Ego2HandsPose: A Dataset for Egocentric Two-hand 3D Global Pose Estimation

Fanqing Lin
Magic Leap, Inc.
7500 W Sunrise Blvd, Plantation, Florida U.S.A
flin@magicleap.com

Tony Martinez
Brigham Young University
Brigham Young University, Provo, UT
martinez@cs.byu.edu

Abstract

Color-based two-hand 3D pose estimation in the global coordinate system is essential in many applications. However, there are very few datasets dedicated to this task and no existing dataset supports estimation in a non-laboratory environment. This is largely attributed to the sophisticated data collection process required for 3D hand pose annotations, which also leads to difficulty in obtaining instances with the level of visual diversity needed for estimation in the wild. Progressing towards this goal, a large-scale dataset Ego2Hands was recently proposed to address the task of two-hand segmentation and detection in the wild. The proposed composition-based data generation technique can create two-hand instances with quality, quantity and diversity that generalize well to unseen domains. In this work, we present Ego2HandsPose, an extension of Ego2Hands that contains 3D hand pose annotation and is the first dataset that enables color-based two-hand 3D tracking in unseen domains. To this end, we develop a set of parametric fitting algorithms to enable 1) 3D hand pose annotation using a single image, 2) automatic conversion from 2D to 3D hand poses and 3) accurate two-hand tracking with temporal consistency. We provide incremental quantitative analysis on the multi-stage pipeline and show that training on our dataset achieves state-of-the-art results that significantly outperforms other datasets for the task of egocentric two-hand global 3D pose estimation.

1. Introduction

Hand pose estimation and tracking is significant for many applications that involve Human-Computer Interaction (HCI), gesture recognition and sign language recognition. Particularly, as VR/AR/MR applications gain rapid development with the trending of metaverse, two-hand tracking is becoming a fundamental feature for an immersive user experience. In addition, due to the overhead cost of multi-camera setups as well as depth cameras, there is motivation to approach this task using a single ubiquitous RGB camera. However, there is limited attention from the community on color-based two-hand application, which is an extremely challenging task requiring steps not needed in single-hand tasks that use cropped images as input: two-hand detection/segmentation, robustness against inter-hand occlusion as well as 3D global hand pose estimation with accurate absolute 3D joint positions.

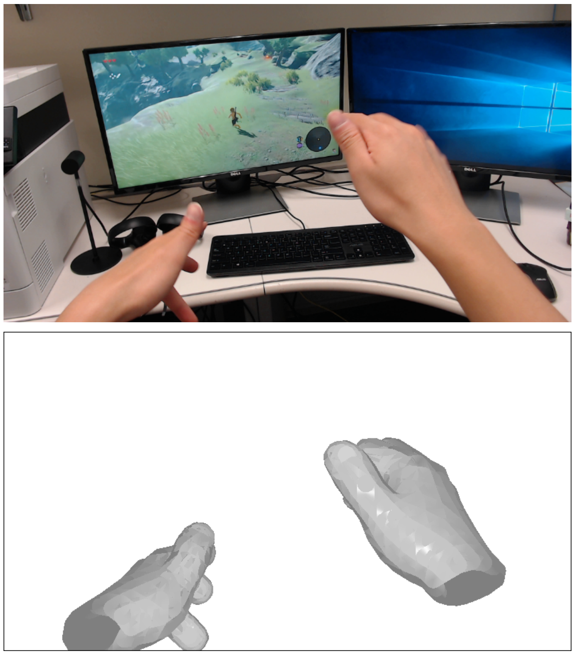


Figure 1: Our method enables 3D hand pose annotation using a single image. We show a sample image from the test set of Ego2Hands (top) and visualization of its corresponding two-hand pose annotation (bottom).

Existing color-based two-hand pose datasets [12, 3] are captured in third-person viewpoints with multiple cameras and laboratory backgrounds. Two-hand datasets [20, 18] with RGB-D data can remove background using depth thresholding or a green screen. However, these data have limited accuracy from depth-based tracking and lack visual diversity due to the small number of participants. As a result, methods trained on existing datasets cannot generalize to the real-world domain. In addition to limited diversity, although data for third-person viewpoint has many applications, it constrains the users to perform gestures in front of a fixed camera. On the other hand, hand tracking in egocentric viewpoint has no such constraint and has increasing demand in VR/AR applications. To the best of our knowledge, there is currently no real-world RGB dataset for egocentric two-hand tracking in the wild.

To enable two-hand applications using a single RGB camera in a domain invariant setting, [10] first proposed a large-scale two-hand segmentation/detection dataset named Ego2Hands. Unlike traditional hand segmentation datasets [2, 19, 9] with limited quantity/diversity and generalization ability, Ego2Hands composites two-hand instances at training time with excellent generalization to unseen environments. However, Ego2Hands does not contain hand pose annotation. In this work, we develop a parametric fitting algorithm ManoFit that can fit the deformable MANO hand model [14] given an arbitrary number of 2D joint locations and minimal manual guidance towards the optimal solution. This is the first tool that enables users to annotate 3D hand poses using a single image, which significantly simplifies the annotation process compared to previous methods (see Section 2) and provides open access for the community to generate additional hand pose data. Using our method, we create a new dataset by manually annotating$\sim 7,000$ selected frames with diversity from the training set and$\sim 2,000$ frames from the test set of Ego2Hands (see Figure 1). We introduce Ego2HandsPose, the first dataset that enables egocentric two-hand 3D global pose estimation in the wild using a single camera.

