Efficient Annotation and Learning for 3D Hand Pose Estimation: A Survey

Takehiko Ohkawa[1] · Ryosuke Furuta[1] · Yoichi Sato[1]

Abstract

In this survey, we present a systematic review of 3D hand pose estimation from the perspective of efficient annotation and learning. 3D hand pose estimation has been an important research area owing to its potential to enable various applications, such as video understanding, AR/VR, and robotics. However, the performance of models is tied to the quality and quantity of annotated 3D hand poses. Under the status quo, acquiring such annotated 3D hand poses is challenging, e.g., due to the difficulty of 3D annotation and the presence of occlusion. To reveal this problem, we review the pros and cons of existing annotation methods classified as manual, synthetic-model-based, hand-sensor-based, and computational approaches. Additionally, we examine methods for learning 3D hand poses when annotated data are scarce, including self-supervised pretraining, semi-supervised learning, and domain adaptation. Based on the study of efficient annotation and learning, we further discuss limitations and possible future directions in this field.

1 Introduction

The acquisition of 3D hand pose annotations[1] has presented a significant challenge in the study of 3D hand pose estimation. This makes it difficult to construct large training datasets and develop models for various target applications, such as hand-object interaction analysis (Boukhayma et al., 2019; Hampali et al., 2020), pose-based action recognition (Iqbal et al., 2017; Tekin et al., 2019; Sener et al., 2022), augmented and virtual reality (Liang et al., 2015; Han et al., 2022; Wu et al., 2020), and robot learning from human demonstration (Ciocarlie & Allen, 2009; Handa et al., 2020; Qin et al.,

---

[1] We denote 3D pose as the 3D keypoint coordinates of hand joints, $P^{3D} \in \mathbb{R}^{J \times 3}$ where J is the number of joints.
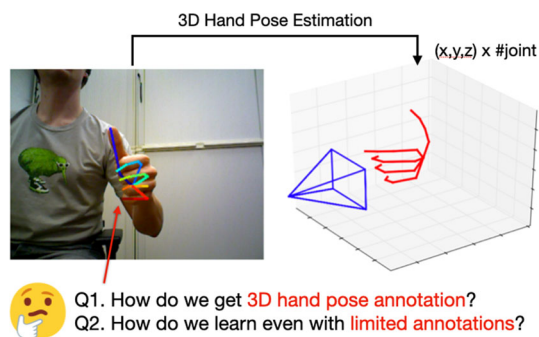
✉ Takehiko Ohkawa
  ohkawa-t@iis.u-tokyo.ac.jp

  Ryosuke Furuta
  furuta@iis.u-tokyo.ac.jp

  Yoichi Sato
  ysato@iis.u-tokyo.ac.jp

[1] Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

2022; Mandikal & Grauman, 2021). In these application scenarios, we must consider methods for annotating hand data, and select an appropriate learning method according to the amount and quality of the annotations. However, there is currently no established methodology that can give annotations efficiently and learn even from imperfect annotations. This motivates us to review methods for building training datasets and developing models in the presence of these challenges in the annotation process.

During the annotations, we encounter several obstacles including the difficulty of 3D measurement, occlusion, and dataset bias. As for the first obstacle, annotating 3D points from a single RGB image is an ill-posed problem. While annotation methods using hand markers, depth sensors, or multi-view cameras can provide 3D positional labels, these setups require a controlled environment, which limits available scenarios. As for the second obstacle, occlusion hinders annotators from accurately localizing the positions of hand joints. As for the third obstacle, annotated data are biased to a specific condition constrained by the annotation method. For instance, annotation methods based on hand markers or multi-view setups are usually installed in laboratory settings, resulting in a bias toward a limited variety of backgrounds and interacting objects.

Fig. 1 Our survey on 3D hand pose estimation is organized from two aspects: (i) obtaining 3D hand pose annotation and (ii) learning even with a limited amount of annotated data. These two issues will be considered in the scenarios of practical applications where we work on dataset construction and model development with limited resources. The figure is adapted from Zimmermann and Brox (2017)

Given such challenges in annotation, we conduct a systematic review of the literature on 3D hand pose estimation from two distinct perspectives: *efficient annotation* and *efficient learning* (see Fig. 1). The former view highlights how existing methods assign reasonable annotations in a cost-effective way, covering a range of topics: the availability and quality of annotations and the limitations when deploying the annotation methods. The latter view focuses on how models can be developed in scenarios where annotation setups cannot be implemented or available annotations are insufficient.

In contrast to existing surveys on network architecture and modeling (Chatzis et al., 2020; Doosti, 2019; Le & Nguyen, 2020; Lepetit, 2020; Liu et al., 2021), our survey delves into another fundamental direction that arises from the annotation issues, namely, dataset construction with cost-effective annotation and model development with limited resources. In particular, our survey includes benchmarks, datasets, image capture setups, automatic annotation, learning with limited labels, and transfer learning. Finally, we discuss potential future directions of this field beyond the current state of the art.

For the study of annotation, we categorize existing methods into manual (Chao et al., 2021; Mueller et al., 2017; Sridhar et al., 2016), synthetic-model-based (Chen et al., 2021; Hasson et al., 2019; Mueller et al., 2017, 2018; Zimmermann & Brox, 2017), hand-marker-based (Garcia-Hernando et al., 2018; Taheri et al., 2020; Yuan et al., 2017), and computational approaches (Hampali et al., 2020; Kulon et al., 2020; Kwon et al., 2021; Moon et al., 2020; Simon et al., 2017; Zimmermann et al., 2019). While manual annotation requires querying human annotators, hand markers automate the annotation process by tracking sensors attached to a hand. Synthetic methods utilize computer graphics engines to render plausible hand images with precise keypoint coordinates. Computational methods assign labels by fitting a hand template model to the observed data or using multi-

view geometry. We find these annotation methods have their own constraints, such as the necessity of human effort, the sim-to-real gap, the changes in hand appearance, and the limited portability of the camera setups. Thus, these annotation methods may not always be adopted for every application.

Due to the problems and constraints of each annotation method, we need to consider how to develop models even when we do not have enough annotations. Therefore, learning with a small amount of labels is another important topic. For learning from limited annotated data, leveraging a large pool of unlabeled hand images as well as labeled images is a primary interest, e.g., in self-supervised pretraining, semi-supervised learning, and domain adaptation. Self-supervised pretraining encourages the hand pose estimator to learn from unlabeled hand images, so it enables building a strong feature extractor before performing supervised learning. While semi-supervised learning trains the estimator with labeled and unlabeled hand images collected from the same environment, domain adaptation further solves the so-called problem of domain gap between the two image sets, e.g., the difference between synthetic data and real data.

The rest of this survey is organized as follows. In Sect. 2, we introduce the formulation and modeling of 3D hand pose estimation. In Sect. 3, we present open challenges in the construction of hand pose datasets involving depth measurement, occlusion, and dataset bias. In Sect. 4, we cover existing methods of 3D hand pose annotation, namely manual, synthetic-model-based, hand-marker-based, and computational approaches. In Sect. 5, we provide learning methods from a limited amount of annotated data, namely self-supervised pretraining, semi-supervised learning, and domain adaptation. In Sect. 6, we finally show promising future directions of 3D hand pose estimation.

