

VIHand: Enhancing 3D Hand Pose Estimation with Visual-Inertial Benchmark

Xinyi Wang
Shanghai Jiao Tong University
Shanghai, China
shirley_w0118@sjtu.edu.cn

Pengfei Ren*
Beijing University of Posts and
Telecommunications
Beijing, China
rpf@bupt.edu.cn

Haoyang Zhang*
Defense Innovation Institute,
Academy of Military Sciences
Beijing, China
haoyang@tju.edu.cn

Xin Sheng
Tianjin University
Tianjin, China
shengxin0216@tju.edu.cn

Da Li
Nankai University
Tianjin, China
2320230830@mail.nankai.edu.cn

Liang Xie
Defense Innovation Institute,
Academy of Military Sciences
Beijing, China
xielnudt@gmail.com

Yue Gao
MoE Key Lab of Artificial Intelligence,
AI Institute, Shanghai Jiao Tong
University
Shanghai, China
yuegao@sjtu.edu.cn

Erwei Yin
1.Shanghai Jiao Tong University
2.Defense Innovation Institute,
Academy of Military Sciences
Shanghai, China
yinerwei1985@gmail.com

Abstract

Accurate and robust 3D hand pose estimation (HPE) plays a crucial role in human-computer interaction. Existing 3D HPE solutions predominantly rely on vision-based or inertial measurement units (IMUs)-based methods. Vision-based methods benefit from rich appearance information for high-accuracy HPE but are sensitive to field of view (FoV), occlusion, motion blur and lighting. IMU-based methods can operate immune to optical sensitivity and FoV constraints but remain vulnerable to cumulative integration errors and drift. Given their complementary strengths, combining dual modalities offers a promising direction for HPE in complex environments. However, the lack of large-scale visual-inertial datasets has limited progress in this area. In this paper, we construct VIHand, the first large-scale glove-worn dataset for visual-inertial HPE, comprising over 1.4 million synchronized RGB-D and IMU frames from 15 subjects. It enables comprehensive research in HPE tasks, such as multimodal fusion and cross-modal knowledge transfer. Building on VIHand, we propose visual-inertial fusion network (VIFNet) for dual-modalities estimation, and its distilled student model (VIFNet-S) for IMU-only evaluation. Experimental results reveal that integrating visual and inertial modalities significantly improves the accuracy and robustness of 3D HPE, particularly under occlusion and motion

blur. In IMU-only inference even sparse IMU configurations, models distilled from visual-inertial supervision achieve substantial performance gains, enabling robust HPE for challenging optical sensitive scenarios. Our dataset and supplementary materials are available on the project website: <https://shirley0118.github.io/VIHand>.

CCS Concepts

• **Computing methodologies** → **Motion capture**; • **Activity recognition and understanding**;

Keywords

Multi-modality, Hand Pose Estimation, IMU, Knowledge Distillation

ACM Reference Format:

Xinyi Wang, Pengfei Ren, Haoyang Zhang, Xin Sheng, Da Li, Liang Xie, Yue Gao, and Erwei Yin. 2025. VIHand: Enhancing 3D Hand Pose Estimation with Visual-Inertial Benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3746027.3758215>

1 Introduction

3D Hand Pose Estimation (HPE) technology holds significant potential for a wide range of applications (HCI), including augmented reality (AR) [41], virtual reality (VR) [48], and human-computer interaction (HCI) [47]. For successful deployment in these domains, 3D HPE approaches must achieve precise estimation, real-time performance, and exhibit robustness under diverse environments.

Existing 3D HPE solutions predominantly rely on vision-based or inertial measurement units (IMUs)-based approaches. Vision-based approaches estimate hand shape and pose from monocular or multiple RGB [11, 20, 33, 36, 37, 44], depth [14–16, 39] or RGB-D [34, 45, 58] cameras. Although accuracy has improved significantly

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
<https://doi.org/10.1145/3746027.3758215>

in recent years, they remain limited by field-of-view (FoV) constraints and are susceptible to occlusion, motion blur, and varying illumination. In contrast, by directly capturing hand motion signals, IMU-based approaches can operate immune to optical sensitivity and FoV constraints [1, 4, 19, 27]. However, they remain vulnerable to cumulative integration errors, magnetic interference, and significant drift. Overall, vision-based approaches encounter privacy concerns and struggle with the range of hand motions. IMU-based techniques fall short in achieving high accurate hand reconstruction. These limitations compromise the robustness and generalizability of mono-modal methods.

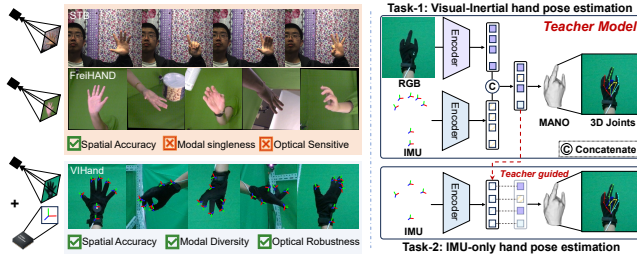


Figure 1: An overview of the VIFNet benchmark.

