InterHandNet: Capturing Two-hand Interaction for Robust Hand-washing Activity Recognition

Yiqing Zhang
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
zhang.yiqing@ist.osaka-u.ac.jp

Takuya Maekawa
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
maekawa@ist.osaka-u.ac.jp

***Abstract*—This study proposes a new deep learning method for hand-washing activity recognition using a series of hand skeleton data extracted from an RGB-D camera. Assessment of hand-washing activity based on recognized hand-washing steps is crucial in both industrial and medical domains, as well as in promoting healthy habits. However, recognizing hand-washing activities presents unique challenges compared to typical activity recognition for a single person due to the specific nature of hand-washing tasks. First, the steps of hand-washing can be better explained by the interaction between objects, i.e., the two hands, such as rubbing palms and fingers. Second, occlusion occurs much more frequently during hand-washing due to the frequent interaction between both hands. Therefore, we propose a new neural network tailored for hand-washing recognition called InterHandNet to address these challenges. To capture the interaction, we propose two novel modules in InterHandNet: Interaction Graph and Interaction Attention. These modules enable to exchange information across skeleton graphs of the two hands within a graph neural network framework and to focus on important keypoints in one hand by referencing the other hand through the query-key-value mechanism, respectively. To address the issue of missing data caused by occlusion, we propose Inter-hand Temporal Fusion, which fills in the missing information by referencing data from the other hand and other time steps within a time window. InterHandNet outperforms other state-of-the-art skeleton-based and RGB-based methods in terms of accuracy, and significantly surpasses RGB-based methods in runtime efficiency on edge devices.**

***Index Terms*—Hand-washing activity recognition, hand skeleton, RGB-D camera**

I. Introduction

1) Background and Goal: With the proliferation and performance enhancement of edge devices with rich sensors such as RGB-D sensors, collecting and processing such data have become more economical and efficient. For example, devices such as Intel RealSense, Microsoft Azure Kinect DK, and Luxonis OAK-D Depth AI Camera are capable of capturing both visual and depth features due to the integrated sensors and power units. Therefore, in the pervasive computing community, human activity recognition (HAR) using skeleton data from such edge devices has become an active research topic [1]–[5], as using optimized skeleton extractor like Mediapipe Hands [6] is more computationally efficient than directly using RGB data, which is demonstrated in the evaluation section. Because analyzing and processing data extracted from activities have many applications in various fields, including the industrial, sports, and medical domains [2], [7], [8], researchers have attempted to recognize activities performed in the domains. After the COVID-19 pandemic, the importance of maintaining healthy habits has become increasingly apparent to the public. Among them, hand-washing is one of the most effective measures to prevent viral transmission, and has significantly contributed to protecting both ourselves and others. Furthermore, hand-washing is crucial in both industrial and medical domains, where high safety standards are required. For example, in many food factories, the employees are required to wash their hands according to prescribed steps. Hand-washing is performed only at a specific place (basin) in many industrial sites such as food factories, indicating its small installation costs. Moreover, although recent work activity recognition methods in industrial domains employ wearable sensors such as smartwatches [9]–[14], in many food factories, wearing accessories such as watches is prohibited for hygiene reasons. Additionally, privacy concerns can be addressed since the camera captures only the hands. Using cameras is also more feasible than wearable devices, especially in factory environments, where ease of installation and maintenance is crucial. So, installing edge devices that automatically recognize/assess hand-washing activities in hand-washing stations is a reasonable solution.

Therefore, the goal of this study is to recognize hand-washing activities by using a sequence of hand skeletons extracted from RGB-D images. As shown in Fig. 1, the World Health Organization (WHO) defined six steps in hand-washing. Thus, in this study, we classify the hand skeletons within a given window size (which typically corresponds to the frames per second of the camera) into one of the six steps. As shown in the examples in Fig. 1, frequent rubbing between the two hands helps clean them more effectively.

2) Challenges: In the field of activity recognition, analyzing the movements of the whole body using extracted skeleton data of the body has been a popular choice [8], [15]–[27]. However, the direct application of networks designed for HAR [18], [19], [21]–[27] to hand-washing recognition cannot fully capture the specific characteristics of hand-washing. Observing each step of hand-washing reveals two significant differences compared to typical human activities. The first difference

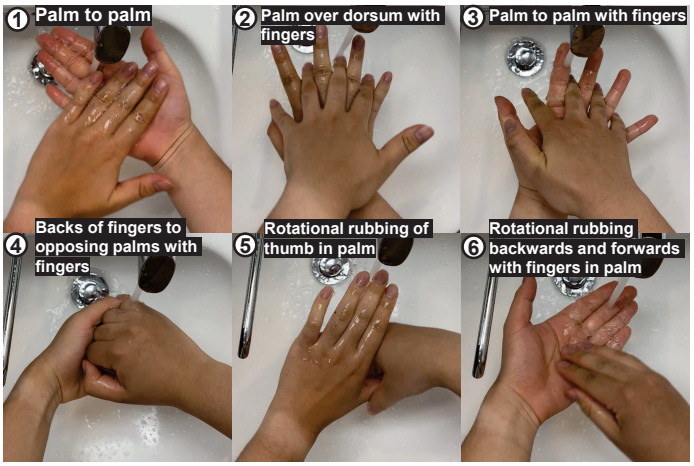

Fig. 1: Six steps of hand-washing defined by the World Health Organization (WHO)

is that HAR primarily focuses on individual movements. However, the steps of hand-washing can be better explained by the interaction between both hands rather than the separate movements of the two hands, as shown in Fig. 1. In other words, the direct application of the methods of HAR cannot consider the interaction between both hands, which is crucial for effective hand-washing recognition. Thus, effectively incorporating the interaction in network design has become the primary challenge.

The second difference is that occlusion occurs much more frequently in hand-washing as shown in Fig. 1, due to the frequent interaction between both hands, compared to typical HAR scenarios. The occlusion poses a significant challenge to skeleton detection models; thus, the ability to infer hidden hand skeleton data is crucial. Although several state-of-the-art skeleton detection methods [28]–[30] exist that are robust to occlusion, their complexity makes them difficult to implement on edge devices. Mediapipe Hands [6] has achieved a great balance between efficiency and performance, making it seem like the best choice for hand-washing recognition. However, its accuracy significantly decreases under extreme occlusion conditions such as hand-washing. Thus, designing a network for hand-washing recognition to enhance the adaptability to occlusion has become the second challenge.

In fact, these two challenges are closely connected. Frequent interaction causes significant occlusion, and both challenges arise from the specific requirements of hand-washing recognition. These factors also distinguish hand-washing recognition from typical HAR tasks. This is why a new model specifically designed for hand-washing recognition is necessary to effectively handle both interaction and occlusion.

3) Approaches: We propose a new neural network tailored to hand-washing recognition called **InterHandNet** to address these two challenges. "Inter" signifies interaction, highlighting the primary challenge and the corresponding approach. This method leverages the interaction between the two hands during hand-washing recognition and addresses the incomplete skeletons caused by occlusion, thereby improving hand-washing recognition accuracy.

