Fine-Grained Egocentric Hand-Object Segmentation: Dataset, Model, and Applications

Lingzhi Zhang*[1], Shenghao Zhou*[1], Simon Stent[2], and Jianbo Shi[1]

[1] University of Pennsylvania
[2] Toyota Research Institute

Abstract. Egocentric videos offer fine-grained information for high-fidelity modeling of human behaviors. Hands and interacting objects are one crucial aspect of understanding a viewer's behaviors and intentions. We provide a labeled dataset consisting of 11,243 egocentric images with per-pixel segmentation labels of hands and objects being interacted with during a diverse array of daily activities. Our dataset is the first to label detailed hand-object contact boundaries. We introduce a context-aware compositional data augmentation technique to adapt to out-of-distribution YouTube egocentric video. We show that our robust hand-object segmentation model and dataset can serve as a foundational tool to boost or enable several downstream vision applications, including hand state classification, video activity recognition, 3D mesh reconstruction of hand-object interactions, and video inpainting of hand-object foregrounds in egocentric videos. Dataset and code are available at: https://github.com/owenzlz/EgoHOS

Keywords: Datasets, Egocentric Hand-Object Segmentation, Egocentric Activity Recognition, Hand-object Mesh Reconstruction

1 Introduction

Watching someone cooking from a third-person view, we can answer questions such as "what food is the person making?", or "what cooking technique is the person using?" First-person egocentric video, on the other hand, can often show much more detailed information of human behaviors, such as "what finger poses are needed to cut a steak into slices?", "what are the procedures to construct a IKEA table with all the pieces and screws?" Thus, egocentric videos are an essential source of information to study and understand how humans interact with the world at a fine level. In these videos, egocentric viewer's hands and interacting objects are incredibly informative visual cues to understand human behaviors. However, existing tools for extracting these cues are limited, due to lack of robustness in the wild or coarse hand-object representation. Our goal is to create data labels and data argumentation tools for a robust fine-grained egocentric hand-object segmentation system that can generalize in the wild. Utilizing the fine-level interaction segmentation, we show how to construct a high-fidelity model that can serve as a foundation for understanding and modeling human hand-object behaviors.

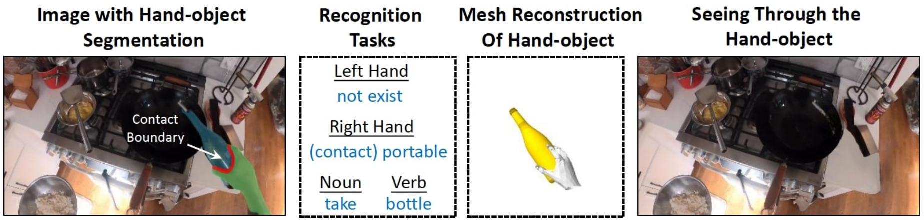
Image with Hand-object Segmentation
Contact Boundary
Recognition Tasks
Left Hand
not exist
Right Hand
(contact) portable
Noun
Verb
take
bottle
Mesh Reconstruction Of Hand-object
Seeing Through the Hand-object

Fig. 1. Leftmost image: our proposed dataset enables us to train a robust hand-object segmentation model. We introduce contact boundaries to model the hand-object interaction explicitly. Right: our hand-object segmentation model is helpful for many vision tasks, including recognizing hand state, activities, mesh reconstruction, and seeing-through the hand-object.

The first and foremost factor in building a robust egocentric hand-object segmentation model is a good-quality labeled dataset. Previous works [1,33,73] have constructed hand segmentation datasets for egocentric videos. However, the collected data are mostly restricted to in-lab settings or to limited scenes, and lack labels for interacting objects. More recently, 100-DOH [58] made a great effort to label large-scale hand and object interactions in the wild, but the labels for hands and objects are at the bounding box level. To bridge the gap and further advance fine-level understanding of hand-object interactions, we propose a new dataset of 11,243 images with per-pixel segmentation labels. A major characteristic is that our dataset contains very diverse hand-object interaction activities and scenarios, where the frames are sparsely sampled from nearly 1,000 videos in Ego4D [16], EPIC-KITCHEN [8], THU-READ [68], and from our own collected GoPro videos. In addition, we also provide fine-grained labels of whether an object is interacted with by the left hand, right hand, or both hands and whether it is being interacted with directly (in touch) or indirectly.

To serve as an out-of-distribution test set for evaluating in-the-wild performance, we sparsely sampled and labeled 500 additional frames from 30 egocentric videos from YouTube. With our new segmentation dataset, we boost the hand segmentation performance significantly compared to the previous datasets [1,33,73]. Our dataset is the first to label interacting hand-object contact boundaries in egocentric videos. We show this label can improve the detection and segmentation of interaction objects. No matter how diverse our dataset is, we will inevitably encounter new domains with very different illumination, objects, and background clutter. We propose a context-aware data augmentation technique that adaptively composites hand-object pairs into diverse but plausible backgrounds. Our experiments show that our method is effective for out-of-domain adaptation.

We view our hand-object segmentation model as a foundation for boosting or enabling many vision applications, of which we demonstrate three, as shown in Fig. 1. First, we show that recognition tasks can get consistent performance improvement by simply adding reliably segmented hand or object masks as inputs.

We experiment with a low-level recognition task to classify the left/right-hand state and a high-level recognition task to understand egocentric video activities by predicting verbs and nouns. Another useful but challenging application is reconstructing hand-object interaction in 3D mesh, which relies on the 2D hand-object masks during optimization. In this application, we integrate our hand-object segmentation model into the mesh reconstruction pipeline [18], and show improvements and generalization for mesh reconstruction of hand-object, compared to its original hand-object segmentation pipeline pretrained on COCO [39]. Finally, we show an interesting application by combining our accurate per-frame hand segmentation and video inpainting [12] to see through hands in egocentric videos, which could help scene understanding models that have not been trained with hands in the foreground. More details of each of these applications are discussed in Section 7.

