공학석사 학위논문

A Wrist Mounted Camera System for Hand Pose Estimation from Single RGB Images

단일 이미지 기반 손 자세 추정을 위한
손목 장착 카메라 시스템

2023 년 12월

서울대학교 대학원

컴퓨터공학부

최연우

A Wrist Mounted Camera System for Hand Pose Estimation from Single RGB Images

지도 교수 주한별

이 논문을 공학석사 학위논문으로 제출함
2023 년 12 월

서울대학교 대학원
컴퓨터공학부
최연우

최연우의 공학석사 학위논문을 인준함
2024 년 1 월

위 원 장 김 선 (인)

부위원장 주한별 (인)

위 원 장병탁 (인)

Abstract

In the field of egocentric vision, the accurate estimation of 3D hand poses is critical, particularly in applications such as augmented and virtual reality, and human-computer interaction. This thesis introduces a methodology to enhance the precision and stability of 3D hand pose estimation from single RGB images, employing a novel pseudo-ground-truth data capture system with wrist and head mounted cameras. Our approach takes advantage of the unique perspectives offered by this multi-camera setup to address the challenges of occlusion and the high degrees of freedom in hand movements.

The core contributions of our thesis are manifold. Firstly, we present an innovative pseudo-ground-truth data capture system utilizing wrist-mounted cameras, to gather accurate hand data. This data is then used to train previous state-of-the-art models, leading to significant performance enhancements. Secondly, we initiate bounding box detection from the wrist cameras rather than the conventional single-head camera view, enabling more consistent and precise tracking of hand movements.

Our experimental results demonstrate the superiority of our approach over current state-of-the-art techniques. We achieve marked improvements in the accuracy of visible point estimation and the stability of occluded point prediction. The employment of wrist-mounted cameras for bounding box detection and the generation of pseudo-ground-truth data from these cameras significantly bolster the robustness of hand pose estimation methods. Our findings herald a future of more precise and stable 3D hand pose estimation in egocentric scenarios, facilitating the development of more natural and intuitive user interfaces in immersive technologies.

Keywords: 3D Hand Pose Estimation, Egocentric Vision, Computer Vision
Student Number: 2022-28614

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

3D hand pose estimation is a fundamental component for interactive applications, critically underpinning advancements in augmented and virtual reality, as well as human-computer interaction. The task is notably challenging due to the hand's complex movements and the occurrence of occlusions that often obscure essential keypoints. Efforts within this research domain aim to refine the precision of visible hand keypoints and to bolster the reliability of occluded point predictions.

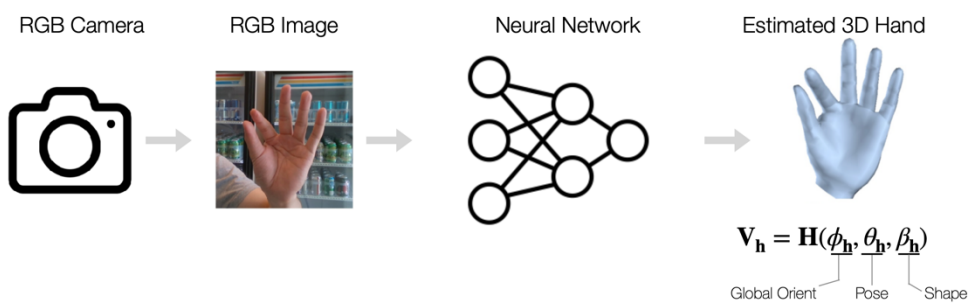

Figure 1.1 Monocular 3D Hand Pose Estimation

This thesis rigorously examines the established pipeline of monocular 3D hand pose estimation, as portrayed in Figure 1.1. The figure depicts the standard pipeline of monocular 3D hand pose estimation, commencing with an RGB camera capturing the image, progressing through neural network analysis, and culminating with the generated 3D hand pose. A neural network's success in this domain is intimately linked to the quality of the training dataset. To this end, we present an innovative data-centric strategy that enhances the data acquisition process integral to monocular 3D hand pose estimation. By integrating a data capture system utilizing wrist-mounted and head cameras, our approach facilitates the collection of pseudo-ground-truth data that is invaluable for training cutting-edge models. Although our annotations may not attain the precision of those obtained from motion capture suits or multi-camera systems, the agility and simplicity of our setup allow for rapid data collection in diverse environments, including those 'in the wild.' This capability significantly expands the breadth and depth of training data, thereby enhancing model performance.

By leveraging the high-fidelity data from our easy-to-deploy capture system, we have recognized noticeable improvements in the accuracy and stability of hand pose estimations, as verified by our empirical results. This advancement not only quantitatively surpasses current state-of-the-art models but also qualitatively enhances the interpretability of hand gestures, enabling for more natural and intuitive user interfaces in a range of immersive technological applications.

Chapter 2

Related Works

2.1 Monocular RGB Hand Motion Capture

Recent explorations into the field of monocular RGB 3D hand pose capture have been motivated by the inherent complexity of depth ambiguity. Zimmermann and Brox [1] develop a CNN that deduces 3D joint coordinates straight from RGB images. A novel approach by Iqbal et al. [2] introduce a 2.5D heatmap approach that merges 2D joint placement with depth cues, substantially enhancing precision. Multiple studies are adopting diverse image datasets [3,4,5] to expand training variability, aiming for broader generalization. Mueller et al. [6] curate an extensive rendered dataset with CycleGAN post-processing to mitigate domain discrepancies. However, their efforts concentrate on pinpointing joint locations without addressing the rotation of joints, a critical factor for animating hand meshes. Attempts to compute joint rotations [7,8], involve fitting a prototypical hand model to the estimations through a repetitive optimization

process, which is computationally intensive and dependent on elaborate energy functions. Proposals to deduce deformable hand mesh parameters directly from images in a unified process [9, 10] have been made, but such rotation estimations are subject to weak supervision, leading to compromised precision. Ge et al. propose GraphCNN [12] to directly infer a hand mesh, although the necessity of a specialized dataset with accurate hand meshes presents a substantial challenge. Rong et al. introduced Frankmocap [40], a system that draws upon earlier advancements in human body mesh recovery [60]. It utilizes the MANO [31] hand model to infer hand pose and shape directly from a single RGB image. Subsequent studies [61, 62] have adopted a more direct method, choosing to estimate the hand mesh vertices directly. Although this approach typically yields results that are more consistent with visual observations, it tends to be less reliable when dealing with occlusions or incomplete views of the hand. Progress in 3D hand pose estimation has spurred advancements in related fields as well, such as simultaneous hand poses and object interaction modeling [63,64], and the intricate task of capturing the interaction between two hands [65,66,67,68]. In our research, we focus on enhancing the reliability of these regression-based methods. With our cost-effective and rapid techniques for collecting pseudo-ground-truth data through our system, we can integrate our findings with previous method. This integration has the potential to significantly boost performance and accuracy in the field of 3D hand pose estimation.

2.2 Datasets for 3D Hand Pose Estimation

Regarding datasets for hand pose estimation, Table 2.1 compiles comprehensive data on the existing 3D hand pose datasets. Early egocentric pose estimation methodologies annotated 2D points on depth snapshots [15] or employed magnetic markers [21]. Nevertheless, the precision of these benchmarks is limited by sensor noise and the extensive labor required for annotations. Hence,

the focus of most research in 3D hand pose estimation has been on utilizing static exocentric cameras [5,6,22,23,24,25,26,27] or leveraging such datasets to refine egocentric hand pose predictions [28]. Multi-camera setups offer clear benefits, as discussed widely in literature [29]. The quantity of images enhances with the number of cameras, exemplified by InterHand2.6M [13] with its extensive view range. Triangulation from multiple 2D keypoints [6,24] or template fitting [5,22,27,30] (like using the MANO model [31]) ensures accurate 3D keypoint annotation. Synchronized egocentric and exocentric camera systems have been introduced in recent activity datasets like Assembly101 [32] and H2O [30], which can significantly alleviate the annotation process.