To capture the interaction, we propose two novel modules in InterHandNet: **Interaction Graph** and **Interaction Attention**. i) For the Interaction Graph, we extend the graph structure from individual hands to both hands. In recent GCN (Graph Convolutional Network)-based method designs [18], the scope of the graph typically considers a single object, such as one person in HAR. To capture the two-hand interaction, we aim to establish a connection between the two objects (two hands) in order to capture their interaction. In other words, we introduce edges between the two graphs of the hands to enhance the expressive power of the graph neural networks, enabling them to capture the complex two-hand interaction.

ii) The Interaction Attention builds upon the outstanding attention [31] mechanism. One major characteristic of the attention mechanism is its ability to automatically focus on the relevant regions of input data. In InterHandNet, we expand the scope to both hands, allowing each hand to automatically identify the relevant features of another hand in order to better recognize the hand-washing steps because the shape of one hand is an important clue to understand that of the other hand in hand washing involving a variety of two-hand interactions. We employ the three main components (Query, Key, Value) of the attention mechanism to understanding the hand-washing features. Specifically, we treat one hand feature as the Query, and the other hand feature as the Key and Value. The product of one hand feature (Query) and the other hand feature (Key) represents the relevance between the two hands. Then, the result of multiplying with the other hand feature (Value) denotes the hand feature fused with the most relevant information from the other hand. In that case, a connection is established between the two hands during hand-washing recognition.

iii) To address the issue of missing data caused by occlusion, we design an additional module named **InterHand Temporal Fusion**. We expand the scope of temporal convolution in ST-GCN (Spatio-Temporal Graph Convolution Network)-based methods to include both hands, rather than just one, which allows us to capture the relationship between the two hands over time during hand-washing. By leveraging the relationship information between both hands along with the data from one hand, it is possible to infer the missing information from the other hand. The inference also relies on the attention mechanism, enabling the neural network to automatically learn which temporal information is most helpful in filling the missing data between the two hands during the training period.

4) Contributions: (i) We proposed a novel method named InterHandNet, specifically optimized for hand-washing recognition. InterHandNet is modular and generalizable, enabling its core mechanisms (Interaction Graph, Interaction Attention, and Inter-hand Temporal Fusion) to be integrated into most STGCN-based methods [18], [19], [21], [22], [24], [25], thereby enhancing their performance in hand-washing recognition tasks. (ii) We designed the Interaction Graph and Interaction Attention mechanisms to leverage hand interactions

in hand-washing recognition. Additionally, we proposed the Inter-hand Temporal Fusion to address the issue of missing data caused by occlusion in the skeleton extractor.

II. Related work

HAR based on video data has been a popular research topic due to its wide range of applications. Generally speaking, activity recognition uses features from the video, which can be divided into two major approaches. The first approach involves directly using the visual features extracted from the video [32], [33] and then applying temporal modeling, such as long short-term memory (LSTM), for activity recognition. The second approach involves using the skeleton data [1], [3], [4], extracted from video frames for activity recognition. Between the two approaches, skeleton-based methods have the advantages of being less affected by variations like appearance, and are capable of capturing 3D features using depth sensors. Specifically, hand-washing activity recognition is a subset of HAR, which primarily focuses on capturing the shape and motion of hands to recognize gestures or activities.

A. Skeleton-based Human Activity Recognition

Human skeletons can be represented by joints and bones, serving as abstractions for capturing human posture information. Compared to RGB frames, the amount of information is significantly smaller, as the skeleton representation utilizes only three coordinates per joint for three dimensions. With the introduction of the Graph Convolutional Network (GCN) [34], the structural similarity between graphs and skeletons has made GCN a popular choice for processing skeleton data, as the joints and bones of skeletons can be treated as the nodes and edges of graphs.

Yan et al. [18] introduced the Spatial Temporal-Graph Convolutional Network (ST-GCN), a deep learning-based method that connects adjacent graphs at different time steps and brings GCN into the field of skeleton-based HAR. They innovatively considered multiple graphs from different frames together to capture the motion information, such as acceleration and velocity, over a period of time, which is essential for activity recognition. Later, Shi et al. [19] introduced adaptive graphs, which improved the accuracy of recognition. Afterwards, more and more methods [21]–[27] based on ST-GCN were introduced, further improving the performance in skeleton-based activity recognition tasks. However, as mentioned in the introduction, the direct application of these methods to the hand-washing recognition cannot fully capture the specific characteristics of hand-washing.

B. Hand-washing Activity Recognition

In the field of IMU-based hand-washing recognition, Burchard et al. [35] employed an LSTM-based neural network to distinguish between routine hand-washing and other activities. Their subsequent work [43] introduced Multi-modal Atmospheric Sensing, which further improved the use of IMU data for activity recognition. Additionally, Wang et al. [44] and Li et al. [45] utilized wearable devices to recognize specific steps of the hand-washing process. Beyond IMU-based methods, several studies, including those by Zhong et al. [38], Kim et al. [37], Cikel et al. [46], Xie et al. [41], Vo et et al. [47] and Wang et al. [48] have focused on recognizing specific steps of hand-washing as defined by the WHO; however, their approaches use RGB and thermal videos as input, which contain significantly more information than skeleton data, making them prohibitive to deploy on edge devices for real-time recognition.

TABLE I: Comparison with different hand-washing datasets

Dataset	Format	Number of videos	Steps guided by WHO
Burchard et al. [35]	IMU	-	-
Kinetics [36]	RGB	916	-
Kim et al. [37]	RGB	176	1
Zhong et al. [38]	RGB	2055	1
Bakshi [39]	RGB	162	3
ICU-MH [40]	**Skeleton**	168	**6**
Xie et al. [41]	RGB	656	**6**
Lulla et al. [42]	RGB	**3185**	**6**

Most similar to our research is the work by Huang et al. [40], where they used the same skeleton extractor, as we did and their method is skeleton-based as well. However, one limitation of their approach is that they did not consider the temporal features during hand-washing, which we believe is a crucial aspect in any tasks of HAR. Instead, they input the skeleton data from each frame into basic classifiers like Support Vector Machines (SVM) to recognize each hand-washing step.

C. Two-hand Interaction

Several previous studies have explored the relationship between the two hands and leveraged attention mechanisms for integration. For instance, Li et al. [49], Li et al. [50] and Yu et al. [51] utilized attention-based methods to model inter-hand interactions, primarily for the task of two-hand reconstruction from an RGB image. In contrast, our study targets hand-washing activity recognition, which involves not only modeling inter-hand interactions but also capturing temporal dynamics. Unlike the hand reconstruction, the temporal dimension is crucial for recognizing complex actions, as it allows the model to understand the sequential nature of hand movements over time. Moreover, integrating temporal information across both hands facilitates the recovery of missing data caused by occlusion. These features make our approach particularly well-suited for dynamic tasks such as hand-washing, where actions unfold across time and often involve overlapping or partially occluded hand movements.

D. Hand-Washing Activity Recognition Datasets

The details of datasets related to hand-washing are shown in Table I. Most of the existing hand-washing datasets are based on videos, while Burchard et al. [35] used IMU data, and Huang et al. [40] employed a skeleton extractor to extract skeleton data from the video dataset. Although the "washing hands" action is included in large HAR datasets, like Kinetics [36], the label typically only indicates whether hand-washing is occurring, without subdividing the process into distinct

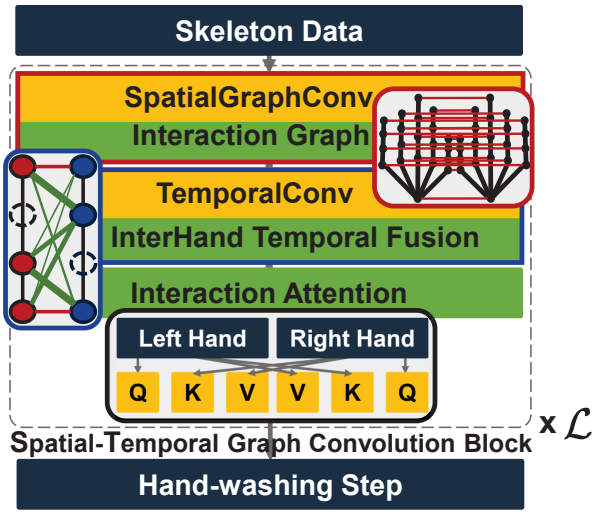

Fig. 2: Architecture of InterHandNet. InterHandNet comprises $\mathcal{L}$ STGC (Spatial-Temporal Graph Convolution) blocks, each integrating Interaction Graph, InterHand Temporal Fusion, and Interaction Attention. Following the STGC blocks, a fully connected layer predicts the action class.

steps. With the spread of the COVID-19 pandemic, researchers like Kim et al. [37] and Zhong et al. [38] began to realize the importance of recognizing the steps of hand-washing, which should follow the guidelines of an authoritative organization. Therefore, they chose to annotate their datasets based on the guidelines provided by the World Health Organization (WHO). Afterwards, Bakshi [39], Huang et al. [40], Xie et al. [41], and Lulla et al. [42] pushed the number of hand-washing step annotations up to six steps. Among these datasets, Lulla et al. [42] provided the largest-scale and real-world hand-washing dataset, labeled according to World Health Organization (WHO) guidelines, which was collected in hospitals during the COVID-19 pandemic.

