Supplementary Material for Detecting Human-Object Contact in Images

Yixin Chen[1†] Sai Kumar Dwivedi[2] Michael J. Black[2] Dimitrios Tzionas[3]

[1]Beijing Institute of General Artificial Intelligence, China
[†]Work done while interning at MPI-IS[2]
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]University of Amsterdam, the Netherlands

In Sec. 1, we introduce the detailed human-body part labels for human-object contact. In Sec. 2, we describe more details for the annotation protocol for "HOT-Annotated" and how we generate pseudo ground truth for "HOT-Generated". In Sec. 3, we report more implementation details. Sec. 4 shows more experimental results in the contact detection task, including failure cases, evaluation under different settings and attention maps, etc. In Sec. 5, we provide more details of the part-specific contact detector that we compare with HOT. In Sec. 6, we report more experiment details and results to illustrate the use of our HOT contact detection for 3D human pose estimation. Sec. 7 includes more details on how the HOT dataset can facilitate 3D contact estimation. Section 8 discusses more potential downstream applications for contact detection and qualitative results on self-contact and human-human contact. The use of existing assets is listed in Sec. 9.

1. Human Part Labels

For the contact estimation task, we want to know if contact takes place in the image, the area in which it takes place, as well as the body part that is involved.

To get the human part labels, we divide the parametric human body model, SMPL-X [7] into 17 parts, i.e.: Head, Chest, L_UpperArm, L_ForeArm, L_Hand, R_UpperArm, R_ForeArm, R_Hand, Buttocks, Hip, Back, L_Thigh, L_Calf, L_Foot, R_Thigh, R_Calf and R_Foot. This is based on the original part segmentation of SMPL-X, but for simplicity we unite certain parts (e.g., parts of the back across the spine), that even human annotators cannot easily differentiate. Figure S.1 shows the color-coded body parts, together with part labels, on the SMPL-X mesh.

2. Dataset Details

2.1. Contact Annotation for "HOT-Annotated"

We hire professional annotators to annotate the contact information for the in-the-wild images. The annotation pipeline is similar to semantic segmentation annotation but with different task requirements. In this section, we describe the instructions given to the annotators in detail.

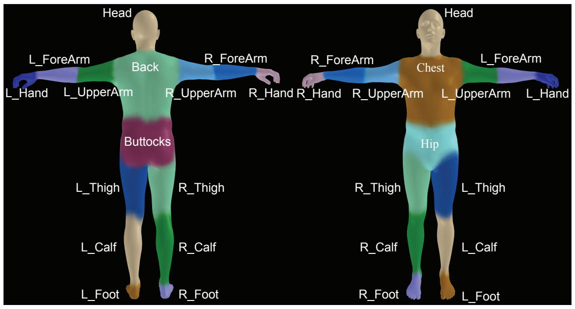

Figure S.1. The color-coded human parts with labels.

The overall annotation process includes two steps: (1) "segmenting" the image area for human-object contacts, and (2) assigning the human part label associated with the contact. In the first step, the annotators are asked to hallucinate the contact area in an image and draw a tight polygon around it. In the second step, the annotators pick a label for the contact area out of our pre-defined 17 human parts.

Determining the exact contact area between a human and an object is non-trivial, especially in the image space. Thus, we first perform a round of trial annotations, in which we test our annotation protocol, as well as train our annotators. We provide the following instructions to annotators:

- Contact areas between humans and objects are always occluded. Annotators should hallucinate the contact area in 3D, and then annotate its projection on the 2D image.
- A polygon annotation should cover only the subset of the human part that is in contact, and not the whole part. Note that this is different from part segmentation.
- There may be multiple contact areas between a single human and a single object.
- Only humans in the foreground should be considered; any humans in the background should be ignored.
- Contact areas that are occluded by another human or object should be ignored.
- Contact for body parts with extreme out-of-frame cropping, e.g., when only a hand is visible, should be ignored.

