WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild

Rolandos Alexandros Potamias[1], Jinglei Zhang[2],
Jiankang Deng[1], Stefanos Zafeiriou[1]
[1]Imperial College London, [2]Shanghai Jiao Tong University

Figure 1. We propose **WiLoR**, a full-stack in-the-**Wi**ld **Lo**calization and 3D hand **R**econstruction method. WiLoR first localizes and defines the handedness of the detected hands which are then lifted to 3D using a transformer-based hand pose estimation module. To aid high-fidelity reconstructions and facilitate image-alignment, we employ a refinement module that extracts localized features to correct miss-aligned poses. WiLoR achieves state-of-the-art performance under different benchmark datasets while boosting the temporal coherence of image-based 3D hand pose estimation methods.

Abstract

In recent years, 3D hand pose estimation methods have garnered significant attention due to their extensive applications in human-computer interaction, virtual reality, and robotics. In contrast, there has been a notable gap in hand detection pipelines, posing significant challenges in constructing effective real-world multi-hand reconstruction systems. In this work, we present a data-driven pipeline for efficient multi-hand reconstruction in the wild. The proposed pipeline is composed of two components: a real-time fully convolutional hand localization and a high-fidelity transformer-based 3D hand reconstruction model. To tackle the limitations of previous methods and build a robust and stable detection network, we introduce a large-scale dataset with over than 2M in-the-wild hand images with diverse lighting, illumination, and occlusion conditions. Our approach outperforms previous methods in both efficiency and accuracy on popular 2D and 3D benchmarks. Finally, we showcase the effectiveness of our pipeline to achieve smooth 3D hand tracking from monocular videos, without utilizing any temporal components. Code, models, and dataset are available on our project page.

1. Introduction

Hand detection and reconstruction has been a long-studied problem due to its numerous applications, ranging from virtual reality [22] to sign language [2, 30] and human behaviour recognition [65]. Given the large variations in hand appearance and articulation [64] along with heavy occlusions and motion blur [24, 26] that are usually present in hand interactions, the task of hand pose estimation is a considerably challenging. Over the years, several methods have been proposed to tackle 3D hand pose estimation [11, 42, 49, 61]. However, despite producing credible results, these methods primarily focus on images containing a fixed number of hands and hence cannot generalize to in-the-wild images.

In the closely related fields of 3D human body and face reconstruction, state-of-the-art methods [7, 17, 21, 23] employ bottom-up pipelines founded on top of high-performance detection models that initially localize human body and face within the image, enabling their generalization to in-the-wild images. Despite the numerous methods that have been proposed to solve the task of human body and face detection, there has been a notable lack in real-time hand detection methods. The importance of hand detectors is further emphasized considering that current 3D hand pose estimation frameworks operate on tight crops around the hand regions [10, 96], high fidelity detectors are essential for the generalisation of such methods in in-the-wild scenarios. Popular hand detection and localisation methods [7, 92], fail significantly to detect multiple hands and challenging poses, while more recent methods [39, 52, 53] albeit producing reasonable results can not operate in real-time. Motivated from the lack of accurate hand detection frameworks, we propose a robust single-state anchor-free detector that can operates in over than 100 frames-per-second (fps). As we experimentally showcase, robust detections can enforce more stable 4D reconstructions and overcome jittering artifacts which is currently one of the main limitations of 3D frame-based pose estimation methods.

In contrast to the relatively unexplored hand detection and localization, 3D hand pose estimation has received significantly more attention. Initial 3D pose reconstruction methods have focused on traditional convolution-based backbones to process and extract image features [5, 11, 36, 49]. Following the success of transformers and their ability to consume large amounts of data [14, 37, 70, 83, 85], several methods have paved the way of utilising transformer architectures scaling up the 3D human body and hand recovery [41, 42, 61]. Recently, Pavlakos *et al*. [61] showcased the effectiveness of vision transformers (ViT) using a simple yet powerful framework trained on a large-scale dataset. The key to the success of this method lies in the scale of its architecture, which is composed of more than 0.5 billion parameters, enabling it to effectively consume large amount of data. However, as shown in the literature [40, 55, 79, 93], regressing the hand parameters from a single image results in bad alignment and incorrect poses. Currently, methods that aim to achieve better image alignment rely on a sub-optimal solutions, such as intermediate heatmap representations [10, 34, 96]. To tackle this, we propose a high-fidelity 3D pose estimation method that decomposes 3D hand reconstruction into two stages. In particular, the decoder first predicts a rough hand estimation that is used to extract multi-scale image-aligned features from our refinement module. By leveraging the rough hand estimation, we can extract meaningful spatial features that lead to better image alignment and state-of-the-art performance on FreiHand [100] and HO3D [24] benchmark datasets. Additionally, in contrast to vertex regression methods [36, 41, 42] that directly regress 3D vertices, our method predicts MANO parameters [71], ensuring both explainable and plausible hand poses.

In this paper, we propose a high fidelity full stack method that can reconstruct 3D hands in real-time. Specifically:

- Based on the limitations of current hand detection benchmark datasets, we collect a large-scale dataset of in-the-wild images that contain multiple hands and introduce a challenging benchmark for hand detection. We make the dataset along with its 2D and 3D publicly available.
- We propose a real-time hand detection method trained on the aforementioned large-scale dataset that outperforms previous hand detection methods by a large margin in both accuracy and efficiency.
- We propose a transformer-based method that facilitates high fidelity 3D reconstructions that tackles the architectural limitations of previous method using an novel refinement module. The proposed method, apart from highly efficient, achieves state-of-the-art performance in Freihand and HO3D benchmark datasets.

