POLITECNICO DI TORINO
Repository ISTITUZIONALE

Beyond pixels: A comprehensive survey on deep learning-based scene understanding for human-centered applications

Original
Beyond pixels: A comprehensive survey on deep learning-based scene understanding for human-centered applications / Ruggieri, R., Marullo, G., Ulrich, L., Grandvalet, Y., Moos, S., Vezzetti, E.. - In: IMAGE AND VISION COMPUTING. - ISSN 0262-8856. - 176:(2026). [10.1016/j.imavis.2026.106227]

Availability:
This version is available at: 11583/3016310 since: 2026-11-25T15:01:41Z

Publisher:
Elsevier

Published
DOI:10.1016/j.imavis.2026.106227

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

27 November 2026

Image and Vision Computing 176 (2026) 106227

Contents lists available at ScienceDirect

Image and Vision Computing

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

Review article

Beyond pixels: A comprehensive survey on deep learning-based scene understanding for human-centered applications

Rossella Ruggieri [a,*], Giorgia Marullo [a], Luca Ulrich [a], Yves Grandvalet [b], Sandro Moos [a], Enrico Vezzetti [a]

[a] Department of Management and Production Engineering, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy
[b] Université de Technologie de Compiègne, CNRS, CS 60 319, Compiègne Cedex, 60 203, France

ARTICLE INFO

Keywords:
Computer vision
Deep learning
Scene understanding
Human-centered
Context-awareness

ABSTRACT

Artificial Intelligence (AI) represents an enabling technology to provide users with adaptive and context-aware support, in line with the Industry 5.0 paradigm. Among AI approaches, Deep Learning (DL) has proven effective in interpreting visual information, which captures environmental features, human activities, and interactions. In this context, high-level scene understanding refers to methods that go beyond object detection or image segmentation, aiming to model interactions, contextual dependencies, and meaningful relationships within visual environments. This capability is essential for supporting people in complex and information-rich scenarios, where low-level visual analysis alone is insufficient. This survey aims to identify the primary contexts where high-level scene understanding supports individuals in their activities, providing context-aware feedback and enhancing safety and situational awareness. Three application areas emerged as most representative: construction environments, where complex and critical scenarios require safety monitoring; surgical environments, characterized by high-risk procedures that demand precise context-aware support; and assistive and inclusive environments for users with special needs, particularly visually impaired individuals, who require a tangible solution for vision support. Across these domains, the review maps domain-specific needs, examines the forms of support via DL technologies, and highlights the constraints that limit real-world adoption. The findings suggest that future progress requires scalable data annotation, the integration of 3D and temporal information, efficient models, and a user-centered, interdisciplinary approach to ensure alignment with actual needs and enhance trust, usability, and real-time applicability.

Contents

1. Introduction 2
2. Methodology 3
3. Results 4
   3.1. Construction environments 5
   3.2. Surgical environments 8
   3.3. Assistive environments for special needs 10
4. Discussion 12
   4.1. Domain-specific needs and technological forms of support 13
   4.2. Domain-independent challenges and future guidelines 14
5. Conclusions 18
   CRediT authorship contribution statement 19
   Declaration of competing interest 19
   Data availability 19
   References 19

* Corresponding author.
E-mail address: rossella.ruggieri@polito.it (R. Ruggieri).

https://doi.org/10.1016/j.imavis.2026.106227
Received 31 January 2026; Received in revised form 28 August 2026; Accepted 26 September 2026
Available online 2 October 2026
0262-8856/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent research has highlighted the essential role of technology as an assistive tool for human activities, emphasizing the utility of methods that enhance human performance and improve safety and situational awareness. This has led to more interest in human-centered solutions, which put users' needs at the center of design and decision-making processes [1]. This trend is further reinforced by the paradigm shift represented by Industry 5.0, which goes beyond the traditional focus on maximizing production and economic benefits of Industry 4.0. While Industry 4.0 centered on automation and productivity through cyber–physical systems and data-driven manufacturing, Industry 5.0 explicitly repositions the human at the center of the industrial process, promoting human-AI collaboration rather than substitution, and addressing social and planetary challenges by placing human well-being at the heart of technological and industrial design. Under this paradigm, decision support systems are expected not only to optimize operations but to adapt to human needs and capabilities, reinforcing the relevance of human-centered scene understanding as an enabling technology for this transition. One of the most relevant challenges in this context is the creation of intelligent human-centric environments that prioritize user needs while maintaining performance and efficiency [2]. In this perspective, the user is defined as the individual to whom the system's feedback is addressed and who directly benefits from it and acts upon it.

Within the framework of human-centered design [2,3], technologies are conceived, developed, and evaluated by focusing on human capabilities, limitations, goals, and values [4]. Systems must automate tasks, enhance cognition, support decisions, and improve safety, usability, and well-being [5,6]. To achieve these goals, human-centered systems must offer adaptive assistance based on user behavior and context, requiring computational methods capable of perceiving and interpreting the environment with a semantic depth comparable to that of humans [7].

Artificial Intelligence (AI) represents a key enabling technology in this scenario, as it allows heterogeneous data to be transformed into information that is useful for understanding the context and generating adaptive support [8]. Among all the forms of data processed by AI systems, visual data, namely, images and videos, have taken a central role thanks to their ability to naturally represent the structure of the environment, objects, postures, dynamics, and interactions. This wealth of information has made Computer Vision (CV) a central technology in the development of user-centered systems [9], and Deep Learning (DL) methods have demonstrated a considerable ability in visual data analysis. DL networks can efficiently and accurately find complex features in images, enabling them to learn both simple and abstract patterns. In recent years, DL has greatly spread the CV field, allowing systems to comprehend and interpret scenes more precisely and efficiently [9,10].

The term scene understanding refers to computational methods that interpret an environment by considering its semantic components, spatial organization, temporal evolution, and relational structures, allowing computers to automatically analyze the content of images and videos, thus emulating human vision and the ability to comprehend scenes [11].

Early research primarily focused on identifying objects through instance-level tasks and on pixel-wise classification via pixel-level tasks, typically using bounding boxes or segmentation masks [12,13]. However, these tasks are insufficient for reasoning about complex, dynamic, and information-rich environments. In such contexts, it is more helpful to adopt an advanced contextual understanding approach that can significantly improve scene analysis and provide more accurate feedback to the user [13,14]. For human-centered assistance, a deeper level of interpretation is essential, where systems can recognize not only what is present in the scene but also how elements relate, interact, and influence activities [14]. To date, the continuous advancement of DL techniques is leading scene understanding to increasingly higher levels of semantics [15,16]. Beyond instance-level and pixel-level tasks, more advanced approaches have emerged that focus on understanding semantic relationships and interactions between entities and their surrounding environment, leading to a more complete comprehension of context [11,17,18]. Research demonstrates that such a deeper contextual understanding significantly helps to improve scene interpretation, especially in complex or ambiguous visual situations [13].

Nowadays, scene understanding techniques are increasingly being integrated across various fields [12], enhancing human–machine interaction, supporting decision making, helping humans in their activities, monitoring environments, and identifying risks. For example, in autonomous driving, they enable vehicles to navigate safely and effectively [19,20]. In agriculture, scene understanding facilitates automated harvesting and crop monitoring [21–23]. Furthermore, these technologies support humans in their daily and work-related activities [12,14,24], particularly for people with special needs [25,26]. Collaborative robots rely on scene understanding to operate safely in environments with humans [26,27]. In smart cities, these technologies allow the analysis of public spaces, supporting security monitoring and resource management [28]. They also contribute to the creation of interactive and immersive entertainment experiences by interpreting visual scenes [29]. Moreover, these techniques are employed to identify changes in land use, deforestation, and dangerous events such as floods and fires [30–32].

While these examples illustrate the breadth of scene understanding applications, this survey deliberately focuses on domains in which scene understanding explicitly models human–object or human–environment interactions in direct support of a human activity. Applications such as autonomous driving, precision agriculture, and smart-city infrastructure monitoring, although methodologically related and increasingly reliant on similar deep learning architectures, predominantly analyze relationships among non-human entities (e.g., vehicles, crops, urban infrastructure) or serve the perception and control needs of an artificial agent itself, rather than modeling how a person relates to their surrounding environment for the purpose of receiving context-aware support. This distinction, rather than the application domain itself, underlies the scope adopted in the remainder of this review. Beyond this scope-defining distinction, a further gap concerns the survey literature itself: despite the widespread adoption, existing surveys typically address individual domains or focus on lower-level semantic tasks, such as object detection and semantic segmentation, without examining how high-level scene understanding contributes to human-centered support across contexts [12,19,33–36]. In 2025, Li et al. [12] reviewed construction scene understanding based on five key scene elements, signals, pixels, objects, relationships, and events, outlining advances in perception and reasoning abilities within this domain. In the same year, Ankareddy and Delhibabu [33] reviewed dense segmentation techniques for urban scene parsing, focusing exclusively on semantic, instance, and panoptic segmentation architectures (e.g., Vision Transformer-based models such as Mask2Former) and their performance on benchmark datasets such as Cityscapes and Mapillary Vistas; the scope of their review is restricted to pixel-level perceptual tasks, with no consideration of relational or human-centered aspects of scene understanding. Awan et al. [34] similarly reviewed deep learning methods for autonomous driving scene understanding, organizing the field into four major perceptual tasks, object detection, semantic segmentation, instance segmentation, and panoptic segmentation, aimed to enable reliable perception of the surrounding environment by autonomous vehicles; the review is centered on vehicular perception and contextual knowledge for autonomous decision-making, with no consideration of human-centered or relational aspects of scene understanding. Dolatyabi et al. [19] extended this focus to traffic scene understanding more broadly, covering classification, detection, and segmentation as well as more advanced applications such as action recognition, tracking, and anomaly detection, primarily oriented toward intelligent transportation systems. In the surgical domain, Khan

Table 1
Comparison of this review with existing surveys on scene understanding.

Survey	Domain coverage	Human-centered focus	Level of understanding
Li et al. [12]	Construction	Yes	High (relationships as core element)
Ankareddy and Delhibabu [33]	Urban/Smart-city	No	Low (segmentation)
Awan et al. [34]	Autonomous Driving	No	Low (detection, segmentation)
Dolatyabi et al. [19]	Traffic/ITS	No	Low (perception, tracking, action recognition)
Khan et al. [35]	Surgical	No	Low (segmentation, instrument tracking)
Liu et al. [36]	Smart Manufacturing	No	Low (architecture-based classification)
This review	**Multiple application domains**	**Yes**	**High (relationships as core element)**

et al. [35] reviewed the integration of state-of-the-art ML and DL technologies to understand the scene in surgical workflows, focusing on tool segmentation, instrument tracking, phase recognition, and anomaly detection. While explicitly framed around improving clinical outcomes and patient safety, the review remains centered on perceptual tasks and does not examine how systems explicitly model relationships between surgical instruments, tissues, or personnel, an aspect addressed by relational approaches such as instrument–tissue interaction detection or scene graph-based surgical reasoning. Finally, Liu et al. [36] reviewed scene understanding methods in smart manufacturing environments, classifying data-driven, end-to-end, and cross-modal approaches. As Table 1 summarizes, these reviews are each confined to a single application domain and, with partial exceptions such as [12], primarily address the perceptual component of scene understanding, namely detection, segmentation, or tracking, without systematically examining the reasoning component and how systems model and exploit human–object or human–environment relationships to support human activities. Moreover, none of these reviews adopts a methodology that explicitly uses human-centered interaction as a literature selection criterion, nor connects multiple application domains under a unified analytical lens.

As a result, there is a lack of unified analyses that connect diverse application areas through the common lens of human–environment interaction and context-aware visual interpretation. To fill this gap, the current survey defines and investigates the primary application domains in which high-level scene awareness can be utilized to assist individuals during real-world activities. In this work, we adopt the term human-centered high-level scene understanding to denote approaches that move beyond isolated perceptual tasks to explicitly model and communicate the relationship between a person and their surrounding environment, or between a person and other entities in the scene, with the aim of providing context-aware feedback that supports a human activity. Fig. 1 clarifies this distinction. Building on this definition, the study systematically examines how such high-level scene understanding is applied across domains to provide successful human-centered support.

It first analyzes domain-specific needs arising from real-world operations. While human-centered systems also involve a range of design choices like feedback modalities, devices, and data acquisition, this review focuses on how these requirements influence the selection of DL models and scene understanding algorithms. By assessing current solutions in connection to their application contexts, the survey identifies the major limitations and research gaps that still restrict the reliable implementation of these technologies in practice. Based on this analysis, the investigation has been structured around three research questions:

- RQ1: What application domains use high-level scene understanding based on human–object and human–environment interactions to support people in their activities?
- RQ2: What specific needs are addressed in each domain, and how do high-level scene understanding algorithms provide support?
- RQ3: What research gaps and future challenges are identified concerning high-level scene understanding technologies across these application domains?

By addressing these research questions, this study clarifies the current state of the art in high-level scene understanding, linking cross-domain applications under the common lens of human–environment relationship and context-aware visual interpretation. The information obtained from this analysis should provide insights for both academic research and industrial practice, promoting the development of more reliable perception systems that can enhance human safety, efficiency, and situational awareness.

The content of this survey is structured as follows: Section 2 describes the methodology for selecting the scientific papers included in the review; Section 3 presents an overview of the studies identified through this process, highlighting the main application domains where they have been applied; Section 4 discusses the key findings, focusing on the domain-specific needs, the forms of support offered by these technologies, and the current limitations; finally, Section 5 concludes the review by synthesizing the insights gained and outlining directions for future research.

2. Methodology

The goal of this review is to map the state-of-the-art in high-level scene understanding applied to human-centered support across real-world domains. The research is structured around three research questions (RQ1–RQ3) formulated in Section 1, which guide the systematic literature collection and the analysis described below.

The electronic databases consulted for this purpose were Scopus and IEEE Xplore. Scopus was selected as a retrieval database, as it represents a curated citation index with well-defined journal and conference selection policies [37], ensuring a consistent level of peer-reviewed quality and broad interdisciplinary coverage. IEEE Xplore was additionally consulted to ensure comprehensive coverage of the computer science and engineering conference literature underlying much of this review. Three main groups of keywords were considered for searching papers, each selected for a specific purpose. The first group comprises the words “Deep Learning”, “Neural Networks”, and “Computer Vision”. This set was chosen to focus the research on visual data, as it constitutes the most informative modality for representing the scene structure, and on deep-learning models, since they have proven to be particularly effective at analyzing this kind of data. The second group, including “Scene Understanding”, “Scene Analysis”, and “Context Awareness”, constrained the search to studies targeting more contextual reasoning, excluding low semantic tasks like object detection or segmentation. The third group includes “Ergonomics”, “Worker”, “Work”, “Human Activity”, “Occupational Health”, “Industry 5.0”, “Human–Machine”, “Human-Centered”, “Smart Environments”, “Safety”, “Human Support”, “Activity Monitoring”, “Decision Making”. It defined human-centered application domains, focusing on ergonomics, safety, well-being, and health in workplaces, and activity support. During the research, each main group was connected by the “AND” operator to ensure that all three were included. Within each group, keywords were combined using the “OR” operator to ensure a comprehensive coverage of each conceptual area. The research was carried out by checking for the words listed above in papers’ titles, abstracts, and keywords.

