ThermoHands: A Benchmark for 3D Hand Pose Estimation from Egocentric Thermal Image

Lawrence (Yunzhou) Zhu

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2024

Abstract

In this work, we present ThermoHands, a new benchmark for thermal image-based egocentric 3D hand pose estimation, aimed at overcoming challenges like varying lighting and obstructions (e.g., handwear). The benchmark includes a diverse dataset from 28 subjects performing hand-object and hand-virtual interactions, accurately annotated with 3D hand poses through an automated process. We introduce a bespoken baseline method, TheFormer, utilizing dual transformer modules for effective egocentric 3D hand pose estimation in thermal imagery. Our experimental results highlight TheFormer's leading performance and affirm thermal imaging's effectiveness in enabling robust 3D hand pose estimation in adverse conditions.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 980730
Date when approval was obtained: 2024-01-30
The participants' information sheet and a consent form are included in the appendix.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Lawrence (Yunzhou) Zhu*)

Acknowledgements

I appreciate Dr. Chris Xiaoxuan Lu, leader of the Mobile Autonomy, Perception, and Sensing (MAPS) Lab, for offering me this invaluable opportunity to work on this project. Without his continuous support and resources that went well beyond the undergraduate level, this project could never have achieved its current quality.

I am grateful to Fangqiang Ding and Xiangyu Wen, my colleagues from the MAPS Lab, who provided me with remarkable mentorship and support during the publication preparation of this work.

I appreciate my hardworking parents for their continuous support over the past 22 years and for funding me through this expensive programme.

I am thankful to all the participants in this project for their time spent performing hand actions and providing feedback, especially to Shikai Geng and Xinrui Xie, who volunteered in the early stage of data collection. They provided me with guidance to design and improve instructions for participants.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Egocentric 3D hand pose estimation is critically important for interpreting hand gestures across various applications, ranging from extended reality (XR) [101, 76, 63, 54], to human-robot interaction [78, 26, 27], and to imitation learning [73, 16, 1]. Its importance has been magnified with the advent of advanced XR headsets such as the Meta Quest series [64] and Apple Vision Pro [2], where it serves as a cornerstone for spatial interaction and immersive digital experiences.

The core objective of robust egocentric hand pose estimation is to develop methods that accurately determine the 3D positions of 21 joints per hand (*cf*. Fig. 1.1), adaptable to the complex environments typical in real-world applications. Estimating the 3D hand pose from 2D images requires the algorithm not only to estimate the 2D pixel coordinates of hand joints on the image frame but also to perform monocular depth estimation to estimate 2.5D (x, y, and depth with respect to the camera center). Subsequently, camera intrinsic parameters are used to project the estimated 2.5D hand pose into 3D world space.

While current research of hand pose estimation primarily focuses on RGB image-based methods [55, 53, 100, 25, 18], these approaches are particularly vulnerable to issues related to lighting variation and occlusions caused by handwear, e.g., gloves or large jewellery [93, 3]. These challenges underscore the imperative for robust egocentric 3D hand pose estimation capable of performing reliably in a variety of common yet complex daily scenarios. The prevailing approach to facilitate robust hand pose estimation in low-light conditions utilizes *near infrared* (NIR) cameras paired with active NIR emitters. This technology, invisible to the human eye, leverages active NIR emitter-receiver configurations for depth estimation through time-of-flight (ToF) or structured lighting. Nevertheless, active NIR systems are more power-intensive compared to passive sensing technologies [38, 37] and are prone to interference from external NIR sources, such as sunlight [91] and other NIR-equipped devices [82]. Consequently, these vulnerabilities restrict the effectiveness of hand pose estimation under bright daylight conditions and in situations where multiple augmented reality (AR) or virtual reality (VR) systems are used for collaborative works.

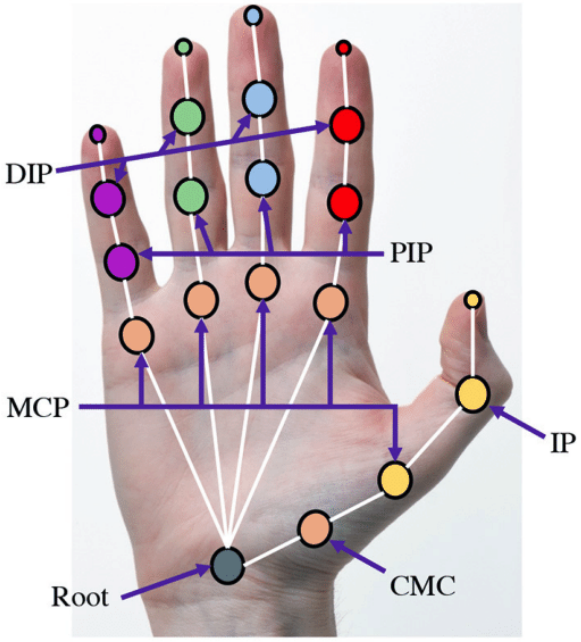

Figure 1.1: Each hand contains 21 joints (including finger tips), in this work we estimation the 3D position of these 21 joints $\times$ 2 hands. [70]

In contrast to NIR-based methods, thermal imaging cameras offer a passive sensing solution for hand pose estimation by capturing long-wave infrared (LWIR) radiation emitted from objects, thereby eliminating reliance on the visible light spectrum [60]. This unique attribute of thermal imaging introduces several benefits for 3D hand pose estimation. Primarily, it accentuates the hand's structure via temperature differentials, negating the effects of lighting variability. Moreover, thermal cameras are capable of detecting hands even under handwear such as gloves by identifying heat transmission patterns. This ability ensures a stable and consistent representation of hands, independent of any coverings, thereby broadening the scope and reliability of hand pose estimation across various scenarios.

Building on the above insights, this study probes the following research question: *Can egocentric thermal imagery be effectively used for 3D hand pose estimation under various conditions (such as different lighting and handwear), and how does it compare to techniques using RGB, NIR, and depth*[1] *spectral imagery?* To answer this, we introduce `ThermoHands`, the first benchmark specifically tailored for egocentric 3D hand pose estimation utilizing thermal imaging. This benchmark is supported by a novel multi-spectral and multi-view dataset designed for egocentric 3D hand pose estimation and is unique in comprising thermal, NIR, depth, and RGB images (*cf*. Tab. 2.1). Our dataset emulates real-world application contexts by incorporating both hand-object and hand-virtual interaction activities, with participation from 28 subjects to ensure a broad representation of actions (*cf*. Fig. 3.1). The performance scene is setup as shown in Fig. 1.2 To offer a thorough comparison across spectral types, we gather data under five distinct scenarios, each characterized by varying environments, handwear,

[1]For readability, we treat depth and NIR as two 'spectra', despite their usual overlap.

Figure 1.2: Data collection scene setup in IF-G.17, School of Informatics, The University of Edinburgh

and lighting conditions (*cf*. Tab. 3.1). Considering the challenges associated with manually annotating large-scale 3D hand poses, we developed an automated annotation pipeline. This pipeline leverages multi-view RGB and depth imagery to accurately and efficiently generate 3D hand pose ground truths through optimization based on the MANO model [75] (*cf*. Fig. 3.7).

Together with the multi-spectral dataset, we introduce a new baseline method named *TheFormer*, specifically designed for thermal image-based egocentric 3D hand pose estimation (*cf*. Fig. 3.9). This approach is notable for its two parallel transformer modules, *i.e*., mask-guided spatial transformer and temporal transformer, which encode spatio-temporal relationship for 3D hand joints without losing the computation efficiency. Our validation process begins with verifying the annotation quality, which averages an accuracy of 1cm (*cf*. Tab. 4.1). We then benchmark *TheFormer* against leading methods (*cf*. Tab. 4.2) and compare the performance of various spectral images (*cf*. Tab. 4.3 and Fig. 4.2). The findings underscore thermal imagery's advantages in difficult lighting conditions and when hands are gloved, showing superior performance and better adaptability to challenging settings than other spectral techniques.

1.2 Contribution

This thesis focuses on egocentric hand pose estimation using thermal cameras, positioning itself within the broader domain of human pose estimation through thermal imaging. The primary aim of this research is to assess whether the intrinsic characteristics of thermal imaging confer distinct advantages for identifying human hands, particularly in conditions where traditional imaging modalities—such as RGB and

depth sensing—may falter due to lighting variations, occlusions, or the presence of handwear. Through a systematic exploration and evaluation, this study aims to unlock the potential of thermal imaging, proposing novel approaches and methodologies that could significantly advance the field of hand pose estimation by leveraging the unique thermal signatures of human hands.

Our main contributions are summarized as follows:

- We introduce the first-of-its-kind benchmark, dubbed `ThermoHands`, to investigate the potential of thermal imaging for egocentric 3D hand pose estimation.
- We collected a diverse dataset comprising approximately 96,000 synchronized multi-spectral, multi-view images capturing hand-object and hand-virtual interactions from 28 participants across various environments. This dataset is enriched with 3D hand pose ground truths through an innovative automatic annotation process.
- We introduce a new baseline method, termed *TheFormer*, and implement two state-of-the-art image-based methods on our dataset for benchmarking.
- Based on the ThermoHands benchmark, we conduct comprehensive experiments and analysis on TheFormer and state-of-the-art methods.
- We will release our dataset, code and models and maintain the benchmark to serve as a new challenge in 3D hand pose estimation. on
- This work is submitted and currently in review for **European Conference on Computer Vision** (ECCV) 2024.

In light of thermal imaging's distinct advantages for egocentric hand pose estimation, its integration into MR devices promises not only a reduction in sensor count and imaging modalities but also a potential decrease in production costs by supplanting the specialized hand tracking apparatuses currently in use. This paradigm shift not only addresses the challenges posed by adverse lighting conditions and obstructive handwear but also aligns with the sustainability and efficiency imperatives of next-generation MR technologies.

Chapter 2

Background

2.1 Related Works

2.1.1 3D Hand Pose Datasets

Datasets with 3D hand pose annotations are imperative for training and evaluating *ad-hoc* models. Existing datasets, according to their approaches of annotation acquisition, can be summarized as four types in general, *i.e.*, marker-based [105, 28, 92, 23], synthetic [69, 34, 68, 67, 111], manual [69, 90] or hybrid [14, 112, 66, 58, 71], and automatic [32, 8, 51, 83] annotated datasets. Marker-based approaches, using magnetic sensor [105, 28] or Mocap markers [92, 23], can alter and induce bias to the hand appearance. Synthetic data [69, 34, 67, 111, 68] suffers from the *sim2real* gap in terms of hand motion and texture features. Introducing human annotators to fully [69, 90] or partly [14, 112, 66, 58, 71] annotate 2D/3D keypoints circumvents the issues above, but it either limits the scale of datasets or manifests costly and laborious in practice. Most similar to ours, some datasets adopt fully automatic pipelines to obtain 3D hand pose annotations [32, 8, 51], which leverage pre-trained models (*e.g.* OpenPose [12]) to infer the prior hand information and rely on optimization to fit the MANO hand model [75].

Despite the existing progress, previous datasets only provide depth [105, 68], RGB images [67, 66, 112, 71] or both of them [28, 92, 34, 111, 69, 90, 14, 58, 32, 8, 51] as the input spectra, unable to support the study of NIR or thermal image-based 3D hand pose estimation. ThermoHands fills the gap by providing a moderate amount of multi-spectral image data, from infrared to visual light, paired with depth images. Moreover, we capture bimanual actions from both egocentric and exocentric viewpoints and design hand-object as well as hand-virtual interaction actions to facilitate a wide range of applications. Tab. 2.1 shows a comprehensive comparison between the existing datasets (in chronological order) and ours.

2.1.2 Image-based 3D Hand Pose Estimation

As a key computer vision task, 3D hand pose estimation from images is highly demanded by applications like XR [101, 76, 63, 54], human-robot interaction [78, 26, 27] and

Table 2.1: Comparison of ThermoHands with existing image-based hand datasets with 3D pose annotations. (*) The number of synchronized frames of images. (**) No markers or sensors are attached to hands during data creation or capture.

