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

Fangqiang Ding*
University of Edinburgh
Edinburgh, United Kingdom
f.ding-1@sms.ed.ac.uk

Yunzhou Zhu*
Georgia Institute of Technology
Atlanta, USA
lawrencezhu@gatech.edu

Xiangyu Wen
University of Edinburgh
Edinburgh, United Kingdom
wenxiangyu2001@gmail.com

Gaowen Liu
Cisco Research
San Francisco, USA
gwliu213@gmail.com

Chris Xiaoxuan Lu†
University College London
London, United Kingdom
xiaoxuan.lu@ucl.ac.uk

Abstract

Designing egocentric 3D hand pose estimation systems that can perform reliably in complex, real-world scenarios is crucial for downstream applications. Previous approaches using RGB or NIR imagery struggle in challenging conditions: RGB methods are susceptible to lighting variations and obstructions like handwear, while NIR techniques can be disrupted by sunlight or interference from other NIR-equipped devices. To address these limitations, we present ThermoHands, the first benchmark focused on thermal image-based egocentric 3D hand pose estimation, demonstrating the potential of thermal imaging to achieve robust performance under these conditions. The benchmark includes a multi-view and multi-spectral dataset collected from 28 subjects performing hand-object and hand-virtual interactions under diverse scenarios, accurately annotated with 3D hand poses through an automated process. We introduce a new baseline method, TherFormer, utilizing dual transformer modules for effective egocentric 3D hand pose estimation in thermal imagery. Our experimental results highlight TherFormer's leading performance and affirm thermal imaging's effectiveness in enabling robust 3D hand pose estimation in adverse conditions.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; • **Computer systems organization → Embedded and cyber-physical systems**; • **Human-centered computing → Ubiquitous and mobile computing**;

Keywords

3D Hand Pose Estimation, Thermal Vision, Hand Pose Dataset

ACM Reference Format:
Fangqiang Ding, Yunzhou Zhu, Xiangyu Wen, Gaowen Liu, and Chris Xiaoxuan Lu. 2025. ThermoHands: A Benchmark for 3D Hand Pose Estimation from Egocentric Thermal Images. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715014.3722058

1 Introduction

Egocentric 3D hand pose estimation is critically important for interpreting hand gestures across various applications, ranging from extended reality (XR) [1–4], to human-robot interaction [5–7], and to imitation learning [8–10]. Its importance has been magnified with the advent of advanced XR headsets such as the Meta Quest series [11] and Apple Vision Pro [12], where it serves as a cornerstone for spatial interaction and immersive digital experiences.

While current research of hand pose estimation primarily focuses on RGB image-based methods [13–17], these approaches are particularly vulnerable to issues related to lighting variation and occlusions caused by handwear, *e.g.*, gloves or large jewellery [18, 19]. 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 [20, 21] and are prone to interference from external NIR sources, such as sunlight [22] and other NIR-equipped devices [23]. 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.

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 [24]. This unique attribute of thermal imaging introduces several benefits for hand pose estimation. Primarily, it accentuates the hand's structure via temperature differentials, negating the effects of lighting variability. Moreover, thermal cameras can detect hands even under

[*] Equal contribution.
[†] Corresponding author: xiaoxuan.lu@ucl.ac.uk

SenSys '25, May 6–9, 2025, Irvine, CA, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1479-5/25/05
https://doi.org/10.1145/3715014.3722058

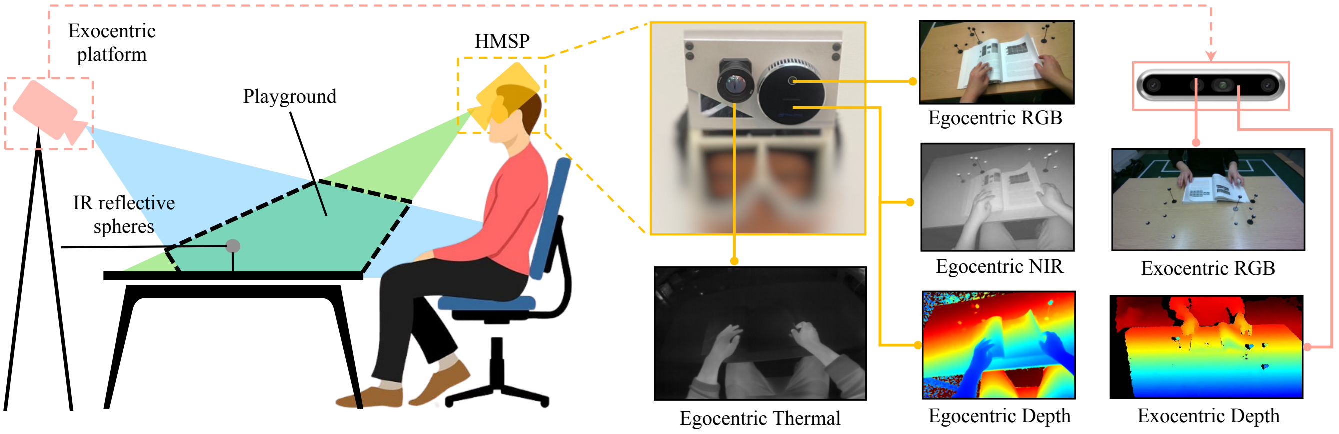

Figure 1: Data capture setup with the customized head-mounted sensor platform (HMSP) and exocentric platform recording multi-view multi-spectral images of two-hand actions performed by participants.

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[2], 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*. Fig. 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. To offer a thorough comparison across spectral types, we gather data under five distinct scenarios, each characterized by varying environments, handwear, and lighting conditions (*cf*. Tab. 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 [25] (*cf*. Fig. 4).

