

A Survey of Hand Pose Estimation and Hand Datasets

Xiao Hu^a, Zhiyuan Wang^a, Wei Jia^{*a}

^aHefei University of Technology, Hefei, China

ABSTRACT

Hand pose estimation aims to detect and reconstruct 3D hand skeletons or mesh models from information media such as RGB images, depth maps, point clouds, and voxels. The relevant technologies are mainly applied in fields like human-computer interaction, motion analysis, augmented reality, and virtual reality. In recent years, with the rapid development of deep learning technology and the widespread application of depth cameras, hand pose estimation has attracted more and more attention. The current state-of-the-art has achieved good performance, but there are still great challenges when facing hand-object interactions, two-handed interactions, and occlusions in outdoor situations. This paper will classify and introduce the characteristics and advantages and disadvantages of various hand pose estimation algorithms according to the type of information media, collect the main datasets and benchmarks used in hand pose estimation, and finally discuss the main challenges and prospects in the field of hand pose estimation.

Keywords: Hand pose estimation, human-computer interaction, deep learning

1. INTRODUCTION

The task of hand pose estimation has different requirements for different application scenarios: in human-computer interaction¹, facing single/multiple RGB cameras or depth cameras, it is necessary to accurately capture the main hand in the scene and infer the meaning of gestures. Background interference, self-occlusion, and dimension loss are the main challenges. In wearable VR/AR^{2,3,4}, joint magnetic sensors map the positions of hand joints to virtual 3D space, and more precise fingertip positioning is required to determine the hand-object interaction relationship in space. With the promotion and popularization of home wearable VR/AR devices by large technology companies such as Google, Microsoft, and Facebook, as well as the wide application of human-computer interaction verification technology in various fields, hand pose estimation has become a research hotspot in the field of computer vision. With the rapid iteration of deep learning solutions, challenges such as self-occlusion, environmental interference, depth blur, and hand-object interaction in recent years' 3D hand pose estimation have been solved in a series.

This paper uses the introduction of deep learning models as a dividing line. It first introduces human hand pose estimation based on traditional vision schemes (Section 2), and then details the development and latest progress of human hand pose estimation based on deep learning schemes, classified according to the medium of information transmission (Sections 3), followed by collecting widely used datasets in hand pose estimation (Section 4), and finally summarize the main challenges and development prospects of hand pose estimation technology (Section 5), aiming to provide ideas and inspiration for researchers in this direction. The detailed organization is shown in Figure 1. The main contributions of this paper are as follows:

- 1) Classify and summarize hand pose estimation methods from 2017 to 2024 according to single-view/multi-view, and depth estimation/deep learning.
- 2) Organize the mainstream datasets and performance evaluation standards of hand pose estimation methods and compare and discuss the advantages and disadvantages of various standards according to the timeline.
- 3) Discuss the key challenges currently faced by hand pose estimation and point out possible future research directions.

*jiawei@hfut.edu.cn

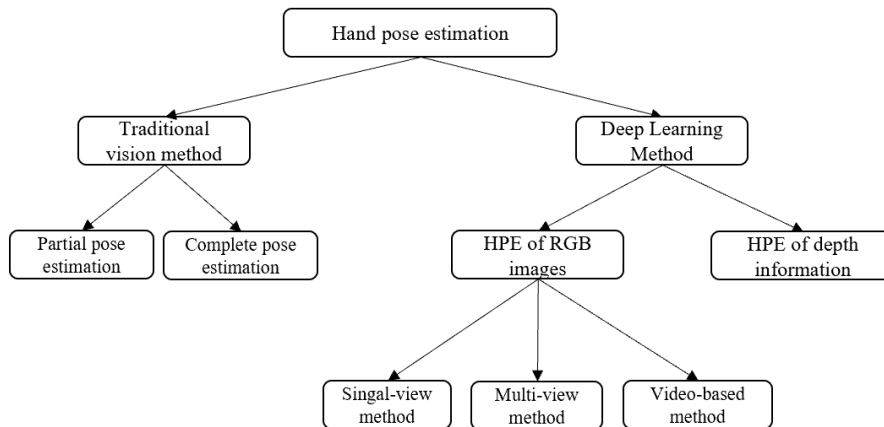


Figure 1. Classification of Hand Pose Estimation (HPE) Methods

2. TRADITIONAL VISION METHOD

Twenty years ago, in human-computer interaction applications, computer vision solutions for obtaining hand poses were more convenient and rapid compared to contact sensors. However, when dealing with high-dimensional hand motion data, traditional vision solutions have certain limitations. It is difficult to balance real-time performance and robustness under limited hardware conditions, and the image conditions are demanding (such as clean backgrounds, static movements, and reduced occlusions). Therefore, traditional vision solutions could only perform limited human-computer interactions. This section, starting from the perspective of performance and based on the types of mainstream methods, categorizes the traditional vision-based approaches for human hand pose estimation into partial pose estimation and tracking-based limited complete pose estimation.

