Supplementary: ContactOpt: Optimizing Contact to Improve Grasps

Patrick Grady[1], Chengcheng Tang[2], Christopher D. Twigg[2], Minh Vo[2],
Samarth Brahmbhatt[3], Charles C. Kemp[1]

[1]Georgia Institute of Technology, [2]Facebook Reality Labs Research, [3]Intel Labs

1. Overview

This document provides additional implementation details and qualitative results to supplement the main paper. Additional details are given about the object subsets used while training networks on ContactPose (Section 2) and HO-3D (Section 3). Training settings for the DeepContact estimation network (Section 4), hyperparameters for the contact optimization procedure (Section 5), and more details on the evaluations (Section 6) are discussed. Finally, we provide more qualitative examples of the full pipeline running (Section 7).

2. Selection of Objects from ContactPose

The ContactPose dataset contains grasps from 50 participants across 25 items in two grasp intents, handoff and use. The dataset additionally includes right-handed, left-handed, and bimanual grasps where both hands are used. In all ContactOpt experiments, only right-handed grasps were used. These criteria result in the use of only 1768 grasps of the dataset's full 2258 grasps.

3. Image-Based Pose Estimator Baseline

The HO-3D dataset contains many pre-grasp and post-grasp poses where the hand is not in contact with the object. Because the ContactOpt is intended for poses in contact, we filter out frames where the minimum distance between the ground truth hand and object surfaces is greater than 2 mm. We perform all experiments using our own train/test split created from the official HO-3D training split, since the labels for the official testing split are not publicly released and the evaluation server does not allow frame filtering. This train/test split divides sequences with filtered frames by an 80/20 ratio. The testing split contains unseen camera angles of objects in the training split.

We use the baseline image-based pose estimator from Hasson *et al*. [2]. Because the accompanying trained model is trained on the entire HO-3D official training split, we re-train this model using the accompanying code at [2] on our training split. The re-trained model achieves an MPJPE of 57.7 mm on our testing split, which is comparable to the 55.2 mm MPJPE achieved by Hasson *et al*.'s released model on the HO-3D official test split. These MPJPE numbers are calculated without doing translation, scale, or rotation alignment.

4. DeepContact

DeepContact is based on PointNet++ by Qi et al [5]. However, the radius of the grouping layers is modified as the size of the hand and object are smaller than the objects PointNet++ was originally trained on. The radius of the first grouping layer ball query is set to $0.1m$, and the second layer is set to $0.2m$.

Instead of directly regressing the contact values from $[0, 1]$, DeepContact estimates the contact value as a classification task. The range $[0, 1]$ is evenly split into 10 bins. The training loss is weighted to account for class imbalance.

Before contact prediction, the hand and object meshes are jointly normalized so that the centroid of the object lies at $(0, 0, 0)$ and the centroid of the hand lies along the vector $(1, 0, 0)$. The network is trained using the ADAM optimizer with a learning rate of 0.01 dropping to 0.001 after 5 epochs.

5. Pose Optimization Hyperparameters

In addition to the hyperparameters listed the main paper, the pose optimization considers the following additional settings:

1. λ_H: Weight of hand contact loss in optimization objective. Qualitatively, hand contact is often more important to grasp reconstruction than object contact. Often grasps on different objects have different object contact maps, but very similar hand contact maps. Additionally, since DeepContact is applied to novel objects but the hand representation does not change across datasets, estimated hand contact has more reliable predictions.

2. λ_O: Weight of object contact loss in optimization objective.

3. λ_{pen}: Weight of penetration cost in optimization objective. This term penalizes interpenetration of the hand and the object as described in the paper.

4. λ_{opt_trans}: Translation optimization weight. This provides a relative weight for the different intrinsic scales of optimized parameters. This controls the amount the root translation vector is updated. High weights allow the optimizer to move the hand root long distances.

5. λ_{opt_rot}: Rotation optimization weight. This provides a relative weight for the different intrinsic scales of optimized parameters. This controls the amount the root rotation angle is updated.

