



ARTICLE

Fuse3D: A Feature Fusion Framework for Human–Object Interaction Pattern Recognition

Danish Hamid¹, Yazeed Alkhrijah², Syed Muhammad Usman^{3,*}, Muhammad Tahir Akram¹,
Fazal ur Rehman⁴, Shehzad Khalid^{5,6,*}, Sulieman S. Alshuhri⁷ and Amal Alomran⁷

¹Department of Creative Technologies, Faculty of Computing and Artificial Intelligence (FCAI), Air University, Islamabad, Pakistan

²Department of Electrical Engineering, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

³Department of Computer Science, Bahria School of Engineering and Applied Sciences, Bahria University, Islamabad, Pakistan

⁴Department of Electrical and Biomedical Engineering, Khawaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan

⁵Department of Computer Engineering, Bahria School of Engineering and Applied Sciences, Bahria University, Islamabad, Pakistan

⁶Computer and Information Sciences Research Center (CISRC), Imam Mohammad ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

⁷Department of Information Technology, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

*Corresponding Authors: Syed Muhammad Usman. Email: smusman.h11@bahria.edu.pk; Shehzad Khalid.

Email: shehzad@bahria.edu.pk

Received: 30 May 2026; Accepted: 03 September 2026; Published: 21 October 2026

ABSTRACT: Human–Object Interaction (HOI) recognition is essential for intelligent applications such as robotics, assistive healthcare, human–computer interaction, augmented reality, and smart environments. Existing HOI recognition methods often rely on appearance-based visual cues, including object texture, color, scene context, and visual motion. Although these approaches can achieve strong recognition performance, they usually require large annotated datasets, computationally expensive feature extraction networks, and specialized hardware, and their predictions may be affected by illumination changes, occlusion, and visual background variability. To address these issues from a hand-kinematic perspective, this paper proposes Fuse3D, a geometry-enriched machine-learning framework for HOI recognition using 3D hand descriptors extracted from egocentric video streams. The proposed method focuses on structural and dynamic hand-motion characteristics rather than directly modeling object appearance. Hand joints and surface vertices are reconstructed in 3D space using the MANO hand model and transformed into a 7D representation by combining Cartesian coordinates with Euclidean norm and planar distance descriptors. For each descriptor channel and frame, mean pooling across the 21 joints or 778 MANO vertices produces compact global temporal signals. The joint-derived and mesh-derived signals are then summarized using temporal statistics and fused to combine complementary whole-hand kinematic evidence obtained from sparse skeletal landmarks and dense surface sampling. The pooled vertex stream is interpreted as a global mesh-derived kinematic summary; it does not preserve local mesh topology or detailed finger-surface morphology. The resulting feature vectors are classified using a lightweight XGBoost-based framework implemented in both two-stage and single-stage HOI recognition architectures. Experiments conducted on the HOI4D benchmark dataset show that the two-stage fused representation achieves an average F1-score of 81% with millisecond-level inference time under the reported evaluation protocol. The results indicate that geometry-enriched hand kinematics can provide an efficient and interpretable representation for HOI recognition, particularly in settings where lightweight inference is preferred. However, the method remains dependent on RGB-based hand-pose reconstruction and does not fully replace appearance-based or multimodal deep models.

KEYWORDS: Human-object interaction recognition (HOI); 3D hand kinematics; geometric feature learning; egocentric vision; machine learning; feature fusion

1 Introduction

Research on human-object interaction (HOI) focuses on understanding how people interact with objects in their environment, often by examining photos, videos, or additional types of information. Past studies used object recognition to identify individuals along with surrounding items they interacted with. The process also accounts for changes in the position and pose of the body. Such considerations help in further elucidating interactions, for instance, when a person reaches for or holds onto an object.

Through the video modality, temporal changes observed during interaction sequences provide valuable information for understanding human behavior. Human-object interaction (HOI) detection is important for the development of robotic systems and assistive technologies. However, existing HOI recognition approaches still face several challenges that limit their robustness and generalizability. Earlier methods often rely heavily on appearance-based object characteristics, such as color, texture, and size [1–3]. A key limitation of such approaches is their sensitivity to illumination variations, which can alter observed color distributions, texture descriptors, and other local image features even when the underlying object remains unchanged. These appearance variations, together with partial occlusion and changes in object viewpoint or presentation, can reduce feature consistency and adversely affect the robustness of HOI recognition systems.

Deep learning based HOIR methods are associated with some major challenges. These techniques relate to all the supervised learning methods, where building the corresponding large, fully labeled data sets is often very expensive and demanding. According to [4], they are quite tedious to assemble. The limitation of such a model is that the understanding of such a model is not advanced enough to tell about subtle interactions that differentiate holding onto something from actually using it [5]. Thus, such limitations imply that such models lack generalization when faced with a novel or otherwise unexpected context. These are the features relative to the technical inefficiency of requiring extremely advanced computational resources in terms of amounts required, combined operating costs in object detection and action recognition tasks which, besides power, require very specialized hardware resources [6,7]. All this has led to the development of other HOIR systems, with their performance being somewhat compromised for computational efficiency.

Advances in egocentric vision have attracted considerable research interest for human-object interaction (HOI) recognition [8]. In contrast to third-person viewpoints, egocentric cameras capture interactions from the wearer's perspective, providing direct observations of the hands, manipulated objects, and task-relevant regions of the scene. This first-person viewpoint can therefore provide useful cues regarding the user's attention, manipulation patterns, and interaction context. Among these cues, hand information is particularly important because the spatial configuration and temporal motion of the hands are closely related to the way objects are grasped, manipulated, and used. Recent developments in computer vision and machine learning have substantially improved hand pose estimation and tracking, enabling detailed modeling of hand dynamics for interaction understanding [9]. Incorporating hand pose information into HOI recognition can provide complementary geometric and motion cues, particularly when object appearance is partially degraded by occlusion, viewpoint variation, or background clutter. Motivated by these advances, the proposed Fuse3D framework models egocentric human-object interactions using 3D hand-pose information derived from hand joint coordinates and mesh vertex representations. These geometric representations capture both the articulated configuration and fine-grained spatial structure of the hand during object manipulation, providing interaction-specific information that complements conventional visual appearance features. Rather than assuming that existing HOI systems rely primarily on shape, color,

or texture, Fuse3D is positioned as a complementary geometry-driven approach to the broader range of contemporary HOI methods, including deep neural networks, graph-based models, temporal models, transformers, scene-graph representations, and multimodal learning frameworks. By focusing on structured 3D hand information, Fuse3D aims to improve the representation of hand-object manipulation patterns while reducing dependence on object appearance alone.

Advancing egocentric cameras toward human object interaction detection has drawn vast research attention [8]. Egocentric vision contrasts with third person views in placing the human object interaction right before the eyes of the observer, this view plays an important role in the relevant correct representation of human object interactions. It shows what is important and identifies the main subject areas of interest and concentration and hence the observer's priorities and objectives. Availability of information on hand improves the overall HOI recognition. This information, with its finer movements, can give a perfect understanding of interactions. With the advancement of computer vision and machine learning, hand pose tracking is becoming a vital part of object recognition using hand dynamics, and is a field being driven forward by these recent developments [9]. Hand pose information for the HOI recognition systems is added to the knowledge of object handling. This ultimately enhances the performance of the models in situations of occlusion and partial visibility of objects in the background clutter. This last decade has seen the inclusion of both image and video in an acceptable amount of progress in hand pose estimation. On top of this progress, we develop a structure of our hand pose data hoi recognition system which undoubtedly will use the hand actions focus. This learns human object interactions from egocentric video and images, through the use of joint positions and vertex points. In this HOI recognition system, the processing efficiency is improved with the minimum overlaps. Other systems primarily assign shape, colour and texture.

Recently, the dynamics of hand poses has been found to be useful features in egocentric HOI recognition and user-centred extended-reality applications [10–12]. Specifically, 3D hand-pose graphics were used for object manipulation pattern recognition, user identification, and immersive interaction analysis, which further demonstrates the importance of hand kinematics in human behavior recognition and HCI. The present study does not claim to preserve dense mesh morphology after pooling. Instead, it examines whether compact global temporal signals derived separately from sparse joints and dense MANO surface sampling provide complementary evidence for segment-level HOI recognition.

In this work, we propose a novel architectural framework, Fuse3D, for 3D human-object interaction (HOI) recognition. The framework is designed as a compact segment-level classifier rather than a topology-aware hand-mesh model. It converts reconstructed joint and vertex coordinates into global channel-wise motion signals, thereby retaining whole-hand trajectory and distributional dynamics while deliberately sacrificing point identity and local mesh connectivity. This information–compactness trade-off is relevant to HOI recognition because many manipulation classes differ in their global displacement, movement range, velocity changes, timing of extrema, and motion energy over the interaction segment. Fuse3D is designed to characterize the dynamic patterns of hand motion associated with object manipulation and to capture discriminative interaction cues from 3D hand representations. In this study, two complementary HOI recognition strategies are investigated: a two-stage model and a single-stage model. In the two-stage approach, the object is first identified and the corresponding interaction is subsequently classified, whereas the single-stage approach performs interaction recognition directly without explicitly separating object identification from interaction classification. [Section 4](#) presents a comprehensive category-wise evaluation and discussion of both approaches. The experimental results demonstrate that Fuse3D achieves a favorable balance between recognition accuracy and computational efficiency, indicating its potential applicability to interactive systems and assistive technologies. Because Fuse3D performs recognition using compact 3D

hand-derived representations rather than directly relying on full-scene visual appearance during classification, it may reduce the amount of extraneous visual information retained or processed at this stage. However, Fuse3D should not be regarded as inherently privacy-preserving, since visual sensing may still be required for hand-pose reconstruction and the framework does not guarantee the absence or protection of privacy-sensitive visual information.

Key Contributions

Listed below are the key contributions of our paper:

- A geometry-enriched HOI recognition framework is proposed using 3D hand-joint and MANO surface-vertex information extracted from egocentric video streams, reducing direct dependence on object appearance descriptors during classification.
- A deterministic coordinate-derived nonlinear expansion is constructed by combining Cartesian coordinates with Euclidean norm and planar distances. The expansion makes radial and coordinate-plane magnitudes explicitly available to the tree classifier, but it is not claimed to be rotation invariant or to introduce independent geometric degrees of freedom.
- A reproducible mean-pooling and temporal feature-extraction pipeline is defined. For every frame and channel, values are averaged across 21 joints or 778 vertices to produce seven global temporal signals per stream; eighteen time-domain statistics then produce compact 1×126 joint and vertex descriptors.
- Joint-only, vertex-only, and fused descriptors are evaluated to determine whether global kinematic signals derived from sparse skeletal landmarks and dense MANO surface sampling provide complementary evidence for HOI recognition.
- Two-stage and single-stage XGBoost-based HOI recognition architectures are compared on the HOI4D dataset to examine the trade-off between classification performance and inference efficiency.

