Recognition of Human Interactions in Still Images using AdaptiveDRNet with Multi-level Attention

Arnab Dey[1], Samit Biswas[2], Dac-Nhoung Le[3]
Computer Science and Technology,
Indian Institute of Engineering Science and Technology, Shibpur, Howrah, 711103, India[1, 2]
Faculty of Information Technology, Haiphong University, Haiphong, 180000, Vietnam[3]

***Abstract*—Human-Human Interaction Recognition (H2HIR) is a multidisciplinary field that combines computer vision, deep learning, and psychology. Its primary objective is to decode and understand the intricacies of human-human interactions. H2HIR holds significant importance across various domains as it enables machines to perceive, comprehend, and respond to human social behaviors, gestures, and communication patterns. This study aims to identify human-human interactions from just one frame, i.e. from an image. Diverging from the realm of video-based interaction recognition, a well-established research domain that relies on the utilization of spatio-temporal information, the complexity of the task escalates significantly when dealing with still images due to the absence of these intrinsic spatio-temporal features. This research introduces a novel deep learning model called AdaptiveDRNet with Multi-level Attention to recognize Human-Human (H2H) interactions. Our proposed method demonstrates outstanding performance on the Human-Human Interaction Image dataset (H2HID), encompassing 4049 meticulously curated images representing fifteen distinct human interactions and on the publicly accessible HII and HIIv2 related benchmark datasets. Notably, our proposed model excels with a validation accuracy of 97.20% in the classification of human-human interaction images, surpassing the performance of EfficientNet, InceptionResNetV2, NASNet Mobile, ConvXNet, ResNet50, and VGG-16 models. H2H interaction recognition's significance lies in its capacity to enhance communication, improve decision-making, and ultimately contribute to the well-being and efficiency of individuals and society as a whole.**

***Keywords*—Human interaction recognition; still images; adaptiveDRNet; multi level attention; human interactions**

I. Introduction

Human-human interactions are the fundamental building blocks of human society, influencing various aspects of our lives, from personal relationships to collaborative efforts in professional settings. The study of these interactions has garnered increasing attention in recent years, driven by advancements in social psychology [1], communication sciences, and technology. Human-human interaction recognition from still images using deep learning has emerged as a compelling and cutting-edge research area with profound implications for a wide range of applications. In an era characterized by the ubiquity of image data and the growing demand for automated systems capable of understanding human behaviors, this field stands at the forefront of technological innovation.

In real-life scenarios, deep learning-based frameworks for recognizing human interactions in still images [2] find application in diverse fields. They play a crucial role in social behavior analysis [3] by detecting and deciphering subtle cues in body language and gestures, facilitating a deeper understanding of human interactions. They also find utility in educational settings, specifically identifying classroom interactions. Analyzing human interactions yields valuable behavioral data, facilitating data-driven decision-making across diverse industries. Enabling systems to understand user behavior and intentions through images can create more intuitive interfaces across applications, from gaming to virtual reality, seamlessly adapting technology to human interaction and preferences. Furthermore, these frameworks assist in automated content tagging on social media and content-sharing platforms, enhancing content discoverability.

The proposed AdaptiveDRNet with Multi-level Attention model represents a significant advancement in human interaction recognition from still images, standing out for its distinctive architectural features. The proposed network combines the multi-level attention mechanism and an adaptive deep residual network to enhance its ability to recognize human-human interactions in images. The term "Adaptive" signifies the model's dynamic ability to adjust its focus and attention within the input data. This adaptiveness is facilitated by the Multi-level self-attention mechanism that allows the model to intelligently prioritize important information, enhancing its capacity to discern complex human interactions. Instead of relying on a single, global attention mechanism, the model employs multiple attention levels. This approach enables the model to adaptively focus on salient image regions, capturing fine-grained spatial dependencies and intricate patterns crucial for accurate recognition. The residual connections help capture and propagate important information through the network. Additionally, the model incorporates depthwise separable convolution, batch normalization, and the Swish activation function. This combination improves computational efficiency and bolsters generalization, enabling the model to adapt effectively to diverse interaction scenarios. The Swish activation enhances non-linearity, and batch normalization aids in stable training, reducing overfitting risks. The proposed model offers a distinct advantage over other existing deep learning models for H2H Interaction recognition from images by incorporating the GELU (Gaussian Error Linear Unit) activation function instead of the traditional ReLU (Rectified Linear Unit) in the fully connected Dense layers. GELU activation, known for its smoothness and non-linearity, provides an essential edge to this model. In contrast to ReLU, which can suffer from vanishing gradients in deeper networks, GELU maintains gradient flow, facilitating the learning process in a deeper architecture.

Combining multi-level self-attention, efficient convolutional operations, and enhanced activations, this model excels in capturing nuanced details and spatial relationships, offering superior generalization for human interaction recognition from images, surpassing existing models in this field.

The proposed model mainly classifies the Human-Human Interaction images into fifteen categories: Celebrating, Dancing, Dining, Handshaking, Hugging, Protesting, Punching, Pointing, Waving, Kicking, Kissing, Highfive, RaisingHands, Talking, Teaching.

The main contributions of this research are: (a) Creation of a novel image dataset for the recognition of Human-Human Interactions (H2HID) with comprehensive data labelling. (b) Recognition of Human-Human Interaction in images is carried out using the proposed AdaptiveDRNet with Multi-level Attention. (c) Introduction of Regularized Categorical Cross-entropy (RCCE) Loss Function. (d) The model proposed in this study is assessed on three related standard benchmark datasets, HII, HIIv2 and Stanford40. (e) Result analysis with various well-established deep learning models based on accuracy and trainable parameters.

The structure of this research paper can be outlined as follows: The comprehensive exploration of related works is discussed in Section II. Section III delves into the intricacies of our proposed method, meticulously detailing our approach, which encompasses data preparation, the network architecture of the proposed model, and details of the loss function utilized. Section IV delves into the experimental results with various standard models on our curated H2HID dataset using various performance metrics and analysis of results with various available related datasets based on accuracy and the number of trainable parameters. The final section summarizes our unique contributions, emphasizing significance and discusses the future scope of this research.

