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

Zhijie Zhang, Jianmin Zheng, and Nadia Magnenat Thalmann

Abstract—With the improvement of the standard of living, humans have the increasing longevity, but improving the quality of life for the elderly remains a challenge. In recent years, many researchers have studied the use of assistive agents in clinical treatment and nursing homes, where the concept of engagement has been proposed and has become an essential indicator. However, traditional engagement estimation often requires expert involvement and is in controlled dyadic interaction environment. In this paper, we use real-world video recordings to estimate old people's engagement in a multiparty human-robot interaction scenario. For this purpose, we place a humanoid robot in a non-profit nursing home, where the elderly are accompanied by nurses and engage in spontaneous conversations with the robot. We propose a framework based on self-attention mechanism and graph attention neural networks, which uses videos as input to automatically generate an engagement value between 0 and 1. The model is built on the components of engagement discussed in geriatric psychiatry (behavioral, affective, and visual) and video representations for interaction understanding (group and individual levels). We test the proposed framework using 22 wild multiparty elderly-robot interaction videos by answering the following questions: What is the engagement value of the elderly? Who makes up the main interaction group? Could the information from other participants help make an estimation? The results show that our model is capable of detecting the key participants and estimating the engagement of the elderly. Moreover, information from the surroundings can considerably improve the estimation results.

Index Terms—Human-Robot Interaction, Engagement Estimation, Affective Computing, Multiparty Interaction, the Elderly.

1 Introduction

The degree and pace of population ageing have increased dramatically in the past decades. According to the World Health Organization, the proportion of the world's population over 60 years will nearly double from 12% in 2015 to 22% in 2050 [1]. In addition, old people usually experience significant declines in physical and cognitive capacities, so round-the-clock medical and psychological care are important for them. Many studies claimed that deploying robots in the healthcare setting is a possible solution, *e.g.*, they support people with dementia (PwD) and their caregivers by reducing costs and improving the independence of the old people [2], [3].

This work centers on the elderly-robot interaction and attempts to come up with an approach to estimating the engagement of the elderly. Here, engagement is defined as the inner state of a participant attributing to being together with the other participants and continuing the interaction [4]. Many studies have approved the importance of engagement in both human-human interaction (HHI) and human-robot interaction (HRI) [5]. By giving robots the ability to recognize engagement state, they are able to generate socially acceptable behavior, form long-term connections, and more importantly to design activities for seniors that allow them to use their skills and feel a sense of belonging.

Automated engagement estimation is comparable to affective computing and behavior recognition in computer vision area, but it goes a step further to probe the inner intention behind the apparent behavior and emotion. Previous works have been proposed to estimate engagement in various scenarios such as general HRI [6], [7], [8], [9], museum tour guide [10], and classroom or distance learning [11], [12], [13], [14], [15], [16], [17], toddlers and seniors healthcare [5], [18], [19], [20]. Conventional approaches use nonverbal cues such as proxemics, body pose, gaze patterns, facial expressions, and context information to build engagement estimation classifiers. However, as shown in Fig. 1, when the research is expanded to some special populations or more complex circumstances, such challenging work is still lacking. This work attempts to confront these challenging issues:

- Q1 Given that the non-verbal signals from **the elderly**, which alter in facial shape and patterns of body behaviors along with aging [21], [22], can we use a pure computer vision approach to accurately estimate engagement state?
- Q2 From dyadic to **multiparty HRI**, how to understand the dynamics and stability of the interaction in such a scenario? Can we propose an automated approach to analyze an old person?
- Q3 In unconstrained **wild** space, how to understand the complex environment (moving people, bad lighting, and confusing objects)?

To this end, we propose a novel deep learning frame-

- *Z. Zhang and J. Zheng are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. E-mail: {zhijie002, asjmzheng}@e.ntu.edu.sg*
- *N. M. Thalmann is with the MIRALab, University of Geneva, Switzerland. E-mail: thalmann@miralab.ch*
- *N. M. Thalmann is with the MIRALab, University of Geneva, Switzerland. E-mail: thalmann@miralab.ch*
- *N. M. Thalmann is with the MIRALab, University of Geneva, Switzerland. E-mail: thalmann@miralab.ch*

Manuscript received April 19, 2005; revised August 26, 2015.

work for automatically estimating engagement from videos of real-world multiparty elderly-robot interactions. Specifically, as shown in Fig. 2, our approach takes video sequences as input and output the estimated engagement state. For each image, we first perform human detection and pose estimation, on the basis of which we implement 3D pose reconstruction. Through the reconstructed scenes and the positions of people in different frames, we explore the relationships between people, and then obtain the main groups interacting with the robot, after which we continue the scale down to the main elderly, which is done through facial recognition. We perform feature extraction and preliminary estimation at each of these three scales. Specifically, on the global scale, we use the ResNet-50 [23] as the backbone with self-attention mechanism [24]. For the group branch, we utilize both image features from pre-trained convolutional networks (CNNs) and high-level social features such as body pose, facial landmarks as the node information of graph attention networks (GATs) [25] where the edges model the interactions among individuals. Individual estimation is similar to group branch but without GATs. The final estimation is a combination of these estimations. The main contributions of the paper are

- We propose an automated approach that analyzes wild multiparty HRI videos and estimates the engagement of the elderly.
- We explore three scales' information and approve that the combination of global, group, and individual information achieves better performance than only using individual features.

