Whose Hands are These?
Hand Detection and Hand-Body Association in the Wild

Supreeth Narasimhaswamy[1], Thanh Nguyen[2], Mingzhen Huang[3], Minh Hoai[1,2]
[1]Stony Brook University, [2]VinAI Research, [3] University at Buffalo

Abstract

We study a new problem of detecting hands and finding the location of the corresponding person for each detected hand. This task is helpful for many downstream tasks such as hand tracking and hand contact estimation. Associating hands with people is challenging in unconstrained conditions since multiple people can be present in the scene with varying overlaps and occlusions.

We propose a novel end-to-end trainable convolutional network that can jointly detect hands and the body location for the corresponding person. Our method first detects a set of hands and bodies and uses a novel Hand-Body Association Network to predict association scores between them. We use these association scores to find the body location for each detected hand. We also introduce a new challenging dataset called BodyHands containing unconstrained images with hand and their corresponding body locations annotations. We conduct extensive experiments on BodyHands and another public dataset to show the effectiveness of our method. Finally, we demonstrate the benefits of hand-body association in two critical applications: hand tracking and hand contact estimation. Our experiments show that hand tracking and hand contact estimation methods can be improved significantly by reasoning about the hand-body association. Code and data can be found at http://vision.cs.stonybrook.edu/~supreeth/BodyHands/

1. Introduction

Hand analysis is an important problem in Computer Vision with applications in human understanding, action, gesture, and sign-language recognition. The visual analysis of human hands is also vital for Augmented & Virtual Reality applications. Although the Computer Vision community has studied problems such as hand detection [7, 13, 23, 24, 30, 39, 49, 58, 59], hand pose estimation [17, 44, 51, 66, 67], hand tracking [48, 52, 53, 55, 63], and hand contact estimation [5, 33, 36, 47], there has been no significant effort in studying hand-body association.

Figure 1. **Hand Detection & Hand-Body Association.** We develop a method to detect hands and their corresponding body locations. Hands and bodies belonging to the same person have bounding boxes in the same color and identification numbers.

In this work, we study the problem of detecting hands in an image and finding the location of the corresponding person for each detected hand. This task is useful for action recognition and scene understanding, especially for multiple-person images and videos. For example, it is helpful to identify people when understanding hand gestures in human-human communication. Another example is to assess the motor and social skills of children with mental disorders by tracking their hands and how hands interact with objects and other people in a tabletop game. Hand-body association helps develop safety applications and assists people working with hand-held tools in manufacturing settings.

Detecting hands and associating them with suitable bodies is challenging in unconstrained conditions. As shown in Fig. 1, an image may contain multiple people with substantial overlaps and occlusions between hands and bodies. One approach is to detect hands and people separately and use heuristics based on their sizes, distances, or overlapping areas to establish correspondences between hands and bodies. However, such methods do not perform well due to the extreme articulation of hands and bodies, leading to tremendous variations in the relative locations and sizes between a hand and the corresponding human body. An alternative method is to use a human pose detector to find

the skeleton of humans and find the hands of each detected pose in the image. However, pose detection by itself is unreliable. For a scene of congregated or interacting people, the hand and arm of one person might be entangled with the skeleton of another person. Furthermore, the pose detector might not detect poses for everyone in the image, especially for people who are partially occluded or partly outside the camera's field of view. Thus we cannot solely rely on pose detection to associate hands with people. Our experiments empirically show that pose-based approaches are unreliable for associating hands and bodies.

This work proposes a novel convolutional architecture that can jointly detect hands and bodies and associate them. Specifically, we build upon MaskRCNN [18], a state-of-the-art object detector, and extend it by adding a novel Hand-Body Association Network module. We first use a Region Proposal Network to generate candidate hand and body proposal boxes. We then use the bounding box regression and mask generation heads to obtain the bounding box and segmentation maps for hands and bodies. The detected hands and bodies are then passed to the Hand-Body Association Network to obtain an association between them.

The Hand-Body Association Network has two novel modules. **The first module** is the Overlap Estimation Module that uses the visual features of hands and bodies to estimate if they can overlap. Intuitively, if a hand and a body have no overlap, they cannot belong to the same person. The converse, however, does not hold; a hand and a body can overlap even though they belong to different people. For instance, in the proposed BodyHands dataset, more than 33% of the people have their hands overlapping with other people. The overlap is a piece of mutual geometric information between two regions. Learning mutual geometric information between hands and bodies using their appearance features allows learning-rich discriminative representations useful for associating hands and bodies. **The second module** is the Positional Density Module that uses hand features to estimate a density over possible body locations for each detected hand. Intuitively, the appearance and location of a hand provide some cues for estimating its body location. However, directly locating the body from the hand can be difficult due to the tremendous variation in relative scales between hands and bodies and mutual occlusions between people. We thus first estimate a density over possible locations and use these density values to find compatible matching for all hand-body pairs using the Hungarian Algorithm.

We also contribute a large-scale dataset of unconstrained images containing annotations for hand locations and their corresponding body locations. The dataset has around 20K images with bounding box annotations for more than 57K hand and 63K body instances. This dataset has numerous images containing multiple people with varying degrees of occlusions and overlap, where it is challenging to detect and associate hands and bodies.

Finally, we demonstrate the benefits of the hand-body association in two crucial downstream tasks: hand tracking and hand physical contact estimation. We show that hand tracking and hand contact estimation methods can be improved by reasoning about the hand-body association.

2. Related Work

Hand Analysis. Hands have been extensively studied by the Computer Vision community and there are methods for hand detection [7, 12, 13, 21, 23, 24, 30, 35, 39, 45, 58, 59], hand pose estimation [9, 10, 17, 22, 25, 28, 44, 46, 51, 60, 66, 67], hand tracking [32, 48, 52, 53, 55, 63], and hand contact estimation [5, 33, 36, 47]. However, previous works do not consider the problem of hand-body association. Existing works mostly focus on constrained scenarios such as ego-centric perspectives with a single subject in a video. In such cases, the full body is not always visible, and it may not be essential to find them. Some works analyze hands in unconstrained conditions [35, 36, 47] but they do not address this problem either. Zhou et al. [64] address the problem of hand-raiser recognition in classroom scenarios. However, their work is developed for indoor classroom environments and is unsuitable for unconstrained outside environments. Lee et al. [26] and Tsutsui et al. [54] study the problem of hand disambiguation in egocentric videos. However, they identify only the person's identity but not their body locations. On the other hand, we try to address the hand-body association problem, and our work focuses mainly on third-person views.