Table 2.1 Existing Datasets for 3D Hand Pose Estimation

	Modality	#img	#views	Annotation approach
InterHand2.6M	RGB-D	2.59M	80-140	Manual + 2D triangulation
Assembly Hands	RGB	2.81M	8+4(ego)	Manual + 3D volume + refinement
EgoDexter	RGB-D	3K	1(ego)	Manual
FreiHAND	RGB	37K	8	Manual + 3D volume + template fitting
HO3D	RGB-D	103K	5	2D + template fitting
DexYCB	RGB-D	508K	8	Manual + template fitting
Panoptic Studio	RGB	15K	31	2D + triangulation
H20	RGB-D	571K	4+1(ego)	2D + template fitting + smoothing

Chapter 3

Method

3.1 System Setup

In our egocentric vision framework, we leverage the distinct perspectives provided by a combination of head-mounted and wrist-mounted cameras to enhance the fidelity of 3D hand reconstruction. With the prevalence of mobile phones and wide-angle action cameras, such as those in the GoPro series, radial lens distortion has become a common artifact. To counteract this, we employ the single-parameter division model [33] to correct for radial distortion, chosen for its simplicity and effectiveness. Our setup comprises three GoPro Hero 11 cameras: one mounted on the user's head and one on each wrist, using adjustable mounts for optimal positioning flexibility. Figure 3.1 illustrates our system setup and displays a sample capture from the three synchronized cameras. After evaluating the trade-offs between output quality and computational expense, we

selected a capture resolution of FHD at 30 frames per second as our optimal recording setting.

Data capture sessions spanned various everyday environments, such as offices, kitchens, bathrooms, dining areas, bedrooms, classrooms, and outdoor locations, with some scenarios depicted in Figure 3.2. During the application of structure-from-motion (SfM) [34] techniques for scene reconstruction and the perspective-n-point (PnP) [35] algorithm for camera localization, challenges arose in feature-scarce scenes that compromised the quality of reconstruction and camera localization. To overcome this, we utilized printed patterns on A4 paper, strategically placed within the scenes to support feature detection based on SIFT [36]. This method substantially improved the robustness of our system. In total, we collected 60 sequences, each 1-2 minutes in duration, resulting in approximately 180,000 frames captured across 15 distinct settings.

Figure 3.1. The setup of our system. The left section depicts the capture environment, the middle section highlights the GoPro cameras attached to their mounts, and the right section demonstrates a synchronized capture from all three cameras.

Figure 3.2 Images showcasing environments captured using our system, displayed sequentially. On the left, a scene from an office setting; in the center, a kitchen environment; and on the right, a storage room. Each image represents the distinct characteristics and details of the respective environments as captured by our system.

3.2 Enhancing Hand Bounding Box Detection

Recent advancements in hand pose estimation techniques [37,38,39,40] primarily focus on detecting a hand's bounding box within an image, subsequently cropping this region for regression. This method is crucial as irrelevant regions can disrupt accurate hand pose estimation. Hence, precise bounding box detecting is critical for effective 3D hand pose estimation. In our pseudo-ground-truth generation pipeline, detailed in the subsequent section, we employ 3D hand pose estimators to estimation the initial 3D hand pose from the captured data. Extensive testing with various pose detectors revealed that the quality of the reconstruction significantly diminishes if the bounding box is inaccurately detected evident in cases of incorrect hand identification (left/right), improper region selection, or failure to detect the hand. Therefore, precise 3D hand pose extraction is vital to our pipeline, and we have addressed these bounding box issues to enhance accuracy.

Our system setup, involving attaching two cameras to the user's wrists, inherently improves the precision of bounding box detection. These cameras consistently face the hands in a fixed position, confining the movement of the hands to a specific area within the camera frame. Furthermore, leveraging the known camera parameters of both head and wrist-mounted cameras, we can

effectively transfer the bounding box detected by the wrist cameras to the head camera, ensuring consistent and accurate hand pose estimation across different viewpoints.

We initially identify the bounding boxes of the hands, denoted as $B_{coord} = \{p1, p2, p3, p4\}$ where each $p_i = (u_i, v_i)$ represents the image pixel coordinates. Specifically, we detect the left hand using the left wrist camera, and the right hand using the right wrist camera, utilizing the 100DOH model by Shan et al. [41]. Subsequently, we apply an optical flow-based tracking method developed by Teed et al. [42] to consistently track these bounding boxes throughout the video sequence. Optical flow-based methods, effective though they are, exhibit certain weaknesses, especially in scenarios involving rapid movements or blurred imagery. Additionally, these methods can be prone to the accumulation of small errors over time, potentially leading to diminished tracking accuracy. However, with the benefits of our system mentioned above, we observed robust tracking throughout the sequence.

With the camera intrinsic parameters I_{wrist} of the wrist cameras, each 2D point is back-projected to form 3D rays in the camera coordinates. We define π the projection function that projects 3D points in camera coordinates back to 2D image coordinates, K the intrinsic matrix of the camera. Introducing the function $R(d) = \begin{bmatrix} x(d) \\ y(d) \\ z(d) \end{bmatrix}$ to represent the 3D coordinates at depth d the optimization problem can be reformulated as in Equation 1.

$$\min_{d} |\pi(K \times [\boldsymbol{R}(d), 1]^T) - [u, v, 1]^T|^2 \quad (1)$$

Now that we have the optimized d, we can now find the 3D bounding box points in the wrist camera's coordinates as in Equation 2.

$$P_{3Dwrist} = d \times \mathrm{I}_{\mathrm{wrist}}^{-1} \times \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (2)$$

We can then transform the 3D bounding box points in the camera coordinates into the world coordinate system using the wrist camera's extrinsic matrix E_{wrists} by Equation 3.

$$P_{3Dworld} = E_{wrist} \times P_{3Dwrist} \quad (3)$$

For perceiving the 3D bounding box from the head camera's viewpoint, the world coordinates are translated to the head camera's coordinate space using its extrinsic matrix, then the 3D points are projected onto the head camera's plane using its intrinsic matrix, I_{head} as shown in Equation 4.

$$B_{head} = I_{head} \times E_{head}^{-1} \times P_{3Dworld} \quad (4)$$

3.3 Generating Pseudo Ground Truth Data

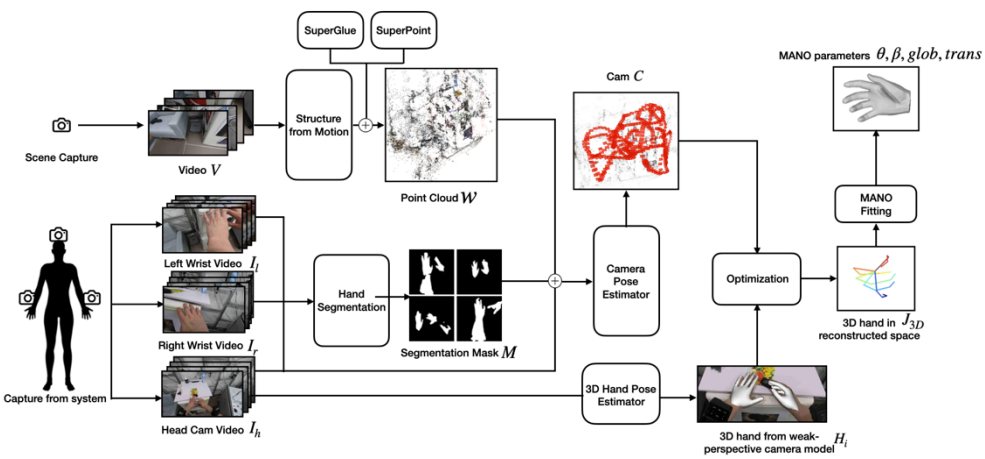

Figure 3.3. Our system's pipeline for generating pseudo ground truth data. Beginning with captured RGB hand data, the process obtains 3D hand joint coordinates and the extraction of corresponding MANO hand pose parameters