In fact, visual and IMU modalities are highly complementary. The former captures rich appearance details, such as texture, contour, shape, etc. IMU provides 3D dynamic hand motion information, including joint rotation, acceleration, and velocity variations. By fusing dual modalities, vision compensates for the lack of semantic richness of IMU, while IMU addresses the optical limitations of vision, thereby enhancing the robustness and generalizability of HPE. However, existing large-scale HPE datasets primarily concentrate on visual modality, while no public visual-inertial datasets for gesture interaction tasks are available, which limits the development of this field.

In this paper, we propose VIFNet, a large-scale visual-inertial dataset with accurate 3D hand joint coordinates and MANO parameters annotations, as illustrated in Fig. 1 (left). To the best of our knowledge, it is the first extensive worn-glove visual-inertial dataset, which contains over 1.4 million multimodal frames from 15 subjects, comprising a wide range of complex gestures. VIFNet aims to fill a critical gap in current research on multimodal HPE. With its rich time-synchronized visual-inertial data and high-quality annotations, VIFNet enables comprehensive research in HPE tasks, including multimodal fusion, cross-modal knowledge transfer and cross-modal data generation, etc. It provides a solid foundation for advancing modality compensation and uni-modal enhancement in real-world applications.

Building on VIFNet, we construct two benchmarks to explore the potential of visual-inertial integration for enhancing 3D hand pose estimation, as shown in Fig.1 (right). We first introduce VIFNet, a fusion framework that integrates appearance features from RGB with 3D kinematics features from IMU signals. This multimodal strategy significantly improves estimation performance under challenging occlusion, motion blur, and lighting variation. To demonstrate the effectiveness of cross-modal knowledge distillation, we propose a student framework, VIFNet-S for IMU-only HPE. Guided by the

VIFNet through semantic supervision, VIFNet-S achieves high estimation accuracy only with IMU, offering a practical and lightweight solution for wearable IMU-based HPE in real-world applications. In summary, our contributions are summarized as follows:

- We construct VIFNet, the first large-scale visual-inertial dataset for 3D HPE, containing over 1.4 million RGB-D and IMU frames with accurate 3D joint and MANO annotations. It enables multimodal fusion, and cross-modal knowledge distillation for enhancing HPE in multiple scenarios.
- We propose benchmarks for multimodal fusion and cross-modal knowledge distillation on VIFNet. VIFNet fuses RGB and IMU data to enhance hand pose estimation. VIFNet-S distills knowledge from VIFNet into an IMU-only student model, achieving high accuracy even with sparse IMU.
- We benchmark state-of-the-art models on the VIFNet dataset, highlighting the challenges posed by complex self-occlusion scenarios to current methodologies. Experimental results demonstrate the effectiveness of multimodal fusion and cross-modal knowledge distillation in enhancing HPE.

2 Related Works

2.1 Visual-based Methods

Visual-based methods aim to estimate hand pose from appearance or geometric information. These methods are commonly categorized into RGB-based methods [11, 20, 36, 37, 44] and depth-based methods [14–16, 39]. This paper focuses on RGB-based methods. Early works [26, 44, 53, 55, 56] predominantly adopt a skeleton regression paradigm, directly predicting 3D joint locations from single RGB images. More recent efforts have shifted towards MANO model-based [2, 28, 29, 36, 37] and mesh-based approaches [6–8]. Despite strong performance on existing datasets, visual-based methods still face limitations in real-world scenarios: (1) the restrictive FoV; and (2) appearance-based models are highly sensitive to occlusion, lighting, motion blur, and background clutter. These challenges limit their generalization and robustness in natural human interaction and dynamic environments.

2.2 IMU-based Methods

As an alternative to vision-based approaches, IMU-based methods have gradually gained attention due to their inherent portability and robustness to visual constraints [10]. Early IMU-based studies primarily focused on gesture recognition [9, 21, 60] and motion trajectory tracking tasks [24, 25]. Subsequently, some researchers explored dense configurations (10–18 IMUs) for 3D HPE [1, 4, 19, 27], achieving strong robustness under occlusion in VR/AR and rehabilitation settings. Although dense IMU configurations can provide more accurate estimation, they are typically accompanied by issues such as inconvenient wearing and bulky hardware, which hinder their practicality in daily use. To improve portability, recent studies have explored sparse IMU on full body [35, 59, 62] and hand pose estimation [43, 46, 51, 63]. For instance, Shen et al. [43] reported a 13.3 cm wrist position error using single wrist-mounted IMU, while LimbMotion [63] further reduced the error to 8.9 cm by introducing acoustic ranging. Although these approaches demonstrate the feasibility of 3D HPE with sparse setups, IMU-based methods still face key limitations: (1) they often suffer from high prediction errors;

Table 1: Comparison of existing 3D HPE datasets and VIHand. For the RGB-D datasets, we report their RGB resolution. To count the number of images, we consider images from multi-view at the same time step as different ones.