We summarize the contributions of this work as follows: 1) We propose a dataset of 11,243 images with fine-grained per-pixel labels of hand and interacting objects, including interacting object masks, enabling hand segmentation models to generalize much better than previous datasets. 2) We introduce the notion of a dense contact boundary to explicitly model the relationship between hands and interacting objects, which we show helps to improve segmentation performance. 3) We propose a context-aware compositional data augmentation technique, which effectively boosts object segmentation. 4) We demonstrate that our system can serve as a reliable foundational tool to boost or enable many vision applications, such as hand state classification, video activity recognition, 3D reconstruction of hand-object interaction, and seeing through hands in egocentric videos. We will release our dataset, code, and checkpoints to the public for future research.

2 Related Work

2.1 Hand Segmentation

Prior to deep learning, several works have attempted to solve the hand segmentation task. Jedynak et al. [22] used a color statistics–based approach to separate the skin region and the background. Ren and Gu [52] proposed a bottom-up motion-based approach to segment hand and object using the different motion patterns between hands and background in egocentric videos. Following up with [52], Fathi et al. [10] further separates hand and interacting object from the whole foreground by assuming a color histogram prior over hand super-pixels, and uses graph-cut to segment hands and objects. Li and Kitani [31,32] first addressed the hand segmentation problem under various illuminations, and proposed to adaptively select a model that works the best under different illumination scenarios during inference time. Zhu et al. [84] proposed a novel approach by estimating a probability shape mask for a pixel using shape-aware structured forests. Beyond the egocentric viewer, Lee et al. [30] studied the problem of hand disambiguation of multiple people in egocentric videos by modeling the spatial, temporal, and appearance coherency constraints of moving hands.

More recently, many works [1,73,4,34,36,37,58,59,61,24,50] have applied deep networks for hand or object segmentation. Bambach et al. [1] introduced a dataset that contains 48 egocentric video clips for people interacting with others in real environments with over 15,000 labeled hand instances. The authors also proposed CNNs to first detect hand bounding boxes and then use GrabCut [56] to segment the hands. Following up on the same dataset, Urooj and Borji [73] used the RefineNet-ResNet101 [38] to achieve the state-of-the-art hand segmentation performance at the time. To alleviate the generalization issue, Cai et al. [4] proposed the use of a Bayesian CNN to predict the model uncertainty and leveraged the common information of hand shapes to better adapt to an unseen domain. There are also some other dataset efforts regarding hand segmentation. Li et al. [34] proposed the Georgia Tech Egocentric Activity Datasets (GTEA), which includes 625 frames with two-hand labeling and 38 frames with binary labeling. Later, Li et al. [33] extended the dataset (EGTEA) with 1,046 frames with two-hand labels and 12,799 frames with binary masks. Lin et al. [36,37] also explored artificially composited hands with various backgrounds to scale up a large-scale synthetic dataset. Urooj et al. [73] recognized the constrained environment as one big limitation of existing datasets and collected an in-the-wild dataset by sampling frames from YouTube videos (EYTH). Though it is more diverse, it is relatively small with around 2,000 frames sampled from only 3 videos. Since frames are selected by simply sampling at a fixed rate, many frames are similar to each other in appearance. In addition to datasets with per-pixel labels, Shan et al. [58] labeled 100K video frames with bounding box labels for hands and interacting objects. More recently, Shan et al. [59] also proposed to learn hand and hand-held objects segmentation from motion using image and hand location as inputs.

Our work differs from previous works in two main aspects. While previous work mainly focus on egocentric hand segmentation, we take a step further to study not only hand segmentation but also interacting object segmentation. In addition, previous datasets were mainly focused on certain constrained scenes and limited activities. Our proposed dataset includes diverse daily activities with hundreds of human subjects. More detailed comparisons are shown in Section 3.

2.2 Hand-Object Interaction

Many works have studied hand-object interaction from different angles other than segmentation. One highly related direction is to model and estimate 3D hand joints [5,44,45,53,60,63,66,72,78,79,80,85,3] and mesh [26,27,51,55,75,83,46], object pose [14,28,29,35,43,57,64,71,76,81], or both [6,17,18,19,69]. A line of works [2,7,21,23,67] have also attempted to generate hand pose conditioned on objects. Mostly related to our work, Hasson et al. [18] and Cao et al. [6] used segmentation masks of hands and interacting objects to compute 2D projection loss in order to optimize the 3D mesh reconstructions of hand-object pairs. However, the instance segmentation model [25] they used to pre-compute hand and object masks are pretrained on COCO [39], which is not tailored to egocentric hand-object segmentation, and thus heavy human intervention is often needed to fix or filter out wrongly predicted masks. Other directions of hand-object interaction involve using hands as probes for object understanding [15], affordance hotspot reasoning [47,49], or even leveraging visual hand-object understanding for robotic learning [41,42,48]. Overall, we view our work as an orthogonal foundational tool for many of these vision tasks.

3 Dataset

Gathering Data from Multiple Sources. A big motivation of this work is that the existing datasets do not support researchers to train a model that generalizes well in the wild. Therefore, we collect data from multiple sources, including 7,458 frames from Ego4d [16], 2,121 frames from EPIC-KITCHENS [8], 806 frames from THU-READ [68], as well as 350 frames of our own collected indoor egocentric videos. This results in a total of 11,243 frames sparsely sampled from nearly 1,000 videos covering a wide range of daily activities in diverse scenarios. We manually select diverse and non-repetitive video frames from the sampled set that contain interesting hand-object interactions to label with per-pixel segments, as shown in Fig. 2. More details on video frame sampling are included in the supplementary materials.

