ContactOpt: Optimizing Contact to Improve Grasps

Patrick Grady[1], Samarth Brahmbhatt[2], Chris Twigg[3], Minh Vo[3], Chengcheng Tang[3], Charlie Kemp[1]

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

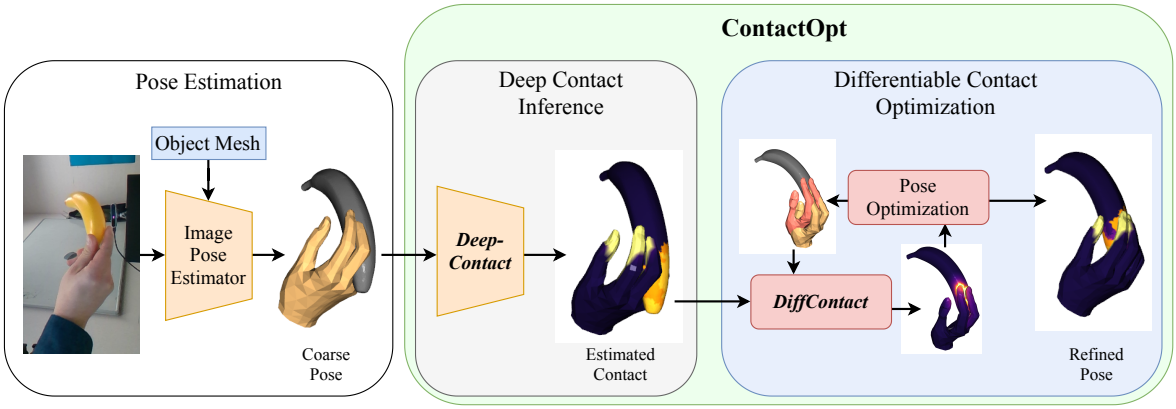

Figure 1: ContactOpt pipeline. Left: A pose estimator generates a hand pose. Middle: DeepContact estimates where contact should occur (target contact). Right: The hand pose is optimized to achieve target contact via a contact model (DiffContact).

Abstract

Physical contact between hands and objects plays a critical role in human grasps. We show that optimizing the pose of a hand to achieve expected contact with an object can improve hand poses inferred via image-based methods. Given a hand mesh and an object mesh, a deep model trained on ground truth contact data infers desirable contact across the surfaces of the meshes. Then, ContactOpt efficiently optimizes the pose of the hand to achieve desirable contact using a differentiable contact model. Notably, our contact model encourages mesh interpenetration to approximate deformable soft tissue in the hand. In our evaluations, our methods resulted in grasps that better matched ground truth contact, had lower kinematic error, and were significantly preferred by human participants. Code for this work will be publicly released.

1. Introduction

The availability of data, hand and body models, and learning algorithms has fueled a growing interest in capturing, understanding, and simulating hand-object interactions [46, 14, 55, 12, 43, 4]. Recent algorithms can predict hand and object pose increasingly accurately from an image. However, inferred poses continue to exhibit sufficient error to cause unrealistic hand-object contact, making downstream tasks in simulation, virtual reality, and other applications challenging.

A key issue is that physical contact is sensitive to small changes in pose. For example, less than a millimeter change in the pose of a fingertip normal to the surface of an object can make the difference between the object being held or dropped on the floor. In addition to physical implausibility, lack of contact and other small-scale phenomena can reduce the perceptual realism of rendered poses.

In this paper we present ContactOpt, an algorithm that improves the quality of hand-object contact by refining hand pose. When given a hand mesh and an object mesh, ContactOpt infers where contact is likely to occur and then optimizes the hand pose to achieve this contact.

As shown in Figure 1, ContactOpt consists of two main components, DeepContact and DiffContact. DeepContact is a deep network that takes pointclouds sampled from the hand and object meshes as input and outputs a value from 0 to 1 for each mesh vertex indicating whether it should be in contact. We trained DeepContact using ground truth contact data for stable grasps from the ContactPose dataset [4]. DiffContact is a differentiable contact model that takes hand and object meshes as input and outputs a value from 0 to 1 for each mesh vertex indicating whether it is in contact. ContactOpt uses gradient-based optimization to find pose, translation, and rotation parameters for the MANO hand model [35] that improve the match between contact estimated by DiffContact and target contact from DeepContact or an alternate source, such as ground truth contact.

Notably, ContactOpt takes into account soft tissue deformation in the hand. The inner surface of a human hand undergoes significant deformation when making contact with objects. For example, the finger pad can deform 2-3 mm under 1 N force and the palm can deform 5 mm [10, 5]. DiffContact permits interpenetration between the hand and object meshes with up to 2 mm of penetration being considered ideal contact. In addition, ContactOpt's gradient-based optimization uses a loss function that only penalizes penetration greater than 2 mm.

In our evaluations, ContactOpt resulted in hand poses that had lower mean per-joint position error (MPJPE), better matched ground truth contact (precision & recall), and were significantly preferred by human participants in two-alternative forced choice (2AFC) tests. ContactOpt also outperformed RefineNet, a recent state-of-the-art network from GrabNet that improves hand poses without explicitly modeling contact [46].

