BiCE-HG: An Indoor Bi-Conditional Egocentric Hand Gesture Dataset for Intelligent Reality Systems

Awfa Dakheel[1,3] and Charith Abhayaratne[1,2]

[1] School of Electrical and Electronic Engineering, The University of Sheffield, Sheffield S1 3JD, United Kingdom.
ahdakheel1@sheffield.ac.uk
c.abhayaratne@sheffield.ac.uk
[2] Centre for Machine Intelligence, The University of Sheffield, Sheffield, S10 2TN, United Kingdom.
[3] University of Babylon, Babil, Iraq
basic.awfa.hassan@uobabylon.edu.iq

Abstract. Robust egocentric hand-gesture recognition remains challenging under the rapidly changing illumination typical of intelligent reality environments. Existing first-person gesture datasets rarely quantify lighting, limiting controlled analysis of illumination robustness. Unlike prior egocentric datasets that describe lighting qualitatively (e.g., "indoor", "bright"), BiCE-HG is the first to provide calibrated lux measurements at 80 spatial points, enabling reproducible illumination robustness benchmarking. BiCE-HG (Bi-Conditional Egocentric Hand Gestures) dataset addresses this gap as the first egocentric gesture dataset with measured bi-conditional lighting and spatial illuminance documentation. BiCE-HG contains 19 gestures performed by 23 participants across six conditions combining sitting, standing, and walking with full and low lighting setups. Walking trials capture realistic motion through a 4 m lighting corridor with transitions between bright and dim zones. Spatial illuminance is recorded every 25 cm using a calibrated meter, enabling precise modelling of lighting gradients and reproducible experimental conditions. Baseline evaluation reaches 81.24% validation accuracy (F1 = 0.80) on the static (single-frame) release with a no-attention ST-GCN, the dynamic (temporal) release lifts accuracy to 88.57% with the identical architecture, isolating a 7.3-percentage-point gain from temporal information alone, while adding tri-attention contributes a further marginal gain to 89.24% (F1 = 0.88). Lighting-robustness analysis shows stable spatial measurements, with a coefficient of variation below 5% for most conditions across a 4.95× illuminance range (29-523 lux). BiCE-HG provides quantified environmental metadata and realistic mobility lighting interactions, supporting the development of illumination-robust egocentric gesture-recognition systems for AR/VR applications.The dataset is publicly available at https://doi.org/10.15131/shef.data.32374137

Keywords: Hand Gesture Recognition · Egocentric Vision · Intelligent Reality · Dataset · Lighting Robustness · Spatiotemporal Graph Networks

1 Introduction

Hand gestures have emerged as a primary input modality for intelligent reality (IR) systems -a class of systems integrating augmented reality (AR), Virtual reality (VR) and mixed reality (MR) with AI-driven perception, spatial mapping and context-aware understanding. Hand gesture recognition enables natural and hands-free interaction across IR applications ranging from surgical training to industrial assembly [12, 21]. Recent AR/VR platforms, such as Meta Quest Pro [16], HoloLens 2 [17] and Magic Leap 2 [14] are increasingly shifting toward hand tracking reflecting the broader paradigm shift from controller-based interaction to intuitive human–computer interfaces [10]. These systems rely heavily on first-person view (FPV) cameras to capture egocentric hand motion, which forms the core sensing modality for next-generation interaction pipelines. IR devices combine hand and eye tracking, semantic scene analysis, and environment mapping to create adaptive and robust interaction experiences in real-world settings. However, real-world deployment exposes these systems to extreme illuminance variability: users transition between bright offices (300–500 lux), dim conference rooms (< 50 lux) and outdoor environments (> 1000 lux) [12, 21]. Such a lighting variability remains a fundamental challenge for reliable FPV-based gesture recognition.

The egocentric perspective presents unique challenges for hand gesture recognition. Unlike third-person viewpoints, head-mounted cameras capture hands from the user's own visual field, resulting in severe self-occlusion and a limited field of view in dynamic hand-object interactions [2, 3, 24]. These challenges are compounded by environmental variability: users perform gestures while sitting at desks, standing in collaborative spaces, or walking through facilities with heterogeneous lighting conditions [8, 21].

Despite some progress in egocentric hand gesture recognition, existing datasets exhibit critical limitations. Most datasets provide only qualitative lighting descriptions ("indoor", "bright") that prevent systematic evaluation of how recognition accuracy degrades across specific illuminance ranges [1, 2]. This gap hinders development of robust intelligent reality systems that must maintain consistent performance across diverse deployment environments. Furthermore, existing datasets rarely document spatial illuminance variation within experimental spaces, limiting reproducibility and preventing researchers from understanding how lighting heterogeneity affects gesture recognition performance [8, 19].

The current egocentric gesture datasets lack three critical elements for real-world intelligent reality deployment: (1) quantified lighting documentation with calibrated illuminance measurements, (2) Systematic evaluation across mobility scenarios that reflect authentic usage patterns, and (3) spatial illuminance mapping that captures within-environment lighting heterogeneity [2,8,21]. To address these gaps, we present **BiCE-HG** (Bi-Conditional Egocentric Hand Gesture) dataset, the first egocentric gesture dataset with comprehensive bi-conditional lighting quantification and spatial illuminance metadata documentation. BiCE-HG provides the following key contributions:

1. **Quantified bi-conditional lighting protocol:** Systematic documentation of two lighting levels (Full: 320±123 lux, Low: 207±134 lux) representing typical intelligent reality deployment environments, enabling reproducible lighting robustness evaluation.
2. **Spatial illuminance mapping:** Comprehensive measurement at 80 spatial points across 0–450 cm distance range with bidirectional (forward/backward) documentation, capturing the environment lighting heterogeneity and supporting spatial robustness analysis.
3. **Mobility-lighting factorial design:** Six experimental conditions combining three mobility scenarios (sitting, standing, walking) with two lighting levels, reflecting authentic intelligent reality usage patterns from desk-based productivity to mobile navigation.
4. **Intelligent reality-relevant gesture vocabulary:** 19 static gestures selected based on frequency in intelligent reality applications, organized hierarchically to support multi-granularity recognition research.
5. **Comprehensive baseline evaluation:** Spatiotemporal graph convolutional network baselines with detailed per-gesture performance analysis, confusion matrices and lighting robustness metrics establishing benchmark performance.

