Engineering, Construction and Architectural Management

1

Intentions prediction for human–robot collaboration in utility tunnel maintenance

Yang Su
School of Design and the Built Environment, Curtin University, Perth, Australia
Jun Wang and Wenchi Shou
School of Engineering, Design and Built Environment, Western Sydney University, Sydney, Australia
Peng Wu
School of Design and the Built Environment, Curtin University, Perth, Australia
Chengke Wu
Chinese Academy of Sciences, Shenzhen Institutes of Advanced Technology, Shenzhen, China, and
Shuyuan Xu
School of Civil Engineering and Architecture, Zhejiang Sci-Tech University, Hangzhou, China and
School of Design and the Built Environment, Curtin University, Perth, Australia

Received 9 November 2024
Revised 11 March 2025
28 September 2025
Accepted 15 November 2025

Abstract

Purpose – This paper introduces a novel hybrid deep learning model aimed at enhancing human intention prediction in human–robot collaboration for utility tunnel maintenance. Recognizing the inherent dangers and the confined nature of utility tunnels, the study aims to advance safety and operational efficiency by improving robot assistants' ability to accurately interpret and predict human-worker intentions. The purpose is to reduce human exposure to hazardous environments and optimize task execution through precise and timely robot actions.

Design/methodology/approach – Our approach involves a hybrid deep learning architecture combining traditional time-series image classification with advanced semantic information extraction. The proposed hybrid intentions prediction deep learning model (HIPM) utilizes convolutional neural networks, long short-term memory networks and the contrastive language–image pre-training model for a comprehensive feature extraction and intention prediction. This integration enables the model to process visual and contextual data from dynamic and challenging tunnel environments, addressing the inadequacies of traditional vision-based and physical-based intention prediction methods.

Findings – Empirical validation conducted on real-world utility tunnel data from Jiangsu Province, China, demonstrated that HIPM significantly outperforms traditional models. HIPM achieved a precision of 91.52% and a recall of 91.20%, indicating a high level of accuracy in predicting human intentions. The results underscore the model's robustness and reliability, confirming its effectiveness in understanding and responding to complex human-worker behaviors in utility tunnel settings.

Originality/value – The originality of this research lies in its novel integration of multimodal deep learning techniques to enhance the interpretability and adaptability of robots in human-robot collaborative environments. The HIPM model's ability to interpret both human actions and environmental contexts presents a significant advancement over existing models, offering a more reliable and efficient approach to managing the safety and

Engineering, Construction and Architectural Management
Vol. 33 No. 15, 2026
pp. 1-21
Emerald Publishing Limited
e-ISSN: 1365-232X
p-ISSN: 0969-9988
DOI 10.1108/ECAM-11-2024-1531

efficiency of utility tunnel maintenance operations. This study contributes a pioneering solution to the challenges of human intention prediction in one of the most hazardous and demanding industrial settings.

Keywords Construction safety, Information and communication technology (ICT) applications, Experimental studies

Paper type Research article

1. Introduction

Utility tunnel maintenance (UTM) is a frequent and hazardous task involving infrastructure inspection, repairs, cleaning, ventilation and security monitoring (Lee *et al*., 2018; Hai *et al*., 2024; Wang, 2021). Conducted in confined underground spaces, UTM exposes workers to risks such as poor air quality, hazardous materials, structural collapses and high-voltage electrical or gas-related dangers (Zhang *et al*., 2020; Sousa and Einstein, 2021). Human–robot collaboration (HRC) is essential to reducing human exposure to these risks, enabling robots to take on high-risk tasks. For effective collaboration, robots must accurately and timely recognize human intentions (Lin and Lukodono, 2021; Pan and Yu, 2024), allowing them to dynamically adjust their actions, provide necessary tools and assist in emergencies. This capability enhances safety, efficiency and coordination while minimizing misunderstandings and operational risks.

Human intention prediction (HIP) refers to inferring a collaborator's near-future goals and actions from observable cues – such as body pose, hand–object interactions and task/environment context – before those actions are completed. Accurate HIP is pivotal in HRC because it enables proactive robot behaviors (e.g. handing the right tool, yielding space or triggering safety responses), thereby improving safety and task efficiency in risk-intensive, time-critical settings like utility tunnels. However, existing human intention prediction (HIP) methods are difficult to apply in UTM scenarios. Current mainstream HIP methods can be divided into physical-based (contact) and vision-based (contactless) approaches (Robinson *et al*., 2023; Semeraro *et al*., 2023).

Physical-based methods use various physical sensors to collect human motion data to help robots understand human intentions. For example, tactile gloves (Zou *et al*., 2024), surface electromyography (EMG) (Zhang *et al*., 2022) and electroencephalography (EEG) (Buerkle *et al*., 2021) have all shown good potential in HIP applications but are almost inapplicable in UTM scenarios. Tactile gloves and other wearable devices are highly sensitive to moist environments (Büscher *et al*., 2015), which are common in underground utility tunnels where handling water pumps or groundwater leakage is frequent. Methods like surface EMG and EEG require heavy auxiliary equipment to function (Soufineyestani *et al*., 2020; Jamil *et al*., 2021), making them unsuitable for long-term use.

Another approach, vision-based HIP, is achieved through the collection and real-time algorithmic processing of images of bodily and environmental features of the target personnel. In recent years, vision-based HIP has demonstrated strong potential in various HIP fields (Razali *et al*., 2021; Cui *et al*., 2021; Dani *et al*., 2020). Despite these advances, their application in UTM scenarios remains underexplored and fraught with challenges. (1) Lack of environmental context integration: Most existing vision-based HIP methods primarily focus on human-centric cues such as body gestures (Laplaza *et al*., 2022, Wang *et al*., 2021) and skeletal movements (Li *et al*., 2020; Wei *et al*., 2021) while neglecting critical environmental contexts. Poor lighting, water vapor, steam or dust can rapidly alter visibility and distort the physical context, making it difficult to rely solely on human-generated signals for accurate predictions. (2) Adaptability to environmental variability. In UTM scenarios, vision-based HIP methods encounter unique challenges stemming from the specific lighting arrangements within these environments (Huang *et al*., 2021). Artificial lighting in UTM is typically spaced at fixed intervals (University of Washington Facilities, 2024; Ministry of Housing and Urban-Rural Development of China, 2015), leading to abrupt changes in lighting conditions as robots navigate the tunnels. These rapid shifts in lighting can significantly impair the performance of

vision sensors, affecting the accuracy and robustness of the system's judgments (Rezaei *et al*., 2015). Since most existing methods are designed for stable, well-controlled environments, they cannot reliably handle the rapid lighting changes of UTM scenarios, where high precision is crucial under variable conditions.

To address these challenges, this paper proposes a hybrid intentions prediction model (HIPM) that integrates sophisticated image processing and semantic analysis technologies. Specifically, HIPM combines convolutional neural networks (CNNs) for visual feature extraction, long short-term memory networks (LSTMs) for temporal dynamics and the CLIP (contrastive language-image pre-training) model (Radford *et al*., 2021) for semantic understanding of environmental context. This design enables HIPM to capture not only human-centric cues such as gestures or skeletal movements but also contextual environmental information – including tools, infrastructure and background elements – that are often overlooked by traditional vision-based methods (Guo *et al*., 2021). Moreover, by leveraging time-series image classification, HIPM adapts to abrupt lighting changes and harsh visibility conditions common in utility tunnels, ensuring robust prediction performance (Canto-Perello and Curiel-Esparza, 2013). To enhance classification precision, HIPM further employs contrastive loss alongside cross-entropy loss, refining feature separability. Experimental evaluation on 2,636 images collected from real underground utility tunnels in Jiangsu Province, China, shows that HIPM achieves an average precision of 91.52% and a recall of 91.20%, significantly outperforming state-of-the-art approaches such as CNN + LSTM (88.22%/82.77%) and Vision Transformer (68.25%/70.41%). These results highlight HIPM's practical applicability and its contribution to advancing human–robot collaboration in safety-critical utility tunnel maintenance by improving both intention recognition accuracy and operational efficiency.

2. Related work

Human–robot collaboration (HRC) has emerged as a transformative approach in industrial and hazardous environments, enabling robots to assist human workers in complex and high-risk tasks (Arents *et al*., 2021; Zhang *et al*., 2023). Despite increasing research in HRC, existing methodologies face significant limitations in UTM scenarios, largely due to the inadequacies of HIP, which hinder effective human–robot interaction in these complex environments. This section critically reviews related works, identifies knowledge gaps and establishes the novelty of the proposed approach.

2.1 Human–robot collaboration (HRC) in utility tunnel maintenance

In tunnel scenarios, maintenance personnel face risks such as poor air quality, confined space hazards and exposure to electrical and gas systems (Zhang *et al*., 2020; Bai *et al*., 2020; Sousa and Einstein, 2021). HRC is therefore critical in enabling robots to undertake high-risk tasks such as structural inspections, leak detection and equipment maintenance, thereby reducing human exposure (Shahrour *et al*., 2020; Che *et al*., 2020). HRC can also enhance maintenance efficiency through intelligent task allocation and automation.

