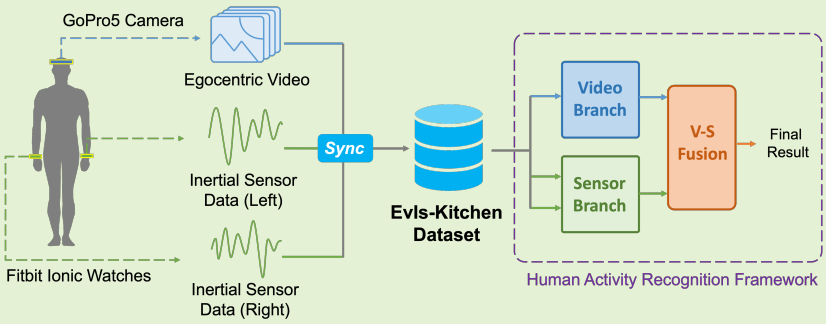
Egocentric Human Activities Recognition with Multi-modal Interaction Sensing

Yuzhe Hao, Asako Kanezaki, *Member, IEEE*, Ikuro Sato, *Member, IEEE*, Rei Kawakami, *Member, IEEE*, Koichi Shinoda, *Senior Member, IEEE*

***Abstract*— Egocentric Human Activity Recognition (ego-HAR) has received attention in fields where human intentions in a video must be estimated. However, the performance of existing methods is limited due to insufficient information about the subject's motion in egocentric videos. To overcome the problem, we proposed to use two hands' inertial sensor data as supplements for egocentric videos to do the ego-HAR task. For this purpose, We construct a publicly available dataset, Evls-Kitchen, which contains well-synchronized egocentric videos and two-hand inertial sensor data and includes interaction-focus actions as recognition targets. We also designed the optimal choices of input combination and component variants through experiments under two-branch late-fusion architecture. The results show our multi-modal set-up outperforms any other single-modal methods on Evls-Kitchen.**



***Index Terms*— Egocentric video, Inertial sensor, Multimodal dataset.**

I. Introduction

HUMAN Activity Recognition (HAR) is a task to infer human behavior from video [1]. It is important in the field of video understanding, as many video applications such as security analysis, health monitoring, and human-computer interaction rely on this technology. In HAR, the interaction between humans and objects is particularly important. It tells us which part of the environment the action subject is focused on, providing the key to estimating the intention of the subject. For example, in cooking instruction, by recognizing what tool the subject is holding and which materials the subject is looking for, the instruction system can provide appropriate suggestions.

Recently, egocentric videos collected from head-mounted cameras have been used in HAR. Unlike traditional third-person-view videos, which observe the action subject from a camera fixed in the environment, egocentric videos contain only the action subject's field of vision and change its scope as the action subject's focus changes. It can track the subject's attention, from which the intention of the subject can be easily estimated. Egocentric videos can provide information exclusive to the subject individual, making it possible to achieve personalized HAR applications, such as medical monitoring applications [2], [3].

However, egocentric video often fails to obtain the subject's movements in an interaction. This is because the subject's body or limbs, which are largely involved in most movements, may not be visible. In addition, the unstable self-centered viewpoint of a moving head camera introduces shake and blur into the RGB data stream. There are several studies [4]–[6] that tried to tackle these problems by introducing attention mechanisms towards the subject's hands. Nevertheless, these problems remain inherently as far as one relies only on the visual data stream.

Combining inertial sensor data with egocentric video is a way to overcome these problems. Below, we refer to the inertial sensor and video streams as *S-stream* and *V-stream*, respectively. Inertial sensors can track the subject throughout the interaction, providing data purely on movement. Several previous studies [7]–[9] combine these two modalities. However, they are limited as only one positional inertial sensor was used, which was placed at the subject's head. They focused on actions that only involve the subject and are not related to other objects. (*e.g.,* sitting, standing up, etc.) These limitations cause two problems. Since human action often involves the movement of multiple body parts, a single sensor may not represent human movement well. The head sensors may overlap with the head-mounted camera's motion and may not be sufficient to capture the subject-object interaction.

The inertial sensor data processing methods in the previous

An earlier version of this paper was presented at the International Conference of Multimodal Modeling 2023 and was published in its Proceedings. (https://doi.org/10.1007/978-3-031-27077-2_29)

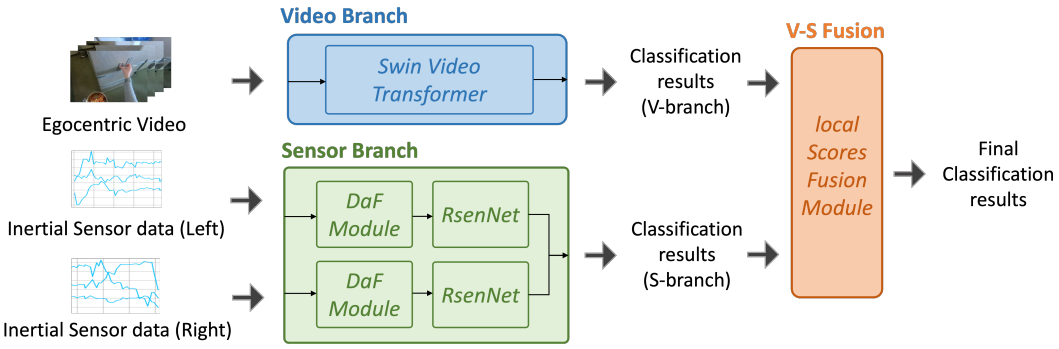


Fig. 1. Overview of our proposed Two-branch Late-fusion architecture.

works also have limitations. The independent streams architecture [8], [9] loses the correlation among different sensor modes (Accelerometer, Gyroscope, etc.), and directly concatenating these sensor modes [10] is not enough to reveal their complicated correlation since raw data is low-dimensional. Besides, a shallow network [9] may not handle fused high-dimensional sensor representation well, and its continuously connected layers structure causes the feature degradation problem when it comes to deeper networks [11].

In this paper, We propose to use inertial sensor data from both hands and egocentric video to solve ego-HAR. Namely, we have *V-S-S-streams* where *V* and *S-S* are for the video stream and the two sensor streams at two hands, respectively. Because the hands are the body parts most associated with human motion, inertial sensor data at both wrists better capture correlated movements and provide necessary supplementary information when hands are missing in the visual data stream. By combining them, we can obtain a precise representation of human activities.

Technically, we propose two innovations paying special attention to the processing of inertial sensor data. The first is the Decomposition-and-Fusion (DaF) module for fusing the data from sensors. Inspired by feature expansion operation in SVM [12], the DaF module incorporates a two-step fusion. It expands the information of each sensor mode to high-dimensional space first and fuses these high-dimensional representations, and applies non-linear activation after each layer to better model their correlation. The second is the Residual Sensor Network (RsenNet), the sensor feature extraction backbone that we designed. It contains the residual structure to prevent semantic information in the deep layers from degradation [13], [14] and processes comprehensive sensor features from the DaF module, which is more informative than raw data. Figure 1 shows the overview of our multi-modal ego-HAR architecture. We apply Swin Video Transformer [15] to process the *V* stream. Two sensor streams are fused by RsenNet, and *V-S* streams are fused by late score fusion.