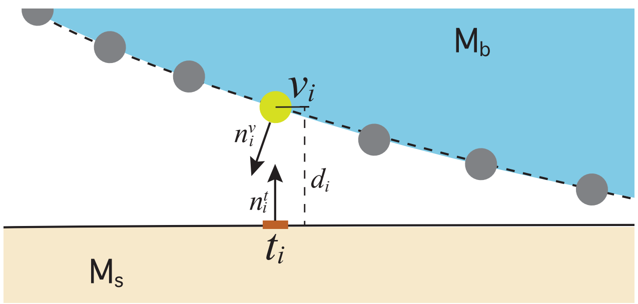

Figure S.2. Illustration of computing the properties involved in the contact annotation between the body mesh $\mathcal{M}_b$ and scene mesh $\mathcal{M}_s$ for "HOT-Generated".

– Human-human and self contact should be ignored.

After a full annotation round, we have two rounds of quality checks. In more detail, for every 3 annotators, there is 1 extra annotator that only conducts quality checks. The quality check verifies if the annotated polygon matches the contact area, if the contact label corresponds to the correct body part, if there are missing contact annotations (false negatives), if there are false positive contact annotations and if contact annotations are consistent across images.

2.2. Contact Generation for "HOT-Generated"

The PROX dataset [3] captures human subjects interacting with static scenes. Briefly, we use the reconstructed 3D human and scene meshes to first compute the human vertices that are in close 3D proximity to scene ones, and consider the former as contact vertices. We then render the respective triangles onto the 2D image to get automatic contact area annotations, as well as the associated body labels.

More specifically, the human pose and shape is represented with the SMPL-X body model with pose parameters, θ, and shape parameters, β. The 3D human mesh is denoted as $\mathcal{M}_b \in \mathbb{R}^{10475\times3}$. Each vertex, $v_i \in \mathbb{R}^3$, has a surface normal n_i^v and an associated human part label c_i. For each frame, given the estimated SMPL-X mesh, $\mathcal{M}_b$, and the scene mesh, $\mathcal{M}_s$, we first calculate the distance $\{d_i\}_{i=1}^{10475}$ from all human vertices $\{v_i\}_{i=1}^{10475}$ to the scene mesh $\mathcal{M}_s$. For each vertex v_i, we also find the closest triangle in $\mathcal{M}_s$, denoted as t_i, with surface normal n_i^t.

Then, a human vertex, v_i, is considered in contact if its distance to the scene, d_i, is below a threshold, and the surface normal, n_i^v, is in the opposite direction to the scene normal, n_i^t. Specifically, both of the following two constraints should be satisfied:

- **Distance constraint:** $d_i \leq \delta_d$, where the distance threshold δ_d is set to be $0.07m$ empirically;
- **Surface normal compatibility:** $\texttt{Angle}(n_i^v, n_i^t) \geq \delta_a$, where the $\delta_a = 110°$ is an angle threshold.

Figure S.2 demonstrates the criteria mentioned above.

Finally, for the contact vertices we find the respective triangles on the 3D body mesh, and render them separately per body part to get dense 2D contact areas. In this way, we automatically create pseudo ground truth for contact.

2.3. Annotation repeatability in "HOT-Annotated"

Annotating contact from images is a very challenging task. To verify the repeatability of the manual annotation, two new trained persons are hired to annotate 200 random images from "HOT-Annotated". We compare the labels to the ones collected by the annotators of Sec. 2.1. The agreement for body-part contact labels is 93.2%, and the agreement for pixel contact labels is 77.1%; this is comparable to the 82.4% agreement of the semantic-segmentation pixel annotations of ADE20K [14] in their experiment for consistency check across annotators.

2.4. Dataset Statistics by Splits

Current Human-Object Interaction (HOI) datasets have many walking, standing-up, or sitting-down poses (foot contact) or grasping poses (hand contact); this naturally biases the data distributions as shown in the main paper. Randomly spliting data into training, validation and testing sets naturally captures such biases, but the statistics are similar across these sets as can be seen from Figs. S.3 and S.4.