Dataset	(*)frames	(**)markless	depth	thermal	real	ego	exo	two-hand	hand-obj int	hand-virt int
BigHands2.2M [105]	2.2M	✗	✓	✗	✓	✓	✓	✗	✗	✓
SynthHands [69]	220K	✓	✓	✗	✗	✓	✓	✗	✓	✗
EgoDexter [69]	3K	✓	✓	✗	✓	✓	✗	✗	✓	✗
FPHA [28]	105K	✗	✓	✗	✓	✓	✗	✗	✓	✗
ObMan [34]	150K	✓	✗	✗	✗	✗	✓	✗	✓	✗
FreiHAND [112]	37K	✓	✗	✗	✓	✗	✓	✗	✓	✓
HO-3D [32]	78K	✓	✓	✗	✓	✗	✓	✗	✓	✗
ContactPose [8]	2.9M	✓	✓	✗	✓	✗	✓	✓	✓	✗
InterHand2.6M [66]	2.6M	✓	✗	✗	✓	✗	✓	✓	✗	✓
H2O [51]	571K	✓	✓	✗	✓	✓	✓	✓	✓	✗
DexYCB [14]	508K	✓	✓	✗	✓	✗	✓	✗	✓	✗
HOI4D [58]	2.4M	✓	✓	✗	✓	✓	✗	✗	✓	✗
AssemblyHands [71]	3.0M	✓	✗	✗	✓	✓	✓	✓	✓	✗
ARCTIC [23]	2.1M	✗	✗	✗	✓	✓	✓	✓	✓	✗
ThermoHands (ours)	96K	✓	✓	✓	✓	✓	✓	✓	✓	✓

imitation learning [73, 16, 1]. Therefore, this field has been extensively explored in previous arts that uses single RGB [40, 89, 104, 17, 44, 55, 66, 53, 87, 99, 106, 34, 57, 6, 4, 107] or depth [29, 65, 102, 108, 109] image as input. These methods can be roughly categorized into two fashions, *i.e.*, model-based and model-free methods. Model-based methods [53, 87, 99, 106, 34, 57, 6, 4, 107] utilize the prior knowledge of the MANO hand model [75] by estimating its shape and pose parameters, while model-free methods [40, 89, 104, 17, 44, 55, 66, 29, 65, 102, 108, 109] learn the direct regression of 3D hand joints or vertices coordinates. Recently there has been growing interest in leveraging the temporal supervision [33, 57, 103, 49, 72] or leveraging sequential images as input [100, 25, 18, 11, 42, 19] for 3D hand pose estimation. In this study, we evaluate existing methods and our baseline method in both single image-based and video-based problem settings, respectively. Apart from the previous approaches, we investigate the potential of thermal imagery for tackling various challenges in 3D hand pose estimation.

2.1.3 Thermal Computer Vision

Thermal cameras achieve imaging by capturing the radiation emitted in the LWIR spectrum and deducing the temperature distribution on the surfaces [60]. Leveraging its robustness to variable illumination and unique temperature information, numerous efforts have been made to address various computer vision tasks, including super-resolution [74, 30, 41], human detection [10, 39, 31], action recognition [5, 21] and pose estimation [62, 15, 88], semantic segmentation [50, 46, 96, 52], depth estimation [85, 61, 84, 45], visual(-inertial) odometry/SLAM [43, 80, 86, 79], 3D reconstruction [56, 77, 81], *etc*. In this work, we focus on 3D hand pose estimation, which is an under-exploited task based on thermal images.

2.2 Machine Learning Basics

2.2.1 Image Feature Extraction

In the realm of computer vision and machine learning, feature extraction stands as a cornerstone technique, pivotal for transforming raw image data into a refined form for analysis and interpretation. This process involves distilling essential characteristics or attributes from images—such as edges, textures, or patterns—that are crucial for various tasks including image classification, object detection, and hand joint identification. It effectively reduces the dimensionality of data, enhancing both the efficiency and performance of subsequent algorithms. Building upon this foundational concept, Residual Networks (ResNet) [36] introduces a transformative approach to deep learning architectures. Developed by He et al. (2016), ResNet addresses the challenge of training very deep networks through the introduction of residual blocks, featuring skip connections that facilitate the flow of gradients and mitigate the vanishing gradient problem. This innovation not only enables the construction of deeper neural networks but also significantly improves their learning capability and performance across a wide spectrum of tasks in computer vision. In the development of our methods, ResNet-18, which denotes the number of layers in the network that have trainable weights, totaling 18 layers, is used to extract hand features from egocentric images. ResNet-18, specifically, is designed to be lighter and faster than its more extensive counterparts (such as ResNet-50, ResNet-101, and ResNet-152) while still retaining a significant capacity for learning. This performance-accuracy balance of ResNet-18 makes it a perfect choice for our task, that requires real-time performance in future deployment.

2.2.2 Visual Transformer and the Attention Mechanism

Transformers, initially introduced for natural language processing (NLP) tasks by Vaswani et al. [95] in their seminal 2017 paper, ”Attention is All You Need,” have revolutionized the way sequential data is processed. At the core of the Transformer architecture is the **self-attention mechanism**, which enables the model to weigh the importance of different parts of the input data relative to each other. Unlike previous models that processed data sequentially, Transformers handle data in parallel, significantly improving efficiency and performance on tasks such as language translation, text generation, and sentiment analysis.

Building upon the success in NLP, researchers extended the Transformer architecture to handle visual data [22], leading to the development of Visual Transformers. This adaptation involves treating images not as grids of pixels but as sequences of smaller image patches. Each patch is encoded into a token, similar to how words are treated in NLP, and these tokens are processed by the Transformer to capture complex relationships and dependencies between different parts of the image.

Visual Transformers mark a significant departure from the conventional convolutional neural network (CNN) approach to image analysis. While CNNs rely on the local connectivity of pixels and hierarchically extract features through convolutional layers, Visual Transformers leverage the global relationships between patches across the entire image. This global perspective allows Visual Transformers to understand the context

and semantic relationships within images more holistically.

In the context of egocentric hand pose estimation, the attention mechanism of Visual Transformers provides a focused approach, especially on hands, making them a suitable baseline model for this task. The ability to dynamically allocate attention across image patches enables these models to adaptively emphasize the regions of interest—primarily the hands—in complex visual scenes. This capability not only enhances the accuracy of pose estimation but also ensures robust performance across varied and challenging scenarios, highlighting the potential of Visual Transformers in advancing the domain of hand pose analysis. In the development of an egocentric hand pose estimator for this thesis, two specific types of attention mechanisms are employed: spatial attention and temporal attention.

2.2.3 Deformable Self-attention

Deformable self-attention is a novel adaptation of the self-attention mechanism, designed to enhance the efficiency and effectiveness of processing sparse and irregular data, such as images, within transformer models. This mechanism selectively focuses on a subset of key sampling points in the input data, allowing for flexible and adaptive attention modeling. The introduction of deformable self-attention represents a significant advancement in the application of transformers to computer vision tasks, where the spatial heterogeneity and complexity of images pose unique challenges.

The foundation of deformable self-attention lies in its ability to dynamically adjust the attention sampling locations based on the content of the input data. Unlike traditional self-attention, which considers all positions in the input sequence or feature map equally, deformable self-attention mechanisms focus on a small set of relevant positions. This approach not only reduces computational complexity but also improves model performance by emphasizing critical features and patterns in the data.

One prominent application of deformable self-attention is in the Detection Transformer (DETR) model [13]. DETR leverages deformable self-attention to efficiently process image data for object detection tasks. By focusing on a sparse set of key points within the image, DETR can identify and localize objects with high precision, without relying on the dense sampling and complex pre- and post-processing steps typical of conventional object detection frameworks. The deformable variant of DETR further enhances this process by allowing the model to adaptively focus its attention on the most informative parts of the image, significantly boosting detection performance, especially in cases involving small or partially occluded objects such as hand when grasping objects.

2.2.4 Spatial Attention

The implementation of spatial attention typically involves generating an attention map that signifies the importance of each location in the input data. This attention map is then used to modulate the input or the feature maps within the network, emphasizing areas deemed more relevant for the task at hand. The generation of the attention map can be conditioned on the input itself or on additional context provided to the model, making spatial attention mechanisms highly adaptable to a wide range of applications.

Spatial attention has been successfully applied in various domains of deep learning, particularly in computer vision tasks such as image classification, object detection, and semantic segmentation. By allowing the model to focus on salient features or regions, spatial attention mechanisms can lead to significant improvements in accuracy and efficiency. For instance, in object detection tasks, spatial attention enables the model to hone in on areas of the image where objects are likely to be located, reducing the computational burden of processing less informative regions and enhancing detection performance.

In the context of egocentric hand pose estimation, incorporating spatial attention mechanisms can markedly improve the model's capability to accurately identify and analyze hand positions and movements. Given the dynamic nature of hand movements and the variety of backgrounds and lighting conditions in which they may occur, spatial attention can guide the model to focus on the hands themselves, disregarding irrelevant background information. This focused approach not only increases the precision of pose estimation but also enhances the model's robustness across diverse environments and scenarios. Through the strategic deployment of spatial attention, models can achieve a more nuanced understanding of spatial dynamics, paving the way for advancements in how machines perceive and interact with their surroundings.

2.2.5 Temporal Attention

Temporal attention mechanisms specialize in processing sequential data, such as a sequence of images (also known as a video clip), allowing models to dynamically prioritize certain time steps over others based on their relevance to the task. By generating a temporal attention map, these mechanisms highlight the importance of specific moments in a sequence, enabling the model to focus on critical temporal features while minimizing the influence of less informative ones. This approach is particularly valuable in tasks where the context and changes over time are crucial for accurate predictions or analysis.

In various domains of deep learning, especially those involving time-series data or sequential inputs such as video analysis, natural language processing, and audio recognition, temporal attention has proven to be highly effective. For example, in video classification or activity recognition tasks, temporal attention enables the model to pay more attention to the keyframes or segments that are indicative of the activity being performed, thereby improving the overall accuracy of the classification.

Applying temporal attention in the context of egocentric hand pose estimation significantly enhances the model's ability to track and analyze hand movements over time. Given the continuous and dynamic nature of hand gestures, temporal attention ensures that the model focuses on the most informative frames, contributing to a deeper understanding of the gesture or action being performed. Moreover, temporal attention aids in pose estimation during instances when hands are occluded in certain frames, by making predictions based on adjacent critical frames. This feature is particularly important in scenarios where hand movements are subtle or involve complex sequences of actions.

2.3 Timeline of the Project

The project timeline is summarized in the following steps:

1. **Background Research and Baseline Establishment (Aug 2023 - Sep 2023):** Engage in comprehensive exploration within the domain of egocentric hand pose estimation on RGB images. Achieve replication of the Hand Tracking Transformer (HTT) as a foundational benchmark method.
2. **HMSP Design, Camera Calibration, and Initial Data Collection (Oct 2023 - Nov 2023):** Design and validate the Hand Motion Sensing Platform (HMSP) and perform camera calibration to determine extrinsic parameters, including translation and rotation between camera frames. In November 2023, set up data collection environments within the MAPS lab, gathering initial test datasets from lab members for the development and refinement of TheFormer.
3. **Annotation Pipeline Development and Comprehensive Data Collection (Oct 2023 - Feb 2024):** Between October 2023 and January 2024, develop and evaluate an automatic handpose annotation pipeline, pending ethical approval. Subsequently, in February 2024, expand the dataset by collecting data from 35 participants, carefully selecting high-quality data for further analysis.
4. **Model Benchmarking and Conference Submission Preparation (Feb 2024 - Mar 2024):** Benchmark the TheFormer against HTT and A2J transformers to evaluate performance and prepare the findings for submission to the European Conference on Computer Vision (ECCV) by March 2024.

Chapter 3

The ThermoHands Benchmark

3.1 Multi-Spectral Hand Pose Dataset

Overview. At the core of our benchmark lies a multi-spectral dataset for 3D hand pose estimation (*cf*. Tab. 3.1), capturing hand actions performed by 28 subjects of various ethnicities and genders. As shown in Fig. 3.1, we develop a customized head-mounted sensor platform (HMSP) and an exocentric platform to record multi-view data. During capture, our participants are asked to perform pre-defined hand-object and hand-virtual interaction actions within the playground above the table. The main part is captured in the well-illuminated office scenario. To facilitate the evaluation under different settings, four auxiliary parts are recorded i) under the darkness, ii) under the sun glare, iii) with gloves on hand, and iv) in the kitchen environment with different actions, respectively.

In creating a realistic office environment for testing, random heat sources, such as servers and chargers, are strategically placed in the background. This approach introduces challenging factors into the thermal images, increasing the realism and complexity of the data captured by the HMSP. Although the thermal camera's center is aligned with the RGB camera's center at the same height on the 3D-printed sensor board, extrinsic parameters obtained from camera calibration are employed to ensure enhanced alignment accuracy.

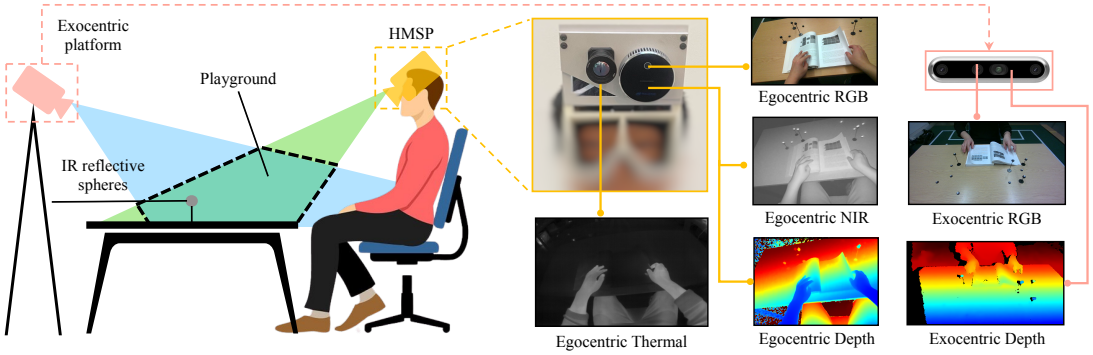

Figure 3.1: **Data capture setup** with the HMSP and exocentric platform recording multi-view multi-spectral images of two-hand actions performed by participants.