2. Related Work

Hand Detection and Tracking. Object detection has been extensively studied in the literature achieving remarkable advancements [44, 47, 67] and setting the foundations for human body [17, 19, 58, 63, 80, 90] and face detection [12, 98] pipelines. In contrast, despite two decades of research efforts [69], hand detection has not yet achieved comparable breakthroughs. Initial approaches used controlled conditions and depth cameras [77, 78, 87, 88, 97] to detect and track human hands. Several efforts have been made to boost hand detection under different skin tones and backgrounds using multi-stage frameworks [48, 62], however, they fail to generalize in challenging environments. Following the success in object detection, several methods have adopted fully convolutional architectures for hand detection [13, 29, 72, 73, 92]. Simon *et al*. [75] introduced a multi-view bootstrapping procedure to annotate in-the-wild data and train a real-time convolutional detector network. Recently, Narasimhaswamy *et al*. [52] proposed an extension of MaskRCNN [28] network to detect in-the-wild hands and identifying their corresponding contact points [53] and body associations [54]. Nevertheless, despite the extensive efforts in the literature, most methods rely on slow backbones and struggle with challenging images. The primary issue is the lack of large-scale training data featuring multiple levels of occlusions and motion blur from in-the-wild scenes. To tackle such limitation, we propose lightweight hand detector that is 45 $\times$ faster compared to previous state-of-the-art detectors, trained on 2M in-the-wild images with diverse environments and occlusion.

3D Hand Pose Estimation. Similar to hand detection, initial approaches for hand pose estimation relied on depth cameras [18, 57, 81] to reconstruct 3D hands. Boukhayma *et al*. [5] introduced the first fully learnable pipeline that directly regresses the parameters of the MANO hand model [71] from RGB images. In a similar manner, several follow-up works used heatmaps [94] and iterative refinement [1] to enforce 2D alignment. Kulon *et al*. [35, 36] introduced an alternative regression method that regresses 3D vertices instead of MANO pose parameters, which significantly outperformed previous methods. Various approaches have been proposed to improve task-specific challenges of 3D pose estimation, including robustness to occlusions [59] and motion blur [56] and reducing inference speed [10, 96]. Recently, Pavlakos *et al*. [61] highlighted the importance of scaling up both the training data and the capacity of the model. Specifically, building on the success of the Vision Transformer (ViT) backbones for body pose estimation [6, 20, 40], they demonstrated that using a simple yet effective large-scale transformer architecture can achieve state-of-the-art performance when trained on a diverse collection of datasets. However, directly regressing MANO parameters from the image in one go may introduce miss-alignments and incorrect poses. To tackle this, we propose a novel refinement layer that deforms hand pose using mesh-aligned multi-scale features.

3. WHIM Dataset

A vital cause behind the lack of high-fidelity hand detection systems lies in the limited amount of in-the-wild datasets with multiple hand annotations. To build a robust hand detection and reconstruction framework, we collected a large-scale dataset with **m**illions of **i**n-the-**w**ild **h**ands (**WHIM**) with diverse poses, illuminations, occlusions, and skin tones.

To collect the proposed dataset, we devised a pipeline to automatically annotate YouTube videos from diverse and challenging in-the-wild scenarios. In particular, we selected more than 1,500 YouTube videos containing hand activities including sign language, cooking, everyday activities, sports, and games with ego- and exo-centric viewpoints, motion blur, different hand scales, and interactions. To accurately detect and annotate the hands on each frame we used a combination of ensemble networks. Firstly, we used VitPose [90] and AlphaPose [17] to detect all humans in the frame and selected the bounding boxes with confidence bigger than 0.65. We then cropped the bounding boxes and fed them to an ensemble hand detection pipeline that consists of MediaPipe [92], OpenPose [7] and ContactHands [52] models. To localize the hand, we used a weighted average between the bounding box positions $\mathbf{b}_i$ of the three detectors d_i, scaled from their corresponding confidence $P(y_i|d_i)$:

$$\hat{y} = \frac{\sum_i P(\mathbf{b}_i|d_i) y_i}{\sum_i P(\mathbf{b}_i|d_i)} \tag{1}$$

where $\mathbf{b}_i$ denotes the estimated hand bounding box.

In addition to the bounding box, we used the estimated 2D landmarks [7, 92], to fit a 3D parametric hand model $\mathcal{M}$ [71]. More specifically, we optimized shape β and pose θ parameters to minimize the re-projection loss $\mathcal{L}_{proj}$ between the regressed $\mathbf{J}_{\mathcal{M}}$ and the estimated landmarks $\hat{\mathbf{J}}_{\mathbf{s}}$:

$$\mathcal{L}_{proj} = ||\mathbf{J}_{\mathcal{M}} - \pi(\hat{\mathbf{J}}_{\mathbf{s}}, K)||_1, \tag{2}$$

where $\pi(\cdot)$ denotes the weak perspective projection transform and K the estimated intrinsic camera matrix.

Given the degrees of freedom of the human hand, optimizing the hand model using joint terms usually results in unnatural poses. To tackle the ambiguities during the optimization process, we followed [76] and included bio-mechanical losses to constrain the optimization. In particular, apart from the re-projection error, we enhanced the fitting process using loss functions that constrain the bone lengths and the angle rotations to feasible ranges, as defined in [76]:

$$\mathcal{L}_{BMC} = \mathcal{L}_{BL} + \mathcal{L}_{A} \tag{3}$$

where $\mathcal{L}_{BL}$ and $\mathcal{L}_{A}$ denote the bone length and the joint angle loss terms, respectively. For additional details of the bio-mechanical constraints, we refer the reader to [76].