Together with the multi-spectral dataset, we introduce a new baseline method named *TherFormer*, specifically designed for thermal image-based egocentric 3D hand pose estimation (*cf*. Fig. 5). This approach is notable for its two consecutive 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. 2). We then benchmark *TherFormer* against leading methods (*cf*. Tab. 3) and compare the performance of various spectral images (*cf*. Tab. 5, Tab. 4 and Fig. 8). 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. Our main contributions are summarized as follows:

- We introduce the first-of-its-kind benchmark, 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 *TherFormer*, and implement state-of-the-art image-based methods on our dataset for benchmarking.
- Based on the ThermoHands benchmark, we conduct comprehensive experiments and analysis on TherFormer and state-of-the-art methods.
- We release our dataset, code and models and maintain the benchmark at https://github.com/LawrenceZ22/ThermoHands.

2 Related Works

2.1 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 [26–29], synthetic [30–34], manual [30, 35] or hybrid [36–40], and automatic [41–44] annotated datasets. Marker-based approaches, using magnetic sensor [26, 27] or Mocap markers [28, 29], can alter and induce bias to the hand appearance. Synthetic data [30–34] suffers from the *sim2real* gap in terms of hand motion and texture features. Introducing human

[1] For readability, we treat depth and NIR as two 'spectra', despite their usual overlap.
[2] Project page: https://thermohands.github.io/.

	Normal 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	-

Table 1: Benchmark Dataset Statistics. The overall duration of our dataset is over 3 hours with ~96K synchronized frame of all types of images collected.

annotators to fully [30, 35] or partly [36–40] 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 [41–43], which leverage pre-trained models (*e.g.* OpenPose [45]) to infer the prior hand information and rely on optimization to fit the MANO hand model [25].

Despite the existing progress, previous datasets only provide depth [26, 32], RGB images [33, 37, 38, 40] or both of them [27, 28, 30, 31, 34–36, 39, 41–43] 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.

2.2 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 [1–4], human-robot interaction [5–7] and imitation learning [8–10]. Therefore, this field has been extensively explored in previous arts that uses single RGB [13, 14, 31, 38, 46–57] or depth [58–62] image as input. These methods can be roughly categorized into two fashions, *i.e.*, model-based and model-free methods. Model-based methods [14, 31, 51–57] utilize the prior knowledge of the MANO hand model [25] by estimating its shape and pose parameters, while model-free methods [13, 38, 46–50, 58–62] learn the direct regression of 3D hand joints or vertices coordinates. Recently there has been growing interest in leveraging the temporal supervision [54, 63–66] or leveraging sequential images as input [15–17, 67–69] 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.

2.3 Thermal Computer Vision

Thermal cameras achieve imaging by capturing the radiation emitted in the LWIR spectrum and deducing the temperature distribution on the surfaces [24]. Leveraging its robustness to variable illumination and unique temperature information, numerous efforts have been made to address various computer vision tasks, including super-resolution [70–72], human detection [73–75], action recognition [76, 77] and pose estimation [78–80], semantic segmentation [81–84], depth estimation [85–88], visual(-inertial) odometry/SLAM [89–92], 3D reconstruction [93–95], *etc.* In this work, we focus on 3D hand pose estimation, which is an under-exploited task based on thermal images.

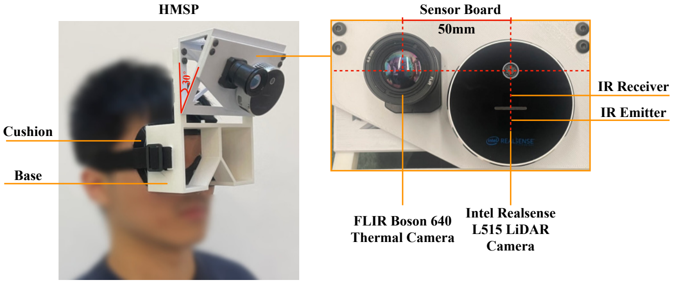
HMSP
Sensor Board
50mm
30°
IR Receiver
IR Emitter
Cushion
Base
FLIR Boson 640 Thermal Camera
Intel Realsense L515 LiDAR Camera

Figure 2: Design of the head-mounted sensor platform and sensor alignment.

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. 1), capturing hand actions performed by 28 subjects of various ethnicities and genders[3]. As shown in Fig. 1, we develop a customized head-mounted sensor platform (HMSP) and an exocentric platform to record multi-view data. During the 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 normal 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.

Sensor Platforms. Apart from the traditional approach of mounting cameras on helmets [43], the design of the HMSP (*cf*. Fig. 2) focuses on simulating an actual Mixed Reality (MR) device and reducing extra weight to allow participants to perform freely. The HMSP consists of three major components: a cushion for comfort, a base component that provides a 30-degree downward tilt, and a sensor board that carries two cameras - an Intel RealSense L515 LiDAR camera [20] that streams egocentric RGB, depth, and NIR images, and a Teledyne FLIR Boson 640 long-wave infrared (LWIR) camera [96] that receives the LWIR to obtain the thermal images.

[3] The study has received the ethical approval from University of Edinburgh, and participant consent forms were signed before the collection.

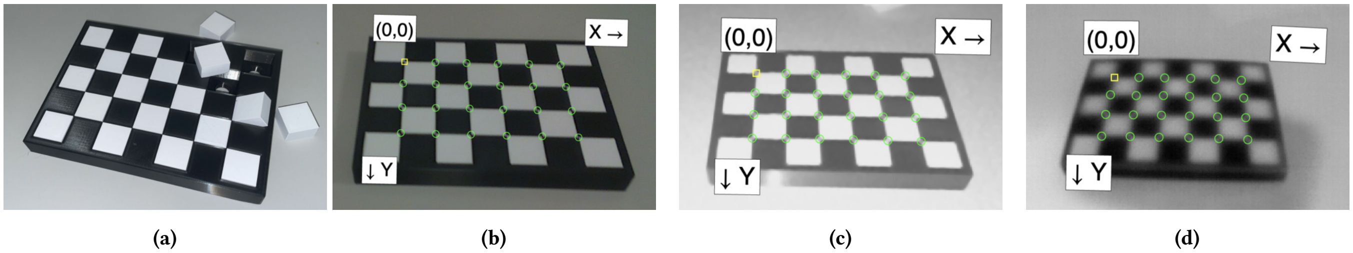

Figure 3: Thermal calibration chessboard containing a black base board and multiple removable white cubes (a). By cooling down the base board, it shows similar patterns and allows automatic corner detection in all (b) RGB, (c) NIR and (d) thermal images.