2 Related Work

In recent years, many learning frameworks have been proposed to enhance human-object interaction (HOI) recognition in various application areas [13,14]. The studies show that the acquisition of learning discriminative representations that reflect the relation between manipulated object and human action is becoming an increasingly important process. Other methods have explored spatial and temporal information to provide more accurate representations of the evolution of interactions through time, especially when viewpoint, object appearance and motion vary over time in complex scenes [15,16]. Later approaches have adopted deep neural networks and multimodal representations, using complementary interaction cues for improving the robustness of recognition, with a growing number of approaches being developed [17,18]. Moreover, a recent line of egocentric HOI literature has focused on the first-person visual information, in which hand motion and object related cues are important cues for fine-grained manipulation activity recognition [19]. All of these changes point towards a move from simple visual features to more informative representations that combine spatial, temporal, geometric and contextual features for HOI recognition. For instance, one of the examples adopts a framework of ontology-based captions of the action of athletes in a sport related format by combining semantic reasoning with the binding of people and objects. The authors of [20,21] developed a hybrid network which integrates graph convolutional networks and a multi layer perceptron to enhance HOI recognition. Likewise, ref. [22] uses Hidden Markov Model (HMM) with precise object localization to realize interaction recognition. Another approach [23] merges object detection with skeletal analysis which is evaluated on several datasets reaching about 98.5% mean accuracy for HOI tasks. This method demonstrates strong predictions and gives a flexible framework that works well on few training data.

The authors of [24] discuss the challenges posed by assembling large scale datasets requiring all human object pairs, acknowledging that such task incurs a great extent of time and computational costs. They propose an HOI detection system that is able to capture new and unseen cases with a two stream framework for visual and textual modality. The stacked transformer for human object interaction further enhances the recognition ability. A pertinent approach by RPNN in [25] uses two graphs: one for human posture and the other for object interaction with body parts. The two interact with each other in an action interpretation module for supporting HOI detection. The procedures suggested in [26] present an auxiliary knowledge graph for a new dataset to recognize objects and surpass existing benchmarks in categorizing objects for new pairings. On similar lines, ref. [27] proposed a neural network based method for the recognition of human movements based on the consecutive frames in a video to predict actions that are occluded, it has a specific module for estimating poses under strong occlusion. Besides that, ref. [28] applies scene graphs to capture the semantic relations in an image dataset, which provide contextual information to infer human object interactions.

Hand-centric egocentric recognition has recently gained attention because hands are directly involved in object manipulation and often provide strong behavioral cues in first-person video. Hamid et al. [11] used 3D hand-pose data for HOI recognition and user identification in extended-reality settings, while Hamza et al. [12] investigated person classification from egocentric HOIs using 3D hand-pose information. Similarly, Hamid et al. [10] demonstrated that egocentric wearable-camera data can support HOI recognition by exploiting hand-joint information. The results of these studies justify the use of hand kinematics as a light-weight alternative or complement to appearance-based representations.

Recent deep HOI models have also investigated the transformer based and multimodal fusion methods. For instance, TransFlow-HOI combines RGB appearance and optical-flow information with multimodal fusion using transformer networks to achieve high performance in HOI recognition tasks [29]. These models are useful for capturing long-range temporal dependency and appearance-motion relationships, though they often involve more complicated visual feature extraction models. The present work instead studies a compact representation in which global temporal signals are derived from reconstructed 3D joints and from the complete MANO vertex set. The vertex stream is mesh-derived but not topology preserving; its role is to summarize the motion of the reconstructed hand envelope for lightweight segment-level classification.

Much of the current work on HOI is based on image and video data that has been visually extracted and analyzed, and a majority of these works use deep neural networks to capture spatial, appearance and temporal interaction features. These approaches have been successful in achieving good recognition results, but they can be computationally intensive, especially in cases of high-dimensional visual features or complex network structures. Additionally, much of the previous HOI research has been created and tested from an exocentric or third person perspective. In contrast, egocentric vision is a first-person view where the hands and manipulated objects are often prominent parts of the visual field, and is well suited to the analysis of fine-grained object manipulation and user-centered interactions [30,31]. In the field of application, egocentric sensing has attracted interests in several areas including augmented reality, virtual reality, wearable computing and human-robot interaction, where understanding of hand-object manipulation is crucial to better understand the activity and intentions of the user. However, recognition from egocentric data remains challenging because of frequent hand-object occlusions, rapid viewpoint changes, background clutter, and substantial variations in object appearance. These challenges highlight the need for HOI representations that exploit structured hand-motion and geometric information while limiting unnecessary dependence on computationally intensive full-scene visual processing.

3 Proposed Method

We propose both two stage and single stage HOI recognition methods. Figs. 1 and 2 show two stage and single stage methods proposed in this research respectively. Fig. 3 presents the detailed flow diagram of the proposed method.

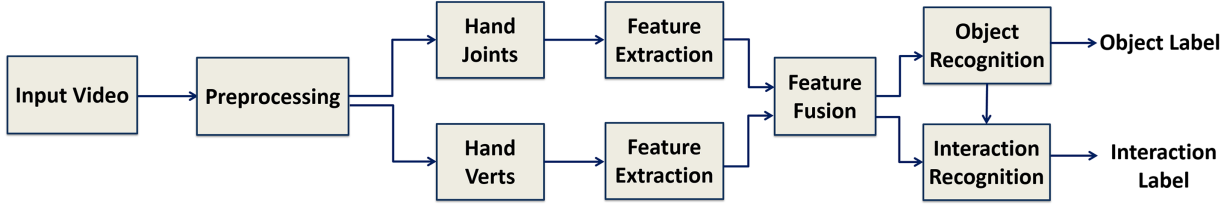


Figure 1: Fuse3D: two stage HOI recognition.

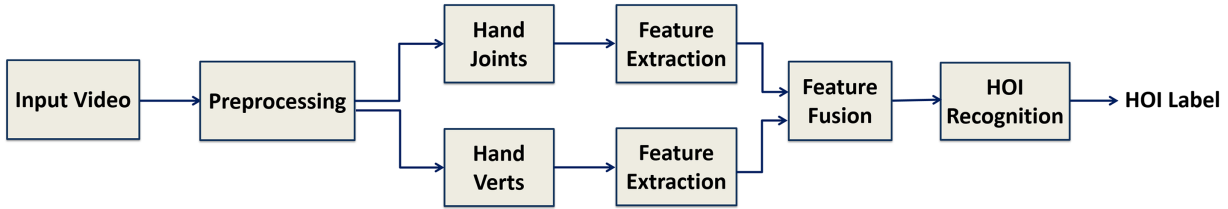


Figure 2: Fuse3D: single stage HOI recognition.

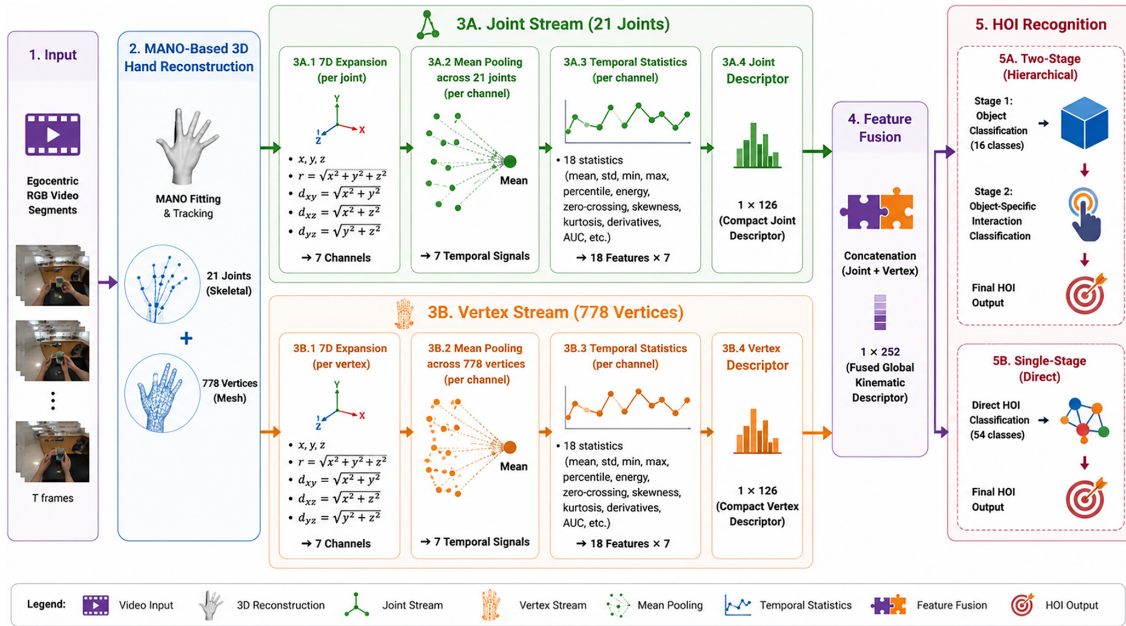


Figure 3: Overview of the proposed Fuse3D framework. Egocentric RGB video segments are first processed using MANO-based 3D hand reconstruction to estimate 21 skeletal joints and 778 mesh vertices. Two parallel processing streams are then constructed. Each stream performs coordinate-derived 7D feature expansion, channel-wise mean pooling across anatomical points, and temporal statistical summarization to generate compact 1×126 descriptors. The joint-derived and mesh-derived global kinematic summaries are concatenated to form a fused 1×252 descriptor, which is subsequently used by either a hierarchical two-stage HOI recognition framework or a direct single-stage HOI recognition framework.

