Article | Received 19 May 2026; Revised 19 August 2026; Accepted 6 September 2026; Published 22 September 2026
https://doi.org/10.55092/sc20260017

YOLO-SafeAttr: improving the ability to identify unsafe conditions of objects in construction scenes by using visual attribute representation

Yichuan Deng[1], Zile Deng[1], Hengyun Zhang[2], Hui Deng[1,*] and Vincent Jielong Gan[3]

[1] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
[2] Institute of Intelligence and Digitalization, CCCC Fourth Harbor Engineering Institute Co., Ltd., Guangzhou, China
[3] College of Design and Engineering NUS, National University of Singapore, Singapore

* Correspondence author; E-mail: hdeng@scut.edu.cn.

Highlights:

- A visual-attribute framework for detecting unsafe object states.
- Visual unsafe attributes bridge perception features and safety semantics.
- VALD dataset with 11,980 images and four construction risk attributes.

Abstract: Unsafe conditions of objects are major contributors to construction accidents such as falls, strikes, and collapses. Although traditional inspections and existing vision methods can detect objects, they often fail to infer their potential risks. To overcome this limitation, this paper proposes a deep identification method integrating visual attribute representations. We introduce the high-level concept of visual unsafe attributes to describe hazards arising from an object's shape, state, and environmental context. Based on accident text analysis, a visual attribute system covering four risk types—fall, strike, collapse, and rollover—is established, and the VALD dataset is built by extending SODA dataset. An integrated detection framework based on YOLOv11 is then developed to achieve end-to-end joint inference of object locations and unsafe attributes. The model is trained and evaluated on VALD. Experiments show mAP@50 of 84.8%, recall of 91%, and F1-score of 0.82, with improvements of 21.1%, 33.6%, and 0.175 over models without attention mechanisms. These results demonstrate that visual attribute representation and attention significantly enhance risk feature extraction. The resulting dynamic risk assessment system quantifies object-strike scenarios spatiotemporally, providing a real-time and interpretable safety monitoring solution for construction sites.

Keyword: construction safety management; visual attribute; computer vision; deep learning

1. Introduction

1.1. Visual attributes for safety management

Construction safety accidents are one of the important factors restricting the development of construction projects due to their high frequency, great harm and difficulty in control. At the same time, the current level of building informatization and intelligence is far lower than the level of building development, which makes the safety management of the construction industry still largely manual management and the overall level of intelligence low, making it difficult for managers to achieve refined management. Therefore, the traditional construction safety management model needs to be changed to adapt to the needs of industry development.

The causes of construction site accidents can be summarized into two categories: unsafe human behavior and unsafe conditions of objects [1]. Among them, the determination of unsafe conditions of objects occupies an important position in the pre-construction safety management system. However, the determination of unsafe conditions of objects in current construction safety management research mainly relies on manual inspection methods. Manual inspection methods have disadvantages such as large workload, strong subjectivity and poor real-time performance. This has obvious limitations in practical applications. Against this background, the application of intelligent technology means has shown a strong development trend, especially visual technology, which has attracted much attention due to its low cost and strong real-time performance. However, the existing visual means in construction safety management only focus on information related to the object itself, without combining deeper semantics. The essence of this problem lies in the semantic gap between the underlying information of construction site images and the high-level semantics of construction safety management.

Visual attributes are abstractions and generalizations of low-level features, providing a generalized semantic description that can be understood by both computers and humans. They serve as a link between high-level semantics and low-level information. Therefore, visual attributes hold the promise of bridging the semantic gap between the low-level information of construction site images and the high-level semantics of construction safety management. However, past research in the field of construction safety management has been limited by the lack of defined “attributes” representing unsafe states of objects, as well as the absence of corresponding datasets and management methods. Furthermore, technological limitations have resulted in the lack of visual attribute-driven risk detection frameworks, making it difficult to effectively detect high-level visual attributes.

1.2. Problem analysis

To meet the needs of intelligent construction safety management in the building industry, the determination of unsafe conditions of objects based on visual attributes needs to overcome the following obstacles, as outlined above:

(1) The determination of unsafe conditions of objects at construction sites relies on manual inspection, which lacks real-time perception capabilities, is highly subjective, and has a lagging dynamic response.

(2) Current visual technologies for identifying unsafe conditions of objects generally focus on extracting superficial features while neglecting the association and mining of deep semantic information.

Furthermore, the key “attributes” characterizing the unsafe condition of an object lack clear definitions and a unified categorization.

(3) Currently, there is a lack of comprehensive labeled datasets and systematic management methods for visual attribute detection of unsafe object states.

The first obstacle is the significant limitation of traditional construction safety management methods in determining the unsafe condition of objects. For example, on the one hand, the limited spatial and temporal coverage of manual inspections makes it difficult to achieve real-time perception of the state of objects on site, resulting in a significant time lag in identifying safety hazards. On the other hand, the inspection process relies heavily on subjective judgment, and differences in the professional experience of different inspectors can easily lead to inconsistencies in judgment standards, thus causing delays in dynamic risk response. This contradiction between the traditional judgment model and the multi-dimensional and dynamic safety management needs of construction sites has become a key bottleneck restricting the intelligent development of construction safety management.