Hand datasets. Although there are several datasets with annotated hand locations, such as [35, 36], they do not have annotations for the corresponding body locations. Another option would be to use human pose datasets [1, 2, 27], which have human body joint locations. However, such datasets do not have bounding box annotations for hands. Zhou et al. [64] propose a dataset containing hand and body locations, but they develop the dataset in indoor environments. Moreover, their dataset is not publicly available. Bambach et al. [3] propose a dataset containing 48 videos of first-person interaction between two people. However, they provide annotations for only hand locations but not body locations. Jin et al. [20] develop the COCO-WholeBody dataset by annotating hand key points for images from the COCO dataset. Compared to this dataset, the proposed BodyHands dataset has a higher number of crowded images with significant overlaps and occlusions between people; 34% people in BodyHands have their hands overlapping with different people compared to 19% from COCO-WholeBody.

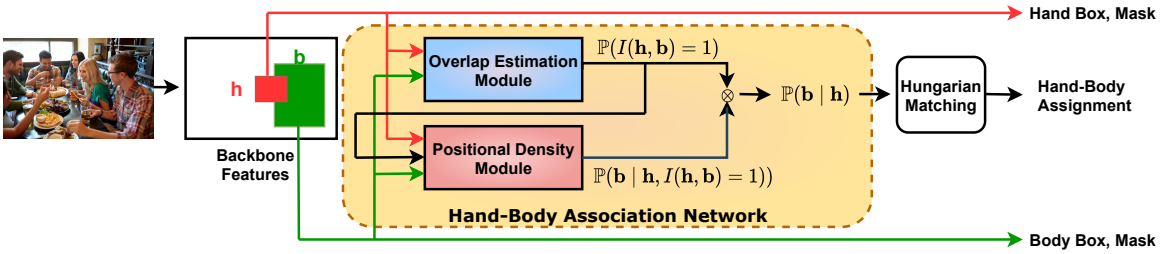

Figure 2. **Proposed Architecture.** A ResNet network extracts the backbone features of the input image. We use the feature maps of hand and body proposal boxes to obtain their bounding boxes and binary segmentation masks. The Overlap Estimation Module uses the feature maps of the hand and body to predict if they can overlap, i.e., $\mathbb{P}(I(\mathbf{h},\mathbf{b}) = 1)$. The Positional Density Module uses the hand features and the output from the Overlap Estimation Module to estimate the conditional likelihood $\mathbb{P}(\mathbf{b} \mid \mathbf{h}, I(\mathbf{h},\mathbf{b}) = 1)$. The outputs from these two modules are combined to obtain the likelihood that the body $\mathbf{b}$ belongs to the hand $\mathbf{h}$, i.e., $\mathbb{P}(\mathbf{b}|\mathbf{h})$. We use the estimated conditional likelihood to find compatible matching for all hand-body pairs using the Hungarian Algorithm (used only during inference).

3. Problem Definition and Proposed Method

This section describes the proposed architecture that can jointly detect hands and bodies and provide an association score between them. We also provide details on the training objective that allows training the proposed architecture end-to-end. In the following subsection, we will first formally define the problem.

3.1. Problem Definition

Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, our goal is to:

1. Detect the bounding box locations $\mathbf{H} = \{\mathbf{h}_i \in \mathbb{R}^4 : 1 \le i \le m\}$ and $\mathbf{B} = \{\mathbf{b}_j \in \mathbb{R}^4 : 1 \le j \le n\}$ for hands and bodies, respectively. Here, m and n denote the number of hands and bodies in the image $\mathbf{I}$. Each bounding box is represented by a 4-dimensional vector for its left, top, right, bottom locations.

2. For each detected hand $\mathbf{h} \in \mathbf{H}$, we need to associate a body $\mathbf{b} \in \mathbf{B}$ such that the following two constraints are satisfied: (1) each hand $\mathbf{h} \in \mathbf{H}$ is associated with exactly one body $\mathbf{b} \in \mathbf{B}$; (2) each body $\mathbf{b} \in \mathbf{B}$ can be associated to at most two hands in $\mathbf{H}$. Note that we consider any visible regions of the human as the body. Therefore when the detector fails to detect any humans, i.e., $\mathbf{B} = \emptyset$, we treat a hand bounding box as its corresponding person bounding box.

3.2. Architecture Overview

We illustrate the proposed architecture in Fig. 2. We build upon a two-stage object detector [15, 16, 18, 43] such as MaskRCNN. Given an input image, we use a ResNet to obtain backbone features and a Region Proposal Network to obtain proposals corresponding to two object classes: hands and bodies. We then use the RoIAlign operation to extract features corresponding to these proposals and perform bounding box regression and mask generation.

For each detected hand $\mathbf{h} \in \mathbf{H}$, we use a novel Hand-Body Association Network to estimate the conditional-likelihood $\mathbb{P}(\mathbf{b}|\mathbf{h})$ over all the detected bodies $\mathbf{b} \in \mathbf{B}$. The conditional $\mathbb{P}(\mathbf{b}|\mathbf{h})$ denotes the probability that the body $\mathbf{b}$ is associated with the hand $\mathbf{h}$. We use $\mathbb{P}(\mathbf{b}|\mathbf{h})$ as weights of a bipartite graph between hands and bodies and pose the hand-body association problem as finding a maximum-weighted assignment satisfying the constraints described by the problem definition in Sec 3.1. We finally use the Hungarian algorithm [34] to obtain a solution for this matching problem. We implement the Hand-Body Association Network as a new branch of MaskRCNN and train this module end-to-end together with other MaskRCNN components.