In addition to Ego2handsPose, as there are existing datasets with only 2D hand pose annotations, we apply ManoFit to automatically convert HIU-DMTL [23], PanHand2D [15] and OneHand10k [21] to 3D hand pose datasets. Manual validation is performed on all generated instances to remove dirty data. For quantitative analysis on the accuracy of our generated hand poses, we evaluate ManoFit on FreiHAND [25] which contains both 2D and 3D hand pose annotation and show convincing results.

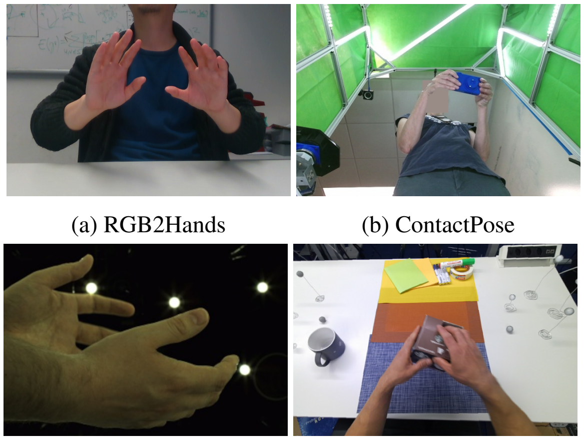
To validate Ego2HandsPose for the task of two-hand 3D tracking, we follow [11] and use a multi-stage pipeline to estimate the segmentation/detection and 2D/3D canonical hand pose for both hands. As the 3D canonical hand pose estimation network uses input in the form of heatmaps, there is no need for the training to be constrained by 3D hand poses of actual images and we take a novel approach by training on generated poses. We introduce this synthetic dataset as MANO3Dhands, which generates 3D hand poses based on our collected real-world pose distribution. Cross-dataset evaluation shows that MANO3Dhands has the best generalization score compared to existing 3D hand pose datasets. For the last stage of the pipeline, we modify ManoFit to achieve temporally consistent two-hand 3D tracking by using the estimated 2D and 3D canonical joint locations. Quantitative analysis shows that training on Ego2HandsPose achieves over 30% improvement compared to the second top dataset for the task of two-hand 2D, 3D canonical and 3D global pose estimation.

2. Related Work

In this section, we introduce relevant RGB-based 3D hand pose datasets for single-hand and two-hand scenarios.

Single-hand 3D pose datasets. Obtaining 3D hand pose annotation on real-world images is a challenging problem that commonly requires extensive manual annotation using multi-view setups or RGB-D data to resolve the depth ambiguity and self-occlusion. Early work [22] proposed the *Stereo Tracking Benchmark* (*STB*) dataset that includes a single participant performing simple gestures with 6 backgrounds. To generate data with quantity and diversity, pioneering work [24] introduced the *Rendered Hand Pose Dataset* (*RHD*) using 20 rendered characters performing 39 static gestures. Using the CMU Panoptic Studio with 10 RGB-D sensors, 480 VGA and 31 HD cameras, [5] proposed the *Panoptic Hand* (*PanHand3D*) dataset labeled using multiview Bootstrapping [15]. To achieve better cross-dataset generalization, [25] proposed *FreiHAND* that was captured using 8 calibrated & synchronized cameras in a green screen setting, which allowed for additional diversity from background replacement.

Two-hand 3D pose datasets. Inter-hand occlusion and interaction can introduce additional challenges for 3D hand pose annotation. To address two-hand pose estimation with object handling, [18] proposed the *Tzionas Dataset* collected using RGB-D data and a combination of a generative linear blend skinning model [8] and a discriminative model trained on manually annotated finger tips. To provide more data with two-hand interaction, [20] proposed *RGB2Hands* that was captured using a RGB-D sensor and labeled with a depth-based two hand tracker [13]. However, the obtained annotation can be erroneous and synthetic data was used to complement this issue. Focusing on two-hand object grasping, [3] proposed *ContactPose* that was captured using 7 calibrated RGB cameras, 3 RGB-D cameras and one thermal camera for contact capture. 3D hand Pose annotation was obtained using estimated 2D keypoints as well as extracted object pose and contact locations. Taking a different direction, [12] focused on two-hand close interaction without objects and proposed *InterHand2.6M*. The data was collected in a multi-view studio consisting of 80-140 cameras and annotated in a two-stage pipeline. The first stage consists of extensive manual annotation of 2D hand poses followed by 3D triangulation. The second stage utilizes an automatic 2D annotator trained from data in the first stage and triangulation for 3D keypoints. Recognizing the importance


(a) RGB2Hands

(b) ContactPose

(c) InterHand2.6M

(d) H2O

Figure 2: Samples from existing two-hand pose datasets.

of egocentric data, *H2O* [7] was captured with 5 RGB-D cameras with 1 mounted on the helmet in 3 environments. Two-hand poses with object manipulation annotation was obtained by fitting the MANO model to multi-view depth-data and estimated 2D poses from OpenPose [4].

3. Ego2HandsPose

There are two major motivations for the introduction of Ego2HandsPose. First, Figure 2 shows that existing datasets for RGB-based two-hand pose estimation all contain images captured in laboratory environments with limited diversity. Consequently, models trained on these datasets cannot generalize to other unseen environments or be applied to practical applications. Second, there is limited RGB data that address two-hand pose estimation in the egocentric viewpoint, which does not constrain the user to be in front of the fixed camera and is essential in applications such as VR/AR/MR. As the only existing RGB-based egocentric two-hand pose dataset, **H2O** has a heavy emphasis on manipulation of 8 objects in 3 scenes with 4 participants, which results in limited pose space and visual diversity for the environment and the target hands.

