Front. Eng. Manag. 2025, 12(1): 99–116
https://doi.org/10.1007/s42524-024-4074-y

RESEARCH ARTICLE

Da HU, Mengjun WANG, Shuai LI

Multi-classifier information fusion for human activity recognition in healthcare facilities

Abstract In healthcare facilities, including hospitals, pathogen transmission can lead to infectious disease outbreaks, highlighting the need for effective disinfection protocols. Although disinfection robots offer a promising solution, their deployment is often hindered by their inability to accurately recognize human activities within these environments. Although numerous studies have addressed Human Activity Recognition (HAR), few have utilized scene graph features that capture the relationships between objects in a scene. To address this gap, our study proposes a novel hybrid multi-classifier information fusion method that combines scene graph analysis with visual feature extraction for enhanced HAR in healthcare settings. We first extract scene graphs, complete with node and edge attributes, from images and use a graph classification network with a graph attention mechanism for activity recognition. Concurrently, we employ Swin Transformer and convolutional neural network models to extract visual features from the same images. The outputs from these three models are then integrated using a hybrid information fusion approach based on Dempster-Shafer theory and a weighted majority vote. Our method is evaluated on a newly compiled hospital activity data set, consisting of 5,770 images across 25 activity categories. The results demonstrate an accuracy of 90.59%, a recall of 90.16%, and a precision of 90.31%, outperforming existing HAR methods and showing its potential for practical applications in healthcare environments.

Keywords human activity classification, scene graph, graph neural network, multi-classifier fusion, healthcare facility

Received May 1, 2024; revised July 1, 2024; accepted July 2, 2024

Da HU
Department of Civil and Environmental Engineering, Kennesaw State University, Marietta, GA 30060, USA

Mengjun WANG, Shuai LI (✉)
Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN 37996, USA
E-mail: sli48@utk.edu

This research was funded by the US National Science Foundation (NSF) via Grant number 2038967. This research also received support from the Science Alliance at the University of Tennessee Knoxville (UTK) via the Joint Directed Research and Development Program. Any opinions, findings, recommendations, and conclusions in this paper are those of the authors and do not necessarily reflect the views of NSF, Kennesaw State University, and UTK.

1 Introduction

Large gatherings, particularly in healthcare settings, are recognized as critical points for pathogen spread. The Centers for Disease Control and Prevention (CDC) reports numerous outbreaks of infectious diseases in hospitals, exerting significant strain on healthcare resources and resulting in a high number of fatalities (CDC, 2024). Annually, approximately 1.7 million people acquire infections during hospital treatment, leading to around 99,000 deaths (Haque et al., 2018). Moreover, hospital-acquired infections incur an estimated annual direct medical cost of about $30 billion and societal costs exceeding $10 billion due to premature mortality and reduced productivity. The recent global COVID-19 pandemic has further emphasized this issue, with over 449 million infections and 6 million deaths worldwide, and the numbers continue to rise due to more transmissible variants (Dong et al., 2020). In healthcare settings, maintaining proper sanitation and implementing effective disinfection practices are crucial in controlling the spread of infectious diseases (Assadian et al., 2021). This is where disinfection robots can play a pivotal role in effectively mitigating disease transmission.

Traditional cleaning and disinfection methods in healthcare facilities are manual, laborious, time-consuming, and prone to inconsistency due to human fatigue, potentially resulting in overlooked or insufficiently sanitized surfaces (Rutala and Weber, 2016). Robots can provide a solution to these challenges, operating

continuously and consistently without endangering humans (Guettari et al., 2021). The COVID-19 pandemic has further emphasized the importance of disinfection robots in performing environmental cleaning tasks (Zemmar et al., 2020). In our prior research, we developed algorithms for object detection (Hu et al., 2023), contaminated area segmentation (Hu et al., 2020), and material classification (Hu and Li, 2022) specifically for disinfection robots in healthcare facilities. However, a significant limitation remains: current disinfection robots lack the ability to recognize human activities within healthcare environments. This limitation undermines their effectiveness and hinders their widespread implementation. Healthcare facilities include a wide range of human activities, each of which influences how a disinfection robot should operate. For example, if a doctor's consultation is in progress, the robot should ideally postpone disinfection of that area to prevent disruptions. Therefore, the accurate recognition of human activities is a crucial prerequisite for the successful integration of disinfection robots into healthcare environments.

Human Activity Recognition (HAR) has been an area of study for over 20 years and has produced various methods for classifying activities from visual data such as videos or snapshots (Chen et al., 2021). Our study focuses on the classification of human activities from static images, aiming to enhance the operational efficiency of disinfection robots in healthcare environments. While video data provides a dynamic sequence of spatial and temporal cues, many actions can be discerned from single images or selected video frames, enabling prompt situational analysis (Guo and Lai, 2014). Additionally, the computational requirements for processing video data are significantly higher compared to still images (Rodríguez-Moreno et al., 2019), making image-based recognition more practical for disinfection robots. However, identifying human activities from a single frame presents distinct challenges, particularly in the presence of environmental noise or complex backdrops.

To address these challenges, our research proposes a novel approach that combines classifier techniques and leverages both scene structure and visual cues. Specifically, we have created a specialized image data set for healthcare environments, carefully annotated with precise activity labels. This data set serves as a valuable resource for testing activity classification models in healthcare settings, addressing the scarcity of specialized data available for HAR in the healthcare field. Additionally, our developed process transforms images into scene graphs, which highlight the contextual relationships within a scene using nodes and edges. This aids in the effective execution and evaluation of scene graph classifiers, as it enables the model to understand and utilize the spatial and relational context of activities. Furthermore, we integrate the strengths of Convolutional Neural Network (CNN), Visual Transformer (ViT), and Graph Neural Network (GNN) to extract a comprehensive range of features from images. Our fusion strategy combines Dempster-Shafer theory (DST) and weighted majority voting, resulting in a significant advancement in multi-classifier fusion for image classification.

The structure of this paper is as follows: Section 2 provides a literature review on the relevant work in the HAR field, while Section 3 presents the proposed model in detail. Section 4 discusses the experiments conducted and the results obtained, while Section 5 evaluates the proposed data set and model by comparing them with state-of-the-art methods. Finally, Section 6 concludes the study and discusses its potential applications.

2 Literature review

HAR plays a crucial role in studying human behavior and enhancing human-robot interfaces, including a wide range of activities from physical movements like jumping to daily tasks such as watching TV or making phone calls. There are two primary techniques in activity recognition: visual-based approaches and sensor-based approaches, each utilizing different types of data (Dang et al., 2020). Visual-based methodologies rely on image and video data captured by camera systems, such as RGB and RGB-D cameras. For instance, Uyguroğlu et al. (2024) investigated Alzheimer's disease classification using the fusion of multiple 3D angular orientations through CNN. On the other hand, sensor-based approaches employ various sensors like accelerometers, WiFi, and radar to detect activities. Studies conducted by Mudiyanselage et al. (2021) and Raza et al. (2023) have demonstrated the potential of sensors in activity recognition and related risk control. In the field of robotics, cameras are commonly used to enable robots to perceive and interact with their surroundings. Visual-based HAR can be categorized into three main types: traditional machine learning techniques, deep learning models, and the fusion of different data types for feature extraction.

2.1 Machine learning for HAR

For several years, machine learning techniques like Support Vector Machine (SVM) and Random Forest have been widely used in the field of identifying human actions from images. SVM aims to find the optimal hyperplane for classification, whereas Random Forest utilizes an ensemble of decision trees. These methods have shown success in less complex scenarios with a smaller number of samples. Earlier studies, such as Wang et al. (2006), employed unsupervised methods to categorize actions by comparing silhouettes in images. Building upon this work, Ikizler et al. (2008) refined the approach by evaluating human poses through shape histograms and reducing data before classification. Yao et al. (2011)

proposed the "Stanford 40 Action" benchmark data set and proposed a method that simultaneously models action attributes and parts using Locality-constrained Linear Coding (LLC) on dense SIFT features and a linear SVM classifier. In a study by Yun et al. (2013), hand-centered image patches were the focal point in classifying activities using SVMs. However, SVM encounters challenges with large-scale data sets due to increasing computational complexity, while the effectiveness of Random Forest may diminish due to the complex nature of decision trees and resource-intensive computational demands. Although these approaches have been effective, they often struggle with larger data sets because of their computational requirements and reliance on manually designed features, which may not scale well (Rashidi Nasab and Elzarka, 2023).

2.2 Advances in deep learning for HAR

With the availability of better computational resources, deep learning has emerged as a prominent method for image-based activity recognition, reducing the need for manual feature engineering (Yao et al., 2022). Oquab et al. (2014) found that CNNs trained on general images could enhance HAR. Following this trend, Gkioxari et al. (2015a) and Zhao et al. (2017b) advanced the field by incorporating contextual information and body part details. However, these methods often require additional manual annotations, which might limit their applicability in real-world scenarios.

To overcome limitations in HAR that arise from manually annotating bounding boxes around individuals, researchers have proposed innovative solutions. Khan et al. (2015) proposed a detection technique that utilizes features extracted by CNNs. Another approach, presented by Siyal et al. (2020), employed CNNs for feature extraction alongside SVMs for classification. Bera et al. (2021) took this further by integrating a CNN that specifically focuses on key areas of an image. They generated keypoints using a SIFT algorithm, grouped them using a Gaussian Mixture Model (GMM) to create salient regions, and then processed these regions with an attentional module for activity classification. Bas and Ikizler-Cinbis (2022) proposed an attentional deep multiple-instance learning network that identifies action-related regions and generates a pixel-level action map without irrelevant pixels. While these methodologies yield promising results, they primarily focus on visual features and neglect the valuable object and relationship information present in images. To address this, our study utilizes scene graphs extracted from images to capture complex relationships and object-specific information. We process these scene graphs using a GNN, a technique that has shown effectiveness in hyperspectral image classification (Hong et al., 2021, Ding et al., 2022, 2023). By combining visual aspects with the complex interplay of objects and relationships, our approach aims for a more holistic and accurate HAR.