The second obstacle is related to the creation of visual attribute labels and detection frameworks for object visual unease: The goal of computer vision is to achieve a real understanding of the scene, which means that the objects, attributes and relationships in the scene can be understood by the computer [2]. Although computer vision is an active research field, there is still a big gap in its scene understanding. The semantic gap between low-level visual features and high-level semantics understood by humans is the main reason. In addition, computers still cannot fully understand architectural scenes, and it has not yet been realized to solve all problems with a single model. Therefore, to achieve high-level semantic understanding of architectural scenes through computer vision, it is necessary to go through the three stages of semantic understanding. Visual attributes are considered to be the key to solving the semantic gap because of their three major advantages: they can be understood by humans, can be calculated by machines, and can be shared across classes [3]. Nowadays, research on visual attributes has progressed from low-level visual attributes to high-level visual attributes, such as emotions, artistic styles, and complex scene semantics. However, it is still difficult to advance from high-level visual attributes to visual scene understanding, mainly due to two limitations: First, the algorithm model relies on manual design or local feature extraction methods, which makes it difficult to understand the function of objects or scene logic. Second, there is a lack of large-scale labeled datasets as a training basis. Furthermore, current assessments of unsafe conditions of objects are still based on manually constructed rules, which suffer from strong subjectivity and poor scalability. Therefore, resolving semantic gaps in construction safety management requires rule construction that incorporates relevant industry knowledge.

The final obstacle relates to the industry’s research on visual attribute detection. Existing construction safety management research focuses more on unsafe human behaviors and less on unsafe conditions of objects, often remaining at the theoretical stage and lacking practical applications. For example, scholars have conducted detailed studies on production safety based on relevant theories, analyzing factors such as people, objects, and relationships. However, they lack early warning and detection methods, focusing only on the factors that cause accidents. Therefore, this leads to a data shortage for visual attribute detection of unsafe conditions of objects.

1.3. Research questions

Meanwhile, the research trend of monitoring unsafe conditions of objects at construction sites and existing buildings is that researchers from different fields use visual attributes to establish the connection between low-level image features and high-level semantics, so as to automatically extract valuable information from images or videos and eliminate semantic gaps, and apply them to the entire life cycle of building projects, including the planning and design stage [4], construction stage [5] and operation and maintenance stage [6]. On this basis, visual attributes are also used to enhance the effect of target detection and to share attributes, so as to enhance the model to achieve real-time, objective and robust detection of unsafe conditions of objects in complex construction dynamic scenarios. Because these technologies provide various advantages, mainly in terms of real-time performance, adaptability and logical reasoning. Therefore, the main research question of this paper is whether visual attributes can provide a basis for solving the above challenges. In order to answer this question, we consider the following sub-questions:

(1) Is it possible to define visual attributes used to characterize the unsafe state of objects at a construction site within the field of construction safety management?

(2) Is it possible to construct an efficient visual attribute perception network model for learning and identifying security-related visual features of a structure from a visual data stream?

(3) Is it possible to establish a semantic mapping mechanism between the visual perception features of unsafe states of objects and engineering safety semantics in the field of construction safety management?

This research focuses on monitoring the unsafe conditions of objects at construction sites, using visual attribute detection as the main framework. We will review and address these issues in this paper. To achieve this, we extend existing concepts of visual attributes to higher-level visual attributes in the field of construction safety management—the visual unsafety of objects. Then, based on three stages—constructing a visual attribute label set, constructing a visual attribute detection dataset, and training and validating a deep learning algorithm model—we build a visual unsafety detection model for objects, proposing a construction safety management framework based on visual attributes. Therefore, depending on the application, it can achieve cognitive alignment and real-time detection of visual perception features and engineering semantics in construction scenarios, enabling data-driven construction safety management.

1.4. Structure of the paper

The next section briefly outlines the application of visual technology in construction safety management (Section 2.1) and reviews the development of the field of visual attributes, discussing the current research status of visual attributes (Section 2.2), providing theoretical background knowledge for the research methodology framework in Section 3. Section 4 verifies the effectiveness of the model through relevant experiments. In Section 5, based on the visual attribute detection results of objects, we conduct an empirical analysis by selecting object impact as a typical case scenario to assess the risks at construction sites, thereby verifying the effectiveness of the proposed method. In Section 6, we discuss the technical contributions and effectiveness of this paper, as well as general findings and future work. Finally, this paper concludes with a summary of its main contents in Section 7.

2. Background and related work

2.1. Application of vision technologies in construction safety management

Computer vision is an interdisciplinary technology derived from the study of human visual systems. It aims to transform visual sensor data into semantic information and enable computers or robots to have visual scene understanding capabilities similar to humans through image recognition, analysis and understanding. Visual perception is an important part of state detection, and it plays an indispensable role in identifying the operational changes of various equipment. Nowadays, in the field of construction safety management, the application of computer vision technology mainly includes the following three aspects: detection, identification and tracking; safety assessment; and hazard prediction [7].

Object detection in the construction field has mainly gone through three stages: geometry and appearance-based methods [8], feature-based methods [9], and deep learning-based methods [10,11]. In the early stages of the construction industry, research mainly used geometry or appearance-based methods for object recognition. For example, Chi *et al.* [8] selected four main geometric and appearance features, including aspect ratio, height-normalized area size, bounding box occupancy percentage, and average gray color, to classify moving equipment and workers. In the transition stage, visual feature detection became the mainstream. Feature methods mainly use the capture of local visual features to represent the information of objects. The main ones are Scale Invariant Feature Transform (SIFT) [12], Oriented Gradient Histogram (HOG) [13] and Speed-Up Robust Feature Transform (SURF) [14]. Experiments have shown that feature-based methods have stronger robustness than previous methods. With the rapid development of neural networks, deep learning has become the mainstream method in the industry. The main methods include Convolutional Neural Network (CNN), R-CNN, Faster R-CNN, *etc.* For example, Kim *et al.* [15] used a region-based fully convolutional network as a classifier to identify heavy equipment on site.