We conducted two types of evaluations to assess ContactOpt's performance. For the first type of evaluation, we evaluated ContactOpt's ability to refine hand pose estimates with *small* inaccuracies. This presents methodological challenges due to limits in the precision of dataset ground truth annotations. To overcome this, we used the ContactPose dataset, which has both pose estimates *and* measured contact data obtained via thermal imagery. We had ContactOpt refine these hand pose estimates with respect to ground truth contact. The refined hand poses better matched ground truth contact and were preferred by human participants, demonstrating that ContactOpt can improve state-of-the-art pose estimates from existing datasets.

For the second type of evaluation, we evaluated ContactOpt's ability to refine hand pose estimates with *large* inaccuracies. First, we confirmed that ContactOpt can improve perturbed hand poses from the ContactPose dataset to better match ground truth contact and reduce kinematic error. Second, we used ContactOpt to refine hand pose estimates from an existing hand pose estimation network (Hasson et al. [16]) applied to the HOnnotate dataset [14]. ContactOpt's refined hand poses had lower kinematic error, were preferred by human participants, and matched more closely to previously observed hand contact patterns (Figure 2). ContactOpt also outperformed RefineNet [46] (an end-to-end grasp refinement neural network) with respect to both measures. This demonstrates ContactOpt's value as a post-processing stage for existing hand-object pose estimation algorithms for which it has not been specifically trained. Since ContactOpt operates on hand and object meshes, it has the potential to improve the output of a variety of recent image-based estimation methods [17, 16, 14, 48], while avoiding potential generalization issues associated with operating on images.

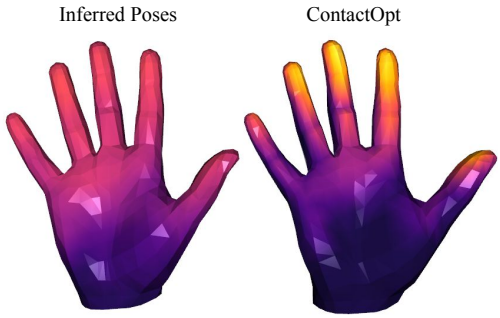

Figure 2: Frequency of hand contact calculated with poses inferred with an image-based pose estimator [16] (left) and after refinement with ContactOpt (right). Note the increase in contact on the finger pads, and the closer match to thermal ground truth hand contact patterns from [4].

In summary, our contributions follow:

- We show that methods that explicitly consider hand-object contact can improve hand pose estimates at both coarse (≈cm) and fine (≈mm) spatial scales.
- We show that optimizing hand-object contact can improve the visual realism of rendered grasps.
- We show that optimizing hand-object contact can reduce kinematic error from hand pose estimation.
- We present DeepContact, a deep network that estimates where contact should occur across the surfaces of inaccurately aligned hand and object meshes.
- We present DiffContact, a differentiable contact model that estimates where contact is occurring between hand and object meshes.
- We present ContactOpt, an algorithm that performs gradient-based optimization to improve hand-object contact by refining hand pose.

2. Related Work

In this work, we use likely contact and a contact model to improve the pose of a hand grasping an object. Applications in computer vision, animation, and robotics have driven interests in hand-object interaction tracking from different angles, e.g., recovering poses from input images or generating grasps based on object pose and geometry. The information about contact is playing an increasingly important role for hand-object interaction tracking, grasp generation and multiple other related applications.

Datasets of hand-object contact. Recently, there has been a focus on collecting datasets that include interac-

tions between hands and objects. FreiHand [55] uses multiple cameras to extract high-quality annotations using the MANO model, but do not include the object pose. HOnnotate [14] optimizes simultaneously for both hand and object poses from RGB-D sensors. FHAB [11] leverages a unique magnetic tracking system to infer the pose of hand and object even under occlusion. GRAB [46] uses professional optical motion capture to collect a dataset of people grasping and manipulating objects. The work additionally infers contact from the proximity of hand and object. However, these estimates may be noisy due to the very high pose accuracy required. There may be error introduced due to shape/scale calibration, marker-offset estimation, and the inability to measure soft-tissue deformation. While these datasets can usefully supervise many tasks like image-based hand-object pose estimation, they do not provide the ground truth measurements needed to reason about the contact interface.

Datasets for contact directly obtained on objects [22, 2] and hands [44] are complementary to datasets on hand-object poses. The ContactPose dataset [4] is unique in capturing both ground-truth thermal contact maps, as well as hand and object pose. The participants held a static grasp for each of 25 objects while being captured using multiple RGB-D cameras. The object was tracked using motion capture, and the hand pose was estimated by aggregating predictions across all frames from an off-the-shelf RGB hand pose estimator [6]. A thermal camera measured the body heat transferred from the participant's hand to the object, providing ground truth about the actual contact. A limitation of the method is that the 3D hand pose accuracy is bounded by the accuracy of the hand pose estimation, so there may be discrepancies between the contact map and the MANO posed hand mesh.

Image-based hand-object pose estimation. There is an extensive body of work on estimating the pose of the hand using a variety of input modalities, including: gloves with markers or sensors [51, 15, 12], depth/RGB-D input [53, 40, 39, 27, 49, 1, 45, 50, 42, 47], and RGB or monochrome images [26, 55, 6, 7, 53, 9, 38], with an increasing focus on hand-object interaction [13, 34, 29, 30, 39, 48, 17, 16, 20, 9, 14]. Researchers have long realized that inferring and enforcing contact is important for hand-object interaction tracking [34], and it remains a challenging task, particularly in the absence of depth data. For RGB-D hand tracking, hand-object contact modeled as fingertip to object distance was part of the energy function during optimization with Gaussian Mixture Models in [39]. For image-based prediction, skeletal hand poses [48, 9] or MANO [35] hand model parameters [17, 16] are predicted jointly with object geometry or pose in an end-to-end manner. Despite sharing a joint latent space, since the output representations for the hand and object are decoupled, there can be *relative* errors in the poses, leading to unrealistic grasps. Even though contact can be encouraged at training time, these networks have no method of **enforcing** alignment at test time. Our work complements these existing methods by leveraging the strength of their joint hand-object pose prediction, but uses explicit contact inference and enforcement to achieve higher quality grasps.