This paper is organized as follows. Sec. 2 presents the related work of engagement estimation in HRI, especially for the elderly. Sec. 3 describes our proposed approach for estimating the engagement in a real-world scenario and the method of detecting main conversation group members. In Sec. 4, we elaborate the dataset we collected and used for our experiments, as well as the annotation process. The details of the implementation, evaluation metrics, and main results are presented and discussed in Sec. 5. Finally, Sec. 6 summarizes the paper and outlines possible future work.

2 Related Work

The estimation of engagement encompasses a wild range of fields from computer vision to psychological science and psychiatric nursing. The related work provided in this section includes the concepts of engagement (Sec. 2.1) and the methodology for automatically estimating engagement (Sec. 2.2).

2.1 What Is Engagement?

2.1.1 Definition of Engagement

Before we discuss engagement in HRI, let's clarify the definition of engagement first. According to Oertel *et al*. [27], the notion of engagement is ambiguous among different research domains, which are regarded as a state or a process. Participants are either engaged or not engaged in terms of state, but, by contrast, a classical process definition was proposed by Sidner *et al*. [28] as the process through which

Fig. 1. Two sample frames from a video recording of real-world multiparty HRI, demonstrating 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, t_a and t_b denoting two time stamps.

interactors establish, maintain, and complete their perceived connection to each other during an interaction.

It is important to note that the term ***state*** is used here to distinguish it from *process*, which represents objectively observed facts in HHI or HRI, *i.e.*, the participants are within interaction or not. This is different from the ***state*** used in the phrases—*engagement state estimation* or *inner state*—depicting participants' the whole mental, feeling, emotional, and cognitive states. In this paper, we adopt the definition of engagement from [4], which is the participant's inner state of being together with other participants and continuing the interaction.

2.1.2 Components of Engagement

According to [3], [11], [14], [26], [29], [30], [31], [32], [33], [34], engagement contains different components. The distinctions and definitions of these components vary across research areas. In general, common components include: behavioral, affective, visual, verbal, social, and cognitive. Moreover, these components are not mutually exclusive but often overlap with each other.

- **Behavioral:** involves observable behaviors such as approaching, touching, avoiding, or hitting, *etc*.
- **Affective:** is defined as the affective 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.

TABLE 1
Comparison of Engagement Estimation Approaches

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

[1] The difference between multiparty and group is that multiparty treats participants separately but group is a whole.
[2] Modalities: vis = visual, dpt = depth, per = personality, aud = audio, phy = physiological, env = environmental, and txt = text.
[3] The symbol & indicates using both and comparing with each other. + 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.

- **Cognitive:** pertains to the psychological investment and effort allocation of the person in order to fully comprehend the situation.
- **Social:** includes social activities such as encouraging or disrupting others.

In our work, since our target is to estimate the engagement of the elderly with a humanoid robot in casual conversation via a pure computer vision approach, we select behavioral, affective, and visual engagement. Because of the input modality, the verbal component is eliminated, while the cognitive and social components are overlooked due to the participant's physical and mental conditions.

2.1.3 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 perform and feel differently in HHI and HRI, the estimation of engagement in these two contexts is similar in methods. For example, in [35], Oertel *et al*. demonstrated the knowledge from the HHI can be applied to HRI. Therefore, we will discuss both in Sec. 2.2.

On the other hand, the application scenarios of engagement estimation are also different, such as everyday conversations, healthcare, and learning situations, among others. In different scenarios, engagement is often expressed differently, while the dominance of its components varies. As a result, the corresponding estimation methods are different and hard to make comparison, let alone find a universal approach.

2.2 Automated Engagement Estimation

In terms of the methodology, traditional approaches [6], [7], [8], [15], [17] extract high-level social features, *e.g*., body pose, facial expressions, gaze, and task-related information, followed by a machine learning classifier. Several papers are heuristic, demonstrating that specific features are meaningful by using and comparing unimodal and multimodal feature combinations. Recently, with the development of computer vision, more and more deep learning methods have been proposed [9], [10], [12], [13], [14], [18], [26]. A summary of the literature is shown in Table 1, grouped by estimation approaches.

2.2.1 Machine Learning Classifiers

In general HRI, Salam *et al*. [6] classified engagement using support vector machines (SVM) and random forests (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 is from Celiktutan *et al*. [7]. Ben-Youssef *et al*. [8] investigated 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 is used.

Another widely investigated situation is online and in-class learning. Monkaresi *et al*. [15] explored engagement where students complete an online writing activity. Heart rate, animation units (AUs), and local binary patterns are 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 class. They used a set of features from wearable and indoor weather sensors to infer students' engagement.