2.1 Partial Pose Estimation

The development of virtual reality VR and augmented reality AR applications is often accompanied by limited resource allocation and real-time requirements. Based on limited computing power, it is necessary to process the high-dimensional problem of hand movement in real time. However, not all application scenarios need to obtain all the information of the hand. In some scenarios (such as when the finger points in a direction), it is only necessary to obtain a rough model of the hand to judge information such as the position of the fingertip, the direction of the finger, or the palm. As shown in Figure 2, only the partial contour of the hand is detected to determine the meaning of the gesture. It can be further divided into hand positioning and gesture classification tasks.



Figure 2. Detecting the meaning of hand pose in a static background⁵

In the hand positioning task, people usually rely on graphics solutions, such as skin color segmentation using lookup tables^{6,7,8}, static background cropping⁹, and particle filtering⁵. These methods generally have strict environmental requirements, such as images with pure color backgrounds, uniform lighting, and single main hands that are easy to segment, or using infrared cameras to crop hand shadows according to body temperature. Another category is object detection methods based on classification, but classifiers are difficult to quickly traverse the entire list of assumptions, so the main work of scholars is concentrated on how to quickly reduce the number of incorrect assumptions^{10,11}, thereby reducing the load of training classifiers.

In the gesture classification task, the position of the hand, the position of the fingertip, and the direction of the finger are the main features. Generally, the position of the hand is determined through the hand silhouette, and the center of gravity or the nearest point farthest from the hand contour¹² is regarded as the position of the hand. The position of the fingertip is processed using circular mask-related technologies^{12,13}, and the connection between the finger and the palm is determined according to the local maximum curvature of the contour boundary^{7,8,9}. On the basis of determining the position of the hand and the fingertip, the direction of the finger can be determined by calculating the main axis of the contour¹⁴. It should be noted that the hand classification task depends on the high-quality hand contour, so it is difficult to ensure the accuracy of the hand classification results in complex image backgrounds.

2.2 Complete Pose Estimation

For hand pose estimation facing full degrees of freedom, model-based tracking methods are generally adopted: in the first frame of the image sequence, search for the template with the smallest matching error in the template library, thereby obtaining a set of parameters closest to the true value. Subsequently, in each frame of the image sequence, the model is calibrated on the basis of the previous frame to minimize the matching error. The matching error is defined as the similarity measurement between model feature groups, and the template library is mainly generated by 3D model projection. During the analysis process, model parameters can be calibrated for feature subsets or specific areas. Complete pose estimation is divided into two main directions based on the model-based tracking architecture. One is to adopt local search, only tracking the best estimation value of each frame. This type of method is difficult to maintain good performance in long-time sequences because the appearance of local minima and frame discontinuities^{15,16} in long-time sequences is inevitable. The other is to adopt multiple hypothesis tracking¹⁷, retaining multiple pose estimation results for each frame. There are alternative solutions to compensate when the best estimation value is not ideal.

The reliability of model-based tracking methods depends on the temporal correlation of the input image sequence. The estimation result of the current frame is determined by the result of the previous frame. When a certain frame is disconnected (such as self-occlusion or rapid movement), the quality of subsequent estimates will be difficult to guarantee. In this regard, it is possible to try to estimate the hand pose of only a single frame, which is similar to the hand positioning task in Section 2.1.

3. DEEP LEARNING METHOD

3.1 Hand Pose Estimation of RGB Images

As the most common and easily accessible visual data, RGB images have attracted widespread attention in the field of hand pose estimation. However, due to the relatively sparse information in RGB images, hand pose estimation based on RGB images faces significant challenges. This section categorizes RGB image inputs into three types according to the scenario: single-view, multi-view, and consecutive frames. Single-view methods are designed for single RGB images and are suitable for pose estimation of data that are easy to obtain, such as outdoor images and images captured by handheld cameras. Multi-view methods integrate information from multiple perspectives at the same time frame in space, mainly used in experimental settings for precise 3D pose capture and recording. Compared with the first two types, methods based on consecutive frames rely on the temporal correlation between frames in motion videos for inference and pose prediction. In addition to the traditional task of 3D spatial position estimation, some methods also demonstrate the ability for model reconstruction (as shown in Figure 3).

3.1.1 Single-View Method

Hand pose estimation and reconstruction methods based on a single RGB image are commonly used in deep learning-based hand pose estimation. Compared with other methods, this method has lower hardware requirements and fewer

restrictions on the implementation scene and has received extensive research and rapid development in the past. However, despite the great success achieved in recent years, estimating and reconstructing hand poses from a single view is still a challenging task due to severe self-occlusion, depth blur, and scarce annotation data.