However, while there are successful demonstrations – such as robots for natural gas pipe leakage inspection (Liu *et al*., 2019) and crack detection in underground tunnels (Gu *et al*., 2022) – these primarily focus on task substitution or monitoring rather than deep collaboration (Lin and Lukodono, 2021; Pan and Yu, 2024; Wang *et al*., 2021). The lack of robust HIP mechanisms limits robots' ability to anticipate and coordinate with human workers. Thus, the state of the art in HRC for UTM demonstrates technical feasibility but falls short of seamless, collaborative interaction due to insufficient intention prediction capabilities.

2.2 Human intention prediction (HIP) in utility tunnel maintenance

HIP is a core enabler of HRC, allowing robots to anticipate and respond to human actions in dynamic settings (Hoffman *et al*., 2023; Liu and Wang, 2017). Existing HIP methods can be

categorized into physical-based (contact) and vision-based (contactless) approaches (Hoffman *et al.*, 2023; Semeraro *et al.*, 2023).

Physical-based HIP methods: Physical-based methods use wearable sensors such as tactile gloves (Yu *et al.*, 2023a, b; Zou *et al.*, 2023), EMG (Wang *et al.*, 2021; Ison and Artemiadis, 2015), and EEG (Buerkle *et al.*, 2021; Liu *et al.*, 2021). These achieve high accuracy in controlled settings but face severe limitations in UTM due to humidity, water exposure and user discomfort (Niu *et al.*, 2020; Jang *et al.*, 2021; Tai *et al.*, 2022; Apak *et al.*, 2022). Wearables also require frequent calibration, reducing practicality for continuous operations.

Vision-based HIP methods: Vision-based HIP techniques leverage camera-based systems to interpret human gestures, body posture and movement trajectories. Early works relied on static features (Li *et al.*, 2020; Du *et al.*, 2025), while later studies introduced CNN–LSTM hybrids to exploit temporal dynamics (Sarabu and Santra, 2021; Saif *et al.*, 2023) and transformer-based models for intent forecasting (Zhang *et al.*, 2024). These advances significantly improve recognition accuracy. Nevertheless, most vision-based HIP studies remain narrowly focused on human-centric cues, paying little attention to contextual environmental information (e.g. tool usage, infrastructure and background activities), which is crucial in UTM. Table 1 presents a comparative analysis of existing vision-based HIP models and the proposed HIPM model, highlighting their respective features.

2.3 Limitations and research gaps

Despite significant progress in HRC and HIP, existing studies reveal key gaps that hinder their effective deployment in UTM:

Challenges of physical-based HIP methods in UTM: Physical-based HIP methods, such as tactile gloves (Yu *et al.*, 2023a, b; Zou *et al.*, 2023), EMG (Wang *et al.*, 2021; Ison and Artemiadis, 2015) and EEG (Buerkle *et al.*, 2021; Liu *et al.*, 2021), face severe limitations in UTM due to the harsh environmental conditions. The high humidity and dust levels (Ma *et al.*, 2023; Zhai *et al.*, 2024) in utility tunnels negatively affect the reliability and durability of these wearable devices, making them unsuitable for prolonged use. Furthermore, UTM tasks are frequently performed over extended periods and at high frequencies, making the use of heavy auxiliary equipment impractical for such high-intensity applications.

Limitations of vision-based HIP methods in UTM: Existing vision-based HIP approaches struggle to achieve accurate predictions in UTM due to poor lighting conditions and high environmental complexity. Most traditional methods concentrate on human-centric cues such

Table 1. Comparative analysis of vision-based HIP models

Models	Features
Two-stream-based CNNs (Li *et al.*, 2020)	Human action is determined by calculating the L2 distance of the positions of the human joints between frames (CNNs)
CLSTM (Sarabu and Santra, 2021)	Present a two-stream network with two CNNs and Convolution Long-Short Term Memory (CLSTM) (CNN + LSTM) a powerful feature extractor in human action recognition in videos
CLSTDN (Saif *et al.*, 2023)	This research proposes Convolutional Long Short-Term Deep Network (CLSTDN) consists of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for Recognition of human intention
Transformer and Bi-LSTM (Zhang *et al.*, 2024)	extract features from human motion trajectories by analyzing changes in human joint distances with a Transformer and a Bi-LSTM, respectively
GRU-CNN (Du *et al.*, 2025)	proposed multi-channel parallel GRU-CNN neural network combines the temporal analysis capabilities of GRU with the spatial feature extraction strengths of CNN through a weight allocation strategy for trajectory prediction and intention recognition

Source(s): Authors' own work

as body gestures (Wang *et al*., 2021; Vianello *et al*., 2021; Park *et al*., 2021) and skeletal angles (Vianello *et al*., 2021; Duan and Zou, 2024). While effective in controlled environments, these approaches become unreliable in utility tunnels, where visibility is often compromised by dust, steam and low-light conditions. More importantly, they largely overlook contextual environmental information – such as tool usage, surrounding infrastructure and background activities – that could substantially enhance prediction accuracy (Wei *et al*., 2021; Girase *et al*., 2021). In addition, the alternating presence of overly bright and dark areas in UTM further degrades the robustness of vision-based HIP methods (Mukherjee *et al*., 2022), making reliable deployment in such scenarios highly challenging.

Therefore, the state of the art demonstrates promising techniques but lacks context-aware, multimodal HIP approaches capable of robustly handling the unique environmental variability of UTM.

2.4 Novelty and contributions

Building on the identified gaps, this study introduces a hybrid intentions prediction model (HIPM) with the following contributions:

(1) Multimodal deep learning integration: HIPM fuses CNN (visual features), LSTM (temporal dependencies) and CLIP (semantic understanding), addressing the limitation of unimodal methods.

(2) Context-aware HIP: By incorporating environmental cues (tools, infrastructure and background activities) through CLIP-based semantic analysis, HIPM overcomes the neglect of contextual data in prior studies.

(3) Robustness to lighting variability: Time-series image classification with uniform frame sampling ensures stable predictions despite abrupt lighting fluctuations.

(4) Empirical validation in real tunnels: HIPM is trained and evaluated on 2,636 images from real UTM environments, achieving superior precision and recall compared with CNN + LSTM and transformer baselines.

In this way, this study directly addresses the key limitations of prior HIP methods, advances the state of the art and provides a practical, validated framework for safe and efficient HRC in utility tunnels.

3. Methodology

The human intention prediction model (HIPM) proposed in this study is a hybrid deep learning model specifically designed for UTM scenarios.

3.1 Model architecture

The model architecture integrates convolutional neural network (CNN), CLIP (contrastive language–image pre-training) model and long short-term memory (LSTM) networks to harness both visual and semantic features for HIP classification performance, as shown in Figure 1. The framework operates as follows:

3.1.1 Input. The input for the HIPM model is key frames uniformly extracted from continuous video footage captured from a simulated robot's perspective during the UTM process. These images are resized to a uniform dimension of 224 × 224 pixels to match the input size expectation for the image and semantic feature extraction part. Then they are converted to PyTorch tensors and transferred into two different processing streams.

3.1.2 Parallel feature extraction. The feature extraction process leverages two powerful models: ResNet-18 and CLIP. In this model, the classic ResNet-18 provides the general image features needed for the HIP classification task. The primary role of CLIP is to offer additional

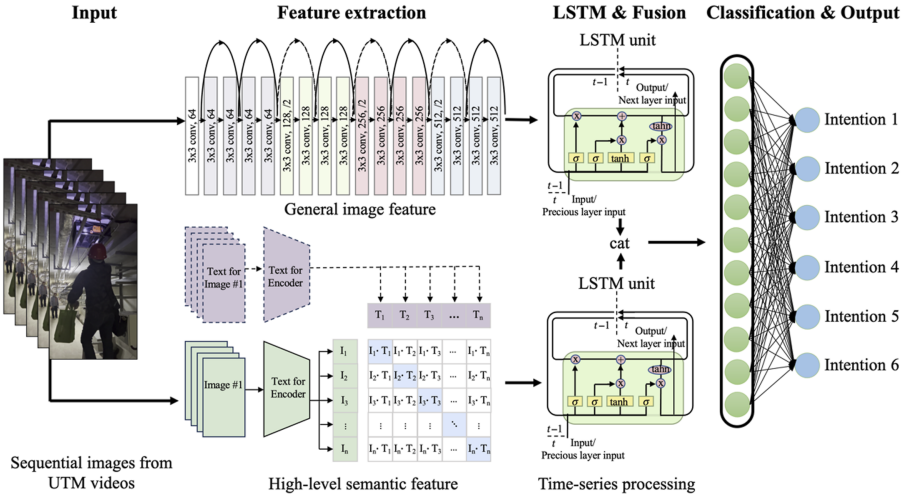

Figure 1. Model architecture of the HIPM. Source: Authors' own work

semantic features, enhancing the model's feature representation capability by combining with the traditional CNN (ResNet-18) features.