6. $n_{restart}$: Number of random restarts. To recover from local minima, for example if the hand is "stuck" inside the object, the optimization is initialized multiple times with random perturbations. The optimization with the lowest final loss is taken.

7. $d_{restart}$: Mean perturbation distance of a random restart. For each restart, normally distributed noise is added to the hand translation. This is set proportionally to the initial translation error.

In the main paper, three tasks are evaluated:

- Refining ContactPose Dataset Poses (**Small**): The already high-quality poses from the ContactPose ground truth annotations are refined further using the measured contact.
- Perturbed ContactPose (**Large**): The original dataset poses are perturbed by adding translation and pose noise. ContactOpt is evaluated in recovering the initial poses. Hand and object contact is inferred using DeepContact.
- Image-Based Pose Estimates (**Image**): Hand pose estimates are generated by a baseline pose estimator on the HO-3D dataset. These poses are the initialization for refinement, and contact is inferred using DeepContact.

The hyperparameters for each evaluation phase are shown in Table 1

6. Evaluations

6.1. Physics Simulation

Multiple prior works have attempted to quantify the realism of grasps using a physics simulation engine [3, 4, 6].

	Small	**Large**	**Image**
λ_H	0.0	2.0	3.0
λ_O	1.0	1.0	1.0
λ_{pen}	600.0	600.0	300.0
λ_{opt_trans}	0.03	0.3	0.3
λ_{opt_rot}	1.0	1.0	1.0
$n_{restart}$	1	8	8
$d_{restart}$	0cm	4cm	2cm

Table 1. Hyperparameters of ContactOpt for evaluation settings in the main paper.

Following these, we obtained results using a PyBullet-based [1] evaluation environment. Results were obtained that were numerically favorable to our approach.

However, we found that the results were sensitive to simulation details, such as global orientation, friction coefficients, and object mass. In particular, the handling of penetration due to hand compliance had a large effect on results.

Due to the sensitivity of the simulation, physics evaluation results were not included. The team instead decided to focus on evaluation methods with more straightforward interpretation.

6.2. Perceptual Evaluation

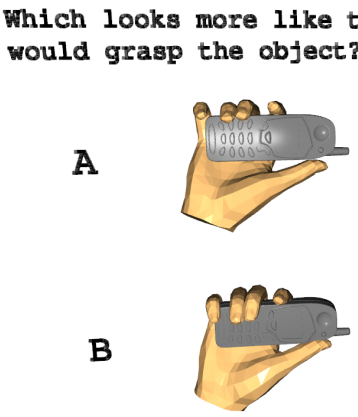

Figure 1. Human evaluators are shown a 3D viewer when judging perceptual grasp quality. The evaluators can rotate/pan/zoom the viewer to fully understand each grasp. Once their choice is made, they press the corresponding key on their keyboard.

To evaluate the holistic grasp quality, we used nine human evaluators to judge quality. This study was approved by the academic institution's Institutional Review Board (IRB). The study recruited participants who were not familiar with the research. Each evaluator judged 75 grasps for each of the evaluation tasks.

The study was conducted as a two-alternative force

choice (2AFC) test. This format was selected instead of a numerical rating test due to the subtle differences between grasps when performing small-scale refinement. We still found that non-experts had difficulty comparing grasps with small differences, so pairs with less than 5mm of MPJPE movement were filtered out. This accounts for 47% of the results for the small-scale evaluation. To ensure that the grasps were judged by pose alone, contact maps were not shown on the hand or object.