Ge et al.¹⁸ proposed a method based on Graph CNN¹⁹ without relying on parametric hand models. This method predicts the 3D coordinates of 1280 hand mesh vertices from a monocular RGB image, simultaneously estimating the spatial positions of hand joints and reconstructing the hand mesh. This approach compensates for the lack of hand shape information reconstruction in previous methods. By creating a large-scale synthetic dataset for fully supervised training and performing weakly supervised fine-tuning on real-world datasets, the method can accurately estimate hand pose and shape without 3D ground truth data.

To break through the bottleneck of 3D hand pose estimation's heavy reliance on annotated data, some works have explored in the direction of self-supervision. Spurr et al.²⁰ inspired by the successful application of contrastive learning in image classification tasks, proposed a new self-supervised 3D hand pose estimation method called PeCLR, which estimates hand pose from monocular RGB images using Equivariant Contrastive Learning. The PeCLR method encourages the learned feature representations to remain invariant under appearance transformations and equivariant under geometric transformations through a contrastive learning objective function, enabling the model to effectively learn feature representations from a large amount of unlabeled data.

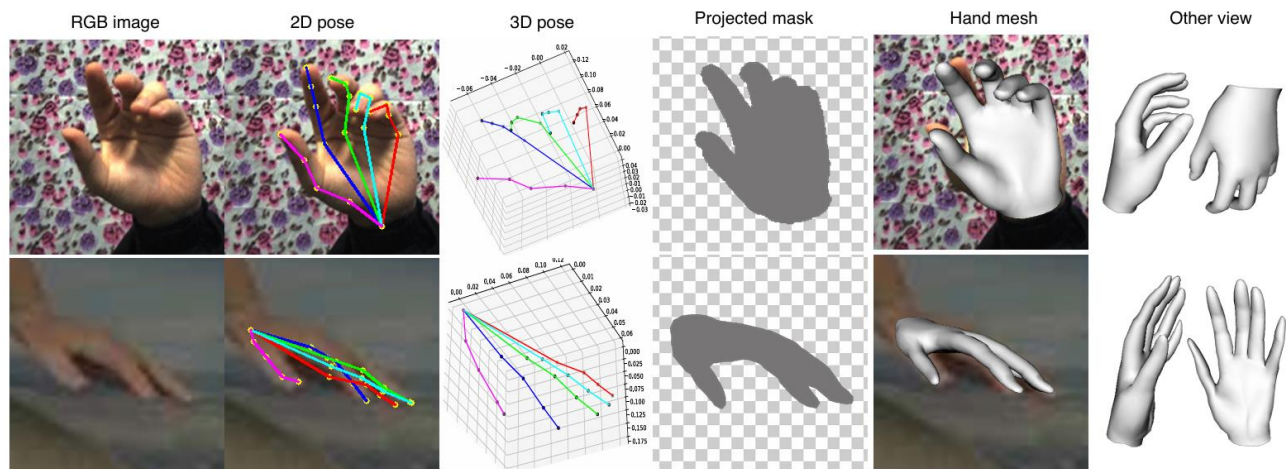


Figure 3. Hand pose estimation and mesh recovery from single-view RGB image²¹

Zhang et al.²¹ introduced an end-to-end hand mesh recovery framework (HAMR) that achieves richer mesh representations through the parametric MANO model²². Specifically, the hand mesh calculates 3D joint positions through linear interpolation and obtains 2D joint positions through the projection of 2D joints. Additionally, a differentiable reprojection loss is defined based on the difference between the defined representation and the ground truth labels, enabling end-to-end training of the framework.

Although the aforementioned methods can provide relatively accurate hand estimates, they perform poorly in complex scenes, especially under strong interactions and occlusions, often leading to optimization failures or local optimal solutions, thereby affecting the accuracy and stability of the estimates.

Reducing annotation information and lowering the difficulty of data acquisition has always been an important and challenging direction in 3D hand pose estimation. Kulon et al.²³ proposed a weakly supervised mesh convolution method for estimating 3D hand poses from monocular RGB images. This method collects a large-scale hand motion YouTube video dataset and uses these data as weak supervision signals. Specifically, OpenPose is first used to detect 2D keypoints, and then the MANO model is iteratively fitted to generate 3D hand meshes. In this way, the method not only improves the model's generalization ability but also significantly outperforms existing methods in public benchmark tests.

The work of Potamias et al.²⁴ represents the current state-of-the-art (SOTA). The paper proposes an end-to-end method for wild 3D hand localization and reconstruction called WiLoR. The framework consists of two parts: real-time fully

convolutional hand localization and a high-fidelity 3D hand reconstruction model based on Transformer. By introducing a large-scale wild hand image dataset WHIM, a lightweight and high-precision hand detection model is trained, which can robustly detect hands at a speed of over 110 FPS under various occlusions and lighting conditions. WiLoR first localizes and defines the orientation of the detected hand and then uses a Transformer-based hand pose estimation module to elevate it to 3D. To assist high-fidelity reconstruction and promote image alignment, a refinement module is introduced to extract local features to correct misaligned poses. WiLoR achieves the most advanced performance on different benchmark datasets and improves the temporal coherence of image-based 3D hand pose estimation methods.