ResNet-18 for general image features: ResNet-18 is a widely used CNN architecture known for its balance between computational efficiency and strong feature extraction capabilities (Pandey and Srivastava, 2023). Compared to deeper architectures like ResNet-50 or ResNet-101, ResNet-18 offers faster inference while maintaining competitive accuracy, making it suitable for real-time processing in UTM scenarios (Ou *et al*., 2019). In the HIPM, the ResNet-18 is initialized with weights pretrained on the ImageNet dataset (Deng *et al*., 2009), which helps in capturing general visual patterns and textures. Then, the network layers extract hierarchical features, starting from low-level features like edges and textures. The final feature vectors obtained from ResNet-18 represent the basic visual content of the images for the human intention prediction task. And the fully connected layer was excluded to focus solely on feature extraction.

CLIP for high-level semantic features: The CLIP model is pretrained on a large dataset of 400 m images and their associated textual descriptions (Radford *et al*., 2021), allowing it to learn rich semantic representations that go beyond low-level visual patterns. Given the complexity of UTM environments, where human actions are influenced by tools and background context, CLIP's ability to capture semantic relationships enhances intention prediction accuracy compared to purely vision-based models (Du *et al*., 2025; Li *et al*., 2020; Saif *et al*., 2023). This training allows CLIP to learn a shared embedding space for images and text. In HIPM, the input images are processed concurrently through ResNet-18 and CLIP for feature extraction. Unlike the general image features extracted by ResNet-18, CLIP captures high-level semantic features. CLIP's training involves visual and textual data, allowing it to understand and embed images in a contextually rich and semantically meaningful way (Krojer *et al*., 2022). The CLIP's embedding captures the underlying concepts and relationships that are present in the images. This dual understanding – combining visual patterns with contextual semantics – enables CLIP to generate features that encompass not only just what the image looks like but also what it represents or means within a broader context. For example, a traditional CNN might excel at identifying the shapes and colors within an image of a cat, whereas CLIP, due to its multimodal training, would recognize the image as a "cat" and understand related concepts such as "pet," "feline" and "animal." This semantic embedding makes CLIP particularly powerful for tasks that benefit from a deeper understanding of image content beyond mere visual details (Sain *et al*., 2023).

By integrating these features, the model benefits from the strengths of both approaches, resulting in a richer and more comprehensive representation of the input data. Multimodal learning has been shown to improve classification accuracy in complex real-world environments (Bayoudh *et al*., 2022), making it particularly beneficial for HIP in UTM scenarios, where environmental factors and human motion cues interact dynamically (Liu *et al*., 2019; Gu *et al*., 2022).

3.1.3 Long-shot term memory and intention prediction. The LSTM network is designed to capture temporal dependencies and patterns in sequential data, which is particularly useful for tasks involving time series or sequences of images. By processing the features through LSTMs, the model can learn to recognize how features evolve over time or across frames in a video sequence (Ullah *et al*., 2017).

In the HIMP, the LSTM networks play a crucial role in processing and understanding sequential data. The visual features from ResNet-18 and the semantic features from CLIP are each fed into their respective LSTM layers. The LSTM layers process these features, retaining relevant information through their gating mechanisms, which help in handling long-term dependencies and mitigating the vanishing gradient problem (Waqas and Humphries, 2024). The outputs of the LSTM layers are then concatenated, combining the temporal dynamics of both the visual and semantic features. This combined representation is richer and more informative, allowing the model to make more accurate predictions by leveraging detailed visual patterns and the high-level semantic context over the sequence.

3.1.4 Loss functions. The HIPM model employed two loss functions as follows:

Cross-entropy loss: Employed for the primary classification task, this loss measures the discrepancy between the predicted probabilities and the actual class labels. For a multi-class classification problem in this study, the cross-entropy loss can be defined as formula (1). Where C is the total number of classes. y_i is the true label for class i, with $i = 1$ indicating that the sample belongs to the class and 0 indicating it does not. p_i is probability predicted by the model that the sample belongs to class i.

$$L_1 = -\sum_{i=1}^{C} y_i \log(p_i) \tag{1}$$

Contrastive loss: This custom-defined loss function enhances the discriminative power of the model by encouraging the minimization of distances between similar pairs and maximization between dissimilar pairs. It is computed based on the Euclidean distance between image and semantic feature representations, as shown in formula (2). Where, DW is the Euclidean distance between the feature representations of two samples. Y is a binary label associated with the pair, if the pair is dissimilar (from different classes), $Y = 1$ and if the pair is similar (from the same class), $Y = 0$. M is a margin, a hyperparameter that defines how far apart the dissimilar pairs should be pushed in the feature space. The margin acts as a threshold beyond which no further loss is accumulated for dissimilar pairs.

$$L_2 = (1 - Y)\frac{1}{2}(DW)^2 + Y\frac{1}{2}max\,(0, M - DW)^2 \tag{2}$$

3.2 Human–robot collaboration intention classification

The essence of vision-based human intention prediction is the classification of images for human-robot collaboration. Therefore, the classification of human intentions in targeted scenarios is crucial. This section determines the specific categories and descriptions for the HIPM by analyzing the daily operational processes of UTM from ISO (International Organization for Standardization) standard (smart community infrastructures – operation and maintenance of utility tunnels) [61].

3.2.1 Maintenance processes. Routine monitoring and inspection. The operation of utility tunnels involves regular monitoring and inspection to ensure the functionality and safety of

infrastructure. Human workers perform checks on electrical systems, structural integrity and other critical components (e.g. induced draft fan and water pump).

Emergency responses. Maintenance teams are also tasked with responding to emergencies such as system malfunctions or safety hazards. These situations require quick, precise actions to mitigate risks and restore normal operations. Repair and upkeep. Scheduled and unscheduled repairs form a significant part of maintenance work, involving the replacement of worn-out parts or the upgrading of systems to ensure continued reliability and efficiency. Logistical operations. This includes the handling and transportation of materials and tools necessary for maintenance tasks, which must be done efficiently to minimize downtime and ensure that workers have the necessary resources when and where they need them.

The operational needs and scenarios identified in these maintenance processes naturally lead to the formulation of specific human–robot collaboration intentions. These intentions are categorized to streamline interactions and enhance mutual understanding between humans and robots during maintenance operations. The categories are specifically designed to cover a comprehensive range of activities that robots are expected to understand and act upon, ensuring that all aspects of maintenance are addressed efficiently and safely.

3.2.2 Collaboration intentions. The intention classifications of HIPM need to provide precise judgment criteria for robots to support an efficient UTM operational process while also conforming to the characteristics of robot (computer) vision judgment to ensure the reliability of intention recognition. In the classification of the HIPM for engineering practices, the following key factors must be considered: (1) Coverage of entire processes: The classification should ensure coverage of all key stages of maintenance and operation to effectively support human workers under all circumstances. It should be detailed enough to identify the needs of all maintenance operations, such as inspection, repair and material handling, to optimize the efficiency and effectiveness of human–robot interaction. (2) Granularity of classification: The level of detail in classification needs to be sufficient to distinguish between different operational demands and contextual environments. For instance, there should be clear distinctions between tasks that require precise operations, like specific repairs and broader tasks like material transportation or patrolling. This helps the robot to recognize and execute tasks more accurately, reducing errors or unnecessary operations, thereby ensuring safety and optimal use of resources. However, classifying scenarios down to the level of specific operational procedures would be exceedingly extensive. Detailed subdivision of procedures with high repetitiveness is not the focus of this study. Therefore, this study only classifies at a coarser granularity to validate the feasibility of the HIPM methodology. (3) Feasibility: The feasibility of the classification should fully consider the complexity of the UTM scenarios and the limitations of current robotic vision capabilities. The classification should be realistic, considering the current state of technology and the practical challenges faced in tunnel environments. It must be designed with a pragmatic approach, ensuring that the tasks assigned to robots are within their operational capabilities while being sufficiently robust to handle the unpredictability of such environments. Additionally, the classification scheme should be flexible enough to adjust and optimize as technology evolves.

Considering these factors, this study divides the specific human operational intentions in HIPM into the following six categories: (1) Currently operating, please wait; (2) need parts/tools, please hand over to me; (3) hold parts/tools for me; (4) assistance needed, please help support; (5) Need to move to the next maintenance point, please follow and (6) No personnel present, please search for personnel. Table 2 provides detailed descriptions of the specific scenarios for each category. Plate 1 illustrates the examples for these different intention scenarios.

3.3 Evaluation metrics

Evaluating the performance of the HIPM model goes beyond a single metric, incorporating these metrics to gain a holistic understanding of the model's behavior in practical applications.

Table 2. Detailed descriptions of each intention classification in HIPM

No.	Intention classification	Specific description
1	Currently operating, please wait	The human worker needs uninterrupted focus for a precise task. The robot should monitor from a distance, avoiding interference, until the operation is complete or assistance is requested
2	Need parts/tools, please hand over to me	The human worker requires specific tools or parts immediately to continue their work without leaving the work site. The robot should promptly deliver these items directly to the worker's location/hand
3	Hold parts/tools for me	During complex tasks, the human worker may need to free up their hands. The robot should act as an additional support by holding or storing tools and materials temporarily, keeping them organized and readily accessible
4	Assistance needed, please help support	The human worker requires additional physical support for tasks that are risky or require multi-hand coordination. The robot should provide the necessary support by stabilizing equipment or assisting in holding materials securely
5	Need to move to the next maintenance point, please follow	The human worker is moving to another maintenance area and requires the robot to carry tools and equipment to the next point or lead the way. The robot should ensure all necessary materials are transported efficiently and safely following the worker
6	No personnel present, please search for personnel	No human worker in the view of robot, the robot should initiate a search protocol to find the worker, ensuring that all personnel are safe and follow human workers closely

Source(s): Authors' own work

Plate 1. Examples of the different intention scenarios. Source: Authors' own work

To thoroughly assess the performance of the HIPM, this study utilizes four popular statistical metrics to evaluate the model's accuracy, robustness and reliability. The evaluation metrics applied include *Precision*, *Recall*, *F1 score* and AUC (Area Under the *ROC (Receiver Operating Characteristic) Curve)*.