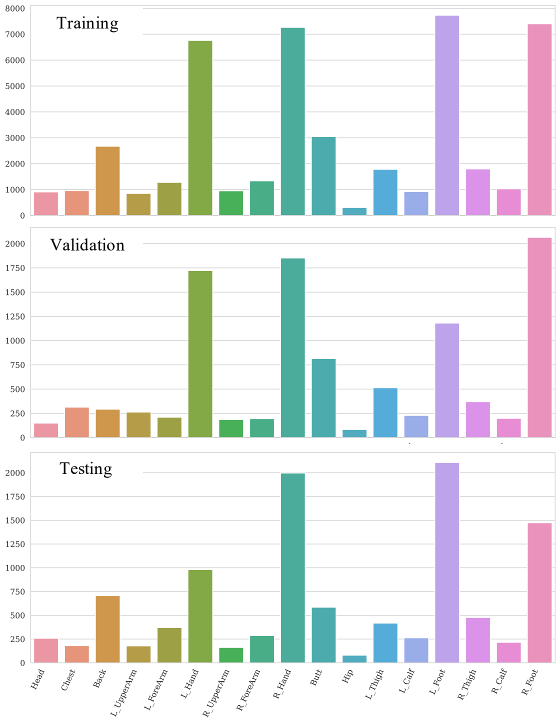

Figure S.3. Distribution of body-part labels for contact in "HOT-Annotated"; number of contact areas (Y-axis) for a certain body part (X-axis) in different data splits.

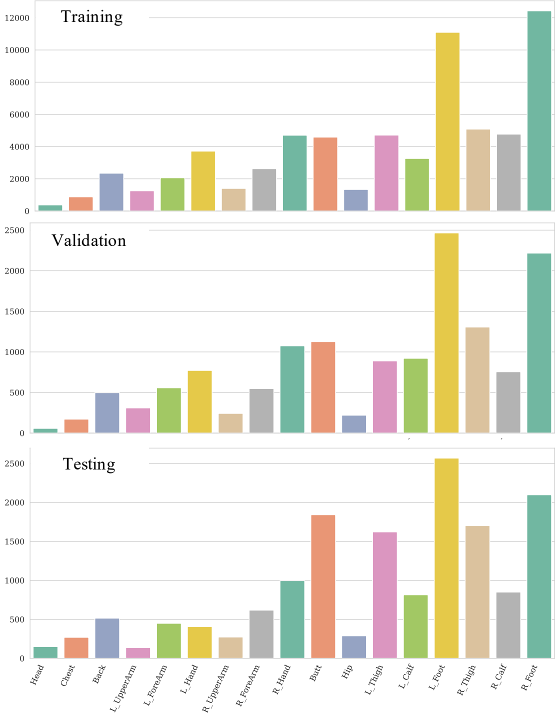

Figure S.4. Distribution of body-part labels for contact in “HOT-Generated”; number of contact areas (Y-axis) for a certain body part (X-axis) in different data splits.

3. Implementation Details

During training, the loss weight for the attention branch λ_a is set to be 0.1 for the first 10 epochs and 0 for the rest of the epochs. The loss weight λ_c for contact estimation is set to be 1. We use a pre-trained dilated ResNet-50 [12] as image encoder backbone. For the attention branch we use 3×3 convolutional layers with batch-norm and ReLU as image decoder, followed by another convolutional layer with kernel size 1 to make pixel-wise human part label classification. For the contact branch, we apply 3×3 convolutional layers with batch-norm and ReLU on the part-specific features, which we further concatenate along the channel axis. The weights of convolutional layers are different across human parts, so that the contact branch learns part-specific features under the attention guidance. Another convolutional layer with kernel size 1 is used to make pixel-wise contact label prediction. Since the background dominates the label ground truth for both human-part segmentation and contact estimation, we assign a smaller weight 0.02 for the background label and 1 for the rest of the labels in the cross-entropy loss. We re-scale all images to have their longer side 400 pixels long, and then pad, if necessary. Random flipping is applied for data augmentation. We train the model for 20 epochs on 4 NVIDIA-A100 GPUs with a batch size of 24. We use the SGD [9] optimizer, with an initial learning rate of 0.02 with polynomial decay following Zhou et al. [14].

