Understanding Human Hands in Contact at Internet Scale

Dandan Shan[1], Jiaqi Geng*[1], Michelle Shu*[2], David F. Fouhey[1]

[1]University of Michigan, [2]Johns Hopkins University

{dandans, jiaqig, fouhey}@umich.edu, msh1@jhu.edu

Abstract

Hands are the central means by which humans manipulate their world and being able to reliably extract hand state information from Internet videos of humans engaged in their hands has the potential to pave the way to systems that can learn from petabytes of video data.

This paper proposes steps towards this by inferring a rich representation of hands engaged in interaction method that includes: hand location, side, contact state, and a box around the object in contact. To support this effort, we gather a large-scale dataset of hands in contact with objects consisting of 131 days of footage as well as a 100K annotated hand-contact video frame dataset. The learned model on this dataset can serve as a foundation for hand-contact understanding in videos. We quantitatively evaluate it both on its own and in service of predicting and learning from 3D meshes of human hands.

1. Introduction

The hand is the key to how humans interact with the world. If machines are to understand our actions and intentions as well as the world we have build for and with our hands, they must have a deep understanding of our hands. For instance, in Figure 1, we can readily recognize that there are two hands (one left and one right), opening a bag and even imagine how one might pull up the flap. The goal of this paper is to build the foundation for studying hands engaged in interaction with objects at Internet scale.

Hand analysis is, of course, an area of long-standing interest in the field with work on pose estimation [41, 34], reconstruction [21, 42], and grasp analysis [30, 7]. These approaches, however, have largely focused on in-lab settings, often with a pre-localized hand or in settings with limited variety. While there has been substantial progress, deploying these on the rich world of Internet videos [3, 39] poses a challenge due to the dizzying diversity in viewpoint and context. A single system must handle data ranging from a fifty pixel high hand in a cooking video to an enormous thousand pixel high hand closeup showing DIY.
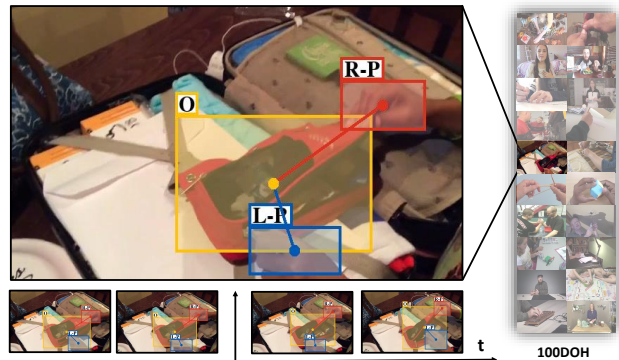

Figure 1. The goal of this paper is to infer a rich representation for helping understand hands engaged in contact with the world at Internet scale. Our system produces a rich output in terms of hand location (boxes), side (left/right), contact state (here – a portable object) and what object each hand is in contact with. To support this, we collect a new large-scale hand video interaction dataset, 100 Days of Hands (*100DOH*). We use this, plus videos from VLOG [14], to make a 100K image dataset annotated with our rich representation.

Our work aims to enable hand analysis at Internet scale and diversity. To this end, we introduce a model, described in Section 4 that identifies, for every single hand in a single RGB image (demonstrated on a wide variety of scales and contexts): a *hand box*; its *side* (left/right); its *contact state* (none / self / other person / non-portable object / portable object); and, for the hand in contact, an *object box* around the object or person in contact. These enable crucial downstream problems like reconstruction and grasp analysis. For example, the detection of hand location and side enables the use of recent mesh reconstruction systems [21, 42].

This system, along with an existing mesh reconstruction method [21], yields a system that can detect hands, their contact state, their 3D reconstruction, and what object they are touching. We believe this output enables large-scale fine-grained learning about human-object interaction. As an example, we introduce a method to identify bad mesh reconstructions (necessary for learning *from* meshes) and provide a demonstration of learning on consumer videos.
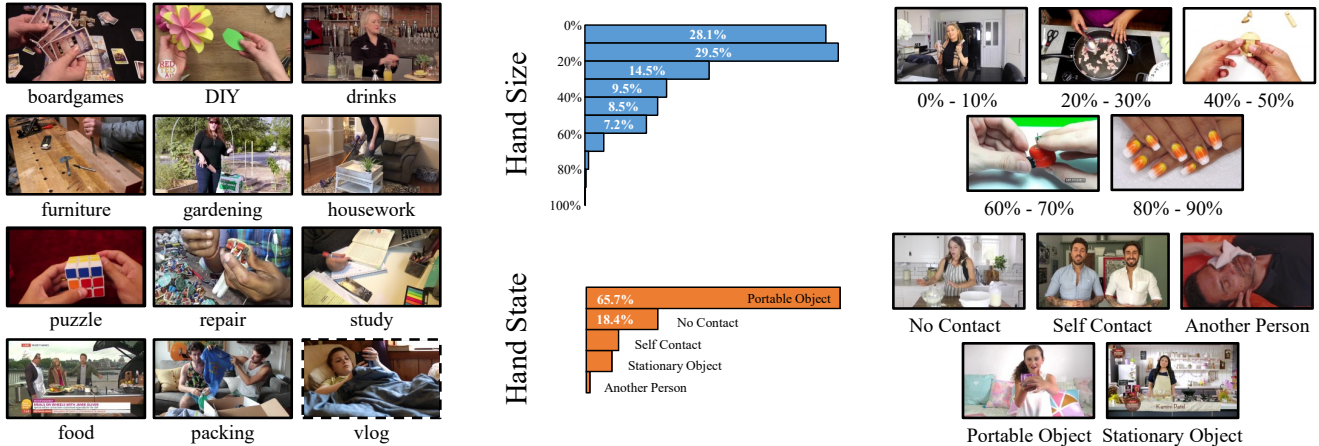
1