The analysis was limited to articles published between January 2021 and September 2025, considering the period in which DL-based

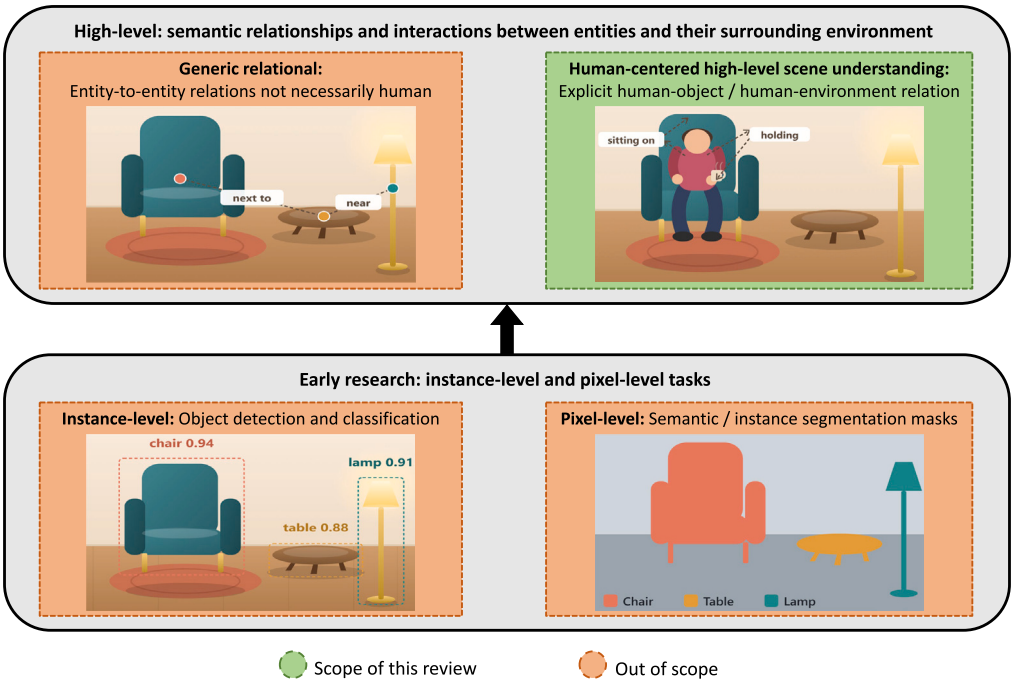


Fig. 1. Taxonomy of scene understanding approaches. Low-level pixel- and instance-level tasks are insufficient for complex, dynamic environments. Among high-level relational approaches, generic methods do not necessarily center on a person, while human-centered high-level scene understanding, the scope of this review, requires an explicit human–object/environment relation and a support-oriented purpose.

scene understanding techniques began to mature rapidly and be systematically applied to human-centered contexts.

The analysis and selection of papers followed the PRISMA 2020 statement guidelines [38]. A set of inclusion and exclusion criteria was established to assess the eligibility of the articles. The inclusion criteria are listed below:

- Papers published between January 2021 and July 2026, inclusive;
- Papers published in journals ranked Q1 or Q2 according to the Scimago Journal Rank (SJR);
- Conference papers with an SJR > 1;
- Studies based entirely or partially on Computer Vision methods;
- Studies whose primary aim is to support human activities;
- Studies considering high-level scene understanding approaches that explicitly model human–object and/or human–environment interactions, rather than relationships confined to non-human entities;
- Studies focusing on the analysis of human–object and human–environment interactions;

The exclusion criteria are reported in the following list:

- Reviews papers, theses, books, web pages;
- Papers without available full text;
- Non-English papers;

The inclusion and exclusion criteria listed above were defined to select the most recent studies relevant to the field of interest, published in high-impact journals according to the SJR index. In this way, the selection aims to obtain a representative and significant sample of the state of the art in high-level scene understanding for human activities support through the analysis of interaction with the environment. Subsequently, titles and abstracts of the obtained papers were examined to assess their overall relevance with respect to the research scope, excluding works that do not address the topic of the analysis. Finally, a full-text examination was performed on the remaining papers to verify their relevance and alignment with the defined inclusion criteria, leading to the final set of selected studies. Fig. 2 illustrates the entire selection procedure.

3. Results

At the end of the selection process, thirty-seven studies were considered for the review analysis. Their main qualitative characteristics are summarized in Table 2, which reports, for each study, the year of publication, the application environment, the support purpose, whether temporal information and 3D data are exploited, and the relational modeling technique. Temporal modeling and 3D representations were explicitly considered, as they have been shown to provide additional capabilities for understanding high-level scenes in many cases. This includes supporting the recognition of activities and events over time and enabling more robust spatial reasoning in complex environments [13]. Observing these dimensions provides a key for interpreting the different strategies adopted in the analyzed works and for comparing how various design choices contribute to scene understanding in different application domains and use cases.

As illustrated in Fig. 3, almost half of the studies concern applications in surgical environments (15 studies), where automatic scene interpretation supports surgeons by providing contextual assistance during operations. Another significant portion of the selected studies focuses on construction environments (12 studies), where a high-level scene understanding is used for safety monitoring and compliance assessment. Finally, a smaller number of studies (10 studies) address assistive environments for visually impaired or disabled users, to enhance their autonomy and environmental awareness through intelligent systems. These three domains represent the main areas of application in which high-level scene understanding techniques are currently employed to support people in their activities. This finding provides a preliminary answer to the first research question, RQ1. Based on this categorization, this Section is organized as follows: Section 3.1 concerns applications in construction environments; Section 3.2 focuses on applications in surgical environments; Section 3.3 examines assistive environments designed to support people with special needs. Each Section analyzes the respective domain in greater detail, highlighting the type of support provided and the DL techniques adopted.

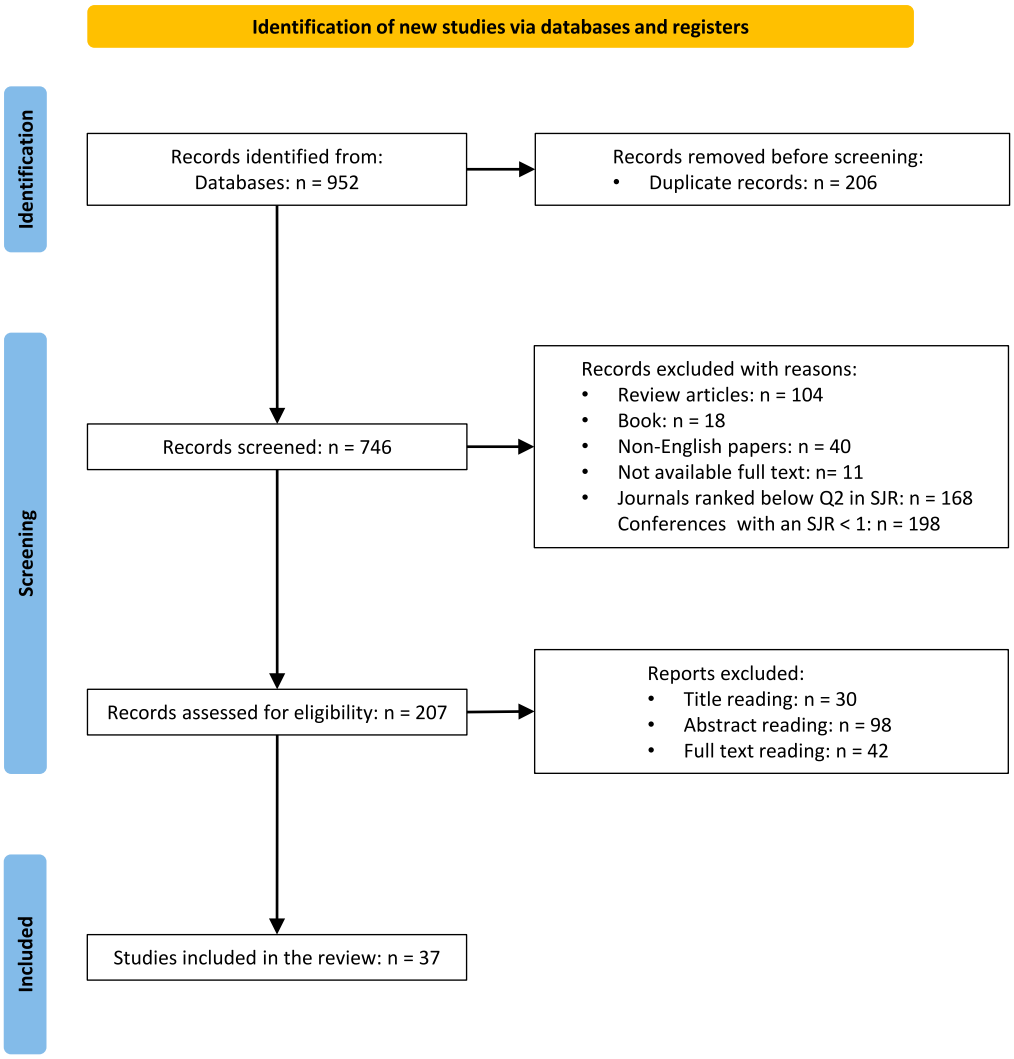


Fig. 2. Flowchart describing the papers' selection procedure, prepared in accordance with the PRISMA template [38].

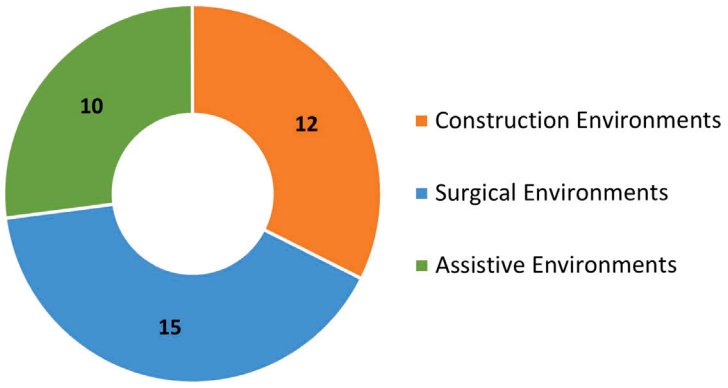


Fig. 3. Pie chart summarizing the analyzed studies, grouped by application domain: 15 studies concern surgical environments (blue), 12 construction environments (orange), and 10 assistive environments for users with special needs (green) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Construction environments

Construction sites are among the most dangerous workplaces, where many potential hazards occur, often with serious or fatal consequences [39]. The main causes of accidents are falls from heights, collisions with vehicles or machinery, inhalation of dust, and lack of protective equipment such as helmets, goggles, or masks. However, most of these accidents could be avoided if all workers correctly followed safety rules and wore personal protective equipment (PPE), e.g., hard hats, gloves, and body harnesses [43]. Nonetheless, even if they have been previ-

Table 2
Summary of the analyzed studies, including authors, application environment, support purpose, indicators for temporal and 3D information usage, and relational modeling technique.

Authors	Application environment	Support purpose	Temp. Info	3D Info	Relational modeling
2021					
Chen and Demachi [39]	Construction	Worker safety	✗	✗	Graph-based
Xu et al. [40]	Surgical	Automatic surgical reporting	✗	✗	Transformer/Attention
2022					
Nwoye et al. [41]	Surgical	Laparoscopy	✗	✗	Transformer/Attention
Pan et al. [42]	Construction	Worker safety	✗	✗	Graph-based
Chen et al. [43]	Construction	Worker safety	✗	✗	Graph-based
Zhang et al. [44]	Construction	Worker safety	✗	✗	Transformer/Attention
Pang et al. [45]	Surgical	Surgical procedure and reporting	✗	✗	Graph-based
2023					
Liu et al. [46]	Assistive	Scene description	✗	✗	Transformer/Attention
Wang et al. [47]	Surgical	Automatic surgical reporting	✓	✗	Graph-based
Liu et al. [48]	Construction	Worker safety	✗	✗	CNN-based
Dessalene et al. [49]	Assistive	Home assistance	✓	✗	Graph-based
Zhong et al. [50]	Construction	Worker safety	✗	✗	Recurrent
Murali et al. [24]	Surgical	Laparoscopy	✗	✗	Graph-based
Xu et al. [51]	Surgical	Surgical procedure	✗	✗	Transformer/Attention
Lin et al. [52]	Surgical	Automatic surgical reporting	✗	✗	Graph + Transformer (hybrid)
2024					
Zhang et al. [53]	Surgical	Laparoscopy	✓	✗	Graph-based
Lin et al. [54]	Surgical	Surgical procedure	✓	✗	Graph-based
Cao et al. [55]	Assistive	Navigation	✗	✗	Transformer/Attention
Su and Chen [56]	Assistive	Home assistance	✗	✗	Graph-based
Özsoy et al. [57]	Surgical	Holistic OR	✓	✓	Graph-based
Yuan et al. [58]	Surgical	VQA for intraoperative procedures	✗	✗	Graph-based
Pei et al. [59]	Surgical	Holistic OR	✗	✓	Transformer/Attention
2025					
Kim et al. [60]	Assistive	Scene description	✗	✗	Transformer + Recurrent (hybrid)
Xie et al. [61]	Construction	Worker safety	✓	✓	Graph-based
Deng et al. [14]	Construction	Worker safety	✗	✗	Recurrent + Attention (hybrid)
Li et al. [62]	Construction	Worker safety	✗	✗	Transformer/Attention
Chan et al. [63]	Construction	Worker safety	✗	✗	Transformer/Attention
Özsoy et al. [64]	Surgical	Holistic OR	✓	✓	Graph-based
Yuan et al. [65]	Assistive	Navigation	✓	✗	Transformer/Attention
Yang et al. [66]	Assistive	Home assistance	✓	✓	Transformer/Attention
2026					
Kim et al. [67]	Construction	Worker safety	✗	✗	Transformer/Attention
Abidi et al. [68]	Construction	Worker safety	✗	✗	Transformer/Attention
Yang et al. [69]	Surgical	Surgical procedure	✓	✗	Transformer/Attention
Huang et al. [70]	Assistive	Home assistance	✓	✓	Transformer/Attention
Sajid et al. [71]	Assistive	Home assistance	✗	✗	Transformer/Attention
Low et al. [72]	Surgical	VQA for intraoperative procedures	✗	✗	Graph + Transformer (hybrid)
Zhao et al. [73]	Assistive	Navigation	✗	✗	Transformer/Attention

ously educated and trained, many workers do not always comply with the rules due to distractions, fatigue, or habit [39]. Since PPE compliance is not perfect and continuous human monitoring is costly and not always feasible, there is a need for an automated monitoring approach to help facilitate the safety monitoring of construction workers and ensure the proper use of PPE [39,43,44]. Currently, several approaches have been investigated for the automated identification of hazards associated with improper PPE use, which can be categorized into two main types: sensor-based approaches and vision-based approaches [39]. Sensor-based approaches primarily rely on wearable electronic devices to locate and monitor PPE or the position of workers. With these methods, it is possible to verify whether PPE is close to the person or whether a tag is present on the person, but they cannot always verify whether the PPE is actually being worn. They are also less practical for managing complex cases involving different types of PPE, multiple people, and continuous movement. They are more invasive and can cause discomfort to the worker, and they require high maintenance costs [39,44]. On the other hand, vision-based approaches are non-intrusive and less device-intensive due to the widespread application of surveillance cameras on construction sites.

Among the CV methods, early vision-based approaches focused on systems based on object detection. They analyze images or videos recording construction sites and automatically recognize whether workers are complying with safety rules, such as wearing a helmet. These methods work accurately for simple cases, such as checking whether PPE is present or not. However, when multiple types of protective equipment need to be checked simultaneously, the model becomes difficult to train and less accurate. To overcome this limitation, some more recent studies combine object detection with human pose estimation algorithms, which reconstruct the position of workers' body joints. This allows the user to understand how people move, what objects are used, and how they are interacting with them. Recently, knowledge of

current safety rules has also begun to be integrated into these systems to enable automatic hazard identification [39].