For two-hand 3D tracking, segmentation and detection of both hands are commonly required prior to pose estimation [17, 13, 11]. Recently, [10] introduced Ego2Hands for the task of two-hand segmentation and detection in the wild. It consists of a training set captured in a green screen setting with$\sim$ $180k$ right hand instances from 22 participants and composites two-hand images at training time by horizontally flipping one right hand to create the left hand. This approach circumvents the issue of data scarcity for two-hand segmentation/detection and allows for sufficient diversity necessary for estimation in the wild. For evaluation, Ego2Hands provides a test set consisting of 8 sequences collected with diverse scenes, lighting and skin tones. Despite enabling models to achieve a promising level of generalization on two-hand segmentation and detection, Ego2Hands does not provide any hand pose annotation.

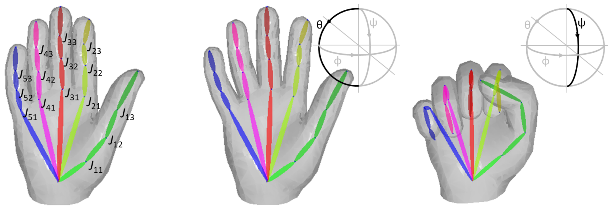


Figure 3: MANO hand representation for joints and axis angles. The first image shows a default hand pose. The second and third image show poses with joints rotated in the two primary directions.

Existing 3D hand pose annotation methods require either RGB-D or calibrated multi-view RGB setups for data collection. Extensive multi-view manual annotation on 2D keypoints as well as triangulation are commonly required to subsequently extract the 3D hand poses. We argue that the resources required by existing annotation methods are not commonly available in the community, which significantly limits the quantity and diversity of hand pose data available in general and consequently impacts hand pose related research potential. To address this issue, we introduce an annotation tool that utilizes a parametric fitting algorithm *ManoFit* with manual guidance to enable 3D hand pose annotation using a single RGB image.

3.1. Supervised ManoFit

The differentiable MANO hand model proposed by [14] is widely used for hand pose estimation and fitting. It is parameterized by $P \in \mathcal{R}^{61}$ where $P = (\alpha, \beta, \gamma)$. $\alpha \in \mathcal{R}^{10}$, $\beta \in \mathcal{R}^{45}$ and $\gamma \in \mathcal{R}^{6}$ represent the shape, articulation and global translation & orientation respectively. Although loss minimization can be applied to directly optimize P given the target 2D and 3D keypoints (obtained from manual annotation in multi-view setups) as in [25], 3D keypoints are not initially available in monocular RGB data. We find that loss computed using the target 2D keypoints alone oftentimes cannot reach a global minimum P_{gt} from a default $P_0 = \mathbf{0}$ using gradient descent. However, with a proper $P_0 = P_{gt} + \epsilon$ where $\epsilon \in \mathcal{R}^{61}$ represents an arbitrary error that is insufficient to deviate the gradient descent to an incorrect local minimum, we can successfully fit the hand model using 2D keypoints from a single RGB image.

To allow effective manual modification of P, we first apply physical constraints to the pose space defined by $\beta \in \mathcal{R}^{45}$, which consists of 3 rotational values (θ, ϕ, ψ) for each of the 15 finger joints and there are 3 joints J_{ij} defined for each finger i where $i = [1, 5]$ and $j = [1, 3]$. Note that each joint of the original MANO model has 3 Degrees of Freedom (DoF) with unlimited range. This is obviously not realistic as all finger joints J_{ij} primarily rotate in the direction of ψ with J_{i1} being able to rotate in the direction of θ as well (Figure 3). Additionally, there is physical limitations for the range of rotation of each joint. To enforce physical plausibility, we define constant vectors β_{min} and β_{max} that clip β within a reasonable range as follows,

$$\beta_c = min(max(\beta,\ \beta_{min}),\ \beta_{max}). \tag{1}$$

To encourage more natural poses, we define $\beta_{mean} = (\beta_{min} + \beta_{max})/2$ and the following regularization loss,

$$\mathcal{L}_{reg} = (\beta_c - \beta_{mean})^2\ \omega \tag{2}$$

where $\omega \in \mathcal{R}^{45}$ applies element-wise scaling to the squared difference between the current pose and the mean pose. In general, $\mathcal{L}_{reg}$ punishes gradients for valid but less common rotations in the direction of θ for J_{i0}.

We formulate the ManoFit algorithm for manual annotation as the minimization problem below,

$$\mathcal{L}_{2d} = \sum_{k \in \mathcal{A}} (\boldsymbol{q}_k - \Pi(\tilde{\boldsymbol{p}}_k))^2 \tag{3}$$

$$\mathcal{L}_{fit} = \mathcal{L}_{2d} + \mathcal{L}_{reg} \tag{4}$$

where $\mathcal{L}_{2d}$ represents the Sum Squared Error (SSE) computed using the user-provided 2D keypoints $\boldsymbol{q} \in \mathcal{R}^2$ and the 2D projection Π from 3D MANO keypoints $\tilde{\boldsymbol{p}} \in \mathcal{R}^3$ over the set of annotated joint indices $\mathcal{A}$. The combined loss $\mathcal{L}_{fit}$ aims to minimize the 2D keypoint error with valid and natural poses. Backpropagation using the Adam optimizer is applied to update β and γ with a stopping criteria that terminates when the loss ceases to improve for 10 iterations. Note that we use the default shape parameter $\alpha = \mathbf{0}$. Although subject-specific α can improve fitting accuracy, shape-fitting requires multi-view data and our goal is to introduce a universal tool for hand pose annotation using monocular RGB. Qualitative and quantitative results in Section 4.3 show that our fitting algorithm achieves excellent accuracy without subject-specific shape finetuning.