Behavior recognition is a research hotspot and a difficult problem in computer vision [7]. The goal of behavior recognition is to understand the behavior of the target. Construction safety accidents are often closely related to human behavior. Li *et al.* [16] used human skeleton key points and spatiotemporal graph convolutional network technology to extract skeleton key points from images to judge the body posture and working status of workers, which helps in the safety management of the construction process and to perceive the abnormal state of workers in the construction process.

Target tracking is equally important in construction. Generally speaking, the task of target tracking is to monitor the state changes of a specific target over time. Target tracking methods are mainly divided into contour tracking methods [17], kernel tracking methods [18] and point tracking methods [19]. For example, Li *et al.* [20] combined the improved YOLOv5s target detection algorithm and the Strong SORT multi-target tracking algorithm to monitor and track whether workers are wearing safety helmets in real time and perceive the safety risks of workers in the construction site.

In contrast, the basis of construction site safety assessment is the detection of targets, the identification of behaviors, and the tracking of objects, which is also a combination of the above technical means [21]. Generally speaking, assessing the safety status of construction sites is a higher-level application of target identification, tracking, and action recognition [7]. The information extracted from the above technologies still needs further processing, often requiring assessment methods based on rules

or ontology-based knowledge models, such as the safety rules designed by Chi *et al.* [22] to detect safety violations at work.

Hazard prediction is the ideal state of safety management. From the perspective of construction safety management, it is possible to perceive the arrival of a safety accident before it occurs. This is called hazard prediction. So far, in the construction industry, vision-based prediction methods are difficult and limited. Zhu *et al.* [23] used Kalman filters to predict the movement of workers and mobile equipment on the construction site to prevent potential safety accidents. In the field of computer vision, Farha *et al.* [24] proposed a prediction method using Recurrent Neural Network (RNN) and CNN, which can predict content up to several minutes in advance, but it has not yet been applied to the construction field. Overall, the application of hazard prediction in the construction industry is limited.

Overall, while computer vision technology has brought new vitality to construction safety management, problems and challenges remain. Even though vision technology has demonstrated powerful effects, single vision technologies are still insufficient to solve safety issues on construction sites, and problems such as the inability to achieve end-to-end integration and barriers to industry knowledge persist. Furthermore, vision methods only focus on information related to the object itself, concentrating on applications at the levels of detection, recognition, tracking, and segmentation, without incorporating deeper semantics.

2.2. Current research on visual attributes

The essence of attributes is a mid-level semantic that connects the low-level semantics that computers can understand and the high-level semantics that humans can understand. Both visual attributes and low-level visual features can describe images. The difference is that features can only be recognized by computers and cannot be understood by humans [25]. Visual attributes are an abstraction of low-level features by humans and are a higher-level description of low-level semantics.

Visual attributes, as a medium connecting low-level image information and high-level semantic attribute labels, have important applications. In the field of image recognition, attribute detection occupies an important position. Siddiquie *et al.* [26] used image attribute streams to achieve image retrieval and sorting, and took the relationship between attributes into the attribute production model, which improved the accuracy of multi-attribute retrieval. In addition, some scholars have enriched the concept of attribute detection, and proposed that images can be described based on generated structured semantic triples < object, action, scene > [27], or a hierarchical tree semantic unit can be used to describe images at different semantic levels (attribute level, category level, *etc.*) [28].

More importantly, visual attribute detection has been further developed in the field of human-computer interaction. In 1996, Gibson [29] proposed the concept of availability, which describes the availability of the environment as either positive or negative support that can be provided for animal behavior [30]. Norman [31] further developed this concept, describing availability as the basic property that determines how to use something, such as whether a stone can be thrown or a button can be pressed. This concept has received widespread attention, especially after being introduced into the field of computer vision, breaking the constraints of traditional visual monitoring tasks, further exploring the properties of objects, and providing new ideas and application scenarios for attribute detection and computer vision [30].

In the field of architecture, visual attributes have been widely applied throughout the entire lifecycle of building projects, including the planning and design phase [32], the construction phase [33–35], and

the operation and maintenance phase [36]. Specifically, in the planning and design phase, the research interests in the planning phase are mainly focused on urban and spatial analysis, while the research interests in the design phase are mainly focused on drawing inspection and 3D modeling. Tomé *et al.* [32] used visual technology to analyze the spatial attributes of buildings, studied the functional status of architectural relics, and promoted the understanding of spatial conditions. In the construction phase, the research interests of scholars tend to be focused on safety management, progress monitoring, and quality inspection. This is also the key stage of attribute detection research, and most of the current research focuses on this.