Grasp synthesis. A related task is to generate plausible grasps of a given object geometry for a human or robot kinematic model. Contact points are first sampled for generating plausible robotic grasps [37, 23, 36] and animation [52, 25]. Most of those focus on generating stable grasps based on heuristic sparse contact points. In GanHand [8], a dataset of affordances and grasps was proposed to generate plausible human grasps based on input images. The works that are most similar to ours are ContactGrasp [3] and GRAB [46]. In ContactGrasp [3], dense ground truth contact maps from ContactDB are used to generate plausible grasps for a given object geometry; they use the less-realistic HumanHand model [24], and because the ContactDB dataset lacks ground truth hand poses they cannot compare against ground truth or condition on images as we do. In GRAB [46] the authors leverage their collected data to generate compelling grasps for a variety of objects. Their RefineNet, which improves the quality of a grasp given an initial pose, has similarities to our approach, but is data-driven with poses and fixed contact patterns aggregated over GRAB rather than optimization-based with contact estimated separately for each grasp. The method only explicitly considers hand geometry, and because it is fully learned, may have less ability to generalize. We show comparisons against this approach when applied to image-base inference tasks in Sec. 4.

Contact in human pose. Besides for hand-object interaction, contact is informative for full human body poses in human-environment interaction [28]. Inferred contact constraints are used in [33] to improve body pose estimation from videos to mitigate artifacts such as feet sliding. Coarse contact points are used in generating human poses interacting with scenes [54, 18, 41]. Our work leveraging fine-grain contact information to improve hand pose in hand-object interaction tracking is in line with and applicable to context-aware full-body pose estimation and generation.

3. Methods

We represent the grasp with an object mesh $\mathbf{O}$ and a MANO [35] hand mesh $\mathbf{H}$. $\mathbf{H}$ is described by parameters $\mathbf{P} = \left(\theta, \beta, {}^{O}t_H, {}^{O}R_H\right)$, consisting of pose, shape, translation, and rotation w.r.t. object respectively. Pose θ is represented as a 15-dimensional PCA manifold, which lowers

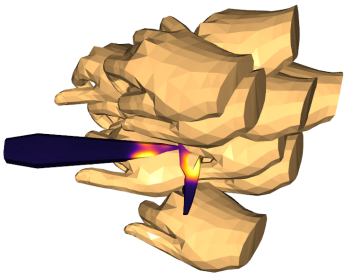

Figure 3: Example of multiple hand poses from Perturbed ContactPose, all generated from a single dataset sample

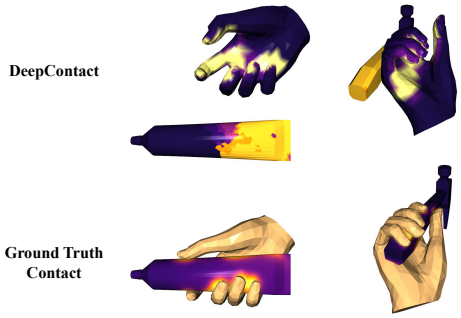

Figure 4: Top: DeepContact predicts contact maps for the hand and object as if they were aligned. Poses from Perturbed ContactPose. Bottom: Ground truth poses and thermal contact.

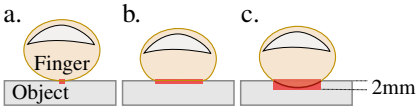

Figure 5: When a human finger contacts a rigid object, point contacts (a) are rare. More commonly, the soft tissue in the finger conforms to the surface (b) so that the contact spreads across a larger area. While the MANO mesh does not locally deform to match the surface, we can encourage the optimizer to create matching area-based contact by marking vertices as "in contact" even when they are 2mm inside the surface (c).

the high-dimensional joint angle representation to a compressed space of normal hand poses.

Given a noisy estimate of $\mathbf{P}$ (which typically comes from a image-based algorithm), we seek a better grasp by exploiting the hand-object contact information. Figure 1 shows an overview of our approach. In the following, we describe our learned contact map estimation module DeepContact (Section 3.1) and our differential contact model DiffContact (Section 3.2) that is iteratively updated according to the optimized hand pose to reproduce the estimated contact (Section 3.3).

3.1. DeepContact: Learning to Estimate Contact

Given an object mesh $\mathbf{O}$ and and hand mesh $\mathbf{H}$ with potentially inaccurate pose $\mathbf{P}$, this module learns to infer contact on the hand and object vertices. The motivations for explicitly predicting contact and choosing this input representation are discussed in Section 1 and empirically validated in Section 4.