To evaluate our method, we also build a new interaction-focused multi-modal ego-HAR dataset, "**E**gocentric **v**ideo and **I**nertial **s**ensor data Kitchen (EvIs-Kitchen)" dataset. It contains well-synchronized egocentric videos and inertial sensor data at the left and right wrists. The recognition targets are interaction-involving actions such as "cut carrot," or "wash spoon." On this dataset, we validate our choice of variants combination for the overall architecture and thoroughly examine the effectiveness of adding inertial sensors through the experiments.

The main contributions of this paper include the following:

- We extend the *V-S* ego-HAR task to the *V-S-S* interaction-focused ego-HAR task and propose a novel *DaF* + *RsenNet* method to handle and fuse the inertial sensor data well.
- We release a new "EvIs-Kitchen" dataset as a benchmark to evaluate the interaction modeling ability of *V-S-S* ego-HAR methods.
- We make a detailed ablation walk-through analysis of input data combinations and structure variants and provide an optimal set-up on our EvIs-Kitchen dataset for the *V-S-S* ego-HAR task.

II. Related Work

A. Egocentric Human Activities Recognition

Deep learning-based methods have been successfully applied in third-person-view HAR. Simonyan *et al.* [16] propose the two-stream network, where one branch extracts features from RGB frames and the other extracts those from optical flow independently, and those two features are concatenated and fed to a final fully-connected layer to obtain the prediction results. Tran *et al.* propose the 3D convolution network [17], which includes spatial and temporal feature extraction in the same convolutional kernel. Later they improve it to (2+1)-D convolution [18], [19], further boosting the performance by decomposing spatial and temporal feature extraction as two separate steps. The residual structure [11] they applied allows the important information of the input data to be retained, preventing degradation problems at deeper layers of the network. These benefits have been proven effective for other than image and video data, such as for speech [14] and for ECOG data [13]. Recently, many transformer-based methods have shown superior spatio-temporal modeling ability, such as ViViT [20], DeiT [21], SwinVT [15], which use the self-attention mechanism to encode spatio-temporal information in the video.

Deep learning also has been applied to ego-HAR. The subject's focus in the egocentric video is important to capture the subject's intention. Many studies use the attention mechanism towards the subject's hands, which play an important role in

Fig. 2. The devices used to collect the data. A GoPro 5 camera is mounted at the subject's head and two Fitbit Ionic watches are installed on two wrists.

Title	**#4** Fruits Salad
Introduction	This recipe is about making a most common fruits salad.
Purpose	In this recipe, there are some "cutting" and "dicing" actions, but unlike recipe#3, we make them into small chunks rather than slices. And we are using fruits as the material rather than vegetables in recipe#3, this could also be a factor used to distinguish them.
Materials	banana, apple, orange, salad sauce
Steps	**<Fruits Salad>:** 1. Peel a banana, cut the it into bite-sized chunks. 2. Wash an apple, then cut it into bite-sized chunks. (could also peel it if desired) 3. Peel an orange, then separate it into pieces. 4. Transfer all the cut materials into a bowl. 5. Put some salad sauce onto them, then slightly mix them together.

Fig. 3. Recipe #4 used in data collection experiments. Basic cooking instruction steps are included in this recipe.

many actions. Singh *et al.* [22] pre-annotated hands area, and includes their masks as inputs for training the network to learn interactions between hands, objects, and eyes. Ma *et al.* [23] adds spatial segmentation and temporal localization modules for hands to accurately recognize the object in interaction. Similarly, Kapidis *et al.* [24] uses a hand track module to obtain the subjects' interaction intention.

However, these methods assume that hands are visible most of the time in egocentric videos. To reduce the dependency on hands, some studies utilize other visual data streams to obtain more effective features than just using an RGB stream. Lu *et al.* [5] uses a Bi-LSTM layer to enhance the temporal representation in their extracted features with the subject's focusing region provided by gaze [25] data stream. Wang *et al.* [6] uses a Faster R-CNN as an object detection module [26] to provide attention from the object side during an interaction.

In addition, extracting features only from visual data streams causes two problems. First, the quality of the estimated optical flow is low because of the movement of the egocentric camera. Second, introducing other visual data streams, such as optical flow and object tracking, brings higher computational costs.

B. Video-sensor Multi-modal Methods

Multi-modal data could capture more information than just a single data stream. Many multi-modal methods, such as studies that combine video and audio [27], [28], RGB video and depth maps [29], or audio and text [30], [31], have effectively improved the performance of corresponding tasks. In order to better capture the interaction in the egocentric videos, use of the lightweight inertial sensors is reasonable since it can provide motion characteristics of the action subject. They are useful especially when hands are not visible in the video.

There are several studies that utilize inertial sensors along with visual data streams for action recognition. Chen *et al.* [32] utilizes inertial sensor data from either the right wrist or the right thigh of a subject along with a third-person-view depth map. However, they use conventional hand-crafted features for recognition. Song *et al.* [8] also utilizes videos and an inertial sensor but to ego-HAR. They collect the inertial sensor data from the head of a subject, and the classification is based on the classical Fisher linear discriminant analysis. They further improved the classification with deep learning, where an LSTM network and a two-stream CNN are applied to inertial sensor data and videos, respectively [7]. The method fuses features at the decision level. Imran *et al.* [9] follows this late fusion approach but uses multiple 1D-CNNs to extract features from the two modalities.

However, these studies are limited because they neglect the correlation among different sensor modes in the inertial sensor processing branch. The independent feature extraction for each sensor mode [7], [9] precludes the use of the correlation between these sensor modes, which makes the S-branch loses the ability to capture complicated sensor-mode-wise characteristics of actions. Although Shavit *et al.* [10] tries to directly fuse raw sensor data through concatenation, this may not reveal correlations among sensor modes because it is limited by the sparse amount of information in each frame. Besides, the continuous layers structure [9] causes feature degradation when the network becomes deeper, while the shallow network may not extract the semantic information in inertial sensor data.

High-dimensional expansion of features can unfold some data characteristics that are hidden at the low dimension. It has been a common method to get features with rich information since the support vector machine (SVM) [12], and recently this approach has been proven effective for GNN [33]. Non-linear layers are widely applied in deep learning models. By combining it, feature expansion can better model complex non-linear characteristics than just linear projections. Since more characteristics are revealed, using unfolded high-dimensional representations can capture the sensor modes correlation easier than using dense and unfolded low-dimensional raw data.

In this paper, two inertial sensors are used, one on the left hand and the other on the right hand, to model human-object interactions more accurately than just using egocentric videos. Besides, we adopted two novel designs the *DaF* module and the *RsenNet* to overcome the corresponding problems.

C. Ego-HAR Related Datasets

Most of the existing datasets for ego-HAR are only based on visual information. EPIC-Kitchen dataset [34], [35] is about kitchen activity and is the largest ego-HAR dataset. Because kitchen activities have high diversity and high relevance to the subject's field of vision, this set of activities is widely chosen for benchmarks for ego-HAR, such as Breakfast dataset

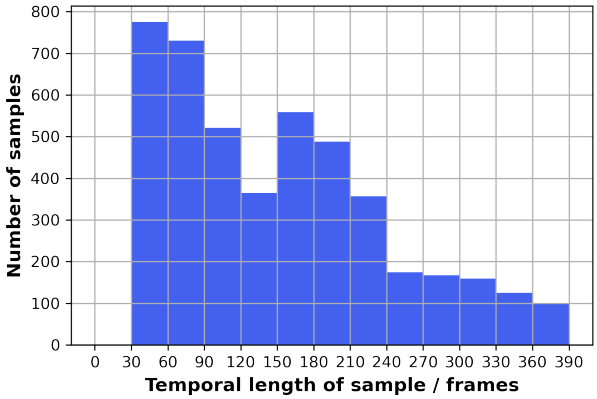


Fig. 4. Histogram of samples' temporal length (number of frames in each sample) in the EvIs-Kitchen dataset.

