Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Engagement Estimation of the Elderly from Wild Multiparty Human-Robot Interaction

Zhijie Zhang[1] | Jianmin Zheng[1] | Nadia Magnenat Thalmann[2]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore

[2]MIRALab – CUI – University of Geneva, Geneva, Switzerland

Correspondence
*Corresponding author name. Email: authorone@gmail.com

Present Address
This is sample for present address text

Abstract

The use of social robots in healthcare systems or nursing homes to assist the elderly and their caregivers will be becoming common, where robots' understanding of engagement of the elderly is important. Traditional engagement estimation often requires expert involvement in a controlled dyadic interaction environment. In this paper, we propose a supervised machine learning method to estimate the engagement state of the elderly in a multiparty human-robot interaction (HRI) scenario from the real-world video recording as input. The method is built upon the basic concept of engagement in geriatric psychiatry and HRI video representations. It adapts pre-trained models to extract behavior, affective and visual signals to form the multi-modal features. These features are then fed into a neural network made of a self-attention mechanism and average pooling for individual learning, a graph attention network for group learning and a fully connected layer to estimate the engagement. We tested the proposed method using 43 wild multiparty elderly-robot interaction videos. The experimental results show that our method is capable of detecting the key participants and estimating the engagement state of the elderly effectively. Also our study demonstrates that the signals from side-participants in the main interaction group considerably contribute to the engagement estimation of the elderly in the multiparty elderly-robot interaction.

KEYWORDS:
Human-robot interaction, Engagement estimation, Affective computing, Multiparty, Machine learning

1 | INTRODUCTION

This paper considers the problem of estimating the engagement of the elderly in wild multiparty human-robot interaction (HRI). With the advance of social robots, deploying robots in the healthcare becomes a possible solution to providing round-the-clock medical and psychological care to the elderly, especially the people with dementia (PwD), and supporting their caregivers as well[1, 2]. Natural elderly-robot interaction helps make the robot a good companion for the elderly who usually experience declines in physical and cognitive capacities. This has great impact since the proportion of people aged 60 years and older in the world will nearly double from 12% in 2015 to 22% in 2050 according to the World Health Organization[3].

During the elderly-robot interaction, if the robot can recognize the engagement state of the elderly, it helps the robot to respond to the elderly properly to maintain long-term interaction or to produce appropriate social behavior for the elderly to feel a sense of belonging. Here engagement refers to the inner state of a participant attributing to being together with the other participants

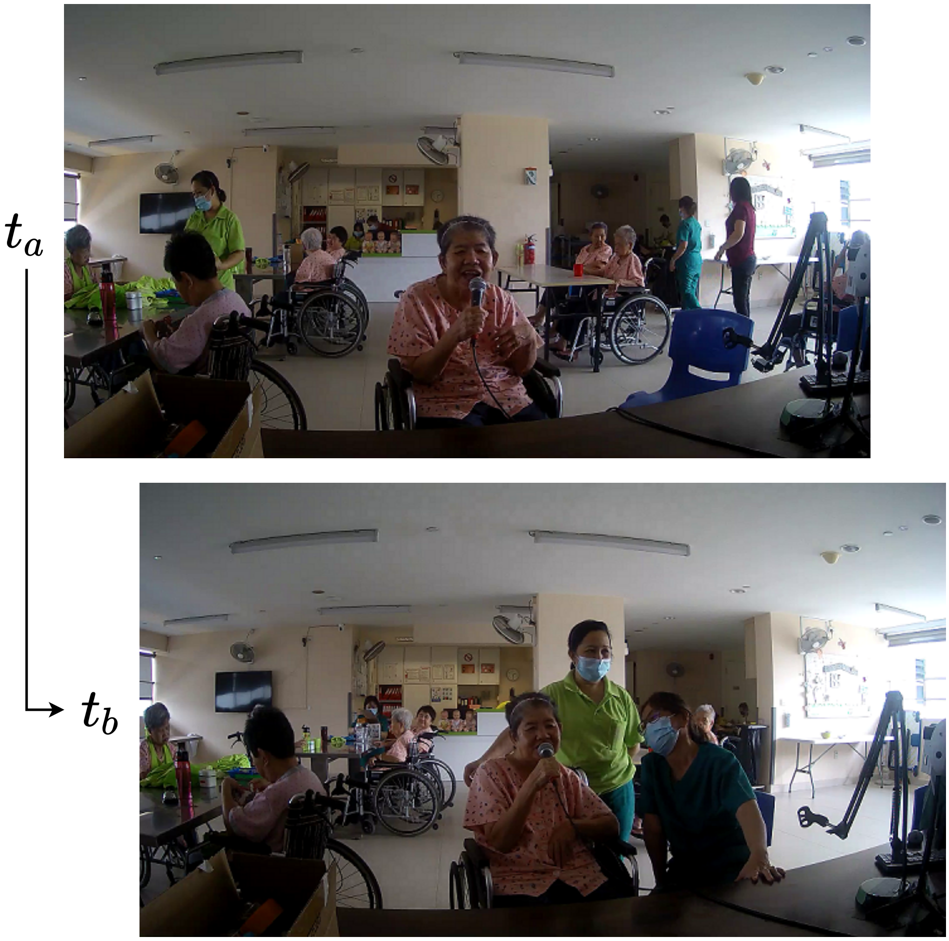


FIGURE 1 Two sample frames from a video recording of real-world multiparty HRI demonstrate conversation dynamics (from one to three participants) and unconstrained environment (open space and free-moving background people). The video is recorded from robot ego-view, and t_a, t_b denote two time stamps.

and continuing the interaction[4]. Many studies have shown that the engagement plays an important role in both human-human interaction (HHI) and human-robot interaction[5].

Engagement estimation is kind of affective computing and behavior recognition, and it goes further to probe the inner intention behind the apparent behavior and emotion. Many methods have been developed to estimate engagement in various scenarios such as general HRI[6–9], museum tour guide[10], classroom or distance learning[11–17], and healthcare[5, 18–20]. Conventional approaches use nonverbal cues such as proxemics, body pose, gaze patterns, facial expressions, and context information to build engagement estimation classifiers. Deep learning approaches have also been developed for engagement estimation[9, 10, 12, 14, 18, 21]. However, most previous work assumes that the interaction is in a laboratory environment or a dyadic situation. When the research is expanded to special populations and more complex circumstances as this paper is (see Fig. 1 for example), not much work has been done before. This may be in part due to the following challenges:

C1 The non-verbal signals from **the elderly** alter in facial shape and patterns of body behaviors along with aging[22, 23]. This challenges the conventional computer vision approaches in accurately estimating engagement state.

C2 From dyadic to **multiparty HRI**, understanding the dynamics and stability of the interaction becomes more complicated.

C3 In unconstrained **wild** space, moving people, bad lighting, confusing objects, *etc*. make it difficult to interpret the complex environment.