3.1.2 Multi-view Method

Multi-view methods enhance the accuracy of pose estimation by integrating images from multiple perspectives, overcoming issues such as viewpoint limitations and depth ambiguity present in single-view scenarios. The following sections introduce several representative methods:

Zheng et al.²⁵ introduced the HaMuCo framework, a significant advancement in the field of multi-view 3D hand pose estimation. This framework employs self-supervised learning to train a monocular hand pose estimator from pseudo 2D labels derived from multi-view images. The core of HaMuCo is the introduction of a cross-view interaction network, which leverages cross-view correlated features and enforces multi-view consistency to facilitate collaborative learning. This approach addresses the issues of noisy labels and "groupthink" effects inherent in self-supervised learning. HaMuCo not only achieves state-of-the-art performance in multi-view self-supervised hand pose estimation but also demonstrates superior performance when applied to multi-view input hand pose estimation compared to previous methods.

However, HaMuCo faces certain limitations when dealing with more challenging scenarios such as hand-object interactions or relaxed synchronization constraints for multi-view inputs. Liu et al.²⁶ proposed a novel solution called S2DHand, which adapts pre-trained monocular estimators to a bi-view setup without the need for multi-view annotations or camera parameters. Unlike HaMuCo, S2DHand does not require expensive multi-view annotations and can handle arbitrary bi-view pairs with unknown camera parameters, making the model adaptable to a diverse range of camera settings. S2DHand is based on stereo constraints, including view consensus and transformation invariance, and generates pseudo labels in a complementary manner to achieve reliable adaptation.

3.1.3 Video-Based Method

Video-based 3D hand pose estimation involves estimating the 3D joint positions and poses of the hand from consecutive video frames. Compared to pose estimation from single images, video data provides rich spatiotemporal information that can effectively address issues such as self-occlusion and depth uncertainty present in single frames, thereby enhancing the accuracy and robustness of pose estimation.

Cai et al.²⁷ proposed a method based on graph convolutional networks¹⁹ that combines spatial dependencies and temporal consistency to improve the accuracy of 3D pose estimation. This method effectively embeds spatial and temporal relationships into the framework for learning. The paper explicitly incorporates domain knowledge of human hand configurations into graph convolution operations and designs a local-to-global network architecture capable of learning multi-scale feature representations.

To handle large-scale unlabeled data, Liu et al.²⁸ introduced a semi-supervised learning framework that uses Transformer for explicit contextual reasoning. This approach transcends the limited 3D annotations within individual images by leveraging spatiotemporal consistency in large-scale hand-object videos as a constraint to generate pseudo labels for semi-supervised learning. The authors first train an initial model and then deploy it on a large-scale video dataset to collect pseudo labels. In this way, the method not only improves the performance of hand pose estimation on challenging real-world datasets but also significantly enhances the accuracy of object pose estimation, especially when there are fewer ground truth values per instance.

Wen et al.²⁹ proposed a hierarchical framework based on Transformer for simultaneously addressing 3D hand pose estimation and action recognition tasks from first-person RGB videos. The framework consists of two cascaded Transformer encoders, with the first encoder utilizing short-term temporal cues for hand pose estimation and the second encoder leveraging long-term temporal cues for action recognition. This design fully exploits the different temporal granularities of the tasks, thereby improving the performance of the model.

3.2 Hand Pose Estimation of Depth Information

Compared to 2D RGB images, depth maps, point clouds, and voxels possess higher-dimensional information, which results in smaller errors during the 3D reconstruction process in hand pose estimation (for example, it is difficult to determine whether a hand is interacting with other objects without depth information). Among these three types of information media, depth images have the lowest acquisition cost, but they can be either single-view or multi-view; point clouds can be understood as 3D pixels, which are the smallest units of spatial partitioning; point clouds are captured by 3D scanners, and each point contains data such as 3D coordinates, color, and reflectance.

This section will introduce the overview of hand pose estimation methods based on depth information using several representative works as examples. Additionally, since the datasets used for hand pose estimation based on depth information are relatively uniform, Table 1 compares the performance of various methods on four datasets (NYU³⁰, ICVL³¹ and MSRA³²).

Ge et al.³³ based their work on point cloud data to regress the 3D coordinates of hand joints. By normalizing the point cloud data, they obtained the global orientation of the hand. On the basis of the originally estimated fingertip coordinates, they refined the coordinates using the PointNet network³⁴, taking the neighboring points of the fingertip coordinates as input to obtain more accurate fingertip coordinates.