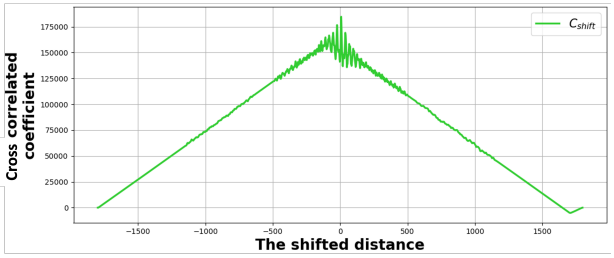


Fig. 5. The cross-correlation coefficient between the L and R jump-patterns of S-streams. The maximum of $C_{\mathbf{shift}}(z)$ represents that two signals reach the best-matched point at the corresponding shifted distance.

[36], [37], and 50 Salad dataset [38]. GTEA Gaze [25] and GTEA Gaze+ [39] datasets are also kitchen-activity datasets. They collect egocentric videos and gaze information with eye-tracking glasses. However, gaze information is still based on visual information; it cannot complement the visual data stream when inferring the subject's motions outside the scope.

There are also a few datasets that involve both video and inertial sensor data. UTD-MHAD dataset [32] provides third-person-view videos and inertial sensor data at the right wrist or right thigh. Egocentric Multi-modal Activity (EMA) dataset [7], [8] provides egocentric video and inertial sensor data collected from the subject's head as EPIC-Tent. Its action classes are all subject-highlighted actions, such as "sit down" and "stand up", which have evident action characteristics and hardly involve the interaction between subjects and objects in the scene. EPIC-Tent dataset [40] is an ego-HAR dataset that takes camping as the activity theme, which involves full-body actions with finer details compared to the EMA dataset.

Due to the device limitation, both EMA and EPIC-Tent datasets suffer from mainly two problems. First, they collect the inertial sensor data of the subject's head, which is hardly involved in most actions. Also, their features largely overlap with those obtained by the egocentric video camera at the same position. Second, they collect the sensor data only from one position of the subject. Since complex interactions often use both hands, data from one position may not be enough to represent the action well, which causes a bottleneck at the data level for the ego-HAR task.

Building a new dataset is necessary to examine the correlation and the complementarity between egocentric video and inertial sensor data. It should satisfy three requirements: (1) Inclusion of multiple inertial sensors installed at multiple positions of the subject; (2) High relation between the sensors' positions and human actions; (3) Good synchronization between egocentric videos and inertial sensor data.

III. EvIs-Kitchen Dataset

A. Overview

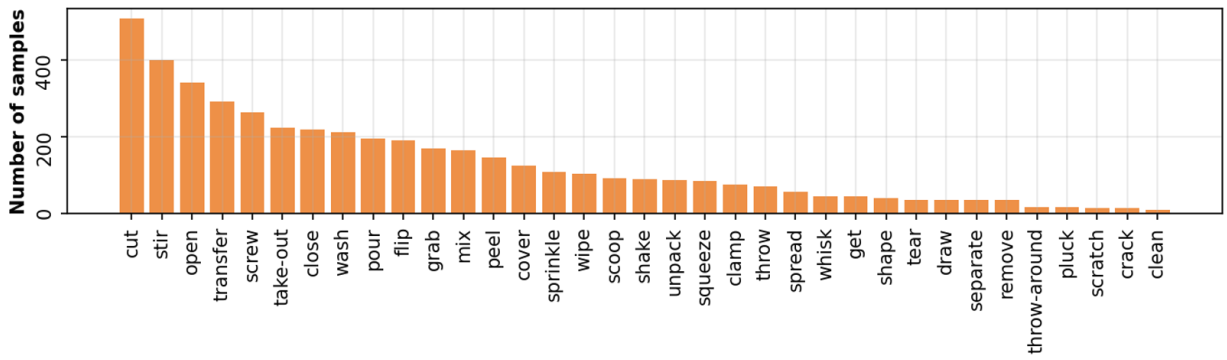
We built **E**gocentric **V**ideo and **I**nertial **S**ensor data Kitchen activity dataset (EvIs-Kitchen dataset), which is the first V-S-S interaction-focused dataset for the ego-HAR task. It consists of sequences of everyday kitchen activities as it involves rich interactions among the subject's body, object, and environment. Different from the EMA [7], [8] and EPIC-Tent [40], which merely use an inertial sensor attached to the subject's head, we use two inertial sensors attached on the left and right wrists. In contrast to the third-person-view HAR V-S multi-modal dataset [32], which includes inertial sensor data only from the right hand (or right thigh), our inertial sensor data are from both wrists. The ego-HAR datasets should include subject-object interaction to match their application scenarios. We choose kitchen activities as the action theme for our dataset because they are typical daily activities that involve rich human-object interactions.

B. Data Collection

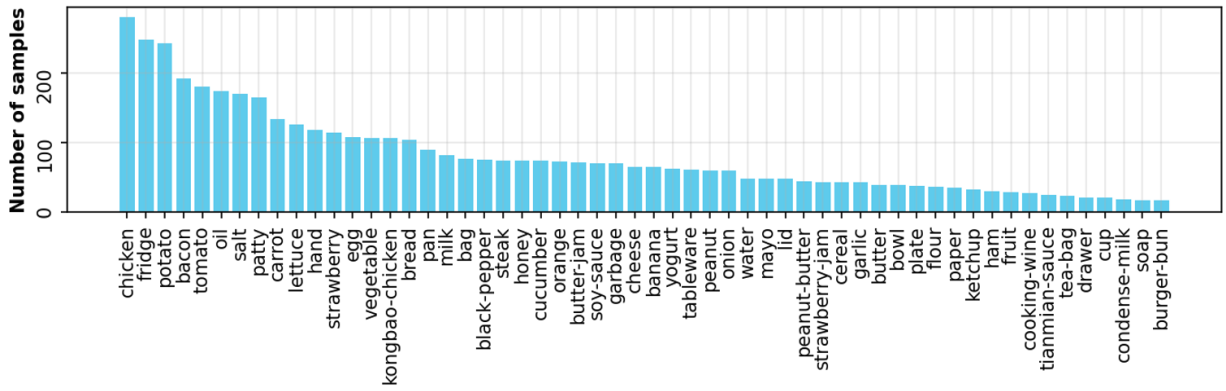
We used a GoPro 5 camera with a headband to collect egocentric video data and used two Fitbit Ionic watches placed on the left and right wrists to collect inertial sensor data, as Fig 2 shows. The watch contains three sensor modes: (1) An accelerometer provides 3-axis linear acceleration (a_x, a_y, a_z), (2) a gyroscope provides 3-axis angular velocity $(\omega_x, \omega_y, \omega_z)$, (3) an orientation sensor provides orientation vectors in 4-digit quaternion form (a, b, c, d), which overcomes the gimbal lock problem in the Euler angle form's orientation vector.

The data are collected from 12 subjects (eight males and four females), each cooking seven recipes. Among these subjects, three are left-handed, and the rest are right-handed. The details of actions vary from subject to subject since they have different cultural backgrounds and can cook in their ways, making our dataset's contents diverse and realistic. We show an example of recipes in Fig 3.