3.3.1 Data Capture, Preprocessing

To generate a sequence of hand data, we initially capture a 1-2 minutes video within the scene of action and extract the frames V_{scene} to reconstruct the world coordinate space. Subsequently, using our tri-camera setup, we capture hand data from the left, right, and head cameras each noted as V_l, V_r, V_h extract frames, and manually synchronize the videos via clapping sound data. Employing V_{scene}, we reconstruct the 3D point cloud space W through SfM. For enhanced scene reconstruction, we utilized the SuperPoint [43] and SuperGlue [44] feature extractor and matcher, respectively. SuperPoint is an advanced deep learning-based method for keypoint detection and description. It efficiently identifies salient points in images, proving robust features that are crucial for accurate matching across different views. SuperGlue, complementing SuperPoint, is a powerful feature matching algorithm. It employs graph neural networks to find correspondences between sets of keypoints, even in challenging conditions where traditional methods like SIFT falter. The combination of SuperPoint and SuperGlue integrated with SfM offers superior performance, as evidenced in Figure 3.5.

Having reconstructed the 3D space with V_l, V_r, V_h , and W, we proceed to

Figure 3.4 Comparison of SIFT feature detection with SuperPoint on the left and SIFT feature matching with SuperGlue on the right. Results show that SuperPoint and SuperGlue show superior performance than SIFT in both feature detection and matching.

determine the camera extrinsics, which include the rotation matrix R and translation matrix t, to localize the cameras within the world coordinate space. This localization is achieved through PnP [45], a geometric algorithm used for estimating the pose of a camera in 3D space given a set of 2D-3D point correspondences. However, applying PnP directly to frames captured by the hand presents challenges. While the RANSAC algorithm [46], a robust method for outlier rejection in data fitting, excludes many outliers from feature matching, it does not entirely eliminate all outliers, such as features originating from hand features. To address this, we use off-the-shelf segmentation model [47] to obtain hand segmentation masks. This model generates hand segmentation masks M, which are then input into PnP, effectively preventing the detection of features within the masked regions, thereby refining the pose estimation process.

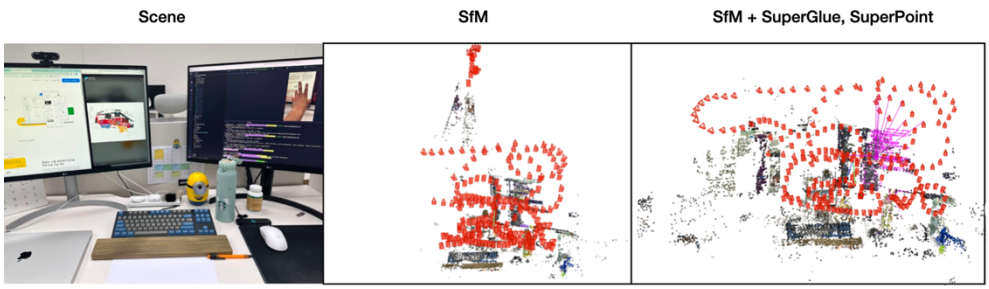

Figure 3.5. Comparison of the quality of scene reconstruction using SfM against SfM with SuperGlue and SuperPoint. Integrating SuperGlue and SuperPoint along with the traditional SfM method led to accurate 3D reconstruction and camera localization.

3.3.2 Pseudo Ground Truth with Processed Data

Upon preprocessing the captured data and localizing the cameras within the global space, we begin extracting the initial hand pose, denoted as $J_{intial} \in \mathbb{R}^{3\times1}$, from the head-view video V_h. This approach is preferred, as egocentric

3D hand pose estimators have demonstrated superior performance with head-view videos compared to wrist-captured videos due to the lack of wrist-captured datasets for training. We utilize an off-the-shelf 3D egocentric hand pose estimator based on a weak perspective model, as elucidated in the work by Rong et al. [40]. This model operates under the assumption that objects, such as hands in our context, are positioned at a sufficient distance from the camera, permitting the perspective effect to be approximated as a mere scaling factor.

However, this assumption introduces a limitation: the scale and position of hands in the reconstructed space may be inaccurately represented due to this simplification. To overcome this and obtain accurate 3D hand in the scene, we optimize scale $\alpha \in \mathbb{R}$ and translation $\beta \in \mathbb{R}^{3\times1}$ parameters. This optimization aims to minimize the 3D reprojection error. It involves the camera's intrinsic $K \in \mathbb{R}^{3\times3}$, rotation $R \in \mathbb{R}^{3\times3}$, translation $t \in \mathbb{R}^{3\times1}$ matrices, and projects J_{intial} into each camera's own coordinate system. The alignment's accuracy is further refined by comparing the L1 distance between the detected 2D joints $J_{2D} \in \mathbb{R}^{21\times2}$ and the projected 3D hand joints in each camera view. The optimization process is encapsulated in Equation 5.

$$\min_{\alpha,\beta} \sum_{i=1}^{N} |J_{2d}^{(i)} - \pi\left(K\left(R\left(\alpha J_{\text{initial}}^{(i)} + \beta\right) + t\right)\right)|_1 \tag{5}$$

After obtaining the raw pseudo-ground-truth 3D hand data, we implement a rigorous filtration process to ensure the high-quality inputs for training. This process is guided by a set of custom-designed criteria aimed at identifying and eliminating suboptimal or misleading data. Our primary criterion is based on the 3D reprojection of hand joints into the image space. Given the wrist camera's view consistency, particularly the restricted range of motion of the hand within the frame, we discard any data where the hand appears outside a predefined bounding box. This strategy is underpinned by the rationale that a hand reprojected outside this box is likely indicative of erroneous data.

Furthermore, we employ temporal coherence as a secondary filter. This involves analyzing the distances between keypoints in consecutive frame reprojections. Considering our data capture rate of 30 frames per second, substantial positional shifts between successive frames are unlikely. Thus, data exhibiting such marked changes are classified as unreliable and excluded. These stringent filtering criteria play a pivotal role in preserving the integrity and reliability of our training dataset. Upon refining the 3D hand joint data to J_{final}, we proceed to derive the hand mesh H_{mesh} by fitting the 3D joints to the MANO model parameters. The MANO hand model provides a parametric representation of hand poses and shapes. For each hand, the corresponding MANO model is loaded. Before optimization, we align the palm joints from the captured data with those of the MANO model using a rigid body registration process [48]. This step is essential for accurate initial positioning of the hand model. This optimization objective is to minimize the discrepancy between the joint positions of the model and those in the captured 3D data. We employ the Dogleg optimization method [49], with the objective function encompassing the difference between corresponding joints in the model and the captured data, alongside a regularization term for the model's shape parameters. Upon completion of this optimization, we extract pose, shape, and transformation parameters, which include joint rotations $\theta \in \mathbb{R}^{(15+3)\times 3}$, shape coefficients $\beta \in \mathbb{R}^{10\times 1}$, and the transformation matrix comprising translation and rotation components. The optimization objective is formulated as follows in Equation 6.

$$\min_{\theta,\beta,T}\left(\sum_{i=1}^{N}|J_{\text{MANO}}^{(i)}(\theta,\beta,T)-J_{\text{final}}^{(i)}|_2^2+\lambda|\beta-\beta_{\text{mean}}|_2^2\right) \tag{6}$$

Our practical framework establishes a robust pipeline for the generation and refinement of pseudo-ground-truth data. By incorporating strict filtering criteria and optimization techniques, we ensure the precision and reliability of our 3D hand pose data. The integration of the MANO model, combined with our optimization strategy, not only enhances the fidelity of the hand mesh

reconstruction but also expands the dataset with detailed pose and shape parameters by obtaining the 3D mesh of the hand for training SOTA models.