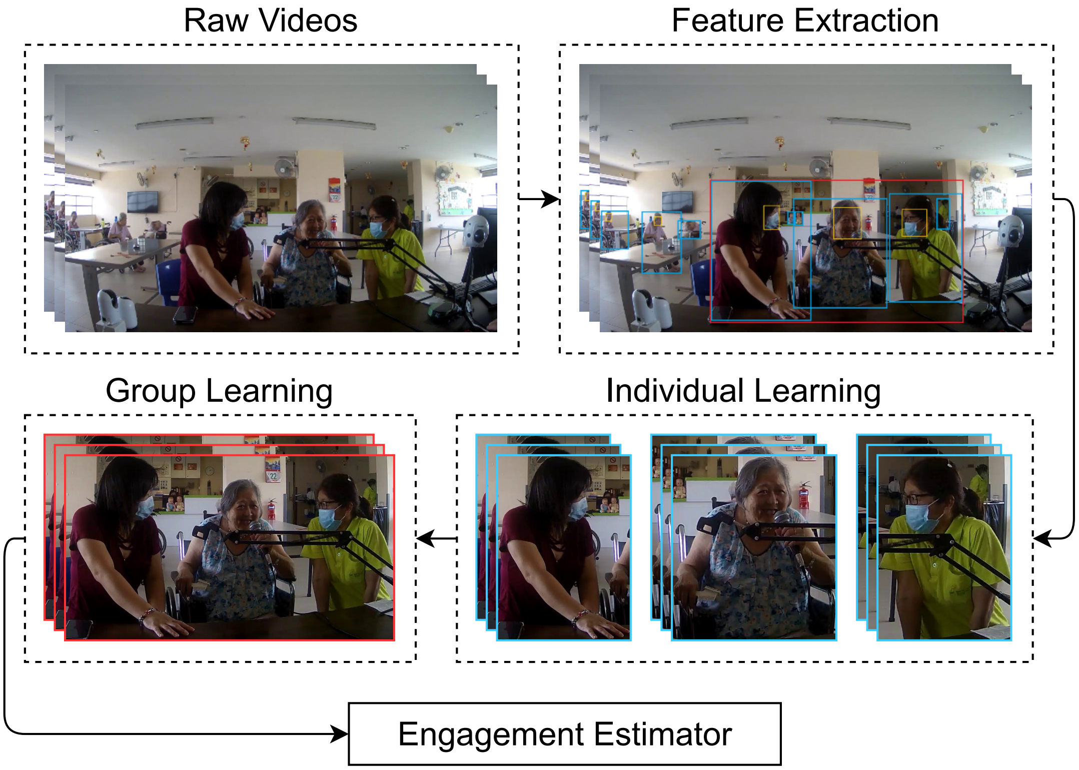


FIGURE 2 Overview of the proposed engagement estimation. The method is composed of four modules: (i) Feature Extraction, (ii) Individual Learning, (iii) Group Learning, and (iv) Engagement Estimation.

To tackle these challenges, in this paper we propose a supervised learning method for estimating engagement from real-world multiparty elderly-robot interaction. It takes video sequences as input and outputs the estimated engagement state. Fig. 2 shows the whole process. For each video clip, we first extract behavioral, affective, and visual feature maps. We adapt ResNet-3D[24] as the backbone to generate behavioral features from spacetime region. Affective and visual representations of participants are extracted using emotion recognition and face analysis tools. Then, we feed these features into an individual learning module, where features are refined by a self-attention mechanism. After that, the refined features are fed to the group learning module, which is a graph attention network learning the relationships among participants. The relationship conveys side participants' information, which helps the engagement estimation of the key elderly. Furthermore, to support the supervision, we label a real-world dataset, which is the video recording of the interaction between the elderly and an intelligent social robot, using the conventional psychological approach. The main contributions of the paper are

- We propose an automated approach to analyze wild multiparty HRI videos and estimate the engagement of the elderly.
- We borrow the concept of engagement components from psychiatry to design our network investigating behavioral, affective, and visual features.
- We build a machine learning model consisting of the self-attention network and the graph attention neural network for individual and group learning, which efficiently uses both individual and group information to improve engagement estimation.
- We create a labeled engagement dataset from a video recording of multiparty elderly-robot interaction in a wild environment.

TABLE 1 Comparison of Engagement Estimation Methods

Paper	Scenario	Participants[1]	Modality(s)[2]	Approach [3]	Output[4]
[8]	HRI	I/G	vis, aud	LR	$\hat{y} \in \{\text{NBrk, Brk}\}$
[17]	HHI	G (age 15-17)	phy, env	LightGBM	$\hat{y} \in [1, 5]$
[15]	HCI	I (age 20-60)	vis	NB	$\hat{y} \in \{\text{Eng, NEng}\}$
[6]	HRI	M	vis, dpt, per	SVM & RF	$\hat{y} \in \{\text{Eng, NEng}\}$
[18]	HRI	I/G (children)	vis, dpt	AlexNet & 2D CNNs	$\hat{y} \in \{\text{Eng, MEng, NEng}\}$
[10]	HRI	I/G	vis	CNNs+LSTM	$\hat{y} \in [0, 1]$
[21]	HHI/HCI	I	vis, aud, txt	GANs	$\hat{y} \in \{\text{Eng, NEng}\}$
[13]	HRI	I (age 4-6)	vis	RL	$\hat{y} \in \{\text{HEng, MEng, LEng}\}$
[9]	HRI	I/G	vis	I3D	$\hat{y} \in \{\text{Eng, NEng}\}$
[20]	HCI	I (PwD)	vis	LSTM	$\hat{y} \in \{\text{Eng, MEng, NEng}\}$
[14]	HHI	G (students)	vis	MLP & LSTM	$\hat{y} \in \{\text{HEng, MEng, LEng}\}$
[12]	HCI	I (age 19-27)	vis	GRU	$\hat{y} \in \{\text{HEng, Eng, BEng, NEng}\}$
Ours	HRI	M (PwD)	vis	ResNet3D+Attention+GAT	$\hat{y} \in [0, 1]$

[1] I, G, and M denote individual, group, and multiparty. The difference between multiparty and group is that multiparty treats participants separately but group treats them as a whole.

[2] Modalities: vis = visual, dpt = depth, per = personality, aud = audio, phy = physiological, env = environmental, and txt = text.

[3] Symbol '&' indicates using both and comparing with each other, and symbol '+' means combining to form a framework.

[4] $\hat{y}$ represents the inferred engagement label or value. For classification, Eng = Engage, Brk = Breakdown. The letters before Eng and Brk are N = Not, H = Highly, B = Barely, and M = Medium.

2 | RELATED WORK

This section briefly reviews some relevant work, especially in engagement estimation and engagement studies on geriatric psychiatry.

2.1 | Engagement Estimation

Traditional engagement estimation[6–8, 15, 17] extracts high-level social features, for example, body pose, facial expressions, gaze and task-related information, followed by a machine learning classifier. These features are intuitive and can be used with unimodal and multimodal combinations. Recently, with great progress of machine learning in computer vision, more and more deep learning methods have been developed for engagement estimation[9, 10, 12–14, 18, 21]. A summary of the estimation methods is given in Table 1, which also includes our proposed method for comparison.

Machine Learning Classifiers. In general HRI, Salam *et al.*[6] classified engagement using support vector machine (SVM) and random forest (RF), depending on predicted personality in a triadic interaction. They advanced the concept of engagement to the group level and claimed that categorization of engagement based on individual and interpersonal features without personality is insufficient. A similar work was also proposed by Celiktutan *et al.*[7]. Ben-Youssef *et al.*[8] studied engagement in HRI from the breakdown perspective, *i.e.*, users leave before the expected end. They extracted nonverbal multimodal data such as the distance to the robot, gaze and head motion, facial expressions, and audio. A logistic regression (LR) classifier was used.

Another widely investigated situation is online or in-class learning. Monkaresi *et al.*[15] explored engagement in the situation where students completed an online writing activity. Heart rate, action units (AUs) and local binary patterns were extracted and fed to a set of classifiers like Naive Bayes (NB). Gao *et al.*[17] predicted high school students' learning engagement including emotional, behavioral, and cognitive engagement in real-world classes. They used a set of features from wearable and indoor weather sensors to infer students' engagement status.

Deep Neural Networks. The aforementioned approaches require expert design of input features and cannot efficiently deal with large feature dimensions. Del Duchetto *et al.*[10] proposed a regression model based on CNNs and Long Short-Term Memory (LSTM) networks, which allows robots to compute the engagement from ego-view HRI videos. The model was built on a long-term dataset from an autonomous tour guide robot in a museum. Zhu *et al.*[12] presented an attention-based Gated Recurrent Unit network to predict engagement of students learning online. Taking the advantage of the published dataset[8], Saleh *et al.*[9] applied Inflated 3D ConvNets architecture to predict engagement state in an end-to-end way.

To estimate the engagement of children with autism spectrum disorder interacting with robots, Anagnostopoulou *et al.*[18] compared AlexNet[25] and 2D CNNs using 2D or 3D poses. Rudovic *et al.*[13] proposed personalized reinforcement learning (RL) to estimate engagement level (low, medium, high) from videos of child-robot interactions. The videos were labeled offline by experts, and used to personalize the policy and engagement classifier to a target child over time.

For HHI, Sumer *et al.*[14] utilized video recordings of classes to get attentional and emotional engagement features, and then applied SVM, RF, multilayer perceptron (MLP), and LSTM to predict students' engagement levels. Guhan *et al.*[21] described a multimodal GAN-based approach, called ABC-Net, to identify engagement from online dyadic HHI recordings. They utilized three-branch networks to gain valence and arousal, from which they generated engagement labels.