2 Overview of 3D Hand Pose Estimation

Task setting. As shown in Fig. 2, 3D hand pose estimation is typically formulated as the estimation from a monocular RGB/depth image (Erol et al., 2007; Supancic et al., 2018; Yuan et al., 2018). The output is parameterized by the hand joint positions with 14, 16, or 21 keypoints, which are introduced in Tompson et al. (2014), Tang et al. (2014), and Qian et al. (2014), respectively. The dense representation of 21 hand joints[2] has been popularly used as it contains more precise information about hand structure. For a single RGB image in which depth and scale are ambiguous, the 3D coordinates of the hand joint relative to the hand root are estimated from a scale-normalized hand image (Cai et al., 2018; Ge et al., 2019; Zimmermann & Brox, 2017). Recent works additionally estimate hand shape by regressing 3D hand pose and

---

[2] Five end keypoints are fingertips, not strictly called joints.
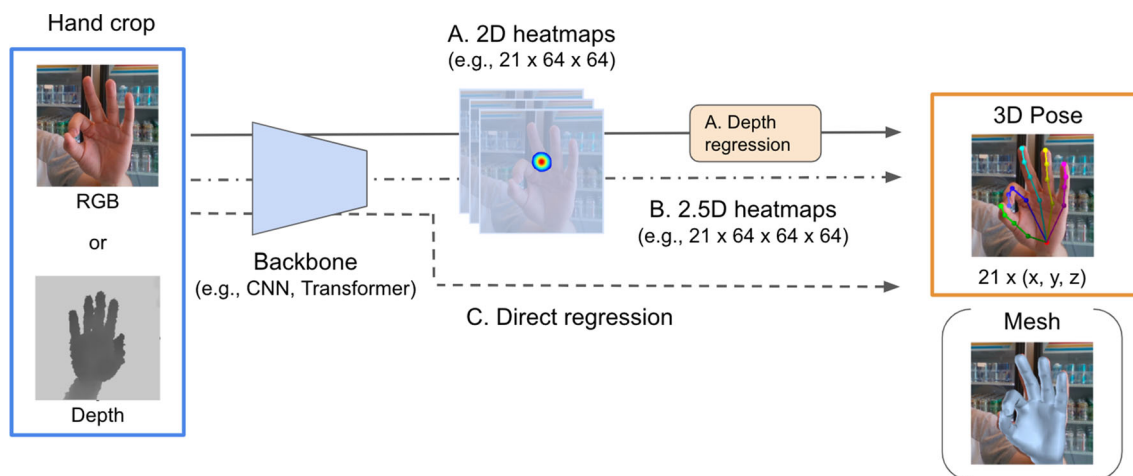
Fig. 2 Formulation and modeling of single-view 3D hand pose estimation. For input, we use either RGB or depth images cropped to the hand region. The model learns to produce a 3D hand pose defined by 3D coordinates. Some works additionally estimate hand shape using a 3D hand template model. For modeling, there are three major designs; **A** 2D heatmap regression and depth regression, **B** extended three-dimensional heatmap regression called 2.5D heatmaps, and **C** direct regression of 3D coordinates

shape parameters together (Boukhayma et al., 2019; Ge et al., 2019; Mueller et al., 2019; Zhou et al., 2016). In evaluation, produced prediction is compared with ground truth, e.g., in the space of world or image coordinates. These two metrics are often used: mean per joint position error (MPJPE) in millimeters, and area under curve of percentage of correct keypoints (PCK-AUC).

Modeling. Classic methods estimate a hand pose by finding the closest sample from a large set of hand poses, e.g., synthetic hand pose sets. Some works formulate the task as nearest neighbor search (Rogez et al., 2015; Romero et al., 2010) while others solve pose classification given predefined hand pose classes and a SVM classifier (Rogez et al., 2014, 2015; Sridhar et al., 2013).

Recent studies have adopted an end-to-end training manner where models learn the correspondence between the input image and its label of the 3D hand pose. Standard single-view methods from an RGB image (Cai et al., 2018; Ge et al., 2019; Zimmermann & Brox, 2017) consist of (A) the estimation of 2D hand poses by heatmap regression and depth regression for each 2D keypoint (see Fig. 2). The 2D keypoints are learned by optimizing heatmaps centered on each 2D hand joint position. An additional regression network predicts the depth distance of detected 2D hand keypoints. Other works use (B) extended 2.5D heatmap regression with a depth-wise heatmap in addition to the 2D heatmaps (Iqbal et al., 2018; Moon et al., 2020), so it does not require a depth regression branch. Depth-based hand pose estimation also utilizes such heatmap regression (Huang et al., 2020; Ren et al., 2019; Xiong et al., 2019). Instead of the heatmap training, other methods learn to (C) directly regress keypoint coordinates (Santavas et al., 2021; Spurr et al., 2018).

For the architecture of the backbone network, CNNs [e.g., ResNet (He et al., 2016)] are a basic choice while recent Transformer-based methods have been proposed (Hampali et al., 2022; Huang et al., 2020). To generate feasible hand poses, regularization is a key trick in correcting predicted 3D hand poses. Based on the anatomical study of hands, biomechanical constraints are imposed to limit predicted bone lengths and joint angles (Spurr et al., 2020; Chen et al., 2021; Liu et al., 2021).

3 Challenges in Dataset Construction

Task formulation and algorithms for estimating 3D hand poses are outlined in Sect. 2. During training, it is necessary to build a large amount of training data with diverse hand poses, viewpoints, and backgrounds. However, obtaining such massive hand data with accurate annotations has been challenging for the following reasons.

Difficulty of 3D annotation. Annotating the 3D position of hand joints from a single RGB image is inherently impossible without any prior information or additional sensors due to an ill-posed condition. To assign accurate hand pose labels, hand-marker-based annotation using magnetic sensors (Garcia-Hernando et al., 2018; Wetzler et al., 2015; Yuan et al., 2017), motion capture systems (Miyata et al., 2004; Schröder et al., 2015; Taheri et al., 2020), or hand gloves (Bianchi et al., 2013; Glauser et al., 2019; Wang & Popovic, 2009) has been studied. These sensors can provide 6-DoF information (i.e., location and orientation) of attached markers and enable us to calculate the coordinates of full hand joints from the tracked markers. However, their setups are

(a) Articulation (b) Viewpoint (c) Object
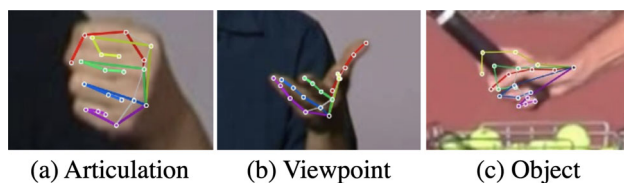
Fig. 3 Difficulty of hand pose annotation in a single RGB image (Simon et al., 2017). Occlusion of hand joints is caused by **a** articulation, **b** viewpoint bias, and **c** grasping objects

Synthetic model Hand marker In-lab setup

Fig. 4 Example of major data collection setups. The synthetic image on the left (ObMan (Hasson et al., 2019)) can be generated inexpensively, but they exhibit unrealistic hand texture. The hand markers on the middle (FPHA (Garcia-Hernando et al., 2018)) enable automatic tracking of hand joints, although the markers distort the appearance of hands. The in-lab setup on the right (DexYCB (Chao et al., 2021)) uses a black background to make it easier to recognize hands and objects, but it limits data variation in environments

expensive and need good calibration, which constrains available scenarios.

On the contrary, depth sensors (e.g., RealSense) or multi-view camera studios (Chao et al., 2021; Hampali et al., 2020; Moon et al., 2020; Simon et al., 2017; Zimmermann et al., 2019) make it possible to obtain depth information near hand regions. Given 2D keypoints for an image, these setups enable annotation of 3D hand poses by measuring the depth distance at each 2D keypoint. However, these annotation methods do not always produce satisfactory 3D annotations, e.g., due to an occlusion problem (detailed in the next section). In addition, depth images are significantly affected by the sensor noise, such as unknown depth values in some regions and ghost shadows around object boundaries (Xu & Cheng, 2013). Due to the limited depth distance that depth cameras can capture, the depth measurement becomes inaccurate when the hands are far from the sensor.

Occlusion. Hand images often contain complex occlusions that distract human annotators from localizing hand keypoints. Examples of possible occlusions are shown in Fig. 3. In figure (a), articulation causes a self-occlusion that makes some hand joints (e.g., fingertips) invisible due to the overlap with the other parts of the hand. In figure (b), such self-occlusion depends on a specific camera viewpoint. In figure (c), hand-held objects induce occlusion that hides the hand joints by the object during the interaction.

To address this issue, hand-marker-based tracking (Garcia-Hernando et al., 2018; Taheri et al., 2020; Wetzler et al., 2015; Yuan et al., 2017) and multi-view camera studios (Chao et al., 2021; Hampali et al., 2020; Moon et al., 2020; Simon et al., 2017; Zimmermann et al., 2019) have been studied. The hand markers offer 6-DoF information during these occlusions, so the hand-maker-based annotation is robust to the occlusion. For multi-camera settings, the effect of occlusion can be reduced when many cameras are densely arranged.

