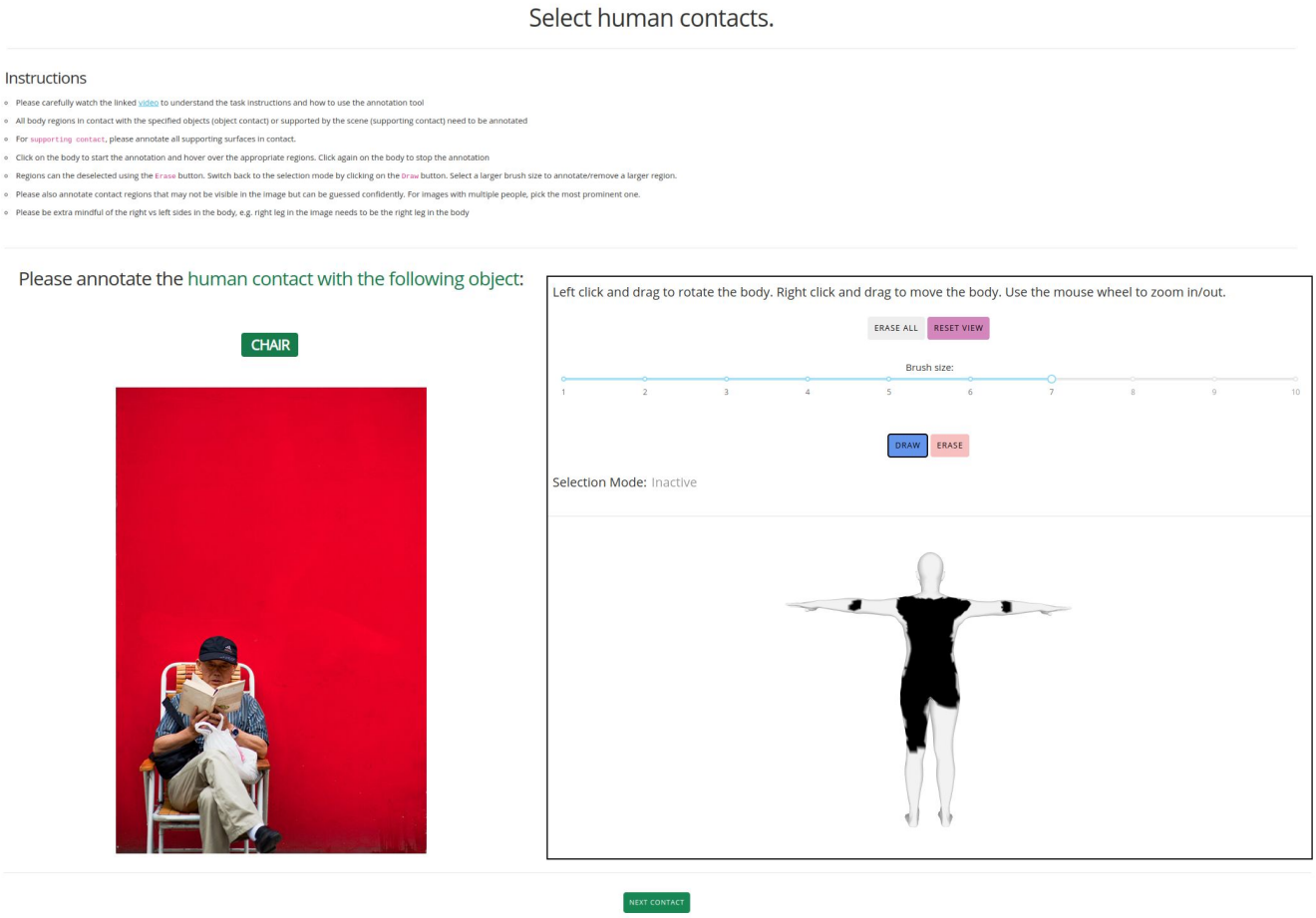

Figure 1: AMT interface design for our annotation tool. We show a an example of an annotator collecting *human-supported* contact for the object label "Book". The application cycles through all available object labels in the image and the *scene-supported* contact. Please refer to the **Supplemental Video** for a detailed description of the tool. **Zoom in**

scenarios with a wide variety of objects and scenes. Please refer to the sunburst plot in Fig. 3 for a full breakdown.