3.3. Hand-Body Association Network

The inputs to the Hand-Body Association Network are the set of detected hands $\mathbf{H}$ and bodies $\mathbf{B}$. For each detected hand instance $\mathbf{h} \in \mathbf{H}$, it outputs the conditional-likelihood $\mathbb{P}(\mathbf{b}|\mathbf{h})$ over all bodies $\mathbf{b} \in \mathbf{B}$. The probability $\mathbb{P}(\mathbf{b}|\mathbf{h})$ is high whenever the body $\mathbf{b}$ belongs to the hand $\mathbf{h}$, otherwise it is low. We show that under some independence assumptions, the term $\mathbb{P}(\mathbf{b}|\mathbf{h})$ can be factorized as a product of two terms involving overlap between $\mathbf{h}$ and $\mathbf{b}$ and positional density over $\mathbf{b}$:

$$\mathbb{P}(\mathbf{b}|\mathbf{h}) = \underbrace{\mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 1)}_{\text{overlap between h \& b}} \cdot \underbrace{\mathbb{P}(\mathbf{b}|\mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 1)}_{\text{density over b}}. \quad (1)$$

To see this, we first note an important relationship between the hands and body that belong to the same person. Since hands are a part of the human body, a hand bounding box and a body bounding box that belong to the same person must have a positive overlap. In other words, if a hand and a body have no overlap, they cannot belong to the same person. The converse, however, does not hold. Hands and bodies can overlap even though they belong to different people. For instance, in the proposed BodyHands dataset, more than 33% of people have their hands significantly overlapping with other people. Formally, if we let $I_{\mathbf{h},\mathbf{b}}$ be an indicator random variable to denote whether $\mathbf{h}$

and $\mathbf{b}$ have any overlap, we have

$$\mathbb{P}(\mathbf{b}|\mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 0) = 0. \quad (2)$$

We can use the law of total probability and condition over possible values of $I_{\mathbf{h},\mathbf{b}} \in \{0, 1\}$ to write:

$$\begin{aligned}\mathbb{P}(\mathbf{b} \mid \mathbf{h}) = \mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 0 \mid \mathbf{h}) \cdot \mathbb{P}(\mathbf{b} \mid \mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 0) \\ + \mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 1 \mid \mathbf{h}) \cdot \mathbb{P}(\mathbf{b} \mid \mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 1), \quad (3)\end{aligned}$$

Combining Eq. (2) and Eq. (3), we get:

$$\mathbb{P}(\mathbf{b} \mid \mathbf{h}) = \mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 1 \mid \mathbf{h}) \cdot \mathbb{P}(\mathbf{b} \mid \mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 1). \quad (4)$$

The independence assumption $\mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 1|\mathbf{h}) = \mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 1)$ reduces Eq. (4) to Eq. (1). We learn the probabilities $\mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 1)$ using the Overlap Estimation Module and $\mathbb{P}(\mathbf{b} \mid \mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 1)$ using the Positional Density Module.

Overlap Estimation Module. This module takes as input the visual features corresponding to the hand bounding box $\mathbf{h}$ and the body bounding box $\mathbf{b}$ and estimate the probability of them overlapping each other. Specifically, we use a neural network $f_{overlap}$ to model $\mathbb{P}(I_{\mathbf{h},\mathbf{b}} = 1) := f_{overlap}(\mathbf{h}, \mathbf{b})$.

We implement $f_{overlap}$ as an additional branch of MaskRCNN using convolutional and fully-connected layers. This network module is computationally light and we learn their parameters together with MaskRCNN during training using the following binary cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{overlap} := &- Y^{(gt)}_{\mathbf{h},\mathbf{b}} \log f_{overlap}(\mathbf{h}, \mathbf{b}) \\ &- \left(1 - Y^{(gt)}_{\mathbf{h},\mathbf{b}}\right) \log\left(1 - f_{overlap}(\mathbf{h}, \mathbf{b})\right). \quad (5)\end{aligned}$$

In the above, $Y^{(gt)}_{\mathbf{h},\mathbf{b}}$ denotes the groundtruth and is equal to 1 if $\mathbf{h}$ and $\mathbf{b}$ overlap and 0 otherwise.

Note that we predict $f_{overlap}(\mathbf{h}, \mathbf{b})$ using the appearance features of the hand and the body rather than computing the overlap between bounding boxes $\mathbf{h}$ and $\mathbf{b}$ directly. This is because the overlap is a piece of mutual geometric information between two regions. Learning mutual geometric information between hands and bodies using their appearance features allows learning-rich discriminative representations useful for associating hands and bodies. We show this empirically in our experiments.

Positional Density Module. We use this module to model the term $\mathbb{P}(\mathbf{b} \mid \mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 1)$ in Eq. (1). Specifically, given any hand $\mathbf{h}$, for any possible body location $\mathbf{b}$ with $I_{\mathbf{h},\mathbf{b}} = 1$, we model this probability using the following distribution:

$$f_{density}\left(\mathbf{b}|\mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 1\right) \propto \exp\left(-\frac{||\mathbf{b}^{\mathbf{h}} - \mu^{\mathbf{h}}_{body}||}{2\sigma^2}\right). \quad (6)$$

In the above equation, $\mu^{\mathbf{h}}_{body} \in \mathbb{R}^4$ is the mean body location relative to the hand $\mathbf{h}$, $\mathbf{b}^{\mathbf{h}}$ is an encoding of the body box coordinates $\mathbf{b}$ relative to the hand $\mathbf{h}$, and σ is a tunable hyperparameter. More specifically, inspired by the bounding box regression formulation in FasterRCNN [42], we use

$$\mathbf{b}^{\mathbf{h}} = \left(\frac{\mathbf{b}_x - \mathbf{h}_x}{\mathbf{h}_w}, \frac{\mathbf{b}_y - \mathbf{h}_y}{\mathbf{h}_h}, \log\frac{\mathbf{b}_w}{\mathbf{h}_w}, \log\frac{\mathbf{b}_h}{\mathbf{h}_h}\right). \quad (7)$$

In the above, $(\mathbf{h}_x, \mathbf{h}_y)$ denotes the (x, y) coordinates of the center of $\mathbf{h}$, $\mathbf{h}_w$ and $\mathbf{h}_h$ denotes the width and height of $\mathbf{h}$. Similarly, $(\mathbf{b}_x, \mathbf{b}_y)$ denotes the (x, y) coordinates of the center of $\mathbf{b}$, $\mathbf{b}_w$ and $\mathbf{b}_h$ denotes the width and height of $\mathbf{b}$. We predict $\mu^{\mathbf{h}}_{body}$ in Eq. (6) using the appearance features and bounding box location of the hand $\mathbf{h}$.