Figure 2. **Snapshots from our dataset. (Left)** Samples from the eleven genres that we use to collect our video dataset *100DOH* as well as a sample from the VLOG dataset that we additionally sample from to produce our 100K frame subset. **(Right)** Statistics about the hand frame dataset in terms of hand size (bounding box diagonal length over image diagonal length) and contact state, as well as illustrative samples of both. Our dataset depicts a wide variety of hands engaged in interaction in a variety of contexts and at a variety of scales.

This effort is backed by a new dataset, *100 Days of Hands*, introduced in Section 3, consisting of a large-scale (131 days+) video dataset of humans engaged in interaction that was gathered implicitly [14] (i.e., with a set of relatively generic genre tags as opposed to particular actions). These videos depict a wide range of activities, viewpoints, and settings. We use frames from this dataset and a similar dataset [14] to create a 100K image dataset which has the rich hand-state annotation to support our model.

We believe that this model and data (along with existing past work, especially in reconstruction) enable the field to collectively tackle new and important problems in human-object interaction in general consumer videos and demonstrate this concretely in Section 5. We show that: (1) Our new hand dataset yields high performance detectors (90% VOC [12] AP) that generalize well across datasets, occasionally outperforming training and testing on the same dataset; (2) Our new hand state model and dataset serves as an enabling technology that lets the community deploy exciting hand-mesh reconstruction systems like [21] on YouTube videos; (3) Our system may provide a stepping stone towards important tasks like grasp analysis by showing how to use it to build a proof-of-concept system that maps objects to 3D meshes of hands engaged in interaction.

2. Related Work

Our work focuses on identifying a rich state of hands in an ordinary RGB image with the goal of using it as a foundation for understanding human-object interaction in Internet videos. It therefore touches on many papers in the area of understanding human-object interaction.

Human-object interaction, and understanding the affordances (opportunities for interaction [16]) has been a long-term interest of computer vision. Recent work has largely taken the approach of recognizing verb-noun pairs [17, 9, 10, 19, 6]. In terms of technical approach, our method is most related to the approach of Gkioxari et al [17]. As an output, however, we propose an alternate representation based on physical contact and interaction.

In the process, we gather a video dataset of humans engaged in interaction that we annotate and learn from. Of the many works in video human-object interaction [11, 39, 33, 24, 40], ours is most clearly related to VLOG [14], which also gathers data of interaction, and AVA [18], which investigates atomic (i.e., base-level) actions. We build on the ideas and part of the data of VLOG, but expand it to a wider and more diverse dataset and far more thoroughly investigate and annotate contact. Like AVA, we also investigate a representation below activities, but is different and complementary (in contact with a box) compared to AVA's semantic actions (e.g., "write", "play instrument").

Hands have long been the subject of study in computer vision. In this area, our data and approach sit between image-based hand detection datasets like [1, 25, 5, 14, 28] and efforts at understanding contact with richer annotation but requiring more specialized devices or more constrained environments such as [7, 35, 30, 15, 36]. We expect that fully understanding hands may require a variety of approaches; our approach tries to strike a balance between potential for scalability and richness of annotation. We note that while there has been progress in full body pose estimation (e.g., [8]), our approach works even in highly truncated settings like Internet videos.

One particularly important line of work in understanding human hands is extracting the pose of hands from images. A full survey of this literature is beyond the scope of this paper and we refer the reader to [41, 34]. Most recently,

Table 1. Comparison of 100DOH with existing datasets for human-object interaction. While only a small fraction of it is labeled compared to more densely annotated datasets, the proposed large-scale video dataset is a rich source for unsupervised learning about hands.

Name	Length	Annotation	Source
100DOH	131D	100K frames per-hand state	YouTube
AVA	2D	3s-level atomic Actions	Movies
HowTo100M [24]	(5.6K)D	None / Captions	YouTube
Moments [26]	34D, 17H	Vid. Class	Misc
VLOG [14]	14D, 8H	Vid Class, Sparse Annots	YouTube
YouCook2 [39]	7D, 8H	Action Segments	YouTube
Charades [33]	3D, 8H	Action Segments	Home
EPIC-KITCH. [11]	2D, 7H	Actions Segments, Object BBs	Home

Table 2. Comparison of 100DOH with hand datasets. Our dataset is far larger and has a rich annotation of contact state with objects.