A representative example is the framework proposed by Chen and Demachi [39]. The system involves object detection to locate PPE instances, and human pose estimation to extract poses, for computing the distance between each worker's body joints and the PPEs' locations. This data is then arranged into a scene graph, a structured representation composed of nodes corresponding to objects, people, PPE, and edges representing the relationships among them. This scene graph is then compared with rule graphs, which are a simplified representation manually derived from construction safety regulations. In this way, the system can identify rule violations and provide a list of hazards. In a subsequent study [43], the same authors addressed limitations of their earlier work by improving detection robustness under occlusion and low-light conditions, and by automating the rule-translation process using Natural Language Processing (NLP) tools for syntactic parsing. This avoids the need for manual rule encoding and enables more comprehensive reasoning across complex scenarios.

Subsequently, the use of more advanced neural networks was also introduced in this context to enable a more accurate understanding of the scene. Zhang et al. [44] exploited the efficiency of Transformers, which allow for obtaining scene graphs by learning how different elements in an image relate to one another, avoiding the explicit calculation of geometric relationships between entities. This model analyzes the entire image simultaneously and determines which entities are connected based on their visual context and appearance. As a result, the system can recognize meaningful semantic relationships like interactions or functional roles, achieving a more comprehensive and flexible interpretation of the scene, as it is not based solely on objective distance measurements but takes into account the image as a whole to infer semantic relationships. Additionally, the rule-translation process has been improved too, through the use of C-BERT (Construction-Bidirectional Encoder Representations from Transformers), a tailored NLP model adapted to the construction domain. This model compares the information in the graph with safety rules expressed in natural language to automatically identify potential risks. Liu et al. [48] presented a similar framework that relies on CNNs to extract visual relationships. The main advantage of their approach is that both images and regulatory texts are encoded as relational triples that comprise subject, predicate, and object, reducing the semantic gap between natural language and scene graph representation.

Other studies have explored how scene understanding may be extended beyond predefined categories. Pan et al. [42] presented Image2Triplets, a framework that automatically updates Knowledge Graphs (KGs) in the construction industry, to address the limitations related to the scarcity of annotated data, the difficulty of detecting new interactions, and the manual updating of KGs. The pipeline combines Zero-Shot Detection (ZSD) and Zero-Shot Human–Object Interaction (ZSHOI) techniques to recognize entities and interactions, including those never seen before. This allows automatic updates and expansion of the knowledge useful for aiding in monitoring and managing construction sites.

In parallel, efforts have focused on generating natural language descriptions to support security documentation and indexing. Zhong et al. [50] developed a system for the automatic understanding and captioning of construction site scenes from videos, exploiting an encoder that extracts visual and semantic features and an LSTM decoder that generates textual descriptions in natural language. The final output provided to the user, therefore, consists of complete sentences describing the objects, relationships, and attributes present in the scene, which are useful for monitoring the safety status of workers and for quickly indexing and retrieving images or videos in safety management systems. Deng et al. [14] developed a similar system, mainly focusing on tower crane lifting activities. To improve the quality of captions, the LSTM decoder uses a spatial attention mechanism, which allows the model to focus on the most relevant regions of the image. Moreover, during the generation of textual descriptions, the beam search method is employed, which explores multiple word sequences simultaneously and selects the most probable one, avoiding shortcut errors typical of the greedy method. This represents an advancement over previous work, where captions were generated using simpler and less accurate strategies. In 2025, Li et al. [62] presented SafetyGPT, whose structure is shown in Fig. 4.

This framework was designed to monitor unsafe behavior on electrical construction sites, an area that has received little attention in the literature. Unlike previous approaches, the system not only combines visual analysis of images with safety regulations knowledge, but also incorporates textual information or requests provided by supervisors, encouraging active interaction with users monitoring the construction sites. More recently, Kim et al. [67] proposed a regulation-guided Visual Question Answering (VQA) framework for ladder safety compliance, a hazard category (falls from height during ladder work) relatively underexplored compared to general PPE detection. Nine safety-related questions, encoding regulatory factors such as ladder type, working height, top-step usage, outriggers, co-worker presence, and PPE, were derived from national safety standards (KOSHA Guide G-130-2020) and used to fine-tune a BLIP-1 model to generate sentence-level, regulation-aware answers for each scene. The framework further incorporates uncertainty-aware outputs (e.g., "not identifiable") instead of forced binary Safe/Unsafe classification, and generates explainable safety reports intended to support, rather than replace, human safety managers. Building on this same VQA-based paradigm, Abidi et al. [68] proposed SafeVision, a vision–language framework that, unlike Kim et al. [67]'s single-pass reasoning over predefined questions, explicitly separates global and local scene interpretation into two modules: a Scene-Aware Reasoner (SAR) classifying site-level context (e.g., construction vs. mining site), and a Region-Focused Neural Reasoner (ReFiNER) verifying PPE compliance and machinery presence within annotated bounding boxes. The two representations are gated-fused and projected into a LoRA-fine-tuned Flan-T5 model, enabling natural language scene descriptions, region-grounded compliance answers, and open-ended VQA within a single framework, with demonstrated deployability on edge hardware (Jetson Orin NX) for unmanned monitoring. Subsequently, Chan et al. [63] presented a framework that leverages semi-automatic annotations generated through advanced multimodal models, overcoming the problem of data scarcity and reducing the time and effort required to prepare accurate datasets. In addition, the system has been refined with a two-stage curriculum learning approach, which allows the model to first learn the general and easy construction site concepts and then focus on more specific safety rules. This gradual training process helps overcome the limitation of traditional methods that attempt to learn everything in a single step, sometimes confusing general and specific information.

A further step forward is introduced by Xie et al. [61], who propose a system that represents the scene as a 3D point cloud and models workers' movements over time using a spatio-temporal network, as shown in Fig. 5.

The goal is not only to recognize objects and positions, but to interpret human–object interactions more realistically, exploiting depth and movement dynamics. This approach allows for a more complete reading of operational activities and potential risks, overcoming the limitations of 2D methods based on single frames.

To conclude, this analysis shows how increasingly advanced vision-based techniques are being applied to achieve a more automatic and comprehensive understanding of construction site scenes. These systems aim to support safety managers by monitoring rule compliance, enhancing situational awareness, reducing the need for traditional sensor-based systems, and taking an important step toward smarter and more proactive safety management on construction sites. Fig. 6 displays the general workflow commonly adopted by construction site safety monitoring pipelines, which can be generally summarized in five phases: visual data acquisition refers to the collection of visual

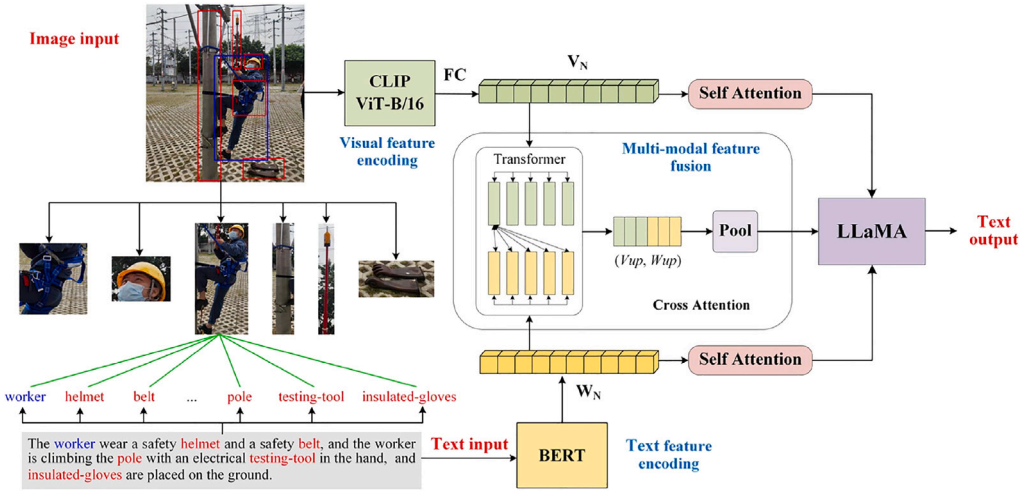


Fig. 4. Illustration of SafetyGPT [62], presented as an example of a framework applied to the construction site domain. The image shows how the system combines the analysis of images with encoded safety rules to identify hazardous situations. Figure licensed under CC BY-NC 4.0.

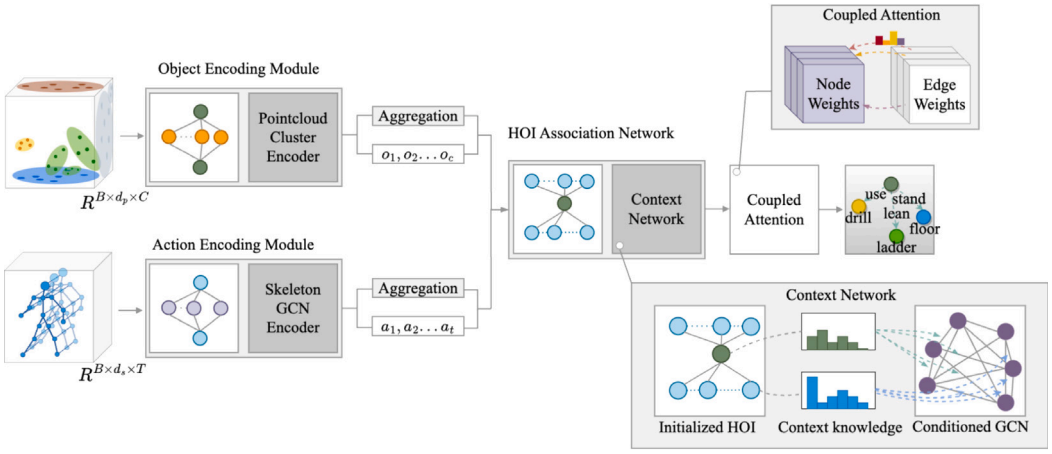


Fig. 5. Representation of the 3D Human–Object Interaction framework [61], as an example of a system applied to the construction environments domain. The image shows how spatio-temporal and geometric features are integrated through graph-based networks to infer human–action–object interactions. Figure licensed under CC BY-NC-ND 4.0.

data from the environment using on-site cameras; feature extraction involves processing the acquired data to identify relevant elements; relational understanding infers interactions and relationships between entities exploiting the extracted features; safety reasoning leverages this understanding of the scene to detect unsafe situations; and safety feedback translates the outcome of the pipeline into information for the user.

3.2. Surgical environments

Scene understanding is increasingly being used to offer intelligent support during operations in surgery, evaluating surgical skills, providing real-time feedback, or automatically generating surgical descriptions and reports. Over time, research in this field has evolved across various levels of detail, from detecting surgical tools or surgical phases to recognizing the steps within each surgical phase and transitions between phases. In recent years, attempts have been made to reach a finer level of recognition that considers the entire context and relationships between entities, and to provide more comprehensive and contextual support to the surgeons [41,45].

Many studies related to understanding the surgical scene focus on the development of support systems to assist surgeons during laparoscopy. Laparoscopy is a minimally invasive technique in which the surgical field is accessed through small incisions and visualized thanks to a camera inserted into the body of the patient. This procedure presents a significant technical difficulty for the surgeon due to indirect vision and the use of laparoscopic instruments. This high level of complexity has greatly driven the development of context-aware support systems for surgery, namely, systems capable of assisting surgeons based on the perception and dynamic understanding of the surgical scene and workflow. For instance, Nwoye et al. [41] introduced Rendezvous (RDV), a Transformer-based model for automatic recognition of surgical actions in images from laparoscopic videos, which are described as instrument–verb–target triplets. The model offers a structured representation that explicitly identifies, in real time, who is doing what and to whom in the surgical scene. To provide more complete and specific feedback on the ongoing laparoscopic operation, Murali et al. [24] proposed Latent Graph Critical View of Safety (LG-CVS). This method was thought for assessing Critical View of Safety (CVS) in laparoscopic cholecystectomy, which is the minimum safety standard for correctly identifying anatomical structures before dissecting or ligating. Through CNNs, scene graphs are extracted, and thanks to a Graph Neural Network (GNN) and two decoders, the model predicts the three CVS criteria, ensuring safer surgical decision-making. Thereafter, Zhang et al. [53] presented TAI-GNN, a GNN-based solution for tracking anatomical landmarks during laparoscopic surgery. During

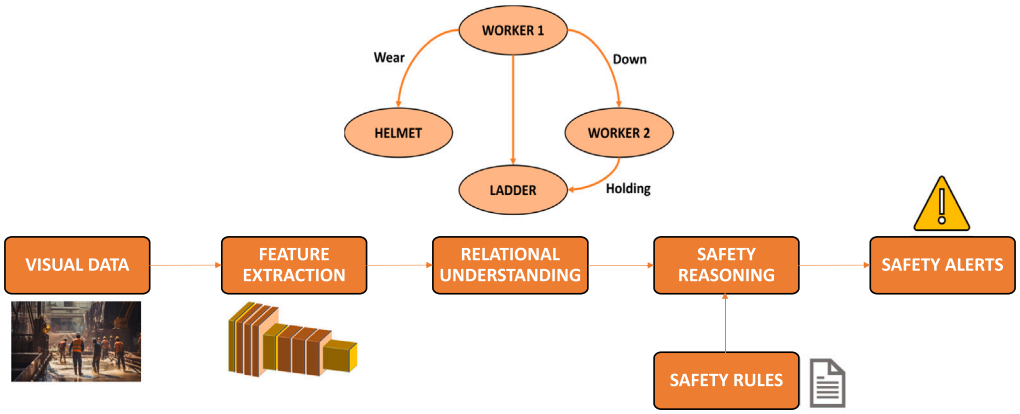


Fig. 6. Overview of a generic pipeline for automated safety monitoring in construction sites. Visual data acquired on site is processed through feature extraction modules, followed by relational understanding components that model interactions within the scene. These structured representations enable higher-level analysis and safety reasoning, comparing the relational understanding of the scene with the safety rules.

this procedure, indeed, the appearance of tissues changes continuously due to interactions between instruments and anatomical structures. Furthermore, tissues are often partially obscured by instruments or other anatomical structures. Unlike previously analyzed methods, this framework also considers the temporal relationships between frames, providing more accurate output on the temporal evolution of the tissues. Similarly, Lin et al. [54] introduced ITIDNet, a framework to detect instrument–tissue interactions in surgical videos, which also integrates temporal information. The framework exploits a Temporal Graph Layer that constructs an interaction graph, modeling intra-frame and inter-frame relationships to predict actions between instruments and tissues over time. The final output provides a fine-grained and temporal representation of the surgical scene. A distinct contribution addressing the temporal dimension of surgical scene understanding at a more foundational level is offered by Yang et al. [69], who introduced SurgVISTA, the first video-level self-supervised pre-training framework for surgical scene understanding. Unlike prior surgical foundation models restricted to image-level pre-training, which decouple spatial representation learning from temporal modeling introduced later during fine-tuning, SurgVISTA jointly learns spatial and temporal representations from 3650 surgical videos through a masked video reconstruction objective, in which randomly selected spatiotemporal regions of the input video are hidden and the model is trained to reconstruct them from the surrounding visual context, requiring no manual annotation at this stage; labeled data is instead reserved for the subsequent fine-tuning phase, applied separately to each downstream task. This pre-training objective is further refined via knowledge distillation from an image-level surgical expert model to preserve fine-grained anatomical detail. Rather than targeting a single downstream task, SurgVISTA is evaluated as a general-purpose backbone across four clinically relevant tasks, surgical workflow recognition, action recognition, triplet recognition capturing instrument–tissue interactions, and skill assessment based on the Critical View of Safety, consistently outperforming both natural-domain and surgery-specific pre-trained models, particularly on out-of-domain procedures (that is, unseen during pre-training). This positions video-level self-supervised pre-training as a scalable alternative to task-specific architectures for capturing the dynamic, relational structure of surgical scenes.