For the annotation of each instance, the user is instructed to 1) modify γ to the approximate values based on visualization of the MANO hand rendering, 2) annotate the wrist joint as well as the 5 finger tip joints, and 3) initiate parametric optimization. Until accurate fitting is achieved, the annotator can repeat the aforementioned steps with the freedom to modify (γ, β) and annotate additional 2D joint locations. We demonstrate in the supplementary video that our annotation tool is well-designed for efficient control and fitting of the MANO hand model. Additional details are provided in the supplementary document.

3.2. Annotated Data

Training data. Although the training set of Ego2Hands consists of$\sim 180k$ frames, since some frames do not contain valid poses and some others have similar poses, we selected $7,033$ frames with diverse articulation from the training set of Ego2hands for manual annotation. For the training of 2D hand pose estimation model, we first follow [10] and composite images at training-time. As illustrated in Figure 4, for each composited image, we randomly select the primary right hand from our annotated Ego2HandsPose training set, which contains the 3D hand pose annotation. For the secondary left hand, we randomly select a horizontally flipped right hand from the complete training set of Ego2Hands, which does not need the pose annotation as its purpose is to merely create a two-hand appearance. The background image is randomly selected from the proposed background set [10]. For data augmentation, we apply random horizontal/vertical translation, color and smoothness augmentation. Quantitative evaluation in Section 6.1 shows that our composited data with pose annotation achieves state-of-the-art results on 2D hand pose estimation for our task.

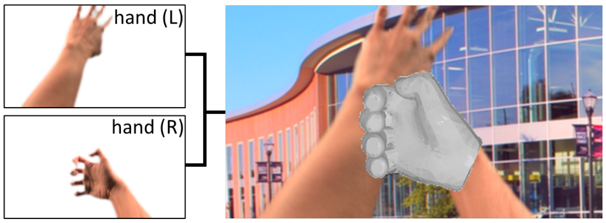


Figure 4: Illustration of two-hand image composition with visualized pose annotation of the primary right hand.

Evaluation data. As there is no evaluation benchmark for egocentric two-hand 3D global pose estimation in the wild, we manually annotate the complete test set of Ego2Hands, which consists of 8 diverse sequences with a total of $2,000$ two-hand images. In Section 6, we provide quantitative evaluation for two-hand 2D, 3D canonical and global hand pose estimation on this test set.

4. 2D Hand Pose Dataset Conversion

Using our annotation tool, we established in Section 3.1 that manual guidance is needed to generate a well-initialized P_0 for parametric fitting on 2D keypoints from a single camera. In addition, 2D keypoints also need to be annotated but it is oftentimes unnecessary to annotate the complete set of 21 joints. In existing 2D hand pose datasets, since instances contain full annotation for the 2D keypoints, we can theoretically train a network to estimate the corresponding 3D canonical hand poses [24] and apply parametric fitting to generate the 3D hand pose annotations for 2D datasets. For the training of this network, as existing fixed-sized 3D hand pose datasets contain pose space with limited size that do not necessarily cover the true data distribution, we create a synthetic dataset MANO3DHands that provides the largest and most diverse pose space sufficient for accurate parametric fitting of any generic 2D hand pose dataset.

4.1. MANO3DHands

Since 3D canonical pose estimation models use heatmaps as input, there is no visual domain gap between heatmaps generated from synthetic and real-world data. However, synthetic hands with unrealistic articulation can lead to a domain gap in the pose space and a naively randomly sampled β_i within the constrained space can still have an unrealistic combination of rotations.

To sample realistic β and γ, we first obtain two real-world data distributions of 3D hand poses from 5 participants using the LeapMotion device [1], which can automatically generate 3D keypoints from egocentric stereo infrared data. For the first distribution, we focus on $\gamma \in \mathcal{R}^6$ in the egocentric view and collect $31,796$ poses that cover a wide range of wrist rotations $\gamma_1 \in \mathcal{R}^3$ and global locations $\gamma_2 \in \mathcal{R}^3$. For the second distribution, we focus on β and collect $42,496$ poses with diverse joint rotations. For both distributions with 3D keypoints, we apply parametric fitting to obtain the matching distributions $\mathcal{B}$ and $\mathcal{G}$ with MANO parameters for β and γ respectively.

We find that egocentric and third-person viewpoints have different data distributions for the global orientation of the hand. In the egocentric viewpoint, we generate MANO hand poses by randomly sampling (β, γ) from $\mathcal{B}$ and $\mathcal{G}$. In the third-person viewpoint, we sample γ from all possible global orientations in $[-\pi, \pi]$. To enable evaluation, we generate two test sets with $50,000$ poses for each viewpoint. Section 6.2 shows that training on MANO3DHands achieves the best generalization score compared to existing 3D hand pose datasets in cross-dataset evaluation.

4.2. Unsupervised ManoFit

With the complete set of annotated 2D keypoints $\boldsymbol{q_{gt}}$ as well as the estimated 3D canonical hand pose $\boldsymbol{p}^*$, we apply the multi-stage parametric fitting algorithm that progressively solves the following problems.

1. Global orientation fitting. We discovered in our experiments that the global orientation γ_1 has the greatest impact on the overall hand pose and should be properly optimized first using the following loss,

$$\mathcal{L}_{\gamma_1} = \lambda_{\gamma_1} \frac{1}{N} \|L \cdot \boldsymbol{p}^* - \tilde{\boldsymbol{p}}\|_2^2. \quad (5)$$

For the number of joints $N = 21$, we minimize the Mean Squared Error (MSE) between the 3D keypoints $\tilde{\boldsymbol{p}}$ from the MANO model and the estimated 3D canonical keypoints $\boldsymbol{p}^*$ scaled by the reference bone length L. We use $\lambda_{\gamma_1} = 1 \times 10^5$ as the scaling constant.