2.3 Integrating multiple data types for HAR

Incorporating data from multiple sources has proven beneficial in interpreting complex human activities, as it provides rich semantic knowledge (Dang et al., 2020). Khaire et al. (2018) merged visual, depth, and structural data to classify actions, demonstrating improved resilience to suboptimal lighting conditions compared to using visual data alone. Guo et al. (2018) developed a technique that combines signals from WiFi and cameras to detect activities, proving particularly useful in challenging indoor environments with low light and obstructions. Singh et al. (2020) proposed a multi-view CNN approach that utilizes motion and depth information to enhance HAR, achieving high accuracy on demanding data sets such as MSR Daily Activity, UTD MHAD, and CAD 60. Although these methods have made significant advancements, they heavily rely on data from supplementary sensors. There is still a gap in leveraging the relationships between objects within an image for activity classification. A method that considers these relationships would provide a more comprehensive understanding of visual content and ultimately enhance the overall performance of HAR.

3 Methodology

This research presents a novel approach to decision fusion that combines visual cues and scene graph data to identify human actions within healthcare settings. The methodology, depicted in Fig. 1, consists of four main steps. First, the robot's camera acquires images. Second, three classifiers are utilized: a CNN using ConvNext (Liu et al., 2022), a ViT employing Swin Transformer (Liu et al., 2021), and a GNN based on an unbiased scene graph generation approach (Tang et al., 2020). These classifiers aim to recognize features from the image, focusing on spatial patterns, global context, and object relationships, respectively. Third, a hybrid decision fusion stage combines the probability distributions (PDs) of the activity categories output by these three classifiers using the DST and weighted majority vote, taking into account the level of conflict among the classifiers' outputs. Finally, the fused output is used to recognize and classify the human activity depicted in the images, providing valuable insights for monitoring and managing healthcare facilities.

3.1 Scene graph-based activity classification

Scene graph features offer a more comprehensive contextual understanding by capturing the relationships between

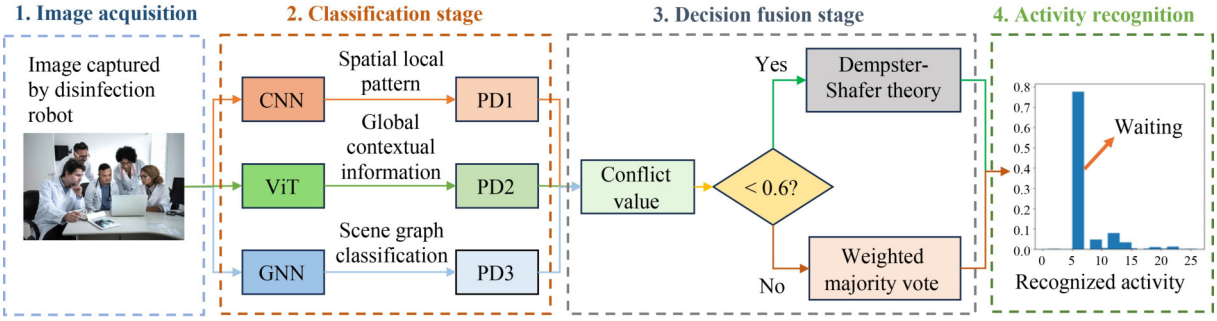

Fig. 1 Methodology overview.

objects. This is crucial for distinguishing relevant activities from background clutter and representing the complex interactions commonly found in healthcare environments. To achieve this, we leverage the unbiased scene graph generation method developed by Tang et al. (2020). This method employs counterfactual causality to identify and mitigate unwanted biases. It integrates with existing scene graph generation models such as MOTIFS (Zellers et al., 2018), Iterative Message Passing (IMP) (Xu et al., 2017), and VCTree (Tang et al., 2019). In our approach, we specifically utilize the MOTIFS model, which has been trained on the extensive Visual Genome data set (Krishna et al., 2017), to generate scene graphs. By combining MOTIFS with causal inference methods, we enhance our ability to accurately detect and interpret scene graphs, allowing us to understand complex visual content without pre-existing biases.

For the construction of scene graphs, we establish connections and define nodes using pairs with confidence levels above 0.5. This approach ensures a balance between detail and computational load. The selection of this threshold is based on common practices in the field and preliminary experiments, which have shown that a threshold of 0.5 effectively filters out less reliable relationships while preserving valuable information.

However, it is important to note that our experiments and training were conducted on a workstation, not on the actual disinfection robots. These robots have significantly lower computational power, and therefore the 0.5 threshold may need to be reconfigured when transferring the system to accommodate their limited capabilities.

To improve the representation of our graph elements and conserve time and computational effort, we utilize the Skip-gram model from Word2Vec. This model is trained on the extensive Google News data set and creates 300-dimensional vectors that capture nuanced language patterns for a wide range of words and phrases. The embeddings derived from Word2Vec enrich our node and edge features, providing a strong foundation for classifying the scene graphs with image categories as labels.

Figure 2 presents sample scene graphs generated from our image data. These examples are limited to the top 10 most confident objects and relationships for clarity. However, our method allows for more comprehensive extraction when necessary. This graph-based approach captures essential interaction data within images, providing valuable insights for effective HAR, such as identifying a group of medical professionals gathered for a conference.

The purpose of our study is to develop a graph classification network that can accurately classify human activity based on generated scene graphs. Figure 3 provides an overview of the flowchart for the GNN, which includes feature encoding and feature classification stages. We have utilized the graph attention mechanism in our network to effectively learn from graph-structured data by focusing on the neighboring nodes of each node. The Graph Attention Network (GAN), proposed by Veličković et al. (2017), has emerged as a popular architecture for representation learning with graphs. The GAN computes node similarity within a graph to uncover the hidden features of the graph nodes. A recent innovation is the dynamic graph attention variant proposed by Brody et al. (2021), which proposes a simple modification to enhance the model's ability to fit the training data. In our study, we have employed the proposed graph attention mechanism to encode the graph features. The GAN operator operates by attending to a node's neighbors to learn a representation that captures the node's local graph structure. This attention mechanism allows the network to focus on the most pertinent components of the graph, thereby facilitating more effective feature learning and classification. The details of the GAN operator are as follows:

The GAN takes node features, $h = \{h_1, h_2, \ldots, h_N\}$, $h_i \in \mathbb{R}^F$, and edge features, $e_{i,j} \in \mathbb{R}^D$ as inputs, where N is the total nodes, F is the node feature count, and D is the edge feature count. $e_{i,j}$ are considered only when connecting node i and node j. The GAN's output is an enhanced node feature set, $h' = \{h'_1, h'_2, \ldots, h'_N\}, h'_i = \mathbb{R}^{F'}$. The GAN process begins with a shared linear transformation applied to each node. In Eq. (1), attention coefficients μ_{ij}, signifying the influence of node j's feature on node i, are computed through concatenation and weighted by $\mathbf{W} \in \mathbb{R}^{F' \times (2F+D)}$.

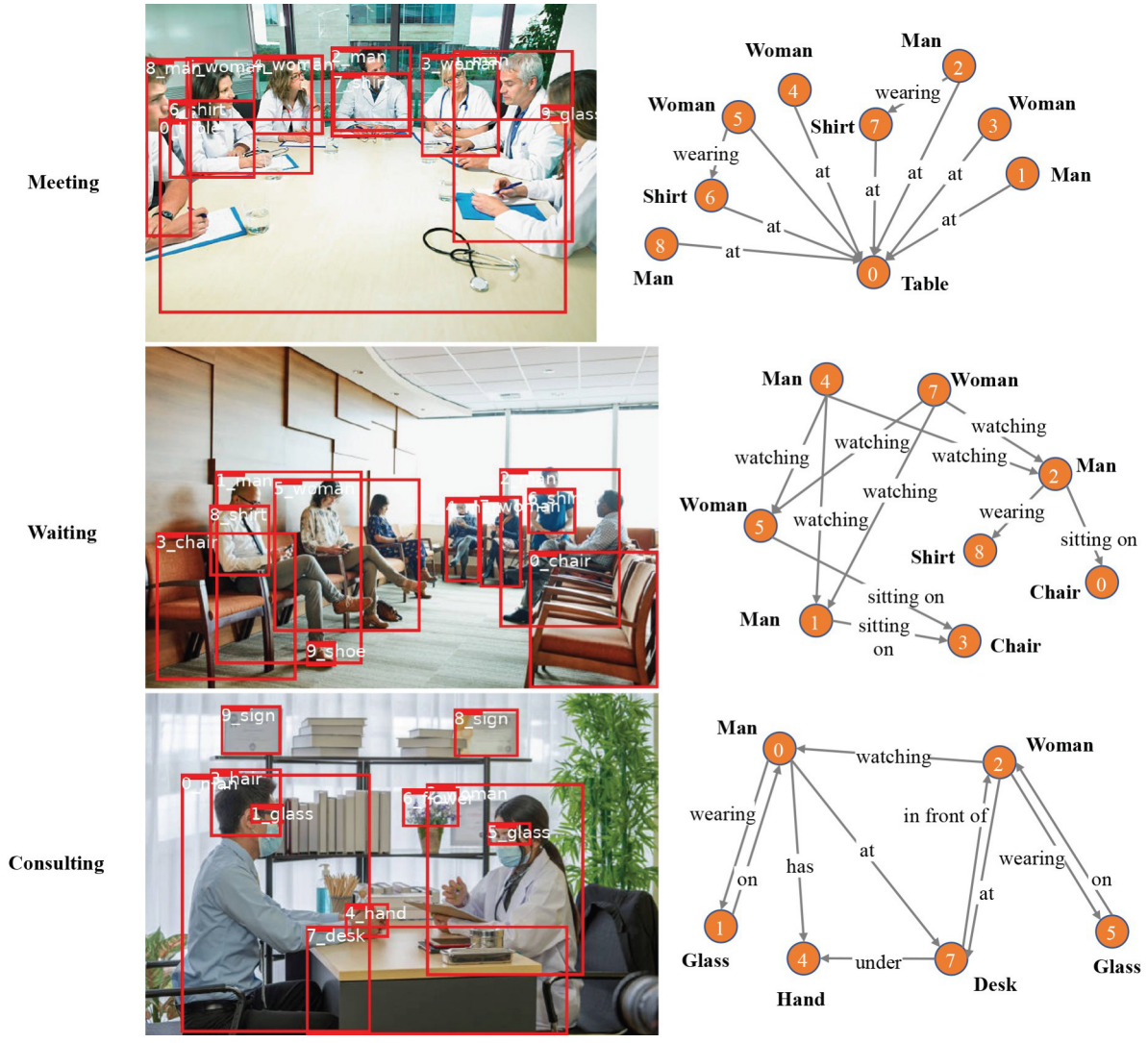

Fig. 2 Example of generated scene graphs for images.

