

Lightweight Dual-View Feature Fusion for Hand-Object Interaction Recognition

Houda Skhoun

School of Artificial Intelligence Nanjing University of Information Science and Technology (NUIST)
Nanjing, China

DOI: <https://doi.org/10.51583/IJLTEMAS.2026.150500198>

Received: 17 May 2026; Accepted: 22 May 2026; Published: 13 June 2026

ABSTRACT

Wrist-worn hand-object interaction (HOI) recognition is a critical capability for wearable rehabilitation systems, assistive technologies, augmented reality and human-computer interaction applications. Compared with fixed external cameras, wrist-worn wearable devices provide user-centered observations that are more suitable for continuous real-world interaction monitoring. However, many existing wrist-worn HOI recognition systems still face important challenges, including incomplete interaction representation caused by single-view observations, viewpoint ambiguity, self-occlusion and the high computational complexity of recent deep learning approaches. To address these limitations, this paper proposes a lightweight dual-view framework for hand-object interaction recognition using synchronized palm-view and back-view RGB images acquired from a wrist-worn dual-camera device. The proposed framework employs a shared MobileNetV2 backbone combined with multi-level feature extraction to jointly capture fine-grained spatial details and high-level semantic representations. To effectively integrate complementary information from different network depths, a view-specific adaptive fusion mechanism is introduced to dynamically balance intermediate and deep feature representations for each visual stream. The fused dual-view representation is subsequently used for interaction classification. Experimental evaluation under the Leave-One-Participant-Out (LOPO) cross-subject protocol demonstrates that the proposed framework achieves a mean accuracy of 82.36% and a mean F1-score of 81.65% while maintaining low computational complexity suitable for real-time wearable applications. Ablation studies further confirm the effectiveness of the proposed multi-level feature extraction and adaptive fusion strategy. The proposed approach provides an effective balance between recognition performance and computational efficiency for lightweight wearable HOI recognition systems.

Keywords: Hand-object interaction recognition; Dual-view learning; Adaptive feature fusion; Lightweight deep learning; Wrist-worn systems

INTRODUCTION

Wrist-worn hand-object interaction (HOI) recognition has emerged as an important capability for wearable rehabilitation systems, assistive technologies, augmented reality (AR/VR), and human-robot collaboration [1–3]. Accurate recognition of object manipulation enables intelligent systems to better understand user intention and provide adaptive responses during real-world interaction scenarios. Recent advances in deep learning have significantly improved visual recognition performance by enabling automatic extraction of hierarchical feature representations directly from image data, overcoming many limitations associated with traditional handcrafted feature approaches [4–8].

Among existing acquisition settings, wrist-worn wearable systems have attracted increasing attention due to their portability, non-intrusive design and suitability for continuous real-world interaction monitoring. Compared with fixed external cameras, wrist-worn devices provide observations that remain closely aligned with the user's hand movements during object manipulation while reducing dependency on environmental setup and camera placement. Their compact and wearable nature makes them particularly suitable for assistive technologies, rehabilitation systems and real-time human-computer interaction

applications.

Existing hand-object interaction recognition methods have explored various sensing configurations, including external cameras, egocentric wearable cameras, RGB-D systems and multi-modal sensing approaches [9–13]. Despite recent progress, robust HOI recognition remains challenging in wrist-worn settings due to several factors. First, many approaches rely on single-view observations, which may fail to capture complete interaction information, including fine-grained finger articulation, object contact regions and global hand posture. Second, viewpoint variation, illumination changes and inter-subject differences in hand shape and manipulation style further degrade recognition performance [2, 14]. Finally, many recent methods depend on computationally expensive architectures or temporal modeling strategies, limiting their suitability for lightweight real-time wearable applications.

To address these limitations, this paper proposes a lightweight dual-view framework for hand-object interaction recognition using synchronized palm-view and back-view observations acquired from a wrist-worn dual-camera device. The proposed acquisition setup provides complementary visual information from two synchronized viewpoints, where the palm-view stream captures detailed finger articulation and object contact regions, while the back-view stream provides global hand posture and movement configuration. By combining both visual streams, the proposed framework improves robustness against self-occlusion and viewpoint ambiguity during fine-grained interaction analysis.

Based on this acquisition setting, a lightweight multi-level deep learning framework is developed using a shared MobileNetV2 backbone combined with adaptive feature fusion. Intermediate and high-level feature representations are extracted from different network depths to capture complementary spatial and semantic information. An adaptive gating mechanism is then employed to dynamically balance the contribution of multi-level features for each visual stream before dual-view fusion and interaction classification. The proposed framework is designed to maintain low computational complexity while achieving strong recognition performance suitable for wearable real-time applications.

The proposed framework is evaluated using a subject-independent Leave-One-Participant-Out (LOPO) protocol on a dual-view hand-object interaction dataset collected using the proposed wrist-worn acquisition system. Experimental results demonstrate that the proposed framework achieves strong recognition performance while maintaining low computational cost under cross-subject evaluation conditions.

The main contributions of this paper are summarized as follows:

1. A lightweight dual-view framework is proposed for hand-object interaction recognition using synchronized palm-view and back-view observations acquired from a wrist-worn dual-camera device.
2. A multi-level feature extraction strategy is developed to integrate intermediate and deep representations for fine-grained interaction analysis.
3. An adaptive view-specific fusion mechanism is proposed to dynamically combine complementary multi-level representations from the two visual streams.
4. Extensive experiments under the LOPO cross-subject evaluation protocol demonstrate the effectiveness, robustness and computational efficiency of the proposed framework for wearable real-time HOI recognition.

RELATED WORK

Hand-Object Interaction Recognition

Hand-object interaction (HOI) recognition has attracted significant attention in computer vision due to its applications in human-computer interaction, rehabilitation systems, robotics and wearable intelligent devices [1–3]. Compared with isolated hand gesture recognition, HOI recognition is more challenging

because interaction understanding depends on both hand configuration and object-related contextual information. In practical scenarios, severe occlusions, viewpoint variations, illumination changes and inter-subject differences further complicate robust recognition [2, 14].

Early HOI recognition methods mainly relied on handcrafted features and conventional machine learning techniques. Traditional approaches focused on hand segmentation object tracking, and geometric modeling using RGB or RGB-D inputs [15–18]. Other works incorporated contextual information such as grasp type and object attributes to better characterize manipulation actions [11, 19]. Although these methods established important foundations for interaction analysis, their dependence on handcrafted representations and multi-stage processing pipelines limited their generalization capability in unconstrained environments.

Recent advances in deep learning have significantly improved HOI recognition by enabling end-to-end learning of discriminative visual representations [4–6]. CNN-based approaches have been widely adopted for egocentric interaction understanding, object localization and hand-object relationship modeling [3,20,21]. In addition, several studies proposed multitask frameworks integrating hand pose estimation, object detection and interaction recognition within unified architectures [12, 22, 23]. Temporal modeling techniques based on recurrent networks, 3D convolutional networks and ConvLSTM architectures further improved dynamic interaction analysis [9, 24–29]. Multi-modal approaches combining RGB, depth and skeleton information have also been explored to improve robustness under challenging conditions [10, 30, 31].