Dataset bias. While hands are a common entity in various image capture settings, the category of objects, including hand-held objects (i.e., foregrounds) and backgrounds, is potentially diverse. In order to improve the generalization ability of hand pose estimators, hand images must be annotated under various imaging conditions (e.g., lighting, viewpoints, hand poses, and backgrounds). However, it is challenging to create such large and diverse datasets nowa-

days due to the aforementioned problems. Rather, existing hand pose datasets exhibit a bias to a particular imaging condition constrained by the annotation method.

As shown in Fig. 4, generating data using synthetic models (Chen et al., 2021; Hasson et al., 2019; Mueller et al., 2017, 2018; Zimmermann & Brox, 2017) is cost-effective, but it creates unrealistic hand texture (Ohkawa et al., 2021). Although the hand-marker-based annotation (Garcia-Hernando et al., 2018; Taheri et al., 2020; Wetzler et al., 2015; Yuan et al., 2017) can automatically track the hand joints from the information of hand sensors, the sensors distort the hand appearance and hinder the natural hand movement. In-lab data acquired by multi-camera setups (Chao et al., 2021; Hampali et al., 2020; Moon et al., 2020; Simon et al., 2017; Zimmermann et al., 2019) make the annotation easier because they can reduce the occlusion effect. However, the variations in environments (e.g., backgrounds and interacting objects) are limited because the setups are not easily portable.

4 Annotation Methods

Given the above challenges concerning the construction of hand pose datasets, we review existing 3D hand pose datasets in terms of annotation design. As shown in Table 1, we categorize the annotation methods as manual, synthetic-model-based, hand-marker-based, and computational approaches. We then study the pros and cons of each annotation method in Table 2.

4.1 Manual Annotation

MSRA (Qian et al., 2014), Dexter+Object (Sridhar et al., 2016), and EgoDexter (Mueller et al., 2017) manually annotate 2D hand keypoints on the depth images and determine the depth distance from the depth value of the images on the 2D point. This method enables assigning reasonable annota-

Table 1 Taxonomy of methods for annotating 3D hand poses

Annotation	Dataset	Year	Modality	Resolution	#Frame	#Subj./#Obj.	#View	Motion	Obj. pose	Sensor/engine
Manual	MSRA (Qian et al., 2014)	2014	Depth	320×240	2K	6/–	1 (3rd)	×	×	Creative Senz3D
	Dexter+Object (Sridhar et al., 2016)	2016	RGB-D	640×480	3K	2/2	1 (3rd)	✓	✓	Creative Senz3D
	EgoDexter (Mueller et al., 2017)	2017	RGB-D	640×480	3K	4/–	1 (1st)	✓	×	RealSense SR300
Synthetic	SynthHands (Mueller et al., 2017)	2017	RGB-D	640×480	220K	–/7	5 (1st)	×	×	Unity
	RHD (Zimmermann & Brox, 2017)	2017	RGB	320×320	43K	20/–	1 (3rd)	×	×	Unity
	GANerated (Mueller et al., 2018)	2018	RGB	256×256	331K	–/–	1 (3rd)	×	×	Unity
	ObMan (Hasson et al., 2019)	2019	RGB-D	256×256	154K	20/3K	1 (3rd)	×	✓	Blender/GraspIt (Miller & Allen, 2005)
	MVHM (Chen et al., 2021)	2021	RGB-D	256×256	320K	–/–	8 (3rd)	×	×	Blender
Hand marker	BigHand2.2M (Yuan et al., 2017)	2017	Depth	640×480	2200K	10/–	1 (1st + 3rd)	✓	×	NDI trakSTAR
	FPHA (Garcia-Hernando et al., 2018)	2018	RGB-D	640×480	105K	6/4	1 (1st)	✓	✓	NDI trakSTAR
	GRAB (Taheri et al., 2020)	2020	MoCap	–	1624K	10/51	–	✓	✓	VICON Vantage 16
Computational Model fitting	ICVL (Tang et al., 2014)	2014	Depth	320×240	180K	10/–	1 (3rd)	×	×	Creative Senz3D
	NYU (Tompson et al., 2014)	2014	RGB-D	640×480	81K	2/–	1 (3rd)	×	×	PrimeSense
	FreiHAND (Zimmermann et al., 2019)	2019	RGB	224×224	37K	32/27	8 (3rd)	×	×	Basler ace
	YouTube3DHands (Kulon et al., 2020)	2020	RGB	640×480	47K	(109)/–	1 (3rd)	✓	×	–
	HO-3D (Hampali et al., 2020)	2020	RGB-D	640×480	103K	10/10	1–5 (3rd)	✓	✓	
	DexYCB (Chao et al., 2021)	2021	RGB-D	640×480	582K	10/20	8 (3rd)	✓	✓	RealSense D415
	H2O (Kwon et al., 2021)	2021	RGB-D	1280×720	571K	4/10	5 (1st & 3rd)	✓	✓	Azure Kinect
Triangulation	Panoptic Studio (Simon et al., 2017)	2017	RGB	1920×1080	15K	–/–	31 (3rd)	✓	×	HD camera
	InterHand2.6M (Moon et al., 2020)	2020	RGB	512×334	2590K	27/–	80–140 (3rd)	✓	×	HD camera
	AssemblyHands (Ohkawa et al., 2023)	2023	RGB/Mono	636×480	3030K	34/–	12 (1st & 3rd)	✓	×	HD camera/Meta Quest

We categorize the annotation methods as manual, synthetic-model-based, hand-marker-based, and computational annotation

Table 2 Pros and cons of each annotation approach

Annotation	Pros	Cons
Manual	Reasonable accuracy	Labor intensive
		Hard to address occlusion
Synthetic	Large scale	Sim-to-real gap
	High diveristy	Hard to simulate motion
	Low cost	
Hand marker	Robust to occlusion	Requires special sensors
	Low annotation cost	Changes visual modality
		Prevents natural motion
Computational	Natural motion	Lacks diversity
	Low annotation cost	Hard to evaluate quality
		Needs multi-camera setups

tions of 3D coordinates (i.e., 2D position and depth) when hand joints are fully visible.

However, it is not extensively available according to the number of frames due to the high annotation cost. In addition, since it is not robust for occluded keypoints, this approach only allows fingertip annotation, instead of full hand joints. For these limitations, these datasets provide a small amount of data (≈ 3K images) used for evaluation only. Additionally, these single-view datasets can produce view-dependent annotation errors because a single-depth camera captures the distance to the hand skin surface, not the true joint position. To reduce such unavoidable errors, subsequent annotation methods based on multi-camera setups provide further accurate annotations (see Sect. 4.4).

4.2 Synthetic-Model-Based Annotation

To acquire large-scale hand images and labels, synthetic methods based on synthetic hand and full-body models (Loper et al. 2015; Rogez et al. 2014; Romero et al. 2017; Šarić 2011) have been proposed. SynthHands (Mueller et al., 2017) and RHD (Zimmermann & Brox, 2017) render synthetic hand images with randomized real backgrounds from either a first- or third-person view. MVHM (Chen et al., 2021) generates multi-view synthetic hand data rendered from eight viewpoints. These datasets have succeeded in providing accurate hand keypoint labels on a large scale. Although they can generate various background patterns inexpensively, the lighting and texture of hands are not well simulated, and the simulation of hand-object interaction is not considered in the data generation process.

To handle these issues, GANerated (Mueller et al., 2018) utilizes GAN-based image translation to stylize synthetic hands more realistically. Furthermore, ObMan (Hasson et al., 2019) simulates the hand-object interaction in data generation using a hand grasp simulator (Graspit (Miller & Allen, 2005)) with known 3D object models (ShapeNet (Chang

et al., 2015)). Ohkawa et al. proposed foreground-aware image stylization to convert the simulation texture in the ObMan data to a more realistic one while separating the hand regions and backgrounds (Ohkawa et al., 2021). Corona et al. attempted to synthesize more natural hand grasps with affordance classification and the refinement of fingertip locations (Corona et al., 2020). However, the ObMan data only provide static hand images with hand-held objects, not hand motion. The hand motion simulation while approaching the object remains an open problem.