2. Pose articulation fitting. After global orientation alignment, we optimize γ_1 and β with the loss below,

$$\mathcal{L}_{\gamma_1 + \beta} = \mathcal{L}_{\gamma_1} + \mathcal{L}_{reg} \quad (6)$$

where we use a combined loss of Equation 5 and 2 with β_{mean} being the mean pose computed using the collected real-world pose distribution $\mathcal{B}$.

3. Global translation fitting. With β and γ_1 properly aligned, we optimize γ_2 using $\mathcal{L}_{\gamma_2} = \mathcal{L}_{2d}$ defined in Equation 3 with $\mathcal{A}$ being the complete set of joint indices.

4. Full pose fitting. After the previous steps, we should have a well-initialized $P_0 = P_{gt} + \epsilon$ for the final fitting. Therefore, we optimize β and γ using $\mathcal{L}_{fit} = \mathcal{L}_{2d} + \mathcal{L}_{reg}$.

We perform optimization for 100 and 300 iterations for stages 1-3 and 4 respectively. The learning rate is set to 1.0, 0.01, 1.0 and 0.01 for stage 1-4. This process can achieve accurate matching between the ground truth 2D keypoints $\boldsymbol{q_{gt}}$ and the projected 2D keypoints $\Pi(\tilde{\boldsymbol{p}})$ from the 3D keypoints of the MANO hand model.

4.3. Fitting Results & Analysis

We select the following 2D hand pose datasets for automatic conversion: HIU-DMTL [23] (41,539 instances), PanHand2D [15] (14,817 instances) and OneHand10k [21] (2,040 instances). Instances without full 2D annotation in OneHand10k are discarded. In addition, since hand side information is necessary for parametric fitting, we manually annotate hand side labels for HIU-DMTL and OneHand10k, which contain both left and right hand instances.

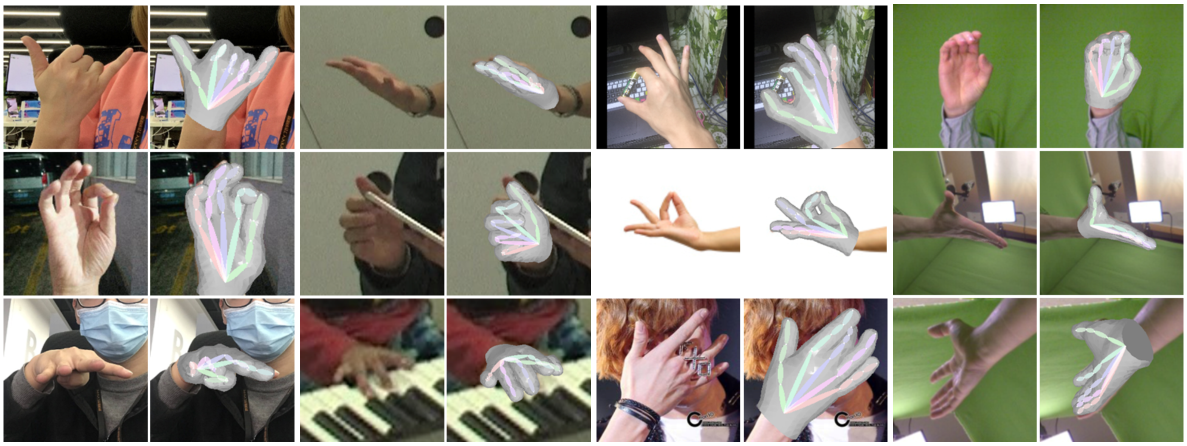
Qualitative examples in Figure 5 show that our algorithm can accurately fit a wide range of poses from different datasets. To quantitatively evaluate our fitting algorithm, we generate 3D hand poses for the training set of FreiHAND, which provides $32,560$ annotated instances with diverse 3D hand poses. We apply scaling using the reference bone length and Cartesian alignment using the ground truth absolute 3D location of the root joint. Our automatic fitting achieves an End Point Error (EPE) of 1.17cm.

Similar to our parametric fitting algorithm, [6] proposed to automatically fit the MANO hand model using 2D keypoints estimated using OpenPose [4] without using the estimated 3D canonical keypoints. We point out that this approach only selects samples that pass the designed heuristic verification and is limited to fit poses with P_0 less susceptible to local minima during gradient descent. In comparison, after manual multi-view validation for all generated instances using our algorithm, we report high acceptance rates of 84.6%, 88.1% and 79.1% for HIU-DMTL, PanHand2D and OneHand10k respectively. Note that a small rejection rate is expected due to the depth ambiguity in 2D keypoints.

5. Two-hand 3D Global Pose Estimation

For comprehensive analysis on our proposed dataset, we follow [11] and use a multi-stage pipeline for the task of two-hand 3D global pose estimation (Figure 6).

Two-hand segmentation and detection. For the first stage, We use the scene-adapted ICNet from [10] to estimate



(a) HIU-DMTL (b) PanHand2D (c) OneHand10k (d) FreiHAND

Figure 5: Qualitative results of our generated 3D poses on 4 datasets using only 2D keypoints from a single image.