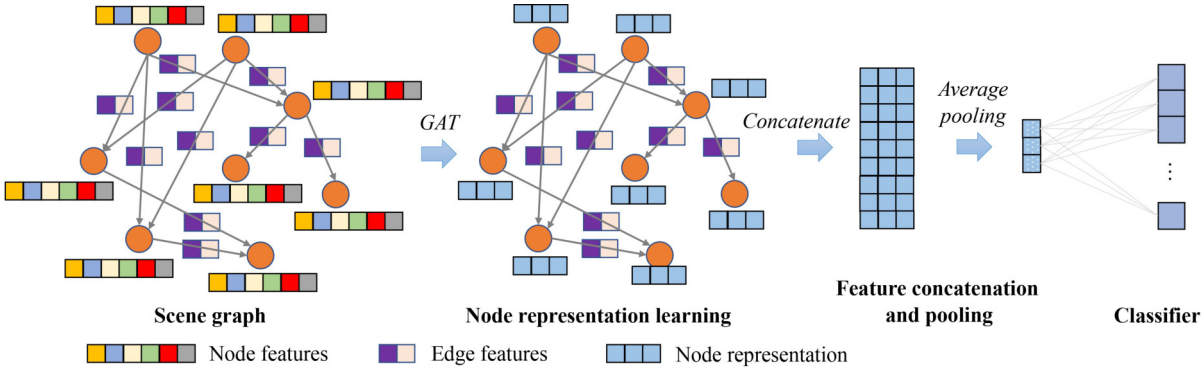

Fig. 3 Framework of the presented GNN.

$$\mu_{ij} = \mathbf{W}(h_i \,\|\, h_j \,\|\, e_{i,j}). \tag{1}$$

These coefficients are calculated for the neighboring nodes $k \in \mathcal{N}_i$ to maintain the graph's structure. Normalization of these coefficients is accomplished using a LeakyReLU activation (with a slope of 0.2) according to Eq. (2), where a $\mathbb{R}^{F'} \times \mathbb{R}^{F'}$ is the learnable weight vector. ensuring a balanced distribution of attention:

$$\alpha_{ij} = \frac{\exp(\mathbf{a}^T \mathrm{LeakyReLU}(\mu_{ij}))}{\sum_{k \in \mathcal{N}_i} \exp(\mathbf{a}^T \mathrm{LeakyReLU}(\mu_{ik}))}. \tag{2}$$

By employing a multi-head attention approach, the model calculates and averages distinct attention outputs to enhance performance and stabilize learning. Equation (3) defines this process, where K represents the attention count and σ signifies the ReLU function:

$$h_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^{k}\mathbf{W}^{k}h_j\right). \quad (3)$$

After the GAN operations, global average pooling is applied to aggregate node features, and a dense layer is used for classification. With a mere 0.11 million parameters, the GNN ensures efficiency in both training and inference times.

3.2 Transformer-based activity classification

Our research utilizes the Swin Transformer model, which is widely recognized for its effectiveness in visual feature learning for activity recognition tasks (Liu et al., 2021). The Swin_L architecture, an expanded version of the Swin Transformer, is summarized in Fig. 4. It follows a patch partitioning mechanism to divide the input image into distinct patches. Each patch's raw pixels form a 48-dimensional vector when using a 4 × 4 patch size. These vectors are then refined through a series of transformer stages. The first stage consists of a linear embedding layer that converts the 48-dimensional vectors into 192-dimensional vectors, followed by two Swin Transformer blocks. Stages 2 through 4 incorporate a patch merging step, which combines features from neighboring patches and reduces dimensions before entering additional transformer blocks. Each stage includes 2, 18, and 2 Swin Transformer blocks, respectively. Swin Transformer blocks utilize a shifted window Mechanism for Self-Attention (MSA) and contain dual integrated Multilayer Perceptron (MLP) layers with GELU activations, all preceded by layer normalization. This hierarchical structure enables comprehensive activity classification through tiered feature synthesis.

3.3 CNN-based activity classification

For activity classification using CNN, we have chosen the State-of-the-Art (SotA) ConvNeXt model. Based on the well-established ResNet framework, ConvNeXt incorporates several refinements to improve performance. Figure 5 presents the ConvNeXt_L setup, which includes an initial convolution, successive stages of ConvNeXt units, and a classification layer. The process begins with a 4 × 4 convolution to reduce the feature map size, followed by four tiers of ConvNeXt units, each consisting

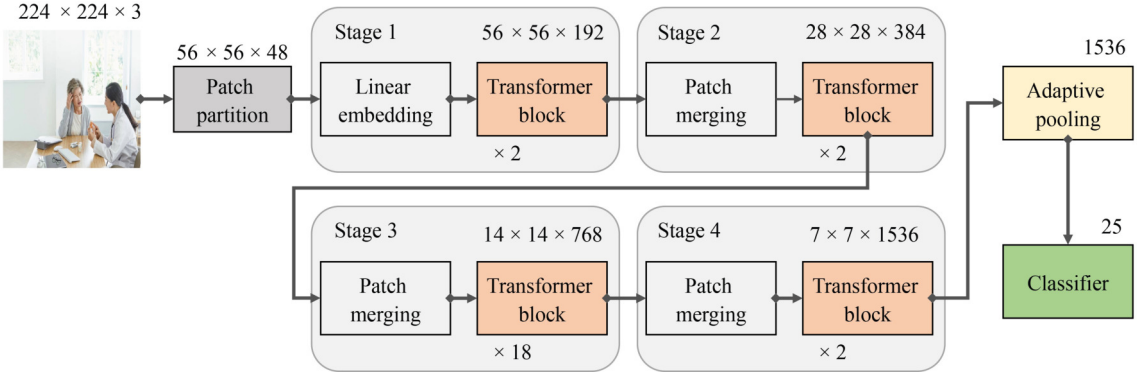

Fig. 4 Architecture of the Swin_L architecture.

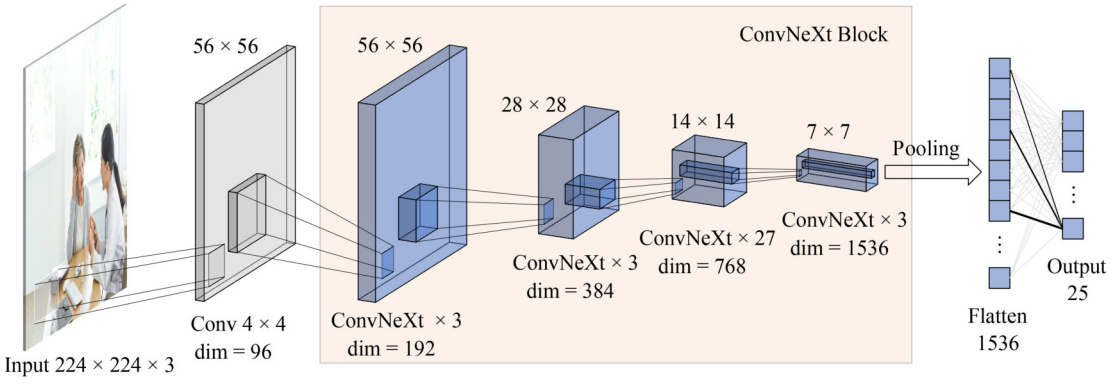

Fig. 5 Architecture of the ConvNeXt_L architecture.

of 3, 3, 27, and 3 units. As the layers deepen, the feature map's resolution decreases while the channel count increases. After the final stage, an adaptive pooling layer condenses the feature map to 1 × 1, leading to a dense layer that performs the final activity classification. For a detailed explanation of the network, please refer to (Liu et al., 2022).

3.4 Decision fusion

This section presents our decision fusion strategy, which effectively combines the outputs of the GNN, Swin_L, and ConvNeXt_L models to enhance robustness, accuracy, and adaptability in varying lighting conditions. The integration of multiple classifiers through multi-classifier fusion leverages their respective strengths. Specifically, this approach capitalizes on the CNN's proficiency in identifying spatial patterns, the ViT's ability to capture global context, and the GNN's effectiveness in understanding object relationships and interactions. We propose a novel fusion method that combines DST with a weighted majority voting system to establish a coherent decision framework. This hybrid methodology incorporates a weighted majority vote mechanism to enhance the decision fusion process by capitalizing on the advantages of both DST and majority voting.

To tailor our decision fusion framework, we have adapted the DST (Rogova, 2008), considering the unique classification strengths of individual classifiers to gauge the reliability of their decisions. Decision templates based on typical classification patterns observed during training are utilized as reference points for evaluating new data. Equation (4) defines the decision template DT_j for class w_j, where $\boldsymbol{Z}$ represents the data set, N_j denotes the number of elements of $\boldsymbol{Z}$ from w_j, DP signifies the decision outputs from the classifier with probabilities, and c represents the total number of classes.

$$DT_j = \frac{1}{N_j}\sum_{z_k \in w_j, z_k \in Z} DP(z_k),\ j = 1,2,\ldots,c. \tag{4}$$

The decision profile $DP(\mathbf{x})$ and decision template DT_j are Lxc-dimensional matrices. L is the number of classifiers. Let DT_j^i denote the ith row the decision template DT_j. Let $D_i(x) = [d_{i,1}(x),\ldots,d_{i,c}(x)]$ denote the ith row of the decision profile $DP(\mathbf{x})$ from the classifier D_i. The proximity between DT_j^i and the output of classifier D_i for the input $\mathbf{x}$ is given in Eq. (5), where $\|\cdot\|$ represents L_1 norm, and $\phi_{j,i}(\mathbf{x})$ is the proximity for class $j = 1,2,\ldots,c$ and classifier $i = 1,2,\ldots,L$.

$$\phi_{j,i}(\mathbf{x}) = \frac{\left(1 + \|DT_j^i - D_i(\mathbf{x})\|\right)^{-1}}{\sum_{k=1}^{c}\left(1 + \|DT_k^i - D_i(\mathbf{x})\|\right)^{-1}}. \tag{5}$$