Intuitively, the appearance features and location of the hand provide some cues on estimating its body location. However, directly locating the body from hand features can be difficult due to the tremendous variation in relative scales between hands and bodies and mutual occlusions between people. We, therefore, first estimate a density over possible locations and use these density values to find compatible matching for all hand-body pairs using the Hungarian Algorithm. If the body $\mathbf{b}$ is far from the estimated mean body location $\mu^{\mathbf{h}}_{body}$, then $\mathbb{P}(\mathbf{b}|\mathbf{h}, I_{\mathbf{h},\mathbf{b}} = 1)$ is small, and therefore according to Eq. (1), $\mathbb{P}(\mathbf{b}|\mathbf{h})$ is also small.

We can efficiently implement the network $f_{density}$ as an additional branch of MaskRCNN using convolutional and fully-connected layers. We train $f_{density}$ together with MaskRCNN end-to-end by minimizing the smooth-$L1$ loss [15] between the predicted $\mu^{\mathbf{h}}_{body}$ and the groundtruth body $\mathbf{b}^{\mathbf{h}}_{(gt)}$ associated with the hand $\mathbf{h}$:

$$\mathcal{L}_{density} := \sum_{i=1}^{4} \text{Smooth-}L_1\left(\mu^{\mathbf{h}}_{body}[i] - \mathbf{b}^{\mathbf{h}}_{(gt)}[i]\right), \quad (8)$$

In the above equation, $\mu^{\mathbf{h}}_{body}[i]$ and $\mathbf{b}^{\mathbf{h}}_{(gt)}[i]$ denote the i^{th} components of four dimensional vectors $\mu^{\mathbf{h}}_{body}$ and $\mathbf{b}^{\mathbf{h}}_{(gt)}$.

3.4. Training Objective

We train the proposed Hand-Body Association network together with the MaskRCNN end-to-end by optimizing the following multi-task loss:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} + \mathcal{L}_{association}. \quad (9)$$

Here, $\mathcal{L}_{cls}$, $\mathcal{L}_{box}$, $\mathcal{L}_{mask}$ denote the classification, the bounding box regression, and the segmentation mask losses for the detection. These are the standard losses used in MaskRCNN [18]. The term $\mathcal{L}_{association}$ denotes the hand-body association loss and is defined as:

$$\mathcal{L}_{association} := \lambda_1 \mathcal{L}_{overlap} + \lambda_2 \mathcal{L}_{density}, \quad (10)$$

In the above, $\mathcal{L}_{overlap}$ denotes the loss for the Overlap Estimation Module and $\mathcal{L}_{density}$ is the loss for the Positional Density Module. These losses are defined in Eq. (5) and Eq. (8). The scaling factors λ_1 and λ_2 are tunable hyperparameters denoting the relative importance between the overlap estimation and positional density estimation.

3.5. Hungarian Hand-Body Assignment

Given a set of detected hands $\mathbf{H} = \{\mathbf{h}_i : 1 \leq i \leq m\}$, bodies $\mathbf{B} = \{\mathbf{b}_j : 1 \leq j \leq n\}$, and the conditional distribution $\mathbb{P}(\mathbf{b}|\mathbf{h})$ estimated from the Hand-Body Association network, we need an assignment strategy to match hands and bodies subject to the constraints described in Sec. 3.1.

We follow the bipartite matching strategy and use $\mathbb{P}(\mathbf{b}|\mathbf{h})$ as the weight between hand $\mathbf{h}$ and body $\mathbf{b}$ in the bipartite graph. We obtain a maximum-weighted assignment between the detected hands $\mathbf{H}$ and bodies $\mathbf{B}$ using the Hungarian Algorithm [34].

Note that the Hungarian algorithm matches each hand with exactly one body, but also it produces an undesirable result: each body can match to at most one hand. However, we need the flexibility to match a body to two hands. We provide a simple solution to this by duplicating $\mathbf{B}$ to ensure that each body is present exactly twice before running the Hungarian algorithm. This ensures that a body can have two hands associated with them.

4. BodyHands Dataset

This section describes BodyHands, the new dataset collected to develop and evaluate hand-body association methods. BodyHands is a large-scale dataset containing unconstrained images with annotations for hand and body locations and correspondences.

Dataset Source. We built the BodyHands dataset starting from the images from the ContactHands dataset [36]. ContactHands is a large-scale dataset containing unconstrained images annotated with hand polygon locations and their contact states. It has images from popular datasets such as MS COCO [27], PASCAL VOC [14], Oxford-Hand [30], TV-Hand [35], and COCO-Hand [35]. We chose the ContactHands dataset for several reasons. First, we wanted to develop a dataset that we can use to train methods that robustly detect and associate hands and bodies regardless of shape, size, skin tone, and motion blur. Second, we want to detect and associate hands and bodies in challenging cases where people have mutual occlusions. Third, we wanted to use hand-body association in existing applications such as hand contact estimation and thus require contact state annotations. The ContactHands dataset has numerous images which satisfy these requirements.