We represent the meshes $\mathbf{H}$ and $\mathbf{O}$ as point clouds, and use PointNet++ [32] to predict contact. The object pointcloud contains 2048 points randomly sampled from the object. The hand point cloud contains all 778 vertices of the MANO mesh. We employ the "mesh" features, training loss, and discrete contact representation of Brahmbhatt et al. [4]. The "mesh" features capture distances from each point to the hand and object, as well as normal information. Additionally, we include a binary per-point feature indicating whether the point belongs to the hand or the object. The network predicts contact as a classification task, where the range $[0, 1]$ is split into 10 bins. We train DeepContact with the standard binary cross-entropy loss.

Similarly to GrabNet [46], we train this module on a dataset of randomly perturbed hand poses from the ContactPose dataset, which we call **Perturbed ContactPose**. The hand mesh is modified by adding noise to the parameters $\Delta\theta \sim \mathcal{N}(0, 0.5)$, $\Delta^O t_H \sim \mathcal{N}(0, 5)$ cm, and $\Delta^O R_H \sim N(0°, 15°)$. Object contact is supervised with ground-truth thermal contact from ContactPose. To generate the target hand contact map, we run DiffContact described in Section 3.2. By applying multiple perturbations to each grasp, a training/testing split of 22k/1.4k grasps is generated.

Figure 3 shows some example perturbations, and Figure 4 shows an example contact prediction. Hand and object poses that are farther from a particular grasp tend to result in larger and more diffuse areas of predicted contact.

3.2. DiffContact: Differentiable Contact Model

To optimize the hand parameters $\mathbf{P}$ to match the fixed contact predicted by the learned contact estimation module, we need to model a hand-object contact map according to the current meshes $\mathbf{O}, \mathbf{H}(\mathbf{P})$. We propose a contact model using *virtual capsules* that is differentiable w.r.t. $\mathbf{P}$, as shown in Figure 6. Our virtual capsules have useful attraction extended beyond the surface (which a binary proximity would not) and approximately model soft hand tissue deformation.

More concretely, we place a virtual capsule at every object vertex $\mathbf{v}_i^O$ and orient it along the object surface normal $\mathbf{n}_i^O$. This capsule has a principal line segment defined by $\mathbf{v}_i^O + \alpha\mathbf{n}_i^O$, $\alpha \in [-c_{\text{bot}}, c_{\text{top}}]$ defines the vertical extent of of capsule. Let $\phi(\mathbf{x})$ be the Euclidean distance from a 3D point $\mathbf{x}$ to this line segment, $\phi(\mathbf{x}) = \min_{\alpha \in [-c_{\text{bot}}, c_{\text{top}}]} ||\mathbf{x} - (\mathbf{v}_i^O + \alpha\mathbf{n}_i^O)||_2$. The contact is defined to be uniformly 1 for points such that $\phi(\mathbf{x}) < c_{\text{rad}}$ and falls off proportionally with distance outside c_{rad} as $\frac{c_{\text{rad}}}{\phi(\mathbf{x})}$. Let $v_j^H(\mathbf{P})$ be the hand vertex at pose $\mathbf{P}$ with the smallest distance ϕ to the object vertex v_i^O. The contact map at the object vertex v_i^O is expressed as:

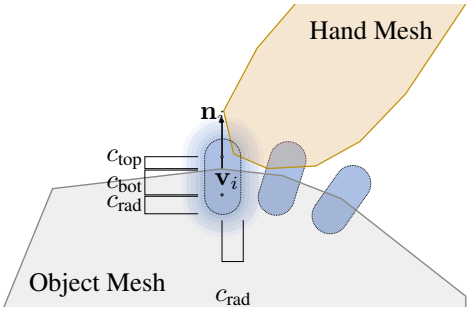

Figure 6: Capsule fields are placed on each vertex of the object, aligned with the vertex normal. If there are any hand vertices inside the capsule, the object point is marked as being in full contact, $c = 1$.

Figure 7: Left: Hand and object from ContactPose dataset. Center: Thermal contact from ContactPose. Right: Contact from the proposed differentiable contact model (Section 3.2).

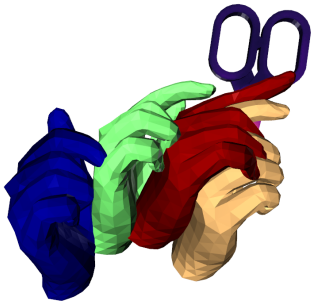

Figure 8: Optimization of hand pose to match the ground truth thermal contact map. From left to right: hand pose at selected iterations during optimization.

$$C_O\left(\mathbf{v}_i^O; \mathbf{P}\right) = \min\left(\frac{c_{\text{rad}}}{\phi(\mathbf{v}_j^H(\mathbf{P}))}, 1\right). \tag{1}$$

The same procedure can be used to calculate the contact map on the hand surface. We choose an asymmetric $c_{\text{bot}} > c_{\text{top}}$ such that the region considered "in contact" extends farther inside the mesh than outside, which approximately models soft hand tissue deformation as shown in Figure 5. In our experiments, $c_{\text{top}} = 0.5$mm, $c_{\text{bot}} = 1$mm, and $c_{\text{rad}} = 1$mm. As the total capsule depth inside the object is $c_{\text{bot}} + c_{\text{rad}} = 2$mm, this conservatively matches the $2-3$mm finger pad deformation found in the biomechanics literature [5, 10].

Figure 7 shows an example of object contact computed with this model. Because the generated contact has a gradual dropoff, this provides gradients for optimization. Additionally, the resulting contact maps have diffuse edges, which appear visually similar to thermal contact maps [2, 4]. The generated contact is an *area* instead of a single point.