The study of visual attributes has also gone through a process from traditional machine learning to deep learning. Traditional machine learning methods mainly include Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Bayesian networks and other methods to detect attributes extensively. In the early stages of machine learning, Piyathilaka *et al.* [37] used SVM to study attributes, learned attributes through geometric features, and used SVM as a classifier to classify attributes by labels. In addition, the KNN method has also been widely used. However, as research deepened, machine learning algorithms could not complete higher-level attribute detection, and deep learning developed accordingly. Due to its better accuracy and stronger adaptability, deep learning has played an important role in the field of visual attribute detection in recent years. Commonly used deep learning methods include CNN, RNN and other methods. Roy *et al.* [38] constructed a multi-branch CNN architecture to achieve attribute segmentation, extracted depth maps, surface normals and semantic segmentation features through independent multi-scale networks, and designed a feature fusion module for pixel-level attribute prediction. Nguyen *et al.* [39] developed an end-to-end training mechanism based on RGBD data, and used a deep convolutional network to directly learn high-order feature representations of availability detection from images.

In recent years, the Transformer architecture has received increasing attention in the computer vision community due to its high performance and low requirement for visual-specific sensory bias [40], which has also made the detection of visual attributes possible. As a powerful deep learning model, the Transformer can effectively capture the complex relationship between images and text by utilizing its self-attention mechanism [41] and extract intrinsic features [42]. In addition, the Transformer architecture can also completely change the traditional sequence modeling method through its self-attention mechanism and parallel computing advantages, overcoming the limitations of recurrent neural network sand long short-term memory networks in terms of long-range dependency and computational efficiency [40]. This also makes end-to-end object detection possible, greatly simplifying the process of visual detection tasks.

Therefore, attribute detection is a medium for computers to perceive and understand the real world. For the application of visual attributes in construction safety management, visual attributes can bridge the gap between underlying visual information and construction safety management. However, the current problem is that the application of existing visual attributes is highly correlated with the development of visual technology. Although many applications already exist, such as attributes like target shape, category, location, path, and posture, there is still a lack of high-level visual attributes that can be integrated with construction management. The attributes representing unsafe states of objects cannot be determined, and there is a lack of corresponding visual attribute detection frameworks.

3. Methods for detecting visual unsafety of objects

This paper builds upon existing research on visual attributes and proposes visual unease as a high-level visual attribute in the field of construction safety management. Using the GLM-4 large language model, relevant labels describing the visual unease of objects are extracted from accident reports, thus constructing the visual attribute label set for this paper. Then, based on the SODA dataset, we construct a new dataset, VALD (Visual Attribute Learning Datasets), for object impact accidents in construction safety scenarios.

For the network model used to detect the visual attribute of unsafe object states, the YOLOv11 model was chosen. YOLOv11's advantages over previous models lie in its optimized network architecture, efficient feature extraction mechanism, and targeted loss function design. This allows it to enhance the extraction of detailed features, providing rich feature support for attribute recognition. Considering the nature of object unsafety and the complexity of construction sites, the model needs strong spatial attention, strong reasoning ability, and fast and accurate computational capabilities. Therefore, YOLOv11 was selected as the detection model for visual unsafe objects in this study, and the overall framework of the method is shown in Figure 1.