Another application of these techniques for understanding the surgical scene is the automatic generation of surgical descriptions and reports. Automatic surgical reports are essential since they help surgeons to focus on the operation and record information about the entire surgical process for post-operative analysis and evaluation, thereby helping to improve clinical practice. For example, Xu et al. [40] proposed a Transformer-based framework to improve the automatic generation of surgical reports where a strong domain shift between datasets occurs. The model combines a Transformer encoder–decoder with gradient reversal adversarial learning to learn domain-invariant features and label smoothing, which improves model calibration and reduces overconfidence. The result is a model that is more robust to domain shift and more reliable in generating tool-tissue relationships. Similarly, Lin et al. [52] presented SGT++, another Transformer-based model to generate reports from endoscopic images, to improve the accuracy of hidden relationship identification The model processes the image together with its scene graph, using an explicit encoder for observable relationships and an implicit encoder with prototype memory for recurring but hidden relationships. In another study, Wang et al. [47] took a further step forward by dynamically updating scene graphs through video frames, capturing temporal interactions via CNNs to produce more comprehensive, time-aware, and clinically consistent reports.

Some other studies tackle complementary essential aspects for robust surgical scene understanding, such as computational efficiency, human-machine interaction intuitivity, and generalization. Pang et al. [45] proposed an end-to-end approach for analyzing surgical scenes that uses Grad-CAM and Grad-CAM++ for localized feature extraction, eliminating traditional object detectors and pipelines. The features thus extracted feed downstream tasks of scene captioning and scene graph generation. The approach reduces computational complexity, minimizes the need for costly annotations, and enables real-time processing. Yuan et al. [58] explored a different direction by introducing the SSG-VQA-Net model for Visual Question Answering in surgery, whose pipeline is represented in Fig. 7.

The system combines visual features with a scene graph that encodes instruments, anatomical structures, and their relationships, enabling it to answer complex natural language questions from the user about actions, locations, or interactions during the procedure. This approach aims to make human-machine interaction more intuitive, allowing surgeons to question the system directly in their own terms. Building on this line of visual question answering for surgery, Low et al. [72] propose SurgRAW, a multi-agent framework that orchestrates VLM agents via structured Chain-of-Thought prompting, dividing reasoning into perceptual (instrument and action recognition) and inferential (outcome and patient-context prediction) streams that explicitly reason over instrument–tissue interactions and spatial relationships. A panel-discussion mechanism cross-checks predictions against a static instrument–action knowledge graph, while a retrieval-augmented module grounds contextual inferences in external clinical knowledge, improving reliability over single-pass VLM reasoning.

Xu et al. [51] focused instead on improving the reliability of model training in surgical contexts. The authors presented P-CBLS (Paced Curriculum By Label Smoothing), a framework designed to overcome problems such as overconfident predictions, weak generalization, and sensitivity to noisy labels. Their method employs a paced curriculum, where

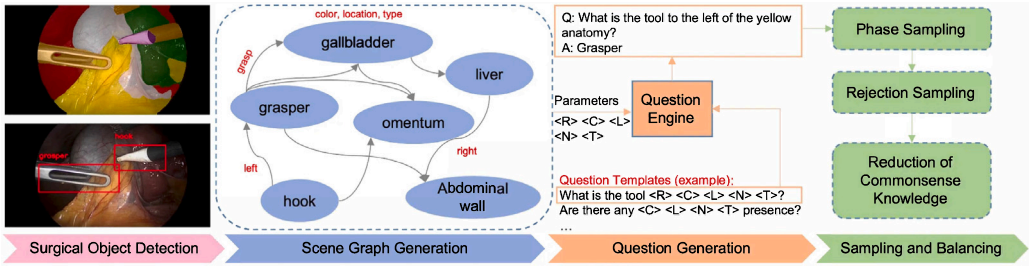


Fig. 7. Overview of the SSG-VQA pipeline developed by Yuan et al. [58] for Visual Question Answering in surgery, to improve intuitive interaction during intraoperative procedures. Figure licensed under CC BY 4.0.

more challenging samples are added only when the model is ready, along with adaptive label smoothing that changes from uniform to spatially variable during training. This approach helps the network handle uncertainty more efficiently, providing more reliable, well-calibrated models for various surgical tasks, such as scene captioning.

The studies cited so far focus on understanding the surgical scene as the area of intervention, recognizing the instruments and actions directly involved in the procedure. However, to provide real support for interventions, it is essential to develop a holistic understanding of the entire operating room, including all people, devices, objects, and their dynamic interactions. Such a comprehensive understanding would allow digital systems to automatically adapt to dynamic procedures, optimizing the workspace and clinical outcomes. In this case, traditional scene graph pipelines based on 2D images provide a useful semantic description. However, they are limited by the lack of geometric and depth information, which are essential for precise spatio-structural relationships in the presence of entities with strong variations in scale and occlusions in operating rooms. On the other hand, 3D representations, while providing information on geometric properties and allowing more spatially sensitive predictions of relationships, remain poor in semantic information. Besides that, annotating 3D instances is difficult and costly. For this reason, the most recent pipelines adopt a bimodal approach that exploits the strengths of both modalities, combining 2D and 3D information to obtain a complete representation of the scene [59]. Özsoy et al. [57] developed a bimodal framework that generates Semantic Scene Graphs to represent people, tools, and their interactions in the operating room, exploiting 3D human and object poses. To improve the accuracy, additional specific strategies for the surgical environment are implemented, such as virtual nodes for small or transparent instruments, which can be hard to localize in the point cloud. Furthermore, this method also considers temporal relationships between frames, allowing a better understanding of temporal interactions during surgical operations. The resulting graphs enable downstream applications, such as clinical role prediction and surgical phase recognition, by directly exploiting the semantic structure of the graph. Fig. 8 shows an overview of the pipeline.

The approach is multi-stage, involving a sequence of pose estimation, object detection, and scene graph construction, in which each module depends on the results of the previous one. However, this approach has two main limitations: (1) The intermediate steps could cause a loss of context, making the global dependencies understanding less accurate. (2) Training and executing multiple modules in cascade implies high computational complexity. To overcome these limitations, Pei et al. [59] proposed S2Former-OR, a single-stage framework based on a bimodal Transformer architecture that directly generates the scene graph in an end-to-end way, without intermediate detection or estimation steps. This end-to-end approach provides a more consistent scene modeling, while reducing computational costs and improving accuracy compared to previous multi-stage models. In another more recent study, Özsoy et al. [64] focused on advancing multimodal scene understanding in surgery by introducing MM2SG, a framework for generating scene graphs from diverse intraoperative data. The model is designed to overcome the limitations of previous approaches that rely solely on video input, such as incomplete scene representation, inability to capture audio or robot signals, and limited context modeling. MM2SG integrates multi-view RGB-D video, high-resolution detail views, audio, speech transcripts, robotic system logs, infrared tracking data, and segmentation masks. This allows the framework to capture both dynamic interactions and spatial relationships between surgical staff, instruments, and equipment. The framework has been evaluated on scene graph generation and downstream clinical tasks, including surgical phase prediction, next action anticipation, and sterility breach detection, demonstrating enhanced holistic understanding of complex surgical environments.

In summary, this analysis reveals how automatic understanding of surgical scenes has been addressed from different perspectives, each focusing on specific clinical or technical objectives. Despite this, all the studies have the same final goal of improving situational awareness and intraoperative support, making artificial vision systems increasingly integrated in the surgical context. Following the literature analysis, the methods related to the surgical field can be categorized into three general workflows according to the kind of support given, as shown in Fig. 9. Specifically, the first two pipelines refer to studies for intraoperative support and automatic report generation, respectively. They share the first three phases, including visual data acquisition, feature extraction, and relational understanding. However, methods for automatic report generation require an additional step performed through a language transformer for textual output generation. On the other hand, approaches for OR analysis, while keeping feature extraction and relational understanding steps, require multiview visual data and sometimes multimodal data in the acquisition phase for a more holistic scene understanding, including the role of the medical staff members and their actions and interactions among them and with the surrounding objects.

3.3. Assistive environments for special needs

For people with special needs, the ability to move around and carry out daily activities independently is essential, which is the reason why research is increasingly focusing on designing tools that promote this autonomy [25,56,74]. Traditional tools, such as canes, guide dogs, or human assistants, offer support but have limitations: canes need the use of hands and may be uncomfortable to use in crowded environments, dogs require training, and they are not always allowed everywhere, and human assistance reduces the independence of the user [74]. To overcome these limits, numerous assistive systems based on wearable and non-wearable sensors have been developed. Wearable devices include accelerometers, gyroscopes, and cameras to process visual data in real time. Non-wearable sensors may include environmental cameras or assistive robots. These devices offer support to users with different needs, allowing them to receive direct assistance or audio descriptions of the scene, detect obstacles, and interact with voice apps for the home [56,74]. Among these applications, systems based on images or videos are promising, as they can provide complete and contextual

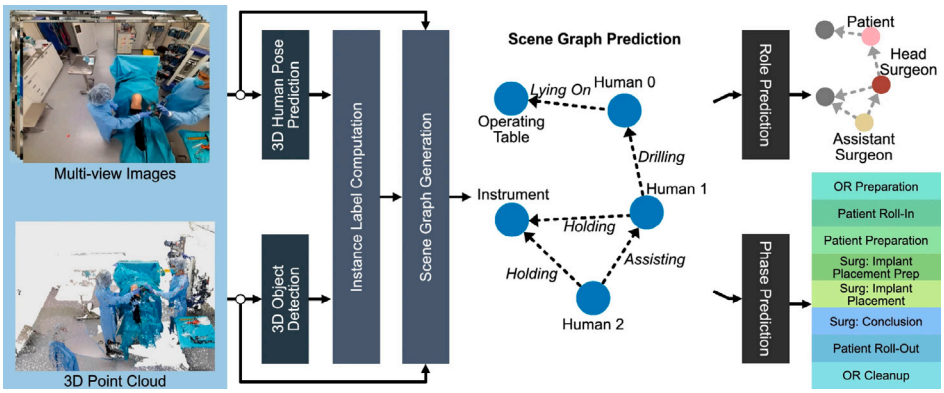


Fig. 8. Overview of the multi-stage pipeline by Özsoy et al. [57], showing how 2D and 3D data are combined to generate Scene Graphs that describe people, instruments, and their interactions in the operating room. Figure licensed under CC BY 4.0.

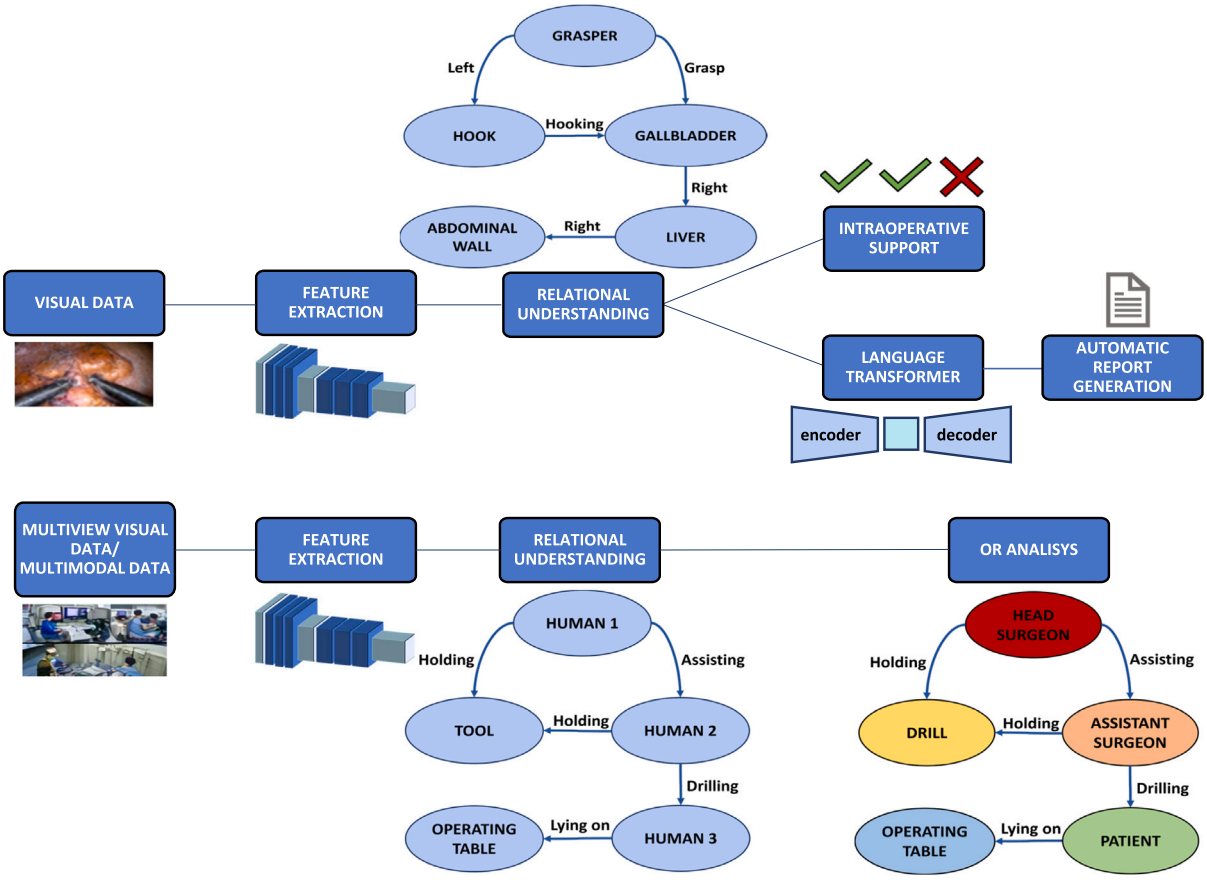


Fig. 9. Schematic overview of the pipelines for surgery support. Visual data acquired during the procedure are processed through feature extraction modules, followed by relational understanding components that model interactions within the surgical scene. These structured representations enable higher-level analysis and support downstream tasks, including intraoperative support, automatic surgical report generation through transformers, and holistic operating room analysis.

visual information. For this reason, many studies in this field focus on video frame analysis to recognize the scene surrounding the user [56]. Early solutions relied on CNNs to extract low-semantic visual features. For example, many systems developed to guide the visually impaired were based on simple obstacle detection, integrated with depth or laser sensors to estimate distance [74,75]. While effective for basic obstacle avoidance, these systems compute spatial relations between the user and surrounding objects through fixed geometric rules applied downstream of detection, rather than modeling them explicitly. Researchers subsequently began to adopt more advanced neural architectures capable of learning richer relational and contextual representations of the scene. Cao et al. [55] introduced DFFTNet, a Transformer-based image captioning system for wearable navigation devices, which replaces traditional algorithms for distance or motion estimation with richer relationship reasoning. The model improves the clarity and precision of scene descriptions and provides users with more consistent feedback. Building on this line, Yuan et al. [65] introduced WalkVLM, a vision-language model for streaming video analysis in blind walking assistance. Unlike prior semantic-based approaches that process single frames, WalkVLM employs chain-of-thought hierarchical planning to structure scene understanding into perception, comprehension, and decision levels, and introduces a Temporal-Aware Adaptive Prediction module that leverages historical frame and state sequences to determine the optimal moment to issue a reminder, reducing redundant outputs

while preserving timely, opportune guidance during continuous walking. While these approaches focus on improving the perceptual and temporal quality of the generated scene descriptions, Zhao et al. [73] addressed a complementary challenge: aligning navigation instructions with how visually impaired users actually interpret and act upon them. The authors proposed LaF-GRPO, se a second LLM that simulates how a visually impaired user would interpret and act on each instruction, and reward the model based on whether that simulated action is correct. This avoids the need for costly trials with real users. The method produces instructions that are clearer and more accurate than those from strong VLMs such as GPT-4o.