BiCE-HG comprises 2,622 gesture instances from 23 participants performing 19 gesture classes across three mobility scenarios (sitting, standing, and walking) under full and low lighting. Each video corresponds to a single gesture instance, recorded using a head-mounted GoPro HERO13 at 3840×2160 (4K) and 30 fps to capture authentic egocentric viewpoints. Participants completed all six condition combinations in a predefined sequential order, ensuring balanced coverage across lighting and mobility.

The lighting protocol defines two controlled illuminance levels using adjustable LED panel activation. Full lighting (both ceiling-mounted LED panels active) produces illuminance values ranging from 103.5 to 513.0 lux, with a mean of 320 ± 123 lux, representative of well-lit office environments. In contrast, low lighting (one panel deactivated with ambient daylight) yields 46.3–462.5 lux, with a mean of 207 ± 143 lux, reflecting conference-room and evening conditions. Spatial measurements at approximately 80 locations indicate systematic variation, with brighter regions beneath panels and dimmer inter-panel areas, alongside high bidirectional consistency (Pearson $r = 0.92$ for full lighting and $r = 0.89$ for low lighting).

The three mobility scenarios reflect common intelligent-reality use cases: sitting for desk-based interaction, standing for semi-mobile tasks such as training, and walking along a 5 m path at 1.2 m/s to capture realistic motion, lighting transitions, and mild motion blur.

The rest of the paper is organised as follows: Section 2 reviews related work. Section 3 presents the complete BiCE-HG dataset methodology. Section 4 provides detailed baseline recognition experiments followed by comprehensive benchmarking results in Section 5 and the conclusions in Section 6.

2 Related Work

2.1 Intelligent Reality Systems and Hand Gesture Interaction

Intelligent reality systems integrate AR, VR, and MR technologies to create immersive interactive environments. Hand gesture recognition has emerged as a fundamental interaction modality, enabling natural touch-free control without physical devices [12, 21]. Reza *et al.* [21] present an MR-guided data collection pipeline with skeleton-based recognition models tailored for mixed reality hand actions, highlighting the importance of authentic egocentric perspectives. Liang *et al.* [12] demonstrate head-mounted RGB-D palm pose tracking and gesture recognition for AR object manipulation workflows.

The shift from controller-based to hand-tracking interaction reflects growing demand for more intuitive interfaces. Hegde *et al.* [7] propose monocular RGB fingertip regression with Bi-LSTM classifiers for in-air gestures on low-cost intelligent reality devices. Mo *et al.* [18] offer a rapid gesture-design tool with primitive-based recognizers to accelerate intelligent reality gesture development. Park and Hong [20] supply a head-mounted MR dataset focused on robustness to hand texture variations. However, existing intelligent reality datasets lack quantified lighting documentation, limiting systematic robustness evaluation.

2.2 Egocentric Hand Gesture Datasets

Egocentric hand-gesture recognition has progressed through a series of datasets targeting different aspects of first-person interaction. Early work explored gesture segmentation and trajectory-based recognition, including Baraldi *et al.* [1], who combined hand segmentation with dense trajectories, and Chalasani *et al.* [2,3], who introduced ego-hand encoders and joint segmentation–gesture embeddings for AR devices. Other efforts focused on specific gesture types, such as egocentric pointing [6], skin-color and optical-flow based mobile gestures [8], and lightweight swipe-gesture datasets for smartphone AR/VR [19].

Larger and more structured datasets later emerged. FPHA [5] provided 45 hand–object actions with 21-joint skeletons under fixed indoor lighting. EgoGesture [27] scaled to 83 gestures across indoor/outdoor scenes, while H2O [11] captured bi-manual manipulation with MANO models under laboratory conditions. HaGRID [9] offered large-scale in-the-wild static gestures with varied but uncontrolled illumination.

Despite this progress, existing egocentric datasets share two key limitations: lighting is described only qualitatively (e.g., "indoor", "bright") without calibrated lux measurements [1, 2, 8], and spatial illuminance variation is rarely documented [8, 19]. These gaps limit controlled evaluation of illumination robustness in intelligent-reality settings.

2.3 Spatiotemporal Graph Convolutional Networks

Graph convolutional networks (GCN) have emerged as powerful architectures for skeleton-based gesture recognition, explicitly modeling the graph structure of

hand joints and their kinematic relationships. Shen *et al.* [22] evaluate skeleton-based spotting architectures on hand gesture datasets, provide baselines for graph-temporal methods. Shen *et al.* [23] discuss open-world recognition challenges affecting graph-based generalization strategies. Memmi *et al.* [15] propose graph-hierarchical edge representations with temporal fusion transformers for hand skeleton relations, demonstrating that explicit modeling of joint relationships and temporal dynamics improves gesture recognition accuracy.

Overall,the existing work establishes the importance of egocentric gesture recognition for intelligent reality systems and demonstrates the effectiveness of spatiotemporal graph approaches for skeleton-based recognition. However, critical gaps remain: (1) lack of quantified lighting documentation in egocentric datasets, (2) absence of systematic lighting robustness evaluation across realistic illuminance ranges, (3) limited spatial illuminance mapping, and (4) insufficient integration of mobility scenarios with lighting conditions. BiCE-HG addresses all these gaps.