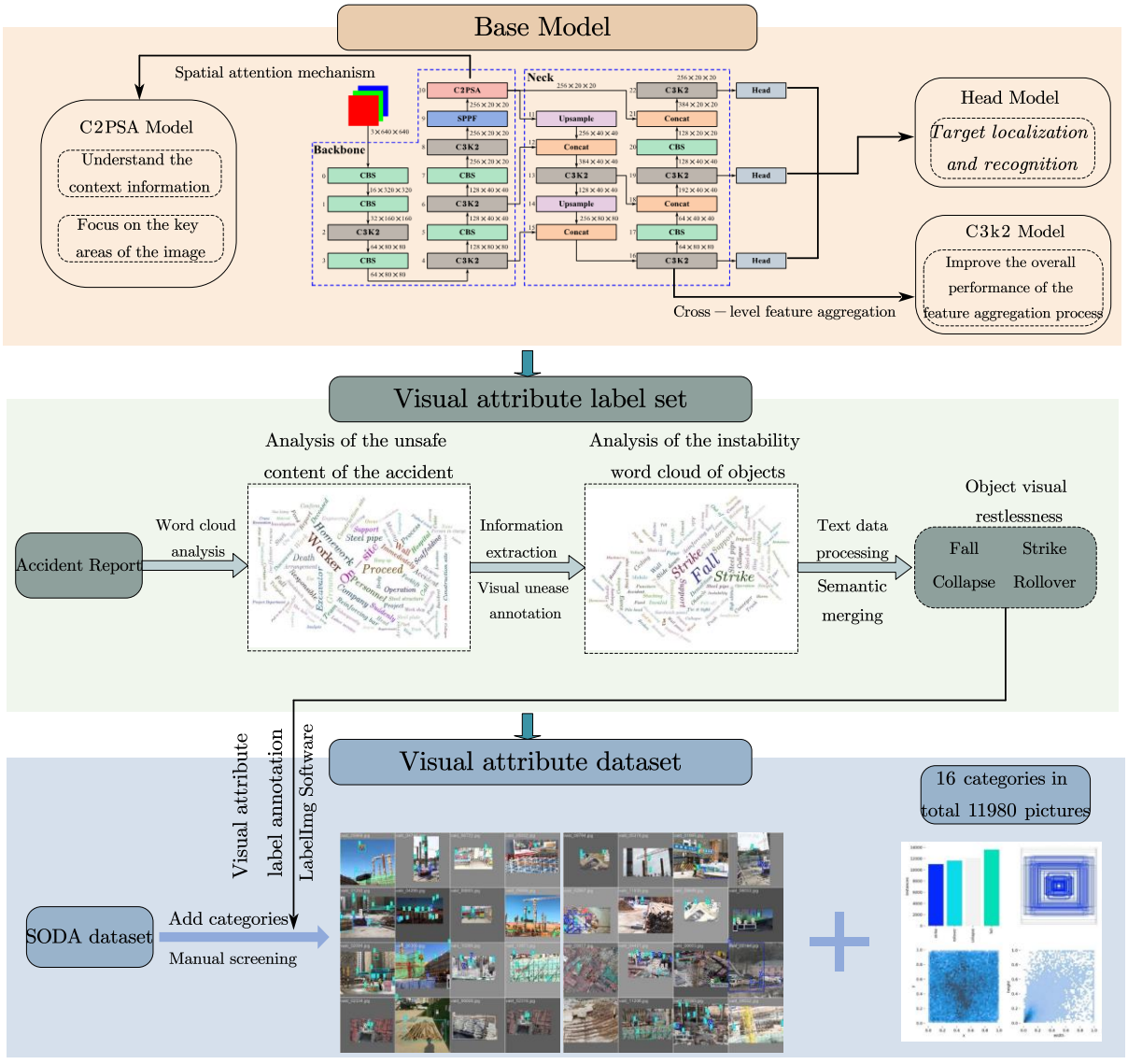

Figure 1. Overall framework.

By definition, visual insecurity is a type of visual availability. The concept of visual availability is much broader, while the insecurity of an object only focuses on information in the field of construction safety. Therefore, we can use the classification method of availability to establish a conceptual framework for the visual insecurity of an object [31]. Based on a summary of existing concepts and application frameworks of visual attributes, we propose an advanced visual attribute in the field of construction safety management—the visual unsafety of objects. We define the visual unsafety of objects as: the degree to which an object, in visual data such as images and videos, may cause accidents, injuries, or health risks due to its physical properties, state, usage, or environment.

To further clarify the conceptual differences between the object visual insecurity proposed in this paper and related research paradigms, this paper compares it with danger-aware object detection, availability/usability detection, and risk-aware visual systems in terms of research focus, output format, context awareness, and knowledge sources, as shown in Table 1.

Table 1. Comparison of visual insecurity of objects and related research paradigms.

Conceptual Dimension	Hazard detection object detection	Availability Detection	Risk perception visual system	Visual unease about objects Safety
Focus of attention	Inherent danger of the object	Object-action relationship	Scene rule adaptation	Dynamic danger state of objects
Output format	This is a dangerous object.	Can perform a certain action part	Violating a certain rule	It may cause some kind of accident.
Context awareness	none		Dependency rules	powerful
Source of knowledge	Category Tags	Action annotation	Expert Rules	Accident text + rule learning

As shown in Table 1, the "visual unsafe attribute of objects" proposed in this paper focuses on characterizing the dynamic risk state of objects and their potential accident types in specific contexts. It models the relationship between the object's state and the environmental context through visual attribute learning, thereby achieving a visual representation of risks in construction scenarios.

Traditional object detection and rule matching methods rely on identifying explicit hazardous objects or rule violations, making it difficult to effectively model implicit risks caused by continuous state changes or structural instability. This paper proposes an object visual insecurity approach that introduces a learnable intermediate semantic layer between object detection and safety inference. By embedding risk states as visual attributes into the detection framework, the model can encode risk semantics during the feature learning stage, achieving a shift from post-detection inference to detection-as-risk representation. This significantly improves risk modeling capabilities in complex construction scenarios.

3.1. YOLO-SafeAttr: methods for detecting the visual unsafety of objects

The YOLO algorithm is more suitable for training small-scale models due to its advantages of high speed, high accuracy and lightweight. The YOLO [43] algorithm pioneered the single-stage object detection algorithm, creatively treating the object detection task directly as a regression problem, combining the candidate region and detection stages into one, and obtaining prediction results from the image end-to-end. The visual attribute detection process of this study is shown in Figure 2.