The whole data collection process includes three stages. First, at the beginning stage, the subjects are requested to do a fixed "stay still - repeat jumping vertically - stay still" process by the moderator's instruction as the data recording starts. This process creates shared signals among all data streams for further synchronization. Then, at the cooking stage, the subjects need to make a dish by following a given recipe, which provides the necessary basic cooking steps, as Fig 3 shows. It is not required to strictly follow the recipes. The subjects can cook in their own way during the data collection, making our dataset robust to real-world application scenarios. Finally, the subjects report the completion of cooking at the



(a) Number of samples of each *Verb* class



(b) Number of samples of each *Noun* class

Fig. 6. Visualization of all classes' number of samples in our EvIs-Kitchen dataset

end stage, the moderator will stop recording raw data.

This process collects three data streams, all of which are in 30 frames per second: (1) RGB data stream, each frame of which is a 3-channel RGB image, (2) left sensor data stream, each frame of which is a 10×1 vector $(a_x, a_y, a_z, \omega_x, \omega_y, \omega_z, a, b, c, d)$, and (3) right sensor data stream, the shape of each frame is the same as the left sensor data.

C. Synchronization

Synchronization is vital in multi-modal tasks, facilitating the alignment of different data streams representing the same action. By ensuring and reinforcing the shared characteristics among data streams, synchronization becomes crucial for comprehending interactions between these modes and a prerequisite for future advanced fusion methods.

However, it is challenging to apply conventional timestamp synchronization to our dataset. Because GoPro 5 cannot obtain accurate timestamps for each video frame. Moreover, the egocentric video and inertial sensor data both suffer from inaccuracies in their timestamps due to unavoidable transfer and hardware delays.

To overcome these problems, we propose a simple but effective synchronization method. We synchronize the different data streams by synchronizing their beginning series of frames, which contain a fixed-process action "stay still - repeat jumping vertically - stay still". This action causes a jump-pattern Ψ, which existed for a fixed period (contains T_j frames) and has evident action characteristics among all data streams. Then we synchronize different data streams by matching their Ψ.

1) Left-Right Sensor Synchronization: Our synchronization utilizes the oscillation waveforms created during the jump actions because they happen simultaneously and have similar representations in all data streams. The accelerometer data in the gravity direction of left and right wrists are utilized as their jump-patterns Ψ^{L} and Ψ^{R} to achieve the synchronization between left and right S-streams. This process is done by calculating the cross-correlation coefficient C_{shift} of them:

$$C_{\mathrm{shift}}(z) = \sum_{t=-T_j}^{T_j} \Psi^{\mathrm{L}}(t-z)\ \Psi^{\mathrm{R}}(t), \tag{1}$$

where z is the shifted distance. At $z \in [-T_j, T_j]$ with maximum $C_{\mathrm{shift}}(z)$ the two jump-patterns of the inertial sensor data of left and right wrists are synchronized. Fig 5 shows the cross-correlation coefficient between them.

Ψ^{S} is averaged from the synchronized Ψ^{L} and Ψ^{R}, which represents the synchronized jump pattern of S-stream and later will be used to synchronize with V-stream.

2) Video-Sensor Synchronization: Because jumping action causes vertical camera motion in V-stream, we utilize the Scale-Invariant Feature Transform (SIFT) features [41] to extract this kind of visual jump-pattern Ψ^{V} in V-stream and use it to synchronize with Ψ^{S}.

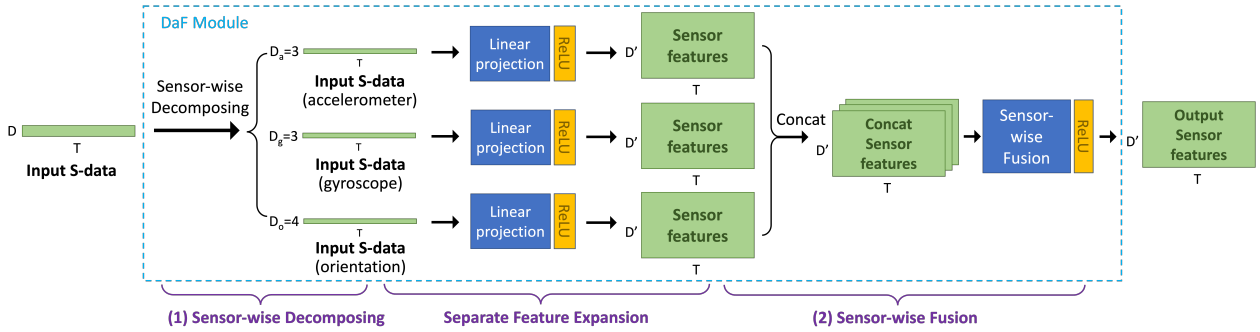


Fig. 7. The DaF module. $\boldsymbol{D = 10}$ means 10 channels in each frame's inertial sensor data; Among them, $\boldsymbol{D_a, D_g, D_o}$ represent accelerometer-related channels ($\boldsymbol{a_x, a_y, a_z}$), gyroscope-related channels ($\boldsymbol{g_x, g_y, g_z}$) and orientation-related channels ($\boldsymbol{o_a, o_b, o_c, o_d}$); $\boldsymbol{D'}$ represents the channel dimension of expanded features.

First, we find the set of all SIFT-matched point pairs for two adjacent frames t and $t+1$:

$$P_t = \{((x_t^{(i)}, y_t^{(i)}), (x_{t+1}^{(i)}, y_{t+1}^{(i)})) \mid i = 0,1,...,n_t\}, \quad (2)$$

where $(x_t^{(i)}, y_t^{(i)})$ is the coordinate of i-th matched pairs in frame t (x denotes the horizontal direction and y denotes the vertical direction). n_t is the number of SIFT-matched point pairs between t and $t+1$ frames.

Then we calculate the average SIFT vertical displacement d_t between frames t and $t+1$:

$$d_t = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{t+1}^{(i)} - y_t^{(i)}). \quad (3)$$

Finally, we obtain the entire jump-pattern of V-stream Ψ^{V} by calculating the d_t for every two adjacent frames in V-stream:

$$\Psi^{\mathrm{V}}(t) = d_t, \ t = 0,1,...,T_j. \quad (4)$$

With the extracted Ψ^{V}, we synchronize V-stream and S-stream by calculating the C_{shift} of Ψ^{V} and Ψ^{S}, which is the same way as synchronizing Ψ^{R} and Ψ^{L}.

In our EvIs-Kitchen dataset, all action samples are cut from this method's synchronized long data sequence.

D. Annotation

We annotate the action samples in our dataset with Verb-Noun labels. The action sample is the short segment that is manually cut from the synchronized long data sequence, which only includes one action in the entire cooking action sequence. Each action sample is annotated with a *Verb* class representing the subject's action and a *Noun* class representing the object the subject is manipulating. The *Verb* classes are mainly related to motion information, while the *Noun* classes are mainly related to visual information.

The sensor data are annotated with additional "Dominant/Non-Dominant" hand labels beside the "Left/Right" hand labels. For the right-handed subjects, the sensor data of the right hand is defined as the "Dominant" hand, while the left hand is defined as the "Dominant" hand for left-handed subjects.