Figure 2 shows the number of images per object label. We see that contact with feet, hands, and the bigger body parts (torso, hips, upper arms) prevails; this makes sense as humans interact with objects mostly with these (e.g., for walking, grasping, sitting, lying down). However, interactions are highly varied, thus, the distribution is long-tailed and includes all body parts.

Workers take on average 3.48 min/image and we pay \$0.5/image. The total cost is \$3313.20 with AMT fees. The DAMON contact annotations are not prohibitively expensive given that it provides a stepping stone for future research.

1.3. Quality Control and Evaluation

We adopt two strategies to ensure quality and avoid noisy annotations in the DAMON dataset. First, we conduct qualification tasks to shortlist high-quality annotator candidates. This qualification task has two parts: (i) watching a detailed tutorial video (see **Supplementary Video**) explaining the task and annotator tool step-by-step by showing three example annotations with varying degrees of contact complexity, (ii) annotating 10 sample images for contact annotations. For the sample images, we had a set of author-annotated *pseudo-ground-truth* (pseudo GT) labels. The responses of candidates were evaluated using Intersection-over-Union (IoU) with the pseudo-GT labels. Workers who responded satisfactorily were allowed to annotate the DAMON dataset images. We qualified 14 out of 100 participants after the qualification round. The second strategy involved hiring Master's students as meta-annotators to visually inspect the quality of contact annotations. Annotations that were flagged as incorrect or low-quality were sent for re-annotation with specific feedback to the annotators on how to avoid mistakes.

We assess the quality of the DAMON dataset by measuring the *label accuracy* and the *level of annotator's agree-*

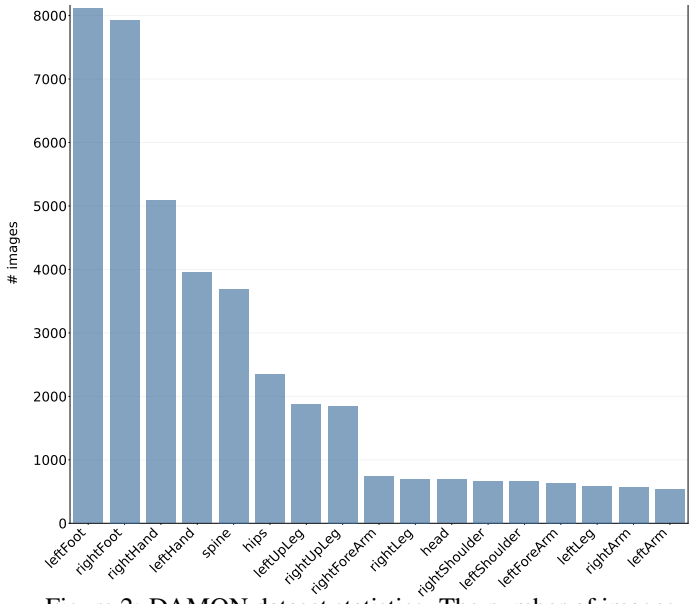

Figure 2: DAMON dataset statistics. The number of images (y-axis) for which each body part (x-axis) has at least 10 vertices in contact. For visualization purposes here we combine the fingers into the hands category, the toes into the feet category, and several spine parts into a single spine category (SMPL has 24 parts but here we show 17 bars). **Zoom in**.

ment.

We evaluate label accuracy by manually selecting 100 images with contact labels from the RICH [5] and PROX [4] datasets. Note that the pseudo-ground-truth contact labels in these datasets are obtained by thresholding the Signed Distance Field (SDF) between the reconstructed human mesh and the 3D scene. We evaluate annotations from qualified workers on these images and compute IoU w.r.t. the pseudo-ground-truth contact labels. With this, we obtain an IOU score of 0.512 on RICH, 0.263 on PROX, and a mean IOU (mIOU) score of 0.450.