Moon et al.³⁵ proposed a voxel-to-voxel network that uses a Transformer-based encoder-decoder model to directly predict the 3D positions of hand joints. They employed a two-stage reasoning method to achieve smooth context translation and introduced a new loss term to improve the accuracy of training and inference.

Xiong et al.³⁶ introduced the Anchor-to-Joint regression network (A2J) for 3D hand and body pose estimation from depth images. As shown in Figure 3, A2J captures global-local spatial context information by setting dense anchors on the depth image and predicts joint positions in an integrated manner to enhance generalization ability. Through background segmentation, hand localization, voxelization of depth images, and data augmentation, the training and computational requirements were significantly reduced.

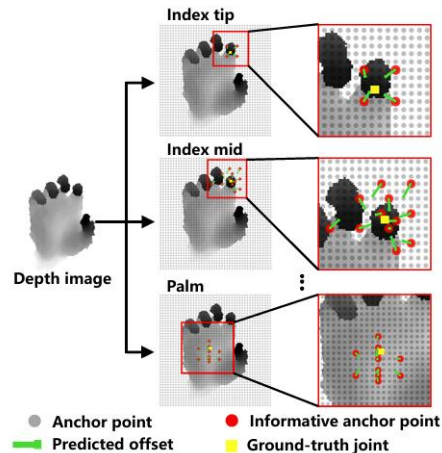


Figure 4. Xiong et al.³⁶ set dense anchors on the hand mesh.

Huang et al.³⁷ combined the advantages of detection-based and regression-based methods, estimating hand joint coordinates through discrete integration of all pixels' dense representations guided by an adaptive weight map. This method achieves fast and robust 3D hand pose estimation from a single depth image and has achieved state-of-the-art performance on multiple datasets, with advantages in practical applications.

Ren et al.³⁸ proposed a strongly self-supervised framework for 3D hand pose and mesh estimation based on depth images. The network aggregate's multi-view information, encoding diverse semantic information from each view into multiple compact nodes and modeling dependencies between different views. Additionally, they introduced a pseudo multi-view training strategy that can extend the framework to more general scenarios using only single-view training data.

Cheng et al.³⁹ proposed an effective virtual view selection and fusion framework to improve the accuracy and robustness of 3D hand pose estimation from a single depth image. Through virtual multi-view rendering and 3D pose fusion of each

view, accuracy is enhanced. Moreover, a confidence network built using multi-head attention mechanisms is used to improve the accuracy of 3D reconstruction during fusion and view selection. To efficiently obtain confidence in each view, Cheng also proposed a lightweight confidence network obtained through network distillation.

Table 1. Comparison of hand pose estimation algorithms based on depth information across three datasets, using average error as the metric.

Methods	Year	NYU ⁴⁰	ICVL ³¹	MSRA ³²
HandPointNet ³³	2018	10.5	6.935	8.5
V2V ³⁵	2018	8.42	6.28	7.49
A2J ³⁶	2019	8.61	6.46	-
AWR ³⁷	2020	7.37	5.98	7.2
MMV ³⁸ (self-supervised)	2022	12.91	14.44	12.97-
VVS ³⁹	2022	6.4	4.79	-

4. DATASETS

This section will introduce 17 influential datasets in the field of human hand pose estimation, including data from various dimensions such as multi-view, depth cameras, interactions between two hands, interactions between hands and objects, and challenging environments. The publication dates, data types, frame counts, and relevant link information for each dataset are presented in Table 2.

4.1 InterHand2.6M

InterHand2.6M⁴¹ is a large-scale dataset containing approximately 2.6 million annotated single-hand and two-hand interaction frames from multiple subjects in various poses. Due to its high-quality captured images and accurate 3D interaction annotations, the InterHand2.6M dataset is widely used in hand pose estimation.

4.2 ICVL

The Imperial College Vision Lab (ICVL)³¹ dataset is one of the oldest in the field of hand pose estimation, comprising 180,000 annotated hand depth image frames with 10 subjects performing 26 different poses. The annotations use a hand model with 16 joints.

4.3 3D Hand Pose

3D Hand Pose⁴² is a multi-view RGB image dataset for hand poses, where each image is annotated with hand bounding boxes, 2D and 3D joint positions. The dataset contains 20,500 frames distributed across 21 sequences, with each frame including four RGB images from different viewpoints and nine different annotations.

4.4 NYU Hand

The New York University⁴⁰ hand dataset includes 8,252 test sets with a total of 72,757 RGB-D images and ground truth annotations. The annotations are derived from three different views of the Kinect RGB-D images and synthetic models of hand poses from different views.

4.5 ARCTIC

Articulated Objects in Free-form Hand Interaction⁴³ dataset contains 2.1 million video frames, with each frame corresponding to precise 3D hand and object mesh model information.