Annotation and Quality Control. We hired several annotation workers to annotate our dataset. For each person with an annotated hand instance in the ContactHands dataset, we asked an annotator to draw a rectangular bounding box around the person and enter an identification number for the hand and the body. The hands and body which belong to the same person have the same identification number and therefore serve as an association between hands and bodies. We asked the annotators to draw the human bounding box to include all visible parts of the person. If an image contains N people who did not have hand location annotations, we asked annotators to annotate body bounding boxes for all N people if $N \leq 5$ and for at least five people otherwise. This can help us use such human bounding boxes as negative pairs with other hand instances. We also instructed the annotators to ensure that each body has at most two hands associated with it, and also each hand is associated with precisely one body. Thus, every hand instance in our dataset has a body associated with it. When hands are the only visible regions of the person, we use the hand bounding box as the human bounding box. There are some human bounding boxes with no associated hands; this is when hands are occluded or not visible. We collected annotations in batches and manually verified the annotation results ourselves.

Figure 3. Representative images from the BodyHands dataset. Hands and bodies belonging to the same person have bounding boxes in the same color and identification numbers.

Statistics. The BodyHands dataset has 20,490 images with 57,898 annotated polygons for hands and 63,095 axis-parallel rectangular bounding boxes for people. There are 19,810 people with one annotated hand, 19,044 people with two annotated hands, and 24,241 annotated people with no annotated hands (because their hands were either occluded or too small). We use the same training and test splits as the ContactHands dataset to be backward compatible. Fig. 3 shows some representative images. We provide more statistics in the supplementary material.

5. Experiments

In this section, we describe two sets of experiments. The first experiments analyze the proposed method's hand-body association performance and benchmark it against several other baseline methods. The second set of experiments demonstrates the benefits of hand-body association for hand tracking and hand contact estimation.

5.1. Hand-Body Association Experiments

This subsection describes the evaluation metrics used and experimental results for the hand-body association task.

5.1.1 Evaluation Metrics

We measure the hand detection performance using the standard VOC Average Precision (AP) metric. To measure the hand-body association performance, we consider two metrics: **(1) Conditional Accuracy** for body association. We define this as the percentage of correctly associated bodies among the correctly detected hand instances. Here we define that a body is correctly associated with the hand if the Intersection over Union (IoU) between the associated body box and the corresponding ground truth body box is greater than 0.5. We call this conditional accuracy since we only consider associated bodies corresponding to the correctly detected hand instances. Note that a hand detection is correct if the IoU between the detected hand bounding box and a ground truth bounding box is greater than 0.5. **(2) Joint AP** for hand detection and body association. In this metric, a detected hand is considered a true positive if: (a) the Intersection over Union (IoU) between the bounding box of the detected hand and a ground truth hand bounding box is greater than 0.5; and (b) the Intersection over Union (IoU) between the body bounding box associated with the detected hand instance and the ground truth body bounding box is greater than 0.5.

5.1.2 Competing Methods and Comparison Results

We conducted several experiments to measure the hand-body association performance and compare them to the proposed method. Note that in the proposed method variant with an option to match the detected hand to itself, we allow the hand box to be its corresponding body box when running the Hungarian Algorithm. We summarize the results in Table 1. The proposed method outperforms other methods by a significant margin. We describe the methods in the comparison below.

2D Human Pose. We run different 2D pose estimation methods such as OpenPose [8, 50, 56], Keypoint Communities [61] and DOPE [57] to obtain hand keypoints and body joints. We obtain the hand bounding boxes and corresponding body bounding boxes using these keypoints and joints. We use a less-stricter evaluation protocol since the detected hand keypoints can be very noisy: we consider a hand to be a true positive if its bounding box has positive IoU with a ground-truth bounding box. These methods do not perform well since obtaining accurate hand and body pose in unconstrained conditions is challenging.

MaskRCNN + X. We train MaskRCNN using a ResNet101 backbone to detect hands and bodies. We then use the Hungarian matching algorithm to match hands to bodies using several cost functions: (1) **Feature Distance** first extracts MaskRCNN's box regression 1024-dimensional feature vectors for hands and bodies and then uses the L_2 distance between these feature vectors; (2) **Feature Similarity** first extracts MaskRCNN's box regression 1024-dimensional feature vectors for hands and bodies, and then uses the inner product between these feature vectors; (3) **Location Distance** uses the L_2 distance between the center of the detected hand and body bounding boxes; (4) **IoU** uses the Intersection over Union (IoU) overlap between the detected hand and body bounding boxes.

Figure 4. **Qualitative results and failure cases.** We visualize hands and bodies that belong to the same person using the same color and identification numbers.

Ablation Studies. We conduct ablation studies to study the effects of different components of the proposed method. Specifically, we train three different models using the training set of BodyHands: (1) the proposed method without the Overlap Estimation Module; (2) the proposed method without the Positional Density Module; and (3) the proposed method using overlap computed from hand and bounding boxes instead of Overlap Estimation Module. The Joint AP on the BodyHands test set of these methods are 59.03%, 50.29%, and 60.34 %, respectively. These results show that both the overlap estimation module and the positional density module are helpful for the hand-body association.

Qualitative Results. Fig. 4 shows some qualitative results and failure cases from our method. Failure cases are mainly due to incorrect hand detections and false body association, especially in crowded images.

5.2. Benefits of Hand-Body Association

The ability to associate each detected hand to a human body is beneficial for many downstream tasks. This subsection demonstrates the benefits of this ability for two such tasks: hand tracking and hand contact estimation.

Dataset	BodyHands			COCO-WholeBody [20]		
Method	Hand AP	Cond. Accuracy	Joint AP	Hand AP	Cond. Accuracy	Joint AP
DOPE	9.09	32.51	2.27	15.02	47.56	9.09
OpenPose	39.69	74.03	27.81	30.22	82.97	18.65
Keypoint Communities	33.62	71.48	20.71	44.39	87.91	40.89
MaskRCNN + Feature Distance	84.82	41.38	23.16	75.92	59.97	38.44
MaskRCNN + Feature Similarity	84.82	39.12	23.30	75.92	53.60	33.72
MaskRCNN + Locaction Distance	84.82	72.83	50.42	75.92	78.47	50.92
MaskRCNN + IoU	84.82	74.52	51.74	75.92	79.53	53.08
Proposed	84.82	83.44	63.48	75.92	88.05	62.87
Proposed (with hand self-association option)	84.82	84.12	**63.87**	75.92	88.69	**62.92**

Table 1. **Hand detection and hand-body association performance** of several methods evaluated on BodyHands and COCO-WholeBody.