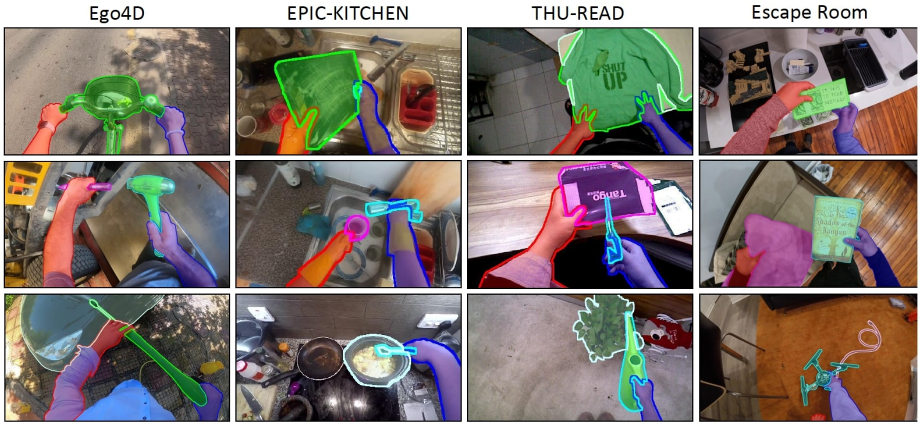

Fig. 2. A selection of images from multiple sources which we label with per-pixel hand and object segments. Color mapping: red → left hand, blue → right hand, green → object interacted by both hands, pink → object interacted by left hand, cyan → object interacted by right hand.

Annotations. For every image in the dataset, we obtained the following per-pixel mask annotations if applicable: (a) left-hand; (b) right-hand; (c) left-hand object; (d) right-hand object; (e) two-hand object. For each type of interacting object, we also provide two levels of interaction: direct and indirect interaction. We define direct interaction between hand and object if the hand touches the objects, such as the blue, cyan, or pink masks in Fig.2. Otherwise, we label the object as indirectly interacted with by the hand if the object is being indirectly interacted with, without touching, such as the light cyan masks in the third row of Fig.2. In this work, we only study directly interacting objects, but we will release the data to support future research into indirect interacting object segmentation. Note that previous works define hand masks in two types: hand only [1,73] and hand with arm [33]. We think both types of labels are useful depending on the application, so we provide both types of hand mask labels for all images in our dataset, where one for hands and another one for the rest of the arms.

Datasets	Label	#Frames	#Hands	#Objects	Objects	Interaction	L/R Hand	#Subjects	#Activities
100-DOH [33]	box.	100K	189.6K	110.1K	Yes	Yes	Yes	-	-
EGTEA [33]	seg.	13,847	-	-	Yes	No	No	32	1
EgoHand[1]	seg.	4,800	15,053	-	Yes	-	Yes	4	4
EYTH [73]	seg.	1,290	2,600	-	No	No	No	-	-
Ours	seg.	11,243	20,701	17,568	Yes	Yes	Yes	100+	300+

Table 1. Egocentric Hand-Object Segmentation Datasets Comparison. Unknown information is denoted with a dash "-". Compared to previous datasets, our proposed datasets cover relatively diverse scenes and activities with fine-grained segmentation labels of both hands and interating objects.

Comparison with Existing Datasets. In Table 1, we compare our proposed dataset with existing labeled datasets. 100-DOH [58] also provide a large volume of labelled images and objects, but its labels are at the bounding box level and not tailored towards egocentric images only. Although 100-DOH [58] has made a great effort to improve the generalization of hand-object bounding box detection, we think that having the segmentation prediction is particularly useful or necessary for many downstream vision applications, such as mesh reconstruction of hand-object interaction and seeing through the hands, as shown in Section 7. Compared to other segmentation datasets, one important characteristic of our dataset is that our images cover diverse activities and many human subjects. Since we do not have frame-level semantic labels, our conservative estimation of the number of human subjects and activity types are 300+ and 100+ respectively, according to the video IDs/names in the datasets [8,16,68]. Both the number of subjects and activities are orders of magnitude larger than previous segmentation datasets. In addition, unlike previous segmentation datasets, we are also the first to provide per-pixel mask labels for the interacting objects.

4 Hand-Object Contact Boundary

A key challenge of hand-object segmentation is the explicit understanding and modeling of the relationship between the hand and the interacting object. Segmenting the object purely based on appearance, as in traditional segmentation tasks, would not properly solve our problem. The reason is that the same object requires segmentation in certain frames but not in the others, depending on whether the hand is in contact with the object. To this end, we propose to explicitly model the interaction relationship between hand and object by introducing the notion of a dense contact boundary.

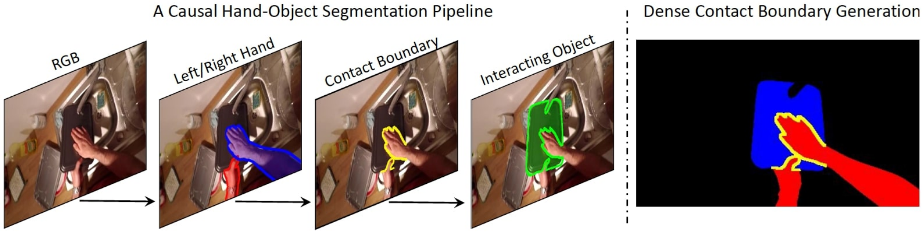
A Causal Hand-Object Segmentation Pipeline
RGB
Left/Right Hand
Contact Boundary
Interacting Object
Dense Contact Boundary Generation

Fig. 3. Left: an overview of our causal hand-object segmentation pipeline. **Right**: a demo to show how dense contact boundary is defined.