Dataset	Accessible	Data Modality			Metadata				Annotations		
		RGB	Depth	IMU	Type	Resolution	Subjects	Frames	Method	Joint Coord	MANO
ICVL [48]	✓	✗	✓	✗	real	320×240	10	18K	track	✓	✗
NYU [49]	✓	✗	✓	✗	real	640×480	2	243K	track	✓	✗
STB [61]	✓	✓	✓	✗	real	640×480	1	36K	manual	✓	✗
RHD [64]	✓	✓	✓	✗	syn	320×320	20	44K	synthetic	✓	✗
HO-3D [18]	✓	✓	✗	✗	real	640×480	10	78K	automatic	✓	✗
FreiHAND [65]	✓	✓	✗	✗	real	224×224	32	134K	semi-auto.	✓	✓
InterHand2.6M [32]	✓	✓	✗	✗	real	512×334	27	2.6M	semi-auto.	✓	✓
RGB2Hands [50]	✓	✓	✓	✗	real	640×480	2	108K	manual	✓	✓
DexYCB [5]	✓	✓	✓	✗	real	640×480	10	582K	manual	✓	✓
H2O [23]	✓	✓	✓	✗	real	1280×720	4	571K	automatic	✓	✓
OakInk [54]	✓	✓	✓	✗	real	848×480	12	230K	manual	✓	✓
Re:InterHand [31]	✓	✓	✗	✗	real	1024×668	10	1.5M	semi-auto.	✓	✓
Kortier et al. [22]	✗	✗	✗	✓	real	N/A	N/A	N/A	marker	✓	✗
Yang et al. [57]	✗	✗	✗	✓	real	N/A	N/A	N/A	marker	✓	✗
HANDMI4 [30]	✓	✗	✗	✓	real	N/A	20	N/A	marker	✓	✗
VIHand	✓	✓	✓	✓	real	1280×720	15	1.4M	semi-auto.	✓	✓

(2) most approaches estimate only wrist or fingertip positions, lacking precise full hand articulation; and (3) signal integration errors accumulate and are highly sensitive to magnetic interference [38].

2.3 3D Hand Pose Estimation Datasets

In 3D HPE, learning-based methods heavily rely on large-scale datasets that capture diverse real-world scenarios [12]. Most existing public datasets [3, 5, 13, 18, 23, 32, 48, 49, 52, 65] primarily focus on visual modalities. IMU-based methods have traditionally relied on handcrafted motion models and integration techniques such as Kalman filters, which reduce the need for large-scale annotated datasets, presenting a major bottleneck for advancing multimodal HPE research. Table 1 shows specification of existing 3D hand pose datasets and VIHand. Early datasets comprise depth-maps [48, 49] motivated by the Kinect device. For more practical applications, STB [61] introduced RGB images with simple hand poses, but the diversity and scale of the dataset is limited. RHD [64] addressed scalability by generating a fully synthetic dataset with broader pose coverage, yet domain gaps hinder real-world applicability. More recent datasets, including FreiHAND [65], HO-3D [18], InterHand2.6M [32], ContactPose [3], and DexYCB [5], provide multi-view images and high-precision 3D annotations, capturing realistic hand deformations and object interactions.

Despite these improvements, vision-based datasets remain limited by the camera’s FoV and intrinsic weaknesses of visual sensing, hindering robust modeling in complex, realistic scenarios. Moreover, those dataset solely focus on bare-hand, overlooking glove-worn conditions commonly encountered in practice. Our dataset aims to fill the gap. In this paper, we present VIHand, a large-scale, multi-modal, and high-fidelity dataset for 3D HPE. VIHand synchronously captures RGB, depth, and IMU data. Each sample includes accurate 3D hand joint coordinates and MANO parameters to support high-resolution pose modeling. It also covers both bare-hand and

glove-worn interactions, aligning more closely with real-world usage scenarios.

3 VIHand Dataset

3.1 Data Capture

Acquisition Setup. VIHand was collected using a multi-sensor platform (Fig. 2), consisting of five Intel RealSense D415i cameras and a data glove equipped with 7 IMU sensors. Specifically, 5 cameras were placed at diverse viewpoints around the hand to synchronously capture high-resolution RGB-D stream at 1280×720 resolution and 30 FPS. The multi-camera system was calibrated using a 3D calibration target [17], with a pixel-wise RMS reprojection error between 0.42 and 0.48 pixels. To enrich illumination and simulate diverse lighting conditions, four external lights were also deployed. 7 IMUs were embedded in the data glove at the wrist, back of the hand, and fingertips of all 5 fingers. Each IMU incorporates a 9-axis ICM-20948 chip that captures 3-axis acceleration, angular velocity, and magnetic field data at 30 FPS. We record these inertial data and real-time quaternion and orientation matrices calculated by an on-board sensor fusion algorithm.

Participants. We recruited 15 participants (8 males and 7 females) with diverse ages and hand shapes. All participants provided informed consent for the use of their data in both research and commercial applications. To protect privacy, the dataset was anonymous to remove personally identifiable information. Our data collection was approved by the local Institutional Review Board (IRB).

Acquisition Protocol. Each participant performed 15 gesture sequences, consisting of 14 predefined sequences and 1 random sequence. The predefined gestures span a broad spectrum of everyday hand movements, including common sign language expressions and varying degrees of finger flexion and extension. They also feature a diverse range of complex gestures, such as finger crossing, interlocking, wiggling, tapping, pinching, bending, snapping, and



Figure 2: Recording setup with 5 temporally synchronized RGB-D cameras and a dataglove intergrated 7 IMUs.

sliding. These performed gestures span a range of categories, including action-driven, pose-oriented, and composite types, exhibiting rich structural variation and semantic diversity. The random gesture sequence consists of approximately one minute of spontaneous and continuous hand motion, intended to simulate natural, unconstrained behavior. VIHand is constructed to capture a representative and naturally diverse distribution of hand poses, rather than optimizing for a specific task or gesture taxonomy. As illustrated in Fig. 3, our dataset exhibits significantly greater 3D pose diversity compared to FreiHAND [65], which is currently the most widely used single-hand dataset.

