# Reconstructing Hands in 3D with Transformers

Georgios Pavlakos[1], Dandan Shan[2], Ilija Radosavovic[1], Angjoo Kanazawa[1], David Fouhey[3], Jitendra Malik[1]
[1]UC Berkeley, [2]University of Michigan, [3]New York University

Figure 1. **Monocular 3D hand mesh reconstruction.** We propose HaMeR, a fully transformer-based approach for **Ha**nd **Me**sh **R**ecovery. HaMeR achieves consistent improvements upon the state-of-the-art for 3D hand reconstruction. We can faithfully reconstruct hands in a wide variety of scenarios, including captures from different viewpoints (third person or egocentric), under occlusion, hands that interact with objects or other hands, hands with different skin tones, with gloves, from art paintings or mechanical hands. We encourage the reader to watch our reconstructions in the Supplemental Video to appreciate the temporal stability.

## Abstract

*We present an approach that can reconstruct hands in 3D from monocular input. Our approach for Hand Mesh Recovery, HaMeR, follows a fully transformer-based architecture and can analyze hands with significantly increased accuracy and robustness compared to previous work. The key to HaMeR's success lies in scaling up both the data used for training and the capacity of the deep network for hand reconstruction. For training data, we combine multiple datasets that contain 2D or 3D hand annotations. For the deep model, we use a large scale Vision Transformer architecture. Our final model consistently outperforms the previous baselines on popular 3D hand pose benchmarks. To further evaluate the effect of our design in non-controlled settings, we annotate existing in-the-wild datasets with 2D hand keypoint annotations. On this newly collected dataset of annotations, HInt, we demonstrate significant improvements over existing baselines. We make our code, data and models available on the project website:* `https://geopavlakos.github.io/hamer/`.

> "It is because of his being armed with hands that man is the most intelligent animal."
>
> *Anaxagoras*

## 1. Introduction

Consider the images of hands interacting with the world in Figure 1. These interactions are happening in 3D, so to interpret them, we also need a system that can automatically perceive hands in 3D from visual input.

Recent developments in computer vision and NLP point

to the direction where advances are achieved by simple, high capacity models, powered by huge amounts of data. This emerging insight has been demonstrated in the context of NLP by Large Language Models, like GPT-3 [3] and GPT-4 [43]. In the context of computer vision, we observe this with models like CLIP [45], Stable Diffusion [47] and SAM [29]. In the area of 3D human mesh recovery, a similar trend has been observed, where the simple, large scale HMR2.0 architecture [17] achieves state-of-the-art results.

In this paper, we take this philosophy and apply it to the problem of 3D hand pose estimation. We propose HaMeR, a robust and accurate approach for **Ha**nd **Me**sh **R**ecovery from images and video frames. HaMeR captures faithful 3D reconstructions of hands in a variety of poses, viewpoints and visual conditions, as shown in Figure 1. This translates to improvements over existing baselines in the standard 3D hand pose benchmarks. More importantly, HaMeR shines when evaluated on challenging in-the-wild images, where we outperform the state-of-the-art by significant margins. Even though HaMeR is a single-frame approach, it recovers temporally smooth and consistent reconstructions when applied on video frames (please see the Supplemental Video for video results).

The key to HaMeR's success lies in scaling up the techniques for hand mesh recovery. More specifically, we scale both the training data and the deep network architecture used for 3D hand reconstruction. For training data, we use multiple available sources of data with hand annotations, including both studio/controlled datasets with 3D ground truth [6, 19, 40, 56, 63, 64], and in-the-wild datasets annotated with 2D keypoint locations [15, 25, 52]. For our network, we use a large-scale transformer architecture [14, 57] which can successfully consume data of this scale. The combination of these two ingredients leads to significant improvements compared to previous work.

Benchmarking progress of these models is challenging and is often constrained on datasets captured in controlled conditions. To encourage evaluation on in-the-wild images, we introduce a new dataset of annotations, HInt, by annotating hands from diverse image sources, including videos from YouTube [9, 51] and egocentric captures [12, 18]. The annotations consist of 2D keypoints for the hand joints, as well as labels of the visibility (occluded or not) for each joint. We provide 2D hand keypoints annotations for 40.4K hands, where 86.7% of them are hands in natural contact. Even though with HInt we can only benchmark the 2D aspect of our 3D reconstruction, this evaluation is complementary to the existing benchmarks due to its diversity of data, and together provide a more holistic picture on the performance of different systems.

We contribute HaMeR, an approach for 3D hand mesh reconstruction from images and video frames. We demonstrate the key effect of scaling up to large scale training data and high capacity deep architectures for the problem of hand mesh recovery. We achieve state-of-the-art results where we obtain 2-3$\times$ improvement in PCK@0.05 on in-the-wild datasets compared to previous works. We also contribute HInt, a dataset of annotations that complements training and evaluation of 3D hand reconstruction approaches. We make our model, code and data available to support future work.