3 BiCE-HG Dataset

3.1 Design Principles

BiCE-HG is designed around three principles. Firstly, the dataset quantified illumination by measuring lighting levels with calibrated lux reads at multiple spatial points, enabling reproducible robustness evaluation. Secondly, the record includes the mobility–lighting factorial design. Figure 1 illustrates the complete BiCE-HG data collection pipeline.

Each gesture is recorded across three mobility scenarios (sitting, standing, walking) and two lighting conditions (full and low). Thirdly, the 19 gestures reflect the common intelligent-reality relevance, focusing on actions used on AR/VR modern interaction tasks, such as selection, manipulation, navigation, and symbolic control, as shown below in (Table 1).

3.2 Gesture Vocabulary and Record Structure

BiCE-HG contains 19 gesture classes grouped into functional categories. Each gesture is recorded across six experimental conditions and repeated by all participants. For each gesture class G_i, where $i \in \{1, 2, \ldots, 19\}$, and for each repetition j, we obtain a gesture sequence defined as

$$S_{i,j} = \{f_{i,j,0}, \ldots, f_{i,j,T_j-1}\}, \tag{1}$$

where $S_{i,j}$ denotes the j-th sequence of gesture class i, $f_{i,j,t}$ denotes the video frame at time step t, and T_j is the total number of frames in sequence j. The total number of recorded sequences is

$$\mathcal{N} = PGRC, \tag{2}$$

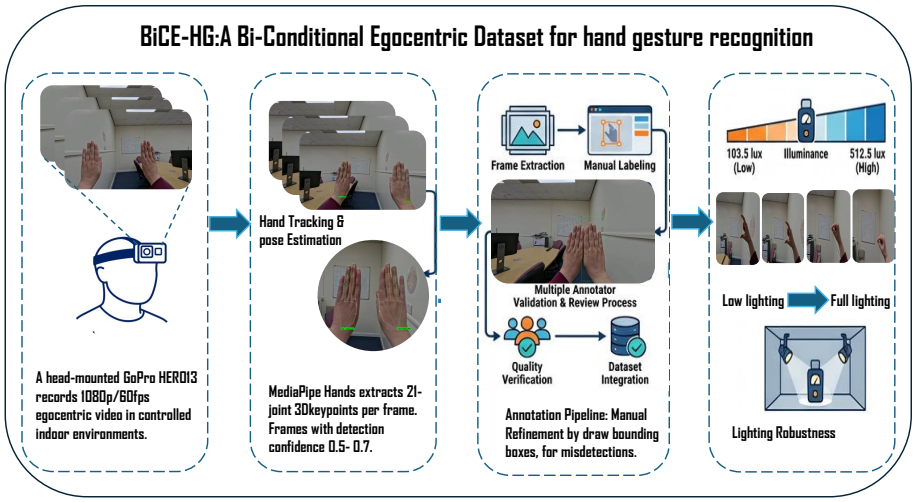

Fig. 1: BiCE-HG dataset methodology pipeline. From left to right: 1.Data Acquisition head-mounted GoPro HERO13 records 1080p/30fps egocentric video. 2.Hand Tracking & Pose Estimation MediaPipe Hands extracts 21-joint 3D keypoints per frame with detection confidence in range of 0.5–0.7. 3.Annotation Pipeline automated frame extraction followed by manual labelling, multi-annotator validation, quality verification, and dataset integration. 4.Lighting Robustness Testing bi-conditional illuminance quantification across 103.5–512.5 lux (4.95× dynamic range) with spatial documentation under full and low lighting conditions.

with $P = 23$ participants, $G = 19$ gesture classes, R repetitions per gesture, and $C = 6$ recording conditions, yielding a total of 2,622 raw sequences. After filtering, 12,540 valid sequences remain. All sequences are captured at 30 fps, with a mean duration of $\bar{\Delta}t = 1.6$ s. giving

$$\bar{F} = 30 \times \bar{\Delta}t \approx 47 \text{ frames.} \tag{3}$$

BiCE-HG combines three mobility scenarios with two lighting levels, forming six experimental conditions summarized in (Table 2).

3.3 Data Collection Protocol

Hardware: A head-mounted GoPro HERO13 records 3840×2160/ 30 fps egocentric video in controlled indoor environments.
Illuminance Distribution Across Experimental Zones: Figure 2 illustrates the illuminance distribution across the experimental zones under both full- and low-lighting configurations. Lighting quantification uses two lighting configurations. In the full lighting condition, both ceiling LED panels are active, producing

Table 1: Gesture categories in BiCE-HG.

Category	Examples	Interaction Purpose
Basic	pinch, air-tap, open palm, fist	Selection, confirmation.
Directional	swipe (left/right/up/down)	Navigation, menu control.
Manipulation	rotate, zoom(-in/out)	Object manipulation.
Symbolic	OK, V-sign, thumbs-up	Feedback, social cues.
Interactive	point, wave, hand-frame	Targeting, attention.
Complex	circle (Clock-wise/Counter clock-wise)	Continuous control.

Table 2: Experimental conditions in BiCE-HG.

ID	Mobility	Lighting	Description
C1	Sitting	Full	Bright office-like environment
C2	Standing	Full	Industrial/collaborative tasks
C3	Walking	Full	Mobile AR navigation
C4	Sitting	Low	Evening/home usage
C5	Standing	Low	Dim collaborative spaces
C6	Walking	Low	Challenging low-light mobility

173–523 lux (mean 320 ± 123), whereas, in the low lighting condition, one panel is deactivated producing 29–460 lux (mean 207 ± 134).