Finally, given that the bio-mechanical prior acts mainly on the joint space, we followed [2] and trained a PCA model on ARCTIC dataset [16] acting as a 3D prior and to model the distribution of feasible hand poses. We formulated the prior loss as the reconstruction error of the 3D mesh $\mathbf{X}$ projected and reconstructed from the PCA space $\mathbf{U}$ as:

$$\mathcal{L}_{prior} = ||\mathbf{X} - [(\mathbf{X} - \boldsymbol{\mu})\mathbf{U}^T]\mathbf{U} + \boldsymbol{\mu}||_2, \tag{4}$$

where $\mathbf{U} \in \mathbb{R}^{N \cdot 3 \times d}$ is the eigenvector basis of d components and $\boldsymbol{\mu}$ is the mean mesh. In Fig. 2

4. Method

4.1. Hand Detection and Localization

Over the past years, fully convolutional networks (FCNs) have shown remarkable efficiency in human detection [86] and object detection [68]. Building on their success, we employ an FCN architecture to achieve both accurate and real-time hand localization. Similar to object detection frameworks, given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ our goal is to detect the bounding boxes $\mathbf{B} = \{\mathbf{b}_j \in \mathbb{R}^4 : 0 \leq j \leq n\}$ of the n hands present in the image along with their hand side label $\mathbf{y}_j$. We follow the commonly used one-stage backbone-neck-head formulation, where we built upon the

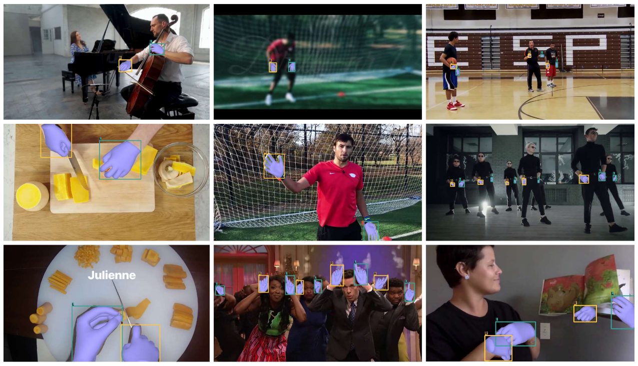

Figure 2. Example of the proposed WHIM in-the-wild dataset.

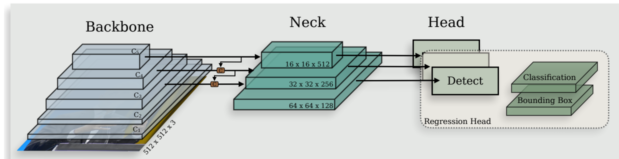

Figure 3. **Detection overview**: The proposed fully convolutional one-stage hand detection method receives an image and extracts multi-resolution feature maps that are then processed by the Path Aggregation Network (PANet). The corresponding features are then fed to three detection heads that predict the hand side, bounding box, and hand joints at different resolutions. We train the network with a multi-task loss for each anchor.

powerful and efficient DarkNet backbone [67]. We extract the last three feature maps $\{C3, C4, C5\}$ of the backbone to generate a multi-scale feature pyramid in the neck module. To enable our model to effectively capture multi-scale features using both top-down and bottom-up pathways across different feature maps, we utilized Path Aggregation Network (PANet) [45], an extension of Feature Pyramid Network [43] that facilitates fine-grained information flow using a bottom-up path augmentation. Finally, we use three detection heads to predict the bounding boxes $\mathbf{b}_j$ and hand side labels $\mathbf{y}_j$ at different anchor resolutions. Following [15], we adopt an anchor-free design to enhance the flexibility of our localization method and directly predict bounding box coordinates without relying on predefined anchor boxes. An overview of the proposed detection network is visualized in Fig. 3. Similar to [12], we observed that joint keypoint supervision significantly improved the performance and the robustness of the detector. The full training objective can be defined as:

$$\mathcal{L} = \lambda_0 \mathcal{L}_{BCE} + \lambda_1 \mathcal{L}_{DFL} + \lambda_2 \mathcal{L}_{CIoU} + \lambda_3 \mathcal{L}_{kpts} \quad (5)$$

where $\mathcal{L}_{BCE}$ is the binary cross entropy loss between the predicted and the ground truth box labels, $\mathcal{L}_{DFL}$ denotes the distributional focal loss [38] which measures the difference between the predicted and the ground truth bounding box distributions, $\mathcal{L}_{CIoU}$ measures the discrepancy between the predicted and the ground truth bounding box [95] and $\lambda_0, \lambda_1, \lambda_2, \lambda_3$ are weights that balance the losses. For additional details about the detection network training, we refer the reader to the supplementary material.

4.2. Hand Reconstruction

Given an image $\mathbf{I_h} \in \mathbb{R}^{H \times W \times 3}$ that contains a human hand, tightly cropped around the hand detectors bounding box, the proposed 3D hand reconstruction method estimates the corresponding hand pose $\theta \in \mathbb{R}^{45}$ and shape $\beta \in \mathbb{R}^{10}$ MANO [71] parameters along with the camera $\mathbf{K}_{cam} = \{\mathbf{t}_{cam}, \mathbf{s}_{cam}\}$ to obtain a 3D hand.

To build a powerful 3D pose estimation network that can scale on large amounts of data, we follow [6, 40, 61], and built our backbone using a pre-trained ViT encoder [85, 90]. The image $\mathbf{I_h}$ is first split into M-size patches $\mathbf{P} \in \mathbb{R}^{\frac{HW}{M^2} \times (M^2 \times 3)}$ and then embedded to high dimensional tokens $\mathbf{T}_{img} \in \mathbb{R}^{\frac{HW}{M^2} \times C}$. To uniquely encode their spatial location, positional embeddings $\mathbf{P}_e$ are added to the image tokens $\mathbf{T}_{img}$ [85]. In addition to the image tokens, we explicitly model hand pose, shape, and camera parameters with three distinct tokens $\mathbf{T}_{pose}, \mathbf{T}_{shape}, \mathbf{T}_{cam}$. We then feed the concatenated tokens to a ViT transformer encoder to obtain a set of updated feature tokens $\mathbf{T'}_{img}, \mathbf{T'}_{pose}, \mathbf{T'}_{shape}, \mathbf{T'}_{cam}$. Using a set of MLP layers we regress a rough estimation of pose θ^c and shape β^c parameters of the MANO model, which will serve as a prior for the refinement network. Similarly, we regress the camera translation and scale parameters $\mathbf{K}_{cam} = \{\mathbf{t}_{cam}, \mathbf{s}_{cam}\}$ from the camera token features.