## 2. Related work

**3D hand pose and shape estimation.** In this section we focus specifically on the works that estimate 3D hand pose and shape from a single RGB image. The earlier efforts [1, 2, 62] take inspiration from related work on human mesh recovery [27] - they use the MANO parametric hand model [48] and regress the hand pose and shape parameters given an RGB image as input. FrankMocap [49] is a good representative of this line of works which adopts a simple design, similar to HMR [27]. Followup work [10, 16, 31, 39] follows a non-parametric approach and directly regresses the vertices of the MANO mesh. This strategy often leads to results that align better with the image evidence, but it is more prone to failure in cases of occlusions and truncations. The improvements in 3D hand pose estimation have also lead to progress in related problems, including joint hand pose and object reconstruction [21, 22, 54, 58] and reconstruction of two interacting hands [28, 32, 37, 38, 46, 55, 60, 61, 65]. More recently, there have been works that address other aspects of the problem. MobRecon [8] focuses on high inference speed, that could potentially be supported on a mobile device. HandOccNet [44] designs an architecture that could offer increased robustness to occlusions. AMVUR [24] proposes a probabilistic approach for hand pose and shape estimation. BlurHand [41] focuses on the problem of motion blur that often exists in footage that captures hand motion. Our work is orthogonal to these approaches. We adopt a simple design and we investigate the effect of scaling up the training data and the capacity of our architecture. Given that our main design is simple, the different choices of previous work could be combined with our HaMeR architecture which could potentially lead to further improvements.

**Hand datasets.** Many of the datasets used to train and evaluate 3D hand pose estimation systems are captured in indoor/studio settings and provide 3D ground truth. FreiHAND [64] is captured in a multi-camera setting and focuses on different hand poses as well as hands interacting with objects. HO-3D [19] and DexYCB [6] are also captured in a controlled setting with multiple cameras but focuses more specifically on cases where hands interact with objects. InterHand2.6M [40] is captured in a studio with a focus on two interacting hands. Hand pose datasets [52, 56] captured in the Panoptic studio [26] also offer 3D hand an-

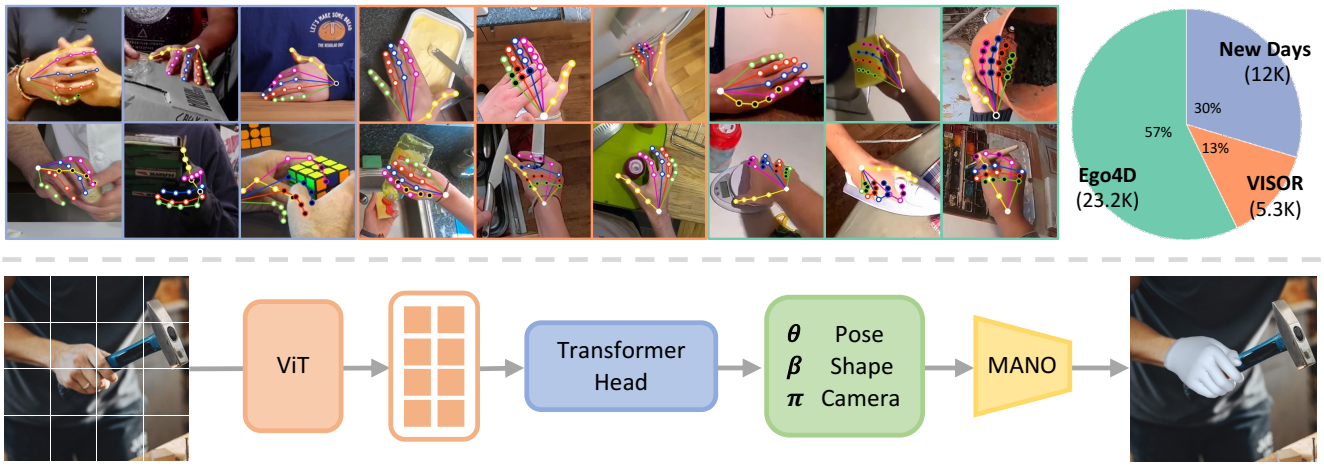


Figure 2. **Dataset and Architecture. (Top)** Hand crops with keypoint annotations from our HInt dataset of annotations for different image sources, Hands23 [9], Epic-Kitchens [12, 13], and Ego4D [18]. We provide location annotations for 21 hand keypoints as well as the "occlusion" label for each joint. Occluded keypoints are marked using solid dot filled with black while non-occluded ones are filled with white. The pie chart shows the distribution and statistics of our dataset. **(Bottom)** The architecture for HaMeR follows a fully transformer-based design. We use a large scale ViT backbone [14] followed by a transformer decoder to regress the parameters of the hand.

notations. AssemblyHands [42] annotated 3D hand poses for synchronized images from Assembly101 [50] which participants assemble and disassemble take-apart toys in a multi-camera setting. In this work, we use these datasets for training and evaluation. However, we also argue that to get a more holistic picture about the accuracy and the robustness of 3D hand pose estimation systems, it is important to evaluate performance on in-the-wild images as well.

While we cannot annotate 3D ground truth poses for in-the-wild images, there is work that annotates 2D keypoint positions. Among the larger scale efforts, COCO-WholeBody [25] provides hand annotations for the people in the COCO dataset [35] and Halpe [15] annotates hands in the HICO-DET dataset [4, 5]. Both of them source images from image datasets that contain very few egocentric images or transitionary moments. In our dataset, HInt, we sourced images from both egocentric and third-person video datasets. Since our annotated hands come from video, they depict more natural interactions with the world.