There are 4,527 action samples with 56 *Noun* and 35 *Verb* classes in our EvIs-Kitchen dataset. The temporal length of the action sample varies from about 1 second (32 frames) to about 13 seconds (390 frames). Fig 4 shows the distribution of all samples' temporal lengths. The number of samples of each class also varies according to their frequency in the kitchen activities. Fig 6 shows the number of samples in each class.

IV. Method

A. Overview

We propose a two-branch late-fusion architecture for the multi-modal ego-HAR task. It consists of three main components: (1) S-branch, which extracts features from the input inertial sensor data; (2) V-branch, which extracts features from the input egocentric video data; and (3) VS-fusion component, which provides the final recognition results through the fusion of the results from the two branches. Fig 1 shows the overview of the proposed architecture.

The S-branch extracts motion information from inertial sensor data. It consists of an input-processing module and a feature extraction backbone. We propose the *Decomposition and Fusion (DaF)* module to obtain comprehensive inertial sensor representations by fusing the information provided by the different modes of inertial sensors, the accelerometer, the gyroscope, and the orientation sensor. As for the sensor feature extraction backbone, we designed the *ResNet-based Sensor Network (RsenNet)*. It processes the information from the two S-streams and outputs a score vector to estimate the actions.

The V-branch focuses on extracting visual information from egocentric videos. We selected Swin Video Transformer [15] in our proposed architecture. It expands the shift-window strategy from image transformers to the temporal dimension as video transformers and is state-of-the-art in the video processing field. Considering the efficiency and information redundancy in the egocentric video, we uniformly sample 32 frames from each sample as the input for the Video Branch.

The fusion component outputs the final result by fusing the classification vector of the V-branch and S-branch. We select local score fusion which can put more weight than global score fusion on informative channels. It is channel-adaptive and can model precise fusion between two branches' results.

B. Decomposition and Fusion (DaF) Module

The DaF module outputs comprehensive sensor representations for further feature extraction. As Fig 7 shows, it consists of a two-stage process.

TABLE I
COMPARISON AMONG SENSOR PROCESSING BACKBONES.

Sensor Backbone	Noun Acc@1(%)	Verb Acc@1(%)
RsenNet (ours)	**22.88**	**52.61**
SenLSTM [7]	21.50	49.95
T4sen [10]	20.85	40.87

TABLE II
S-BRANCH ABLATION EXPERIMENTS.

ID	DaF	Input	Norm	Jitter	Noun Acc@1(%)	Verb Acc@1(%)
1	-	AGO	-	-	21.22	43.37
2	✓	AGO	-	-	22.88	52.61
3	✓	AG	-	-	27.69	55.36
4	✓	AG	✓	-	**28.36**	**57.41**
5	✓	AG	✓	✓	19.71	46.44

1) Decomposition Stage: During this stage, we first split the raw inertial sensor data I^{S} into independent parts $I^{\mathrm{S}}_{\mathrm{acc}}, I^{\mathrm{S}}_{\mathrm{gyro}}, I^{\mathrm{S}}_{\mathrm{ori}}$, each containing data from a single sensor mode, where I^{S} is the input inertial sensor data, and its dimension is $\mathbb{R}^{D\times T\times 1}$ where D is the dimension of the sensor information, T is the number of frames, and 1 is for spatial dimension. $I^{\mathrm{S}}_{\mathrm{acc}}, I^{\mathrm{S}}_{\mathrm{gyro}}, I^{\mathrm{S}}_{\mathrm{ori}}$ are the decomposed sensor data, where their dimensions are $\mathbb{R}^{D_a\times T\times 1}$, $\mathbb{R}^{D_g\times T\times 1}$, and $\mathbb{R}^{D_o\times T\times 1}$, and $D_a + D_g + D_o = D$.

Then each part of inertial sensor data goes through an independent linear projection layer and gets activated by a ReLU layer to obtain the representation at the higher channel dimension:

$$\begin{aligned} R^{\mathrm{S}}_{\mathrm{acc}} &= \mathrm{ReLU}(\mathrm{Linear}_{\mathrm{acc}}(I^{\mathrm{S}}_{\mathrm{acc}})),\\ R^{\mathrm{S}}_{\mathrm{gyro}} &= \mathrm{ReLU}(\mathrm{Linear}_{\mathrm{gyro}}(I^{\mathrm{S}}_{\mathrm{gyro}})),\\ R^{\mathrm{S}}_{\mathrm{ori}} &= \mathrm{ReLU}(\mathrm{Linear}_{\mathrm{ori}}(I^{\mathrm{S}}_{\mathrm{ori}})), \end{aligned} \tag{5}$$

where $R^{\mathrm{S}}_{\mathrm{acc}}, R^{\mathrm{S}}_{\mathrm{gyro}}, R^{\mathrm{S}}_{\mathrm{ori}}$ are the representations for each sensor. Their dimensions are all $\mathbb{R}^{D'\times T\times 1}$, where D' means the higher channel dimension.

2) Fusion Stage: All the sensor representations from the previous stage are concatenated at a new dimension first, then go through a linear projection layer and a non-linear ReLU layer to fuse three channels into a single comprehensive channel, obtaining the sensor representations with correlation among different modes of sensors (Eq 6).

$$R^{\mathrm{S}}_{\mathrm{fused}} = \mathrm{ReLU}(\mathrm{Linear}(\mathrm{Concat}(R^{\mathrm{S}}_{\mathrm{acc}}, R^{\mathrm{S}}_{\mathrm{gyro}}, R^{\mathrm{S}}_{\mathrm{ori}}))), \tag{6}$$

where $R^{\mathrm{S}}_{\mathrm{fused}}$ is the fused sensor representation. Its dimension after the concatenation is $\mathbb{R}^{3\times D'\times T\times 1}$, then becomes $\mathbb{R}^{D'\times T\times 1}$ after the fusion.

C. Residual Sensor Network (RsenNet)

Residual Sensor Network (RsenNet) extracts features from sensor representations from the DaF module. The architecture of RsenNet is based on the ResNet-18. The spatial kernel size of all the convolutional blocks is set to 1×1 since sensor representations do not contain any spatial structure. It also has two sub-branches to process the inertial sensor data from two hands separately. An adaptive pooling layer is applied at the output layer to fuse the result of two sub-branches into one final output vector. The final output of the S-branch is a classification vector $\boldsymbol{f}^{\mathrm{S}}$:

$$\boldsymbol{f}^{\mathrm{S}} = [f^{\mathrm{S}}_1, f^{\mathrm{S}}_2, \cdots, f^{\mathrm{S}}_i, \cdots, f^{\mathrm{S}}_C]^{\top}, \tag{7}$$

where C is the number of classes, and element f^{S}_i represents the classification score of i-th class.

D. Local Score Fusion

The local score fusion fuses the classification vector of the V-branch and S-branch with a set of learnable weights. The score of each class in fused classification vector $\boldsymbol{f}^{\mathrm{F}} = [f^{\mathrm{F}}_1, f^{\mathrm{F}}_2, \cdots, f^{\mathrm{F}}_i, \cdots, f^{\mathrm{F}}_C]^{\top}$ is the weighted sum of corresponding elements in $\boldsymbol{f}^{\mathrm{V}}$ and $\boldsymbol{f}^{\mathrm{S}}$, and each class has its own specified pair of weights.

$$f^{\mathrm{F}}_i = w^{\mathrm{V}}_i f^{\mathrm{V}}_i + w^{\mathrm{S}}_i f^{\mathrm{S}}_i, \quad i = 1, 2, 3, ..., N, \tag{8}$$

where N is the number of classes in the classification vector.