2.2 | Engagement in Geriatric Psychiatry

In geriatric psychiatry, the concept and measurement of engagement is well established. For example, Cohen-Mansfield *et al.* proposed an Observational Method of Engagement for PwD[26], which was one of the most well-known tools that many studies have used to measure engagement[27]. Following this concept, Jones *et al.*[5] developed the Engagement of a Person with Dementia Scale (EPWDS) towards psychosocial activities by assessing the behavioral and emotional expressions and responses. Perugia *et al.*[2] presented an affective computing framework which specifies the components of engagement in HRI.

Moreover, robotic and computer assistance has been shown to be an effective intervention. Moyle *et al.*[28] designed a robot seal for PwD. They found that the robot seal was more effective than usual care in improving mood states and agitation and participants were more engaged with it than with a toy. Similarly, Feng *et al.*[29] introduced an interactive system involving a display and a robotic sheep to engage PwD. They claimed that multimodal stimuli played a significant role in promoting engagement.

However, all the previously mentioned methods require expert involvement for engagement estimation. To achieve automated estimation, Steinert *et al.*[20] proposed a vanilla LSTM model to predict emotional engagement based on visual facial features (extracted by OpenFace[30] and VGGFace[31]) and contextual information (daytime, wellbeing, *etc.*). Similar to Steinert *et al.*'s work[20], our work is also an automated method. We use not only visual facial features, but also affective features, behavior features and the relation among all participants in the main interaction group.

3 | CONCEPT OF ENGAGEMENT

Engagement is generally regarded as a state or a process. According to Oertel *et al.*[32], this notion is ambiguous across different domains. While in terms of the state participants are either engaged or not engaged, by process the concept emphasizes how interactors establish, maintain, and complete their perceived connection to each other during an interaction[33]. Note that the term state represents objectively observed facts in HHI or HRI, which is whether the participants are within interaction or not. It is used to distinguish itself from *process*. In this paper, we adopt Poggi's definition of engagement[4], which refers to the participant's inner state of being together with other participants and continuing the interaction.

3.1 | Elements of Engagement

In general, engagement contains several elements[2, 11, 14, 21, 34–39]. They usually include behavioral, affective, visual, verbal, social, and cognitive signals.

- **Behavioral** involves observable behaviors such as approaching, touching, avoiding, and hitting.
- **Affective** is defined as the reactions that are usually represented by the valence and arousal.

- **Visual** encompasses actions involving the eyes and head such as maintaining contact or appearing inattention to others or materials.
- **Verbal** reflects the sounds and semantic information towards other participants.
- **Cognitive** refers to psychological investment and effort allocation of the person in order to fully comprehend the situation.
- **Social** includes the activities of encouraging or disrupting others.

Moreover, these elements are not mutually exclusive but often overlap with each other.

In this work, our goal is to estimate the engagement of the elderly interacting with a humanoid robot in casual conversation via a computer vision approach, so we pay attention to behavioral, affective, and visual engagement. The verbal element is eliminated due to the input modality, and the cognitive and social elements are overlooked due to the participant's physical and mental conditions.

3.2 | Engagement in Different Scenarios

Engagement estimation is studied in many disciplines and interaction scenarios. A simple taxonomy is based on the type of interactors: engagement in HHI or HRI. Although participants may behave differently in HHI and HRI, the estimation of engagement in these two disciplines is similar in terms of methodology. Thus the knowledge from the HHI could be applied to HRI[40].

In addition, the application scenarios of engagement estimation could be different. Examples are everyday conversations, healthcare and learning situations. In different scenarios, engagement may have different dominance of its elements. As a result, the corresponding estimation methods are different, and it is difficult to make fair comparison of different methods. There is no universal approach.

4 | PROPOSED METHOD

Our problem can be described as follows. The input is a raw video clip of the elderly-robot interaction in a wild multiparty environment. This video clip has a duration of about 10 seconds and contains only one elderly as the main participant and possibly a few other participants. It is assumed that within this duration the elderly's engagement state is fixed and corresponds to a value in $[0, 1]$. The value 0 represents the lowest level of engagement and the value 1 represents the highest level of engagement. Our goal is to estimate the value of engagement. In real applications, a human-robot interaction usually contains several interaction sessions and each session can be further divided into a sequence of clips. If we can estimate the value of engagement for each clip, we can obtain the engagement state over the entire interaction session.

In this section, we present a supervised learning method for estimating the value of engagement of the elderly given a video clip. The method consists of four modules. The first module is feature extraction. We use some pre-trained network models to obtain spatio-temporal representations of the input videos, from which behavior, affective and visual face features are extracted. The second module is individual learning, which refines individual features by adding a self-attention mechanism. Because the facial features and body movements of older adults are difficult to recognize, we introduce the attention mechanism to enrich individual features. The third module is group learning. We construct a graph network to learn the response from nurses as well as the relationships of participants within the group, which further helps to understand the engagement of the elderly. The last module is engagement estimation, which generates a value representing the elderly's engagement state. This is done just simply by a fully connected (FC) estimator. That is, we treat our task as a regression problem, and let the final layer of the network be a fully connected layer. The loss function for training is based on the mean squared error (MSE):

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2, \tag{1}$$

where y_i is the predicted engagement value, $\hat{y}_i$ is the ground truth, and M is the number of video clips. Fig. 3 gives the overall architecture of the proposed method. We next elaborate on the first three modules in detail.

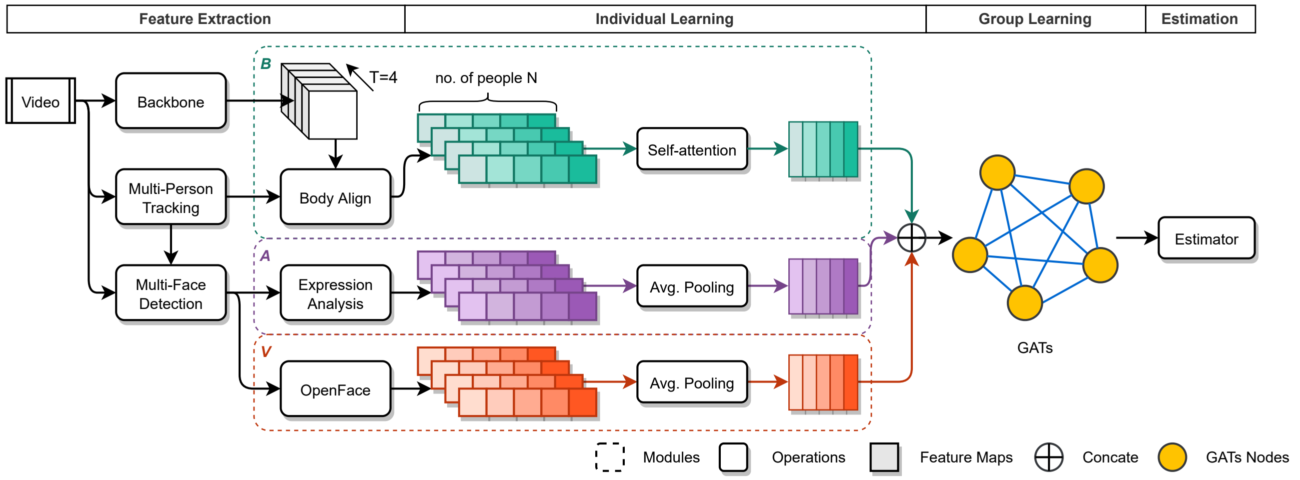


FIGURE 3 Architecture of the proposed network. A ResNet-3D model is adapted for extracting spatio-temporal features. Multi-person tracking and multi-face detection are used to get the bounding boxes in order to align and slice out corresponding body and face feature maps, which are then pooled for individual learning. The self-attention mechanism or average pooling is applied to refine the behavioral ($\mathcal{B}$), affective ($\mathcal{A}$), and visual ($\mathcal{V}$) features. The concatenation of these learned three components gives the representation of an individual, which is then further improved via group learning. Finally a fully connected layer estimates the elderly's engagement state.