Name	# Im	# Hands	Side	Contact	Objects	Source
100DOH	100K	189.6K	✓	✓	✓	YouTube
VLOG [14]	5K	26.1K	X	Per-Image	X	YouTube
VIVA [1]	5.5K	13.2K	✓	X	X	Capture
Ego [5]	4.8K	15K	✓	X	X	Capture
VGG [25]	2.7K	4.2K	X	X	X	Flickr, TV
TV-Hand [28]	9.5K	8.6K	X	X	X	TV
COCO-Hand [28]	26.5K	45.7K	X	X	X	Flickr

this has taken the form of systems that can, given a cropped hand with known side, such as [21, 42], infer a mesh via a low-dimensional model like MANO [31]. Our approach provides the necessary input for this reconstruction and thus enables the large-scale deployment of these techniques to Internet videos (including a self-supervised learning-based system that can detect reconstruction failures).

One of the applications we demonstrate is mapping an image of an object not being interacted with to a 3D mesh of the hand. This work has tackled previously using RGBD sensors [20, 2] or thermal data [7]; our work is able to learn this simply by mining examples via the rich representation our approach can infer. Most existing work in this area that can learn from Internet videos [27, 13] focuses on interaction hotspots, while our work infers a mesh.

3. Dataset

We gathered a large and rich dataset of everyday interactions from YouTube that serves as a basis for our subsequent investigation. This dataset consists of two parts that play complementary roles: (i) a massive, unlabeled video dataset that is source for unsupervised learning; and (ii) a 100K frame subset that has been labeled.

We follow the principles outlined in [14], where we search *implicitly* for hands engaged in interactions rather than explicitly. We see a few advantages to frames from implicitly gathered video data: (a) still photos require an intentional decision to take and upload the photo, meaning that the transitional fossils of daily life (e.g., a half-ajar refrigerator with a hand resting on it) usually go undocumented, a form of selection bias [37]; (b) explicitly gathered data (e.g., searching "playing tuba" for Kinetics [23]) tends to capture unusual activities since these are easy to find.

3.1. Gathering a Large-Scale Video Dataset

Gathering implicitly consists of two rough stages: identifying an overcomplete set of candidate videos using generic queries and filtering out irrelevant videos.

Generating a set of query candidates: We began with a set of 11 categories: boardgames, DIY, making drinks, making food, furniture assembly, gardening, doing housework, packing, doing puzzles, repairing, and studying. We generated 13.2K queries using frequent words, Wordnet hyponyms and templated queries (e.g., "DIY cookies home 2014"), and searched YouTube. These queries yield ∼6.5M video responses (an estimated 86 years), which we must filter for containing hands interacting with objects.

Filtering: Manually screening such a large dataset is impractical and we therefore use a learned model based on video thumbnails. In particular, we use three learning targets: (a) what fraction of 100 evenly-spaced frames have high responses from a Faster-RCNN hand detector trained on [14]; and (b) what fraction of frames are judged as containing interaction by human workers; (c) whether the frames are cartoons.

These cannot be evaluated on the whole dataset, so we train models (see supplemental material) that map thumbnails to a prediction of each. This can be rapidly evaluated at scale on our full dataset's thumbnails and our final dataset is the intersection between the datasets that are likely to contain hands and depict interaction, with likely cartoons removed. These systems are not used in the future and all subsequent annotations are at a frame-level and are independent of video-level filtering mechanism.

3.2. Image Dataset

This yields a video dataset – 100 Days of Hands (100DOH) – of 27.3K videos across 11 categories with 131 days of footage of everyday interaction. We use this to build a new 100K frame-level dataset, that is primarily (∼ 85%) a subset of 100DOH and (∼ 15%) a 3× extended and relabled version of the hand dataset in [14]. We chose randomly among frames, filtering out (and retaining for later use) images containing no hands. We include VLOG because we tried building off of VLOG, but realized that we needed more diverse underlying data.

Annotation: For every hand in each image, we obtained the following annotations: **(a)** a bounding box around the hand; **(b)** side: left / right, which is crucial for mesh reconstruction; **(c)** the hand contact state ({no contact, self-
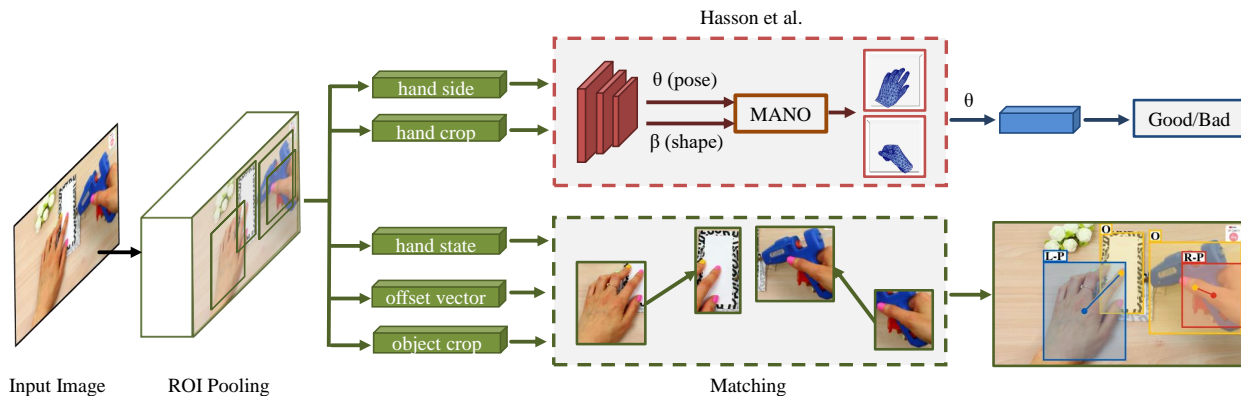
Figure 3. Our system can act as a foundation to understand interacting hands on the Internet. Our system takes a single RGB image and detects hands (irrespective of scale) and for every hand predicts: a box, side, contact state, and a box around the object it is touching. We can then (1) obtain a parse of hand state; and (2) use the hand box and side to feed a reconstruction system like [21]. To help make better use of Internet reconstructions, we introduce a self-supervised system that assesses mesh quality that we train on our data.