## 3. Technical approach

In this section, we describe HaMeR, our approach for hand mesh recovery from monocular input. We follow a simple, fully "transformerized" design that focuses on scaling up the training data and the deep model architecture.

### 3.1. MANO parametric hand model

We adopt the MANO parametric model of the human hand [48]. MANO takes as input the pose parameters $\theta \in \mathbb{R}^{48}$ and shape parameters $\beta \in \mathbb{R}^{10}$ and defines a function $\mathcal{M}(\theta, \beta)$ that returns the mesh of the hand $M \in \mathbb{R}^{V \times 3}$, with $V = 778$ vertices. MANO additionally returns the joints $X \in \mathbb{R}^{K \times 3}$ of the hand, for a total of $K = 21$ joints.

### 3.2. Hand mesh recovery

Given an RGB image of a hand, $I$, our goal is to reconstruct the 3D hand surface. We approach this problem by estimating the MANO pose and shape parameters for the hand in the image. Similar to previous work in the parametric human [17, 27] and hand [49, 62] reconstruction, we use a network to learn the mapping $f$ from image pixels to MANO parameters. Our regressor also estimates camera parameters $\pi$. The camera $\pi$ corresponds to a translation $t \in \mathbb{R}^3$ that allows us to project the 3D mesh and the 3D joints to the image. Given fixed camera intrinsics $K$, the projection of the 3D joints $X$ is: $x = \pi(X) = \Pi_K(X + t)$. Eventually, we learn the mapping $f(I) = \Theta$, where the regressed parameters are $\Theta = \{\theta, \beta, \pi\}$.

### 3.3. Architecture

HaMeR adopts a simple architecture with a fully transformer-based design (Figure 2, bottom), similar to [17]. We use a Vision Transformer (ViT) [14] as the backbone, followed by a transformer head that regresses the hand and camera parameters. We first convert the input RGB image to patches, which are fed as input tokens to ViT which follows the "huge" design, *i.e.*, ViT-H. The ViT backbone processes the image patches and returns a series of output tokens. The transformer head is a transformer decoder that processes a single token while cross-attending to the ViT output tokens. The output of the head returns the parameters $\Theta$ for the input image.

### 3.4. Losses

For our training losses, we follow best practices for parametric human and hand reconstruction [17, 27, 30, 49] and supervise our model with a combination of 2D and 3D losses. For the images that provide 3D ground truth, we can directly apply a loss on the model parameters, $\theta$ and $\beta$. Simultaneously, we can encourage consistency in the actual 3D space, and supervise on the level of the 3D joints $X^*$:

$$\mathcal{L}_{\texttt{3D}} = ||\theta - \theta^*||_2^2 + ||\beta - \beta^*||_2^2 + ||X - X^*||_1. \quad (1)$$

To enable training with 2D annotations, we also apply a reprojection loss between the projection $x$ of the 3D joints $X$ and the ground truth 2D keypoint annotations $x^*$:

$$\mathcal{L}_{\texttt{2D}} = ||x - x^*||_1. \quad (2)$$

We apply this loss, even when 3D ground truth is available, since it promotes consistency on the output image space.

Finally, if only 2D keypoints are available, it is possible to recover an unnatural pose that still reprojects well to the image. To encourage the reconstruction of natural hands, we train discriminators $D_k$ for a) the hand shape $\beta$, b) the hand pose $\theta$, and c) each hand joint angle separately [27]. Then, we can apply an adversarial loss:

$$\mathcal{L}_{\texttt{adv}} = \sum_k (D_k(\Theta) - 1)^2. \quad (3)$$

### 3.5. Training data

To train our model, we consolidate multiple datasets that provide 2D or 3D hand annotations. Specifically, we use FreiHAND [64], HO3D [19], MTC [56], RHD [63], InterHand2.6M [40], H2O3D [19], DEX YCB [6], COCO WholeBody [25], Halpe [15] and MPII NZSL [52]. This results to 2.7M training examples, which is $4\times$ larger than the training set of the popular FrankMocap system [49]. The majority of this data is collected in controlled environments (*e.g.*, studio or multi-camera setup), while only 5% of the training examples (COCO WholeBody, Halpe and MPII NZSL) include images from in-the-wild datasets.

## 4. HInt: Hand Interactions in the wild

In this section, we describe the dataset we contribute, with the goal to complement existing datasets used for training and evaluation. Since we focus on **H**and **Int**eractions in the wild, we call our dataset HInt. HInt annotates 2D hand keypoint locations and occlusion labels for each keypoint. We built off of Hands23 [9] (using an early copy otained from the authors), Epic-Kitchens [12], and Ego4D [18].

By sourcing from video datasets, we harvest more transitional moments and natural poses, compared with sourcing from image data. For HInt, we source frames from three video datasets. In Hands23, we choose from the New Days subset [9] containing YouTube video frames of humans engaging in daily activities. In Epic-Kitchens, we choose frames from VISOR [13] containing frames extracted from cooking actions. In Ego4D [18], we choose frames from the critical frames (pre45, pre30, pre15, pre-frame, contact-frame, point-of-no-return frame, and post-frame) in the FHO (Forecasting Hands and Objects) task.