II. Related Works

The study of human-human interactions (HHI), human-object interactions (HOI), human-computer [4] interaction (HCI) and human actions has been a prominent research focus [5] in the analysis of video sequences. However, it is noteworthy that there has been a noticeable decrease in the volume of research dedicated to exploring these topics when shifting the focus from video sequences to static still images. Tanisik et al. [6] introduced a range of facial region-based descriptors in their research. Their experiments revealed that while these facial descriptors offer valuable insights, their standalone use yields less effective results. However, when integrated with global scene features, particularly deep features, the proposed facial descriptors exhibited enhanced recognition performance and demonstrated the capability to recognize human interactions in static images. The authors have attained 80.11% accuracy using their collected Human Interaction image dataset. Gong et al. [7] introduced a new image dataset containing four distinct categories of human interactions. Li et al. [8] proposed a new method for transferring knowledge from images to videos, which adapts well to video data with limited training samples. They employ class-specific spatial attention maps within Convolutional Neural Networks (CNNs) to transform video frames into a condensed feature representation. Their approach incorporates a new Siamese EnergyNet framework, optimizing two loss functions to enhance attention maps aligned with ground truth concepts. They have attained 96.8% accuracy on HII data using the fine-tuned ResNet101 model. In another study, Tanisik et al. [9] delved into the significance of human poses in discerning human interactions within still images. Their novel approach introduces a multi-stream convolutional neural network architecture, harmonizing diverse human pose information to enhance human interaction recognition. Various pose-based representations are scrutinized, and extensive experimentation on an expanded benchmark dataset validates the efficacy of their multi-stream pose CNN in distinguishing a broad spectrum of human interactions and poses. When coupled with contextual information, it serves as a valuable tool for discriminative insights into human-human interactions. They have attained 92.78% accuracy in recognizing human interactions. Verma et al. [10] have employed a feature-based neural network to identify human interactions in images. Tang et al. [11] devised a novel approach to enhance vision-based safety compliance checks by explicitly categorizing worker-tool interactions. Their human-object interaction recognition model, built upon this detector and dataset, also delivered impressive results. On the other hand, Zhou et al. [12] tackled the detection and recognition of Human-object interactions (HOI) in images. They introduced a cascaded parsing net (CP-HOI) that employs a multi-stage, structured approach for HOI understanding. CP-HOI refines HOI proposals through instance detection and structured interaction reasoning (SIR) modules, utilizing a graph parsing neural network (GPNN) to represent HOI structures as interconnected graphs, enhancing contextual information extraction for better interaction comprehension.

On the contrary, there is a substantial body of literature dedicated to the recognition of human interactions in video streams. Numerous references, such as Zhou et al. [13] and Tapaswi et al. [14], have extensively explored this domain by utilizing traditional classification techniques. Nguyen et al. [15] utilize handcrafted features in conjunction with a three-layer convolutional neural network for model training. Yan et al. [16] introduce a CNN-based network to extract features, while Shu et al. introduce Hierarchical Long Short-Term Memory (HLSTM) network [17] to handle temporal information. Alazrai et al. [18] introduced an H2H interaction video dataset having 12 interactions giving more focus on the Pointing, Kicking and Punching interaction. Lee et al. [19] have recognized eight different interactions. Guerdelli et al. [20] provided a comprehensive overview of interpersonal relation recognition datasets and well-established methods, aiding researchers in gaining deeper insights into their characteristics.

Human interaction recognition is a specific area within the broader field of human action recognition. Thus, in addition to our focus on human interaction, we have also listed some notable and relevant existing research and studies in the field of action recognition. Zhang et al. [21] introduced an approach for recognizing actions in static images while minimizing the need for extensive annotations. Luo et al. [22] employed an improved EfficientNet framework to recognize human behaviours. In another study, Yu et al. [23] have proposed a deep ensemble learning model to recognize human actions in still images. In our prior research endeavours focused on Action Recognition, our primary objective was to predict various workout actions from still images [24], subsequently

classifying them into ten distinct workout categories. It's worth highlighting that the proposed WorkoutNet model showcased a remarkable validation accuracy of 92.75% when rigorously assessed on our WAId dataset. Siyal et al. [25] have proposed a Residual CNN model for feature extraction and SVM as a classifier to categorize human actions in still images. Saif et al. [26] have proposed an InceptionResNetV2-based CNN-LSTM model to recognize actions in videos.

Existing standard image datasets HII [27] and HIIv2 [6] have been pivotal in driving progress in the field of human interaction recognition. However, these publicly available datasets often lack specificity in terms of interaction types. The publicly available HII [27] dataset initially focused on four distinct human interaction categories, while the HIIv2 [6] dataset expanded this scope to include ten different human interaction classes. To further enhance the resources available for researchers and practitioners in the field, we have introduced the Human-Human Interactions Image dataset (H2HId), which comprises fifteen diverse types of human interactions. Our primary objective is to furnish the research community and professionals with valuable assets that can be utilized to develop and assess models designed to classify and recognize human interactions depicted in images accurately. This dataset will contribute to enhancing the generalization capabilities and real-world applicability of Human-human interaction recognition models, particularly in the context of interaction recognition.

The proposed AdaptiveDRNet model revolutionizes human interaction recognition from still images with multi-level self-attention, efficient convolution, and GELU activation. It dynamically adapts focus, captures fine details, and maintains gradient flow, outperforming traditional models.

III. Proposed Method

The method commenced by gathering human interaction images and standardizing them to a consistent size with augmentation. Subsequently, the curated dataset underwent division into training, validation, and test sets. Data augmentation was employed to enhance the training data. The proposed model, integrated with Callback functions, facilitated model training and extraction of image features. Ultimately, human-human interactions were categorized using the last layer of the proposed model, specifically the Softmax layer. The illustration of the proposed method can be visualized from Fig. 1.

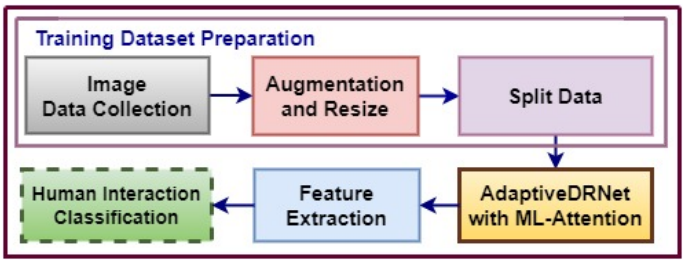

Fig. 1. Illustration of the proposed approach's workflow.

Recognizing the need for research methodologies that can perform efficiently with limited computational resources, this study introduces the AdaptiveDRNet with Multi-level Attention to recognize various human interactions in still images. The subsequent sections provide detailed insights into the following aspects: a) Training dataset preparation, b) *AdaptiveDRNet with ML-Attention*: Model Architecture, c) Description of the employed loss function, and d) Advantage of AdaptiveDRNet ML-Attention model.

A. Dataset Preparation

The dataset capturing Human-human interactions (H2HId) has been meticulously curated, enhanced through augmentation techniques, and then resized all images to a uniform resolution of 224 x 224. The collected images are in RGB. There are a total of 15 different human interaction categories. Some sample images of the collected H2HId dataset are highlighted in Fig. 2. The dataset details are given in Section IV(A).

Fig. 2. H2HId data sample.