contact, other person contact, in contact with portable object}), which provides insights into what the person is doing; and **(d)** a bounding box around the object the person is contacting *irrespective of name*. The annotation of non-named bounding boxes is crucial: in-the-wild data is known to have a heavy tail, dooming categorization.

Annotation began by counting hands, then marking hand bounding boxes and sides simultaneously; in addition to standard qualification, consensus, and sentinel techniques, hand bounding boxes were annotated, then verified, then re-annotated if hands were missing. Catching these missing boxes is crucial since with data of this scale, hand detection performance can reach a mAP of 90%. Then hand state and object bounding box were annotated, again using qualifications, consensus, and sentinels. We only included images on which we could get conclusive judgments from workers. In total, in the 100K images, there are 189.6K hands annotated which are in contact with 110.1K objects.

Splits: We split by YouTube uploader id to make a 80/10/10% train/val/test split where each uploader appears in only one split. This split is backwards compatible with the VLOG hand data: no VLOG test appears in the trainset.

Comparison to existing data: We compare our dataset in raw statistics with other comparable datasets of videos in Table 1 and image-based datasets for studying hands in Table 2 (empirical cross-dataset evaluations demonstrating the utility of the dataset appear in Section 5). While unlabeled, our video dataset provides a valuable source of large-scale demonstrations of hands engaged in interaction. Our image dataset fills a gap of providing object contact and side information at vast scale. Additionally, as seen in the paper, 100DOH hands appear at a wide variety of scales.

4. Finding Hands & Objects in Interaction

Equipped with this data, we show how to build a system that can produce a fine-grained understanding of the scene. Our base system (Sec. 4.1) can predict, from a single image: (1) a box around any visible human hands in the scene as well as their *side* (left-vs-right) and *contact state* (none/self/person/portable/non-portable); (2) the box of an object the hand is in contact with; and (3) a *link* between each hand and an object it is in contact with. Our outputs can be directly plugged into machinery for hand reconstruction (Sec. 4.2) [21, 31], also enabling (4) a 3D mesh reconstruction; (5) and whether that reconstruction was likely correct. We believe this output can enable many exciting downstream applications, and we show a proof of concept of mapping objects to grasps (Sec. 4.3).

4.1. Hand and Object Detection

We build our system on top of a standard object detection system, Faster-RCNN [29] (FRCNN) by adding auxiliary predictions and losses per-bounding box. We deliberately chose FRCNN for its reputation as a standard foundation for detection tasks; we see additional improvements to the base network as orthogonal to our contributions. Specifically, we build on FRCNN trained to identify two objects – human hands and contacted objects. As in standard Faster-RCNN, the network predicts, for each anchor box, whether the anchor box is an object, what its category is, and bounding box regression adjustments to the anchor box; these remain unchanged. We predict a series of auxiliary outputs directly from the same ROI-pooled features as the standard classification outputs. We now report these outputs and the losses we use to train these additional layers:

We predict hand side $\mathbf{s} \in \mathbb{R}^2$ and contact state $\mathbf{c} \in \mathbb{R}^5$ via two additional fully connected layers. The outputs rep-
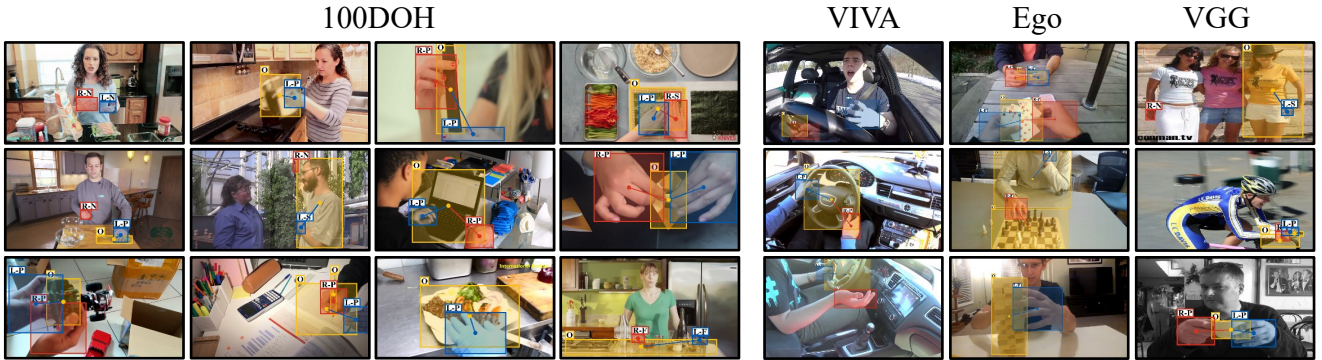
Figure 4. Selected results from our full hand state detection system. Here we show our results on 100DOH as well as generalizing (untrained) to VIVA [1], EgoHands [5], and VGG [25]. Our system is able to reliably extract hands at a variety of scales, poses, and contexts as well as identify contact state and which object is in contact.