3.1 Data Acquisition

We present a new method Fuse3D for HOI recognition that has been applied on an egocentric dataset with a first-person view. There are a number of datasets available for 3 person view, but very few comprehensive datasets. There are many 2D datasets which have been converted synthetically into 3D hand representations [32], whereas, HOI4D dataset consists of annotations from real images of hands [33] which aligns with our objective of this research. Four different cameras have been used to record videos in the present data set in an extensive indoor setting. It contains around 2.4 million RGB-D frames recorded from an egocentric perspective, 4000 interaction sequences by nine different participants and 16 different object categories. The data set contains substantial variability in the conditions under which the data were acquired: room layout, furnishings, participants, viewpoints, backgrounds and complexity of interactions. Such diversity provides challenging scenarios for HOI recognition, including variations in environmental configuration, object manipulation patterns, occlusion, and interaction complexity. The availability of both RGB and depth information further enables the analysis of appearance and three-dimensional geometric cues associated with human-object interactions.

3.2 Segmentation and Handpose Extraction

HOI recognition method involves preprocessing to video sequencing before feature extraction. In the first step, an optimal non overlapping window of five seconds has been selected empirically and applied on videos to divide them into segments. After segmentation, each window is annotated using the video clips of original recordings. Hand pose information is then extracted from each segmented frame of 2D RGB array. Hand joint and vertex information extracted is then aggregated. 2D hand pose keypoints and landmarks are extracted and passed to regressor that estimates initial shape and pose parameters which are projected into 3D space using the MANO parametric hand model [34]. These parameters are refined to reduce the reprojection error which is then used in MANO framework to get the consistent 3D coordinates for hand joints and vertices. This 3D hand joint array has a dimension of 21×3 and 3D hand vertex has dimension of 778×3 . Fig. 4 shows a sample visualization of the 3D hand joint and vertex structure derived from the MANO framework.

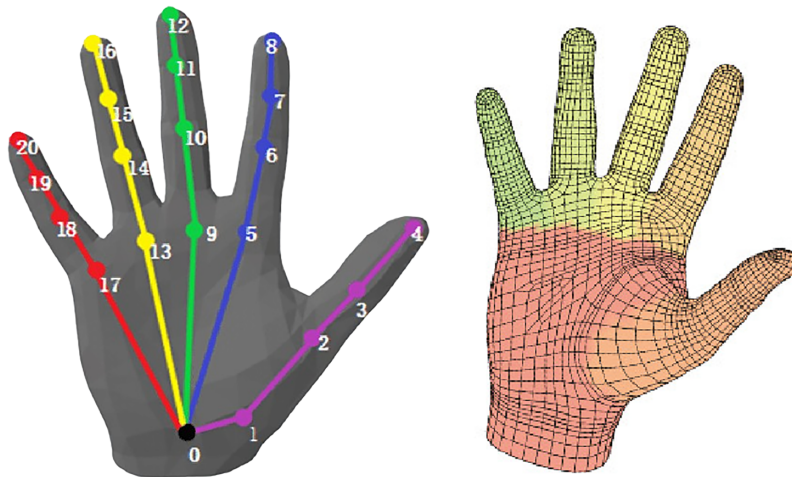


Figure 4: Representation of the MANO 3D hand joints and vertices model.

3.3 Hand Joint and Vertex Transformation

The reconstructed 3D hand representations consist of the 778 mesh vertices provided by the MANO model together with a 21-keypoint hand representation. Standard MANO defines 16 skeletal joints; therefore, following the commonly adopted MANO-based 21-keypoint convention, the 16 MANO joints are augmented with five fingertip landmarks obtained from corresponding mesh vertices, yielding a total of 21 hand keypoints [35]. Each 3D joint or vertex coordinate triplet (x, y, z) is subsequently transformed into a geometry-enriched 7D representation by appending four distance-based measures: the Euclidean norm $\sqrt{x^2 + y^2 + z^2}$ and the three planar distances $\sqrt{x^2 + y^2}$, $\sqrt{x^2 + z^2}$, and $\sqrt{y^2 + z^2}$. Accordingly, the transformed hand-keypoint representation for each frame has a dimension of 21×7 , while the transformed mesh-vertex representation has a dimension of 778×7 .

For a point $\mathbf{p} = (x, y, z)$, the expanded descriptor is

$$\mathbf{q} = [x, y, z, r, d_{xy}, d_{xz}, d_{yz}], \quad (1)$$

where $r = \sqrt{x^2 + y^2 + z^2}$, $d_{xy} = \sqrt{x^2 + y^2}$, $d_{xz} = \sqrt{x^2 + z^2}$, and $d_{yz} = \sqrt{y^2 + z^2}$. These appended terms are deterministic nonlinear functions of (x, y, z) ; they do not add independent geometric degrees of freedom. In particular,

$$d_{xy}^2 + d_{xz}^2 + d_{yz}^2 = 2r^2. \quad (2)$$

The representation is not generally rotation invariant because the Cartesian coordinates and planar distances change under arbitrary 3D rotations. It is also not generally translation invariant unless the coordinates are first expressed relative to a selected origin. Its purpose is therefore not invariance. Rather, the expansion exposes radial and coordinate-plane magnitudes directly to the axis-aligned tree classifier, which would otherwise need multiple partitions to approximate these nonlinear relations from the Cartesian coordinates alone. Throughout the revised manuscript, “geometry-enriched” is used only in this restricted sense of a coordinate-derived nonlinear feature expansion.

Each descriptor channel is min–max normalized to range $[0, 1]$ before classification based on the min–max of the training split of each fold, and applied to the validation/test split of each fold. The normalization is done fold-wise, thus preventing the leakage of information from test data and maintaining the same numerical scale of the Cartesian coordinates, Euclidean norms and planar distances. Table 1 shows the mathematical expressions of all the time-domain handcrafted features extracted in this research.

3.4 Feature Extraction

Temporal statistical descriptors are extracted from the transformed 7D hand-joint and hand-vertex sequences to summarize motion dynamics over each five-second video segment as shown in Fig. 5. For each frame, the flattened transformed joint descriptor has dimension $1 \times (21 \times 7) = 1 \times 147$, whereas the flattened transformed vertex descriptor has dimension $1 \times (778 \times 7) = 1 \times 5446$. These are frame-level spatial descriptors before temporal aggregation.

Table 1: Extracted time-domain features and their mathematical descriptions.

Feature	Mathematical Description
Total Squared Amplitude	$E_x = \sum_{i=1}^M x(i)^2$
Signal Minimum	$x_{\min} = \min(x)$
Average Absolute First-Order Difference	$D_1 = \frac{1}{M} \sum_{i=1}^{M-1} x(i+1) - x(i) $
Maximum-Value Sample Index	$\tau_{\max} = \arg \max_i \{x(i)\}$
Skewness	$m_2 = \frac{1}{M} \sum_{i=1}^M (x(i) - \mu_x)^2, \quad m_3 = \frac{1}{M} \sum_{i=1}^M (x(i) - \mu_x)^3, \quad S_x = \begin{cases} m_3/m_2^{3/2}, & m_2 > 0, \\ 0, & m_2 = 0. \end{cases}$
Full Amplitude Range	$P_{pp} = x_{\max} - x_{\min}$
Mean Squared Amplitude	$\tilde{E}_x = \frac{E_x}{M}$
Sample Average	$\mu_x = \frac{1}{M} \sum_{i=1}^M x(i)$
Normalized Absolute Second-Order Difference	$\tilde{D}_2 = \frac{1}{M} \sum_{i=2}^{M-1} \left \frac{x(i+1) - 2x(i) + x(i-1)}{\max x } \right $
Signal Maximum	$x_{\max} = \max(x)$
Extrema-to-Extrema Gradient	$S_{pp} = \frac{P_{pp}}{T_{pp}}$
Fourth-Moment Peakedness	$K_x = \frac{E[(x(i) - \mu_x)^4]}{(E[(x(i) - \mu_x)^2])^2}$
Extrema Latency Ratio	$L_{abs} = \frac{\tau_{\min}}{\tau_{\max}}$
Root-Mean-Square Deviation	$\sigma_x = \sqrt{\frac{1}{M} \sum_{i=1}^M (x(i) - \mu_x)^2}$
Extrema Time Separation	$T_{pp} = t(x_{\max}) - t(x_{\min})$
Normalized Absolute First-Order Difference	$\tilde{D}_1 = \frac{1}{M} \sum_{i=1}^{M-1} \left \frac{x(i+1) - x(i)}{\max x } \right $
Minimum-Value Sample Index	$\tau_{\min} = \arg \min_i \{x(i)\}$
Average Absolute Second-Order Difference	$D_2 = \frac{1}{M} \sum_{i=2}^{M-1} x(i+1) - 2x(i) + x(i-1) $

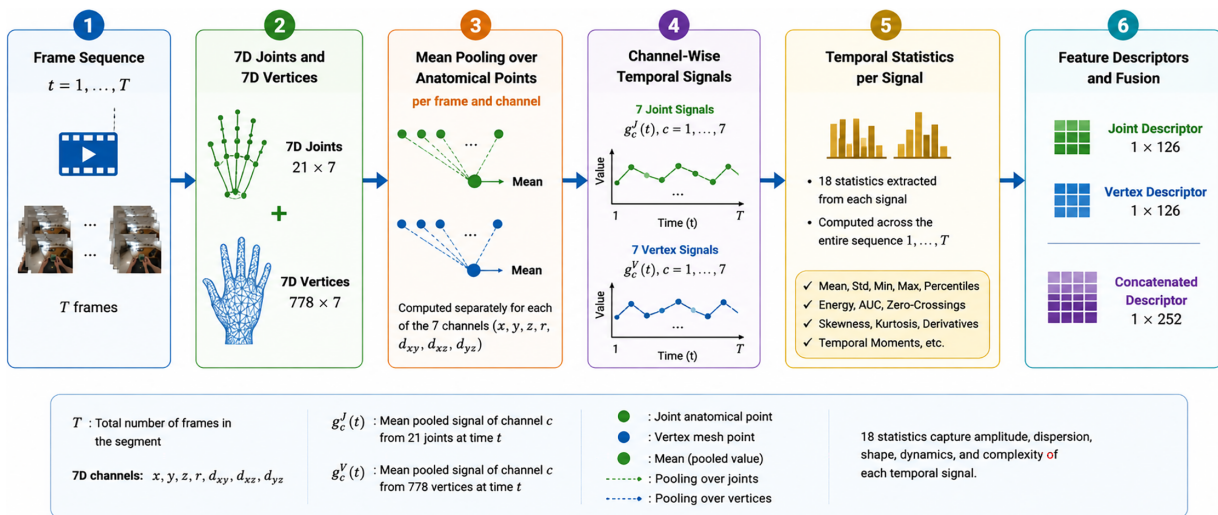
**Figure 5:** Feature-extraction sequence of the proposed Fuse3D framework. For each frame, the reconstructed hand joints and MANO vertices are converted into seven coordinate-derived descriptor channels. Channel-wise mean (Continued)

Figure 5: (continued) pooling is then performed separately across the 21 joints and 778 vertices, producing seven joint-derived temporal signals and seven mesh-derived temporal signals. Eighteen temporal statistical descriptors are computed from each signal across the complete frame sequence, producing a 1×126 joint descriptor and a 1×126 vertex descriptor. The two descriptors are concatenated to obtain the final 1×252 fused feature representation.