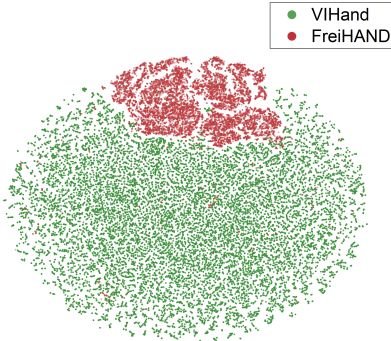


Figure 3: t-SNE visualization for hand pose distribution of VIHand and FreiHAND [65].

3.2 Data Annotation

VIHand follows the standard 21-joint annotation scheme [61], which includes the fingertip and three phalangeal joint centers for each finger, along with the wrist joint. However, annotating large-scale hand pose datasets is both time-consuming and labor-intensive. To balance annotation accuracy and scalability, we adopt a two-stage pipeline that combines automatic machine annotation with manual refinement. In the first stage, we use automatic machine annotation. Given the significant challenges posed by severe global occlusion and diminished visual contrast in our glove-worn datasets, state-of-the-art RGB-based machine annotation techniques face challenges in accurately annotating joint positions. Furthermore, traditional marker-based methods, not only disrupt natural motion but also

visually contaminate the hand’s surface, thereby substantially diminishing realism and adversely affecting the generalizability of downstream models. In contrast, we employed multi-view depth image-based self-supervised annotation [40], which effectively addresses issues related to visual texture ambiguity and artifacts induced by markers. Specifically, a self-supervised keypoint detector is applied to unlabeled multi-view sequences following the Cross-View Fusion Network (CVFN) strategy. Subsequently, 3D keypoints are reconstructed using RANSAC-based triangulation. This automated stage produces accurate and consistent 3D joint coordinates as well as MANO parameters for most frames. We tested this method on the held-out evaluation set, and the mean per joint position error is 3.86 mm. In the second stage, a team of professional annotators manually review and refine the machine-generated results. Low-confidence or erroneous predictions are re-annotated to ensure high-quality ground truth. The final annotations integrate high-confidence automatic estimates with manual corrections, achieving both precision and scalability. Based on this workflow, VIHand is annotated with 14,012,258 frames of synchronized RGB, depth, and IMU data, along with accurate 3D hand joint coordinates and MANO parameters. Visual examples are shown in Fig. 1 (left), with further details provided in the supplementary material.

4 Benchmark

In this section, to explore the potential of VIHand for advancing multitask HPE, we propose VIFNet and VIFNet-S, which demonstrate the enhancement of multimodal fusion and cross-modal knowledge distillation, respectively. The overall pipeline is depicted in Fig. 4.

4.1 VIFNet

To perform visual-inertial fusion and achieve high-accuracy 3D HPE, VIFNet utilizes the RGB and IMU data in parallel. It consists of three stages: visual-inertial feature extraction, multimodal feature fusion, and hand pose regression.

For the visual modality, a ResNet-18 backbone encodes the RGB image into initial feature maps $F_c \in \mathbb{R}^{H \times W \times C}$. These feature maps are refined by a spatial attention module, which re-weights F_c to produce refined features F_r , emphasizing hand-salient regions while attenuating background noise. The F_r is then structured into seven groups, $\{F_{rgb}^{(k)}\}_{k=1}^7$, allowing each group to specialize in different spatial aspects of the hand. Simultaneously, seven corresponding groups of IMU signals (3-axis acceleration and 9-axis orientation), $\{S_{imu}^{(k)}\}_{k=1}^7$, are processed. Each group $S_{imu}^{(k)} \in \mathbb{R}^{T \times D_{raw}}$, where T is the sequence length and D_{raw} is the raw feature dimension per sensor reading, is projected into a higher-dimensional embedding space $E_{imu}^{(k)} \in \mathbb{R}^{T \times D_{embed}}$ via a learned embedding layer. These embedded sequences are processed by a transformer encoder, which captures temporal dependencies and inter-sensor correlations. This produces seven robust group-specific IMU feature vectors, $\{G_{imu}^{(k)}\}_{k=1}^7$.

VIFNet employs a cross-modal attentive mechanism (CMAM) to integrate the visual and inertial features. The seven grouped visual features, $\{F_{rgb}^{(k)}\}_{k=1}^7$, and corresponding IMU features, $\{G_{imu}^{(k)}\}_{k=1}^7$, are processed by a CMAM module, as shown in Fig. 4. Then, the multi-head self-attention (MHSA) module captures complex intra-modal relationships within each modality. This allows the model