Figure 5 visualizes the DAMON annotation earning the lowest IoU scores. Scanned datasets that rely on thresholding SDF values for estimating contact labels fail to take into account the soft-tissue deformation of the human body when it interacts with rigid objects. The vertices in the "soft" body parts such as buttocks, thighs, etc interpenetrate far enough from the scan surface to overshoot the heuristic threshold, leading to noisy GT annotation and a "ring" like contact profile. DAMON is annotated by human annotators and therefore does not suffer from this issue. This produces a mismatch between these two types of ground truth. Note that DAMON ground truth is closer to reality.

We also compare annotations on a randomly-selected set of 10 images from all the qualified workers against author-annotated labels, resulting in mIOU = 0.510.

To determine the agreement between annotators, qualified workers annotate the same set of 10 images and we report the *Fleiss' Kappa (κ)* metric. Fleiss' Kappa is a statistical measure used to evaluate the agreement level among a fixed number of annotators when assigning categorical labels to data. It considers the possibility of chance agreement and provides a standardized measure of inter-rater reliability that ranges from 0 (no agreement) to 1 (perfect agreement). In this study, we obtain a Fleiss' Kappa $\kappa = 0.656$ which is considered "substantial agreement" between workers [3]. Note, κ of 1 means "perfect agreement", 0 means "chance agreement" and -1 means "perfect disagreement". To build intuition on the significance of κ, Fig. 4 shows example annotations with low and high κ scores.

2. DECO Experiments

2.1. Implementation Details

For training DECO, we resize input images, the scene segmentation mask and the part segmentation mask such that $\mathbf{I} \in \mathbb{R}^{3\times256\times256}$, $\boldsymbol{X}_s \in \mathbb{R}^{133\times256\times256}$ and $\boldsymbol{X}_p \in \mathbb{R}^{25\times256\times256}$. $\mathbf{F}_s$ and $\mathbf{F}_s$ are of size $\mathbb{R}^{480\times64\times64}$S. We determine the loss weights in Eqn. 4 empirically and set it to $w_c = 10.0$, $w_{pal} = 0.05$, $w_s = 1.0$ and $w_p = 1.0$. We use the Adam optimizer with a learning rate of 5×10^{-5} and batch size of 4, and training takes 12 epochs ($\sim$ 31 hours) on an Nvidia Tesla V100 GPU.

For evaluation on RICH-test in Tab. 1 in main, we subsample every 10th frame from the released test set.

The base model without context branches has 90.19M parameters. Adding context branches ($\mathcal{L}_s^{2D}$ and $\mathcal{L}_p^{2D}$) adds another 853K parameters. This improves the geodesic error by $\sim$24% (see Tab. 1 in main), at the cost of $\sim$1% increase in complexity. We will release both models, with and without context branches.

2.2. Additional Qualitative results

Figure 6 shows DECO estimated contact and comparison with baseline methods from the test subset of DAMON. Figure 7 shows DECO contacts on some randomly sampled images from the internet.