The *Precision*, *Recall*, *F1 score* are derived from the counts of samples classified into four categories by the model: True Positive (*TP*), True Negative (*TN*), False Positive (*FP*) and False Negative (*FN*), based on the actual classification results, as shown in Table 3. The specific formulas used to calculate *Precision*, *Recall* and *F1 score* are presented in equations (3, 4, and 5), respectively. The ROC curve is generated by varying the decision threshold and computing the True Positive Rate (TPR) and False Positive Rate (FPR) at each point. It plots FPR against

Table 3. Definition of *TP*, *TN*, *FP* and *FN*

		Predicted class Positive	Negative
Actual class	Positive	True Positive (*TP*)	False Negative (*FN*)
	Negative	False Positive (*FP*)	True Negative (*TN*)

Source(s): Authors' own work

TPR to visualize the model's discrimination ability. Considering is a multi-class classification model, the One-vs-Rest (OvR) approach was applied, where a separate ROC curve is plotted for each class by treating it as the positive class while considering all other classes as negative. The final multi-class AUC is obtained by averaging the AUC values across all classes, using either macro or weighted averaging.

$$Precision = TP / (TP + FP) \tag{3}$$

$$Recall = TP / (TP + FN) \tag{4}$$

$$F1\ Score = 2 * Precision * Recall / (Precision + Recall) \tag{5}$$

4. Experiments

4.1 Data collection and processing

For the experimental part of this study, 2,636 frames were collected in a practical utility tunnel environment in Jiangsu Province, China. The dataset includes key frames extracted from continuous video streams captured from a simulated robot perspective, covering various maintenance activities and worker interaction scenarios. Figure 2 illustrates the data collection and processing steps. To ensure dataset variability, data was collected under diverse environmental conditions, including variations in lighting, worker positions and tunnel layouts. Furthermore, temporal sequences were incorporated to allow the model to capture dynamic interactions, enhancing its adaptability to real-world UTM scenarios.

4.1.1 Collection steps. Scene selection: Multiple typical tunnel maintenance scenes were selected to ensure a wide range of environmental conditions and tasks, including inspections of lighting and ventilation systems, equipment and structural repairs and emergency handling, as mentioned in section 3.2.

Video recording: High-resolution cameras were used to record from the simulated robot's working perspective (manually). Each video lasted at least 30 min, ensuring the capture of comprehensive action sequences and interaction details. Key frame extraction: Key frames

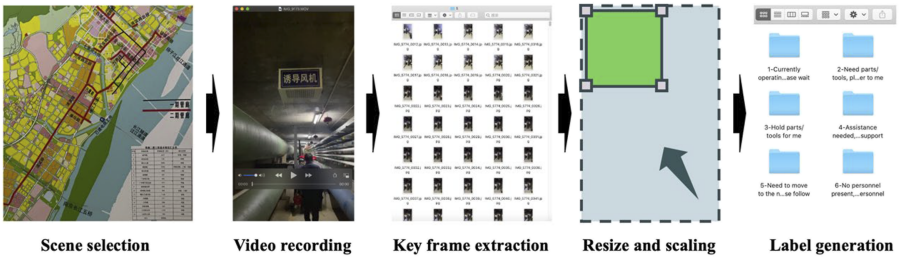

Figure 2. Data collection and processing. Source: Authors' own work

were uniformly extracted from the recorded videos. The selection of key frames was based on the importance and representativeness of the activities, with at least five frames extracted for each key action to capture sufficient action information.

4.1.2 Data processing. Image preprocessing: The extracted key frames were resized to ensure consistency and quality of the input data. Additionally, the images were normalized, scaling pixel values to between 0 and 1 to reduce the impact of lighting variations and background differences on model training. Label generation: Each frame was annotated by professional tunnel maintenance engineers, defining the human work intentions present. Labels might include six categories: (1) Currently operating, please wait; (2) Need parts/tools, please hand over to me; (3) Hold parts/tools for me; (4) Assistance needed, please help support; (5) Need to move to the next maintenance point, please follow and (6) No personnel present; please search for personnel.

This thorough data collection and processing method resulted in a high-quality and representative dataset, providing a solid foundation for subsequent model training and validation. These data not only help in training robots to better understand and predict human work intentions but also provide valuable experimental material for studying interaction patterns between humans and robots in complex environments.

4.2 Experiment results

To evaluate the effectiveness of the proposed HIPM model, a series of controlled experiments were conducted by comparing it with CNN, LSTM, CNN + LSTM and Vision Transformer. The models were implemented in PyTorch and trained using an NVIDIA RTX 3090 GPU with 24 GB VRAM. The experimental setup was designed to ensure fair comparisons across all models.

Hyperparameter tuning and model training: (1) Dataset split: The dataset was divided into 80% training, 10% validation and 10% test to ensure robust model evaluation. (2) Optimization algorithm: All models were trained using the Adam optimizer with an initial learning rate of 0.0001. (3) Batch size and training duration: All models were trained for 50 epochs with a batch size of 32.

Model selection justification. The baseline models used for comparison were selected based on findings from our literature review, which indicates that vision-based HIP models predominantly utilize CNN, CNN + LSTM or transformer-based architectures. Several studies have demonstrated the effectiveness of these architectures in intention recognition and HRC: (1) CNN-based models have been widely used in HIP due to their strong spatial feature extraction capabilities for recognizing human gestures and postures (Wang *et al.*, 2021; Vianello *et al.*, 2021; Duan and Zou, 2024). (2) LSTM models have been introduced to enhance intention prediction by capturing temporal dependencies in human motion sequences (Waqas and Humphries, 2024). (3) CNN + LSTM hybrids are frequently employed in HIP research, combining spatial and temporal features to improve prediction accuracy (Ravipati *et al.*, 2023). (4) Vision Transformers represent a newer approach in HIP, leveraging self-attention mechanisms to model complex relationships between human actions (Domínguez-Vidal and Sanfeliu, 2024).

Given the prevalence of these architectures in prior studies, our experimental comparison against CNN, LSTM, CNN + LSTM and Vision Transformer ensures that HIPM is benchmarked against the most relevant and widely adopted HIP methodologies. By validating our model against these baselines, we provide a fair assessment of HIPM's advantages in multimodal feature integration and adaptability to dynamic UTM environments.

4.2.1 Experimental setup. The experiment compared several mainstream deep learning models. The models were evaluated using *Precision*, *Recall* and *F1 score* to evaluate their performance.

4.2.2 Model performance comparison. To more comprehensively evaluate the performance of various models in predicting tunnel maintenance worker intentions, comparisons were

made among CNN, LSTM, CNN + LSTM, Vision Transformer and the proposed HIPM model. Table 4 and Figure 3 shows the performance of each model in the experiment. The following provides a detailed analysis of each model's performance:

CNN: Leveraging its robust image processing capabilities, CNN performed well in this image classification tasks, achieving a precision of 80.33 and 77.15% of recall. However, due to its lack of ability to handle temporal data, its performance in dynamic scenes was limited, with an F1 score of 75.89%. CNN achieved an AUC of 92.61%, significantly higher than LSTM (77.15%). This indicates that despite some Recall limitations, CNN maintains strong overall classification performance with a robust ability to distinguish between worker intention classes.