An extra exocentric platform equipped with an Intel RealSense D455 [21] is leveraged to support multi-view annotation (*cf*. Sec. 4) as well as provide the RGB-D image data from the third-person viewpoint. As exhibited in Fig. 1, we place the two depth sensors outside each other's field of view (FoV) to minimize interference caused by their NIR emitters [20, 21].

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 well-aligned multi-spectral, multi-view data samples as our released data.

Egocentric Calibration. For accuracy, factory-calibrated intrinsic parameters are used for the Intel RealSense D455 and L515 cameras [20, 21]. As seen in Fig. 2, although the thermal camera's center is aligned with the RGB camera's center at the same height on the 3D-printed sensor board, we still calibrate the extrinsic parameters between these two egocentric cameras to ensure enhanced alignment accuracy. To accommodate the unique imaging mechanism of thermal camera, we self-design a modular calibration chessboard as shown in Fig. 3. 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. In this way, we can simultaneously calibrate the intrinsic parameter of the thermal camera and its extrinsic parameter w.r.t. the D455 RGB-D camera.

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 [97]. 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 [98]. 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 [99] 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.

Dataset Statistics. As seen in Tab. 1, our dataset consists of approximately 96K synchronized multi-spectral multi-view frames (*cf*. Fig. 1) and 352 independent sequences in total. The main part is collected under the normal office scenario, where each participant[4] performs 7 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.

4 Hand Pose Annotation

To avoid employing tedious human efforts for annotation, we implement a fully automatic annotation pipeline, similar to the approaches in [41–43], to obtain the 3D hand pose ground truth for our dataset. In particular, we use the MANO statistical hand model [25] 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 [102], 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:

$$\begin{aligned}\theta^* = \arg\min_{\theta} \lambda_{j2d}\mathcal{L}^{\bullet}_{j2d} + \lambda_{mask}\mathcal{L}^{\bullet}_{mask} + \lambda_{j3d}\mathcal{L}^{\bullet}_{j3d} \\ + \lambda_{mesh}\mathcal{L}^{\bullet}_{mesh} + \lambda_{reg}\mathcal{L}^{\bullet}_{reg}\end{aligned} \tag{1}$$

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. 4.

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

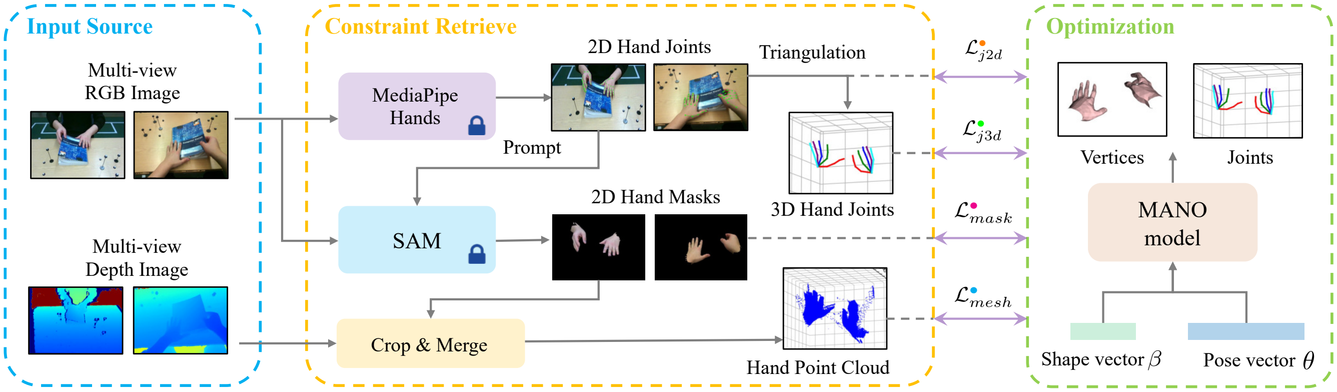

Figure 4: Automatic annotation pipeline of 3D hand pose. We utilize the multi-view RGB and depth images as the input source and retrieve constraint information with off-the-shelf MediaPipe Hands [100] and SAM [101]. Various error terms are formulated to optimize the MANO parameters.

Initialization. For each sequence, we optimize the shape parameter β[5] together with θ only at the first frame using Eq. (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 $\mathcal{J}^{2D}$ with MediaPipe Hands [100] and define the 2D joint error as:

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

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 [101] with the 2D hand joints $\mathcal{J}^{2D}$ and the bounding box derived from it. We penalize the distance between the hand mesh vertices $\mathcal{V}(\theta)$ and the binary hand mask $\mathcal{M}_c$ as:

$$\mathcal{L}^{\bullet}_{mask} = \sum_{c=1}^{N_C} \alpha_c \sum_{i=1}^{N_\mathcal{V}} \min_j ||\mathcal{M}_{c,j} - \pi_c(\mathcal{V}(\theta)_i)|| \tag{3}$$

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

3D Joint Error $\mathcal{L}^{\bullet}_{j3d}$. 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:

$$\mathcal{L}^{\bullet}_{j3d} = \sum_{i=1}^{N_\mathcal{J}} ||\mathcal{J}^{3D}_i - \mathcal{J}(\theta)_i|| \tag{4}$$

3D Mesh Error $\mathcal{L}^{\bullet}_{mesh}$. 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:

$$\mathcal{L}^{\bullet}_{mesh} = \sum_{i=1}^{N_\mathcal{V}} \min_j ||\mathcal{P}_j - \mathcal{V}(\theta)_i|| \tag{5}$$

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

$$\mathcal{L}^{\bullet}_{reg} = \sum_{i=1}^{45} (\max(\underline{\theta_i} - \theta_i, 0) + \max(\theta_i - \overline{\theta_i}, 0)) \tag{6}$$

Note that we also impose regularization to the shape parameter β during initialization. For more implementation details of our dataset annotation process, please refer to Appendix B.