Let $q_{t,j,c}^J$ and $q_{t,v,c}^V$ denote descriptor channel $c \in \{1, \dots, 7\}$ at frame t for joint $j \in \{1, \dots, 21\}$ and vertex $v \in \{1, \dots, 778\}$, respectively. The implemented anatomical-point aggregation is mean pooling. The channel-wise signals are therefore

$$g_c^J(t) = \frac{1}{21} \sum_{j=1}^{21} q_{t,j,c}^J, \quad g_c^V(t) = \frac{1}{778} \sum_{v=1}^{778} q_{t,v,c}^V. \quad (3)$$

This operation produces seven temporal signals for the joint stream and seven temporal signals for the vertex stream. Eighteen temporal statistics are computed independently from each signal. Consequently, the joint descriptor and vertex descriptor each have dimension $1 \times (7 \times 18) = 1 \times 126$, and their concatenation has dimension 1×252 .

Mean pooling removes joint or vertex identity and does not preserve local anatomical relations, MANO mesh connectivity, detailed finger articulation, or dense surface morphology. The vertex stream is therefore interpreted as a global mesh-derived kinematic summary rather than a topology-preserving morphological descriptor. Nevertheless, it remains scientifically relevant to segment-level HOI recognition. The pooled Cartesian channels describe whole-hand displacement over time, while the pooled norm and planar-distance channels summarize the average radial and coordinate-plane distribution of the complete reconstructed mesh. For example,

$$\bar{r}(t) = \frac{1}{778} \sum_{v=1}^{778} \|\mathbf{p}_v(t)\|_2, \quad (4)$$

and, in general, $\frac{1}{778} \sum_v \|\mathbf{p}_v\|_2 \neq \left\| \frac{1}{778} \sum_v \mathbf{p}_v \right\|_2$; hence the nonlinear pooled channels are not reducible to the Cartesian centroid alone. The temporal statistics of the five-second segment carry the information of magnitude, range, variability, energy, first and second order change, extrema timing, the skewness and peakedness of these global signals. These descriptors can distinguish interactions in terms of whole-hand trajectory, extent of movement, temporal phase and even motion energy, but don't capture local mesh structure.

The spatial support of the joint and vertex streams is also different: the joint signal is an average of 21 sparse anatomical landmarks, while the vertex signal is an average of 778 surface samples over the palm and fingers. Their fusion is then assessed by the integration of a global skeletal derived and mesh derived kinematic evidence, rather than by evidence of the maintenance of dense morphology. Another rationale that is non-experimentally proved to be robust is that mean pooling serves as a low order summarization that highlights coherent whole-hand movement and can flatten out fluctuations at the point level.

The selected temporal descriptors capture complementary properties of the hand-motion signal, including amplitude range, central tendency, dispersion, distribution shape, latency, temporal variation, and signal energy. They were selected because they provide a lightweight summary of dynamic hand behavior without requiring sequence-learning models or computationally expensive visual feature extraction.

3.5 Classification Task

We propose two approaches of classification for HOI recognition in our Fuse3D framework i.e.; Two-Stage HOI Recognition and Single-Stage HOI Recognition. Two stage methodology consists of two phases. In the first step, object classification is done with the help of the fused feature vector which is formed by combining characteristics of hand joints and vertices. HOI label is assigned to the detected object. For instance, the proposed framework identifies the object as “Bottle”, whereas, the interaction “pick and place” is detected after that. This two stage method uses a hierarchical approach in which first 16 objects are classified followed by the interaction with these objects in the second stage. On the other hand, in the Single stage method, the proposed algorithm detects the interaction along with objects directly. For instance, in single stage method, “Pick and Place Bottle” is directly detected. In this approach, 54 distinct classes are detected for HOI recognition. Classification is performed using XGBoost for all objects along with interactions. In the two stage approach, the proposed method trains 17 classifiers including one for object recognition and other sixteen for recognition of interactions with the object. However, in case of single stage approach, only one classifier is trained for HOI recognition that include objects along with interaction classification.

XGBoost was selected because it is well suited to structured tabular descriptors, supports nonlinear decision boundaries, and provides efficient inference for compact handcrafted feature vectors. In this study, the classifier was trained using the default XGBoost configuration for multi-class classification, including a tree-based booster, 100 estimators, maximum tree depth of 6, and learning rate of 0.3. The multi-class soft-probability objective was used, and the final class label was assigned using the maximum predicted probability. These settings provide a standard baseline configuration for evaluating the discriminative capability of the proposed descriptors without extensive hyperparameter tuning.

4 Results Assessment and Discussion

We evaluate two HOI recognition strategies: a two-stage hierarchical approach and a single-stage direct classification approach. Two stage approach: Every video segment has its corresponding object and interaction labels. The object category is first classified and then the corresponding interaction is recognized in the predicted object category. The single-stage approach is different in that it considers each object–interaction pair as a separate HOI class and predicts the combined class in one classification step. In the single-stage recognition setting 54 different HOI classes derived from 16 object categories are considered using this formulation. Several experimental set-ups were explored for evaluating hand-joint and hand-mesh-vertex discriminative contribution. Three feature-fusion strategies investigated—mean pooling, max pooling, min pooling, and feature concatenation—are explored. No fusion operator is found to be universally superior, as the best performing fusion operator is different for each experiment and interaction class. A desktop computer with an Intel(R) Core(TM) processor and 8 GB of memory was used for all experiments, with the exception of the object specific interaction-recognition experiment, in which the average F1 score was obtained. The performance of the models was assessed based on the accuracy, precision, recall, and F1-score. The average inference time per sample was evaluated to check the practical feasibility of the proposed framework for time-sensitive applications. The dataset was evaluated using 5-fold cross-validation, with normalization parameters estimated exclusively from the training partition of each fold and subsequently applied to the corresponding test partition. Because the present experiments follow this 5-fold protocol, a fully subject-independent evaluation remains an important direction for future work to further assess cross-subject generalization.

4.1 Analysis of Two Stage HOI Recognition

In the two-stage HOI recognition framework, the first stage classifies the object category, while the second stage identifies the interaction associated with the classified object.

4.1.1 Stage 1: Analysis of Object Recognition

In the first stage, the fused feature vector is passed to the XGBoost classifier to classify the 16 object categories. Table 2 presents the object-classification performance in terms of F1-score for the individual hand-joint and hand-vertex representations and the different feature-fusion strategies. Considering the concatenation-based fusion results, strong classification performance was obtained for several object categories, including Stapler (0.81), Scissors (0.87), Laptop (0.85), Kettle (0.89), Chair (0.90), Bowl (0.90), and Bottle (0.89). Among these categories, Chair and Bowl achieved the highest F1-score of 0.90 under concatenation fusion. In contrast, comparatively lower F1-scores were observed for Safe (0.66), Toy Car (0.64), and Furniture (0.66), indicating that these categories were more challenging to distinguish using the concatenated representation. Overall, concatenation fusion achieved an average F1-score of 0.79 across the 16 object categories. The results also demonstrate that the optimal fusion strategy varies across object classes; for example, minimum fusion produced the highest overall average F1-score of 0.80, while individual categories attained their best performance under different fusion operators.

Table 2: Object recognition outcomes across categories (C).

Object	Hand Joints F1	Hand Vertices F1	Fusion Concat F1	Fusion Mean F1	Fusion Max F1	Fusion Min F1	Fusion Sum F1
C1:Stapler	0.70	0.68	0.81	0.77	0.81	0.81	0.80
C2:Lamp	0.68	0.68	0.78	0.72	0.79	0.81	0.83
C3:Scissors	0.81	0.84	0.87	0.89	0.87	0.86	0.84
C4:safe	0.59	0.60	0.66	0.72	0.77	0.75	0.77
C5:Mug	0.74	0.77	0.77	0.80	0.75	0.73	0.84
C6:Pliers	0.65	0.65	0.75	0.71	0.71	0.73	0.69
C7:Laptop	0.81	0.80	0.85	0.87	0.88	0.91	0.92
C8:Toy Car	0.60	0.58	0.64	0.60	0.66	0.67	0.69
C9:Furniture	0.70	0.68	0.66	0.58	0.79	0.85	0.79
C10:Trash Can	0.70	0.71	0.77	0.77	0.71	0.77	0.73
C11:Kettle	0.78	0.79	0.89	0.88	0.88	0.86	0.80
C12:Chair	0.89	0.88	0.90	0.85	0.92	0.89	0.94
C13:Bucket	0.69	0.70	0.76	0.77	0.70	0.71	0.72
C14:Bowl	0.77	0.76	0.90	0.93	0.79	0.79	0.79
C15:Bottle	0.87	0.88	0.89	0.83	0.82	0.83	0.79
C16:Knife	0.67	0.68	0.77	0.76	0.81	0.80	0.78
Average	0.73	0.73	0.79	0.78	0.79	0.80	0.79

Among the evaluated fusion strategies, concatenation preserves the joint-derived and mesh-derived descriptors as separate global kinematic sources. Mean, sum, maximum, and minimum fusion combine corresponding dimensions and may suppress stream-specific variation. Table 2 does not show universal

superiority of concatenation: its average object-recognition F1-score is 0.79, while minimum fusion reaches 0.80. Concatenation is nevertheless competitive and improves over either individual stream; it also obtains the highest average F1-score in the reported object-specific interaction experiment (Table 3). These findings support complementarity between the two global streams, but they do not demonstrate preservation of dense surface morphology.

Table 3: Experimental results for interaction recognition (where ‘C’ represents object category and ‘T’ represents corresponding task).