For our validation and test set, we randomly sample frames to keep data distribution the same as source datasets. For the training set, our goal is to include more challenging samples to compensate for other existing 2D keypoints datasets. Thus, for New Days and VISOR, we chose half of the samples using random sampling and forcing the other half to contain hand-object or hand-hand interaction. For Ego4D, we still randomly sample frames since the critical frames already typically focus on interactions.

Annotating hand keypoints from scratch can be time-consuming. Similar to [25], we initialize the annotation procedure with an existing keypoint detection model [11] to get rough keypoint locations. Given the annotation instructions (details in the Supplemental Material), workers are asked to correct the keypoint locations (see annotation samples in Figure 2, top). Additionally, each keypoint is annotated with an "existence" and an "occlusion" label. Existence indicates whether the keypoint exists within the image frame or not. Occlusion indicates whether the keypoint is occluded or not. To the best of our knowledge, HInt is the first dataset to provide "occlusion" annotations for 2D hand keypoints. We believe this can lead to a more fine-grained analysis of the pose estimation systems.

In total, we annotate 40.4K hands with keypoints, 12.0K for New Days, 5.3K for VISOR, and 23.2K for Ego4D. In our Ego4D subset, we annotate 9.3K hands from sequences, which could help future evaluation of temporal tasks.

Finally, we perform an annotation consistency check, by having 90 valid images annotated twice. Across this subset, 90.5% of the occlusion labels and 100% of "existence" labels are consistent. In terms of keypoint locations, 94.6% visible keypoints have offset distance within $0.25\times$ of the palm length (see details about the data annotation process and analysis in the Supplemental Material).

## 5. Experiments

In this section, we present the quantitative and qualitative evaluation of our system. First, we evaluate the 3D pose accuracy (subsection 5.1) and the 2D pose accuracy (subsection 5.2) of HaMeR. Then, we ablate some characteristics of our system (subsection 5.3) and present qualitative results and comparisons (subsection 5.4).

### 5.1. 3D pose accuracy

To evaluate the 3D accuracy of HaMeR, we use two standard benchmarks for 3D hand pose estimation, FreiHAND [64] and HO3Dv2 [19]. Both datasets are collected in controlled multi-camera environments and provide 3D ground truth annotations in the form of 3D hand meshes (using the MANO model). To be comparable with previous work, we follow the typical protocols [34, 44], and we report metrics that evaluate 3D joint and 3D mesh accuracy. These metrics include PA-MPJPE and $AUC_J$ (3D joints evaluation), PA-MPVPE, $AUC_V$, F@5mm and F@15mm (3D mesh evaluation).

We present the complete results for FreiHAND in Table 1 and for HO3Dv2 in Table 2. We compare with many baselines that estimate the 3D hand mesh from a single image in parametric or non-parametric form (*i.e.*, regressing hand model parameters or hand model vertices respectively). We observe that our HaMeR approach achieves state-of-the-art results and consistently outperforms the previous work across the majority of the metrics.

### 5.2. 2D pose accuracy

Although the 3D hand pose datasets provide accurate 3D ground truth for evaluation, they are typically collected in controlled settings, which limits the variety of subjects, viewpoints, objects of interactions, environments, etc. To better analyze the properties of the different hand pose estimation systems, we also propose to evaluate on our HInt benchmark that is closer to real in-the-wild conditions, compared to the previous 3D benchmarks. The annotations of HInt are in the form of 2D keypoints. Metrics based on 2D only evaluate reprojection accuracy of 3D methods, but due to the nature of the images (*i.e.*, in the wild), we can get complementary evidence about the performance of our method. For evaluation, we report results with the commonly used PCK metric [59], computed at different thresholds. Given the form of HInt, we provide a more detailed analysis, reporting separate results for images coming from New Days [9], VISOR [12] and Ego4D [18]. Moreover, we provide more fine-grained results, considering all the joints, considering only the joints that have been annotated as visible (non-occluded), or considering only the joints that have been annotated as occluded.

The complete results are presented in Table 3. Here, we compare with a number of recent 3D hand mesh estimation approaches that provide publicly available code. Similarly with the results on FreiHAND and HO3D, we observe that our method outperforms the previous baselines. However, on these datasets we observe much larger improvements. This highlights the clear improvement in the robustness of our approach which performs consistently across a variety of benchmarks. Performance on FreiHAND and HO3D tends to be more saturated and it is not surprising that the margin of improvement for our approach on these datasets is smaller. In contrast, performance on in-the-wild datasets is more representative of the robustness of the approaches in different visual conditions, different viewpoints and different interactions, *e.g.*, contacts with surrounding objects.

| Method | PA-MPJPE ↓ | PA-MPVPE ↓ | F@5 ↑ | F@15 ↑ |
|---|---|---|---|---|
| I2L-MeshNet [39] | 7.4 | 7.6 | 0.681 | 0.973 |
| Pose2Mesh [10] | 7.7 | 7.8 | 0.674 | 0.969 |
| I2UV-HandNet [7] | 6.7 | 6.9 | 0.707 | 0.977 |
| METRO [33] | 6.5 | 6.3 | 0.731 | 0.984 |
| Tang *et al*. [53] | 6.7 | 6.7 | 0.724 | 0.981 |
| Mesh Graphormer [34] | 5.9 | 6.0 | 0.764 | 0.986 |
| MobRecon [8] | **5.7** | 5.8 | 0.784 | 0.986 |
| AMVUR [24] | 6.2 | 6.1 | 0.767 | 0.987 |
| Ours | 6.0 | **5.7** | **0.785** | **0.990** |