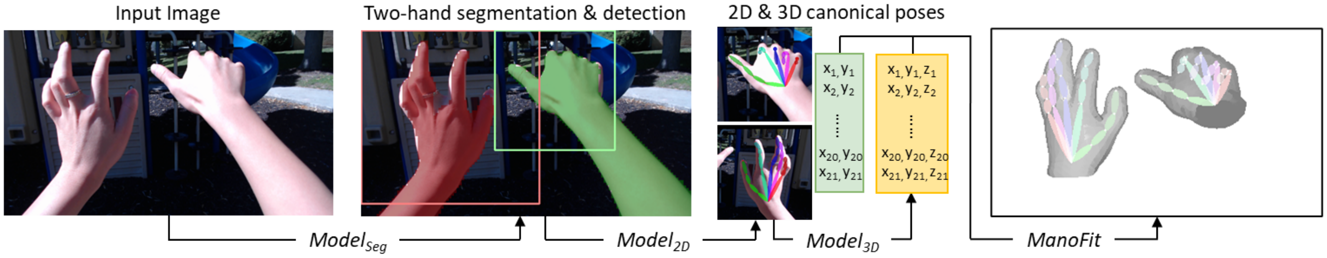


Figure 6: Overview of our two-hand 3D global pose estimation pipeline. Given an input image in the wild, we first extract hand bounding boxes and segmentation. 2D heatmaps are estimated using the cropped input and are used for 3D canonical pose estimation. MANO hand models in the global coordinate system are optimized using the 2D and 3D canonical poses.

the segmentation as well as the activation energy for both hands. Segmentation is used to address inter-hand occlusion and hand energy that excludes the arm is used for hand detection even when occluded. We apply a binary threshold using $\tau = 0.5$ to the estimated energy and perform a close operation with kernel size of 3 for noise removal.

2D hand pose estimation. Using the energy mask, we obtain the cropped images for 2D hand pose estimation. In addition to the three-channel RGB input, we concatenate the cropped binary segmentation mask for the other hand to encode occlusion information. We train HRNet-W32 [16] on the training set of Ego2HandsPose with a batch size of 16 for $40,000$ iterations. The Adam optimizer is used with an initial learning rate of 0.0001 (decreasing with a rate of 0.5 per $10,000$ iterations). Using input images resized to 224×224, the output heatmaps have a resolution of 56×56. To simplify the pose space, we horizontally flip the left hand so the network trains on the right hand only.

3D canonical hand pose estimation. We use a compact ResNet10 with 3 fully connected layers to regress the root-relative 3D joint locations. We generate training instances online using data distributions $\mathcal{B}$ and $\mathcal{G}$ from the proposed MANO3DHands and train for 400k iterations with a batch size of 1 and a learning rate that decreases with a rate of 0.5 per 100k iterations.

Hand tracking via Manofit. For each hand's reappearance, we first use the projection algorithm proposed by [11] to set the global translation parameter γ_2. The Manofit algorithm introduced in Section 4.2 is then applied for hand tracking. For continuous tracking, as subsequent frames contain poses with gradual changes, we do not reset the MANO parameters P or the internal state of the Adam optimizer prior to optimization, which conveniently leads to temporally smooth pose estimation. For each optimization stage, a properly selected threshold value is used on the corresponding loss as the stopping criteria.

6. Quantitative Benchmarking

In this section, we demonstrate that Ego2HandsPose enables state-of-the-art results on egocentric two-hand 3D global pose estimation in the wild. First, we evaluate on the test set of Ego2HandsPose and perform isolated studies on each stage of our proposed pipeline. Second, we evaluate using the complete pipeline on its ability to track both hands in the global coordinate system.

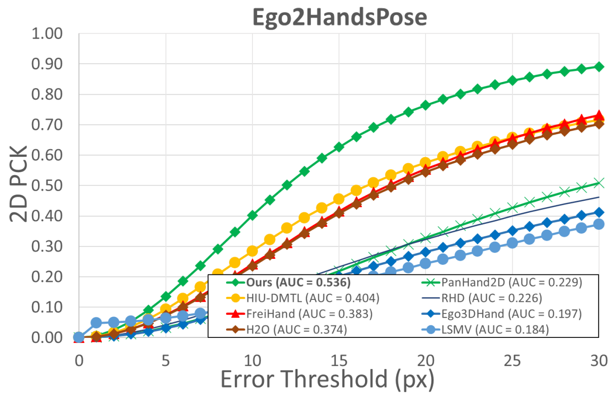


Figure 7: Quantitative comparison of 2D hand pose estimation on the proposed dataset.

6.1. 2D Hand Pose Estimation

We use the ground truth two-hand energy to obtain the cropped hand images as input and evaluate using HRNet-W32 trained on 8 large-scale datasets with data augmentation. Figure 7 shows that training on our dataset significantly outperforms others and enables the top $\text{AUC}_{2D} = 0.536$. We theorize the performance gap is caused by the difference in pose space (third-person for HIU-DMTL, FreiHAND) and the lack of diversity (H2O). Note that datasets with synthetic data or laboratory backgrounds have the lowest accuracy. Although it is expected for a network to perform well on the dataset it was trained on, we argue that since the test set of Ego2Hands contain subjects and scenes not present in its training set, the score obtained by Ego2HandsPose demonstrates its ability to enable accurate estimation on unseen data. Despite the fact that cross-dataset evaluation is commonly performed to show the generalization ability of a dataset, we do not perform this experiment in this stage as it is not our goal to generalize to third-person or synthetic visual data, but to achieve the best accuracy on the task of egocentric two-hand tracking.