4.3 Hand-Marker-Based Annotation

As shown in Fig. 5, hand-marker-based annotation automatically tracks attached hand markers and then calculates the coordinates of hand joints. Initially, Wetzler et al. attached magnetic sensors to fingertips that provide 6-DoF information of the markers (Wetzler et al., 2015). While this scheme can annotate fingertips only, recent datasets, Big-Hand2.2M (Yuan et al., 2017) and FPHA (Garcia-Hernando et al., 2018), use these sensors to offer the annotation of the full 21 hand joints. Figure 6 shows how to compute the joint positions given six magnetic sensors. It uses inverse kinematics to infer all 21 hand joints, which fits a hand skeleton
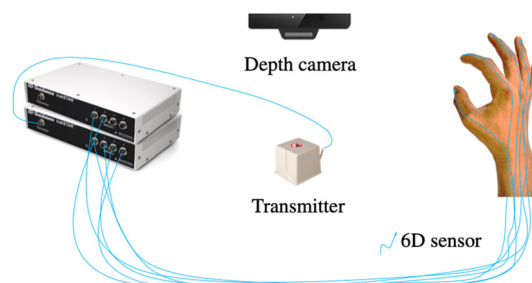
Fig. 5 Illustration of a hand marker setup (Yuan et al., 2017)
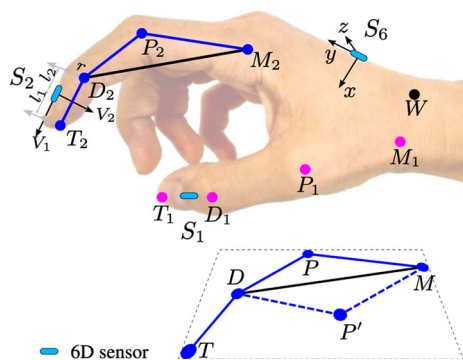
Fig. 6 Calculation of joint positions from tracked markers (Yuan et al., 2017). S_i denotes the position of the markers, and W, M_i, P_i, D_i, and T_i are the positions of hand joints listed from the wrist to the fingertips

Fig. 7 Illustration of a multi-camera setup (Zimmermann et al., 2019)

Fig. 8 Illustration of a many-camera setup (Wuu et al., 2022). This setup has about 100 synchronized cameras and is used to create the InterHands2.6M dataset (Moon et al., 2020)

with the constraints of the maker positions and user-specific bone length manually measured beforehand.

However, these sensors obstruct natural hand movement and distort the appearance of the hand. Due to the changes in hand appearance, these datasets have been proposed for the benchmark of depth-based estimation, not the RGB-based task. On the contrary, GRAB (Taheri et al., 2020) is built with a motion capture system for human hands and body, but it does not possess visual modality, e.g., RGB images.

4.4 Computational Annotation

Computational annotation is categorized into two major approaches: *hand model fitting* and *triangulation*. Unlike hand-marker-based annotation, these methods can capture natural hand motion without attaching hand markers.

Model fitting (depth). Early works of computational annotation utilize *model fitting* on depth images (Supancic et al., 2018; Yuan et al., 2018). Since a depth image provides 3D structural information, their works fit a 3D hand model, from which joint positions can be obtained, to the depth image. ICVL (Tang et al., 2014) fits a convex rigid body model by solving a linear complementary problem with physical constraints (Melax et al., 2013). NYU (Tompson et al., 2014) uses a hand model defined by spheres and cylinders and formulates the model fitting as a kind of particle swarm optimization (Oikonomidis et al., 2011, 2012). The use of other cues for the model fitting is also studied (Ballan et al., 2012; Lu et al., 2003), such as edges, optical flow, shading, and collisions. Sharp et al. paint hands to obtain hand part labels by color segmentation on RGB images and the proxy cue of hand parts further helps the depth-based model fitting (Sharp et al., 2015).

Using these depth datasets, several more accurate labeling methods have been proposed. Rogez et al. gave manual annotation to a few joints and searched the closest 3D pose from a pool of synthetic hand pose data (Rogez et al., 2014).
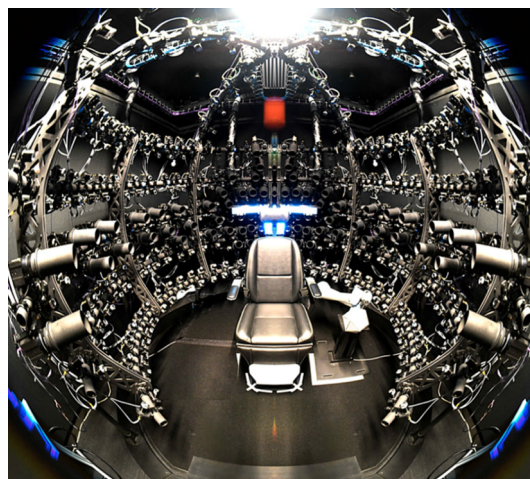
Oberweger et al. considered model fitting with temporal coherence (Oberweger et al., 2016). This method selects reference frames from a depth video and asks annotators for manual labeling. Model fitting is done separately for annotated reference frames and unlabeled non-reference frames. Finally, all sequential poses are optimized to satisfy temporal smoothness.

Triangulation (RGB). For the annotation of RGB images, a multi-camera studio is often used to compute 3D points by multi-view geometry, i.e., *triangulation* (see Fig. 7). Panoptic Studio (Simon et al., 2017) and InterHand2.6M (Moon et al., 2020) triangulate a 3D hand pose from multiple 2D hand keypoints provided by an open source library, OpenPose (Hidalgo et al., 2018), or human annotators. The generated 3D hand pose is reprojected onto the image planes of other cameras to annotate hand images with novel viewpoints. This multi-view annotation scheme is beneficial when many cameras are installed (see Fig. 8). For instance, the InterHand2.6M manually annotates keypoints from 6 views
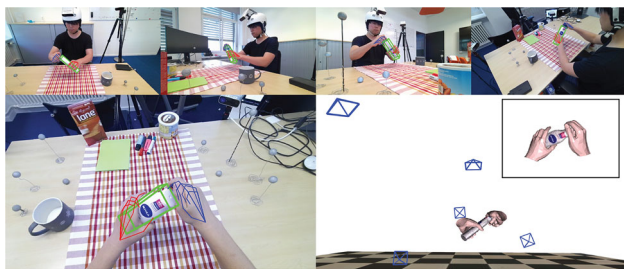
Fig. 9 Synchronized multi-camera setup with first-person and third-person cameras (Kwon et al., 2021)

and reprojects the triangulated points to the other many views (100+). This setup can produce over 100 training images for every single annotation. The InterHand2.6M has million-scale training data.

This point-level triangulation method works quite well when many cameras (30+) are arranged (Moon et al., 2020; Simon et al., 2017). However, the AssemblyHands setup (Ohkawa et al., 2023) has only eight static cameras, and then the predicted 2D keypoints to be triangulated tend to be suboptimal due to hand-object occlusion during the assembly task. To improve the accuracy of triangulation in such sparse camera settings, Ohkawa et al. adopt multi-view aggregation of encoded features by the 2D keypoint detector and compute 3D coordinates from constructed 3D volumetric features (Bartol et al., 2022; Iskakov et al., 2019; Ohkawa et al., 2023; Zimmermann et al., 2019). This feature-level triangulation provides better accuracy than the point-level method, achieving an average keypoint error of 4.20 mm, which is 85% lower than the error of the original annotations in Assembly101 (Sener et al., 2022).

Model fitting (RGB). Model fitting is also used in RGB-based pose annotation. FreiHAND (Zimmermann et al., 2019, 2021) utilizes a 3D hand template (MANO (Romero et al., 2017)) fitting to multi-view hand images with sparse 2D keypoint annotation. The dataset increases the variation of training images by randomly synthesizing the background and using captured real hands as the foreground. YouTube3DHands (Kulon et al., 2020) uses the MANO model fitting to estimated 2D hand poses in YouTube videos. HO-3D (Hampali et al., 2020), DexYCB (Chao et al., 2021), and H2O (Kwon et al., 2021) jointly annotate 3D hand and object poses to facilitate a better understanding of hand-object interaction. Using estimated or manually annotated 2D keypoints, their datasets fit the MANO model and 3D object models to the hand images with objects.