Conceptually, the dense contact boundary is defined to be the contact region between the hand and the interacting object. In our implementation, we first dilate both the labeled hand and the object masks in an image, then find the overlapped region between the two dilated masks, and finally binarize the overlapped region as our pseudo-ground truth for contact boundary, as shown in the yellow region in Fig. 3. With such a pipeline, we automatically generate supervision on the contact boundary for all images, where we could train the network to make prediction for it with standard binary cross entropy loss.

The advantages of explicitly predicting a dense contact boundary for interacting object segmentation are: 1) the contact boundary could provide a cue as to whether there is an interacting object for a given hand mask; 2) it also provides a clearer hand-object separation cue to improve segmentation accuracy. Our experiments show that the contact boundary helps the segmentation model to achieve a higher averaged object mask mIoU, and more ablation studies are shown in Section 6.2. Other advantages of the contact boundary besides boosting segmentation performance include: 1) the contact boundary segmentation contains crucial information for many downstream tasks, such as activity recognition and 3D mesh modeling of hand and object; 2) it could also provide potential metrics for evaluating segmentation, specifically for object-hand segmentation during an interaction.

We experiment with one hand-object segmentation pipeline that uses dense interaction boundary as an intermediate stage output. We sequentially predict first the left/right hand, then the contact boundary, and finally the interacting object in three stages, as shown in the left of Fig. 3. In each stage, we concatenate the outputs from previous stages as additional inputs. For example, the left/right hand masks are concatenated with the RGB image as inputs to predict the contact boundary; and in the last stage, the RGB image, hand masks, and contact boundary masks are concatenated as inputs to predict the interacting object masks. Our model is built by sequentially stacking networks, which we tried both a convolutional architecture (ResNet-18 backbone [20] and HRNet head [74]) and a transformer architecture (Swin-L backbone [40] and UperNet head [77]). Note that we do not focus on the architecture and loss design in this work, and more training details are described in the supplementary materials.

5 Context-Aware Compositional Data Augmentation

Copying-and-pasting foreground instances at different locations into different background scenes has shown to be a simple and effective data augmentation technique for object detection and instance segmentation, as shown in [9,13,82]. In order to further expand the dataset and improve our model performance, we build a context-aware compositional data augmentation pipeline such that the new composite image has semantically consistent foreground (hand-object) and background context.

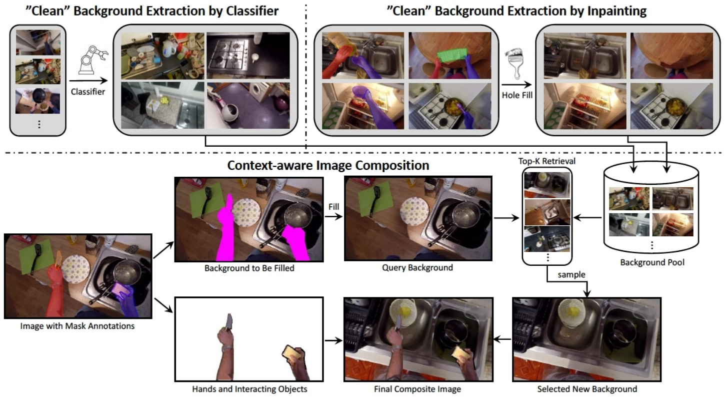
"Clean" Background Extraction by Classifier
Classifier
"Clean" Background Extraction by Inpainting
Hole Fill
Context-aware Image Composition
Top-K Retrieval
Fill
Background to Be Filled
Query Background
Background Pool
sample
Image with Mask Annotations
Hands and Interacting Objects
Final Composite Image
Selected New Background

Fig. 4. An overview of our context-aware compositional data augmentation pipeline.

Our overall pipeline design is shown in Fig. 4. In the first step, we need to find the "clean" background scenes that do not contain any hands or interacting objects. The reason is that the image should only contain one egocentric viewer's hands and interacting objects after the composition. To this end, we propose two ways to generate "clean" background. The first is to build a simple binary classifier that finds the frames with no hands from a large pool of video frames, as shown in the top left of Fig. 4. The second way is to remove the existing hand-object using an image inpainting model [65] and the labeled segmentation masks, as shown in the top right of Fig. 4. Both approaches enable us generate a large pool of "clean" background candidates. On the other hand, when given an image with hand-object segmentation masks, we first inpaint the hand-object regions using the inpainting model [65] to generate the "clean" query background, and then use it to retrieve the top-K similar background scenes from the "clean" background candidate pool based on deep features extracted by [62]. Finally, multiple background scenes are sampled from the top-K retrieved background images, as shown in the bottom of Fig. 4. Overall, our designed context-aware image composition pipeline allows us to generate semantically consistent hand-object and context as much as needed. In the experiments, we show the effectiveness of our proposed data augmentation technique.

6 Experiments on Hand-Object Segmentation

In this section, we first make comparison studies with the existing datasets on hand segmentation, and then we discuss the benchmark performance of the hand-object segmentation with an ablation study. In order to evaluate the in-the-wild segmentation performance, we sparsely sampled 500 frames from 30 collected Youtube egocentric videos to label as our out-of-distribution test set. In the following segmentation experiments, all results are evaluated on this test set unless otherwise specified. All of our models are trained and evaluated using the MMSegmentation codebase[1].

6.1 Two-Hand Segmentation

Previous hand segmentation datasets have different definitions of hand labels, such as left/right hand [1], binary hand [73] or binary hand + arm [33]. Since our datasets provide all these types of hand labels, we compare individually with the previous datasets in their settings. For a fair comparison, we train the same ResNet-18 backbone [20] and HRNet head [74] on each dataset, select the best checkpoints based on the validation set, and finally compute the results on the same held-out test set.