(a) **(b)** **(c)**

Figure 3.2: The continuous improvement of HMSP based on participants' feedback. Panel (a) demonstrates the initial prototype of HMSP; panel (b) shows cushions added; and panel (c) depicts a 30-degree downward tilt added to the sensor platform.

Sensor Platforms. The head-mounted sensor platform (HMSP) represents a significant advancement in simulating actual Mixed Reality (MR) devices, focusing on minimizing extra weight to facilitate unencumbered participant movement. Unlike traditional methods that involve mounting cameras on helmets [51], the HMSP is designed with comfort and functionality in mind. It comprises three main components: a cushion for comfort, a base providing a fixed 30-degree downward tilt, and a sensor board. This board is equipped with an Intel RealSense L515 LiDAR depth camera [38] for streaming egocentric RGB, depth, and NIR images, and a Teledyne FLIR Boson 640 long-wave infrared (LWIR) camera [94] for capturing thermal images.

The HMSP's modular design enhances its ability to simulate XR devices effectively. To further support multi-view annotation (*cf*. Sec. 3.2) and provide additional RGB-D image data from a third-person viewpoint, an exocentric platform equipped with an Intel RealSense D455 [37] is utilized. This setup strategically places the two depth sensors outside each other's field of view (FoV) to minimize interference caused by their NIR emitters, as demonstrated in Fig. 3.1.

The development of HMSP is characterized by a process of continuous improvement, informed by feedback from participants. As illustrated in Fig. 3.1, the initial prototype of HMSP consisted only a sensor board attached to a hard 3D-printed PLA base. Participants complained discomfort due to the majority of the device's weight resting on their cheeks. To address this issue, a cushion was incorporated to alleviate the pressure. Additionally, the original design required participants to lower their heads to manually align the sensor board to their hands, complicating the data collection process. In response, the final version of HMSP was modified to include a fixed 30-degree downward tilt, enhancing user comfortness and simulation to real use cases.

Synchronization. We use a single PC to simultaneously gather data streams from two sensor platforms, ensuring the synchronization of their timestamps. After collection, we synchronize six types of images, each with distinct frame rates, w.r.t. the timestamps of thermal images (8.5fps), thereby generating synchronized multi-spectral, multi-view data samples as our released data.

Camera Calibration. For accuracy, we use factory-calibrated intrinsic parameters

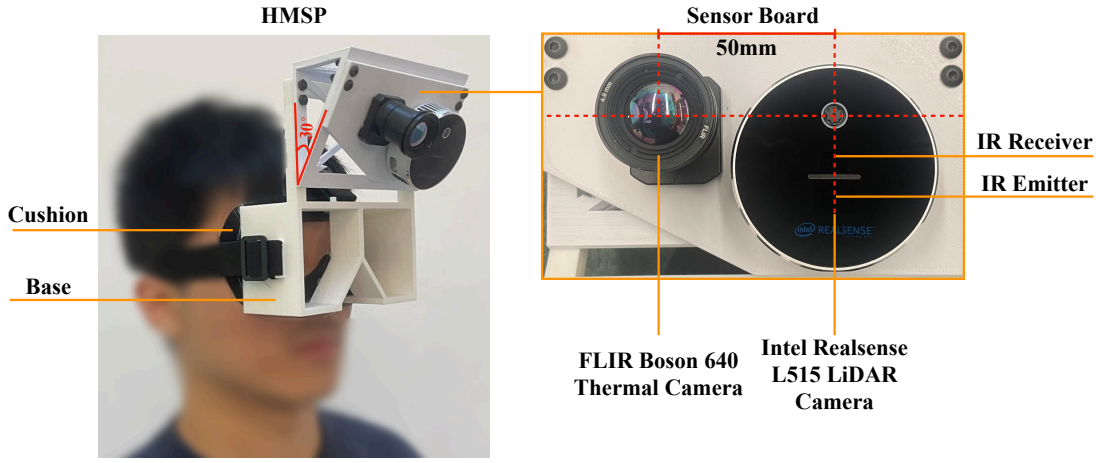

Figure 3.3: Design of the head-mounted sensor platform and sensor alignment.

Table 3.1: **Benchmark Dataset Statistics**. The overall duration of our dataset is over 3 hours with ∼96K synchronized frame of all types of images collected.

	Well-illuminated office (Main)				Other settings				Total
Setting	train	val	test	sum	darkness	sun glare	gloves	kitchen	
#frames	47,436	12,914	24,002	84,352	3,188	2,508	3,068	2,808	95,924
#seqs	172	43	86	301	12	12	12	14	352
#subjects	16	4	8	28	1	1	1	2	-

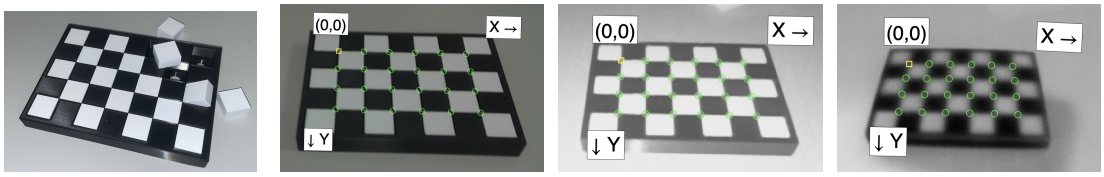

Figure 3.4

Figure 3.5: **Thermal calibration chessboard** containing a black base board and multiple removable white cubes (a). By cooling down the base board, it shows similar patterns and allows automatic corner detection in all (b) RGB, (c) NIR and (d) thermal images.

from the D455 and L515 cameras [38, 37]. To better calibrate the parameters of the thermal camera, we self-design a modular calibration chessboard as shown in Fig. 3.5. Before calibration, we cool down the black base board while keeping the white cubes at room temperature to create a visible chessboard pattern in all three spectra, which can be seen in Fig. 3.5. In this way, we can simultaneously calibrate the intrinsic parameter of the thermal camera and its extrinsic parameter w.r.t. the L515 camera. To enable the calibration between the two viewpoints, we place 11 IR reflective spheres at random locations and heights above the table. At the first frame of each sequence, we manually annotate their 2D locations from two viewpoints, retrieve their depth values and compute the transformation between two viewpoints by solving the PnP [24]. For the subsequent frames, the KISS-ICP [97] odometry method is used to track the motion of the ego-head using the point clouds converted from the egocentric depth images.

Cross-View Calibration. To calibrate the cameras between the egocentric and exocentric viewpoints, we place 11 IR reflective spheres at random heights within the playground, ensuring they are visible from both viewpoints. Initially, we plan to detect these spheres automatically from the NIR images and track them with the Kalman Filter [98]. However, we find this approach leads to many false positive detections, which severely affect the tracking accuracy. To ensure the calibration accuracy, we only annotate the sphere markers manually from two viewpoints in the first frame. The 3D positions of markers can be computed using the egocentric depth image and the cross-view transformation can be computed by solving the PnP [24]. For the subsequent frames, we can calculate the transformation by only tracking the pose of the egocentric camera as the exocentric platform keeps stationary during collection. We leverage the state-of-the-art odometry method KISS-ICP [97] and track the motion of the egocentric camera with the point clouds converted from depth images. Specifically, we set the `initial_threshold` and `min_motion_th` parameters of KISS-ICP as 0.0001 to make it capable of catching subtle motion. The input point cloud range is set as [0.2, 2.0] meters while the `max_points_per_voxel` is 30.

Spectrum Coverage and Interference Avoidance. The multi-spectral data encompasses imaging from the visible spectrum (400-700nm), near-infrared (NIR) spectrum (850 nm $\pm$ 10 nm), and the LWIR spectrum (8-14 μm). Images captured across different spectra contain unique information and serve varied purposes. For instance, RGB images offer semantic information, facilitating a deeper understanding of human-environment interaction. NIR lights can be actively emitted by our depth cameras to obtain the depth measurements via ToF or structured lighting. The LWIR frame, capturing temperature information, readily isolates uniform heat emitters like human hands. On the head-mounted sensor platform, the L515 LiDAR depth camera [38] emits NIR lasers at a wavelength of 860 nm, which falls outside the thermal camera's receptive range (8-14 μm), thereby eliminating any potential interference between cameras on the HMSP. Conversely, the exocentric RGB-D camera [37] necessitates structured lighting employing NIR at a wavelength identical to the IR emitter on the L515 LiDAR camera [38]. To prevent interference and image corruption, the exocentric RGB-D camera and the egocentric NIR LiDAR are strategically positioned outside each other's receptive fields during data collection. The specification of all sensor frames we collect are shown in Tab. 3.2.

Table 3.2: The specification of sensor frames captured in data collection.

Sensor frames	Sensor Type	Resolution			Fov			FPS
		Range	Horizontal	Vertical	Range	Horizontal	Vertical	
RGB (ego)	Intel RS L515 [38]	-	1280	720	-	69	42	30
NIR (ego)	Intel RS L515 [38]	-	640	480	-	70	55	30
Depth (ego)	Intel RS L515 [38]	$<$ 5mm @ 1m	640	480	0.25m to 9m	70	55	30
Thermal (ego)	FLIR Boson 640 [94]	$\leq$ 60 mK	640	512	-	95	-	8.5
RGB (exo)	Intel RS D455 [37]	-	1280	720	-	87	58	30
Depth (exo)	Intel RS D455 [37]	$<$ 2% at 4m	848	480	0.6m to 6m	87	58	30

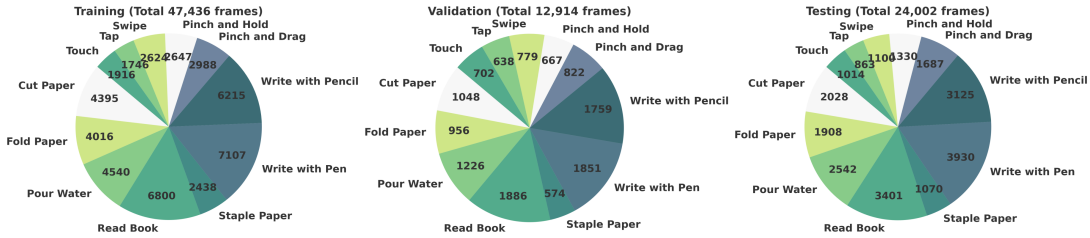

Figure 3.6: Distribution of data from the main part over hand actions.

Dataset Statistics. As shown in Tab. 3.1, our dataset consists of approximately 96K synchronized multi-spectral multi-view frames (*cf*. Fig. 3.1) and 352 independent sequences in total. The main part is collected under the well-illuminated office scenario, where each participant[1] performs 7 scenario-specific hand-object interaction actions: *cut paper, fold paper, pour water, read book, staple paper, write with pen, write with pencil*, and 5 hand-virtual interaction actions: *pinch and drag, pinch and hold, swipe, tap, touch*, with two hands. This main part is divided into the training, validation and testing splits by subjects with a ratio of 4:1:2. We also collect four auxiliary testing sets by asking one subject to perform the aforementioned 12 actions in the darkness, sun glare and gloves settings individually, and two subjects to perform 7 scenarios-specific interaction actions: *cut, spray, stir, wash hands, wash mug, wash plate, wipe* in the kitchen environment. Please refer to the supplementary for more dataset details.

Dataset Distribution. We show the data distribution of our main part over hand actions in Fig. 3.6. Among hand-object interaction actions, *write with pen*, *write with pencil* and *read book* have more frames than others due to the complexity of these actions, making them tend to last longer than others. As seven of our participants did not perform the hand-virtual actions, we have fewer frames from them than their hand-object counterparts. Please see our supplementary video for more visualization of our collected data.

3.2 Hand Pose Annotation

To refrain from employing tedious human efforts for annotation, we implement a fully automatic annotation pipeline, similar to the approaches in [32, 8, 51], to obtain the 3D hand pose ground truth for our dataset.

[1]Due to their limited time, 7 participants only perform the hand-object actions.

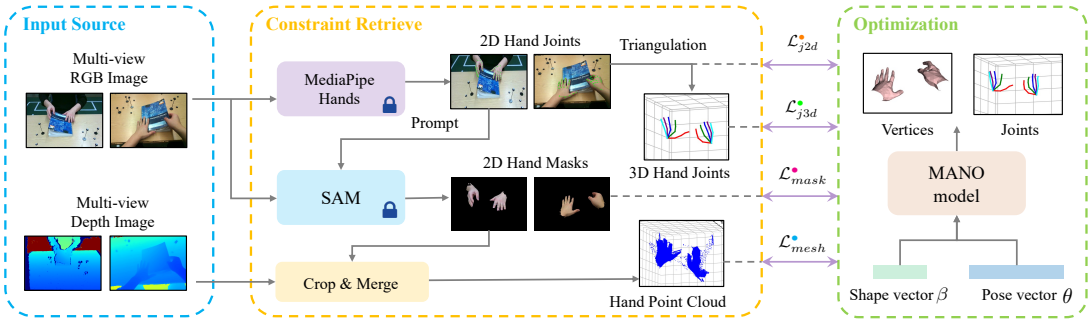

Figure 3.7: **Automatic annotation pipeline** of 3D hand pose. We utilize the multi-view RGB and depth images as the input source and retrieve both 2D and 3D constraint information. Various error terms are formulated to optimize the MANO parameters.