Table 1. **Comparison with the state-of-the-art on the FreiHAND dataset [64].** We use the standard protocol and report metrics for evaluation of 3D joint and 3D mesh accuracy. PA-MPVPE and PA-MPJPE numbers are in mm.

| Method | $AUC_J$ ↑ | PA-MPJPE ↓ | $AUC_V$ ↑ | PA-MPVPE ↓ | F@5 ↑ | F@15 ↑ |
|---|---|---|---|---|---|---|
| Liu *et al*. [36] | 0.803 | 9.9 | 0.810 | 9.5 | 0.528 | 0.956 |
| HandOccNet [44] | 0.819 | 9.1 | 0.819 | 8.8 | 0.564 | 0.963 |
| I2UV-HandNet [7] | 0.804 | 9.9 | 0.799 | 10.1 | 0.500 | 0.943 |
| Hampali *et al*. [19] | 0.788 | 10.7 | 0.790 | 10.6 | 0.506 | 0.942 |
| Hasson *et al*. [21] | 0.780 | 11.0 | 0.777 | 11.2 | 0.464 | 0.939 |
| ArtiBoost [58] | 0.773 | 11.4 | 0.782 | 10.9 | 0.488 | 0.944 |
| Pose2Mesh [10] | 0.754 | 12.5 | 0.749 | 12.7 | 0.441 | 0.909 |
| I2L-MeshNet [39] | 0.775 | 11.2 | 0.722 | 13.9 | 0.409 | 0.932 |
| METRO [33] | 0.792 | 10.4 | 0.779 | 11.1 | 0.484 | 0.946 |
| MobRecon[8] | - | 9.2 | - | 9.4 | 0.538 | 0.957 |
| Keypoint Trans [20] | 0.786 | 10.8 | - | - | - | - |
| AMVUR [24] | 0.835 | 8.3 | 0.836 | 8.2 | 0.608 | 0.965 |
| Ours | **0.846** | **7.7** | **0.841** | **7.9** | **0.635** | **0.980** |

Table 2. **Comparison with the state-of-the-art on the HO3D dataset [19].** We use the HO3Dv2 protocol and report metrics that evaluate accuracy of the estimated 3D joints and 3D mesh. PA-MPVPE and PA-MPJPE numbers are in mm.

### 5.3. Ablation analysis

Having demonstrated the effectiveness of HaMeR, we further ablate different options for our system.

**Effect of large scale data and deep model.** One of the key aspects of HaMeR is that a simple design can achieve strong performance if we scale up, *i.e.*, train with large scale data and use a large scale model for the hand reconstruction. We evaluate these choices using different models on HInt and we present the complete results in Table 5. More specifically, we start from a basic design (2nd row of Table 5), that follows the choices of [49] (1st row of Table 5), using a ResNet50 architecture [23] and a relatively small training set (only a quarter of the examples we use to train HaMeR). This basic design is indeed very close to [49] in terms of quantitative results. Then, by keeping the architecture the same, we increase the volume of training examples, using our complete training set. This model (3rd row of Table 5)

| | Method | New Days | | | VISOR | | | Ego4D | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | @0.05 | @0.1 | @0.15 | @0.05 | @0.1 | @0.15 | @0.05 | @0.1 | @0.15 |
| **All Joints** | FrankMocap [49] | 16.1 | 41.4 | 60.2 | 16.8 | 45.6 | 66.2 | 13.1 | 36.9 | 55.8 |
| | METRO [33] | 14.7 | 38.8 | 57.3 | 16.8 | 45.4 | 65.7 | 13.2 | 35.7 | 54.3 |
| | MeshGraphormer [34] | 16.8 | 42.0 | 59.7 | 19.1 | 48.5 | 67.4 | 14.6 | 38.2 | 56.0 |
| | HandOccNet (param) [44] | 9.1 | 28.4 | 47.8 | 8.1 | 27.7 | 49.3 | 7.7 | 26.5 | 47.7 |
| | HandOccNet (no param) [44] | 13.7 | 39.1 | 59.3 | 12.4 | 38.7 | 61.8 | 10.9 | 35.1 | 58.9 |
| | Ours | **48.0** | **78.0** | **88.8** | **43.0** | **76.9** | **89.3** | **38.9** | **71.3** | **84.4** |
| **Visible Joints** | FrankMocap [49] | 20.1 | 49.2 | 67.6 | 20.4 | 52.3 | 71.6 | 16.3 | 43.2 | 62.0 |
| | METRO [33] | 19.2 | 47.6 | 66.0 | 19.7 | 51.9 | 72.0 | 15.8 | 41.7 | 60.3 |
| | Mesh Graphormer [34] | 22.3 | 51.6 | 68.8 | 23.6 | 56.4 | 74.7 | 18.4 | 45.6 | 63.2 |
| | HandOccNet (param) [44] | 10.2 | 31.4 | 51.2 | 8.5 | 27.9 | 49.8 | 7.3 | 26.1 | 48.0 |
| | HandOccNet (no param) [44] | 15.7 | 43.4 | 64.0 | 13.1 | 39.9 | 63.2 | 11.2 | 36.2 | 60.3 |
| | Ours | **60.8** | **87.9** | **94.4** | **56.6** | **88.0** | **94.7** | **52.0** | **83.2** | **91.3** |
| **Occluded Joints** | FrankMocap [49] | 9.2 | 28.0 | 46.9 | 11.0 | 33.0 | 55.0 | 8.4 | 26.9 | 45.1 |
| | METRO [33] | 7.0 | 23.6 | 42.4 | 10.2 | 32.4 | 53.9 | 8.1 | 26.2 | 44.7 |
| | MeshGraphormer [34] | 7.9 | 25.7 | 44.3 | 10.9 | 33.3 | 54.1 | 8.3 | 26.9 | 44.6 |
| | HandOccNet (param) [44] | 7.2 | 23.5 | 42.4 | 7.4 | 26.1 | 46.7 | 8.0 | 26.1 | 45.7 |
| | HandOccNet (no param) [44] | 9.8 | 31.2 | 50.8 | 9.9 | 33.7 | 55.4 | 9.6 | 31.1 | 52.7 |
| | Ours | **27.2** | **60.8** | **78.9** | **25.9** | **60.8** | **80.7** | **23.0** | **56.9** | **76.3** |