Despite these advances, many existing HOI recognition methods rely on large annotated datasets, temporal video modeling or computationally expensive architectures, which may limit their applicability in lightweight wearable systems and real-time environments [3, 12, 23]. Furthermore, single-view observations often provide incomplete interaction representations under occlusion and viewpoint variation, motivating the exploration of multi-view learning strategies.

Multi-View Learning and Feature Fusion

Multi-view learning has been widely investigated to improve recognition robustness under viewpoint variation and partial occlusion [10]. By combining complementary observations from multiple viewpoints, multi-view systems can capture more complete spatial information and reduce ambiguity during visual recognition tasks. Previous studies demonstrated that multi-view observations improve gesture and action recognition performance by exploiting complementary visual cues across different perspectives [33–35].

To integrate information from multiple view-points, several fusion strategies have been proposed. Early fusion combines visual information before feature extraction, enabling the network to directly learn correlations between views [36, 37]. Feature-level fusion merges intermediate representations extracted independently from each view-point while preserving discriminative characteristics from individual views [38, 39]. Late fusion combines predictions generated from separate visual streams [36, 40]. More recent approaches employ attention-based or adaptive fusion mechanisms to dynamically emphasize the most informative representations according to the input characteristics [41–43]. These methods demonstrated improved robustness in recognition tasks affected by viewpoint ambiguity and occlusion.

However, many existing multi-view approaches rely on computationally intensive architectures or dense temporal processing. In addition, several methods primarily focus on global representations without explicitly exploiting complementary information from different feature levels.

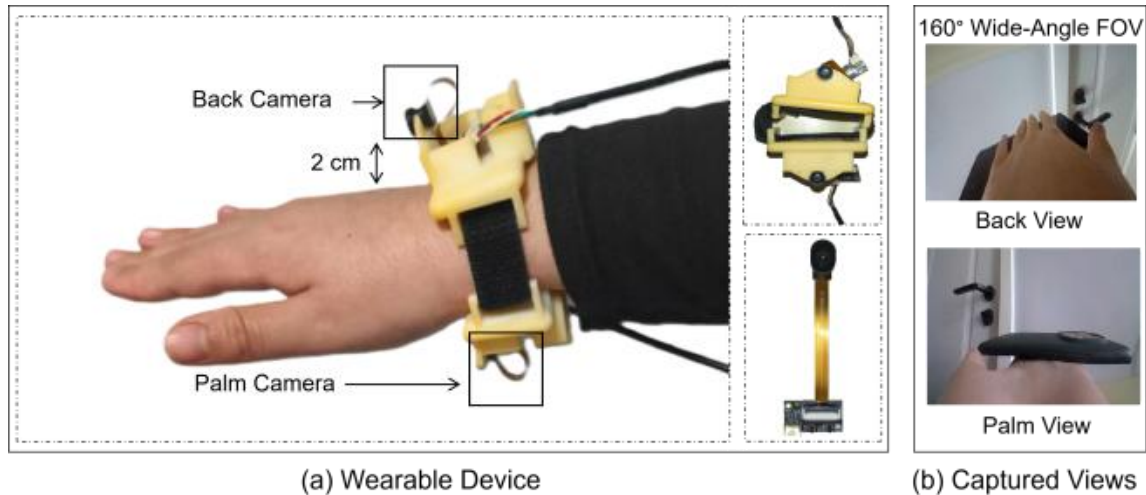


Figure 1: (a) Custom-designed wearable dual-camera device showing the positioning of the back and palm cameras at approximately 2 cm from the wrist. (b) Examples of captured images from the two viewpoints, illustrating the back view and palm view with a 160° wide-angle field of view.

Multi-Level Feature Learning

Multi-level feature learning has become an important component of modern deep visual recognition systems. CNNs naturally learn hierarchical representations, where shallow and intermediate layers capture local spatial patterns, while deeper layers encode high-level semantic information. Residual learning and feature pyramid strategies demonstrated that combining features from multiple network depths improves robustness in complex visual recognition tasks [44, 45].

For HOI recognition, multi-level feature integration is particularly important because subtle variations in finger articulation, grasp configuration and object contact regions often determine the interaction category. However, many existing approaches mainly rely on deep semantic representations while ignoring intermediate spatial features that may contain important fine-grained interaction cues. Moreover, several methods continue to depend on single-view observations or computationally expensive temporal architectures, limiting their suitability for lightweight wearable applications and real-time deployment.

Another important challenge concerns subject-independent generalization. Variations in hand shape, manipulation style and interaction execution across users can significantly affect visual appearance and reduce recognition robustness. Consequently, there remains a need for efficient HOI recognition frameworks capable of integrating complementary multi-view information while maintaining low computational complexity and strong cross-subject generalization capability.

To address these limitations, this work proposes a lightweight dual-view multi-level feature fusion framework for hand-object interaction recognition using synchronized palm-view and back-view observations. The proposed framework combines multi-level feature extraction with adaptive view-specific fusion to effectively exploit complementary visual information while maintaining computational efficiency suitable for wearable real-time applications.

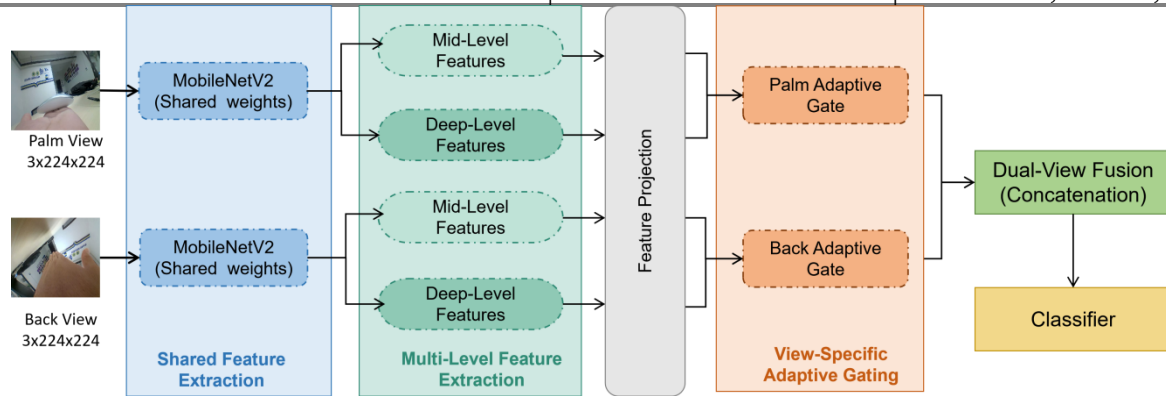


Figure 2: Overview of the proposed dual-view multi-level feature fusion framework for hand object interaction recognition. The framework integrates dual-view inputs, multi-level feature extraction, adaptive gating fusion and final classification

METHODOLOGY

Data Acquisition

The proposed framework employs a lightweight wrist-worn dual-camera device designed to capture complementary visual information during hand-object interactions. As illustrated in Figure 1, two synchronized wide-angle RGB cameras with an approximate 160° field of view are mounted on a wearable wrist support to simultaneously acquire palm-view and back-view observations of the same interaction instance. The back-view camera captures global hand posture and interaction context, whereas the palm-view camera provides detailed information related to finger articulation, contact regions and manipulated objects. The cameras are positioned approximately 2 cm from the wrist surface to maximize viewpoint complementarity while maintaining a compact wearable configuration.