LSTM: As a typical model for handling temporal data, LSTM has unique advantages in capturing behavioral dynamics. Nevertheless, the LSTM model's precision was only 47.79% and 59.74% of recall. This may be because a standalone LSTM struggles to extract sufficient

Table 4. Performance comparison of various deep learning models across precision, recall, F1 score and AUC metrics. The table highlights the effectiveness of different architectures, including CNN, LSTM, their combination (CNN + LSTM), Vision Transformer and the proposed HIMP model. The HIMP model outperforms all others, achieving the highest scores in all evaluation metrics, particularly with an F1 score of 91.20% and an AUC of 95.36%

Models	Precision (%)	Recall (%)	F1 score (%)	AUC (%)
CNN	80.33	77.15	75.89	92.61
LSTM	47.79	59.74	51.92	77.15
CNN + LSTM	88.22	82.77	84.15	91.11
Vision Transformer	68.25	70.41	68.39	84.55
HIMP (our)	*91.52*	*91.20*	*91.20*	*95.36*

Source(s): Authors' own work

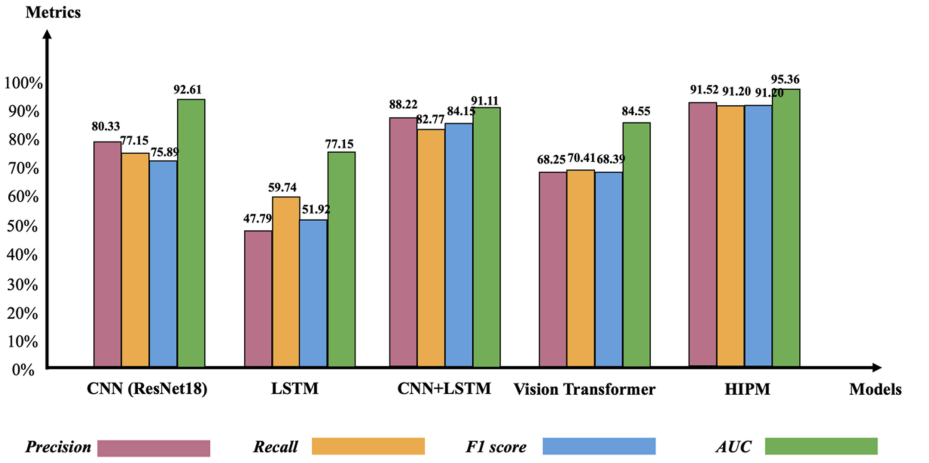

Figure 3. Performance comparison of different deep learning models – CNN, LSTM, CNN + LSTM, Vision Transformer and the proposed HIMP model – evaluated using precision, recall, F1 score and AUC. The bar chart visually illustrates the effectiveness of each model, with the HIMP model demonstrating superior performance across all metrics, achieving the highest F1 score (91.20%) and AUC (95.36%). These results highlight the advantage of the HIMP model over traditional architectures in this task. Source: Authors' own work

spatial features from static images. LSTM achieved an AUC of 77.15%, the lowest among all models. This suggests that its decision boundary is unstable, likely because it does not have strong spatial feature extraction capabilities, making it prone to misclassifications.

CNN + LSTM: Combining the spatial feature extraction capabilities of CNN with the temporal sequence processing abilities of LSTM, this hybrid model significantly improved performance, achieving a precision of 88.22% and an F1 score of 84.15%. This indicates that the combined model outperforms single models in handling complex scenarios involving both temporal and spatial information, which is a very popular approach to improve. The CNN + LSTM model achieved an AUC of 91.11%, slightly lower than CNN alone but still significantly better than LSTM. This suggests that while CNN + LSTM enhances sequential understanding, the temporal dependencies introduced by LSTM slightly impact its overall decision boundary stability.

Vision Transformer: As a popular emerging model that utilizes self-attention mechanisms for image tasks, the Vision Transformer underperformed compared to the CNN + LSTM combination in this experiment. It achieved a precision of 68.25% and an F1 score of 68.39%. This could be attributed to the Vision Transformer's requirement for large amounts of data to reach optimal performance, which might not have been met in the data-limited experimental setup. Vision Transformer obtained an AUC of 84.55%, which is lower than CNN and CNN + LSTM. This suggests that its overall classification capability is weaker and that its decision boundaries fluctuate more. This is consistent with the well-known limitation of transformers: they require large-scale data to reach their full potential, which was not fully met in this experiment.

HIPM: The proposed HIPM model performed exceptionally well on the dataset, with all metrics exceeding 91%. This demonstrates that the HIPM model effectively integrates temporal and spatial features, enhancing the accuracy and robustness of intention prediction. HIPM achieved the highest AUC of 95.36%, outperforming all other models. This suggests that HIPM not only classifies samples accurately but also maintains the most stable decision boundary, even in varied data distributions. A higher AUC implies that HIPM is more reliable in distinguishing between worker intentions, making it the most robust model for practical applications.

4.3 Summary of results

The experiment results showed that the proposed HIPM model significantly outperforms all other comparison models across all metrics, achieving over 91% in *Precision*, *Recall* and *F1 score*. This underscores the HIPM model's excellent performance in predicting human–robot interaction intentions in tunnel maintenance scenarios. By combining the strengths of traditional convolutional neural networks and long short-term memory networks and incorporating contrastive loss functions to optimize the learning process with the contrastive language-image pre-training component, the HIPM model effectively enhances classification accuracy and robustness in the human intention prediction task under the UTM scenario.

5. Discussion

5.1 Why HIMP works?

The proposed HIPM model represents a technological advancement in predicting human intentions in hazardous and dynamic UTM environments. By integrating CNN, CLIP and LSTM, HIPM achieves a synergistic fusion of visual, semantic and temporal information. This section critically examines the interactions among these components and compares HIPM's effectiveness with existing methodologies.

Complementary strengths: Visual, semantic and temporal dimensions. The integration of CNN, CLIP and LSTM allows HIPM to process data across multiple dimensions. CNNs are well-documented for their ability to extract detailed visual features (Wang *et al.*, 2021;

Vianello *et al*., 2021; Duan and Zou, 2024), making them essential for identifying objects and contextual elements within images. However, CNNs alone often lack higher-order semantic understanding, a limitation that CLIP effectively mitigates. CLIP enhances CNN's output by aligning visual features with textual semantics (Radford *et al*., 2021), which is crucial for understanding human activities and their implicit intentions. LSTM further augments this framework by capturing temporal dependencies in sequences of images, a technique proven effective in activity recognition (Waqas and Humphries, 2024). The combination of these methods enables HIPM to achieve a level of interpretability and predictive accuracy beyond what each component could provide individually.

Fusing information for robust predictions. Unlike traditional ensemble approaches that simply concatenate features, HIPM integrates CNN, CLIP and LSTM outputs in a manner that maximizes complementary strengths. CNNs often struggle with ambiguous or occluded visuals (Nath and Behzadan, 2020), a challenge mitigated by CLIP's ability to infer missing information through semantic understanding. Moreover, CLIP + CNN alone achieves high precision (88.60%), but the absence of temporal awareness, as seen in the CLIP + LSTM combination (25.95% precision), highlights the necessity of all three components. Our findings, summarized in Table 5 and Figure 4, confirm that the holistic fusion in HIPM significantly outperforms other model combinations, achieving a precision of 91.52%, a recall of 91.20%, F1 score of 91.20% and an AUC of 95.36%. These results align with existing literature demonstrating that hybrid models incorporating both spatial and temporal feature extraction outperform static models (Tripathi and Verma, 2023).

A comparative analysis further underscores the necessity of HIPM's design. While CNN + LSTM (precision: 88.22%) captures spatial and temporal features well, the lack of semantic integration leads to suboptimal performance in ambiguous environments. Similarly, CLIP + CNN, despite performing well, does not account for sequential dependencies, limiting its applicability in dynamic environments where intention evolves over time. These comparisons illustrate how HIPM's architectural synergy provides superior predictive capabilities.

5.2 Limitations

The HIPM model, despite its effectiveness, has several limitations. First, data dependency. The model's performance is highly dependent on data quality and diversity. Regarding dataset size, while 2,636 images may seem limited, prior studies in human intention recognition and industrial task prediction have successfully trained deep learning models on smaller or similar

Table 5. Performance comparison of different model combinations, including CNN, LSTM, CLIP and their hybrid architectures, evaluated using precision, recall, F1 score and AUC. The table demonstrates how different combinations impact performance, with models like CLIP + LSTM showing lower scores, whereas CNN + LSTM and CLIP + CNN provide improvements. The proposed HIMP model outperforms all others, achieving the highest scores across all metrics, particularly with an F1 score of 91.20% and an AUC of 95.36%. These results emphasize the robustness of HIMP compared to traditional and hybrid architectures

Models	Precision (%)	Recall (%)	F1 score (%)	AUC (%)
CNN	80.33	77.15	75.89	92.61
LSTM	47.79	59.74	51.92	77.15
CLIP	45.72	53.00	48.00	62.51
CLIP + LSTM	25.95	50.94	34.38	53.44
CLIP + CNN	88.60	88.85	88.40	93.56
CNN + LSTM	88.22	82.77	84.15	91.11
HIMP (our)	*91.52*	*91.20*	*91.20*	*95.36*

Source(s): Authors' own work

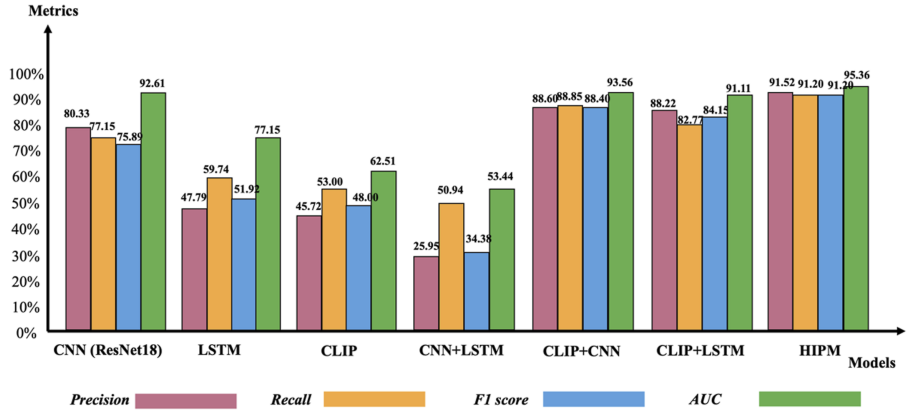

Figure 4. Performance comparison of different model combinations – CNN, LSTM, CLIP and their hybrid architectures (CLIP + CNN, CLIP + LSTM, CNN + LSTM and HIMP) – evaluated using precision, recall, F1 score and AUC. The bar chart visually highlights the impact of different model architectures on performance, with hybrid models demonstrating varying levels of effectiveness. The CLIP + LSTM model shows relatively lower performance, whereas CNN + LSTM and CLIP + CNN exhibit stronger results. The proposed HIMP model achieves the highest scores across all metrics, notably with an F1 score of 91.20% and an AUC of 95.36%, confirming its superiority over other combinations. Source: Authors' own work

dataset sizes (Kong and Fu, 2022). However, this dependency introduces operational risks in UTM environments, where unexpected conditions such as structural anomalies, water leakage or unique tunnel configurations may challenge the model's adaptability (Zhou *et al.*, 2023). To mitigate these risks, domain adaptation techniques and continuous model retraining with diverse datasets should be considered.