2.2.2 Deep Neural Networks

The aforementioned approaches require expert design of input features and cannot deal with large feature dimensions efficiently, *e.g.*, when pixel values from face images are used as input. Del Duchetto *et al*. [10] propose 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 is 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 (GRU) network to predict engagement of students learning online. Taking the advantage of the published dataset from [8], Saleh *et al*. [9] applied Inflated 3D ConvNets (I3D) architecture to predict engagement state in an end-to-end way.

In addition, Anagnostopoulou *et al*. [18] estimated the engagement of children with autism spectrum disorder when they interact with robots. They compared architectures of AlexNet, CNNs, and LSTM using 2D or 3D poses. Rudovic *et al*. [13] proposed a personalized reinforcement learning (RL) approach to estimate engagement level (low, medium, high) from videos of child-robot interactions, where queried videos are 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 offline 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*. [26] 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 first and then generate engagement labels.

2.3 Engagement for the Elderly

In geriatric psychiatry, research on engagement is well established. In [36], the authors described an Observational Method of Engagement (OME) for persons with dementia, which is one of the most well-known tools many studies used to measure engagement [37]. Followed by 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. In [3], ENGAGE-DEM, an affective computing framework is presented, specifying the components of engagement in HRI.

Moreover, robotic and computer assistance has been approved as an effective intervention. Moyle *et al*. [38] developed a robot seal for PwD. They found that participants were more engaged with it compared to a toy, and robot seal is more effective than usual care in improving mood states and agitation. Similarly, Feng *et al*. [39] introduced an interactive system involving a display and a robotic sheep to engage PwD. They claimed that multimodal stimuli play a significant role in promoting engagement.

However, all the previously mentioned engagement estimation methods require expert involvement. For automated estimation methods, Steinert *et al*. [20] presented a vanilla LSTM model to predict emotional engagement based on visual (extracted OpenFace [40] and VGGFace [41] facial features) and contextual information (daytime, wellbeing, *etc*.).

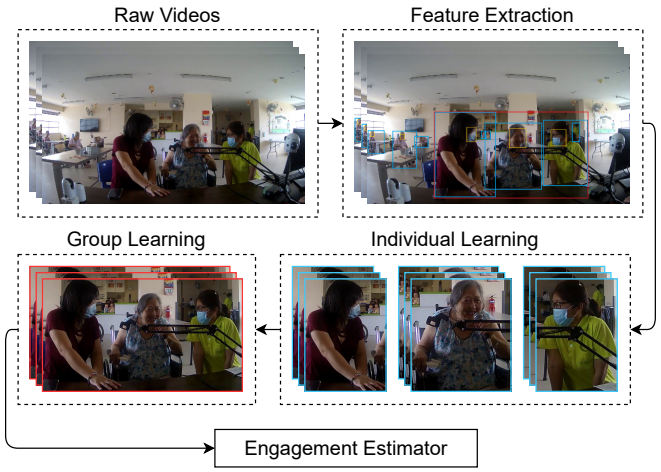

Fig. 2. Overview of the proposed engagement estimation architecture. This architecture is composed of four modules, (i) Feature Extraction, (ii) Individual Learning, (iii) Group Learning, and (iv) Engagement Estimator.

3 Engagement Estimation of the Elderly

In this section, we describe our proposed engagement estimation approach of the elderly from wild multiparty HRI. We use raw video recordings of the elderly-robot interaction, described in Sec. 4. Formally, we define the dataset as $\mathcal{V} = \{V_1, ..., V_m, ..., V_M\}$, where V_m is video recordings of interaction session i. For every interaction session, normally there is only one old participant p. Moreover, each interaction session includes K video clips, denoted as $V_m = \{v_m^1, ..., v_m^k, ..., v_m^K\}$, where K may vary per session. Lastly, each video clip is associated with a target ground truth $y_m^k \in [0, 1]$, corresponding to the elderly's engagement state. From 0 to 1, the output value indicates an increase in engagement. Given these data, our goal is to output an estimated value of engagement $\hat{y}_m^k$ that is as close as possible to the ground truth.

As mentioned before, our task is very challenging and complicated, because we contend with the background clutter and accurately capture the inner engagement state of a specific old person from a dynamic multiparty interaction. To address them, we design a novel framework for engagement estimation of the elderly. The overview of the architecture is illustrated in Fig. 2. First, we use the pre-trained I3D as the backbone to obtain spatio-temporal representations of the input videos. Then, we design an individual learning module for refining individual features by adding a self-attention mechanism. Following that, we construct GATs to learn the relationships within the main group and group-level information. Finally, we make inference through an estimator outputting a value that reflects the elderly's engagement state.

In comparison to other related studies, the main advantage of our approach is that we attempt to provide answers the three questions raised in Sec. 1. First, because the facial features and body movements of older adults

are difficult to recognize, we design an individual-group structure, using the attention mechanism to improve the quality of the representations of individual features, and construct a graph network to learn the response from nurses as well as the relationships, which further helps us understand the engagement of the elderly. Second, in our approach, we use reconstructed 3D skeleton information and image information to detect the main interaction group in a video clip, thus improving the accuracy of excluding other distractions. The detailed explanation is elaborated in the following subsections.

3.1 Feature Extraction

We use I3D as the backbone to capture the spatio-temporal context of an input video clip. In I3D, ImageNet [42] pre-trained convolutional kernels are expanded into 3D, allowing it to seamlessly learn effective spatio-temporal representations. Motivated by the promising performance of I3D models in a wide range of video-related benchmarks, we exploit the feature representations offered by this backbone pre-trained on Kinetics 400 [43] at multiple resolutions. More specifically, we use the deep spatio-temporal feature maps ($X^G \in \mathbb{R}^{1024\times8\times7\times7}$) extracted from the final convolutional layer as a rich semantic representation describing the entire video clip. These deeper features provide low-resolution yet high-level representations that encode a summary of the video. In addition, accurate recognition of individuals' action rely on finer details which are often absent in very deep coarse representations. To extract fine spatio-temporal representations for the individuals, we use the higher resolution feature maps ($X^M \in \mathbb{R}^{832\times16\times14\times14}$) from the intermediate Mixed-4f layer of I3D. As depicted in Fig. 3, from this representation we extract three feature maps, *i.e.*, the beginning, temporally-centered, and ending feature maps, corresponding to the beginning, center, and ending frames of the input video clip.

In addition, to eliminate the interference of redundant background information on learning, we conduct multi-person tracking and multi-face detection on the input video clips. To do that, we adopt the state-of-the-art methods ByteTrack [44] and RetinaFace [45] to gain the bounding boxes of the bodies and faces. ByteTrack is a multi-object tracking method associating every detection box instead of only the high score ones, while the RetinaFace performs pixel-wise face localization on various scales of faces by taking advantages of joint extra-supervised and self-supervised multitask learning. Given the bounding boxes, we use RoIAlign [46] to project the coordinates on the frames' feature maps and slice out the corresponding features for each individual's bounding box. We use an average pooling to calculate the individuals' feature maps. Formally,

$$\mathbf{F}_m^k = AP\left(RoI\left(E\left(v_m^k\right)\right)\right) \tag{1}$$

where $\mathbf{F}_m^k$ denoting the extracted individuals' feature maps from video clip v_m^k, includes participants of number P_m^k. AP, RoI, and E are the average pooling, RoI Align, and feature extraction operations, respectively. We conduct two-level RoI align separately, *i.e.*, $RoI = \{RoI_{body}, RoI_{face}\}$. In addition, for simplifying the notations, we will eliminate m and k in the following sections.

3.2 Individual Learning

Despite being localized to the body and face bounding boxes, these representations still lack emphasis on visual clues that play a crucial role in understanding the underlying information, *e.g.*, a person's body posture and facial expressions, which related to the behavioral, affective, and visual components of engagement. To overcome this, we adopt the self-attention mechanism [47] to directly learn the interactions between any two feature positions of an individual's feature representation and accordingly leverage this information to refine each individual's body and face feature maps.

The self-attention module computes the response at a position in a sequence by attending to all positions and taking their weighted average in an embedding space. This mechanism was originally introduced in machine translation, but in computer vision tasks, attention mechanisms were designed to discover the important spatial regions in an image or the critical frames in a video. In our framework the self-attention module functions as a non-local operation [24] and computes the response at each position by attending to all positions in an individual's feature map. The output of the self-attention module contextualizes the input bounding box feature map with visual clues and thus, enriches the individual's representation by highlighting the most informative features. As substantiated by ablation studies in Sec. 5, capturing such fine details significantly contribute to the estimation performance.

We utilize three separate self-attention modules that take the averaged body features and features as inputs. These three modules are designed to capture the behavioral ($\mathcal{B}$), affective ($\mathcal{A}$), and visual ($\mathcal{V}$) engagement. Sequentially, a fully connected (FC) layer is used to project the revised feature maps to the same size. In this step, what we expect is to learn a new representation of three engagement components from body and face maps. Finally, we concatenate each individual's features as the input of the next group learning module. The refined individual feature map is defined as

$$\mathbf{H} = \left[\alpha_{\mathcal{B}}\left(\mathbf{F}_{body}\right), \alpha_{\mathcal{A}}\left(\left[\mathbf{F}_{body}, \mathbf{F}_{face}\right]\right), \alpha_{\mathcal{V}}\left(\mathbf{F}_{face}\right)\right]. \tag{2}$$

Here, $\mathbf{H} = \{h_1, ..., h_p, ..., h_P\}$ is the output from individual module and $[\cdot,\cdot]$ denotes concatenation. α is the attention operation.

3.3 Group Learning

Engagement estimation of the elderly relies on uncovering subtle interactions among individuals present in a multi-party HRI scenario. Based on the studies of aging faces and postures, it is not robust to estimate engagement solely from the old participant. However, it is worth noting that, in the elderly-nurse-robot interaction scenario, nurses are not only the auxiliaries and participants of the interaction, but as the people who are in daily contact with these elderly people, they have the most prompt judgment about the minor expressions of the elderly people, and these judgments will be conveyed in their behaviors. Therefore, in the HRI scenario described above, we suggest a hypothesis that ***analyzing all participants in the main interaction group and the relationships between them helps us to estimate the***

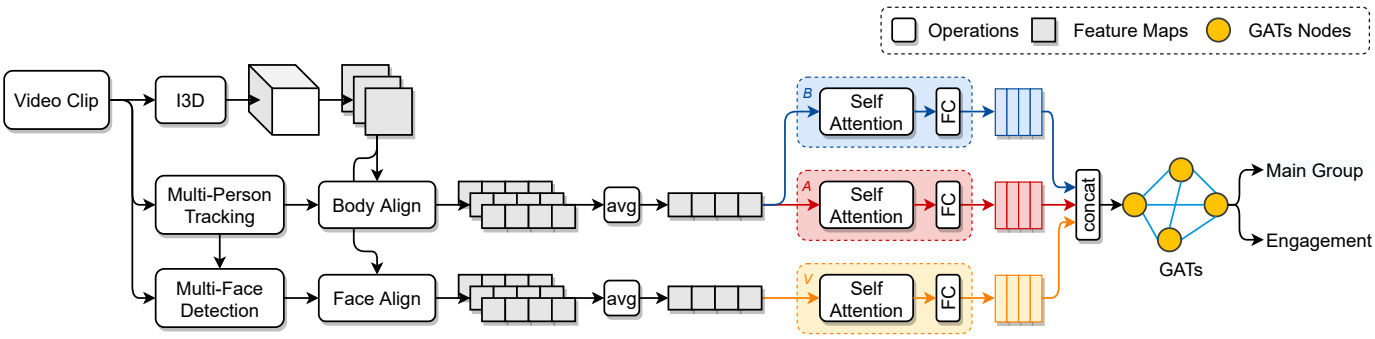

Fig. 3. Illustration of the feature extraction and individual learning modules. We employ a I3D model for extracting individual representations. Multi-person tracking and multi-face detection are used to get the bounding boxes in order to align and slice out corresponding body and face feature maps, which are then pooled for individual learning. Three-branch self-attention mechanism is applied to refine the feature maps, *i.e.*, learning the behavioral ($\mathcal{B}$), affective ($\mathcal{A}$), and visual ($\mathcal{V}$) engagement, respectively. The concatenation of these learned three components is the final representation of an individual.

engagement of individual elder person. If this hypothesis is proved true, then we can provide an answer to Q1.

Furthermore, how to identify the main interaction group in a video clip of dynamic multiparty interaction in the wild involves Q2 and Q3. According to studies for human conversation, a widely adopted classification of roles is proposed by Goffman [48]. He presented a participation structure model where each participant is assigned a participation role, *i.e.*, speaker, addressee, and side-participant (person is part of the group of possible speakers but who currently are taking on a listening role). Clark [49] further made a distinction between participants and non-participants. The former includes speaker, addressee, and side-participant, while the latter includes bystanders and overhearers. By using this concept to classify the detected people into main group and background, we are able to solve Q2 and Q3 if we find an automated approach.

In order to solve the above questions, we propose a group learning module. This problem can elegantly be modeled by a graph. Graph neural networks (GNNs) [50] allow graph representations to be learned with neural networks, which were originally developed for structured data, but have recently been generalized to various computer vision tasks. In a graph, the nodes represent refined individuals' feature maps and the edges represent the interactions among individuals. Typically, graph-based models employ the information from neighbor points according to the characteristics of the specific task, *i.e.*, compute the hidden representations of each node in the graph, by attending over its neighbors. We adopt the recently proposed GATs [25] to directly learn the underlying interactions and seamlessly capture the global activity context. GATs flexibly allow learning attention weights between nodes through parameterized operations based on a self-attention strategy and have successfully demonstrated state-of-the-art results by outperforming existing counterparts.

We construct our graph attention layer (GAL) as following. The input of our GAL is $\mathbf{H} = \{h_1, ..., h_p, ..., h_P\}$, where P is the number of nodes. The layer produces a new set of node features $\mathbf{H}' = \{h'_1, ..., h'_p, ..., h'_P\}$ as its output. Initially, a shared linear transformation with the weight matrix W is applied to every node. We then perform a shared self-attention mechanism a to compute attention

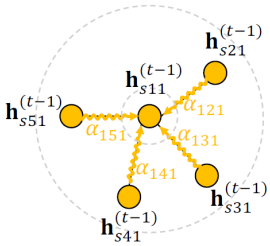

Fig. 4. An illustration of Graph Attention Networks by node 1 on its neighborhood. Different arrow styles and colors denote independent attention computations. The aggregated features from each head are concatenated or averaged to obtain h'_1.

coefficients

$$e_{ij} = a\left(Wh_i, Wh_j\right) \tag{3}$$

that indicate the importance of node j's features to node i. The model allows every node to attend on other nodes. To make coefficients easily comparable across different nodes, we normalize them across all choices of j using the softmax function:

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}} \exp(e_{ik})}. \tag{4}$$

In our experiments, the attention mechanism a is a single-layer feedforward neural network, parameterized by a weight vector $\mathbf{a}$, and applying the LeakyReLU nonlinearity (with negative input slope $\alpha = 0.2$). Fully expanded out, the coefficients computed by the attention mechanism (illustrated by Fig. 4) may then be expressed as:

$$\alpha_{ij} = \frac{\exp(\mathrm{LeakyReLU}(\mathbf{a}^T[Wh_i, Wh_j]))}{\sum_{k\in\mathcal{N}} \exp(\mathrm{LeakyReLU}(\mathbf{a}^T[Wh_i, Wh_k]))}. \tag{5}$$

Once obtained, the normalized attention coefficients are used to compute a linear combination of the features corresponding to them, to serve as the final output features for every node (after potentially applying a nonlinearity, σ):

$$h'_i = \sigma\left(\sum_{j\in\mathcal{N}} \alpha_{ij} W h_j\right). \tag{6}$$

To stabilize the learning process of self-attention, we have found extending our mechanism to employ multi-head

attention to be beneficial. Specifically, K independent attention mechanisms execute the transformation of Equation 4, and then their features are concatenated, resulting in the following output feature representation:

$$h_i' = \left[\sigma \left(\sum_{j \in \mathcal{N}} \alpha_{ij}^k W^k h_j \right) \right]_{k=1 \to K} . \quad (7)$$

α_{ij}^k is normalized attention coefficients computed by the k-th attention mechanism (a^k), and W^k is the corresponding input linear transformation's weight matrix.

Specially, if we perform multi-head attention on the final (prediction) layer of the network, concatenation is no longer sensible—instead, we employ averaging, and delay applying the final nonlinearity (usually a softmax or logistic sigmoid for classification problems) until then

$$h_i' = \sigma \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in \mathcal{N}} \alpha_{ij}^k W^k h_j \right) . \quad (8)$$

The aggregation process of a multi-head graph attentional layer is illustrated by Fig. 4.

GATs compute attention coefficients for every possible pair of nodes, which can be represented in an adjacency matrix $\hat{O}^{\alpha}$.

3.4 Engagement Estimator

After the model is trained, we can perform two types of group-based re-id tasks. Specifically, the graph-level representations from the readout attention module are directly employed for group re-id. For group-aware person re-id, the node-level features already contain discriminative context information as they receive messages from both intra-group and inter-group members. In addition, we can also utilize the person correspondence learning module to further reduce the ambiguity between people with similar appearances. The inference of group re-id and group-aware person re-id can be jointly computed in our framework. The results of group re-id and group-aware person re-id are discussed in Sec. 5, respectively.

In our framework, the GAT module consumes the individuals' feature maps obtained from the self-attention component, encodes inter-node relations, and generates an updated representation for each individual. We acquire the group representation by max-pooling the enriched individuals' feature map and adding back a linear projection of the holistic video's features obtained from the I3D backbone. A classifier is then applied on this representation to generate group activity scores denoted by $\hat{O}^G$. Similarly, another classifier is applied on the individuals' representation to govern the individual action scores denoted by $\hat{O}_n^I$. The associated operations provide a fully differentiable mapping from the input video clip to the output predictions, allowing the framework to be trained in an end-to-end fashion by minimizing the following objective function,

$$\mathcal{L} = \mathcal{L}_{gp}(O_G, \hat{O}^G) + \lambda \sum \mathcal{L}_{ind}(O_n^I, \hat{O}_n^I) \quad (9)$$

where $\mathcal{L}_{gp}$ and $\mathcal{L}_{ind}$ respectively denote the cross-entropy loss for group activity and individual action classification. Here, $\hat{O}^G$ and $\hat{O}^I$ n represent the ground-truth group activities and individual actions, n identifies the individual and λ is the balancing coefficient for the loss functions.

4 Datasets

It is still challenging for researchers to adequately investigate and explore the interaction between old people and a robot. The barriers can be divided into three parts. First, in order to understand how older people interact with robots, researchers must have a robot with basic build-in vision and language capabilities. Many small research groups cannot afford the cost. Second, researchers need to find sufficient samples (elderly people) participating in the experiments. Data collection is also very difficult due to safety and privacy concerns. Third, there is no publicly available dataset or benchmark in our research community, which makes the comparison and discussion difficult to achieve.

We have 21 interaction sessions in a span of 4 months with over 15 older people attending the experiments. The IMI-BHEH dataset has the following contributions

- IMI-BHEH dataset is collected in the free old-people-robot interaction scenario. There is no constraint imposed on the participants. The elderly talk to a humanoid robot while the nurses occasionally provide help.
- The participants of IMI-BHEH dataset are people with dementia. To the best of our knowledge, there is no publicly available dataset aiming to study their behaviors and interaction with robots.
- IMI-BHEH dataset is collected in a wild, dynamic, and multiparty environment. In the background, nurses may pass through the scene; some old people may sit near to or far from the interaction group; the interaction participants may leave and join at any time. All of these make the data more realistic.

Our dataset includes annotations as described in Table. 2. 2D people bounding box is automatically annotated by Yolo5[1]. 2D skeleton information and reconstructed 3D skeleton information are extracted using [5]. Facial landmarks and AUs are extracted from [6]. It is noteworthy that previously mentioned features are manually checked. The main interaction group participant ID and bounding box are annotated by professional volunteers. The label of the data is the engagement score, which is gained from a revised version of the engagement psychometric scale of people with dementia [7]. The details can be found in Appendix A.

5 Experiments

5.1 Implementation Details

In our model, we use an I3D backbone which is initialized with Kintetics-400 [32] pre-trained model. We utilize ROI-Align with crop size of 5×5 on extracted feature-map from Mixed-4f layer of I3D. We perform self-attention on each individuals' feature map with query, key and value being different linear projections of individuals' feature map with output sizes being $1/8$, $1/8$, 1 of the input size. We then

1. https://github.com/ultralytics/yolov5

TABLE 2
Details of IMI-BHEH Dataset

Raw Data	Extracted Features	Label
Videos	2D people bounding box 2D skeleton Reconstructed 3D skeleton Facial landmarks and AUs Main group participant ID Main group 2D bounding box	Engagement

learn a 1024-dim feature map for each individual features obtained from self-attention module. Aligned individuals' feature maps are fed into our single-layer, multi-head GAT module with 8 heads and input-dim, hidden-dim, output-dim being 1024 and dropout probability of 0.5 and $\alpha = 0.2$ [60]. We utilize ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ following [65]. We train the network in two stages. First, we train the network without the graph attention module. Then we fine-tune the network end-to-end including GAT. For the Volleyball dataset, we train the network in 200 epochs with a mini-batch size of 3 and a learning rate ranging from 10^{-4} to 10^{-6} and $\lambda_1 = 8$. For the CAD, we train the network in 150 epochs with a mini-batch size of 4 and a learning rate ranging from 10^{-5} to 10^{-6} and $\lambda_1 = 10$. Input video clips to the model are 17 frames long, with the annotated key frame in the center. At test time, we perform our experiments based on two widely used settings in group activity recognition literature namely ground truth-based and Detection-based settings. In the ground truth-based setting, ground-truth bounding boxes of individuals are given to the model to infer the individual action for each box and the group activity for the whole scene. In the detection based setting, we fine-tune a Faster-RCNN [46] on both datasets and utilize the predicted boxes for inferring the individuals' action and group activity.

6 Conclusions

Human behavior and social signal understanding have achieved good performance in the computer vision area. However, most of them are based on the analysis of an individual subject and are achieved in experimental settings. In this paper, we use real-world and open-world image data to estimate the engagement of old people in a multi-party human-robot interaction scenario. For this purpose, we place a humanoid robot in a non-profit nursing home, where the elderly are accompanied by nurses and engage in spontaneous conversations with the robot. We perform the analysis and feature extraction on three scales, namely global, group, and individual, which gradually reduce via people detection, main group detection, and the old people localization. We propose a multi-branch network with attention and graphic convolutional layers combining three-level information. Our experiments show that the proposed framework achieves good performance in the engagement estimation task.

Appendix

We used a revised version of EPWDS for artificial engagement annotation. The form for annotators are appended in Fig. 5.

References

[1] "Ageing and health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, World Health Organization, oct 2021.

[2] M. Ghafurian, J. Hoey, and K. Dautenhahn, "Social robots for the care of persons with dementia: A systematic review," *ACM Transactions on Human-Robot Interaction*, 2021.

[3] G. Perugia, M. Díaz-Boladeras, A. Català-Mallofré, E. I. Barakova, and M. Rauterberg, "ENGAGE-DEM: A model of engagement of people with dementia," *IEEE Transactions on Affective Computing*, 2020.

[4] I. Poggi, "Mind, hands, face, and body: A sketch of a goal and belief view of multimodal communication," in *Body - Language - Communication*, 2013.

[5] C. Jones, B. Sung, and W. Moyle, "Engagement of a person with dementia scale: Establishing content validity and psychometric properties," *Journal of Advanced Nursing*, 2018.

[6] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, 2017.

[7] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, 2019.

[8] A. Ben Youssef, C. Clavel, and S. Essid, "Early detection of user engagement breakdown in spontaneous human-humanoid interaction," *IEEE Transactions on Affective Computing*, 2019.

[9] K. Saleh, K. Yu, and F. Chen, "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.

[10] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," *Frontiers in Robotics and AI*, 2020.

[11] A. Ben-Eliyahu, D. Moore, R. Dorph, and C. D. Schunn, "Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts," *Contemporary Educational Psychology*, 2018.

[12] B. Zhu, X. Lan, X. Guo, K. E. Barner, and C. Boncelet, "Multi-rate attention based gru model for engagement prediction," in *Proceedings of the International Conference on Multimodal Interaction*, 2020.

[13] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "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.

[14] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *arXiv:2101.04215*, 2021.

[15] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, 2017.

[16] A. Abedi and S. Khan, "Affect-driven engagement measurement from videos," *arXiv:2106.10882*, 2021.

[17] N. Gao, W. Shao, M. S. Rahaman, and F. D. Salim, "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.

[18] D. Anagnostopoulou, N. Efthymiou, C. Papailiou, and P. Maragos, "Engagement estimation during child robot interaction using deep convolutional networks focusing on asd children," in *IEEE International Conference on Robotics and Automation*, 2021.

[19] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders," *Science Robotics*, 2020.

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									

Fig. 5. Engagement Estimation Annotation Form

[20] L. Steinert, F. Putze, D. Küster, and T. Schultz, "Towards engagement recognition of people with dementia in care settings," in *Proceedings of the International Conference on Multimodal Interaction*, 2020.

[21] G. Guo, R. Guo, and X. Li, "Facial expression recognition influenced by human aging," *IEEE Transactions on Affective Computing*, 2013.

[22] M. Fölster, U. Hess, and K. Werheid, "Facial age affects emotional expression decoding," *Frontiers in Psychology*, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proceedings of the International Conference on Learning Representations*, 2017.

[26] P. Guhan, M. Agarwal, N. Awasthi, G. Reeves, D. Manocha, and A. Bera, "Abc-net: Semi-supervised multimodal GAN-based engagement detection using an affective, behavioral and cognitive model," *arXiv:2011.08690*, 2020.

[27] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, 2020.

[28] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, 2005.

[29] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *Proceedings of the International Conference on Multimodal Interfaces*, 2009.

[30] S. Christenson, A. L. Reschly, C. Wylie *et al.*, *Handbook of research on student engagement*. Springer, 2012.

[31] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, 2008.

[32] J. Cohen-Mansfield, M. S. Marx, L. S. Freedman, H. Murad, N. G. Regier, K. Thein, and M. Dakheel-Ali, "The comprehensive process model of engagement," *The American Journal of Geriatric Psychiatry*, 2011.

[33] I. Archambault and V. Dupéré, "Joint trajectories of behavioral, affective, and cognitive engagement in elementary school," *The Journal of Educational Research*, 2017.
[34] L. J. Corrigan, C. Peters, D. Küster, and G. Castellano, "Engagement perception and generation for social robots and virtual agents," in *Toward Robotic Socially Believable Behaving Systems*, 2016.
[35] C. Oertel, P. Jonell, D. Kontogiorgos, K. F. Mora, J.-M. Odobez, and J. Gustafson, "Towards an engagement-aware attentive artificial listener for multi-party interactions," *Frontiers in Robotics and AI*, 2021.
[36] J. Cohen-Mansfield, M. Dakheel-Ali, and M. S. Marx, "Engagement in persons with dementia: The concept and its measurement," *The American Journal of Geriatric Psychiatry*, 2009.
[37] M. A. Trahan, J. Kuo, M. C. Carlson, and L. N. Gitlin, "A systematic review of strategies to foster activity engagement in persons with dementia," *Health Education & Behavior*, 2014.
[38] W. Moyle, C. J. Jones, J. E. Murfield, L. Thalib, E. R. A. Beattie, D. K. H. Shum, S. T. O'Dwyer, M. C. Mervin, and B. M. Draper, "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.
[39] Y. Feng, G. Perugia, S. Yu, E. I. Barakova, J. Hu, and G. W. M. Rauterberg, "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.
[40] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2018.
[41] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, 2015.
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
[43] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *arXiv:1705.06950*, 2017.
[44] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," *arXiv:2110.06864*, 2021.
[45] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
[48] E. Goffman, *Forms of talk*. University of Pennsylvania Press, 1981.
[49] H. H. Clark, *Using language*. Cambridge university press, 1996.
[50] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, 2008.

PLACE PHOTO HERE

Zhijie Zhang Biography text here.

PLACE PHOTO HERE

Jianmin Zheng is a professor in the School of Computer Science and Engineering at Nanyang Technological University, Singapore. He received the BS and Ph.D. degrees from Zhejiang University, China. His recent research focuses on T-spline technologies, digital geometric processing, reality computing, AR/VR, and AI assisted part design for 3D printing. He is currently the program director for the research pillar of ML/AI under the HP-NTU Digital Manufacturing Corporate Lab. He is also a member of executive committee of AsiaGraphics Association.

PLACE PHOTO HERE

Nadia Magnenat Thalmann (Member, IEEE) received her bachelor's and master's degrees in psychology, biology, chemistry, and computer science, and the Ph.D. degree in quantum physics from the University of Geneva. She is currently the Director of the Virtual Humans and Social Robotics Research Laboratory, MIRALab, University of Geneva. From 2009 to 2021, she was the Director of the Institute for Media Innovation, NTU, Singapore. In NTU, she revolutionized social robotics by unveiling the first social robot Nadine that can have mood and emotions and remember people and actions. She is a Life Member of the Swiss Academy of Engineering Sciences. She received honorary doctorates from Leibniz University of Hannover and the University of Ottawa, Canada, and several prestigious other awards as the Humboldt Research Award in Germany. She is the Editor-in-Chief of The Visual Computer and the Co-Editor-in-Chief of Computer Animation and Virtual Worlds.