While most methods capture hands from static third-person cameras, H2O and AssemblyHands install first-person cameras that are synchronized with static third-person cameras (see Fig. 9). With camera calibration and head-mounted camera tracking, such camera systems can offer 3D hand pose annotations for first-person images by pro-

jecting annotated keypoints from third-person cameras onto first-person image planes. This reduces the cost of annotating first-person images, which is considered expensive because the image distribution changes drastically over time and the hands are sometimes out of view.

These computational methods can generate labels with little human effort, although the camera system itself is costly. However, assessing the quality of the labels is still difficult. In fact, the annotation quality depends on the number of cameras and their arrangement, the accuracy of hand detection and the estimation of 2D hand poses, and the performance of triangulation and fitting algorithms.

5 Learning with Limited Labels

As explained in Sect. 4, existing annotation methods have certain pros and cons. Since perfect annotation in terms of amount and quality cannot be assumed, training 3D hand pose estimators with limited annotated data is another important study. Accordingly, we introduce learning methods using unlabeled data in this section, namely self-supervised pretraining, semi-supervised learning, and domain adaptation.

5.1 Self-Supervised Pretraining and Learning

Self-supervised pretraining aims to utilize massive unlabeled hand images and build an improved encoder network before supervised learning with labeled images. As shown in Fig. 10, recent works (Spurr et al., 2021; Zimmermann et al., 2021) first pretrain an encoder network that extracts image features by using contrastive learning [e.g., MoCo (He et al., 2020) and SimCLR (Chen et al., 2020)] and then fine-tune the whole network in a supervised manner. The core idea of contrastive learning is to push a pair of *similar* instances closer together in an embedding space while unrelated instances are pushed apart. This approach focuses on how to define the *similarity*
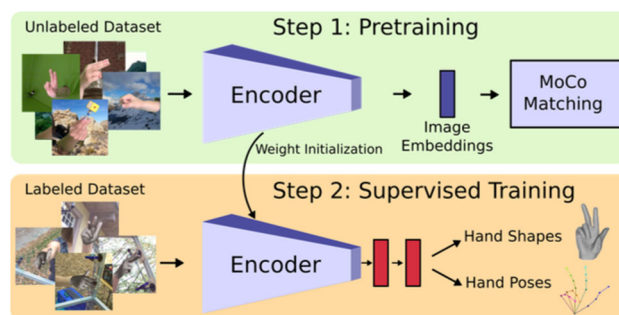
Fig. 10 Self-supervised pretraining of 3D hand pose estimation (Zimmermann et al., 2021). The pretraining phase (step 1) aims to construct an improved encoder network by using many unlabeled data before supervised learning (step 2). The work uses MoCo (He et al., 2020) as a method of self-supervised learning

of hand images and how to design embedding techniques. Spurr et al. proposed to geometrically align two features generated from differently augmented instances (Spurr et al., 2021). Zimmermann et al. found that multi-view images representing the same hand pose can be effective pair supervision (Zimmermann et al., 2021).

Other works utilize the scheme of self-supervised learning that solves an auxiliary task, instead of the target task of hand pose estimation. Given the prediction on an unlabeled depth image, (Oberweger et al., 2015; Wan et al., 2019) render a synthetic depth image and penalize the matching between the input image and the one generated from the prediction. This auxiliary loss by image synthesis is informative even when annotations are scarce.

5.2 Semi-Supervised Learning

As shown in Fig. 11, semi-supervised learning is used to learn from small labeled data and large unlabeled data simultaneously. Liu et al. proposed a pseudo-labeling method that learns unlabeled instances with pseudo-ground-truth given from the model's prediction (Liu et al., 2021). This pseudo-label training is applied only when its prediction satisfies spatial and temporal constraints. The spatial constraints check the correspondence of a 2D hand pose and the 2D pose projected from 3D hand pose prediction. In addition, they include a constraint based on bio-mechanical feasibility, such as bone lengths and joint angles. The temporal constraints indicate the smoothness of hand pose and mesh predictions over time.

Yang et al. proposed the combination of pseudo-labeling and consistency training (Yang et al., 2021). In pseudo-labeling, the generated pseudo-labels are corrected by fitting the hand template model. In addition, the work enforces consistency losses between the predictions of differently augmented instances and between the modalities of 2D hand poses and hand masks.

Spurr et al. applied adversarial training to a sequence of predicted hand poses (Spurr et al., 2021). The encoder network is expected to be improved by fooling a discriminator that distinguishes between plausible and invalid hand poses.
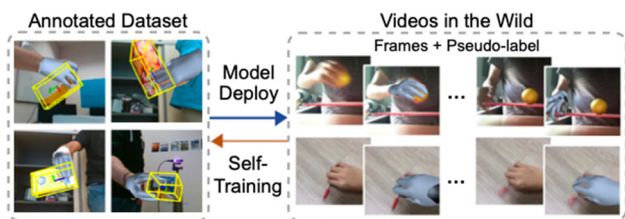

Fig. 11 Semi-supervised learning of 3D hand pose estimation (Liu et al., 2021). The model is trained jointly on annotated data and unlabeled data with pseudo-labels

5.3 Domain Adaptation

Domain adaptation aims to improve model performance on target data by learning from labeled source data and target data with limited labels. This study has addressed two types of underlying domain gaps: *between different datasets* and *between different modalities*.

The former problem *between different datasets* is a common domain adaptation problem where the source and target data are sampled from two datasets with different image statistics, e.g., sim-to-real adaptation (Jiang et al., 2021; Tang et al., 2013) (see Fig. 12). The model has access to readily available synthetic images with labels and target real images without labels. The latter problem *between different modalities* is characterized as *modality transfer* where the source and target data represent the same scene, but their modalities are different, e.g., depth vs. RGB (see Fig. 13). This aims to utilize information-rich source data, e.g., depth images contain 3D structural information, for inferring easily available target data (e.g., RGB images).

To reduce the gap between the two datasets, two major approaches have been proposed: *generative methods* and *adversarial methods*. In generative methods, Qi et al. proposed an image translation method to alter the synthetic textures to realistic ones and train a model on generated real-like synthetic data (Qi et al., 2020).

Adversarial methods enforce matching two domains' features so that the feature extractor can encode features even from the target domain. However, in addition to the domain gap in an input space (e.g., the difference in backgrounds), the gap in a label space also exists in this task, which is
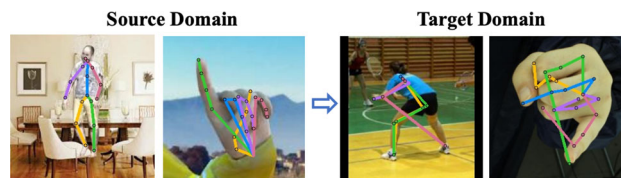

Fig. 12 Poor generalization to an unknown domain (Jiang et al., 2021). The models trained on synthetic images (source) exhibit a limited capacity for inferring poses on real images (target)
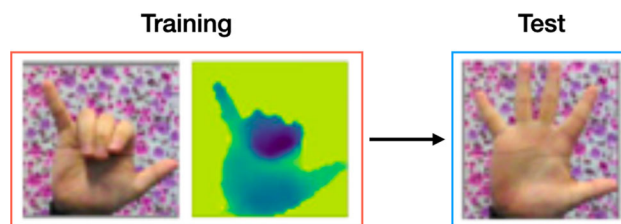

Fig. 13 Example of modality transfer. During training, RGB and depth images are accessible and RGB images are given in the test phase. The training aims to utilize the support of depth information to improve RGB-based hand pose estimation

not assumed in typical adversarial methods (Ganin & Lempitsky, 2015; Tzeng et al., 2017). Zhang et al. developed a feature matching method based on Wasserstein distance and proposed adaptive weighting to enable matching only for features related to hand characteristics, except for label information (Zhang et al., 2020). Jiang et al. utilized an adversarial regressor and optimized the domain disparity by a minimax game (Jiang et al., 2021). Such minimax of disparity is effective in domain adaptation of regression tasks, including hand pose estimation.