Equation (6) calculates the belief degree.

$$b_j(D_i(\mathbf{x})) = \frac{\phi_{j,i}(\mathbf{x})\prod_{k\neq j}\left(1-\phi_{k,i}(\mathbf{x})\right)}{1-\phi_{j,i}(\mathbf{x})\prod_{k\neq j}\left(1-\phi_{k,i}(\mathbf{x})\right)}. \tag{6}$$

Equation (7) calculates the final degrees of support, where β is the normalizing constant.

$$\mu_j(\mathbf{x}) = \beta\prod_{i=1}^{L} b_j(D_i(\mathbf{x})). \tag{7}$$

For high-conflict scenarios within our decision fusion process, we shift to a weighted majority vote to bolster the dependability of our fusion method. The conflict value K is defined in Eq. (8), where $D_i^{k_i}$ denotes the ith row and k_ith column of the decision profile DP(x).

$$K = \sum_{k_1\cap\cdots\cap k_c=\varnothing} D_1^{k_1}(\mathbf{x})\cdot D_2^{k_2}(\mathbf{x})\cdot\cdots\cdot D_L^{k_c}(\mathbf{x}). \tag{8}$$

Under the assumption that a conflict value (K) greater than 0.6 indicates a significant conflict between different classifiers (Daniel and Lauffenburger, 2011), the weighted majority vote is given by Eq. (9), where $\mathbf{M}$ represents the $1\times L$ weight matrix calculated based on the accuracy of the classifier on the validation set, $DP^k(\mathbf{x})$ represents the kth column of the decision profile $DP(\mathbf{x})$. The threshold of 0.6 was determined empirically through extensive experimentation, balancing sensitivity and accuracy. The weighted vote is calibrated based on each classifier's validation set accuracy, with classifiers demonstrating higher accuracy receiving greater influence in the final fusion decision.

$$\mu_j(\mathbf{x}) = \mathbf{M}\times DP^k(\mathbf{x}). \tag{9}$$

In a demonstrative scenario using synthetic data, we present three decision templates (given in Eq. (10)) to clarify the fusion approach.

$$DT_1 = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.7 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.1 \end{bmatrix},\quad DT_2 = \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.7 & 0.1 \end{bmatrix},$$
$$DT_3 = \begin{bmatrix} 0.1 & 0.3 & 0.6 \\ 0.2 & 0.2 & 0.6 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}. \tag{10}$$

The decision profile for input $\mathbf{x}$ the three classifiers is described in Eq. (11)

$$DP(\mathbf{x}) = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.3 & 0.6 & 0.1 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}. \tag{11}$$

0.55 is the calculated result of the conflict value K using Eq. (8), which is smaller than 0.6 where DST will be used. Using Eq. (5), the proximities for each decision

template are calculated and shown in Table 1.

Table 2 shows the belief degrees and final degree of supports calculated using the Eqs. (6) and (7). In this example, w_2 performs slightly better on the fusion method.

4 Experiment and results

4.1 Data set description

Our data set creation involved two stages: downloading images and cleaning the data. We identified 25 common activities in healthcare settings, each with implications for robotic disinfection, as detailed in Table 3. For example, the activity of "comforting" involves close physical contact, indicating a need for thorough disinfection in areas where such interactions are frequent. In contrast, during "doctor meetings," the disinfection robot could focus on other high-risk areas due to the gathering of multiple individuals. Similarly, "infusing," a routine procedure, requires regular disinfection in the areas where it takes place to maintain hygiene standards.

Images were sourced from Getty and Shutterstock,

Table 1 Proximity measures

Class	$\phi_{j,1}(\mathbf{x})$	$\phi_{j,2}(\mathbf{x})$	$\phi_{j,3}(\mathbf{x})$
w_1	0.3600	0.3139	0.3222
w_2	0.3600	0.4036	0.2636
w_3	0.2800	0.2825	0.4142

Table 2 Belief and final degrees of support

Class	$b_j(D_1(\mathbf{x}))$	$b_j(D_2(\mathbf{x}))$	$b_j(D_3(\mathbf{x}))$	$\mu_j(\mathbf{x})$
w_1	0.2058	0.1637	0.1701	0.3352
w_2	0.2058	0.2499	0.1244	0.3741
w_3	0.1374	0.1388	0.2609	0.2907

Table 3 Selection and implications of activities in healthcare facilities for robot disinfection

Activity	Selection reason	Implication for robot disinfection
Measuring blood pressure	High frequency activity with close patient interaction	Areas where this activity is performed need frequent disinfection due to potential for pathogen transmission
Comforting	Involves close physical contact between healthcare workers and patients	Increased risk of pathogen transmission requires careful disinfection
Carrying patients	High-risk activity due to physical contact when transporting through various parts of the facility	Frequent disinfection needed in pathways and patient handling areas
Consulting	Universal healthcare interaction, potentially involving multiple parties	Regular disinfection of consulting rooms needed due to consistent human presence
Scanning	High frequency activity involving use of shared equipment	Equipment and surrounding areas need regular disinfection
Doctor meeting	Regular occurrence for patient updates and care planning	Meeting rooms need regular disinfection due to the number of individuals present
Analyzing samples	Crucial activity for diagnosis and treatment planning	Laboratories require stringent disinfection due to potential for pathogen presence in samples
Doctor sleeping	Necessary for healthcare providers in settings with extended shifts	Living quarters require regular disinfection to ensure providers' health
Eating	Universal activity with potential for droplet spread	Cafeterias or eating areas require frequent disinfection
Family visiting	Frequent activity that brings in external individuals	Visiting areas require regular disinfection due to the number of different individuals present
Cleaning	Essential activity for maintaining a healthy environment	Avoiding cleaning overlap and improving overall efficiency
Patient sitting in a wheelchair	Common scenario for mobility-impaired individuals	Wheelchairs and associated areas require frequent disinfection
Patient sitting in a wheelchair with assistance	High-risk activity due to physical contact	Assistive devices and associated areas need frequent disinfection
Infusing	Regular occurrence in hospital settings	Infusion areas require stringent disinfection due to direct pathogen entry risk
Injecting	High frequency activity in many healthcare contexts	Injection areas require stringent disinfection due to direct pathogen entry risk
Lying in a bed	Universal scenario for hospitalized patients	Patient rooms require regular disinfection due to continuous patient presence
Online meeting	Common occurrence for remote consultations or internal discussions	Devices used for online meetings need regular disinfection
Performing surgery	High-risk activity with strict disinfection standards	Operating rooms require stringent disinfection protocols
Patient walking	Common activity in physical therapy or recovery phases	Walkways need frequent disinfection due to potential for droplet spread
Working in pharmacy	Regular activity for medication management	Pharmacies need regular disinfection due to the handling of shared items
Sitting on a bed	Common scenario for patients or visitors	Patient rooms require regular disinfection due to potential for pathogen transmission
Discussing	Universal activity with potential for droplet spread	Meeting or discussion areas require regular disinfection

which offer extensive image libraries with textual descriptions corresponding to visual content. Approximately 1,000 images were acquired for each hospital activity. The data set was further refined through a verification process, where mislabeled images were either removed or reassigned to the correct categories. We also discarded images of poor quality, such as blurry or unclear ones that did not clearly depict the intended human activity. Despite these exclusions, our data set retains substantial diversity in terms of background, human pose, and appearance within each category. To improve the data set, the final selection of images was reviewed by an independent human labeler. The completed data set contains between 180 and 396 images per class, totaling 5,770 images.

Table 4 provides a comprehensive breakdown of the data set used in this research. It categorizes the images into different classes, along with the quantity of images in each class. Additionally, brief descriptions are included for each class to highlight their unique characteristics and attributes. This detailed breakdown enhances the reader's understanding of the data set's scope and diversity.

Figure 6 shows a curated collection of images that represent the wide range of activities taking place in healthcare settings. Each of the 25 identified activities is depicted by a carefully chosen representative image from our data set. These selected images vividly portray the specific environments and scenarios associated with each activity, emphasizing the diversity and nuances captured in our study.

In our analysis of activity patterns, we employed the

Table 4 Activity description and statistics

Activity	Description	Shortened	Number of people	Count
Measuring blood pressure	Medical staff is helping a patient to measure blood pressure.	Measuring	2	250
Comforting	Medical staff is comforting or expressing concern to patients.	Comforting	2 and above	189
Carrying patients	Critically ill patients are being carried by first responders, doctors, or nurses.	Carrying	2 and above	231
Consulting	Patients are consulting a professional doctor or nurse about their condition.	Consulting	2 and above	197
Scanning	Doctors are helping patients with MRIs or full-body scans in large instruments.	Scanning	1 and above	220
Doctor meeting	Doctors are having an in-person academic meeting around the table	Meeting	2 and above	203
Analyzing samples	Doctors are analyzing blood or drug sample in hospital laboratory analysis room.	Analyzing	1 and above	247
Doctor sleeping	Doctors or first responders are temporarily sleeping and resting in the hospital.	Sleeping	1 and above	293
Eating	Inpatients in wheelchairs or beds are eating.	Eating	1 and 2	198
Family visiting	Family members are visiting a patient who lying or sitting on the bed.	Visiting	3 and above	202
Cleaning	Cleaning staff is disinfecting and cleaning within a hospital.	Cleaning	1 and above	262
Patient sitting in a wheelchair	A patient is sitting alone in a wheelchair.	Wheelchair_1	1	253
Patient sitting in a wheelchair with assistance	A patient is sitting in a wheelchair with an or multiple assistant or nurse next to it.	Wheelchair_2	2 and above	202
Infusing	A patient is being infused while lying or sitting on the bed.	Infusing	1 and above	165
Injecting	A patient is being injected by a doctor.	Injecting	2 and above	197
Lying in a bed	Patients are lying in the nursing bed.	Lying	1 and above	203
Online meeting	A doctor or paramedic is meeting online using a computer.	O-meeting	1	205
Performing surgery	Doctors and nurses are performing surgery for a patient in a professional operating room.	Operating	2 and above	193
Patient walking	A patient is walking alone in hospital hallway or with one/couple of nursing staff.	Walking	1 and above	180
Working in pharmacy	Doctors in pharmacy are sorting, helping clients, or recording.	Pharmacy	1 and above	248
Sitting on a bed	Patients are sitting on the nursing bed.	Sitting	1	194
Discussing	A group of doctors or paramedics is standing together and discussing.	Discussing	2 and above	181
Checking temperature	Nursing staff is using thermometer guns to measure the temperature of the patients.	Checking	2 and above	343
Patient waiting	Patients are waiting in a waiting room or hallway for treatment or examination.	Waiting	1 and above	396
Examining X-ray	Doctors are examining the X-ray pictures by himself/herself or with a colleague or nurse.	Examining	1 and 2	318

Fig. 6 Example images in the proposed data set.