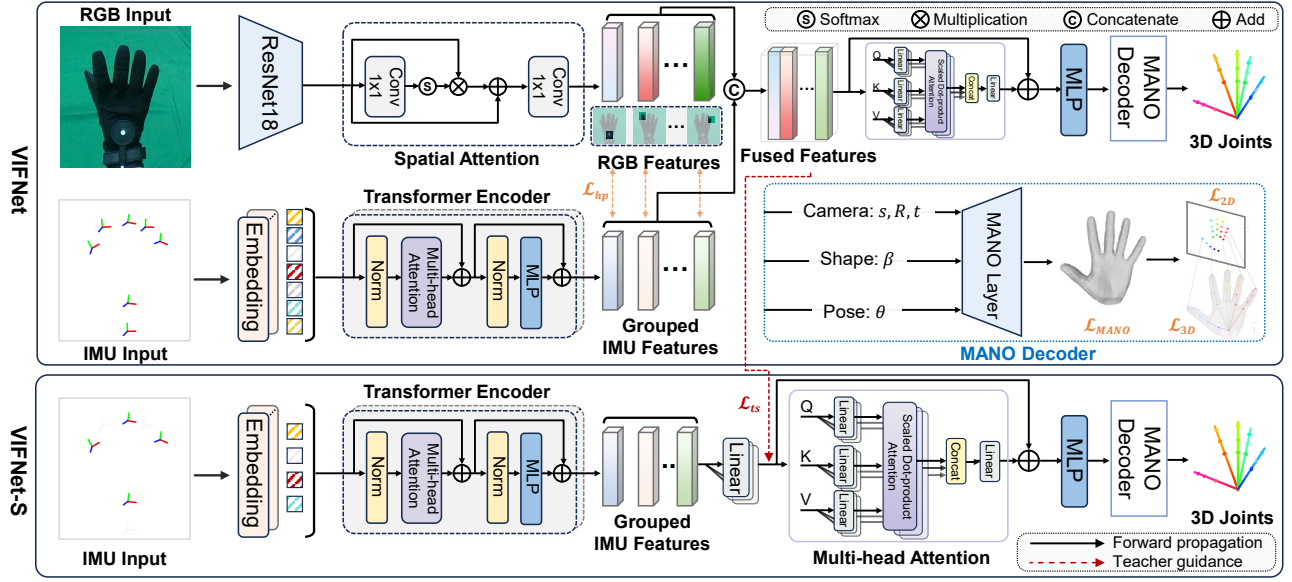


Figure 4: Overall framework. The fusion model (VIFNet) takes a monocular RGB and IMU data as input, extracts spatial and temporal features, and fuses them via cross attention mechanism. The fused features are then decoded by the MANO layer to estimate 3D hand pose. The student model (VIFNet-S), trained on IMU data alone, learns to mimic the VIFNet’s fused features representation through knowledge distillation.

to dynamically adjust the contribution of each modality and group, capturing complementary cues and addressing ambiguities inherent in individual modalities. The output is a unified set of fused features, F_{fused} , which encode both spatial and temporal characteristics of hand motion. These features are passed through a Multi-Layer Perceptron (MLP) to predict the MANO [42] parameters $\Psi = \{\theta, \beta, R, t, s\}$, which are then fed into the MANO Decoder to compute the 3D hand mesh vertices $V \in \mathbb{R}^{778 \times 3}$ and the joint locations $J \in \mathbb{R}^{21 \times 3}$. The VIFNet model is trained end-to-end by minimizing a composite loss function, $\mathcal{L}_{\text{VIFNet}}$, which aggregates several key losses for accurate 3D hand pose estimation:

$$\mathcal{L}_{\text{VIFNet}} = \lambda_{3D} \mathcal{L}_{3D} + \lambda_{2D} \mathcal{L}_{2D} + \lambda_{\text{MANO}} \mathcal{L}_{\text{MANO}} + \lambda_{\text{hp}} \mathcal{L}_{\text{hp}} \quad (1)$$

The details of $\mathcal{L}_{\text{VIFNet}}$ are provided in the supplementary material.

4.2 VIFNet-S

IMU-only hand pose estimation is appealing for real-world applications due to its portability. However, the low signal-noise-ratio and absence of appearance cues limit the accuracy. Leveraging VIFNet, which captures the synchronized appearance and motion data, we propose VIFNet-S using cross-modal knowledge distillation strategy. This approach exploits the strong semantic correspondence between hand appearance and motion patterns, enabling the transfer of rich representations learned from visual-inertial supervision to IMU-only models even with sparse IMU configurations.

We first align the intermediate representations of the student model (VIFNet-S) and teacher model (VIFNet). Specifically, VIFNet’s Fused Features (F_{fused}^T) are used as a supervisory signal to guide the student’s IMU-based features (F_{imu}^S). To align the student’s IMU-based features with the teacher’s fused representations, we

employ a contrastive distillation loss $\mathcal{L}_{\text{ts}}$. This loss encourages the student’s feature to be close to its corresponding teacher feature while remaining distant from others in the batch. This alignment helps the student capture richer, and more discriminative information. The VIFNet-S model is optimized using a composite loss function, denoted as $\mathcal{L}_{\text{VIFNet-S}}$, which synergistically integrates direct supervision and knowledge distillation:

$$\mathcal{L}_{\text{VIFNet-S}} = \alpha_{3D} \mathcal{L}_{3D} + \alpha_{2D} \mathcal{L}_{2D} + \alpha_{\text{MANO}} \mathcal{L}_{\text{MANO}} + \alpha_{\text{ts}} \mathcal{L}_{\text{ts}} \quad (2)$$

The details of $\mathcal{L}_{\text{VIFNet-S}}$ are provided in the supplementary material.