III. InterHandNet

A. Overview

To achieve hand-washing recognition, we designed a pipeline consisting of three main steps. First, we use an RGB-D camera to capture depth video of the hand-washing process. Then, we employ a skeleton extractor to extract 3D skeleton data from the video. Finally, our approach, InterHandNet, is applied to identify the hand-washing steps.

Figure 2 shows the main architecture of InterHandNet. InterHandNet features three novel modules designed to enhance the accuracy of hand-washing recognition. InterHandNet consists of $\mathcal{L}$ STGC (Spatial-Temporal Graph Convolution) blocks, where $\mathcal{L}$ corresponds to the number of blocks in the backbone network. Each STGC block is composed of the spatial graph convolution integrated with Interaction Graph, the temporal convolution integrated with InterHand Temporal Fusion, and Interaction Attention that processes outputs of the temporal convolution. After passing through several STGC blocks, the action class is predicted by a fully connected layer. Moreover, these modules including **Interaction Graph**, **InterHand Temporal Fusion**, and **Interaction Attention** can be easily integrated into most STGCN-based methods. We explain these three modules one by one after formulating this problem.

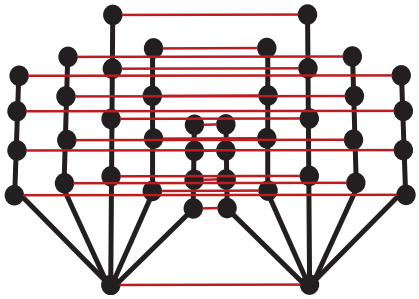

Fig. 3: Structure of original two-hand graphs and Interaction Graph

B. Data and Notation Definitions

The input of InterHandNet is a window of time-series 3D skeleton data f_0 extracted from Mediapipe Hands, with the shape of $3 \times \mathcal{T} \times N$ where $\mathcal{T}$ denotes the length of the window and N denotes the number of nodes. A normalized $N \times N$ adjacency matrix A that represents the physical structure of the two hands is also fed into InterHandNet, which is depicted in Fig. 3 as black nodes and edges. Here, $N/2 = 21$, meaning that each hand has 21 keypoints.[1] The input node feature of the l-th STGC block is f_l, with the shape of $C \times \mathcal{T} \times N$, where C denotes the number of channels. f_{l+1} is the output of the l-th STGC block and becomes the input for $(l+1)$-th STGC block. The node feature at time t within f_l is denoted as $f_{l,t}$. The features f_l^S, f_l^T and f_l^A correspond to the outputs of the spatial graph convolution, temporal convolution and Interaction Attention, respectively, for the l-th STGC block. The node features for the left and right hands within f_l^S are represented as $f_l^{S,L}$ and $f_l^{S,R}$, respectively. The output of InterHandNet is an action class estimate for the input window.

C. Interaction Graph

Figure 3 illustrates how the Interaction Graph is constructed on a visual level. The combination of black nodes and lines represents the structure of a hand and the adjacency matrix A describes the connections, while the red lines, which connect the two hands, form the Interaction Graph. The connection method between the two hands follows a pattern of fingers to fingers, palms to palms, and corresponding parts from each hand. These connections capture information, including the distance between the two hands, providing additional reference for the network to determine the current step of the hand-washing process. From the data perspective, Fig. 4 shows the data flow, where a circle denotes the left/right hand features $C \times N/2$ at a specific time step, and the direction of the

[1]For each hand, the keypoints are indexed starting from the palm, followed by the base of the thumb to its tip, and ending with the pinky finger.

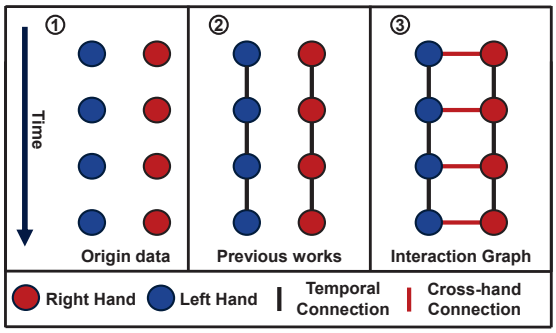

Fig. 4: Data flow under the Interaction Graph module. (1) The original data for four frames. (2) Recent approaches, such as ST-GCN, connect features across time steps for action recognition. (3) After applying the Interaction Graph module, features between hands are exchanged to capture inter-hand interactions during hand-washing.

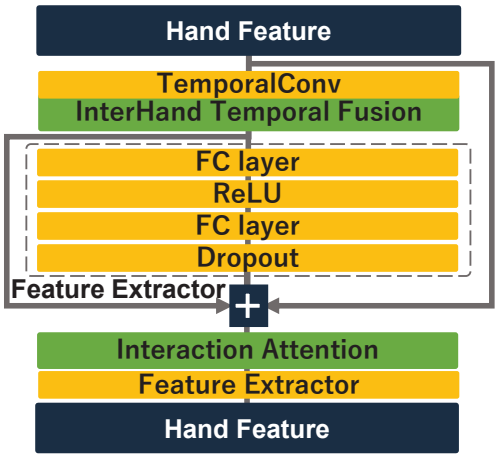

Fig. 5: Overviews of InterHand Temporal Fusion and InterHand Attention in InterHandNet

arrow indicates the direction of time. Fig. 4 (1) illustrates the original data for four frames, where the blue circles represent the left hand features, and the red circles represent the right hand features. The same row indicates that both hands appear at the same time step. In recent approaches, starting from ST-GCN, features are connected across different time steps to recognize action classes. This can be seen in Fig. 4 (2) in the context of hand-washing recognition. Fig. 4 (3) shows the data flow after applying the Interaction Graph module, where the features of each hand can be exchanged during hand-washing. The details of the connections are already shown in Fig. 3.

Before explaining the equation of Interaction Graph, we introduce the standard spatial graph convolution in STGCN-based methods, which can be generally summarized as follows:

$$f_{l,t}^{S} = W f_{l,t}(A + M), \tag{1}$$

where W is a $C_{l+1} \times C_l$ learnable weight matrix; M is a $N \times N$ matrix that can be interpreted flexibly depending on the specific method used. For ST-GCN, M can be explained as an $N \times N$ matrix that indicates the importance of each vertex; in 2s-AGCN [19], M can be explained as the sum of a similar importance matrix of ST-GCN and the data-dependent graph, which learns a unique graph for each sample; in MS-AAGCN [21], weighting factors are added to M.

To leverage interaction during hand-washing recognition, we change Eq. (1) to spatial interaction graph convolution as:

$$f_{l,t}^{S} = W f_{l,t}(A_{IG}D + A + M). \tag{2}$$

The primary modification involves adding $A_{IG}D$ to the original Eq. (1), where A_{IG} represents the $N \times N$ adjacency matrix of the Interaction Graph, and D denotes the distance matrix at time t. Retaining the form of Eq. (1) significantly enhances the compatibility of our approach, allowing the interaction graph to be easily integrated into a spatial graph convolution of any STGC-based methods following Eq. (1), such as CTR-GCN [24] and FR Head [25].

A spatial graph is constructed from an undirected graph $G = (V, E)$. In previous skeleton-based HAR methods [18], [19], [21]–[27], if there are multiple objects, for example, two, G is divided into two separate graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. In these two graphs, the node sets $V_j = \{v_{ji} \mid j = 1, 2, i = 1, \ldots, N/2\}$ and edge sets $E_j = \{e_{jik} \mid j = 1, 2, i = 1, \ldots, N/2, k = 1, \ldots, N/2\}$ are connected within each object.