Recognizing the significant demand for ample training data in deep learning models, we employed the data augmentation technique to bolster the overall performance of the AdaptiveDRNet with ML-Attention model. This approach entailed enlarging the training dataset and mitigating the risk of overfitting. We introduced random transformations, including a zoom range of 0.18, a contrast range of 0.23, rotations ranging from 0 to 25 degrees, and horizontal flips. These random transformations effectively generated additional training data, thereby exposing the training model to a wider array of potential data distribution characteristics. Furthermore, the H2HId dataset has been methodically split into three separate subsets following a precise distribution: 61% of the data is assigned for training, 20% for validation, and 19% for testing. This partitioning strategy adheres to established best practices in deep learning.

B. AdaptiveDRNet with ML-Attention: Model Architecture

The proposed Adaptive Deep Residual Network (AdaptiveDRNet) with Multi-level Attention model architecture comprises an initial convolutional layer with 32 filters, followed by three sets of residual blocks with attention mechanisms. The model begins with an input layer configured to accept RGB images with dimensions of 224x224 pixels. The initial layer is a 2D convolutional layer with a 3x3 kernel, utilizing 32 filters and applying the GELU activation function, followed

by batch normalization. Subsequently, three residual blocks with Self-Attention mechanisms are employed. Each block encompasses a depthwise convolution layer with a 3x3 kernel and either a stride of 1 (for the first block) or 2 (for subsequent blocks), maintaining spatial dimensions with 'same' padding. Batch normalization and Swish activation functions follow. The block also features a 1x1 convolutional bottleneck with varying filter sizes (64, 128, and 256, respectively, for each block) to capture and transform features. Self-Attention layers are introduced to capture long-range dependencies within the data. The model also comprises a Global Average Pooling (GAP) layer, succeeded by two dense layers consisting of 128 and 256 units, both utilizing GELU activation and dropout with a rate of 0.15 for regularization. Finally, the output layer is a dense layer employing the softmax activation function for classification into 15 human interaction classes. The model has 419,279 trainable parameters. This architecture offers a versatile and customizable framework for image classification tasks, with the flexibility to fine-tune hyperparameters based on specific datasets and requirements. The layered diagram representing the proposed AdaptiveDRNet with ML-Attention model, which is employed to implement the proposed work, is illustrated in Fig. 3.

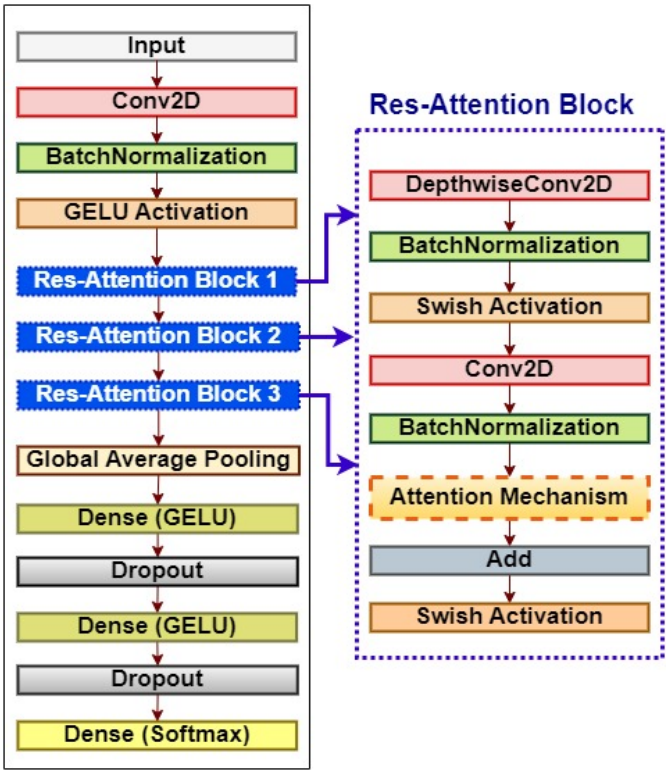

Fig. 3. Proposed adaptiveDRNet ML-attention model.

The proposed AdaptiveDRNet ML-Attention model architecture is outlined as follows:

Input Layer: The model takes input images with a shape of 224x224 pixels and three color channels (RGB).

Initial Convolution Layer: The input images are processed by a convolutional layer with 3x3 kernels having 32 filters, batch normalization (BatchNorm) layer, and a GELU activation function. This layer extracts basic features from the input images. The choice of a GELU activation function is pivotal, as it captures non-linearities more effectively than traditional activation functions. This layer's primary role is to extract low-level features from input images while maintaining spatial dimensions. Batch normalization aids in faster convergence and mitigating overfitting, ensuring smoother training. The first convolution layer applies a set of filters to the input, followed by batch normalization and a GELU activation function, which introduces non-linearity. The 2D convolution operation is represented as illustrated in Eq. (1). Here, $y(q,r)$ is the output at position (q,r) denotes the feature map, $x(q,r)$ represents the input feature map, $w(l,m)$ are the convolutional filter weights and b denotes the bias term.

$$y(q,r) = \sum_{l=0}^{L-1}\sum_{m=0}^{M-1} x(q+l, r+m)\cdot w(l,m) + b \quad (1)$$

Batch Normalization Layer (BatchNorm): Batch Normalization is a method employed to standardize the output of a layer. This process involves subtracting the mean of the batch and dividing it by the batch standard deviation. This process can be mathematically represented as illustrated in Eq. (2). In this equation, y_B represents the normalized output, I is the input to the BatchNorm layer, μ denotes the batch mean, σ^2 stands for the batch variance, β denotes the shifting parameter (learnable), ϵ is a small constant added for numerical stability, γ represents scaling parameter (learnable).

$$y_B = \frac{I-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma + \beta \quad (2)$$

Res-Attention Block: The proposed model comprises three Residual Attention (Res-Attention) blocks that serve as a powerful feature extraction unit. This block consists of several key layers, starting with a depthwise separable convolutional layer, followed by the BatchNorm layer and the Swish activation. Each block follows a consistent pattern, commencing with depthwise convolution, which performs spatial convolutions independently for each channel, preserving spatial dimensions. This is particularly beneficial for capturing fine-grained details in the image.

The DepthwiseConv2D layer performs a depthwise convolution operation, where each input channel is convolved separately with its own set of learnable filters. The DepthwiseConv2D operation can be represented as illustrated in Eq. (3) The operation is performed independently for each input channel, resulting in an output with the same number of channels as the input. Here, $Y_{i,j,k}$ is the value of the output feature map at position (i,j) and channel k, $X_{(i+d-1),(j+e-1),k}$ is the value of the input feature map at position $(i+d-1, j+e-1)$ and channel k, $K_{d,e,k}$ is the value of the depthwise convolution filter at position (d,e) and channel k, B_k is the bias term for channel k and the summation is over the filter size D and both spatial dimensions.

$$Y_{i,j,k} = \sum_{d=1}^{D}\sum_{e=1}^{D} X_{(i+d-1),(j+e-1),k}\cdot K_{d,e,k} + B_k \quad (3)$$

The proposed model incorporates Swish activation functions [28], a choice made due to their well-known smoothness

characteristics and effectiveness in enhancing overall model performance. Subsequently, a 1x1 convolutional bottleneck further processes the features. The distinctive feature is the Self-Attention Layer, which calculates attention scores between spatial locations, allowing the block to weigh the significance of different regions and capture global dependencies. Like the residual blocks, the Attention Block includes a residual connection that adds the Self-Attention output to the original input feature map. This combination of residual connections and Self-Attention enables the block to effectively capture both local and long-range contextual information, making it a valuable building block for image recognition tasks where understanding spatial relationships and context is crucial for accurate classification.