In particular, we use the MANO statistical hand model [75] to represent 3D hand pose. The MANO model parameterizes the hand mesh vertices $\mathcal{V}(\beta,\theta)$ into two low-dimensional embeddings, *i.e.*, the shape parameters $\beta \in \mathbb{R}^{10}$ and the pose parameters $\theta \in \mathbb{R}^{51}$, consisting of 45 parameters accounting for 15 hand joint angles (3 DoF for each) plus the rest for global rotation and translation. Following the MANO PyTorch version in [35], we denote the $N_\mathcal{J} = 21$ hand joints mapped from the hand parameters as $\mathcal{J}(\beta,\theta)$. The MANO fitting is performed by minimizing the following optimization objective per frame for each hand:

$$\theta^* = \arg\min_{\theta} \lambda_{j2d}\mathcal{L}^{\bullet}_{j2d} + \lambda_{mask}\mathcal{L}^{\bullet}_{mask} + \lambda_{j3d}\mathcal{L}^{\bullet}_{j3d} + \lambda_{mesh}\mathcal{L}^{\bullet}_{mesh} + \lambda_{reg}\mathcal{L}^{\bullet}_{reg} \tag{3.1}$$

where $\lambda_{j2d}, \lambda_{mask}, \lambda_{j3d}, \lambda_{mesh}, \lambda_{reg}$ are used to balance the weight of different errors. The diagram illustrating our annotation process is shown in Fig. 3.7.

Initialization. For each sequence, we optimize the shape parameter β[2] together with θ only at the first frame using Eq. (3.1) until convergence and fix its values for the subsequent frames. We initialize the pose parameter θ for each frame using the optimization result from the last frame. This helps to accelerate the convergence as well as keep the temporal consistency.

2D Joint Error $\mathcal{L}^{\bullet}_{j2d}$. Given RGB images from N_C viewpoints, we infer 2D hand joints $\mathcal{J}^{2D}$ using MediaPipe Hands [59] and define the 2D joint error as:

$$\mathcal{L}^{\bullet}_{j2d} = \sum_{c=1}^{N_C} \alpha_c \sum_{i=1}^{N_\mathcal{J}} ||\mathcal{J}^{2D}_{c,i} - \pi_c(\mathcal{J}(\theta)_i)|| \tag{3.2}$$

where $\pi_c(\cdot)$ returns the 2D projection location for 3D position in the c-th camera viewpoint, and α_c is hyperparameter used to weigh different viewpoints.

2D Mask Error $\mathcal{L}^{\bullet}_{mask}$. To generate the high-quality 2D hand mask, we prompt the prevalent Segment Anything Model [48] with the 2D hand joints $\mathcal{J}^{2D}$ and the bounding box derived from it. We penalize the distance between the hand mesh vertices $\mathcal{V}(\theta)$

[2]For simplicity, we omit β in Eq. (3.1) and subsequent equations.

and the 2D binary hand mask $\mathcal{M}_c$ as:

$$\mathcal{L}_{mask}^{\bullet} = \sum_{c=1}^{N_C} \alpha_c \sum_{i=1}^{N_{\mathcal{V}}} \min_j ||\mathcal{M}_{c,j} - \pi_c(\mathcal{V}(\theta)_i)|| \tag{3.3}$$

where $\mathcal{M}_{c,j}$ is the coordinate of j-th non-zero pixel in the mask $\mathcal{M}_c$.

3D Joint Error $\mathcal{L}_{j3d}^{\bullet}$. We triangulate the 2D joints from multiple views to lift them to 3D joints $\mathcal{J}^{3D}$ and measure their difference to $\mathcal{J}(\theta)$, which is written as:

$$\mathcal{L}_{j3d}^{\bullet} = \sum_{i=1}^{N_{\mathcal{J}}} ||\mathcal{J}_i^{3D} - \mathcal{J}(\theta)_i|| \tag{3.4}$$

3D Mesh Error $\mathcal{L}_{mesh}^{\bullet}$. To better supervise the hand mesh and fasten the optimization, we generate the 3D hand pose cloud $\mathcal{P}$ by cropping the depth image using the 2D hand mask and merging all views together. The 3D mesh error term compensates for the distance between the hand mesh and point cloud:

$$\mathcal{L}_{mesh}^{\bullet} = \sum_{i=1}^{N_{\mathcal{V}}} \min_j ||\mathcal{P}_j - \mathcal{V}(\theta)_i|| \tag{3.5}$$

Regularization $\mathcal{L}_{reg}^{\bullet}$. To alleviate irregular hand articulation, we constrain the joint angles to pre-defined lower and upper boundaries $\underline{\theta}$ and $\overline{\theta}$:

$$\mathcal{L}_{reg}^{\bullet} = \sum_{i=1}^{45} (\max(\underline{\theta_i} - \theta_i, 0) + \max(\theta_i - \overline{\theta_i}, 0)) \tag{3.6}$$

Acquisition of 2D Keypoint and Mask. We leverage the open-source MediaPipe Hands pipeline [59] to infer the 2D hand keypoints from both egocentric and exocentric RGB images. To improve the recall of hand detection, we modify the hyperparameters `min_detection_confidence` from the default 0.5 to 0.1. Sequential RGB images are fed into the MeidaPipe Hands pipeline to obtain the 2D keypoints for two hands. Particularly, we distinguish between two hands according to their relative locations on images. Given 2D keypoints estimated from the task-specific MeidiaPipe Hands model, we employ the versatile Segment-Anything Model (SAM) [52] to infer the 2D hand masks. To ensure the high quality of the generated 2D hand mask, we utilize the largest version of SAM, *i.e.*, ViT-L SAM model and prompt it with both the 2D hand keypoints and the bounding boxes defined by them. As a result, we acquire 2D hand keypoints and masks of two hands for each frame.

3D Keypoint Triangulation. After inferring the 2D hand keypoints from two views, we can obtain the positions of 3D hand keypoints using triangulation. To implement this, we use the OpenCV function `cv2.triangulatePoints` [7].

3D Hand Point Cloud Generation. To generate 3D hand point cloud, we first index the hand pixels with the 2D hand mask on the depth image and then convert them into 3D points with the camera intrinsic parameters. Points generated from the exocentric

view are transformed into the egocentric camera space. By merging points from two viewpoints, we can obtain a dense 3D hand point cloud for each hand.

Joint Angle Limitation. Following [32, 51, 58], we optimize the MANO model in the joint angle space instead of the PCA space so that we can constrain the joint angles explicitly. As said in the main paper, the joint angles are limited by a set of empirical boundary values during optimization. Specifically, we use the same parameters of the upper and lower joint angle boundaries as [51].

Shape Regularization. To avoid unrealistic hand shape, we also add regularization to the shape parameter β when optimizing it for the first frame of each sequence, which can be written as:

$$\mathcal{L}_{shape} = \sum_{i=1}^{10} ||\beta_i|| \tag{3.7}$$

MANO Fitting. We utilize the Pytorch version of the MANO layer proposed in [35]. The Adam optimizer [47] is used to minimize the optimization objective mentioned in the main paper. To obtain accurate 3D hand pose annotation, we optimize the MANO parameters individually for each frame instead of considering batches of them. For the first frame of each sequence, we run the optimization for 500 iterations with an initial learning rate of as 0.1 decayed by 0.9 for each 50 frames. For the subsequent frames, as we initialize with the results from the last frame, we only optimize for 60 iterations and initialize the learning rate as 0.05 to speed up the convergence.

Example Figures. We show more visualization examples of our 3D hand pose annotation in Fig. 3.8. As can be seen, our dataset provides high-fidelity and accurate hand pose annotations for various actions. Please refer to our supplementary videos for more annotation results.

Figure 3.8: Examples of 3D hand pose annotations. We show the left (blue) and right (red) hand 3D joints projected onto egocentric RGB images. From the same viewpoint, we also visualize the corresponding hand mesh annotation. These figures are selected from different subjects and meticulously *cropped* to highlight the hands on images.

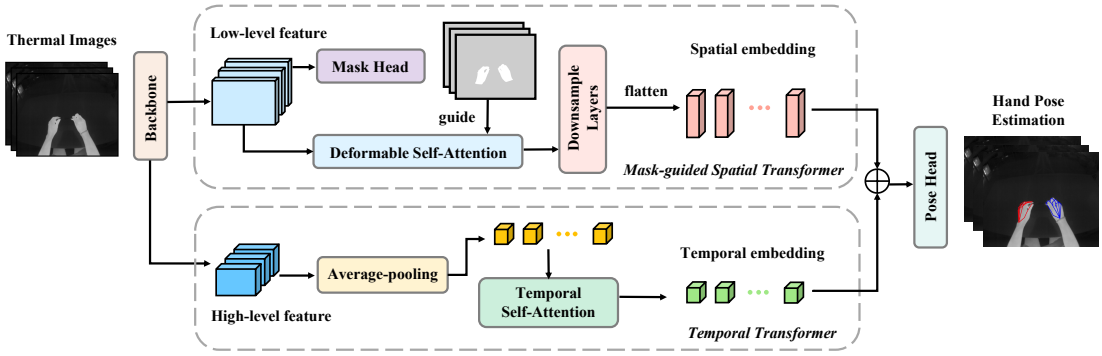

Figure 3.9: **Overall Framework of TheFormer.** Given thermal images, we extract both low-level and high-level features for each frame, which are input to the mask-guided spatial transformer and temporal transformer. Frame-wise spatial and temporal embeddings are concatenated and fed into the pose head to regress the 3D hand pose.

3.3 TheFormer: A Baseline Method

Our proposed baseline method, dubbed TheFormer, learns 3D hand pose estimation from egocentric thermal images based on Transformer, as exhibited in Fig. 3.9. The baseline network features in its two parallel transformer modules, *i.e.*, mask-guided spatial transformer and temporal transformer, to model the spatio-temporal relationship of hand joints while being computationally efficient.

3.3.1 Problem Definition

We consider two problem settings: single image-based and video-based egocentric 3D hand pose estimation for our benchmark. In the former setting, we aim to estimate the 3D joint positions $\mathcal{J}_t$ for two hands given the single thermal image I_t captured for the t-th frame. For the video-based one, our input is a sequence of thermal images $\mathcal{S} = \{I_i\}_{i=1}^{T}$ and we estimate the per-frame 3D hand joint positions $\mathcal{J}_i$ together. Compared to the single image-based counterpart, the video-based setting allows for a fully temporal interaction among sequential images, potentially offering a better performance. Nevertheless, the accumulation of sequential images leads to a higher latency for online inference, limiting its usage to applications that demand real-time hand tracking.

3.3.2 Network Architecture

Without losing the generality, here we illustrate our network architecture for the video-based setting. Note that our network can be flexibly adapted to the single image-based problem by setting $T = 1$. As seen in Fig. 3.9, given sequential thermal images, our architecture extracts the multi-level 2D feature using the backbone network per frame. We design a mask-guided spatial transformer module to enhance the spatial representation while enforcing the network to focus on the hand area. In parallel, we apply a temporal transformer module to reason the feature interaction along the temporal dimension. Both two transformers output per-frame feature embeddings that encode the

spatial and temporal information, respectively. We concatenate them and use a pose head to regress the 3D coordinates of hand joints for two hands.

3.3.3 Backbone

For efficiency, we use the lightweight ResNet-18 [36] as our backbone network to extract multi-level features from thermal images. Specifically, we reserve both the low-level and high-level image features and input them to the spatial and temporal transformer modules individually. Low-level features carry fine-grained spatial details, which is crucial for accurate mask generation and spatial reasoning in our mask-guided spatial transformer. High-level features provide a broader context, helping temporal transformers to better understand the overall scene dynamics and the hand's role within it.

The input to our backbone is consecutive thermal images $\mathcal{S} = \{I_i\}_{i=1}^{T}$ resized to 320×256 during data loading, where $T = 1$ for the single image setting and $T > 1$ for the video setting. We initialize the ResNet-18 [36] network with the ImageNet-1K [20] pre-trained weights for the multi-scale features extraction. The ResNet-18 [36] backbone returns feature at 5 levels, with feature sizes as follows given the thermal images as input:

syntax: $Feature([channel, width, height], \dots)$

$$Features([64,160,128],[64,80,64],[128,40,32],[256,20,16],[512,10,8]) \tag{3.8}$$

Features at the third and fifth levels are chosen for future processing. The lower-level feature is fed to the mask-guided spatial transformer module while the higher-level feature is used in the temporal transformer module.

3.3.4 Mask-guided Spatial Transformer

Human hands are highly articulated objects that can adopt a wide variety of poses, often against complex backgrounds. We propose a mask-guided spatial transformer module to accurately identify and focus on the intricacies of hand poses during spatial feature interaction. Given low-level features, we first utilize a mask head to estimate the binary hand mask in the thermal image. Then, we leverage the deformable self-attention [110] to refine the hand spatial features under the guidance of the hand mask. Specifically, we only take feature elements whose spatial locations are within the hand area as queries and sample keys from only the hand area and its surrounding locations. In this way, we not only reduce the computation waste on the irrelevant region but also increase the robustness to background clutter. Lastly, we reduce the spatial dimensions of the spatial features via several downsample convolutional layers and flatten it into 1D spatial feature embedding per frame.