4.1 | Feature Extraction

We use ResNet-3D[24] as the backbone to capture spatio-temporal context of an input video clip. This is motivated by the promising performance of ResNet-3D models in a wide range of video-related benchmarks[41]. We utilize the feature representations extracted by this backbone pre-trained on Kinetics 400[42]. The model details are listed in Table 2. The output, $X_{r4}^{B} \in \mathbb{R}^{1024 \times 4 \times 14 \times 14}$ from res_4 layer as the spatio-temporal feature maps, represents the behavioral engagement. Here 1024 is the channel depth, 4 is the temporal dimension and 14×14 is the map size. As depicted in Fig. 3, we extract four feature maps in time positions from this representation.

TABLE 2 The structure of our backbone model.

Layer name	Architecture	Output size
conv1	$5 \times 7 \times 7$, stride 2, 2, 2	$16 \times 112 \times 112$
maxpool1	$2 \times 3 \times 3$, stride 2, 2, 2	$8 \times 56 \times 56$
res2	$\begin{bmatrix} 3 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	$8 \times 56 \times 56$
maxpool2	$2 \times 1 \times 1$, stride 2, 1, 1	$4 \times 56 \times 56$
res3	$\begin{bmatrix} 3 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	$4 \times 28 \times 28$
res4	$\begin{bmatrix} 3 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 6$	$4 \times 14 \times 14$

Meanwhile, we conduct multi-person tracking on each input video clip. The purpose of this step is threefold: (i) the main interaction participants are identified based on the bounding boxes; (ii) we use these bounding boxes to eliminate the interference of redundant background information; and (iii) the tracked bounding boxes are used as constraints for face tracking as a way to ensure the consistency of face and body information.

To do this, we first perform multi-person tracking (MPT) using ByteTrack[43] to gain the bounding boxes (BBX) of the detected bodies. We identify the main group members intuitively based on the frequency of a person's appearance in the temporal axis and the distance from the camera in the spatial axis. We evaluate the distance by the size of the BBX. We set the frequency threshold to be 20% and the minimum BBX size to be 5000.

Given these, we use RoI Align[44] to project the coordinates on the frames' feature maps and slice out the corresponding features for each individual. After that the behavioral feature maps are refined to $X^B \in \mathbb{R}^{N\times1024\times4\times7\times7}$, where N is the number of detected participants. Formally,

$$X^B = RoI\left(E\left(v\right), \mathrm{BBX}\right) \tag{2}$$

where v represents the video clip, and RoI and E are the RoI align and feature extraction operations.

To extract affective and visual engagement features, we first perform multi-face detection by RetinaFace[45]. In order to ensure the consistency of the face and body in the temporal dimension, we keep $T = 4$ in the temporal direction. The detected $N \times 4$ faces are used for affective and visual feature extraction.

Specifically, we apply a pre-trained emotion recognition model, DMUE[46], on the cropped and aligned faces. To keep more facial information, we do not use the original 8 emotion classification results of DMUE, but instead, we remove the final linear projection layer and use the mid-output to represent affective engagement, which gives the affective features $X^A \in \mathbb{R}^{N\times4\times512}$.

Also, we use OpenFace[30] to extract the visual features. Since visual engagement is highly related to head and eyes behaviors, we select head pose and gaze features. Particularly, 6 head pose, 6 gaze directions and 2 gaze angles, 112 two-dimensional eye region landmarks, and 168 three-dimensional landmarks are extracted, which together form the visual engagement features $X^V \in \mathbb{R}^{N\times4\times294}$.

4.2 | Individual Learning

For the behavior features extracted in the previous module, though they are localized to the bounding boxes, they lack detailed body posture and action information, which actually plays an important role in understanding behavioral engagement. To overcome this issue, we introduce a self-attention mechanism[47] to refine the behavioral features. We hope that the attention mechanism can learn the interaction between any two feature positions in spatio-temporal dimensions and accordingly leverage this information to improve the feature representations, *i.e.*, focusing more on the important body regions in the spatial domain and the critical frames in the temporal domain. As demonstrated by the ablation study in Sec. 5, capturing such fine details contributes to the improvement of estimation performance.

In our implementation, the self-attention mechanism is a non-local operation, which calculates the response at a given position as a weighted sum of the features at all positions. That is, the self-attention block receives behavioral feature maps X^B extracted from the previous module as input and outputs the updated representations highlighting the most informative features. The non-local block is shown in Fig. 4. X^B is fed into three separate convolutions to embed the feature map. The non-local operation f, together with g, a simple linear embedding, computes the relationship between different locations. Then a residual connection is applied, followed by an average pooling to down sample feature maps to the size of $N \times 1024$.

For affective and visual features, because of the relative small feature dimension, we do not apply self-attention mechanism on them. Instead, we simply apply an average pooling (AP), which works on the temporal dimension, to make affective and visual features have appropriate size with the behavioral features.

Finally, we concatenate the three individual's features to derive the refined individual feature map:

$$\mathbf{H} = \left[\alpha\left(X^B\right), \mathrm{AP}\left(X^A\right), \mathrm{AP}\left(X^V\right)\right]. \tag{3}$$

where α is the attention operation.

Note that in Eq. 3, $\mathbf{H}$ is represented in terms of feature types. We reorganize it according to the detected people. Without ambiguity, we still use the notation $\mathbf{H}$, *i.e.*, $\mathbf{H} = \left[h_1, ..., h_N\right]$, where each h_i contains behavioral, affective, and visual features obtained from individual learning, particularly h_1 is for the elderly, and $[\cdot, \cdot]$ denotes concatenation. This $\mathbf{H}$ is then used as the input to the next module: group learning.

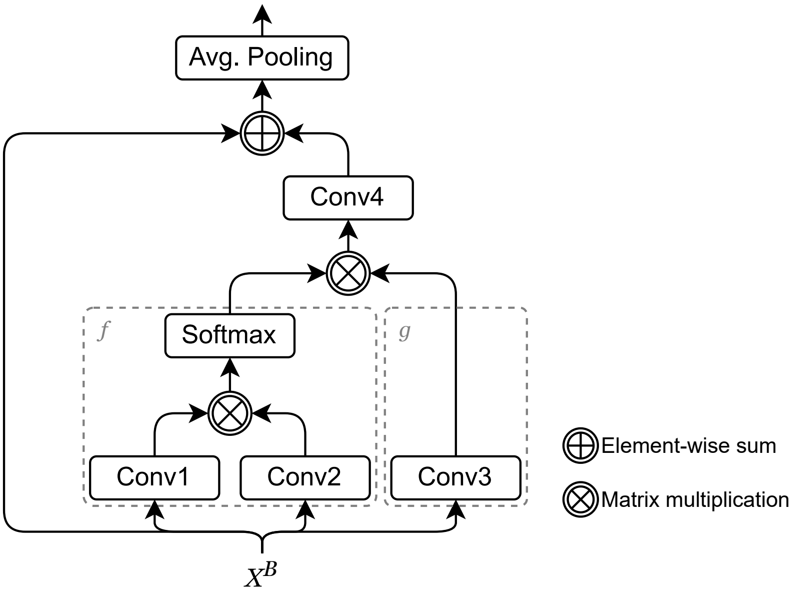


FIGURE 4 Self-attention block. The convolutional layers are all with kernel size of 1 × 1 × 1, but have different weights.

4.3 | Group Learning

Engagement estimation of the elderly relies on subtle interactions among individuals present in a multiparty HRI scenario. It has been known that estimating engagement solely from the elderly is not very reliable. In fact, in the elderly-nurse-robot interaction scenario, nurses are not just the auxiliaries and participants of the interaction. They are also the people who are in daily contact with the elderly and hence they have the prompt judgment about the expressions of the elderly. These judgments are conveyed in their behaviors.

Generally in human conversation, each participant plays a specific role: speaker, addressee, or side-participant who is part of the group of potential speakers but is currently taking on a listening role[48, 49]. In a wild dynamic multiparty interaction, we define the main interaction group to be composed of these participants including speaker, addressee and side-participants. The rest, who may be bystanders and overhearers, is called the background. In the elderly-nurse-robot interaction scenario described above, we hypothesize that *analyzing all participants in the main interaction group and their relationships helps to estimate the engagement of the individual elderly.*

To represent the main interaction group, we propose to use the graph structure where each node corresponds to a participant and stores his/her feature map, and each edge represents the interaction between the participants of the two nodes. We further build graph neural networks (GNNs)[50] to learn the graph representation, which is to compute the hidden representation of each node in the graph by attending over the rest. Specifically, we adapt a graph attention network (GAT)[51] to learn the underlying interactions between nodes by computing attention weights for each edge. The input to our network is $\mathbf{H} = \{h_1, ..., h_N\}$ that are derived from individual learning. The network outputs a new set of transformed node features $\mathbf{H'} = \{h'_1, ..., h'_N\}$.