Attention Mechanism: The role of the attention mechanism is pivotal in augmenting the performance of the proposed AdaptiveDRNet model. The attention mechanism, based on the scaled dot-product attention mechanism, is employed to selectively emphasize and de-emphasize features within the intermediate representations of the network. Its importance lies in its ability to capture and focus on crucial information while discarding less relevant details. The attention mechanism is represented using mathematical equations illustrated in Eq. (4), Eq. (5) and Eq. (6).

At first, the Query (Q), Key (K), and Value (V) are obtained by projecting the input (X) using learnable weight matrices W_q, W_k, and W_v, respectively.

$$Q = X \cdot W_q K = X \cdot W_k V = X \cdot W_v \tag{4}$$

The attention logits are calculated as the dot product of Q and K, scaled by the square root of the dimension of K. Here, d_k denotes the dimension of the key vectors.

$$Att_{Logits} = \frac{Q \cdot K^T}{\sqrt{d_k}} \tag{5}$$

The attention weights are computed by applying the softmax function to the attention logits (Att_{Logits}) as illustrated in Eq. (6).

$$Att_{Weights} = \text{softmax}(Att_{Logits}) \tag{6}$$

Finally, the attention output is obtained by taking the weighted summation of the values (V) utilizing the attention weights ($Att_{Weights}$) as illustrated in Eq. (7).

$$Att_{Output} = \text{Att Weights} \cdot V \tag{7}$$

Here, the "Multi-level Attention" aptly characterizes the attention mechanism within our proposed model due to its ability to operate on multiple levels or scales of information. This means that it can simultaneously focus on fine-grained details and broader context within the input data. It's not limited to a single level of attention but rather incorporates various levels, making it a versatile tool for capturing nuanced relationships and patterns in the data. The attention mechanism's significance in the proposed model is multifaceted. Firstly, it aids in the model's ability to focus on relevant regions of the input, thereby improving feature selection. Secondly, it enhances feature refinement by amplifying important information. Lastly, it facilitates gradient flow during training, as it provides a clear path for gradients to propagate through the network.

Global Average Pooling (GAP) Layer: After the residual attention blocks, GAP is performed, which reduces the spatial dimensions to a single vector for each feature map while retaining essential information. This operation helps in creating a fixed-size representation of the features extracted by the previous layers. GAP calculates the average value of each feature map, effectively summarizing the presence of specific features across the entire image. This is essential for making the model translation-invariant, allowing it to recognize objects regardless of their position in the image. GAP contributes to reducing the model's computational complexity and parameters, making it more efficient. The GAP operation is represented using Eq. (8). Here, $GAP(Y_3)[c]$ represents the spatial average value of channel c in the feature map Y_3, which is the output of Res-Attention Block 3 in the model. K and L are the spatial dimensions of the feature map, and C is the number of channels. The GAP operation is applied independently to each channel, calculating the spatial average of values across the entire spatial region of Y_3.

$$GAP(Y_3)[c] = \frac{1}{K \cdot L} \sum_{i=1}^{K} \sum_{j=1}^{L} Y_3[i, j, c] \tag{8}$$

Fully Connected Layers: The proposed model comprises two dense layers (fully connected) with GELU activation. The initial dense layer comprises 128 units, succeeded by a dropout layer configured having a 0.15 dropout rate, strategically employed to mitigate overfitting. The subsequent dense layer consists of 256 units, accompanied by yet another dropout layer with an identical dropout rate. These layers enable the model to learn complex patterns and relationships in the feature representations produced by the earlier layers. The dropout layers further prevent overfitting by randomly deactivating a fraction of neurons during training.

The GELU activation function applied to the fully connected dense layer is mathematically represented using Eq. (9). This activation function introduces a layer of non-linearity into the network, enabling it to capture intricate relations and patterns within the data. Let's denote the input to the fully connected layer as I, which is a vector of activations from the previous layer. The GELU activation denoted by G(I) is applied element-wise to each element of the input I. In Eq. (9) and Eq. (10), I represents the input to the fully connected layer, $\tanh$ is the hyperbolic tangent function, $\sqrt{\frac{2}{\pi}}$ is a constant and 0.044715 is a constant. GELU helps with the vanishing gradient problem and is utilized in the proposed model for improved performance.

$$G(I) = \frac{1}{2} I \left(1 + \tanh(R1)\right) \tag{9}$$

$$R1 = \left(\sqrt{\frac{2}{\pi}} \left(I + 0.044715 I^3\right)\right) \tag{10}$$

Output Layer: The ultimate layer of the proposed model consists of a dense layer comprising 15 units, as there are 15 human interaction categories. It employs softmax activation to produce probabilities for each human interaction class, making it suitable for multi-class classification.

The illustration in Fig. 4 emphasizes the filter outputs stemming from an image belonging to the 'Pointing' interaction class. Specifically, it highlights the outputs originating from the 1st DepthwiseConv2D layer (1st Res-Attention block), the 2nd DepthwiseConv2D layer (2nd Res-Attention block), and the 3rd Conv2D layer (2nd Res-Attention block). It's worth noting that the numbering of these layers aligns with the sequence provided in the layered diagram of the model, which is depicted in Fig. 3.

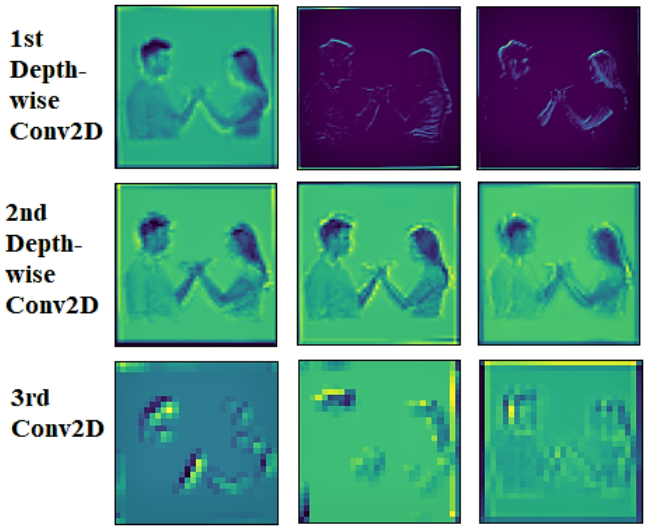

Fig. 4. Feature map visualization of some layers.