7. Further Qualitative Results

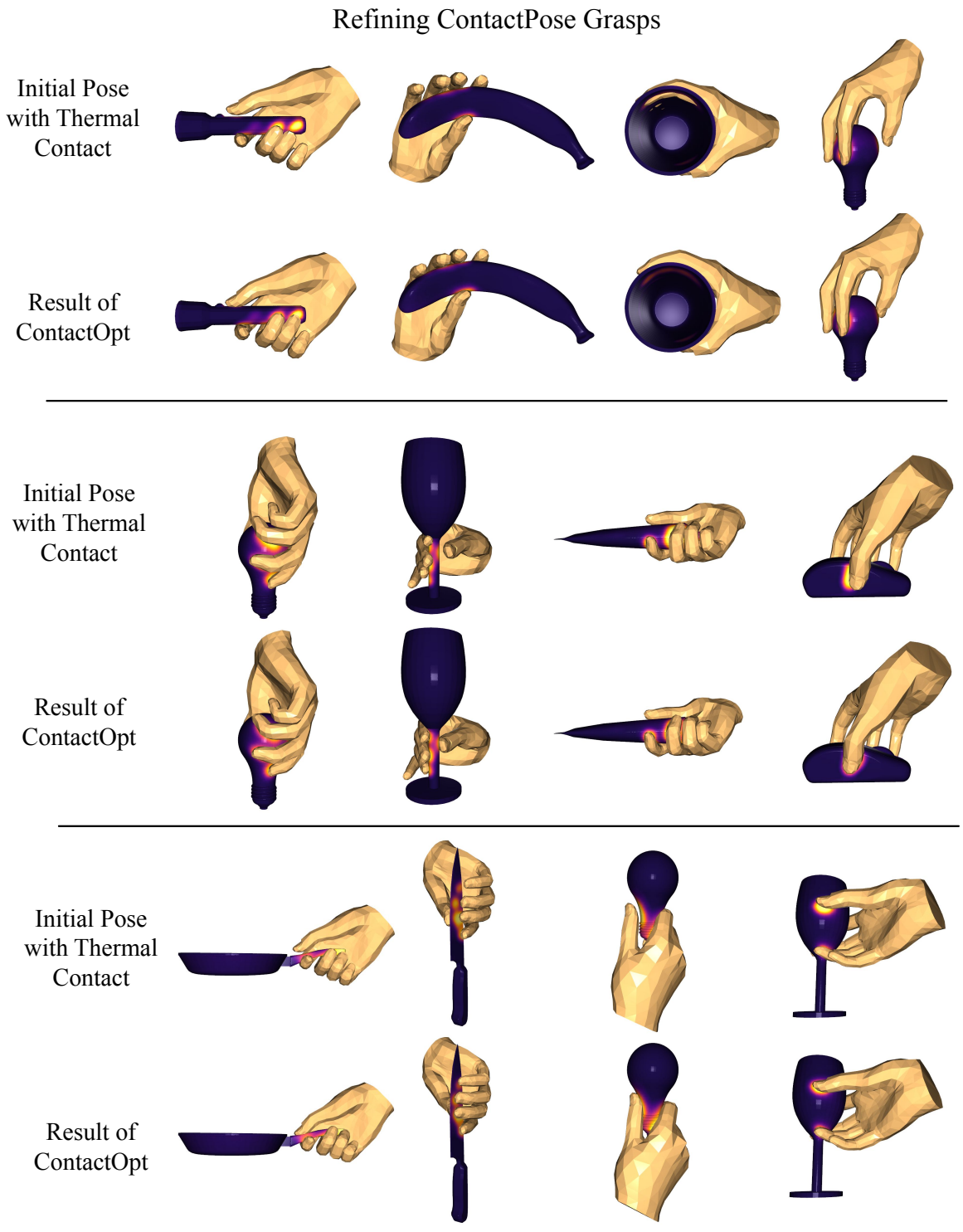

Figure 2. Examples of refining hand poses when the ground-truth thermal contact map is given. Better agreement between the hand pose and object pose is shown, in addition to the resolution of heavy interpenetration or gaps between the hand and object.