Multi-Scale Pose Refinement Module. In order to get better image alignment and more accurate hand pose, we introduce a fully differentiable refinement module that predicts pose and shape residuals of the rough hand estimation. To do so we leverage the image features extracted from the ViT backbone to serve as the 2D feature cues in our refinement module. In particular, we reshape image feature tokens $\mathbf{T'}_{img}$ to form a low resolution feature map $\mathbf{F}_0 \in \mathbb{R}^{\frac{H}{M} \times \frac{W}{M} \times C}$ and project the rough hand estimation $\mathcal{M}_l$ to feature map using the estimated $\mathbf{t}_{cam}, \mathbf{s}_{cam}$ camera parameters. Then, using bilinear interpolation we sample from $\mathbf{F}_0$ a feature vector $\mathbf{f}_0^{\mathbf{v}}$ for each projected vertex $\mathbf{v}$:

$$\mathbf{f}_0^{\mathbf{v}} = \pi(\mathbf{v}, \mathbf{K}_{cam}) \quad (6)$$

where $\pi(\cdot)$ denotes the weak perspective projection.

Note that we project the whole hand mesh $\mathcal{M}_l$ to the feature map, instead of just the hand joints, as we aim to acquire better shape and pose image alignment. The image-aligned vertex features are then aggregated to form a global feature vector that is used to regress pose and shape residu-

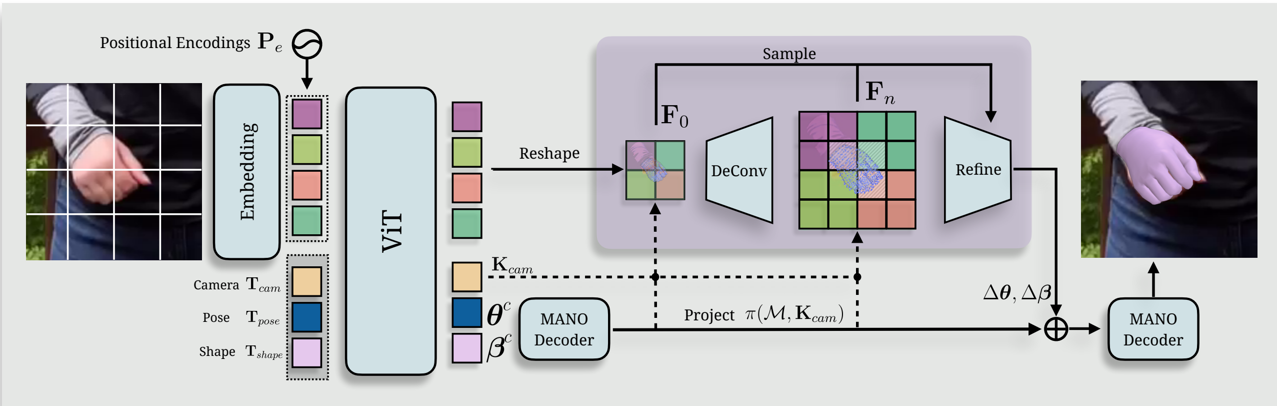

Figure 4. **Overview of the proposed 3D hand pose estimation method**: Given an image $\mathbf{I}_h$ represented as a series of feature tokens $\mathbf{T}_{img}$ along with a set of learnable camera $\mathbf{T}_{cam}$, pose $\mathbf{T}_{pose}$ and shape $\mathbf{T}_{shape}$ tokens, we initially predict a rough estimation of the MANO [71] and camera $\mathbf{K}_{cam}$ parameters using a ViT backbone (light blue). The updated image tokens are then reshaped and upsampled through a series of deconvolutional layers to form a set of multi-resolution feature maps $\{\mathbf{F}_0, ..., \mathbf{F}_0\}$. We then project the estimated 3D hand to the generated feature maps and sample image-aligned multi-scale features through a novel refinement module (purple). The sampled features are used to predict pose and shape residuals $\Delta\theta, \Delta\beta$ that refine the coarse hand estimation. Using this coarse-to-fine pose estimation strategy we facilitate image alignment and achieve better reconstruction performance.

als:

$$\begin{aligned} \Delta\beta &= MLP_\beta(\square_{\mathbf{v}\in\mathcal{M}_l}\mathbf{f}_0^{\mathbf{v}}) \\ \Delta\theta &= MLP_\theta(\square_{\mathbf{v}\in\mathcal{M}_l}\mathbf{f}_0^{\mathbf{v}}) \end{aligned} \tag{7}$$

where $\square$ denotes the aggregation function, *e.g.*, mean, max, sum.

Given that the initial feature map is very low-dimensional, we use a set of deconvolutional layers to upsample $\mathbf{F}_0$ to multiple higher resolution feature maps $\mathbf{F}_0, \mathbf{F}_1, ..., \mathbf{F}_n$ that will serve as multi-scale features for the proposed refinement module. Intuitively, low-dimensional feature maps will provide global and structural residuals of the hand shape while more high-resolution features could provide finer details of the hand pose.

Loss function. The method proposed in this paper is trained with supervision for 3D vertices $\hat{\mathbf{V}}_{3D}$, 2D joints $\hat{\mathbf{J}}_{2D}$ and MANO parameters $\hat{\theta}, \hat{\beta}$, when available. Additionally, following [33, 61] we utilize a discriminator network D to enforce plausible hand poses and shapes and penalize irregular articulations. The full loss function can be defined as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{3D} + \mathcal{L}_{2D} + \mathcal{L}_{mano} + \mathcal{L}_{adv}, \\ \mathcal{L}_{3D} &= ||\mathbf{V}_{3D} - \hat{\mathbf{V}}_{3D}||_1, \\ \mathcal{L}_{2D} &= ||\pi(\mathbf{J}_{3D}, \mathbf{K}_{cam}) - \hat{\mathbf{J}}_{2D}||_1 \\ \mathcal{L}_{mano} &= ||\theta - \hat{\theta}||_2^2 + ||\beta - \hat{\beta}||_2^2, \\ \mathcal{L}_{\text{adv}} &= ||D(\theta, \beta) - 1||_2. \end{aligned} \tag{8}$$

5. Experiments

In this section, we first evaluate the proposed hand detection network using established benchmarks to assess its performance. Subsequently, we conduct an extensive qualitative and quantitative analysis of the proposed 3D hand pose estimation method. Finally, we demonstrate the critical role of precise hand localization in enhancing the accuracy of 4D hand reconstruction.