As for the *modality transfer* problem, Yuan et al. and Rad et al. attempted to use depth images as the auxiliary information during training and test the model on RGB images (Rad et al., 2018; Yuan et al., 2019). They observed that learned features from depth images could support RGB-based hand pose estimation. Park et al. transferred the knowledge from depth images to infrared (IR) images that have less motion blur (Park et al., 2020). Their training is facilitated by matching two features from paired images, e.g., (RGB, depth) and (depth, IR). Baek et al. newly defined the domain of hand-only images where a hand-held object is removed. The work translates hand-object images to hand-only images by using GAN and mesh renderer (Baek et al., 2020). Given a hand-object image with an unknown object, this method can generate hand-only images, from which hand pose estimation is more tractable.

6 Future Directions

6.1 Flexible Camera Systems

We believe that hand image capture will feature more flexible camera systems, such as using first-person cameras. To reduce the occlusion effect without the need for hand markers, recently published hand datasets have been acquired by multi-camera setups, e.g., DexYCB (Chao et al., 2021), InterHand2.6M (Moon et al., 2020), and FreiHAND (Zimmermann et al., 2019). These setups are static and not suitable for capturing dynamic user behavior. To address this, a *first-person camera* attached to the user's head or body is useful because it mostly captures close-up hands even when the user moves around. However, as shown in Table 1, existing first-person benchmarks have a very limited variety due to heavy occlusion, motion blur, and a narrow field-of-view.

One promising direction is a joint camera setup with first-person and third-person cameras, such as H2O (Kwon et al., 2021) and AssemblyHands (Ohkawa et al., 2023). This results in flexibly capturing the user's hands from the first-person camera while taking the benefits of multiple third-person cameras (e.g., mitigating the occlusion effect). However, the first-person camera wearer doesn't always have to be alone. Image capture with multiple first-person camera

wearers in a static camera setup will advance the analysis of multi-person cooperation and interaction, e.g., game playing and construction with multiple people.

6.2 Various Types of Activities

We believe that increasing the type of activities is an important direction for generalizing models to various situations with hand-object interaction. A major limitation of existing hand datasets is the narrow variation of users' performing tasks and grasping objects. To avoid object occlusion, some works did not capture hand-object interaction (Moon et al., 2020; Yuan et al., 2017; Zimmermann & Brox, 2017). Others (Chao et al., 2021; Hasson et al., 2019; Hampali et al., 2020) used pre-registered 3D object models (e.g., YCB (Çalli et al., 2015)) to simplify in-hand object pose estimation. User action is also very simple in these benchmarks, such as *pick and place*.

From an affordance perspective (Hassanin et al., 2021), diversifying the object category will result in increasing hand pose variation. Potential future works will capture goal-oriented and procedural activities that naturally occur in our daily life (Damen et al., 2021; Grauman et al., 2022; Sener et al., 2022), such as cooking, art and craft, and assembly.

To enable this, we need to develop portable camera systems and robust annotation methods for complex backgrounds and unknown objects. In addition, occurring hand poses are constrained to the context of the activity. Thus, pose estimators conditioned by actions, objects, or textual descriptions of the scene will improve estimation in various activities.

6.3 Towards Minimal Human Effort

Sections 4 and 5 separately explain efficient annotation and learning. To minimize the effort of human intervention, utilizing findings from both annotation and learning perspectives is one of the promising directions. Feng et al. exploited *active learning* that optimizes which unlabeled instance should be annotated and semi-supervised learning that jointly utilizes labeled data and large unlabeled data (Feng et al., 2021). However, this method is constrained to triangulation-based 3D pose estimation. As we mentioned in Sect. 4.4, another major computational annotation is *model fitting*; thus, we still need to consider such a collaborative approach in the annotation based on model fitting.

Zimmermann et al. also proposed a framework of human-in-loop annotation that inspects the annotation quality manually while updating annotation networks on the inspected annotations (Zimmermann & Brox, 2017). However, this human check will be a bottleneck in large dataset construction. The evaluation of annotation quality on the fly is a

necessary technique to scale up the combination of annotation and learning.

6.4 Generalization and Adaptation

Increasing the generalization ability across different datasets or adapting models to a specific domain is a remaining issue. The bias of existing training datasets hinders the estimators from inferring test images captured under very different imaging conditions. In fact, as reported in Han et al. (2020); Zimmermann et al. (2019), models trained on existing hand pose datasets poorly generalize to other datasets. For real-world applications (e.g., AR), it is crucial to transfer models from indoor hand datasets to outdoor videos because common multi-camera setups are not available outdoors (Ohkawa et al., 2022). Thus, aggregating multiple annotated yet biased datasets for generalization and robustly adapting to very different environments are important future tasks.

7 Summary

We presented the survey of 3D hand pose estimation from the standpoint of efficient annotation and learning. We provided a comprehensive overview of this task and modeling, and open challenges during dataset construction. We investigated annotation methods categorized as manual, synthetic-model-based, hand-marker-based, and computational approaches, and examined their respective strengths and weaknesses. In addition, we studied learning methods that can be applied even when annotations are scarce, namely self-supervised pretraining, semi-supervised learning, and domain adaptation. Finally, we discussed potential future advancements in 3D hand pose estimation, including next-generation camera setups, increased object and action variation, jointly optimized annotation and learning techniques, and generalization and adaptation.

References

Baek, S., Kim, K. I., & Kim T.-K. (2020). Weakly-supervised domain adaptation via GAN and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 6120–6130).

Ballan, L., Taneja, A., Gall, J., Gool, L. V., & Pollefeys, M. (2012). Motion capture of hands in action using discriminative salient points. In *Proceedings of the European conference on computer vision (ECCV)* (Vol. 7577, pp. 640–653).

Bartol, K., Bojanić, D., Petković, T. & Pribanić T. (2022). Generalizable human pose triangulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11018–11027).

Bianchi, M., Salaris, P., & Bicchi, A. (2013). Synergy-based hand pose sensing: Optimal glove design. *The International Journal of Robotics Research (IJRR), 32*(4), 407–424.

Boukhayma, A., de Bem, R., & Torr, P. H. S. (2019). 3D hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10843–10852).

Cai, Y., Ge, L., Cai, J., & Yuan, J. (2018). Weakly-supervised 3D hand pose estimation from monocular RGB images. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 678–694).

Çalli, B., Walsman, A., Singh, A., Srinivasa, S. S., Abbeel, P., & Dollar, A. M. (2015). Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robotics Automation Magazine, 22*(3), 36–52.

Chang, A. X., Funkhouser, T. A., Guibas, L. J., Hanrahan, P., Huang, Q.-X., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L. & Yu, F. (2015). Shapenet: An information-rich 3d model repository. *CoRR*, arXiv:1512.03012

Chao, Y.-W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y. S., Van Wyk, K., Iqbal, U., Birchfield, S., Kautz, J. & Fox, D. (2021). DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 9044–9053).

Chatzis, T., Stergioulas, A., Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2020). A comprehensive study on deep learning-based 3d hand pose estimation methods. *Applied Sciences, 10*, 6850.

Chen, L., Lin, S.-Y., Xie, Y., Lin, Y.-Y. & Xie, X. (2021). MVHM: A large-scale multi-view hand mesh benchmark for accurate 3d hand pose estimation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 836–845).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the international conference on machine learning (ICML)* (Vol. 119, pp. 1597–1607).

Chen, Y., Tu, Z., Kang, D., Bao, L., Zhang, Y., Zhe, X., Chen, R., & Yuan, J. (2021). Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10451–10460).

Ciocarlie, M. T., & Allen, P. K. (2009). Hand posture subspaces for dexterous robotic grasping. *The International Journal of Robotics Research (IJRR), 28*(7), 851–867.

Corona, E., Pumarola, A., Alenyà, G., Moreno-Noguer, F., & Rogez, G. (2020). GanHand: Predicting human grasp affordances in multi-

object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 5030–5040).

Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W. & Wray, M. (2021). Rescaling egocentric vision. *International Journal of Computer Vision (IJCV)*, early access.

Doosti, B. (2019). Hand pose estimation: A survey. *CoRR*, arXiv:1903.01013

Erol, A., Bebis, G., Nicolescu, M., Boyle, R. D., & Twombly, X. (2007). Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding (CVIU), 108*(1–2), 52–73.