3.3. Contact Optimization

We now iteratively optimize hand mesh parameters $\mathbf{P}$ to minimize the difference between the current contact maps $C_H(\mathbf{P})$ and $C_O(\mathbf{P})$ computed using DiffContact (Section 3.2), and the target contact maps $\hat{C}_H$ and $\hat{C}_O$ as predicted by DeepContact, Sec. 3.1, or from ground truth thermal contact. Our contact loss for the object surface is:

$$E_O(\mathbf{P}) = \begin{cases} \lambda|C_O(\mathbf{P}) - \hat{C}_O| & \text{if } C_O(\mathbf{P}) < \hat{C_O} \\ |C_O(\mathbf{P}) - \hat{C}_O)| & \text{otherwise} \end{cases} \tag{2}$$

Here we use $\lambda > 1$ to penalize "missing" contacts (where the target contact map is higher than the value estimated by DiffContact from the current pose $\mathbf{P}$) more heavily than "unexpected" contacts. This is based on the empirical observation that it is visually worse for the hand to "hover" over the object than to be slightly interpenetrating. We apply a corresponding loss $E_H(\mathbf{P})$ to penalize differences between the target hand contact map $\hat{C}_H$ and $C_H(\mathbf{P})$. We use $\lambda = 3$ in both cases.

We also include a term that discourages penetrations beyond c_{pen}. For each object vertex $\mathbf{v}_i^O$, object surface normal $\mathbf{n}_i^O$, and nearest hand vertex $\mathbf{v}_i^H(\mathbf{P})$, the penetration loss is defined as

$$E_{pen}(\mathbf{P}) = \sum_i \max\left(0, (\mathbf{v}_i^O - \mathbf{v}_i^H(\mathbf{P})) \cdot \mathbf{n}_i^O - c_{pen}\right), \tag{3}$$

where $c_{pen} = 2$ mm, following [5]. The final loss is $E(\mathbf{P}) = E_H(\mathbf{P}) + \lambda_O E_O(\mathbf{P}) + \lambda_{pen} E_{pen}(\mathbf{P})$.

The loss is minimized by the ADAM optimizer [21] using gradients computed by PyTorch Autograd automatic differentiation [31]. We use a learning rate of 0.01 and optimize for 250 iterations. Optimizing a minibatch of 64 $\{\mathbf{H}, \mathbf{O}\}$ hand-object pairs takes approx. 4s (roughly 62ms for each). We reweight the gradients for the different components of $\mathbf{P}$ to match their intrinsic scales. See the supplementary material for more details.

Random Restarts. Since the contact optimization is local, a poor initialization (*e.g.* initial hand position on the wrong side of an object) can result in the optimizer settling into a bad local minimum. We avoid this by applying the pose optimization to several perturbations of the provided pose and select the result with the lowest loss.

4. Evaluation

We evaluate ContactOpt on the ContactPose and HOnnotate datasets. In each case, the refined hand mesh is eval-

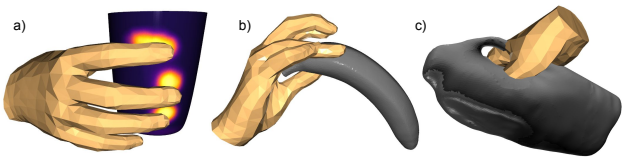

Figure 9: Examples of contact inaccuracy in dataset ground truth: (a) ContactPose [4] (alignment offset), (b) HOnnotate [14] (hand self-penetration, hand-object gap), and (c) FHAB [11] (hand-object penetration).

uated using the following metrics.

- **Intersection Volume** (cm^3): The intersection volume of meshes $\mathbf{H}$ and $\mathbf{O}$ is calculated from their mesh union. The per-sample standard deviation is also shown.
- **Mean Per-Joint Position Error (MPJPE)** (mm): The average L2 per-joint kinematic error with respect to the ground truth hand [19].
- **Contact Coverage** ($\%$): The percentage of hand points between -2mm and +2mm of the object surface.
- **Contact Precision/Recall** ($\%$): This quantifies how well the contact from the refined hand mesh matches the thermal contact map. A binary object contact map is obtained by considering the object points within ± 2 mm of the hand surface to be in contact. Precision and recall are calculated by comparing this to the thermal contact map thresholded at 0.4, following [4].
- **Perceptual Evaluation** ($\%$): Six evaluators who were unfamiliar with the research were recruited to judge the relative quality of grasps in two-alternative forced choice tests (2AFC). Each participant was shown two hand-object pairs and asked to judge "Which looks more like the way a person would grasp the object?". In pilot studies, we found that non-experts had difficulty comparing grasps with small differences, so pairs with less than a 5 mm MPJPE difference were removed. For each method, the evaluators judged 75 pairs of grasps with an equal number randomly selected for each object. The mean and 95% confidence intervals are shown. More details of this evaluation can be found in the supplementary material.

4.1. Refining Small Inaccuracies

Recent hand-object interaction datasets use a variety of techniques to capture hand and object pose, such as magnetic trackers (FHAB [11]), multi-view reconstruction from RGB-D cameras (HOnnotate [14]) and/or motion capture systems (ContactPose [4], GRAB [46]). Despite using high quality sensors, errors on the centimeter-level are not uncommon (Figure 9).

However, when considering the realism of grasps, *millimeters matter*. Gaps between the hand and object result in unstable grasps, but are also visually unsatisfying. Similarly, unrealistic penetration is physically impossible. Notably, a few millimeters of these errors is inconspicuous in a euclidean error metric, but results in an impossible grasp.