Datasets	mIoU	mPrec	mRec	mF1
EgoHand[1]	10.68/33.28	43.61/43.20	12.39/59.16	19.30/49.93
100-DOH[58] + BoxInst[70]	36.30/37.51	50.06/61.63	56.91/48.94	53.27/54.55
Ours	76.29/77.00	83.39/87.06	89.97/86.95	86.55/87.00
+ CCDA	79.73/82.17	84.26/90.38	93.68/90.04	88.72/90.21

Table 2. Left/Right Hand Segmentation.

[1] MMSegmentation github: https://github.com/open-mmlab/mmsegmentation

Datasets	mIoU	mPrec	mRec	mF1
EgoHand[1]	56.51	76.33	68.52	72.22
100-DOH[58]+ BoxInst[70]	69.50	84.80	79.67	82.00
EYTH[73]	75.94	85.17	87.51	86.32
Ours	83.18	89.34	92.34	90.82
+ CCDA	85.45	90.11	94.3	92.15

Table 3. Binary Hand Segmentation.

Datasets	mIoU	mPrec	mRec	mF1
EGTEA [33]	33.26	38.24	71.87	49.92
Ours	92.46	96.67	95.50	96.08
+ CCDA	95.20	97.68	97.40	97.54

Table 4. Binary Hand + Arm Segmentation.

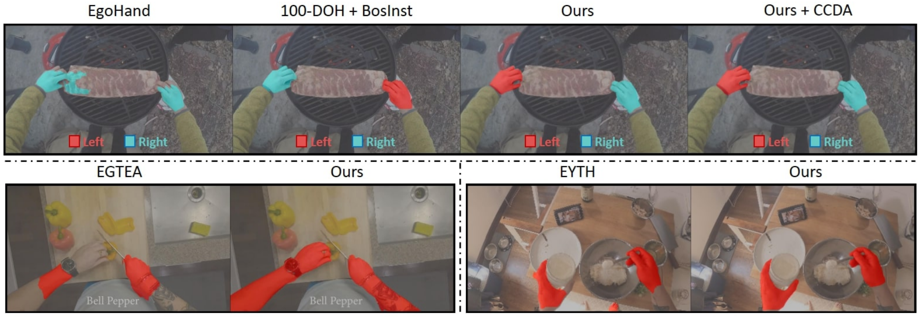

Fig. 5. A qualitative comparison between segmentation models trained on previous datasets and our proposed dataset. The **top row** shows the comparison with EgoHand and 100-DOH + BoxInst in Left/Right Hand segmentation, where red and cyan indicate left hand and right hand respectively. The **bottom left** shows the comparison with EGTEA on binary Hand + Arm segmentation. The **bottom right** shows the comparison with EYTH on binary Hand segmentation.

In the first type of labeling, EgoHand [1] labeled the left and right hands of people in egocentric activities. Similarly, 100-DOH [58] also labeled large-scale left/right hands but with only bounding box annotations. We compare with 100-DOH by training a weakly supervised segmentation model, BoxInst [70], which learns to segment objects given bounding box annotation. To make the hand segmentation performance as good as possible for 100-DOH, we pre-trained BoxInst on 2000 frames sampled from EPIC-KITCHEN. As shown in Table 2, the model trained on our dataset significantly outperforms the model trained on EgoHand and 100-DOH with BoxInst. From the visual results, as shown in the first row of Fig. 5, we observe that models trained on EgoHand and 100-DOH often generate wrong mask categorical labels, which causes significantly lower mIoU (mean Intersection over Union) for left/right hand segmentation. When we binarize the predicted left/right hand masks of EgoHand and 100-DOH and evaluate them on the binary hand segmentation task, the performance gap bridges closer to us, as shown in Table 3. This again shows that mis-classification of left/right hand is indeed a major issue that causes low mIoU in Table 2 for the models trained on [1,58].

The other two datasets EYTH [73] and EGTEA [33] provide only binary mask labels for both hands without differentiating between left or right. EYTH [73] labeled only the hand region, and EGTEA [33] labeled both hand and arm regions. In Table 3 and Table 4, the quantitative results show that the model trained on our datasets also outperforms previous datasets by an obvious margin in both "hand" and "hand + arm" settings, and visual comparisons are shown in the bottom of Fig. 5. In all these hand segmentation settings, we observe that our context-aware compositional data augmentation (CCDA) consistently improves the hand segmentation performance quantitatively. More qualitative comparisons are included in the supplementary materials.

6.2 Hand and Interacting Object Segmentation

Since our dataset is the first to provide mask labels for interacting objects, we discuss the benchmark performance of hand-object segmentation with an ablation study in this section. In this task, we assign hands to left and right categories, and objects to three categories based on the interacting hand: left-hand object, right-hand object, and two-hand object. A naive solution is to train a segmentation network that decodes five channels of outputs in parallel, as shown in the 1^{st} row of Table 5. However, this might not be ideal, since parallel decoding of outputs does not leverage any explicit understanding of the hand-object relationship, as discussed in Section 4. Thus, we propose to try to sequentially decode the hand first, and then use predicted left/right hand mask information to explicitly guide the interacting object segmentation, as shown in the 3^{rd} row of Table 5. We also studied adding contact boundary (CB) as intermediate guide information, and found that it effectively boosts the object segmentation performance, as shown in the comparison between 3^{rd} and 5^{th} rows. More details about contact boundary are discussed in Section 4. Finally, we evaluated the effectiveness of our context-aware compositional data augmentation (CCDA) by integrating it on top of both parallel and sequential models. As shown in the comparison between rows $1^{st}, 3^{rd}, 5^{th}$ and rows $2^{nd}, 4^{th}, 6^{th}$, CCDA slightly improves the left/right hand segmentation and significantly boosts the object segmentation performance. We think the reasons are that compositional augmentation enables the network to learn the pixel grouping of objects more easily when placing them into many different background. More details are on how we choose the quantity of augmented images are are discussed in the supplementary materials. The qualitative results for dense contact boundary prediction and hand-object segmentation in diverse activities are shown in the supplemental materials.

