Expert Systems With Applications 266 (2025) 126147

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Human intention recognition using context relationships in complex scenes

Tong Tong, Rossitza Setchi [*], Yulia Hicks

Research Centre in AI, Robotics and Human-Machine Systems (IROHMS), Cardiff University, Cardiff CF24 3AA, UK

ARTICLE INFO

Keywords:
Human intention recognition
Dynamic scene graph dataset
Attention mechanism
Spatial temporal graph attention informer neural network
Context

ABSTRACT

Recognizing human intentions is a key challenge in human-robot interaction research. Much of the current work in this area centers on identifying human intentions within specific activities, often relying on a limited set of features. In contrast, this paper introduces a more versatile framework for intention recognition and introduces a novel model: the Spatial-Temporal Graph Attention Informer Neural Network (STGAIN). To recognize intentions, this model leverages spatial relationships between humans and objects in different scenes, along with their temporal evolution. In addition, to address an existing research gap, this research developed a new dataset called Dynamic Scene Graph (DSG) with representative dynamic relationships, derived from 471 videos covering 20 categories of human intentions. This dataset represents people and objects in different scenes, and the relationships between them. The model was tested rigorously at different points in the videos to track how the scenes evolved and to assess prediction accuracy, comparing the results to a range of advanced algorithms. Our findings clearly demonstrate that STGAIN outperforms these models, showcasing its potential for advanced human intention recognition applications. This model represents a significant advance toward creating more human-centered robots, capable of understanding and adapting to human intentions in real-world situations.

1. Introduction

Human intention recognition is the key to implementing advanced human robot interactions. It is an emerging field of research and application that seeks to understand the intricate cognitive processes underlying human behavior. Comprehending human intention, which pertains to the mental state driving human actions and decisions, has far-reaching implications across diverse domains such as technology, artificial intelligence (AI), psychology, social sciences, healthcare, and education (Liu et al., 2021). Recognizing human intentions plays a crucial role in social cognition (Selvakumar, 2024). The ability to recognize human intentions holds the potential to refine decision-making processes, optimize resource allocation, enhance user experiences, and deepen our understanding of human behavior in various settings. In recent years, there has been notable interest in human intention recognition and prediction due to the enormous capacity of artificial intelligence (AI) to analyze vast datasets, simulate decision-making processes, and utilize context-aware strategies.

Human intention recognition is both a challenging and high impact research domain within artificial intelligence. Human intentions are multifaceted, and it is not uncommon for individuals to have different intentions in similar circumstances, making the prediction of those intentions inherently complex. In that context, AI has an increasingly proactive role to play in human robot interactions (Li et al., 2022b).

Recognizing human intentions is pivotal for fostering seamless interactions between humans and robots (Tong et al., 2022). Proactive human-robot collaborations are the new norm, heavily dependent on anticipating human intentions to execute tasks. Numerous factors can influence human intention. In psychology, intention is the concept or plan an individual aims to implement (Bagheri et al., 2021). Although it involves recognizing actions, gestures, or gaze, intention is about comprehensively understanding context, i.e. the environment and the objects within it. Traditional computer vision models are good at identifying human actions and objects but often fall short in connecting these actions to the surrounding environment and objects (Kim et al., 2017).

Recognizing humans' next actions in different scenarios has always been a challenging problem (Kong & Fu, 2022). In this paper, human intention' specifically refers to what actions of individuals are likely to perform next. Our focus is on predicting these forthcoming actions to better understand and model human intention in different scenes.

Recognizing and predicting human intentions requires not only the analysis of fundamental human features, such as gaze and gestures, but also a thorough understanding of the surrounding environmental context, which is pivotal to this task (Singh et al., 2020). However,

* Corresponding author.
E-mail addresses: TongT6@Cardiff.ac.uk (T. Tong), Setchi@Cardiff.ac.uk (R. Setchi), HicksYA@Cardiff.ac.uk (Y. Hicks).

https://doi.org/10.1016/j.eswa.2024.126147
Received 20 September 2024; Received in revised form 14 November 2024; Accepted 9 December 2024
Available online 12 December 2024

current research on contextual integration remains limited, often lacking a comprehensive explanation of the interactions between humans and objects that can directly impact human intentions. Consequently, there is a need for incorporating richer contextual information to address this gap effectively.

This paper is focused on the non-verbal recognition of human intention by understanding the relationships between humans and objects. Therefore, the knowledge gap addressed in this research is the development of computational means for modelling spatial and temporal relationships between humans and objects and using them to predict human intention within a given context. An additional challenge is the lack of a suitable dataset for validation. In summary, the key objectives of this work are:

1) to develop a novel framework for human intention recognition, which is based on perception, context representation and contextual reasoning.
2) to create a new dataset for the needs of human intention recognition containing a variety of relationships between human and objects within videos representing different contexts.
3) to develop a novel model to learn both the spatial relationships between humans and objects, and the temporal features across successive video frames.

2. Related work

2.1. Human intention recognition

Human intention recognition has garnered significant attention among researchers, which is evidenced by recent studies focused on identifying human actions through hand-object interactions (Fan et al., 2021) and gaze for discerning human intentions (Singh et al., 2020). The value of predicting human intention is evident across a spectrum of studies, including determining human motion intentions in warehouses (Petković et al., 2018), intention recognition during engine assembly (Wang et al., 2018) and predicting pedestrian intentions in autonomous driving, which achieved 77 % accuracy (Yang et al., 2022).

In human-robot collaboration, accurate intention recognition is essential to minimize misunderstandings, reduce coordination errors, and enable robots to provide proactive assistance, particularly in environments where seamless interaction is paramount, such as manufacturing, healthcare, and service industries (Li et al., 2021). This capability is crucial to ensuring that robots can effectively support humans in complex and dynamic scenes.

Human intentions are influenced significantly by the interaction between people and objects. For instance, holding a pot in a kitchen setting generally does not suggest an intention to work. Consequently, understanding the relationship between individuals and objects within a scene is essential. Singh et al. utilized eye-view interactions with chess pieces to identify human intentions in a multiplayer board game, demonstrating the role of visual cues in intention recognition (Singh et al., 2020). Belardinelli approached this from a psychological perspective, emphasizing that gaze and actions are crucial in assessing human intentions. The specific objects with which individuals engage provide valuable clues for inferring intention (Belardinelli, 2024). Liu et al. identified human intentions through the spatiotemporal visual features of hand-object interactions with the F1 score of 0.25, further highlighting the role of such relationships in intent recognition (Liu et al., 2020). In human communication, understanding and predicting intentions often relies on contextual information derived from the environment or objects an individual interacts with (Özdel et al., 2024).

Additionally, existing research emphasizes the importance of human-centered features and environmental context in human intention recognition. Chen and Hou used gaze data captured through a head-mounted VR device to recognize human intentions in a scenario involving identifying and moving objects (Chen & Hou 2022). Similarly, Sabab et al. utilized real-time gaze data from a standard webcam, highlighting the importance of human-centered features, such as gaze, and the role of environmental context in predicting intentions (Sabab et al., 2022). In shared control, Jarrasse et al. demonstrated that object interactions provide valuable cues for recognizing human intentions in interactive locomotion tasks shared by robots and humans (Jarrassé et al., 2013).

Applications of intention recognition span beyond human–machine collaboration to areas like autonomous driving, where understanding drivers' and pedestrians' intentions enhances safety. Yang et al. fused RGB image sequences, semantic segmentation masks, and speed data to predict pedestrian intentions, noting that reliance solely on visual data can omit key real-time information (Yang et al., 2022). For drivers, Shangguan et al. used a long short-term memory (LSTM) neural network with time-series data to predict lane-change intentions, underscoring the role of temporal features (Shangguan et al., 2022). Xu et al. proposed an intent-driven Human-Object Interaction (iHOI) framework that leverages body joint-object distances and gaze to focus on contextual regions for recognition in weakly supervised settings (Xu et al., 2020). Collectively, these studies underscore the need for human intention recognition to incorporate person-object interactions and diverse features for improved predictive accuracy. To summarize, the study of human intention recognition holds paramount importance in the evolving landscape of artificial intelligence.

However, despite the existing body of research on human intention prediction, there remain certain challenges that need addressing: 1) Many studies target intentions within the context of narrowly defined specific scenarios, such as handovers of components and tools or actions such as standing up or sitting down. These specialized models often lack broad generalization capabilities, limiting their applicability in real-world settings. 2) The vast majority of the existing models center around individual interactions like gestures or gaze. In real life situations, human intentions are recognized from a multitude of features, particularly in complex scenarios. 3) There is an underutilization of spatial features. In practical contexts, spatial elements in the environment play a pivotal role in predicting human intentions. 4) Temporal features, too, are not maximally exploited. Human intentions exhibit continuity, with subsequent intentions largely influenced by prior states and contexts.

2.2. Scene graph

Scene graph generation plays a crucial role in various fields, including robotics, autonomous vehicles, and augmented reality. This technology interprets images by identifying objects and their interrelations, transforming them into structured visual representations. This process is vital for creating a detailed understanding of visual scenes, essential for numerous advanced applications (Robinson et al., 2015).

Graph-based datasets are particularly effective in this context. They model the relationships between entities (or nodes) in the scene, enhancing the learning process and providing greater interpretability. In these graphs, nodes represent not just individual objects but can encompass a range of entities. The connections or edges between these nodes depict the relationships, capturing various attributes and nuances. This allows for a more comprehensive use of spatial and temporal relationships within the scene (Kong et al., 2020). A notable contribution in this area is by Yang et al., who introduced a technique to extract relationships between objects in images using a Relation Proposal Network (Yang et al., 2018).