Second, interpretability. The HIPM model's complex structure creates a "black box" effect, limiting transparency and reducing trust in its predictions (Hassija *et al.*, 2024). In practical UTM environments, this lack of interpretability poses operational risks, as maintenance workers may be reluctant to rely on AI-driven assessments without clear justification. A lack of understanding of why the model predicts certain intentions or risks may lead to hesitation, misinterpretation or even human override of correct AI recommendations, potentially compromising safety and efficiency. To address this, explainable AI techniques should be integrated into the system to provide real-time, user-friendly explanations. Methods such as feature importance analysis (e.g. SHAP (Lundberg, 2017)) and saliency-based methods (e.g. Grad-CAM (Selvaraju *et al.*, 2017)) can help clarify how the model arrives at its predictions. Additionally, human factors influencing robot acceptance should be carefully considered. Training programs that educate maintenance workers on interpreting AI outputs, along with visualization tools that allow for HRC feedback loops, can help build trust and adoption (Pinto *et al.*, 2022).

Third, computational complexity. The hybrid model's computational demands are high, potentially restricting real-time applications or use in environments with limited resources. In practical UTM operations, where efficiency is critical, excessive computation time could delay maintenance tasks or require costly high-performance hardware. Potential mitigation strategies include model optimization techniques such as edge computing integration (Zhou *et al.*, 2020) to ensure faster inference while maintaining accuracy.

6. Conclusion and future works

This study introduced HIPM, a novel hybrid deep learning model designed to enhance human intention prediction in UTM through an integrated approach combining CNNs, LSTM and the

CLIP model. By leveraging both image features and high-level semantics, HIPM effectively addresses key challenges in existing HIP methods, particularly in environmental adaptability and multimodal feature fusion.

The experimental results demonstrate that HIPM significantly outperforms mainstream models, achieving over 91% in precision, recall and F1 score. These findings underscore its superior predictive capabilities in complex and hazardous UTM environments, where traditional vision-based methods often struggle due to lighting variability, environmental complexity and the lack of contextual information integration. By incorporating temporal dynamics and semantic context, HIPM ensures more robust and reliable intention recognition compared to conventional approaches. This research contributes to both theoretical advancements and practical applications in HRC: Theoretically, HIPM integrates multimodal deep learning techniques to enhance the accuracy and robustness of human intention prediction in dynamic environments, setting a new benchmark for future research. Practically, HIPM improves the responsiveness of robotic assistants in UTM settings, enhancing safety by reducing human exposure to hazards while optimizing task efficiency and collaboration.

To improve the HIPM model for utility tunnel maintenance, future research could focus on two key areas: (1) Explainability and trust: Integrate explainable AI techniques, such as attention mechanisms and saliency maps, to make HIPM's predictions more interpretable. This will help build trust with maintenance workers by clarifying the model's decision-making process, thus enhancing human-robot collaboration. Specifically, as shown in Figure 5, SHAP (Lundberg, 2017) could be used to quantify the impact of input features on predictions, while Grad-CAM (Selvaraju *et al.*, 2017) could visualize the most influential regions in image-based decision-making. Integrating attention maps in the LSTM component would help illustrate how temporal dependencies contribute to human intention recognition. Additionally, incorporating user-friendly visualization tools that provide real-time feedback on model predictions could enhance human trust and facilitate better human–robot interaction. (2) Human–robot interaction studies: Conduct field trials and user studies to assess HIPM's effectiveness in real-world settings. Understanding worker interactions, perceptions of the model, and impacts on efficiency and safety will provide insights for refining the model and optimizing its role in maintenance.

In conclusion, HIPM represents a significant step forward in improving safety, efficiency and collaboration in hazardous maintenance environments. By bridging the gap between

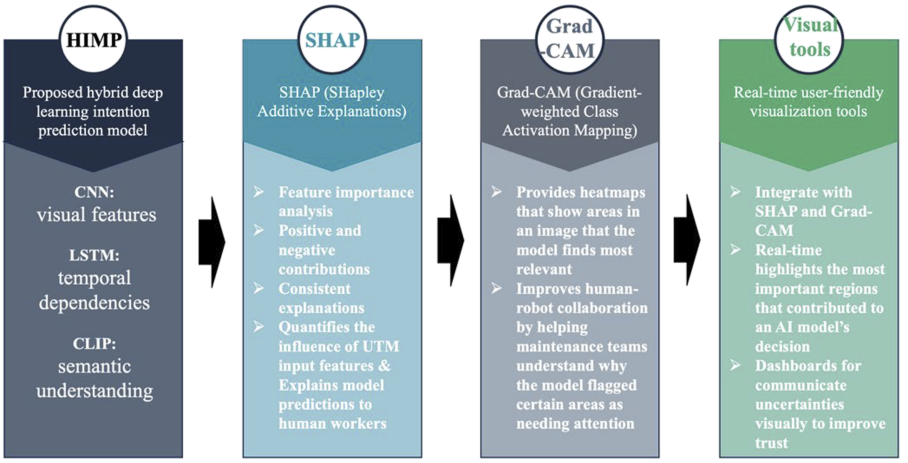

Figure 5. Roadmap for future work regarding interpretability

human intention prediction and robotic assistance, this research paves the way for more adaptive, intelligent and reliable HRC systems in industrial settings, ultimately contributing to the advancement of intelligent robotics in safety-critical operations.

References

Apak, M.Y., Ozen, H., Calis, M., Golgeli, B. and Ataoglu, S. (2022), "Applications of utility tunnels for natural gas pipelines", *Tunnelling and Underground Space Technology*, Vol. 122, 104243, doi: 10.1016/j.tust.2021.104243.

Arents, J., Abolins, V., Judvaitis, J., Vismanis, O., Oraby, A. and Ozols, K. (2021), "Human–robot collaboration trends and safety aspects: a systematic review", *Journal of Sensor and Actuator Networks*, Vol. 10 No. 3, p. 48, doi: 10.3390/jsan10030048.

Bai, Y., Zhou, R. and Wu, J. (2020), "Hazard identification and analysis of urban utility tunnels in China", *Tunnelling and Underground Space Technology*, Vol. 106, 103584, doi: 10.1016/j.tust.2020.103584.

Bayoudh, K., Knani, R., Hamdaoui, F. and Mtibaa, A. (2022), "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets", *The Visual Computer*, Vol. 38 No. 8, pp. 2939-2970, doi: 10.1007/s00371-021-02166-7.

Buerkle, A., Eaton, W., Lohse, N., Bamber, T. and Ferreira, P. (2021), "EEG based arm movement intention recognition towards enhanced safety in symbiotic human-robot collaboration", *Robotics and Computer-Integrated Manufacturing*, Vol. 70, 102137, doi: 10.1016/j.rcim.2021.102137.

Büscher, G.H., Kõiva, R., Schürmann, C., Haschke, R. and Ritter, H.J. (2015), "Flexible and stretchable fabric-based tactile sensor", *Robotics and Autonomous Systems*, Vol. 63, pp. 244-252, doi: 10.1016/j.robot.2014.09.007.

Canto-Perello, J. and Curiel-Esparza, J. (2013), "Assessing governance issues of urban utility tunnels", *Tunnelling and Underground Space Technology*, Vol. 33, pp. 82-87, doi: 10.1016/j.tust.2012.08.007.

Che, H., Shi, C., Hu, H., Wang, W., Ren, F., Li, J. and Xu, X. (2020), "Research on navigation and location system of inspection robot for urban utility tunnel", *2020 4th International Conference on Robotics and Automation Sciences (ICRAS)*, pp. 11-14, doi: 10.1109/ICRAS49812.2020.9135051.