5.2.1 Hand-Body Association for Hand Tracking

Hand tracking is essential with many applications, including gesture recognition and skill evaluation. We hypothesize that the ability to associate hands with human bodies can improve tracking results. Intuitively, by associating hands with human bodies and linking human bodies across frames, we can establish correspondence between detected instances of the same hand across different frames, reducing identity switches in tracking.

Proposed hand tracking method and other baselines. Tracking hands is a multi-object tracking (MOT) problem, and a popular approach to address this problem is tracking by detection. This approach consists of two main steps: (1) detecting hands in individual video frames and (2) linking the detected hands between frames to form hand tracks. We adopt this tracking by detection approach in this work. For detection, we use our network trained for hand detection and hand-body association. For linking, we use the Hungarian algorithm [34] to optimize for the best set of one-to-at-most-one correspondence between a set of detected hands in frame t and a set of previously established hand tracks up until frame $t-1$. The matching outcome by the Hungarian algorithm depends on the affinity/cost matrix that defines the compatibility/cost for matching a hand to a hand tracklet. A popular approach is to define the affinity based on the Intersection over the Union (IoU) value between two detected objects (i.e., hands in this case). We will refer to this as the **Hand-IoU** baseline. However, hands are fast-moving objects, and the location and size of a hand can change drastically from one frame to the next. Thus, linking hands using Hand-IoU leads to incorrect identity switches in many cases. We consider a simple approach for linking based on hand-body association that treats a hand-and-body pair as a single identity. We define their affinity for two hand-body pairs detected at two different frames based on the weighted sum of the hand IoU and the body IoU. We refer to this method as **Hand-&-Body-IoU**. We also consider several other linking methods as follows. In **Re-ID**, the matching cost between two detected hands is defined based on the distance between the corresponding embedding vectors. In **Pose-based**, we use LightTrack [37] to detect and track skeleton keypoints and associate each detected hand instance to a skeleton based on the distances between the predicted wrist keypoint and center of the detected hand bounding box. **Flow-based** is the method that uses optical flows to link detections. Here, we use the average optical flows for pixels inside the detected object to link it with detection in the previous frame.

Evaluation dataset. There were no publicly available datasets for tracking hands in unconstrained environments. Most of the existing datasets [11, 31, 38, 40, 41] for hand tracking was captured in constrained environments such as ego-centric perspectives and contained only one or two hands. To evaluate hand tracking methods in unconstrained conditions, we collected 20 videos from YouTube and manually annotated hand bounding boxes and their trajectories. Specifically, we annotated every 15 frames, and altogether the dataset has 3299 annotated frames, 8893 hand instances, and 131 hand trajectories. We call this dataset YoutubeHands-20, and this dataset has many videos that contain multiple people interacting in the scene, so tracking hands in such cases is challenging. YoutubeHands-20 has now been expanded to a larger dataset YoutubeHands containing 200 videos [19].

Evaluation metric. To evaluate hand tracking performance, we use the standard multi-object-tracking evaluation metrics [4, 29]: False Positives (FP), False Negatives (FN), Identity Switches (IDs), and Multiple Object Tracking Accuracy (MOTA). MOTA is the combined metric, and it is considered the most crucial metric to quantify the overall detection and tracking performance.

Tracking results. Table 2 compares the tracking results of all methods. CenterTrack [65] and FairMot [62] are end-to-end methods in which object detection and association are performed together. To the best of our knowledge, there is no publicly available large-scale hand tracking datasets to train these methods. We do our best to train these two methods: first, we use static images from TVHand [35] and COCOHand [35] datasets to pre-train these methods. We then use the VIVAHandTracking [41] dataset to fine-

	FP↓	FN↓	IDs↓	MOTA↑
FairMOT [62]	412	3859	114	8.6
CenterTrack [65]	376	3909	2	10.7
MPNTrack [6] (offline)	1192	1074	545	41.4
CenterTrack (our detection)	458	1553	750	42.5
Re-ID	681	1284	817	42.0
Hand-IoU	681	1284	624	46.1
Flow-based	681	1284	882	40.7
Pose-based	681	1284	591	46.8
Hand-&-Body-IoU (proposed)	681	1284	436	**50.0**

Table 2. **Hand tracking results.**

tune them to perform hand tracking. We also conducted experiments by replacing the detection component of CenterTrack with the proposed hand detector. MPNTrack [6] is an offline tracking method. We pre-train MPNTrack on the VIVA [41] dataset and then use hands detected from our method as inputs to the tracker. These methods do not work well on hands, perhaps because they are geared towards less deformable classes such as pedestrians and vehicles.

The methods Re-ID, Hand-IoU, Flow-based, Pose-based, and Hand-&-Body-IoU use the same hand detector, so they have the same FP and FN. The main differences are how we link the detected hands into tracks. As seen, using both hands and bodies for linking yields the highest MOTA.

5.2.2 Hand-Body Assoc. for Physical Contact Analysis

We now demonstrate the benefits of hand-body association for recognizing the physical contact state of a hand, which could be: (1) No-Contact, (2) Self-Contact, (3) Person-Contact, and (4) Object-Contact. These conditions are not mutually exclusive, and a hand can be in more than one state. Recognizing the physical contact states of hands has many applications in human understanding, augmented reality, and virtual reality.

Contact state recognition is a complex problem in general, and the most challenging category to recognize is Person-Contact, with the current state-of-the-art result being 39.51% Average Precision (AP) [36]. This is due to the difficulty of distinguishing between Person-Contact and Self-Contact. The visual appearance of a hand and its surrounding local context can determine if the hand is touching a body part. However, it is not easy to know if this body part is part of the same person (Self-Contact) or a different person (Person-Contact). Next, we will describe two approaches to improve the performance of Person-Contact recognition by reasoning about the hand-body association.