The primary aim of scene graph generation is to extract and understand the semantic information about the entities in an image and their relationships to one another (Jia et al., 2021). This data is very useful in applications such as image captioning, visual question answering, and image retrieval. The process begins with the identification of objects within the image, followed by an analysis of their relative positions and

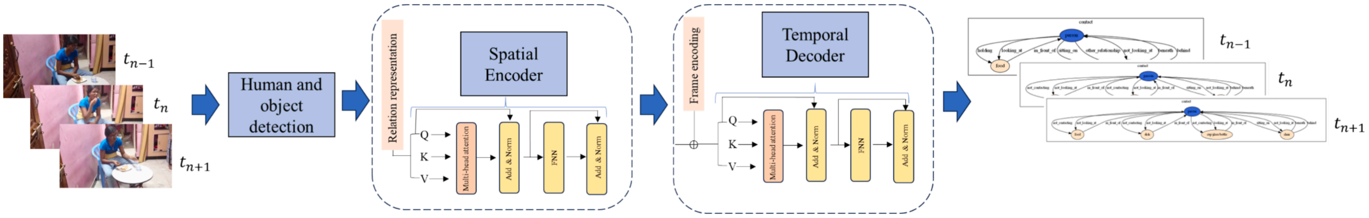

Fig. 1. Generation of the DSG dataset.

interactions. For instance, in a scene graph of a kitchen, the graph might illustrate the spatial relationships like "the refrigerator is to the left of the stove" or "the sink is in front of the window.".

The creation of scene graphs involves diverse methods, from simple rule-based systems to complex machine learning models. These models are trained on large, labeled image datasets to produce accurate and detailed scene graphs for new images (Xu et al., 2017). This is key to improving machines' understanding of visual contexts. However, generating scene graphs for complex scenes with many objects and relationships is challenging. Researchers use methods like pruning and clustering to simplify the graph (Xu et al., 2020). Another issue is unclear object relationships, like whether a person is holding a baseball bat or just leaning on it. To solve this, techniques for probabilistic inference have been developed, helping to determine the most likely relationships (Li et al., 2022a). Contrary to existing research, this study uses scene graphs to understand context and human intentions in complex situations.

2.3. Graph neural network

Graphs, with their nodes and edges, are great at showing relationships between entities. Traditional convolutional neural networks (CNNs) struggle with graph data, but graph neural networks (GNNs) excel in using these relationships, making them more interpretable. GNNs have been used for tasks like pedestrian intention recognition, as they can pick up subtle features missed by other methods, giving them better generalization (Ye & Ji, 2023).

The Graph Attention Network (GAT) is a standout type of GNN, they focus on important nodes to learn complex relationships and provide clear understanding through attention coefficients, showing the importance of neighboring nodes. They have been highly effective in node classification, link prediction, and recommendation systems (Wu et al., 2022). GATs are used in various areas like trajectory prediction (Li et al., 2022c), traffic flow forecasting (Wu et al., 2018), and social network analysis (Jiang et al., 2022). Their attention mechanisms highlight key neighboring nodes, aiding in learning detailed node relationships (Dong et al., 2022). A study uses heterogeneous graph attention neural networks to capture the underlying relationship between user travel choices and transportation reliability and influence human travel intentions. (Huang et al., 2024). This is particularly useful in human intention recognition, where GATs can better contextualize the role of the environment and objects in the execution of human actions.

2.4. Spatial temporal scene graph

Scene graph generation has evolved with the realization that single images are insufficient when considering temporal activities such as traffic flow prediction. This led to the development of spatial–temporal scene graphs (STSG), which use video footage for graph generation. For example, Wang et al. (Wang et al., 2020) used temporal graph neural networks for traffic prediction, focusing on temporal data relationships. Similarly, Yang et al. (Yang et al., 2020) advanced re-identification with temporal graphs, and Min et al. (Min et al., 2021) applied Spatial-Temporal Graph Neural Networks (STGNNs) to social network analysis, capturing spatial–temporal aspects. STGNNs are particularly effective in complex tasks like human intention recognition due to their ability to integrate both spatial and temporal data.

STGNNs have several extensions. The Spatial-Temporal Graph Attention Network (STGAT) (Huang et al., 2019) combines temporal and spatial attention, focusing on node proximities for spatial relationships. Other examples include STGCN (Han et al., 2021) for Point-of-interest (POI) recommendations and Dynamic Spatial–temporal Graph Recurrent Neural Networks to predict traffic flow, emphasizing the importance of dynamic space (Xia et al., 2024). Wu et al. proposed an architecture based on STGCN for human node analysis, leveraging the spatiotemporal dimension to achieve a deep and precise identification of human activities (Wu et al., 2023). To address human-robot collaboration (HRC) across diverse scenarios, Semeraro et al. applied LSTM and STGCN models independently to identify joint human activities, preprocessing the collected data in HRC settings to mitigate challenges in data generation (Semeraro et al., 2023). Intention recognition plays a vital role in collaborative assembly as well; Liu et al. utilized STGCN Plus (STGCNPP) as the core model, enhancing the stability of parallel time processing in human–machine collaboration, thereby improving task performance (Liu et al. 2023).

In essence, graph data, capturing both human-object relationships and their spatial–temporal dynamics, significantly enhances the modelling of human intention. The next section discusses how to generate a new graph dataset for human intent recognition to represent spatial and temporal dynamics in images or videos.

3. Dynamic scene graph generation

3.1. Dataset preparation

As emphasized above, a dataset combining both spatial and temporal information about humans and objects is required. However, there is no existing dataset, which represents the relationships between people and objects in videos and is suitable for human intention prediction. Previous attempts included annotating single images (Sigurdsson et al., 2016). However, predicting human intention through a single image was proven inadequate. To our best knowledge, our dataset is the first graph dataset, which was specifically developed for human intention recognition.

Our dynamic scene graph (DSG) dataset is a graph dataset which contains nodes (human and objects) and edges (relationships). It is based on the Action Genome (AG) dataset (Ji et al., 2020), which offers scene graph labels for individual frames and comprises annotations for 476,229 bounding boxes across 35object classes (excluding humans). Additionally, it includes 1,715,568 labelled relationships that span over 25 relationship classes across 234,253 frames. These relationships are segmented into three categories: (1) Attention relationships, which signify if a human's gaze is directed towards an object. (2) Positional relationships, which outline the positions between a human and objects. (3) Contact relationships that depict the 3 ways a human interacts with objects. The AG dataset covers many different relationships between humans and objects, making it exceptionally suited for human intention prediction.

Although the AG dataset is not used for human intention recognition, many of the scripts accompanying the videos indicate human intentions.

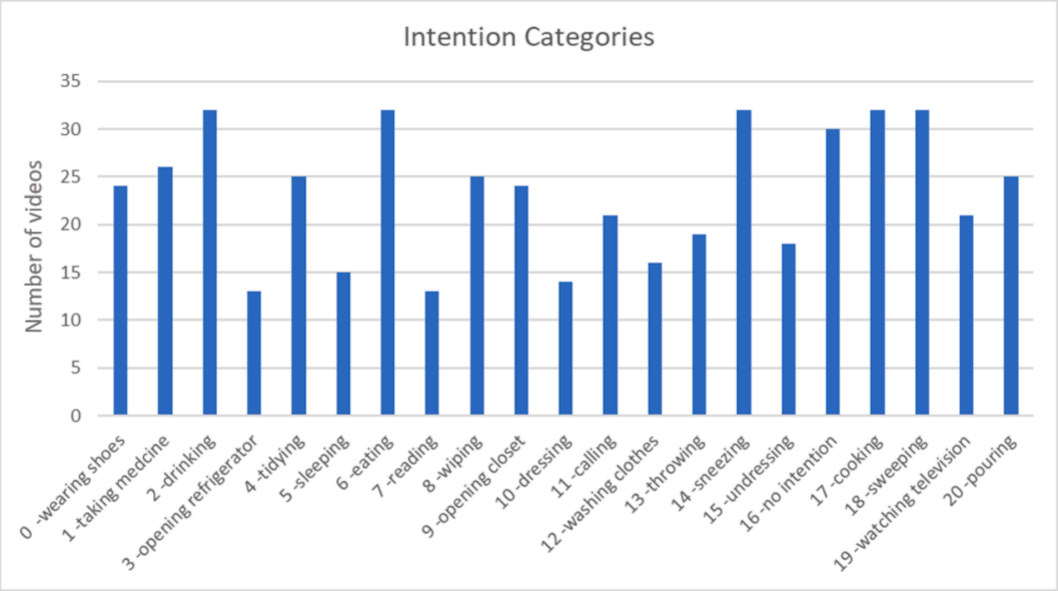

Fig. 2. Intention categories.

Further analysis of the videos showed that in many cases the human intention in the videos progresses into an action. Therefore, each of the selected videos were trimmed to ensure that the intended actions are removed from the videos. video frames without humans were also removed. Therefore, the graph dataset discussed next in this paper has been developed on the basis of the trimmed videos.

3.2. Dynamic scene graph generation

Our approach to the dynamic scene graph generation was inspired by the spatial–temporal transformer (STTran), which is based on the method by Cong et al. (Cong et al., 2021).

The procedure of generating a dynamic scene graph includes several steps (Fig. 1). The first step entails deploying algorithms for the detection of objects and humans to identify and classify these entities within each frame of the video. Using the information gathered, an elementary scene graph is formed, wherein nodes symbolize humans and objects, and edges represent their relationships. To incorporate the temporal dynamics, the nodes are interconnected across successive frames, which serves to record the interactions of the humans with the objects. Next, the multi-head attention transformer is used to extract the spatial features. The input X, which is represented by each frame, $X \in \mathbb{R}^{(}N \times D)$ that has N entries of D dimensions, is transformed into queries $\left(Q = XW_Q,\ W_Q \in \mathbb{R}^{(}D \times\ D_q)\right)$, keys $(K = XW_K, W_K \in \mathbb{R}^{(}D \times D_k))$ and values $(V = XW_V, W_V \in \mathbb{R}^{(}D \times D_v)$ through linear transformations. W_Q, W_K, W_V are trainable parameter matrices which add the input to get the Q,K,V. Each entry is influenced by other entries through the attention defined by:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(QK^T \Big/ \sqrt{D_k}\right) \tag{1}$$

This transformer layer can explore the frame level, which works on a single frame, and temporal dependencies that work on a sequence, respectively.