Feng, Q., He, K., Wen, H., Keskin, C., & Ye, Y. (2021). Active learning with pseudo-labels for multi-view 3d pose estimation. *CoRR*, arXiv:2112.13709

Ganin, Y. & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the international conference on machine learning (ICML)* (pp. 1180–1189).

Garcia-Hernando, G., Yuan, S., Baek, S. & Kim, T.-K. (2018). First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 409–419).

Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J. & Yuan, J. (2019). 3D hand shape and pose estimation from a single RGB image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10833–10842).

Glauser, O., Wu, S., Panozzo, D., Hilliges, O., & Sorkine-Hornung, O. (2019). Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (ToG), 38*(4), 41:1-41:15.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S. K., Ryan, F., Sharma, J., Wray, M., Xu, M.g, Xu, E. Zhongcong, Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty,M.,Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Fuegen, C., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P. R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G. M., Ghanem, B., Ithapu, V. K., Jawahar, C. V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H. Soo, Rehg, J. M., Sato, Y., Shi, J., Shou, M. Z., Torralba, A., Torresani, Lo, Yan, M.i, & Malik, J. (2022). Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 18973–18990).

Hampali, S., Rad, M., Oberweger, M. & Lepetit, V. (2020). Honnotate: A method for 3D annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 3196–3206).

Hampali, S., Sarkar, S. D., Rad, M. & Lepetit, V. (2022) 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 (CVPR)* (pp. 11080–11090).

Han, S., Liu, B., Cabezas, R., Twigg, C. D., Zhang, P., Petkau, J., Yu, T.-H., Tai, C.-J., Akbay, M., Wang, Z., Nitzan, A., Dong, G., Ye, Y., Tao, L., Wan, C., & Wang, R. (2020). MEgATrack: Monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG), 39*(4), 87.

Han, S., Wu, P.-C., Zhang, Y., Liu, B., Zhang, L., Wang, Z., Si, W., Zhang, P., Cai, Y., Hodan, T., Cabezas, R., Tran, L., Akbay, M., Yu, T.-H., Keskin, C. & Wang, R. (2022). UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *Proceedings of the ACM SIGGRAPH Asia conference* (pp. 50:1–50:9).

Handa, A., Wyk, K. V., Yang, W., Liang, J., Chao, Y.-W., Wan, Q., Birchfield, S., Ratliff, N. & Fox, D. (2020) DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (pp. 9164–9170).

Hassanin, M., Khan, S., & Tahtali, M. (2021). Visual affordance and function understanding: A survey. *ACM Computing Survey, 54*(3), 47:1-47:35.

Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M. J., Laptev, I. & Schmid, C. (2019). Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11807–11816).

He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. B. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 9726–9735).

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 770–778).

Hidalgo, G., Cao, Z., Simon, T., Wei, S.-E., Raaj, Y., Joo, H. & Sheikh, Y. (2018). OpenPose. https://github.com/CMU-Perceptual-Computing-Lab/openpose

Huang, L., Tan, J., Liu, J., & Yuan, J. (2020). Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation. In *Proceedings of the European conference on computer vision (ECCV)* (Vol. 12370, pp. 17–33).

Huang, W., Ren, P., Wang, J., Qi, Q. & Sun, H. (2020). AWR: Adaptive weighting regression for 3D hand pose estimation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (pp. 11061–11068).

Iqbal, U., Garbade, M. & Gall, J. (2017). Pose for action–action for pose. In *Proceedings of the IEEE international conference on automatic face & gesture recognition, FG* (pp. 438–445).

Iqbal, U., Molchanov, P., Breuel, T. M., Gall, J. & Kautz, J. (2018). Hand pose estimation via latent 2.5D heatmap regression. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 125–143).

Iskakov, K., Burkov, E., Lempitsky, V. & Malkov, Y. (2019). Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 7718–7727).

Jiang, J., Ji, Y., Wang, X., Liu, Y., Wang, J. & Long, M. (2021). Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 6780–6789).

Kulon, D., Güler, R. A., Kokkinos, I., Bronstein, M. M. & Zafeiriou, S. (2020). Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4989–4999).

Kwon, T., Tekin, B., Stühmer, J., Bogo, F. & Pollefeys, M. (2021). H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 10118–10128).

Le, V.-H., & Nguyen, H.-C. (2020). A survey on 3d hand skeleton and pose estimation by convolutional neural network. *Advances in Science, Technology and Engineering Systems Journal (ASTES), 5*(4), 144–159.

Lepetit, V. (2020). Recent advances in 3d object and hand pose estimation. *CoRR*, arXiv:2006.05927

Liang, H., Yuan, J.G., Thalmann, D. & Magnenat-Thalmann, N. (2015). AR in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In *Proceedings of the ACM international conference on multimedia (MM)* (pp. 743–744).

Liu, S., Jiang, H., Xu, J., Liu, S. & Wang, X. (2021). Semi-supervised 3D hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 14687–14697).

Liu, Y., Jiang, J. & Sun, J. (2021). Hand pose estimation from rgb images based on deep learning: A survey. In *IEEE international conference on virtual reality (ICVR)* (pp. 82–89).

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (ToG), 34*(6), 24:81-24:816.

Lu, S., Metaxas, D. N., Samaras, D. & Oliensis, J. (2003). Using multiple cues for hand tracking and model refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 443–450).

Mandikal, P. & Grauman, K. (2021). DexVIP: Learning dexterous grasping with human hand pose priors from video. In *Proceedings of the conference on robot learning (CoRL)* (pp. 651–661).

Melax, S., Keselman, L., & Orsten, S. (2013). Dynamics based 3d skeletal hand tracking. In *Proceedings of the graphics interface (GI)* (pp. 63–70).

Miller, A., & Allen, P. (2005). Graspit!: A versatile simulator for robotic grasping. *IEEE Robotics and Automation Magazine (RAM), 11*, 110–122.

Miyata, N., Kouchi, M., Kurihara, T. & Mochimaru, M. (2004). Modeling of human hand link structure from optical motion capture data. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 2129–2135).

Moon, G., Yu, S.-I., Wen, H., Shiratori, T. & Lee, K. M. (2020). InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 548–564).

Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D. & Theobalt, C.(2018). GANerated Hands for real-time 3D hand tracking from monocular RGB. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 49–59).

Mueller, F., Davis, M., Bernard, F., Sotnychenko, O., Verschoor, M., Otaduy, M. A., Casas, D., & Theobalt, C. (2019). Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG), 38*(4), 49:1-49:13.

Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D. & Theobalt, C. (2017). Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 1163–1172).

Oberweger, M., Riegler, G., Wohlhart, P. & Lepetit, V. (2016). Efficiently creating 3d training data for fine hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4957–4965).

Oberweger, M., Wohlhart, P. & Lepetit, V. (2015). Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 3316–3324).

Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L. & Keskin, C. (2023). AssemblyHands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Ohkawa, T., Li, Y.-J., Fu, Q., Furuta, R., Kitani, K. M. & Sato, Y. (2022). Domain adaptive hand keypoint and pixel localization in the wild. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 68–87).

Ohkawa, T., Yagi, T., Hashimoto, A., Ushiku, Y., & Sato, Y. (2021). Foreground-aware stylization and consensus pseudo-labeling for domain adaptation of first-person hand segmentation. *IEEE Access, 9*, 94644–94655.

Oikonomidis, I., Kyriazis, N. & Argyros, A. A. (2011). Efficient model-based 3d tracking of hand articulations using kinect. In *Proceedings of the British machine vision conference (BMVC)* (pp. 1–11).

Oikonomidis, I., Kyriazis, N. & Argyros, A. A. (2012). Tracking the articulated motion of two strongly interacting hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (pp. 1862–1869).

Park, G., Kim, T.-K. & Woo, W. (2020). 3d hand pose estimation with a single infrared camera via domain transfer learning. In *Proceedings of the IEEE international symposium on mixed and augmented reality (ISMAR)* (pp. 588–599).

Qi, M., Remelli, E., Salzmann, M. & Fua, P. (2020). Unsupervised domain adaptation with temporal-consistent self-training for 3d hand-object joint reconstruction. *CoRR*, arXiv:2012.11260

Qian, C., Sun, X., Wei, Y., Tang, X.& Sun, J. (2014). Realtime and robust hand tracking from depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 1106–1113).