Following the approach of[51], we first apply a learnable transformation parameterized by a shared weight matrix W to every node feature h_i in order to obtain higher-level feature Wh_i.

Then we compute the score e_{ij} of attention from node j to node i by

$$e_{1j} = \mathbf{a}_1 \cdot Wh_1 + \mathbf{b}_1 \cdot Wh_j \tag{4}$$

$$e_{ij} = \mathbf{a}_2 \cdot Wh_i + \mathbf{b}_2 \cdot Wh_j, \quad \text{for } i \neq 1 \tag{5}$$

where $\mathbf{a}_k$ and $\mathbf{b}_k$ $(k = 1, 2)$ are the weight vectors to be learnt, and "·" represents the dot product of vectors. Here $\mathbf{a}_1$ and $\mathbf{b}_1$ are for the attention from any node to node 1 only, and $\mathbf{a}_2$ and $\mathbf{b}_2$ are shared for all other situations. This special design is due to the fact that the elderly is the main participant among all the participants in the group.

Next we apply the LeakyReLU nonlinearity (with negative input slope $\alpha = 0.2$) to the scores. They are further passed through a softmax operation to generate the normalized weights:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}(e_{ij})\right)}{\sum_{k=1}^{N} \exp\left(\text{LeakyReLU}(e_{ik})\right)}. \tag{6}$$

Finally, the normalized weights are used to compute the new node feature as a linear combination of the old features:

$$h_i' = \sum_{j=1}^{N} \alpha_{ij} W h_j. \tag{7}$$

To stabilize the learning process, we employ multi-head attention, where $K(> 1)$ independent attention mechanisms execute the transformation. The updated node features on the last layer of group learning are averaged, which gives

$$h_i' = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{N} \alpha_{ij}^{k} W^{k} h_j \tag{8}$$

where α_{ij}^{k} is the normalized weights computed by the k-th attention mechanism and W^k is the corresponding weight matrix.

5 | EXPERIMENTS AND RESULTS

This section reports our experiments, which include dataset labeling, implementation detail, main results, and ablation studies.

5.1 | BHEH Dataset

To the best of our knowledge, there is no publicly available labeled dataset for learning elderly-robot interaction, not to mention that in a multiparty scenario. In our experiments, we used BHEH dataset, which was collected in a project that studied the interaction of the elderly with a socially intelligent humanoid robot at Bright Hill Evergreen Home (BHEH), Singapore. The BHEH dataset is a video recording of real elderly-robot interaction where 29 participants aged 60 years and above with dementia participated. There was no constraint imposed on the participants. The elderly talked to a humanoid robot while the nurses occasionally provided help. Moreover, the dataset was collected in a wild, dynamic, and multiparty environment. In the background, nurses might pass through the scene; some old people might sit near to or far from the interaction group; and the interaction participants might leave and join at any time. More detailed information of the dataset and the research setting can be found in[52, 53].

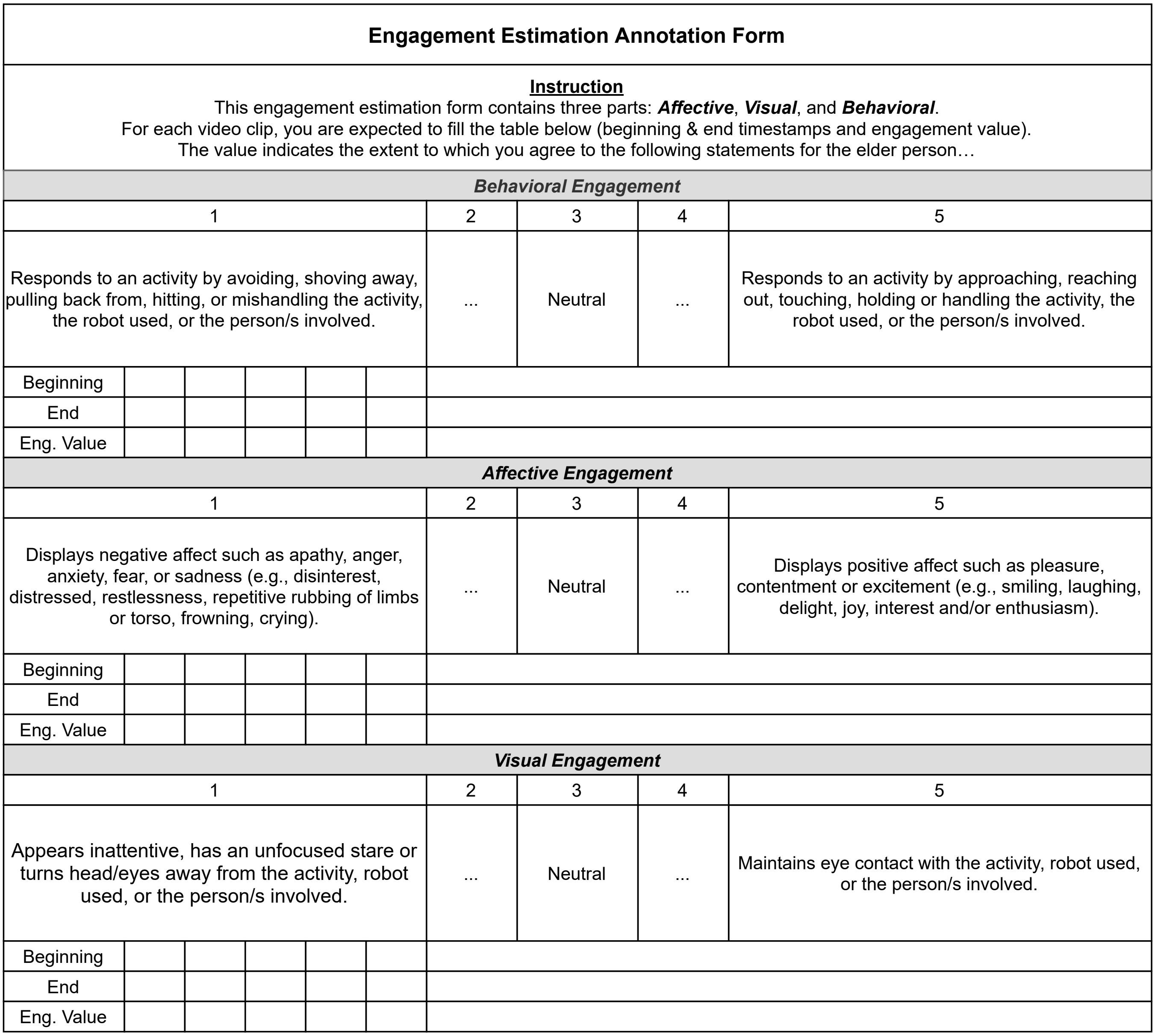
We annotated 43 interaction sessions. The length of the vide is between 3 and 38 minutes (over 560 minutes in total). The number of participants for each session is from 2 to 6.

The label of the data is the engagement score, which was obtained by normalizing the EPWDS values[5] to [0, 1]. We actually removed two components, which were less relevant to our problem, to simplify the process of EPWDS for artificial engagement annotation. The detail of the annotation form is shown in Fig. 5. Each interaction session was annotated at least by two experts. Fig. 6 illustrates the labelled engagement statistics.

5.2 | Implementation Details

The original video was collected in 15 fps. To extract video features by the pre-trained networks, we sampled the videos by selecting one frame from every 5 frames. As a result, for each video clip that records the interaction for about 10.67 seconds, we got 32 frames that were used as the input to our method.