The next step involves using Faster R-CNN (Ren et al., 2017) to extract relationship representation for each frame in a given video at time t. Next, the relationship representations exchange spatial and temporal information. The spatial information and semantic embeddings are formulated as:

$$x_t^k = \left\langle \left(W_s v_t^i, W_o v_t^j, W_{u\varphi}\left(u_t^{ij} \oplus f_{box}\left(b_t^i, b_t^j,\right)\right)\right), s_t^i, s_t^j \right\rangle \tag{2}$$

where $\langle . \rangle$ is concatenation operation, φ is flattening operation and ⊕ is addition.Ws, Wo, and Wu represent the linear matrices for dimension compression, u_t^{ij} indicates the feature map of the union box, while f_{box} is the function transforming the bounding boxes of subject and object to an entire feature. The semantic embedding vectors s_t^i, s_t^j are determined by the object categories of subject and object.

Finally, a spatial encoder, frame encoder and temporal decoder are used to generate the dynamic graph. The spatial encoder concatenates the two frames next to each other. The frame encoder concatenates every frame, while the temporal decoder classifies each frame and generates the scene graph. The spatial encoder is formulated as follows:

$$X_t^{(n)} = Att_{enc.}\left(Q = K = V = X_t^{(n-1)}\right) \tag{3}$$

where X_t is a single input after spatial encoding, and the queries Q, keys K, values V share the same input and output of n encoder layer. To be more specific, $X_t^{(n)}$ is the output of $X_t^{(n-1)}$ through the attention encoding.

For capturing temporal dependencies between the video frames, it uses a sliding window η approach to batch frames, ensuring only adjacent frames interact, reducing irrelevant information. The contextualized representations $[X1, ..., XT]$ and the $i-th$ generated input batch is presented as:

$$Z_i = [X_i, ..., X_{i+\eta-1}], i \in \{1, ..., T - \eta + 1\} \tag{4}$$

where T is the video length. Similar to the encoder, the decoder consists of the self-attention layer $Q\ = K\ = Z_i + E_f$, $V\ = Z_i$, $\widehat{Z_i}\ = Att_{dec.}(Q,K,V)$, where E_f is the frame encoding. This completes the task of generating spatial temporal scene graphs.

3.3. Dataset summary

This paper is focused on human intention recognition before the intended action has been executed by a human. For example, when a person is detected holding a cup and drinking water, the act of drinking water is not regarded as intention, because this action has already been executed. Human intention should be recognized before the person starts drinking. The DSG dataset created in this article is aimed at different scenes. The relationships between humans and objects are represented through a spatial temporal graph, and then processed through the model proposed. The graph data labels are automatically

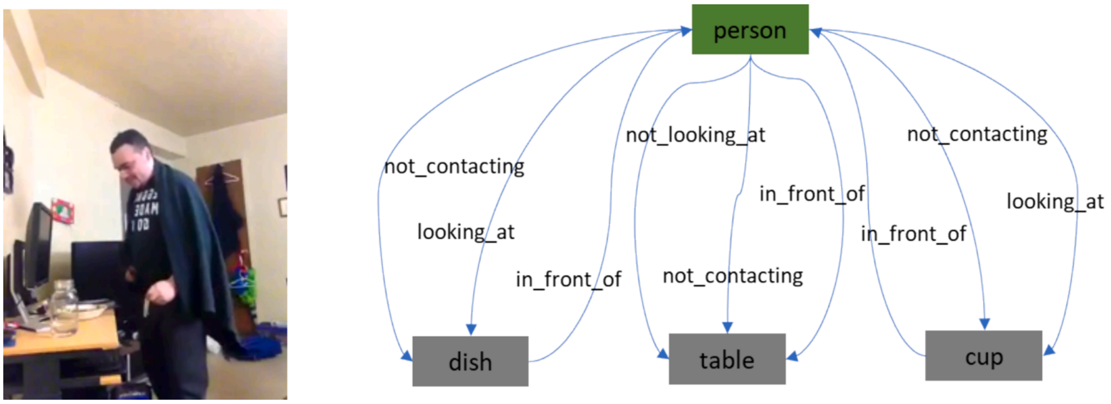

Fig. 3. An illustrative example based on one frame and its corresponding graph scene. The intention indicated in the script is to drink.

generated using the descriptive scripts of the original dataset. The scene graph generation algorithm used to generate graph relationships and labels between people and objects, achieved an accuracy of 82 % (Cong et al., 2021). However, it is important to acknowledge that this limitation affects the generalization capability of the proposed method. In the future, Generative AI (GenAI) could be utilized to enhance generalization. As highlighted by Ye et al. (2024), GenAI has the potential to address challenges associated with complex graph structures, such as knowledge graphs and scene graphs (Ye et al., 2024). The method proposed is designed to effectively handle varying graph structures, enhancing flexibility across scenarios and improving the generalization capability of human intention recognition.

All videos in the dataset are accompanied by a script, which provides a clear indication of the human intention. For example, in the narrative “A person stands in front of a cabinet and starts taking medicine”, the word “starts” indicates a clear intention, which can help to filter out videos without clear human intentions. The intentions are extracted by capturing the gerund after “start” or “begin”. As shown in Fig. 2. the final dataset contains 20 categories of intention, such as taking medicine, drinking, opening refrigerator, etc. Fig. 3 shows an example frame and its graph in the dataset.

The dataset compiled includes many scenarios in a home environment, and is not limited to kitchens, bedrooms, and bathrooms. Interestingly, despite the diversity in scenarios, the same intention such as drinking can be observed both in the kitchen and bedroom. Conversely, a similar scenario may reveal different intentions. For instance, in the kitchen, one individual may have the intention to drink, while another may have the intention to tidy up. This complexity enhances the versatility of the dataset, enabling the model to effectively generalize across some wider contexts.

In reality, different individuals exhibit divergent habits. To illustrate, when considering the intention of sleeping, the variations are abundant. While some individuals express a preference for resting in a bed, others may opt for a sofa or even a floor. Consequently, even within a single scene and same intention, it is also possible that internal features in the dataset are different. This consideration significantly contributes to the generic nature of the dataset, enhancing its applicability across diverse contexts.

Furthermore, the dataset contains instances where similar nodes and edges exist within a scenario, but it would generate different intentions. As an illustration, consider a kitchen setting represented by three nodes: individual, refrigerator, and food. Despite the same edge characteristics, the intention can differ, such as the intention to open the refrigerator or eat. This complexity enhances the opportunity for practical implications.

In summary, the DSG dataset that was created contains 20 different intention categories, 27 edge attributes and 39 object types. The categories of intentions include eating, drinking, cooking, pouring water, wiping, tidying, etc. The relationship between people and objects includes looking at/not looking at, contacting/not contacting, holding, in

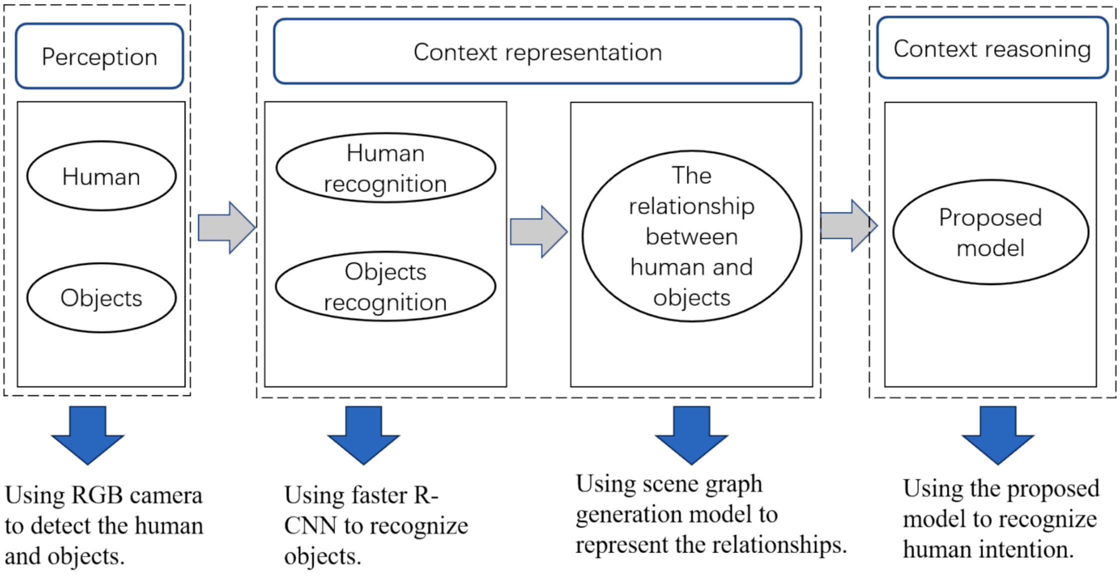
Perception
Human
Objects
Context representation
Human recognition
Objects recognition
The relationship between human and objects
Context reasoning
Proposed model
Using RGB camera to detect the human and objects.
Using faster R-CNN to recognize objects.
Using scene graph generation model to represent the relationships.
Using the proposed model to recognize human intention.

Fig. 4. Human intention recognition framework.