5.1. Evaluation of Hand Detection and Localization

Training. We train the proposed hand detection network using the curated WHIM dataset that consists of over 2M in-the-wild images of multiple hands and scales. To further boost the generalization and robustness of our network, we follow several data augmentations during training. Particularly, we introduce random rotations in the range of $[-60°, 60°]$ and translations in the range of $[-0.1, 0.1]$ along with random masking and cropping in the image. Additionally, in each training batch, we follow mosaic and mixup augmentation, which significantly affects the robustness to diverse hand scales.

Evaluation. To compare our network, we employ popular baselines such as OpenPose [7] and Mediapipe [92], which are widely used across the community [64, 66], along with more recent hand detection pipelines such as ContactHands [52] and ViTDet [39]. All methods are evaluated under three criteria: i) the inference speed in terms of frames per second (FPS), ii) the detection performance in terms of average precision (AP) at IoU = 0.5 and mean AP at different IoU=0.5:0.05:0.95 thresholds and iii) the model size measured in Mb. An optimal hand detection system should be lightweight to ensure compatibility with mobile devices, operate in real-time to avoid impacting the runtime of a 3D pose estimation pipeline while achieving precise detections. In Tab. 1, we evaluate the proposed and the baseline methods on three datasets: the proposed WHIM dataset along with the benchmark Coco-WholeBody [32] and Oxford-Hands [48] dataset. As can be easily seen, the proposed detector can run in over than 130 FPS while achieving up to 21% improvement on AP compared to previous state-of-the-art models. In addition, compared to previous state-of-the-art model, ContactHands [52], the proposed detector is 45× faster and has 32× reduced model size which enables the utilization of the proposed detector in mobile applications and heavy pipelines without posing any significant overhead. It is important to note that despite the varying resolutions and hand scales across the three datasets, the proposed model consistently outperforms the baseline methods. This is particularly evident on the COCO-WholeBody dataset [32], an extension of the COCO dataset that includes full-body images, where the hands are relatively small compared to the overall image size. **Ablation.** The efficiency and accuracy of the proposed hand detection method are mainly attributed to the selection of the backbone architecture and the utilization of a large-scale training dataset. We further evaluate the contribution of each component using an ablation study on OxfordHands [48] and WHIM datasets where we utilized different backbone networks and training datasets. As can be seen from Tab. 2, trained the detection network with different backbones, apart from achieving similar detection performance, significantly degraded the inference speed of the network. The importance of the proposed large-scale in-the-wild WHIM dataset is also validated in Tab. 2, where we can observe a significant performance drop when the model was trained with significantly less data, *e.g.*, *Proposed w. 0.25M, Proposed w. 0.5M, Proposed w. 1M*. An interesting observation highlighting the versatility of the WHIM dataset is that the proposed model achieves better performance when trained on WHIM compared to the OxfordHands dataset. Finally, we evaluate the contribution of the proposed augmentation strategy and the use of landmark regression loss. The augmentation strategy significantly contributes to cross-dataset generalization, achieving 14% increase on mAP. Similarly, we can observe that incorporating landmark regression loss enhances the detector's precision, leading to more robust detections.

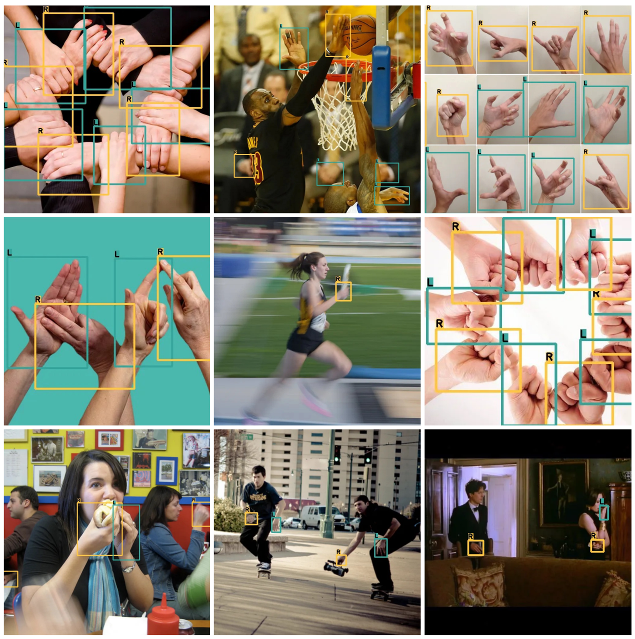

Figure 5. **Qualitative Evaluation** of the proposed hand detection network on in-the-wild images. The proposed model demonstrates robustness across various lighting conditions, resolutions, hand scales, and even in the presence of motion blur.

			Coco-Whole		Oxford-Hands		WHIM	
Method	Size (Mb)	FPS	AP0.5	mAP	AP0.5	mAP	AP0.5	mAP
MediaPipe [92]	25	25	15.43	3.72	8.72	1.80	53.09	12.01
OpenPose [7]	141	29	37.05	9.06	20.74	4.41	76.8	34.25
ContactHands [53]	819	3	50.29	16.67	70.02	36.41	93.42	49.44
ViTDet [39]	1400	1	17.17	3.16	7.49	1.44	38.81	8.12
Proposed-S	**7**(×3.5 ↓)	**175**(×6 ↑)	46.96	18.56	75.21	38.16	91.80	46.50
Proposed-M	25	138	**62.48**	**25.97**	**82.64**	**48.98**	**96.06**	**53.79**

Table 1. **Comparison with the state-of-the-art hand detection methods on COCO-Whole [32], Oxford-Hands [48] and the proposed WHIM dataset.** For each method we report the average precision (AP) at IoU=0.5 along with the mean average precision (mAP). We also compare the performance of each method in terms of model size, measured in Mb, and speed, measured in frames per second (FPS).