Let I^p and I^b denote the palm-view and back-view RGB inputs, respectively. Each synchronized image pair represents the same interaction captured from complementary viewpoints. This dual-view representation enables the framework to exploit both fine-grained local interaction cues and global contextual information while reducing ambiguities caused by self-occlusion and viewpoint variation. For compatibility with the MobileNetV2 backbone, all input images are resized to 224×224 pixels. During training, data augmentation techniques including random resized cropping, color jittering, affine transformations, Gaussian blurring and random grayscale conversion are applied to improve robustness and reduce overfitting. Since the two images form a synchronized pair, identical augmentation parameters are applied to both views to preserve spatial correspondence. The processed images are finally normalized using ImageNet mean and standard deviation values before being passed to the network.

Network Overview

The proposed framework performs hand-object interaction (HOI) recognition using synchronized palm-view and back-view observations. As illustrated in Figure 2, the framework consists of dual visual streams, multi-level feature extraction, adaptive feature fusion and a lightweight classification module. Let $I^p \in \mathbb{R}^{3 \times H \times W}$ and $I^b \in \mathbb{R}^{3 \times H \times W}$ denote the palm-view and back-view RGB inputs, respectively. Both visual streams are processed using a shared MobileNetV2 backbone $\Phi(\cdot)$ to extract hierarchical feature representations F^p and F^b . To preserve both local spatial details and high-level semantic information, multi-level features are extracted from intermediate and deep network layers. An adaptive gating module then generates compact view-specific representations v^{palm} and v^{back} for the palm-view and back-view streams, respectively. The resulting representations are subsequently fused to generate the final interaction representation v , which is passed to the classification module for HOI prediction. The proposed framework effectively exploits complementary multi-view information while maintaining computational efficiency adapted to wearable real-time applications.

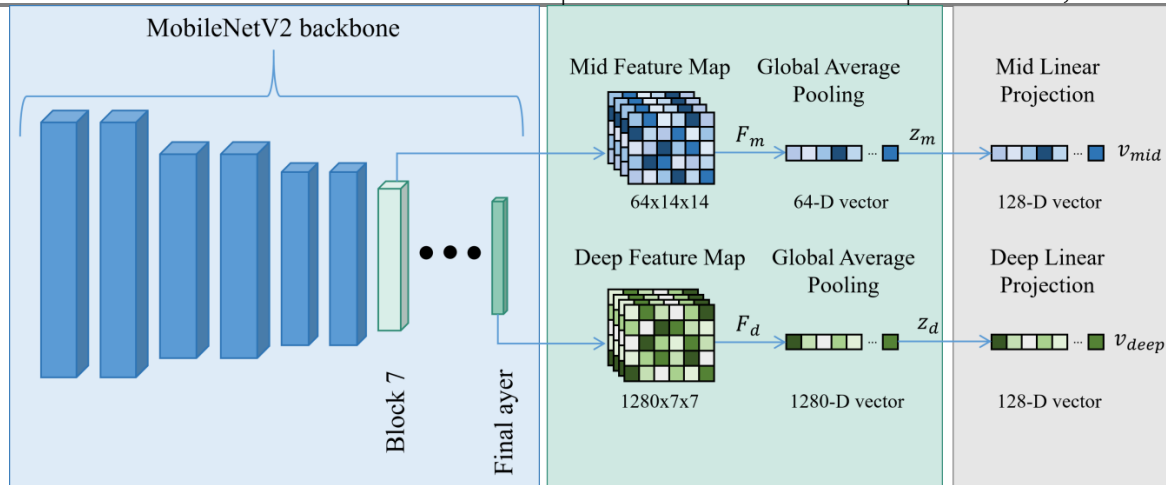


Figure 3: Multi-level feature extraction and token generation from the MobileNetV2 backbone

Multi-Level Feature Extraction

To obtain discriminative yet efficient visual representations, the proposed framework employs MobileNetV2 as a shared lightweight backbone for both palm-view and back-view streams. MobileNetV2 is selected due to its favorable trade-off between recognition performance and computational efficiency, making it compatible with wearable real-time applications.

Given the synchronized inputs I^p and I^b , the shared backbone $\Phi(\cdot)$ extracts hierarchical feature representations for each visual stream:

$$F^p = \Phi(I_p), \quad F^b = \Phi(I_b) \quad (1)$$

Instead of relying only on the final deep representation, the proposed framework extracts features from multiple network depths to preserve both local spatial details and high-level semantic information. As illustrated in Figure 3, intermediate feature maps capture fine-grained interaction patterns related to finger articulation, object boundaries and contact regions, whereas deeper feature maps encode higher-level interaction semantics. For each visual stream, the backbone produces an intermediate feature map F_m and a deep feature map F_d . Global average pooling is then applied to obtain compact feature vectors:

$$z_m = \text{GAP}(F_m), \quad z_d = \text{GAP}(F_d) \quad (2)$$

Since the extracted representations have different dimensions, lightweight linear projection layers are used to map them into a unified embedding space:

$$v_{mid} = W_m z_m + b_m, \quad v_{deep} = W_d z_d + b_d \quad (3)$$

where W_m and W_d denote learnable projection matrices and b_m and b_d are bias terms. The projected representations v_{mid} and v_{deep} are subsequently used as inputs to the adaptive fusion module.

Adaptive Feature Fusion

The discriminative importance of intermediate and deep feature representations may vary depending on the interaction type, viewpoint conditions and object appearance. To dynamically balance the contribution of different representation levels, the proposed framework employs an adaptive gating module for each visual stream, as illustrated in Figure 4. Given the projected intermediate and deep representations v_{mid} and v_{deep} , adaptive gating weights are computed as:

$$\omega = \text{MLP}([v_{mid}; v_{deep}]) \quad (4)$$

where $[\cdot]$ denotes feature concatenation and $\omega = [\omega_{mid}; \omega_{deep}]$ represents the learned fusion weights associated with the intermediate and deep representations, respectively. To obtain normalized adaptive weights, a softmax operation is applied:

$$[\alpha_{mid}, \alpha_{deep}] = \text{Softmax}(\omega) \quad (5)$$

The final view-specific representation v^{view} is then computed as a weighted combination of the two feature levels:

$$v^{view} = \alpha_{mid} \odot v_{mid} + \alpha_{deep} \odot v_{deep} \quad (6)$$