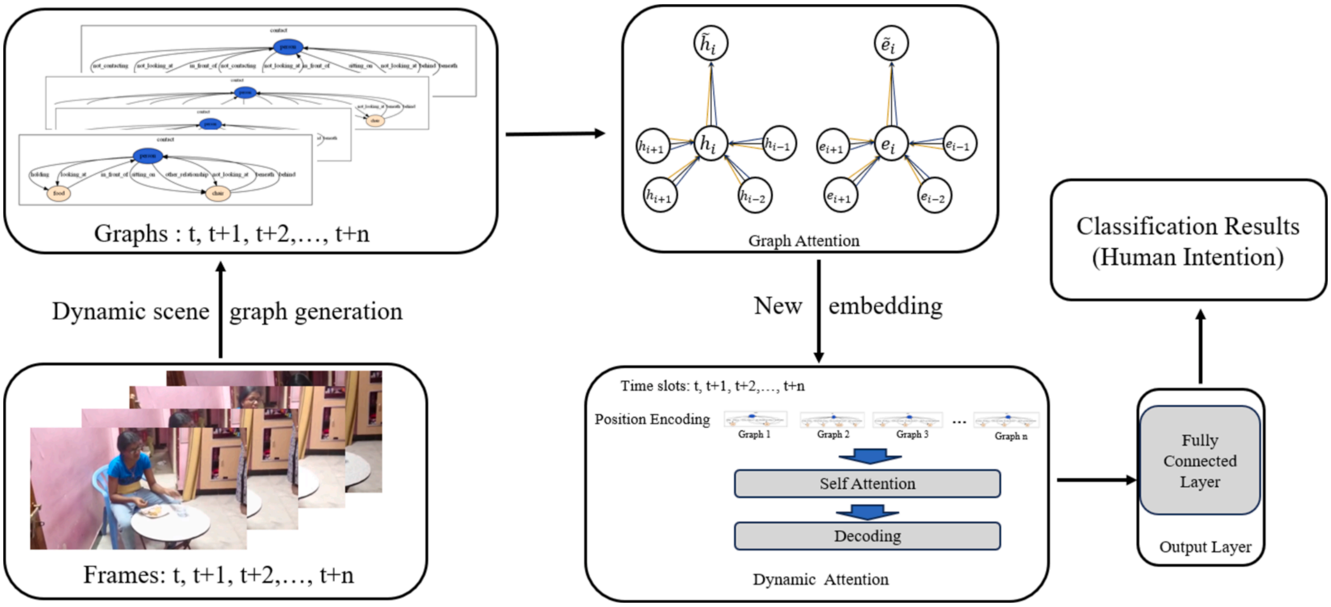

Fig. 5. The structure of STGAIN. Initially, the video is decomposed into individual frames. A graph representation is generated for each frame and subsequently processed through a Graph Attention Network to re-embed its nodes and edges, resulting in encodings that incorporate attention coefficients. These new graph representations are then endowed with positional encoding, allowing temporal features to be learned via a self-attention mechanism. Finally, the fully connected layer is employed to classify human intentions.

front of, behind, etc. Objects are object recognition pre-trained based on the COCO dataset. It is extracted from 471 videos; each video contains 3–8 frames, and each frame contains 3–15 edges. The data type is graph data because it can better express and visualize the relationship between people and objects. The DSG was developed to ensure that the predictions of the model are focused specifically on human intentions, and not actions in general.

4. Spatial temporal graph attention informer neural network

4.1. Human intention recognition framework

The proposed human intention recognition framework is shown in Fig. 4. It is composed of three main modules. The perception module is used to detect and classify humans and objects in each frame. Subsequently, a spatial–temporal scene graph is generated based on the video, forming context representations. They represent the human and object, and the relationship between them, which can support the model to learn the spatial and temporal features. Finally, the context employed by the spatial temporal graph attention informer neural network (STGAIN) to learn features of the nodes and edges in the generated graph.

The perception module includes detecting as many objects surrounding the human as possible because the method relies on having rich contextual information. The context representation module uses Faster R-CNN to recognize these objects as candidates for the DSG dataset. The relationships between humans and objects are important and may include a rich set of modalities such as gaze, position, and physical touch/contact. Using different modalities has an important impact on disambiguation. The context reasoning module uses the proposed Spatial-Temporal Graph Attention Informer Neural Network (STGAIN) model developed in this research. STGAIN is a type of graph neural network that has been specifically designed for processing DSG datasets for intention recognition. It models complex relationships in a spatial–temporal graph, like objects or human actions, which evolve over time.

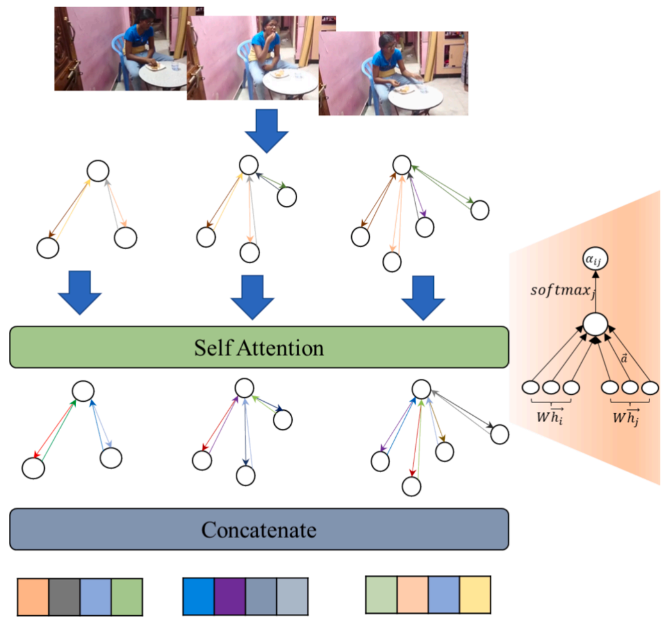

Fig. 6. The spatial part of the STGAIN. It can learn the structural information from the graph data produced in each frame. By employing graph networks with attention mechanisms, the system learns from the nodes and edges within the graph structure, subsequently deriving novel encodings. The different colors represent the new embedding after calculating the attention coefficients. These encodings are then reassembled to form a new representation of the graph structure.

4.2. STGAIN model

The proposed Spatial-temporal Graph Attention Informer Neural Network (STGAIN) is designed for graph-structured data. As shown in Fig. 5, STGAIN includes a spatial part and temporal part. The nodes and edges learn the graph features, whilst the temporal part learns the time features in the consequent frames.

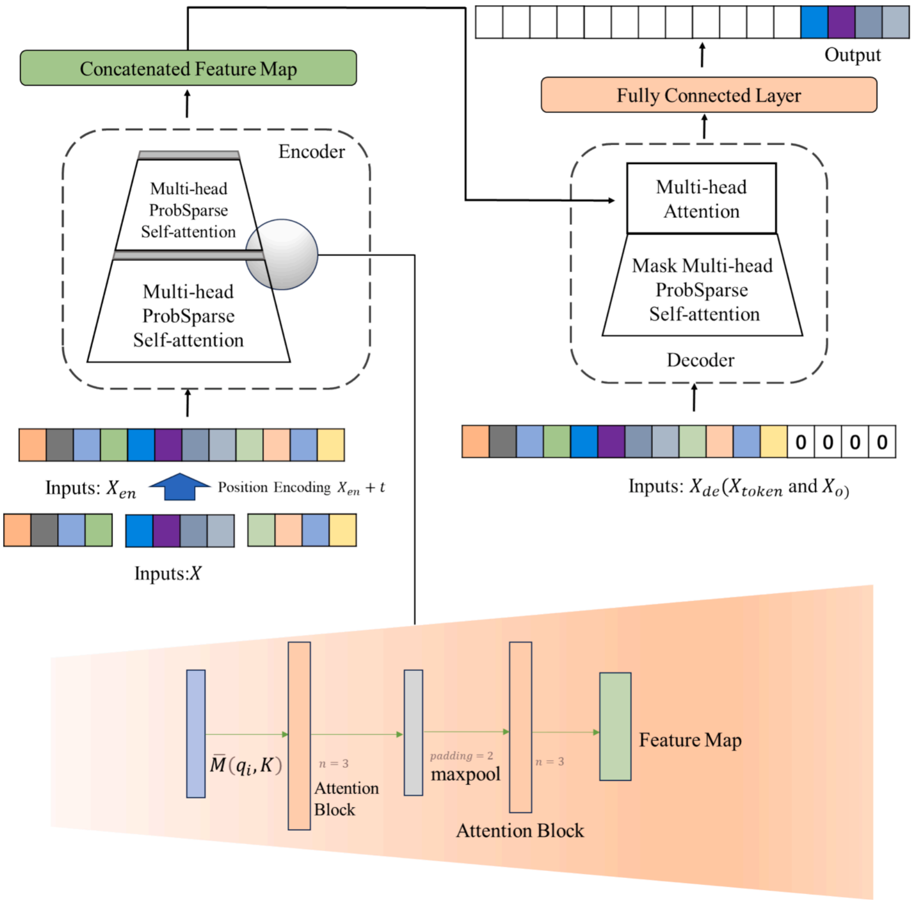

Fig. 7. The temporal part of the STGAIN. The features of successive frames were treated as a temporal sequence. Each frame undergoes positional encoding to embed temporal order characteristics. Through the encoding and decoding process involving multi-head attention mechanisms, the algorithm assimilates the temporal features of the data.

4.2.1. Spatial graph

Fig. 6 shows the spatial part of STGAIN which uses the self-attention mechanism that allows the model to learn to attend selectively to different nodes in the graph, capturing more important nodes and edges (Veličković et al., 2018). It The self-attention mechanism allows the model to capture complex relationships between nodes in the graph, including both global and local information. In the spatial part of the STGAIN algorithm, each node in the graph is represented as a feature vector. The feature vector of each node is passed through a multi-layer perceptron (MLP) to produce a hidden representation. The hidden representation of each node is then used to compute attention coefficients for each node in the graph.

$$\alpha_{ij} = exp\left(LeakyReLU\left(\vec{a}^T\left[W\vec{h}_i \| W\vec{h}_j\right]\right)\right) \Big/ \sum\nolimits_{k\in N_i} exp\left(LeakyReLU\left(\vec{a}^T\left[W\vec{h}_i \| W\vec{h}_k\right]\right)\right) \tag{5}$$

The attention coefficients (Equation (5) are computed based on the dot product between the hidden representation of each node and a learnable parameter vector. W is the training weight. || is used to denote a concatenation of two vectors. $\vec{a}$ is used for transforming the size of vector. The attention coefficients are then normalized using a softmax function to ensure that they sum to one. The normalized attention coefficients are used to compute a weighted sum of the hidden representations of the nodes in the graph. The weighted sum is the output of the algorithm. Then, it will update the nodes from $\widetilde{h}_i$ to $\widetilde{h}_j$ to aggregate the features applying a nonlinearity, σ:

$$\widetilde{h}_i = \sigma\left(\sum\nolimits_{j\in N_i}\alpha_{ij}W\widetilde{h}_j\right) \tag{6}$$

The use of multi-head attention as an extension of our self-attention mechanism is advantageous in ensuring a more stable learning process. In particular, K distinct attention mechanisms carry out the transformation in Equation (7). Subsequently, the features they produce are joined together, leading to the generation of the following output feature representation:

$$\widetilde{h}_i = \|_{k=1}^{K}\sigma\left(\sum\nolimits_{j\in N_i}\alpha_{ij}^k W^k\widetilde{h}_j\right) \tag{7}$$