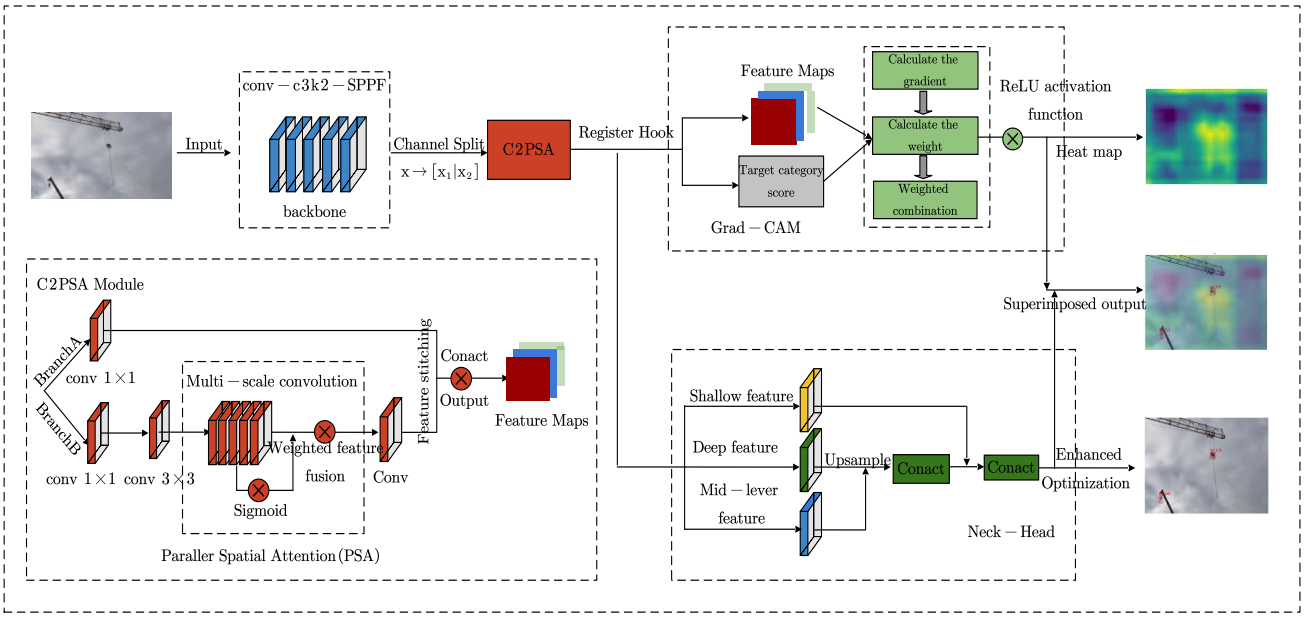

Figure 2. Flowchart of visual attribute detection.

The YOLOv11 model framework mainly consists of three parts: Backbone, Neck, and Head [44]. Among them, the Backbone network undertakes the core function of visual feature extraction, and realizes multi-level feature extraction of input images through convolutional neural networks. It introduces a new Cross-Stage Partial Spatial Attention (C2PSA) block [45]. The C2PSA module can enhance spatial attention in feature maps, enabling the model to focus more effectively on important regions in the image. By pooling spatial features, the C2PSA module enables YOLOv11 to focus on specific regions of interest, thereby improving the detection accuracy of objects of different sizes and positions.

Secondly, the feature fusion module adopts cross-level feature aggregation technology to construct a feature pyramid adapted to multi-scale detection tasks. Innovatively, the C3k2 block is introduced in this module, which replaces the C2f block used in previous versions [45]. The C3k2 block is a more efficient implementation of the cross-stage part (CSP) bottleneck, aiming to be faster and more efficient, thereby improving the overall performance of the feature aggregation process. The detection head module performs bounding box regression and class probability prediction based on the optimized feature map, and finally completes the dual tasks of target localization and recognition.

In terms of the loss function, YOLOv11 employs a multi-task learning objective composed of three major components:

$$L_{\text{total}}=L_{\text{box}}+L_{\text{cls}}+L_{\text{dfl}} \tag{1}$$

The L_{box} loss is used to optimize the position of the bounding box. YOLOv11 typically uses Complete Intersection over Union (CIoU) loss, which considers the overlapping area of the bounding boxes, the distance between the center points, and the aspect ratio, enabling the model to predict a more stable bounding box that better fits the target shape.

$$L_{\text{box}}=1\text{-}CIoU=1\text{-}\left(IoU\text{-}\frac{\rho^2(b_{\text{pred}},b_{\text{gt}})}{c^2}\text{-}\alpha v\right) \tag{2}$$

L_{cls}class loss measures the accuracy of a model's prediction of an object's class. For multi-class classification problems, cross-entropy loss is typically used to support multi-label detection scenarios,

reduce classification errors, thereby improving the accuracy of class identification and ensuring that the model can correctly distinguish between different class targets.

$$L_{cls}=-[y\log(p)+(1-y)\log(1-p)] \tag{3}$$

The Distributed Focus Loss (DFL) is a loss function in the YOLO series. Its main purpose is to address the class imbalance problem in object detection and improve the model's performance when handling small objects and difficult samples.

$$L_{dfl}=-\sum_i y_i \log(p_i) \tag{4}$$

Overall, YOLOv11's core breakthrough in visual attribute detection lies in its newly added C2PSA spatial attention module and multi-layered collaborative loss function system. The C2PSA module enables the model to more effectively understand the scene context and the relationships between objects, significantly enhancing its robustness in recognizing small targets and partially occluded objects, thus achieving a leap from detecting objects to assessing risks in conjunction with the environment. Simultaneously, the model significantly improves detection accuracy, robustness, and convergence speed by integrating multiple losses such as localization, classification, and boundary fitting for joint optimization. These two improvements enable YOLOv11 to more accurately and stably complete visual attribute recognition and safety status determination in complex construction scenarios.

3.2. Construction of visual attribute label set based on large language model

For the study of visual unsafety of objects, only accurate visual attribute labels can effectively measure the unsafe state of an object. The research objects of visual unsafety are those that may cause safety accidents at construction sites; therefore, visual attribute labels related to the objects can be compiled from accident reports.

First, we manually collected 100 reports of object strike accidents from Guangdong Province, China. Simultaneously, we extracted necessary information from these reports, manually extracting the text content under the Accident Description heading and saving it to an Excel file. We then performed word cloud analysis on the 100 Accident Description sections. As shown in Figure 3a, words related to on-site construction, such as site, worker, operation, and conducting, appeared frequently in the collected text, indicating a high correlation between the collected data and on-site construction operations. It also reveals that the accidents involved a complex and diverse range of entities, actions, and behaviors, thus requiring further data processing.