Cui, A., Casas, S., Sadat, A., Liao, R. and Urtasun, R. (2021), "Lookout: diverse multi-future prediction and planning for self-driving", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16107-16116, available at: https://openaccess.thecvf.com/content/ICCV2021/papers/Cui_LookOut_Diverse_Multi-Future_Prediction_and_Planning_for_Self-Driving_ICCV_2021_paper.pdf

Dani, A.P., Salehi, I., Rotithor, G., Trombetta, D. and Ravichandar, H. (2020), "Human-in-the-loop robot control for human-robot collaboration: human intention estimation and safe trajectory tracking control for collaborative tasks", *IEEE Control Systems Magazine*, Vol. 40 No. 6, pp. 29-56, doi: 10.1109/MCS.2020.3019725.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009), "Imagenet: a large-scale hierarchical image database", *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Domínguez-Vidal, J.E. and Sanfeliu, A. (2024), "Exploring transformers and visual transformers for force prediction in human-robot collaborative transportation tasks", *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3191-3197, doi: 10.1109/ICRA57147.2024.10611205.

Du, J., Lu, D., Li, F., Liu, K. and Qiu, X. (2025), "Trajectory prediction and intention recognition based on CNN-GRU", *IEEE Access*, Vol. 13, pp. 26945-26957, doi: 10.1109/ACCESS.2025.3539931.

Duan, K. and Zou, Z. (2024), "Morphology agnostic gesture mapping for intuitive teleoperation of construction robots", *Advanced Engineering Informatics*, Vol. 62, 102600, doi: 10.1016/j.aei.2024.102600.

Girase, H., Gang, H., Malla, S., Li, J., Kanehara, A., Mangalam, K. and Choi, C. (2021), "Loki: long term and key intentions for trajectory prediction", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9803-9812, available at: https://openaccess.thecvf.com/content/ICCV2021/papers/Girase_LOKI_Long_Term_and_Key_Intentions_for_Trajectory_Prediction_ICCV_2021_paper.pdf

Gu, J., Wu, L., Chen, J., Cai, R., Wan, H., Shi, W. and Lv, X. (2022), "Intelligent monitoring of subsidence cracks in underground power utility tunnel", *Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing*, Vol. 12166, pp. 848-852, doi: 10.1117/12.2616361.

Guo, B.H., Zou, Y., Fang, Y., Goh, Y.M. and Zou, P.X. (2021), "Computer vision technologies for safety science and management in construction: a critical review and future research directions", *Safety Science*, Vol. 135, 105130, doi: 10.1016/j.ssci.2020.105130.

Hai, N., Gong, D. and Dai, Z. (2024), "Target spectrum-based risk analysis model for utility tunnel O&M in multiple scenarios and its application", *Reliability Engineering and System Safety*, Vol. 242, 109777, doi: 10.1016/j.ress.2023.109777.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. and Hussain, A. (2024), "Interpreting black-box models: a review on explainable artificial intelligence", *Cognitive Computation*, Vol. 16 No. 1, pp. 45-74, doi: 10.1007/s12559-023-10179-8.

Hoffman, G., Bhattacharjee, T. and Nikolaidis, S. (2023), "Inferring human intent and predicting human action in human–robot collaboration. Annual review of control", *Robotics and Autonomous Systems*, Vol. 7 No. 1, pp. 73-95, doi: 10.1146/annurev-control-071223-105834.

Huang, M.Q., Ninić, J. and Zhang, Q. (2021), "BIM, machine learning and computer vision techniques in underground construction: current status and future perspectives", *Tunnelling and Underground Space Technology*, Vol. 108, 103677, doi: 10.1016/j.tust.2020.103677.

Ison, M. and Artemiadis, P. (2015), "Multi-directional impedance control with electromyography for compliant human-robot interaction", *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*, pp. 416-421, doi: 10.1109/ICORR.2015.7281235.

Jamil, N., Belkacem, A.N., Ouhbi, S. and Lakas, A. (2021), "Noninvasive electroencephalography equipment for assistive, adaptive, and rehabilitative brain–computer interfaces: a systematic literature review", *Sensors*, Vol. 21 No. 14, p. 4754, doi: 10.3390/s21144754.

Jang, S., Choi, J.Y., Yoo, E.S., Lim, D.Y., Lee, J.Y., Kim, J.K. and Pang, C. (2021), "Printable wet-resistive textile strain sensors using bead-blended composite ink for robustly integrative wearable electronics", *Composites Part B: Engineering*, Vol. 210, 108674, doi: 10.1016/j.compositesb.2021.108674.

Kong, Y. and Fu, Y. (2022), "Human action recognition and prediction: a survey", *International Journal of Computer Vision*, Vol. 130 No. 5, pp. 1366-1401, doi: 10.1007/s11263-022-01594-9.

Krojer, B., Adlakha, V., Vineet, V., Goyal, Y., Ponti, E. and Reddy, S. (2022), "Image retrieval from contextual descriptions", arXiv, doi: 10.48550/arXiv.2203.15867.

Laplaza, J., Moreno-Noguer, F. and Sanfeliu, A. (2022), "Context and intention aware 3D human body motion prediction using an attention deep learning model in handover tasks", *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4743-4748, doi: 10.1109/IROS47612.2022.9981465.

Lee, P.C., Wang, Y., Lo, T.P. and Long, D. (2018), "An integrated system framework of building information modelling and geographical information system for utility tunnel maintenance management", *Tunnelling and Underground Space Technology*, Vol. 79, pp. 263-273, doi: 10.1016/j.tust.2018.05.010.

Li, S., Zhang, L. and Diao, X. (2020), "Deep-learning-based human intention prediction using RGB images and optical flow", *Journal of Intelligent and Robotic Systems*, Vol. 97 No. 1, pp. 95-107, doi: 10.1007/s10846-019-01049-3.

Lin, C.J. and Lukodono, R.P. (2021), "Sustainable human–robot collaboration based on human intention classification", *Sustainability*, Vol. 13 No. 11, p. 5990, doi: 10.3390/su13115990.

Liu, H. and Wang, L. (2017), "Human motion prediction for human-robot collaboration", *Journal of Manufacturing Systems*, Vol. 44, pp. 287-294, doi: 10.1016/j.jmsy.2017.04.009.

Liu, Y., Habibnezhad, M. and Jebelli, H. (2021), "Brainwave-driven human-robot collaboration in construction", *Automation in Construction*, Vol. 124, 103556, doi: 10.1016/j.autcon.2021.103556.

Liu, C., Wang, D., Guo, Y., Zhang, S., Wang, H. and He, R. (2019), "Research on diffusion behaviors of leaked natural gas in urban underground utility tunnels", *2019 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 2076-2081, doi: 10.1109/ICMA.2019.8816276.

Lundberg, S. (2017), "A unified approach to interpreting model predictions", arXiv preprint arXiv: 1705.07874, available at: https://arxiv.org/abs/1705.07874

Ma, H., Zhou, X. and Huang, J. (2023), "Effect of ventilation on thermal and humidity environment of the underground utility tunnel in the plum rain season in Southern China: field measurement and CFD simulation", *Underground Space*, Vol. 13, pp. 301-315, doi: 10.1016/j.undsp.2023.02.016.

Ministry of Housing and Urban-Rural Development of China (2015), "Technical specifications for underground utility tunnel projects (GB 50838-2015)", available at: http://www.weboos.cn:8078/assets/basicStandard/std_620206.pdf

Mukherjee, D., Gupta, K., Chang, L.H. and Najjaran, H. (2022), "A survey of robot learning strategies for human-robot collaboration in industrial settings", *Robotics and Computer-Integrated Manufacturing*, Vol. 73, 102231, doi: 10.1016/j.rcim.2021.102231.

Nath, N.D. and Behzadan, A.H. (2020), "Deep convolutional networks for construction object detection under different visual conditions", *Frontiers in Built Environment*, Vol. 6, p. 97, doi: 10.3389/fbuil.2020.00097.

Niu, Y., Liu, H., He, R., Li, Z., Ren, H., Gao, B., Guo, H., Genin, G.M. and Xu, F. (2020), "The new generation of soft and wearable electronics for health monitoring in varying environment: from normal to extreme conditions", *Materials Today*, Vol. 41, pp. 219-242, doi: 10.1016/j.mattod.2020.10.004.

Ou, X., Yan, P., Zhang, Y., Tu, B., Zhang, G., Wu, J. and Li, W. (2019), "Moving object detection method via ResNet-18 with encoder–decoder structure in complex scenes", *IEEE Access*, Vol. 7, pp. 108152-108160, doi: 10.1109/ACCESS.2019.2931922.

Pan, Z. and Yu, Y. (2024), "Learning multi-granular worker intentions from incomplete visual observations for worker-robot collaboration in construction", *Automation in Construction*, Vol. 158, 105184, doi: 10.1016/j.autcon.2023.105184.

Pandey, G.K. and Srivastava, S. (2023), "ResNet-18 comparative analysis of various activation functions for image classification", *2023 International Conference on Inventive Computation Technologies (ICICT)*, pp. 595-601, doi: 10.1109/ICICT57646.2023.10134464.

Park, K.B., Choi, S.H., Lee, J.Y., Ghasemi, Y., Mohammed, M. and Jeong, H. (2021), "Hands-free human–robot interaction using multimodal gestures and deep learning in wearable mixed reality", *IEEE Access*, Vol. 9, pp. 55448-55464, doi: 10.1109/ACCESS.2021.3071364.