Thus, this model architecture incorporates convolutional layers for feature extraction, deep residual blocks with attention mechanisms to capture important features, global average pooling to reduce spatial dimensions, and fully connected layers with dropout enable classification. The attention mechanism between the convolutional layers allows the model to focus on relevant image regions, which can be crucial for tasks with complex visual patterns. The final output layer provides class probabilities for making predictions.

The proposed AdaptiveDRNet model architecture is meticulously designed to capture features at different levels, from low-level details to high-level contextual information with Multi-level Self Attention. The Self-Attention mechanism is a key innovation, enhancing the model's ability to capture spatial relationships, making it well-suited for image classification tasks where context and object dependencies are critical. The Multi-level Attention mechanism enables the model to adaptively focus on important features within the data, enhancing feature selection and representation learning.

C. Loss Function Details

In addressing the multi-class nature of this study, the Regularized categorical cross-entropy (RCCE) loss function is employed to quantify the error that the model seeks to decrease throughout its training. The Standard categorical cross-entropy loss ($SCCE_{Loss}$) is expressed as illustrated in Eq. (11). Here, $y_{\text{tp}}^{(k)}$ represents actual probability of class k and $y_{\text{pp}}^{(k)}$ represents the anticipated probability of class k.

$$SCCE_{Loss} = -\sum_{k=1}^{M=15} y_{\text{tp}}^{(k)} \cdot \log(y_{\text{pp}}^{(k)}) \tag{11}$$

The $RCCE_{Loss}$ function adds an extra term, which is the L2 regularization [29] term (LT), to the standard loss as mathematically illustrated in Eq. (12).

$$RCCE_{Loss} = SCCE_{Loss} + LT \tag{12}$$

The L2 regularization term (LT) is mathematically expressed as shown in Eq. (13). Here, y_{tp} denote the true probability distribution (one-hot encoded labels) of class membership, y_{p} denote the predicted probability distribution (model's output) of class membership and value of M is 15 which denotes the number of interaction classes in the dataset.

$$LT = \sum_{k=1}^{M=15} (y_{\text{tp}}^{(k)} - y_{\text{pp}}^{(k)})^2 \tag{13}$$

The Nadam optimizer is then utilized with an initial learning rate set at 0.001 to effectively reduce the error function linked to the proposed model.

D. Advantage of AdaptiveDRNet ML-Attention Model

In the proposed model, three residual blocks with attention having varying filter sizes (64, 128, 256) help in hierarchical feature extraction. The use of multiple residual blocks with increasing filter sizes allows the model to learn hierarchical features from low-level to high-level representations. Smaller filter sizes (64) in the initial layers help capture fine-grained details and edges, while larger filter sizes (128 and 256) in subsequent layers capture more abstract and complex features. Each residual block typically includes a convolutional layer with a stride greater than 1, which downsamples the spatial dimensions of the feature maps. Starting with a stride of 2 in the first block and possibly increasing it in later blocks helps reduce the spatial resolution of the feature maps. This downsampling reduces computational complexity and increases the receptive field of the network. By using progressively larger filter sizes, the model increases its capacity to learn more complex patterns and features in the data. The skip connections in residual networks enable the reuse of features from prior layers. With multiple blocks, each incorporating its attention mechanism, the model can selectively leverage features from different network stages. This allows the model to focus on fine-grained and high-level features, improving prediction. Therefore, the design choice of three residual blocks with attention and varying filter sizes facilitates efficient feature learning and representation across different scales and complexities.

IV. Experimental Results

This proposed research is conducted in Google Colab using Python programming, emphasising resource efficiency without utilizing any GPU. It leverages an Intel-Xeon CPU with 2.3GHz clock speed, 13GB of RAM, and approximately 80GB of disk space, diverging from the current practice of relying on GPUs for deep learning tasks. In this section, we have provided the details of our curated H2HId dataset, including available datasets, thorough performance evaluation, and comprehensive analysis of the results.

A. Dataset Details

Several publicly accessible datasets related to human interactions include HII [27], HIIv2 [6], and Stanford40 [30]. In this work, we have introduced a substantial dataset called the Human-Human Interaction Image dataset (H2HId) [1], a comprehensive collection of images sourced from a variety of online platforms and social media. To ensure effective categorization, images depicting diverse human interaction scenarios were systematically organized into distinct directories. The H2HId dataset is noteworthy for its unbiased representation of human interactions across various individuals and scenarios, aiming for universality without demographic, ethnic, or regional bias. The proposed dataset encompasses a diverse range of human interactions, encompassing 15 distinct categories and containing a total of 4,049 images. The human interaction categories considered are *Celebrating*, *Dancing*, *Dining*, *Handshaking*, *Hugging*, *Protesting*, *Punching*, *Pointing*, *Waving*, *Kicking*, *Kissing*, *Highfive*, *RaisingHands*, *Talking*, *Teaching*.

B. Performance Evaluation

The effectiveness of the proposed research is assessed through a range of metrics, including measures like training and validation accuracy, F1-score ($F1_s$), AUC-score, classification performance, and the analysis of the confusion matrix.

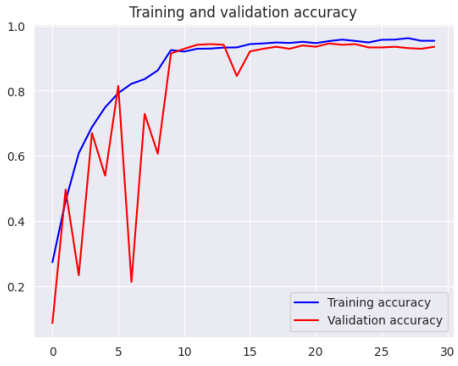

Fig. 5. Accuracy curve of proposed adaptiveDRNet attention model.

Fig. 5 illustrates the proposed model's training and validation accuracy curve. It is seen from the curve that the proposed model shows 99.57% train accuracy and 97.20% validation accuracy. After 28 epochs, the performance of the proposed model did not improve much, so we stopped the training process at that point by utilizing the callback functions. The hyperparameters of the proposed AdaptiveDRNet ML-Attention model are highlighted in Table I.