V. EXPERIMENTS AND ANALYSIS

A. Settings

1) Task Definition: Our experiments divide the ego-HAR task into two sub-tasks: -*Verb* recognition and -*Noun* recognition. This sub-task separation allows us to evaluate the differences in the model's abilities to recognize interaction motions and interacted objects. The models used for these two sub-tasks are the same, except the number of classes in the final output layer is different. The overall top-1 accuracies on -*Noun* task and -*Verb* task are set as the evaluation metric.

2) Dataset Arrangement: Our proposed EvIs-Kitchen dataset is applied to all models evaluated. To keep the class frequency distributions, gender ratio (Male : Female), and dominant hand ratio (Left-dominated : Right-dominant) similar between the train set and test set, we select the samples of Subjects 4, 10, 11 (1,167 samples in total) as the test set, and keep the rest as the training set (3,360 samples in total). Both sets' gender ratios are 2 : 1, and their dominant hand ratios are also 2 : 1.

3) Sensor-only Experiments: All samples are zero-padded to the same temporal length (400 frames) for inertial sensor data to keep a unified input shape to S-branch. All of the Accelerometer, Gyroscope, and Orientation data are used if it is not mentioned. D in I^{S} is 10, and D_a, D_g, D_o in $I^{\mathrm{S}}_{\mathrm{acc}}, I^{\mathrm{S}}_{\mathrm{gyro}}, I^{\mathrm{S}}_{\mathrm{ori}}$ are $3, 3, 4$, respectively. The channel dimension of sensor representation (Eq 6) D' is set to 64 in all experiments.

4) Video-only Experiments: For egocentric video data, the frame shape is resized to 112×112, and 32 frames are uniformly sampled from the original data as the input for the V-Branch to reduce computational costs.

To train the V-branch backbone, we loaded the pre-trained weights provided by the authors. They are both pre-trained on the Kinetics-400 dataset (a massive third-person-view HAR dataset). Then we further fine-tuned it on our EvIs-Kitchen dataset.

TABLE III
COMPARISON BETWEEN SINGLE(S) AND MULTIPLE(S-S) SENSORS (APPLIED THE SETTING IN EXPERIMENT II-4)

ID	S-Data	Noun Acc@1(%)	Verb Acc@1(%)
1	Left	22.19	44.82
2	Right	26.73	49.79
3	Dominant	28.11	51.33
4	Non-Dominant	22.02	45.67
5	Both-hand	**28.36**	**57.41**

TABLE IV
COMPARISON BETWEEN DIFFERENT V-BACKBONE

V-Backbone	S-Data	Fusion	Noun Acc@1(%)	Verb Acc@1(%)
R(2+1)D-18 [18]	-	-	77.72	77.72
SwinVT-tiny [15]			86.72	85.09
R(2+1)D-18 [42]	Both-Hands	Local	80.38	80.29
SwinVT-tiny (ours)			**87.40**	**87.49**

5) Multi-modal Experiments: For the multi-modal method, we first loaded the trained weights from separate single-modal experiments for V- and S-branches in our architecture, then froze these two branches and just trained and fine-tuned the last fusion module.

B. Sensor-only Experiments

1) S-branch Backbone Comparison: We compare the performance among three different inertial sensor processing backbones: (1) Our proposed RsenNet, (2) Sensor LSTM network (SenLSTM) [7] with two layers and 128 feature channels, and (3) Transformer for sensor (T4sen) [10].

Table I shows the results of the comparison among S-branch backbones. Benefits from the residual structure of RsenNet, semantic information is accumulated through the well-inherited and enhanced representations from multiple levels. Thus it can acquire accurate predictions with meaningful features.

The performance of T4sen is inferior, and we believe insufficient training is the reason. Massive pre-training is usually required for a transformer structure, especially when modeling interaction in the egocentric video, which is more difficult than the subject-only action recognition task applied in [10].

2) S-branch Ablation: Based on the best-performed RsenNet model, we have three ablation experiments for exploring the best setting of the S-branch component:

Structure level: We compared the models with and without the DaF module to confirm the effectiveness of this module.

Input level: We compared the different input combinations (A-G-O and A-G, where A means Accelerometer, G means Gyroscope, and O means Orientation sensor) for the S-branch.

Data-Augmentation level: Our data augmentation includes two options: Max normalization and Gaussian jittering. For the max normalization, the max value of the accelerometer and gyroscope data of the train set is applied to normalize all sensor data into a range between 0 and 1. For the Gaussian jittering, we applied Gaussian noise with $\mu = 0$ and $\sigma = 0.1$.

Table II shows the results of S-branch components ablation experiments. In this table, each following experiment is based on the better setting of the previous experiment. We use Experiments ID to refer to the corresponding line in this table for convenience.

By comparing Experiments 1 and 2 in Table II, we can see the significant improvement the DaF module brought. This indicates that the DaF module successfully disentangled different modes of sensors and well modeled their correlation, providing informative motion representations for further feature extraction. However, their correlation is only about the motion, and they do not contain visual information, causing the improvement on *noun* task is not as significant as *verb* task.

By comparing Experiments 2 and 3 in Table II, it surprisingly improved the performance when the input removed the orientation sensor data. This is caused by the different mathematics meanings between Accelerometer+Gyroscope (A+G) and Orientation (O) data. A and G data are 3D orthogonal vector sequences representing spatial coordinates, while O data is a quaternion sequence representing a rotation in complex space. Fusing them may cause some contradiction and confusion. The orientation information could be characterized by the accumulation of G data (since it is angular velocity), so removing O data can also reduce information redundancy.

Experiment 4 in Table II added normalization to the input data. Scaling all samples into the range between 0 and 1 reduces the difficulty of convergence, which improves the performance on both *Verb* and *Noun* tasks. The similar level of improvement on both tasks indicates that normalization is working as a mathematics strategy. It does not change the semantic quality of representations. Experiment II-5 added Gaussian jittering to the input data, which damaged the performance badly, indicating that the semantic consistency of inertial sensor data is sensitive to noises.

3) Single(S)/Multiple(S-S) Sensor Experiments: We did experiments using single and multiple sensors to confirm whether multiple positions' sensors can better capture the interaction. For the single-hand experiments, besides the left/right-hand settings, we also did dominant/non-dominant-hand settings (This label is mentioned in Sec. III-D)

Table III shows the results of the comparison between S and S-S settings. Using two-hand inertial sensor data can improve performance, which is especially significant on the *Verb* task, indicating that two-hand sensor data successfully better modeled the interaction.

One thing that needs to be noted is Experiment 4 in Table II, where the single dominant-hand S-branch performs as well as Experiment 5 in Table III on the *Noun* task. This highlight on the *Noun* task is because the dominant hand's motions match particular objects. Nevertheless, this advantage may disappear if the dataset involves more classes and samples, where the "lucky matching" will be reduced, and more complicated motion patterns are required to achieve one-on-one matching between motion patterns and objects.