In contrast, in the Interaction Graph $IG = (V, E)$, the edge set is defined as $E = \{e_{ik} \mid i = 1, \ldots, N/2, k = i + N/2\}$, establishing connections between the two objects as depicted as the red edges in Fig. 3. To facilitate better interaction between the two graphs (hands), we introduce the distance matrix D, with the shape of $(N \times N)$, that contains the Euclidean distance in the 3D space between two corresponding keypoints in the different hands, i.e., the length of the red edge in Fig. 3. Therefore, the product $A_{IG}D$ contains the distance information between the two hands. $Wf A_{IG}D$ represents the interaction, including distance information between the two hands, and this additional information helps InterHandNet recognize the steps of hand-washing more accurately.

D. InterHand Temporal Fusion

InterHand Temporal Fusion fills in missing information by referencing data from the other hand and other time steps within a time window. Similar to Interaction Graph, InterHand Temporal Fusion can be integrated into most STGCN-based methods by simply modifying each original temporal convolution as shown in Fig. 2, leading to improved accuracy in hand-washing recognition. Fig. 5 shows how the InterHand Temporal Fusion module is added to the InterHandNet in detail. First, the hand feature output by the spatial graph convolution is passed into the temporal convolution with InterHand Temporal Fusion. Then, the feature undergoes a combination of fully connected (FC) layers, ReLU, and dropout layers to extract the feature (Feature Extractor), which is then element-wise added to the original hand feature to produce the input to the Interaction Attention module.

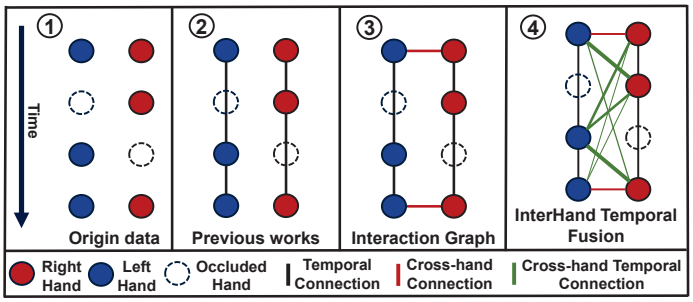

Fig. 6: Occluded data flow under the Interaction Graph and InterHand Temporal Fusion module. (1) An example where only one hand skeleton is extracted in some frames. (2) Previous STGCN-based methods connect hand features across time steps (black lines) but ignore occluded hand features. (3) The Interaction Graph module cannot fully utilize its potential when only one hand is detected. (4) After applying InterHand Temporal Fusion, connections are extended across all frames within the window, enabling interactions between two hands even in occlusion scenarios.

Due to the complex environment of hand-washing, not all hand coordinates can be detected by a skeleton extractor in every frame, resulting in missing node features in occluded nodes. For a clear illustration, we provide Fig. 6, which is similar to the previous Fig. 4. Fig. 6 shows the data composition under occlusion, where the dashed circle indicates that the hand cannot be detected by a skeleton extractor; Fig. 6 (1) shows an example of an extreme condition where, in some frames, only one hand skeleton can be extracted; Fig. 6 (2) shows the previous STGCN-based methods, where the black line connecting hand features across different time steps ignores the occluded hand feature; Fig. 6 (3) shows that the Interaction Graph module cannot fully utilize its potential, as the interaction cannot be established in frames where only one hand is detected; Fig. 6 (4) shows the connections after applying InterHand Temporal Fusion, where the connections extend from two hands within the same frame to two hands across all frames within the window size.

We assume that one blue circle in Fig. 4 (2) represents $f_{l,t}^{L}$, so the receptive field for $f_{l,t}^{L}$ includes $\left\{f_{l,t-1}^{L}, f_{l,t+1}^{L}\right\}$.[2] After applying Interaction Graph in Fig. 4 (3), the receptive field for $f_{l,t}^{L}$ further includes $\left\{f_{l,t-1}^{L}, f_{l,t+1}^{L}, f_{l,t}^{R}\right\}$, where the connection between $f_{l,t}^{R}$ and $f_{l,t}^{L}$ captures the interaction from the other hand. However, if $f_{l,t}^{L}$ cannot be detected, as shown by the blue dashed circle in Fig. 6, in such a case, the missing feature can only be inferred through its adjacent available features $\left\{f_{l,t-1}^{L}, f_{l,t+1}^{L}\right\}$ as shown in Fig. 6 (2) and (3), which degenerates to the scenario prior to the application of the Interaction Graph, as illustrated in Fig. 4 (2).

[2]The receptive field refers to the region of inputs to the convolution operation. In this case, $\left\{f_{l,t-1}^{L}, f_{l,t+1}^{L}\right\}$ as well as $f_{l,t}^{L}$ are used as the inputs to calculate node features for the left hand at time t.

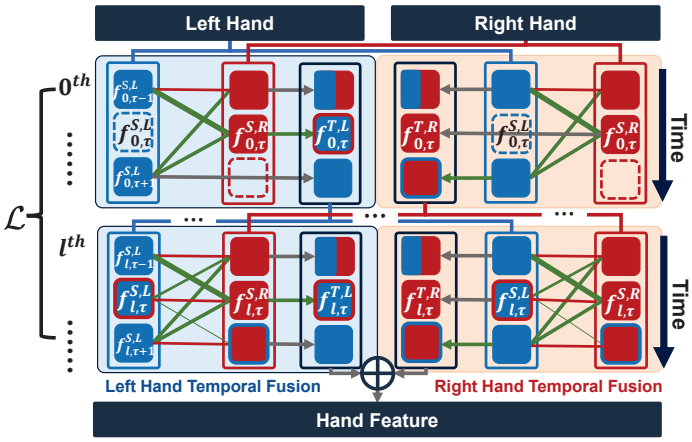

Fig. 7: Architecture of InterHand Temporal Fusion

To address the incompleteness of the data and take full advantage of Interaction Graph, we introduce InterHand Temporal Fusion. Because InterHand Temporal Fusion is integrated into temporal convolution in STGCN-based methods, we briefly introduce the temporal convolution before explaining the equation of InterHand Temporal Fusion, as shown in the following formula:

$$f_{l,t}^{T} = \sum_{k=0}^{K_t-1} F_k f_{l,t+k}^{S}, \tag{3}$$

where F_k is the convolutional filter at temporal position k, K_t is the size of the temporal kernel, determining how many adjacent time steps are considered in a single temporal convolution operation.

Here we explain how InterHand Temporal Fusion uses the attention mechanism to fill missing information. We use the left hand input $f_{l}^{S,L}$ as an example. In Fig. 7, we assume that $f_{0,\tau}^{S,L} = 0$, which means the left hand is occluded at time τ. By employing the attention mechanism, right hand feature $f_{0,\tau}^{S,R}$ automatically selects the most relevant features from left hand features $\left\{f_{0,\tau-1}^{S,L}, f_{0,\tau+1}^{S,L}\right\}$ to infer the occluded feature $f_{0,\tau}^{T,L}$. Next, we output $f_{0}^{T,L}$ within a time window for the first temporal convolution. After passing through l STGC blocks, the occluded feature $f_{l,\tau}^{T,L}$ is continuously refined, rebuilding the connection between the two hands. Meanwhile, the feature $f_{l}^{T,L}$ within the time window, which serves as the input for the $(l+1)$-th STGC block, is fused with the features of the other hand $f_{l}^{S,R}$ across the temporal dimension. The same operation is applied to produce the right hand output $f_{l}^{T,R}$.

Overall, in InterHand Temporal Fusion, node features $f_{l,t}^{T,L}$ are calculated based on node features of the other hand $f_{l,t}^{S,R}$ and node features of the same hand within the time window $f_{l}^{S,L}$ by using the attention mechanism, and vice versa for $f_{l,t}^{T,R}$. Finally, we output $\left\{f_{l}^{T,L}, f_{l}^{T,R}\right\}$ within a time window.