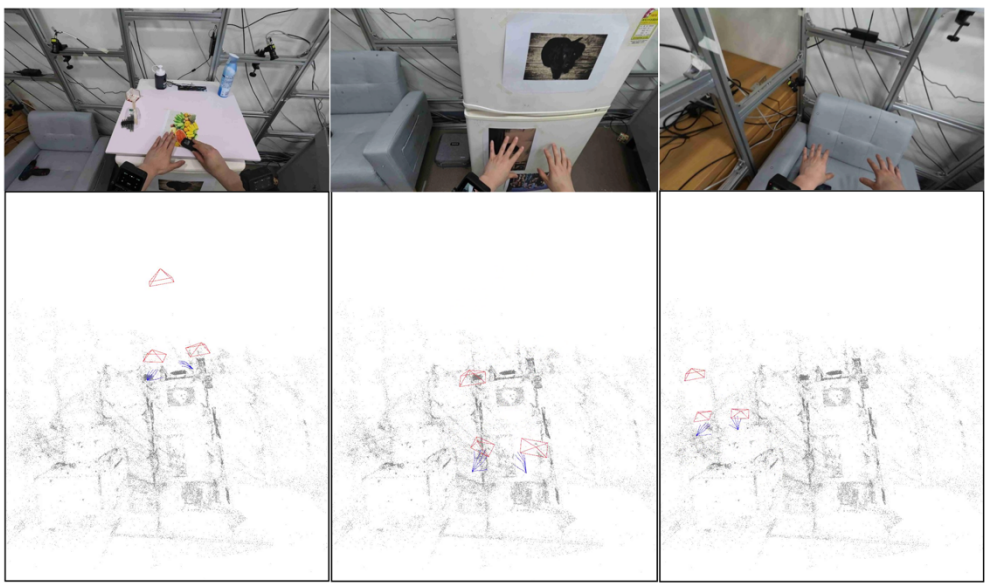

Figure 3.6. Visualization of head, wrist cameras and the generated pseudo-ground-truth 3D hand data in reconstructed space.

Figure 3.7. 3D reprojection of generated pseudo-ground-truth joints in image space.

Chapter 4

Experiments

4.1 Implementation Details

Our evaluation of the generated pseudo-ground-truth data involves training on current state-of-the-art models. We adopt the ACR [50] network architecture, implementing it within the PyTorch framework [51]. The model takes a RGB image as input and employs a feature map encoder. This encoder is responsible for extracting various maps, including hand-center maps, part-segmentation maps, cross-hand prior maps, and parameter maps. These are then aggregated to form the final feature set for hand model regression.

4.1.1 Training Stage Details

For benchmarking across multiple datasets, we trained our model using the Adam optimizer [52] with an initial learning rate of 5×10^{-5} across eight epochs. Our setup utilized dual NVIDIA 3090 GPUs with a batch size of 16. To accelerate the training phase, we initialized the network with a pre-trained HRNet-32W [53] backbone. The backbone feature map is set to 128×128, and the four pixel-aligned output maps are sized at 64×64. Augmentation technique es such as random scaling, rotation, flipping, and color jittering were applied during training to enhance robustness as in the original ACR paper.

We enhance our training by using different hand datasets featuring both 2D and 3D annotations. Specifically, our compilation includes part of the Assembly Hands [14], InterHand2.6M [13] datasets, and the full DexYCB [18], FreiHAND [16], HO3D [17] datasets, and additional 180,000 3D hand data from our pseudo-ground-truth system collection. The size of the training set is approximately $3 \times$ of the original ACR model. Proportionally, approximately 35% of the images are sourced from uncontrolled, in-the-wild environments, while the remaining data is gathered in controlled settings.

4.1.2 Testing Stage Computation

In our testing phase, HRNet-32W serves as the default backbone model. For comparative analysis against state-of-the-art models, the full official test set was employed. A confidence threshold was configured at 0.25 to detect a maximum of one left hand and one right hand per image. This setting aligns with our data's structural composition, which consistently presents only a single left and right hand across training images captured from the wrist cameras.

To ensure our model's resilience against varied real-world scenarios, our test suite comprises images featuring diverse lighting conditions, background clutter, and hand-object interactions. This diversity is critical for testing the model's

ability to generalize from our pseudo-ground-truth data to unpredictable in-the-wild settings.

Our framework includes stress tests against occlusions and rapid hand movements to validate the model's temporal consistency and resilience to transient visual obstructions. These scenarios are crucial for applications in dynamic environments where occlusion and motion blur are common. This comprehensive testing framework is designed to validate the hypothesis that our pseudo-ground-truth data, combined with our novel camera system, significantly enhances the hand pose estimation accuracy in comparison to the current state-of-the-art methodologies.

4.2 Datasets

A robust dataset serves as the foundation of any machine learning task, more so in the realm of computer vision where the nuances of data diversity and volume directly influence model performance. For enhancing 3D hand pose estimation from single RGB images, we have selected and utilized a variety of datasets, each serving a strategic purpose in both training and testing phases. These datasets range from capturing intricate hand-object interactions to varied egocentric hand movements, providing comprehensive coverage of the challenges inherent in hand pose estimation. In this section, we describe each dataset's unique characteristics, the scope of their annotations, and their specific applications in our research, detailing how they contribute in synergy with our collected pseudo-ground-truth data on the model performance.

InterHand2.6M [13] is a unique resource for two-hand interaction, offering accurate mesh annotations across 1.36M training frames and 850K testing frames, subdivided into interacting hands (IH) and single hands (SH). We employ the IH subset, down-sampled to 5fps, for our testing purposes.

FreiHAND [16] provides a 3D single-hand pose estimation dataset featuring 32 subjects, coupled with MANO and 3D keypoint annotations. It encompasses 132,560 training samples and 3,960 for evaluation, the entirety of which is utilized in our testing set.

HO-3D [17] dataset, a compilation of hand-object interactions captured in both multi-camera and single-camera configurations, includes 66,034 training and 11,524 test images across 68 sequences. We leverage the evaluation set from version 3 for our tests.

Assembly Hands [14] offers a large-scale, egocentric dataset with precise 3D hand pose annotations pertinent to complex hand-object interactions, sourced from the Assembly101 dataset [55]. It features 3.0M annotated images, including 490K from an egocentric perspective. Our testing employs the entire evaluation set.

DexYCB [18] caters to 3D hand pose and shape estimation during hand-object interactions, with a collection of over 582,000 frames from 10 subjects, annotated with 3D joint positions and shapes via MANO. The full evaluation set is incorporated into our testing framework.

Custom Bounding Box Detection Dataset due to the absence of suitable data for hand bounding box detection from a multi-view camera setup. This dataset comprises 3 sequences with manually annotated bounding boxes for hands, capturing the coordinates at multiple camera angles. It totals 4,500 frames and is exclusively used for testing our bounding box transfer methodology.

4.3 Evaluation Metrics

Mean Per-Joint Position Error (MPJPE) and **Mean Per-Vertex Error (MPVPE)**: These metrics compute the average Euclidean distance between predicted and ground truth (GT) 3D joint locations and mesh vertices in

millimeters, post-alignment to the root joint, providing a localized measure of geometric accuracy.

Procrustes Aligned MPJPE (PA-MPJPE) and **Procrustes Aligned MPVPE (PA-MPVPE)**: By performing Procrustes analysis prior to computing errors, these variations neutralize global transformation effects, offering a refined analysis of the model's positional accuracy independent of pose variations.

Mean Relative-Root Position Error (MRRPE): This metric evaluates the relative translation accuracy between paired hands by measuring the 3D distance between the predicted and GT root joint positions of the right hand relative to the left.