We also report the model size for fair performance comparison during the experiments. Our model has a total of 50.2 million trainable parameters, whereas ResNet+PPM [13] has 46.7 million and ResNet+UperNet [11] has 64.2 million.

4. More Contact Detection Results

4.1. Failure Cases

Figure S.5 shows some examples of failure cases. We see that our model might struggle with occlusions, multiple persons or fine-grained contact areas. We also observe that the model sometimes fails in distinguishing left and right for the body parts. These point out that contact detection may benefit from future work on adding human pose information, multi-resolution reasoning and differentiating human-object contact with self-contact and person-person contact, but these are currently out of our scope.

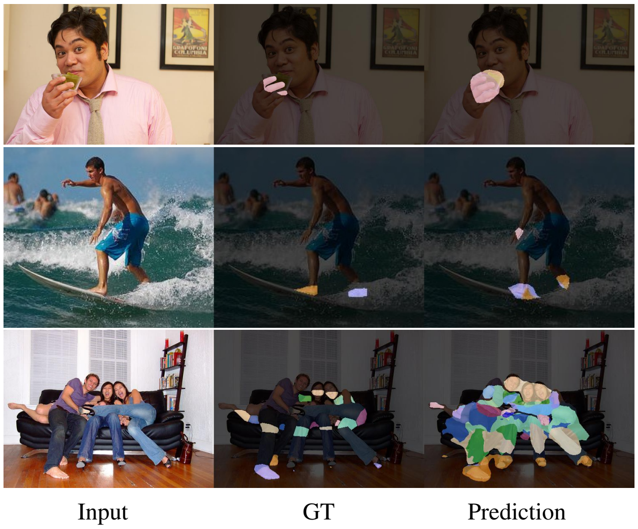

Figure S.5. Representative failure cases for our contact detector.

4.2. Model Performance under Various Settings

To better diagnose the model’s performance under different settings, we conduct the following two experiments.

1) The contact detection for different body parts. Quantitative results are shown in Tab. S.1. We can see that our methods performs better on the body parts with more data, e.g., hand, foot and butt, and fails in the body parts that naturally have less contact, e.g., hip and calf. This shows the importance of data balance when developing a general purpose contact detector.

2) We also evaluate the model’s performance with various contact area sizes, i.e., *small*, *medium* and *large*. The size thresholds are 0.052% and 0.22% based on the size distribution, which can be seen in the main paper. The quantitative

body part	Head	Chest	Back	L_UpperArm	L_ForeArm	L_Hand	R_UpperArm	R_ForeArm	R_Hand	Butt	Hip	L_Thigh	L_Calf	L_Foot	R_Thigh	R_Calf	R_Foot	Mean
SC-Acc. ↑	54.9	27.4	62.0	29.3	11.4	43.1	5.07	2.86	69.5	57.0	3.77	12.0	20.3	47.5	11.3	7.95	36.4	40.7
mIoU ↑	0.532	0.252	0.558	0.199	0.092	0.215	0.047	0.026	0.430	0.374	0.034	0.173	0.138	0.334	0.090	0.070	0.262	0.260

Table S.1. Contact estimation performance by different body parts on “HOT-Annotated”.