Heuristic method. We consider a simple post-processing heuristic to improve the performance of an off-the-shelf contact estimation network [36] as follows. Given a detected hand H and its person-contact score s obtained by running the pre-trained hand-contact network of [36], our simple heuristic method will adjust s while leaving the scores of other contact states unchanged. We provide more details about this in the supplementary material. The heuristic improves the AP for detecting other person-contact from 39.51% to 40.89%. This heuristic is simple, but is only possible because we have a network that tells us who is the self person among the set of detected people. Next, we will describe an end-to-end trainable network that jointly performs contact state estimation and body association.

	NC	SC	OC	PC	mAP
Previous SoTA [36]	62.48	54.31	73.34	39.51	57.41
Leveraging hand-body association					
Heuristic	62.48	54.31	73.34	40.89	57.56
End-to-end	64.74	56.12	74.32	47.09	60.56

Table 3. **Hand contact estimation results.** The states NC, SC, PC, OC, denotes No-Contact, Self-Contact, Person-Contact, and Object-Contact, respectively. We can advance the state-of-the-art by leveraging the ability to associate detected hands to bodies.

End-to-end method. We build a new architecture that extends the proposed method in Sec. 3 with an additional branch to estimate the contact state of a detected hand. The inputs to this new branch are the RoI feature maps of the detected hand and the corresponding body. We concatenate the RoI features and use fully-connected layers to obtain the contact state scores for the hand. We train this new architecture end-to-end using the following multi-task loss:$\mathcal{L} := \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} + \mathcal{L}_{association} + \mathcal{L}_{contact}$. The losses $\mathcal{L}_{cls}, \mathcal{L}_{box}, \mathcal{L}_{mask}, \mathcal{L}_{association}$ are the same as described in Eq. (9). The term $\mathcal{L}_{contact}$ is the loss for contact state of the hand. Following [36], we define $\mathcal{L}_{contact}$ to be the sum of four independent binary cross-entropy losses corresponding to four possible contact states. We train this architecture on the training set of ContactHands [35] and evaluate its performance on the test set of ContactHands. This method improves the AP for Person-contact from 39.51% to 47.09%. We summarize the results in Table 3.

6. Conclusions, limitation, and societal impact

We investigated a new problem of detecting hands and associating them with their corresponding bodies. We introduced a novel architecture based on MaskRCNN, and we also contributed a large-scale dataset of images annotated with hand locations and corresponding body locations. Finally, we demonstrated the benefits of this new problem in two tasks, hand tracking and hand contact estimation.

The potential negative social impacts of the work are similar to most action and activity recognition applications, where privacy could be a concern. Our code will be available for research usage. However, one must be cautious when using it to support the making of accusations or decisions since our method still makes many mistakes.

Acknowledgements. This work was partially supported by DARPA PTG HR00112220001 award. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[3] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the International Conference on Computer Vision*, 2015. 2

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 7

[5] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *The European Conference on Computer Vision (ECCV)*, August 2020. 1, 2

[6] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020. 8

[7] Patrick Buehler, Mark Everingham, Daniel P Huttenlocher, and Andrew Zisserman. Long term arm and hand tracking for continuous sign language tv broadcasts. In *Proceedings of the British Machine Vision Conference*, 2008. 1, 2

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[9] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[10] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020. 7

[12] Xiaoming Deng, Yinda Zhang, Shuo Yang, Ping Tan, Liang Chang, Ye Yuan, and Hongan Wang. Joint hand detection and rotation estimation using cnn. *IEEE Transactions on Image Processing*, 2018. 2

[13] Eng-Jon Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2004. 1, 2

[14] Mark Everingham, Luc Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2009. 5

[15] Ross Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision*, 2015. 3, 4

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3

[17] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision*, 2017. 2, 3, 4

[19] Mingzhen Huang, Supreeth Narasimhaswamy, Saif Vazir, Haibin Ling, and Minh Hoai. Forward propagation, backward regression, and pose association for hand tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 7

[20] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 7

[21] Leonid Karlinsky, Michael Dinerstein, Daniel Harari, and Shimon Ullman. The chains model for detecting parts by their context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2

[22] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[23] Mathias Kolsch and Matthew Turk. Robust hand detection. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2004. 1, 2

[24] Pavan. M. Kumar, Andrew Zisserman, and Philip. H. S. Torr. Efficient discriminative learning of parts-based models. In *Proceedings of the International Conference on Computer Vision*, 2009. 1, 2

[25] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[26] Stefan Lee, Sven Bambach, David J. Crandall, John M. Franchak, and Chen Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014. 2, 5

[28] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[29] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv arXiv:1603.00831*, 2016. 7

[30] Arpit Mittal, Andrew Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. In *Proceedings of the British Machine Vision Conference*, 2011. 1, 2, 5

[31] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1284–1293, 2017. 7

[32] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[33] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[34] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32–38, March 1957. 3, 5, 7

[35] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the International Conference on Computer Vision*, 2019. 2, 5, 7, 8

[36] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 5, 8

[37] Guanghan Ning and Heng Huang. Lighttrack: A generic framework for online top-down human pose tracking. *Proceedings of CVPRW 2020 on Towards Human-Centric Image/Video Synthesis and the 4th Look Into Person (LIP) Challenge*, 2020. 7

[38] Tomas Pfister, James Charles, Mark Everingham, and Andrew Zisserman. Automatic and efficient long term arm and hand tracking for continuous sign language tv broadcasts. In *British Machine Vision Conference*, 2012. 7

[39] Pramod Kumar Pisharady, Prahlad Vadakkepat, and Ai Poh Loh. Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 2013. 1, 2

[40] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 7

[41] Akshay Rangesh, Eshed Ohn-Bar, Mohan M Trivedi, et al. Driver hand localization and grasp analysis: A vision-based real-time approach. In *IEEE International Conference on Intelligent Transportation Systems*, 2016. 7, 8

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2015. 4

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 3

[44] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, 2017. 1, 2

[45] Kankana Roy, Aparna Mohanty, and Rajiv Ranjan Sahay. Deep learning based hand detection in cluttered environment using skin segmentation. In *Proceedings of ICCV Workshops*, 2017. 2