7 Applications

7.1 Boosting Hand State Classification and Activity Recognition

Understanding the hand state and recognizing types of activities in egocentric videos are important for human behavior analysis. Similarly to 100-DOH [58],

Models	Left Hand	Right Hand	Left-Hand Object	Right-Hand Object	Two-Hand Object
Para. Decode	69.08	73.50	48.67	36.21	37.46
Para. Decode + CCDA	77.57	81.06	54.83	38.48	39.14
Seq. Decode	73.17	80.56	54.83	38.48	39.14
Seq. Decode + CCDA	87.70	88.79	58.32	40.18	46.24
Seq. Decode + CB	77.25	81.17	59.05	40.85	49.94
Seq. Decode + CB + CCDA	87.70	88.79	62.20	44.40	52.77

Table 5. A quantitative ablation study on the hand-object segmentation.

we define hand states for both left and right hands as the following: (contact with) portable, (contact with) stationary, no-contact, self-contact, and not-exist. The goal of this task is to classify a correct state for each of the two hands of the egocentric viewer, where we use two classification heads to handle this. To this end, we labeled the hand states of 3,531 frames from EPIC-KITCHEN [8] dataset with diverse hand-object interaction. During training, we adopt 8:1:1 ratio to split train, val, and test sets. As shown in Table 6, by adding hand mask and hand-object masks into the input channel, a classifier with the same backbone [62] could effectively improve its classification performance compared to the baseline that uses RGB images only. For the video activity recognition, we used a subset data of EPIC-KITCHEN Action Recognition benchmark [8] as well as its evaluation protocol. With the SlowFast network [11], we show that by adding hand masks into training, the top-1 classification accuracy of "verbs/nouns" boosts from 23.95%/36.77% to 25.98%/37.04%. A visual illustration of these two tasks is shown in the supplemental.

Models	Accuracy	Precision	Recall	F1 score
Baseline	78.53%/74.29%	43.02%/37.70%	36.86%/30.60%	37.53%/32.02%
+ Hand Mask	84.18%/83.33%	64.33%/66.16%	57.31%/56.34%	59.42%/59.00%
+ Hand & Object Mask	86.72%/83.33%	68.18%/69.04%	57.12%/56.08%	60.35%/59.64%

Table 6. Quantitative results for left/right hand state classification.

7.2 Improved 3D Mesh Reconstruction of Hand-Object Interaction

Mesh reconstruction of hand-object interaction is a useful but very challenging task. One mainstream approach [6,18] to solve this task is to jointly optimize the 3D scale, translation and rotation of a given 3D object model, as well as the MANO parameters [54] for hand. Such optimization process often relies on the estimated hand and object segmentation masks to compute the 3D mesh to 2D projection error. In previous works, researchers [6,18] leverage the 100-DOH's

[58] detector to localize the hand and interacting object at bounding box level, and then use PointRend [25] pre-trained on COCO [39] to segment the hand and object masks. The interacting object mask is assigned by a heuristic that the object mask with highest confidence score is the one in interaction.

In this work, we integrate our robust hand-object segmentation model into the previous mesh reconstruction pipeline [18]. Since our hand-object segmentation could generalize better than the previous segmentation component, we enable the hand-object reconstruction generalize in more diverse scenarios with higher visual fidelity. As shown in the first row of Fig. 6, the previous segmentation pipeline oftentimes fails to segment the complete object, and thus the object was optimized into a wrong 3D pose, while our accurate hand-object segmentation enables the object mesh reconstruction to be more accurate. In the second row of Fig. 6, we observe that the previous segmentation pipeline sometimes completely misses the interacting object at the bounding box detection stage, and thus no segmentation and 3D mesh could be generated. In contrast, our pipeline provides higher recall on the object detection, and thus is able to recover the object mesh, as shown in the bottom right of Fig. 6.

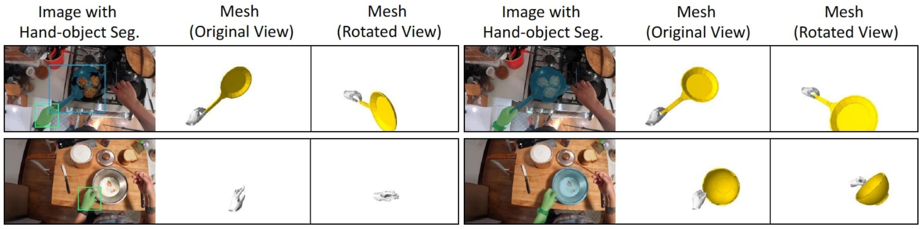

Fig. 6. Visual comparison between 3D mesh reconstruction of hand-object interaction. The **left** results are from the original code of [18], where they use 100-DOH [58] detector with PointRend [25] to compute hand-object masks. The **right** results are computed by integrating our hand-object segmentation into [18] for mesh optimization.

7.3 Seeing Through the Hand in Egocentric Videos

Finally, in this work, we propose a new interesting application, where the goal is to see through the hand in egocentric videos. With our robust per-frame segmentation of hand masks, we use the recent flow-guided video inpainting algorithm [12] to completely remove the hands such that we could see the original content occluded by hands in the videos. A visual example of this application is shown in supplemental, where the hand is removed and the bottles and fridge layers that are originally occluded can now been visualized in every video frame. More video results are included in the supplementary materials. In the egocentric videos, since hands are prevalent and almost moving all the time, they create large occlusions of visual contents. The practical use of our "hand see through" system is that we could potentially enable the vision system analyze more previously occluded information, for example, in the future AR system.