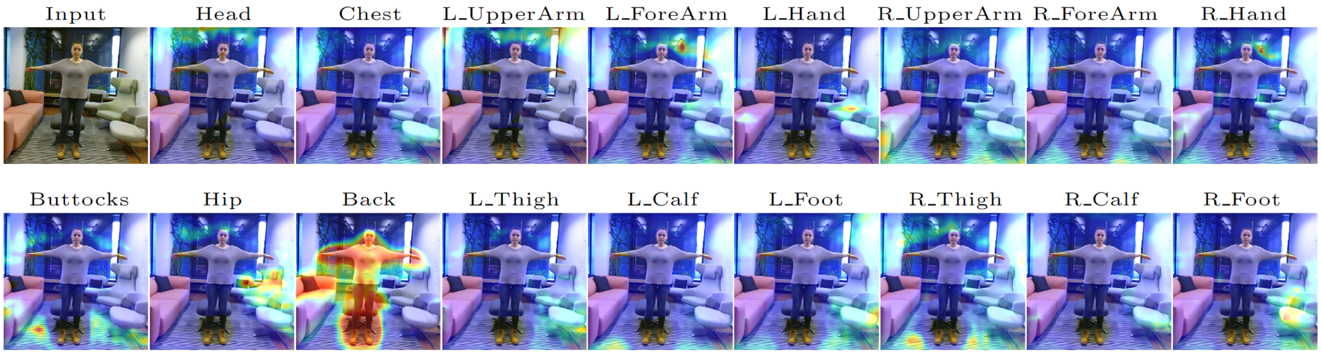

Figure S.6. Attention maps for “Ours$_{\text{pure\_att}}$”, visualized separately per body part.

results in Tab. S.2 show our model has decent performance on contacts with *medium* and *large* sizes, but cannot distinguish fine-grained contact with *small* areas. This indicates that contact detection will benefit from multi-resolution reasoning for different types of human-object contact.

Contact area	Sc-Acc.↑	mIoU↑	wIoU↑
small	21.6	0.020	0.025
medium	39.7	0.253	0.301
large	53.4	0.381	0.494
all	40.7	0.215	0.260

Table S.2. Contact estimation performance by contact area sizes on “HOT-Annotated”.

4.3. Attention without Human Part Supervision

Figure S.6 shows the learned attention maps for “Ours$_{\text{pure\_att}}$”. In this setting, no supervision is applied for the attention branch, which functions as an unsupervised pure soft-attention module. In contrast to “Ours$_{\text{Full}}$” where the attention focuses on areas around each human part (see Fig. 5 in the main paper), for “Ours$_{\text{pure\_att}}$” certain parts (e.g., the “Back” in this case) attend to the full body, while others can get distracted by the background.

5. Part-Specific Contact Detectors

5.1. Foot-Contact Detector

“ContactDynamics” [8] is a physics-based trajectory optimization method that generates physically-plausible motions. To this end, an intermediate step detects contact for the *toe* and *heel* joints of each foot. The authors use MoCap sequences to generate ground-truth contact for training such a detector using heuristics. The contact detector is a multilayer perceptron (MLP) that takes as input lower-body 2D joints in a temporal window, and outputs four contact labels (left/right toe, left/right heel) for the central frames.

For evaluation on PROX’s test set (aka “quantitative set”), we use OpenPose [1] to generate 2D keypoints and feed these into the pre-trained foot contact model. For a fair comparison with our HOT contact detector, we consider a foot to be in contact when at least one joint (either *toe* or *heel*) is in contact. Our detector achieves similar performance (HOT **59.2**% vs ContactDynamics [8] 58.6%); see the related discussion in Sec. 5.2 (i) of the main paper.

5.2. Hand-Contact Detector

“ContactHands” [6] detects hands as bounding boxes and classifies their contact state as “self-contact”, “person-person”, or “person-object” (hand-object) contact. Here we only consider the hands with hand-object contact label in the model output.

During evaluation, a detected hand-object contact from “ContactHands” is considered as a true positive if the hand bounding box and the ground-truth hand contact area overlap. For HOT, we consider our predicted hand contact area as a true positive if the Intersection-over-Union (IoU) with the ground-truth hand contact area is larger than 0.4. Experimental results show that our detector achieves similar performance (HOT **63.5**% vs ContactHands [6] 62.2%); see the related discussion in Sec. 5.2 (ii) of the main paper.