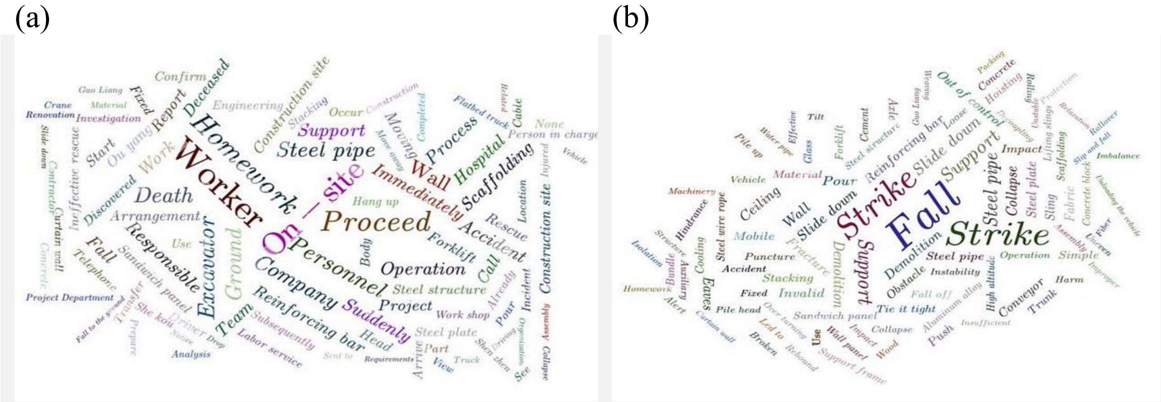

Figure 3. Extraction of visual attribute label information.Main title. **(a)** Analysis of the content word cloud of the accident process; **(b)** Analysis of the instability word cloud of the accident subject and the involved objects.

This paper selects the GLM-4 large language model to extract accident safety information. The prompt word project adopts the ICIO Prompt framework, and the prompt words are shown in Figure 4. Key information is accurately extracted from construction safety accident reports, the objects involved are listed, and their visual unsafety is labeled, thus obtaining the unsafety of all accident subjects and related objects, as shown in Figure 5. In addition, to ensure the accuracy and reliability of the model extraction results, this paper constructs a manual review mechanism. Researchers check the model output results one by one, correcting and unifying obvious semantic deviations or label errors. After manual verification, a final set of visual unsafety attribute labels is formed. Through the manual review process, the semantic drift problem that may be generated by the large language model is effectively avoided, and the stability and credibility of the label system are improved. Then, word cloud analysis is performed on the unsafety of all the collected accident subjects and related objects, as shown in Figure 3b. It can be seen that "falling" is the most frequently occurring attribute, and other attributes such as "hit", "slip", and "collapse" also appear frequently. At the same time, the corresponding accident types account for a high proportion in the accident statistics, reflecting the relevance and authenticity of the collected accident report data.

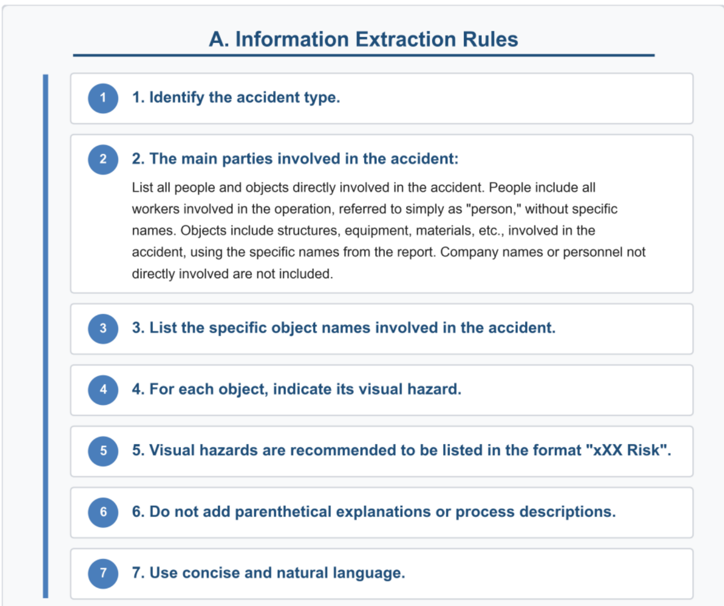

Figure 4. Prompt words used in the GLM-4 extraction process.

