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

XINYI WANG, Shanghai Jiao Tong University, China
PENGFEI REN*, Beijing University of Posts and Telecommunications, China
HAOYANG ZHANG*, Defense Innovation Institute, Academy of Military Sciences, China
XIN SHENG, Tianjin University, China
DA LI, Nankai University, China
LIANG XIE, Defense Innovation Institute, Academy of Military Sciences, China
YUE GAO, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
ERWEI YIN, 1.Shanghai Jiao Tong University 2.Defense Innovation Institute, Academy of Military Sciences, China

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 Supplementary Materials. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746027.3758215

A VIHand Dataset Construction Details

VIHand are captured under a wide range of complex pose, exhibiting complex inter-finger interaction such as finger crossing, interlocking, wiggling, tapping, pinching, bending, snapping, and sliding. Fig. 1 presents a diverse set of hand poses captured in the VIHand dataset. Each row showcases variations in hand pose, demonstrating the accuracy and consistency of our 3D keypoint annotation pipeline. These samples highlight the rich diversity in hand motion and appearance, which is critical for robust hand pose estimation under real-world conditions. In particular, the annotations remain accurate across self-occlusions and varying compose gestures, showcasing the reliability of our depth-based multi-view self supervision annotation framework.

To capture the dynamic inertial data of hand motion, 7 IMUs were embedded in the data glove at the wrist, palm, thumb, index, middle finger, ring finger and little finger. Fig. 2 visualizes both acceleration and orientation signals from the data glove during a continuous hand motion sequence. Each row corresponds to one IMU sensor, with the

*Corresponding author.

Authors' Contact Information: 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 and 2.Defense Innovation Institute, Academy of Military Sciences, Shanghai, China, yinerwei1985@gmail.com.