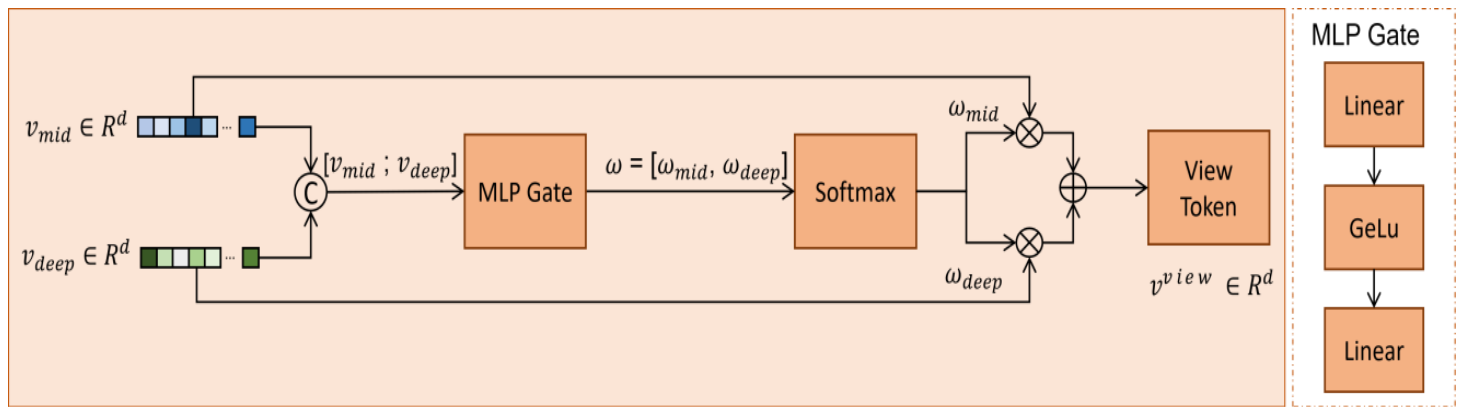


Figure 4: View-specific adaptive gating mechanism for combining mid-level and deep-level feature representations

where $\odot$ denotes element-wise multiplication. This adaptive formulation enables the framework to dynamically emphasize either fine-grained spatial information or high-level semantic representations according to the characteristics of the interaction. After adaptive fusion, the palm-view and back-view representations are concatenated to generate the final interaction representation:

$$v = [v^{palm}; v^{back}] \quad (7)$$

By adaptively balancing multi-level representations for each visual stream, the proposed fusion strategy effectively exploits complementary information from both viewpoints while maintaining low computational complexity appropriate for lightweight wearable applications.

Classification

The fused representation v is passed to a lightweight classification module composed of two fully connected layers with GELU activation and dropout regularization. The classifier transforms the fused representation into the final hand-object interaction prediction scores. The network is trained end-to-end using the cross-entropy loss function. The resulting output scores correspond to the predicted hand-object interaction categories.

EXPERIMENTS AND RESULTS

Experimental Setup

Dataset Description

The experiments are conducted on a dual-view hand-object interaction dataset collected using the proposed wrist-worn acquisition system. Each interaction instance consists of synchronized palm-view and back-

view RGB image pairs captured simultaneously from complementary viewpoints. Example interaction samples from the proposed dataset are illustrated in Figure 5. The dataset contains 13 hand-object interaction categories involving four commonly manipulated objects: pen, mouse, book and phone. The interaction types include grasp, hold, pinch and support actions. Data were collected from 10 participants under varying hand configurations and interaction styles to ensure cross-subject diversity. Overall, the dataset contains 26,894 synchronized interaction samples with a near-uniform class distribution, as summarized in Table 1. The diversity of subjects, viewpoints and interaction patterns provides a challenging evaluation scenario for cross-subject hand-object interaction recognition under varying occlusion and viewpoint conditions.

Table 1: Class-wise distribution of interaction samples in the proposed dual-view hand object interaction dataset

Class ID	Interaction	#Samples
C01	Grasp Pen	2092
C02	Hold Book	2052
C03	Hold Mouse	2059
C04	Hold Pen	2059
C05	Hold Phone	2051
C06	Pinch Book	2071
C07	Pinch Mouse	2095
C08	Pinch Pen	2077
C09	Pinch Phone	2060
C10	Support Book	2052
C11	Support Mouse	2059
C12	Support Pen	2110
C13	Support Phone	2057
Total	—	26,894

Evaluation Protocol

To evaluate the cross-subject generalization capability of the proposed framework, a Leave-One-Participant-Out (LOPO) cross-validation protocol is adopted. The dataset contains interaction samples collected from 10 participants. In each evaluation fold, the samples from one participant are used for testing, while the remaining participants are used for training. This process is repeated across all participants and the final performance is reported as the average across the 10 folds. The LOPO protocol ensures strict subject-independent evaluation by testing the model on unseen users, providing a reliable assessment of the framework under varying hand shapes, interaction styles and manipulation patterns.

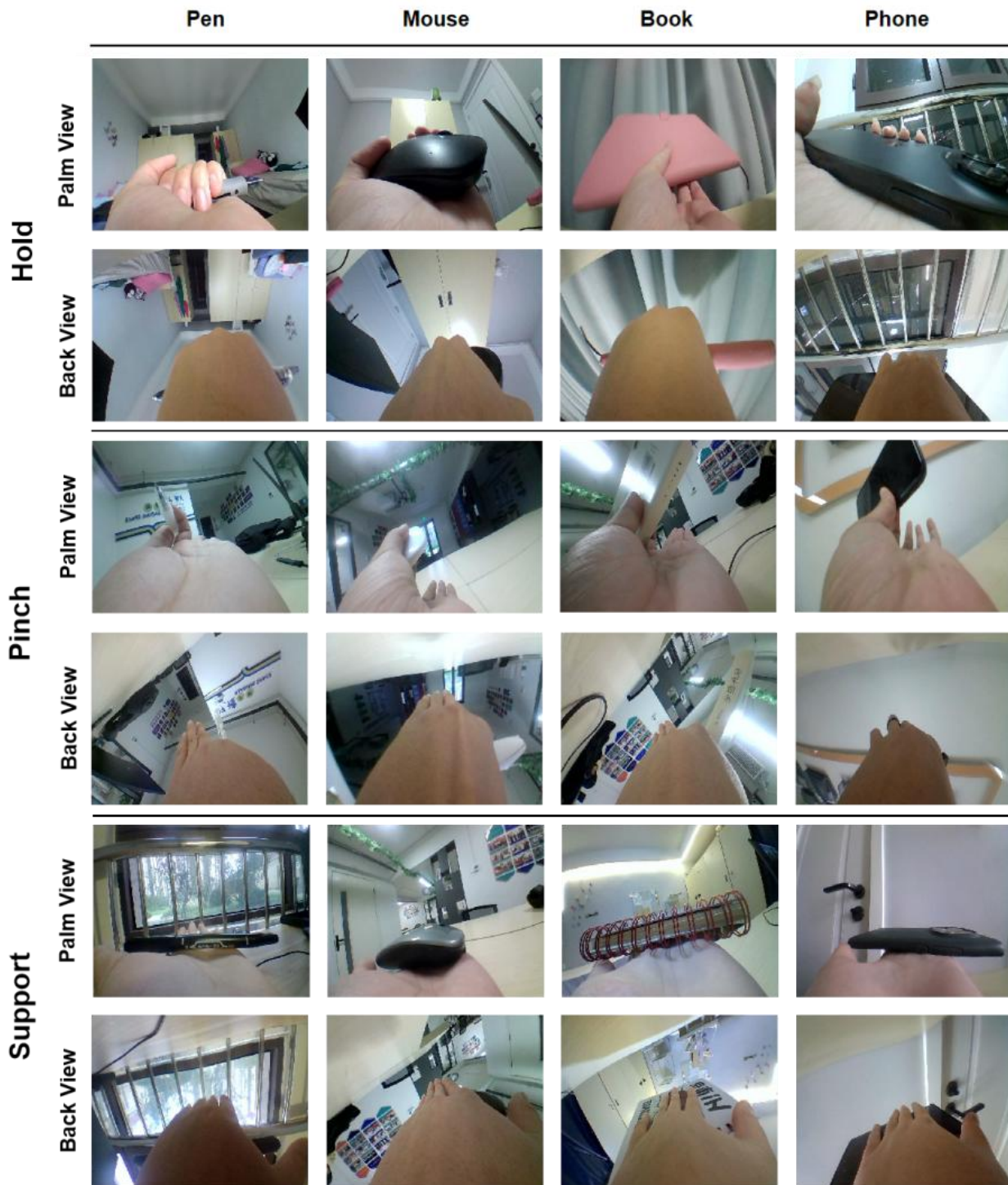


Figure 5: Each interaction instance is represented by synchronized palm and back views, highlighting complementary visual information for different grasp types and manipulated objects