We randomly selected 80% of clips as the training set and the rest was for testing. In our model, we adapted a pretrained ResNet-3D as the backbone, followed by RoI Align with crop size of 7 × 7. We performed the self-attention on individuals' feature map by a spatio-temporal non-local block with embedded Gaussian bottleneck. Individuals' feature maps were fed into a single-layer, 3-head GATs module with hidden size 64, dropout rate 0.5, and $\alpha = 0.2$. We trained our model in two stages. The individual learning module was trained first, and then we fine-tuned the network end-to-end including GATs. Both steps were trained using Adam optimizer in 120 epochs with initial learning rate 10^{-3}, divided by 10 every 50 epochs. The MSE loss is used in the training process.



Engagement Estimation Annotation Form

Instruction

This engagement estimation form contains three parts: ***Affective***, ***Visual***, and ***Behavioral***.
For each video clip, you are expected to fill the table below (beginning & end timestamps and engagement value).
The value indicates the extent to which you agree to the following statements for the elder person…

Behavioral Engagement

1	2	3	4	5
Responds to an activity by avoiding, shoving away, pulling back from, hitting, or mishandling the activity, the robot used, or the person/s involved.	...	Neutral	...	Responds to an activity by approaching, reaching out, touching, holding or handling the activity, the robot used, or the person/s involved.

Beginning						
End						
Eng. Value						

Affective Engagement

1	2	3	4	5
Displays negative affect such as apathy, anger, anxiety, fear, or sadness (e.g., disinterest, distressed, restlessness, repetitive rubbing of limbs or torso, frowning, crying).	...	Neutral	...	Displays positive affect such as pleasure, contentment or excitement (e.g., smiling, laughing, delight, joy, interest and/or enthusiasm).

Beginning						
End						
Eng. Value						

Visual Engagement

1	2	3	4	5
Appears inattentive, has an unfocused stare or turns head/eyes away from the activity, robot used, or the person/s involved.	...	Neutral	...	Maintains eye contact with the activity, robot used, or the person/s involved.

Beginning						
End						
Eng. Value						

FIGURE 5 The simplified annotation form of EPWDS.

5.3 | Results

We attempted to compare our proposed method with the state-of-the-art methods in engagement estimation. Note that there are no publicly available datasets and benchmarks for our problem. Moreover, the scenarios of HRI involved in the state-of-the-art are very different and the results on different datasets may appear different. Hence it is very difficult to perform a fair comparison. We took a compromise approach by applying the prior art methods to our dataset. It should be pointed out that the prior art methods we compared with do not publish their codes, so we implemented them ourselves.

For 2D CNNs[18], as our inputs were just RGB videos without depth information, we adopted its 2D body pose as features and used AlexNet as the comparison model, which achieved similar performance reported in its paper. Inception V2[9] had the same input format as ours, so the implementation just followed its light-inception model architecture. In[20], LSTM was used to classify engagement based on extracted OpenFace and VGG features. To make fair comparisons, we used the same frame length. For

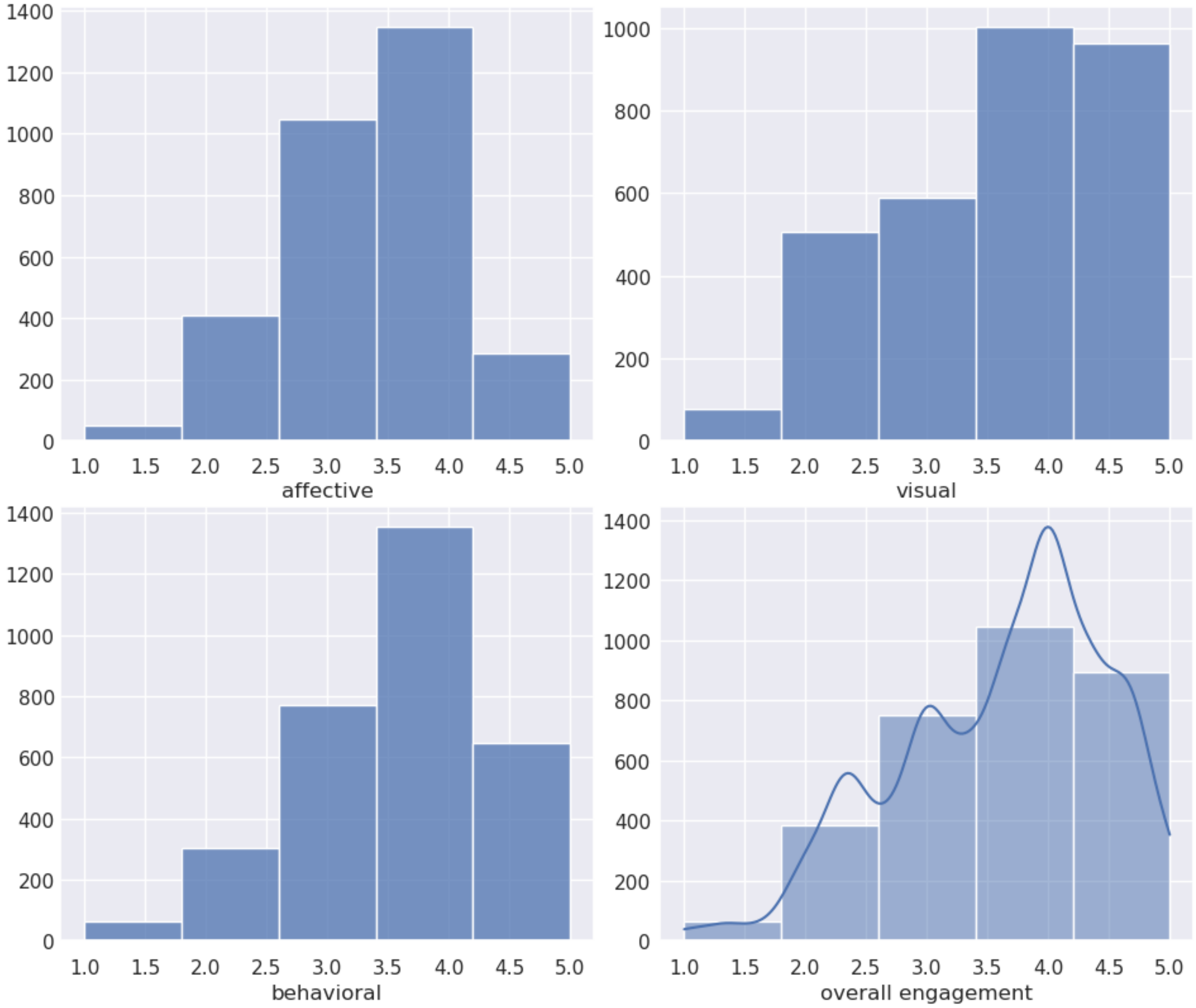


FIGURE 6 Overview of the engagement annotation. The horizontal axis and vertical axis represent the EPWDS engagement value and the video frame count, respectively.

TABLE 3 Engagement estimation of the elderly.

	MSE	**MAE**
2D CNNs[18]	0.1104	0.4123
Inception V2[9]	0.0235	0.1401
LSTM[20]	0.1471	0.3170
Ours	**0.0142**	**0.0773**

evaluation, we used two metrics, MSE of Eq. 1 and mean absolute error (MAE) defined as follows:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|y_i - \hat{y}_i| \tag{9}$$

where y_i is the predicted engagement value, $\hat{y}_i$ is the ground truth, and M is the number of sample video clips.