For spatial feature refinement, we implement a mask-guided attention mechanism that emphasizes regions of interest within the thermal images, specifically focusing on the hands' positions. This approach utilizes predicted masks from a segmentation head with two convolution layers from the low-level feature to guide the attention mechanism.

The dimension of the spatial self-attention module we used is 1024, for which we use an additional one-stride convolution to lift the number of channels of the original low-level feature. Two transformer encoder layers are leveraged and the transformer head number M is set to be 8. The deformable attention mechanism introduced in deformable DETR [110] is used and the number of sampling points per attention head K is 4:

$$\text{DeformAttn}(z_q, p_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk} \cdot W'_m \left(x(p_q + \Delta p_{mqk}) \right) \right] \tag{3.9}$$

where x is the masked low-level feature, m indexes the attention head, Δp_{mqk} and A_{mqk} denote the sampling offset and the learnable attention weight for the k-th sampling point in the m-th attention head respectively. Particularly, A_{mqk} is a scalar attention weight lying between 0 and 1, and normalized such that the sum across all K points is 1, Δp_{mqk} is a pair of 2D real numbers representing an unconstrained range. Noted that 2-D reference point p_q is selected from the hand area derived from the predicted mask and Δp_{mqk} and A_{mqk} are obtained via linear projection over the query feature z_q. In our self-attention, z_q is the added feature for x and the corresponding position encoding. Following deformable DETR [110], the first $2M \times K$ channels encode the sampling offsets and the remaining $M \times K$ channels are for the softmax operation to obtain the attention weights.

The output spatial embedding of the deformable mask-guided spatial attention module has the same size as the input feature, which is [1024, 40, 32]. Three convolution layers with a stride of 2 are used to downsample the feature map to [128, 10, 8]. This higher-level feature is then flattened, and a linear layer is used to project this spatial embedding to [1,512] for future concatenation with the temporal embedding.

3.3.5 Temporal Transformer

To efficiently model temporal relationships, we apply the average-pooling to the high-level features to get the global feature vector per frame. The self-attention [95] is then employed to explicitly attend to the feature vector of every frame for a comprehensive temporal interaction. Similar to the spatial transformer, the output is frame-wise feature embedding that encompasses the temporal information, as mentioned previously in Sec. 2.2.5.

Our network also incorporates a temporal multi-head self-attention module, designed to process sequences of thermal images capturing the temporal dynamics inherent being performed. This module is crucial for understanding temporal patterns over consecutive frames, utilizing a transformer-based approach. A convolution layer is used for the high-level feature from the backbone to lift the channel dimension to 512, which is the channel dimension used in the temporal transformer. Then an average pooling is applied to aggregate the spatial space information. The resulting feature $x_t \in \mathbb{R}^{T \times 512}$ is then fed to the temporal attention encoder:

$$\text{Attention}(z_t, x_t) = \sum_{m=1}^{M} W_m \left[\sum_{k \in \Omega_k} A_{mtk} \cdot W'_m (x_t) \right] \tag{3.10}$$

where M stands for the number of attention heads, which is 8 in our implementation. $k \in \Omega_k$ indexes a key element with representation feature $x_k \in \mathbb{R}^C$, where $C = 512$, is the attention feature dimension. The resulting temporal embedding [T,512] is then rearranged, and the temporal dimension is flattened with the batch dimension and forms a feature vector [1,512]. We concatenated the temporal and spatial embeddings and fed this feature to an MLP head for the position prediction of 42 hand joints.

$$MLP(1024 \rightarrow 1024 \rightarrow 1024 \rightarrow 42 \times 3)$$

3.3.6 Loss Functions

Two losses are used for joint learning of hand segmentation and pose estimation during the training process. The L1 hand pose regression loss is defined as:

$$L_{Hand} = \frac{1}{42}(\|P^{2D} - P_{gt}^{2D}\|_1 + \lambda\|P^{depth} - P_{gt}^{depth}\|_1) \tag{3.11}$$

where P^{2D}, P_{gt}^{2D} are the 2D projection of estimated hand joint positions and ground truth hand joint positions. P^{depth} and P_{gt}^{depth} are the corresponding depth values for the hand joints. λ is the weight parameter and set as 100 for our training. The hand mask binary cross entropy loss is formulated as follows:

$$L_{Mask} = -\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left[w_{pos}\cdot M_{wh}\cdot\log(\hat{M}_{wh}) + w_{neg}\cdot(1-M_{wh})\cdot\log(1-\hat{M}_{wh})\right] \tag{3.12}$$

where M_{wh} is the binary ground truth label at pixel location (w,h) and $\hat{M}_{wh}$ is the predicted probability for the estimated mask. w_{pos} and w_{neg} are the weights to cope with the foreground-background imbalance. We set them to 30 : 1 during our training. Besides, we also leverage the binary cross-entropy loss to supervise the hand segmentation with the mask ground truth rendered from the annotated hand mesh (*cf*. Sec. 3.2). For 3D hand joint positions, we measure the *L1* distance of its 2D projection and depth to that of the ground truth separately.

3.3.7 Pose Regression

We concatenate the frame-wise spatial and temporal embedding from two transformer modules and get the spatio-temporal representations for each frame. In the pose head, we use the MLP to project the representations to the output space and obtain the per-frame 3D joint output $\mathcal{J}_i$.

For the video-based setting, our baseline method is named as TheFormer-V, which accepts sequential images as input. The variant of our baseline to the single image-based setting is called TheFormer-S, where the temporal transformer degrades to a global transformer that only encodes the global information for a single frame. Please see the supplementary materials for more parameters, design and training details of our baseline method.

Chapter 4

Experiments

4.1 Evaluation of the Annotation Method

As the first step, we validate the accuracy of our 3D hand pose annotation and analyze the impact for optimization results. For evaluation, we manually annotate two random sequences from our main dataset: one involving hand-object interaction and the other hand-virtual interaction, with a total of over 600 frames. To that end, we first annotate the 2D joint locations from both egocentric and exocentric images and obtain the 3D positions by triangulation. We calculate the average 3D joint errors across all frames to measure the accuracy.

As shown in Tab. 4.1, our annotation method achieves an average joint error of nearly 1cm, comparable to the results of [51, 32, 112]. The multi-view setting shows remarkably better precision than the ego-view only optimization, demonstrating the necessity of multi-camera capture. We also observe that only combining $\mathcal{L}_{mask}$ and $\mathcal{L}_{j2d}$ can already provide a plausible accuracy since they fit the projection of the 3D hand pose to two heterogeneous views. $\mathcal{L}_{mesh}$, though it hardly improves the joint accuracy, can result in more natural hand mesh. Adding $\mathcal{L}_{j3d}$ further refines the joints as it induces the explicit constraint to their positions. We showcase some annotation examples in Fig. 3.8 and Fig. 4.1. As can be seen, both hand joint and mesh can be accurately annotated across different actions despite the presence of occlusion and the variance in subject ethnicities.

Table 4.1: Evaluation of annotation results. The average 3D joint errors across all frames are reported (in cm). $\mathcal{L}_{reg}$ is used for all to mitigate irregular hand poses.

	Ego-view optimization		Multi-view optimization			
Errors	$\mathcal{L}_{mask} + \mathcal{L}_{j2d}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d} + \mathcal{L}_{mesh}$	$\mathcal{L}_{mask}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d} + \mathcal{L}_{mesh}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d} + \mathcal{L}_{mesh} + \mathcal{L}_{j3d}$
mean (std)	37.29 ($\pm$ 18.02)	7.03 ($\pm$ 2.57)	8.13 ($\pm$ 0.57)	1.29 ($\pm$ 0.43)	1.28 ($\pm$ 0.43)	1.01 ($\pm$ 0.34)

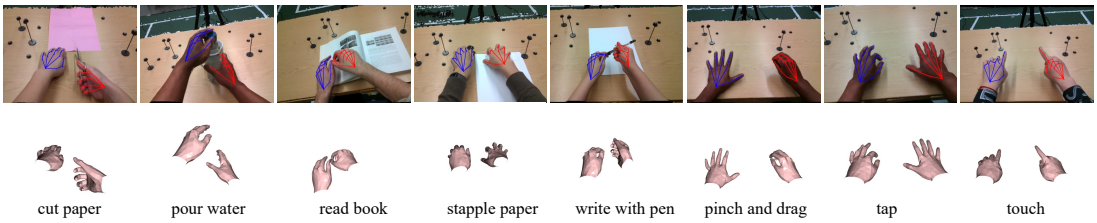

Figure 4.1: Examples of 3D hand pose annotations. Top row: left (blue) and right (red) hand 3D joints projected onto egocentric RGB images. Bottom row: visualization of hand mesh annotation (defined by MANO [75]).

4.2 Benchmark Setup

Dataset Preparation. We utilize our own dataset for experiments as it uniquely contains egocentric images from multiple spectra, essentially for our benchmark experiments. We annotate the main part of our dataset (*cf*. Tab. 2.1) in an automatic way following Sec. 3.2, of which the training and validation sets serve as the foundation for the training of all network models.

Method and Implementation. To provide sufficient baselines for follow-up works, we selected two state-of-the-art methods in 3D hand pose estimation, *i.e.*, HTT [100] and A2J-Transformer [40], and reproduce them on our dataset for benchmark experiments. HTT [100] is a video-based method thus enabling the evaluation in both two problem settings while A2J-Transformer [40] only works under the single image-based setting. For a feasible comparison, we use the same sequence length, *i.e.*, $T = 8$, for HTT and TheFormer-V baseline. We also exclude the additional action block used by HTT [100] for hand action prediction, focusing solely on hand pose estimation. We adjusted the anchor initialization phase of the A2J-Transformer [40] to better accommodate our dataset, without altering the density of its anchors. All trained models are tested on a single NVIDIA RTX 4090 GPU for a fair comparison of their inference speed.

Evaluation Metrics. We evaluate the accuracy of 3D hand pose estimation with two metrics: *Percentage of Correct Keypoints* (PCK) and *Mean End-Point Error* (MEPE) [111], in both camera space and root-aligned (RA) space [66]. For RA space, we align the estimated wrist with its groundtruth position before measurement. For PCK, we report the corresponding *Area Under the Curve* (AUC) over the 0-50mm/80mm error thresholds for the camera/RA space.

4.3 Benchmark Results

4.3.1 Comparison against State of the Arts.

We compare the TheFormer against the state-of-the-art methods for thermal-based 3D hand pose estimation, as shown in Tab. 4.2. On the MEPE and AUC metrics, TheFormer-S outforms two competing methods under the single image-based setting, while TheFormer-V surpasses the counterpart HTT [100] given the same sequential images as input. Such an improvement stems from our mask-guided spatial attention

Table 4.2: Benchmark results of TheFormer and state-of-the-art methods on thermal-based 3D hand pose estimation. Results are reported on the main testing set. ↑ denotes larger values are better, and vice versa. The fps displayed for sequence input indicates the number of sequences that models can process per second.

Method	Input	MEPE (mm) ↓	AUC ↑	MEPE-RA (mm) ↓	AUC-RA ↑	fps ↑
A2J-Transformer [40]	Single	51.68	0.474	20.76	**0.603**	34
HTT [100]	Single	49.09	0.489	**20.69**	0.599	**211**
TheFormer-S	Single	**48.25**	**0.510**	22.60	0.565	120
HTT [100]	Sequence	47.07	0.512	**17.49**	**0.659**	**129**
TheFormer-V	Sequence	**46.57**	**0.519**	19.62	0.619	67

Table 4.3: Comparison between different spectra. Models are trained with their corresponding spectrum images from the training set. We test them on the main testing set.

Method	Spectrum	MEPE (mm) ↓	AUC ↑	MEPE-RA (mm) ↓	AUC-RA ↑
HTT - Sequence [100]	RGB	43.30	0.542	**15.43**	**0.697**
	Depth	41.62	0.559	17.66	0.654
	NIR	**41.57**	**0.562**	16.41	0.679
	Thermal	46.57	0.519	17.49	0.659
A2J-Transformer - Single [40]	RGB	**40.21**	**0.577**	**16.77**	**0.675**
	Depth	43.80	0.552	18.26	0.647
	NIR	46.46	0.513	18.00	0.648
	Thermal	51.68	0.474	20.76	0.603

design that can adaptively encode the spatial interaction among hand joints with the guidance of the hand mask (*cf*. Sec. 3.3). A performance gap can be seen between single image-based and video-based settings for both HTT [100] and TheFormer. We credit this to their usage of temporal information that helps to tackle the occlusion cases and solve ambiguities. We can also observe that TheFormer has slightly inferior results in the RA space relative to its opponents, leaving room for improvement in localizing individual hand joints once hand roots are correctly identified. This also suggests that our advancements are primarily attributed to our accurate localization of the overall hand especially the wrists. Thanks to our lightweight network design, TheFormer is highly efficient to run with fps of 120 and 67 for single and sequence input respectively, ensuring its real-time application to resource-constrained devices.