Implementation Details

The proposed framework is implemented using the PyTorch deep learning library and trained using GPU acceleration. The network is trained for 20 epochs with a batch size of 32 using the AdamW optimizer. The initial learning rate is set to 3×10^{-5} , with a weight decay of 5×10^{-5} . A cosine annealing learning rate scheduler is employed to gradually reduce the learning rate during training. Cross-entropy loss with label smoothing 0.1 is adopted to improve generalization.

During training, synchronized data augmentation is applied to both visual streams. The images are normalized using the ImageNet mean and standard deviation values. For evaluation, images are resized to 224×224 , normalized using ImageNet statistics and tested without augmentation. Experiments are

conducted under the Leave-One-Participant-Out (LOPO) cross-validation protocol, where each participant is used once as the test subject while the remaining participants are used for training.

Table 2: Cross-subject performance of the proposed method under the Leave-One-Participant-Out (LOPO) protocol

Fold	Acc (%)	F1 (%)	Precision (%)	Recall (%)
1	89.95	90.57	89.89	89.94
2	87.26	89.04	87.38	87.00
3	71.50	81.31	72.02	70.63
4	76.49	81.91	76.75	76.71
5	91.28	92.57	91.57	91.06
6	86.73	89.21	86.73	86.32
7	63.09	77.65	63.65	59.42
8	86.99	88.63	86.71	86.20
9	84.63	88.86	84.85	83.61
10	85.68	88.58	85.77	85.58
Mean±std	82.36 ± 8.61	86.83 ± 4.55	82.53 ± 8.43	81.65 ± 9.42

Evaluation Metrics

The performance of the proposed framework is evaluated using accuracy, macro-averaged precision, recall and F1-score. Accuracy measures the overall proportion of correctly classified interaction samples, while precision, recall and F1-score provide a more comprehensive evaluation of the recognition performance across all interaction categories. The macro-average strategy computes each metric independently for every class and then averages the results, ensuring equal contribution from all categories regardless of sample frequency. This evaluation protocol provides a balanced assessment of the framework under the cross-subject LOPO setting. In addition, confusion matrix analysis is performed to investigate class-wise recognition behavior and identify common misclassification patterns between visually similar hand-object interaction categories.

Quantitative Results

LOPO Cross-Subject Performance

To evaluate the cross-subject generalization capability of the proposed framework, experiments are conducted using the Leave-One-Participant-Out (LOPO) evaluation protocol described in the previous section. The model is trained using data from nine participants and evaluated on the remaining unseen participant across all 10 folds.

The quantitative results are summarized in Table 2. The proposed framework achieves a mean accuracy of 82.36% and a mean F1-score of 81.65%, demonstrating strong recognition performance under the subject-independent evaluation setting. The framework also achieves high precision and recall values, indicating

stable classification performance across different interaction categories and participants.

The results demonstrate that the proposed dual-view framework effectively captures complementary interaction information from the palm and back views while maintaining robustness to variations in hand appearance, manipulation style and viewpoint conditions. Despite the challenging cross-subject evaluation setting, the framework maintains consistent recognition performance across most participants, highlighting the effectiveness of the proposed multi-level feature extraction and adaptive fusion strategy.

Comparison with Baseline Methods

To evaluate the effectiveness of the proposed framework, comparisons are conducted with several widely used lightweight convolutional neural network architectures, including MobileNetV2, MobileNetV3, ShuffleNetV2, Ghost-Net, EfficientNet-B0, MobileViT, SqueezeNet and MnasNet. All models are evaluated under the same LOPO cross-subject protocol using identical training settings and preprocessing strategies to ensure a fair comparison.

Table 3: Performance comparison between the proposed method and baseline convolutional neural network architectures in terms of mean accuracy (mAcc), mean F1-score (mF1), mean precision, mean recall and test loss under the LOPO cross-subject evaluation protocol

Method	mAcc (%)	mF1 (%)	mPrec (%)	mRec (%)	Loss
MobileNetV2	80.60 ± 10.60	79.58 ± 11.80	85.62 ± 6.16	80.82 ± 10.43	1.03 ± 0.27
MobileNetV3-Small	68.74 ± 12.27	66.73 ± 13.46	74.71 ± 9.06	68.99 ± 12.26	1.33 ± 0.28
MobileNetV3-Large	76.33 ± 11.75	75.06 ± 12.69	82.63 ± 8.14	76.57 ± 11.54	1.17 ± 0.32
ShuffleNetV2	75.04 ± 12.63	73.57 ± 14.11	79.58 ± 11.30	75.24 ± 12.43	1.16 ± 0.32
GhostNet	73.88 ± 16.95	71.99 ± 18.83	78.72 ± 13.32	74.21 ± 16.81	1.22 ± 0.44
EfficientNet-B0	77.75 ± 13.18	76.18 ± 14.72	83.08 ± 9.35	77.97 ± 13.05	1.11 ± 0.35
MobileViT	77.07 ± 12.29	75.73 ± 14.12	82.81 ± 6.61	77.40 ± 12.12	1.14 ± 0.34
SqueezeNet	75.35 ± 14.72	73.86 ± 16.24	78.93 ± 12.50	75.67 ± 14.57	1.20 ± 0.37
MnasNet	77.69 ± 10.77	76.83 ± 11.80	83.33 ± 7.31	77.94 ± 10.74	1.17 ± 0.26
Proposed Method	82.36 ± 8.61	81.65 ± 9.42	86.83 ± 4.55	82.53 ± 8.43	1.01 ± 0.27

Since the proposed framework operates on synchronized dual-view inputs, all baseline architectures are adapted to the same two-view setting. Feature representations extracted from the palm-view and back-view streams are combined using feature concatenation before classification.

The quantitative comparison results are summarized in Table 3. The proposed framework achieves the best overall performance among the evaluated lightweight architectures, obtaining a mean accuracy of 82.36%, a mean F1-score of 81.65% and a mean precision of 86.83%. In comparison, MobileNetV2 achieves 80.60% accuracy, while EfficientNet-B0 and MobileViT achieve 77.75% and 77.07%, respectively. The proposed framework also achieves the lowest test loss among most evaluated methods.

These results demonstrate that the proposed multi-level feature extraction and adaptive fusion strategy effectively improve recognition performance under the cross-subject evaluation setting. Despite its lightweight design, the proposed framework achieves superior recognition performance.