resenting left-vs-right and {none / self / other / portable / non-portable}. Both are trained by minimizing standard cross-entropy losses L_{side} and L_{state}.

To link up boxes between hands and objects, we predict an association from a hand to an object, similar to Gkioxari et al. [17], by predicting an *offset vector*, factored into a unit vector $\mathbf{v} \in \mathbb{R}^2$ plus a magnitude $m \in \mathbb{R}$ by two fully connected layers. Given the ground-truth vector between the center of the bounding box of a hand to the center of the bounding box of the object the hand is contacting, we write it as a unit vector $\mathbf{v}' \in \mathbb{R}^2$ and magnitude $m \in \mathbb{R}$. We minimize the distance between the two vectors $L_{\text{ori}}(\mathbf{v}, \mathbf{v}') = ||\mathbf{v} - \mathbf{v}'||_2^2$ as well as the squared difference between the magnitudes $L_{\text{mag}}(m, m') = (m - m')^2$. Formulating the relationship as predicting an object per-hand allows multiple hands to contact the same object; while it does preclude a hand contacting multiple objects, we find this is rarer and leave it to future work.

We obtain a final discrete parse in terms of a set of hands in contact/correspondence with a set of objects through a greedy optimization on network output. Given a new image, we infer all the hand and object boxes, as well as their side and contact scores and association vector. We convert these soft predictions into a discrete prediction by suppressing unlikely hand/object detections and then associating each confident hand with the object whose center closest matches the hand's bounding box center plus its offset vector.

Training details. The standard FRCNN losses are minimized as usual; we minimize L_{side}, L_{state} over anchor boxes corresponding to ground-truth hands and L_{ori} and L_{mag} over anchor boxes corresponding to ground-truth hands in contact. We scale the loss terms to handle wide variance in the loss scale but otherwise did not tune loss scales (details in supplemental material). We use a ResNet-101 [22] backbone, initialized with Imagenet [32] and train it for 8 epochs with a learning rate of 10^{-3} with batch size of 1.

4.2. Applications to Reconstruction

Our system, out-of-the-box, directly enables the large-scale automatic deployment of techniques for mapping hands to 3D meshes which supplement our outputs. As a concrete demonstration, we build off of the technique of [21] that maps images to the MANO [31] low-dimensional parameterization of hands via a Resnet-18 [22]; this parameterization comprises $[\boldsymbol{\theta}, \boldsymbol{\beta}]$ representing hand pose and shape, which can be converted to a 3D mesh via the differentiable MANO model. Our system provides the necessary inputs (locations *plus side*); building a more complex system that integrates with the detection system is an interesting future direction and technically feasible but beyond the scope of a single paper.

While this enables many interesting downstream tasks, these tasks would be harmed by incorrect reconstructions and so we present a simple technique for recognizing these failures. We use the ideas of checking a network's equivariance as a signal for confidence from [4]. Specifically, given an image, we reconstruct the hand from six rotated copies of the image, reproject joints, rotate them and computed the the mean L2 distance of corresponding joints. We generate these for 3 frames per training video (70.9K images), sort by consistency and set the examples in the top 30% as positives and the bottom 30% as negatives. We train a two layer multilayer perceptron (hidden layer sizes 100, 50) on $[\boldsymbol{\theta}]$, minimizing the binary cross-entropy; this classifier can be run at inference time to quickly identify poorly estimated frames. We quantify its effectiveness in Section 5.3.

4.3. Proof of Concept: From Object to Grasp

Once we can identify hands in contact in videos and reconstruct them, we can generate training data for identifying *how* hands might contact an object. After associating hands to tracks, we search our training set for moments in time where a hand makes contact with an object. On either

Table 3. Cross-dataset performance: training Faster-RCNN [29] on our dataset ensures near-equal performance to training and testing on other datasets; other datasets usually generalize poorly. For each train set, we report the minimum ratio (across test sets) between its performance and the best-performer.

Test → Train ↓	100 DOH	VLOG [14]	VIVA [1]	Ego [5]	VGG [25]	TV+Co [28]	min ratio
100DOH	90.1	86.4	86.5	90.8	73.9	65.4	92.9
VLOG	78.6	77.5	76.6	83.2	64.6	59.2	81.1
VIVA	23.6	27.7	90.8	56.8	21.5	10.1	14.5
Ego	40.7	32.6	44.9	90.7	17.4	8.0	11.5
VGG	61.4	61.7	78.8	63.0	79.6	66.6	68.1
TV+Co	56.2	61.5	74.9	62.4	77.4	69.6	62.4

Table 4. Average Precision when we vary definitions of correct detection. Using 15K samples dramatically degrades performance on any category, and 45K samples produces a large drop on getting all outputs correct.