5 Experiments

5.1 Setup

Dataset split strategy. VIFNet is partitioned on a per-subject basis, where 12 out of 15 gesture sequences are assigned to the training set and the remaining 3 sequences are reserved for testing. This fixed split is applied consistently across all participants and used in all experiments to ensure fair comparison and reproducibility.

Implementation details. Experiments are implemented on the PyTorch with NVIDIA RTX 4090 GPU. To prepare the input data for training and inference, the RGB images are resized to 256×256 first and then center cropped into 224×224 , respectively. During training, the batch size is set to 32, and lasts for 80 epochs. The initial learning rate is set to 0.0001, and is decayed by a factor of 0.1 after 20 epochs. We train all networks using the Adam optimizer.

Evaluation metrics. We use mean per joint position error (MPJPE) and mean per vertex position error (MPVPE) as our evaluation metrics, which measure Euclidean distance (mm) between estimated coordinates and ground-truth. Before calculating the metrics, we align the translation of the root wrist joint.

5.2 Benchmark Evaluation

We benchmark VIHand with state-of-the-art IMU-based methods ([59], [35]) and Visual-based methods ([36], [37]). As shown in Table 2, IMU-based methods (TransPose [59] and FusePose [35]), relying on 7 IMUs, struggle with accurately reconstructing the 3D structure and fine-grained poses. Visual-based methods (HaMeR [36] and WiLoR [37]), despite leveraging rich image information, suffer from performance degradation due to joint self-occlusion. On the contrary, by effectively fusing RGB and IMU information, our VIFNet robustly estimates the 3D hand pose, achieves the lowest MPJPE and MPVPE results, respectively. Fig. 5 further validates that our VIFNet produces much better results than previous methods on VIHand, even under the complex self-occlusion gestures. In addition, we provide ablation study in the supplementary material.

Table 2: Benchmark of state-of-the-art methods with MPJPE (mm) and MPVPE (mm). † indicates ground-truth tested methods, * denotes additional training on VIHand.

Methods	RGB	IMU	#IMUs	MPJPE	MPVPE
TransPose* [59]	✗	✓	7	21.84	24.76
FusePose* [35]	✗	✓	7	18.56	22.32
HaMeR† [36]	✓	✗	-	11.98	13.67
WiLoR† [37]	✓	✗	-	10.49	11.74
VIFNet	✗	✓	7	16.93	19.41
VIFNet	✓	✗	-	8.73	9.85
VIFNet-S	✗	✓	7	13.69	16.53
VIFNet	✓	✓	7	7.86	8.94

5.3 Evaluation under IMU-only Settings

We further evaluate the cross-modal knowledge distillation performance of the VIHand under IMU-only inference, including 7-IMU, 3-IMU (wrist, thumb, middle finger), and 2-IMU (wrist, middle finger) setups. As shown in Table 3, when the number of sensors was cut from 7 to 3, even 2, methods not using knowledge distillation saw a significant drop in performance. This shows that traditional training methods struggle to keep up the original accuracy with limited IMU input. In contrast, our VIFNet-S, distilled from the teacher model, maintains strong performance under the same sparse setting. While achieving MPJPE of 13.69 mm with 7 IMUs, it still attains 18.26 mm using 3 IMUs and 20.72 mm only using 2 IMUs, significantly outperforming other methods. The results of MPVPE show a similar trend to those of MPJPE, which demonstrate that cross-modal distillation effectively transfers semantic representations, enabling accurate pose estimation even with sparse IMUs.

6 Conclusion

We propose VIHand, a large-scale visual-inertial dataset for 3D HPE, comprising over 1.4 million synchronized RGB-D and IMU frames from 15 subjects. It fills the gap of high-quality visual-inertial data and builds a bridge for deep semantic alignment and information

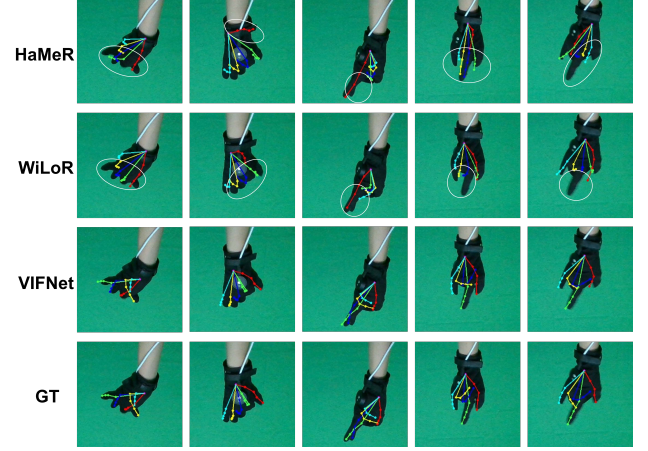


Figure 5: Qualitative results for HaMeR[36], WiLoR [37], VIFNet and ground-truth (GT) on VIHand from top to bottom.

Table 3: Benchmark of state-of-the-art methods using MPJPE (mm) and MPVPE (mm) under different IMU settings.

