DimHand: A Large-Scale Benchmark for Hand Pose Estimation under Dynamic Low-Light Conditions

ACM Reference Format:
. 2026. DimHand: A Large-Scale Benchmark for Hand Pose Estimation under Dynamic Low-Light Conditions. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3767308.3835262

In the supplementary material, we provide a video demo in Sec. A, more experiments in Sec. B, and more dataset details in Sec. C.

A Video Demo

In the supplementary material, we provide a video file to further illustrate visualization details in the DimHand dataset and the sequential comparative visualizations of the proposed DimNet against state-of-the-art methods.

The video presents further comparisons between previous datasets and our DimHand dataset. We provide representative visualizations of 15 collected gesture sequences, detailing the RGB images, depth maps, and high-precision annotations. These performed gestures encompass action-driven, pose-oriented, and composite types, alongside continuous spontaneous motions intended to capture a naturally diverse distribution of unconstrained hand behavior. Additionally, we provide sequential comparisons demonstrating the effectiveness of the proposed DimNet under hand self-occlusion and self-similarity, compared with WiLor, HVI enhancement with WiLor, and the ground truth.

B More Experiments

B.1 More Ablation Studies

To thoroughly evaluate our architectural configurations and hyperparameter settings, we conduct detailed ablation studies under the Cross-Action (CA) protocol, with comprehensive quantitative results presented in Table 1. Regarding the backbone architecture, ResNet-34 achieves the optimal balance between representation capacity and computational efficiency. Shallower networks like ResNet-18 lack sufficient capacity to extract robust semantic representations from degraded low-light images. Conversely, while deeper architectures such as ResNet-50 possess larger parameter spaces, they introduce significant computational overhead without yielding commensurate performance improvements. Our results indicate that the representation power of ResNet-34 [**?**] is already sufficient to capture complex hand structures effectively, making it the most optimal choice for our framework. Furthermore, the proposed KPC relies on the codebook size to discretize the continuous pose space and provide structural constraints. We evaluate different codebook dimensions and observe that $K = 1024$ yields the best performance. A smaller codebook limits the representation capacity for complex and fine-grained hand articulations. On the other hand, an excessively large codebook introduces representational redundancy and increases the risk of matching visual features to noisy or incorrect latent codes under low-light ambiguities. These optimal architectural and hyperparameter choices ensure that DimNet maintains precision while effectively decoupling the underlying kinematic structures from dynamic illumination.

Table 1: Ablation studies on the backbone architecture and KPC codebook size under the CA protocol. Errors are reported in mm.

KPC Codebook Size	P-MPJPE ↓	P-MPVPE ↓
Codebook ($K = 512$)	8.55	7.68
Codebook ($K = 1024$)	**7.75**	**7.03**
Codebook ($K = 2048$)	7.80	7.12

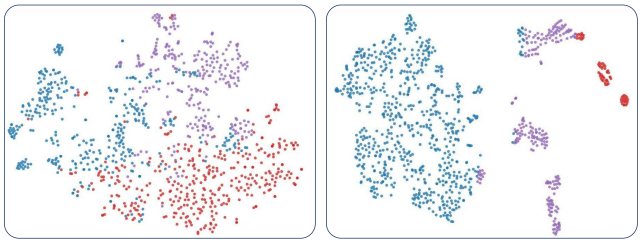
To further investigate the effectiveness of the proposed Kinematic Prior Codebook (KPC), we visualize the feature distributions before and after the codebook mapping using t-SNE. As shown in Figure 1(a), the initial query features extracted from the backbone are highly entangled and noisy, showing no clear organization under severe low-light degradation. In contrast, the refined features processed by the KPC (Figure 1(b)) spontaneously self-organize into multiple compact and cohesive clusters. While explicit semantic labels are not assigned to these individual subgroups, this distinct unsupervised separation demonstrates a crucial methodological advantage. It proves that the KPC effectively avoids representation collapse—a common failure mode where deep networks produce homogenous outputs when facing degraded inputs. Instead, by mapping continuous, corrupted features to these discrete geometric anchors, the KPC successfully constructs a structured and discriminative feature space, which is essential for resolving spatial ambiguities and estimating accurate 3D hand poses.