	Hand	Obj	H+Side	H+State	H+O	All
Full	89.6	63.9	78.9	64.0	46.9	38.5
45K	88.4	61.0	77.3	62.7	39.3	31.3
15K	80.9	54.2	66.8	53.3	27.3	19.7

side are a timestamp t_{before} where a hand is not in contact and a timestamp t_{after} where the hand is in contact. At t_{after}, our system provides side, bounding box for both hand and object, a mesh (via [21]), and our self-supervised mesh assessment score. We can use the object box at t_{after} to crop the image pre-contact at time t_{before}. We apply a number of filters, including removing examples with overlapping hands and scenes where the object appeared to move (detected by change in appearance).

We can then learn a mapping from an image of an uncontacted object to a hand-in-contact. We use 203.4K training samples to build a system. We fine-tuned an Imagenet-pretrained [32] Resnet-18 [22] capped by a MLP to predict, for each mesh, hand pose and side, supervised by standard L2 losses along with supervision from hand vertices similar to [21]. We found that this, like many regression formulations (e.g., see [38]), averaged out between the multiple modes. To prevent averaged hands, we generated a 10-hand codebook from training samples, represented each hand with the nearest of 10 classes and predicted these. For simplicity, trained another Resnet-18 to predict these classes, minimizing a cross-entropy loss.

5. Experiments

We conduct a series of experiments that aim to quantify: (a) how well the dataset allows the hand detection relative to other hand detection datasets? (b) how well our full model for hand-state works? (c) how much our full hand model assists reconstruction? (d) how well our model infer interacting hands for isolated objects? We conduct our experiments on both our newly introduced datasets, as well as on other hand detection datasets [14, 1, 25, 5].

5.1. Hand Bounding-Box

We evaluate the merit of our new dataset of hands by evaluating cross-dataset performance for hand detection with a standard fixed detector, and a comparison with full-body pose estimation.

Cross-dataset hand detection analysis: We begin by training the same base model – a standard F-RCNN [29] with a Resnet-101 [22] backbone – on a number of datasets and evaluating same and cross-dataset performance. We use F-RCNN for simplicity and due to its widespread use as a commodity detection system. We only evaluate on datasets where *all* hands in a frame are annotated with boxes [1, 5, 25, 14]. We do not compare with [15] since it only annotates *one* hand (and note [11] has no hand boxes). We evaluate hand detection results using an IoU of 0.5 (i.e., PASCAL [12]) since unlike objects with clear boundaries e.g., fire hydrants, cars, precise boundary between hand and wrist is unclear in the wild.

Table 3 shows that a model trained on 100DOH generalizes well across datasets and nearly matches performance obtained training and testing on every other hand datasets: at worst, it obtains 92.9% of the mAP of training and testing on the same data (on VGG). VLOG and VGGHands (both gathered from large-scale diverse data) generalize reasonably well, but far worse than 100DOH with minimum relative mAP of 81.1% and 68.1% respectively; models trained on VIVA and EgoHands perform well on egocentric datasets but generalize poorly to non-egocentric views (which is unsurprising but worth quantifying). The annotation format of VGG and TV is a quadrilateral rather than an axis-aligned rectangle, and we preprocess the labels to be axis-aligned to make all annotations consistent; this diminishes results slightly. We show some sample cross-dataset results in Figure 4.

Comparison to full-body pose estimation: One common question (also asked by [28]) is whether we need specialized hand-detection given the success of full-body pose estimation systems and datasets [8]. We evaluate this by computing precision/recall for hand detection using OpenPose [8]. We convert body joint configuration and hand detection to a common evaluation scheme by defining true positives as: ([8]) when a pose detector places the hand (estimated as $w + 0.2(w − e)$ where w and e are wrist and elbow locations) within a ground truth box with the same center but twice the width and height; (*ours*) when the bounding box detector puts its center point inside the ground-truth box (a higher accuracy standard). Due to truncated people, [8] achieves a low average precision of around 43.0% and ef-

fectively maxes out recall at 49.5%. At this recall, a Faster RCNN still has a precision of 99.7%. While current pose estimators trained on current datasets are effective when the body is mainly visible, dedicated hand detectors appear to still be necessary.

Statistical Baselines: We additionally test whether the dataset can be solved with simple statistical baselines. We computed the median box for all hands as well as a median box for left and right hands. These get an AP of 0.08% and 0.11%, which shows that the hands are widely distributed across the image and vary in size.

5.2. Full Hand State

We next evaluate our full hand state detection system in isolation. Here, we show that the scale of our datasets is beneficial by comparing with results obtained by training on smaller subsets of the data.

Qualitative Results: We show some results of the full system in Figures 4 and 5. In general, our system does a good job at recognizing hands and sides despite a wide variety of scales and contexts present in the data. While it often gets the full state right, this is a clearly harder task with lots of room for improvement.