4.6 First-Person Hand Action Benchmark

Garcia-Hernando et al.⁴⁴ provides a first-person RGB-D dynamic hand action dataset, consisting of over 100,000 frames of 45 common hand actions and 26 different hand objects. Garcia-Hernando used six magnetic sensors to capture real movements and combined inverse kinematics algorithms to infer the 3D positions of 21 joints in the hand model, while also recording 6D object poses.

4.7 MSRA Hands

Microsoft Research Asia Hands³² is a hand tracking dataset constructed by the Intel Creative Interactive Hand Camera, capturing the right hands of six subjects. Each subject performs various actions in a sequence of 400 frames, with annotations for the 3D positions of 21 hand joints.

4.8 EgoDexter

EgoDexter³⁰ includes 3,190 video sequences and corresponding 2D and 3D pose annotations from four test sequences. The videos were recorded from an egocentric perspective using an RGB-D onboard camera, with factors such as cluttered backgrounds, rapid camera movements, and object interactions. The annotations include 3D position information for 21 keypoints.

4.9 ContactArt

ContactArt⁴⁵ hand-object interaction dataset provides a more convenient system, allowing operators to record motion data using smartphones. The dataset records over 550,000 RGB images, with annotations including hand poses, hand bounding boxes, multi-view images, and hand-object interaction labels.

4.10 BigHand2.2M

BigHand2.2M⁴⁶ dataset contains 2.2 million depth images. Yuan et al. installed six magnetic sensors on the hand (one on each finger and one on the back of the hand) to capture precise 6D information, and then used inverse kinematics to reconstruct a human hand model with 31 degrees of freedom.

4.11 DexYCB

The hand grasping dataset DexYCB⁴⁷ was recorded by eight RealSense cameras from eight viewpoints, capturing depth images of ten subjects interacting with 20 objects. The dataset includes over 1,000 sequences and 580,000 frames.

4.12 HOI4D

Human-Object Interaction⁴⁸ dataset contains 2.4 million frames of egocentric RGB-D video data, with over 4,000 sequences recorded by four subjects interacting with 800 different object instances from 610 different indoor rooms across 16 categories. It also provides full-view segmentation, motion segmentation, 3D hand poses, category-level object poses, and hand action frame annotations, as well as reconstructed object meshes and scene point clouds.

4.13 HO-3D

Hand and Object Poses⁴⁹ is an unlabelled RGB hand-object interaction image dataset with 3D annotations for hands and objects. It contains 68 sequences with a total of 77,558 frames (66,034 training set and 11,524 test set), involving 10 people and 10 categories of objects.

Table 2. Statistical information of the thirteen datasets

Dataset	Release Date	Data Type	Frames	Link
InterHand2.6M ⁴¹	2020	RGB	2.6M	https://mks0601.github.io/InterHand2.6M/
ICVL ³¹	2014	Depth	18K	https://labicvl.github.io/Datasets_Code.html
3D Hand Pose ⁴²	2019	RGB	20K	https://www.rovit.ua.es/dataset/mhpdataset/
NYU ⁴⁰	2014	Depth	243K	https://jonathantompson.github.io/NYU_Hand_Pose_Dataset.htm

ARCTIC ⁴³	2023	RGB	2.1M	https://arctic.is.tue.mpg.de/
First-Person Hand Action ⁴⁴	2018	RGB	100K	https://kcvl-kaist.github.io/FPHA/
MSRA ³²	2015	Depth	76K	https://jimmysuen.github.io/
EgoDexter ³⁰	2017	RGB+Depth	3K	https://handtracker.mpi-inf.mpg.de/projects/OccludedHands/EgoDexter.htm
ContactArt ⁴⁵	2024	RGB	550K	https://zehaozhu.github.io/ContactArt/
BigHand2.2M ⁴⁶	2017	Depth	2.2M	https://kcvl-kaist.github.io/bighand2.2M/
DexYCB ⁴⁷	2021	Depth	580K	https://dex-ycb.github.io/
HOI4D ⁴⁸	2022	RGB+Depth	2.4M	https://hoi4d.github.io/
HO-3D ⁴⁹	2020	RGB	76K	https://www.tugraz.at/index.php?id=40231

5. CONCLUSION

The advent of deep learning has revolutionized the approach to hand pose estimation, enabling more accurate and robust methods that can handle complex scenarios such as occlusions, varying lighting conditions, and dynamic backgrounds. The development of large-scale datasets has further fueled progress by providing the necessary data for training in sophisticated models. These datasets, combined with advances in multi-view and depth-based techniques, have led to significant improvements in the accuracy and generalization of hand pose estimation algorithms.

Despite these advancements, several challenges remain. Handling occlusions, ensuring real-time performance, and enhancing the generalization capabilities of models to new environments and unseen data are areas that require further research. Additionally, the estimation of poses for multiple interacting hands and the integration of cross-modal data present opportunities for future work.