ContactOpt can be used to resolve these types of errors when applied to already high-quality poses provided by dataset annotations.

Refining ContactPose Dataset Poses: Millimeter-scale refinement is demonstrated by refining the ContactPose annotated hand meshes. Rather than estimating contact using DeepContact, the ground truth object contact map is used as a target. Hand contact is not used. Table 1 and Figure 11 show the results of this refinement.

Both contact recall and precision metrics increase, demonstrating that ContactOpt improves the self-consistency between ground truth contact and mesh poses. There is both less unwanted contact as well as excess contact.

However, it is difficult to quantify the holistic quality of a grasp. We perform a Perceptual Evaluation where human participants choose the most natural-looking grasp. As shown in Table 1, participants favored the refined grasps at over a 2:1 ratio. ContactOpt is able to consistently resolve cases of millimetric penetration or under-shoot and pull the fingers into realistic contact with the object, which is noticed by the participants.

This demonstrates that contact and accurate poses can be used together to achieve even higher quality than is possible with pose alone.

4.2. Refining Large Inaccuracies

4.2.1 Perturbed ContactPose

We test the full ContactOpt pipeline on Perturbed ContactPose (Section 3.1), which contains poses with a mean error of $8cm$. This tests the ability to improve hand poses with large errors. Results are shown in Figure 12 and Table 1.

Despite being initialized from a heavily misaligned hand pose, the pipeline is still able to reduce kinematic error (MPJPE) by approximately 70% and improves perceptual grasp quality dramatically. Additionally, the refined meshes are more consistent with the ground truth contact maps, even though they are not provided to the algorithm.

However, some kinematic error remains. Qualitatively, this is because the objects have many valid grasp modes (*i.e.* grasping an apple in any rotation), which are not possible to recover from the inaccurate initial pose. Although most refined meshes are visually high quality, often a slight translation results in a large kinematic error.

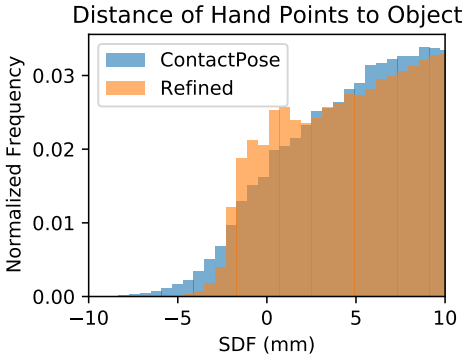

Figure 10: Distance of hand points to object surface, before and after refinement of ContactPose. Note that unrealistic deep interpenetrations (negative) have been mostly eliminated while the fraction of vertices near the surface of the object $[-2\text{mm}, 2\text{mm}]$ has increased.

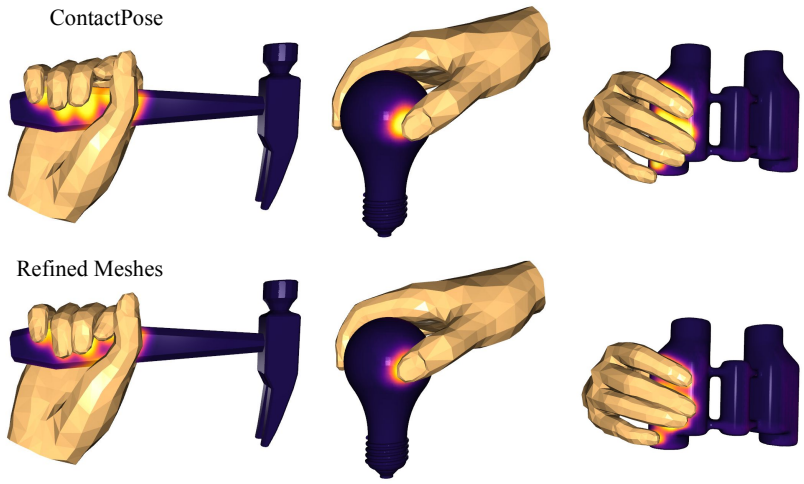

Figure 11: Top: Original meshes from ContactPose with misalignment between hands and contact maps. Bottom: After refinement using ContactOpt. See Sec. 4.1.

Dataset	ContactOpt Refinement	Intersection Volume (cm^3) ↓	MPJPE (mm) ↓	Score (%) ↑			
				Perceptual	Coverage	Precision	Recall
ContactPose[4]	×	2.45 ± 1.99	0.00	31.1 ± 4.7	6.9	64.6	34.0
	✓	1.35 ± 0.90	8.06	**68.9 ± 4.7**	8.9	**75.9**	**50.0**
Perturbed ContactPose	×	8.46 ± 16.49	79.89	N/A	2.3	9.9	11.5
	✓	12.83 ± 8.00	25.05	N/A	19.7	**38.7**	**54.8**

Table 1: Effect of ContactOpt refinement on the ContactPose ground-truth (top 2 rows) and Perturbed ContactPose dataset (bottom 2 rows). The precision and recall scores quantify (Sec. 4) agreement with the measured contact map. ContactOpt improves both perceptual quality and contact agreement.