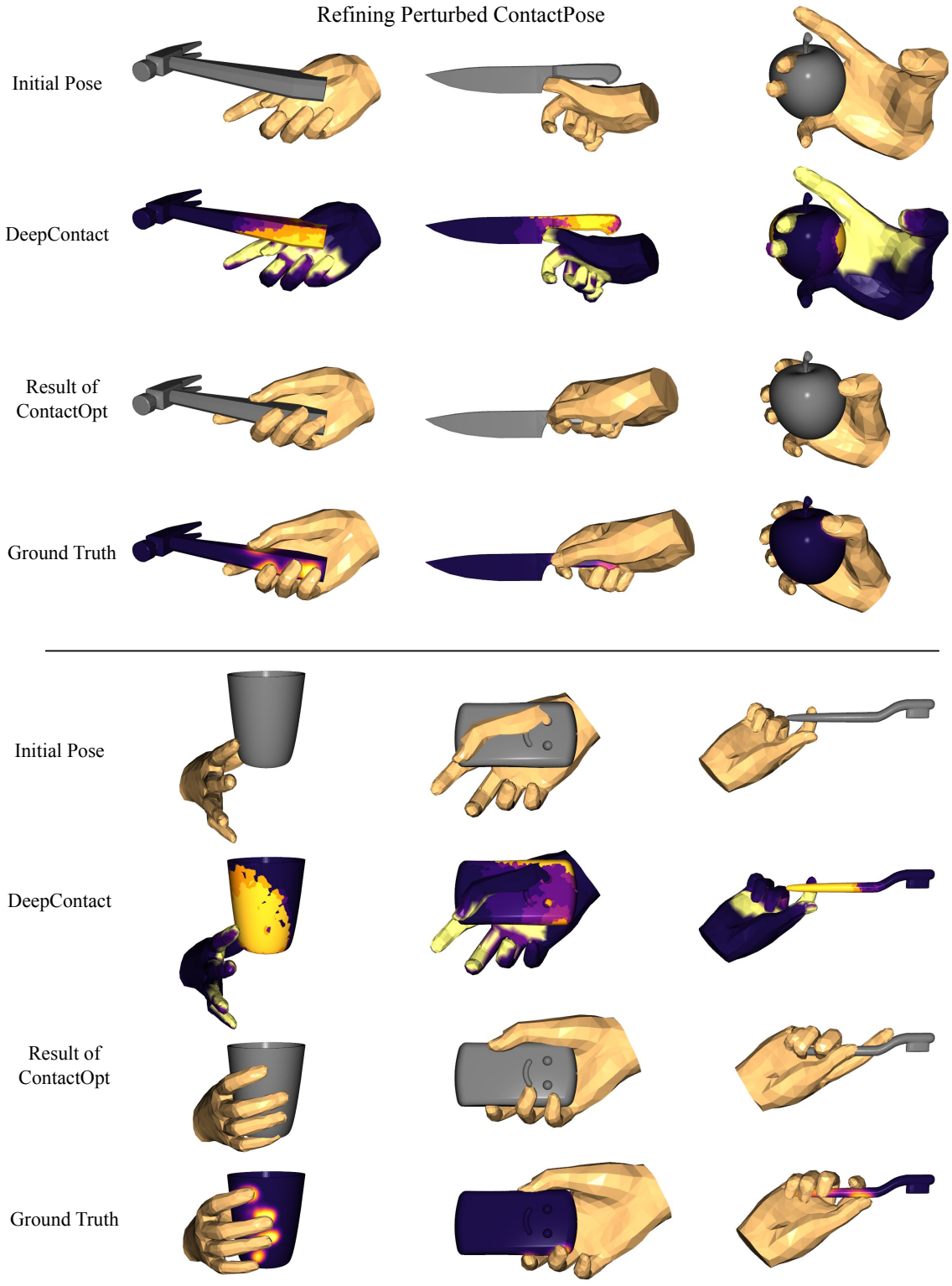

Figure 3. Examples of refining poses from the Perturbed ContactPose dataset. This dataset has been generated by heavily perturbing the ContactPose annotations. The last case represents a failure that is typical on thin objects, where the small volume leads to low penetration cost and causes unrealistic interpenetration.