Objects	Interactions	Hand Joints F1	Hand Vertex F1	Fusion Concat F1	Fusion Sum F1	Fusion Mean F1	Fusion Max F1	Fusion Min F1
C:Stapler	T:Bindthepaper	0.89	0.89	0.96	0.91	0.92	0.93	0.94
	T:Pickandplace	0.85	0.86	0.95	0.89	0.91	0.92	0.93
C:Toy Car	T:Pickandplace	0.80	0.75	0.86	0.84	0.82	0.79	0.83
	T:Push toy car	0.63	0.62	0.68	0.63	0.68	0.63	0.73
	T:Putitinthedrawer	0.47	0.45	0.63	0.67	0.45	0.43	0.48
	T:Takeitoutofthedrawer	0.74	0.69	0.77	0.77	0.62	0.67	0.74
C:Scissor	T:CutSomething	0.95	0.95	0.99	0.91	0.94	0.92	0.94
	T:Pickandplace	0.83	0.86	0.97	0.71	0.85	0.82	0.85
C:Safe	T:Openandclosethedoor	0.75	0.76	0.87	0.77	0.72	0.71	0.72
	T:Putsomethinginit	0.66	0.65	0.80	0.65	0.71	0.68	0.68
	T:Takesomethingoutofit	0.54	0.53	0.74	0.56	0.59	0.67	0.61
C:Mug	T:Fillwithwaterbyakettle	0.60	0.56	0.59	0.56	0.96	0.83	0.77
	T:Pickandplace	0.76	0.74	0.70	0.75	0.87	0.84	0.85
	T:Pickandplace(withwater)	0.67	0.63	0.47	0.45	0.40	0.40	0.47
	T:Pourwaterintoanothermug	0.75	0.71	0.68	0.85	0.89	0.81	0.85
	T:Putitinthedrawer	0.50	0.52	0.40	0.45	0.54	0.46	0.57
	T:Takeitoutofthedrawer	0.82	0.79	0.82	0.83	0.79	0.64	0.85
C:Pliers	T:Clampsomething	0.85	0.86	0.84	0.83	0.82	0.86	0.74
	T:Pickandplace	0.88	0.83	0.91	0.88	0.77	0.76	0.77
	T:Putitinthedrawer	0.52	0.48	0.50	0.48	0.50	0.68	0.52
	T:Takeitoutofthedrawer	0.70	0.66	0.76	0.74	0.73	0.67	0.63
C:Laptop	T:Openandclosethedisplay	0.92	0.90	0.93	0.95	0.95	0.94	0.96
	T:Pickandplace	0.87	0.83	0.88	0.92	0.90	0.89	0.93
C:Lamp	T:Pickandplace	0.90	0.83	0.95	0.97	0.86	0.89	0.86
	T:Turnandfold	0.97	0.95	1.00	1.00	1.00	1.00	1.00
	T:Turnonandturnoff	0.89	0.87	0.94	0.97	0.81	0.84	0.80
C:Furniture	T:Openandclosethedoor	0.81	0.79	0.78	0.79	0.76	0.77	0.63
	T:Openandclosethedrawer	0.85	0.82	1.00	0.89	0.82	0.87	0.79
	T:Putthedrinkinthedoor	0.73	0.66	0.66	0.69	0.83	0.84	0.74
	T:Putthedrinkinthedrawer	0.76	0.70	0.81	0.73	0.81	0.83	0.81
C:Kettle	T:Pickandplace	0.96	0.96	0.94	0.92	0.96	0.97	0.95
	T:Pourwaterintoamug	0.95	0.96	0.93	0.91	0.96	0.97	0.95
C:Chair	T:Pickandplacetoeanewposition	0.54	0.52	0.51	0.45	0.57	0.53	0.45
	T:Pickandplacetoooriginalposition	0.54	0.48	0.54	0.44	0.61	0.64	0.51
C:Bucket	T:Pickandplace	0.77	0.77	0.90	0.87	0.86	0.88	0.89
	T:Pourwaterintoanotherbucket	0.84	0.84	0.93	0.91	0.89	0.91	0.92

(Continued)

Table 3 (continued)

Objects	Interactions	Hand Joints F1	Hand Vertex F1	Fusion Concat F1	Fusion Sum F1	Fusion Mean F1	Fusion Max F1	Fusion Min F1
C:Bowl	T:Pickandplace	0.86	0.85	0.91	0.90	0.88	0.83	0.91
	T:Pickandplace(withball)	0.83	0.83	0.71	0.71	1.00	1.00	1.00
	T:Putitinthedrawer	0.83	0.79	0.91	0.88	0.83	0.81	0.84
	T:Puttheballinthebowl	0.77	0.69	0.73	0.85	1.00	0.95	0.84
	T:Takeitoutofthedrawer	0.71	0.68	0.77	0.72	0.78	0.70	0.82
	T:Take the ball out of the bowl	0.76	0.82	0.78	0.82	0.92	0.84	0.90
C:Bottle	T:Pickandplace	0.72	0.72	0.84	0.88	0.63	0.62	0.61
	T:Pickandplace(withwater)	0.61	0.67	0.84	0.57	0.86	0.81	0.74
	T:Pourallthewaterintoamug	0.37	0.41	0.47	0.42	0.62	0.44	0.57
	T:Putitinthedrawer	0.45	0.43	0.61	0.68	0.49	0.40	0.56
	T:Repositionthebottle	0.75	0.73	0.82	0.78	0.79	0.81	0.81
	T:Takeitoutofthedrawer	0.58	0.65	0.73	0.69	0.68	0.77	0.76
C:Knife	T:Cutapple	0.86	0.85	0.92	0.90	0.92	0.93	0.93
	T:Pickandplace	0.93	0.92	0.96	0.94	0.91	0.93	0.93
	T:Putitinthedrawer	0.62	0.63	0.65	0.65	0.52	0.60	0.56
	T:Takeitoutofthedrawer	0.64	0.67	0.83	0.70	0.67	0.81	0.69
Average		0.77	0.75	0.82	0.78	0.80	0.80	0.79

The results indicate that some object categories are more difficult to classify because their associated hand-motion patterns may exhibit similarities across different interactions. Fig. 6 presents the confusion matrix obtained using the hand-joint feature set. Consistent with the quantitative results reported in Table 2, Bottle achieved a relatively high F1-score of 0.87, whereas Toy Car and Bucket showed lower performance, with F1-scores of 0.60 and 0.69, respectively. The Mug and Bowl categories achieved F1-scores of 0.74 and 0.77, respectively, using the hand-joint representation. Fig. 7 presents the confusion matrix obtained using the hand-vertex feature set. In this representation, Bottle and Mug perform well with F1-scores of 0.88 and 0.77, respectively, while Safe is still relatively hard to classify with an F1-score of 0.60. The results indicate generally that hand-joint and hand-vertex representations are discriminative in different ways depending on the object categories.

The confusion matrix for concatenation-based fusion feature set is shown in Fig. 8. The relatively high diagonal responses suggest good discriminative performance for several object categories such as Bottle, Bowl and Mug. Additionally, the confusion matrix indicates that some categories are misclassified with others, such as Bowl and Bucket, implying that the hand-motion and geometric representation of these two have some commonalities. The Furniture category is also more difficult to classify, with its features being less distinguishing when represented as an interaction. As a whole, the concatenation-based fusion configuration gets an average F1 score of 0.79 in the object classification task. It is important to note that this setup is not always better than the other fusion strategies as minimum fusion (MF) results in a slightly better average F1 score (0.80) as reported in Table 2. These results indicate that the effectiveness of a fusion strategy varies across object categories and that no single fusion operator consistently provides the best performance.

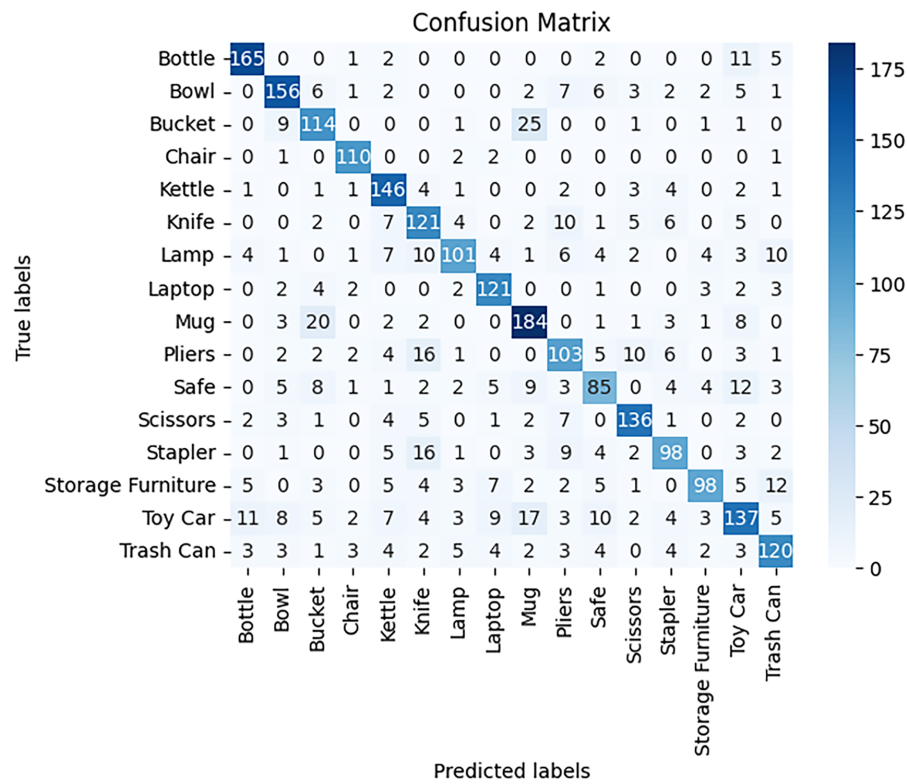


Figure 6: Hand-joint-based object recognition: confusion matrix.