The edges are also very important in a human and objects relationships dataset. The attention coefficients are calculated from the graph

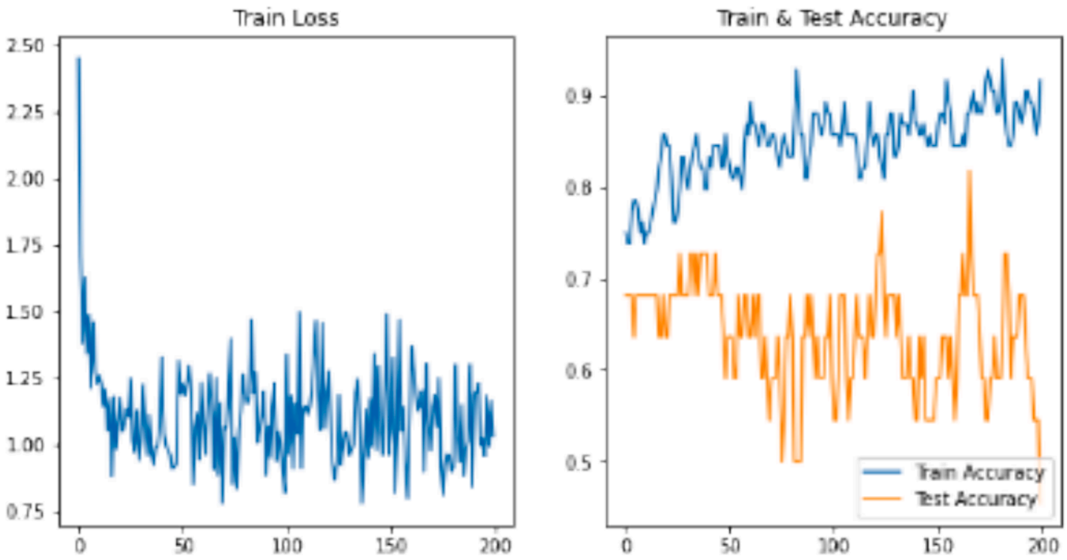

Fig. 8. Calculating without the edge coefficients.

and embedded in the edges.

$$e_{ij} = exp\left(LeakyReLU\left(\overrightarrow{b}^T\left[W\overrightarrow{e}_i \| W\overrightarrow{e}_j\right]\right)\right) \Big/ \sum\nolimits_{k \in N_i} exp\left(LeakyReLU\left(\overrightarrow{b}^T\left[W\overrightarrow{e}_i \| W\overrightarrow{e}_k\right]\right)\right) \tag{8}$$

Where W is the training weight, $\|$ is used for concatenating two vectors, and $\overrightarrow{b}$ is used for transforming the size of vector.

In conclusion, the spatial part algorithm of STGAIN is a useful tool for analyzing graph-structured data. By using a self-attention mechanism, this algorithm can capture both global and local information in the graph. It has the potential to transform many fields that rely on graph-structured data. In the case of the dataset of human objects relationships developed in this research, it can learn the graph nodes and edges features using the self-attention mechanism to calculate the attention coefficient from the neighbor nodes and edges.

4.2.2. Dynamic graph

The temporal part of STGAIN (Fig. 7) is a deep learning model designed for time series forecasting that can handle irregularly sampled data. The temporal part can handle missing values, variable-length sequences, and irregularly sampled data. The temporal part is based on a transformer architecture that includes an encoder-decoder framework with a self-attention mechanism (Zhou et al., 2021). The self-attention mechanism allows the model to capture long-term dependencies and relationships between different time steps in the time series data.

Since our graph data is a sequence of coherent frames, represented by a different vector in the spatial part, we use position encoding to encode each frame so that the algorithm can learn the temporal features.

The self-attention mechanism operates by performing a scaled dot-product on the queries, keys, and values of all nodes. It then computes attention scores for each position by applying the softmax function to the compiled data. The attention is determined in the following manner:

$$A(Q,K,V) = Softmax\left(\overline{Q}K^T \Big/ \sqrt{d}\right)V \tag{9}$$

Where the $Q \in \mathbb{R}^{l_Q \times d_{in}}$, $K \in \mathbb{R}^{l_K \times d_{in}}$, and d_{in} is the input dimension. If q_i, k_i, v_i represent the i_{th} row in Q, K, V, then the i_{th} query attention is:

$$A(q_i,K,V) = \sum_j \left(k(q_i,k_i) \Big/ \sum\nolimits_l k(q_i,k_i)\right) v_j = \sum\nolimits_j p(k_j|q_i) v_j$$
$$= E_{p(k_j|q_i)}\left[v_j\right] \tag{10}$$

Each frame has a q and k value, which is derived from the attention mechanism. Although dot multiplication is required for each frame, in reality, not every frame is important, so important frames need to be selected for dot multiplication to improve prediction accuracy and calculation efficiency. Query's attention towards other keys is notated as a probability $p(k_j|q_i)$. If the $p(k_j|q_i)$ more like the $\frac{1}{l_K}$, it should be the element which can be ignored.

$$KL(q\|p) = \ln \sum_{j=1}^{l_K} e^{(q_i k_l^T / \sqrt{d_{in}})} - \frac{1}{l_K} \sum_{j=1}^{l_K} (q_i K_j^T \Big/ \sqrt{d_{in}}) - lnl_K \tag{11}$$

Dropping the constant, the i-th query's sparsity measurement is:

$$M(q_i,K) = \ln \sum_{j=1}^{l_K} e^{(q_i k_l^T / \sqrt{d_{in}})} - \frac{1}{l_K} \sum_{j=1}^{l_K} (q_i K_j^T \Big/ \sqrt{d_{in}}) \tag{12}$$

If the i-th query is the $M(q_i,K)$, it should be the more important frame.

The important part in the temporal part is ProbSparse, which saves the computing resources for the attention. ProbSparse removes the insignificant values and then calculates the attention coefficient. It will randomly sample some key in the K, and calculate the (q,K) to reduce the computing.

In fact, calculating each M() is more complicated. Therefore, there is $\ln l_k \leq M(q_i,K) \leq \max_j(q_i K_j^T / \sqrt{d_{in}}) - \frac{1}{l_k}\sum_{j=1}^{l_k}(q_i K_j^T / \sqrt{d_{in}}) + \ln l_k$, so the max mean measurement is:

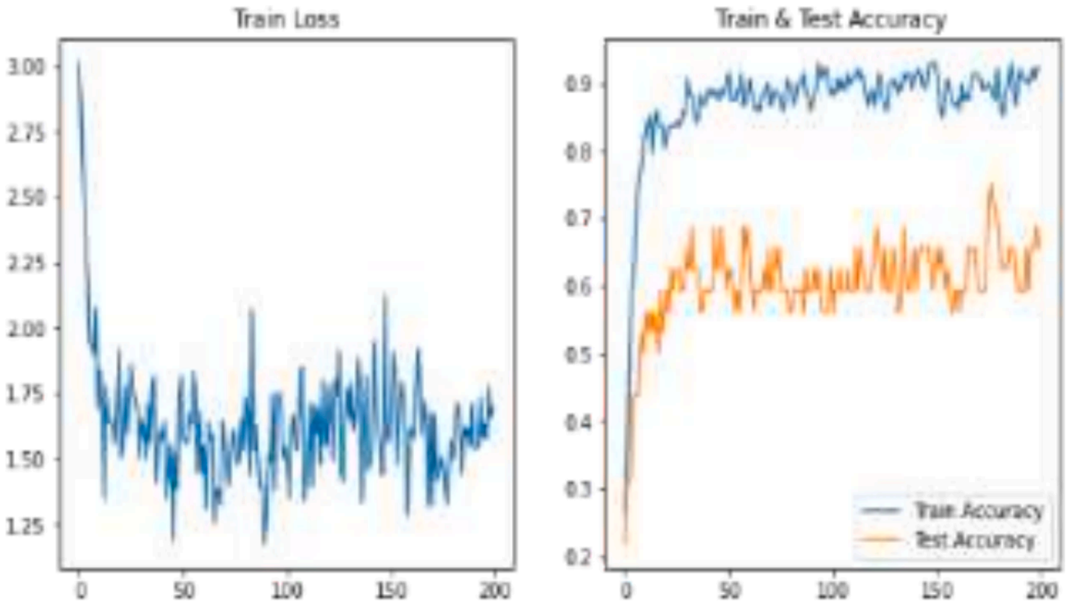

Fig. 9. Calculating with the edge coefficients.

$$\overline{M}(q_i, K) = \max_j \left\{ q_i k_j^T \Big/ \sqrt{d} \right\} - \frac{1}{L_K} \sum_{j=1}^{L_K} (q_i k_j^T \Big/ \sqrt{d}) \tag{13}$$

We have the ProbSparse self-attention by allowing each key to only attend to the dominant queries:

$$A(Q, K, V) = Softmax\left(\overline{Q}K^T \Big/ \sqrt{d}\right)V \tag{14}$$

where $\overline{Q}$ is a sparse matrix that matches the size of Q and it only includes the top-u queries as determined by the sparsity measure $M(q, K)$.

$$X_{j+1}^t = MaxPool\left(ELU\left(Conv1d\left(\left[x_j^t\right]_{AB}\right)\right)\right) \tag{15}$$

where X_{j+1}^t is the output of the multi-headed ProbSparse self-attention layer in this layer; $\left[x_j^t\right]_{AB}$ is the calculation result of the multi-headed ProbSparse self-attention layer in the previous layer; ELU is the activation function.

For fast prediction, we put the input X_{de} into the decoder layer.

$$X_{de}^t = Concat\left(X_{token}^t, X_0^t\right) \tag{16}$$

For the decoder layer, where X_0 is a placeholder of the target to fill the values with zeros, X_{token} is the start input which is same as the encoder layer. A fully connected layer acquires the final output. Till now, the kernel of temporal part has been created.

Our algorithm does not forget past frames and can handle information of different lengths due to position encoding. This fits the requirements of our dataset.