TABLE I. Hyperparameters of Proposed Model

Parameter(s)	Considered Value
Input Shape	224X224
Batch Size	5
Epochs	28
No. of Classes	15
Kernel	3X3
Initial Learning Rate	0.001
Loss Function	Regularized CCE
Metrics	Accuracy
Optimizer	Nadam

The proposed work has been evaluated with several benchmark deep learning models, all of which were trained from scratch. Table II presents the training accuracy (Train Acc) and validation accuracy (Val. Acc) and F1-Score ($F1_s$) of VGG-16, ResNet50, InceptionResNetV2, NASNet Mobile, ConvXNet, EfficientNet [22], and our proposed model on the H2HId dataset. The proposed model significantly outperforms the benchmark deep models in terms of accuracy and achieves an impressive 97.20% accuracy on the validation data, surpassing the performance of the EfficientNet [22], NASNet Mobile, ConvXNet, InceptionResNetV2, ResNet50 and VGG-16 model. Additionally, the InceptionResNetV2, NASNet Mobile, ConvXNet and EfficientNet [22] models also exhibited strong performance in classifying the images of the H2HId dataset, achieving validation accuracies (Val. Acc) of 95.30%, 94.76%, 94.60%, and 96.52% respectively.

TABLE II. Assessment of Deep Learning Models' Performance on H2HId Dataset

Model	Train Acc	Val. Acc	$F1_s$	AUC
VGG-16	96.21%	94.27%	0.92	0.95
ResNet50	97.08%	94.54%	0.93	0.96
InceptionResNetV2	97.61%	95.30%	0.95	0.98
NASNet Mobile	97.85%	94.76%	0.94	0.97
ConvXNet	98.28%	94.60%	0.94	0.97
EfficientNet [22]	98.70%	96.52%	0.95	0.98
Proposed Model	**99.57%**	**97.20%**	**0.96**	**1.0**

When considering the F1-score ($F1_s$), the Proposed Model performed exceptionally well, achieving the highest $F1_s$ of 0.96, whereas the EfficientNet [22], and InceptionResNetV2 model achieved the second highest $F1_s$ of 0.95 on our H2HId dataset. This indicates that the proposed model exhibits a balanced combination of precision and recall score, making it highly effective in human interaction classification. However, it's worth noting that the proposed model maintains its high $F1_s$ while simultaneously achieving the highest validation accuracy, demonstrating its consistency and robustness in performance. In evaluating deep learning models on the H2HId dataset, their performance was further assessed using Area Under Curve (AUC) scores [31], a crucial metric for measuring their ability to discriminate between positive and negative instances. The results revealed that the Proposed Model stands

[1]H2HId Dataset Link: https://sites.google.com/view/h2hid/home

out as the top performer with a perfect AUC score 1.0. This achievement showcases its exceptional capability in effectively distinguishing between classes. Following closely, the EfficientNet [22] and InceptionResNetV2 model exhibited strong performance with an AUC score of 0.98. In comparison, the other models, including InceptionResNetV2, NASNet Mobile, ConvXNet, ResNet50, and VGG-16, demonstrated descending levels of AUC scores, with the VGG-16 model having the lowest AUC score at 0.95. These findings emphasize the superiority of the proposed model and EfficientNet [22] in terms of their ability to discriminate between positive and negative instances on the H2HId dataset.

Confusion Matrix: It provides a clear and informative snapshot of a model's predictions aligning with actual ground truth values. In this visualization, the rows represent the true or actual classes, while the columns depict the predicted classes. Each cell in the matrix indicates the number of instances that fall into a particular category. The confusion matrix of the H2HId dataset utilizing the proposed model is depicted in Fig. 6. It serves as a powerful tool for fine-tuning and optimizing classification models.

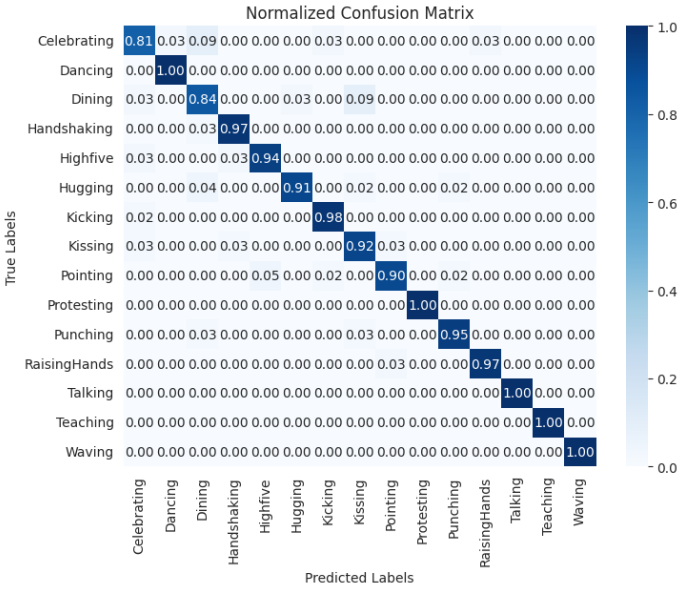

Fig. 6. Confusion matrix of H2HId dataset utilizing proposed model.

Classification Performance: The performance of our Proposed model in classifying human interactions is visually depicted in Fig. 7, showcasing the F1-scores ($F1_s$) for the H2HId dataset.

It is observed that most of the interaction categories in the H2HId dataset exhibit high $F1_s$, exceeding 0.95. However, a few categories, such as "Celebrating" (0.85 $F1_s$), "Dining" (0.90 $F1_s$), "Hugging" (0.93 $F1_s$), and "Pointing" (0.92 $F1_s$), demonstrate slightly lower but still respectable $F1_s$. The proposed AdaptiveDRNet with Multi-level Attention model demonstrates outstanding classification performance on the H2HId dataset. The proposed AdaptiveDRNet model has accurately predicted all the human-human interaction test image samples taken for further evaluation as depicted in Fig. 8.

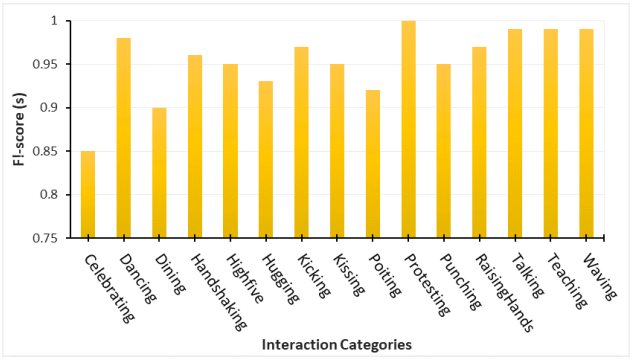

Fig. 7. Classification score of H2HId dataset utilizing proposed model.

Fig. 8. Test result of proposed model.

C. Result Analysis

The analysis of results with the proposed AdaptiveDRNet with Multi-level Attention model on various Optimizers is depicted in Table III. In this context, the remaining model parameters, including a fixed learning rate (LR) of 0.001, the utilization of the Regularized Cross-entropy loss (RCCE) as the loss function (Loss Func.), a consistent number of training epochs set to 28, and a batch size of 5, were maintained unchanged throughout the optimizer analysis.