5 TherFormer: A Baseline Method

For our benchmark, we setup a baseline method, dubbed TherFormer, for thermal image-based 3D hand pose estimation. As exhibited in Fig. 5, TherFormer features in its two consecutive transformer modules to model the spatio-temporal relationship of hand joints while being computationally efficient. Note that TherFormer is also capable of processing other spectral images due to their format consistency with thermal images. However, our primary objective is to establish a baseline method for future research, rather than focusing exclusively on methods for thermal images.

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 $\mathcal{I}_t$ captured for the t-th frame. For the video-based one, our input is a sequence of thermal images $\mathcal{S} = \{\mathcal{I}_i\}_{i=1}^{T}$ and we estimate the per-frame 3D hand joint positions $\mathcal{J}_i$ together. Without losing the generality, here we illustrate our network architecture for the video-based setting, *i.e.*, TherFormer-V. In practice, our network can be flexibly adapted to the single image-based version (*i.e.*, TherFormer-S) by setting $T = 1$.

[5] For simplicity, we omit β in Eq. (1) and subsequent equations.

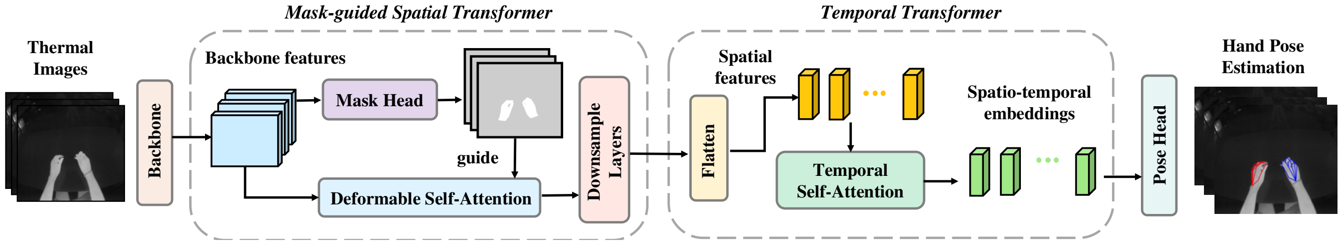

Figure 5: Overall Framework of TherFormer. Backbone features are input to the mask-guided spatial transformer and temporal transformer to enhance the spatial representation and temporal interaction. Spatio-temporal embeddings are fed into the pose head to regress the 3D hand pose.

	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 (± 18.02)	7.03 (± 2.57)	8.13 (± 0.57)	1.29 (± 0.43)	1.28 (± 0.43)	1.01 (± 0.34)

Table 2: 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.

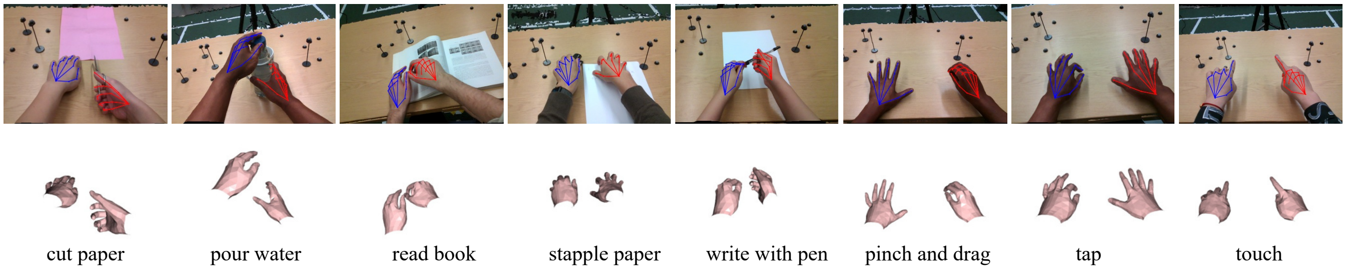

Figure 6: Examples of 3D hand pose annotations. Top row: left (blue) and right (red) hand 3D joints projected onto egocentric RGB images. Bottom row: visualization of hand mesh annotation.

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 backbone features, we first utilize a mask head to estimate the binary hand mask in the thermal image. Then, we leverage the deformable self-attention [103] to refine the hand spatial features under the guidance of the estimated 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 with a series of convolutions layers for efficiency.

Temporal Transformer. Temporal information is crucial for 3D hand pose estimation when coping with occlusion and solving ambiguities. To model temporal relationships, we first flatten the spatial features into 1D feature vectors and then employed the temporal self-attention [104] to explicitly attend to the feature vector of every frame. The output is frame-wise spatio-temporal feature embeddings. Note that the temporal self-attention degrades to an MLP for single-image based setting.

In the pose head, we use the MLP to project the embeddings to the output space and obtain the per-frame 3D joint $\mathcal{J}_i$. We leverage the binary cross-entropy loss to supervise the hand segmentation with the mask ground truth rendered from the annotated hand mesh (*cf*. Sec. 4). For 3D hand joint positions, we measure the *L1* distance of its 2D projection and depth to that of the ground truth separately.

6 Experiment

6.1 Evaluation of the Annotation Method

As the first step, we validate the accuracy of our 3D hand pose annotation (*cf*. Sec. 4) and analyze the impact on optimization results. For evaluation, we manually annotate two random sequences from our main dataset, with a total of over 600 frames. To that end, we first annotate the 2D joint locations on RGB images from two viewpoints and obtain the 3D joint positions by triangulation. We calculate the average 3D joint errors across all frames to measure the accuracy.