4.3.2 Comparison between Spectrum.

To compare the efficacy of different spectra (*i.e*., RGB, depth, NIR and thermal), we assess their performance by applying them as varying inputs to the HTT [100] and A2J-Transformer [40] method. As seen in Tab. 4.3, RGB spectrum yields favourable performance for both two methods as they provide rich multi-channel semantics, allowing for accurate hand pose reasoning. Specifically, A2J-Transformer [40] achieves comparable performance to HTT [100] with sequential input. This can be attributed to its usage of the larger ResNet50 [36] backbone that is pretrained with RGB images. However, this also increases the computation overhead, making A2J-Transformer [40]

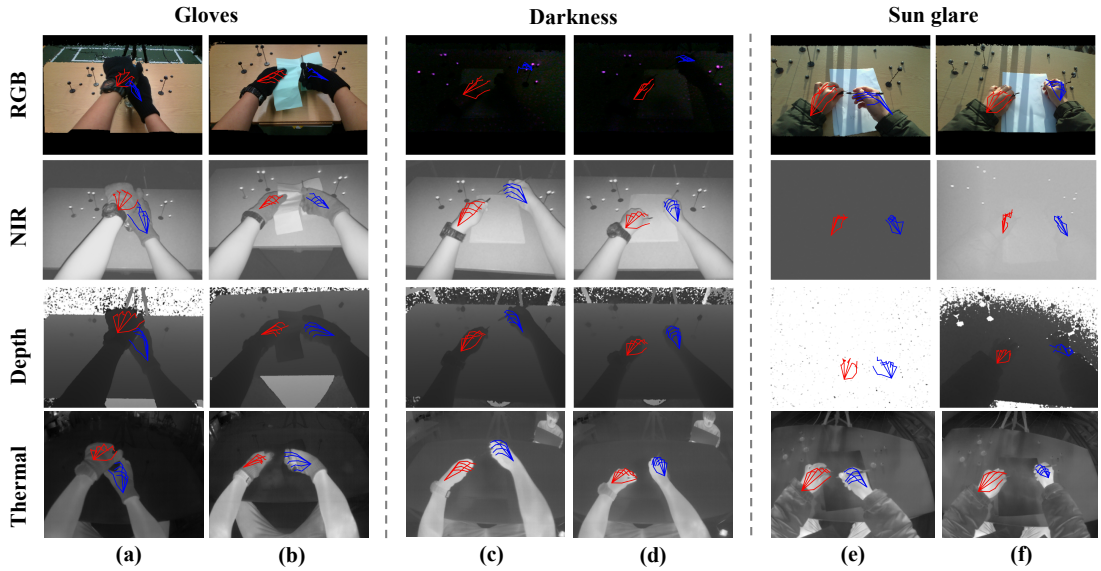

Figure 4.2: Examples of hand pose estimation results with various spectra under three challenging settings. For visualization, we show the projection of the estimated left (red) and right (blue) 3D hand joints on corresponding images.

6× slower than HTT [100] (*cf*. Tab. 4.2). Depth and NIR spectra show a marginal decrease in performance compared to RGB in well-illuminated conditions. However, the active illumination design of NIR sensors enables them to serve as supplements to RGB in dark conditions, as illustrated in Fig. 4.2. The thermal spectrum, operating independently of any active light source, still achieves plausible hand pose estimation accuracy, with a small discrepancy compared to other spectra. In addition to the above evaluation on the main testing set, we show more comparison of different spectra to highlight the value of thermal images under challenging conditions for 3D hand pose estimation.

4.4 Performance under Challenging Conditions

To justify the advantages of using thermal cameras for egocentric hand pose estimation under challenging scenarios, we conduct a comparison of four spectra on three of our auxiliary testing sets, *i.e*., gloves, darkness and sun glare (*cf*. Tab. 3.1). Our automatic annotation pipeline (*cf*. Sec. 3.2) becomes infeasible since hands appear corrupted in either RGB or depth images. Therefore, we conduct a qualitative analysis of their performance and present some representative examples in Fig. 4.2. As can be seen, RGB-based methods fail with gloves wearing (*cf*. Fig. 4.2 (a-b)) and in darkness (*cf*. Fig. 4.2 (c-d)). Gloves change how human hands look and hide their natural colors and textures. Since RGB algorithms rely on skin's texture and color to identify hand parts and joints, gloves, particularly those with solid colors or textures unlike skin, can interfere with this identification process. Contrary to RGB techniques, as shown in Fig. 4.2 (a-b), thermal imaging methods excel in identifying hands by leveraging the principles of heat conduction through gloves, effectively bypassing the limitations imposed by color and texture variations.

Table 4.4: Ablation study of two key transformer modules in TheFormer. We conduct this experiment in a single-image based setting where $T = 1$.

Method	MEPE (mm) ↓	AUC ↑	MEPE-RA (mm) ↓	AUC-RA ↑
TheFormer w.o temporal module	48.65	0.498	**19.93**	**0.612**
TheFormer w.o spatial module	49.09	0.489	20.69	0.599
TheFormer	**48.25**	**0.510**	22.60	0.565

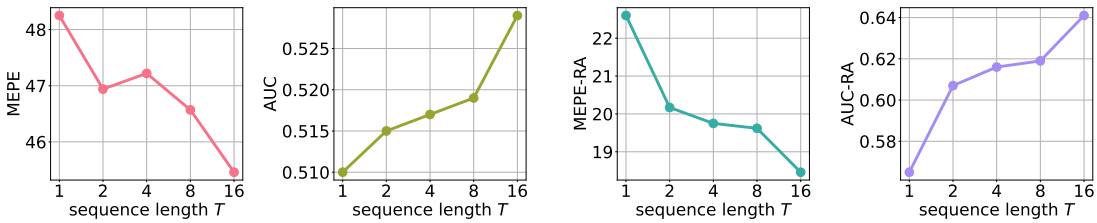

Figure 4.3: Ablation study of the temporal sequence length T in the TheFormer. The plots show all four metrics MEPE (mm) ↓, AUC ↑, MEPE-RA (mm) ↓, and AUC-RA ↑ against 5 sequence length settings, *i.e.*, $T \in \{1, 2, 4, 8, 16\}$.

NIR sensors are significantly disrupted by strong sunlight, affecting both NIR imaging and depth map creation, as shown in Fig. 4.2 (e-f). Conversely, thermal imaging is immune to sunlight and outdoor conditions. The temperature difference between hands and their surroundings in the thermal spectrum facilitates effortless identification of the hands, leaving the thermal-based estimator unaffected. The thermal camera's ability to consistently capture hand features in diverse lighting conditions positions it as a suitable option for future XR applications. Please see the supplementary materials for more figure examples and demo videos under challenging conditions.

4.4.1 Ablation Study

Here we conduct ablation studies to analyse the impact of two key transformer modules and the sequence length T on the performance of our baseline method. **Impact of Transformer Modules.** As seen in Tab. 4.4, both spatial and temporal transformer modules contribute to the performance growth of TheFormer, illustrating the effectiveness of our transformer module designs. Under the single-image based setting, our temporal transformer module degrades to a 'global' transformer module as it only applies self-attention to the global feature vector of a single frame. Therefore, the spatial module yields a larger enhancement than the temporal one. We also observe that leveraging either of the two transformer modules brings a drop to the performance in the RA space. We leave this to the future works for further investigation.

Impact of Sequence Length. We show the change of our performance against the sequence length in Fig. 4.3. It can be observed that the performance improves when the sequence length becomes larger. This emphasizes the ability of temporal information to elevate the 3D hand pose estimation accuracy.

4.4.2 Qualitative Results

Here we show more qualitative results and comparisons between spectra under different settings. To illustrate the property of different spectrum inputs We use the same method, *i.e.*, HTT [100], across the different spectrums for a fair comparison. Please see our supplementary video for more demo results in the video format.

Main Setting. We exhibit in Fig. 4.4 some qualitative results for different spectra on our main testing set, which is collected under the well-illuminated office scenario. As can be seen, each spectrum can provide reliable testing results, close to the ground truth annotations, validating the capability of our dataset to support 3D hand pose estimation research based on various spectra.

Kitchen Setting. To justify the generalization ability of our trained model to the new application environment, we test them on our kitchen testing set and show some qualitative results in Fig. 4.5. Specifically, we compare the results between the RGB and thermal spectrum. The thermal camera shows better performance than the RGB camera when generalized to an unseen environment. This can be credited to the unique attribute of thermal cameras that accentuates the hand's structure via temperature differentials, alleviating the effects of background variability. We believe that thermal image-based solutions could also generalize well to other environments, such as the dining hall and bathroom.

Glove Setting. In Fig. 4.6, we show more qualitative results for RGB and thermal spectrum in our glove testing set. As the hand appearance, including colour and textures, is greatly altered by the handwear like gloves in the RGB images, the RGB image-based solution fails to accurately estimate the 3D hand pose in our examples. In contrast, thermal cameras are able to correctly detect hands even under handwear, such as gloves by identifying heat transmission patterns.

Darkness Setting. More examples showing qualitative testing results under the darkness setting can be seen in Fig. 4.7. From the RGB images captured in the darkness, we can hardly recognize the hand contour even by human eyes. As a result, the estimated hand joints either exhibit irregular articulation or deviate significantly from their actual locations. Independent of the visible light, thermal cameras are unaffected by the variation of lighting conditions, successfully estimating 3D hand pose in the darkness.

Sun Glare Setting. We show more examples of testing results for NIR and thermal spectrum under the sun glare setting in Fig. 4.8. It can be seen that NIR images are prone to interference from the sunlight, which consists of the NIR lighting component. In comparison, thermal images are less affected and thus yield a better performance under the strong sun glare.

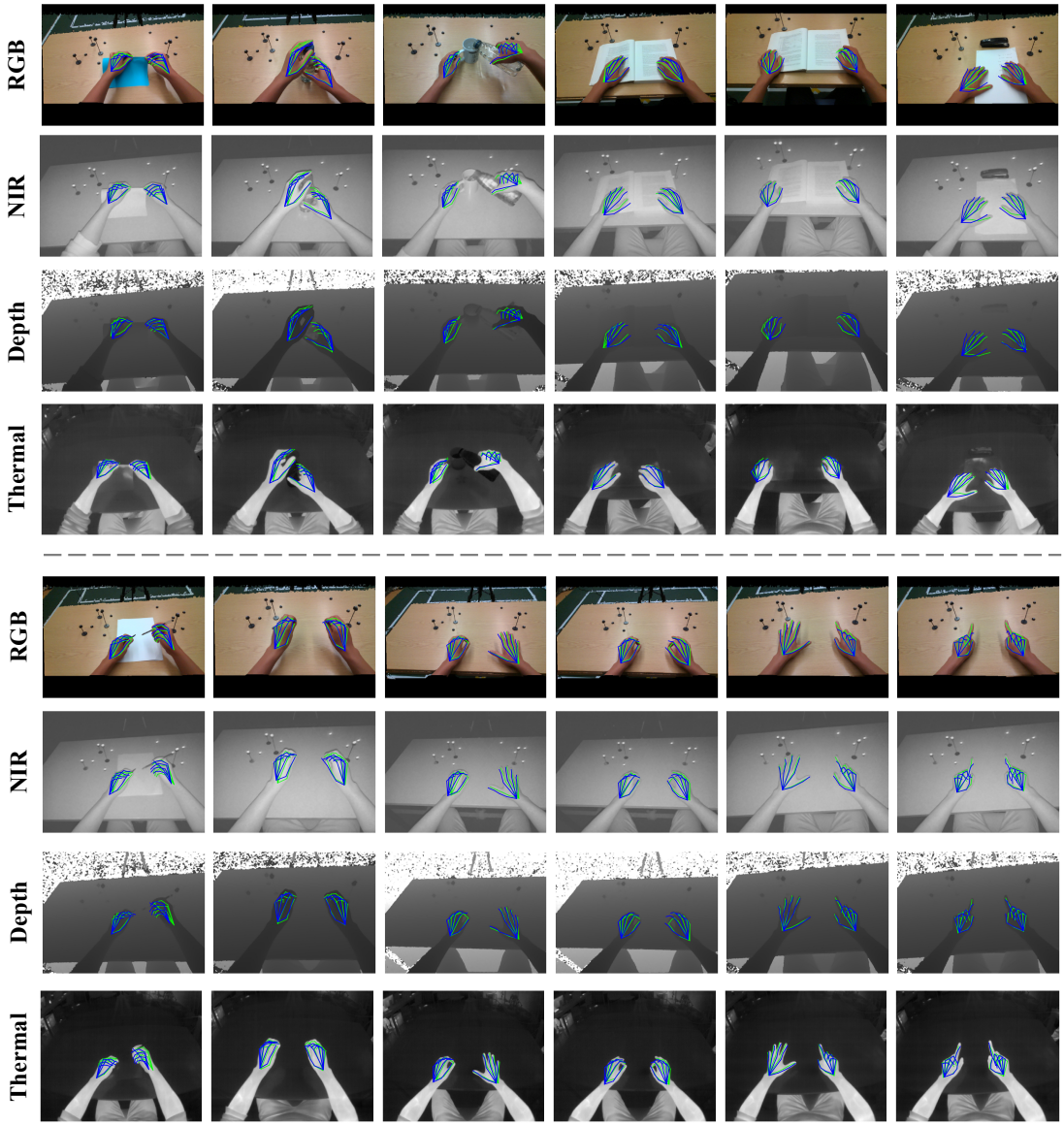

Figure 4.4: Qualitative results for different spectra under the **well-illuminated office (main)** setting. 3D hand joints are projected onto 2D images for visualization. Ground truth hand pose is shown in green while the prediction results in blue.