			OxfordHands		WHIM	
Method	Size (Mb)	FPS	AP0.5	mAP	AP0.5	mAP
Proposed-w. 0.25M	-	-	49.15	26.69	75.75	38.48
Proposed-w. 0.5M	-	-	58.32	35.15	83.03	42.11
Proposed-w. 1M	-	-	69.21	43.04	88.37	47.92
Proposed-w.OxfordHands	-	-	68.15	40.34	70.14	35.29
Proposed-w. ResNet50	118	34	74.13	47.34	95.43	51.85
Proposed-w. HRNet	132	30	**84.82**	**49.83**	**97.23**	**54.12**
Proposed-w/o Augmentations	-	-	70.76	42.17	91.13	49.93
Proposed-w/o Landmark Loss	-	-	72.57	45.96	92.43	51.44
Proposed-M	**25** (↓ 5×)	**138** (↑ 4×)	82.64	48.98	96.06	53.79

Table 2. **Ablation study**: Evaluation of individual components in the proposed detection pipeline on OxfordHands and WHIM datasets. We use – to denote identical network architecture and performance.

5.2. 3D Hand Pose Estimation

Training. Following [6, 40, 61], we trained the proposed hand regressor using a combination of datasets to improve robustness to diverse poses, illuminations and occlusions. Particularly, we utilized a set of datasets containing both 2D and 3D annotations namely FreiHAND [100], HO3D [24],

Method	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↑
I2L-MeshNet [49]	7.4	7.6	0.681	0.973
Pose2Mesh [11]	7.7	7.8	0.674	0.969
I2UV-HandNet [9]	6.7	6.9	0.707	0.977
METRO [41]	6.5	6.3	0.731	0.984
Tang *et al.* [82]	6.7	6.7	0.724	0.981
Mesh Graphormer [42]	5.9	6.0	0.764	0.986
MobRecon [10]	5.7	5.8	0.784	0.986
AMVUR [31]	6.2	6.1	0.767	0.987
HaMer [61]	6.0	5.7	0.785	0.990
Proposed	**5.5**	**5.1**	**0.825**	**0.993**

Table 3. **Comparison with the state-of-the-art on the FreiHAND dataset [100].** 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.* [46]	0.803	9.9	0.810	9.5	0.528	0.956
HandOccNet [60]	0.819	9.1	0.819	8.8	0.564	0.963
I2UV-HandNet [9]	0.804	9.9	0.799	10.1	0.500	0.943
Hampali *et al.* [24]	0.788	10.7	0.790	10.6	0.506	0.942
Hasson *et al.* [27]	0.780	11.0	0.777	11.2	0.464	0.939
ArtiBoost [91]	0.773	11.4	0.782	10.9	0.488	0.944
Pose2Mesh [11]	0.754	12.5	0.749	12.7	0.441	0.909
I2L-MeshNet [49]	0.775	11.2	0.722	13.9	0.409	0.932
METRO [41]	0.792	10.4	0.779	11.1	0.484	0.946
MobRecon[10]	-	9.2	-	9.4	0.538	0.957
Keypoint Trans [25]	0.786	10.8	-	-	-	-
AMVUR [31]	0.835	8.3	0.836	8.2	0.608	0.965
Hamer	0.846	7.7	0.841	7.9	0.635	0.980
Proposed	**0.851**	**7.5**	**0.846**	**7.7**	**0.646**	**0.983**

Table 4. **Comparison with the state-of-the-art on the HO3D dataset [24].** 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.

Method	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↑
Proposed w. FastViT	6.5	6.3	0.741	0.967
Proposed w/o ViTPose	5.9	5.7	0.795	0.989
Proposed w. Single-Scale	6.0	5.9	0.793	0.991
Proposed w/o Refinement	6.1	5.8	0.795	0.991
Proposed w. FreiHAND [100]	6.1	5.8	0.793	0.990
Proposed w. Datasets [61]	5.9	5.7	0.805	0.992
Proposed Full	**5.5**	**5.1**	**0.825**	**0.993**

Table 5. **Ablation study on the FreiHAND dataset [100].** We report ablations on the backbone and the training data used along with the novel refinement module.

MTC [89], RHD [99], InterHand2.6M [50], H2O3D [24], DEX YCB [8], COCO WholeBody [32], Halpe [17] MPII NZSL [75], BEDLAM [4], ARCTIC [16], Re:InterHand [51] and Hot3D [3]. In total we utilized 4.2M images, 55% more than previous state-of-the-art. Including egocentric datasets such as ARCTIC [16] and Hot3D [3] significantly improve the reconstruction performance of the proposed method on egocentric views.

Evaluation. We compare the proposed method with state-of-the-art methods including METRO [41], Mesh Graphormer [42], AMVUR [31], MobRecon [10], HaMeR [61] and SimpleHand [96]. In Tab. 3 and Tab. 4 we report the reconstruction results the popular benchmark FreiHAND [100] and HO3Dv2 [24] datasets. Following the common protocol [100], we measure the reconstruction performance in terms of Procrustes Aligned Mean per Joint and Vertex Error (PA-MPJPE, PA-MPVPE) along with the fraction of poses with less than 5mm and 15mm error (F@5, F@15). Additionally, we report Area Under the Curve for 3D joints and vertices (AUC_J, AUC_V) for HO3D dataset. As can be observed the proposed method achieves state-of-the-art performance and outperforms previous methods under all metrics on both benchmark datasets.

Ablation. To further investigate the contributions of each component of the proposed method we conducted an ablation study. In Tab. 5, we assess the contribution of the backbone architecture, the training datasets used along with the refinement module. As can be observed, swapping the ViT backbone with the recent efficient FastViT [84] architecture *(Proposed w. FastViT)* results in significant degradation of performance despite the runtime efficiency. Similarly, training the backbone from scratch without using the pre-trained weights of ViTPose [90] *(Proposed w/o ViTPose)* also results in a performance drop. To evaluate the effect of the proposed refinement module we trained a model that directly regresses the MANO and camera parameters from the ViT output tokens without using any refinement module *(Proposed w/o Refinement)*. Additionally, we trained a model with a single-scale refinement module that samples features from a single feature map *(Proposed w. Single-Scale)*. Both architectural choices deteriorate the reconstruction performance of the proposed model which highlights the effect of the proposed multi-scale refinement module. Finally, we examine the effect of the large-scale training set by training two derivatives of the proposed model using only the FreiHAND dataset [100] *(Proposed w. FreiHAND [100])*, similar to [10, 41, 42, 96] and a model trained on the datasets used in [61] *(Proposed w. Datasets [61])*.