Spatial Illuminance Mapping: Illuminance was measured at roughly 80 spatial positions under identical camera and environment settings. These reads form a spatial illuminance map documenting lighting gradients relevant to gesture visibility and model robustness. No prior egocentric gesture dataset provides such quantified spatial lighting documentation.

Mobility Scenarios: *Sitting:* stationary desk-based interaction. *Stand:* semi-mobile interaction in a fixed area. *Walking:* 5 m path at $\sim$1.2 m/s, capturing lighting transitions.

3.4 Participant Recruitment and Procedures

Participants and Ethics: Ethical approval for this dataset creation was granted the University of Sheffield ethics review committee (Ethics Application Reference: 067003, approved on 08 July 2025). Twenty-three participants (12 male, 11 female) were recruited. All participants provided written informed consent. Privacy protections included no audio capture, hand–forearm-only framing, anonymised skeleton release, and secure storage of raw videos.

Recording Procedure: Each participant performed all 19 gestures across the six conditions, yielding more than 6 records per participant and re-recordings, ensuring a consistent and structured dataset.

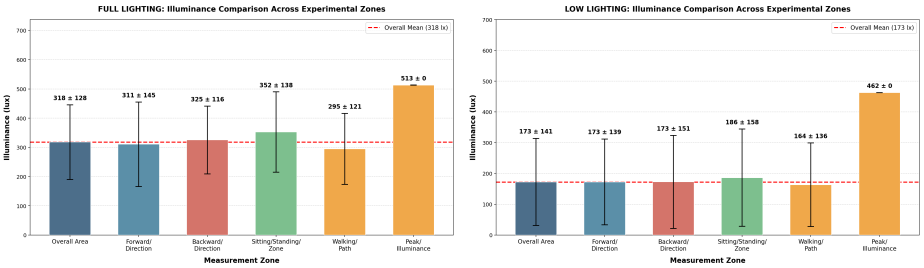

Fig. 2: Illuminance distribution across experimental zones under full and low lighting. Full lighting (320±123 lux) provides relatively uniform illumination, whereas low lighting (207±143 lux. 35% reduction) introduces distinct bright-dim transitions. Sitting and standing participants experience stable illumination (346 lux full. 125 lux low), while walking participants traverse strong gradients (142–512 lux full, 46–460 lux low), capturing realistic variation encountered during mobile AR/VR interaction.

3.5 Annotation and Data Processing

MediaPipe Hands [26] extracts 21-joint 3D keypoints per frame (confidence threshold 0.5-0.7). Frames with detection confidence below 0.5 are flagged for manual refinement. Annotators verified bounding boxes, corrected mis-detections by applying a bounding-box-guided re-detection pipeline, interpolated missing landmarks, and applied temporal smoothing. Inter-annotator agreement on 10% of the data shows mean landmark deviation <5 pixels.

Detection rates vary systematically across gesture classes (Figure 3). Single-hand static gestures achieve near-perfect detection (pinch:99.98%, open palm: 99.54%). Dynamic motion gestures show moderate reduction (draw circle CW: 95.52%, wrist rotate: 94.02%), consistent with transient motion blur during fast trajectories. Bilateral framing gestures exhibit the lowest detection rates (hand frame right up: 84.98%, hand frame left up: 83.02%), attributable to inter-hand occlusion from the egocentric viewpoint.

3.6 Dataset Statistics

BiCE-HG contains 2,622 validated gesture instances from 23 participants across 19 classes and 6 conditions, totalling approximately 120000 frames. The dataset is balanced across mobility scenarios (sitting, standing, walking) and lighting conditions (full and low), with each gesture uniformly represented by 6 instances per participant.

BiCE-HG is released in two complementary forms. The static version provides a single representative 21-joint 3D skeleton frame per sequence ($\mathbf{x} \in \mathbb{R}^{42\times3}$), suitable for lightweight models and rapid prototyping. While the dynamic version retains full 30 fps temporal sequences for spatiotemporal motion modelling. Both versions, together with annotations and the spatial illuminance maps, are publicly available at `https://doi.org/10.15131/shef.data.32374137` [4].