4.2.2 Image-Based Pose Estimates

One of the applications of ContactOpt is to refine the predictions from an image-based pose estimator. In this task, 3D hand and object pose are often estimated using CNNs. For approaches that operate on single-frame RGB images, errors in the multiple-centimeter range are typical, leading to physically implausible grasps. We apply ContactOpt to the poses generated by a single-frame RGB pose estimator. Note that in this setting, there are no image-based constraints placed on the optimization, thus allowing total freedom of pose refinement.

We use the baseline pose estimation network from Hasson et al. (2020) [16] and retrain it on a training split of the HOnnotate dataset. As the network's object predictions are often unstable, the object shape and pose are taken from ground truth. Additionally, poses where the ground truth is not in contact are filtered out. More details can be found in the supplementary material.

We demonstrate that DeepContact is able to generalize well to new datasets. Despite being trained on the Perturbed ContactPose dataset, it can still perform acceptably on HOnnotate, which has both different objects and features dynamic grasps. Generally, since hand and object geometry is mostly consistent across datasets, the domain gap is smaller than modalities such as RGB, where learned methods often must be completely retrained. We qualitatively find that DeepContact is able to transfer hand contact more reliably than object contact, as the hand representation (MANO) is consistent across datasets.

Results from this task are found in Table 2. Human evaluators favored the refined grasps over the initial grasp estimates by a ratio of 6:1. Additionally, the frequency of contact across the hand for the refined grasps (Figure 2) is similar to ground truth frequencies of contact, while the frequency of contact for originally inferred grasps lacks key features such as greater frequency of contact with the tips of the fingers.

As the dataset contains shapes with many grasp modes (*i.e.* boxes), DeepContact may have difficulty predicting the correct grasp location from a low quality inferred grasp. Figure 12 shows a refined grasps with high perceptual quality but a large MPJPE error metric. Despite this, ContactOpt

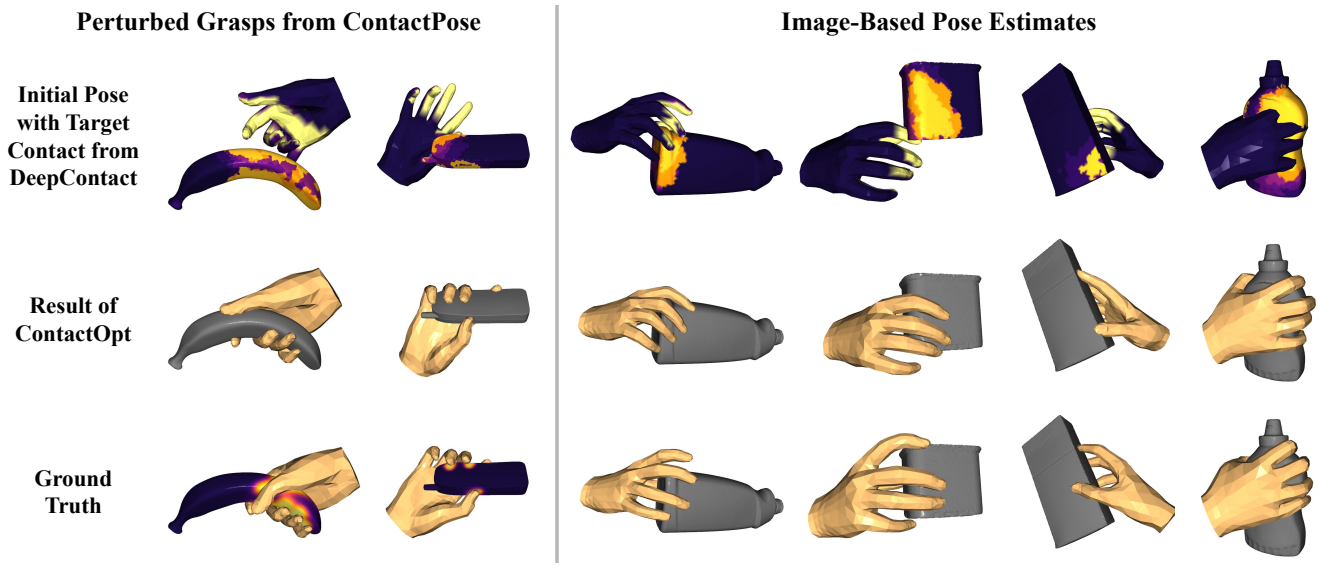

Figure 12: Application of ContactOpt to poses from Perturbed ContactPose and poses generated by an image-based pose estimator. Top: Hand pose inferred by pose estimator with hand and object contact inferred by DeepContact. Middle: Resulting poses after refinement with ContactOpt. Bottom: Ground truth annotations from ContactPose/HOnnotate dataset. The first column presents an example where the refined grasp is of higher perceptual quality, but as DeepContact estimated a different grasp mode, the grasp has high kinematic error. More examples are available in the supplementary material.

Method	**Intersection Volume** (cm^3) ↓	**MPJPE** (mm) ↓	**Score (%)** ↑	
			Perceptual	**Coverage**
Image Pose Estimator [16]	15.3 ± 21.1	57.7	*reference*	4.4
RefineNet (n=3) [46]	13.8 ± 19.0	56.3	70.8 ± 4.3	5.3
RefineNet (n=10) [46]	11.6 ± 18.5	64.1	N/A	3.9
ContactOpt (ours)	6.0 ± 6.7	**48.1**	**84.9 ± 3.4**	14.7
HOnnotate Ground Truth [14]	1.9 ± 2.8	0.0	N/A	2.5

Table 2: Effect of RefineNet and ContactOpt algorithms on the hand pose predicted by Hasson et al. [16] on the HOnnotate dataset. The perceptual studies compare refined poses against the original image-based estimates. The ContactOpt refinement achieves the lowest MPJPE and is favored by human evaluators.

is still able to lower the mean kinematic joint error by 20%.