Therefore, the temporal convolution after implementing In-

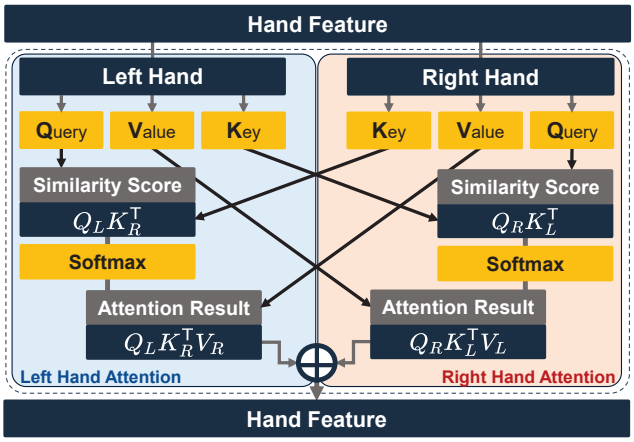

Fig. 8: Architecture of Interaction Attention

terHand Temporal Fusion is described by the equation below:

$$f_{l,t}^{T} = \sum_{k=0}^{K_t-1} F_k[(f_{l,t+k}^{S,R} \circledast f_l^{S,L}) \| (f_{l,t+k}^{S,L} \circledast f_l^{S,R})], \quad (4)$$

where $\|$ indicates the concatenation operation, $\circledast$ denotes the attention multiplication, which will be explained in detail in the Eq. (6) and (7) of the Interaction Attention section.

E. Interaction Attention

Similar to InterHand Temporal Fusion, the hand feature output by the InterHand Temporal Fusion module is passed into the InterHand Attention module and feature extractor as shown in Fig. 5. In this section, T denotes transpose operation.

Inspired by the SOTA attention mechanism, we designed Interaction Attention, which fuses the spatial-temporal features from both hands. In an attention mechanism, Query, Key, and Value are the three essential elements, and it can be represented by the following formula:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\mathsf{T}}{\sqrt{d_k}}\right)V, \quad (5)$$

where Q denotes the vector used to retrieve relevant information; K denotes the vector used to match against Q; QK^T denotes the relevance between Q and K; V denotes the actual information associated with each K; $QK^\mathsf{T}V$ represents the vector that aggregates the most relevant information from V, weighted by the relevance between Q and K; and d_k denotes the dimension of K. In Interaction Attention, we allocate the hand features to Q, K, and V:

$$f_R^A = f_R \circledast f_L = \text{Softmax}\left(\frac{f_R {f_L}^\mathsf{T}}{\sqrt{d_f}}\right) f_L, \quad (6)$$

$$f_L^A = f_L \circledast f_R = \text{Softmax}\left(\frac{f_L {f_R}^\mathsf{T}}{\sqrt{d_f}}\right) f_R. \quad (7)$$

We use Eq. (6) to explain the principle of Interaction Attention. For simplicity, f_R shows node features for the right hand output by temporal convolution and feature extractor as shown in Fig. 5, i.e., the input to the Interaction Attention. When we treat f_R and f_L as Q and K, the product of f_R and ${f_L}^\mathsf{T}$ represents the relevance between the left and right hand features. Then, we use $f_R {f_L}^\mathsf{T}$ to perform a weighted summation on f_L, which serves as V in the attention mechanism, to integrate the most relevant left hand features into the right hand features. After applying Interaction Attention, the features of one hand now include integrated information from the features of the other hand, leveraging the interaction between the two hands. Furthermore, with the help of the attention mechanism, the right hand features will automatically select the most important information from the left hand to predict the hand-washing steps. The above operation is concisely formulated as $f_R \circledast f_L$, and vice versa for Eq. (7).

Figure 8 shows the details inside the InterHand Attention module. The hand feature is separated into the right hand feature and left hand feature. Then, the hand feature (Query) undergoes matrix multiplication with the feature from the other hand (Key), resulting in the similarity score between the two hands, which corresponds to QK^T in the attention mechanism. This operation is related to the similarity score $Q_L {K_R}^\mathsf{T}$ for the left hand attention and $Q_R {K_L}^\mathsf{T}$ for the right hand attention. Next, the result is divided by the dimension of the hand feature and goes through a softmax layer. Subsequently, the result is multiplied by the other hand feature (Value) again to obtain the fused features of both hands, which relates to $QK^\mathsf{T}V$ operation. The operation above is related to the attention result $Q_L {K_R}^\mathsf{T} V_R$ for the left hand attention and $Q_R {K_L}^\mathsf{T} V_L$ for the right hand attention, Afterwards, the left hand feature fused with the right hand feature is concatenated with the right hand feature fused with the left hand feature. The hand feature is then processed by the feature extractor again as shown in Fig. 5 and its output f_l^A is fed into the next STGC Block as f_{l+1} for further processing.

F. Strong Compatibility

Another important advantage of our approach that we want to emphasize is its strong compatibility. With improvement of performance for HAR, the architecture has become increasingly complex and advanced. One advantage of InterHandNet is that the backbone network can be easily changed; it only requires the new backbone to be an STGCN-based method, which includes spatial graph convolution and temporal convolution, as shown in the yellow blocks of Fig. 2. In such a case, the performance of InterHandNet can continuously improve the accuracy of hand-washing recognition.

IV. EVALUATION

A. Dataset

This study focuses on the task of hand-washing recognition, following the hand-washing steps outlined by World Health Organization. Fig. 1 illustrates the six main steps of hand-washing. We chose the dataset captured by Lulla et al. [42] in Table I. They have created a dataset of 3,185 videos (hand-washing sessions), totaling over 23 hours in duration, making it the largest hand-washing dataset currently annotated according

to WHO guidelines. The dataset created from Xie et al. [41] is also used to evaluate InterHandNet because some RGB-based methods have been evaluated on their dataset. All datasets are resampled to 30 frames along the temporal dimension for each batch tensor, allowing the model to predict one hand-washing step per second in a 30 fps camera. The choice of one-second window size follows a common practice in hand-washing recognition, as seen in prior works [37], [38], [40]. Since our approach is based on skeletons, we extract skeletons from these two datasets by using Mediapipe Hands [6], which offers lightweight skeleton extraction, making it the best choice for edge devices. While Mediapipe Hands is used as the primary skeleton extractor, we also employed InterWild [29], a more advanced skeleton extraction method, to address scenarios where Mediapipe may be less applicable or accurate caused by occlusion or lighting conditions.

B. Evaluation Methodology

In the dataset [42], there are six different camera settings, from camera 100 to camera 105. We apply 5-fold cross-validation to each camera setting and obtain six groups of results. That is, for each camera setting, we use four of the five folds for training and the remaining one for validation. To avoid data leakage, we divide the data based on temporal order, ensuring that the validation set is not immediately after the training set in the temporal sequence. A weighted average is used to calculate the final result, balancing the different dataset sizes across camera settings.

The approaches used for comparison can be divided into two categories: i) SOTA STGCN-based methods and ii) RGB-based methods:
i) The selected STGCN-based methods include spatial graph convolution and temporal convolution modules, and demonstrate impressive HAR performance. These approaches, ordered by the publication year, include ST-GCN [18], 2s-AGCN [19], MS-AGCN [21], STA-GCN [22], CTR-GCN [24], MS-G3D [20], EfficientGCN-B4 [26], HD-GCN [27], and FR-Head [25]. ii) To better evaluate the performance of InterHandNet in the task of hand-washing recognition, we also introduced RGB-based methods for comparison. These methods include the efficient and lightweight image classifier models MobileNetV2 [32] and Xception [33], as well as approaches from Kim et al. [37], Zhong et al. [38], Xie et al. [41], and Cikel et al. [46], whose studies mainly focus on hand-washing recognition.

The evaluation metrics consist of accuracy, precision, recall, and F1 score, which are four widely used metrics for classification tasks.