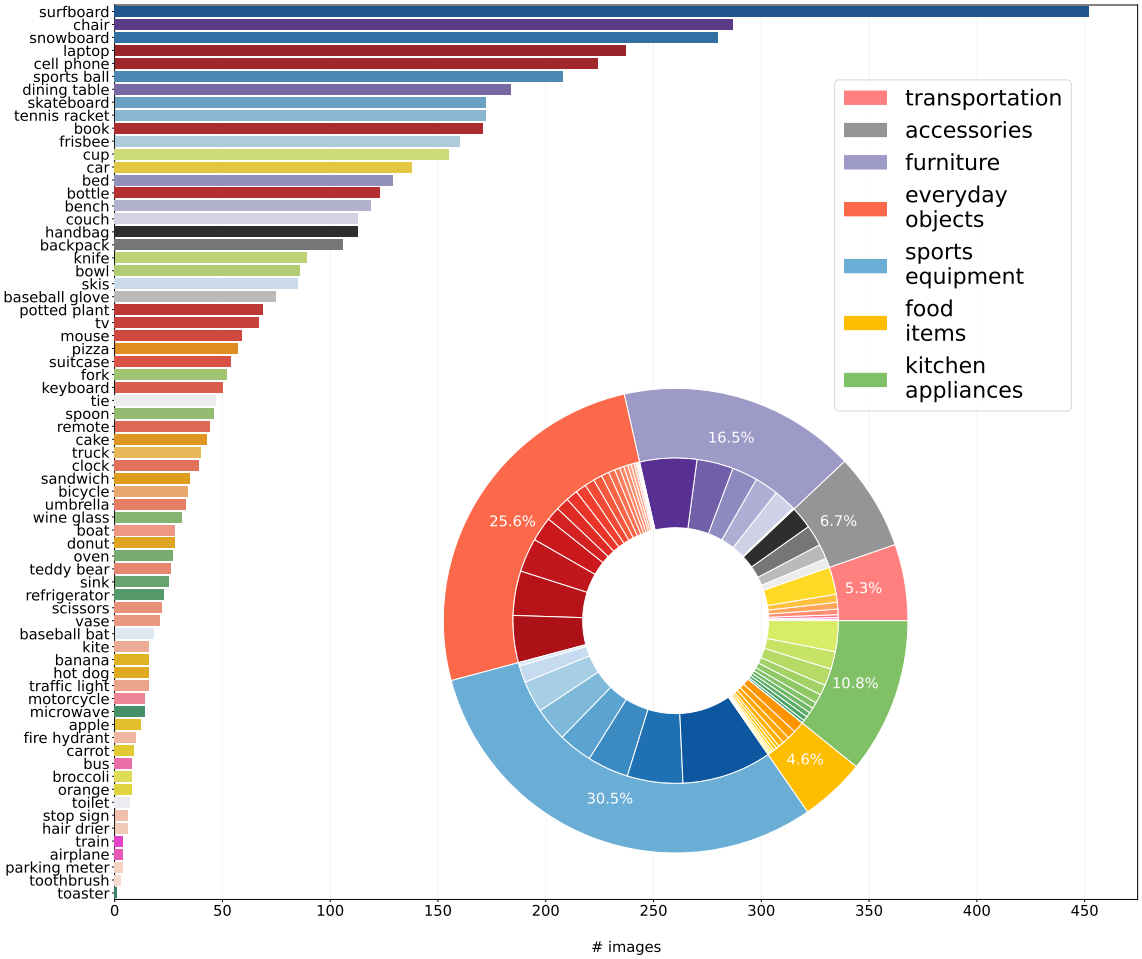

Figure 3: *Full version* plot for DAMON dataset statistics (Fig. 3 in Main). **Histogram**: contact object labels (y-axis) and the number of images in which they are present (x-axis). **Pie chart**: object labels are grouped into 7 main categories; inner colors correspond to the colors in the histogram. 🔍 **Zoom in**.

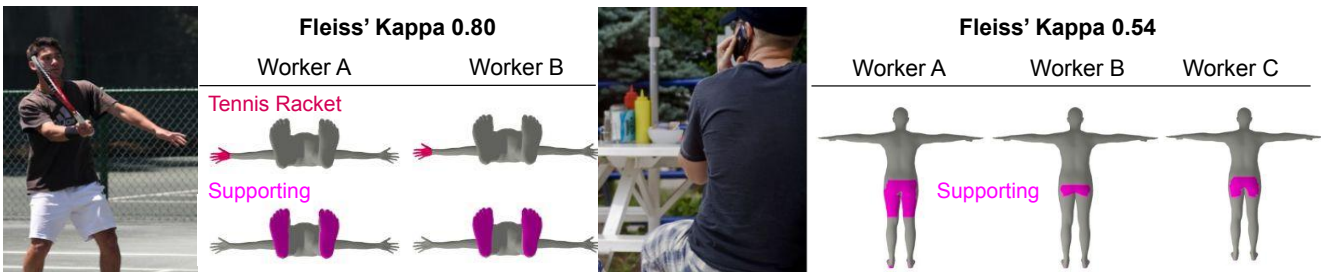

Figure 4: Annotator agreement indicated by κ 🔍 **Zoom in**.