The temporal part of the STGAIN algorithm uses the Prob self-attention mechanism. ProbSparse self-attention is based on the proposed measurement. As stated, one of the key features of the temporal part of STGAIN algorithm is its ability to handle missing data. In many real-world scenarios, time series data may have missing values due to sensor failures, data transmission errors, or other reasons.

5. Evaluation

5.1. Experiment setting

The metrics used to test the model are recognition accuracy, loss, m-F1 score and confusion matrix. To bolster the reliability of the validation process, we implemented a strategy of randomly partitioning the dataset into distinct sets of training (80 %) and testing (20 %) data five separate times. This approach ensures variability in each training–testing split, providing a more robust test of the model's predictive power. In addition, we ensured that the training and testing data include all categories of intention. The final accuracy is determined by calculating the average of the results.

Cross-Entropy (CEL) and NLL are used to calculate the loss values.

Table 1
Evaluation results.

Method	Cross-entropy loss			NLL loss		
	Loss value	Test Acc	m-F1	Loss value	Test Acc	m-F1
STGCN	1.75	0.70	0.69	1.72	0.66	0.64
STGAT	1.73	0.75	0.73	1.67	0.72	0.71
STGAIN	1.74	0.81	0.82	1.66	0.77	0.76

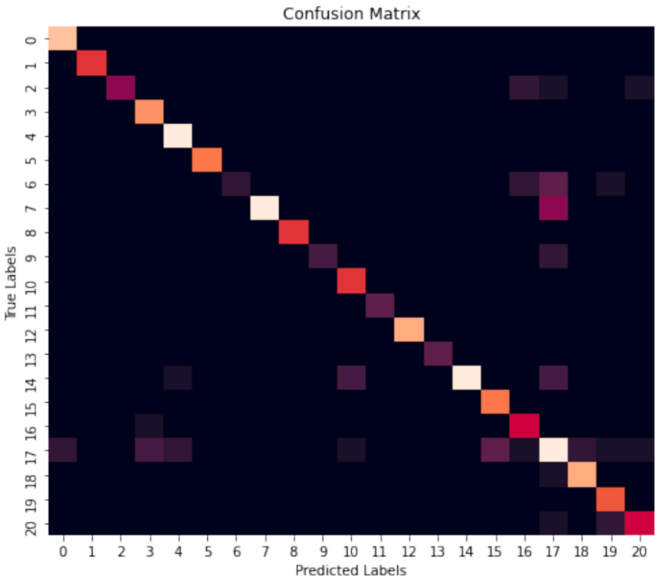

Fig.10. Confusion matrix.

Table 2
Confidence interval (CI) results.

Model	Epoch 30		Epoch 90		Epoch 120		Epoch 150		Epoch 180	
	Test Acc (%)	95 % CI (%)	Test Acc (%)	95 % CI (%)	Test Acc (%)	95 % CI (%)	Test Acc (%)	95 % CI (%)	Test Acc (%)	95 % CI (%)
STGAIN	70.54	[65.68, 76.64]	73.86	[69.25, 77.56]	77.63	[74.93, 80.05]	79.39	[77.10, 81.59]	81.37	[79.04, 82.98]

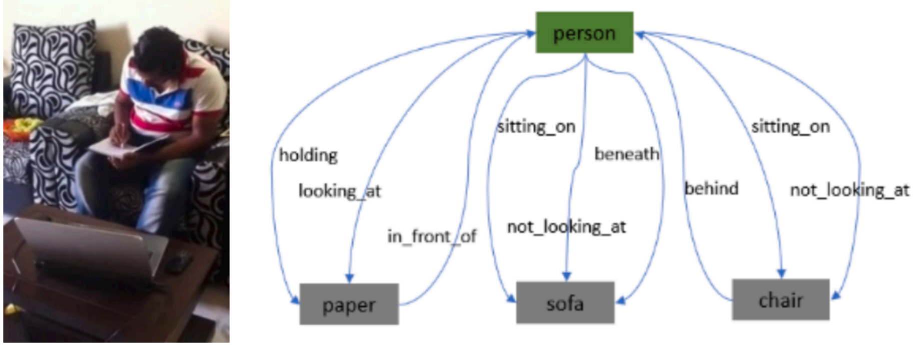

Fig. 11. First frame and its corresponding graph.

Furthermore, two experiments are set up to determine whether it is beneficial to calculate the attention coefficients of the edges. This is used to prove the importance of the relationship between human and objects in human intentions prediction. In addition, the model developed was compared with STGCN and STGAT, which represent the existing state-of-the-art in temporal graph networks.

Furthermore, the Cosine Annealing Scheduler was employed to dynamically adjust the learning rate during training, alongside an early stopping mechanism to prevent overfitting. Data augmentation techniques were applied to limit the impact of data imbalance on results. These included random edge dropping, noise addition, and various other augmentations, such as node feature perturbation, edge perturbation, and random swaps, even considering graph structure perturbations to enhance the model's robustness.

5.2. Experimental results

Fig. 8 and Fig. 9 show the comparison of the graph classification with and without edge coefficients, which correspond to the relationships between human and objects.

The comparison clearly demonstrates that when edge coefficient calculations (relationships between human and objects) are incorporated, the accuracy of classification is notably superior compared to scenarios where edge calculations are omitted. This proves our hypothesis, underscoring the significance of the relationships between humans and objects in the recognition of human intentions.

Table 1 shows the results of the evaluation conducted, which compares the performance of the proposed model STGAIN with two other advanced models, STGCN and STGAT. The results indicate that STGAIN achieved an average accuracy rate of 0.81, showing a better performance in comparison to the other methods. The m-F1 score achieved of 0.82 is good, especially for multi-class classification tasks. This means that the model's performance is relatively balanced, and the model is able to take into account predictions of different categories. The primary distinctions of the new model STGAIN, proposed in this paper lie in their core mechanisms, despite STGCN and STGAT are also designed for spatiotemporal graph data. The model introduced here applies self-attention mechanisms to both nodes and edges in spatial dimensions. Edges represent the crucial relationships between people and objects, which is an essential aspect for accurate human intention recognition. This approach enhances the model's ability to capture and prioritize significant information effectively. Furthermore, in handling temporal features, this paper incorporates position encoding and the informer architecture, which allows the model to preserve long-range dependencies and therefore more accurately learn key frames within sequences. This design enables flexible handling of sequences of varying lengths, increasing the model's adaptability to different data requirements.

Fig. 10 indicates a confusion between intention No. 2 (drinking), No. 20 (pouring) and No. 17 (cooking). This ambiguity predominantly arises

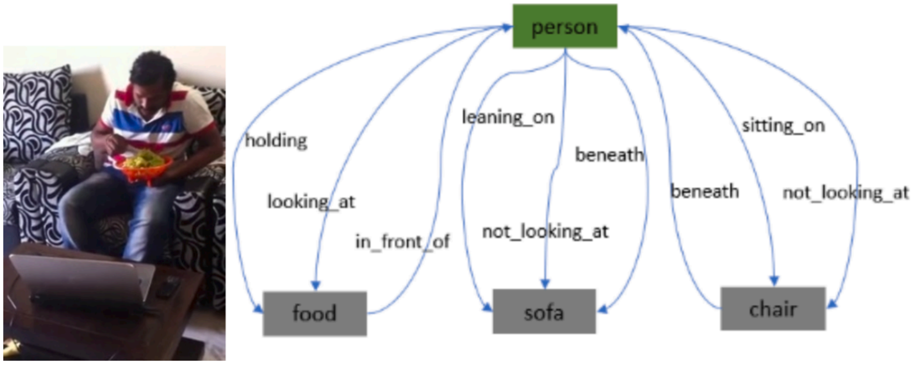

Fig. 12. Fourth frame and its corresponding graph.

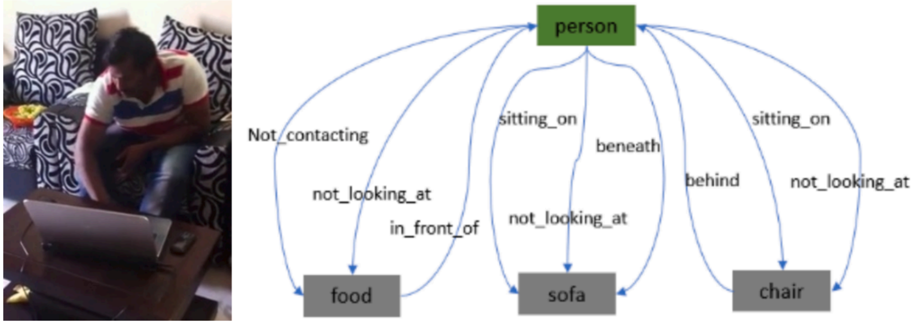

Fig. 13. Eighth frame and its corresponding graph.

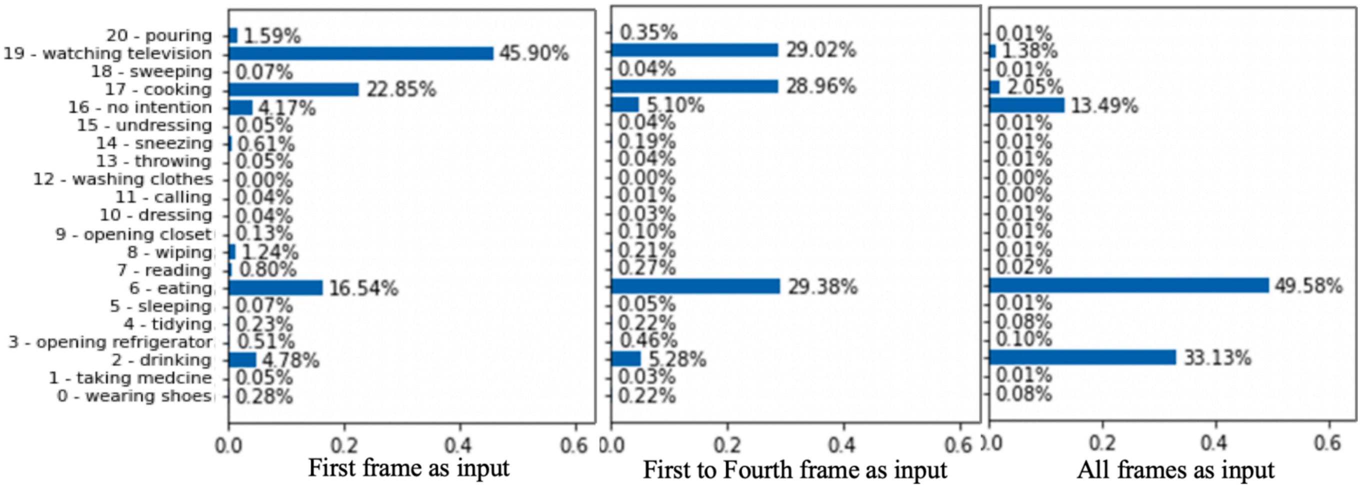

Fig. 14. Comparison of prediction probabilities for different number of frames.