TABLE III. ASSESSMENT OF PROPOSED MODEL ON VARIOUS OPTIMIZERS

Optimizer	Epochs	Loss Fun.	Val. Acc	MCC Val.
SGD	28	RCCE	94.20%	0.9447
RMS	28	RCCE	94.60%	0.9509
Adagrad	28	RCCE	95.35%	0.9517
Adam	28	RCCE	96.43%	0.9640
Nadam	**28**	**RCCE**	**97.20%**	**0.9726**

In the assessment of the proposed model's performance on the H2HId dataset using various optimizers, a nuanced understanding of the optimizer's impact on model accuracy becomes evident. Starting with Stochastic Gradient Descent

(SGD), which achieved a Validation accuracy (Val. Acc) of 94.20%, we observed steady progress with Root Mean Square Propagation (RMS) and Adagrad, reaching Val. accuracies of 94.60% and 95.35%, respectively. However, the Adam optimiser notably improved the model's accuracy, achieving 96.43%. It is observed that the best performance is produced by the use of the Nadam optimizer, which stood out among the optimizers with a remarkable accuracy of 97.20%. These results emphasize the significance of optimizer selection in fine-tuning deep learning models, with Nadam proving to be the most effective choice for maximizing the model's accuracy on the H2HId dataset. Further, the optimizers' performance in the proposed model is evaluated utilizing the Matthews Correlation Coefficient (MCC) values [32] to gauge the quality of multi-class human-human interaction classifications. This metric is crucial in evaluating the model's ability to provide precise predictions while considering both false positives and false negatives. Upon examining the MCC values (MCC Val.) for each optimizer, a clear trend emerged. The Nadam optimizer outshone others, securing the highest 0.9726 MCC value. This achievement signifies a remarkable level of agreement between the model's predictions and the true labels. The Nadam optimizer stands out as the top performer, demonstrating exceptional classification capabilities. This emphasizes the significant influence of optimizer selection on a model's ability to generalize and make accurate predictions.

The proposed technique is compared with various standard CNN models to test the efficacy, as demonstrated in Table IV. It depicts the accuracy results of different deep learning-based models with various benchmark datasets, namely HII [27], HIIv2 [6] and our H2HId dataset. Among all the models considered for evaluation, the proposed model attained the top-most accuracy of 96.38% on the HII dataset, followed by the EfficientNet [22] and EnsembleNet [33] models with 95.83% and 95.37% accuracy, respectively. Our method demonstrated high classification performance on the two benchmark datasets, namely the HII and HIIv2, yielding remarkable accuracies.

TABLE IV. COMPARATIVE EVALUATION OF DEEP LEARNING MODELS IN HUMAN INTERACTION RECOGNITION

Method(s)	HII	HIIv2	H2HId	TP (M)
VGG-16	93.20%	81.36%	94.27%	134.70
ResNet-50	93.56%	81.20%	94.54%	25.60
MobileNetv2	94.54%	82.16%	94.67%	3.50
DELVS1 [23]	94.20%	82.28%	95.82%	>140
IncepResNetV2	95.24%	82.67%	95.30%	55.90
EnsembleNet[33]	95.37%	82.40%	96.21%	–
EfficientNet [22]	95.83%	83.17%	96.52%	1.03
Proposed Model	**96.38%**	**83.42%**	**97.20%**	**0.41**

Notably, the proposed model achieved the highest accuracy across all three datasets, showcasing its exceptional ability to accurately identify human interactions. The proposed model attained 96.38% on HII, 83.42% on HIIv2, and 97.20% on H2HId datasets. On the HIIv2 dataset, EfficientNet [22] and EnsembleNet [33] attain 83.17% and 82.40% accuracy respectively. The InceptionResNetV2 (IncepResNetV2), DELVS1 [23], and EfficientNet [22] model attain 95.30%, 95.82% and 96.52% accuracy, respectively on our H2HId dataset. let's consider the comparison based on the number of trainable parameters (TP). The proposed model features a comparatively streamlined design, comprising merely 0.41 million (M) TP. In contrast, some of the benchmark models have significantly larger numbers of trainable parameters. For instance, VGG-16 and ResNet50 have 134.70 million and 25.60 million TP, respectively. DELVS1 [23] model has 140+ million TP, IncepResNetV2 has 55.90 M, MobileNetv2 has 3.50 M, and EfficientNet [22] has 1.03 million TP. The proposed model's advantage lies in its ability to achieve superior accuracy while maintaining significantly smaller trainable parameters than these well-established benchmark models. This is particularly important in real-world applications where computational efficiency and memory constraints are critical factors.

The F1-Score ($F1_s$) on both the Stanford40 [30] and the HII [27] datasets demonstrates strong classification performance across multiple established deep learning models, as depicted in Fig. 9. Stanford40 is a still image action dataset with 40 action categories, which is also considered in the analysis.

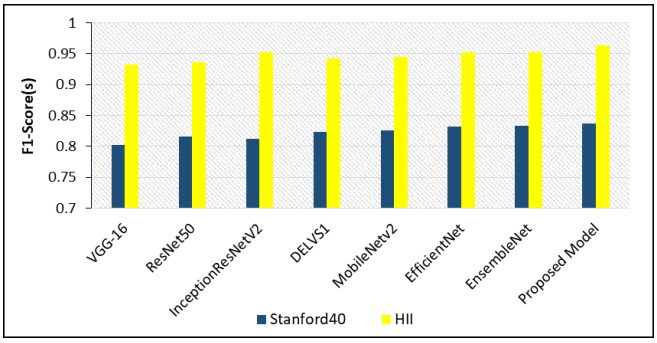

Fig. 9. Performance analysis of various models on stanford40 and HII datasets.

The proposed model emerges as the top performer, showcasing its outstanding classification performance with 0.83 $F1_s$ on the Stanford40 dataset and an impressive 0.96 $F1_s$ on our HII dataset. This model consistently demonstrates superior classification capabilities in recognizing human interactions and actions. Both the EfficientNet [22] and EnsembleNet [33] model attains 0.95 $F1_s$ on HII dataset. The InceptionResNetV2 model achieves 0.81 $F1_s$ on Stanford40, whereas the EnsembleNet [33] model achieves 0.83 $F1_s$ on Stanford40 dataset, and both these model attains $F1_s$ above 0.95 on HII dataset.