	Method	Input	MEPE (mm) ↓	AUC ↑	MEPE-RA (mm) ↓	AUC-RA ↑	fps ↑
(a)	HaMeR* [105]	Single	-	-	20.88	0.598	118
(b)	A2J-Transformer [46]	Single	51.68	0.474	20.76	0.603	34
(c)	HTT [15]	Single	49.09	0.489	20.69	0.599	**211**
(d)	TherFormer-S	Single	**44.64**	**0.539**	**18.34**	**0.643**	136
(e)	(c) w/o spatial transformer	Single	48.79	0.491	20.15	0.609	174
(f)	(c) w/o mask guidance	Single	48.83	0.494	18.89	0.625	141
(g)	HTT [15]	Sequence	47.07	0.512	17.49	0.659	**129**
(h)	TherFormer-V	Sequence	**43.36**	**0.549**	**17.36**	**0.661**	52

Table 3: Comparison between TherFormer and state of the arts on thermal camera-based 3D hand pose estimation. The fps indicates the number of inference steps that models can run per second. HaMeR* estimates 3D hand pose in a root-aligned manner by default; thus, we can only evaluate MEPE-RA and AUC-RA metrics.

As shown in Tab. 2, our annotation method achieves an average joint error of nearly 1cm, comparable to the results of [37, 41, 43]. 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. 6. 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 subjects' hand color and shape. Please see our **project page** for more visualization of annotation.

6.2 Experiment 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. 1) automatically following Sec. 4, of which the training and validation sets serve as the foundation for the training of all network models. Our automatic annotation pipeline (*cf*. Sec. 4) becomes infeasible under challenging scenarios, *i.e.*, gloves, darkness and sun glare, since hands appear corrupted in either RGB or depth images. To facilitate the quantitative evaluation under challenging settings, we manually annotate the ground truth for a few sequences collected in the glove and sun glare scenarios (*cf*. Tab. 1). Specifically, we first annotate the 2D keypoints from two viewpoints and then use triangulation to obtain their 3D positions.
Method and Implementation. To provide sufficient baselines for follow-up works, we selected three state-of-the-art methods in 3D hand pose estimation: HTT [15], A2J-Transformer [46] and HaMeR [105], and reproduced them on our dataset for experiments. HTT [15] is a video-based method thus enabling the evaluation in both two problem settings while A2J-Transformer [46] only works under the single image-based setting. HaMeR [105] is a model-based method that takes a single image as input and reconstructs a MANO-based hand mesh. It estimates each hand individually from a cropped image patch centered on the hand, removing the need for explicit localization. For comparison, we extract 3D keypoints from its predicted mesh as the estimated pose. We use the same sequence length, *i.e.*, $T = 8$, for HTT and TherFormer-V baselines. We exclude the additional action block used by HTT [15], focusing solely on hand pose estimation. We adjusted the anchor initialization phase of A2J-Transformer [46] to better accommodate our dataset, without altering the density of its anchors. All trained models are tested on a single NVIDIA RTX 4090 GPU to fairly compare their inference speed.
Evaluation Metrics. Following HTT [15], we evaluate the accuracy of 3D hand pose estimation with two metrics: *Percentage of Correct Keypoints* (PCK) and *Mean End-Point Error* (MEPE) [34], in both camera space and root-aligned (RA) space. For the RA space, we align the estimated wrist with its groundtruth position before measuring the joint errors. In terms of PCK, we report the corresponding *Area Under the Curve* (AUC) over the 0-80mm/50mm error thresholds for the camera/RA space.

6.3 Thermal Image-based Results

Main Results. To assess TherFormer's performance in thermal image-based 3D hand pose estimation, we compare it against state-of-the-art methods [15, 46] on the main testing set, as shown in Tab. 3. TherFormer-S outforms three competing methods under the single image-based setting, while TherFormer-V surpasses the counterpart HTT [15] given the same sequential images as input. Such an improvement mainly stems from our mask-guided spatial attention design that can adaptively encode the spatial interaction among hand joints with the guidance of the hand mask (*cf*. Sec. 5). A performance gap can be observed between single image-based and video-based settings for both HTT [15] and TherFormer. We credit this to their usage of temporal information that helps to tackle the occlusion cases and solve ambiguities. Thanks to our lightweight network design, TherFormer is highly efficient to run with fps of 136 and 52 for single and sequence input respectively, ensuring its real-time application to resource-constrained devices. Moreover, TherFormer-V also improves the performance over HTT [15] for other spectra (*cf*. Tab. 5), proving its adaptability to different inputs.
Ablation Studies. We ablate our proposed mask-guided attention mechanism and the entire spatial transformer to observe their impact. As can be seen in Tab. 3, the mask-guided attention mechanism notably contributes to TherFormer's performance growth (row (c) vs. (e)), demonstrating its effectiveness in mitigating the effect of background clutter. Other components in the spatial transformer bring a large performance gain in the RA space (row (e) vs.

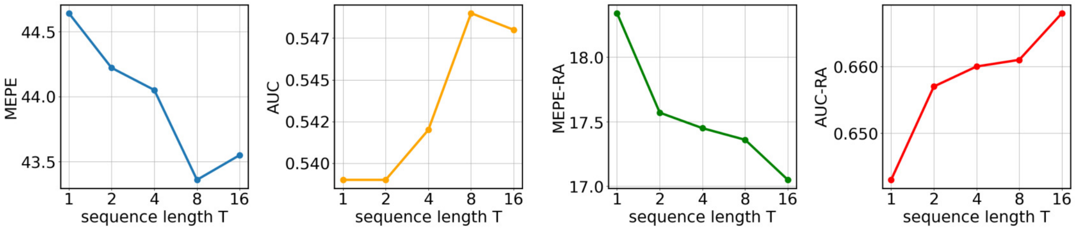

Figure 7: Impact of the temporal sequence length T to TherFormer. The plots show all four metrics MEPE (mm) ↓, AUC ↑, MEPE-RA (mm) ↓, and AUC-RA ↑ against 5 sequence length settings, *i.e.*, $T \in$ {1, 2, 4, 8, 16}.