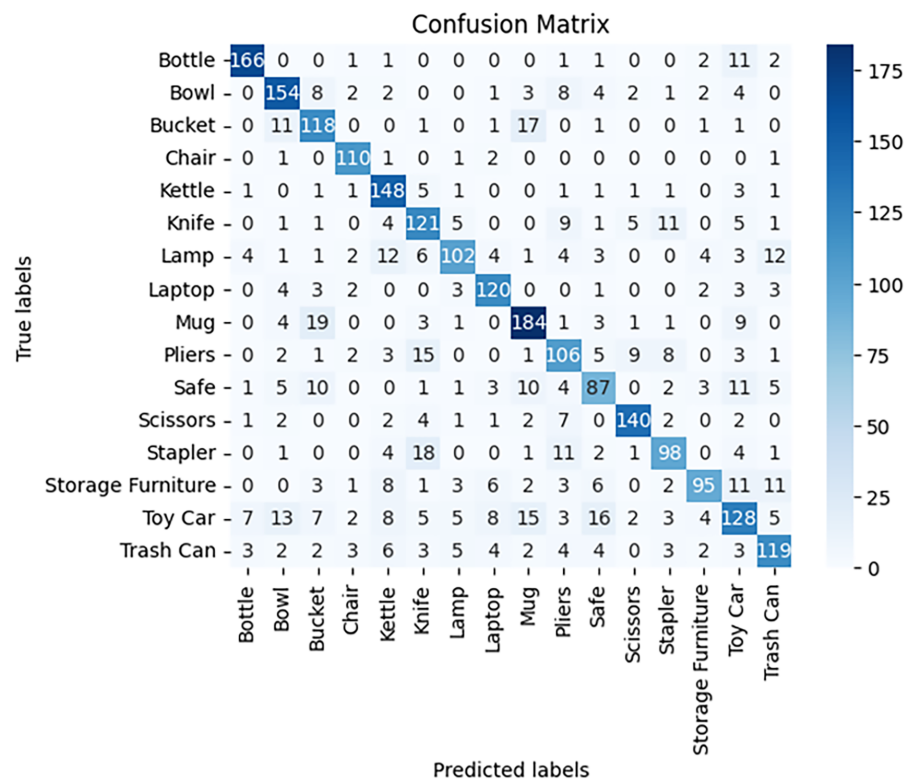


Figure 7: Confusion matrix for object recognition (hand vertex).

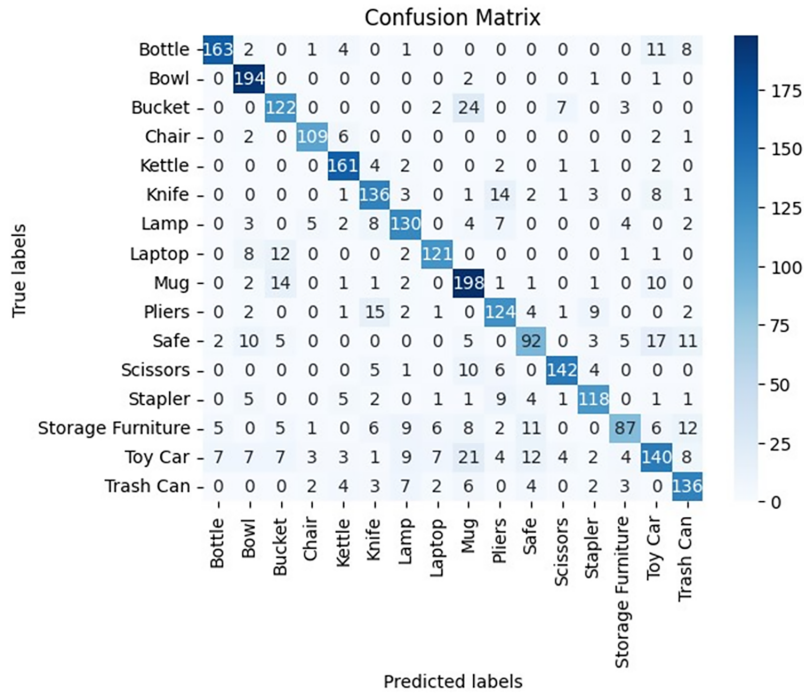


Figure 8: Confusion matrix for object recognition (fusion concatenation).

4.1.2 Stage 2: Analysis of Interaction Recognition

In the first step, objects are detected, whereas, in the next step, interactions between objects are classified. Object label detected from the stage 1 along with the feature vector is then passed to XGBoost classifier for HOI recognition. Table 3 presents the results obtained through experimental setups for object specific interaction recognition. These findings conclude that the classification accuracy for interactions involving “Bucket”, “Kettle”, “Knife”, “Laptop”, “Lamp”, “Scissors”, and “Stapler” categories was high. Specifically, the “Pick and place” interaction achieved high average F1-score across multiple object categories with 91% F1 score on “Bowl”, 90% for “Bucket”, whereas, “Kettle” achieved 94%, “Knife” 96%, and “Lamp” achieved 95%.

Certain interaction classes exhibit comparatively low recognition performance under specific object and feature configurations. For example, for the Mug category, the interaction “Put it into the drawer” achieves an F1-score of 0.40 using concatenation-based fusion, as reported in Table 3. On the other hand, this interaction is not uniform in its performance, ranging between 0.40 and 0.57 for the Mug category in F1-scores. Furthermore, the same interaction demonstrates different performance when coupled with different object categories, suggesting that the difficulty in recognizing the interaction depends not only on the interaction-object combination, but also on the feature representation selected. The differences observed indicate that similar or temporal overlapping hand-movement patterns might be more challenging to differentiate with the extracted geometric features. Within the two-stage approach, errors from the first object-classification stage can also be carried over to the second interaction-classification stage, which can have an impact on the final prediction of the HOI. Fig. 9 presents the average F1-score obtained across the interactions associated with each object category.

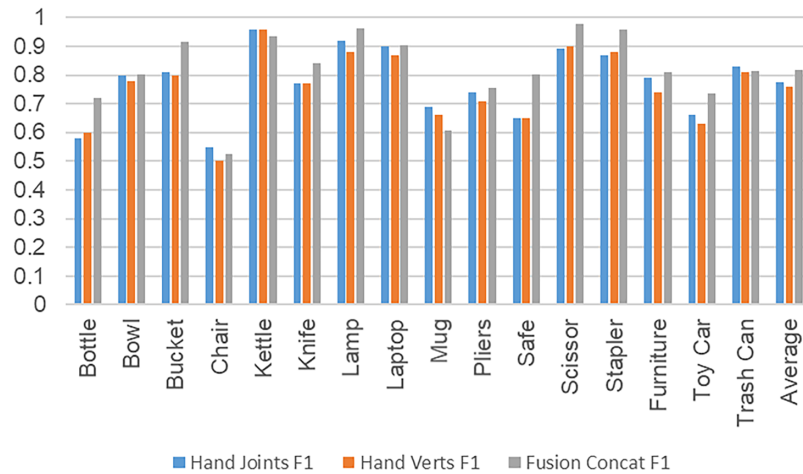


Figure 9: Object based F1-score for interaction classification.

The two-stage architecture also introduces a potential error-propagation effect because incorrect object predictions in the first stage can affect the subsequent object-specific interaction classifier. This issue is particularly relevant for object categories associated with similar grasping or manipulation patterns, such as bowl/bucket-like movements or drawer-related interactions. Therefore, although the two-stage design improves average F1-score by reducing the number of classes handled by each interaction classifier, its final HOI prediction remains dependent on the reliability of the object-recognition stage.

4.2 Analysis of Single Stage HOI Recognition

In the single-stage HOI recognition approach, the combined object–interaction class is predicted directly without first performing a separate object-classification stage. Features extracted from the hand joints and hand vertices are provided to the XGBoost classifier to recognize 54 distinct HOI classes. Table 4 summarizes the results obtained using the individual feature representations and different fusion strategies. Under concatenation-based fusion, interactions such as ‘Scissors: Cut something’ and ‘Lamp: Turn and fold’ achieved F1-scores of 0.89 and 0.92, respectively. The concatenation-based configuration achieved an average F1-score of 0.64 across all 54 HOI classes. However, the results indicate that performance varies across feature-fusion strategies, with minimum fusion yielding the highest overall average F1-score of 0.65. Several interaction classes involving drawer-related activities exhibited comparatively low F1-scores, suggesting that interactions with similar hand-motion patterns are more difficult to distinguish in the single-stage setting. These results indicate that the discriminative difficulty of HOI classes depends on both the object–interaction combination and the selected feature representation, and no single fusion strategy is consistently superior across all classes.

Table 4: Experimental results for single stage HOI recognition (where ‘C’ represents object category and ‘T’ represents corresponding task).