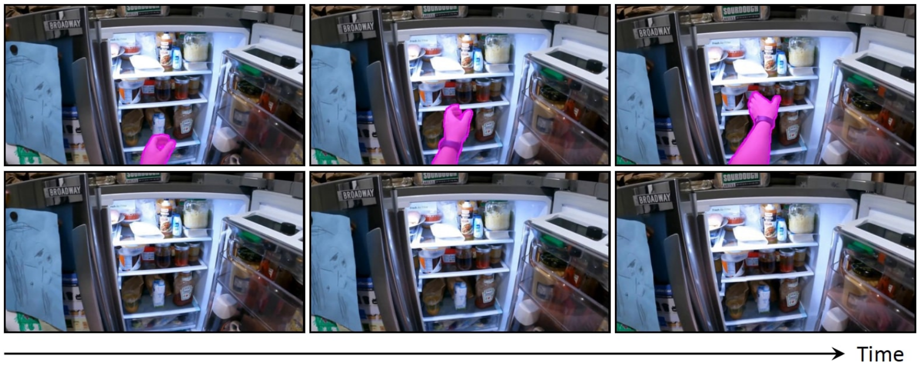

Fig. 7. A qualitative demo to show the application of seeing through the hand in egocentric videos. This application is enabled by our robust per-frame hand segmentation together with the video inpainting model [12]. The top row are the frames with predicted hand segmentation masks, and the bottom row shows the "see through" frames at the corresponding timestamp. More video results are shown in the supplemental.

8 Conclusion

We created a fine-grained egocentric hand-object segmentation dataset and synthetic data augmentation method to 1) enable robustness against out-of-distribution domain change and 2) support downstream tasks. Our labeled dataset of 11,243 images contains both per-pixel segmentation labels of hand and interacting objects and dense contact boundaries. Our context-aware compositional data augmentation technique significantly improves segmentation performance, especially for interacting objects. We show that our robust hand-object segmentation model can serve as a foundational tool for several vision applications, including hand state classification, activity recognition, 3D mesh reconstruction of hand-object interaction, and seeing through the hand in egocentric videos.

Acknowledgment. This research is based on work supported by Toyota Research Institute and Adobe Gift Fund. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsors.

References

1. Bambach, S., Lee, S., Crandall, D.J., Yu, C.: Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1949–1957 (2015) 2, 4, 6, 9, 10, 11
2. Brahmbhatt, S., Handa, A., Hays, J., Fox, D.: Contactgrasp: Functional multi-finger grasp synthesis from contact. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2386–2393. IEEE (2019) 4
3. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: Contactpose: A dataset of grasps with object contact and hand pose. In: European Conference on Computer Vision. pp. 361–378. Springer (2020) 4
4. Cai, M., Lu, F., Sato, Y.: Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14392–14401 (2020) 4
5. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3d hand pose estimation from monocular rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 666–682 (2018) 4
6. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12417–12426 (2021) 4, 12
7. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: Ganhand: Predicting human grasp affordances in multi-object scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5031–5041 (2020) 4
8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 720–736 (2018) 2, 5, 6, 12
9. Fang, H.S., Sun, J., Wang, R., Gou, M., Li, Y.L., Lu, C.: Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 682–691 (2019) 8
10. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: CVPR 2011. pp. 3281–3288. IEEE (2011) 3
11. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019) 12
12. Gao, C., Saraf, A., Huang, J.B., Kopf, J.: Flow-edge guided video completion. In: European Conference on Computer Vision. pp. 713–729. Springer (2020) 3, 13, 14
13. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2918–2928 (2021) 8
14. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9785–9795 (2019) 4
15. Goyal, M., Modi, S., Goyal, R., Gupta, S.: Human hands as probes for interactive object understanding. arXiv preprint arXiv:2112.09120 (2021) 5
16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. arXiv preprint arXiv:2110.07058 (2021) 2, 5, 6

17. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 571–580 (2020) 4
18. Hasson, Y., Varol, G., Schmid, C., Laptev, I.: Towards unconstrained joint hand-object reconstruction from rgb videos. In: 2021 International Conference on 3D Vision (3DV). pp. 659–668. IEEE (2021) 3, 4, 12, 13
19. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11807–11816 (2019) 4
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 8, 9
21. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. arXiv preprint arXiv:2104.03304 (2021) 4
22. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. International journal of computer vision **46**(1), 81–96 (2002) 3
23. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 2020 International Conference on 3D Vision (3DV). pp. 333–344. IEEE (2020) 4
24. Kim, S., Chi, H.G.: First-person view hand segmentation of multi-modal hand activity video dataset. BMVC 2020 (2020) 4
25. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020) 4, 13
26. Kulon, D., Guler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4990–5000 (2020) 4
27. Kulon, D., Wang, H., Güler, R.A., Bronstein, M., Zafeiriou, S.: Single image 3d hand reconstruction with mesh convolutions. arXiv preprint arXiv:1905.01326 (2019) 4
28. Kundu, A., Li, Y., Rehg, J.M.: 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3559–3568 (2018) 4
29. Kuo, W., Angelova, A., Lin, T.Y., Dai, A.: Mask2cad: 3d shape prediction by learning to segment and retrieve. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 260–277. Springer (2020) 4
30. Lee, S., Bambach, S., Crandall, D.J., Franchak, J.M., Yu, C.: This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 543–550 (2014) 3
31. Li, C., Kitani, K.M.: Model recommendation with virtual probes for egocentric hand detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2624–2631 (2013) 3
32. Li, C., Kitani, K.M.: Pixel-level hand detection in ego-centric videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3570–3577 (2013) 3

33. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 619–635 (2018) 2, 4, 6, 9, 10, 11
34. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 287–295 (2015) 4
35. Lim, J.J., Pirsiavash, H., Torralba, A.: Parsing ikea objects: Fine pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2992–2999 (2013) 4
36. Lin, F., Martinez, T.: Ego2hands: A dataset for egocentric two-hand segmentation and detection. arXiv preprint arXiv:2011.07252 (2020) 4
37. Lin, F., Wilhelm, C., Martinez, T.: Two-hand global 3d pose estimation using monocular rgb. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2373–2381 (2021) 4
38. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1925–1934 (2017) 4
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 3, 4, 13
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021) 8
41. Mandikal, P., Grauman, K.: Dexvip: Learning dexterous grasping with human hand pose priors from video. In: 5th Annual Conference on Robot Learning (2021) 5
42. Mandikal, P., Grauman, K.: Learning dexterous grasping with object-centric visual affordances. In: IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021. pp. 6169–6176. IEEE (2021) 5
43. Michel, F., Kirillov, A., Brachmann, E., Krull, A., Gumhold, S., Savchynskyy, B., Rother, C.: Global hypothesis generation for 6d object pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 462–471 (2017) 4
44. Moon, G., Chang, J.Y., Lee, K.M.: V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE conference on computer vision and pattern Recognition. pp. 5079–5088 (2018) 4
45. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Ganerated hands for real-time 3d hand tracking from monocular rgb. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 49–59 (2018) 4
46. Muller, L., Osman, A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9990–9999 (2021) 4
47. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8688–8697 (2019) 5
48. Nagarajan, T., Grauman, K.: Shaping embodied agent behavior with activity-context priors from egocentric video. Advances in Neural Information Processing Systems **34** (2021) 5

49. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 163–172 (2020) 5
50. Narasimhaswamy, S., Nguyen, T., Nguyen, M.H.: Detecting hands and recognizing physical contact in the wild. Advances in neural information processing systems **33**, 7841–7851 (2020) 4
51. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019) 4
52. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3137–3144. IEEE (2010) 3
53. Romero, J., Kjellström, H., Kragic, D.: Hands in action: real-time 3d reconstruction of hands in interaction with objects. In: 2010 IEEE International Conference on Robotics and Automation. pp. 458–463. IEEE (2010) 4
54. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017) 12
55. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. arXiv preprint arXiv:2008.08324 (2020) 4
56. Rother, C., Kolmogorov, V., Blake, A.: " grabcut" interactive foreground extraction using iterated graph cuts. ACM transactions on graphics (TOG) **23**(3), 309–314 (2004) 4
57. Sahasrabudhe, M., Shu, Z., Bartrum, E., Alp Guler, R., Samaras, D., Kokkinos, I.: Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 4
58. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9869–9878 (2020) 2, 4, 6, 9, 10, 11, 13
59. Shan, D., Higgins, R., Fouhey, D.: Cohesiv: Contrastive object and hand embedding segmentation in video. Advances in Neural Information Processing Systems **34** (2021) 4
60. Sharp, T., Keskin, C., Robertson, D., Taylor, J., Shotton, J., Kim, D., Rhemann, C., Leichter, I., Vinnikov, A., Wei, Y., et al.: Accurate, robust, and flexible real-time hand tracking. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems. pp. 3633–3642 (2015) 4
61. Shilkrot, R., Narasimhaswamy, S., Vazir, S., Hoai, M.: Workinghands: A hand-tool assembly dataset for image segmentation and activity mining. In: BMVC. p. 258 (2019) 4
62. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9, 12
63. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive markerless articulated hand motion tracking using rgb and depth data. In: Proceedings of the IEEE international conference on computer vision. pp. 2456–2463 (2013) 4
64. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In:

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2974–2983 (2018) 4

65. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2149–2159 (2022) 8, 9
66. Tagliasacchi, A., Schröder, M., Tkach, A., Bouaziz, S., Botsch, M., Pauly, M.: Robust articulated-icp for real-time hand tracking. In: Computer Graphics Forum. vol. 34, pp. 101–114. Wiley Online Library (2015) 4
67. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision. pp. 581–600. Springer (2020) 4
68. Tang, Y., Tian, Y., Lu, J., Feng, J., Zhou, J.: Action recognition in rgb-d egocentric videos. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3410–3414. IEEE (2017) 2, 5, 6
69. Tekin, B., Bogo, F., Pollefeys, M.: H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2019) 4
70. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5443–5452 (2021) 9, 10
71. Tulsiani, S., Gupta, S., Fouhey, D.F., Efros, A.A., Malik, J.: Factoring shape, pose, and layout from the 2d image of a 3d scene. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 302–310 (2018) 4
72. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision **118**(2), 172–193 (2016) 4
73. Urooj, A., Borji, A.: Analysis of hand segmentation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4710–4719 (2018) 2, 4, 6, 9, 10, 11
74. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020) 8, 9
75. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10974 (2019) 4
76. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017) 4
77. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018) 8
78. Yang, L., Yao, A.: Disentangling latent hands for image synthesis and pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9877–9886 (2019) 4
79. Ye, Q., Yuan, S., Kim, T.K.: Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In: European conference on computer vision. pp. 346–361. Springer (2016) 4

80. Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J.Y., Lee, K.M., Molchanov, P., Kautz, J., Honari, S., Ge, L., et al.: Depth-based 3d hand pose estimation: From current achievements to future goals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2636–2645 (2018) 4
81. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision. pp. 34–51. Springer (2020) 4
82. Zhang, L., Wen, T., Min, J., Wang, J., Han, D., Shi, J.: Learning object placement by inpainting for compositional data augmentation. In: European Conference on Computer Vision. pp. 566–581. Springer (2020) 8
83. Zhou, Y., Habermann, M., Xu, W., Habibie, I., Theobalt, C., Xu, F.: Monocular real-time hand shape and motion capture using multi-modal data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5346–5355 (2020) 4
84. Zhu, X., Jia, X., Wong, K.Y.K.: Pixel-level hand detection with shape-aware structured forests. In: Asian Conference on Computer Vision. pp. 64–78. Springer (2014) 3
85. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE international conference on computer vision. pp. 4903–4911 (2017) 4