B.2 More Qualitative Experiments

To further evaluate our method, we provide extended qualitative comparisons in Figure 3. In addition to prior methods, we supplement our evaluation with HaMeR [**?**], a model based on a Vision Transformer trained on large-scale datasets. A row-by-row analysis illustrates the limitations of these models under low-light conditions. The failure modes can be broadly categorized into single-finger self-occlusions and visual ambiguities (rows 1 to 4), and complex multi-finger occlusions (rows 6 to 8). Specifically, within the first category, the 1st, 3rd, and 4th rows show that WiLoR [**?**] and HaMeR exhibit structural errors when handling single-finger occlusions. The 2nd row further highlights the issue of visual self-similarity, where WiLoR and HaMeR fail to distinguish between the

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835262



(a) Initial Features (b) Features After Codebook Learning

Figure 1: t-SNE visualization of features before and after the proposed KPC. (a) Initial query features are highly entangled under low-light conditions. (b) Refined features form distinct cohesive clusters. This demonstrates that KPC effectively avoids representation collapse and constructs a discriminative discrete space.

Table 2: Gender distribution of subjects in the DimHand.

Gender	Subject IDs
♂ Male	Subject 02, 03, 06, 07, 10, 12, 13, 14, 15
♀ Female	Subject 01, 04, 05, 08, 09, 11

interacting index and middle fingers. In all these instances, DimNet infers the hidden joint positions correctly. Regarding multi-finger occlusions, the 6th and 7th rows demonstrate that severe overlapping fingers cause estimation errors in WiLoR and HaMeR. The 8th row presents a spatial configuration involving a spread thumb and index finger combined with overlapping index and middle fingers. This specific combination of articulation and occlusion disrupts the spatial inference of WiLoR and HaMeR, causing incorrect joint predictions, yet DimNet maintains kinematic correctness.

Furthermore, the visualizations indicate that using a low-light enhancement preprocessing step introduces additional errors. As shown in the 5th and 7th rows, the artifacts and sensor noise introduced by the HVI [**?**] enhancement process alter the visual semantics. Specifically in the 5th row, these artifacts cause HaMeR to predict a flipped hand orientation, while WiLoR also exhibits joint misalignments. In contrast, DimNet constructs 3D hand structures directly, demonstrating the effectiveness of its explicit geometric constraints and illumination adaptation mechanisms. All evaluations across the compared models are computed under the premise of successful hand detection to ensure a standard assessment of their hand pose estimation capabilities.

C More Dataset Details

C.1 Gesture Definition

In this work, we propose DimHand, a large-scale dataset for 3D hand pose estimation in dynamic low-light conditions. Due to double-blind review restrictions on including external links, we commit to releasing the dataset and corresponding benchmarks publicly at a later stage. As illustrated in Table 3, we define 15 meticulously designed continuous pose sequences to capture a naturally

Figure 2: Visual comparison of hand morphological diversity across 15 subjects in the DimHand dataset, illustrating the variance in hand shape, overall size, and the natural bending states of the fingers.

diverse distribution of hand behavior. These sequences are systematically categorized into four dimensions based on structural variation and semantic diversity. The Global dimension includes calibration motions where the open hand freely rotates and translates, capturing global Six Degrees of Freedom (6-DoF) variations. The Pose-oriented category focuses on continuously evolving and sequentially structured poses, intentionally designed to introduce severe hand self-occlusions and complex topological structures. The Action-driven dimension comprises fundamental dynamic articulations representing meaningful everyday actions, such as joint-by-joint bending, tapping, and pinching. Finally, the Composite dimension encompasses complex interactions characterized by the simultaneous movement of all fingers combined with varying degrees of lateral opening and closing. The inherent properties of these defined continuous sequences, ranging from subtle localized articulations to severe structural self-occlusions, guarantee an exceptionally diversity in hand poses and ensure a comprehensive representation of natural hand behavior.

C.2 Hand Morphological Diversity

We present a visual gallery of the participating subjects in the DimHand dataset to demonstrate the rich morphological diversity of the captured hands. As shown in Figure 2, we provide representative samples from all 15 subjects captured under a consistent open hand pose. This comprehensive visualization highlights the significant variance in overall hand size, finger length ratios, palm width, and structural profiles across the recruited population. Furthermore, the natural resting states and joint flexibilities differ remarkably among individuals, naturally introducing a wide spectrum of kinematic constraints. By encompassing such a broad manifold of anatomical variations, our dataset provides a highly diverse and representative foundation. This rich morphological coverage closely reflects the natural physiological differences found in real-world scenarios, supporting the development and evaluation of 3D hand pose estimation models across a wider range of user demographics.

Figure 3: Extended qualitative comparisons on the DimHand dataset. We evaluate the proposed DimNet against state-of-the-art baselines including WiLoR and HaMeR. The notation HVI *enh.* indicates that the low-light images are preprocessed using the HVI low-light enhancement network, yielding the HVI+WiLoR and HVI+HaMeR pipelines. While DimNet consistently maintains robust structural integrity, the baseline methods suffer from severe estimation failures caused by either extreme dark degradation or the misleading artifacts introduced by the image enhancement process.