Methods	RGB	IMU	#IMUs	MPJPE	MPVPE
TransPose [59]	✗	✓	7	21.84	24.76
FusePose [35]	✗	✓	7	18.56	22.32
VIFNet	✗	✓	7	16.93	19.41
VIFNet-S	✗	✓	7	13.69	16.53
TransPose [59]	✗	✓	3	28.01	32.76
FusePose [35]	✗	✓	3	24.91	29.38
VIFNet	✗	✓	3	22.15	26.24
VIFNet-S	✗	✓	3	18.26	22.67
TransPose [59]	✗	✓	2	30.47	34.52
FusePose [35]	✗	✓	2	27.38	31.12
VIFNet	✗	✓	2	25.59	28.75
VIFNet-S	✗	✓	2	20.72	23.79

complementarity between dual modalities. To validate the application potential of VIHand in HPE, we propose VIFNet, a visual-inertial fusion network that achieves high-precision and robust 3D HPE through an efficient cross-modal interaction strategy. Additionally, to validate the potential of VIHand in cross-modal distillation, we introduce VIFNet-S, a cross-modal knowledge distillation method, which transfers the fusion feature of visual-inertial to an IMU-only student network, significantly enhancing its accuracy even with sparse IMU. Extensive experiments demonstrate VIHand advantages in multimodal fusion and cross-modal distillation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62332019, 62406039), the National Key R&D Program of China (Nos. 2023YFF1203900, 2023YFF1203903), the China Postdoctoral Science Foundation (Nos. 2023TQ0039, 2024M750257, GZC20230320), and the Beijing Nova Program (No. 20240484513).

References

- [1] Tommaso Lisini Baldi, Stefano Scheggi, Leonardo Meli, Mostafa Mohammadi, and Domenico Prattichizzo. 2017. GESTO: A glove for enhanced sensing and touching based on inertial and magnetic sensors for hand tracking and cutaneous feedback. *IEEE Transactions on Human-Machine Systems* 47, 6 (2017), 1066–1076.
- [2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 2019. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10843–10852.
- [3] Samarth Brahmabhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. 2020. ContactPose: A dataset of grasps with object contact and hand pose. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII* 16. Springer, 361–378.
- [4] Hsien-Ting Chang and Jen-Yuan Chang. 2019. Sensor glove based on novel inertial sensor fusion control algorithm for 3-D real-time hand gestures measurements. *IEEE Transactions on Industrial Electronics* 67, 1 (2019), 658–666.
- [5] 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. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9044–9053.
- [6] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. 2022. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20544–20554.
- [7] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. 2021. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13274–13283.
- [8] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. 2020. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII* 16. Springer, 769–787.
- [9] Agastasya Dahiya, Dhruv Wadhwa, Rohan Katti, and Luigi G Occhipinti. 2024. Efficient Hand Gesture Recognition using Artificial Intelligence and IMU based Wearable Device. *IEEE Sensors Letters* (2024).
- [10] Bin Fang, Fuchun Sun, Huaping Liu, and Di Guo. 2017. Development of a wearable device for motion capturing based on magnetic and inertial measurement units. *Scientific Programming* 2017, 1 (2017), 7594763.
- [11] Kun Gao, Xiaolong Liu, Pengfei Ren, Haoyu Chen, Tao Zhen, Liang Xie, Zhongkui Li, Ye Yan, Haoyang Zhang, and Erwei Yin. 2024. Progressively global-local fusion with explicit guidance for accurate and robust 3d hand pose reconstruction. *Knowledge-Based Systems* 304 (2024), 112532.
- [12] Kun Gao, Haoyang Zhang, Xiaolong Liu, Xinyi Wang, Liang Xie, Bowen Ji, Ye Yan, and Erwei Yin. 2024. Challenges and solutions for vision-based hand gesture interpretation: A review. *Computer vision and image understanding* Nov. (2024), 248.
- [13] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. 2018. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 409–419.
- [14] Lihao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 2016. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3593–3601.
- [15] Lihao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 2017. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1991–2000.
- [16] Lihao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 2019. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10833–10842.
- [17] Hyowon Ha, Michal Perdoch, Hatem Alismail, In So Kweon, and Yaser Sheikh. 2017. Deltile Grids for Geometric Camera Calibration. In *2017 IEEE International Conference on Computer Vision (ICCV)*.
- [18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2020. Homnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3196–3206.
- [19] Bingcheng Hu, Tian Ding, Yuxin Peng, Li Liu, and Xu Wen. 2020. Flexible and attachable inertial measurement unit (IMU)-based motion capture instrumentation for the characterization of hand kinematics: A pilot study. *Instrumentation Science & Technology* 49, 2 (2020), 125–145.
- [20] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. 2018. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European conference on computer vision (ECCV)*. 118–134.
- [21] Shuo Jiang, Bo Lv, Weichao Guo, Chao Zhang, Haitao Wang, Xinjun Sheng, and Peter B Shull. 2017. Feasibility of wrist-worn, real-time hand, and surface gesture recognition via sEMG and IMU sensing. *IEEE Transactions on Industrial Informatics* 14, 8 (2017), 3376–3385.
- [22] Henk G Kortier, Jacob Antonsson, H. Martin Schepers, Fredrik Gustafsson, and Peter H Veltink. 2015. Hand Pose Estimation by Fusion of Inertial and Magnetic Sensing Aided by a Permanent Magnet. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23, 5 (2015), 796–806.
- [23] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. 2021. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10138–10148.
- [24] Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee. 2021. Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Science Robotics* 6, 58 (2021), eabe1315.
- [25] Yongjun Lee, Myungsun Kim, Yongseok Lee, Junghan Kwon, Yong-Lae Park, and Dongjun Lee. 2018. Wearable finger tracking and cutaneous haptic interface with soft sensors for multi-fingered virtual manipulation. *IEEE/Asme Transactions on Mechatronics* 24, 1 (2018), 67–77.
- [26] Moran Li, Yuan Gao, and Nong Sang. 2021. Exploiting learnable joint groups for hand pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 1921–1929.
- [27] Bor-Shing Lin, I-Jung Lee, Shu-Yu Yang, Yi-Chiang Lo, Junghsi Lee, and Jean-Lon Chen. 2018. Design of an inertial-sensor-based data glove for hand function evaluation. *Sensors* 18, 5 (2018), 1545.
- [28] Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1954–1963.
- [29] Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12939–12948.
- [30] Francesca Mongelli and Matteo Menolotto. 2023. HANDM4 (HAND Motion capture data for Industry 4.0). doi:10.21227/c6t8-ge47
- [31] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. 2023. A dataset of relighted 3D interacting hands. *Advances in Neural Information Processing Systems* 36 (2023), 17689–17701.
- [32] Gyeongsik Moon, Shouo-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX* 16. Springer, 548–564.
- [33] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. Generated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 49–59.
- [34] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2017. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE international conference on computer vision*. 1154–1163.
- [35] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. 2023. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.
- [36] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9826–9836.
- [37] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2024. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259* (2024).
- [38] Adnan Rashid and Osman Hasan. 2019. Wearable technologies for hand joints monitoring for rehabilitation: A survey. *Microelectronics Journal* 88 (2019), 173–183.
- [39] Pengfei Ren, Haifeng Sun, Jiachang Hao, Qi Qi, Jingyu Wang, and Jianxin Liao. 2022. A dual-branch self-boosting framework for self-supervised 3D hand pose estimation. *IEEE Transactions on Image Processing* 31 (2022), 5052–5066.
- [40] Pengfei Ren, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, and Jianxin Liao. 2022. Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20555–20565.
- [41] Javier Romero, Hedvig Kjellström, and Danica Kragic. 2009. Monocular real-time 3D articulated hand pose estimation. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 87–92.
- [42] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).
- [43] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a smartwatch and i can track my user's arm. In *Proceedings of the 14th annual international conference on Mobile systems, applications, and services*. 85–96.