Table 3. **Evaluation on our HInt benchmark.** We report results using PCK scores at three different thresholds. All methods are 3D and we evaluate the scores through the 2D projection of 3D joints. We report separate results for the three subsets of HInt, *i.e.*, New Days of Hands [9], Epic- Kitchens VISOR [13] and Ego4D [18]. We also report separate results considering all joints (first set of rows), considering only the joints annotated as visible (second set of rows), or considering only the joints annotated as occluded (third set of rows).

| | Method | New Days | | | VISOR | | | Ego4D | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | @0.05 | @0.1 | @0.15 | @0.05 | @0.1 | @0.15 | @0.05 | @0.1 | @0.15 |
| **All** | Ours | 48.0 | 78.0 | 88.8 | 43.0 | 76.9 | 89.3 | 38.9 | 71.3 | 84.4 |
| | Ours* | **51.6** | **81.9** | **91.9** | **56.5** | **88.1** | **95.6** | **46.9** | **79.3** | **90.4** |
| **Vis.** | Ours | 60.8 | 87.9 | 94.4 | 56.6 | 88.0 | 94.7 | 52.0 | 83.2 | 91.3 |
| | Ours* | **62.9** | **89.4** | **95.8** | **66.5** | **92.7** | **97.4** | **59.1** | **87.0** | **94.0** |
| **Occl.** | Ours | 27.2 | 60.8 | 78.9 | 25.9 | 60.8 | 80.7 | 23.0 | 56.9 | 76.3 |
| | Ours* | **33.2** | **68.4** | **84.8** | **42.6** | **79.0** | **91.3** | **33.1** | **69.8** | **84.9** |

Table 4. **Effect of training with HInt.** We compare our general model (Ours) with the model trained on HInt as well (Ours*). We report PCK scores on the test set of HInt. Using the training set of HInt can be helpful particularly to improve performance on egocentric data (VISOR and Ego4D).

achieves already consistent improvements over the previous baseline. Similarly, if we use the small training set of the basic design, but adopt a large scale architecture, here ViT-H [14] (4th row of Table 5), we also see improvements over the basic design. Finally, we can combine the two independent updates, *i.e.*, increase the volume of training examples while using a high capacity architecture, which effectively is the design of HaMeR. This version (5th row of Table 5) outperforms by a large margin the other versions, demonstrating the effect of both large data and large deep model in our design.

**Training with HInt.** When comparing with previous work, we avoided training with the training set of HInt. However, here we provide a direct comparison when training with this data too. In Table 4 we present the detailed results on HInt. We observe a clear improvement on VISOR and Ego4D, the two egocentric datasets included in HInt. This can be explained by the fact that there have been little to no egocentric data with hand annotations in the wild before, so using some form of annotations for training can help improve our model. Besides this, we also observe an improvement in New Days. The improvement is smaller, given that New Days mainly includes third-person videos, but it is still consistent across all metrics.

### 5.4. Qualitative results

We show qualitative results of our approach in Figure 1, while we do a more detailed analysis in Figure 4, where we show side and top views of our 3D hand reconstructions. Our approach is robust to different viewpoints, different skin tones or hand appearance (*e.g.* wearing different types of gloves) as well as different objects of interaction that can create various degrees of occlusion. Moreover, in Figure 3, we show more detailed comparisons with previous baselines. We compare with METRO [33], MeshGraphormer [34] and FrankMocap [49]. Following the trend of the quantitative comparison, HaMeR is consistently more robust and precise than the previous work.