Computational Complexity Analysis

To evaluate the efficiency of the proposed framework, a computational complexity comparison is conducted using the number of parameters, FLOPs and inference latency. The results are summarized in Table 4. The proposed framework requires only 2.53M parameters and 0.65 GFLOPs, remaining comparable to lightweight architectures such as MobileNetV2 and Mnas-Net while being substantially less complex than EfficientNet-B0 and MobileViT. The framework achieves an average inference latency of 16.94 ms, supporting real-time execution for wearable hand-object interaction recognition. Although SqueezeNet achieves lower latency due to its compact architecture, its recognition performance is significantly lower. In contrast, models such as GhostNet and EfficientNet-B0 exhibit higher computational cost and inference latency without providing superior recognition performance. These results indicate that the proposed framework provides an effective trade-off between recognition accuracy and computational efficiency for lightweight dual-view HOI recognition.

Table 4: Computational complexity and efficiency comparison of the evaluated models in terms of parameters, FLOPs and inference latency

Method	Params	FLOPs	Latency
MobileNetV2	2.88	0.65	16.19 ± 2.00
MobileNetV3-Small	1.23	0.12	15.21 ± 0.90
MobileNetV3-Large	3.47	0.47	19.24 ± 3.20
ShuffleNetV2	1.78	0.30	23.91 ± 1.74
GhostNet	4.56	0.31	32.51 ± 1.89
EfficientNet-B0	4.67	0.77	32.06 ± 2.09
MobileViT	5.27	2.84	29.55 ± 3.73
SqueezeNet	0.99	0.53	6.16 ± 0.40
MnasNet	3.76	0.67	15.45 ± 0.65
Proposed Method	2.53	0.65	16.94 ± 1.30

Ablation Study

To evaluate the contribution of the different components of the proposed framework, an ablation study is conducted under the LOPO cross-subject evaluation protocol. The study investigates the impact of multi-level feature extraction and different feature fusion strategies while maintaining the same backbone architecture and training configuration across all variants.

The quantitative results are summarized in Table 5. Using only deep features achieves a mean accuracy of 79.64% and an F1-score of 78.87%, indicating that high-level semantic representations alone are insufficient to fully capture fine-grained interaction patterns. Incorporating both intermediate and deep features improves recognition performance, with concatenation and summation strategies achieving 80.36% and 81.63% accuracy, respectively.

Table 5: Ablation study evaluating the impact of multi-level feature extraction and fusion strategies on cross-subject recognition performance under the LOPO protocol

Variant	mAccuracy (%)	mF1-score (%)	Params (M)	Latency (ms)
Deep only	79.64	78.87	2.46	16.95 ± 1.99
Mid+Deep+Concat	80.36	79.92	2.50	17.58 ± 2.44
Mid+Deep+Sum	81.63	80.89	2.47	17.31 ± 2.59
Mid+Deep+Shared Gate	79.82	78.93	2.50	16.30 ± 1.46
Mid+Deep+Separate Gate(Proposed)	82.36	81.65	2.53	16.94 ± 1.30

The shared adaptive gating mechanism achieves 79.82% accuracy, suggesting that a single gating function is insufficient to model the distinct characteristics of the palm-view and back-view streams. In contrast, the proposed separate gating mechanism achieves the best overall performance with 82.36% accuracy and an F1-score of 81.65%, while maintaining low computational complexity and real-time inference latency.

These results demonstrate that the proposed view-specific adaptive fusion strategy enables more effective integration of complementary multi-level representations for dual-view hand-object interaction recognition.

Qualitative Analysis

Confusion Matrix Analysis

To further investigate the classification behavior of the proposed framework, the aggregated confusion matrix obtained from the 10-fold LOPO evaluation is presented in Figure 6. Most predictions are strongly concentrated along the main diagonal, indicating effective recognition performance across the majority of interaction categories.

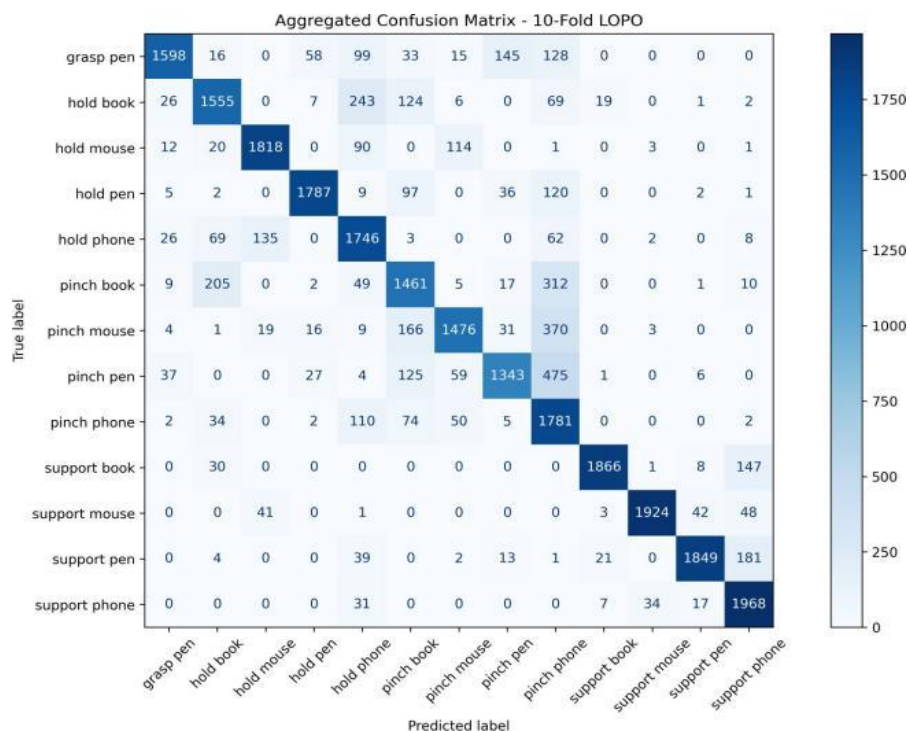


Figure 6: Aggregated confusion matrix obtained from the 10-fold Leave-One-Participant-Out (LOPO) evaluation, illustrating the classification performance of the proposed method across the 13 hand object

interaction categories, with correct predictions concentrated along the main diagonal

Several categories exhibit particularly high recognition accuracy. For example, *support mouse* (1924), *support phone* (1968) and *support book* (1866) achieve a large number of correct predictions. Similarly, strong recognition performance is observed for *hold mouse* (1818), *hold pen* (1787) and *hold phone* (1746). These interactions are generally characterized by more stable hand configurations and clearer object visibility.

Despite the overall strong performance, several informative misclassification patterns are observed. In particular, confusion occurs among visually similar pinch-related interactions, especially between *pinch pen* and *pinch phone* (475 samples), as well as between *pinch mouse* and *pinch phone* (370 samples). Additional confusion is observed between *hold* and *pinch* interactions involving the same object category, such as *pinch book* misclassified as *hold book* (205 samples). These patterns reflect the fine-grained nature of the task and the difficulty of distinguishing subtle differences in finger articulation and object appearance under partial occlusion.

Overall, the confusion matrix demonstrates that the proposed framework effectively distinguishes most interaction categories while maintaining robustness under cross-subject and multi-view evaluation conditions.