Failure modes: Common failure modes include: getting the precise contact state right, especially when the hand is near an object, which video may improve; associating the correct object with the hand, especially with multiple people interacting with multiple objects, which a more complex inference technique might improve; and getting the correct hand side for egocentric hands (e.g., on [11]), which more egocentric training data may improve.

Metrics: We evaluate the full prediction using mAP by modifying the criterion for a true positive. We begin by evaluating hand (*Hand*) and interacted object (*Obj*) individually. We then count a hand as a true positive only if it also has the correct hand side (*H+Side*), state (*H+State*), and has the correct object associated with it (*H+O*). Finally, we count a hand as a true positive only if it has all correct (*All*).

Quantitative Results: We are the first to tackle this problem, and therefore there are no methods to compare with. We therefore test to what extent our large-scale data is important and compare to the same method trained on a subsets of data: *Full* (90K trainval), *45K* (50% of trainval); and *15K* (17% of trainval). We report results in Table 4. In any category, tripling from 15K to 45K produces large gains, while further doubling to 90K produces more incremental gains. However, when looking to correctly identify all hand state, these mistakes combine to yield a steep performance hit (7% AP) compared to using the full data. Together, these underscore the need for large-scale data, especially as one looks to tackle tasks requiring correct estimation of *multiple*
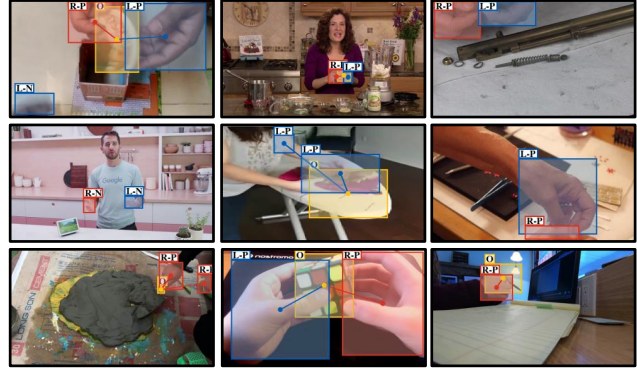

Figure 5. **Random** results on video frames from our 100 DOH dataset. Our approach detects hands reliably in a variety of scales, configurations, and contexts.

aspects of hands.

Analysis as a Function of Scale: We evaluate the performance on different hand scales. We separate images into different bins according to the average hand size (measured as the square-root of the percent of pixels) and evaluate each bin. Tiny hands are naturally harder to find and hand AP rapidly goes up from 78.2% to 90.3% as scale goes from 10% to 20%; performance however remains stable until 70%, where it slightly drops. Additional results appear in the supplemental.

5.3. Hand State for Reconstruction

One of the exciting outcomes of having a system that can reliably identify hand state is that it directly enables automatically applying mesh reconstruction techniques to consumer videos. We present two experiments that assess our method's contributions to this future – identifying side and our self-supervised mesh assessment technique. We use human judgments to evaluate success.

Data: We use 2K images from the test set. Our system produced 3,861 detections, which we reconstructed using [21] for both the correct hand side and the incorrect hand side resulting in 7,722 meshes. Crowdsourced workers re-annotated the detected hand to preclude mistakes and then assessed each mesh five times as correct/incorrect (definitions in the supplemental). Workers were deliberately **not** told to inspect sides of hands. Workers passed a qualification test; we used sentinels to monitor performance; and results from all reconstructions were annotated simultaneously and in randomized order.

Quantitative Results (Side): We first tested whether having the side was important – an alternate hypothesis is that MANO might repurpose thumbs for pinkies, for instance. We show a few select qualitative examples in Figure 6. Unsurprisingly, despite not being told to examine side, workers
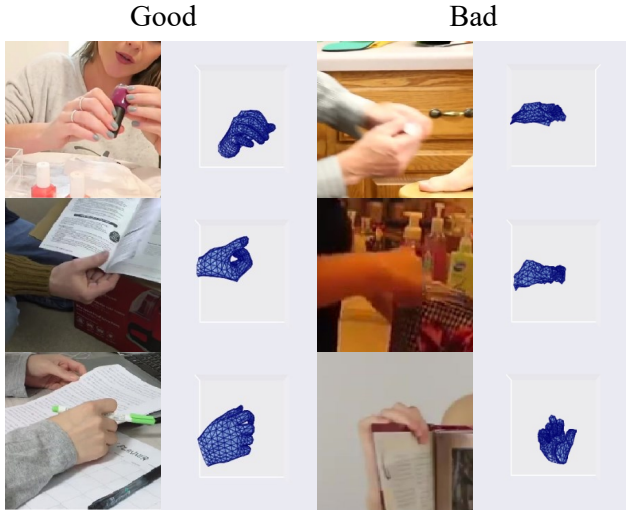
Figure 6. Mesh reconstruction and assessment. Our system detects the location and side of hands in Internet videos, that [21] uses to predict a mesh. Many predictions do not succeed. Our self-supervised system for mesh assessment identifies plausible (useful for downstream tasks) and implausible hands (worth discarding).
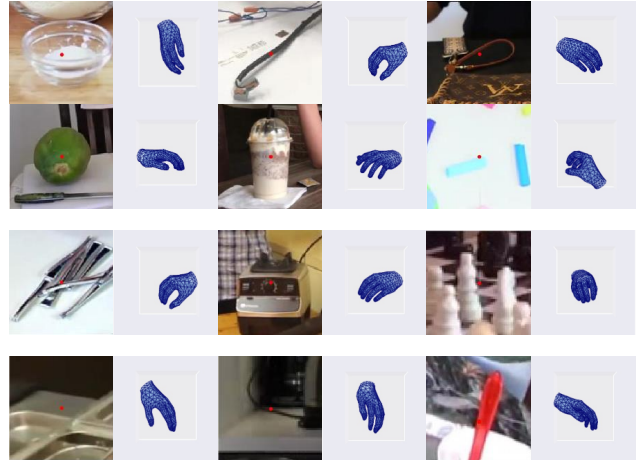