TABLE V
COMPARISON BETWEEN DIFFERENT SCORE FUSION MODULES

V-Backbone	S-Data	Fusion	Noun Acc@1(%)	Verb Acc@1(%)
SwinVT-tiny	Both-Hands	Global	86.98	86.63
		Local	**87.40**	**87.49**

TABLE VI
COMPARISON AMONG MULTI-MODAL SETTINGS

V-Backbone	S-Data	Fusion	Noun Acc@1(%)	Verb Acc@1(%)
SwinVT-tiny	Left	Local	87.92	87.66
	Right		87.75	87.40
	Dominant		87.75	**88.17**
	Non-Dominant		**87.92**	87.49
	Both-Hands		87.40	87.49
R(2+1)D-18	Left	Local	78.84	79.01
	Right		79.18	79.95
	Dominant		79.78	79.78
	Non-Dominant		78.92	79.78
	Both-Hands [42]		**80.38**	**80.29**

C. Video-only and Multi-modal Experiments

1) Comparison between Video-only and Multi-modal: We compared the performance of ego-HAR with the Video-only and Multi-modal methods. We expect to see the improvement will be mainly on the *Verb* task because inertial sensor data bring only the motion information. However, the accuracy of the *Noun* task also improved as Table IV shows. We believe the correlated interaction patterns between motion and object are the reason. In kitchen activities, some *Noun* classes always appear with certain *Verb* classes. For example, "fridge" always appears with "open", if the motion is detected as "cut", then "fridge" would not likely be the correction prediction. By these correlations, knowing the motion information also benefits the *Noun* recognition task.

2) Comparison between Different V-Branch Backbone: We trained and evaluated two V-branch backbone candidates, R(2+1)D-18 [18], [42] and SwinVT-tiny, on our EvIs-Kitchen dataset. Table IV also shows the results of comparisons among different V-branches candidates. The first two rows in Table IV are the results of video-only experiments. SwinVT-tiny outperforms a lot more than R(2+1)D-18, which V-branch used in the previous paper [42] The following two rows in Table IV further compare the Multi-modal fusion performance with these two different V-branch backbones. Similar to the Video-only experiments, the fusion models with SwinVT-tiny generally perform better than those using R(2+1)D-18. And the performance of multi-modal methods is better than video-only methods for both V-branch backbones.

These experiments results indicate two things: (1) SwinVT-tiny has a stronger visual feature extraction ability, and its advantages can be inherent to the multi-modal method. (2) Inertial sensor data do provide some complementary information with egocentric videos, leading to a performance improvement on multi-modal methods regardless of which V-branch backbone is applied.

3) Comparison between Local and Global score fusion: We compared the local and global score fusion modules with SwinVT-tiny as V-branch and RsenNet as S-branch.

Global Score Fusion The global score fusion directly applies a pair of manually-set weights for V-branch and S-branch to fuse their classification vectors:

$$\boldsymbol{f}^{\mathrm{F}} = w^{\mathrm{V}}\boldsymbol{f}^{\mathrm{V}} + w^{\mathrm{S}}\boldsymbol{f}^{\mathrm{S}}, \tag{9}$$

where $\boldsymbol{f}^{\mathrm{F}}$ is the final fused classification vector, $\boldsymbol{f}^{\mathrm{V}}, \boldsymbol{f}^{\mathrm{S}}$ are the outputs from two branches. It constrains the fusion on the overall level. Converging is easier since it just requires modeling one pair of weights.

As the experiment results shown in Table V, the local score fusion can achieve better performance than global score fusion, indicating that local score fusion could model a more complicated relationship between two score vectors.

4) Comparison between V-S and V-S-S: We also explore the fusion performance with different V-S combinations. Table VI shows the comparison among V-S-S and multiple V-S settings. Although the performance varies according to sensor settings, the differences are not as significant as the difference between video-only and multi-modal methods, and introducing inertial sensor data always improve the performance regardless of the sensor setting details.

We expect to see V-S-S can achieve better performance than V-S settings. To our surprise, the results perform differently with R(2+1)D-18 and SwinVT-tiny. When V-branch applies R(2+1)D-18, the advantage of better interaction modeling ability with two hands' inertial sensor data remains. V-S-S outperforms any V-S settings as expected. However, the V-S-S performs worse than V-S when V-branch applies SwinVT-tiny. The non-dominant hand V-S setting performs the best in *Noun* task, and the dominant hand V-S setting performs the best in *Verb* task. Furthermore, V-S performs better than V-S-S with SwinVT as V-branch in most cases. We believe there are mainly three reasons for this:

(1) The information provided by the S-branch overlaps with that from the pre-trained V-branch. S-S enhances the focus on the correlation between two hands, which does increase the number of correctly predicted samples, as Table III shows. However, these samples may overlap with those that SwinVT would have been able to predict correctly, so the advantage of S-S does not show up. In contrast, R(2+1)D-18 is less overlapping because it is a relatively weaker V-branch and can still reflect some advantages of S-S over S.

(2) Information bias in the data is enhanced by the single-hand S-branch. In most cases, people use the dominant hand to do the actions and the non-dominant hand to hold the objects, which caused information bias. When we just apply the single-hand sensor data, the information bias is enhanced. Action-related bias in the dominant hand's inertial sensor data benefits more on the *Verb* task, while the object-related bias in the non-dominant hand's data benefits the *Noun* task more.

(3) Biased single-hand inertial sensor data helps to solve some over-difficult samples. Single-hand S-branch emphasizes single-hand information, which is especially useful for some single-hand actions. Meanwhile, SwinVT cannot predict those

samples where hands do not appear in the video. When combining those two branches, the single-hand S-branch could provide the information on the emphasis of the missing hand in the video, which is precisely V-branch needs. So the single-hand S-branch makes up for the lack of the V-branch, and the shortcoming of Single hand S-branch can be filled by the strong prediction ability of the SwinVT V-branch, which causes a better performance eventually.

VI. Conclusion

In this paper, we provide a detailed introduction to our V-S-S-synchronized multi-modal dataset named EvIs-Kitchen, which focuses on describing interactions between humans and the environment with egocentric video and inertial sensor data. With inertial sensor data from two hands, it provides informative materials to represent the interaction than existing ego-HAR datasets, making space to explore how to deal with the V-S-S correlation for other future research.

We also designed a choice for combinations of components and inputs under the two-branch late-fusion architecture. Through experiments, we give solid comparison analysis and show "SwinVT-tiny + RsenNet(AG) + local-score-fusion" achieved the best performance on our EvIs-Kitchen dataset for the V-S-S ego-HAR task, which improves *Noun task* by 0.68 percent points and *Verb task* by 2.40 percent points compared to the video-only method.

Improving the performance balance and complementarity between V- and S-branch will be our future direction. Firstly, the relatively poor performance of the S-branch is the bottleneck in our multi-modal method. In the future, we will explore how to further improve its performance with the limited amount of inertial sensor data we have by transferring the knowledge from other datasets or applying self-supervised learning. Besides, according to our experiment results in Table VI, we sometimes need to emphasize one of the hand's inertial sensor data to reach better performance. We will explore how to apply dynamic weights between the left and right S-branch according to the content of the input egocentric video, making S-branch supply the information that V-branch exactly is lacking.

Acknowledgment

This work is an outcome of a research project, Development of Quality Foundation for Machine-Learning Applications, supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Tokyo Tech.). It was also supported by JST CREST JPMJCR1687 and JSPS KAKENHI JP23H00490.

References

[1] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41-42, pp. 30509–30555, 2020.

[2] D. Hossain, M. H. Imtiaz, T. Ghosh, V. Bhaskar, and E. Sazonov, "Real-time food intake monitoring using wearable egocnetric camera," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4191–4195, IEEE, 2020.