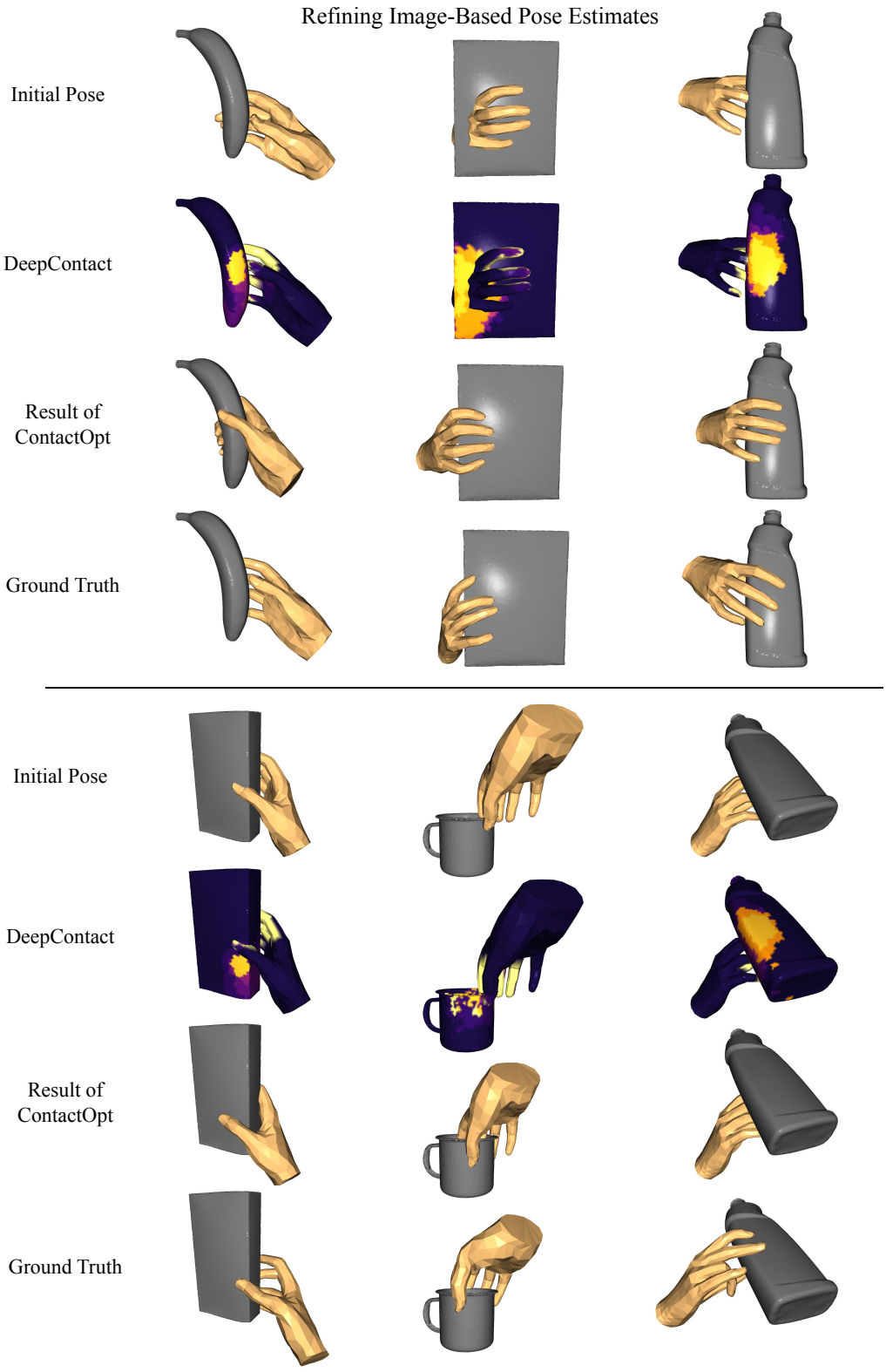

Figure 4. Examples of refining poses generated by an image-based pose estimator. The objects and poses are from the HO-3D dataset, which DeepContact was not trained on. Note that the contact predictions are often more sparse. The last sample represents a failure where the starting pose was too far away for the optimizer to produce a good result.

References

[1] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2019. 2

[2] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 571–580, 2020. 1

[3] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019. 2

[4] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *arXiv preprint arXiv:2008.04451*, 2020. 2

[5] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 1

[6] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016. 2