Further studies have focused on assistance during daily activities, where systems such as domestic robots collaborate with the users and simplify their tasks. For example, Dessalene et al. [49] developed a framework to understand and anticipate actions in egocentric videos based on hand–object interactions. The relationships between hands and objects are estimated through CNNs. Depth information is also exploited thanks to a 3D CNN, while a Graph Convolutional Network (GCN) models temporal relationships between states, and an LSTM projects sequences into the future to anticipate upcoming actions. The method enables both prediction and anticipation of actions. Subsequently, Su and Chen [56] proposed a framework to recognize more general daily activities using scene graphs that are elaborated by a GNN to predict human activities and the objects involved. The framework can be integrated with smart home devices, enabling real-time monitoring and support for daily activities. In this way, a more efficient and generalizable approach based on the analysis of human–object interactions is offered. A substantially more ambitious effort toward long-term daily activity assistance is presented by Yang et al. [66], who introduce EgoLife. The associated EgoButler system combines EgoGPT, an omni-modal (visual–audio) egocentric captioning model, with EgoRAG, a retrieval-augmented module indexing captions into a hierarchical memory (clip-, hour-, and day-level) for question answering over ultra-long temporal contexts. Unlike prior assistive systems addressing isolated daily tasks, EgoLife explicitly targets relational understanding sustained over days, including recurring object usage and location, past events, personal habits, interpersonal interaction patterns among household members, and task-relevant reminders, demonstrating that useful life assistance requires reasoning about relationships between multiple people and their environment over time. More recently, Huang et al. [70] proposed EgoDynamic4D, an end-to-end framework for egocentric 4D scenes that fuses instance-aware point features with a Fourier-based temporal encoder and a cross-attention camera embedding module, projecting the resulting representation into an LLM via LoRA fine-tuning. Similarly to EgoLife, the model explicitly reasons over human–object and human–human spatial relations as they evolve across time, capturing trajectories, distances, and interactions between the wearer and surrounding agents and objects within domestic environments. More recently, Sajid et al. [71] proposed TransFlow-HOI, a transformer-based framework that fuses RGB and optical-flow features extracted via I3D through multi-head self-attention, followed by an ensemble of SVM, Random Forest, and XGBoost classifiers, evaluated on the egocentric HOI4D dataset. Compared to the spatio-temporal relational models of Huang et al. [70] and Yang et al. [66], this method targets fine-grained discrimination between visually and motion-wise similar manipulation actions on the same object (e.g., picking up, repositioning, or placing an item inside a container), showing that attention-driven cross-modal fusion of appearance and motion cues can improve recognition of subtle interaction variants relevant to daily-activity assistance.

Finally, two works focusing on scene recognition in a broader sense were also identified, aimed at interpreting and describing the surrounding scene to blind users. In 2023, Liu et al. [46] developed OpenSU, an Open Scene Understanding system that addresses limitations such as ambiguous object localization and incomplete scene descriptions. The framework exploits a Transformer architecture to analyze a single image and create a detailed scene description and accurate object masks, helping users understand and interact with the environment. Later on, Kim et al. [60] designed a framework inspired by how humans comprehend scenes. It combines a biologically inspired feature extractor, a Vision Transformer (ViT)-based encoder, and an LSTM decoder optimized through reinforcement learning. The output consists of voice messages that help blind users comprehend the scene around them. Fig. 10 represents the process of the described image captioning model.

In summary, computer vision for scene understanding in the field of assistive technology is still in its early stages, but it shows great potential for future developments. Current studies address the problem from different perspectives, from independent navigation to home assistance, with the common goal of promoting users' autonomy. The general pipeline can be summarized in four phases, as represented in Fig. 11. Specifically, in the acquisition phase, visual data collected can be combined with additional sensor data to provide more accurate support, while in the feature extraction phase, the acquired data are processed to identify relevant information. Then, the following step exploits this knowledge to obtain a relational understanding between entities, on which the final feedback provided to the user is based.

4. Discussion

Recent advances in AI have enabled a significant evolution in how visual data is analyzed and interpreted. DL in particular has shown a considerable ability to efficiently and accurately identify complex features in images, leading to a complete and accurate understanding of the scene. Early scene understanding methods focused on low-level semantic tasks, like object detection or segmentation. Nowadays, research increasingly focuses on finding a higher-level meaning in visual data, shifting to modeling interactions and relationships between entities and their surrounding environment. This evolution reflects the growing need for systems capable of reasoning about complex environments rather than merely recognizing isolated visual elements. This shift enables a more holistic comprehension of context, which has been shown to significantly enhance scene interpretation, particularly in visually complex or ambiguous situations [11,13].

Since human activities are central to most real-world scenarios, a growing area of research focuses on using these technologies to enhance individuals' performance in daily and professional contexts. The goal is not only to recognize objects or actions but to provide meaningful support that improves safety, efficiency, and situational awareness [1]. In contexts where individuals perform complex tasks in dynamic, information-rich environments, it is essential to adopt an advanced contextual understanding approach that can significantly improve scene understanding and provide more accurate feedback to the user [13,14].

The conducted analysis focused on this aspect, namely, to explore the areas in which these scene understanding techniques are most widely adopted to provide support to people. It identified the domain-specific needs that motivate the adoption of these technologies and examines how this support is provided based on those needs. Additionally, the analysis identified critical issues and recurring challenges across domains, highlighting aspects that currently limit efficacy or real-world implementation.

The main application domains that emerged from our analysis are, in order of prominence, surgical environments, construction environments, and assistive environments for people with special needs. This finding directly answers the first research question (RQ1) regarding the primary domains of application of such techniques. The findings reveal that the use of high-level scene understanding is evolving along different paths depending on the domain, due to specific technological and operational requirements. Despite their differences, the three domains share challenges related to the adoption of DL and computer vision.

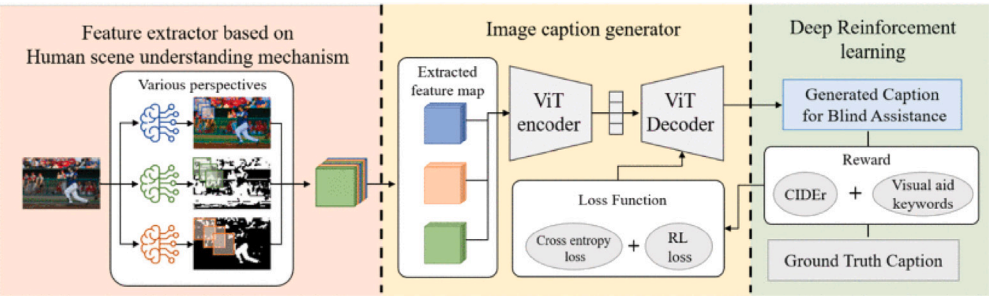


Fig. 10. Process flow of the image captioning framework by Kim et al. [60] presented as an example of a framework for assisting the visually impaired. Figure licensed under CC BY 4.0.

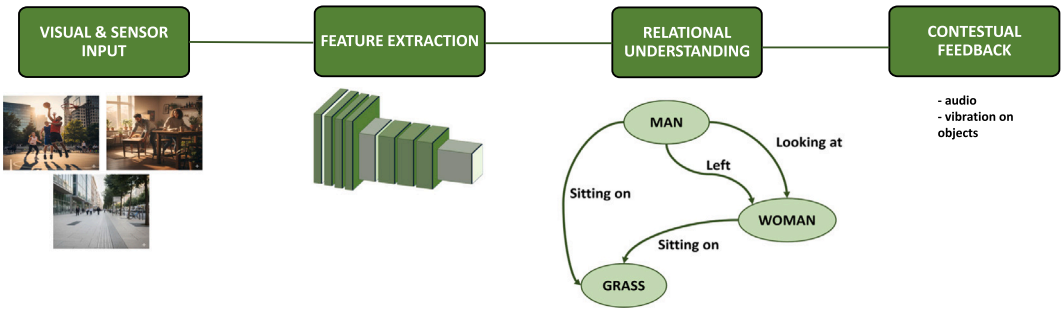


Fig. 11. Overview of a generic pipeline to assist users with special needs. Visual data acquired, often accompanied by sensor data, is processed through feature extraction modules, followed by relational understanding components that model interactions within the scene. These structured representations enable a higher-level understanding of the scene.

The following sections will examine these aspects in detail, exploring the specific needs of each domain and how they offer targeted support to users. At the same time, the analysis uncovers critical limitations and gaps associated with current approaches. This dual perspective addresses the second research question (RQ2), which concerns the alignment between user needs and technological support and is discussed in Section 4.1, and the third research question (RQ3), focusing on identifying unresolved issues and weaknesses that may guide future research, as examined in Section 4.2.

4.1. Domain-specific needs and technological forms of support

The requirements for supporting human activities vary considerably across construction, surgical, and assistive environments for people with special needs. However, they all share the common goal of increasing safety and efficiency. Fig. 12 summarizes the domain-specific needs identified in the analyzed studies.

Construction sites are among the most dangerous workplaces, with a high risk of accidents that are often serious or fatal [39]. Such accidents could be prevented if all workers correctly complied with safety regulations and wore PPE. However, factors such as distraction, fatigue, or habit often lead to violation. Since continuous human monitoring is expensive and difficult to implement on a large scale, there is a growing demand for automated monitoring [39,43]. To address this necessity, systems have been developed that integrate real-time video monitoring from cameras installed on construction sites with scene understanding models and NLP techniques. Many of these systems follow a similar pipeline structure. The first phase consists of visual perception, in which objects, workers, and PPE are detected from images or video streams acquired by on-site cameras. This stage uses models for object detection or segmentation, such as YOLO, Mask R-CNN, or ResNet-based encoders. In some pipelines, this phase also includes human pose estimation or spatiotemporal modeling to capture worker postures, movements, and interactions more accurately. These visual entities and relationships are then organized into a structured representation of the scene in the form of scene graphs or relational triplets, using methods like geometric reasoning, Graph Neural Networks, Transformers, or spatio-temporal graph models. This intermediate level of understanding

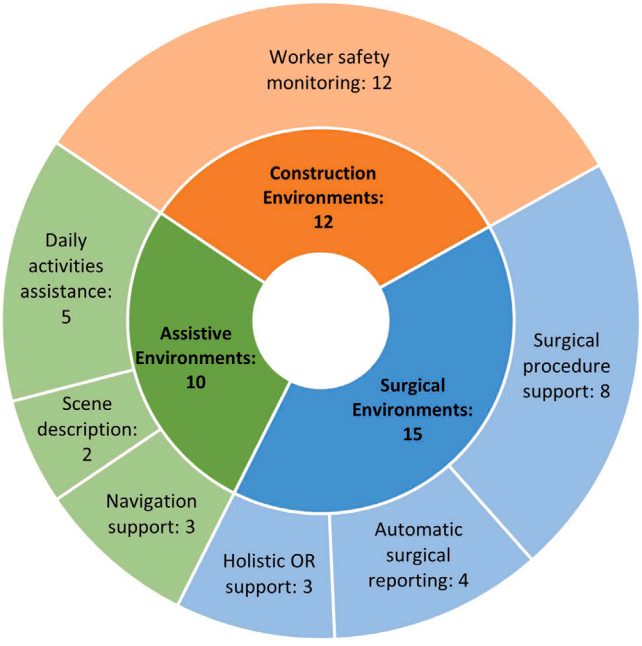


Fig. 12. Pie chart summarizing the analyzed studies, grouped by application domain and in turn by domain-specific needs. More specifically, the inner ring groups the studies into the three application domains (surgical environments, construction environments, and assistive environments for people with special needs), while the outer ring represents the specific needs identified within each domain.

enables the system to move beyond simple object presence and reason more about spatial relationships and unsafe configurations. Finally, the system checks for safety by integrating knowledge, like safety rules and supervisor questions, expressed in natural language. NLP models are used to encode this textual information and align it with the visual scene representation. By comparing the observed scene with expected safety conditions, the system can find violations of PPE, unsafe behaviors, or other non-compliance issues [14,42–44,48,50,61–63]. The goal of these systems is not to replace human supervision, but to enhance it, reducing the reliance on continuous human monitoring and increasing the timeliness of reports. In fact, especially in more recent works [62,67,68], the focus is shifting toward the design of more intuitive and interactive interfaces, where supervisors are not mere observers but can actively interact with the system to request checks or further information.

In the surgical field, the main focus is on developing intraoperative support systems to improve the surgeon's situational awareness. These pipelines typically start from visual data acquired during surgery, from where semantic information about instruments, anatomical structures, and surgical workflow is extracted. Surgeries, particularly laparoscopy, have limited visibility and a high degree of complexity, requiring a precise understanding of instruments, tissues, and the progression of the procedure. For this reason, surgeons need assistance during complex operations through real-time feedback on instruments and tissue interactions. To do so, a common pipeline relies on convolutional backbones for feature extraction, such as ResNet-based architectures, followed by modules for instrument detection, action recognition, or phase identification. These pipelines often model instrument–tissue relationships or safety-relevant criteria, such as the Critical View of Safety, using structured representations and relational reasoning modules, including graph-based models and Transformers [24,41,45,51,53,54,58,69,72]. Another need in this field is the automatic generation of surgical reports, which allow surgeons to focus on the operation and record information about the entire surgical process for post-operative analysis and evaluation. In this case, visual or graph-based representations of the scene are combined with Transformer-based language decoders, enabling the generation of structured textual descriptions that summarize the surgical process [40,45,47,52]. Recently, the focus has expanded to understanding the entire operating room scene, rather than just the limited field of view during surgery. This holistic perspective aims to capture the interactions between all personnel and equipment, dynamically adapting the information provided to the surgeon and contributing to clinical improvements. To achieve this, these pipelines integrate multi-view visual data, and in some cases multimodal signals [57,59,64].

Regarding assistive technologies for users with special needs, the main requirement is to promote independence and improve environmental perception [25,56]. This type of assistance is provided in three main ways: supporting autonomous navigation for blind people to improve their orientation and safety [55,65,73]; providing comprehensive and detailed descriptions of the user's surroundings to help them understand the environment around them [46,60]; assisting users with daily activities, where systems such as domestic robots collaborate with them to simplify tasks [49,56,70,71]. Typically, these pipelines begin with a perception stage, where visual data are acquired through cameras, and visual features are extracted from the acquired images using CNNs or ViTs. The next stage focuses on semantic understanding, where the extracted features are processed to identify relevant entities and relationships. Here, DL architectures such as CNN, transformer encoders, and GNN are frequently employed to capture contextual and relational information. In some systems, 3D and temporal cues are integrated to produce a more coherent representation of the environment [66,70]. This step is essential for activity assistance and anticipatory systems, where temporal modeling like LSTMs or temporal convolutional networks enables the prediction of future actions or events. A related but distinct need is temporal opportuneness of the feedback itself: Yuan et al. [65] show that navigation guidance must not only be semantically correct but also issued at the right moment, addressing this through a dedicated module that suppresses redundant reminders based on historical frame and state sequences. In the case of systems designed for visually impaired people, contextual feedback is provided via audio or vibration on objects. Furthermore, in the case of systems to assist navigation, cameras to capture the scene are often integrated with other sensors, such as depth sensors or IMU sensors, to provide more accurate support based on more complete and contextual information [55]. These technologies are not limited to detecting and classifying visual elements, but must also synthesize the context in a meaningful way, often through multimodal pipelines that combine visual inputs, semantic modeling, and, when necessary, language generation. Fig. 13 summarizes, for each domain and year, which relational modeling mechanism underlies each of the studies discussed above, categorized into graph-based approaches, which represent relationships explicitly as graph structures processed via message passing; transformer- or attention-based approaches, which instead learn relational dependencies through self- or cross-attention without an explicit graph structure; CNN-based mechanisms; recurrent mechanisms; and hybrid combinations of these techniques. It should be noted that, given the literature search was conducted with a cutoff in July 2026, the year 2026 is only partially represented. Accordingly, apparent declines or plateaus in 2026 should be interpreted with caution, as they are likely to reflect this partial coverage rather than an actual slowdown in adoption. Graph-based approaches are the most frequently adopted mechanism in the surgical domain throughout the reviewed period, while transformer- and attention-based approaches become the prevailing mechanism in construction and assistive environments from 2025 onward. CNN-based and recurrent mechanisms, by contrast, appear only in earlier studies (2021–2023) across all three domains. The choice of mechanism reflects the specific architectural strategy adopted by each study rather than a hierarchy of relational sophistication in itself; the level of relational understanding achieved depends on how each mechanism is integrated within the broader framework, as discussed in the following section.