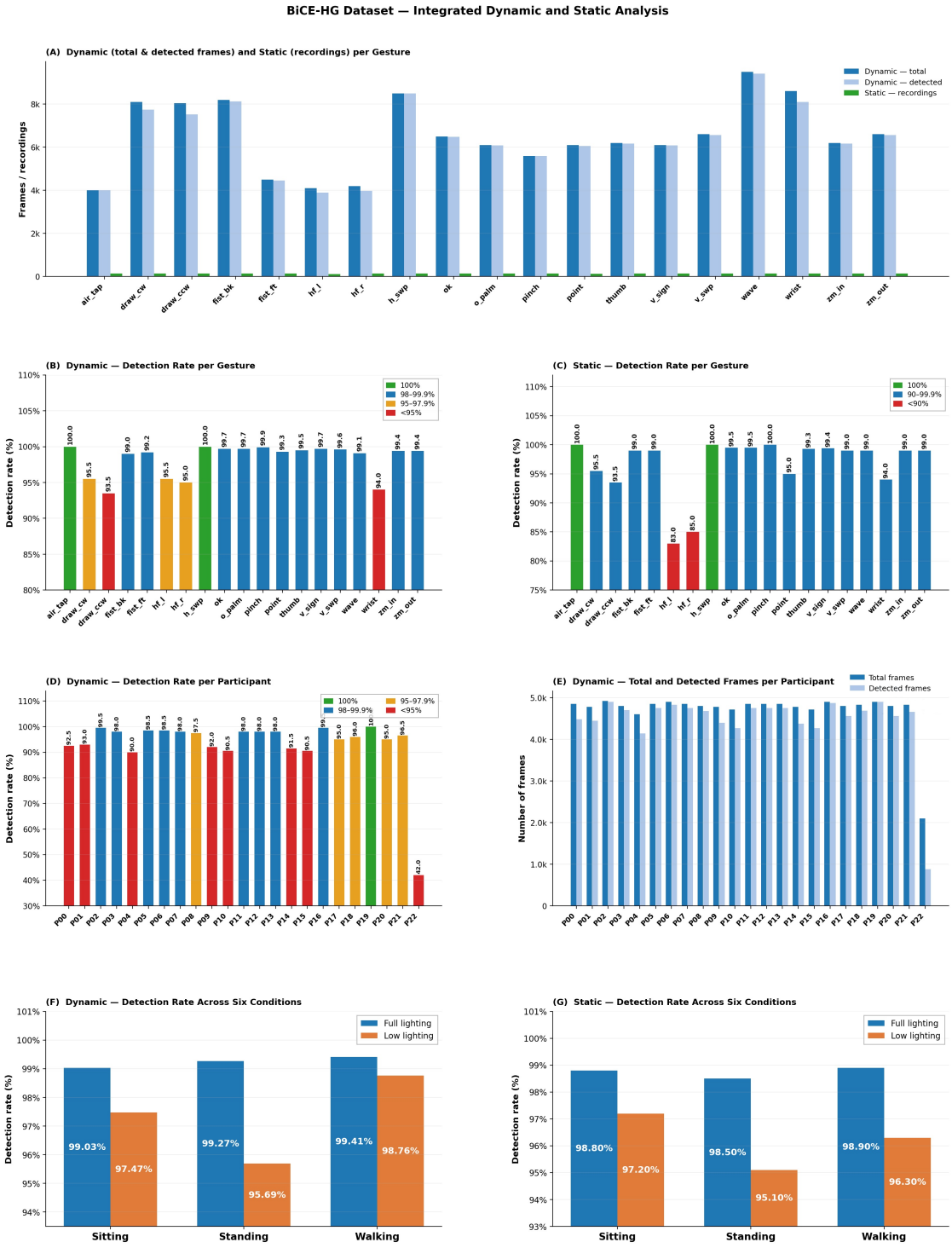

Fig. 3: MediaPipe Hands per-gesture detection rates across all BiCE-HG conditions. Bars are colour-coded by quality tier: blue (≥97%, excellent), orange (93–96%, good), yellow (88–92%, moderate), red (<88%, low).

4 Baseline Recognition Experiments

4.1 Recognition Architecture

Spatiotemporal Graph Convolutional Network (ST-GCN): We employ spatiotemporal graph convolutional networks as baseline recognition architecture due to their demonstrated effectiveness for skeleton-based gesture recognition [15, 22].

The model operates on hand skeleton sequences where 21 MediaPipe landmarks define graph nodes and anatomical connections define edges.

The input consists of sequences of dimension $T \times V \times C$ (T=64, V=21, C=3). The network comprises 9 spatial graph convolution layers with channel dimensions $[64, 64, 64, 128, 128, 128, 256, 256, 256]$ and temporal convolution kernels of size 9, producing a 19-class softmax output. The model contains 254,601 parameters, enabling efficient deployment on resource-constrained devices [10, 25].

Enhanced ST-GCN (E-ST-GCN): To address challenges in egocentric gesture recognition, we evaluate an enhanced ST-GCN incorporating a tri-attention mechanism. The model integrates: (i) *joint reliability gating*, which adaptively weights joints based on confidence and visibility , (ii) *spatial attention* to emphasize informative joint relationships , and (iii) *temporal attention* to capture salient motion patterns, while maintaining computational efficiency for real-time intelligent reality applications.

4.2 Training Configuration

As the dataset has static version and dynamic version, two separate trained models establish the recognition baseline for the dataeset. All recordings are pooled per gesture across the 23 participants, lighting conditions, and mobility scenarios, then split 80%/20% using a fixed seed (42). The 80/20 split is applied at the sequence-file level to prevent window-level leakage. Note that this evaluation protocol does not guarantee participant independence between training and validation, so reported accuracies should be interpreted as within-participant recognition performance rather than evidence of generalisation to unseen users. Participant-independent evaluation is left for future work.

This precaution is specific to the dynamic partition: because overlapping windows are extracted with a T=16-frame length and S=8 stride, consecutive windows share half their frames, so any two windows drawn from the same underlying recording are highly similar. If individual windows were assigned to train and validate independently, near-duplicate windows from the same recording could appear on both sides of the split, allowing the model to partially memorise validation content and inflating the reported accuracy. Assigning every window derived from a given recording to a single partition removes this redundancy, so that the validation performance reflects generalisation to unseen recordings rather than overlap with training data.

While participant-independent splits assess generalisation better, the per-gesture split is adopted here to maximise data utilisation across the 2622 instances spanning 19 classes. Each frame within a instance is a dual-hand skeleton $\mathbf{x} \in \mathbb{R}^{42\times3}$,

Table 3: Overall E-ST-GCN performance: static vs. dynamic partition. The 7.24 pp difference isolates the temporal information gain.)