## 6. Conclusion

We present HaMeR, an approach for 3D hand mesh reconstruction from monocular input. HaMeR is simple, without bells and whistles and demonstrates the importance of two

| | Method | Large Data | Large Model | New Days @0.05 | New Days @0.1 | New Days @0.15 | VISOR @0.05 | VISOR @0.1 | VISOR @0.15 | Ego4D @0.05 | Ego4D @0.1 | Ego4D @0.15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| All | FrankMocap [49] | ✗ | ✗ | 16.1 | 41.4 | 60.2 | 16.8 | 45.6 | 66.2 | 13.1 | 36.9 | 55.8 |
| | Base design | ✗ | ✗ | 16.9 | 43.6 | 62.6 | 17.5 | 47.5 | 67.3 | 13.9 | 37.8 | 56.0 |
| | + large data | ✓ | ✗ | 31.3 | 65.7 | 81.8 | 29.9 | 65.0 | 81.7 | 24.7 | 56.1 | 73.9 |
| | + large model | ✗ | ✓ | 25.9 | 58.9 | 76.9 | 24.1 | 62.5 | 81.2 | 19.4 | 51.6 | 71.1 |
| | HaMeR | ✓ | ✓ | **48.0** | **78.0** | **88.8** | **43.0** | **76.9** | **89.3** | **38.9** | **71.3** | **84.4** |
| Visible | FrankMocap [49] | ✗ | ✗ | 20.1 | 49.2 | 67.6 | 20.4 | 52.3 | 71.6 | 16.3 | 43.2 | 62.0 |
| | Base design | ✗ | ✗ | 21.2 | 51.5 | 70.4 | 21.4 | 54.5 | 73.5 | 17.4 | 45.0 | 63.7 |
| | + large data | ✓ | ✗ | 38.5 | 75.0 | 88.0 | 36.6 | 73.2 | 86.9 | 30.4 | 64.8 | 80.9 |
| | + large model | ✗ | ✓ | 33.1 | 69.3 | 85.0 | 29.2 | 72.8 | 88.9 | 24.3 | 62.3 | 81.3 |
| | HaMeR | ✓ | ✓ | **60.8** | **87.9** | **94.4** | **56.6** | **88.0** | **94.7** | **52.0** | **83.2** | **91.3** |
| Occluded | FrankMocap [49] | ✗ | ✗ | 9.2 | 28.0 | 46.9 | 11.0 | 33.0 | 55.0 | 8.4 | 26.9 | 45.1 |
| | Base design | ✗ | ✗ | 9.4 | 29.8 | 48.8 | 11.8 | 35.6 | 57.4 | 9.2 | 27.4 | 44.8 |
| | + large data | ✓ | ✗ | 19.0 | 49.4 | 70.7 | 19.1 | 51.6 | 72.5 | 17.0 | 44.5 | 64.3 |
| | + large model | ✗ | ✓ | 14.7 | 41.6 | 63.2 | 16.3 | 47.9 | 69.4 | 14.5 | 40.1 | 59.9 |
| | HaMeR | ✓ | ✓ | **27.2** | **60.8** | **78.9** | **25.9** | **60.8** | **80.7** | **23.0** | **56.9** | **76.3** |

Table 5. **Effect of large scale data and deep model.** We evaluate the effect of different design choices when testing on HInt. We start from a basic design that follows FrankMocap [49], using a ResNet50 architecture and a small training set (2nd row). Increasing the amount of training data by $4\times$ (3rd row) or adopting a high capacity ViT-H architecture (4th row) results in clear and consistent improvements in 2D accuracy over the base model. Combining the data scale and high capacity architecture, which is the proposed HaMeR (5th row), obtains the best results by large margins.