As we look ahead, the continued development of more efficient and accurate hand pose estimation methods will play a crucial role in the advancement of human-computer interaction, virtual and augmented reality, and robotics. By addressing the current challenges and leveraging the potential of emerging technologies, researchers can pave the way for more natural, intuitive, and immersive interactive experiences. The future of hand pose estimation holds great promise, and we anticipate that ongoing research will lead to breakthroughs that will further bridge the gap between human and machine interactions.

REFERENCES

- [1] Sharma, Adwait, Joan Sol Roo, and Jürgen Steimle. "Grasping Microgestures: Eliciting Single-Hand Microgestures for Handheld Objects." Paper presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.
- [2] Lee, Taehee, and Tobias Hollerer. "Multithreaded Hybrid Feature Tracking for Markerless Augmented Reality." *IEEE transactions on visualization and computer graphics* 15, no. 3 (2009): 355-68.
- [3] Piumsomboon, Thammathip, Adrian Clark, Mark Billingham, and Andy Cockburn. "User-Defined Gestures for Augmented Reality." In *Chi'13 Extended Abstracts on Human Factors in Computing Systems*, 955-60, 2013.
- [4] Guleryuz, Onur G, and Christine Kaeser-Chen. "Fast Lifting for 3d Hand Pose Estimation in Ar/Vr Applications." Paper presented at the 2018 25th IEEE International Conference on Image Processing (ICIP), 2018.

- [5] MacCormick, John, and Michael Isard. "Partitioned Sampling, Articulated Objects, and Interface-Quality Hand Tracking." Paper presented at the Computer Vision—ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II 6, 2000.
- [6] Martin, Jérôme, Vincent Devin, and James L Crowley. "Active Hand Tracking." Paper presented at the Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998.
- [7] O'Hagan, RG, Alexander Zelinsky, and Sebastien Rougeaux. "Visual Gesture Interfaces for Virtual Environments." *Interacting with Computers* 14, no. 3 (2002): 231-50.
- [8] Sato, Yoichi, Makiko Saito, and Hideki Koike. "Real-Time Input of 3d Pose and Gestures of a User's Hand and Its Applications for Hci." Paper presented at the Proceedings IEEE Virtual Reality 2001, 2001.
- [9] Segen, Jakub, and Senthil Kumar. "Gesture Vr: Vision-Based 3d Hand Interace for Spatial Interaction." Paper presented at the Proceedings of the sixth ACM international conference on Multimedia, 1998.
- [10] Wu, Ying, Qiong Liu, and Thomas S Huang. "An Adaptive Self-Organizing Color Segmentation Algorithm with Application to Robust Real-Time Human Hand Localization." Paper presented at the Proceedings of Asian Conference on Computer Vision, 2000.
- [11] Ong, Eng-Jon, and Richard Bowden. "A Boosted Classifier Tree for Hand Shape Detection." Paper presented at the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings., 2004.
- [12] Grudin, Jonathan. "Integrating Paper and Digital Information on Enhanceddesk: A Method for Realtime Finger Tracking on an Augmented Desk System." *ACM Transactions on Computer-Human Interaction (TOCHI)* 8, no. 4 (2001): 307-22.
- [13] Oka, Kenji, Yoichi Sato, and Hideki Koike. "Real-Time Tracking of Multiple Fingertips and Gesture Recognition for Augmented Desk Interface Systems." Paper presented at the Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, 2002.
- [14] Utsumi, Akira, and Jun Ohya. "Multiple-Hand-Gesture Tracking Using Multiple Cameras." Paper presented at the Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), 1999.
- [15] Deutscher, Jonathan, Ben North, Benedicte Bascle, and Andrew Blake. "Tracking through Singularities and Discontinuities by Random Sampling." Paper presented at the Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999.
- [16] Heap, Tony, and David Hogg. "Wormholes in Shape Space: Tracking through Discontinuous Changes in Shape." Paper presented at the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), 1998.
- [17] Arulampalam, M Sanjeev, Simon Maskell, Neil Gordon, and Tim Clapp. "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking." *IEEE Transactions on signal processing* 50, no. 2 (2002): 174-88.
- [18] Ge, Liuhao, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. "3d Hand Shape and Pose Estimation from a Single Rgb Image." Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.
- [19] Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst. "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering." *Advances in neural information processing systems* 29 (2016).
- [20] Spurr, Adrian, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. "Self-Supervised 3d Hand Pose Estimation from Monocular Rgb Via Contrastive Learning." Paper presented at the Proceedings of the IEEE/CVF international conference on computer vision, 2021.
- [21] Zhang, Xiong, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. "End-to-End Hand Mesh Recovery from a Monocular Rgb Image." Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
- [22] Romero, Javier, Dimitrios Tzionas, and Michael J Black. "Embodied Hands: Modeling and Capturing Hands and Bodies Together." *arXiv preprint arXiv:2201.02610* (2022).