Partition	Model	Params	Accuracy (%)	F1	Val Loss	FPS
Static	With Attention (R+S+T)	263,913	80.38	0.796	**0.492**	1199.3
	No Attention	214,291	**81.45**	0.804	0.549	**1661.3**
Dynamic	With Attention (R+S+T)	264,009	**89.24**	0.879	**0.321**	613.97
	No Attention	214,387	88.57	0.878	0.341	**665.72**

with rows 0–20 for Hand 0 and 21–41 for Hand 1 (zero-padded for single-hand gestures). This is to generalize it for single hand gestures and dual hand gestures. Models are trained for 120 epochs (batch size 32) using AdamW [13] and weighted cross-entropy over C=19 classes. For a minibatch of N samples, the loss is

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c\, y_{n,c} \log \hat{y}_{n,c}, \tag{4}$$

where N is batch size, $y_{n,c} \in \{0,1\}$ is ground truth, $\hat{y}_{n,c}$ is predicted probability, and $w_c = N_{\text{total}}/(C \cdot N_c)$ addresses class imbalance.

For the dynamic training, the 2622 sequences have 113,000 frames and, thus, the following parameters are assigned for the training to accommodate the dynamism of the version. Each sequence is segmented into overlapping T=16-frame windows (stride S=8) and the 3-channel skeleton input is extended to 6 channels by appending a mid-frame displacement vector $\delta = x_{\lfloor T/2 \rfloor} - x_0$, where $x_{T/2}$ is the mid-frame and x_0 is the initial frame of the sequence, encoding dominant motion direction to distinguish between mirrored gestures.

5 Benchmarking Results and Analysis

5.1 Overall Recognition Performance

Training and Convergence: E-ST-GCN trained on mixed-lighting BiCE-HG data demonstrates stable convergence over 120 epochs for both versions. The static version's accuracy reaches 80.38%, while dynamic version's accuracy reaches 89.24% as shown in (table 3). Due to the motion of the gesture, the dynamic version surpasses the static one in recognising certain gestures with similar skeletons and different motions. In both cases, F1-scores reflect strong class-level performance while FPS on an RTX 3090 confirms real-time deployment feasibility for intelligent-reality applications [10, 25]. The no-attention model performs better on the static partition, where the absence of temporal dynamics limits the utility of temporal attention. Conversely, full attention provides a modest advantage on the dynamic partition by leveraging motion cues to better distinguish trajectory-dependent gesture classes.

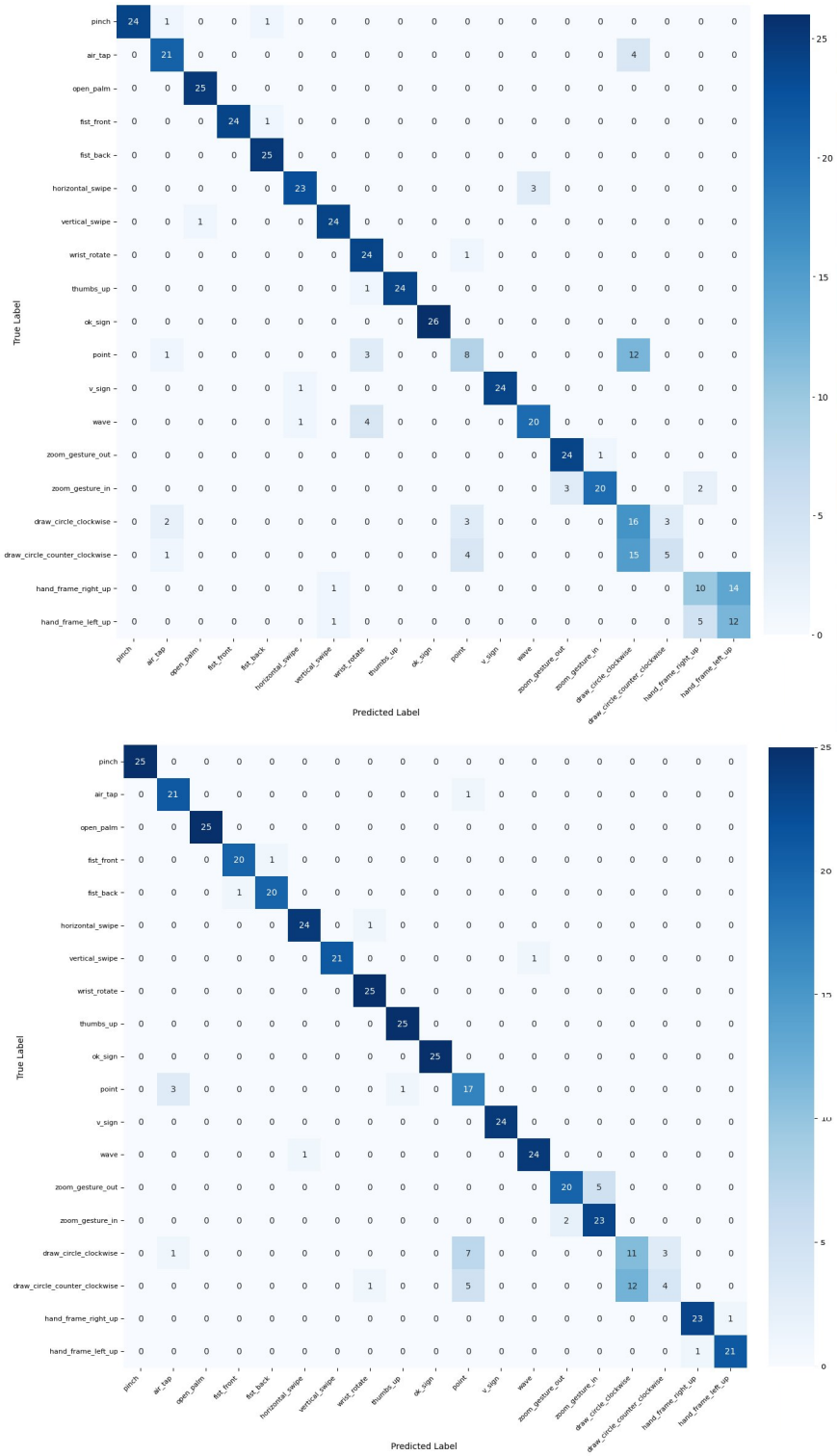

Fig. 4: Confusion matrices for E-ST-GCN on BiCE-HG. (**Top**) Static partition. (**Bottom**) Dynamic partition. The main confusion occurs among *draw circle clockwise*, *draw circle counter clockwise*, and *point*, with counter-clockwise being the most challenging class. Temporal information substantially improves *hand frame right/left up* recognition, increasing accuracy from 40.0%/66.7% (static) to 95.8%/95.5% (dynamic).