Feature Representation Analysis

To further analyze the learned feature representations, a t-SNE visualization is conducted using feature embeddings extracted from the proposed model at different training stages under the LOPO evaluation protocol. The visualization includes samples from all 13 interaction categories and is illustrated in Figure 7.

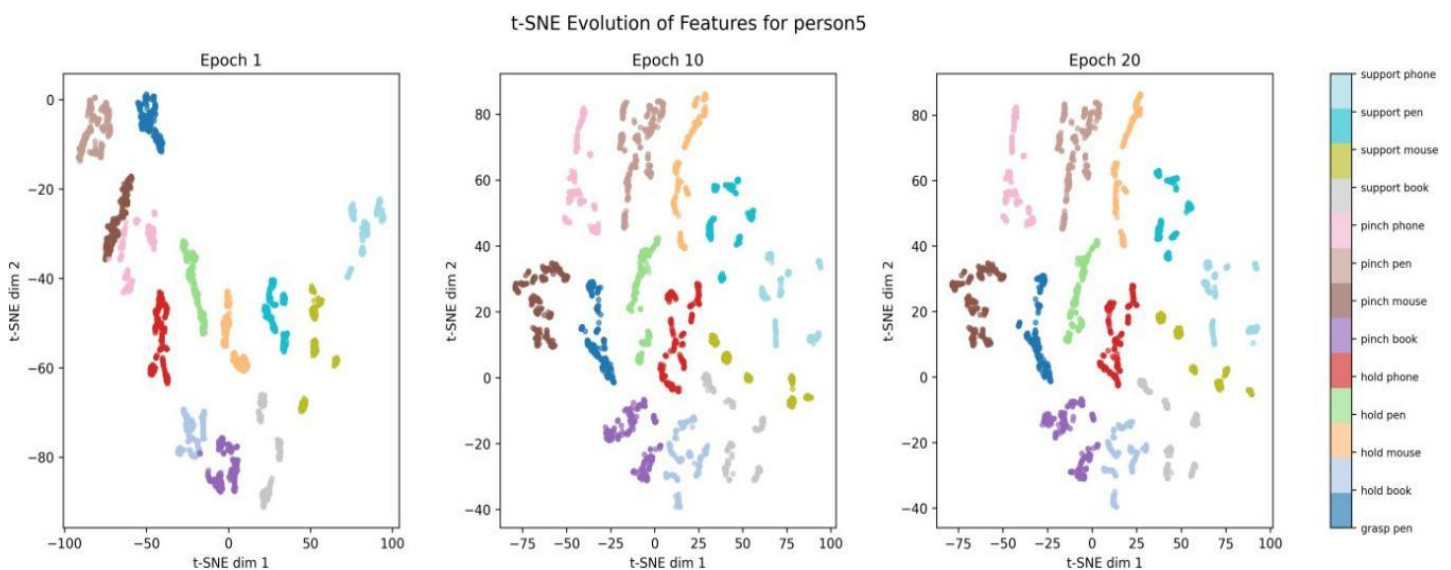


Figure 7: t-SNE visualization of the learned feature representations at different training stages: Epoch 1, Epoch 10 and Epoch 20 for a representative subject under the LOPO evaluation protocol

At Epoch 1, the feature representations already exhibit an initial level of class separation, indicating that the framework quickly captures meaningful interaction patterns from the dual-view inputs. However, several clusters remain relatively dispersed with partial overlap between visually similar interaction categories.

As training progresses, the feature distributions become progressively more structured and compact. At Epoch 10, clearer separation between interaction categories is observed, indicating improved discriminative representation learning. By Epoch 20, the feature embeddings form well-defined and compact clusters with reduced inter-class overlap, demonstrating that the proposed framework learns semantically meaningful representations for fine-grained hand-object interaction recognition.

Scalability Across Backbone Architectures

To further investigate the scalability of the proposed framework, additional experiments are conducted by replacing the lightweight MobileNetV2 backbone with the higher-capacity ConvNeXtV2-Tiny architecture while preserving the proposed multi-level feature extraction and adaptive fusion strategy. In addition, the standalone ConvNeXtV2-Tiny backbone is evaluated under the same LOPO protocol to analyze the contribution of the proposed framework beyond the backbone architecture itself.

The results are summarized in Table 6. The standalone ConvNeXtV2-Tiny backbone achieves a mean accuracy of 86.95% and a mean F1-score of 86.25%, demonstrating strong capability for fine-grained hand-object interaction recognition. When integrated into the proposed framework, the performance further improves to 87.84% mean accuracy and 86.98% mean F1-score, highlighting the effectiveness of the proposed multi-level feature extraction and adaptive fusion strategy in exploiting complementary dual-view information. Compared with the lightweight MobileNetV2-based configuration, the ConvNeXtV2-Tiny-based framework achieves higher recognition performance, but with increased computational complexity, as the number of parameters rises from 2.53M to 28.07M. These results demonstrate that the proposed framework is flexible and scalable across different backbone architectures, enabling adaptation to various computational and performance requirements.

CONCLUSION

This paper presented a lightweight dual-view framework for hand-object interaction recognition using synchronized palm-view and back-view observations acquired from a wrist-worn device. The proposed approach combines multi-level feature extraction and adaptive feature fusion to effectively capture both fine-grained interaction details and high-level semantic information from complementary viewpoints.

To improve representation learning, the framework integrates intermediate and deep features extracted from a shared MobileNetV2 backbone and employs a view-specific adaptive gating mechanism to dynamically balance their contributions. Experimental results under the Leave-One-Participant-Out (LOPO) cross-subject evaluation protocol demonstrate that the proposed framework achieves strong recognition performance while maintaining low computational complexity suitable for real-time wearable applications.

The ablation study confirms the effectiveness of the proposed multi-level fusion strategy and the importance of the separate adaptive gating mechanism for palm-view and back-view streams. Additional experiments further demonstrate the scalability of the framework across different backbone architectures.

Overall, the proposed framework provides an effective and computationally efficient solution for dual-view hand-object interaction recognition and shows strong potential for future wearable intelligent systems, human-computer interaction and assistive technology applications.

Table 6: Performance comparison of the standalone ConvNeXtV2-Tiny backbone and the proposed dual-view framework with different backbone architectures under the LOPO protocol

Configuration	mAcc (%)	mF1 (%)	Params (M)
ConvNeXtV2-Tiny	86.95	86.25	28.26
Proposed (MobileNetV2 backbone)	82.36	81.65	2.53
Proposed (ConvNeXtV2 backbone)	87.84	86.98	28.07

REFERENCES

1. Ohn-Bar E, Trivedi M M. Hand gesture recognition in real time for automotive in- terfaces[J].

IEEE Transactions on Intelligent Transportation Systems, 2014, 15(6): 2368–2377.