To assess the feasibility of deploying InterHandNet on edge devices, we measured its computational latency. The start time was recorded after loading the model before making the prediction for the first sample, while the end time was captured upon generating the prediction for the final sample. We processed 1,800 samples, each with the shape of $(1, 3, 30, 42)$, representing one second of two-hand skeleton data (42 joints with 3D coordinates) captured at 30 frames per second by a camera. By dividing the total processing time by the number of samples, we obtained the average computation time for one-second data. The initial five samples were excluded to eliminate the effects of the warm-up phase. In comparison to existing RGB-based approaches [37], [38], [41], [46], we developed a compact 3D convolutional neural network architecture with two convolutional layers and a classifier, which is significantly simpler than the aforementioned methods.

TABLE II: Comparison with state-of-the-art STGCN-based methods. * indicates RGB-based method. **Bold** indicates the best result. Underline indicates the second best.

Method	Accuracy	Precision	Recall	F1 Score
ST-GCN [18]	0.5339	0.5051	0.4049	0.3849
MobileNetV2* [32]	0.6403	-	-	-
Xception* [33]	0.6683	-	-	-
2s-AGCN [19]	0.6711	0.6834	0.6212	0.6197
HD-GCN [27]	0.7192	0.7315	0.7251	0.7254
MS-AAGCN [21]	0.7393	0.7616	0.6933	0.7009
EfficientGCN-B4 [26]	0.7517	0.7298	0.7048	0.6967
CTR-GCN [24]	0.7605	0.7547	0.7140	0.7038
FR Head [25]	0.7660	0.7572	0.7084	0.7133
STA-GCN [22]	0.7803	0.7659	0.7334	0.7353
MS-G3D [20]	0.7922	0.7931	0.7604	0.7569
InterHandNet(FR Head)	0.8106	**0.8195**	0.7841	0.7824
InterHandNet(STA-GCN)	**0.8170**	0.8121	**0.7964**	**0.7951**

C. Implementation Details

The platform we use to run the entire task is the NVIDIA Jetson Orin Nano, with the Luxonis OAK-D Depth AI Camera connected via a USB cable. The Jetson is a small-sized edge device equipped with a custom chip designed to accelerate the execution of deep learning networks. For the skeleton extractor, we selected Mediapipe Hands, which is specifically designed for edge devices and installed on Jetson. To deploy InterHandNet on Jetson, we use NVIDIA TensorRT, an SDK for high-performance deep learning inference on NVIDIA GPUs, which is also optimized for Jetson's GPU architecture.

From a hardware perspective, we initially trained InterHandNet on an NVIDIA GeForce GTX TITAN X 12GB. The training process ran for 50 epochs with a learning rate of 0.01. The optimizer used was SGD, with momentum set to 0.9 and weight decay set to 0.0005. The loss function was Cross-Entropy, and the batch size ranged from 32 to 128, depending on the dataset's scale and the baseline's complexity. After the training phase, we saved InterHandNet's parameters from the epoch with the best evaluation metrics in ONNX (Open Neural Network Exchange) format, a universal model format compatible with different programming languages. Next, we wrote interface code to run inference on the Jetson using TensorRT, converting the model from ONNX format to a TensorRT engine file, which enables fast execution on edge devices like the Jetson. The entire workflow involves capturing RGB-D videos using a depth camera, extracting 3D hand skeletons with Mediapipe Hands, and then processing them through the TensorRT-optimized InterHandNet to predict the hand-washing step on Jetson in real time, every second.

TABLE III: Improvement after applying three modules with state-of-the-art STGCN-based methods. IG, IA, ITF indicate Interaction Graph, Interaction Attention and InterHand Temporal Fusion.

Method	Accuracy	Precision	Recall	F1 Score
ST-GCN [18]	53.3	50.5	40.4	38.4
+ IG/IA	59.7↑6.4	63.3↑12.8	**58.0↑17.6**	**53.9↑15.5(40%)**
+ IG/IA/ITF	**61.3↑7.8**	**69.1↑15.8**	54.6↑14.2	53.6↑15.2(40%)
2s-AGCN [19]	67.1	68.3	62.1	61.9
+ IG/IA	71.2↑4.1	73.0↑4.7	67.7↑5.6	67.0↑5.1(8%)
+ IG/IA/ITF	**74.4↑7.3**	**73.7↑5.4**	**71.9↑9.7**	**71.6↑9.7(16%)**
MS-AAGCN [21]	73.9	76.1	69.3	70.0
+ IG/IA	**76.8↑2.9**	**77.5↑1.4**	72.7↑3.4	72.6↑2.6(4%)
+ IG/IA/ITF	76.3↑2.4	76.2↑0.1	**74.1↑4.8**	**73.9↑3.9(6%)**
STA-GCN [22]	78.0	76.5	73.3	73.5
+ IG/IA	77.5↓0.5	79.0↑2.5	73.6↑0.3	73.7↑0.2(0.3%)
+ IG/IA/ITF	**81.7↑3.7**	**81.2↑4.7**	**79.6↑6.3**	**79.5↑6.0(8%)**
CTR-GCN [24]	76.0	75.4	71.4	70.3
+ IG/IA	74.4↓1.6	76.2↑0.8	71.4→	71.4↑1.1(2%)
+ IG/IA/ITF	**78.2↑2.2**	**77.5↑2.1**	**76.1↑4.7**	**76.2↑5.9(8%)**
FR Head [25]	76.6	75.7	70.8	71.3
+ IG/IA	77.4↑0.8	79.7↑4.0	75.2↑4.4	75.1↑3.8(5%)
+ IG/IA/ITF	**81.0↑4.4**	**81.9↑6.2**	**78.4↑7.6**	**78.2↑6.9(10%)**

D. Results & Discussion

1) Performance of InterHandNet: In Table II, we compare InterHandNet with SOTA STGCN-based methods and RGB-based methods. Specifically, we selected two backbones for InterHandNet: FR Head and STA-GCN. Table II shows that InterHandNet significantly outperforms other methods across all four evaluation metrics. InterHandNet (FR Head) achieved the best precision score, while InterHandNet (STA-GCN) recorded the highest scores in accuracy, recall, and F1 score. Surprisingly, InterHandNet outperformed MS-G3D by about 0.04 in F1 score. InterHandNet also surpassed the two RGB-based methods, MobileNetV2 and Xception, which were evaluated by Lulla et al. [42].

2) Effectiveness of InterHand Temporal Fusion and Backbone Compatibility: Table III indicates the improvements achieved by integrating the proposed modules into STGCN-based methods. The improvements are clearly demonstrated, with up to a 40% increase in performance for ST-GCN. The improvement by the Interaction Graph and Interaction Attention was limited in both STA-GCN and CTR-GCN. This suggests that the connection between two hands interfere with the core mechanisms of these two networks. However, the improved results obtained after applying InterHand Temporal Fusion reveal the limitations of these networks when dealing with occlusion problems and highlight the effectiveness of InterHand Temporal Fusion.

Our model, which heavily relies on temporal data to recover missing information caused by the occlusion, may fail in extreme scenarios where data from both hands is missing. However, Mediapipe Hands ensures that at least one hand can be detected in hand-washing activities in most cases. Instances where both hands are entirely occluded during hand-washing are negligible in real-world scenarios, minimizing the impact of such extreme cases on overall performance.