5.2 Per-Gesture Performance Analysis

High-Performing Gestures: For both versions, several gestures achieve perfect and near-perfect recognition accuracy. For the static version, pinch, open palm, both fists front and back, vertical swipe, wrist rotate, thumbs up, OK sign, V sign, and zoom out gesture have achieved $> 90\%$ accuracy. The dynamic version, on the other hand, the gesture which have $> 90\%$ accuracy are pinch, open palm, both fists, both swipes, wrist, thumbs up, OK sign, V sign, wave, zoom in gesture, and hand frames. Obviously, the dynamic version has an edge over the static version due to the motion information of the gestures themselves and the displacement function which add more information to the model.
Challenging Gestures: Several gestures exhibit lower recognition accuracy, especially in the static version. In general, the dynamic version has struggled with drawing circle action, both clockwise (50 % accuracy) and anticlockwise (18 % accuracy), confusing both actions themselves along with the point gesture. The static version suffers from the same issue in addition to the mutual confusion for hand frame gestures (the right one (40 % accuracy) and the left one (66 % accuracy)). Confusion matrices for both version are shown in Figure 4.

5.3 Lighting Robustness Analysis

Spatial Illuminance Characteristics. Full lighting configuration exhibits mean illuminance of 310.0 lux (forward) and 326.4 lux (backward), with ranges of 101.8–510.0 lux and 173.0–511.5 lux respectively. The average Coefficient of variation averages 5.29% (forward) and 3.07% (backward), suggesting consistent illuminance measurements across spatial positions. These statistics provide a controlled basis for evaluating recognition performance under varying lighting conditions.
Illuminance Dynamic Range Evaluation: BiCE-HG data achieves 89.24% validation accuracy (dynamic partition, full attention) across the full bi-conditional illuminance range (29–523 lux), demonstrating that skeleton-based recognition remains robust across the spatial illuminance range.

5.4 Computational Performance

E-ST-GCN achieves efficient real-time performance across both static and dynamic settings. As shown in Table 3, the model attains up to 1661 FPS without attention and 1199 FPS with attention under static conditions, and 666 FPS and 614 FPS, respectively, under dynamic conditions. Despite the additional attention modules, the computational overhead remains modest, with model sizes ranging from approximately 214K to 264K parameters. These results demonstrate that E-ST-GCN maintains high efficiency while incorporating attention mechanisms, making it suitable for deployment on resource-constrained intelligent reality devices [10, 25].

Table 4: Comparison of BiCE-HG vs existing datasets

Dataset	View	Subj.	Classes	Lighting	Mobility	Lux points
EgoGesture [27]	Ego	50	83 (stat.+dyn.)	6 scenes, qualitative	sit+stand	–
FPHA [5]	Ego	6	45	Lab only, qualitative	–	–
HoloGesture [20]	Ego	20	10	Lab only, qualitative	–	–
H2O [11]	Ego	4	36	Lab only, qualitative	–	–
HaGRID [9]	3rd	552	18 (stat.)	Varied, qualitative	–	–
BiCE-HG (Ours)	**Ego**	**23**	**19 (stat.+dyn.)**	**Bi-conditional, quantified**	**sit,stand, walk**	**80**

5.5 Comparison with Existing Datasets

Table 4 shows that BiCE-HG is the only egocentric gesture dataset providing both controlled multi-mobility conditions and calibrated illuminance documentation. In general, existing datasets describe lighting qualitatively at best, for example, EgoGesture uses broad scene labels such as "indoor" or "facing a window", while FPHA and HoloGesture offer no lighting characterisation at all. Particularly, none document spatial illuminance variation, making systematic robustness evaluation across illuminance levels impossible. H2O, while influential for hand–object interaction, is confined to a controlled lab setting with four subjects. BiCE-HG directly addresses these gaps by providing quantified bi-conditional lighting across a 4.95× dynamic range with three mobility scenarios, establishing a foundation for reproducible illumination robustness research in intelligent reality gesture recognition.

In contrast to all existing datasets, BiCE-HG introduces quantified lighting variation, which reveals robustness challenges that remain hidden in qualitative or uncontrolled illumination settings. This highlights the need to evaluate gesture recognition models under systematically varied conditions and motivates lighting-aware model design.

6 Conclusions

In this paper, we have presented BiCE-HG, the first egocentric hand gesture dataset with bi-conditional lighting quantification and dense spatial illuminance documentation. BiCE-HG addresses critical gaps in existing egocentric gesture recognition by providing systematic environmental metadata, a 3×2 factorial mobility lighting design, and an intelligent-reality-relevant gesture vocabulary. The dataset includes 2,622 gesture instances from 23 participants performing 19 gestures across six controlled conditions with calibrated illuminance measurements at approximately 80 spatial points spanning 0–450 cm.Comprehensive baseline experiments using spatiotemporal graph convolutional networks establish benchmark performance on BiCE-HG. Enhanced ST-GCN with tri-attention mechanisms achieves 80.38% accuracy with an F1-score of 0.79 on static data and 89.24% accuracy and F1=0.87 on dynamic sequences, while the non-attention

variant yields 81.24% and 88.57%, respectively. These results indicate that attention mechanisms provide benefits under dynamic conditions while maintaining competitive performance under static settings.