Barnes-Hut t-SNE method (Maaten and Hinton, 2008) to visualize the raw image data in a two-dimensional plane. t-SNE is widely recognized for its effectiveness in reducing the dimensionality of high-dimensional data and projecting it onto a lower-dimensional space. The process involves transforming the similarity between image samples into conditional probabilities and minimizing the Kullback-Leibler divergence between these low-dimensional embedding conditional probabilities and the high-dimensional data. Considering the large number of image features, we initially used Principal Component Analysis (PCA) to reduce the dimensionality to 50, following the recommendation of Kobak and Berens (2019). These 50-dimensional features were then inputted into the t-SNE algorithm to convert them into a two-dimensional representation. It is important to note that the input images were resized to 224 × 224 pixels with three channels, resulting in a total of 150,528 features for each image. The perplexity value for the t-SNE analysis was set at 30.

Figure 7 illustrates the t-SNE visualization, which displays the two-dimensional distribution of image data from a hospital setting. It reveals a convergence of features from different activities, indicating similarities in human activities that could pose challenges for automated recognition systems. This similarity is primarily attributed to the images being captured within healthcare facilities, which often share similar environmental backgrounds.

4.2 Implementation details

Our deep learning models were developed using the PyTorch framework (Paszke et al., 2019) and trained on an Ubuntu workstation equipped with two NVIDIA Quadro P5000 graphics cards. We initialized the Swin_L and ConvNeXt_L models by using pretrained weights from ImageNet. To ensure consistency, we employed a uniform set of training parameters for all models, including the GNN. The Stochastic Gradient Descent (SGD) optimizer (Ruder, 2017) was used for network optimization, with a weight decay of 5e-4 and a momentum of 0.9. These parameter values were chosen based on a combination of recommendations from the literature and empirical testing to achieve optimal performance. The initial learning rate for SGD was set at 1e-3, and it was reduced by half every 20 epochs to strike a balance between training convergence speed and stability. Considering hardware limitations, we set the batch size to 24 and trained the networks for a total of 50 epochs. During training, model weights were adjusted using cross-entropy loss, a commonly employed choice for classification models. The hospital activity data set was randomly divided into training (80%), validation (10%), and testing (10%) sets. To assess the effectiveness of our method, we generated five random splits and selected the most promising results from the validation phase to inform decision fusion. Finally, the performance of the fusion model was evaluated on the test set using standard evaluation

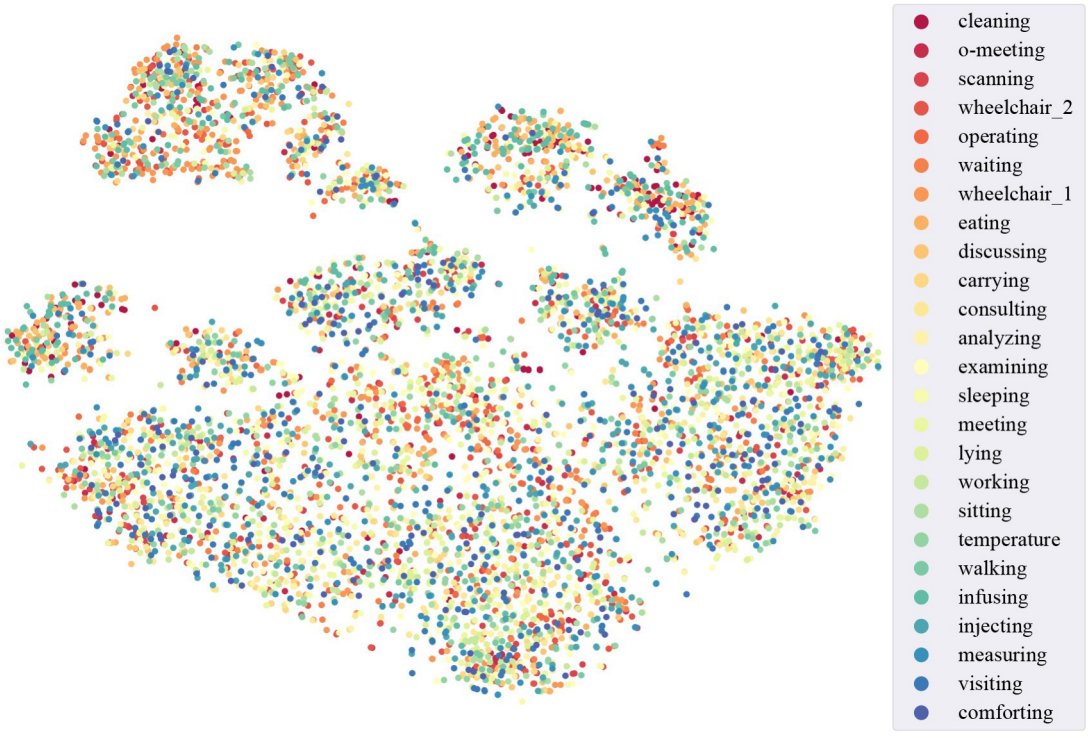

Fig. 7 t-SNE visualization of hospital activity data sets. Each point represents an image, and its associated color represents activity category.

metrics, including accuracy, recall, precision, and the F1 score (Powers, 2010).

4.3 Results

In this section, we present the results of our experiments. Figure 8 illustrates the progression of loss values for the Swin_L, ConvNeXt_L, and GNN models during training and validation. The loss trends for ConvNeXt_L and Swin_L converge, reaching low loss values by the 30th epoch, indicating successful learning of activity representation features. To enhance this learning process and improve model generalizability, we incorporated data augmentation techniques such as varying crops and image transformations to diversify the training data. The GNN, although initially exhibiting higher loss, consistently improved over the course of 50 epochs, suggesting potential benefits from extending the training duration. During the validation phase, ConvNeXt_L attained lower loss values earlier than Swin_L, with GNN stabilizing slightly later. This indicates an opportunity for performance optimization in the GNN when compared to the other models.

Table 5 presents the performance metrics of each classifier, including our proposed method, on the testing set. These metrics, such as mean and standard deviation, are derived from the results of five random splits. The provided metrics are calculated from five random splits of the data set to ensure robust cross-validation. Among the classifiers, ConvNeXt_L stands out in performance, exhibiting high scores across accuracy, F1, precision, and recall metrics of 89.51%, 89.50%, 90.17%, and 89.30%, respectively. However, the GNN records lower scores in comparison, with F1 score, precision, and recall of 63.84%, 62.78%, 63.54%, and 63.16%, respectively. This difference may be attributed to the similar environmental contexts across different categories in healthcare facilities, leading to similarities in the scene graphs representing these images.

Figure 9 displays the confusion matrix for the three classifiers on the testing data set. The confusion matrix aggregates the results across five random splits to account for variability among different splits. It is row-normalized, with the diagonal representing the predictive power of the model for positive classes. For the GNN, recall varies significantly across different human activities, with ‘measuring blood pressure’ and ‘comforting patients’ being the most frequently misclassified activities, having recalls of 0.26 and 0.33, respectively. In contrast, ‘online meeting’ shows the best performance under GNN, with a recall of 0.95. Swin_L’s classifier stands out, showing perfect recall in activities like ‘infusing,’ ‘operating,’ and several others, demonstrating its robustness in accurate activity classification. Similarly, ConvNeXt_L achieves a recall of 1 in a broad range of activities, from ‘infusing’ to ‘analyzing,’ exhibiting its capability in consistently recognizing varied actions within healthcare settings.

Table 4 illustrates the average performance of our

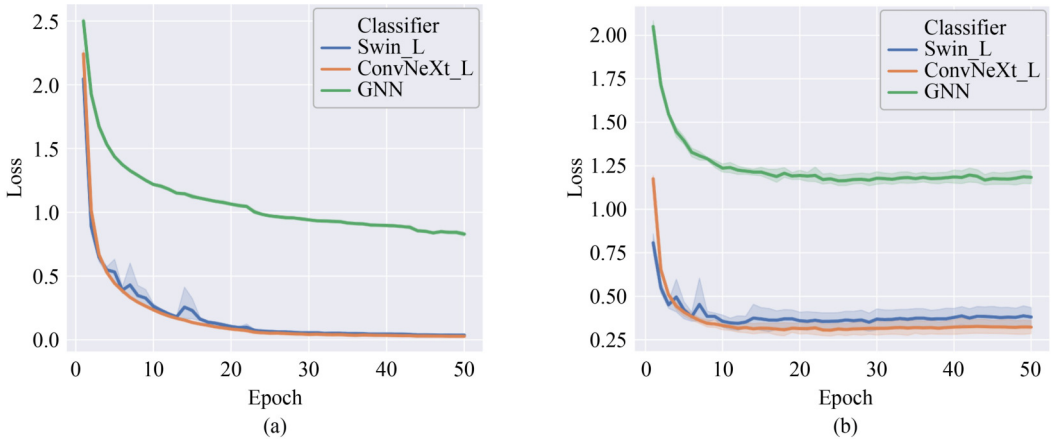

Fig. 8 Training and validation loss trends during model training. Solid lines represent the mean loss across five data set shuffles, with shaded areas showing the variance: (a) Training; (b) Validation.