Table 3
Evaluation results.

Model	$\lambda = 50\%$	$\lambda = 75\%$	$\lambda = 100\%$
STGCN	57.66 %	65.37 %	69.18 %
STGAT	**62.44 %**	68.40 %	75.94 %
STGAIN	60.31 %	**73.81 %**	**81.26 %**

due to the involvement of a common relational aspect between the individual and the cup in both scenarios. Additionally, there is a notable misinterpretation occurring between intention No. 6 (eating), and intentions No. 17 (cooking) and No. 19 (watching television). Such confusion is attributable to the recurrent incorporation of food in intentions 17 and 19. This is part of the challenge of dealing with more complex scenarios.

To address the uncertainty in the experimental results, the model performance was evaluated using confidence intervals. It can be seen from Table 2 that the confidence interval is wider in the early stage, such as [65.68, 76.64] and narrowed in the later stage, such as [79.04, 82.98]. This shows that as training proceeds, the model's performance on the test set gradually stabilizes. It also proves that the accuracy of the model falls within this range in most cases, which is an ideal situation.

5.3. Analysis of the results

For a better understanding of the results of the model, we have conducted an in-depth analysis of the results. Specifically, our objective is to discern the variations in the model's predictions. To illustrate this, we have selected an intention (eating) as a representative example. Since CEL loss function performance is better, the following experiments all use CEL as the loss function.

This chosen video comprises a total of eight frames. To scrutinize the variations in prediction, we sequentially input the initial frame, the range from the first to the fourth frame, and the range from the first to the eighth frame (all frames). Our objective is to investigate the fluctuations in predictions throughout the process. In other words, we aim to examine if and at what point the predicted outcomes undergo changes.

Figs. 11-13 show example frames of a video from the dataset; they include the initial, fourth, and eighth frames and the corresponding graphs generated. The figures show clearly the pertinent nodes and edges of these graphs.

As indicated by the script, the ground truth is eating intention (No. 6), but the prediction after the first frame is No. 19 (Fig. 14). Further analysis helps to understand this inaccurate prediction. As shown in Fig. 11, the nodes include person, sofa, and the relationships are sitting_on, not_looking_at, etc. Watching (No. 19) and eating (No.6) both involve these relationships. The probability associated with the intention eating is not high initially due to the absence of any detected food object in this frame. This results in an inaccurate prediction on the basis of a single frame. However, when the first four frames are taken into consideration, the outcome is correct, though the probability is not high. In these frames, food has been detected. The relationships between the food and the individual for the duration of these frames are "not_looking_at" and "not_contacting". Furthermore, numerous videos depicting the intention of watching (No. 19) incorporate food. This is why the probabilities of these two intentions are very close. When all frames of the videos are used, the result is notably more accurate. The relationships between the food and the individual are "looking_at" and "holding". Both of these relationships are pivotal for the intention eating. This is accompanied by a sudden increase in the probability of intention No. 2

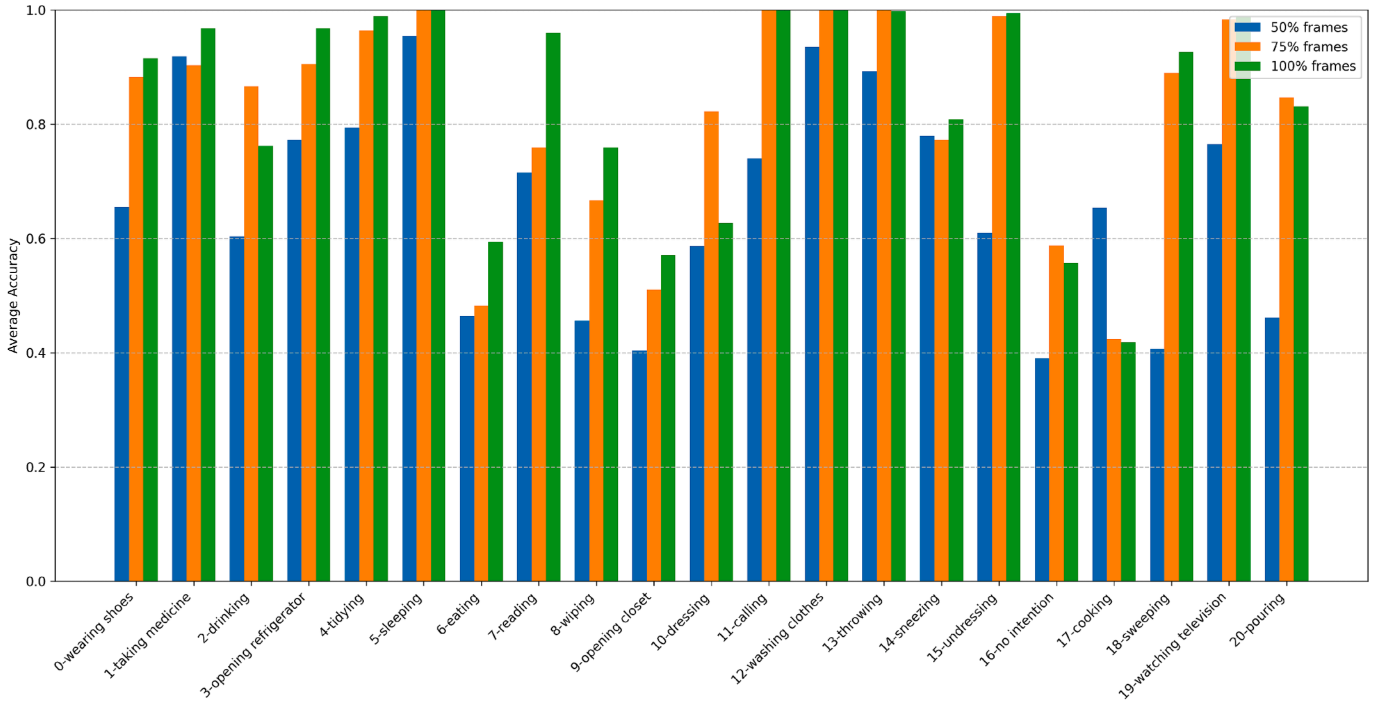

Fig. 15. Overall Prediction Accuracy.

(drinking), which is understandable since food is commonly observed during drinking.

To evaluate our model and observe the changes in the final recognition results, we used as input different number of frames (e.g 50 %, 75 %, and 100 %).

Table 3 shows that the test accuracy for different frames using three models (STGAIN in comparison with STGCN and STGAT), where λ represents the percentage of video frames used as input. As expected, the model improves its performance as more frames are used. As observed in Fig. 15, using only 50 % of the data frames results in inferior prediction outcomes compared to utilizing a larger number of frames. There is a noticeable improvement in prediction performance as the number of data frames increases.

The availability of a larger number of data frames is crucial. It enables the model to learn more critical features such as the more significant nodes (objects) and edges (relationships), contributing positively to the learning of spatial and temporal features. This is the advantage of our model. Moreover, more frame data can prevent the model from making premature predictions before the appearance of key nodes and edges, thereby improving the reliability of the results.

However, the overall prediction effect for some intentions is not very good, such as No. 9 (opening closet), since closets feature in various other intentions such as organizing, reading, and wiping. This overlap presents a challenge in predicting human intentions in complex scenarios, which leads to ambiguity. However, this ambiguity reduces as more distinctive features become available for analysis.

In summary, it is evident that representing scenes by graphs, and objects and relationships with nodes and edges is a useful mechanism in human intention recognition and prediction. The intentions frequently correlate with contextual and environmental factors, the accuracy of intention predictions is enhanced when specific nodes and relationships manifest. By an increased number of nodes and relationships, we can preemptively identify human intentions prior to their actual occurrence.

6. Conclusion

The proposed human intention recognition framework leverages visual cues and human-object relationships extensively to identify intentions, addressing the current underemphasis on human-object interactions within the field. At the data level, incorporating a diverse array of human intentions and scenario-specific data establishes a robust foundation for more generalized human intention recognition, mitigating the limitations associated with single, fixed-scene recognition. Additionally, the model is adept at capturing local features and learning critical human-object relationships, a capability that is essential given the inherent complexity and variability of human intention recognition. This approach enhances the stability and generalizability of human intention recognition across diverse environments.

The key contributions of this work are as follows:

1) a new conceptual framework for human intention recognition, integrating perception, context-based representation and context-based reasoning.
2) a new graph dataset DSG (Dynamic Scene Graph), which represents the person and different objects as nodes and records the relationship between them as edges. It is the first graph dataset for human intention recognition extracted from videos.
3) a novel STGAIN model, which extracts the spatial information from the environment and learns features from a consecutive sequence of frames in the video.

This research provides a new direction for research on human intention recognition and prediction. Obtaining as much temporal and spatial information as possible may make prediction of human intentions more reliable and accurate. The proposed method demonstrates advantages in comparative experiments, underscoring its potential value in future robotic applications. Specifically, for the complex task of human intention recognition, the method effectively captures critical contextual features, thereby enhancing system stability. Moreover, the method is expected to improve the adaptability of robots in dynamic, unstructured environments. Importantly, the method accommodates inputs of varying sequence lengths while preserving sequence dependencies, a capability essential for enabling more informed and accurate decision-making in real-world robotic tasks.