Manuscript submitted to ACM

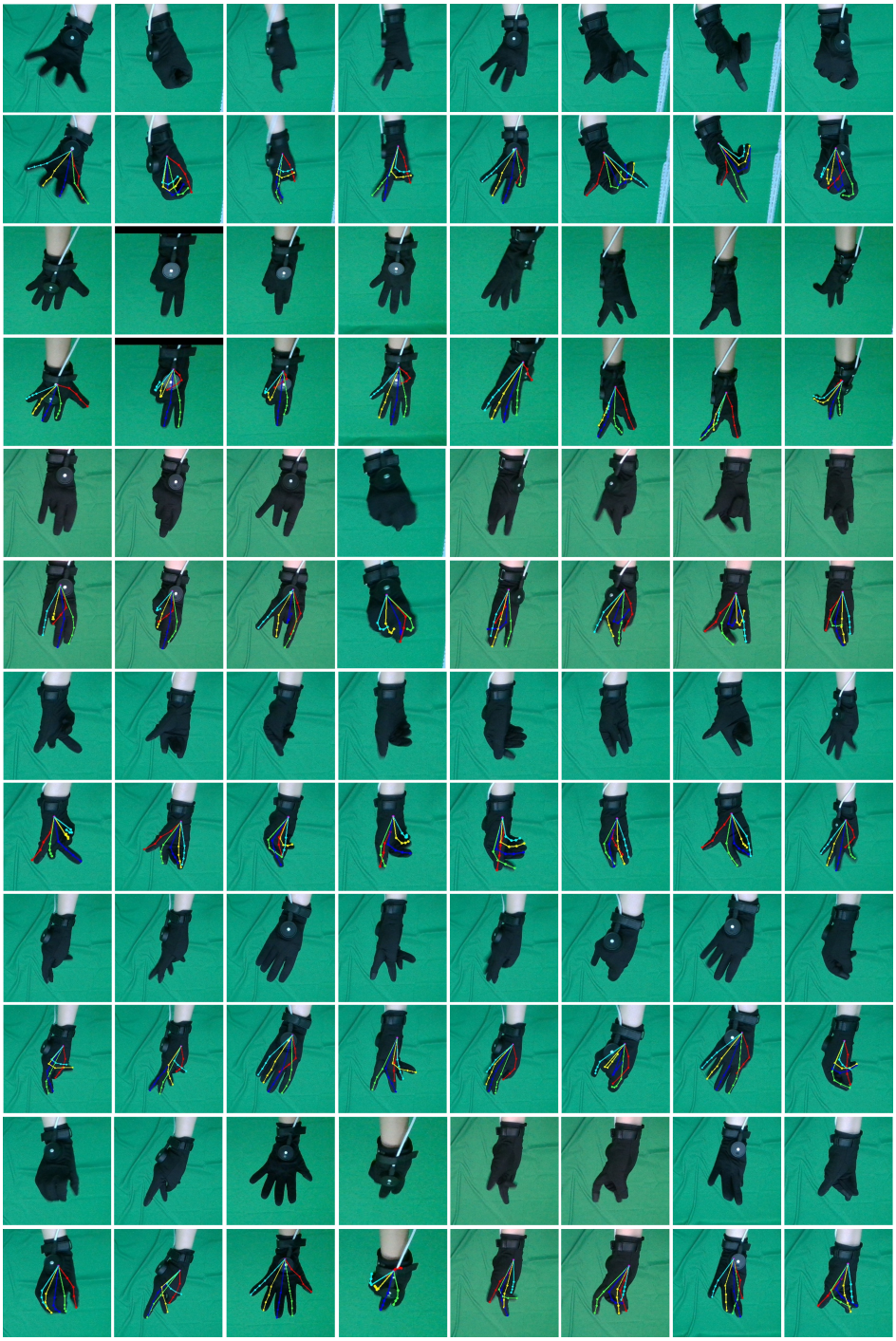

Fig. 1. RGB sample data of various hand poses along with their corresponding high-quality visual annotations.