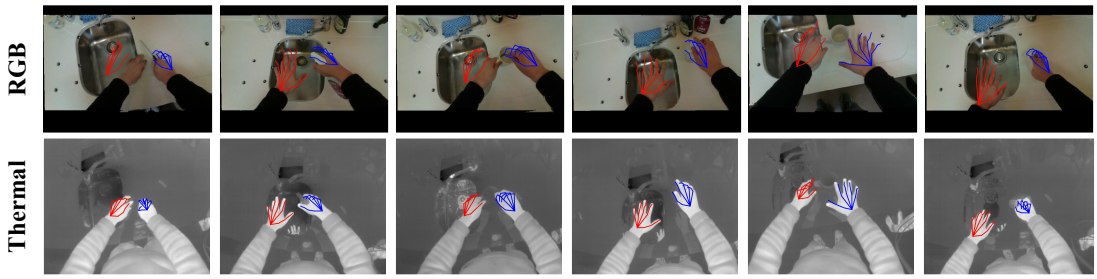

Figure 4.5: Qualitative results for RGB and thermal images when tested in the **kitchen** scenario with different actions. The left (red) and right (blue) hand 3D joints are projected onto egocentric images for visualization.

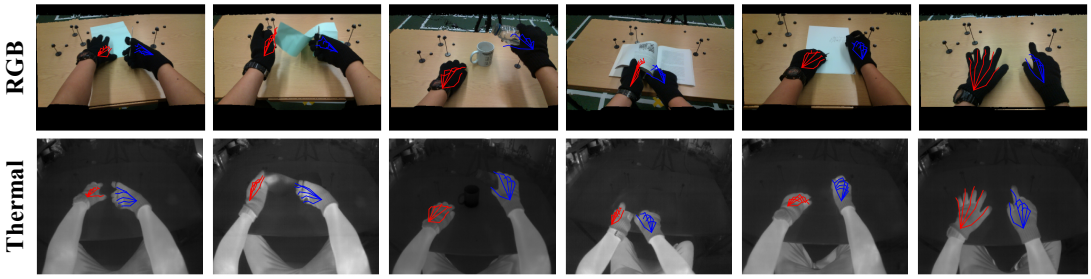

Figure 4.6: Qualitative results for RGB and thermal images when tested in the **glove** scenario where the participant wears a pair of black gloves.

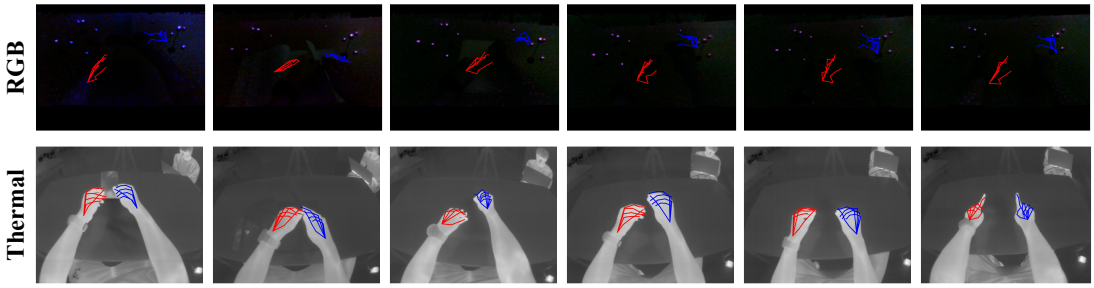

Figure 4.7: Qualitative results for RGB and thermal when tested in the **darkness** scenario.

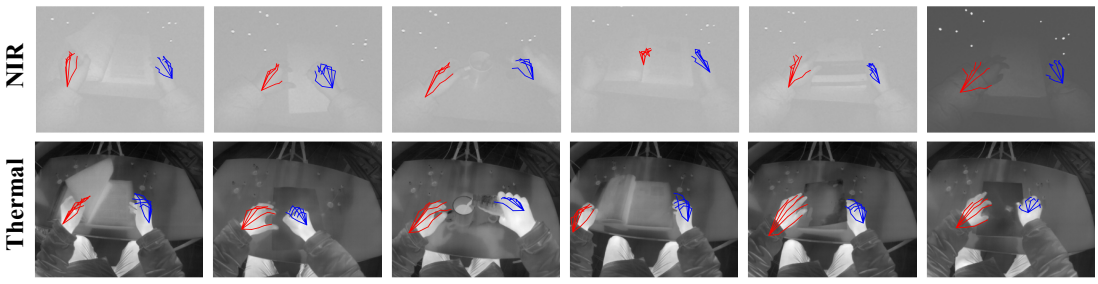

Figure 4.8: Qualitative results for NIR and thermal when tested in the **sun glare** scenario.

Chapter 5

Discussions and Conclusion

5.1 Limitation

This research, while marking a significant step forward in thermal-based egocentric 3D hand pose estimation, highlights key areas ripe for future exploration and improvement. Our dataset, albeit comprehensive, pales in comparison to the vast scale of existing datasets [23, 71, 58, 66] which feature millions of images (*cf*. Tab. 2.1). This limitation underscores an imminent need to expand our dataset to include a wider variety of scenarios and hand actions, aiming to improve model performance and adaptability to diverse real-world settings. Future work will also benefit from diversifying the dataset's utility beyond 3D hand pose annotation, to encompass fine-grained hand action annotations within sequences, annotating heat residuals on objects to generate contact heatmaps [8], and extending ground truth annotations to include both rigid [51] and articulated objects [23]. Such enhancements are poised to significantly increase the dataset's relevance and applicability across a spectrum of computer vision tasks involving intricate human hand interactions.

In addition, the reliance on thermal imaging presents unique challenges in data interpretation under variable environmental temperatures, where the differentiation between objects and the human body can become less distinct. Future work should explore adaptive algorithms that can dynamically adjust to these temperature variations, ensuring consistent performance across all conditions. Additionally, the current focus on hand pose estimation may overlook the integration of full-body gestures and their context, limiting the system's understanding of nuanced human-computer interactions. Expanding the research to include these aspects could vastly enhance the system's applicability and user experience in real-world applications, opening new avenues for comprehensive gesture-based control systems in various technology sectors.

5.2 Future Work

As thermal imaging sensors and AI-driven analysis tools continue to advance, there lies an massive potential to improve our methodology and achieve better accuracy and

reliability in hand pose estimation. The exploration of combination between thermal imaging and other spectral data could uncover novel strategies to overcome current limitations, particularly in complex environments. Moreover, the non-intrusiveness of thermal cameras enables multi-camera collaboration without interference. This unique feature of thermal cameras opens the door to transformative application in fields ranging from multi-MR device collaboration, human computer interaction and educational technology. The deployment of our methods to edge-devices, using state of the art methods for model compression, such as quantisation [9], could lead to significant advancements in consumer electronics, assistive technologies, etc, leading to a new era of intuitive and accessible ways of interaction.

5.3 Conclusion

In this thesis, we presented a comprehensive exploration into the advantages of thermal imaging for egocentric 3D hand pose estimation, a critical area of research with wide-ranging applications from extended reality (XR) to human-computer interaction. Our work includes the development of `ThermoHands`, the first dedicated benchmark for evaluating thermal imaging's efficacy in this domain, and the introduction of TheFormer, a novel baseline method optimized for thermal imagery.

The uniqueness of our approach lies in its resilience to the common challenges that affect traditional RGB and depth-based hand pose estimation methods, notably lighting variability and occlusions caused by handwear. By exploiting thermal imaging's ability to capture temperature differentials, we effectively overcome these challenges, demonstrating a robust solution applicable in diverse and adverse conditions. Our benchmark, `ThermoHands`, well illustrated this capability, featuring an extensive dataset that not only includes thermal, NIR, depth, and RGB spectra but also includes a wide variety of hand-object and hand-virtual interactions performed by participants from diverse backgrounds.

The experiments demonstrate TheFormer's outstanding performance compared with existing techniques, highlighting thermal imaging's potential to significantly extend hand pose estimation use cases. The dual transformer modules of TheFormer effectively encode the spatial and temporal dynamics of hand movements, setting a new standard in thermal imaging. Furthermore, there is massive potential for expanding `ThermoHands` to include a broader spectrum of activities and environmental conditions, further solidifying its position as the definitive benchmark for thermal-based hand pose estimation. Additionally, refining our annotation process and exploring new algorithmic improvements could unlock even higher levels of precision and real-time performance, benefiting the adoption in next-generation XR systems, robotics, and beyond.

In conclusion, our work represents a innovative use of thermal imaging for hand pose estimation. By addressing the limitations of current methods and introducing robust solutions, we open the door to new possibilities in interaction design, virtual reality, and automated systems. Our contributions not only advance the state of the art but also provide a solid foundation for future innovations in this rapidly evolving field.

Bibliography

[1] Dafni Antotsiou, Guillermo Garcia-Hernando, and Tae-Kyun Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *ECCVW*, 2018.

[2] Apple. Apple Vision Pro. `https://www.apple.com/apple-vision-pro/`, 2024. Accessed: 2024-02-23.

[3] Apple. Use gestures with Apple Vision Pro. `https://support.apple.com/en-us/117741`, 2024. Accessed: 2024-02-23.

[4] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019.

[5] Ganbayar Batchuluun, Dat Tien Nguyen, Tuyen Danh Pham, Chanhum Park, and Kang Ryoung Park. Action recognition from thermal videos. *IEEE Access*, 7:103893–103917, 2019.

[6] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019.

[7] Gary Bradski, Adrian Kaehler, et al. OpenCV: Open source computer vision library. `https://opencv.org/`, 2020.

[8] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, pages 361–378. Springer, 2020.

[9] Kaiwen Cai, Zhekai Duan, Gaowen Liu, Charles Fleming, and Chris Xiaoxuan Lu. Self-adapting large visual-language models to edge devices across visual modalities. *arXiv preprint arXiv:2403.04908*, 2024.

[10] Kaiwen Cai, Qiyue Xia, Peize Li, John Stankovic, and Chris Xiaoxuan Lu. Robust Human Detection under Visual Degradation via Thermal and mmWave Radar Fusion. In *EWSN*, 2023.

[11] Yujun Cai, 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. In *ICCV*, October 2019.

[12] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Open-

Pose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *PAMI*, 43(1):172–186, 2021.

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[14] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021.

[15] I-Chien Chen, Chang-Jen Wang, Chao-Kai Wen, and Shiow-Jyu Tzou. Multi-person pose estimation using thermal images. *IEEE Access*, 8:174964–174971, 2020.

[16] Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-DexHands: Towards Human-Level Bimanual Dexterous Manipulation. *PAMI*, 2023.

[17] Wencan Cheng, Jae Hyun Park, and Jong Hwan Ko. Handfoldingnet: A 3d hand pose estimation network using multiscale-feature guided folding of a 2d hand skeleton. In *ICCV*, pages 11260–11269, 2021.

[18] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-Based Unified Recognition of Two Hands Manipulating Objects. In *CVPR*, pages 4769–4778, 2023.

[19] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, pages 1964–1973, 2021.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[21] Meng Ding, Yuanyuan Ding, Li Wei, Yiming Xu, and Yunfeng Cao. Individual Surveillance Around Parked Aircraft at Nighttime: Thermal Infrared Vision-Based Human Action Recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1084–1094, 2022.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *CVPR*, pages 12943–12954, 2023.

[24] Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A

Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[25] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M. Kitani. Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation. In *ICCV*, pages 23600–23611, October 2023.

[26] Qing Gao, Yongquan Chen, Zhaojie Ju, and Yi Liang. Dynamic hand gesture recognition based on 3D hand pose estimation for human–robot interaction. *IEEE Sensors Journal*, 22(18):17421–17430, 2021.

[27] Qing Gao, Jinguo Liu, Zhaojie Ju, and Xin Zhang. Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation. *IEEE Transactions on Industrial Electronics*, 66(12):9663–9672, 2019.

[28] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, pages 409–419, 2018.

[29] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3D hand pose estimation from single depth images using multi-view CNNs. *IEEE Transactions on Image Processing*, 27(9):4422–4436, 2018.

[30] Honey Gupta and Kaushik Mitra. Toward unaligned guided thermal super-resolution. *TIP*, 31:433–445, 2021.

[31] Ali Haider, Furqan Shaukat, and Junaid Mir. Human detection in aerial thermal imaging using a fully convolutional regression network. *Infrared Physics & Technology*, 116:103796, 2021.

[32] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020.

[33] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, pages 571–580, 2020.

[34] 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*, pages 11807–11816, 2019.

[35] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[37] Intel RealSense. Depth Camera D455. `https://www.intelrealsense.com/depth-camera-d455/`, 2023. Accessed: 2024-02-27.