4.2. Domain-independent challenges and future guidelines

The analysis of construction environments, surgical environments, and assistive environments for people with special needs domains shows a common trend toward increasingly sophisticated levels of scene understanding. In the case of construction sites, for example, the earliest studies analyzed base scene understanding on relatively simple neural networks, such as those for object detection or human pose estimation, from which geometric relationships are then derived to construct graphs related to images [39,43]. Subsequently, technological evolution has led to the adoption of more advanced architectures, including networks based on multiple CNNs, GCNs, or Transformers, capable of extracting more complex and representative information [14,42,44,48,50,61–63]. A similar dynamic can be found in assistive technologies for people with special needs. For example, the DeepNavi system [74] uses simple obstacle and scene recognition models, while interactions between the user and other entities in the environment are estimated using depth or motion sensors. The growing need to obtain a more accurate and complete understanding of the environment has driven research towards the use of more advanced neural architectures capable of modeling deep and complex relationships between multiple entities [55]. A similar evolution can be observed in the field of surgical environments, with a gradual shift from understanding only the operating area to a more holistic view of the entire operating room [57,59,64]. In this context, work has also begun with three-dimensional data, which allows for the integration of geometric and depth information, improving the ability to capture the spatial placement of instruments, tissues, and operators. Recently, a shift towards 3D has also been observed in the field of construction environments.

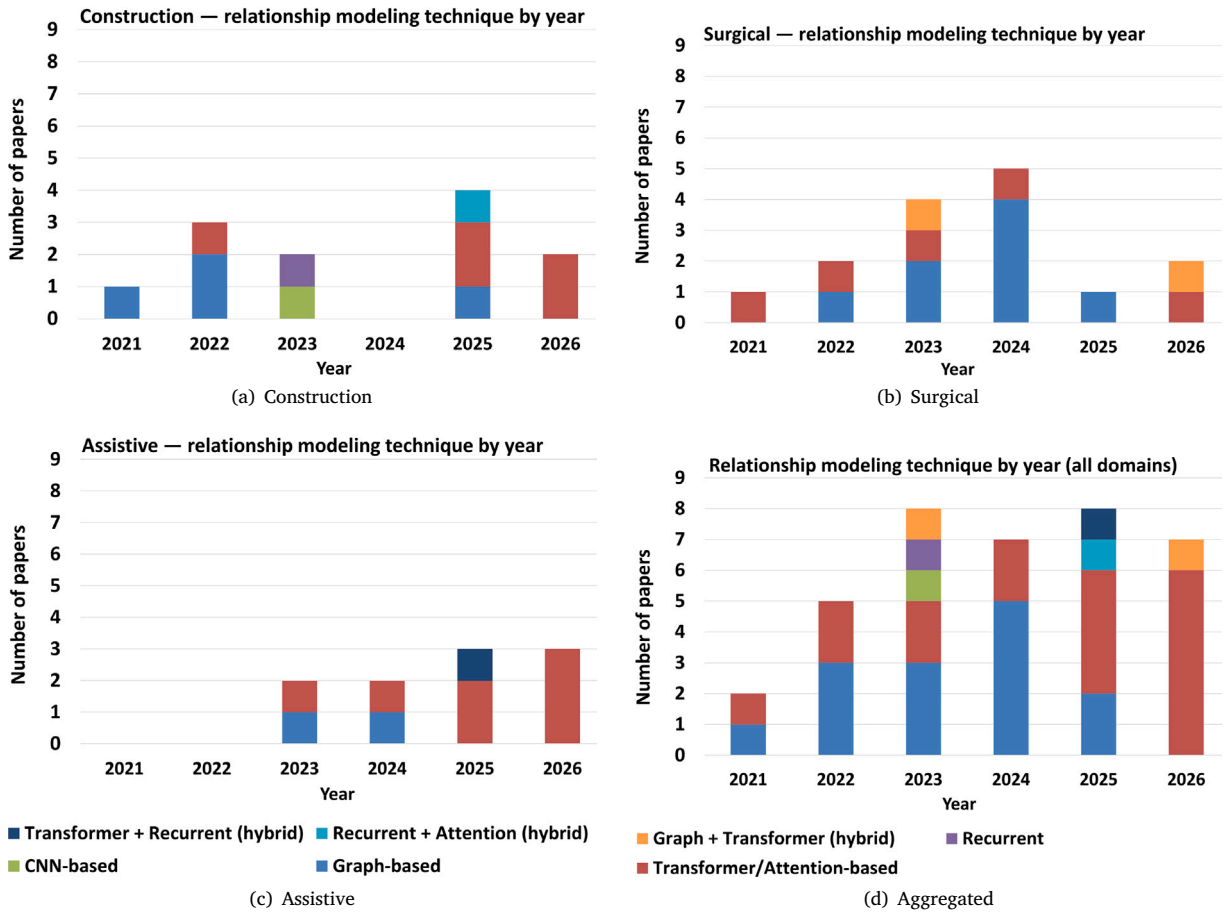
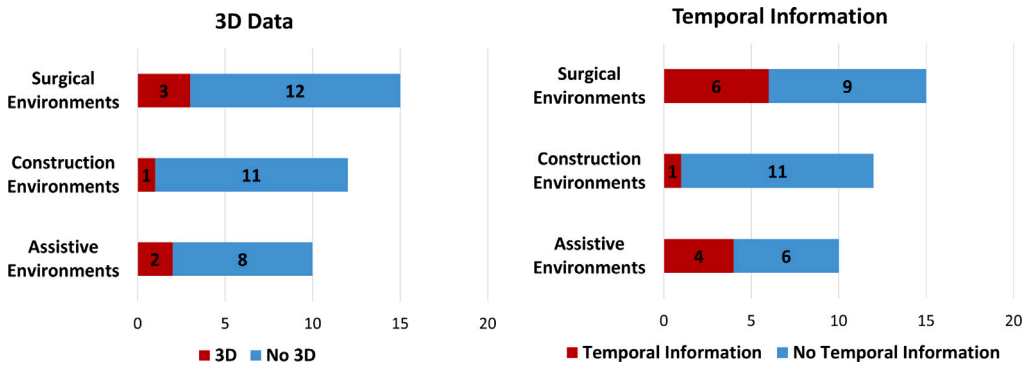


Fig. 13. Relational modeling technique adopted across surveyed papers, by year. (a)–(c) Per-domain breakdown for Construction, Surgical, and Assistive environments, respectively, with bar segments stacked and color-coded by technique family (graph-based, transformer/attention-based, CNN-based, recurrent, and hybrid variants). (d) Aggregated distribution across all three domains combined.

The framework developed by Xie et al. [61] in 2025 represents the first example of an explicit transition from 2D to 3D, with accurate results in detecting positions and spatial relationships, reducing ambiguities related to perspective or occlusions. In practice, across the reviewed corpus, this 3D data is predominantly represented as point clouds, typically derived from RGB-sensors, often complemented by 3D human and object pose estimation. Point clouds support scene graph generation in holistic operating room modeling by capturing the spatial layout of instruments, tissues, and personnel [57,59,64]; support spatio-temporal human–object interaction graphs in construction monitoring [61]; and, in egocentric assistive settings, they allow for obtaining fine-grained modeling of human–object and human–human spatial relations [70]. This representation is particularly suited to these contexts because it preserves accurate geometric and spatial relationships among multiple entities. Despite the advantages it would bring, the integration of 3D data has still been little explored in all three areas, mainly due to the limited availability and high cost of labeled datasets, and to the increased computational complexity of 3D processing. As represented in Fig. 14(a), this information has been utilized in only three studies in the field of surgical environments, in one study in the field of construction environments, and in two studies in the field of assistive environments for individuals with special needs.

Another aspect that is still largely overlooked is the temporal dimension, whose adoption across the application domains is illustrated in Fig. 14(b): only one study uses this information in the field of construction environments, while it has been more explored in surgical environments and in assistive environments, with six and four studies exploiting it respectively. Across the reviewed corpus, temporal modeling is most commonly achieved through graph-based approaches that extend spatial scene graphs with temporal edges or dedicated temporal layers, particularly in surgical and construction settings; other studies instead rely on recurrent architectures such as LSTMs, joint spatiotemporal pre-training via video transformers, retrieval over extended temporal windows, or explicit frequency-domain temporal encoding, as detailed below for each domain. The papers analyzed show that only recent contributions incorporate this component, despite the understanding that doing so would enhance the accuracy of predictions regarding what is occurring in the scene being examined. In the context of construction environments, Xie et al. [61] used a Spatio-Temporal Graph Convolutional Network (ST-GCN) to capture movement sequences and enhance the temporal understanding of the scenario. This represented a significant advancement over previous methods that rely solely on isolated images. In assistive environments, temporal modeling remains limited, though a small but growing set of recent studies begins to explore it from different angles. Yang et al. [66]'s EgoLife system relies fundamentally on long-context temporal retrieval, spanning up to a full week, to support habit discovery and interpersonal interaction understanding, illustrating that extended temporal reasoning may be particularly valuable for life-assistance applications requiring memory of recurring patterns rather than only immediate obstacle or scene detection. Huang et al. [70], by contrast, target short-horizon temporal reasoning, employing a Fourier-based temporal encoder to model fine-grained human–object and human–human spatial relations as they evolve over seconds within a single egocentric sequence, complementing EgoLife's long-range perspective with moment-to-moment relational precision. Yuan et al. [65] address a third, complementary



(a) Summary of the studies that utilized 3D information for each category

(b) Summary of the studies that utilized temporal information for each category

Fig. 14. Summary graphs illustrating the results of the paper analysis. Chart 14(a) indicates, for each category, the number of studies that employed 3D data. Chart 14(b) highlights, for each category, the number of studies that utilized temporal information.

facet of the temporal dimension: rather than modeling how relations or habits evolve, they predict a trigger probability from historical frames and past danger-level states, invoking the reminder pipeline only when the estimated risk changes, using temporal information to time the feedback itself, rather than to deepen scene understanding. In the surgical domain, however, temporal information has been explored more closely, reasonably due to the intrinsically dynamic nature of surgical procedures, where many relevant events and interactions can only be identified and correctly interpreted through their temporal evolution. Zhang et al. [53] used it to track anatomical landmarks during laparoscopy, highlighting that it can provide more accurate output on the temporal evolution of the tissues considered. Lin et al. [54] integrated a Temporal Graph Layer that models intra-frame and inter-frame relationships between instruments and tissues, obtaining a more detailed representation of the surgical scene. As regards the automatic generation of surgical reports, Wang et al. [47] leveraged the temporal information to produce more comprehensive, time-aware, and clinically consistent reports. This growing adoption of temporal modeling in the surgical domain is further reinforced by Yang et al. [69], who show that explicit spatiotemporal pre-training, rather than post-hoc temporal aggregation of image-level features, is critical for capturing dynamic surgical contexts, including instrument–tissue interaction temporality and causality. The frameworks developed by Özsoy et al. [57,64] also exploit temporal relationships in the holistic analysis of the operating room, improving understanding of interactions that occur over time. Fig. 15 further breaks down the adoption of 3D data and temporal information by year, revealing that both have been incorporated almost exclusively in studies published from 2023 onward, with a marked concentration in the surgical domain. It should be noted that, given the literature search was conducted with a cutoff in July 2026, the year 2026 is only partially represented. Accordingly, apparent declines in 2026 should be interpreted with caution, as they are likely to reflect this partial coverage rather than an actual slowdown in adoption.

Beyond this progressive integration of spatial and temporal information, the most recent literature also points to a further, but still narrowly represented, direction: the fusion of sensing modalities beyond RGB and depth imagery. In holistic operating room modeling, Özsoy et al. [64]'s MM2SG framework combines multi-view RGB-D video, audio, speech transcripts, robotic system logs, and infrared tracking to capture surgical context that would be inaccessible from visual data alone. A comparable, though distinct, form of sensor fusion appears in assistive environments, where the previously discussed EgoLife system additionally draws on audio and mmWave sensing alongside egocentric video to support its long-context reasoning [66], and where navigation-support systems more commonly combine wearable cameras with depth or inertial (IMU) sensors to improve distance and motion estimation [55]. Such fully multimodal pipelines remain the exception rather than the rule within the reviewed corpus, but they illustrate a direction in which richer sensing, rather than more sophisticated visual architectures alone, is used to compensate for the limitations of single-modality perception. Another related direction concerns how scene understanding is trained and expressed, rather than which sensors feed it. Rather than training a model directly on task-specific supervision, Yang et al. [69]'s SurgVISTA adopts large-scale, self-supervised pre-training to obtain general-purpose surgical scene representations, subsequently fine-tuned across multiple downstream tasks. Several other recent studies instead build on vision-language models to express scene understanding directly as natural-language output: in construction, VQA-based frameworks couple visual perception with regulation-aware language reasoning [67,68]; in assistive environments, vision-language models are similarly used to generate walking guidance [65] and, combined with reinforcement learning, to tune navigation instructions to how a user is expected to interpret and act on them [73]. A further step in this direction is taken by Low et al. [72]'s SurgRAW, which coordinates multiple specialized reasoning agents through a chain-of-thought workflow for surgical video analysis. Self-supervised pre-training, language-based reasoning, and agentic coordination remain represented by only one or a few studies each within the reviewed corpus, but together they point to architectures that move beyond task-specific pipelines toward more general-purpose and interpretable forms of scene understanding.