6.2. 3D Canonical Hand Pose Estimation

In this stage, we use heatmaps as input and evaluate cross-dataset performance using 6 large-scale datasets with diverse 3D hand pose annotations, including our proposed MANO3DHands with egocentric and third-person data distributions. The complete datasets are used in training and evaluation for more comprehensive analysis. As a result, the non-diagonal scores should be emphasized to analyze the generalization ability, which correlates with the pose space quality each dataset provides. Table 1 shows that our proposed MANO3DHands datasets achieve the highest generalization scores with average rankings of 2.3. This finding indicates that the generated pose space using our collected distributions ($\mathcal{B}$ and $\mathcal{G}$) better represent the true pose space compared to other datasets. In addition, it is important to recognize the difference between egocentric and third-person pose space. We show that MANO3DHands_{3rd} and MANO3DHands_{ego} achieve the best generalization scores for third-person and egocentric datasets respectively. Therefore, we claim that it is best to train on hand pose data with the matching viewpoints for different applications. In this work, we use MANO3DHands_{3rd} for the annotation tool and MANO3DHands_{ego} for egocentric two-hand tracking.

Dataset	Ours	Ours*	H2O*	Frei	LSMV	Ego3D*
Ours	**0.735**	0.732	0.743	0.731	0.613	0.776
Ours*	0.611	**0.790**	**0.827**	0.643	0.522	0.790
H2O*	0.473	0.655	0.821	0.494	0.412	0.671
Frei	0.686	0.705	0.744	**0.738**	0.607	0.743
LSMV	0.528	0.482	0.490	0.541	**0.646**	0.493
Ego3D*	0.523	0.679	0.688	0.551	0.404	**0.834**
Rank	2.3	2.3	4.8	2.7	4.7	4.2

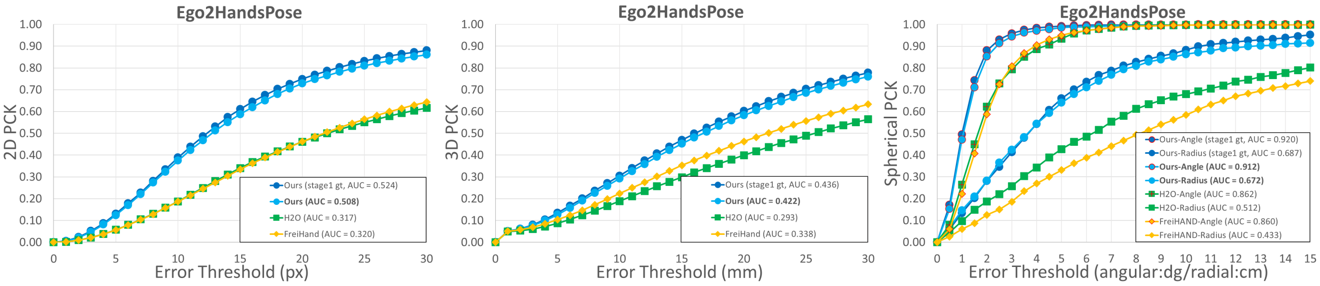
Table 1: Cross-dataset evaluation on 3D canonical pose estimation. We report AUC computed using PCK of root-relative 3D keypoints in an interval from 0.0 to 1.0. Egocentric datasets are labeled with *. The top 3 scores on each evaluation dataset (shown in columns) are marked as **first**, second and third.

Unlike traditional datasets with fixed sizes, the training set of MANO3DHands dynamically generates instances using the collected distribution and does not have a static size. For this reason, the represented pose space can be significantly higher in quantity and diversity. For example, MANO3DHands_{ego} can generate 1.35×10^9 unique poses with the collected $\mathcal{B}$ and $\mathcal{G}$. Consequently, the top evaluation scores achieved by our datasets on our generated test sets (50k instances) also reflect strong generalization ability.

6.3. Two-hand Global Pose Estimation

For the complete cascaded pipeline, we use the scene-adapted ICNet and evaluate subsequent stages using input obtained from the previous stages. For 3D global hand pose estimation that requires the global 3D hand location, we follow [11] and compute the PCK on the root joint in the spherical coordinate system, which measures the directional and distance accuracy. To show that existing datasets are insufficient for our task, we select models trained on H2O and FreiHAND with good performance in Section 6.1 and 6.2 for a comparison in the complete pipeline.

Figure 8a shows that we achieve an $\text{AUC}_{2d} = 0.508$ for 2D hand pose estimation. Accurate results in this stage are necessary for two-hand tracking since both the third and final stage heavily depend on the estimated 2D keypoints. For 3D canonical hand pose estimation, we achieve an $\text{AUC}_{3d} = 0.422$ in Figure 8b. Finally, after applying ManoFit using the estimated 2D/3D hand poses, Figure 8c shows that we obtain top $\text{AUC}_{angle} = 0.912$ and



(a) 2D Pose Estimation

(b) 3D Canonical Pose Estimation

(c) Spherical PCK for Global Estimation

Figure 8: Quantitative results using the complete pipeline trained on different datasets. Results obtained using the ground truth segmentation and detection (stage1 ground truth) are provided to study the impact of Model$_{seg}$ on the overall accuracy.