Intersection over Union (IoU): Employed for bounding box precision, IoU measures the overlap between predicted and GT bounding boxes, with higher values indicating greater accuracy.

Mean Average Precision (mAP): Utilized to assess the accuracy of the object detection model, mAP averages the precision at different levels of recall across multiple thresholds.

Dropout Rate: This reliability measure quantifies the frequency of missed hand detections in frames where a hand is present, with lower rates indicating more consistent detection.

4.4 Qualitative Evaluation

We present a qualitative analysis comparing our trained model's performance with current state-of-the-art methods across a variety of challenging scenarios, substantiated by visual comparisons in the subsequent figures.

ACR [50]: Figure 4.1 showcases a side-by-side comparison between our refined ACR model, trained with our novel pseudo-ground-truth data, and the original ACR. Our method manifests superior fidelity in the reconstruction of intricate

single and dual-hand poses, indicating significant strides in structural detail and pose accuracy.

Interwild [56]: In scenarios characterized by occlusions and motion-induced blur, our model demonstrates exceptional performance (refer to Figure 4.2). It consistently outstrips Interwild in the precision of hand pose estimations, affirming its robustness against partial visibility and dynamic conditions.

Frankmocap [40]: Comparative results depicted in Figure 4.3 illustrate our model's superiority over Frankmocap in rendering plausible hand orientations and movements, thereby reinforcing our method's adeptness at capturing complex hand dynamics.

100DOH [41]: Figure 4.4 underscores the precision of our model in bounding box detection, with green and red boxes delineating left and right hands, respectively. Notably, our model upholds a high detection accuracy even in instances of noticeable occlusion, underscoring the resilience of our prediction capabilities.

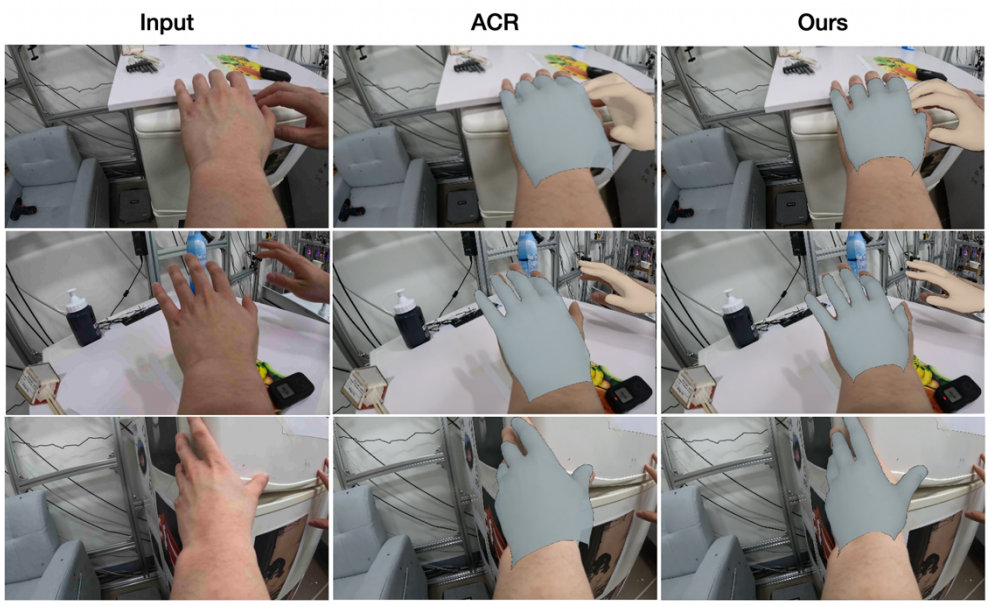

Figure 4.1 Qualitative comparison on our collected test set against the original ACR model. Our approach generates better results in single, two-hand reconstruction.

These qualitative assessments illuminate our model's advanced capabilities, especially in handling intricate and real-world scenarios, strengthening its suitability for deployment in a multitude of practical applications.

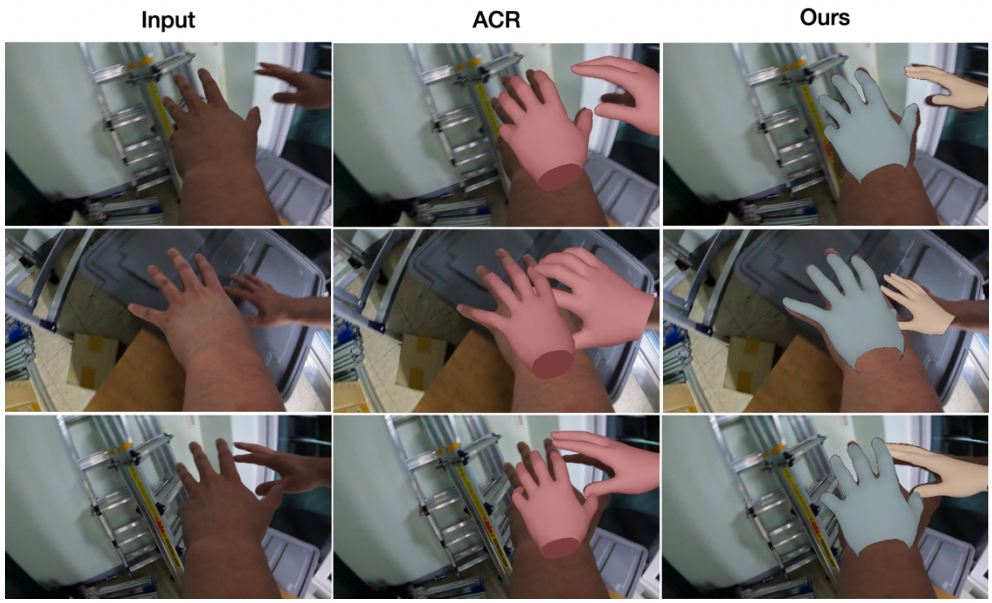

Figure 4.2 Qualitative comparison on our collected test set against Interwild. Our approach generates better results in challenging scenarios such as occlusion and blurry cases.

Figure 4.3. Qualitative comparison on our collected test set against Frankmocap. Our approach generally produces more plausible results and accurate hand orientations.

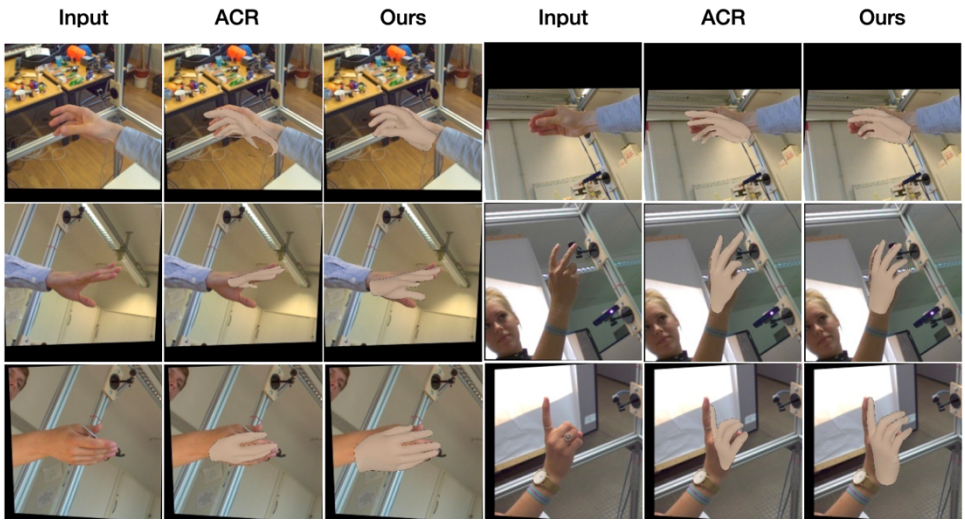

Figure 4.4. Qualitative comparison using the FreiHAND evaluation set against ACR.