Per-gesture analysis shows strong performance for many gestures (e.g., open palm, fists, OK sign: >90%), particularly in the dynamic setting, while motion-sensitive gestures such as point and circular drawing remain challenging (18–50%), along with hand-frame confusion in the static setting (40–66%),underscoring interface design considerations for intelligent reality deployment. Spatial illuminance measurements remain highly stable, with a coefficient of variation below 5% for 85% of measurement conditions, supporting reproducible lighting-robustness evaluation. Looking forward, BiCE-HG's quantified lighting metadata enables research on lighting-robust domain adaptation, including adversarial and self-training approaches. Furthermore, context-aware and meta-learning architectures that adapt to mobility state and illumination conditions offer promising directions for improving real-world robustness. Overall, BiCE-HG provides a foundation for developing next-generation gesture recognition models that operate reliably across diverse and dynamic intelligent-reality environments.

References

1. Baraldi, L., Paci, F., Serra, G., Benini, L., Cucchiara, R.: Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In: CVPRW (2014)
2. Chalasani, T., Ondrej, J., Smolic, A.: Egocentric gesture recognition for head-mounted ar devices. In: IEEE ISMAR-Adjunct (2018)
3. Chalasani, T., Smolic, A.: Simultaneous segmentation and recognition: Towards more accurate ego gesture recognition. In: IEEE/CVF ICCV Worksh. (2019)
4. Dakheel, A., Abhayaratne, C.: BiCE-HG (A bi-conditional egocentric hand gesture dataset under controlled environmental conditions)) (6 2026). `https://doi.org/10.15131/shef.data.32374137`
5. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: CVPR. pp. 409–419 (2018)
6. Garg, G., Hegde, S., Perla, R., Jain, V., Vig, L.: Drawinair: A lightweight gestural interface based on fingertip regression. In: ECCV Workshops. Springer (2018)
7. Hegde, S., Garg, G., Perla, R., Hebbalaguppe, R.: A fingertip gestural user interface without depth data for mixed reality applications. In: IEEE ISMAR-Adjunct (2018)
8. Hegde, S., Perla, R., Hebbalaguppe, R., Hassan, E.: Gestar: Real time gesture interaction for ar with egocentric view. In: IEEE ISMAR-Adjunct (2016)
9. Kapitanov, A., Makhlyarchuk, A., Kvanchiani, K.: Hagrid – hand gesture recognition image dataset. In: IEEE/CVF WACV. pp. 4572–4581 (2024)
10. Karelin, A., Brazhenko, D., Kliukovkin, G., Chernenko, Y.: Real-time hand tracking and collision detection for immersive mixed-reality boxing training on apple vision pro. Sensors **25**(16), 4943 (2025)
11. Kwon, T., Kim, S., Kim, S., Lee, K.: H2O: Two hands manipulating objects for first person interaction recognition. In: ICCV. pp. 10138–10148 (2021)

12. Liang, H., Yuan, J., Thalmann, D., Thalmann, N.: Ar in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In: ACM MM (2015)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
14. Magic Leap: Magic leap 2: Product overview and specs. `https://www.magicleap.com/magic-leap-2` (2022), magic Leap Inc.
15. Memmi, M., Slama, R., Berretti, S.: Optimizing hand gesture recognition with an accurate graph model and mixture of experts for joint relationships and temporal dynamics. IEEE Trans. Biometrics, Behavior, Identity Sci. (2025)
16. Meta: Meta quest pro: Product specifications. `https://www.meta.com/quest/quest-pro/` (2022), meta Platforms
17. Microsoft: HoloLens 2 (2019), mixed reality head-mounted display, released November 7, 2019
18. Mo, G., Dudley, J., Kristensson, P.: Gesture knitter: A hand gesture design tool for head-mounted mixed reality applications. In: ACM CHI (2021)
19. Mohatta, S., Perla, R., Gupta, G., Hassan, E., Hebbalaguppe, R.: Robust hand gestural interaction for smartphone based ar/vr applications. In: IEEE WACV Worksh. (2017)
20. Park, J., Hong, J.: Hologesture: A multimodal dataset for hand gesture recognition robust to hand textures on head-mounted mixed-reality devices. In: ICIP (2024)
21. Reza, S., Zhang, Y., Camps, O., Moghaddam, M.: Towards seamless egocentric hand action recognition in mixed reality. In: IEEE ISMAR-Adjunct (2023)
22. Shen, J., Dudley, J., Mo, G., Kristensson, P.: Gesture spotter: A rapid prototyping tool for key gesture spotting in virtual and augmented reality applications. IEEE TVCG **28**(11), 3868–3878 (2022)
23. Shen, J., Lange, M., Xu, X., Zhou, E., Tan, R.: Towards open-world gesture recognition. In: IEEE ISMAR (2024)
24. Walunj, S., Kumar, A., Singh, R.: Human action recognition using spatiotemporal deep features. In: ICIP. pp. 1–6 (2025)
25. Xie, H., Wang, J., Baitao, S., Gu, J., Li, M.: Le-hgr: A lightweight and efficient rgb-based online gesture recognition network for embedded ar devices. In: IEEE ISMAR-Adjunct (2019)
26. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: MediaPipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214 (2020)
27. Zhang, W.: Unified learning approach for egocentric hand gesture recognition and fingertip detection. PR **121**, 108200 (2022)