[38] Intel RealSense. LiDAR Camera L515. https://www.intelrealsense.com/lidar-camera-l515/, 2023. Accessed: 2024-02-27.

[39] Marina Ivašić-Kos, Mate Krišto, and Miran Pobar. Human detection in thermal imaging using YOLO. In *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, pages 20–24, 2019.

[40] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2J-Transformer: Anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *CVPR*, pages 8846–8855, 2023.

[41] Priya Kansal and Sabari Nathan. A multi-level supervision model: A novel approach for thermal image super resolution. In *CVPR*, pages 94–95, 2020.

[42] Leyla Khaleghi, Alireza Sepas-Moghaddam, Joshua Marshall, and Ali Etemad. Multi-view video-based 3D hand pose estimation. *IEEE Transactions on Artificial Intelligence*, 2022.

[43] Shehryar Khattak, Christos Papachristos, and Kostas Alexis. Keyframe-based thermal–inertial odometry. *Journal of Field Robotics*, 37(4):552–579, 2020.

[44] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *ICCV*, pages 11189–11198, 2021.

[45] Namil Kim, Yukyung Choi, Soonmin Hwang, and In So Kweon. Multispectral transfer network: Unsupervised depth estimation for all-day vision. In *AAAI*, 2018.

[46] Yeong-Hyeon Kim, Ukcheol Shin, Jinsun Park, and In So Kweon. MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. *RA-L*, 6(4):6497–6504, 2021.

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[48] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[49] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *CVPR*, June 2020.

[50] Zülfiye Kütük and Görkem Algan. Semantic segmentation for thermal images: A comparative survey. In *CVPR*, pages 286–295, 2022.

[51] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, pages 10138–10148, 2021.

[52] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation. *TNNLS*, 32(7):3069–3082, 2020.

[53] 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*, pages 2761–2770, 2022.

[54] Hui Liang, Junsong Yuan, Daniel Thalmann, and Nadia Magnenat Thalmann. Ar in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In *ACM MM*, pages 743–744, 2015.

[55] Fanqing Lin, Connor Wilhelm, and Tony Martinez. Two-hand global 3d pose estimation using monocular rgb. In *WACV*, pages 2373–2381, 2021.

[56] Ruoshi Liu and Carl Vondrick. Humans as Light Bulbs: 3D Human Reconstruction from Thermal Reflection. In *CVPR*, pages 12531–12542, 2023.

[57] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, pages 14687–14697, 2021.

[58] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *CVPR*, pages 21013–21022, 2022.

[59] Google LLC. MediaPipe Hands. `https://github.com/google/mediapipe`, 2020. Accessed: 2024-02-13.

[60] J Michael Lloyd. *Thermal imaging systems*. Springer Science & Business Media, 2013.

[61] Yawen Lu and Guoyu Lu. An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image. In *WACV*, pages 3833–3843, 2021.

[62] Marcos Lupión, Aurora Polo-Rodríguez, Javier Medina-Quero, Juan F Sanjuan, and Pilar M Ortigosa. 3D Human Pose Estimation from multi-view thermal vision sensors. *Information Fusion*, 104:102154, 2024.

[63] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *TVCG*, 22(12):2633–2651, 2015.

[64] Meta. Meta Quest. `https://www.meta.com/gb/quest/`, 2024. Accessed: 2024-02-23.

[65] Gyeongsik Moon, 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. In *CVPR*, pages 5079–5088, 2018.

[66] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, pages 548–564. Springer, 2020.

[67] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, pages 49–59, 2018.

[68] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *TOG*, 38(4):1–13, 2019.

[69] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCV*, pages 1154–1163, 2017.

[70] Author names not provided. Single shot corrective cnn for anatomically correct 3d hand pose estimation. *Frontiers in Artificial Intelligence*, 5, 2022. Lab: Haptics lab.

[71] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation. In *CVPR*, pages 12999–13008, 2023.

[72] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, pages 1496–1505, 2022.

[73] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*, pages 570–587. Springer, 2022.

[74] Rafael E Rivadeneira, Angel D Sappa, Boris X Vintimilla, Dai Bin, Li Ruodi, Li Shengye, Zhiwei Zhong, Xianming Liu, Junjun Jiang, and Chenyang Wang. Thermal Image Super-Resolution Challenge Results-PBVS 2023. In *CVPR*, pages 470–478, 2023.

[75] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *TOG*, 36(6), 2017.

[76] K Martin Sagayam and D Jude Hemanth. Hand posture and gesture recognition techniques for virtual reality applications: a survey. *Virtual Reality*, 21:91–107, 2017.

[77] Agata Sage, Daniel Ledwoń, Jan Juszczyk, and Paweł Badura. 3D Thermal Volume Reconstruction from 2D Infrared Images—a Preliminary Study. In *Innovations in Biomedical Engineering*, pages 371–379. Springer, 2021.

[78] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. Pose forecasting in industrial human-robot collaboration. In *ECCV*, pages 51–69. Springer, 2022.

[79] Muhamad Risqi U Saputra, Pedro PB de Gusmao, Chris Xiaoxuan Lu, Yasin Almalioglu, Stefano Rosa, Changhao Chen, Johan Wahlström, Wei Wang, Andrew Markham, and Niki Trigoni. Deeptio: A deep thermal-inertial odometry with visual hallucination. *RA-L*, 5(2):1672–1679, 2020.

[80] Saputra, Muhamad Risqi U. and Lu, Chris Xiaoxuan and de Gusmao, Pedro

Porto B. and Wang, Bing and Markham, Andrew and Trigoni, Niki. Graph-based thermal–inertial slam with probabilistic neural networks. *TRO*, 38(3):1875–1893, 2022.

[81] Sebastian Schramm, Phil Osterhold, Robert Schmoll, and Andreas Kroll. Combining modern 3D reconstruction and thermal imaging: Generation of large-scale 3D thermograms in real-time. *Quantitative InfraRed Thermography Journal*, 19(5):295–311, 2022.

[82] Lucas Adams Seewald, Vinicius Facco Rodrigues, Malte Ollenschläger, Rodolfo Stoffel Antunes, Cristiano André da Costa, Rodrigo da Rosa Righi, Luiz Gonzaga da Silveira Jr, Andreas Maier, Björn Eskofier, and Rebecca Fahrig. Toward analyzing mutual interference on infrared-enabled depth cameras. *Computer Vision and Image Understanding*, 178:1–15, 2019.

[83] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, pages 21096–21106, 2022.

[84] Ukcheol Shin, Kyunghyun Lee, Seokju Lee, and In So Kweon. Self-supervised depth and ego-motion estimation for monocular thermal video using multi-spectral consistency loss. *RA-L*, 7(2):1103–1110, 2021.

[85] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep Depth Estimation From Thermal Image. In *CVPR*, pages 1043–1053, June 2023.

[86] Young-Sik Shin and Ayoung Kim. Sparse depth enhanced direct thermal-infrared SLAM beyond the visible spectrum. *RA-L*, 4(3):2918–2925, 2019.

[87] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *TOG*, 39(6):1–14, 2020.

[88] Javier Smith, Patricio Loncomilla, and Javier Ruiz-Del-Solar. Human Pose Estimation using Thermal Images. *IEEE Access*, 2023.

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

[90] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, pages 294–310. Springer, 2016.

[91] Jesus Suarez and Robin R Murphy. Using the kinect for search and rescue robotics. In *SSRR*, pages 1–2. IEEE, 2012.

[92] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600. Springer, 2020.

[93] Trenton's Tech. Testing the Apple Vision Pro in Pitch Black Environments. YouTube, 2024. https://www.youtube.com/watch?v=wlPpo_QFIRU.

[94] Teledyne FLIR. *Boson - Uncooled, Longwave Infrared (LWIR) OEM Thermal Camera Module*, 2024. Accessed: 2024-02-13.

[95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017.

[96] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *IROS*, pages 8461–8468. IEEE, 2020.

[97] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. *RA-L*, 8(2):1029–1036, 2023.

[98] Eric A Wan and Rudolph Van Der Merwe. The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 153–158. Ieee, 2000.

[99] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *TOG*, 39(6):1–16, 2020.

[100] Yilin Wen, 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. In *CVPR*, pages 21243–21253, 2023.

[101] Min-Yu Wu, Pai-Wen Ting, Ya-Hui Tang, En-Te Chou, and Li-Chen Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *J. Vis. Commun. Image Represent.*, 70:102802, 2020.

[102] Fu Xiong, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019.

[103] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. SeqHAND: RGB-sequence-based 3D hand pose and shape estimation. In *ECCV*, pages 122–139. Springer, 2020.

[104] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *CVPR*, pages 9877–9886, 2019.

[105] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *CVPR*, pages 4866–4874, 2017.

[106] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *ICCV*, pages 11354–11363, 2021.

[107] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019.

[108] Zhaohui Zhang, Shipeng Xie, Mingxiu Chen, and Haichao Zhu. HandAugment: A simple data augmentation method for depth-based 3D hand pose estimation. *arXiv preprint arXiv:2001.00702*, 2020.

[109] Tianqiang Zhu, Yi Sun, Xiaohong Ma, and Xiangbo Lin. Hand Pose Ensemble Learning Based on Grouping Features of Hand Point Sets. In *ICCVW*, 2019.

[110] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.

[111] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017.

[112] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019.

Appendix A

Participants' information sheet

Participant Information Sheet

Project title:	Transformer Approach for Egocentric Thermal Hand Pose Estimation
Principal investigator:	Chris Xiaoxuan Lu
Researcher collecting data:	Lawrence Zhu, Fangqiang Ding
Funder (if applicable):	

This study was certified according to the Informatics Research Ethics Process, reference number 980730 Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

UG student Lawrence Zhu, PhD student Fangqiang Ding and their supervisor Chris Xiaoxuan Lu, are the researchers and will have access to the data.

What is the purpose of the study?

This study aims to develop a transformer-based hand pose estimation method using egocentric thermal hand pose images and to construct the first egocentric hand pose dataset featuring hand actions performed by a diverse group of volunteers. Human participants are required to perform hand actions while wearing a head-mounted sensor platform, which also obscures their facial data from external cameras. Their motions will be captured by a head-mounted egocentric sensor platform and a exocentric RGB-D platform. The dataset will include RGB and thermal images of the participants' hands, but facial data will not be included. These egocentric images will then be used to estimate hand poses, which are represented as skeletons with 21 joints per hand, using existing algorithms. These hand poses will serve as the ground truth for training the hand pose estimation algorithm.

Why have I been asked to take part?

You have been invited to take part in the study because you volunteered to contribute to data collection during this study.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, up until June 2024 without giving a reason. After this point, personal data will be deleted and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI. We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

Specify:

- Kinds of data being collected: egocentric RGB-D, IR and thermal images, exocentric RGB, IR images.
- Means of collection: A RGB-D camera platform that placed in front of the participant and a head-mounted sensor platform equipped with a thermal camera and an Intel RealSense L515 LiDAR camera.
- Duration of session: 25-40 minutes per person.
- If participant audio/video is being recorded: Yes.
- How often, where, when: only one time, at G.17 Informatics Forum, any day-time you would like.
- Data publication: the well-curated dataset based on the recorded data will be made public to other researchers. The dataset will include the egocentric hand pose data and the exocentric images of participants performing specific hand actions.
- Identity protection: The head-mounted sensor platform is a VR like wearable device that obscures participants' facial data from external cameras.

Compensation. [only required if applicable]

N/A.

Are there any risks associated with taking part?

There are no significant risks associated with participation.

Are there any benefits associated with taking part?

No.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any

information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a maximum of 4 years All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymization.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher/research team, including Lawrence Zhu, Fangqiang Ding and Chris Xiaoxuan Lu.

All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Dr.Chris Xiaoxuan Lu (xiaoxuan.lu@ed.ac.uk).

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be made available on http://web.inf.ed.ac.uk/infweb/research/study-updates.

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Lawrence Zhu (s2020514@ed.ac.uk)

General information.

For general information about how we use your data, go to: edin.ac/privacy-research

Appendix B

Participants' consent form

Participant Consent Form

Project title:	Transformer Approach for Egocentric Thermal Hand Pose Estimation
Principal investigator (PI):	Chris Xiaoxuan Lu
Researcher:	Lawrence Zhu, Fangqiang Ding
PI contact details:	xiaoxuan.lu@ed.ac.uk

By participating in the study you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.
- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I consent to my anonymised data being used in academic publications and presentations.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

		Yes	No
1.	I agree to be recorded by exocentric RGBD cameras when performing hand actions.	√	
2.	I agree to wear a hand-mounted sensor platform with a thermal camera and Intel RealSense L515 LiDAR Camera to record egocentric hand actions when performing.	√	
3.	I allow the researchers to generate the mesh data and the skeleton of my hand with both egocentric and exocentric data.	√	
4.	I allow my data to be used for research proposals and made public by the researchers as a part of an egocentric thermal hand pose dataset.	√	
5.	I agree to take part in this study.	√	

Name of person giving consent	Date dd/mm/yy	Signature

Name of person taking consent	Date dd/mm/yy	Signature
Lawrence Zhu		

THE UNIVERSITY of EDINBURGH
informatics