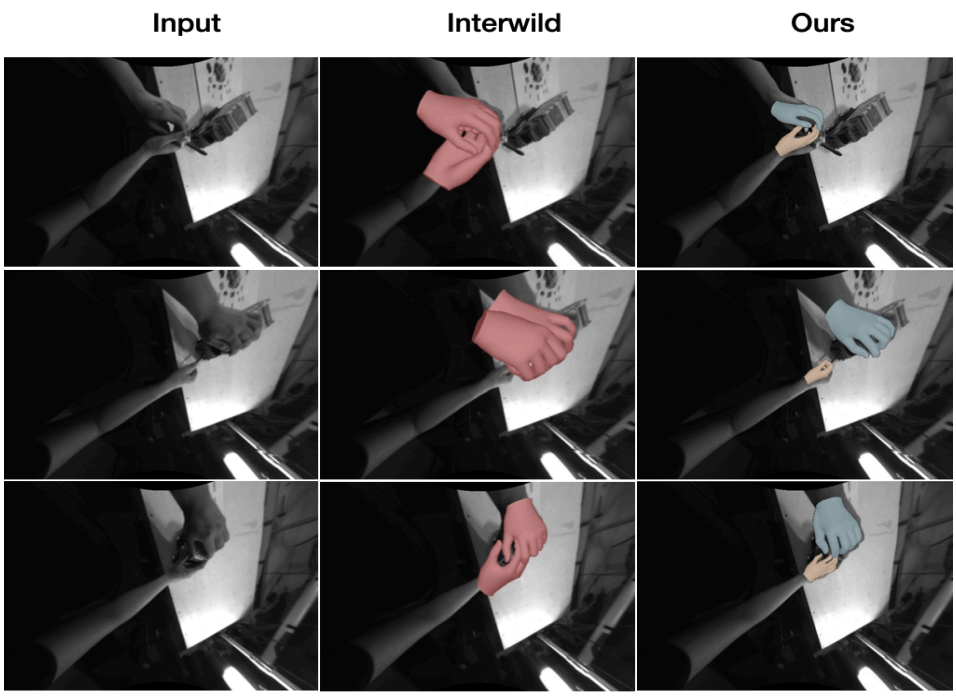

Figure 4.5. Qualitative comparison using the Assembly Hands evaluation set against Interwild.

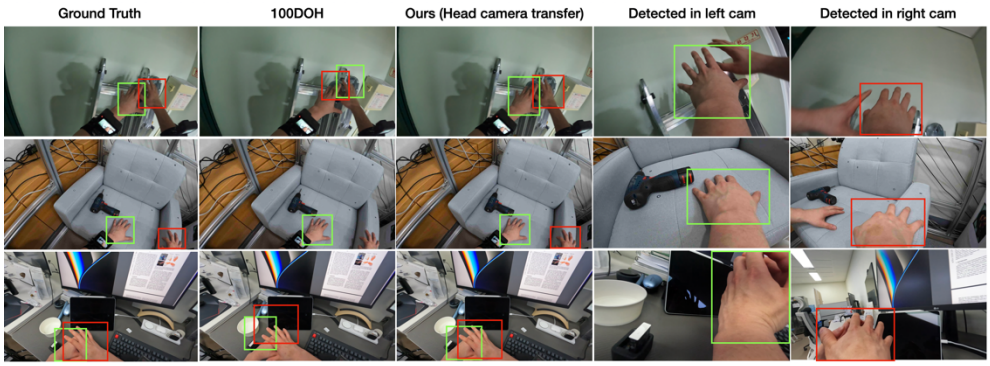

Figure 4.6. Qualitative comparison using manually annotated data. In the figure, green, red bounding boxes represent left, right hand detections respectively. Our approach excels in predicting more accurate hand locations, even in challenging scenarios with severe occlusion..

4.5 Quantitative Evaluation

This section delineates a rigorous quantitative evaluation of our proposed model against prevailing state-of-the-art methodologies, using a series of benchmark datasets. We have meticulously evaluated performance on the FreiHAND, Assembly Hands, Interhand 2.6M, Ho-3D, DexYCB, and a Custom Bounding Box dataset.

FreiHAND Dataset: Known for its complex and varied hand poses, the FreiHAND dataset provides a rigorous testbed for our model. Our model showcases marked improvements in scenarios with external occlusions and truncations, as evidenced by the results tabulated in Table 4.1, thus underscoring its robustness.

Assembly Hands Dataset: The Assembly Hands dataset, with its emphasis on hand-object interaction, mirrors the complexities of real-world application scenarios. Our model's superior performance is validated in Table 4.2, where it is seen to outshine baseline models, confirming its adeptness at complex interaction contexts.

Interhand 2.6M Dataset: The expansive and diverse collection of hand poses in the Interhand 2.6M dataset facilitates an extensive validation. Our model maintains a leading performance, as documented in Table 4.3, highlighting its scalability and generalization capabilities.

Ho-3D Dataset: This dataset's focus on hand-object interactions poses unique challenges in modeling dynamic interplays. Our model's capability to accurately estimate hand poses within this context is substantiated by the results in Table 4.4, revealing its comparative excellence in dynamic interaction settings.

DexYCB Dataset: Tailored for hand-object manipulation assessment, the DexYCB dataset evaluates the intricacy of hand grasping actions. Our model showcases its proficiency, delivering high accuracy in hand pose estimation across the dataset's diverse interactions, as detailed in Table 4.5. This accentuates the model's versatility and precision in real-life mimicked interactions.

Custom Bounding Box Dataset: Our custom dataset, annotated for bounding box detection in a multi-view environment, features scenarios with rapid hand movements and frequent occlusions. Our approach evidences commendable stability and tracking performance, with a notably low dropout rate highlighted in Table 4.6. This attests to the model's resilience and reliability, cementing its efficacy in dynamic and visually complex environments.

Table 4.1. Evaluation results on the FreiHAND test set. Our approach generates better results in reconstruction, particularly in challenging cases such as external occlusion, truncation.

	MPJPE↓	MPVPE↓	PA-MPJPE↓	PA-MPVPE↓	MRRPE↓
Interwild	17.2	24.5	14.4	20.8	11.2
ACR [50]	16.1	22.4	13.3	19.3	12.1
HTT [39]	18.1	22.5	15.2	20.1	**10.1**
DIR [57]	17.7	23.1	14.2	**18.4**	11.3
Frankmocap [40]	20.1	28.1	16.5	25.1	11.4
Ours	**15.8**	**21.3**	**12.8**	18.6	11.8

Table 4.2. Evaluation results on the Assembly Hands test set. Our method demonstrates superior performance in complex hand-object interaction scenarios, outperforming established baselines.

	MPJPE↓	MPVPE↓	PA-MPJPE↓	PA-MPVPE↓	MRRPE↓
Interwild	17.6	25.0	14.7	21.2	11.6
ACR [50]	16.4	22.7	**13.6**	19.6	12.4
HTT [39]	18.4	23.0	15.5	20.5	**10.4**
DIR [57]	18.0	23.4	14.5	21.0	11.7
Frankmocap [40]	20.4	28.5	16.7	25.5	11.6
Ours	**16.1**	**21.2**	13.7	**19.1**	10.7

Table 4.3. Evaluation results on the Interhand 2.6M test set. Our approach showcases a strong competitive edge in accurately capturing diverse hand poses within this extensive, large-scale dataset.