V. CONCLUSION AND FUTURE SCOPE

Human-human interaction recognition from still images using deep learning is a rapidly evolving field with vast implications across various domains. Recognizing complex human interactions from static visual cues using the proposed model finds applications in social behavior analysis, surveillance, market research, education, content tagging, and human-robot interactions. The proposed network in this research combines multi-level attention mechanisms and residual networks to

enhance its ability to recognize human interactions, allowing it to focus on relevant features within images automatically. This approach aims to improve accuracy and effectiveness in recognizing intricate human-human interactions, making it valuable in computer vision and video analysis. The study's main contributions include the development of a new dataset for Human-Human Interaction Recognition (H2HID) with comprehensive labelling, the introduction of a novel AdaptiveDRNet Multi-level Attention Network for recognizing human interactions in images, and an extensive result analysis involving various well-established deep learning models based on accuracy and the number of trainable parameters, utilizing established related benchmark datasets for a comprehensive evaluation. The Multi-level Attention in our model excels by operating on multiple information scales, enabling simultaneous focus on fine-grained details and broader context. Its versatility lies in its ability to encompass attention at various levels, capturing nuanced relationships and data patterns. The suggested model is suitable for operation on devices with limited resources since it employs minimal trainable parameters. This research underscores the potential of deep learning in advancing our understanding of human interactions from visual data and its wide-ranging applications in diverse fields.

The future scope of this research encompasses several promising avenues for further exploration and enhancement. Firstly, there is a significant potential for expanding the dataset size used in this study. A larger dataset would not only increase the diversity and representativeness of human interactions but also enable the model to generalize better across various scenarios. Furthermore, the inclusion of additional and more diverse human interaction categories within the dataset would enhance the model's capability to recognize a broader spectrum of interactions. Another important direction for future research involves the creation of improved classifiers that prioritize energy efficiency to ensure the feasibility and scalability of the proposed work in real-world settings.

REFERENCES

[1] T. M. Newcomb, R. H. Turner, and P. E. Converse, *Social psychology: The study of human interaction*. Psychology Press, 2015.

[2] A. Stergiou and R. Poppe, "Analyzing human–human interactions: A survey," *Computer Vision and Image Understanding*, vol. 188, p. 102799, 2019.

[3] P. Khaire and P. Kumar, "Deep learning and rgb-d based human action, human–human and human–object interaction recognition: A survey," *Journal of Visual Communication and Image Representation*, vol. 86, p. 103531, 2022.

[4] N. A. Mashudi, M. A. M. Izhar, and S. A. M. Aris, "Human-computer interaction in mobile learning: A review," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, 2022.

[5] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai," *International Journal of Human–Computer Interaction*, vol. 39, no. 3, pp. 494–518, 2023.

[6] G. Tanisik, C. Zalluhoglu, and N. Ikizler-Cinbis, "Facial descriptors for human interaction recognition in still images," *Pattern Recognition Letters*, vol. 73, pp. 44–51, 2016.

[7] W. Gong, J. Gonzàlez, J. M. R. S. Tavares, and F. X. Roca, "A new image dataset on human interactions," in *Articulated Motion and Deformable Objects*. Springer, 2012, pp. 204–209.

[8] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Attention transfer from web images for video recognition," in *Proceedings of the 25th ACM international conference on multimedia*, 2017, pp. 1–9.

[9] G. Tanisik, C. Zalluhoglu, and N. Ikizler-Cinbis, "Multi-stream pose convolutional neural networks for human interaction recognition in images," *Signal Processing: Image Communication*, vol. 95, p. 116265, 2021.

[10] A. Verma, T. Meenpal, and B. Acharya, "Multiperson interaction recognition in images: A body keypoint based feature image analysis," *Computational Intelligence*, vol. 37, no. 1, pp. 461–483, 2021.

[11] S. Tang, D. Roberts, and M. Golparvar-Fard, "Human-object interaction recognition for automatic construction site safety inspection," *Automation in Construction*, vol. 120, p. 103356, 2020.

[12] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2827–2840, 2021.

[13] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian, "Interaction part mining: A mid-level approach for fine-grained action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3323–3331.

[14] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "Storygraphs: Visualizing character interactions as a timeline," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 827–834.

[15] N. Nguyen and A. Yoshitaka, "Human interaction recognition using independent subspace analysis algorithm," in *IEEE International Symposium on Multimedia*, 2014, pp. 40–46.

[16] Y. Yan, B. Ni, and X. Yang, "Predicting human interaction via relative attention model," in *International Joint Conference on Artificial Intelligence (IJCAI-17)*, 2017, pp. 3245–3251.

[17] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1110–1118, 2021.

[18] R. Alazrai, A. Awad, A. Baha'A, M. Hababeh, and M. I. Daoud, "A dataset for wi-fi-based human-to-human interaction recognition," *Data in brief*, vol. 31, p. 105668, 2020.

[19] N.-G. Cho, S.-H. Park, J.-S. Park, U. Park, and S.-W. Lee, "Compositional interaction descriptor for human interaction recognition," *Neurocomputing*, vol. 267, pp. 169–181, 2017.

[20] H. Guerdelli, C. Ferrari, and S. Berretti, "Interpersonal relation recognition: a survey," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 417–11 439, 2023.

[21] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5479–5490, 2016.

[22] C.-Y. Luo, S.-Y. Cheng, H. Xu, and P. Li, "Human behavior recognition model based on improved efficientnet," *Procedia computer science*, vol. 199, pp. 369–376, 2022.

[23] X. Yu, Z. Zhang, L. Wu, W. Pang, H. Chen, Z. Yu, and B. Li, "Deep ensemble learning for human action recognition in still images," *Complexity*, vol. 2020, 2020.

[24] A. Dey, A. Dutta, and S. Biswas, "Workoutnet: A deep learning model for the recognition of workout actions from still images," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2023, pp. 1–8.

[25] A. R. Siyal, Z. Bhutto, S. M. S. Shah, A. Iqbal, F. Mehmood, A. Hussain, and A. Saleem, "Still image-based human activity recognition with deep representations and residual learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, 2020.

[26] A. S. Saif, E. D. Wollega, and S. A. Kalevela, "Spatio-temporal features based human action recognition using convolutional long short-term deep neural network," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023.

[27] "Hii dataset [online]," https://zenodo.org/record/831923, Accessed August 19, 2023.

[28] A. Nader and D. Azar, "Evolution of activation functions: An empirical investigation," *ACM Trans. Evol. Learn. Optim.*, vol. 1, no. 2, jul 2021.

[29] X. Li, D. Chang, T. Tian, and J. Cao, "Large-margin regularized softmax cross-entropy loss," *IEEE Access*, vol. 7, pp. 19 572–19 578, 2019.

[30] "Stanford40 [online]," http://vision.stanford.edu/Datasets/40actions.html, Accessed August 19, 2023.

[31] T. Yang and Y. Ying, "Auc maximization in the era of big data and ai: A survey," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–37, 2022.

[32] G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of mcc and cen error measures in multi-class prediction," *PLOS ONE*, vol. 7, no. 8, pp. 1–8, 08 2012.

[33] K. Hirooka, M. A. M. Hasan, J. Shin, and A. Y. Srizon, "Ensembled transfer learning based multichannel attention networks for human activity recognition in still images," *IEEE Access*, vol. 10, pp. 47 051–47 062, 2022.