6. HOT for 3D HPS Estimation

In the main paper, we replace the heuristic contact in PROX [3] with our contact detection when estimating 3D humans from a color image. This tests the usefulness of our contact estimates for human pose estimation. In Tab. S.3 we

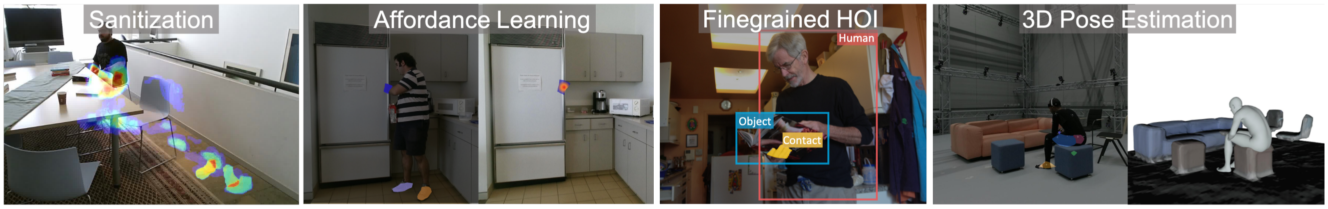

Figure S.7. Example downstream applications of contact detection.

report the full-performance comparison on PROX's "quantitative set"; *"All Contact"* considers all body vertices to be in contact.

Importantly, note that *V2V* is the most appropriate "pose" metric for *surface contact*, as vertices lie *on surfaces* that come in contact with objects. V2V numbers in Tab. S.3 show that detecting contact in images is promising and can be used to replace PROX's hand-crafted contact heuristics.

The rest of the metrics do *not* capture contact; they are reported for completeness. Procrustes (Pr.) *factors out global translation and rotation* to focus only on articulation; "pr.PJE" and "pr.V2V" are irrelevant for contact. Skeleton joints (PJE) lie *under the surface* of the body.

Method	*PJE* ↓	*pr.PJE* ↓	*V2V* ↓	*pr.V2V* ↓
No Contact	180.2	74.0	183.3	65.2
PROX [3]	170.9	**72.3**	174.0	63.4
All Contact	175.4	73.4	176.3	**64.0**
Predicted Contact	**171.3**	73.6	**172.3**	64.9
GT Contact	161.9	71.8	163.0	63.3

Table S.3. Contact-driven human pose and shape (HPS) estimation – results on PROX's "quantitative set". "Predicted Contact" refers to the contact label predicted by our HOT contact detector and "GT Contact" is the ground-truth contact label. "PROX" refers to use of PROX's manually annotated contact vertices. "PJE" refers to the Per-Joint Error, "pr" is Procrustes alignment, and "V2V" is the Vertex-to-Vertex error.

7. HOT for 3D Contact Estimation

In the main paper, we show that our HOT dataset facilitates dense 3D contact estimation on the human body from an image [4], by helping such models generalize better to in-the-wild images. Below we report how we generate the pseudo ground-truth for 3D contact using 2D HOT annotations, and discuss more experimental details.

Pseudo ground-truth generation: For "HOT-Annotated", we annotate (see Sec. 2.1) contact areas as 2D polygons in images and the body part that is involved in contact (see part segmentation in Sec. 1). For the annotated body part, for this experiment we consider all its vertices (see Fig. S.1) as contact vertices. The only exception is the hands and feet; we only consider the vertices on the inner palm and the sole of foot to capture the most common contact in daily life. The above results in a coarse pseudo ground-truth 3D contact map on the human body; for examples see Fig. S.8. We denote the pseudo ground-truth 3D contact for "HOT-Annotated" as "HOT-pGT".