	MPJPE↓	MPVPE↓	PA-MPJPE↓	PA-MPVPE↓	MRRPE↓
Interwild	17.8	**20.9**	14.8	**17.1**	11.7
ACR [50]	16.5	24.9	13.7	19.7	12.5
HTT [39]	18.5	22.9	16.7	20.6	10.5
DIR [57]	18.2	23.6	14.6	21.3	11.8
Frankmocap [40]	21.2	24.5	16.9	25.7	11.8
Ours	**16.1**	21.1	**13.4**	18.2	**10.1**

Table 4.4 Evaluation results on the HO3D v3 dataset. The results highlight our model's effectiveness in hand-object interaction scenarios, significantly enhancing pose estimation accuracy.

	MPJPE↓	MPVPE↓	PA-MPJPE↓	PA-MPVPE↓	MRRPE↓
Interwild	19.2	**13.4**	15.7	17.9	16.1
ACR [50]	15.8	19.1	20.4	20.1	14.5
HTT [39]	18.1	21.4	22.3	22.3	18.3
DIR [57]	19.6	20.4	17.9	21.4	15.8
Frankmocap [40]	21.2	27.8	13.1	20.2	16.8
Ours	**13.9**	17.1	**10.8**	**14.9**	**11.0**

Table 4.5. Evaluation results on the DexYCB test set. The results show that our trained model excels in this dataset, particularly in the intricate scenarios of hand-object manipulations.

	MPJPE↓	MPVPE↓	PA-MPJPE↓	PA-MPVPE↓	MRRPE↓
Interwild	18.1	19.3	20.8	21.4	13.9
ACR [50]	14.2	19.1	17.1	19.9	**11.5**
HTT [39]	16.7	19.0	17.7	18.9	12.5
DIR [57]	18.6	19.1	18.5	23.9	12.2
Frankmocap [40]	19.3	21.5	19.6	28.7	17.8
Ours	**13.8**	**18.3**	**16.4**	**17.7**	12.4

Table 4.6. Comparison of hand bounding box detection in our manually annotated test dataset. Our approach significantly outperforms others in challenging scenarios, demonstrating a remarkably low dropout rate.

	IoU↑	mAP↑	Dropout Rate↓
100DOH [41]	71.8	78.0	5.0
BodyHands [58]	67.4	79.5	4.2
MediaPipe [59]	65.9	67.8	4.5
Ours	**87.1**	**91.2**	**1.8**

4.6 Ablation Study

To describe the contributions of the individual components within our pipeline, we have executed a detailed ablation study, with a particular emphasis on evaluating the impact of integrating wrist-mounted camera data with traditional head-mounted camera data on model accuracy.

Our approach capitalizes on the distinctive vantage points wrist-mounted cameras provide. These perspectives are crucial for capturing complex hand dynamics that may elude head-mounted devices. The results of our rigorous experimentation are in Table 4.7. The study evaluated the ACR model's performance using three distinct training data on the FreiHAND dataset: one with exclusively wrist camera data, one with only head camera data, and a third employing a composite of both wrist and head data, in conjunction with ACR's original datasets.

The results from our study yield interesting insights. The ACR model, when trained solely with additional head data, performed comparably to the baseline ACR model. However, the integration of wrist data contributed substantially to the model's accuracy, with the most marked improvement observed when both wrist and head data were utilized together. This fusion approach significantly outperformed the standalone datasets, affirming the collaborative benefit of this holistic data integration in the training regime.

Table 4.7. Effectiveness of wrist data for improving accuracy of model on the FreiHAND test set.

	MPJPE↓	MPVPE↓	PA-MPJPE↓	PA-MPVPE↓	MRRPE↓
ACR [50]	16.15	22.41	13.32	19.38	12.19
Head Data	16.13	22.36	13.30	19.31	12.12
Wrist Data	15.47	22.35	12.86	18.29	10.12
Wrist + Head Data	**15.44**	**21.31**	**12.83**	**18.21**	**10.01**

Chapter 5

Discussions

This work outlines an integrated approach to enhance the precision and consistency of 3D hand pose estimation within egocentric vision frameworks. Our methodology, capitalizing on a novel multi-camera system that includes both head and wrist-mounted cameras, addresses the inherent shortcomings of conventional hand pose estimation methods. The unique vantage points afforded by the wrist-mounted cameras have been instrumental in refining bounding box detection and hand pose estimation.

We presented a dual contribution to the field: firstly, the augmentation of existing 3D hand reconstruction models with pseudo-ground-truth data harvested from our adapted tri-camera system, which has markedly elevated the reconstruction quality over traditional models. Secondly, our innovative utilization of wrist cameras for the initial detection phase has improved the standard hand pose estimation pipeline, yielding enhanced consistency and precision in hand tracking, particularly in dynamic scenarios involving rapid movements and occlusions.

Empirical evaluations against benchmark state-of-the-art techniques have validated the means of our system, both quantitatively and qualitatively. The robustness of our approach, especially against the typical pitfalls of optical flow-based methods such as error propagation and motion blur, underscores the soundness of our design and algorithmic decisions.

While our findings are promising, we acknowledge the limitations inherent in our system, particularly in the context of scalability and the processing overhead introduced by multiple camera streams. Future work could focus on optimizing the computational efficiency of our framework and extending its applicability to more complex and diverse datasets. Additionally, exploring the integration of deep learning techniques to further refine the pseudo-ground-truth generation process could yield even more precise hand pose estimations.

In summary, our research contributes significantly to the advancement of egocentric vision and 3D hand pose estimation, offering a robust solution that holds potential for expansive applications in HCI, AR, and VR. We are optimistic that the insights gathered from our work will spur further innovation in this area, leading to more continuous and natural user interaction prototypes in immersive technological environments.

References

1. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: ICCV, 2017.
2. Iqbal, U., Molchanov, P., Gall, T.B.J., Kautz, J.: Hand pose estimation via latent 2.5 d heatmap regression. In: ECCV, 2018.
3. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: CVPR, 2018.
4. Yang, L., Li, S., Lee, D., Yao, A.: Aligning latent spaces for 3d hand pose estimation. In: ICCV, 2019.
5. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3d hand pose estimation from monocular rgb images. In: ECCV, 2018.
6. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Ganerated hands for real-time 3d hand tracking from monocular rgb. In: CVPR, 2018.
7. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: CVPR, 2019.
8. Panteleris, P., Oikonomidis, I., Argyros, A.: Using a single rgb frame for real time 3d hand pose estimation in the wild. In: WACV, 2018.
9. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular rgb image. In: ICCV, 2019.
10. Chen, Y., Tu, Z., Ge, L., Zhang, D., Chen, R., Yuan, J.: So-handnet: Self-organizing
network for 3d hand pose estimation with semi-supervised learning. In: ICCV, 2019.
11. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3d hand shape and pose estimation from a single rgb image. In: CVPR, 2019.
12. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. NeurIPS, 2016.
13. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: ECCV, 2020.
14. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands:

Towards egocentric activity understanding via 3d hand pose estimation. In: CVPR, 2023.
15. Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., Theobalt, C.: Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In: ICCV, 2017.
16. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: CVPR, 2019.
17. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR, 2020.
18. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: CVPR, 2021.
19. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV, 2015.
20. Kwon, T., Tekin, B., St¨uhmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: ICCV, 2021
21. Fleishman, S., Kliger, M., Lerner, A., Kutliroff, G.: Icpik: Inverse kinematics based articulated-icp. In: CVPRW, 2015.
22. Ge, L., Liang, H., Yuan, J., Thalmann, D.: Real-time 3d hand pose estimation with 3d convolutional neural networks. TPAMI, 2018.
23. Glauser, O., Wu, S., Panozzo, D., Hilliges, O., Sorkine-Hornung, O.: Interactive hand pose estimation using a stretch-sensing soft glove. TOG, 2019.
24. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610, 2022.
25. Schr¨oder, M., Maycock, J., Ritter, H., Botsch, M.: Real-time hand tracking using synergistic inverse kinematics. In: ICRA, 2014.
26. Sridhar, S., Mueller, F., Zollh¨ofer, M., Casas, D., Oulasvirta, A., Theobalt, C.: Real-time joint tracking of a hand manipulating an object from rgb-d input. In: ECCV, 2016.
27. Tagliasacchi, A., Schr¨oder, M., Tkach, A., Bouaziz, S., Botsch, M., Pauly, M.: Robust articulated-icp for real-time hand tracking. In: Computer graphics forum,