- [23]Kulon, Dominik, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. "Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild." Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.
- [24]Potamias, Rolandos Alexandros, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. "Wilor: End-to-End 3d Hand Localization and Reconstruction in-the-Wild." arXiv preprint arXiv:2409.12259 (2024).
- [25]Zheng, Xiaozheng, Chao Wen, Zhou Xue, Pengfei Ren, and Jingyu Wang. "Hamuco: Hand Pose Estimation Via Multiview Collaborative Self-Supervised Learning." Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.
- [26]Liu, Ruicong, Takehiko Ohkawa, Mingfang Zhang, and Yoichi Sato. "Single-to-Dual-View Adaptation for Egocentric 3d Hand Pose Estimation." Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
- [27]Cai, Yujun, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. "Exploiting Spatial-Temporal Relationships for 3d Pose Estimation Via Graph Convolutional Networks." Paper presented at the Proceedings of the IEEE/CVF international conference on computer vision, 2019.
- [28]Liu, Shaowei, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. "Semi-Supervised 3d Hand-Object Poses Estimation with Interactions in Time." Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
- [29]Wen, Yilin, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. "Hierarchical Temporal Transformer for 3d Hand Pose Estimation and Action Recognition from Egocentric Rgb Videos." Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023.
- [30]Mueller, Franziska, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. "Real-Time Hand Tracking under Occlusion from an Egocentric Rgb-D Sensor." Paper presented at the Proceedings of the IEEE international conference on computer vision, 2017.
- [31]Tang, Danhang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. "Latent Regression Forest: Structured Estimation of 3d Articulated Hand Posture." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.
- [32]Qian, Chen, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. "Realtime and Robust Hand Tracking from Depth." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.
- [33]Ge, Liuhao, Yujun Cai, Junwu Weng, and Junsong Yuan. "Hand Pointnet: 3d Hand Pose Estimation Using Point Sets." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
- [34]Qi, Charles R, Hao Su, Kaichun Mo, and Leonidas J Guibas. "Pointnet: Deep Learning on Point Sets for 3d Classification and Segmentation." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
- [35]Moon, Gyeongsik, Ju Yong Chang, and Kyoung Mu Lee. "V2v-Posenet: Voxel-to-Voxel Prediction Network for Accurate 3d Hand and Human Pose Estimation from a Single Depth Map." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern Recognition, 2018.
- [36]Xiong, Fu, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. "A2j: Anchor-to-Joint Regression Network for 3d Articulated Pose Estimation from a Single Depth Image." Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
- [37]Huang, Weiting, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. "Awr: Adaptive Weighting Regression for 3d Hand Pose Estimation." Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence, 2020.
- [38]Ren, Pengfei, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, and Jianxin Liao. "Mining Multi-View Information: A Strong Self-Supervised Framework for Depth-Based 3d Hand Pose and Mesh Estimation." Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022.

- [39] Cheng, Jian, Yanguang Wan, Dexin Zuo, Cuixia Ma, Jian Gu, Ping Tan, Hongan Wang, Xiaoming Deng, and Yinda Zhang. "Efficient Virtual View Selection for 3d Hand Pose Estimation." Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence, 2022.
- [40] Tompson, Jonathan, Murphy Stein, Yann Lecun, and Ken Perlin. "Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks." *ACM Transactions on Graphics (ToG)* 33, no. 5 (2014): 1-10.
- [41] Moon, Gyeongsik, Shou-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." Paper presented at the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, 2020.
- [42] Gomez-Donoso, Francisco, Sergio Orts-Escolano, and Miguel Cazorla. "Large-Scale Multiview 3d Hand Pose Dataset." *Image and Vision Computing* 81 (2019): 25-33.
- [43] Fan, Zicong, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. "Arctic: A Dataset for Dexterous Bimanual Hand-Object Manipulation." Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
- [44] Garcia-Hernando, Guillermo, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. "First-Person Hand Action Benchmark with Rgb-D Videos and 3d Hand Pose Annotations." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
- [45] Zhu, Zehao, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. "Contactart: Learning 3d Interaction Priors for Category-Level Articulated Object and Hand Poses Estimation." Paper presented at the 2024 International Conference on 3D Vision (3DV), 2024
- [46] Yuan, Shanxin, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. "Bighand2. 2m Benchmark: Hand Pose Dataset and State of the Art Analysis." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
- [47] Chao, Yu-Wei, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, et al. "Dexycb: A Benchmark for Capturing Hand Grasping of Objects." Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
- [48] Liu, Yunze, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, et al. "Hoi4d: A 4d Egocentric Dataset for Category-Level Human-Object Interaction." Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
- [49] Hampali, Shreyas, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. "Honnotate: A Method for 3d Annotation of Hand and Object Poses." Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.