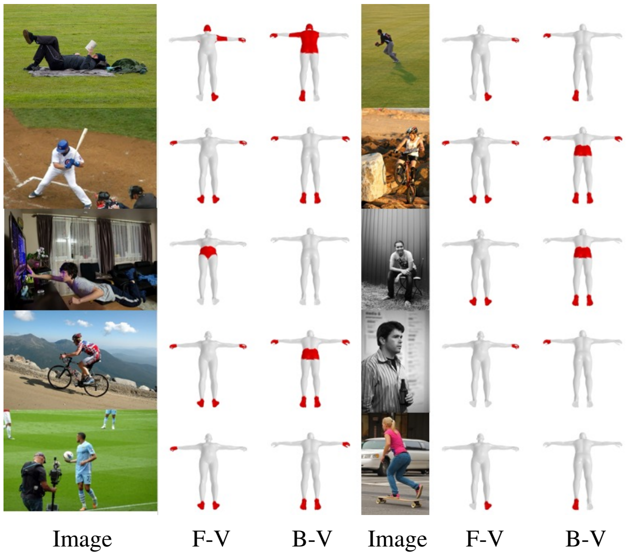

Figure S.8. Examples of the pseudo ground-truth 3D contact generated from "HOT-Annotated", i.e., HOT-pGT. F-V represents front-view and B-V represents back-view.

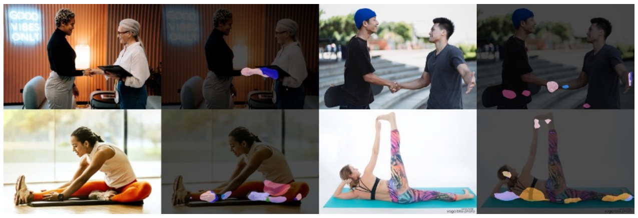

Figure S.9. Qualitative results of testing our model on self-contact and human-human contact.

Experimental details: The recent RICH dataset and BSTRO model [4] focus on dense 3D contact estimation on the human body from an image. To show the usefulness of our HOT dataset for this task, we employ the BSTRO model and extend its training dataset RICH with HOT-pGT. When training on RICH and HOT-pGT, we combine all the images from the training set of RICH and HOT, following their original training/validation/testing split. For faster convergence, we use the pre-trained model of BSTRO and fine-tune on the combination of RICH and HOT-pGT for 20 epochs. The learning rate is set to be 0.0001 and the the batch size is set to be 32. The rest of the network architecture and hyperparameters are the same as original BSTRO training [4]. We compare with the original BSTRO model, which is trained only on RICH. Each model is evaluated on the test set with the best performer from the validation set.

8. Contact Detection Applications

Contact detection is important for applications in many domains such as AR/VR, activity recognition, affordance detection, fine-grained human-object interaction detection (beyond bounding boxes), 3D human pose estimation and populating scenes with interacting avatars. Here we showcase several examples in Fig. S.7. For instance, one possible future direction is to extend the triplet definition of HOI <human/action/object> by adding contact as <human-part/contact-area/object>, which supports finer-grained HOI reasoning. Another application is detecting in videos the areas that people contact, and guiding human cleaners (AR) or robots with heatmaps for sanitization or contamination prevention.

We also test our human-object detector on images with self-contact and human-human contact; see some qualitative results in Fig. S.9. Although our model was not designed for such interaction scenarios, sometimes it can produce meaningful results, and sometimes it expectedly fails; this is a challenging and open problem. How to effectively combine different contacts and build a general-purpose contact detector would be interesting future work.

9. Use of Existing Assets

Our dataset HOT collects image data from PROX [3], V-COCO [2], HAKE [5] and Watch-n-Patch [10]. PROX is licensed under the terms of the Software Copyright License for non-commercial scientific research purposes. V-COCO is licensed under the terms of the CC-BY 4.0 License and HAKE is licensed under the terms of the MIT License.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. 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] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv:1505.04474*, 2015.

[3] 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.

[4] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022.

[5] 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.

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

[7] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[8] 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)*, volume 12350, pages 71–87, 2020.

[9] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[10] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-Patch: Unsupervised understanding of actions and relations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4362–4370, 2015.

[11] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, volume 11209, pages 418–434, 2018.

[12] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 472–480, 2017.

[13] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[14] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision (IJCV)*, 127(3):302–321, 2019.