2015.
28. Oberweger, M., Wohlhart, P., Lepetit, V.: Training a feedback loop for hand pose estimation. In: ICCV, 2015.
29. Oberweger, M., Wohlhart, P., Lepetit, V.: Hands deep in deep learning for hand pose estimation. arXiv preprint arXiv:1502.06807, 2015.
30. Li, S., Lee, D.: Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In: CVPR, 2019.
31. Oberweger, M., Wohlhart, P., Lepetit, V.: Generalized feedback loop for joint hand-object pose estimation. TPAMI, 2019.
32. Oikonomidis, I., Kyriazis, N., Argyros, A.A., et al.: Efficient model-based 3d tracking of hand articulations using kinect. In: BMVC, 2011.
33. Fitzgibbon, A.W.: Simultaneous linear estimation of multiple view geometry and lens distortion. In: CVPR, 2001.
34. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR, 2016.
35. Li, S., Xu, C., Xie, M.: A robust o (n) solution to the perspective-n-point problem. TPAMI, 2012.
36. Lowe, G.: Sift-the scale invariant feature transform. Int. J, 2004.
37. Jiang, Z., Rahmani, H., Black, S., Williams, B.M.: A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In: CVPR, 2023.
38. Wang, C., Zhu, F., Wen, S.: Memahand: Exploiting mesh-mano interaction for single image two-hand reconstruction. In: CVPR, 2023.
39. Wen, Y., Pan, H., Yang, L., Pan, J., Komura, T., Wang, W.: Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In: CVPR, 2023.
40. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. ICCVW, 2020.
41. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: CVPR, 2020.
42. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV, 2020.
43. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: CVPRW, 2018.

44. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR, 2020.
45. Lepetit, V., Moreno-Noguer, F., Fua, P.: Ep n p: An accurate o (n) solution to the p n p problem. IJCV, 2009.
46. Chum, O., Matas, J., Kittler, J.: Locally optimized ransac. In: Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25, 2003.
47. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643, 2023.
48. Ashburner, J., Friston, K.J.: Rigid body registration. Statistical parametric mapping: The analysis of functional brain images, 2007.
49. Bellavia, S., Macconi, M., Pieraccini, S.: Constrained dogleg methods for nonlinear systems with simple bounds. Computational Optimization and Applications, 2012.
50. Yu, Z., Huang, S., Fang, C., Breckon, T.P., Wang, J.: Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In: CVPR, 2023
51. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIIPS, 2019.
52. Zhang, Z.: Improved adam optimizer for deep neural networks. In: IWQos, 2018.
53. Fan, Z., Spurr, A., Kocabas, M., Tang, S., Black, M.J., Hilliges, O.: Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In: 3DV, 2021.
54. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: ICAR, 2015.
55. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: CVPR, 2022
56. Moon, G.: Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In: CVPR, 2023
57. Ren, P., Wen, C., Zheng, X., Xue, Z., Sun, H., Qi, Q., Wang, J., Liao, J.: Decoupled

iterative refinement framework for interacting hands reconstruction from a single rgb image. In: ICCV, 2023
58. Narasimhaswamy, S., Nguyen, T., Huang, M., Hoai, M.: Whose hands are these? hand detection and hand-body association in the wild. In: CVPR, 2022
59. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172, 2019.
60. Kanazawa, Angjoo, et al. "End-to-end recovery of human shape and pose. In: CVPR, 2021
61. Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D hand shape and pose estimation from a single RGB image. In CVPR, 2019.
62. Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In ECCV, 2020.
63. Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In CVPR, 2019.
64. Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In CVPR, 2022.
65. Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3D interacting hand pose estimation by hand de-occlusion and removal. In ECCV, 2022.
66. Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In CVPR, 2022.
67. Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single RGB image. In ICCV, 2023.
68. Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In ICCV, 2023.

국문 초록

단일 이미지 기반 손 자세 추정을 위한 손목 장착 카메라 시스템

최연우
컴퓨터공학부
서울대학교 대학원

이 논문은 증강 현실, 가상 현실 및 인간-컴퓨터 상호 작용과 같은 응용 분야에서 중요한 자기 중심적 (egocentric) 비전 분야에서 3D 손 자세의 정확한 추정을 소개합니다. 이 연구는 손목 및 머리에 장착된 카메라를 사용한 새로운 유사 진실 데이터 (pseudo-ground-truth) 캡처 시스템을 사용하여 단일 RGB 이미지에서 3D 손 자세 추정의 정밀도와 안정성을 향상시키는 방법론을 제시합니다. 우리의 접근 방식은 이 다중 카메라 설정이 제공하는 독특한 관점을 활용하여 손의 움직임에서의 가림과 자유도가 높은 도전을 해결합니다.

우리 논문의 주요 기여는 다양합니다. 첫째, 우리는 정확한 손 데이터를 수집하기 위해 손목 장착 카메라를 사용하는 혁신적인 유사 진실 데이터 캡처 시스템을 제시합니다. 이 데이터는 기존의 최신 모델 (state of the art models)을 훈련시켜 성능을 크게 향상시킵니다. 둘째, 우리는 일반적인 단일 머리 카메라 관점이 아닌 손목 카메라에서 바운딩 박스 감지를 시작하여 손 움직임의 추적을 더 일관되고 정확하게 만듭니다.

우리의 실험 결과는 현재 최신 기술보다 우리 접근법의 우수성을 입증합니다. 우리는 보이는 관절 추정의 정확성과 가려진 관절 예측의 안정성에서 뚜렷한 개선을 달성합니다. 손목 장착 카메라를 사용하여 바운딩 박스 감지를 향상시키며, 이러한 시스템에서 손 데이터를 생성함으로써 손 자세 추정 방법의 강건성을 크게 강화합니다. 우리의 발견은 자기 중심 시나리오에서 더 정확하고 안정적인 3D 손 자세 추정에 기여하고, 응용 분야에서 더 자연스럽고 직관적인 경험을 사용자에게 제공하는 기술 개발에 기여합니다.

주요단어: 3D 손 자세 추정, 컴퓨터 비전, 자기중심적 비전
학번 : 2022-28614

감사의 글

우선 2 년의 석사 과정 동안 큰 가르침을 주신 주한별 지도 교수님께 진심으로 감사드립니다. 부족한 저를 마지막 까지 아낌없이 격려해 주시고 이끌어 주셔서 학위논문을 잘 마무리할 수 있었습니다. 이곳에서 배운 교수님의 가르침에 부끄럽지 않도록 저 또한 끊임없이 배움을 놓지 않고, 사회에 이바지할 수 있는 인재가 되도록 노력하겠습니다.

연구실 생활 동안 의지가 되어준 비주얼 컴퓨팅 연구실 동기들에게도 감사를 표합니다. 좋은 동기들을 만날 수 있었던 것 또한 저에게 행운이라 생각합니다. 연구실 생활에 우여곡절이 많았는데, 어려움을 함께 극복해 나가며 우정을 깊게 쌓을 수 있었던 것에 감사드립니다.

제 가족들에게 감사를 전합니다. 항상 배움의 자세를 놓지 않고 본이 되어 주시는 아버지, 어머니께 감사드립니다. 부모님의 가르침을 삶의 나침반으로 삼아 더욱 성장하는 모습을 보여드리겠습니다.

끝으로, 석사 과정을 하는 동안 힘이 되어준 나의 친구들에게 감사를 전하고 싶습니다. 여러분에게 받은 지지와 격려를 평생 잊지 않고, 저 또한 선한 영향력을 끼치는 사람이 되도록 노력하겠습니다.