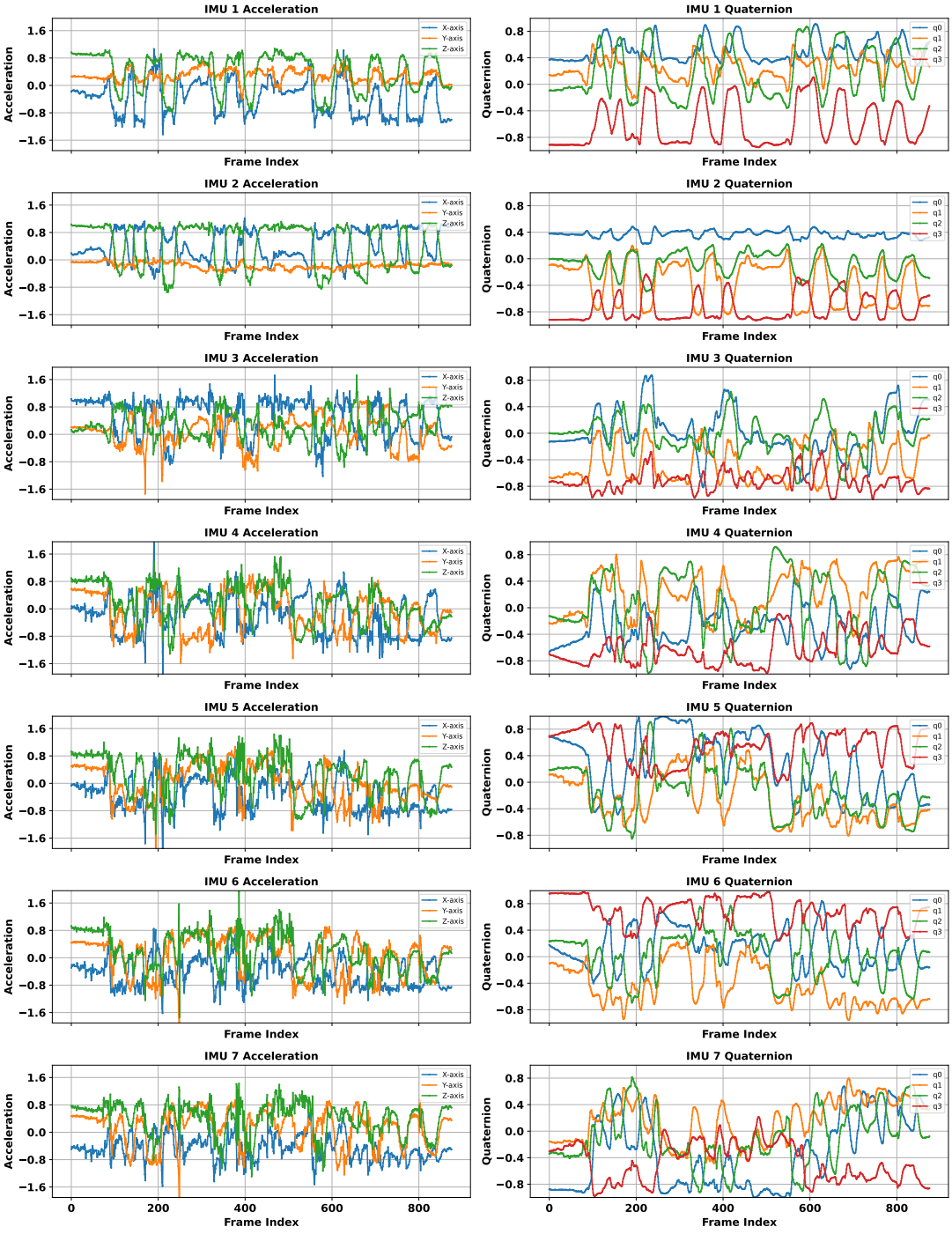

Fig. 2. Synchronized IMU acceleration and quaternion trajectories from all 7 IMUs in a continuous gesture sequence. Left: 3-axis acceleration(X, Y, Z). Right: quaternion (q0, q1, q2, q3).

left subplot showing 3-axis acceleration (X, Y, Z) and the right subplot showing unit quaternion (q0, q1, q2, q3). The acceleration signals exhibit sharp transitions and fine-grained variations, especially on finger-mounted IMUs, capturing rapid local dynamics. In contrast, palm-mounted and wrist-mounted IMUs show smoother transitions, reflecting global hand movement. The corresponding quaternion trajectories maintain consistent temporal alignment and capture coherent rotational patterns across sensors. These high quality multimodal data confirms the stability, precision, and responsiveness of our synchronized visual-inertial recording system, supporting reliable kinematic modeling and multiple gesture interacting tasks.

B More Details in Experiment

The VIFNet model is trained end-to-end by minimizing a composite loss function, $\mathcal{L}_{\text{VIFNet}}$, which aggregates several key terms for accurate 3D hand pose and shape estimation:

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

Here, the 3D Joint Loss ($\mathcal{L}_{\text{3D}}$) measures the difference between predicted and ground-truth 3D joint locations using MSE Loss. The 2D Joint Loss ($\mathcal{L}_{\text{2D}}$) ensures consistency with the 2D joint annotations by comparing the 2D projections of the predicted 3D joints with the ground-truth 2D joints using L1 loss. The MANO Parameter Loss ($\mathcal{L}_{\text{MANO}}$) directly supervises the MANO parameters ($\Psi = \{\theta, \beta, R, t, s\}$), encouraging the model to learn accurate 3D hand representations, with MSE loss applied. The Heatmap Loss ($\mathcal{L}_{\text{hp}}$) guides the visual-inertial feature extractor by minimizing the difference between the RGB and IMU heatmaps, helping the network localize hand joints, and is also computed using MSE loss. The weighting factors λ_{3D}, λ_{2D}, λ_{MANO}, and λ_{heatmap} balance the contribution of each loss term during training, which are set to 5.0, 1.0, 0.5, and 0.1, respectively.