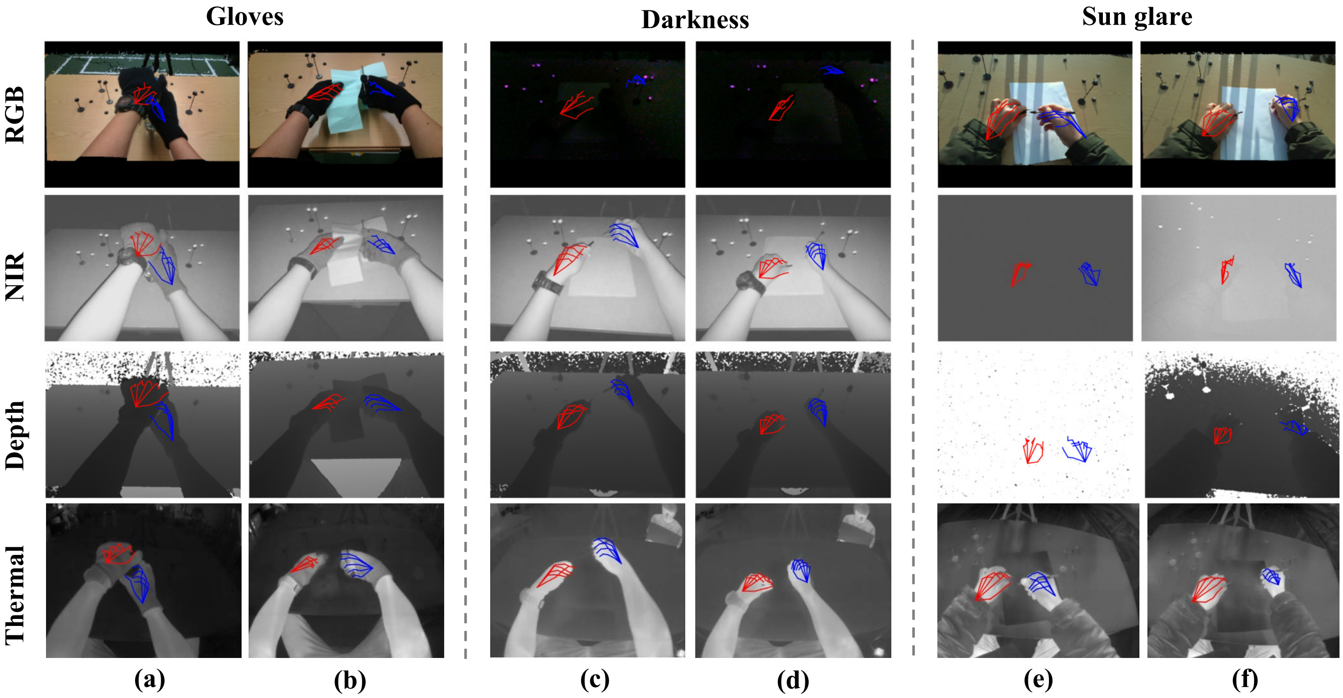

Figure 8: Examples of results with various spectra under three challenging settings. For visualization, we show the projection of the estimated left (red) and right (blue) 3D hand joints on images.

	TherFormer-V (glove)		TherFormer-V (sun glare)	
Spectrum	MEPE-RA (mm) ↓	AUC ↑	MEPE-RA (mm) ↓	AUC ↑
RGB	51.94	0.141	38.24	0.252
Depth	45.96	0.206	42.27	0.254
NIR	39.83	0.282	90.84	0.093
Thermal	**39.23**	**0.302**	**32.56**	**0.363**

Table 4: Comparison between different spectra under two challenging conditions, *i.e.*, glove and sun glare.

(d)), confirming the importance of spatial feature enhancement for fine-grained hand joints localization.

Impact of Sequence Length. Here we conduct experiments to analyze the impact of the sequence length T on the performance of our baseline method. We present the variation in our performance with respect to sequence length in Fig. 7. All experiments are conducted using the same hyperparameters, except for the $T = 16$ model, which uses a smaller batch size due to GPU memory constraints. It can be observed that performance generally improves with increased sequence length, particularly in root-aligned metrics. There is a slight performance decrease in the camera space when the sequence length T is increased to 16 but the performance in the root-aligned space is improved. The general trend underscores the importance of temporal information in enhancing the accuracy of 3D hand pose estimation.

6.4 Comparison between Spectrum

6.4.1 Qualitative Results under Challenging Conditions. To justify the advantages of using thermal images for egocentric hand pose estimation under challenging scenarios, we conduct a comparison

	HTT (Sequence)		TherFormer-V		TherFormer-S		Best	
Spectrum	MEPE (mm) ↓	AUC ↑	MEPE (mm) ↓	AUC ↑	MEPE (mm) ↓	AUC ↑	MEPE (mm) ↓	AUC ↑
RGB	43.30	0.542	43.50	0.542	44.61	0.529	43.30	0.542
Depth	41.62	0.559	**39.70**	**0.581**	**39.84**	**0.579**	**39.70**	**0.581**
NIR	**41.57**	**0.562**	40.79	0.575	40.98	0.575	40.79	0.575
Thermal	47.07	0.512	43.36	0.549	44.64	0.539	43.36	0.549

Table 5: Comparison between different spectra under normal conditions. Models are trained with their corresponding spectrum images from the training set. We test them on the main testing set.

between four spectra on sequences collected in challenging scenarios (*cf*. Tab. 1). To intuitively show the results, we first conduct a qualitative analysis of their performance and present some representative examples in Fig. 8. We use the same model, *i.e.*, HTT [15], trained on different spectra for a fair comparison. Please refer to our **project page** for more figure examples and demo videos under challenging conditions.