C.3 Data Annotation Details

Obtaining precise hand bounding boxes is fundamental for accurate 3D hand pose annotation. While conventional RGB-based detection is highly susceptible to photometric degradation and motion blur in low-light environments, the active infrared depth modality provides a robust alternative by preserving stable, illumination-invariant

Table 3: All gesture classes and their corresponding descriptions defined in the DimHand dataset.

Gesture Id	Gesture Name	Type	Class Description
00	Calibration	Global	Open hand rotating and translating freely in all directions
01	Finger Counting	Pose-oriented	Sequential hand poses representing digits from 0 to 9
02	Finger Bending	Action-driven	Individual and pairwise joint-by-joint finger bending towards the palm
03	Finger Tapping	Action-driven	Individual and pairwise finger tapping motions
04	Finger Pinching	Action-driven	Pairwise, three-finger, and four-finger pinching combinations
05	Thumb Tap Palm	Action-driven	Thumb sequentially tapping various joints on the palmar side
06	Thumb Tap Back	Action-driven	Thumb sequentially tapping various joints on the dorsal side
07	Finger Walking	Composite	Staggered finger walking motions simulating piano playing
08	Finger Snapping	Composite	Snapping motions by rubbing the thumb against each finger
09	Thumb Sliding Palm	Composite	Thumb continuously sliding across fingers on the palmar side
10	Thumb Sliding Back	Composite	Thumb continuously sliding across fingers on the dorsal side
11	Finger Wiggling	Action-driven	Continuous horizontal wiggling motions of individual fingers
12	Finger Crossing	Pose-oriented	Fingers stacking and crossing to create severe self-occlusions
13	Finger Interlocking	Pose-oriented	Thumb interlocking and weaving through gaps between fingers
14	Random Gestures	Composite	Continuous, spontaneous, and unconstrained natural hand motions

geometric contours. To leverage this advantage, we construct a dedicated depth-guided hand detection model. To acquire training labels without manual annotation, we employ a cross-modal supervision strategy using an auxiliary dataset captured under standard well-lit conditions. In this scenario, a pre-trained YOLOv7 [**?**] extracts high-confidence bounding box centers from RGB images. These centers serve as the absolute ground truth to supervise a customized model that takes only the corresponding paired depth maps as input. By regressing centers directly from structural depth gradients, the model learns highly robust spatial representations. During actual low-light acquisition, this depth-only model reliably detects 2D hand centers to obtain the spatial ground truth, entirely bypassing the degraded RGB modality. These stable detections provide accurate spatial priors for hand cropping and downstream 3D annotation. Consequently, all experimental evaluations across the compared methods are strictly computed only on successfully detected frames, ensuring that missed detections do not confound the assessment of hand pose estimation accuracy.

D Datasheet for DimHand Dataset

D.1 Composition

What do the instances that comprise the dataset represent? DimHand consists of synchronized multi-modal instances that represent human hand gestures under dynamic low-Light conditions. Each instance includes RGB-D images captured from five Intel RealSense D415i cameras under dynamic illumination ranging from 0.1 to 10 lux. 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? DimHand contains over 745K synchronized multi-modal instances. Specifically, it comprises more than 745K sets of RGB and depth images captured from five Intel RealSense D415i cameras. 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? It is the full dataset we collected. The full dataset will be released after acceptance to support the following research.

What data does each instance consist of? Each instance in DimHand consists of raw multi-modal data, including five synchronized RGB images and five depth maps from Intel RealSense D415i cameras under dynamic illumination ranging from 0.1 to 10 lux.

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

Are there recommended data splits? [Yes]. The dataset includes a recommended training and testing split to prevent leakage. Details are in Section 5.1. *Experimental Setup.*

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

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

Does the dataset contain 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].To ensure strict privacy protection and compliance with ethical standards, the data collection protocol of DimHand explicitly prohibits the capture of facial features or any personally identifiable information. The camera arrays are strictly focused on the hand and forearm regions of the subjects.
Does the dataset contain sensitive data? [No].As detailed in Table 2 and Figure 2, we only record metadata regarding gender and general hand morphological profiles to guarantee the demographic diversity of the dataset while maintaining subject anonymity.
Any other comments? [N/A].

D.2 Collection Process

How was the data acquired? All data was directly collected through controlled experiments. Participants performed gestures underdynamic low-light conditions. RGB-D images were captured by five RealSense D415i cameras. Labels were obtained via a two-stage pipeline.
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 1–31, 2026, including setup and testing.
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]. It is supported at any time.
Any other comments? [No].

D.3 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.
Any other comments? [No].

D.4 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, 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.5 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.
Any third-party restrictions? [No].
Any export controls? [No].