4.2.3 Baseline Hand Refinement

We also compared ContactOpt to a baseline hand pose refinement method. RefineNet [46] is an end-to-end model trained on the GRAB dataset [46] to refine initial coarse grasp proposals. Given a hand and object mesh, the network predicts pose, rotation, and translation updates. As RefineNet is an iterative method, it is benchmarked with 3 iterations (following the paper) and 10 iterations.

5. Conclusion

We introduce ContactOpt, a method to refine coarsely aligned hand and object meshes. DeepContact estimates likely contact on both the hand and the object. DiffContact then estimates contact based on the current mesh pose. The error between these two estimates is used to optimize hand pose to achieve the target contact.

We show that ContactOpt is able to refine both dataset-quality meshes when ground truth thermal contact is provided, as well as improve pose estimations from images, even when tested on a novel object set.

Additionally, we demonstrate experimentally that the generated grasps are of high quality and physically plausible. Our method increases agreement with the target contact map, has lower kinematic error, and produces visually satisfying grasps for human evaluators.

References

[1] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. *arXiv preprint arXiv:2007.11432*, 2020. 3

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2019. 3, 5

[3] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. *arXiv preprint arXiv:1904.03754*, 2019. 3

[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 *The European Conference on Computer Vision (ECCV)*, August 2020. 1, 2, 3, 4, 5, 6, 7

[5] John-John Cabibihan, Deepak Joshi, Yeshwin Mysore Srinivasa, Mark Aaron Chan, and Arrchana Muruganantham. Illusory sense of human touch from a warm and soft artificial hand. *IEEE Transactions on neural systems and rehabilitation engineering*, 23(3):517–527, 2014. 2, 5

[6] Z Cao, G Martinez Hidalgo, T Simon, SE Wei, and YA Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5041, 2020. 3

[9] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020. 3

[10] Brygida M. Dzidek, Michael J. Adams, James W. Andrews, Zhibing Zhang, and Simon A. Johnson. Contact mechanics of the human finger pad under compressive loads. *J. R. Soc. Interface*, 14, 2017. 2, 5

[11] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 3, 6

[12] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 38(4), 2019. 1, 3

[13] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1475–1482. IEEE, 2009. 3

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3196–3206, 2020. 1, 2, 3, 6, 8

[15] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 3

[16] 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. 2, 3, 7, 8

[17] 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, 3

[18] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics (TOG)*, 39(4):53–1, 2020. 3

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 6

[20] 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. 3

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[22] Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey, and Holly Rushmeier. Tactile mesh saliency. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 3

[23] Quoc V Le, David Kamm, Arda F Kara, and Andrew Y Ng. Learning to grasp objects with multiple contact points. In *2010 IEEE International Conference on Robotics and Automation*, pages 5062–5069. IEEE, 2010. 3

[24] A. T. Miller and P. K. Allen. GraspIt! a versatile simulator for robotic grasping. *IEEE Robotics Automation Magazine*, 11(4):110–122, 2004. 3

[25] Igor Mordatch, Zoran Popović, and Emanuel Todorov. Contact-invariant optimization for hand manipulation. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 137–144, 2012. 3

[26] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand

tracking from monocular RGB. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 3

[27] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1284–1293, 2017. 3

[28] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[29] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*, pages 2088–2095. IEEE, 2011. 3

[30] Paschalis Panteleris, Nikolaos Kyriazis, and Antonis A Argyros. 3d tracking of human hands in interaction with unknown objects. In *BMVC*, pages 123–1, 2015. 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019. 5

[32] 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. 4

[33] 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*, pages 71–87. Springer, 2020. 3

[34] Javier Romero, Hedvig Kjellström, and Danica Kragic. Hands in action: Real-time 3D reconstruction of hands in interaction with objects. In *2010 IEEE International Conference on Robotics and Automation*, pages 458–463. IEEE, 2010. 3

[35] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245, 2017. 1, 3

[36] Carlos Rosales, Josep M Porta, and Lluıs Ros. Global optimization of robotic grasps. *Proceedings of robotics: science and systems VII*, 2011. 3

[37] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008. 3

[38] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. *arXiv preprint arXiv:2003.09282*, 2020. 3

[39] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from RGB-D input. In *European Conference on Computer Vision*, pages 294–310. Springer, 2016. 3

[40] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013. 3

[41] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39(4):54–1, 2020. 3

[42] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 824–832, 2015. 3

[43] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019. 1

[44] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698, 2019. 3

[45] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-ICP for real-time hand tracking. *Symposium on Geometry Processing (Computer Graphics Forum)*, 2015. 3

[46] 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)*, 2020. 1, 2, 3, 4, 6, 8

[47] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3786–3793, 2014. 3

[48] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019. 2, 3

[49] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (ToG)*, 35(6):1–11, 2016. 3

[50] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169, 2014. 3

[51] Robert Y Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, 28(3):1–8, 2009. 3

[52] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (TOG)*, 31(4):41, 2012. 3

[53] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. BigHand2.2M benchmark: Hand pose dataset

and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4866–4874, 2017. 3

[54] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6194–6204, 2020. 3

[55] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 3