A common issue across all three domains concerns the high cost of annotations, which limits the volume and diversity of available training scenarios. As a result, training datasets cannot realistically provide representative coverage of the large variety of objects, interactions, and situations encountered in real-world environments. This lack of representativeness is a key motivation for open-vocabulary scene understanding, which seeks to recognize scene elements not explicitly represented during training. In applications for blind users, the annotation burden is particularly evident in captioning systems, where descriptions must be rich and detailed to provide effective support. In construction environments, annotation is challenging because of the dynamic and chaotic nature of the environment, with numerous simultaneous actions and complex relationships among the entities involved. In surgery, inter-patient and inter-institutional variability make it difficult to gather annotated data that cover the range of situations encountered in practice models. Several technical strategies attempted to address the problem. In surgical environments, Pang et al. [45] avoided traditional detectors by using Grad-CAM and Grad-CAM++ instead, reducing both the annotation requirements and the computational complexity of the model. More directly addressing the limited coverage of training data, open-vocabulary scene understanding

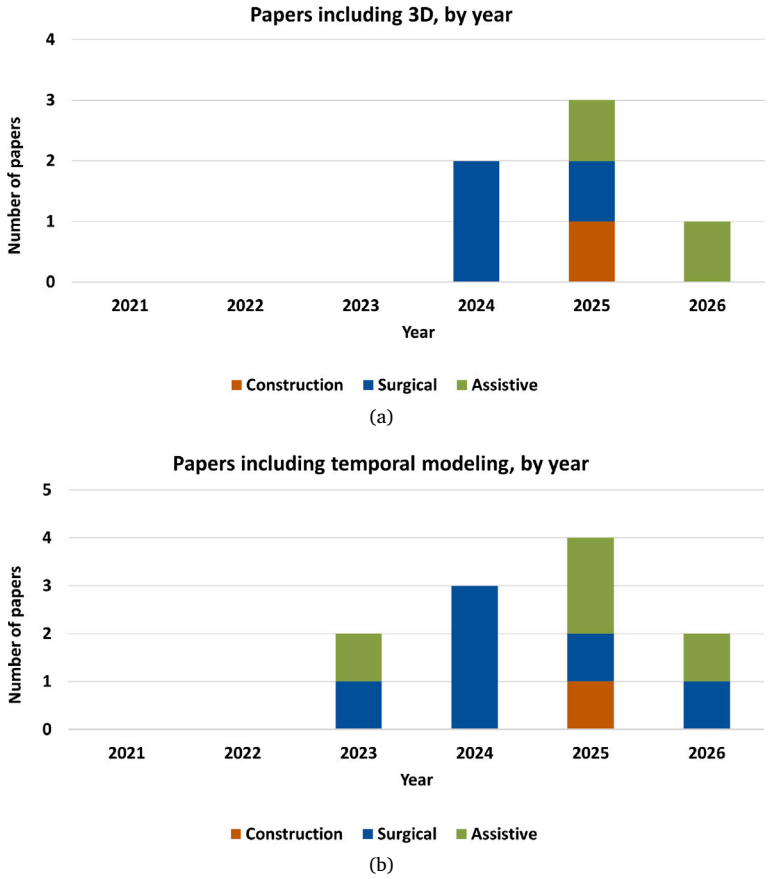


Fig. 15. Temporal and spatial modeling coverage across application domains. (a) Number of papers incorporating 3D reconstruction or pose/scene estimation, by year and domain (Construction, Surgical, Assistive). (b) Number of papers incorporating temporal modeling, by year and domain. Bar segments are stacked and color-coded by domain.

aims to recognize objects, interactions, or scene elements beyond a fixed training taxonomy. Within the reviewed corpus, this direction is illustrated by Pan et al. [42]'s use of Zero-Shot Detection and Zero-Shot Human–Object Interaction techniques in construction environments, combined with a Knowledge Graph that is automatically updated by the framework whenever the system recognizes new situations. Similarly, Liu et al. [46]'s OpenSU system explicitly targets open scene understanding to address ambiguous object localization in assistive environments. A complementary but conceptually distinct strategy is offered by Yang et al. [69], who address annotation scarcity via label-free self-supervised pre-training on a large and procedurally diverse surgical corpus, reserving annotation for fine-tuning alone; unlike the open-vocabulary approaches above, this generalization occurs within a fixed downstream task taxonomy rather than through an open vocabulary. Performance improved with pre-training scale and diversity even on procedures with 0% overlap, suggesting cost can be mitigated at the representation-learning stage, rather than via detector elimination or zero-shot approaches. Overall, these approaches address different consequences of annotation costs, but none fully resolves the fundamental limitation that manually annotated training data cannot provide representative coverage of the diversity of real-world situations. This limitation provides a strong motivation for scene-understanding approaches capable of generalizing beyond the objects, interactions, and situations explicitly represented in their annotated training data.

Another common limitation concerns the high computational complexity of the models. In all the analyzed domains, it is necessary to maintain short inference times without sacrificing accuracy. This is essential on construction sites, where a delay in detecting a hazardous situation can have serious consequences; in assistive technologies, where support must be provided quickly to ensure user safety; and in surgery, where any error or latency could result in risks for the patient. Concrete solutions to balance accuracy and speed have primarily emerged in the surgical field. In addition to the aforementioned work by Pang et al. [45], Pei et al. [59] have proposed a single-stage architecture for the holistic understanding of operating rooms, greatly simplifying the pipeline compared to previous multi-stage approaches [57] and improving computational efficiency. Efficiency-oriented techniques increasingly explored in the broader deep learning literature, such as knowledge distillation for model compression or visual token reduction for transformer-based architectures, both particularly relevant for real-time deployment on resource-constrained devices, remain largely underexplored within the reviewed corpus; even where knowledge distillation is employed [69], it is used with the main goal of improving feature quality during self-supervised pre-training rather than to reduce model size or inference cost.

Finally, another cross-cutting weakness that emerged from the analysis concerns the poor consideration of end-user perspectives in the design and evaluation of technologies. Despite significant technical advances, these solutions are sometimes developed without adequate involvement of users or professionals working in application contexts, resulting in a lack of validation in real-world scenarios. This limitation appears in several ways. In the surgical field, although recent work proposes promising frameworks for comprehensive operating room modeling, their evaluation is still limited to simulated or experimental settings, with no confirmed implementation in real-world clinical practice. Validation by surgeons in real-world clinical settings is important in high-risk and time-critical settings, where effective collaboration and shared decision-making between human professionals and digital

Table 3
Overview of the current state of research across the three analyzed domains, highlighting data representations and modeling capabilities.

Aspect	Surgical environments	Construction environments	Assistive environments
3D integration	Actively explored to improve spatial reasoning, but only in holistic OR studies	Very recent transition from 2D to 3D	Limited to few studies
Temporal modeling	Increasingly investigated to obtain more accurate time-aware results	Mostly neglected, few recent attempts	Emerging but limited
Annotation cost and data scarcity	High inter-patient and inter-institution variability limits generalization	Dynamic and cluttered environments increase annotation costs	Detailed annotations required for effective feedback and captioning
Computational complexity	Errors or latency may endanger patients. Efficiency-oriented architectures	Delays in detecting hazardous situations may have serious consequences	Rapid support required for user safety
User-centered validation	Mainly simulated and experimental	Scarce co-design and real-world longitudinal testing	User studies with limited sample diversity, one study relies entirely on simulated feedback

systems are essential [57]. In the case of assistive systems for people with special needs, it is important that individual differences, such as the specific type of visual impairment, level of digital experience, or sensory preferences, are considered. Although many of the works analyzed were evaluated with real users, such trials often present limited sample variability. Consequently, the same studies recognize the need for future evaluations involving a larger and more heterogeneous number of users [55]. Lastly, the lack of studies based on co-design or longitudinal testing in real-world contexts limits the ability to adapt technologies to out-of-distribution scenes within actual operating conditions [61]. Zhao et al. [73] replaced real user feedback entirely with an LLM that simulates how a visually impaired user would respond, explicitly noting that validation with real users remains necessary for future work. Bridging these gaps requires deeper integration of user feedback, field testing, and a greater focus on customization.

To provide a clear overview of the current state of research and the challenges discussed in the domains of construction environments, surgical environments, and assistive environments for people with special needs, Table 3 summarizes key aspects such as the use of 3D and temporal information, data availability, computational requirements, and user-centered validation.

Although the three domains analyzed show a gradual evolution towards more advanced forms of scene understanding, several future challenges remain that limit their applicability in real-world contexts. The high cost of data annotation is a common challenge across all the domains considered. Future research should therefore prioritize strategies that reduce reliance on manual annotation, including zero-shot and few-shot approaches, or the integration of a priori knowledge through structured models or knowledge graphs. As highlighted above, open-vocabulary and self-supervised strategies remain confined to single studies within individual domains; a concrete step forward would be their systematic adoption in construction and assistive environments, currently lagging behind surgery in this respect. Another challenge is the limited use of three-dimensional and temporal information. In particular, the integration of 3D data has shown significant benefits, especially in complex and highly dynamic scenarios, such as construction sites and the holistic analysis of operating rooms: being these contexts characterized by the simultaneous presence of numerous entities, human operators, objects, and multiple interactions, three-dimensional information allows for more accurate modeling of the geometric properties of the scene and less ambiguous predictions of spatial relationships. In these cases, extending the use of 3D information represents a promising direction for future research. The exploitation of temporal information has also shown significant advantages. While in the surgical and assistive domains the adoption of temporal models is growing, in the construction environment applications, the temporal dimension has been explored to a more limited extent so far, despite available results indicating that the integration of motion sequences can significantly improve scenario understanding and the recognition of dangerous situations. For these reasons, future research should explore more temporal information in these application fields. Computational complexity represents an additional future challenge in scenarios where real-time feedback is required. The literature has already started focusing on simplified architectures, single-stage pipelines, and efficiency-oriented solutions, especially in the surgical domain. Similar efforts should also be extended to construction environments and to assistive technologies. Furthermore, techniques such as knowledge distillation for model compression or visual token reduction, both particularly relevant for real-time deployment on resource-constrained devices, remain unexplored for this purpose within the reviewed corpus; their adoption represents a concrete and currently untapped direction for improving deployability across all three domains. Beyond these architectural considerations, the recent shift toward vision-language models, foundation model pre-training, and agentic reasoning, discussed above, suggests a promising complementary direction: rather than developing task-specific pipelines from scratch in each domain, future work could explore adapting these general-purpose, language-interpretable backbones across domains, particularly given that they are currently concentrated in only a subset of the studies reviewed. Finally, closer collaboration among artificial intelligence researchers, domain experts, and end users would help ensure that the developed systems are aligned with real operational needs, promote trust in their use, and effectively support decision-making processes. This is particularly pressing given that, within the reviewed corpus, user validation is in some cases substituted with simulated feedback rather than input from real users, underscoring the need for longitudinal, real-world testing with diverse user samples across all three domains.

5. Conclusions

High-level scene understanding refers to computational methods that go beyond object recognition or image segmentation, aiming instead to model interactions, contextual dependencies, and meaningful relationships within a visual environment. These techniques allow systems to interpret scenes more closely to human reasoning. As real-world environments become more complex, dynamic, and information-rich, such contextual reasoning is increasingly essential for any AI system that should support human activities safely and effectively. This review aims to identify the main contexts in which high-level scene understanding supports people in activities, providing context-aware feedback to improve safety, efficiency, and situational awareness through the analysis and interpretation of human–objects and human–environment interactions. Applying a structured selection process and a set of inclusion criteria, three application areas emerged as the most

representative in recent literature: construction environments, surgical environments, and assistive environments for users with special needs. Across these fields, the review mapped domain-specific needs and domain-independent challenges, examined the forms of support provided, and highlighted the constraints that currently limit real-world adoption. Future progress will require scalable data annotation to increase model generalization and diversity, the integration of 3D and temporal information to accurately capture spatial and dynamic context, computationally efficient models to ensure real-time applicability in safety-critical environments, and a user-centric interdisciplinary approach to align system design with real needs and improve trust and usability through the active involvement of industry experts and end users.

CRediT authorship contribution statement

Rossella Ruggieri: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Giorgia Marullo:** Writing – review & editing, Visualization, Supervision, Methodology, Investigation, Conceptualization. **Luca Ulrich:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Yves Grandvalet:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Sandro Moos:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition. **Enrico Vezzetti:** Writing – review & editing, Resources, Project administration, Funding acquisition.

Funding

This work was funded by the AI-SURG — human centric AI – behaviour modeling for smart SURGery Ecosystem project, developed within the framework of the "BANDO AGGREGAZIONI R&S – SALUTE" call. The project was co-financed by the European Regional Development Fund (ERDF) under the Valle d'Aosta Regional Programme 2021–2027.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

References

[1] W. Zeng, Toward human-centric deep video understanding, APSIPA Trans. Signal Inf. Process. 9 (2020) e1, http://dx.doi.org/10.1017/atsip.2019.26.

[2] E. Coronado, T. Kiyokawa, G.A.G. Ricardez, I.G. Ramirez-Alpizar, G. Venture, N. Yamanobe, Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0, J. Manuf. Syst. 63 (2022) 392–410, http://dx.doi.org/10.1016/j.jmsy.2022.04.007.

[3] M. Cooley, Human-centered design, Inf. Des. (2000) 59–81, http://dx.doi.org/10.13140/RG.2.2.18296.24329.

[4] G. Marullo, C. Innocente, L. Ulrich, A. Lo Faro, A. Porcelli, R. Ruggieri, B. Vecchio, E. Vezzetti, Home-based mirror therapy in phantom limb pain treatment: the augmented humans framework, Multimedia Tools Appl. (2025) 1–33, http://dx.doi.org/10.1007/s11042-025-20628-1.

[5] I. Göttgens, S. Oertelt-Prigione, The application of human-centered design approaches in health research and innovation: A narrative review of current practices, JMIR Mhealth Uhealth 9 (2021) e28102, http://dx.doi.org/10.2196/28102.

[6] L. Ulrich, G. Carmassi, P. Garelli, G. Lo Presti, G. Ramondetti, G. Marullo, C. Innocente, E. Vezzetti, Signify: Leveraging machine learning and gesture recognition for sign language teaching through a serious game, Futur. Internet 16 (2024) http://dx.doi.org/10.3390/fi16120447, URL https://www.mdpi.com/1999-5903/16/12/447.

[7] R. Bevilacqua, T. Bailoni, E. Maranesi, G. Amabili, F. Barbarossa, M. Ponzano, M. Virgolesi, T. Rea, M. Illario, E.M. Piras, M. Lenge, E. Barbi, G. Sakellariou, Framing the human-centered artificial intelligence concepts and methods: Scoping review, JMIR Hum. Factors 12 (2025) e67350, http://dx.doi.org/10.2196/67350.

[8] G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, Designing and evaluating human-centred AI systems: Best-practices from a multidisciplinary view, in: Workshops of the Annual ACM Conference on Intelligent User Interfaces, in: CEUR Workshop Proceedings, vol. 3957.

[9] V. Dhanya, A. Subeesh, N. Kushwaha, D.K. Vishwakarma, T.N. Kumar, G. Ritika, A. Singh, Deep learning based computer vision approaches for smart agricultural applications, Artif. Intell. Agric. 6 (2022) 211–229, http://dx.doi.org/10.1016/j.aiia.2022.09.007.

[10] J. Ni, Y. Chen, G. Tang, J. Shi, W. Cao, P. Shi, Deep learning-based scene understanding for autonomous robots: A survey, Intell. Robot. 3 (2023) 374–401, http://dx.doi.org/10.20517/ir.2023.22.

[11] H. Li, G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, X. Zhao, S.A.A. Shah, M. Bennamoun, Scene graph generation: A comprehensive survey, Neurocomputing 566 (2024) 127052, http://dx.doi.org/10.1016/j.neucom.2023.127052.

[12] H. Li, H. Deng, Y. Deng, Towards worker-centric construction scene understanding: Status quo and future directions, Autom. Constr. 171 (2025) 106005, http://dx.doi.org/10.1016/j.autcon.2025.106005.

[13] X. Wang, Z. Zhu, Context understanding in computer vision: A survey, Comput. Vis. Image Underst. 229 (2023) 103646, http://dx.doi.org/10.48550/arXiv.2302.05011.

[14] H. Deng, K. Fu, B. Yu, H. Li, R. Duan, Y. Deng, J. r. Lin, Enabling high-level worker-centric semantic understanding of onsite images using visual language models with attention mechanism and beam search strategy, Buildings 15 (2025) 959, http://dx.doi.org/10.3390/buildings15060959.

[15] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, A. Hauptmann, A comprehensive survey of scene graphs: Generation and application, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2021) 1–26, http://dx.doi.org/10.48550/arXiv.2104.01111.

[16] K. Guo, Y. Huang, S. Sun, X. Song, M. Feng, Z. Liu, H. Song, T. Wang, J. Li, N. Akhtar, et al., Beyond human perception: Understanding multi-object world from monocular view, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2025, pp. 3751–3760.

[17] K. Guo, Y. Huang, T. Jia, X. Song, S. Sun, H. Wei, X.-F. Han, S. Huang, N. Strisciuglio, S. Li, Visual grounding in 2d and 3d: A unified perspective and survey, Inf. Fusion 126 (2026a) 103625.

[18] K. Guo, Y. Huang, H. Wei, S. Sun, X. Song, M. Feng, T. Wang, S. Yu, H. Song, From perception to cognition: Unifying multi-object 3d visual grounding and dense captioning in monocular images, Expert Syst. Appl. 325 (2026b) 132657, http://dx.doi.org/10.1016/j.eswa.2026.132657.