Figure 7. Hand prediction results. Our system enables extracting pairs of images of uncontacted objects and *good* reconstructed meshes. We show results of a system trained to infer meshes from images. (Rows 1,2): Selected Results. (Row 3,4): Random results that crowd workers liked/preferred over a random grasp (Row 3) or did not like (preferring random grasp over it).

were far more likely to think hands reconstructed with the *detected* side were correct (57.8%) compared to the opposite side (29.1%).

Quantitative Results (Quality): We then used this data to evaluate whether we can successfully identify correct reconstructions. We binarized worker judgments by majority vote and computed AUROC. The proposed method (a MLP trained on positives/negatives identified by self-consistency) obtains an AUROC of 90% on this data. We compared with a two baselines to put this result in context. Gaussian Näive Bayes on the same training data does similarly (89%), showing that the positive/negative labels are important, not the learning algorithm. Simply fitting a multivariate Gaussian on all generated hands and using the log-likelihood does far worse (60%), which underscores the importance of the labels.

5.4. Future prediction

We took our networks trained on the training set of 100DOH and tested them on videos from the test set, finding points at which contact changes. We then reconstruct 3K examples, showing both qualitative results and computing quantitative results via human judgment.

Qualitative Results: We show a few select qualitative examples in Figure 7. Overall we observe that our method often does a good job of identifying the angle from which the hand should grasp the object. While our approach often finds plausible grasps, the myriad of ways a human *can* grasp an object and difficulty of predicting a full mesh

makes this a challenging task.

Human Judgment: We showed 3K results to crowd-workers in a two-choice test, comparing the result to a random hand from the training set to examine if our system extracts the signal. Note that random is very frequently correct by chance since usually very many grasps suffice (consider how many ways your hand can touch a soda can). Workers selected which they thought was more plausible given the image. Presentation order was randomized, and we employed qualifications and sentinels; examples where workers could not come to an agreement were considered ties. Of the 60% with a conclusive result (some inconclusive results are due to the input not depicting a clear object), our system was preferred 72% of the time.

6. Conclusion

We have presented a method for obtaining information about hand contact state in the scene, a large-scale dataset for training this method, and demonstrated applications of our technique. We are barely scratching the surface in terms of what can be learned in the world of large-scale Internet video and we hope that our rich representation can help the field collectively explore this area.

References

[1] The vision for intelligent vehicles and applications (VIVA) challenge, laboratory for intelligent and safe automobiles, UCSD. http://cvrr.ucsd.edu/vivachallenge/. 2, 3, 5, 6

[2] Shuichi Akizuki and Yoshimitsu Aoki. Tactile logging for understanding plausible tool use based on human demonstration. In *BMVC*, 2018. 3

[3] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 1

[4] Yuval Bahat and Gregory Shakhnarovich. Confidence from invariance to image transformations. *CoRR*, abs/1804.00657, 2018. 5

[5] Sven Bambach, Stefan Lee, David Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015. 2, 3, 5, 6

[6] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *ECCV*, 2018. 2

[7] Samarth Brahmbhatt, Cusuh Ham, Charlie Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 1, 2, 3

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 6

[9] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018. 2

[10] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 2

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2, 3, 6, 7

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 2, 6

[13] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018. 3

[14] David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. From lifestyle VLOGs to everyday interactions. In *CVPR*, 2018. 1, 2, 3, 6

[15] 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 *CVPR*, 2018. 2, 6

[16] J. Gibson. *The ecological approach to visual perception*. Boston: Houghton Mifflin, 1979. 2

[17] Georgia Gkioxari, Ross Girshick, Piotr Dollar, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 2, 5

[18] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2

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

[20] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010. 3

[21] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 3

[24] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *arXiv preprint arXiv:1906.03327*, 2019. 2, 3

[25] A. Mittal, A. Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. In *BMVC*, 2011. 2, 3, 5, 6

[26] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. 3

[27] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019. 3

[28] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *ICCV*, 2019. 2, 3, 6

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4, 6

[30] G. Rogez, J. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 1, 2

[31] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3, 4, 5

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 1–42, April 2015. 5, 6

[33] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2, 3

[34] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 1, 2

[35] 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 *ECCV*, 2016. 2

[36] Bugra Tekin, Federica Bobo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 2

[37] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 3

[38] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 6

[39] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1, 2, 3

[40] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 2

[41] C. Zimmerman and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 1, 2

[42] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 3