5.3. Dynamic Reconstruction

A key challenge for 3D pose estimation methods is to achieve stable and robust 4D reconstructions without being trained on a dynamic setting [74]. Traditionally, methods for 3D pose estimation from single image suffer from low temporal coherence and jittering effects across frames and can not generalize well to videos, setting a huge burden on their generalization to real-world reconstruction. To effectively evaluate the temporal coherence of the proposed method, we reconstruct frame-wise the 4D sequence and

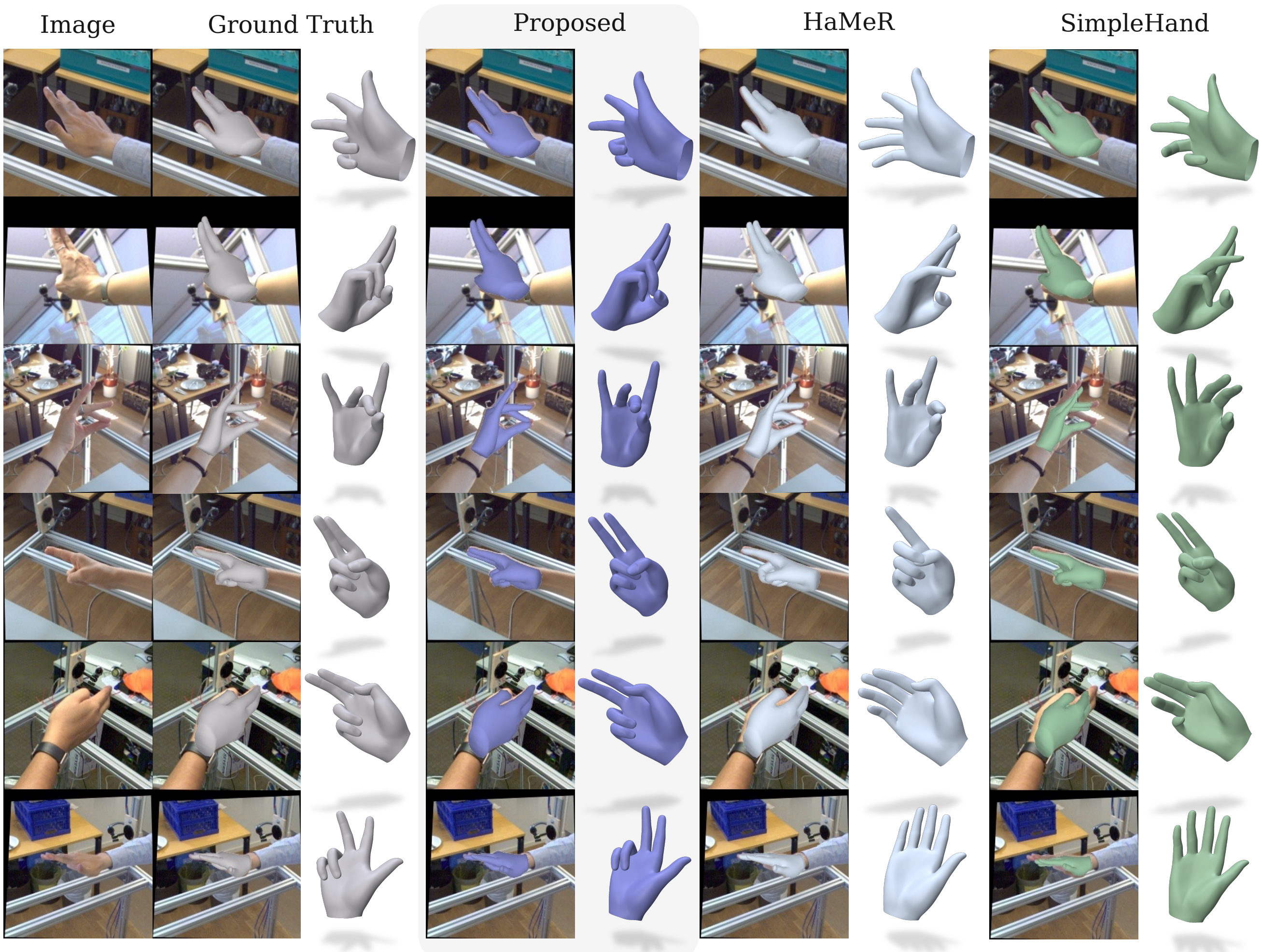

Figure 6. **Qualitative Evaluation** of the proposed hand detection network on in-the-wild images. The proposed model demonstrates robustness across various lighting conditions, resolutions, hand scales, and even in the presence of motion blur.

measure the jittering between frames. In particular, we calculate the mean per frame Euclidean distance of the 3D vertices (MPFVE) and joints (MPFJE) between consecutive frames. Additionally, similar to [74], we measure the jerk (Jitter) of the 3D hand joints motion along with the global Root Translation Error (RTE) that measures the displacement of the wrist across frames. In Tab. 6, we report the reconstruction results fro the best performing methods on HO3D [24] dataset. WiLoR outperforms baseline methods in temporal coherence without relying on any temporal module based on the robust stability of the detections.

Method	MPFVE ($\times 100$) ↓	MPFJE ($\times 100$) ↓	Jitter ↓	RTE ↓
MeshGraphormer [42]	21.86	4.99	41.16	7.92
MobRecon [10]	22.18	6.09	40.25	8.03
SimpleHand [96]	19.72	5.12	38.53	6.04
HaMeR [61]	10.60	1.768	20.43	2.92
Proposed	**4.43**	**0.762**	**5.92**	**0.07**

Table 6. **Reconstruction of dynamic 3D Hands.** We evaluate the temporal coherence and the jittering of the reconstruction for the proposed and the baseline methods on the HO3D dataset. We include for reference the ground truth values.