Table 5 Classifier performance on the test data for hospital activity recognition

Model	Accuracy		F1		Precision		Recall	
	mean	st.d.	mean	st.d.	mean	st.d.	mean	st.d.
GNN	63.84	2.33	62.78	2.21	63.54	2.02	63.16	2.20
Swin_L	89.34	1.19	89.27	1.25	90.00	1.06	89.00	1.36
ConvNeXt_L	89.51	0.91	89.50	0.98	90.17	1.01	89.30	0.94
Proposed	**90.59**	1.18	**90.54**	1.33	**91.16**	1.22	**90.31**	1.38

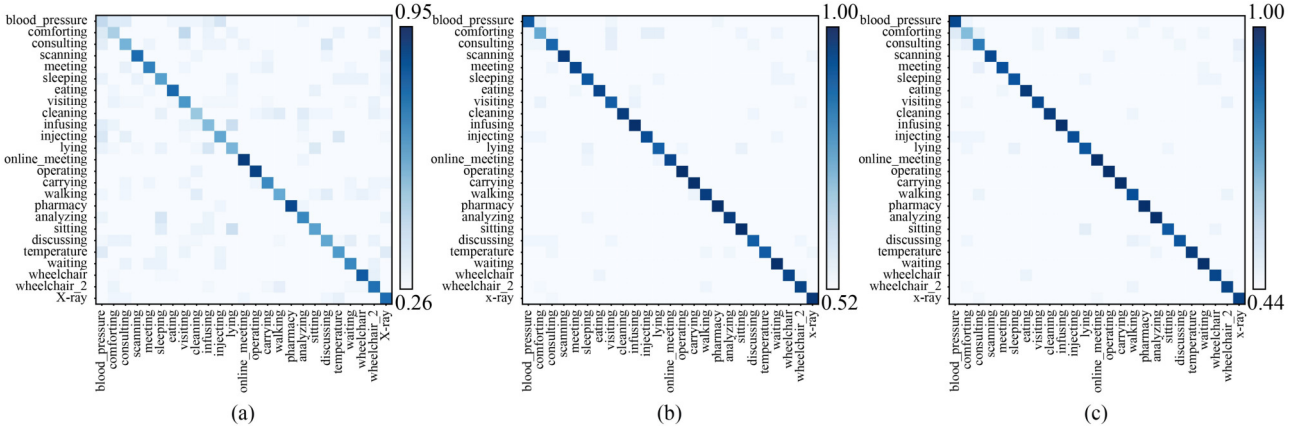

Fig. 9 Confusion matrix of three classifiers: (a) GNN; (b) Swin_L; and (c) ConvNeXt_L.

proposed approach, surpassing 90% in accuracy, F1 score, precision, and recall, as computed from test data across different splits. This reflects a slight but consistent enhancement in all metrics compared to the ConvNeXt_L model. Figure 10 displays the confusion matrix of our proposed method, revealing a recall spectrum ranging from moderate to perfect across activities, with 'comforting patients' being on the lower end. This suggests that activities involving close patient interaction, such as 'comforting,' are more challenging to distinguish due to their contextual similarity to other patient-care actions.

4.4 Ablation study

In the conducted ablation study, we evaluated the individual and collective contributions of classifier pairs using our fusion method. The results in Table 6 demonstrate that combining Swin_L and ConvNeXt_L achieves an impressive accuracy rate of 90.49%, outperforming each classifier alone by a margin of 0.98%. The introduction of GNN to either Swin_L or ConvNeXt_L leads to slight accuracy gains, indicating the complementary nature of these classifiers. Moreover, the combined use of all three classifiers highlights the strength of collaborative fusion

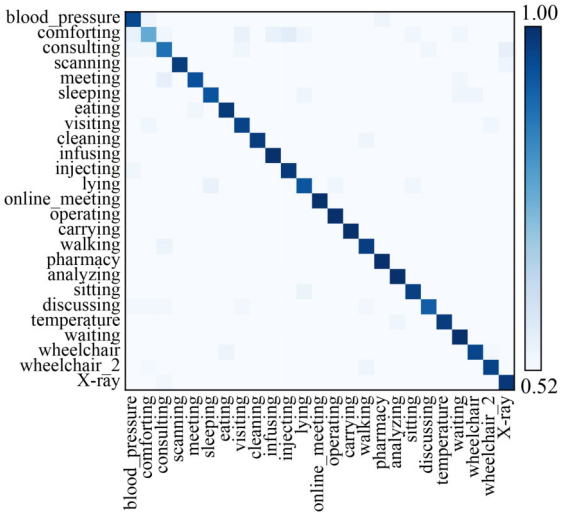

Fig. 10 Confusion matrix of the proposed method.

Table 6 Results of ablation study

Model	Accuracy		F1		Precision		Recall		Infer. (ms)
	mean	st.d.	mean	st.d.	mean	st.d.	mean	st.d.	
GNN	63.84	2.33	62.78	2.21	63.54	2.02	63.16	2.20	1.46
Swin_L	89.34	1.19	89.27	1.25	90.00	1.06	89.00	1.36	31.97
ConvNeXt_L	89.51	0.91	89.50	0.98	90.17	1.01	89.30	0.94	27.16
GNN + Swin_L	89.47	1.00	89.43	1.07	90.16	0.95	89.12	1.17	46.17
GNN + ConvNeXt_L	89.61	0.93	89.54	0.98	90.09	1.05	89.36	0.95	41.36
Swin_L + ConvNeXt_L	90.49	1.16	90.44	1.30	91.11	1.15	90.21	1.35	71.87
Proposed	**90.59**	1.18	**90.54**	1.33	**91.16**	1.22	**90.31**	1.38	73.33

in improving classification results.

However, it is important to note that despite GNN's fast inference time of 1.46 ms, the scene graph generation process takes approximately 100 ms, making the overall process slower compared to other methods. Nevertheless, the enhanced representation of scenes provided by GNN justifies its inclusion. This study demonstrates the potential of combining diverse classifier types in a hybrid model, where each classifier captures unique aspects of image data, resulting in a more comprehensive scene analysis. Furthermore, this research could serve as a foundation for future efforts to refine and optimize this approach, potentially leading to notable accuracy enhancements. The balance between accuracy and processing time can be adjusted to cater to specific use case requirements. Additionally, future work could explore faster scene graph generation methods, such as CV-SGG (Jin et al., 2023), to enhance the overall efficiency of the system.

5 Discussion

5.1 Comparison with state-of-the-arts

This study proposes a novel approach for recognizing human activities in hospitals. We achieved this by constructing a comprehensive data set and generating scene graphs from images. Our approach includes a specialized GNN for scene graph analysis and an innovative amalgamated multi-classifier fusion approach that synthesizes results from GNN, Swin_L, and ConvNeXt_L. The system demonstrates promising results on a dedicated hospital activity data set. Furthermore, we evaluated our model's performance using two benchmark data sets:

Stanford 40 action data set. This data set consists of 9,532 images across 40 action categories, depicting various everyday human activities such as running, walking, and applauding. Each category contains 180 to 300 images, ensuring a diverse and extensive collection. The data set provides a predefined train-test split, with each training set comprising 100 images per action category.

PASCAL VOC 2012. This data set includes 10 human actions, such as playing an instrument, riding a bike, and using a computer. It consists of 2,296 training images and 2,292 validation images. To align with the methods being compared against, we evaluate our model's performance on the validation set. Notably, images depicting multiple actions are excluded from both the training and validation sets, ensuring that our networks are trained and tested on images associated with a single action.

Performance on Stanford 40 action data set. Table 7 compares our method with SotA approaches on the Stanford 40 action data set. Many existing action recognition methods rely on bounding-box annotations, as provided by data sets like Stanford 40, to enhance their performance. These methods utilize contextual information and semantic parts around or within person bounding boxes to improve results. In contrast, our method does not require bounding-box annotations or the detection of humans and objects in images. Instead, it autonomously learns visual and graph features for action representation. Our model achieves the highest performance on the Stanford 40 data set, with a mean average precision (mAP) of 96.6%, surpassing the SotA Attend and Guide method by 0.4%. The Multi-Scale Context Network (MSCNet), which employs learning human bodies and action-specific semantic parts, achieves the second-best performance with a mAP of 94.6%. Notably, our approach shows a significant improvement over MSCNet, with a 2% increase in mAP. Furthermore, our model outperforms other methods by a substantial margin, ranging from 5.4% upwards.

Performance on PASCAL VOC 2012 action data set. Table 8 presents a comparison between our method and SotA approaches. Many existing action recognition methods heavily rely on manually annotated bounding boxes to enhance their predictive power. For example, R*CNN achieves a mAP of 87.9% by incorporating annotated human bounding boxes, but its performance drops to 84.9% without such annotations. However, depending on manual annotations is impractical in real-world applications. Significant improvements in performance on the PASCAL VOC 2012 data set were only achieved with the introduction of the Top-Down + Bottom-Up Attention network (Bas and Ikizler-Cinbis, 2022), which achieves a mAP of 95.0%. Our method matches the performance of this network, placing it among the SotA methods. Notably, our approach outperforms the Top-Down + Bottom-Up Attention network by 5.6% on the Stanford 40 action data set. This comparison with SotA methods highlights the effectiveness and efficiency of our approach for human action recognition tasks.