Pinto, A., Sousa, S., Simões, A. and Santos, J. (2022), "A trust scale for human-robot interaction: translation, adaptation, and validation of a human computer trust scale", *Human Behavior and Emerging Technologies*, Vol. 2022 No. 1, pp. 6437441-12, doi: 10.1155/2022/6437441.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. and Sutskever, I. (2021), "Learning transferable visual models from natural language supervision", *International Conference on Machine Learning*, pp. 8748-8763, available at: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf

Ravipati, A., Kondamuri, R.K., Posonia, M. and J, A.M. (2023), "Vision based detection and analysis of human activities", *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1542-1547, doi: 10.1109/ICOEI56765.2023.10125829.

Razali, H., Mordan, T. and Alahi, A. (2021), "Pedestrian intention prediction: a convolutional bottom-up multi-task approach", *Transportation Research Part C: Emerging Technologies*, Vol. 130, 103259, doi: 10.1016/j.trc.2021.103259.

Rezaei, M., Terauchi, M. and Klette, R. (2015), "Robust vehicle detection and distance estimation under challenging lighting conditions", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16 No. 5, pp. 2723-2743, doi: 10.1109/TITS.2015.2421482.

Robinson, N., Tidd, B., Campbell, D., Kulić, D. and Corke, P. (2023), "Robotic vision for human-robot interaction and collaboration: a survey and systematic review", *ACM Transactions on Human-Robot Interaction*, Vol. 12 No. 1, pp. 1-66, doi: 10.1145/3570731.

Saif, A.S., Wollega, E.D. and Kalevela, S.A. (2023), "Spatio-temporal features based human action recognition using convolutional long short-term deep neural network", *International Journal of Advanced Computer Science and Applications*, Vol. 14 No. 5, doi: 10.14569/IJACSA.2023.0140501.

Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T. and Song, Y.Z. (2023), "Clip for all things zero-shot sketch-based image retrieval, fine-grained or not", *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2765-2775, available at: https://openaccess.thecvf.com/content/CVPR2023/papers/Sain_CLIP_for_All_Things_Zero-Shot_Sketch-Based_Image_Retrieval_Fine-Grained_or_CVPR_2023_paper.pdf

Sarabu, A. and Santra, A.K. (2021), "Human action recognition in videos using convolution long short-term memory network with spatio-temporal networks", *Emerging Science Journal*, Vol. 5 No. 1, pp. 25-33, doi: 10.28991/esj-2021-01254.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017), "Grad-cam: visual explanations from deep networks via gradient-based localization", *Proceedings of the IEEE international conference on computer vision*, pp. 618-626, doi: 10.1109/iccv.2017.74, available at: https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf

Semeraro, F., Griffiths, A. and Cangelosi, A. (2023), "Human–robot collaboration and machine learning: a systematic review of recent research", *Robotics and Computer-Integrated Manufacturing*, Vol. 79, 102432, doi: 10.1016/j.rcim.2022.102432.

Shahrour, I., Bian, H., Xie, X. and Zhang, Z. (2020), "Use of smart technology to improve management of utility tunnels", *Applied Sciences*, Vol. 10 No. 2, p. 711, doi: 10.3390/app10020711.

Soufineyestani, M., Dowling, D. and Khan, A. (2020), "Electroencephalography (EEG) technology applications and available devices", *Applied Sciences*, Vol. 10 No. 21, p. 7453, doi: 10.3390/app10217453.

Sousa, R.L. and Einstein, H.H. (2021), "Lessons from accidents during tunnel construction", *Tunnelling and Underground Space Technology*, Vol. 113, 103916, doi: 10.1016/j.tust.2021.103916.

Tai, C., Tian, G. and Lei, W. (2022), "Measurement of indoor environmental parameters and analysis of the condensation phenomenon in urban utility tunnels", *Indoor and Built Environment*, Vol. 31 No. 4, pp. 1091-1106, doi: 10.1177/1420326X211056485.

Tripathi, R. and Verma, B. (2023), "CLIP-LSTM: fused model for dynamic hand gesture recognition", *2023 IEEE 20th India Council International Conference (INDICON)*, pp. 926-931, doi: 10.1109/INDICON59947.2023.10440820.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M. and Baik, S.W. (2017), "Action recognition in video sequences using deep bi-directional LSTM with CNN features", *IEEE Access*, Vol. 6, pp. 1155-1166, doi: 10.1109/ACCESS.2017.2778011.

University of Washington Facilities (2024), *Utility Tunnels, Trenches, and Manholes*, University of Washington Facilities Design Standard (Section: Tunnels), available at: https://facilities.uw.edu/files/media/uwf-ds-tunnels.pdf

Vianello, L., Mouret, J.B., Dalin, E., Aubry, A. and Ivaldi, S. (2021), "Human posture prediction during physical human-robot interaction", *IEEE Robotics and Automation Letters*, Vol. 6 No. 3, pp. 6046-6053, doi: 10.1109/LRA.2021.3086666.

Wang, M. (2021), "Ontology-based modelling of lifecycle underground utility information to support operation and maintenance", *Automation in Construction*, Vol. 132, 103933, doi: 10.1016/j.autcon.2021.103933.

Wang, W., Li, R., Chen, Y., Sun, Y. and Jia, Y. (2021), "Predicting human intentions in human–robot hand-over tasks through multimodal learning", *IEEE Transactions on Automation Science and Engineering*, Vol. 19 No. 3, pp. 2339-2353, doi: 10.1109/TASE.2021.3074873.

Waqas, M. and Humphries, U.W. (2024), "A critical review of RNN and LSTM variants in hydrological time series predictions", *MethodsX*, Vol. 13, 102946, doi: 10.1109/ACCESS.2017.2778011.

Wei, D., Chen, L., Zhao, L., Zhou, H. and Huang, B. (2021), "A vision-based measure of environmental effects on inferring human intention during human robot interaction", *IEEE Sensors Journal*, Vol. 22 No. 5, pp. 4246-4256, doi: 10.1109/JSEN.2021.3139593.

Yu, H., Kamat, V.R., Menassa, C.C., McGee, W., Guo, Y. and Lee, H. (2023a), "Mutual physical state-aware object handover in full-contact collaborative human-robot construction work", *Automation in Construction*, Vol. 150, 104829, doi: 10.1016/j.autcon.2023.104829.

Yu, H., Kamat, V.R., Menassa, C.C., McGee, W., Guo, Y. and Lee, H. (2023b), "Grip state recognition for enabling safe human-robot object handover in physically collaborative construction work", in *Computing in Civil Engineering 2023*, pp. 787-795, doi: 10.1061/9780784485224.095.

Zhai, C., Shi, R., Xi, P. and Cai, G. (2024), "Study of the leakage mechanism of a utility tunnel waterstop based on a fluid–solid coupled method", *Journal of Pipeline Systems Engineering and Practice*, Vol. 15 No. 1, 05023005, doi: 10.1061/JPSEA2.PSENG-1466.

Zhang, T., Sun, H. and Zou, Y. (2022), "An electromyography signals-based human-robot collaboration system for human motion intention recognition and realization", *Robotics and Computer-Integrated Manufacturing*, Vol. 77, 102359, doi: 10.1016/j.rcim.2022.102359.

Zhang, X., Tian, S., Liang, X., Zheng, M. and Behdad, S. (2024), "Early prediction of human intention for human–robot collaboration using transformer network", *Journal of Computing and Information Science in Engineering*, Vol. 24 No. 5, 051003, doi: 10.1115/1.4064258.

Zhang, Y., Wu, D., Kong, Q., Li, A., Li, Y., Geng, S., Chen, P. and Liu, Y. (2020), "Exposure level and distribution of airborne bacteria and fungi in an urban utility tunnel: a case study", *Tunnelling and Underground Space Technology*, Vol. 96, 103215, doi: 10.1016/j.tust.2019.103215.

Zhang, M., Xu, R., Wu, H., Pan, J. and Luo, X. (2023), "Human–robot collaboration for on-site construction", *Automation in Construction*, Vol. 150, 104812, doi: 10.1016/j.autcon.2023.104812.

Zhou, C., Gao, Y., Chen, E.J., Ding, L. and Qin, W. (2023), "Deep learning technologies for shield tunneling: challenges and opportunities", *Automation in Construction*, Vol. 154, 104982, doi: 10.1016/j.autcon.2023.104982.

Zhou, C., Lin, Z., Du, C., Wang, Z. and Li, F. (2020), "Research on key technologies of tunnel robot based on cloud edge collaboration", *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology, ICCASIT*, pp. 661-666, doi: 10.1109/ICCASIT50869.2020.9368679.

Zou, R., Liu, Y., Li, Y., Chu, G., Zhao, J. and Cai, H. (2023), "A novel human intention prediction approach based on fuzzy rules through wearable sensing in human–robot handover", *Biomimetics*, Vol. 8 No. 4, p. 358, doi: 10.3390/biomimetics8040358.

Zou, R., Liu, Y., Zhao, J. and Cai, H. (2024), "Multimodal learning-based proactive human handover intention prediction using wearable data gloves and augmented reality", *Advanced Intelligent Systems*, Vol. 6 No. 4, 2300545, doi: 10.1002/aisy.202300545.

Corresponding author
Jun Wang can be contacted at: jun.wang@westernsydney.edu.au