[46] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[47] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[48] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Eyal

Krupka, Andrew Fitzgibbon, Shahram Izadi, and Pushmeet Kohli. Accurate, robust, and flexible real-time hand tracking. In *ACM Conference on Human Factors in Computing Systems*, 2015. 1, 2

[49] Roy Shilkrot, Supreeth Narasimhaswamy, Saif Vazir, and Minh Hoai. WorkingHands: A hand-tool assembly dataset for image segmentation and activity mining. In *Proceedings of the British Machine Vision Conference*, 2019. 1

[50] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[51] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[52] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2

[53] Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[54] Satoshi Tsutsui, Yanwei Fu, and David J. Crandall. Whose hand is this? person identification from egocentric hand gestures. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2

[55] Robert Y. Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics*, 28(3), 2009. 1, 2

[56] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

[57] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *Proceedings of European Conference on Computer Vision*, 2020. 6

[58] Ying Wu, Qiong Liu, and Thomas S. Huang. An adaptive self-organizing color segmentation algorithm with application to robust real-time human hand localization. In *Proceedings of the Asian Conference on Computer Vision*, 2000. 1, 2

[59] Xiaojin Zhu, Jie Yang, and A. Waibel. Segmenting hands of arbitrary color. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2000. 1, 2

[60] Linlin Yang, Shicheng Chen, and Angela Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[61] Duncan Zauss, Sven Kreiss, and Alexandre Alahi. Keypoint communities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[62] Yifu Zhan, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*, 2020. 7, 8

[63] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 1, 2

[64] Huayi Zhou, Fei Jiang, and Ruimin Shen. Who are raising their hands? hand-raiser seeking based on object detection and pose estimation. In *Proceedings of The 10th Asian Conference on Machine Learning*, 2018. 2

[65] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020. 7, 8

[66] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the International Conference on Computer Vision*, 2017. 1, 2

[67] 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 *Proceedings of the International Conference on Computer Vision*, 2019. 1, 2

Supplementary Material

Whose Hands are These? Hand Detection and Hand-Body Association in the Wild

Supreeth Narasimhaswamy[1], Thanh Nguyen[2], Mingzhen Huang[3], Minh Hoai[1,2]
[1]Stony Brook University, [2]VinAI Research [3], University of Buffalo

1. BodyHands Dataset

BodyHands is a new dataset collected to develop and evaluate hand-body association methods. It is a large-scale dataset containing unconstrained images with annotations for hand and body locations and correspondences. Table 1 provides some statistics for the proposed BodyHands dataset.

				# people with annotation for		
	#images	#hands	#people	2 hands	1 hand	0 hand
Train	18861	51915	56047	17402	17111	21534
Test	1629	5983	7048	1642	2699	2707
All	20490	57898	63095	19044	19810	24241

Table 1. **Statistics of the BodyHands dataset**

2. Hand Tracking Evaluation Dataset

To evaluate hand tracking methods in unconstrained conditions, we collect 20 videos from YouTube and manually annotate hand bounding boxes and their trajectories. Specifically, we annotate every 15 frames, and altogether the dataset has 3299 annotated frames, 8,893 hand instances, and 131 hand trajectories. We call this dataset YoutubeHands-20, and this dataset has many videos that contain multiple people interacting in the scene, so tracking hands in such cases is challenging. Note that YoutubeHands-20 has now been expanded by Huang et al. [2] to a larger dataset YoutubeHands containing 200 videos. Fig. 1 shows some representative images from this dataset.

3. Implementation Details

We use Detectron2 [4] to implement the proposed architecture. We set the loss weights λ_1 for the Overlap Estimation Module and λ_2 for the Positional Density Module to be 0.1. We train the network using SGD with an initial learning rate 0.0001 for 20 epochs. We reduce the learning rate by a factor of 10 at 10^{th} and 15^{th} epochs. We train our network on NVIDIA RTX 2080 using a batch size of one.

When conducting the ablation studies for the proposed model, we set the probabilities corresponding to the removed components to be 1, and we do not include the option to match the hand to itself when running the Hungarian Algorithm.

4. Heuristic Method: Hand-Body Association for Hand-Contact Estimation

We consider a simple post-processing heuristic to improve the performance of an off-the-shelf contact estimation network [3] as follows. Given a detected hand H and its corresponding person-contact score s obtained by running the pre-trained hand-contact network of [3], our simple heuristic method will adjust s while leaving the scores of other contact states unchanged. First, we use the hand-body association network developed in this paper to detect hands and obtain the associated human bodies for each detected hand; let $\{(A_i, B_i)\}$ denote the set of hand-body pairs obtained. If H does not overlap with any A_j, we will terminate this process and leave the person-contact score s unchanged. Otherwise, we will associate H with A_j that has the highest IoU with H and subsequently associate H to the body B_j. Second, we use a pre-trained MaskRCNN [1] to detect all people in the image; let $\mathcal{P}$ denote this set. We then associate the body B_j with the person $P_k \in \mathcal{P}$ with the highest IoU with B_j. Third, we consider all detected people in $\mathcal{P}$ different from P_k and determine the overlapping region between them and the hand H. If none of the overlapping regions is larger than 15% of the hand area, we heuristically determine that this hand has a low probability of contact with another person. We then decrease the person-contact score using the formula: $s^{new} = \max(s - 0.5, 0)$. This heuristic improves the average precision for detecting other person-contact from 39.51% to 40.89%. This heuristic is simple, but it is only possible because we have a network that tells us who is the self person among the set of detected people. Note that this heuristic only adjusts the person-contact score.

Sample frames from HandTracking Evaluation Dataset

Figure 1. **Hand tracking evaluation dataset.** Sample frames from the videos used for evaluating hand tracking performance. Each row contains frames from the same video.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision*, 2017. 1

[2] Mingzhen Huang, Supreeth Narasimhaswamy, Saif Vazir, Haibin Ling, and Minh Hoai. Forward propagation, backward regression, and pose association for hand tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[3] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020. 1

[4] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1