Table 7 Comparison to SotA methods on Stanford 40 action data set

Method	mAP (%)
R*CNN (Gkioxari et al., 2015b)	90.9
Person Detection (Khan et al., 2015)	75.4
Action Masks (Zhang et al., 2016)	82.6
Top-Down Pyramid (Zhao et al., 2016)	80.6
AttSPP-Net (Feng et al., 2017)	81.6
Semantic Part Action (Zhao et al., 2017a)	91.2
Multi-Branch Attention (Yan et al., 2018)	85.2
Attend and Guide (Bera et al., 2021)	96.2
Top-Down + Bottom-Up Attention (Bas and Ikizler-Cinbis, 2022)	91.0
MSCNet (Zheng et al., 2022)	94.6
This work	96.6

Table 8 Comparison to SotA methods on PASCAL VOC 2012 action data set

Method	mAP (%)
R*CNN (Gkioxari et al., 2015b)	87.9
Action Part (Gkioxari et al., 2015a)	80.4
Action Masks (Zhang et al., 2016)	82.2
AttSPP-Net (Feng et al., 2017)	76.2
Semantic Part Action (Zhao et al., 2017b)	90.0
Generalized Symmetric Pair Model (Zhao et al., 2017a)	71.1
Multi-Branch Attention Network (Yan et al., 2018)	87.1
Top-Down + Bottom-Up Attention (Bas and Ikizler-Cinbis, 2022)	95.0
This work	95.0

5.2 Performance of GNN

The performance of GNN is significantly lower compared to the other two classifiers, Swin_L and ConvNeXt_L. Specifically, GNN achieves accuracies of 63.84%, 59.38%, and 66.15% on our hospital activity data set, Stanford 40, and PASCAL VOC 2012, respectively. The relatively lower performance on our hospital activity data set can be attributed to the similarities in the environment of certain activities. For instance, activities such as “Infusing” and “Injecting” occur in similar healthcare settings, where a healthcare professional interacts with a patient near a bed or medical workstation, with essential medical equipment present like an IV stand, needles, or syringes. These common elements are captured by the scene graph, posing a challenge for the GNN model to differentiate between these activities. To demonstrate this, we specifically selected ‘scanning’, ‘Eating’, ‘Operating’, ‘Carrying’, ‘Analyzing’, and ‘Wheelchair_1’ for classification, as they have distinct environmental backgrounds. The results indicate that GNN achieves an average accuracy of 88.81% on this subset of activities. The relatively lower performance on the Stanford 40 data set is due to certain inherent characteristics of this data set. Actions like ‘blowing bubbles’, ‘smoking’, ‘running’, or ‘waving hands’ mainly focus on a single human with minimal background context. From the perspective of the scene graph, these scenarios primarily capture the human figure, providing limited additional contextual information for effective classification. The same issue is observed with the PASCAL VOC 2012 data set, where the lack of contextual information in many images similarly hampers the performance of the GNN model.

In summary, the lower performance of the GNN model across different data sets can be attributed to various

factors. For our hospital activity data set, the GNN model struggles to effectively differentiate certain activities due to their environmental similarities. Similarly, the Stanford 40 and PASCAL VOC 2012 data sets pose challenges because they often feature images with a single human figure and minimal background context, which deprives the scene graph of crucial additional information for accurate classification. Moving forward, our work suggests potential avenues for future research. Exploring methods to improve activity differentiation in similar environments and enhancing the handling of images with minimal background information could lead to better performance of scene-graph based HAR models. Additionally, developing and integrating advanced techniques for interpreting and contextualizing human figures and their actions in scene graphs holds promise for future research. Despite some limitations, incorporating scene graph analysis into activity recognition presents a promising research direction that offers a structured approach to understanding and interpreting complex visual data. With further refinement and development, these techniques have the potential to significantly advance the field of automated activity recognition.

5.3 Model performance under varied lighting conditions

To assess the robustness of our HAR system, it is crucial to evaluate its sensitivity to varying lighting conditions. Hence, we extended our evaluation to examine the impact of lighting on our proposed model. To assess the model's performance in different lighting scenarios, we conducted tests on the first split of our data set, which was originally one of five random splits used to calculate the mean performance value. This split was modified with gamma correction, which changes the pixel values of an image using a power-law expression (Eq. (12)) to simulate different lighting conditions.

$$I_{\text{out}} = I_{\text{in}}^{\gamma} \tag{12}$$

Where I_{in} is the input value (ranges from 0 to 1), I_{out} is the adjusted value, and γ represents the gamma value. If the gamma value is less than 1, the image becomes lighter, while the image becomes darker if the gamma value is greater than 1. Figure 11 shows the example test images under different gamma values.

The results, presented in Table 9, demonstrate the confusion matrix of our system's performance across various altered scenarios. The confusion matrix provides valuable insight into the model's accuracy in processing images affected by different gamma adjustments, thereby simulating a wide range of realistic lighting environments. By systematically modifying the gamma settings and analyzing the resulting changes in the confusion matrix, we can derive conclusions about the model's robustness to lighting variations. This experimental setup enables us to identify any potential degradation in recognition performance caused by changes in image brightness and contrast, both of which are critical factors for practical applications in diverse and dynamic real-world conditions.

5.4 Future work

This study has made significant advancements in HAR for healthcare applications. However, there are several areas for future research that could enhance system performance. One such area is the refinement of scene graph extraction. Currently, our scene graphs contain unnecessary details that may impact the performance of the GNN. Developing a more sophisticated method to filter out these details could enhance the accuracy and speed of our system. Additionally, expanding our existing graph classification network to a multi-layer network could improve our ability to recognize activities. However, this expansion may give rise to challenges such as nodes becoming too similar and a learning bottleneck resulting from excessive information compression (Li et al., 2018, Alon and Yahav, 2020). Therefore, future work could focus on addressing these issues to strike a

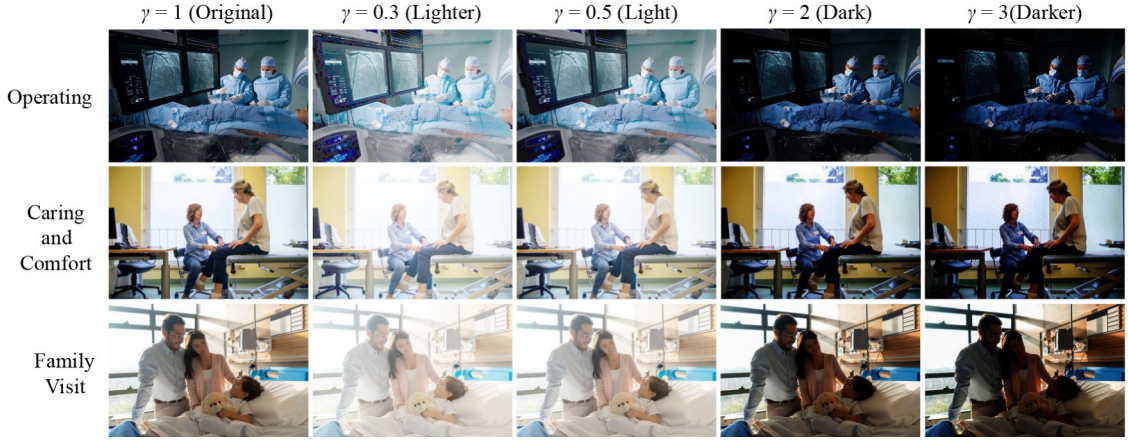

Fig. 11 Example of test images under different lighting.

Table 9 Testing of proposed model on images with different lighting

Gamma	Accuracy	F1	Precision	Recall
1	91.17	91.35	91.99	91.13
0.3	87.95	88.0	88.5	88.01
0.5	90.15	90.25	90.93	90.02
2	90.32	90.37	90.9	90.22
3	89.13	89.22	89.9	88.99

balance between network depth and learning efficiency.

In conclusion, addressing the challenges of variation and noise in images used for HAR is crucial. To tackle these challenges, we propose adapting techniques from other fields, such as the Augmented Linear Mixing Model (Hong et al., 2019), and applying them to enhance the stability of recognition algorithms. Testing and implementing our proposed methods in real-world, uncontrolled environments, particularly in healthcare settings, could lead to more efficient procedures, including cleaning and disinfecting. To validate and refine our method for live environments, we emphasize the importance of enhancing our data set with real images from healthcare settings. Additionally, we are exploring methods for robots to accurately determine their location (Aggarwal, 2020a), using a variety of evaluation measures (Aggarwal, 2020b, Xiao et al., 2023). Implementing these systems in real-time within healthcare facilities will provide practical insights and opportunities for improvement. To enhance computational efficiency, we propose applying optimization techniques such as structured pruning, model quantization, and knowledge distillation to our Swin Transformer and CNN models. These techniques have the potential to reduce computational load without significant performance loss. Furthermore, we suggest utilizing sparse computations for graph analysis, Fast Fourier Transforms, transfer learning, and specialized hardware such as GPUs and TPUs to improve efficiency. In summary, our research aims to advance HAR systems, making them more precise, efficient, and applicable in various settings, particularly in complex environments like healthcare facilities.

6 Conclusions

In this study, we present a composite decision fusion strategy for HAR that effectively integrates information from scene graphs and visual data. We have created a comprehensive hospital activity data set comprising 5,770 images, including 25 diverse categories of activities. To analyze these images, we have applied a technique to transform them into scene graphs and utilized a GNN for accurate classification. Simultaneously, we have utilized two advanced deep learning architectures, namely Swin_L and ConvNeXt_L, for visual feature-based HAR. Our decision fusion method combines the DST with a weighted majority vote to consolidate the decisions obtained from the GNN, Swin_L, and ConvNeXt_L models. This approach has yielded an overall accuracy of 90.59%, with a precision of 90.31% and similar metrics for F1 score and recall. However, the significance of our approach goes beyond these performance metrics. By accurately recognizing human activities, our system can be seamlessly integrated with existing medical interventions, forming a comprehensive strategy to mitigate pathogen transmission in healthcare facilities. This integration enables enhanced infection control measures by providing healthcare professionals with detailed insights into patient and staff movements and interactions. Ultimately, our approach has the potential to play a crucial role in reducing the spread of infections and improving patient safety in healthcare environments, thereby demonstrating its extensive practical implications.

CRediT authorship contribution statement Da Hu: Methodology, Software, Validation, Investigation, Data curation, Writing - Original Draft. Mengjun Wang: Methodology, Data curation, and Validation. Shuai Li: Conceptualization, Methodology, Writing, Review and Editing, Supervision, Project administration, and Funding acquisition.

Competing Interests The authors declare that they have no competing interests.