TABLE IV: Ablation study for ST-GCN by InterWild extractor

Method	Accuracy	Precision	Recall	F1 Score
ST-GCN [18]	75.0	76.9	72.1	71.7
+IG	78.7↑3.7	79.3↑2.4	77.1↑5.0	77.7↑6.0
+IA	80.5↑5.5	81.1↑4.2	78.7↑6.6	79.4↑7.7
+IG/IA	83.8↑8.8	85.0↑8.1	82.7↑10.6	83.5↑11.8

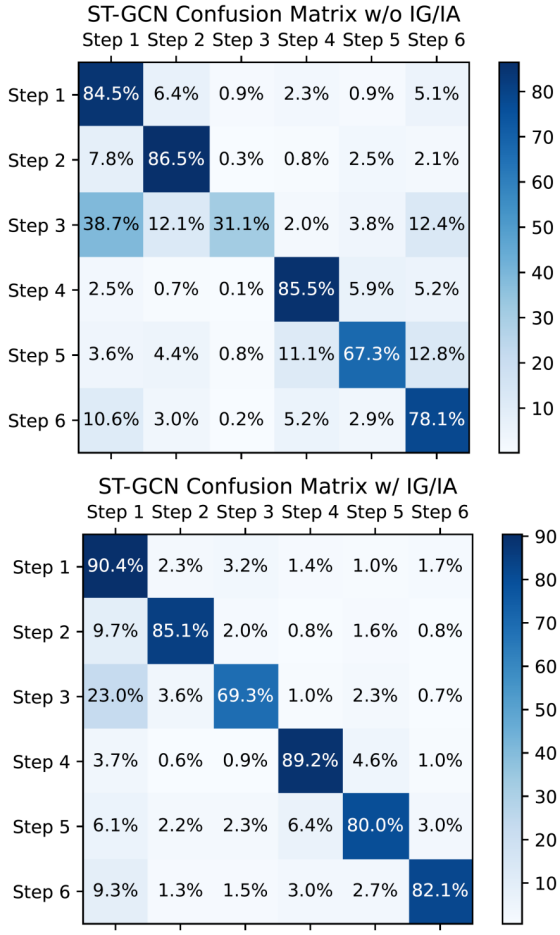

Fig. 9: Confusion Matrix w/ and w/o IG/IA for ST-GCN

3) Effectiveness of Interaction Graph and Interaction Attention: As shown in Table IV, we conducted an ablation study to verify the effectiveness of the Interaction Graph and Interaction Attention modules. We used InterWild [29] to extract the skeleton data instead of Mediapipe Hands in order to avoid the impact of incomplete data. InterWild [29], one of the most advanced skeleton extractors, captures more complete skeleton data even in challenging scenarios. This minimizes noise from missing or occluded keypoints, enabling more accurate evaluation of the Interaction Graph and Interaction Attention mechanisms. However, due to its complexity, the runtime efficiency is insufficient for deployment on edge devices for real-time recognition.

As shown in Table IV, by adding the Interaction Graph and Interaction Attention modules, the F1 score improves by about 0.1 when optimal skeleton extraction results are available.

We also present the confusion matrices in Fig. 9, with the improvement for step 3 reaching nearly 40%. Most steps

show a significant improvement in accuracy after applying Interaction Graph and Interaction Attention. Additionally, the model shows better performance in distinguishing steps 1, 2, 3, and 6. As observed in Fig. 1, there is a similarity between steps 1 and 6, as well as between steps 2 and 3. The two modules capture the subtle differences in interaction features of these steps, which might be easily overlooked when relying solely on skeleton features.

4) Comparison of Skeleton Extractors: Skeleton extractor performance plays a crucial role in hand-washing recognition accuracy. As shown in Table III and Table IV, results using InterWild outperform those based on Mediapipe Hands by over 20% under the same ST-GCN backbone.

The performance gap between InterWild and Mediapipe Hands underscores the importance of skeleton data quality in hand-washing recognition. While InterWild offers superior robustness and completeness, its computational complexity limits its suitability for real-time edge device deployment. In contrast, Mediapipe Hands provides a lightweight and efficient alternative but struggles with occlusions and challenging environments, affecting recognition accuracy.

5) Comparison with RGB-based Methods: Table V shows results of RGB-based methods [37], [38], [41], [46] evaluated by Xie et al. [41] on their own dataset. We extracted the skeleton data from this dataset by using Mediapipe Hands and applied InterHandNet with the STA-GCN and FR Head backbones and evaluated by the same methodology as described by Xie et al. [41]. The results show that InterHandNet (FR Head) excels in each evaluation metric, further enhancing the effectiveness of InterHand Temporal Fusion. While RGB-based methods have a natural advantage in handling occlusion because image features capture all details, skeleton extraction can be incomplete due to the limitations of Mediapipe Hands. Even under these conditions, our approach still outperforms RGB-based methods with significantly less data, further highlighting the performance of InterHandNet in hand-washing recognition. The evaluation metrics are approximately 10% higher than those presented in Table II because Lulla et al.'s dataset [42] contains more videos, covering a greater variety of hand-washing conditions, as shown in Table I.

6) Computational Efficiency: Table VI presents the average computational latency of InterHandNet running on the Jetson platform for processing one-second skeleton data by a 30fps camera, utilizing the ST-GCN, STA-GCN, and FR Head baselines. These latency measurements are compared with those of the original versions of each baseline. It can be observed that latency increases as model complexity rises. Since the entire system operates in two stages, Mediapipe Hands incurs a latency of approximately 600 ms for processing one second of video data on the Jetson platform captured at a resolution of 1280×720 pixels, i.e., 20 ms per frame. In the second stage, InterHandNet requires only $1/10$ of the computational resources compared to the skeleton extractor. These results ensure that InterHandNet-based hand-washing recognition operates in real time. Specifically, a basic 3D CNN requires six seconds to process a one-second RGB sequence.

TABLE V: The performance comparison of various handwashing recognition RGB-based approaches on the dataset from Xie et al. [41].

Method	Accuracy	Precision	Recall	F1 Score
I3D [37]	0.6865	0.7045	0.6888	0.6888
CNN+LSTM [46]	0.8590	0.8634	0.8592	0.8588
Two-stream CNN [38]	0.8665	0.8721	0.8698	0.8670
CNN+self-attention [41]	0.8910	0.8957	0.8913	0.8896
InterHandNet(STA-GCN)	0.8926	0.8942	0.8934	0.8927
InterHandNet(FR Head)	**0.9067**	**0.9108**	**0.9075**	**0.9075**

TABLE VI: One-second computational latency on Jetson

Method	Computational latency (ms)
Mediapipe Hands [6]	594.000
ST-GCN [18]	2.048
STA-GCN [22]	4.081
FR Head [25]	18.483
InterHandNet(ST-GCN)	7.433
InterHandNet(STA-GCN)	23.952
InterHandNet(FR Head)	56.906
3D CNN-based [37], [38], [41], [46]	>6498.790

Consequently, more complex 3D CNN-based methods [37], [38], [41], [46] are expected to incur longer processing times due to the much larger scale of input data compared to skeleton data, making them impractical for real-time recognition.

V. Limitations & Future Works

One limitation of our method is its reliance on edge devices specifically designed for deep learning, such as NVIDIA Jetson. While this ensures efficient deployment in real-time applications, it may restrict the applicability of our approach to cost-sensitive or power-constrained scenarios, such as wearable devices or low-power IoT systems. Addressing this limitation in future work could involve optimizing the model for lightweight deployment on less specialized hardware, thereby broadening its accessibility.

Another limitation lies in the lack of relevant datasets, which creates uncertainty regarding the model's performance in certain specific scenarios. For example, the performance of the model when processing inputs from only one hand, hands with missing fingers (e.g., in the case of individuals with disabilities), or hands interacting with objects remains unknown. These challenges highlight the need for more diverse datasets that represent a broader range of hand configurations and interactions.

Despite these limitations, InterHandNet's unique features suggest that its applicability can be extended to other activities involving frequent interactions and occlusions, similar to hand-washing. Examples include tasks such as washing clothes, playing the piano, or other hand-intensive activities. However, investigating this potential requires appropriate datasets for training and evaluation. Future work will focus on collecting such datasets and enhancing the model's adaptability to handle scenarios with greater variability, enabling its application to a wider range of dynamic and complex tasks.

VI. Conclusion

We proposed InterHandNet, a neural network specifically optimized for hand-washing recognition, incorporating three novel modules. The Interaction Graph and Interaction Attention modules effectively capture the interactions between the two hands, while the InterHand Temporal Fusion module reconstructs missing data caused by hand occlusions. InterHandNet outperforms existing skeleton-based and RGB-based methods in evaluations and exhibits sufficiently low computational latency to enable real-time operation on edge devices, surpassing RGB-based approaches. Additionally, the baseline architecture of InterHandNet can be continuously updated to enhance its performance due to its strong compatibility with existing STGCN-based models. We will release the code for InterHandNet, including the modules for InterHand Temporal Fusion and Interaction Attention, on our GitHub repository: https://github.com/yqzhang99/InterHandNet.