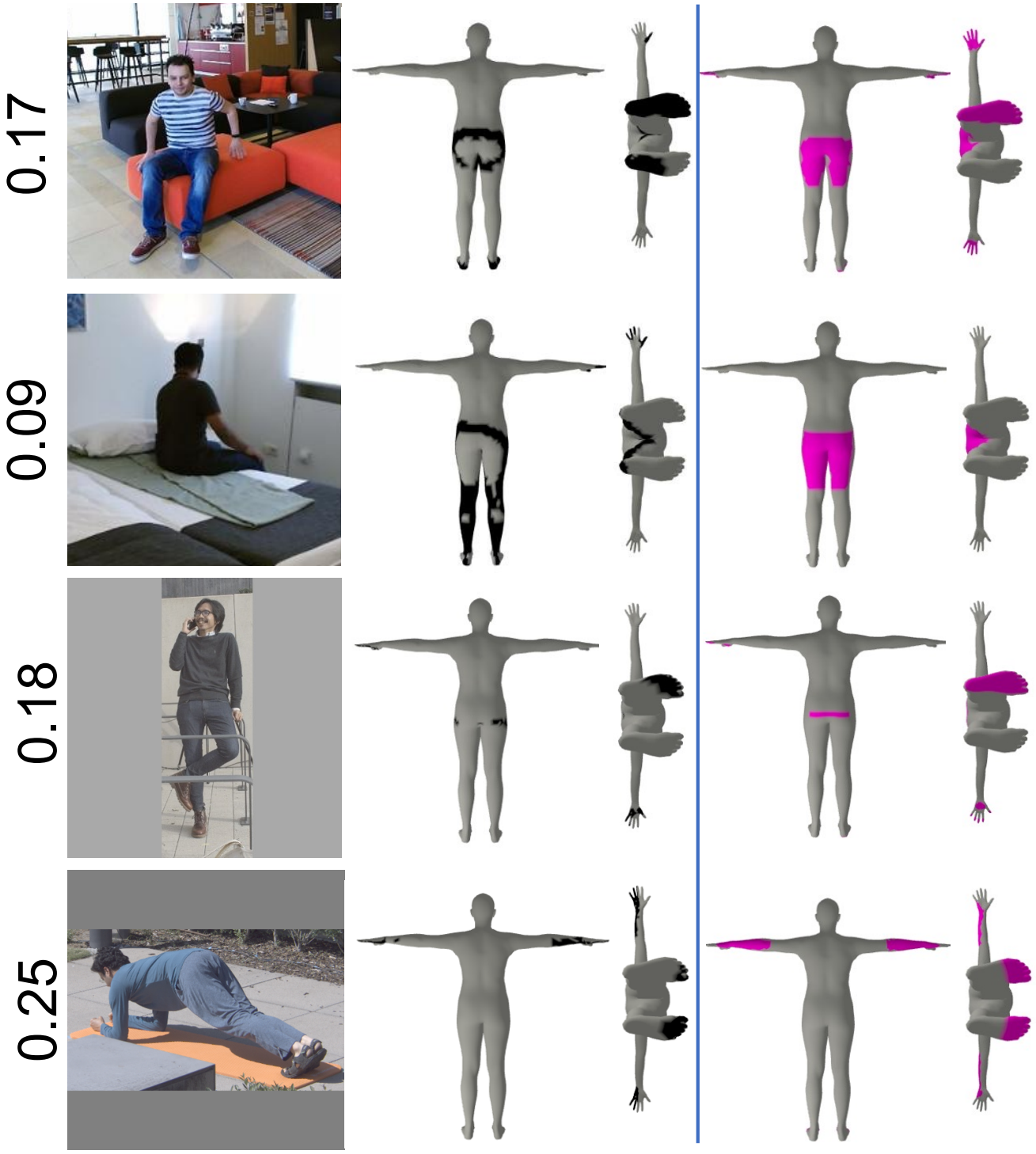

Figure 5: DAMON annotations (in magenta) earning the lowest IOU scores compared to GT contact in PROX and RICH (in **black**). IOU scores are reported to the left of RGB images for each row. Scanned datasets (e.g. PROX/RICH) infer contact by thresholding the SDF between body and scene, which can be sub-optimal due to soft-tissue deformation of the body (see text).