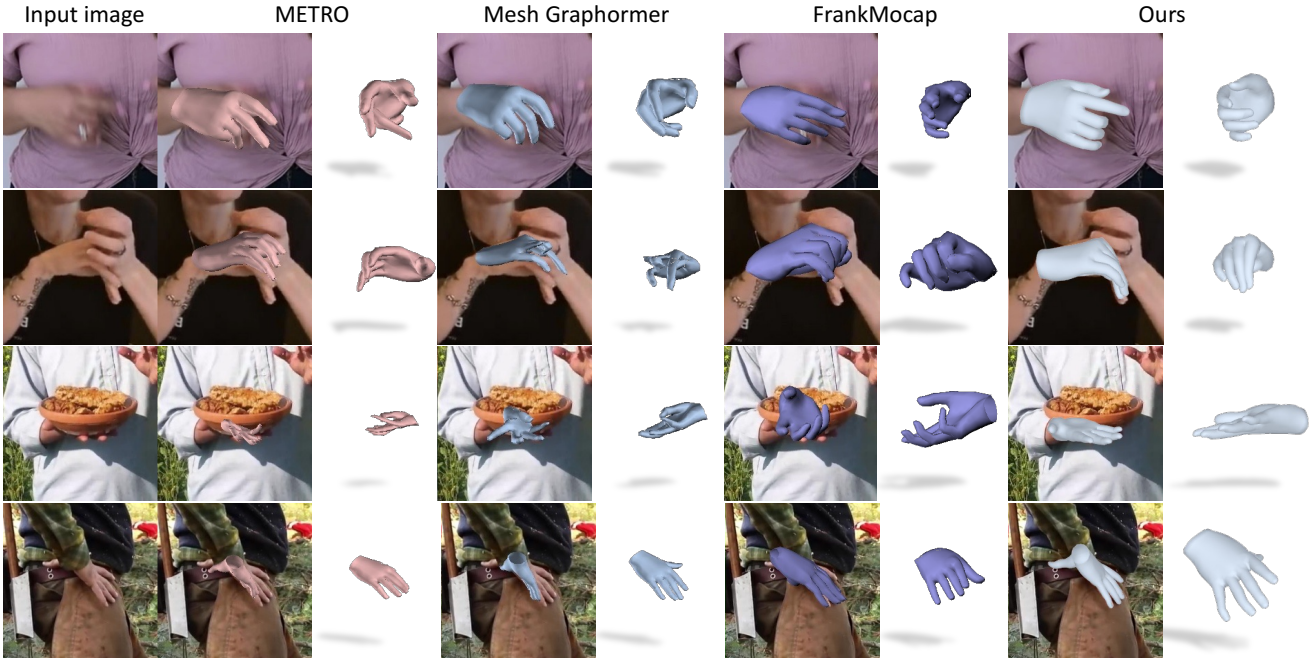


Figure 3. **Qualitative comparison.** We compare our approach qualitatively with state-of-the-art methods for hand mesh reconstruction. The previous baselines include METRO [33], Mesh Graphormer [34] and FrankMocap [49]. METRO and Mesh Graphormer are non-parametric methods (regressing MANO vertices directly), while FrankMocap and HaMeR (ours) are parametric methods (regressing MANO parameters). The reconstructions from HaMeR are consistently better, particularly on more challenging examples, *e.g.*, cases with motion blur, or images with hand-hand or hand-object interaction. We encourage the reader to also watch the Supplemental Video for more comparisons over time.

design choices — scaling up the hand mesh recovery models in terms of a) the training data and b) the architecture we use for 3D hand reconstruction. By consolidating multiple datasets with hand annotations (either 2D or 3D) and adopt-


Figure 4. **Qualitative results.** We present qualitative results of our approach on the test set of HInt. We include images from New Days (row 1-2), VISOR (row 3-4), Ego4D (row 5-6), as well as various Internet images (row 7-8). HaMeR is particularly robust and can gracefully handle cases with heavy occlusion and interactions with objects or other hands.

ing a high capacity deep model (ViT-H [14]), we are able to outperform previous work on traditional 3D hand pose benchmarks. Additionally, we contribute 2D keypoint annotations for datasets with diverse hands, coming from egocentric [12, 18] views or YouTube videos [9]. Evaluation on this challenging new HInt benchmark demonstrates the even bigger improvements that our approach achieves compared to previous baselines. We hope that the robustness and the precision of our approach will ignite the interest for further use of our system in applications that 3D hand estimation is important, including, but not limited to, robotics, action recognition and sign language understanding.

**Acknowledgements** We thank members of the BAIR community for helpful discussions and StabilityAI for supporting us through a compute grant. This work was supported by BAIR/BDD sponsors, ONR MURI (N00014-21-1-2801), and the DARPA MCS program. DF and DS were supported by the National Science Foundation under Grant No. 2006619.

## References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*, 2019.

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

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

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[6] 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 *CVPR*, 2021.

[7] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2UV-HandNet: Image-to-UV prediction network for accurate and high-fidelity 3D hand mesh modeling. In *ICCV*, 2021.

[8] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022.

[9] Tianyi Cheng, Dandan Shan, Ayda Sultan, Jiaqi Geng, Richard EL Higgins, and David F Fouhey. Towards a richer 2D understanding of hands at scale. In *NeurIPS*, 2023.

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020.

[11] MMPose Contributors. OpenMMLab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *ECCV*, 2018.

[13] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR benchmark: VIdeo Segmentations and Object Relations. In *NeurIPS Track on Datasets and Benchmarks*, 2022.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[15] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. *PAMI*, 2022.

[16] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 2019.

[17] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

[18] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020.

[20] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3D pose estimation. In *CVPR*, 2022.

[21] 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 *CVPR*, 2019.

[22] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020.

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

[24] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3D hand reconstruction from a single RGB image. In *CVPR*, 2023.

[25] 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 *ECCV*, 2020.

[26] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015.

[27] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[28] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *ICCV*, 2021.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

[30] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[31] 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.

[32] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, 2022.

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021.

[34] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[36] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*, 2021.

[37] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3D interacting hand pose estimation by hand de-occlusion and removal. In *ECCV*, 2022.

[38] Gyeongsik Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *CVPR*, 2023.

[39] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020.

[40] 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.

[41] Yeonguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee. Recovering 3D hand mesh sequence from a single blurry image: A new dataset and temporal unfolding. In *CVPR*, 2023.

[42] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *CVPR*, 2023.

[43] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[44] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *CVPR*, 2022.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[46] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single RGB image. In *ICCV*, 2023.

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[48] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017.

[49] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *ICCV*, 2021.

[50] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, 2022.

[51] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020.

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

[53] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3dD hand-mesh reconstruction. In *ICCV*, 2021.

[54] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *CVPR*, 2022.

[55] Congyi Wang, Feida Zhu, and Shilei Wen. MeMaHand: Exploiting mesh-mano interaction for single image two-hand reconstruction. In *CVPR*, 2023.

[56] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019.

[57] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022.

[58] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. ArtiBoost: Boosting articulated 3D hand-object pose estimation via online exploration and synthesis. In *CVPR*, 2022.

[59] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 2012.

[60] Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P Breckon, and Jue Wang. ACR: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *CVPR*, 2023.

[61] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3D pose and shape reconstruction from single color image. In *ICCV*, 2021.

[62] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *ICCV*, 2019.

[63] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017.

[64] 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.

[65] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*, 2023.