To align the student's IMU-based features with the teacher's fused representations, we employ a contrastive distillation loss defined as:

$$\mathcal{L}_{\text{ts}} = \sum_{j \in I} -\log \frac{\exp\left(f_j^S \cdot f_j^T / \tau\right)}{\sum\limits_{k \in I} \exp\left(f_j^S \cdot f_k^T / \tau\right)}, \tag{2}$$

where f_j^S and f_j^T denote the normalized feature vectors from the student and teacher models, respectively, and τ is a temperature hyperparameter, which is set to 0.1.

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:

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

where $\mathcal{L}_{\text{ts}}$ denotes the contrastive distillation loss described above. The coefficients α_{3D}, α_{2D}, α_{MANO}, and α_{ts} balance the relative contribution of each term during training, which are set to 5.0, 1.0, 0.5, and 5.0, respectively.

C Ablation Study

The Effectiveness of Components. For a comprehensive understanding of our model, we conduct ablation experiments to demonstrate the improvements obtained by three important components, including group module, transformer

encoder, Multi-head Attention. We ensemble each module to the backbone network step-by-step, and compare the estimation performance on the VIHand. The results are illustrated in Table 1.

Table 1. Ablation study of the key components in VIFNet.

Methods	Input Modality	MPJPE (mm)	MPVPE (mm)
w/o group module	RGB + IMU	8.43	9.62
w/o transformer encoder	RGB + IMU	7.95	9.16
w/o cross-attention mechanism	RGB + IMU	8.19	9.47
VIFNet	RGB + IMU	**7.86**	**8.94**

The comprehensive VIFNet model achieves an outstanding performance with an MPJPE of 7.86 mm. However, when the group module is removed, the MPJPE increases to 8.43 mm. Similarly, eliminating the cross-attention mechanism results in an MPJPE of 8.19 mm, and dismissing the transformer encoder leads to an MPJPE of 7.95 mm. The results of MPVPE show a similar trend to those of MPJPE. These observations clearly demonstrate that the integration of the group module, Transformer encoder, and cross-attention mechanism plays a crucial role in enhancing the model's overall performance. Specifically, the group module effectively captures local joint features, providing a robust foundation for the model's accuracy. The transformer encoder, on the other hand, excels in extracting temporal and spatial related information from IMU features, thereby enriching the model's understanding of the data. The model's capabilities further enhanced by designing cross-attention-based multimodal fusion strategy. Collectively, these key components work synergistically to enable VIFNet to estimate joint positions with remarkable precision, thereby underscoring their indispensable role in the model's architecture.

D Datasheet for VIHand Dataset

D.1 Motivation

For what purpose was the dataset created? We contribute a large-scale visual-inertial hand pose dataset to the community in hope to advance multi-modal hand pose estimation under complex scenarios. By addressing the limitations of vision-only and IMU-only approaches, our dataset aims to bridge the gap between high-quality visual and inertial data, and foster research on deep semantic alignment and complementary information fusion across modalities.

Who created the dataset? Team members at the School of Intelligence and Computing, Tianjin University, China, and the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China, collected, updated, and maintained the dataset.

Who funded the creation of the dataset? The collection of the dataset is funded by the Tianjin Binhai Artificial Intelligence Innovation Center with the National Natural Science Foundation of China under Grant 62332019.

Any other comments? [N/A].

D.2 Composition

What do the instances that comprise the dataset represent? The VIHand dataset consists of synchronized multi-modal instances that represent human hand gestures under both bare-hand and glove-wearing conditions. Each instance includes RGB-D images captured from five Intel RealSense D415i cameras and inertial measurements from seven IMU sensors embedded in a wearable data glove. These instances represent fine-grained hand poses performed by 15 subjects

across a diverse set of predefined and spontaneous gesture sequences. Each sample is annotated with 3D hand joint coordinates and MANO model parameters.

How many instances are there in total? The VIHand dataset contains over 1.4 million synchronized multi-modal instances. Specifically, it comprises more than 1.4 million sets of RGB and depth images captured from five Intel RealSense D415i cameras and corresponding inertial measurements from seven IMU sensors. Each instance is further annotated with 3D hand joint positions and MANO model parameters.