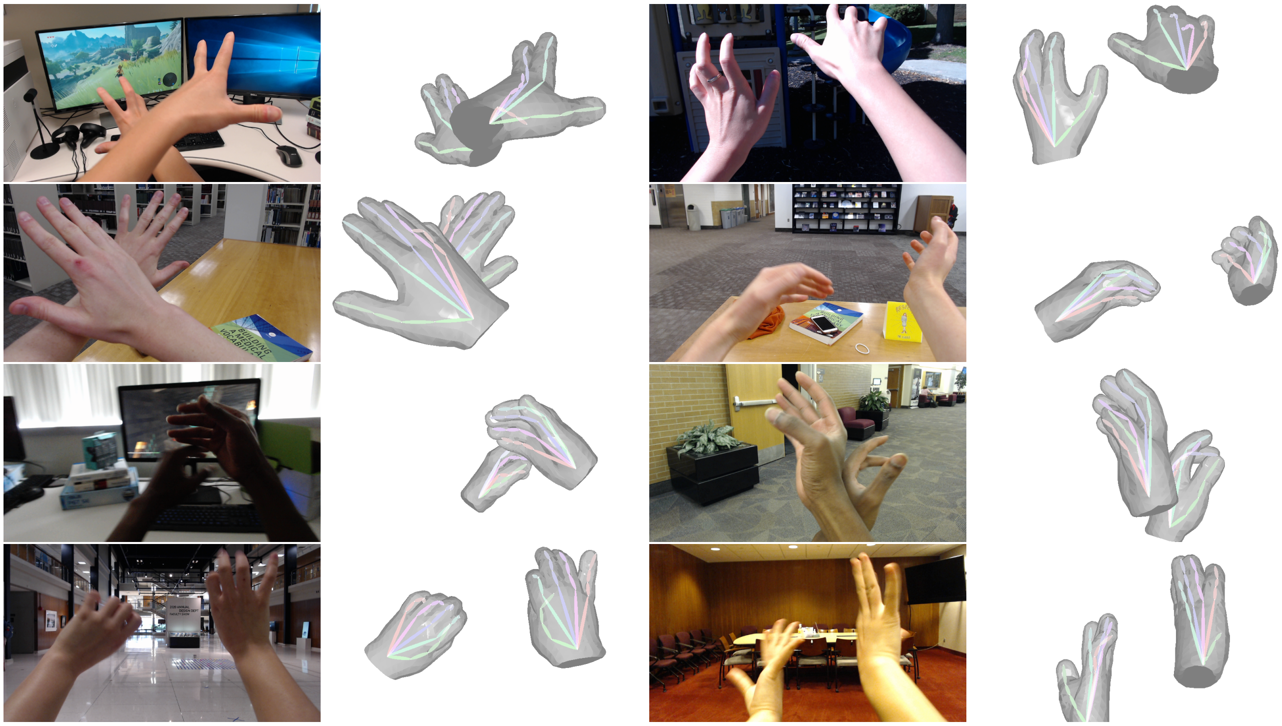


Figure 9: Qualitative results obtained using our multi-stage method trained on Ego2HandsPose. Odd columns show the input images from various sequences in the test set of Ego2HandsPose. Even columns show visualization of the ManoFit output.

AUC$_{radius}$ $= 0.672$. To isolate the impact of Model$_{seg}$ in the first stage, we also provide results obtained using the ground truth segmentation and detection, which is not used in comparison with the other datasets. Figure 9 shows qualitative examples of our two-hand 3D tracking on the test sequences of Ego2HandsPose. Additional results are provided in the supplementary material.

We note that there is a trade-off between fitting accuracy and inference time and our complete pipeline does not currently run in real-time. We report an average inference time of 19.7ms, 50.1ms and 2.4ms for our selected Model$_{seg}$, Model$_{2d}$ and Model$_{3d}$ respectively. We utilize high-performance models for the challenging task of two-hand tracking using a single camera in the wild. Future work includes further optimizations in efficiency.

7. Conclusion

In this work, we propose a set of parametric fitting algorithm that enables 3D hand pose annotation using a single image and automatic conversion from 2D to 3D hand poses. We propose the first dataset, Ego2HandsPose, that tackles two-hand 3D global pose estimation in the wild using a monocular RGB. Results obtained using our multi-stage pipeline shows that training on the proposed dataset significantly outperforms existing datasets. We hope our work can push color-based two-hand applications towards unconstrained environments for practical applications.

References

[1] Leap Motion Controller. *https://www.ultraleap.com/product/leap-motion-controller/*, 2022.

[2] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. *In ICCV*, 2015.

[3] Samarth Brahmbhatt, Chengcheng Tang, Chris Twigg, Charlie Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020.

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

[5] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *In CVPR*, 2018.

[6] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. *In CVPR*, 2020.

[7] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. *In ICCV*, 2021.

[8] JP Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. *In SIGGRAPH*, 2000.

[9] Yin Li, Miao Liu, and James M. Rehg. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. *In ECCV*, 2018.

[10] Fanqing Lin, Brian Price, and Tony Martinez. Ego2hands: A dataset for egocentric two-hand segmentation and detection. In *arXiv:2011.07252*, 2020.

[11] Fanqing Lin, Connor Wilhelm, and Tony Martinez. Two-hand global 3d pose estimation using monocular rgb. *In WACV*, 2021.

[12] Gyeongsik Moon, Shoou i Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020.

[13] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *In ACM Transactions on Graphics*, 2019.

[14] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *In SIGGRAPH Asia*, 2017.

[15] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. *In CVPR*, 2017.

[16] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *In CVPR*, 2019.

[17] J. Taylor, V. Tankovich, D. Tang, C. Keskin, D. Kim, P. Davidson, A. Kowdle, and S. Izadi. Articulated Distance Fields for Ultra-Fast Tracking of Hands Interacting. *In SIGGRAPH Asia*, 2017.

[18] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. In *IJCV*, 2016.

[19] Aisha Urooj and Ali Borji. Analysis of Hand Segmentation in the Wild. *In CVPR*, 2018.

[20] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video. In *ACM Transactions on Graphics (TOG)*, 2020.

[21] Yangang Wang, Cong Peng, and Yebin Liu. Mask-pose cascaded cnn for 2d hand pose estimation from single color images. *In TCSVT*, 2019.

[22] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. *In ICIP*, 2017.

[23] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. *In ICCV*, 2021.

[24] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. *In ICCV*, 2017.

[25] 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 ICCV*, 2019.