[19] P. Dolatyabi, J. Regan, M. Khodayar, Deep learning for traffic scene understanding: A review, IEEE Access (2025) http://dx.doi.org/10.1109/ACCESS.2025.3529289.

[20] H. Wei, R. Wang, H. Hu, S. Sun, X. Song, M. Feng, Monocular multi-object 3d visual language tracking, IEEE Trans. Image Process. (2026).

[21] Y. Park, H.I. Son, Visual scene understanding for efficient cooperative control of agricultural dual-arm robots, in: International Conference on Control, Automation and Systems, ICCAS, 2024, pp. 1471–1472, http://dx.doi.org/10.23919/ICCAS63016.2024.10773238.

[22] Z. Bao, W. Li, J. Chen, H. Chen, V. John, C. Xiao, Y. Chen, Predicting and visualizing citrus color transformation using a deep mask-guided generative network, Plant Phenomics 5 (2023) 0057, http://dx.doi.org/10.34133/plantphenomics.0057.

[23] S. Wu, F. Teng, H. Shi, Q. Jiang, K. Luo, K. Wang, K. Yang, Quadreamer: Controllable panoramic video generation for quadruped robots, 2026, URL https://arxiv.org/abs/2605.23478. arXiv:2605.23478.

[24] A. Murali, D. Alapatt, P. Mascagni, A. Vardazaryan, A. Garcia, N. Okamoto, D. Mutter, N. Padoy, Latent graph representations for critical view of safety assessment, IEEE Trans. Med. Imaging 43 (2023) 1247–1258, http://dx.doi.org/10.48550/arXiv.2212.04155.

[25] H. Son, A. Vora, G. Pandey, J. Weiland, Infrastructure enabled guided navigation for visually impaired, IEEE Trans. Intell. Transp. Syst. (2025) http://dx.doi.org/10.1109/TITS.2025.3538317.

[26] L. Brunke, Y. Zhang, R. Römer, J. Naimer, N. Staykov, S. Zhou, A.P. Schoellig, Semantically safe robot manipulation: From semantic scene understanding to motion safeguards, IEEE Robot. Autom. Lett. (2025) http://dx.doi.org/10.48550/arXiv.2410.15185.

[27] J. Fan, P. Zheng, S. Li, Vision-based holistic scene understanding towards proactive human–robot collaboration, Robot. Comput.-Integr. Manuf. 75 (2022) 102304, http://dx.doi.org/10.1016/j.rcim.2021.102304.

[28] M.R. Ibrahim, J. Haworth, T. Cheng, Understanding cities with machine eyes: A review of deep computer vision in urban analytics, Cities 96 (2020) 102481, http://dx.doi.org/10.1016/j.cities.2019.102481.

[29] P. Mora, C. Garcia, E. Ivorra, M. Ortega, M.L. Alcañiz, Virtual experience toolkit: An end-to-end automated 3D scene virtualization framework implementing computer vision techniques, Sensors 24 (2024) 3837, http://dx.doi.org/10.3390/s24123837.

[30] X. Yan, Y. Zhu, Z. Wang, B. Xu, L. He, R. Xia, Intelligent flood scene understanding using computer vision-based multi-object tracking, Water 17 (2025) 2111, http://dx.doi.org/10.3390/w17142111.

[31] Y. Ye, S. Luo, M. Jing, Y. Ding, K. He, L. Zuo, Graph neural network-based structured scene graph generation for efficient wildfire detection, in: International Conference on Intelligent Computing, 2024, pp. 356–367, http://dx.doi.org/10.1007/978-981-97-5588-2_30.

[32] Y. Wu, Z. Xie, Z. Ma, M. Chen, F. Zhang, Z. Shi, W. Li, S. Zhai, An object-oriented semi-supervised land-use/land-cover change detection method based on siamese autoencoder graph attention network, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17 (2024) 6478–6500, http://dx.doi.org/10.1109/JSTARS.2024.3373039.

[33] R. Ankareddy, R. Delhibabu, Dense segmentation techniques using deep learning for urban scene parsing: a review, IEEE Access (2025) http://dx.doi.org/10.1109/ACCESS.2025.3543944.

[34] M. Awan, T.K. Whangbo, J. Shin, Deep learning methods for autonomous driving scene understanding tasks: A review, Expert Syst. Appl. (2025) 128098, http://dx.doi.org/10.1016/j.eswa.2025.128098.

[35] U. Khan, U. Nawaz, A. Qayyum, S. Ashraf, Y. Xie, M.H. Khan, M. Bilal, J. Qadir, Surgical scene understanding in the era of foundation AI models: A comprehensive review, 2025, http://dx.doi.org/10.48550/arXiv.2502.14886, arXiv preprint arXiv:2502.14886.

[36] Y. Liu, S. Wang, J. Liu, Q. Zhang, A review of scene understanding in smart manufacturing environments, in: 2024 29th International Conference on Automation and Computing, ICAC, IEEE, 2024, pp. 1–6.

[37] J. Baas, M. Schotten, A. Plume, G. Côté, R. Karimi, Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies, Quant. Sci. Stud. 1 (2020) 377–386, http://dx.doi.org/10.1162/qss_a_00019.

[38] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) http://dx.doi.org/10.1136/bmj.n71.

[39] S. Chen, K. Demachi, Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph, Autom. Constr. 125 (2021) 103619, http://dx.doi.org/10.1016/j.autcon.2021.103619.

[40] M. Xu, M. Islam, C.M. Lim, H. Ren, Learning domain adaptation with model calibration for surgical report generation in robotic surgery, in: International Conference on Robotics and Automation, (ICRA), 2021, pp. 12350–12356, http://dx.doi.org/10.1109/ICRA48506.2021.9561569.

[41] C.I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, N. Padoy, Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos, Med. Image Anal. 78 (2022) 102433, http://dx.doi.org/10.1016/j.media.2022.102433.

[42] Z. Pan, C. Su, Y. Deng, J. Cheng, Image2Triplets: A computer vision-based explicit relationship extraction framework for updating construction activity knowledge graphs, Comput. Ind. 137 (2022) 103610, http://dx.doi.org/10.1016/j.compind.2022.103610.

[43] S. Chen, K. Demachi, F. Dong, Graph-based linguistic and visual information integration for on-site occupational hazards identification, Autom. Constr. 137 (2022) 104191, http://dx.doi.org/10.1016/j.autcon.2022.104191.

[44] L. Zhang, J. Wang, Y. Wang, H. Sun, X. Zhao, Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge, Autom. Constr. 142 (2022) 104535, http://dx.doi.org/10.1016/j.autcon.2022.104535.

[45] W. Pang, M. Islam, S. Mitheran, L. Seenivasan, M. Xu, H. Ren, Rethinking feature extraction: Gradient-based localized feature extraction for end-to-end surgical downstream tasks, IEEE Robot. Autom. Lett. 7 (2022) 12623–12630, http://dx.doi.org/10.1109/LRA.2022.3221310.

[46] R. Liu, J. Zhang, K. Peng, J. Zheng, K. Cao, Y. Chen, K. Yang, R. Stiefelhagen, Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments, in: International Conference on Computer Vision (ICCV), 2023, pp. 1857–1867, http://dx.doi.org/10.48550/arXiv.2307.07757.

[47] H. Wang, Y. Jin, L. Zhu, Dynamic interactive relation capturing via scene graph learning for robotic surgical report generation, in: International Conference on Robotics and Automation, (ICRA), 2023, pp. 2702–2709, http://dx.doi.org/10.48550/arXiv.2306.02651.

[48] X. Liu, X. Jing, Q. Zhu, W. Du, X. Wang, Automatic construction hazard identification integrating on-site scene graphs with information extraction in outfield test, Buildings 13 (2023) 377, http://dx.doi.org/10.3390/buildings13020377.

[49] E. Dessalene, C. Devaraj, M. Maynord, C. Fermüller, Y. Aloimonos, Forecasting action through contact representations from first person video, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2021) 6703–6714, http://dx.doi.org/10.1109/TPAMI.2021.3055233.

[50] B. Zhong, L. Shen, X. Pan, L. Lei, Visual attention framework for identifying semantic information from construction monitoring video, Saf. Sci. 163 (2023) 106122, http://dx.doi.org/10.1016/j.ssci.2023.106122.

[51] M. Xu, M. Islam, B. Glocker, H. Ren, Confidence-aware paced-curriculum learning by label smoothing for surgical scene understanding, IEEE Trans. Autom. Sci. Eng. 21 (2023) 3170–3181, http://dx.doi.org/10.48550/arXiv.2212.11511.

[52] C. Lin, Z. Zhu, Y. Zhao, Y. Zhang, K. He, Y. Zhao, SGT++: Improved scene graph-guided transformer for surgical report generation, IEEE Trans. Med. Imaging 43 (2024) 1337–1346, http://dx.doi.org/10.1109/TMI.2023.3335909.

[53] J. Zhang, Y. Wang, S. Zhou, H. Zhao, C. Wan, X. Cai, H. Ding, Graph-based spatial reasoning for tracking landmarks in dynamic laparoscopic environments, IEEE Robot. Autom. Lett. 9 (2024) 8459–8466, http://dx.doi.org/10.1109/LRA.2024.3445654.

[54] W. Lin, Y. Hu, H. Fu, M. Yang, C.-B. Chng, R. Kawasaki, C. Chui, J. Liu, Instrument-tissue interaction detection framework for surgical video understanding, IEEE Trans. Med. Imaging 43 (2024) 2803–2813, http://dx.doi.org/10.1109/TMI.2024.3381209.

[55] Z. Cao, J. Xia, M. Zhou, De-confounding feature fusion transformer network for image captioning in assistive navigation applications for the visually impaired, IEEE Trans. Instrum. Meas. (2025) http://dx.doi.org/10.1109/TIM.2025.3553260.

[56] P. Su, D. Chen, Adopting graph neural networks to analyze human–object interactions for inferring activities of daily living, Sensors 24 (2024) 2567, http://dx.doi.org/10.3390/s24082567.

[57] E. Özsoy, T. Czempiel, E.P. Örnek, U. Eck, F. Tombari, N. Navab, Holistic or domain modeling: a semantic scene graph approach, Int. J. Comput. Assist. Radiol. Surg. 19 (2024) 791–799, http://dx.doi.org/10.1007/s11548-023-03022-w.

[58] K. Yuan, M. Kattel, J.L. Lavanchy, N. Navab, V. Srivastav, N. Padoy, Advancing surgical VQA with scene graph knowledge, Int. J. Comput. Assist. Radiol. Surg. 19 (2024) 1409–1417, http://dx.doi.org/10.48550/arXiv.2312.10251.

[59] J. Pei, D. Guo, J. Zhang, M. Lin, Y. Jin, P.-A. Heng, S^2Former-OR: Single-stage bi-modal transformer for scene graph generation in OR, IEEE Trans. Med. Imaging 44 (2024) 361–372, http://dx.doi.org/10.1109/TMI.2024.3444279.

[60] J.-H. Kim, S.-W. Park, J.-H. Huh, S.-H. Jung, C.-B. Sim, Human scene understanding mechanism based image captioning for blind assistance, IEEE Access (2025) http://dx.doi.org/10.1109/ACCESS.2025.3564991.

[61] L. Xie, S. Misra, N. Suresh, J. Soza-Soto, T. Furuhata, K. Shimada, Learning 3D human-object interaction graphs from transferable context knowledge for construction monitoring, Comput. Ind. 164 (2025) 104171, http://dx.doi.org/10.1016/j.compind.2024.104171.

[62] W. Li, F. Ma, Z. Zuo, R. Jia, B. Wang, A.M. Alharbi, SafetyGPT: An autonomous agent of electrical safety risks for monitoring workers' unsafe behaviors, Int. J. Electr. Power Energy Syst. 168 (2025) 110672, http://dx.doi.org/10.1016/j.ijepes.2025.110672.

[63] C.-F. Chan, P.K.-Y. Wong, X. Guo, J.C. Cheng, J.P.-C. Chan, P.-H. Leung, X. Tao, Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring, Autom. Constr. 177 (2025) 106305, http://dx.doi.org/10.1016/j.autcon.2025.106305.

[64] E. Özsoy, C. Pellegrini, T. Czempiel, F. Tristram, K. Yuan, D. Bani-Harouni, U. Eck, B. Busam, M. Keicher, N. Navab, MM-OR: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 19378–19389, http://dx.doi.org/10.48550/arXiv.2503.02579.

[65] Z. Yuan, T. Zhang, Y. Deng, J. Zhang, Y. Zhu, Z. Jia, J. Zhou, J. Zhang, Walkvlm: Aid visually impaired people walking by vision language model, 2025, arXiv preprint arXiv:2412.20903.

[66] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang, B. Ouyang, Z. Lin, M. Cominelli, Z. Cai, B. Li, Y. Zhang, P. Zhang, F. Hong, J. Widmer, F. Gringoli, L. Yang, Z. Liu, Egolife: Towards egocentric life assistant, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2025, pp. 28885–28900.

[67] T. Kim, H. Park, M. Park, T. Oh, Explainable vqa-based ladder safety monitoring for fall-risk prevention on construction sites, J. Saf. Res. 97 (2026) 646–661, http://dx.doi.org/10.1016/j.jsr.2026.05.008.

[68] S.M.H. Abidi, S.M. Raza, S.Y. Shin, Safevision: Vision–language reasoning for context-aware safety monitoring, Neurocomputing 669 (2026) 132479, http://dx.doi.org/10.1016/j.neucom.2025.132479.

[69] S. Yang, F. Zhou, L. Mayer, F. Huang, Y. Chen, Y. Wang, S. He, Y. Nie, X. Wang, Ö. Sümer, Y. Jin, H. Sun, S. Xu, A.Q. Liu, Z. Li, J. Qin, J.Y. Teoh, L. Maier-Hein, H. Chen, Large-scale self-supervised video foundation model for intelligent surgery, Npj Digit. Med. 9 (2026) 220, http://dx.doi.org/10.1038/s41746-026-02403-0.

[70] J. Huang, S. Hao, B.-C. Hu, H. Wang, G. Wang, Understanding dynamic scenes in egocentric 4d point clouds, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2026, pp. 5031–5039.

[71] M.S.B. Sajid, Y. Alkhrijah, D. Hamid, M.T. Akram, S.M. Usman, S. Khalid, M.A. Alawad, Transflow-hoi: Transformer-driven multimodal fusion for high-performance human–object interaction recognition, IEEE Access 14 (2026) 12717–12732, http://dx.doi.org/10.1109/ACCESS.2026.3655585.

[72] C.H. Low, Z. Wang, T. Zhang, Z. Zhu, Z. Zeng, E.B. Mazomenos, Y. Jin, Surgraw: Multi-agent workflow with chain of thought reasoning for robotic surgical video analysis, IEEE Robot. Autom. Lett. (2026) http://dx.doi.org/10.1109/LRA.2026.3665443.

[73] Y. Zhao, S. Wang, J. Li, Laf-grpo: In-situ navigation instruction generation for the visually impaired via grpo with llm-as-follower reward, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2026, pp. 34994–35002.

[74] B. Kuriakose, R. Shrestha, F.E. Sandnes, Deepnavi: A deep learning based smartphone navigation assistant for people with visual impairments, Expert Syst. Appl. 212 (2023) 118720, http://dx.doi.org/10.1016/j.eswa.2022.118720.

[75] H. Liu, R. Liu, K. Yang, J. Zhang, K. Peng, R. Stiefelhagen, Hida: Towards holistic indoor understanding for the visually impaired via semantic instance segmentation with a wearable solid-state lidar sensor, in: International Conference on Computer Vision (ICCV), 2021, pp. 1780–1790, http://dx.doi.org/10.1109/ICCVW54120.2021.00204.