RGB vs. Thermal. As can be seen, RGB-based methods fail with gloves wearing (*cf*. Fig. 8 (a-b)) and in darkness (*cf*. Fig. 8 (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. Contrarily, as shown in Fig. 8 (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.

NIR, Depth vs. Thermal. NIR sensors are significantly disrupted by strong sunlight, affecting both NIR imaging and depth map creation, as shown in Fig. 8 (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.

6.4.2 Quantitative Results under Challenging Conditions. To provide the numerical results, we evaluate our TherFormer-V models trained for different spectra on our manually annotated sequences (*cf*. Sec. 6.2) and calculated the quantitative results presented in Tab. 4. As can be seen, thermal imaging-based approaches demonstrate the best performance among different spectra in challenging settings, underscoring thermal imagery's advantages in difficult lighting conditions and when hands are occluded. Notably, the conclusion drawn from Tab. 4 resonate with the visualization in Fig. 8. The RGB spectrum performs poorly in scenarios involving gloves, where the color and texture of human hands are significantly altered. Intense sunlight introduces severe artifacts in NIR images, markedly impairing performance under sun glare.

6.4.3 Quantitative Results under Normal Conditions. To validate the versatility of thermal-based methods, we also evaluate their performance compared to other spectra (*i.e.*, RGB, depth, NIR and thermal) under normal conditions for general-purpose use cases. We apply different spectral images as input to multiple baselines methods and report the results in Tab. 5. Not surprisingly, depth image-based methods yield the best performance on average, as they directly utilize the depth information to provide a detailed 3D structure of the hand. The NIR spectrum shows a marginal decrease in performance compared to depth but outperforms both RGB and thermal spectra. This can be attributed to NIR camera's active sensing ability, which leads to more consistent and reliable imaging regardless of the variability of external illuminations (vs. RGB) and temperature (vs. thermal). However, active NIR sensors are prone to interference from external NIR sources like sun glare (*cf*. Fig. 8). Thermal-based methods, despite using single-channel heat capture, achieves comparable performance to those using RGB images that contain rich color and textures. Such results suggest that thermal images can not only serve as supplements in challenging cases but also can be a viable alternative to other spectra in normal conditions for hand pose estimation.

7 Limitation and Future Work

As the first exploration in this field, this work has certain limitations that can guide our future work. First, our dataset primarily focuses on controlled indoor environment and challenging scenarios for RGB/NIR spectrum. Data collection in other challenging scenarios, especially where thermal imaging can be beneficial or limited, is currently scarce. In future works, we plan to expand our dataset beyond the current settings to incorporate a wider range of real-world challenges, including grasping hots objects, reflective surfaces to thermal radiation, ambient heating objects, and more outdoor environments. Paired data should be collected by including and removing these challenging factors to quantify their impact to thermal imaging-based 3D hand pose estimation. Second, we only annotate 3D hand pose for our dataset, which limits its usage of evaluation to other tasks relevant to human hands. Further efforts could be annotating the fine-grained hand action splits and hand-object contact [42].

8 Conclusion

This paper introduces ThermoHands, the first benchmark for egocentric 3D hand pose estimation using thermal images. ThermoHands features a multi-spectral, multi-view dataset with automatically annotated 3D hand poses and a novel baseline method, TherFormer, utilizing dual transformer modules for encoding spatio-temporal relationships. We demonstrate near 1cm annotation accuracy, show that TherFormer surpasses existing methods in thermal-based 3D hand pose estimation, and confirm thermal images' effectiveness in challenging lighting and obstruction scenarios. We believe our foundational endeavour could set the stage for further research in thermal-based 3D hand pose estimation and its wide application.

Acknowledgments

This research is partially supported by the Engineering and Physical Sciences Research Council (EPSRC) under the Centre for Doctoral Training in Robotics and Autonomous Systems at the Edinburgh Centre of Robotics (EP/S023208/1) and the grants from the Cisco Research.

Appendix

The appendix is organized as follows:

- Appendix A illustrates more details about our hand pose dataset, in the aspects of spectrum coverage, interference avoidance, and dataset distribution.
- Appendix B presents more implementation details of our automatic 3D hand pose annotation method.
- Appendix C introduce more architecture design, parameter and loss details of TherFormer.

We also provide more visualization in the format of both images and videos for our data, annotation and results on our **project page**.

A Multi-spectral Hand Pose Dataset

Spectrum Coverage and Interference Avoidance. The specification of all sensor frames we collect are shown in Tab. 1. Note that we keep the faces of subjects outside of the FoV of the exocentric camera to avoid any personally identifiable information. The multi-spectral data encompasses imaging from the visible spectrum (400-700nm), near-infrared (NIR) spectrum (850 nm ± 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 [20] 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 [21] necessitates structured lighting employing NIR at a wavelength identical to the IR emitter on the L515 LiDAR camera [20]. 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. In the office environment, random heat sources, *e.g.*, servers and chargers, are strategically placed in the background to increase realism and introduce challenging factors into thermal images.

Dataset Distribution. We show the data distribution of our main part over hand actions in Fig. 9. 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 project page for more visualization of our collected data.

B Dataset Annotation

Acquisition of 2D Keypoint and Mask. We utilize the open-source MediaPipe Hands pipeline [100] 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) [84] 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` [106].

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.

Joint Angle Limitation. Following [39, 41, 43], 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 [43].

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:

$$\mathcal{L}_{shape} = \sum_{i=1}^{10} ||\beta_i|| \tag{7}$$

MANO Fitting. We utilize the Pytorch version of the MANO layer proposed in [102]. The Adam optimizer [107] 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.