Future work involves experimenting with more features and studying their importance as “triggers” of intention to make human intention recognition and prediction more accurate. The developed model will be deployed in a real-world environment to test the hypothesis that a robot with enhanced context awareness and abilities for intention recognition will exhibit higher degree of trustworthiness. This will be a significant step towards developing robots with awareness inside, which will be able to understand intentions, feel empathy, adapt and explain their actions, decisions and consequences. This will allow a step-up in engineering complex systems, making them more resilient, self-developing and human centric.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the support of the Centre for Artificial Intelligence, Robotics and Human-Machine Systems (IROHMS), operation C82092, part-funded by the European Regional Development Fund (ERDF) through the Welsh Government.

Data availability

Data is available at https://github.com/Tong0720/STGAIN.

References

Bagheri, E., Roesler, O., Cao, H. L., & Bram, V. (2021). A reinforcement learning based cognitive empathy framework for social robots. *International Journal of Social Robotics, 13*, 1079–1093.

Belardinelli, A. (2024). Gaze-based intention estimation: Principles, methodologies, and applications in HRI. *ACM Transactions on Human-Robot Interaction, 13*(3), 1–30.

Chen, X.-L., & Hou, W.-J. (2022). Gaze-Based Interaction Intention Recognition in Virtual Reality. *Electronics, 11*(10), 1647.

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. *In Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*. 16372-16382.

Dong, Y., Liu, Q., Du, B., & Zhang, L. (2022). Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Trans. Image Process., 31*, 1559–1572.

Fan, H., Zhuo, T., Yu, X., Yang, Y., & Kankanhalli, M. (2021). Understanding atomic hand-object interaction with human intention. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(1), 275–285.

Han, H., Zhang, M., Hou, M., Zhang, F., Wang, Z., Chen, E., Wang, H., Ma, J., & Liu, Q. (2021). STGCN: A spatial-temporal aware graph learning method for POI recommendation. In *In 2020 IEEE International Conference on Data Mining (ICDM)* (pp. 1052–1057).

Huang, Y., Bi, H., Li, Z., Mao, T., & Wang, Z. (2019). STGAT: Modeling spatial-temporal interactions for human trajectory prediction. In *In Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 6272–6281).

Huang, Y., Li, D., Han, B., Xu, E., & Medeossi, G. (2024). Multimodal transportation recommendation: Embedding travel intention and transit reliability by heterogeneous graph attention network. *Expert Systems with Applications, 255*, Article 124579.

Jarrassé, N., Sanguineti, V., & Burdet, E. (2013). Slaves no longer: Review on role assignment for human–robot joint motor action. *Adaptive Behavior, 22*(1), 70–82.

Ji, J., Ranjay, K., Li, F., & Juan, C. N. (2020). Action genome: Actions as compositions of spatiotemporal scene graphs. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),10236–10247*.

Jia, M., Wu, Z., Reiter, A., Cardie, C., Belongie, S., & Lim, S. N. (2021). Intentonomy: A dataset and study towards human intent understanding. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12986–12996).

Jiang, N., Jie, W., Li, J., Liu, X., & Jin, D. (2022). Gatrust: A multi-aspect graph attention network model for trust assessment in osns. *IEEE Transactions on Knowledge and Data Engineering, 35*, 5865–5878.

Kim, S., Yu, Z., & Lee, M. (2017). Understanding human intention by connecting perception and action learning in artificial agents. *Neural Networks, 92*, 29–38.

Kong, X., Xing, W., Wei, X., Bao, P., Zhang, J., & Lu, W. (2020). STGAT: Spatial-temporal graph attention networks for traffic flow forecasting. *IEEE Access, 8*, 134363–134372.

Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision, 130*(5), 1366–1401.

Li, S., Wang, R., Zheng, P., & Wang, L. (2021). Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm. *Journal of Manufacturing Systems, 60*, 547–552.

Li, W., Zhang, H., Bai, Q., Zhao, G., Jiang, N., & Yuan, X. (2022). Ppdl: Predicate probability distribution based loss for unbiased scene graph generation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19447-19456.

Li, P., Zheng, S., Liu, Z., Wang, X., Wang, V., Zheng, L., & Wang, L. (2022b). Proactive human–robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives. *Rob. Comput. Integr. Manuf, 81*, Article 102510.

Li, C., Yang, H., & Sun, J. (2022c). Intention-Interaction Graph Based Hierarchical Reasoning Networks for Human Trajectory Prediction. *IEEE Transactions on Multimedia*, 1–12.

Liu, C., Li, X., Li, Q., Xue, Y., Liu, H., & Gao, Y. (2021). Robot recognizing humans intention and interacting with humans based on a multi-task model combining ST-GCN-LSTM model and YOLO model. *Neurocomputing, 430*, 174–184.

Liu, M., Tang, S., Li, Y., & Rehg, J. M. (2020). Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In *In 16th European Conference Computer Vision (ECCV)* (pp. 704–721).

Liu, Z., Liu, Q., Xu, W., Wang, L., & Ji, Z. (2023). Adaptive real-time similar repetitive manual procedure prediction and robotic procedure generation for human-robot collaboration. *Advanced Engineering Informatics, 58*, Article 102129.

Min, S., Gao, Z., Peng, J., Wang, L., Qin, K., & Fang, B. (2021). STGSN—A Spatial–Temporal Graph Neural Network framework for time-evolving social networks. *Knowledge Management System, 214*, Article 106746.

Özdel. S., Rong, Y., Berat Mert Albaba, Kuo, Y.-L., Wang, X., & Enkelejda Kasneci. (2024). Gaze-Guided Graph Neural Network for Action Anticipation Conditioned on Intention. arXiv (Cornell University).

Petković, T., Puljiz, D., Marković, I., & Hein, B. (2018). Human intention estimation based on hidden Markov model motion validation for safe flexible robotized warehouses. *Robotics and Computer-Integrated Manufacturing, 57*, 182–196.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph databases: New opportunities for connected data*. O'Reilly Media Inc.

Sabab, S. A., Kabir, M. R., Hussain, S. R., Mahmud, H., Rubaiyeat, H. A., & Hasan, M. K. (2022). VIS-iTrack: Visual Intention Through Gaze Tracking Using Low-Cost Webcam. *IEEE Access, 10*, 70779–70792.

Selvakumar, S. (2024). An effective framework of human abnormal behaviour recognition and tracking using multiscale dilated assisted residual attention network. *Expert Systems with Applications, 247*, Article 123264.S.

Semeraro, F., Carberry, J., & Cangelosi, A. (2023). Towards Multi-User Activity Recognition through Facilitated Training Data and Deep Learning for Human-Robot Collaboration Applications. In *In 2023 International Joint Conference on Neural Networks (IJCNN)* (pp. 01–09).

Shangguan, Q., Fu, T., Wang, J., Fang, S., & Fu, L. (2022). A proactive lanechanging risk prediction framework considering driving intention recognition and different lane-changing patterns. *Accident Analysis & Prevention, 164*, Article 106500.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *In Proceedings of the European conference on computer vision (ECCV)* (pp. 510–526).

Singh, R., Miller, T., Newn, J., Velloso, E., Vetere, F., & Sonenberg, L. (2020). Combining gaze and AI planning for online human intention recognition. *Artificial Intelligence, 284*, Article 103275.

Tong, T., Setchi, R., & Hicks, Y. (2022). Context change and triggers for human intention recognition. *Procedia Computer Science, 207*, 3826–3835.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph Attention Networks. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Wang, W., Li, R., Chen, Y., & Jia, Y. (2018). Human intention prediction in human-robot collaborative tasks. In *HRI '18: ACM/IEEE International Conference on Human-Robot Interaction* (pp. 279–280).

Wang, X., Ma, Y., Wang, Y., Jin, Y., Wang, X., Tang, J., & Yu, J. (2020). Traffic flow prediction via spatial temporal graph neural network. In *In Proceedings of the web conference* (pp. 1082–1092).

Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: A survey. *ACM COMPUTING SURVEYS Home, 55*(5), 1–37.

Wu, T., Chen, F., & Wan, Y. (2018). Graph attention LSTM network: A new model for traffic flow forecasting. In *In 2018 5th International Conference on Information Science and Control Engineering (ICISCE)* (pp. 241–245).

Wu, W., Tu, F., Niu, M., Yue, Z., Liu, L., Wei, S., Li, X., & Yin, S. (2023). STAR: An STGCN ARchitecture for Skeleton-Based Human Action Recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers, 70*(6), 2370–2383.

Xia, Z., Zhang, Y., Yang, J., & Xie, L. (2024). Dynamic spatial–temporal graph convolutional recurrent networks for traffic flow forecasting. *Expert Systems with Applications, 240*, Article 122381.

Xu, B., Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. (2020). Interact as you intend: intention-driven human-object interaction detection. *IEEE Transactions on Multimedia, 22*(6), 1423–1432.

Xu, D., Zhu, Y., Choy, C. B., & Li, F. (2017). Scene graph generation by iterative message passing. In *In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5410–5419).

Xu, P., Chang, X., Guo, L., Huang, P., Chen, X., & Hauptmann, A. (2020). A survey of scene graph: Generation and application. *IEEE Transactions on Neural Networks and Learning Systems, 1–1*, 3385.

Yang, D., Zhang, H., Yurtsever, E., Redmill, K., & Ozguner, U. (2022). Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, 1.

Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A., & Özgüner, Ü. (2022). Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles, 7*(2), 221–230.

Yang, J., Lu, J., Lee, S., Batra, D., & Parikh, D. (2018). Graph r-cnn for scene graph generation. In *In Proceedings of the European conference on computer vision (ECCV)* (pp. 670–685).

Yang, J., Zheng, W. S., Yang, Q., Chen, Y. C., & Tian, Q. (2020). Spatial-temporal graph convolutional network for video-based person re-identification. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3289–3299).

Ye, Y., Hao, J., Hou, Y., Wang, Z., Xiao, S., Luo, Y., & Zeng, W. (2024). Generative ai for visualization: State of the art and future directions. *Visual Informatics*.

Ye, Y., & Ji, S. (2023). Sparse graph attention networks. *IEEE Transactions on Knowledge and Data Engineering, 35*(1), 905–916.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *In Proceedings of the AAAI conference on artificial intelligence., 35*(12), 11106–11115.