HOI	Hand Joints F1	Hand Vertex F1	Fusion Concat F1	Fusion Sum F1	Fusion Mean F1	Fusion Max F1	Fusion Min F1
C:Stapler T:Bindthepaper	0.5	0.61	0.65	0.60	0.78	0.75	0.76
C:Trash Can T:Throwsomethinginit	0.58	0.61	0.68	0.59	0.49	0.51	0.54
C:Lamp T:Turnonandturnoff	0.75	0.72	0.81	0.77	0.85	0.70	0.76
C:Scissors T:Pickandplace	0.53	0.72	0.87	0.73	0.77	0.80	0.67
C:Pliers T:Clampsomething	0.73	0.74	0.61	0.69	0.69	0.73	0.66
C:Mug T:Pickandplace(withwater)	0.62	0.58	0.50	0.48	0.33	0.39	0.44
C:Safe T:Takesomethingoutofit	0.51	0.5	0.59	0.55	0.56	0.28	0.35
C:Kettle T:Pourwaterintoamug	0.81	0.78	0.80	0.86	0.84	0.87	0.83
C:Toy Car T:Push toy car	0.46	0.48	0.39	0.36	0.58	0.66	0.61
C:Storage Furniture T:Putthedrinkinthedoor	0.68	0.59	0.54	0.66	0.59	0.68	0.53
C:Bowl T:Puttheballinthebowl	0.87	0.84	0.71	0.71	0.89	1.00	1.00
C:Knife T:Cutapple	0.67	0.71	0.69	0.65	0.82	0.76	0.69
C:Laptop T:Openandclosethedisplay	0.73	0.73	0.71	0.73	0.85	0.88	0.86
C:Bucket T:Pickandplace	0.6	0.58	0.61	0.68	0.72	0.64	0.71
C:Chair T:Pickandplacetotheoriginalposition	0.47	0.49	0.48	0.43	0.41	0.45	0.58
C:Bottle T:Pickandplace	0.68	0.68	0.65	0.76	0.55	0.52	0.53
C:Mug T:Pourwaterintoanothermug	0.59	0.58	0.62	0.56	0.80	0.75	0.79
C:Storage Furniture T:Openandclosethedoor	0.52	0.45	0.47	0.46	0.35	0.33	0.37
C:Pliers T:Putitinthedrawer	0.41	0.39	0.22	0.13	0.41	0.35	0.48
C:Scissors T:Cutsomething	0.41	0.83	0.89	0.76	0.86	0.88	0.79
C:Lamp T:Pickandplace	0.67	0.66	0.78	0.79	0.71	0.59	0.74
C:Knife T:Takeitoutofthedrawer	0.61	0.55	0.64	0.69	0.50	0.64	0.77
C:Bowl T:Pickandplace	0.64	0.67	0.76	0.74	0.70	0.70	0.72
C:Toy Car T:Pickandplace	0.58	0.55	0.64	0.55	0.61	0.53	0.55
C:Safe T:Openandclosethedoor	0.5	0.51	0.67	0.65	0.56	0.56	0.41
C:Mug T:Putitinthedrawer	0.38	0.44	0.20	0.35	0.29	0.34	0.33
C:Chair T:Pickandplacetoanewposition	0.48	0.49	0.37	0.28	0.36	0.39	0.45
C:Storage Furniture T:Putthedrinkinthedrawer	0.67	0.65	0.58	0.76	0.61	0.63	0.75
C:Kettle T:Pickandplace	0.81	0.81	0.85	0.81	0.87	0.86	0.79
C:Bottle T:Takeitoutofthedrawer	0.54	0.59	0.75	0.83	0.71	0.79	0.76
C:Pliers T:Pickandplace	0.52	0.51	0.64	0.62	0.58	0.63	0.55
C:Mug T:Takeitoutofthedrawer	0.59	0.65	0.49	0.51	0.68	0.63	0.66
C:Lamp T:Turnandfold	0.8	0.8	0.92	0.82	0.95	0.80	0.90
C:Bowl T:Takeitoutofthedrawer	0.75	0.73	0.77	0.80	0.79	0.78	0.83
C:Toy Car T:Takeitoutofthedrawer	0.48	0.43	0.48	0.59	0.49	0.46	0.62
C:Knife T:Pickandplace	0.74	0.74	0.79	0.80	0.75	0.79	0.69
C:Safe T:Putsomethinginit	0.54	0.55	0.57	0.55	0.59	0.58	0.55
C:Trash Can T:Openandclose	0.66	0.69	0.80	0.68	0.67	0.54	0.64
C:Mug T:Fillwithwaterbyakettle	0.64	0.59	0.50	0.44	0.73	0.82	0.75
C:Bottle T:Repositionthebottle	0.7	0.66	0.78	0.72	0.66	0.60	0.68
C:Storage Furniture T:Openandclosethedrawer	0.79	0.72	0.75	0.92	0.92	0.81	0.79
C:Pliers T:Takeitoutofthedrawer	0.53	0.55	0.67	0.70	0.58	0.59	0.54
C:Bowl T:Take the ball out of the bowl	0.72	0.8	0.77	0.79	0.85	0.86	0.95
C:Toy Car T:Putitinthedrawer	0.36	0.36	0.30	0.28	0.19	0.36	0.32
C:Laptop T:Pickandplace	0.81	0.78	0.84	0.84	0.83	0.89	0.87
C:Knife T:Putitinthedrawer	0.46	0.49	0.41	0.43	0.56	0.65	0.48

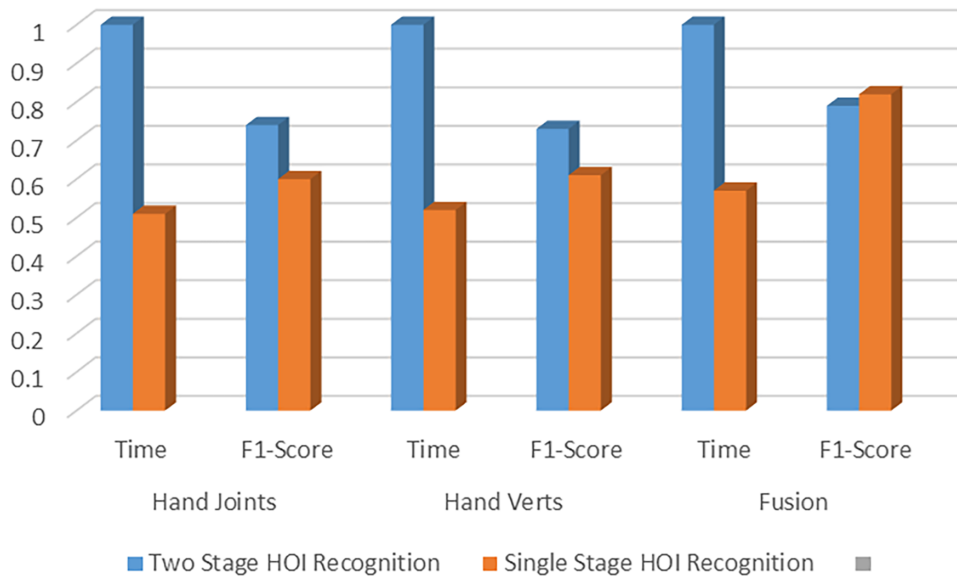
(Continued)

Table 4 (continued)

HOI	Hand Joints F1	Hand Vertex F1	Fusion Concat F1	Fusion Sum F1	Fusion Mean F1	Fusion Max F1	Fusion Min F1
C:Bucket							
T:Pourwaterintoanotherbucket	0.66	0.66	0.69	0.66	0.67	0.73	0.72
C:Bottle T:Pourallthewaterintoamug	0.32	0.4	0.45	0.36	0.49	0.33	0.42
C:Bowl T:Takeitoutofthedrawer	0.66	0.68	0.59	0.58	0.50	0.51	0.67
C:Bottle T:Putitinthedrawer	0.43	0.44	0.41	0.58	0.46	0.37	0.36
Average	0.60	0.61	0.64	0.62	0.64	0.63	0.65

5 Time and Performance Analysis: Single Stage HOI Recognition vs. Two Stage HOI Recognition

We have compared performance in terms of accuracy, F1 score and computational time per sample. Fig. 10 presents this comparison of the computational time with performance measures. In the two-stage approach, object classification is performed first and is followed by object-specific interaction recognition in the second stage. Due to these two stages, the inference time is 1.29 ms per sample for hand joint, whereas, 1.08 ms for hand vertex and 1.30 ms for fused feature vector of both hand joints and vertices. This approach achieved an average F1-scores of 73% for hand joint data, 73% for hand vertex data, and 81% for fused feature vector. On the other hand, the single-stage approach recognizes the interaction directly without object detection and it has a less inference time of 0.51 ms for hand joint data, 0.52 ms for hand vertex data, and 0.57 ms for fused feature set. However, it has been observed that although the inference time is reduced but the accuracy and F1 scores are also reduced. With single stage, proposed method achieved F1-scores of 60%, 61%, and 64% for hand joint, hand vertex, and fused feature vector respectively. The single-stage method demonstrates reduced performance due to its simultaneous classification of all 54 classes, whereas the two-stage method benefits from fewer classes per classification step, facilitating improved differentiation between object categories and their corresponding HOIs.

**Figure 10:** Time and performance analysis: two stage HOI recognition vs. single stage HOI recognition.

Generally, training and classifying interactions within specific object categories proves more manageable compared to global recognition approaches. Joint interaction labels may exhibit significant inter-class similarities across different categories, leading to misclassification and subsequently reducing methodological performance. The experimental evaluations show that the two stage process is much better than the single stage process in situations where the output label is very accurate. On the other hand, the single-stage technique is very effective in applications that require fast decision making and suitable for real-time applications where computational speed is important.

The proposed HOI recognition method is based on compact statistical descriptors rather than deep visual feature extraction. As a result, the feature extraction and inference pipeline is lightweight under the reported experimental setting. However, this study does not provide a direct runtime, memory, or energy comparison with representative transformer-based, graph-based, or multimodal deep HOI models. Therefore, computational advantages should be interpreted within the scope of the reported implementation and hardware configuration rather than as a universal claim over all deep-learning methods.

6 Computational Complexity Analysis

Let T denote the number of frames in a video segment, J the number of hand joints, V the number of hand vertices, C the number of descriptor channels, and S the number of temporal statistical features. The 7D transformation is linear in the number of points and frames, requiring $O(T(J + V)C)$ operations. The temporal aggregation step computes statistical descriptors over the channel-wise temporal signals and scales approximately as $O(TCS)$ after pooling over joints and vertices. XGBoost inference depends on the number of trees N and tree depth D , giving an approximate prediction complexity of $O(ND)$ per sample. This analysis indicates that the proposed pipeline scales linearly with the number of frames and hand points during feature construction, while classifier inference remains efficient for compact segment-level descriptors.

7 Limitations

Although Fuse3D demonstrates the usefulness of geometry-enriched hand kinematics for HOI recognition, several limitations should be noted. First, the method depends on MANO-based 3D hand reconstruction from RGB observations; therefore, reconstruction errors, occlusion, and inaccurate hand-pose estimation may propagate into the extracted descriptors and affect classification performance. A dedicated noise-perturbation or MANO reconstruction-error sensitivity analysis was not included in the present study and will be considered in future work.

Second, channel-wise mean pooling introduces an explicit information–compactness trade-off. Pooling across the 21 joints or 778 MANO vertices yields stable fixed-length temporal signals and substantially reduces dimensionality, but it removes point identity, local anatomical relationships, mesh connectivity, and finger-level surface structure. Fuse3D should therefore not be interpreted as a topology-aware or dense-morphology method. Its vertex stream contains global coordinate and motion statistics derived from dense surface sampling. Interactions that depend on fine local articulation, similar global hand trajectories, or object appearance may remain difficult to distinguish. Topology-aware mesh models, anatomically partitioned pooling, and finger-specific descriptors are important future extensions.

Third, the proposed representation intentionally focuses on hand kinematics and does not directly use object appearance, texture, color, or scene context. This design improves compactness and interpretability, but it can limit discrimination between object categories that involve similar grasping and motion patterns. This limitation is visible in relatively challenging categories and drawer-related interactions, where hand motion alone may not provide sufficient object-specific evidence.

Fourth, the experiments were performed on the HOI4D data set using the 5-fold evaluation protocol reported. In the present study cross-dataset validation and fully subject-independence evaluation were not conducted. These assessments are significant for generalization to unseen subject, environment and data sets and will be discussed later in this work.