The results are reported in Table 3. The testing losses are shown in Fig. 7. The performance of our method (with MSE=0.0142) outperforms the prior art. For better understanding, we provide three estimation examples in Fig. 8 which shows the center

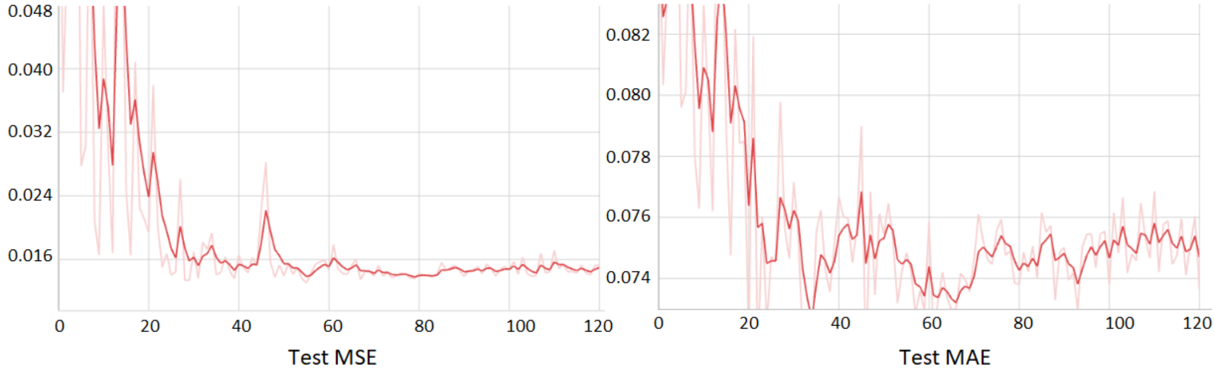
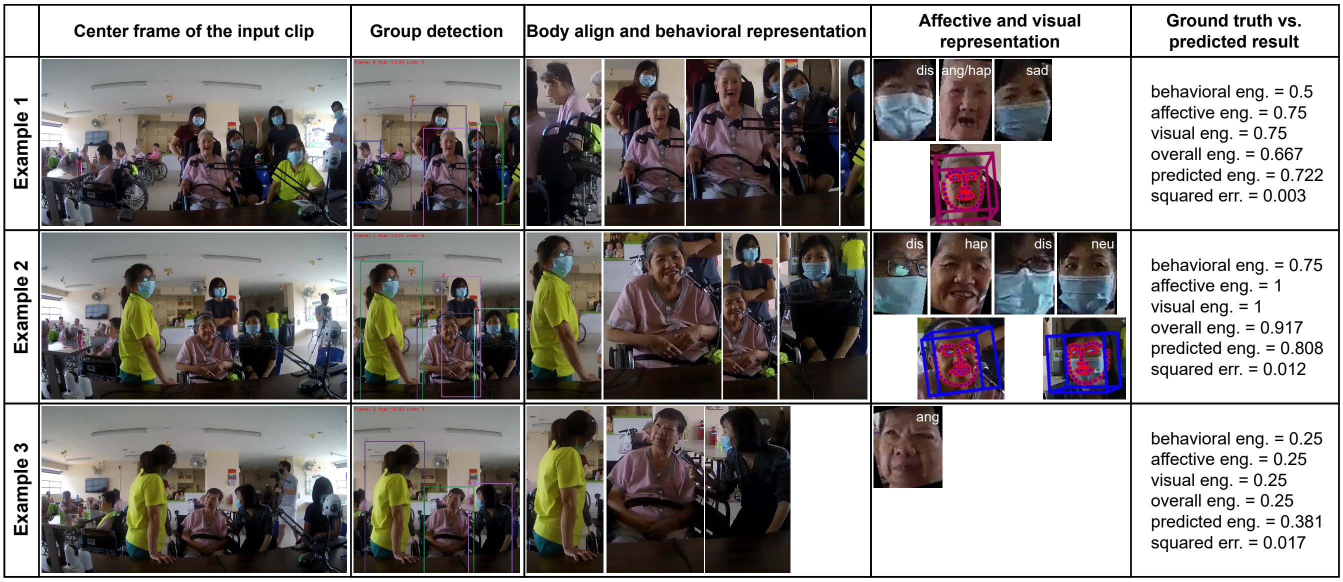


FIGURE 7 Losses of MSE and MAE on the testing set.



	Center frame of the input clip	Group detection	Body align and behavioral representation	Affective and visual representation	Ground truth vs. predicted result
Example 1					behavioral eng. = 0.5 affective eng. = 0.75 visual eng. = 0.75 overall eng. = 0.667 predicted eng. = 0.722 squared err. = 0.003
Example 2					behavioral eng. = 0.75 affective eng. = 1 visual eng. = 1 overall eng. = 0.917 predicted eng. = 0.808 squared err. = 0.012
Example 3					behavioral eng. = 0.25 affective eng. = 0.25 visual eng. = 0.25 overall eng. = 0.25 predicted eng. = 0.381 squared err. = 0.017

FIGURE 8 Visualization of the engagement estimation results.

frame of the example clip, group detection, behavioral representation from body align, affective and visual representation, and estimation results from left column to right column. It can be seen that our method achieves good results even under challenging conditions. Particularly, in examples 1 and 5, participants' bodies are detected and used for feature extraction, but the leftmost person is not desired. This is a counterexample of the main group detection. In contrast, examples 2 and 3 detect all the participants successfully. In terms of affective and visual engagement representation, some inconsistency and instability occur due to the masks and senescent faces. For instance, the elderly in example 1 could not make meaningful expressions and the visual features in example 3 are also missed. This may explain why those methods only involving facial information often fail to produce good results. In addition, the elderly from examples 1 and 3 is not good at body language, so the information from side participants helps in the estimation, *e.g.*, the body representation captures the raised hand of the nurse.

5.4 | Ablation Studies

We ran a number of ablations to analyze our method. The results are reported in Table 4. $\mathcal{B}$, $\mathcal{A}$, and $\mathcal{V}$ are the results of using a single engagement element of behavioral, affective, and visual. It can be seen that the results are inferior to that produced by the proposed multi-element method. The self-attention module also helps improve the performance by 0.0425 in MSE and 0.1357 in MAE. By employing the GATs in our group learning module, the results have 0.0171 increase in MSE, which means that the signals from side-participants contribute to the estimation in our multiparty elderly-robot interaction.

TABLE 4 Ablation results.

	MSE	**MAE**
$\mathcal{B}$	0.0451 (↓0.0309)	0.1750 (↓0.0977)
$\mathcal{A}$	0.1235 (↓0.1093)	0.3690 (↓0.2917)
$\mathcal{V}$	0.1567 (↓0.1425)	0.4184 (↓0.3411)
w/o self-attention	0.0567 (↓0.0425)	0.2130 (↓0.1357)
w/o GATs	0.0313 (↓0.0171)	0.1091 (↓0.0318)

6 | CONCLUSIONS

This paper has presented an automatic method for analyzing wild multiparty human-robot interaction (HRI) videos and estimating the engagement state of the elderly–the main participant–in the HRI. The method adapts pre-trained models to extract behavioral, affective and visual features of the participants in the main interaction group from real-world videos of such interactions. A supervised machine learning model consisting of a self-attention network for individual learning and a graph attention network for group learning is designed to take these multi-modal features of all participants as input and output the predicted engagement state of the elderly. To support the supervision, we have created a labeled engagement dataset from a video recording of multiparty elderly-robot interaction in a wild environment. By utilizing multi-modal features and exploiting individual and group learning, our method can effectively predict the engagement and outperform the prior art, as confirmed by the experimental results.

References

1. Ghafurian M, Hoey J, and Dautenhahn K. Social Robots for the Care of Persons with Dementia: A Systematic Review. ACM Transactions on Human-Robot Interaction. 2021;**10**(4):1–31.

2. Perugia G, Díaz-Boladeras M, Català-Mallofré A, Barakova EI, and Rauterberg M. ENGAGE-DEM: A Model of Engagement of People With Dementia. IEEE Transactions on Affective Computing. 2022;**13**(2):926–943.

3. Ageing and health; 2021. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

4. Poggi I. Mind, hands, face, and body: A sketch of a goal and belief view of multimodal communication. In: Body - Language - Communication. vol. 1. De Gruyter Mouton; 2013. p. 627–647.

5. Jones C, Sung B, and Moyle W. Engagement of a Person with Dementia Scale: Establishing content validity and psychometric properties. Journal of Advanced Nursing. 2018;**74**(9):2227–2240.

6. Salam H, Celiktutan O, Hupont I, Gunes H, and Chetouani M. Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions. IEEE Access. 2017;**5**:705–721.

7. Celiktutan O, Skordos E, and Gunes H. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. IEEE Transactions on Affective Computing. 2019;**10**(4):484–497.

8. Ben Youssef A, Clavel C, and Essid S. Early Detection of User Engagement Breakdown in Spontaneous Human-Humanoid Interaction. IEEE Transactions on Affective Computing. 2019;**12**(3):776–787.

9. Saleh K, Yu K, and Chen F. Improving Users Engagement Detection Using End-to-End Spatio-Temporal Convolutional Neural Networks. In: Companion of the ACM/IEEE International Conference on Human-Robot Interaction; 2021. p. 190–194.