Usage Limitations. Thanks to the well-established RGB-based tools like [100, 101], we can extract accurate 2D hand information, which, combined with depth images, enable MANO-based

		Resolution			Fov			FPS
Sensor frames	Sensor Type	Range	Horizontal	Vertical	Range	Horizontal	Vertical	
RGB (ego)	Intel RS L515 [20]	-	1280	720	-	69	42	30
NIR (ego)	Intel RS L515 [20]	-	640	480	-	70	55	30
Depth (ego)	Intel RS L515 [20]	< 5mm @ 1m	640	480	0.25m to 9m	70	55	30
Thermal (ego)	FLIR Boson 640 [96]	≤ 60 mK	640	512	-	95	-	8.5
RGB (exo)	Intel RS D455 [21]	-	1280	720	-	87	58	30
Depth (exo)	Intel RS D455 [21]	< 2% at 4m	848	480	0.6m to 6m	87	58	30

Table 6: The specification of sensor frames captured in data collection.

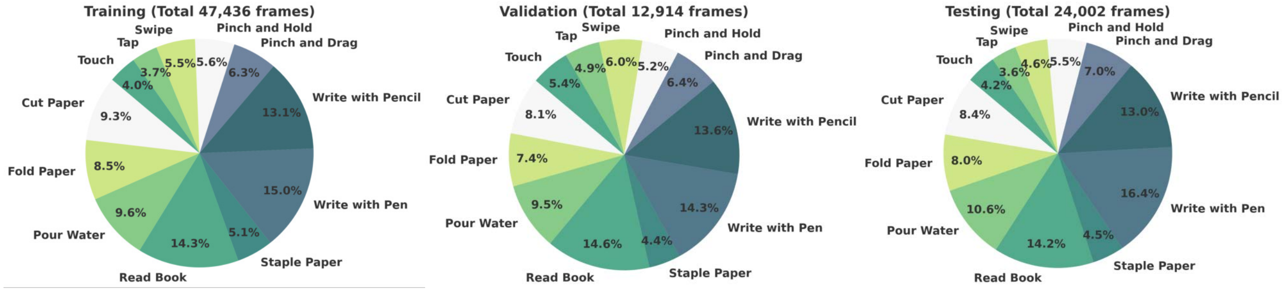

Figure 9: Distribution of data from the main part over hand actions.

3D hand pose annotation. However, this approach relies on adequate illumination and minimal occlusion, as hands may appear degraded in RGB or depth images under challenging conditions such as gloves, darkness, and sun glare, and RGB-based tools lack robustness against these factors. Moreover, while our annotation pipeline provides accurate 3D hand pose labels, it is impractical for real-time egocentric estimation due to its reliance on multi-view input, high computational cost, and iterative MANO optimization. In contrast, 3D hand pose estimation methods are designed for real-time inference from a single image view, more suitable for practical deployment.

C Baseline Method

Backbone. The input to our backbone is consecutive thermal images $\mathcal{S} = \{I_i\}_{i=1}^{T}$ resized to 320×256 after data loading, where $T = 1$ for the single image setting and $T > 1$ for the video setting. We initialize the ResNet-18 [108] network with the ImageNet-1K [109] pre-trained weights for the multi-scale features extraction. The ResNet-18 [108] backbone returns feature at five levels, with feature sizes as follows given the thermal images as input:

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

$$Features([64, 160, 128], [64, 80, 64], [128, 40, 32],$$

$$[256, 20, 16], [512, 10, 8])$$

We chose the third level features for future processing considering they carry fine-grained spatial details, which is crucial for accurate mask generation and spatial reasoning in our spatial transformer.

Mask-guided Spatial Transformer Module. 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 mid-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 mid-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 [103] is used and the number of sampling points per attention head K is 4:

$$\text{DA}(z_q, p_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk} \cdot W'_m \left(x(p_q + \Delta p_{mqk}) \right) \right] \quad (8)$$

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 [103], 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].

Temporal Transformer Module. This module is crucial for understanding temporal patterns over consecutive frames, utilizing a transformer-based approach. Given per-frame downsample feature map from the spatial transformer module, we flattened them and apply a linear layer to project this spatial embedding to [1,512]. Per-frame spatial feature vectors are then fed to the temporal self-attention, which attends to the feature vector of every frame. The number of attention layers is 2 and eight heads are used for each attention. The resulting spatio-temporal embedding has the same dimension as the input spatial features, which is $[T,512]$.

Pose Regression. In the pose head, we use the MLP to project the spatio-temporal embedding to the output space and obtain the per-frame 3D joint $\mathcal{J}_i$ for each frame individually.

$$MLP(512 \rightarrow 512 \rightarrow 512 \rightarrow 42 \times 3)$$

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:

$$L_{Hand} = \|P^{2D} - P^{2D}_{gt}\|_1 + \lambda_1 \|P^{depth} - P^{depth}_{gt}\|_1 + \lambda_2 \|P^{3D} - P^{3D}_{gt}\|_1 \quad (9)$$

where P^{3D}, P^{3D}_{gt},P^{2D}, P^{2D}_{gt} are the 3D/2D projection of estimated hand joint positions and ground truth hand joint positions. P^{depth} and P^{depth}_{gt} are the corresponding depth values for the hand joints. λ_1 and λ_2 are the weight parameters set as 100 in our implementation. The hand mask binary cross entropy loss is formulated as follows:

$$\begin{aligned} L_{Mask} = -\frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} [\, w_{pos} \cdot M_{wh} \cdot \log(\hat{M}_{wh}) \\ + w_{neg} \cdot (1 - M_{wh}) \cdot \log(1 - \hat{M}_{wh}) \,] \end{aligned} \quad (10)$$

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.

Training. We apply the Adam optimizer for our baseline training with an initial learning rate of 1e-4, decayed by a coefficient of 0.95 every epoch. All experiments are conducted using two RTX 3090 GPUs, with a batch size of 16 per GPU. Models are trained for 25 epochs.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

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

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

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

[23] 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.

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

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *PAMI*, 43(1):172–186, 2021.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

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

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

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

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

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

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

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

[82] 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.

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

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

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

[90] 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.

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

[92] 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.

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

[94] 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.

[95] 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.

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

[97] 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.

[98] 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.

[99] 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.

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

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

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

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

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

[109] 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.