- [44] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. 2018. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 89–98.
- [45] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. 2013. Interactive markerless articulated hand motion tracking using RGB and depth data. In *Proceedings of the IEEE international conference on computer vision*. 2456–2463.
- [46] Paul Strelci, Rayan Armani, Yi Fei Cheng, and Christian Holz. 2023. HOOV: Hand out-of-view tracking for proprioceptive interaction using inertial sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [47] Mingkui Tan, Zhuangwei Zhuang, Sitao Chen, Rong Li, Kui Jia, Qicheng Wang, and Yuanqing Li. 2024. EPMF: Efficient perception-aware multi-sensor fusion for 3D semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [48] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. 2014. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3786–3793.
- [49] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. 2014. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)* 33, 5 (2014), 1–10.
- [50] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. 2020. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)* 39, 6 (2020), 1–16.
- [51] Wenchuan Wei, Keiko Kurita, Jilong Kuang, and Alex Gao. 2021. Real-time 3D arm motion tracking using the 6-axis IMU sensor of a smartwatch. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 1–4.
- [52] Zhenyu Wu, Duc Hoang, Shihyao Lin, Yusheng Xie, Liangjian Chen, Yenyu Lin, Zhangyang Wang, and Wei Fan. 2020. MM-Hand: 3D-Aware Multi-Modal Guided Hand Generation for 3D Hand Pose Synthesis. (2020).
- [53] Linlin Yang, Shicheng Chen, and Angela Yao. 2021. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11364–11373.
- [54] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. 2022. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20953–20962.
- [55] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. 2019. Aligning latent spaces for 3d hand pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2335–2343.
- [56] Linlin Yang and Angela Yao. 2019. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9877–9886.
- [57] Zhicheng Yang, Bert Jan F. Van Beijnum, Bin Li, Shenggang Yan, and Peter H. Veltink. 2020. Estimation of Relative Hand-Finger Orientation Using a Small IMU Configuration. *Sensors* 20, 14 (2020), 4008.
- [58] Yuan Yao and Yun Fu. 2012. Real-time hand pose estimation from RGB-D sensor. In *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 705–710.
- [59] Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)* 40, 4 (2021), 1–13.
- [60] Dian Zhang, Zexiong Liao, Wen Xie, Xiaofeng Wu, Haoran Xie, Jiang Xiao, and Landu Jiang. 2021. Fine-grained and real-time gesture recognition by using IMU sensors. *IEEE Transactions on Mobile Computing* (2021).
- [61] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 2016. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214* (2016).
- [62] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. 2024. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1889–1899.
- [63] Han Zhou, Yi Gao, Xinyi Song, Wenxin Liu, and Wei Dong. 2019. Limbmotion: Decimeter-level limb tracking for wearable-based human-computer interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.
- [64] Christian Zimmermann and Thomas Brox. 2017. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*. 4903–4911.
- [65] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. 2019. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 813–822.