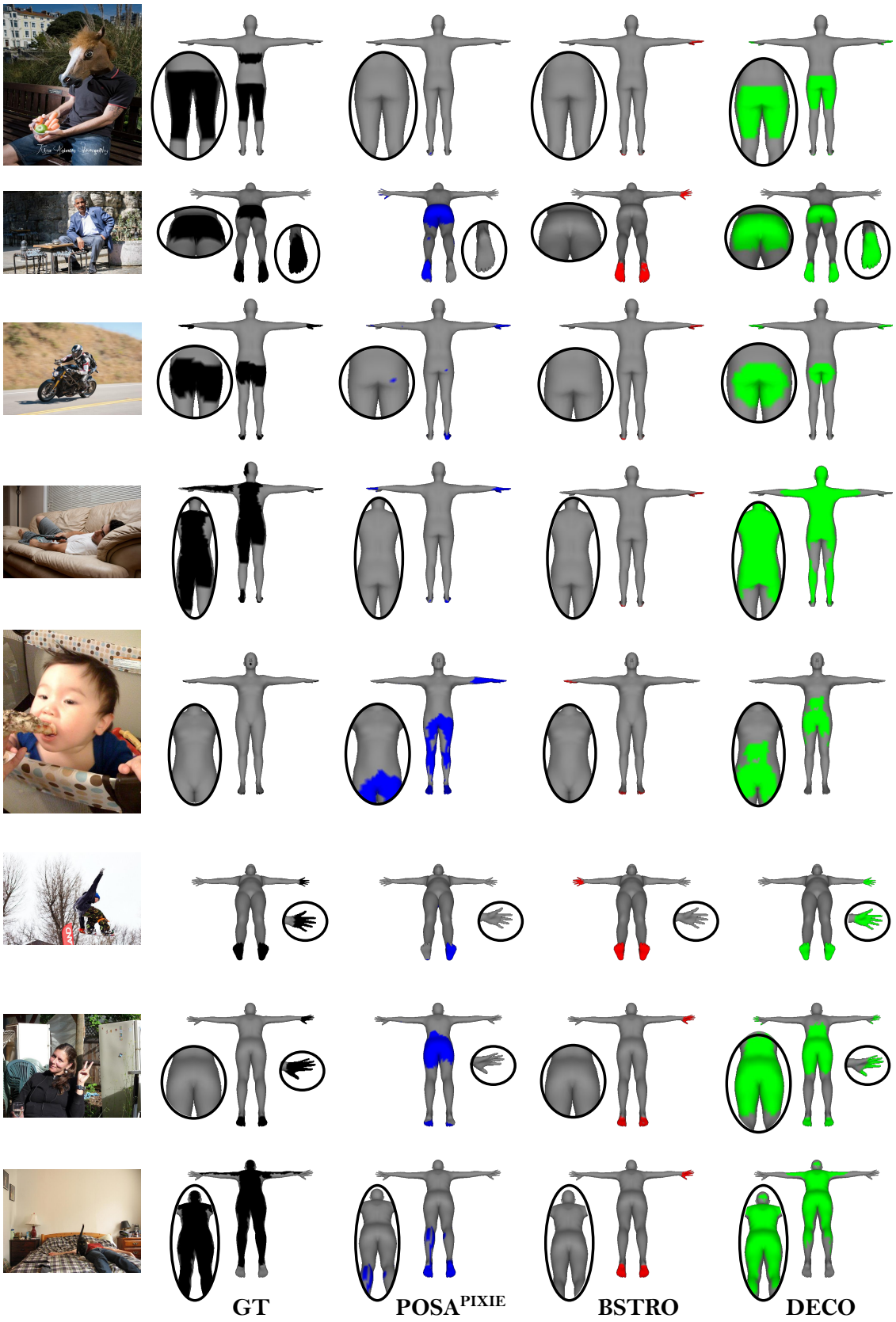

Figure 6: Additional qualitative evaluation of DECO (green), BSTRO (red) and POSA$^{\text{PIXIE}}$ (blue), alongside Ground Truth (**black**) on images from the DAMON dataset.