6. Conclusion

In this work, we propose the first full-stack hand detection and 3D pose estimation framework. Using a large-scale in-the-wild dataset we train a light-weight yet highly accurate hand detector model that can robustly detect hands under different occlusions and illuminations at over 110 FPS. Additionally, we propose a high fidelity 3D hand pose estimation model build on top of our novel refinement module, that overcomes the limitations of previous methods and mitigates the alignment issues of previous methods. Under a series of experiments, we showcase that the proposed method can outperform previous state-of-the-art on two benchmark

datasets and show robust performance on challenging cases.
Acknowledgements S. Zafeiriou was supported by EPSRC Project DEFORM (EP/S010203/1) and GNOMON (EP/X011364). R.A. Potamias was supported by EPSRC Project GNOMON (EP/X011364).

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 3

[2] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1985–1995, 2024. 1, 3

[3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 7

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 7

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2, 3

[6] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 4, 6

[7] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 2, 3, 5, 6

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 7

[9] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12929–12938, 2021. 7

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 2, 3, 7, 8

[11] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 769–787. Springer, 2020. 1, 2, 7

[12] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 2, 4

[13] 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*, 27(4):1888–1900, 2017. 2

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[15] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 4

[16] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 7

[17] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3, 7

[18] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3593–3601, 2016. 3

[19] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3582–3589, 2014. 2

[20] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 3

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

In *International Conference on Computer Vision (ICCV)*, 2023. 2

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[23] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 2

[24] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 1, 2, 6, 7, 8

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 7

[26] 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. 1

[27] 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/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 7

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[29] T Hoang Ngan Le, Yutong Zheng, Chenchen Zhu, Khoa Luu, and Marios Savvides. Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pages 46–53, 2016. 2

[30] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11087–11096, 2021. 1

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 758–767, 2023. 7

[32] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 6, 7

[33] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 5

[34] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, and Wonjun Kim. Sampling is Matter: Point-guided 3d human mesh reconstruction. In *CVPR*, 2023. 2

[35] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. *arXiv preprint arXiv:1905.01326*, 2019. 3

[36] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 2, 3

[37] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021. 2

[38] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 4

[39] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 2, 5, 6

[40] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 2, 3, 4, 6

[41] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 2, 7

[42] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 1, 2, 7, 8

[43] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[44] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[45] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8759–8768, 2018. 4

[46] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 7

[47] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 2

[48] Arpit Mittal, Andrew Zisserman, and Philip HS Torr. Hand detection using multiple proposals. In *Bmvc*, page 5. Citeseer, 2011. 2, 6

[49] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 1, 2, 7

[50] 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 *European Conference on Computer Vision (ECCV)*, 2020. 7

[51] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mize Mallorie, Philippe Bree, Tomas Simon, Bo Peng, Shubham Garg, Kevyn McPhail, and Takaaki Shiratori. A dataset of relighted 3D interacting hands. In *NeurIPS Track on Datasets and Benchmarks*, 2023. 7

[52] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9567–9576, 2019. 2, 3, 5, 6

[53] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. *Advances in neural information processing systems*, 33:7841–7851, 2020. 2, 6

[54] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4889–4899, 2022. 2

[55] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. 2

[56] Yeounguk 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[57] Iason Oikonomidis, Nikolaos Kyriazis, Antonis A Argyros, et al. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, page 3, 2011. 3

[58] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017. 2

[59] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[60] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 7

[61] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[62] 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*, 101:403–419, 2013. 2

[63] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012. 2

[64] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4670–4680, 2023. 1, 5

[65] Jing Qi, Li Ma, Zhenchao Cui, and Yushu Yu. Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems*, 10(1): 1581–1606, 2024. 1

[66] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 54–71. Springer, 2020. 5

[67] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2, 4

[68] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 3

[69] James M Rehg and Takeo Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In *Proceedings of 1994 IEEE workshop on motion of non-rigid and articulated objects*, pages 16–22. IEEE, 1994. 2

[70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[71] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2, 3, 4, 5

[72] Kankana Roy, Aparna Mohanty, and Rajiv R Sahay. Deep learning based hand detection in cluttered environment using skin segmentation. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 640–649, 2017. 2

[73] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2

[74] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 7, 8

[75] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 2, 7

[76] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 3

[77] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013. 2

[78] Björn Stenger, Arasanathan Thayananthan, Philip HS Torr, and Roberto Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1372–1384, 2006. 2

[79] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[80] Min Sun and Silvio Savarese. Articulated part-based model for joint object detection and pose estimation. In *2011 International Conference on Computer Vision*, pages 723–730. IEEE, 2011. 2

[81] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer graphics forum*, pages 101–114. Wiley Online Library, 2015. 3

[82] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11698–11707, 2021. 7

[83] Michail Tarasiou, Rolandos Alexandros Potamias, Eimear O'Sullivan, Stylianos Ploumpis, and Stefanos Zafeiriou. Locally adaptive neural 3d morphable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1867–1876, 2024. 2

[84] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5785–5795, 2023. 7

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[86] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 3

[87] Robert Y Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, 28(3):1–8, 2009. 2

[88] 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 Asian Conference on Computer Vision*, pages 1106–1111, 2000. 2

[89] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 7

[90] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 2, 3, 4, 7

[91] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2750–2760, 2022. 7

[92] 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. 2, 3, 5, 6

[93] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from

monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023. 2

[94] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 3

[95] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics*, 52(8):8574–8586, 2021. 4

[96] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2024. 2, 3, 7, 8

[97] Xiaojin Zhu, Jie Yang, and Alex Waibel. Segmenting hands of arbitrary color. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 446–453. IEEE, 2000. 2

[98] Yanjia Zhu, Hongxiang Cai, Shuhan Zhang, Chenhao Wang, and Yichao Xiong. Tinaface: Strong but simple baseline for face detection. *arXiv preprint arXiv:2011.13183*, 2020. 2

[99] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 7

[100] 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 IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2, 6, 7