2. Cheng H, Yang L, Liu Z. Survey on 3D hand gesture recognition[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015, 26(9): 1659–1673.
3. Fan H, Zhuo T, Yu X, et al. Understanding atomic hand-object interaction with human intention[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, 32(1): 275–285.
4. Chung H Y, Chung Y L, Tsai W F. An efficient hand gesture recognition system based on deep CNN[C]// *IEEE International Conference on Industrial Technology*. IEEE, 2019: 853–858.
5. Lin H I, Hsu M H, Chen W K. Human hand gesture recognition using a convolution neural network[C]// *IEEE International Conference on Automation Science and Engineering*. IEEE, 2014: 1038–1043.
6. Li G, Tang H, Sun Y, et al. Hand gesture recognition based on convolution neural network[J]. *Cluster Computing*, 2019, 22(S2): 2719–2729.
7. Ozcan T, Basturk A. Transfer learning-based convolutional neural networks with heuristic optimization[J]. *Neural Computing and Applications*, 2019, 31(12): 8955–8970.
8. Sahoo J P, Prakash A J, Pl-awiak P, et al. Real-time hand gesture recognition using fine-tuned convolutional neural network[J]. *Sensors*, 2022, 22(3): 706.
9. Tran D S, Ho N H, Yang H J, et al. Real-time hand gesture spotting and recognition using RGB-D camera and 3D convolutional neural network[J]. *Applied Sciences*, 2020, 10(2):722.
10. Mahmud H, Morshed M M, Hasan M K. A deep learning-based multimodal depth-aware dynamic hand gesture recognition[EB/OL]. arXiv:2107.02543, 2021.
11. Ishihara T, Kitani K M, Ma W C, et al. Recognizing hand-object interactions in wearable camera videos[C]// *IEEE International Conference on Image Processing*. IEEE, 2015: 1349–1353.
12. Tekin B, Bogo F, Pollefeys M. Unified egocentric recognition of 3D hand-object poses[C]// *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2019: 4511–4520.
13. Garcia-Hernando G, Yuan S, Baek S, et al. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations[C]// *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018: 409–419.
14. Ahmad A, Migniot C, Dipanda A. Tracking hands in interaction with objects: A review[C]// *International Conference on Signal-Image Technology and Internet-Based Systems*. IEEE, 2017: 360–369.
15. Romero J, Kjellström H, Kragic D. Hands in action: Real-time 3D reconstruction of hands[C]// *IEEE International Conference on Robotics and Automation*. IEEE, 2010: 458–463.
16. Hamer H, Schindler K, Koller-Meier E, et al. Tracking a hand manipulating an object[C]// *IEEE International Conference on Computer Vision*. IEEE, 2009: 1475–1482.
17. Kang B, Tan K H, Jiang N, et al. Hand segmentation for hand-object interaction from depth map[C]// *IEEE Global Conference on Signal and Information Processing*. IEEE, 2017: 259–263.
18. Sridhar S, Mueller F, Zollhöfer M, et al. Real-time joint tracking of a hand manipulating an object[C]// *European Conference on Computer Vision*. Springer, 2016: 294–310.
19. Cai M, Kitani K M, Sato Y. Understanding hand-object manipulation with grasp types and object attributes[C]// *Robotics: Science and Systems*. 2016.
20. Bertasius G, Park H S, Yu S X, et al. First person action-object detection with egonet[EB/OL]. arXiv:1603.04908, 2016.
21. Schroder M, Ritter H. Hand-object interaction detection with fully convolutional networks[C]// *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2017: 18–25.
22. Yan W, Gao Y, Liu Q. Human-object interaction recognition using multitask neural network[C]// *International Symposium on Autonomous Systems*. IEEE, 2019: 323–328.
23. Kwon T, Tekin B, Stühmer J, et al. H2O: Two hands manipulating objects for interaction recognition[C]// *IEEE International Conference on Computer Vision*. IEEE, 2021: 10138–

10148.

24. K`opu`klu` O, Gunduz A, Kose N, et al. Real- time hand gesture detection and classification using convolutional neural networks[C]*IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019: 1–8.
25. Mujahid A, Awan M J, Yasin A, et al. Real- time hand gesture recognition based on deep learning YOLOv3 model[J]. *Applied Sciences*, 2021, 11(9): 4164.
26. Lai K, Yanushkevich S N. CNN+RNN depth and skeleton based dynamic hand gesture recognition[C]// *International Conference on Pattern Recognition*. IEEE, 2018: 3451–3456.
27. Pigou L, Van Den Oord A, Dieleman S, et al. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recog- nition[J]. *International Journal of Computer Vision*, 2018, 126(2): 430–439.
28. Molchanov P, Gupta S, Kim K, et al. Hand gesture recognition with 3D convolutional neural networks[C]// *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2015: 1–7.
29. Zhang L, Zhu G, Shen P, et al. Learning spa- tiotemporal features using 3DCNN and Con- vLSTM for gesture recognition[C]// *IEEE International Conference on Computer Vi- sion Workshops*. IEEE, 2017: 3120–3128.
30. Gao Q, Chen Y, Ju Z, et al. Dynamic hand gesture recognition based on 3D hand pose estimation[J]. *IEEE Sensors Journal*, 2021, 22(18): 17421–17430.
31. Miah A S M, Hasan M A M, Shin J. Dynamic hand gesture recognition using graph neural networks[J]. *IEEE Access*, 2023, 11: 4703–4716.
32. Sun S. A survey of multi-view machine learn- ing[J]. *Neural Computing and Applications*, 2013, 23(7): 2031–2038.
33. Shukla D, Erkent O` , Piater J. A multi-view hand gesture RGB-D dataset for human- robot interaction scenarios[C]// *IEEE Inter- national Symposium on Robot and Human In- teractive Communication*. IEEE, 2016: 1084–1091.
34. Wang L, Ding Z, Tao Z, et al. Generative multi-view human action recognition[C]// *IEEE International Conference on Computer Vision*. IEEE, 2019: 6212–6221.
35. Zhang Z, Wang C, Xiao B, et al. Cross-view action recognition using contextual maximum margin clustering[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, 24(10): 1663–1668.
36. Arnold E, Dianati M, De Temple R, et al. Cooperative perception for 3D object detec- tion in driving scenarios[J]. *IEEE Transac- tions on Intelligent Transportation Systems*, 2020, 23(3): 1852–1864.
37. Teepe T, Wolters P, Gilg J, et al. EarlyBird: Early fusion for multi-view tracking in bird’s- eye view[C]// *IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 2024: 102–111.
38. Gao Y, Maggs M. Feature-level fusion in per- sonal identification[C]// *IEEE Computer So- ciety Conference on Computer Vision and Pattern Recognition*. IEEE, 2005: 468–473.
39. Fadadu S, Pandey S, Hegde D, et al. Multi-view fusion of sensor data for im- proved perception in autonomous driv- ing[C]// *IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 2022: 2349–2357.
40. Seeland M, M`ader P. Multi-view classification with convolutional neural networks[J]. *PLoS One*, 2021, 16(1): e0245230.
41. Cheng J, Yin W, Wang K, et al. Adaptive fu- sion of single-view and multi-view depth for autonomous driving[C]// *IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*. IEEE, 2024: 10138–10147.
42. Zheng D, Zheng X, Yang L T, et al. Multi-view feature fusion network for cam- ouflaged object detection[C]// *IEEE/CVF Winter Conference on Applications of Com- puter Vision*. IEEE, 2023: 6232–6242.
43. Ezati A, Dezyani M, Rana R, et al. A lightweight attention-based deep network via multi- scale feature fusion for multi- view facial expression recognition[EB/OL]. arXiv:2403.14318, 2024.
44. He K, Zhang X, Ren S, et al. Deep resid- ual learning for image recognition[C]// *IEEE*

Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770–778.

45. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017: 2117–2125.