Qin, Y., Wu, Y.-H., Liu, S., Jiang, H., Yang, R., Fu, Y., & Wang, X. (2022). DexMV: Imitation learning for dexterous manipulation from human videos. In *Proceedings of the European conference on computer vision (ECCV)* (Vol. 13699, pp. 570–587).

Rad, M., Oberweger, M., & Lepetit, V. (2018). Domain transfer for 3d pose estimation from color images without manual annotations. In *Proceedings of the Asian conference on computer vision (ACCV)* (Vol. 11365, pp. 69–84).

Ren, P., Sun, H., Qi, Q., Wang, J. & Huang, W. (2019). SRN: Stacked regression network for real-time 3D hand pose estimation. In *Proceedings of the British machine vision conference (BMVC)*.

Rogez, G., Khademi, M., Supancic, J. S., III., Montiel, J. M. M., & Ramanan, D. (2014). 3d hand pose detection in egocentric RGB-D images. In *Proceedings of the European conference on computer vision workshops (ECCVW)* (Vol. 8925, pp. 356–371).

Rogez, G., Supancic III, J. S. & Ramanan, D. (2015). First-person pose recognition using egocentric workspaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4325–4333).

Rogez, G., Supancic III, J. S. & Ramanan, D. (2015). Understanding everyday hands in action from RGB-D images. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 3889–3897).

Romero, J., Kjellström, H. & Kragic, D. (2010). Hands in action: Realtime 3d reconstruction of hands in interaction with objects. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)* (pp. 458–463).

Romero, J., Tzionas, D., & Black, M. J. (2017). Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG), 36*(6), 245:1-245:17.

Santavas, N., Kansizoglou, I., Bampis, L., Karakasis, E. G., & Gasteratos, A. (2021). Attention! A lightweight 2d hand pose estimation approach. *IEEE Sensors, 21*(10), 11488–11496.

Šarić, M. (2011). Libhand: A library for hand articulation. Version 0.9.

Schröder, M., Maycock, J. & Botsch, M. (2015). Reduced marker layouts for optical motion capture of hands. In *Proceedings of the ACM SIGGRAPH conference on motion in games (MIG)* (pp. 7–16). ACM.

Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R. & Yao, A. (2022). Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 21096–21106).

Sharp, T., Keskin, C., Robertson, D. P., Taylor, J., Shotton, J., Kim, D., Rhemann, C., Leichter, I., Vinnikov, A., Wei, Y., Freedman, D., Kohli, P., Krupka, E., Fitzgibbon, A. W. & Izadi, S. (2015). Accurate, robust, and flexible real-time hand tracking. In *Proceedings*

of the SIGCHI conference on human factors in computing systems (CHI) (pp. 3633–3642).

Simon, T., Joo, H., Matthews, I. & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4645–4653).

Spurr, A., Dahiya, A., Wang, X., Zhang, X. & Hilliges, O. (2021). Self-supervised 3d hand pose estimation from monocular RGB via contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 11210–11219).

Spurr, A., Iqbal, U., Molchanov, P., Hilliges, O. & Kautz, J. (2020). Weakly supervised 3D hand pose estimation via biomechanical constraints. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 211–228).

Spurr, A., Molchanov, P., Iqbal, U., Kautz, J. & Hilliges, O. (2021). Adversarial motion modelling helps semi-supervised hand pose estimation. *CoRR*, arXiv:2106.05954

Spurr, A., Song, J., Park, S. & Hilliges, O. (2018). Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 89–98).

Sridhar, S., Mueller, F., Zollhoefer, M., Casas, D., Oulasvirta, A. & Theobalt, C. (2016). Real-time joint tracking of a hand manipulating an object from RGB-D input. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 294–310).

Sridhar, S., Oulasvirta, A. & Theobalt, C. (2013). Interactive markerless articulated hand motion tracking using RGB and depth data. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 2456–2463).

Supancic, J. S., III., Rogez, G., Yang, Y., Shotton, J., & Ramanan, D. (2018). Depth-based hand pose estimation: Methods, data, and challenges. *International Journal Computer Vision (IJCV), 126*(11), 1180–1198.

Taheri, O., Ghorbani, N., Black, M. J. & Tzionas, D. (2020). GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 581–600).

Tang, D., Chang, H. J., Tejani, A. & Kim, T.-K. (2014). Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 3786–3793).

Tang, D., Yu, T.-H. & Kim, T.-K. (2013). Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 3224–3231).

Tekin, B., Bogo, F. & Pollefeys, M. (2019). H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4511–4520).

Tompson, J., Stein, M., LeCun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG), 33*(5), 169:1-169:10.

Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2962–2971).

Wan, C., Probst, T., Gool, L. V. & Yao, A. (2019). Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10853–10862).

Wang, R. Y., & Popovic, J. (2009). Real-time hand-tracking with a color glove. *ACM Transactions on Graphics (ToG), 28*(3), 63.

Wetzler, A., Slossberg, R. & Kimmel, R. (2015). Rule of thumb: Deep derotation for improved fingertip detection. In *Proceedings of the British machine vision conference (BMVC)* (pp. 33.1–33.12).

Wu, M.-Y., Ting, P.-W., Tang, Y.-H., Chou, E. T., & Fu, L.-C. (2020). Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *Journal of Visual Communication and Image Representation, 70*, 102802.

Wuu, C., Zheng, N., Ardisson, S., Bali, R., Belko, D., Brockmeyer, E., Evans, L., Godisart, T., Ha, H., Hypes, A., Koska, T., Krenn, S., Lombardi, S., Luo, X., McPhail, K., Millerschoen, L., Perdoch, M., Pitts, M. Richard, A., Saragih, J. M., Saragih, J., Shiratori, T., Simon, T., Stewart, M., Trimble, A., Weng, X., Whitewolf, D., Wu, C., Yu, S. & Sheikh, Y. (2022). Multiface: A dataset for neural face rendering. *CoRR*, arXiv:2207.11243

Xiong, F., Zhang, B., Xiao, Y., Cao, Z., Yu, T., Zhou, J. T. & Yuan, J. (2019). A2J: Anchor-to-joint regression network for 3D articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 793–802).

Xu, C. & Cheng, L. (2013). Efficient hand pose estimation from a single depth image. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 3456–3462).

Yang, L., Chen, S. & Yao, A. (2021). Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 11364–11373).

Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J. Y., Lee, K. M., Molchanov, P., Kautz, J., Honari, S., Ge, L., Yuan, J., Chen, X., Wang, G., Yang, F., Akiyama, K., Wu, Y., Wan, Q., Madadi, M., Escalera, S., Li, S., Lee, D., Oikonomidis, I., Argyros, A. A. & Kim, T-K. (2018). Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2636–2645).

Yuan, S., Stenger, B. & Kim, T.-K. (2019). Rgb-based 3d hand pose estimation via privileged learning with depth images. In *Proceedings of the IEEE/CVF international conference on computer vision workshops (ICCVW)*.

Yuan, S., Ye, Q., Stenger, B., Jain, S. & Kim, T.-K. (2017). Big-Hand2.2M benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2605–2613).

Zhang, Y., Chen, L., Liu, Y., Zheng, W. & Yong, J. (2020). Adaptive wasserstein hourglass for weakly supervised RGB 3d hand pose estimation. In *Proceedings of the ACM international conference on multimedia (MM)* (pp. 2076–2084).

Zhou, X., Wan, Q., Zhang, W., Xue, X. & Wei, Y. (2016). Model-based deep hand pose estimation. In *Proceedings of the international joint conference on artificial intelligence (IJCAI)* (pp. 2421–2427).

Zimmermann, C., Argus, M., & Brox, T. (2021). Contrastive representation learning for hand shape estimation. In *Proceedings of the DAGM German conference on pattern recognition (GCPR)* (Vol. 13024, pp. 250–264).

Zimmermann, C. & Brox, T. (2017). Learning to estimate 3D hand pose from single RGB images. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 4913–4921).

Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M. J. & Brox, T. (2019). 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 (ICCV)* (pp. 813–822).