Does the dataset contain all possible instances or is it a sample? The dataset provided with the link to the reviewer is the full dataset we collected. The full dataset will also be released after acceptance to support the following research.

What data does each instance consist of? Each instance in the VIHand dataset consists of raw multi-modal data, including five synchronized RGB images and five depth maps from Intel RealSense D415i cameras, and inertial data from seven IMU sensors.

Is there a label or target associated with each instance? [Yes]. The labels are described in *Annotation* part of Section 3.2.

Is any information missing from individual instances? [N/A].

Are relationships between individual instances made explicit? [No].

Are there recommended data splits? [Yes]. The dataset includes a recommended training and testing split with session-wise separation to prevent leakage. Details are in the benchmark link.

Are there any errors, sources of noise, or redundancies in the dataset? [Yes]. IMU signals may suffer from drift or interference. Visual data may be affected by occlusion, blur, or lighting changes. Minor annotation errors and redundant frames during held gestures may exist.

Is the dataset self-contained? [Yes]. All data and benchmark scripts are included via the provided link.

Does the dataset contain confidential data? [No]. All data are anonymized.

Does the dataset contain offensive or threatening content? [N/A].

Does the dataset relate to people? [Yes].

Does the dataset identify any subpopulations? [N/A].

Is it possible to identify individuals from the dataset? [No].

Does the dataset contain sensitive data? [No].

Any other comments? [N/A].

D.3 Collection Process

How was the data acquired? All data was directly collected through controlled experiments. Participants performed gestures under bare and glove-wearing conditions. RGB-D images were captured by five cameras and IMU data from seven sensors. Labels were obtained via a two-stage pipeline.

What mechanisms or procedures were used to collect the data? Detailed in Section 3, using Intel RealSense D415i cameras, wearable IMUs, and synchronized recording scripts.

Is the dataset a sample? [No].

Who was involved and how were they compensated? PhD and Master's research students, supported by scholarships and research assistant salaries.

Over what timeframe was the data collected? January 3–27, 2025, including setup and testing.

Was there ethical review? [Yes]. Reviewed and approved by Tianjin University.

Was consent obtained? [Yes]. All participants signed informed consent forms and were notified of the use and publication of data.
Can consent be revoked? [Yes]. Revocation is supported at any time.
Any other comments? [No].

D.4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling done? [Yes]. Detailed in Section 3.
Was raw data saved? [Yes]. Raw and processed data are both included for extensibility.
Is the preprocessing software available? [Yes]. Available on the project repository.
Any other comments? [No].

D.5 Uses

Has the dataset been used? [No]. This paper is the first.
Is there a repository of works using it? [No].
Other potential tasks? Gesture recognition, action classification, quality assessment, sensor fusion, cross-modal learning, and domain adaptation.
Any limits on future use due to design or cleaning? [No].
Any tasks it should not be used for? [Yes]. The dataset is restricted to non-commercial research use.
Any other comments? [No].

D.6 Distribution

Will the dataset be distributed externally? [No].
How will it be distributed? GitHub repository and direct download link.
When will it be distributed? Upon paper acceptance/publication.
Under what license? CC BY-NC 4.0.
Any third-party restrictions? [No].
Any export controls? [No].

D.7 Maintenance

Who maintains the dataset? Tianjin University and Shanghai Jiao Tong University.
How can they be contacted? Contact the first and corresponding authors via email (in paper and repository).
Erratum available? [Yes]. Via GitHub.
Will the dataset be updated? [Yes]. With new annotations and benchmarks.
Any limits on retention? [No].
Will older versions be maintained? [N/A].
Can others contribute? [Yes]. Contributions are welcome via GitHub.
Any other comments? [No].

D.8 Author Statement

The authors confirm all responsibility in case of rights violation and agree to the dataset license (CC BY-NC 4.0). All code is released under the MIT license.