Appendix

Abbreviation	Meaning
CDC	US Centers for Disease Control and Prevention
HAR	Human activity recognition
CNN	Convolutional neural network
ViT	Visual transformer
GNN	Graph neural network
DST	Dempster-shafer theory
SVM	Support vector machine
LLC	Locality-constrained linear coding
GMM	Gaussian mixture model
PDs	Probability distributions
GAN	Graph attention network
MSA	Multi-head self-attention
MLP	Multilayer perceptron
PCA	Principal component analysis
SGD	Stochastic gradient descent
mAP	Mean average precision
SotA	State-of-the-art
MSCNet	Multi-scale context network
R*CNN	Region-based convolutional neural network
CV-SGG	Computer vision scene graph generation

References

Aggarwal A K (2020a). Enhancement of GPS position accuracy using machine vision and deep learning techniques. Journal of Computational Science, 16(5): 651–659

Aggarwal A K (2020b). Fusion and enhancement techniques for processing of multispectral images. In: Ran A & Teiji W (Eds.), Unmanned Aerial Vehicle: Applications in Agriculture and Environment Cham: Springer International Publishing. 159–175

Alon U, Yahav E (2020). On the bottleneck of graph neural networks and its practical implications. ArXiv: 2006.05205

Assadian O, Harbarth S, Vos M, Knobloch JK, Asensio A, Widmer AF (2021). Practical recommendations for routine cleaning and disinfection procedures in healthcare institutions: A narrative review. Journal of Hospital Infection, 113: 104–114

Bas C, Ikizler-Cinbis N (2022). Top-down and bottom-up attentional multiple instance learning for still image action recognition. Signal Processing Image Communication, 104: 116664

Bera A, Wharton Z, Liu Y, Bessis N, Behera A (2021). Attend and guide (AG-Net): A keypoints-driven attention-based deep network for image recognition. IEEE Transactions on Image Processing, 30: 3691–3704

Brody S, Alon U, Yahav E (2021). How attentive are graph attention networks? 2022 International Conference on Learning Representations (ICLR)

CDC (2024). [Internet]. Available from the wobsit of CDC

Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. ACM Comput. Surv., 54: 77:1–77:40

Daniel J, Lauffenburger J P (2011). Conflict management in multisensor dempster-shafer fusion for speed limit determination. 2011 IEEE Intelligent Vehicles Symposium (IV). p. 987–992.

Ding Y, Zhang Z, Zhao X, Hong D, Cai W, Yang N, Wang B (2023). Multi-scale receptive fields: Graph attention neural network for hyperspectral image classification. Expert Systems with Applications, 223: 119858

Ding Y, Zhang Z, Zhao X, Hong D, Cai W, Yu C, Yang N, Cai W (2022). Multi-feature fusion: Graph neural network and CNN combining for hyperspectral image classification. Neurocomputing, 501: 246–257

Dong E, Du H, Gardner L (2020). An interactive web-based dashboard to track COVID-19 in real time. Lancet. Infectious Diseases, 20(5): 533–534

Feng W, Zhang X, Huang X, Luo Z (2017). Attention focused spatial pyramid pooling for boxless action recognition in still images. In: Lintas A, Rovetta S, Verschure PFMJ, and Villa AEP, eds. Artificial Neural Networks and Machine Learning – ICANN 2017 Cham: Springer International Publishing. 574–581

Gkioxari G, Girshick R, Malik J (2015a). Actions and attributes from wholes and parts. In: 2015 IEEE International Conference on Computer Vision (ICCV). 2470–2478

Gkioxari G, Girshick R, Malik J (2015b). Contextual action recognition with R* CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). 1080–1088

Guettari M, Gharbi I, Hamza S (2021). UVC disinfection robot. Environmental Science and Pollution Research International, 28(30): 40394–40399

Guo G, Lai A (2014). A survey on still image based human action recognition. Pattern Recognition, 47(10): 3343–3361

Guo L, Wang L, Liu J, Zhou W, Lu B (2018). HuAc: Human activity recognition using crowdsourced WIFI signals and skeleton data. Wireless Communications and Mobile Computing, 2018: 1–15

Haque M, Sartelli M, McKimm J, Abu Bakar M B (2018). Health care-associated infections—An overview. Infection and Drug Resistance, 11: 2321–2333

Hong D, Gao L, Yao J, Zhang B, Plaza A, Chanussot J (2021). Graph convolutional networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 59(7): 5966–5978

Hong D, Yokoya N, Chanussot J, Zhu X (2019). An augmented linear mixing model to address spectral variability for hyperspectral unmixing. IEEE Transactions on Image Processing, 28: 1923–1938

Hu D, Li S (2022). Recognizing object surface materials to adapt robotic disinfection in infrastructure facilities. Computer-Aided Civil and Infrastructure Engineering, 37(12): 1521–1546

Hu D, Li S, Wang M (2023). Object detection in hospital facilities: A comprehensive dataset and performance evaluation. Engineering Applications of Artificial Intelligence, 123: 106223

Hu D, Zhong H, Li S, Tan J, He Q (2020). Segmenting areas of potential contamination for adaptive robotic disinfection in built environments. Building and Environment, 184: 107226

Ikizler N, Cinbis R G, Pehlivan S, Duygulu P (2008). Recognizing actions from still images. In: 2008 19th International Conference on Pattern Recognition (ICPR). 1–4

Jin T, Guo F, Meng Q, Zhu S, Xi X, Wang W, Mu Z, Song W (2023). Fast contextual scene graph generation with unbiased context augmentation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 6302–6311

Khaire P, Kumar P, Imran J (2018). Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognition Letters, 115: 107–116

Khan F S, Xu J, Van De Weijer J, Bagdanov A D, Anwer R M, Lopez A M (2015). Recognizing actions through action-specific person detection. IEEE Transactions on Image Processing, 24(11): 4422–4432

Kobak D, Berens P (2019). The art of using t-SNE for single-cell transcriptomics. Nature Communications, 10(1): 5416

Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L J, Shamma D A, Bernstein M S, Fei-Fei L (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1): 32–73

Li Q, Han Z, Wu X (2018). Deeper insights into graph convolutional networks for semi-supervised learning. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 3538–3545

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 9992–10002

Liu Z, Mao H, Wu C Y, Feichtenhofer C, Darrell T, Xie S (2022). A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on

Computer Vision and Pattern Recognition. 11966–11976

Dang L M, Min K, Wang H, Piran M J, Lee C H, Moon H (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition, 108: 107561

Mudiyanselage S E, Nguyen P H D, Rajabi M S, Akhavian R (2021). Automated workers' ergonomic risk assessment in manual material handling using semg wearable sensors and machine learning. Electronics, 10(20): 2558

Oquab M, Bottou L, Laptev I, Sivic J (2014). Learning and transferring mid-level image representations using convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. 1717–1724

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019). PyTorch: An imperative style, high-performance deep learning library

Powers D M W (2010). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. ArXiv

Rashidi Nasab A, Elzarka H (2023). Optimizing machine learning algorithms for improving prediction of bridge deck deterioration: A case study of ohio bridges. Buildings, 13(6): 1517

Raza Usmani A, Kotowski S E, Davis K G (2023). The impact of hospital bed height and gender on fall risk during bed egress. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1): 2434–2436

Rodríguez-Moreno I, Martínez-Otzeta J M, Sierra B, Rodriguez I, Jauregi E (2019). Video activity recognition: State-of-the-Art. Sensors, 19(14): 3160

Rogova G (2008). Combining the results of several neural network classifiers. In: RR Yager and L Liu, editor. Classic Works of the Dempster-Shafer Theory of Belief Functions Berlin, Heidelberg: Springer Berlin Heidelberg. p. 683–692

Ruder S (2017). An overview of gradient descent optimization algorithms. ArXiv

Rutala W A, Weber D J (2016). Monitoring and improving the effectiveness of surface cleaning and disinfection. American Journal of Infection Control, 44(5): e69–e76

Singh R, Khurana R, Kushwaha A K S, Srivastava R (2020). Combining CNN streams of dynamic image and depth data for action recognition. Multimedia Systems, 26(3): 313–322

Siyal A R, Bhutto Z, Muhammad S, Iqbal A, Mehmood F, Hussain A, Ahmed S (2020). Still image-based human activity recognition with deep representations and residual learning. International Journal of Advanced Computer Science and Applications, 11(5): 471–477

Tang K, Niu Y, Huang J, Shi J, Zhang H (2020). Unbiased scene graph generation from biased training. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3713–3722

Tang K, Zhang H, Wu B, Luo W, Liu W (2019). Learning to compose dynamic tree structures for visual contexts. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6612–6621

Uyguroğlu F, Toygar Ö, Demirel H (2024). CNN-based Alzheimer's disease classification using fusion of multiple 3D angular orientations. Signal, Image and Video Processing, 18(3): 2743–2751

Maaten L van der, Hinton G (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9: 2579–2605

Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2017). Graph attention networks

Wang Y, Jiang H, Drew M S, Li Z N, Mori G (2006). Unsupervised discovery of action classes. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1654–1661

Xiao J, Aggarwal A K, Rage U K, Katiyar V, Avtar R (2023). Deep learning-based spatiotemporal fusion of unmanned aerial vehicle and satellite reflectance images for crop monitoring. IEEE Access: Practical Innovations, Open Solutions, 11: 85600–85614

Xu D, Zhu Y, Choy C B, Fei-Fei L (2017). Scene graph generation by iterative message passing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. 3097–3106.

Yan S, Smith J S, Lu W, Zhang B (2018). Multibranch attention networks for action recognition in still images. IEEE Transactions on Cognitive and Developmental Systems, 10(4): 1116–1125

Yao B, Jiang X, Khosla A, Lin A L, Guibas L, Fei-Fei L (2011). Human action recognition by learning bases of action attributes and parts. In: 2011 IEEE International Conference on Computer Vision. 1331–1338

Yao J, Cao X, Hong D, Wu X, Meng D, Chanussot J, Xu Z (2022). Semi-Active convolutional neural networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 99: 1–1

Yun Y, Gu I Y H, Aghajan H (2013). Riemannian manifold-based support vector machine for human activity classification in images. In: 2013 20th IEEE International Conference on Image Processing. 3466–3469

Zellers R, Yatskar M, Thomson S, Choi Y (2018). Neural motifs: Scene graph parsing with global context. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5831–5840

Zemmar A, Lozano A M, Nelson B J (2020). The rise of robots in surgical environments during COVID-19. Nature Machine Intelligence, 2(10): 566–572

Zhang H B, Lei Q, Zhong B N, Du J X, Peng J (2016). A survey on human pose estimation. Intelligent Automation & Soft Computing, 22(3): 483–489

Zhao Z, Ma H, Chen X (2016). Semantic parts based top-down pyramid for action recognition. Pattern Recognition Letters, 84: 134–141

Zhao Z, Ma H, Chen X (2017a). Generalized symmetric pair model for action classification in still images. Pattern Recognition, 64: 347–360

Zhao Z, Ma H, You S (2017b). Single image action recognition using semantic body part actions. In: 2017 IEEE International Conference on Computer Vision. 3411–3419

Zheng X, Gong T, Lu X, Li X (2022). Human action recognition by multiple spatial clues network. Neurocomputing, 483: 10–21