Fifth, despite the reported low computational complexity, it is not confirmed that the proposed method is lightweight in terms of run-time, memory usage, and energy consumption compared to recently reported HOI recognition systems based on trans-formers, graphs, or multimodality. Hence, statements of computational efficiency should be restricted to the time needed for the model to make an inference in the proposed method and cannot be regarded as a direct superiority statement over any other deep-learning method.

Finally, the privacy preserving nature of the method needs to be taken with a pinch of salt. Fuse3D minimizes reliance on features of the object's appearance in the classification process but it depends on RGB video input for hand-pose reconstruction. Thus, the method can alleviate some of the appearances-related dependency but cannot ensure privacy without any visual sensing nor ensure the absence of visual sensing.

8 Conclusion and Future Directions

This study proposed Fuse3D, a geometry-enriched hand-kinematic framework for human-object interaction recognition using 3D hand joints and MANO surface vertices extracted from egocentric video streams. The seven-channel descriptor is a deterministic coordinate-derived nonlinear expansion, not a rotation-invariant representation. Mean pooling across joints or vertices generates global temporal signals, after which statistical descriptors and lightweight XGBoost classifiers are applied. The pooled vertex descriptor summarizes kinematics derived from the complete MANO surface sampling but does not preserve mesh topology or detailed local morphology.

Experimental evaluation on HOI4D showed that fused joint and vertex descriptors improve recognition performance compared with individual streams, and the two-stage framework achieved an average F1-score of 81% under the reported 5-fold evaluation protocol. The single-stage framework provided faster inference but lower recognition performance, indicating a trade-off between computational speed and classification accuracy. The results support the discriminative value and complementarity of compact global joint-derived and mesh-derived kinematic signals for segment-level HOI recognition; they should not be interpreted as evidence that local hand-surface morphology is retained.

Future work will focus on subject-independent and cross-dataset validation, direct comparison with recent transformer-based, graph-based, and multimodal HOI methods, and sensitivity analysis of MANO reconstruction errors through controlled perturbation experiments. Additional ablation studies will also be conducted to isolate the contribution of the original 3D coordinates, the appended geometric descriptors, temporal feature groups, and different fusion strategies. Future work will additionally investigate anatomically partitioned pooling, finger-specific trajectories, and topology-aware mesh models to retain local articulation when required. Further research will also explore integrating object appearance, object pose, and sensor-based tracking to improve recognition of interactions that cannot be reliably distinguished from hand motion alone.

Acknowledgement: Not applicable.

Funding Statement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2603).

Author Contributions: Conceptualization, Danish Hamid, Yazeed Alkhrijah, Syed Muhammad Usman, Muhammad Tahir Akram, Fazal ur Rehman, Shehzad Khalid, Sulieman S. Alshuhri and Amal Alomran; methodology, Danish

Hamid, Yazeed Alkhrijah, Syed Muhammad Usman and Muhammad Tahir Akram; software, Danish Hamid, Yazeed Alkhrijah and Muhammad Tahir Akram; validation, Danish Hamid, Yazeed Alkhrijah, Syed Muhammad Usman and Muhammad Tahir Akram; formal analysis, Danish Hamid, Yazeed Alkhrijah, Syed Muhammad Usman and Muhammad Tahir Akram; investigation, Danish Hamid, Yazeed Alkhrijah, Syed Muhammad Usman, Muhammad Tahir Akram and Fazal ur Rehman; resources, Syed Muhammad Usman, Shehzad Khalid, Sulieman S. Alshuhri and Amal Alomran; data curation, Danish Hamid, Yazeed Alkhrijah and Muhammad Tahir Akram; writing—original draft preparation, Danish Hamid, Yazeed Alkhrijah, Syed Muhammad Usman, Muhammad Tahir Akram, Fazal ur Rehman, Shehzad Khalid, Sulieman S. Alshuhri and Amal Alomran; writing—review and editing, Danish Hamid, Yazeed Alkhrijah, Syed Muhammad Usman, Muhammad Tahir Akram, Fazal ur Rehman, Shehzad Khalid, Sulieman S. Alshuhri and Amal Alomran; visualization, Danish Hamid and Yazeed Alkhrijah; supervision, Syed Muhammad Usman and Shehzad Khalid; project administration, Syed Muhammad Usman and Shehzad Khalid. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and/or analyzed during this study are available in the following Github repository at <https://hoi4d.github.io/>.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hou Z, Yu B, Qiao Y, Peng X, Tao D. Affordance transfer learning for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Virtual. p. 495–504. doi:10.1109/cvpr46437.2021.00056.
2. Gupta S, Malik J. Visual semantic role labeling. arXiv:1505.04474. 2015. doi:10.48550/arxiv.1505.04474.
3. Li Q, Xie X, Zhang J, Shi G. Few-shot human-object interaction video recognition with transformers. Neural Netw. 2023;163(3):1–9. doi:10.1016/j.neunet.2023.01.019.
4. Chao YW, Wang Z, He Y, Wang J, Deng J. HICO: a benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile. p. 1017–25.
5. Sadhu A, Gupta T, Yatskar M, Nevatia R, Kembhavi A. Visual semantic role labeling for video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Virtual. p. 5589–600.
6. Zhou T, Wang W, Qi S, Ling H, Shen J. Cascaded human-object interaction recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 4263–72.
7. Li Y, Ouyang W, Zhou B, Shi J, Zhang C, Wang X. Factorizable net: an efficient subgraph-based framework for scene graph generation. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany. p. 335–51.
8. Bansal S, Wray M, Damen D. HOI-Ref: hand-object interaction referral in egocentric vision. arXiv:2404.09933. 2024.
9. Cai M, Kitani K, Sato Y. Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes. arXiv:1807.08254. 2018.
10. Hamid D, Haq MEU, Yasin A, Murtaza F, Azam MA. Enhancing recognition of human-object interaction from visual data using egocentric wearable camera. Future Internet. 2024;16(8):269. doi:10.3390/fi16080269.
11. Hamid D, Haq MEU, Yasin A, Murtaza F, Azam MA. Using 3D hand pose data in recognizing human-object interaction and user identification for extended reality systems. Information. 2024;15(10):629. doi:10.3390/info15100629.
12. Hamza M, Hamid D, Akram MT. Person identification from egocentric human-object interactions using 3D hand pose. CCF Trans Pervasive Comput Interact. 2026;8(2):270–86. doi:10.1007/s42486-025-00215-x.

13. Wan B, Zhou D, Liu Y, Li R, He X. Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 9469–78.
14. Zhong X, Qu X, Ding C, Tao D. Glance and gaze: inferring action-aware points for one-stage human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Nashville, TN, USA. p. 13234–43.
15. Wang T, Yang T, Danelljan M, Khan FS, Zhang X, Sun J. Learning human-object interaction detection using interaction points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 4116–25.
16. He T, Gao L, Song J, Li YF. Exploiting scene graphs for human-object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 15984–93.
17. Nagarajan T, Feichtenhofer C, Grauman K. Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 8688–97.
18. Qiao T, Li R, Li FWB, Kubotani Y, Morishima S, Shum HPH. Geometric visual fusion graph neural networks for multi-person human-object interaction recognition in videos. *Expert Syst Appl*. 2025;290(6):128344. doi:10.1016/j.eswa.2025.128344.
19. Alkhrijah Y, Hamid D, Haq MEU, Usman SM, Khalid S, Alawad MA. EgoVision a YOLO-ViT hybrid for robust egocentric object recognition. *Sci Rep*. 2025;15(1):34723. doi:10.1038/s41598-026-36489-z.
20. Ye Q, Xu X, Li R. Human-object behavior analysis based on interaction feature generation algorithm. *Int J Adv Comput Sci Appl*. 2023;14(8):540–9. doi:10.14569/ijacsa.2023.0140859.
21. Yang N, Zheng Y, Guo X. Efficient transformer for human-object interaction detection. In: Proceedings of the Sixth International Conference on Computer Information Science and Application Technology; 2023 May 26–28; Hangzhou, China. p. 536–42.
22. Zaib MH, Khan MJ. An HMM-based approach for human interaction using multiple feature descriptors. *SSRN*. 2023:4656240.
23. Ozaki H, Tran DT, Lee JH. Effective human–object interaction recognition for edge devices in intelligent space. *SICE J Control Meas Syst Integr*. 2024;17(1):1–9. doi:10.1080/18824889.2023.2292353.
24. Gkioxari G, Girshick R, Dollár P, He K. Detecting and recognizing human-object interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–22; Salt Lake City, UT, USA. p. 8359–67.
25. Zhou P, Chi M. Relation parsing neural network for human-object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 843–51.
26. Kato K, Li Y, Gupta A. Compositional learning for human object interaction. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany. p. 247–64.
27. Xie X, Bhatnagar BL, Pons-Moll G. Visibility aware human-object interaction tracking from single RGB camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. p. 4757–68.
28. Purwanto D, Chen YT, Fang WH. First-person action recognition with temporal pooling and Hilbert–Huang transform. *IEEE Trans Multimed*. 2019;21(12):3122–35. doi:10.1109/tmm.2019.2919434.
29. Sajid MSB, Alkhrijah Y, Hamid D, Akram MT, Usman SM, Khalid S, et al. TransFlow-HOI: transformer-driven multimodal fusion for high-performance human-object interaction recognition. *IEEE Access*. 2026;14:12717–32.
30. Liu T, Zhao R, Jia W, Lam KM, Kong J. Holistic-guided disentangled learning with cross-video semantics mining for concurrent first-person and third-person activity recognition. *IEEE Trans Neural Netw Learn Syst*. 2024;35(4):5211–25. doi:10.1109/tnnls.2022.3202835.
31. Kwon T, Tekin B, Stühmer J, Bogo F, Pollefeys M. H₂O: two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 10138–48.

32. Leonardi R, Ragusa F, Furnari A, Farinella GM. Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario. *Comput Vis Image Underst.* 2024;242(4):103984. doi:10.1016/j.cviu.2024.103984.
33. Liu Y, Liu Y, Jiang C, Lyu K, Wan W, Shen H, et al. HOI4D: a 4D egocentric dataset for category-level human-object interaction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2022 Jun 18–24; New Orleans, LA, USA. p. 21013–22.
34. Romero J, Tzionas D, Black MJ. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans Graph.* 2017;36(6):1–17.
35. Hasson Y, Varol G, Tzionas D, Kalevatykh I, Black MJ, Laptev I, et al. Learning joint reconstruction of hands and manipulated objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2019 Jun 15–20; Long Beach, CA, USA. p. 11807–16.