10. Del Duchetto F, Baxter P, and Hanheide M. Are You Still With Me? Continuous Engagement Assessment from a Robot's Point of View. Frontiers in Robotics and AI. 2020;**7**:116.

11. Ben-Eliyahu A, Moore D, Dorph R, and Schunn CD. Investigating the Multidimensionality of Engagement: Affective, Behavioral, and Cognitive Engagement across Science Activities and Contexts. Contemporary Educational Psychology. 2018;**53**:87–105.

12. Zhu B, Lan X, Guo X, Barner KE, and Boncelet C. Multi-Rate Attention Based GRU Model for Engagement Prediction. In: Proceedings of the International Conference on Multimodal Interaction; 2020. p. 841–848.

13. Rudovic O, Park HW, Busche J, Schuller B, Breazeal C, and Picard RW. Personalized Estimation of Engagement from Videos Using Active Learning with Deep Reinforcement Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019. p. 217–226.

14. Sümer Ö, Goldberg P, D'Mello S, Gerjets P, Trautwein U, and Kasneci E. Multimodal Engagement Analysis from Facial Videos in the Classroom. arXiv:2101.04215; 2021.

15. Monkaresi H, Bosch N, Calvo RA, and D'Mello SK. Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. IEEE Transactions on Affective Computing. 2017;**8**(1):15–28.

16. Abedi A, and Khan S. Affect-Driven Engagement Measurement from Videos. arXiv:2106.10882; 2021.

17. Gao N, Shao W, Rahaman MS, and Salim FD. N-Gage: Predicting in-Class Emotional, Behavioural and Cognitive Engagement in the Wild. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2020;**4**(3):1–26.

18. Anagnostopoulou D, Efthymiou N, Papailiou C, and Maragos P. Engagement Estimation During Child Robot Interaction Using Deep Convolutional Networks Focusing on ASD Children. In: IEEE International Conference on Robotics and Automation; 2021. p. 3641–3647.

19. Jain S, Thiagarajan B, Shi Z, Clabaugh C, and Matarić MJ. Modeling Engagement in Long-Term, in-Home Socially Assistive Robot Interventions for Children with Autism Spectrum Disorders. Science Robotics. 2020;**5**(39):eaaz3791.

20. Steinert L, Putze F, Küster D, and Schultz T. Towards Engagement Recognition of People with Dementia in Care Settings. In: Proceedings of the International Conference on Multimodal Interaction; 2020. p. 558–565.

21. Guhan P, Agarwal M, Awasthi N, Reeves G, Manocha D, and Bera A. ABC-Net: Semi-Supervised Multimodal GAN-based Engagement Detection Using an Affective, Behavioral and Cognitive Model. arXiv:2011.08690; 2020.

22. Guo G, Guo R, and Li X. Facial Expression Recognition Influenced by Human Aging. IEEE Transactions on Affective Computing. 2013;**4**(3):291–298.

23. Fölster M, Hess U, and Werheid K. Facial age affects emotional expression decoding. Frontiers in Psychology. 2014;**5**.

24. Hara K, Kataoka H, and Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 6546–6555.

25. Krizhevsky A, Sutskever I, and Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;**25**.

26. Cohen-Mansfield J, Dakheel-Ali M, and Marx MS. Engagement in Persons With Dementia: The Concept and Its Measurement. The American Journal of Geriatric Psychiatry. 2009;**17**(4):299–307.

27. Trahan MA, Kuo J, Carlson MC, and Gitlin LN. A systematic review of strategies to foster activity engagement in persons with dementia. Health Education & Behavior. 2014;**41**(1_suppl):70S–83S.

28. Moyle W, Jones CJ, Murfield JE, Thalib L, Beattie ERA, Shum DKH, et al. Use of a Robotic Seal as a Therapeutic Tool to Improve Dementia Symptoms: A Cluster-Randomized Controlled Trial. Journal of the American Medical Directors Association. 2017;**18**(9):766–773.

29. Feng Y, Perugia G, Yu S, Barakova EI, Hu J, and Rauterberg GWM. Context-Enhanced Human-Robot Interaction: Exploring the Role of System Interactivity and Multimodal Stimuli on the Engagement of People with Dementia. International Journal of Social Robotics. 2021;**14**(3):807–826.

30. Baltrusaitis T, Zadeh A, Lim YC, and Morency LP. OpenFace 2.0: Facial Behavior Analysis Toolkit. In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition; 2018. p. 59–66.

31. Parkhi OM, Vedaldi A, and Zisserman A. Deep Face Recognition. In: Proceedings of the British Machine Vision Conference; 2015. p. 41.1–41.12.

32. Oertel C, Castellano G, Chetouani M, Nasir J, Obaid M, Pelachaud C, et al. Engagement in Human-Agent Interaction: An Overview. Frontiers in Robotics and AI. 2020;**7**:92.

33. Sidner CL, Lee C, Kidd CD, Lesh N, and Rich C. Explorations in engagement for humans and robots. Artificial Intelligence. 2005;**166**(1-2):140–164.

34. Castellano G, Pereira A, Leite I, Paiva A, and McOwan PW. Detecting User Engagement with a Robot Companion Using Task and Social Interaction-Based Features. In: Proceedings of the International Conference on Multimodal Interfaces; 2009. p. 119–126.

35. Finn JD, and Zimmer KS. Student Engagement: What Is It? Why Does It Matter? In: Christenson SL, Reschly AL, and Wylie C, editors. Handbook of Research on Student Engagement. vol. 840. Springer; 2012. p. 97–131.

36. O'Brien HL, and Toms EG. What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science and Technology. 2008;**59**(6):938–955.

37. Cohen-Mansfield J, Marx MS, Freedman LS, Murad H, Regier NG, Thein K, et al. The Comprehensive Process Model of Engagement. The American Journal of Geriatric Psychiatry. 2011;.

38. Archambault I, and Dupéré V. Joint trajectories of behavioral, affective, and cognitive engagement in elementary school. The Journal of Educational Research. 2017;**19**(10):859–870.

39. Corrigan LJ, Peters C, Küster D, and Castellano G. Engagement perception and generation for social robots and virtual agents. In: Toward Robotic Socially Believable Behaving Systems. vol. 19. Springer International Publishing; 2016. p. 29–51.

40. Oertel C, Jonell P, Kontogiorgos D, Mora KF, Odobez JM, and Gustafson J. Towards an Engagement-Aware Attentive Artificial Listener for Multi-Party Interactions. Frontiers in Robotics and AI. 2021;**8**.

41. Chen CFR, Panda R, Ramakrishnan K, Feris R, Cohn J, Oliva A, et al. Deep analysis of cnn-based spatio-temporal representations for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 6165–6175.

42. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al.. The Kinetics Human Action Video Dataset. arXiv:1705.06950; 2017.

43. Zhang Y, Sun P, Jiang Y, Yu D, Yuan Z, Luo P, et al.. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. arXiv:2110.06864; 2021.

44. He K, Gkioxari G, Dollár P, and Girshick R. Mask R-CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 2961–2969.

45. Deng J, Guo J, Ververas E, Kotsia I, and Zafeiriou S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2020. p. 5203–5212.

46. She J, Hu Y, Shi H, Wang J, Shen Q, and Mei T. Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 6248–6257.

47. Wang X, Girshick R, Gupta A, and He K. Non-Local Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 7794–7803.

48. Goffman E. Forms of talk. University of Pennsylvania Press; 1981.

49. Clark HH. Using language. Cambridge University Press; 1996.

50. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, and Monfardini G. The graph neural network model. IEEE Transactions on Neural Networks. 2008;**20**(1):61–80.

51. Velič P, Cucurull G, Casanova A, Romero A, Lio P, and Bengio Y. Graph Attention Networks. arXiv:1710.10903; 2017.

52. Mishra N, Tulsulkar G, Li H, Thalmann NM, Er LH, Ping LM, et al. Does elderly enjoy playing bingo with a robot? a case study with the humanoid robot nadine. In: Computer Graphics International Conference; 2021. p. 491–503.

53. Tulsulkar G, Mishra N, Thalmann NM, Lim HE, Lee MP, and Cheng SK. Can a humanoid social robot stimulate the interactivity of cognitively impaired elderly? A thorough study based on computer vision methods. The Visual Computer. 2021;**37**:3019–3038.

How to cite this article: