Joint Engagement Classification using Video Augmentation Techniques for Multi-person HRI in the wild

Yubin Kim
MIT Media Lab
Cambridge, Massachusetts
ybkim95@media.mit.edu

Huili Chen
MIT Media Lab
Cambridge, Massachusetts
hchen25@media.mit.edu

Sharifa Algohwinem
MIT Media Lab
Cambridge, Massachusetts
sharifah@media.mit.edu

Cynthia Breazeal
MIT Media Lab
Cambridge, Massachusetts
cynthiab@media.mit.edu

Hae Won Park
MIT Media Lab
Cambridge, Massachusetts
haewon@media.mit.edu

ABSTRACT

Affect understanding capability is essential for social robots to autonomously interact with a group of users in an intuitive and reciprocal way. However, the challenge of multi-person affect understanding comes from not only the accurate perception of each user's affective state (e.g., engagement) but also the recognition of the affect interplay between the members (e.g., joint engagement) that presents as complex, but subtle, nonverbal exchanges between them. Here, we present a novel hybrid framework for identifying a parent-child dyad's joint engagement by combining a deep learning framework with various video augmentation techniques. Using a dataset of parent-child dyads reading storybooks together with a social robot at home, we first train RGB frame- and skeleton-based joint engagement recognition models with four video augmentation techniques (General Aug, DeepFake, CutOut, and Mixed) applied datasets to improve joint engagement classification performance. Second, we demonstrate experimental results on the use of trained models in the robot-parent-child interaction context. Third, we introduce a behavior-based metric for evaluating the learned representation of the models to investigate the model interpretability when recognizing joint engagement. This work serves as the first step toward fully unlocking the potential of end-to-end video understanding models pre-trained on large public datasets and augmented with data augmentation and visualization techniques for affect recognition in the multi-person human-robot interaction in the wild. Our code and detailed experimental results are available at https://github.com/ybkim95/multi_person_joint_engagement

KEYWORDS

nonverbal communication; multi-person affect understanding; joint engagement recognition; multi human-robot interaction

ACM Reference Format:
Yubin Kim, Huili Chen, Sharifa Algohwinem, Cynthia Breazeal, and Hae Won Park. 2023. Joint Engagement Classification using Video Augmentation Techniques for Multi-person HRI in the wild. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 10 pages.

1 INTRODUCTION

Affective communication is essential for human-human interaction [42] and has strong links to learning [43], persuasion [25], and a variety of other functions. The ability of a socially interactive robot to perceive human nonverbal cues, social signals, and emotions is critical for engaging with humans in an intuitive, natural, and reciprocal manner [7, 51]. This affect understanding capacity has been identified as a basic robot capability required for higher-level competencies in human-robot interactions [56] and contributes to a robot's user profiling and behavior adaptability capabilities [45]. Consequently, the objective of our research is not only to improve the affective recognition and understanding capabilities of social robots in multi-person human-robot interaction (HRI) scenarios but also to assess the quality of the model for deployment.

Given the impact of parent-child affective exchanges in children's development and growth, parent-child interaction will be the primary application domain for our research. High-quality, reciprocal relationships between parents and children promote children's social, emotional, cognitive, and linguistic development [54]. However, not all children have equal access to socially and emotionally rich dialogic interactions, such as being prompted by open-ended questions and back-and-forth conversations [46]. Social robots have a compelling potential to facilitate human-human connection and conversation [29], and they can be designed to mediate parent-child dialogic interactions by encouraging and demonstrating best pedagogical practices. In order to reach such capacity, **affect understanding in a multi-person interaction context** should support robots to comprehend the interpersonal dynamics between a parent and a child and provide timely and appropriate actions to mediate the flow of the interaction and maintain the engagement of the dyad.

The nonverbal cues people display in a group interaction present excellent data sources for the development of contact-free, unobtrusive affect recognition systems that can comprehend human-human affective dynamics. In parent-child interactions, the parent and child's nonverbal behaviors, such as head/body movement, gestures, and postures, are particularly important indicators to gauge interaction quality including synchronization, engagement, attachment, and shared affect [12, 23, 33]. However, automatic perception of these affective dynamics between the members of a group via nonverbal indicators is still much underexplored.

In an effort to further this field of study, we propose a novel hybrid method for identifying parent-child joint engagement by combining

a deep learning framework with a range of video augmentation techniques – General Aug [40], DeepFake [10], CutOut [15], and Mixed [67]). Using the state-of-the-art action recognition algorithms such as SlowFast [20] as base models, we applied video augmentation techniques to improve the pre-trained model's representation learning in order to raise its sensitivity to capture the subtle social cues that leads to understanding the joint engagement states of parent-child dyads. Even though SlowFast and other base models have been largely used to detect human activity and motion in prior works [32, 62, 70], we show that they also produce high performance in social cue understanding tasks by fine-tuning with video augmentation techniques.

Using our novel evaluation metric, we also demonstrate the interpretability of the model's learned representation. The model interpretability is particularly crucial for humans to understand how the model learns the continuous dynamics of human social cues. In order to visualize where our model attends to in the video frames to identify joint engagement, we compare Gradient-weighted Class Activation Mapping (Grad-CAM) results to the social cues humans attend to gauge joint engagement, i.e., annotation guidelines given to human annotators.

In summary, our work contributes the following :

(1) Adapt end-to-end and skeleton-based deep learning models to joint engagement recognition task in multi-person HRI setting. The models are pre-trained on large public datasets of human action and activity recognition task and fine-tuned with video augmentation techniques for recognition of psychological processes, i.e., joint engagement, that involves changes in subtle social cues;
(2) Conduct a comprehensive analysis with experimental results to demonstrate the use of pre-trained models in an affective communication context, and provide a competitive affect recognition baseline for future multi-person affect understanding;
(3) Implement a new metric for evaluating the learned representation that compares human annotation guidelines and the visual regions the model attends to.

2 RELATED WORKS

2.1 Social-Affective cue understanding in Human-Robot Interaction (HRI)

Developing affect signal perception systems for social robots has been extensively investigated [4, 37, 66] and shown to have empirical significance, particularly in early childhood development. Social robots with this affect understanding capability have been found to promote children's learning more effectively than those without, e.g., [22, 41]. Prior works have integrated user social and affective signals into either its behavior policy or cognitive model as human feedback on the robot's newly executed action to deliver real-time personalized interaction [22, 41]. Similarly, affective signals have been incorporated into the robot's user cognitive and skill estimation models to improve user model accuracy, such as a student vocabulary acquisition [52]. Overall, affect recognition enhances the effectiveness of robots to provide timely interventions to individual users and improve their interaction experience. Nonetheless, the majority of affect-aware social robots were only designed for one-on-one interactions with humans. When interacting with a single person, it is sufficient for an intelligent system to recognize the social and affective cues directed at the interaction task or the robot. In contrast, in dyadic human interaction, the technology must be able to recognize the affective and social dynamics between the two users.

To date, the vast majority of current affect recognition models, particularly commercial affect extraction tools, are only applicable in single-person settings. Only a handful of previous Multi-person HRI field studies, e.g., [53, 59], investigated how to equip a robot with a perception system to recognize the social-affective dynamics of a human group. The perception system in [59], for instance, estimates user positions and body orientations using a Kinect to track participants and control the orientation and gaze of the robot. In a separate study [53], a prediction model for a participant's social dominance in a group human-robot interaction was developed but trained on the nonverbal behavioral features that were manually handcrafted by human coders offline, e.g., utterance type, gaze, interruptions. One previous study [50] focused on the automatic analysis and classification of engagement based on humans' and robots' personality profiles using a dataset gathered in the context of triadic human-human-robot interaction.

Due to this limitation in the robot's perception system, the majority of current multi-party HRI research employs wizard-of-oz paradigms to teleoperate a robot or an oversimplified behavior policy (e.g., simple rule-based or tablet-based behavior triggers) that does not depend on the robot's affective perceptual capacity. In a minority of studies, machine learning or reinforcement learning was used to train the robot's behavior policy [30, 60]. Even fewer studies have equipped robots with affective perception that can guide their interactive behavior. Research on developing affect recognition models for dyadic human groups would unlock the potential for a robot to engage in and even enhance inherently complex human-human interactions. Therefore, the advancement of multi-person affect recognition would catalyze the development of fully autonomous social robots in multi-person HRI settings.

2.2 Deep Learning Approach to Affect Recognition

The majority of affect recognition models, and commercial affect extraction tools, in particular, are primarily concerned with single-person or single-modality affect detection. In recent years, deep learning has been used extensively to develop affect detection models trained on human behavioral cues in audiovisual recordings, such as facial expression, speaking style, speech prosody, linguistics sentiment, and head and body movement (see the review [21]). Deep belief networks (DBNs) [69], attentively-coupled long-short term memory (ACLSTM) [28], and multitask LSTM augmented with two-stream auto-encoder for deep feature extraction [26] are examples of deep learning techniques used in previous research. The audio-video input features both handcrafted and deep features have been used as model input for recognition (e.g., [26]), while the predicted affective states range from valence and arousal to engagement (e.g., [47]).

In contrast to the widespread application of deep learning to the detection of individual affect, affect detection in multi-person interactions is significantly less studied. Using deep learning models, a small number of studies have investigated dyadic dynamics, primarily from a single modality, e.g., [8, 27]. A number of multi-person interaction

perception models focus on human action recognition such as handshakes, e.g.[65]. For instance, deep features from the full body and body parts of both individuals, as well as handcrafted motion and posture features, were extracted to train deep learning models for action recognition models, such as [65]. Using a combination of deep features, a graph network, and a logic-aware module, the relationship and interaction between two individuals in still images were analyzed in [63]. Personality recognition in the dyadic interaction context was also modeled using the transformer-based method that utilizes multimodal deep features extracted and individuals' sociodemographic profiles (e.g. gender, relationship status, mood) [39]. In a previous study, Zhang and Radke [68] developed a temporal fusion of multimodal features extracted from vision, audio and text to recongize each participant's social role in a four-person meeting.

To the best of our knowledge, only a handful of recent works have begun to develop deep models for recognizing affect in multi-person interactions (e.g., [9, 34]). For instance, Chen and colleagues [9] recently utilized end-to-end deep learning methods augmented with attention mechanisms to identify each individual's affective expression in an audio stream containing the utterances of two speakers. Another work by [27] developed a framework to identify an individual's engagement in the context of a two-way conversation. In their study, a hybrid approach of deep models and Bayesian networks was used to predict interpersonal dynamics in the dyadic interaction, including back channeling, speaking turns, gender, and face, hand movement, speech, and context data. The audiovisual recordings were captured separately for each interlocutor.

Overall, very little research has been conducted on the application of a deep learning approach to multi-person affect recognition in which all the interlocutors interact with one another in a single audiovisual recording, limiting the development of affect-aware robots suited for dyadic human interactions in the real world. In addition, advanced techniques in deep learning research, such as video augmentation techniques and explainable visualization in deep learning, have been applied to advance multi-person affect perception models for social robots. Consequently, our work proposes a novel framework that employs these cutting-edge deep learning techniques to significantly enhance a robot's ability to comprehend the affective dynamics of multi-person HRI in the wild.

2.3 Video Classification Models and Data Augmentation Techniques

State-of-the-art video classification models were mostly developed for action recognition, a central task in video understanding [19, 20, 62]. Among various modalities (e.g., RGB frames, optical flow, human skeleton, and audio waves) used for feature representation, RGB-based and Skeleton-based action recognition models have been the mainstream approach in recent years [3, 19, 20, 62]. RGB frames are the most basic and typical modality to be used for model training in action recognition tasks. The recently proposed TimeSformer [3] network is only built on self-attention over space and time. It adapts the Transformer architecture to video by enabling spatiotemporal feature learning directly from a sequence of frame-level patches. X3D [19] is a family of efficient video networks that continuously expand a small 2D image classification architecture along multiple network axes (space, time, width, and depth). I3D [62] is a 2D ConvNet inflation-based model, in which the filters and pooling kernels of deep image classification ConvNets are expanded into 3D. In SlowFast [20], a dual-pathway structure is proposed to combine the benefits of a slow pathway for static spatial features and a fast pathway for dynamic motion features.

On the other hand, human skeletons in a video provide a sequence of joint coordinate lists which emphasizes its action-focusing nature and compactness. Recently proposed Channel-wise Topology Refinement Graph Convolution Network (CTR-GCN) [11] dynamically learns different topologies and effectively aggregates joint features in different channels. The multi-Scale aggregation Scheme (MS-G3D) [36] disentangles the importance of nodes in different neighborhoods for effective long-range modeling. It utilizes dense cross-spacetime edges as skip connections for direct information propagation across the spatial-temporal graph. STGCN [17] adopts Graph Convolution Neural (GCN) Networks for skeleton processing. Based on all the above works, Transformer, CNN, and Graph Convolutional models have achieved breakthrough results in action recognition tasks. With the great power of understanding general human activities, pre-trained action recognition models make it suitable for the models to solve the task of joint engagement recognition with the fine-tuning process.

3 METHODS

In this section, we introduce the robot-parent-child interaction dataset our models were trained on, and four different video augmentation methods. The behavior-based objective metric to evaluate the learned representation is also presented.

3.1 Dataset and Annotation

The dataset we used to train our affect model originated from our previous deployment [6]. In the study, a social robot was deployed and teleoperated remotely in the homes of 12 families with 3-7-year-old children to engage in a triadic story-reading activity with the parent and child over the course of six 25-minute sessions and 3 to 6 weeks in total. In particular, the robot had roles of being 1) a moderator and mediator such as by asking story-related questions or vocabulary and 2) a listener by backchanneling and encouraging. For each triadic session, audiovisual recordings were captured and subsequently used to annotate the quality of parent-child engagement. We chose the Joint Engagement Rating Inventory (JERI) to measure parent-child engagement [1], as it has been utilized and validated in previous parent-child interaction studies [1, 5]. This engagement metric quantifies and classifies both the verbal and nonverbal behaviors associated with a child's interaction with its parent.

To annotate the parent-child joint engagement in the audio-visual recordings of the parent-child-robot interactions, we recruited two trained annotators with a psychology or education background. The coding scheme, the choice of the video interval threshold, and the annotation protocol were derived from previous work on joint engagement in parent-child interaction [5]. Specifically, the annotators gave ratings every five non-overlapping seconds on a five-point ordinal scale [-2,2], with two corresponding to cases in which the parent-child pair displayed clear signs of high joint engagement valence and -2 corresponding to cases in which the parent-child pair displayed clear signs of low joint engagement valence. The video

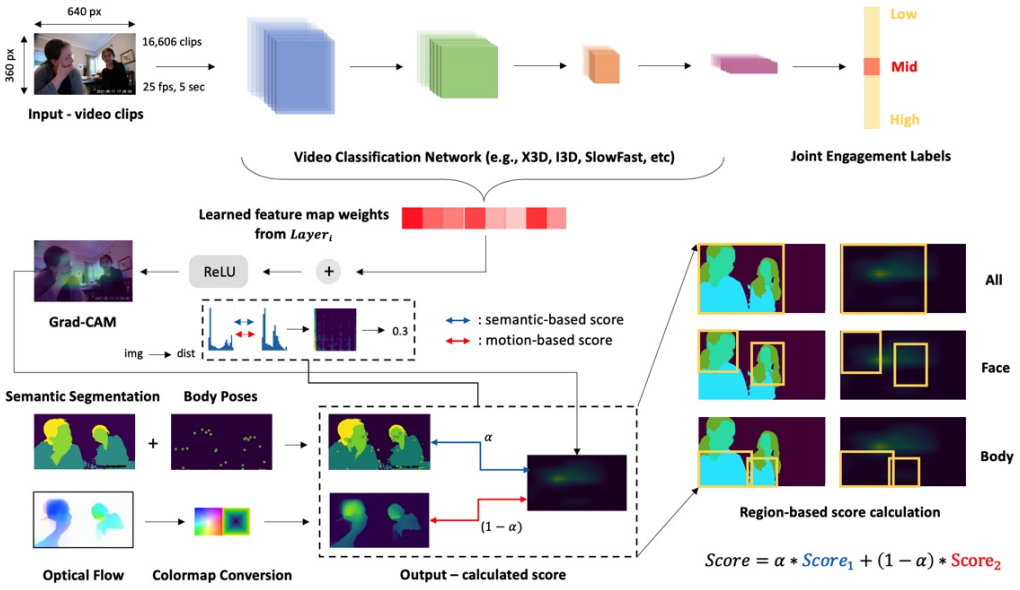


Figure 1: Proposed framework for evaluating the learned representation (Grad-CAM) with modified Optical Flow and skeleton information combined Semantic Segmentation as references. With the fine-tuned models, we generate Grad-CAM for each video clip and evaluate its quality. We calculate the evaluation score based on two sub-scores (semantic-based and motion-based) which are obtained by applying mutual information and cross-entropy.

fragment interval of target audio-visual recordings used to generate continuous quality scales was determined to be five seconds.

Using the intra-class correlation (ICC) type (3,1) for average fixed raters, the agreement among the three annotators was measured. Given these evaluation criteria, the annotation quality with ICC= 0.95 exceeded the threshold for very good quality ($0.75 \leq ICC \leq 1.0$). After recordings were independently coded by the annotators, the final score for each recording fragment was determined by averaging the ratings assigned to each scale by the two annotators. We have 3-scale levels (low: 8.49%, medium: 49.68%, and high: 41.83%) for model training and testing. Strictly following the annotation protocol in [5], we annotated 16,606 five-second video clips with 1517.08 ± 309.34 fragments from each family on average.

3.2 Model Training and Evaluation

In this section, we explain the details of model training and evaluation. We split the video clips and pose datasets (N=24,749) into the train, valid, and test by assigning 8, 2, and 2 different families respectively and conduct Leave-One-Family-Out Cross-Validation. For all models, we measure Top-k (k=1) accuracy with cross-entropy loss. More details about RGB frame- and skeleton-based action recognition models are described in Table 1 and the following sections.

3.2.1 Action Recognition Models. For RGB frame-based models, we use a base learning rate of 0.1 and it is step-wisely decayed for every 20 epochs with a total of 50 epochs. During fine-tuning, we freeze the pre-trained weight by specifying the number of layers in each backbone. The most of hyperparameters are kept the same as the default configuration provided by MMAction2 [13].

3.2.2 Skeleton-based Action Recognition Models. For skeleton-based models, we use PoseDataset as input which is extracted from NTU pose extraction [18]. This dataset has the format of keypoint, keypoint_score, frame_dir, label, img_shape, original_shape, and total_frames. We set the initial learning rate is 0.1, the batch size to 128, and train models for 15 epochs with the CosineAnnealing learning rate scheduler. For the optimizer, we set the momentum to 0.9, weight decay to 5×10^{-4}, and use the Nesterov momentum. The rest of the hyperparameters are kept the same as the default configuration provided by Pyskl [17].

3.3 Video Augmentation techniques

Our dataset has imbalanced label distribution (see Section 3.1) and this poses the classification task very challenging [55], which may be compounded by sample size, label noise, etc. The imbalanced label distribution motivates applying video augmentation techniques. The details about each video augmentation technique will be described in the following subsections.

3.3.1 Baseline. In Section 3.1, we briefly introduce our dataset's imbalanced labels and the total number of video clips. To ensure a fair comparison with the proposed video augmentations techniques, we apply oversampling to the original dataset which duplicated 8,143 video clips from low and high joint engagement labeled video clips to make all the labels have the same ratio. In total, we prepared 24,749 video clips for Baseline and this contained 8,249 video clips per label.

3.3.2 General Augmentations. To prevent the model from overfitting by "fixating" on irrelevant patterns (e.g. backgrounds), we have applied various kinds of augmentation techniques which is why

Table 1: Comparisons of detailed model configuration with state-of-the-art models.

Models	Year	Inputs	Data Modality	# of Params	Backbone	Epochs
TimeSformer	2021	3×32×224×224	RGB frame	121.4M	TimeSformer	15
X3D	2020	3×16×224×224	RGB frame	3.76M	X3D_M	50
I3D	2017	3×32×224×224	RGB frame	28.0M	ResNet50	50
SlowFast	2019	3× 32×224×224	RGB frame	34.6M	ResNet50-4×16	50
CTR-GCN	2021	16× 2×100×17×3	PoseDataset	1.43M	CTRGCN	15
MS-G3D	2020	16× 2×100×17×3	PoseDataset	3.17M	MSG3D	15
ST-GCN++	2022	16× 2×100×17×3	PoseDataset	3.08M	STGCN	15
ST-GCN	2018	16× 2×100×17×3	PoseDataset	1.39M	STGCN	15

we call them General Augmentations. This technique is applied to diversify the background (replace the background with RGB color, random indoor image, and blur the background), encouraging the model's robust learning by adding noise to the whole frame, randomly rotating an image, applying horizontal flipping, and lastly, giving hints of semantics in the frame by applying semantic segmentations. These different types of simple but effective techniques enabled us to supplement more features in the dataset and in total, we gathered 24,749 video clips that have all labels with the same ratio.

3.3.3 DeepFake. DeepFake was applied for dyads' faces to overcome the small populations in the original dataset and also for debiasing purposes. We used SimSwap [10] for multi-person face swapping in videos. To feed a diverse set of target face images, we also utilized AI-generated face dataset (https://generated.photos/faces) which supports realistic customizations (e.g., race, gender, age, accessories, and hair type). As we can see in Table. 2, this generates quite natural video clips according to its target face images. In total, we gathered 24,749 video clips as same in the General Augmentation case.

3.3.4 Mixed. Finally, we also wanted to see if combining the datasets that showed performance improvement individually (see Table. 2) would make even more performance improvements once combined. To do this, we randomly sampled video clips from both General Aug and DeepFake while keeping the same ratio from each dataset. So in total, we kept 24,749 video clips for Mixed.

3.3.5 CutOut. CutOut is a well-known but simple regularization technique that randomly masks out square regions of input during training (spatial prior dropout in input space) [15]. This can be used to improve the robustness and overall performance when conducting classification tasks, and in this work, CutOut is used to validate the model's representation learning without the core information in the scenes (i.e. face). To apply CutOut, we utilized the face detection module to detect the parent's and child's faces and cut out the corresponding regions, which are then replaced by black boxes. In total, we gathered 24,749 video clips by oversampling towards the largest number of labels (Mid, N=8,249) in the dataset.

3.4 Evaluation Metric

3.4.1 Gradient-weighted Class Activation Mapping. Grad-CAM is a generalization of Class Activation Mapping (CAM) which combines the class-specific property of CAM [61]. This supports the intuitive visualization of the model's attention in an image and this technique has been used in various HRI works [16, 24, 31]. Apart from Grad-CAM's effective visualization capability, our purpose is to validate the quality of learned representations. Following the definitions in [61], Class Activation Map (CAM) and Grad-CAM are defined as follows.

Definition 1. CAM. Consider a model f with a global pooling layer l after the last convolution layer $l-1$ and before the last fully connected layer $l+1$. Given a class c of interest, the CAM is defined as:

$$L^c_{CAM} = ReLU(\sum_k \alpha^c_k A^k_{l-1}) \quad (1)$$

where

$$\alpha^c_k = w^c_{l,l+1}[k] \quad (2)$$

Definition 2. Grad-CAM. Consider a convolution layer l in a model and given a class c selected by a model, Grad-CAM is defined as:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k_l) \quad (3)$$

where

$$\alpha^c_k = GP(\frac{\partial Y^c}{\partial A^k_l}) \quad (4)$$

here, $GP(\cdot)$ denotes the Global Pooling operation.

To evaluate the quality of learned representation (Grad-CAM) which is in video format, we break this down into image frames and apply two different image-matching techniques with two references which will be explained in the following sections.

3.4.2 Evaluation References. In the previous section, the concept of Grad-CAM is defined and we particularly use this as a distribution learner for evaluating with two references. Here, we convert the images into distributions and calculate the mutual information and cross-entropy. The two references are 1) modified Optical Flow and 2) skeleton information combined with Semantic Segmentation. Having the two scores from each reference, we calculate the weighted average score with the weight α which has a value between 0 and 1 (See Fig. 2).

The primary hypothesis supporting our framework is based on the annotation coding scheme for joint engagement (see Section 3.1); subtle display (both duration and intensity) of social cues is crucial when evaluating human joint engagement (e.g., shared gaze, contingent smiling, finger pointing, etc).

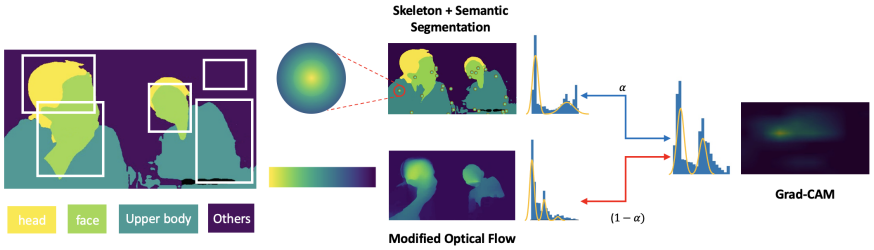
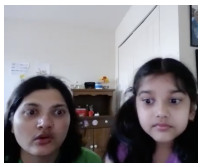
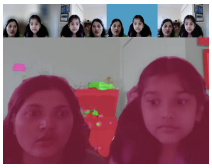
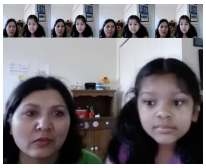
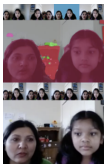
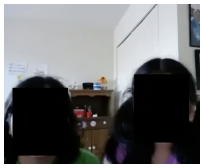


Figure 2: Pipeline for image matching metric which converts images into distributions and calculates mutual information and cross-entropy (between skeleton information combined with semantic segmentation, modified optical flow) and Grad-CAM.

Table 2: Overview of joint engagement recognition task results from state-of-the-art end-to-end (top) and skeleton-based (bottom) action recognition models. General Aug and DeepFake significantly outperformed the performance in all end-to-end models.

	Baseline	General Aug	DeepFake	Mixed	CutOut
Input					
TimeSformer Cls (%)	62.2 (+0.0)	66.7 (+4.5)	**70.1 (+7.9)**	63.0 (+0.8)	60.9 (-1.3)
X3D Cls (%)	62.6 (+0.0)	65.1 (+2.5)	**67.8 (+5.2)**	63.5 (+0.9)	61.0 (-1.6)
I3D Cls (%)	60.1 (+0.0)	63.4 (+3.3)	**66.9 (+6.8)**	60.7 (+0.6)	55.8 (-4.3)
SlowFast Cls (%)	61.2 (+0.0)	**63.8 (+2.6)**	62.2 (+1.0)	58.4 (-2.8)	50.0 (-10.2)
Input					
CTR-GCN Cls (%)			**64.3**		
MS-G3D Cls (%)			63.3		
ST-GCN++ Cls (%)			64.1		
ST-GCN Cls (%)			59.4		

Accordingly, to recognize the subtle motion changes, we apply Optical Flow from [14]. After that, we modify the original Optical Flow which displays the color by their orientations, but instead, we discard this orientation-based colormap and follow the colormap used in Grad-CAM to focus on the motion changes itself (See Fig. 1). Also, to ensure the learned representation is focusing on the proper regions, we considered skeleton information combined Semantic Segmentation as the other reference. We first extract the skeleton information by applying the pose extractor described in [18] and combine this with the Semantic Segmentation results [35]. For Semantic Segmentation, we specify each body segment with pre-defined colors (see Fig. 2), and when combining, the Gaussian heatmap is centered on each of the body poses (see Fig. 2). This is based on the insights from our annotation process where we evaluate the dyad's joint engagement level by focusing on social touch, body closeness, heading angle, and smiling. Since our dataset mostly contains the face and upper parts of the body, we put more importance on the head and the upper body parts rather than the bottom parts of the body.

3.4.3 Image Matching Techniques. First, we adopt mutual information, a dimensionless quantity metric that measures the mutual dependence between two variables. The metric is high when the attention map signal is highly concentrated in a few histogram bins, and low when the signal is spread across many bins. Mutual information is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log(\frac{p(x,y)}{p(x)p(y)}) \tag{5}$$

Here, we convert the image into a distribution by flattening the image arrays and then compute the 2D histogram of two image array samples. The second metric is cross-entropy, which comes from the Kullback-Leibler divergence. This is a widely used metric for calculating the difference between two distributions, and this is defined as:

$$H(p,q) = -E_p[log\, q] \tag{6}$$

where $E_p[\cdot]$ is the expected value operator with respect to the distribution p. Here, we first normalize the pixel values in images

Algorithm 1 Grad-CAM evaluation algorithm

Input: Images I_0, I_1, I_2, Model $F(I)$, MI(H), CE(I;I), CROP($\cdot$)
Output: Score

Score $\leftarrow$ []
while I_0 and I_1 and I_2 **do**
  BoundingBoxes $\leftarrow F(I_0)$
  score $\leftarrow dict()$
  for role, part, bbox in BoundingBoxes **do**
    $I_{new0}, I_{new1}, I_{new2} \leftarrow CROP(I_0,I_1,I_2,bbox)$
    $mi_1 \leftarrow MI(hist_{2d}(I_{new0}, I_{new1}))$
    $mi_2 \leftarrow MI(hist_{2d}(I_{new0}, I_{new2}))$
    $mi \leftarrow (\alpha \cdot mi_1 + (1-\alpha) \cdot mi_2)$ ▷ mutual information

    $I_{p0}, I_{p1}, I_{p2} \leftarrow Log-Softmax(I_0,I_1,I_2)$
    $I_{new0}, I_{new1}, I_{new2} \leftarrow CROP(I_{p0},I_{p1},I_{p2},bbox)$
    $ce_1 \leftarrow CE(log-softmax(I_{new0},I_{new1}))$
    $ce_2 \leftarrow CE(log-softmax(I_{new0},I_{new2}))$
    $ce \leftarrow (\alpha \cdot ce_1 + (1-\alpha) \cdot ce_2)$ ▷ cross-entropy

    $score[mi][role][part] \leftarrow mi$
    $score[ce][role][part] \leftarrow ce$
  $Score.Add(score)$
return $Score$

and then pass this through log-softmax to convert images into distributions. Then we apply cross-entropy as explained in Equation 6. Given that we have all the bounding boxes for each parent's and child's face and body, we could separate these values based on their bounding box coordinates (See Fig. 1). The overall process is summarized in Algorithm 1. where the model $F(I)$ is the face, age, and gender detector, $MI(H)$ is mutual information function, $CE(I;I)$ is a cross-entropy function, and $CROP(\cdot)$ is a image crop function.

4 EXPERIMENT AND RESULTS

In this section, we conduct experiments to evaluate the effectiveness of the proposed video augmentation techniques; General Aug, DeepFake, Mixed, and CutOut to see their capability to improve joint engagement recognition on state-of-the-art action recognition models. Also, we compare the joint engagement classification performance between RGB frame-based and skeleton-based models to see the effect of different inputs in this task. For the implementation, we utilized MMAction2 and Pyskl, an open-source toolbox for video understanding based on PyTorch and all of the models were trained on 8 NVIDIA 1080Ti GPUs

4.1 Joint Engagement Recognition Evaluation

As shown in Table. 2, applying General Aug and DeepFake consistently outperformed the baselines in all end-to-end models. Particularly, TimeSformer and I3D lead the accuracy by up to **7.9%** and **6.8%** respectively. On the other hand, using CutOut did not improve the performance compared to the baseline performance for all end-to-end models. Also, in Mixed, the performance increases except for the case of SlowFast. This shows all of the end-to-end models here, lack diverse features to generalize compared to the case in General Aug and DeepFake. For the comparison between General Aug and DeepFake, General Aug is focusing on randomizing the external parts (e.g., background, light, orientation, etc) whereas Deepfake is focusing on diversifying human-related factors (e.g., ethnicity, age, gender) and in the case of the joint engagement classification task, DeepFake showed more improvements except for the case with SlowFast. These results give us an insight into the generalization of our video augmentation techniques for joint engagement recognition.

The results from the skeleton-based models (see Table. 2) however, show lower top-1 accuracy (max acc: 64.3%) in all cases compared to the baseline in end-to-end models. These graph convolution network-based models attempt to find Spatio-temporal patterns from the human skeleton information. However, unlike human action recognition tasks, joint engagement recognition requires the model to catch the affective state (e.g., facial features) of people which RGB frames particularly include and this makes skeleton-based models challenging to learn joint engagement without the core information.

4.2 Visualizing Interaction Representations

To have a deeper understanding of what the models have learned, we use Grad-CAM to visualize the Spatio-temporal regions that contribute the most to classify into certain joint engagement classes on the dataset (see Fig. 3). We observe that the learned representations focus on either small regions in the "face and body" or get distracted by the backgrounds. For example, in Fig. 3, both the Grad-CAM in the General Aug and DeepFake groups match well with either parent's or child's face regions, but in the Baseline, the heatmap is not actively paying attention to dyads but looking at the corner in the scene. This implies that training on a diversified set of populations or backgrounds encourages the model to be more robust about the change of backgrounds, clothing, etc, and focus more on dyads' faces and bodies. Also, in CutOut, the model did not fully focus on the core parts of the scene but got distracted by the backgrounds. This is because it could not refer to the face part which includes the core information during the training.

5 DISCUSSION AND CONCLUSION

The performance and visualization of the state-of-the-art end-to-end video classification models for recognizing joint engagement demonstrated their potential to recognize complex human-human joint affective states with limited training data. The fine-tuned end-to-end models initially pre-trained for general video understanding (e.g., SlowFast, I3D) performed more effectively on joint engagement recognition than the models trained on human skeleton features (e.g., CTR-GCN, ST-GCN). The video augmentation techniques enhanced the model's performance even further. The visualization of the learned representations in the end-to-end deep learning models revealed their sensitivity to subtle social cues indicative of parent-child interaction. Altogether, these findings and insights indicate that end-to-end models were able to learn the representation of parent-child joint engagement in an interpretable manner.

Due to their superior performance and interpretability, end-to-end models pre-trained on action recognition tasks provide new opportunities for developing autonomous robots for multi-person human-robot interaction. When skeleton-based models are used for real-time affect perception in the wild, typically multiple layers

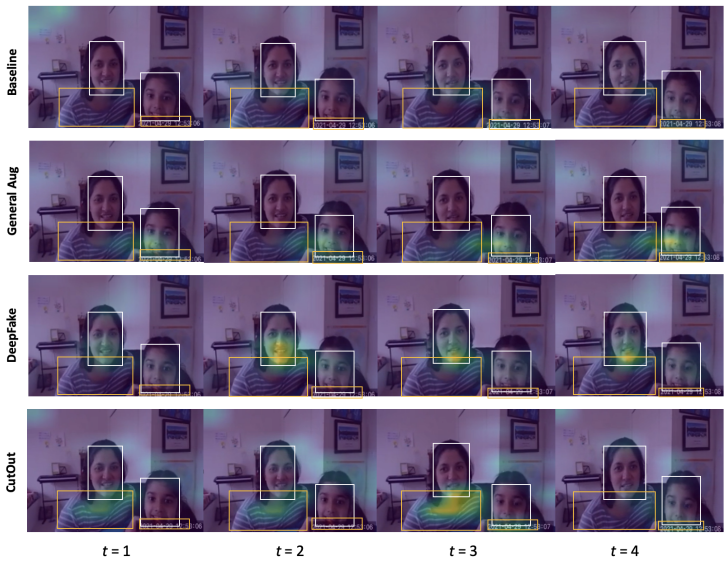


Figure 3: Grad-CAM visualization of a sample test video clip generated from fine-tuned I3D with different video augmentation techniques (Mixed is excluded here since it was only used for the performance comparison).

of models are necessary for human identification, human skeleton feature extraction, and affect prediction (e.g., [2]). In contrast to skeleton-based models, end-to-end models accept image frames as input and generate an affect prediction as output. After being trained, they can be utilized to make predictions in real-time without requiring excessive computational resources. The real-time version of the SlowFast model, for instance, can function with minimal computational resources [64]. Similarly, the visualization method we proposed provides a means of demonstrating how recognition models extract crucial information from videos for the learning task. This interpretable visualization could be used to support diverse methods of robot learning in real-time from human input or human teachers, such as the "social scaffolding for exploration" method and socially guided machine learning [57].

We acknowledge that the relatively small population size of the dyads may limit the applicability of the proposed framework for multi-person affect recognition. However, the small size of the dataset motivates the use of a framework that can leverage various data augmentation techniques and deep learning models pre-trained on larger data corpora for other similar tasks. Indeed, the performance of our proposed framework on the small dataset demonstrates its potential benefits for real-world multi-person HRI, as the real-world interaction datasets used to train a robot's perception model may not be as large as datasets parsed from the Internet or generated by simulation. We also acknowledge that the analysis in the visualization of the model's learned representation does not pinpoint the specific social cues and behavioral characteristics that guide the model's recognition of joint engagement. Our work only aims to demonstrate that the visualization of the model's learned representation can reveal semantically and interpretable insights that can be valuable to humans and can potentially allow humans to correct the model, challenging the widely held belief that end-to-end models have limited interpretability in comparison to skeleton-based models.

In the future, we plan to use additional parent-child interaction datasets [9] to investigate how to generalize this framework across datasets and to quantify how various social and behavioral cues contribute to the learned representations of end-to-end models. Also, the additional dataset is beneficial to deal with angle changes for mobile robots since the videos were collected from multiple cameras with different angles and distances. Additionally, our proposed framework is extensible in multiple ways. First, the proposed data augmentation and visualization can be applied to multiple data modalities, such as audio and video, to jointly learn the joint engagement. Utilizing multiple modalities in the dataset would increase the applicability of multi-person affect models to challenging real-world situations, such as missing data in one modality, and further enable the model to learn more holistic and human-like representations of joint engagement.

Our proposed framework can also be expanded to account for individual differences in affect across dyads by adding the final layer trained on individual human groups. It has been empirically demonstrated that personalized and culture-sensitive affect models outperforms one-size-fits-all generic models when recognizing affect of individuals (e.g., [49], [48]). Thus, personalized or culture-sensitive joint engagement models may further enhance the model prediction performances in the multi-person context.

For potential application, this work can be applied to group interactions that involve human-agent and human-human interactions in different scenarios including online education, group therapy and museum guidance [38, 44, 58].

In conclusion, this work serves as the first step toward fully unlocking the potential of state-of-the-art end-to-end video understanding models pre-trained on large public datasets and augmented with data augmentation and visualization techniques for robot's affect recognition in the multi-person human-robot interaction in the wild.

ACKNOWLEDGMENTS

This work was supported by the Inclusive AI Literacy and Learning gift grant from DP World and IITP grant funded by the Korea government (MSIT) (No.2020-0-00842, Development of Cloud Robot Intelligence for Continual Adaptation to User Reactions in Real Service Environments). Yubin Kim was supported by Korean Ministry of Trade, Industry, and Energy (MOTIE), under Human Resource Development Program for Industrial Innovation (Global, P0017311) supervised by the Korea Institute for Advancement of Technology (KIAT).

REFERENCES

[1] L.B. Adamson, R. Bakeman, and K. Suma. 2018. The Joint Engagement Rating Inventory. *Technical Report 25, 2nd ed.* (2018).

[2] Sharifa Alghowinem, Huili Chen, Hae Won Park, and Cynthia Breazeal. 2021. Body Gesture and Head Movement Analyses in Dyadic Parent-Child Interaction as Indicators of Relationship. In *IEEE International Conference on Automatic Face and Gesture Recognition 2021*. IEEE.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095 [cs.CV]

[4] Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*. 2–9.

[5] Huili Chen, Sharifa Mohammed Alghowinem, Soo Jung Jang, Cynthia Breazeal, and Hae Won Park. 2022. Dyadic Affect in Parent-child Multi-modal Interaction: Introducing the DAMI-P2C Dataset and its Preliminary Analysis. *IEEE Transactions on Affective Computing* (2022), 1–1. https://doi.org/10.1109/TAFFC.2022.3178689

[6] Huili Chen, Anastasia K. Ostrowski, Soo Jung Jang, Cynthia Breazeal, and Hae Won Park. 2022. Designing Long-term Parent-child-robot Triadic Interaction at Home through Lived Technology Experiences and Interviews. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 401–408. https://doi.org/10.1109/RO-MAN53752.2022.9900834

[7] Huili Chen, Hae Won Park, and Cynthia Breazeal. 2020. Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education* 150 (2020), 103836. https://doi.org/10.1016/j.compedu.2020.103836

[8] Huili Chen, Yue Zhang, Felix Weninger, Rosalind Picard, Cynthia Breazeal, and Hae Won Park. 2020. Dyadic Speech-Based Affect Recognition Using DAMI-P2C Parent-Child Multimodal Interaction Dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 97–106.

[9] Huili Chen, Y. Zhang, F. Weninger, Rosalind Picard, C. Breazeal, and H. W. Park. 2020. Dyadic Speech-based Affect Recognition using DAMI-P2C Parent-child Multimodal Interaction Dataset. *ArXiv* abs/2008.09207 (2020).

[10] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *MM '20: The 28th ACM International Conference on Multimedia*.

[11] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13359–13368.

[12] Vivienne M Colegrove and Sophie S Havighurst. 2017. Review of nonverbal communication in parent–child relationships: Assessment and intervention. *Journal of Child and Family Studies* 26, 2 (2017), 574–590.

[13] MMAction2 Contributors. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.

[14] MMFlow Contributors. 2021. MMFlow: OpenMMLab Optical Flow Toolbox and Benchmark. https://github.com/open-mmlab/mmflow.

[15] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[16] Fethiye Irmak Dogan, Gaspar I Melsion, and Iolanda Leite. 2021. Leveraging explainability for comprehending referring expressions in the real world. *arXiv preprint arXiv:2107.05593* (2021).

[17] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. 2022. PYSKL: Towards Good Practices for Skeleton Action Recognition. https://arxiv.org/abs/2205.09443

[18] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. 2021. Revisiting Skeleton-based Action Recognition. arXiv:2104.13586 [cs.CV]

[19] Christoph Feichtenhofer. 2020. X3D: Expanding Architectures for Efficient Video Recognition. arXiv:2004.04730 [cs.CV]

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*. 6202–6211.

[21] Kexin Feng and Theodora Chaspari. 2020. A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science* 2 (2020), 9.

[22] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. 2016. Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. In *Proc. of AAAI*. AAAI Press, Phoenix, Arizona, 3951–3957.

[23] Tsfira Grebelsky-Lichtman. 2014. Parental patterns of cooperation in parent–child interactions: The relationship between nonverbal and verbal communication. *Human Communication Research* 40, 1 (2014), 1–29.

[24] Paras Gulati, Qin Hu, and S Farokh Atashzar. 2021. Toward deep generalization of peripheral emg-based human-robot interfacing: A hybrid explainable solution for neurorobotic systems. *IEEE Robotics and Automation Letters* 6, 2 (2021), 2650–2657.

[25] Geoffrey Haddock, Gregory R. Maio, Karin Arnold, and Thomas Huskinson. 2008. Should Persuasion Be Affective or Cognitive? The Moderating Effects of Need for Affect and Need for Cognition. *Personality and Social Psychology Bulletin* 34, 6 (2008), 769–778. https://doi.org/10.1177/0146167208314871 arXiv:https://doi.org/10.1177/0146167208314871 PMID: 18344496.

[26] Man Hao, Wei-Hua Cao, Zhen-Tao Liu, Min Wu, and Peng Xiao. 2020. Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. *Neurocomputing* 391 (2020), 42–51.

[27] Alexander Heimerl, Tobias Baur, and Elisabeth André. 2020. A Transparent Framework towards the Context-Sensitive Recognition of Conversational Engagement. (2020).

[28] Jia-Hao Hsu and Chung-Hsien Wu. 2020. Attentively-Coupled Long Short-Term Memory for Audio-Visual Emotion Recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1048–1053.

[29] Malte F Jung. 2017. Affective grounding in human-robot interaction. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 263–273.

[30] Simon Keizer, Mary Ellen Foster, Zhuoran Wang, and Oliver Lemon. 2014. Machine Learning for Social Multiparty Human–Robot Interaction. *ACM Trans. Interact. Intell. Syst.* 4, 3, Article 14 (Oct. 2014), 32 pages. https://doi.org/10.1145/2600021

[31] Matthias Kerzel, Jakob Ambsdorf, Dennis Becker, Wenhao Lu, Erik Strahl, Josua Spisak, Connor Gäde, Tom Weber, and Stefan Wermter. 2022. What's on Your Mind, NICO? *KI-Künstliche Intelligenz* (2022), 1–18.

[32] Myeongjun Kim, Taehun Kim, and Daijin Kim. 2020. Spatio-temporal slowfast self-attention network for action recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2206–2210.

[33] C Leclère, M Avril, S Viaux-Savelon, N Bodeau, C Achard, S Missonnier, M Keren, R Feldman, M Chetouani, and David Cohen. 2016. Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3D reconstruction. *Translational psychiatry* 6, 5 (2016), e816.

[34] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. 2021. Improving the Movement Synchrony Estimation with Action Quality Assessment in Children Play Therapy. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 397–406. https://doi.org/10.1145/3462244.3479891

[35] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2020. Self-Correction for Human Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). https://doi.org/10.1109/TPAMI.2020.3048039

[36] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 143–152.

[37] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI* 7 (2020), 92.

[38] Anastasia K Ostrowski, Jenny Fu, Vasiliki Zygouras, Hae Won Park, and Cynthia Breazeal. 2022. Speed Dating with Voice User Interfaces: Understanding How Families Interact and Perceive Voice User Interfaces in a Group Setting. *Frontiers in Robotics and AI* (2022), 375.

[39] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. 2020. Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1–12.

[40] Zoe Papakipos and Joanna Bitton. 2022. AugLy: Data Augmentations for Robustness. arXiv:2201.06494 [cs.AI]

[41] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. 2019. A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education. In *Proc. of AAAI*. AAAI Press, Honolulu, Hawaii, 687–694.

[42] Rosalind W. Picard. 2003. Affective Computing: Challenges. *Int. J. Hum.-Comput. Stud.* 59, 1–2 (July 2003), 55–64. https://doi.org/10.1016/S1071-5819(03)00052-1

[43] R W Picard, S Papert, W Bender, B Blumberg, C Breazeal, D Cavallo, T Machover, M Resnick, D Roy, and C Strohecker. 2004. Affective Learning — A Manifesto. *BT Technology Journal* 22, 4 (2004), 253–269. https://doi.org/10.1023/B:BTTJ.0000047603.37042.33

[44] Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R Cowan, and Naomi Harte. 2022. RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2517–2527.

[45] Silvia Rossi, François Ferland, and Adriana Tapus. 2017. User profiling and behavioral adaptation for HRI: A survey. *Pattern Recognition Letters* 99 (nov 2017), 3–12. https://doi.org/10.1016/j.patrec.2017.06.002

[46] Meredith Rowe. 2008. Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *J. Child Lang.* 35, 1 (2008).

[47] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018), eaao6760.

[48] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W. Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018). https://doi.org/10.1126/scirobotics.aao6760

[49] O. Rudovic, Y. Utsumi, J. Lee, J. Hernandez, E. C. Ferrer, B. Schuller, and R. W. Picard. 2018. CultureNet: A Deep Learning Approach for Engagement Intensity Estimation from Face Images of Children with Autism. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 339–346.

[50] Hanan Salam, Oya Çeliktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2017. Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions. *IEEE Access* 5 (2017), 705–721. https://doi.org/10.1109/ACCESS.2016.2614525

[51] Samuel Spaulding, Huili Chen, Safinah Ali, Michael Kulinski, and Cynthia Breazeal. 2018. A Social Robot System for Modeling Children's Word Pronunciation: Socially Interactive Agents Track. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS'18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1658–1666.

[52] Samuel Spaulding, Goren Gordon, and Cynthia Breazeal. 2016. Affect-Aware Student Models for Robot Tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 864–872.

[53] Sarah Strohkorb, Iolanda Leite, Natalie Warren, and Brian Scassellati. 2015. Classification of Children's Social Dominance in Group Interactions with Robots. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 227–234. https://doi.org/10.1145/2818346.2820735

[54] Catherine S Tamis-LeMonda, Marc H Bornstein, and Lisa Baumwell. 2001. Maternal responsiveness and children's achievement of language milestones. *Child Dev.* 72, 3 (2001).

[55] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition* 118 (2021), 107965.

[56] Andrea Thomaz, Guy Hoffman, and Maya Cakmak. 2016. Computational Human-Robot Interaction. *Foundations and Trends in Robotics* 4, 2-3 (2016), 104–223. https://doi.org/10.1561/2300000049

[57] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. 2016. Computational human-robot interaction. *Foundations and Trends® in Robotics* 4, 2-3 (2016), 105–223.

[58] Marynel Vázquez, Elizabeth J Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E Hudson. 2017. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 42–52.

[59] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson. 2017. Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 42–52.

[60] M. Vázquez, A. Steinfeld, and S. E. Hudson. 2016. Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 36–43.

[61] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 24–25.

[62] Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang, and Shanshan Hao. 2019. I3d-lstm: A new model for human action recognition. In *IOP Conference Series: Materials Science and Engineering*, Vol. 569. IOP Publishing, 032035.

[63] Zhenhua Wang, Jiajun Meng, Jin Zhou, Dongyan Guo, Guosheng Lin, Jianhua Zhang, Javen Qinfeng Shi, and Shengyong Chen. 2020. LAGNet: Logic-Aware Graph Network for Human Interaction Understanding. *arXiv preprint arXiv:2011.10250* (2020).

[64] Dafeng Wei, Ye Tian, Liqing Wei, Hong Zhong, Siqian Chen, Shiliang Pu, and Hongtao Lu. 2022. Efficient dual attention SlowFast networks for video action recognition. *Computer Vision and ImageUnderstanding* 222 (2022), 103484.

[65] Zhengkui Weng, Wuzhao Li, and Zhipeng Jin. 2021. Human activity prediction using saliency-aware motion enhancement and weighted LSTM network. *EURASIP Journal on Image and Video Processing* 2021, 1 (2021), 1–23.

[66] Zerrin Yumak, Bram van den Brink, and Arjan Egges. 2017. Autonomous social gaze model for an interactive virtual character in real-life settings. *Computer Animation and Virtual Worlds* 28, 3-4 (2017), e1757.

[67] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.

[68] Lingyu Zhang and Richard J Radke. 2020. A Multi-Stream Recurrent Neural Network for Social Role Detection in Multiparty Interactions. *IEEE Journal of Selected Topics in Signal Processing* 14, 3 (2020), 554–567.

[69] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2017. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2017), 3030–3043.

[70] Yixiao Zhang, Baihua Li, Hui Fang, and Qinggang Meng. 2021. Current Advances on Deep Learning-based Human Action Recognition from Videos: a Survey. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 304–311.