Despite these contributions, future work could explore the integration of additional data modalities, such as IMU or depth information, to further improve robustness under extreme occlusion scenarios. Moreover, the development of lightweight versions of InterHandNet could expand its usability in power-constrained environments, such as wearable devices or IoT systems. Lastly, we plan to collect more diverse datasets to address underrepresented scenarios, including tasks involving one-handed operations, missing fingers, hands interacting with objects or other hand-intensive activities. These directions aim to enhance the versatility and applicability of InterHandNet across a wider range of dynamic and complex tasks.

Acknowledgement

This study is partially supported by JSPS KAKENHI Grant Number JP21H05299.

References

[1] S. Nirjon, C. Greenwood, C. Torres, S. Zhou, J. A. Stankovic, H. J. Yoon, H.-K. Ra, C. Basaran, T. Park, and S. H. Son, "Kintense: A robust, accurate, real-time and evolving system for detecting aggressive actions from streaming 3d skeleton data," in *Proceedings of 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 2–10, 2014.

[2] N. Yoshimura, J. Morales, T. Maekawa, and T. Hara, "Openpack: A large-scale dataset for recognizing packaging works in iot-enabled logistic environments," in *Proceedings of 2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 90–97, IEEE, 2024.

[3] T. Kumrai, J. Korpela, T. Maekawa, Y. Yu, and R. Kanai, "Human activity recognition with deep reinforcement learning using the camera of a mobile robot," in *Proceedings of 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10, IEEE, 2020.

[4] V. Narayanan, B. M. Manoghar, V. S. Dorbala, D. Manocha, and A. Bera, "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation," in *Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8200–8207, IEEE, 2020.

[5] P. Kouris, M. Sionti, C. Korfitis, and S. Markantonatou, "Motion capture of modern greek verbs: Measuring aspects and relations among actions," in *Adjunct proceedings of 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–7, IEEE, 2020.

[6] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[7] P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang, "Rtmo: Towards high-performance one-stage real-time multi-person pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1491–1500, 2024.

[8] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019, 2016.

[9] T. Maekawa, D. Nakai, K. Ohara, and Y. Namioka, "Toward practical factory activity recognition: Unsupervised understanding of repetitive assembly work in a factory," in *Proceedings of UbiComp '16*, p. 1088–1099, Association for Computing Machinery, 2016.

[10] Q. Xia, A. Wada, J. Korpela, T. Maekawa, and Y. Namioka, "Unsupervised factory activity recognition with wearable sensors using process instruction information," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, p. 60, 2019.

[11] Q. Xia, J. Korpela, Y. Namioka, and T. Maekawa, "Robust unsupervised factory activity recognition with body-worn accelerometer using temporal structure of multiple sensor data motifs," vol. 4, Sep 2020.

[12] N. Yoshimura, T. Maekawa, T. Hara, A. Wada, and Y. Namioka, "Acceleration-based activity recognition of repetitive works with lightweight ordered-work segmentation network," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, Jul 2022.

[13] J. Morales, N. Yoshimura, Q. Xia, A. Wada, Y. Namioka, and T. Maekawa, "Acceleration-based human activity recognition of packaging tasks using motif-guided attention networks," in *Proceedings of the 2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–12, 2022.

[14] J. Morales, Q. Xia, N. Yoshimura, H. Oshima, M. Fukuda, Y. Namioka, and T. Maekawa, "Multilevel transfer learning for complex work activity recognition in logistic domain," in *Proceedings of the 2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2025.

[15] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[16] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1831–1840, 2017.

[17] W. Shi, D. Li, Y. Wen, and W. Yang, "Occlusion-aware graph neural networks for skeleton action recognition," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 10, pp. 10288–10298, 2023.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[19] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, 2019.

[20] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2969–2978, 2022.

[21] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[22] K. Shiraki, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Spatial temporal attention graph convolutional networks with mechanics-stream for skeleton-based action recognition," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[23] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 143–152, 2020.

[24] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13359–13368, 2021.

[25] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 10608–10617, 2023.

[26] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1474–1488, 2022.

[27] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10444–10453, 2023.

[28] G. Moon, J. Y. Chang, and K. M. Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5079–5088, 2018.

[29] G. Moon, "Bringing inputs to shared domains for 3d interacting hands recovery in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17028–17037, 2023.

[30] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "Handoccnet: Occlusion-robust 3d hand mesh estimation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1496–1505, 2022.

[31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[35] R. Burchard, P. M. Scholl, R. Lieb, K. Van Laerhoven, and K. Wahl, "Washspot: Real-time spotting and detection of enacted compulsive hand washing with wearable devices," in *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pp. 483–487, 2022.

[36] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[37] M. Kim, J. Choi, and N. Kim, "Fully automated hand hygiene monitoring\\in operating room using 3d convolutional neural network," *arXiv preprint arXiv:2003.09087*, 2020.

[38] C. Zhong, A. R. Reibman, H. M. Cordoba, and A. J. Deering, "Hand-hygiene activity recognition in egocentric video," in *Proceedings of 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, 2019.

[39] R. Bakshi, "Hand hygiene video classification based on deep learning," *arXiv preprint arXiv:2108.08127*, 2021.

[40] W. Huang, J. Huang, G. Wang, H. Lu, M. He, and W. Wang, "Exploiting cctv cameras for hand hygiene recognition in icu," in *Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[41] T. Xie, J. Tian, and L. Ma, "A vision-based hand hygiene monitoring approach using self-attention convolutional neural network," *Biomedical Signal Processing and Control*, vol. 76, p. 103651, 2022.

[42] M. Lulla, A. Rutkovskis, A. Slavinska, A. Vilde, A. Gromova, M. Ivanovs, A. Skadins, R. Kadikis, and A. Elsts, "Hand-washing video dataset annotated according to the world health organization's hand-washing guidelines," *Data*, vol. 6, no. 4, 2021.

[43] R. Burchard and K. Van Laerhoven, "Multi-modal atmospheric sensing to augment wearable imu-based hand washing detection," *arXiv preprint arXiv:2410.03549*, 2024.

[44] C. Wang, Z. Sarsenbayeva, X. Chen, T. Dingler, J. Goncalves, V. Kostakos, *et al.*, "Accurate measurement of handwash quality using sensor armbands: Instrument validation study," *JMIR mHealth and uHealth*, vol. 8, no. 3, p. e17001, 2020.

[45] H. Li, S. Chawla, R. Li, S. Jain, G. D. Abowd, T. Starner, C. Zhang, and T. Plötz, "Wristwash: towards automatic handwashing assessment using a wrist-worn device," in *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pp. 132–139, 2018.

[46] K. Cikel, M. Arzamendia Lopez, D. Gregor, D. Gutiérrez, and S. Toral, "Evaluation of a cnn+ lstm system for the classification of hand-washing steps," in *Proceedings of XIX Conference of the Spanish Association for Artificial Intelligence (CAEPIA)*, 2021.

[47] H. Q. Vo, T. Do, V. C. Pham, D. Nguyen, A. T. Duong, and Q. D. Tran, "Fine-grained hand gesture recognition in multi-viewpoint hand hygiene," in *Proceedings of 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1443–1448, IEEE, 2021.

[48] C. Wang, W. Jiang, K. Yang, Z. Sarsenbayeva, B. Tag, T. Dingler, J. Goncalves, and V. Kostakos, "Use of thermal imaging to measure the quality of hand hygiene," *Journal of Hospital Infection*, vol. 139, pp. 113–120, 2023.

[49] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu, "Interacting attention graph for single image two-hand reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2761–2770, 2022.

[50] L. Li, L. Tian, X. Zhang, Q. Wang, B. Zhang, L. Bo, M. Liu, and C. Chen, "Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20395–20405, 2023.

[51] Z. Yu, S. Huang, C. Fang, T. P. Breckon, and J. Wang, "Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12955–12964, 2023.