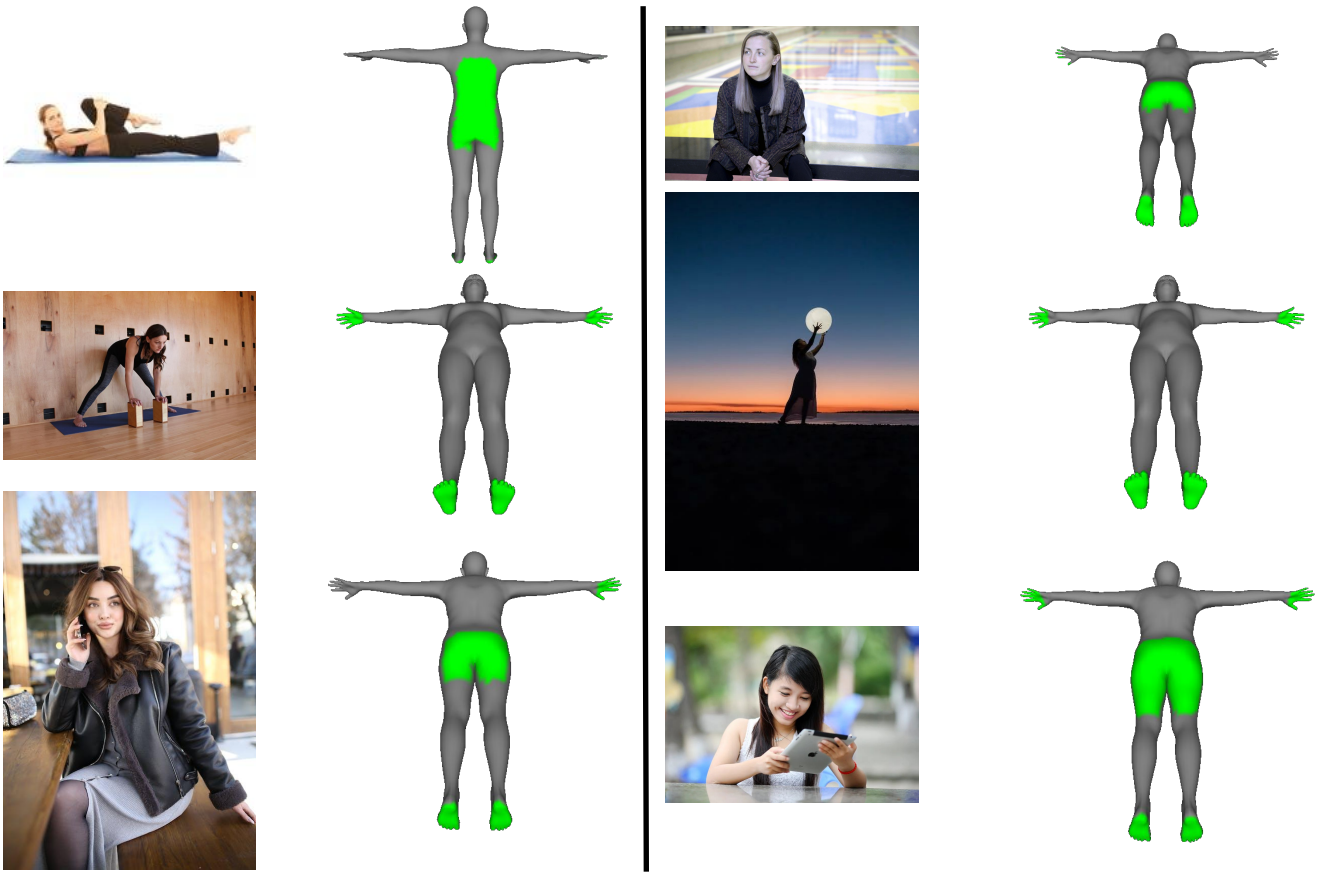

Figure 7: DECO predictions (in green) on Internet images, not seen during training.

References

[1] 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. 1

[2] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv:1505.04474*, 2015. 1

[3] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014. 3

[4] 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. 3

[5] 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. 3

[6] 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. 1

[7] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1