[3] I. González-Díaz, V. Buso, J. Benois-Pineau, G. Bourmaud, G. Usseglio, R. Mégret, Y. Gaestel, and J.-F. Dartigues, "Recognition of instrumental activities of daily living in egocentric video for activity monitoring of patients with dementia," in *Health Monitoring and Personalized Feedback using Multimedia Data*, pp. 161–178, Springer, 2015.

[4] S. Sudhakaran, S. Escalera, and O. Lanz, "Lsta: Long short-term attention for egocentric action recognition," in *CVPR*, pp. 9954–9963, 2019.

[5] M. Lu, Z.-N. Li, Y. Wang, and G. Pan, "Deep attention network for egocentric action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3703–3713, 2019.

[6] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Symbiotic attention with privileged information for egocentric action recognition," in *AAAI*, vol. 34, pp. 12249–12256, 2020.

[7] S. Song, V. Chandrasekhar, B. Mandal, L. Li, J.-H. Lim, G. Sateesh Babu, P. Phyo San, and N.-M. Cheung, "Multimodal multi-stream deep learning for egocentric activity recognition," in *CVPR Workshops*, pp. 24–31, 2016.

[8] S. Song, N.-M. Cheung, V. Chandrasekhar, B. Mandal, and J. Liri, "Egocentric activity recognition with multimodal fisher vector," in *ICASSP*, pp. 2717–2721, 2016.

[9] J. Imran and B. Raman, "Multimodal egocentric activity recognition using multi-stream cnn," in *ICVGIP*, pp. 1–8, 2018.

[10] Y. Shavit and I. Klein, "Boosting inertial-based human activity recognition with transformers," *IEEE Access*, vol. 9, pp. 53540–53547, 2021.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.

[12] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[13] H. Mirzabagherian, M. B. Menhaj, A. A. Suratgar, N. Talebi, M. R. A. Sardari, and A. Sajedin, "Temporal-spatial convolutional residual network for decoding attempted movement related eeg signals of subjects with spinal cord injury," *Computers in Biology and Medicine*, p. 107159, 2023.

[14] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," *Proc. Interspeech 2019*, pp. 1691–1695, 2019.

[15] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *arXiv preprint arXiv:1406.2199*, 2014.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, pp. 4489–4497, 2015.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, pp. 6450–6459, 2018.

[19] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *CVPR*, pp. 12046–12055, 2019.

[20] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, pp. 10347–10357, PMLR, 2021.

[22] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *CVPR*, pp. 2620–2628, 2016.

[23] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *CVPR*, pp. 1894–1903, 2016.

[24] G. Kapidis, R. Poppe, E. Van Dam, L. Noldus, and R. Veltkamp, "Egocentric hand track and object-based human action recognition," in *SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, pp. 922–929, 2019.

[25] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *ECCV*, pp. 314–327, 2012.

[26] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 284–293, 2019.

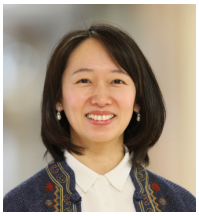
[27] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *CVPR*, pp. 5492–5501, 2019.

[28] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.
[29] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *IROS*, pp. 681–687, 2015.
[30] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, pp. 631–648, 2018.
[31] Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou, "Multi-stream deep neural networks for rgb-d egocentric action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3001–3015, 2018.
[32] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*, pp. 168–172, 2015.
[33] J. Sun, L. Zhang, G. Chen, P. Xu, K. Zhang, and Y. Yang, "Feature expansion for graph neural networks," in *International Conference on Machine Learning*, pp. 33156–33176, PMLR, 2023.
[34] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *ECCV*, pp. 720–736, 2018.
[35] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *IJCV*, vol. 130, no. 1, pp. 33–55, 2022.
[36] H. Kuehne, A. B. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, 2014.
[37] H. Kuehne, J. Gall, and T. Serre, "An end-to-end generative framework for video segmentation and recognition," in *Proc. IEEE Winter Applications of Computer Vision Conference (WACV 16)*, (Lake Placid), Mar 2016.
[38] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 729–738, 2013.
[39] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *CVPR*, pp. 287–295, 2015.
[40] Y. Jang, B. Sullivan, C. Ludwig, I. Gilchrist, D. Damen, and W. Mayol-Cuevas, "Epic-tent: An egocentric video dataset for camping tent assembly," in *ICCV Workshops*, pp. 0–0, 2019.
[41] P. C. Ng and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function," *Nucleic acids research*, vol. 31, no. 13, pp. 3812–3814, 2003.
[42] Y. Hao, K. Uto, A. Kanezaki, I. Sato, R. Kawakami, and K. Shinoda, "Evis-kitchen: Egocentric human activities recognition with video and inertial sensor data," in *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*, pp. 373–384, Springer, 2023.

Yuzhe Hao received his B.S. degree in Information and Communication Engineering from the University of Electronic Science and Technology of China in 2019, and the M.S. degree in Computer Science from the Tokyo Institute of Technology in 2021. He is currently a doctoral student in Artificial Intelligence at Tokyo Institute of Technology. His research interests include action recognition, multi-modal fusion, and computer vision.

Asako Kanezaki received her B.S., M.S. and Ph.D. degrees in Information Science and Technology from the University of Tokyo in 2008, 2010, and 2013, respectively. In 2010, she was a visiting researcher at Intelligent Autonomous Systems Group, Technische Universität München. From 2013 to 2016, she was an assistant professor at the University of Tokyo. She was working at the National Institute of Advanced Industrial Science and Technology (AIST) from 2016 to 2020. Since 2020, she is currently an associate professor at Tokyo Institute of Technology. Her research interests include object detection, 3D shape recognition, and robot applications such as semantic mapping and visual navigation. She is a member of IEEE.

Ikuro Sato Ikuro Sato received his Ph.D. in physics from the University of Maryland in 2005. He was a postdoctoral fellow at Lawrence Berkeley National Laboratory until 2007. He works for Denso IT Laboratory since 2008 until now. He has been concurrently appointed as a specially appointed associate professor at Tokyo Institute of Technology since April, 2020. He works on research problems of computer vision and machine learning especially for autonomous driving applications. He is a member of IEEE.



Rei Kawakami Rei Kawakami received her B.S., M.S., and Ph.D. degrees in information science and technology from the University of Tokyo in 2003, 2005, and 2008, respectively. She is currently an Associate Professor at Tokyo Institute of Technology, Tokyo, Japan. Before that, she worked as a Senior Researcher at Denso IT Laboratory. Her research interests are in computer vision and image processing. She is a member of IEEE, ACM, IEICE, and IPSJ.

Koichi Shinoda Koichi Shinoda received the BS and MS degrees from the University of Tokyo, Tokyo, Japan in 1987 and 1989, respectively, both in physics, and the DEng degree in computer science from the Tokyo Institute of Technology, Japan, in 2001. In 1989, he joined NEC Corporation, Japan, where he was involved in research on automatic speech recognition. From 1997 to 1998, he was a visiting scholar with Bell Labs, Lucent Technologies, Murray Hill, NJ. From October 2001 to March 2003, he was an associate professor at the University of Tokyo, Japan. He is currently a professor at the Tokyo Institute of Technology. His research interests include speech recognition, video information retrieval, and machine learning. He was the chair of SIG-SLP (Spoken Language Processing) in Information Processing Society of Japan from 2014 to 2016 and one of the general co-chairs of APSIPA 2021. He is a fellow of IEICE, and a senior member of IEEE and IPSJ.