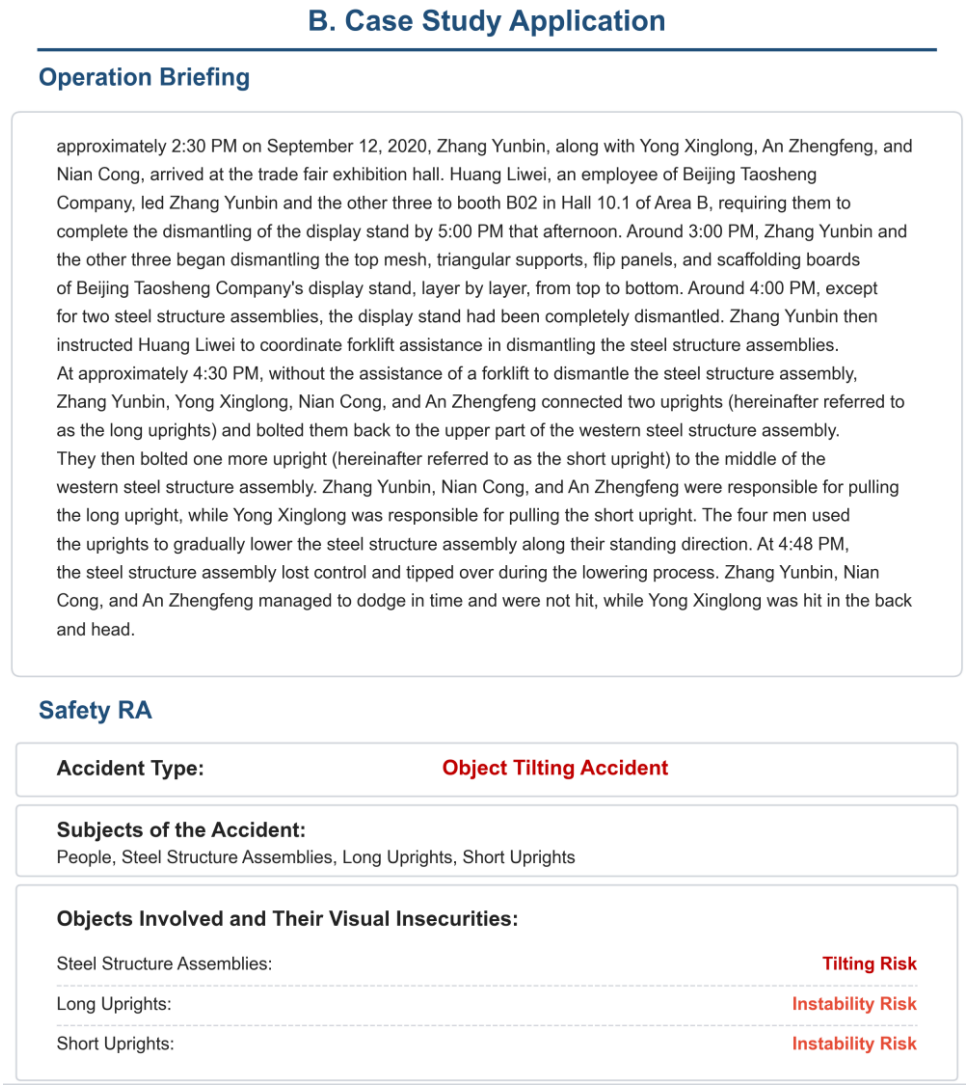

Figure 5. Example of attribute extraction for the GLM-4 large language model.

Further processing of the text data removed the main accident events, leaving only the insecurity of the objects we needed, and analyzed the instability of these objects. After constructing 29 initial visual attributes, this study further optimized the attribute system and abstracted risk levels. The initial attributes were derived from semantic extraction and visual recognizability screening of the accident text, enabling fine-grained descriptions of the abnormal states of objects at the construction site. However, system analysis revealed that these attributes exhibited significant aggregation characteristics along the accident evolution path; different appearances often corresponded to similar accident outcomes, which could be summarized into a few typical forms of structural instability.

Meanwhile, to avoid the difficulty of identification caused by too many attributes and to address the issue of inconsistent synonyms (such as “decoupling” and “loosening”), we semantically merged the original 29 visual insecurity attributes, combining attributes with the same meaning into a single attribute to eliminate ambiguity, save computational resources, and facilitate subsequent management. The specific correspondence is shown in Table 2 below. As can be seen from the table, the visual attribute label set includes four attributes: falling, impact, collapse, and rollover. These attributes can cover most safety incidents in object strike scenarios, characterize the visual hazard of objects, and serve as annotation standards for subsequent visual attribute detection datasets.

Table 2. Semantic Correspondence of Visual unsafety.

Consolidated Attribute Type	Included Original Attributes	Count
Fall	Fall	24
	Sliding	7
	Detachment	2
	Loosening	2
	Unhooking	2
	Instability	3
Strike	Strike	16
	Piercing	3
	impact	2
	Collision	2
	injury	1
Collapse	Collapse	4
	Toppling	4
	overturning	2
	Structural Failure	1
	Tilting	1
Rollover	Rollover	1
	Loss of Control	3
	Unstable	1
	Imbalance	1
	Rolling	1

It should be noted that the 100 accident reports were used to construct a visual unsafety attribute system for object-strike scenarios, rather than to infer the overall statistical distribution of all construction accident types. Since these reports were collected from object-strike accidents in Guangdong Province, China, and were processed through GLM-4 extraction and manual verification, they were able to cover the recurrent unsafe object states within the scope of this study. Although the 29 initially extracted attributes were fine-grained in expression, their accident evolution logic mainly corresponded to four consequence categories: falling, impact, collapse, and rollover. Therefore, the semantic merging in Table 2 was based on consequence consistency, visual recognizability, and annotation stability. Attributes with similar meanings, similar visual manifestations, or the same risk consequence were merged into the same category, forming a four-class visual attribute system suitable for image annotation and model training.

3.3. VALD visual attributes dataset

Since the research object of this paper is a common object in the construction site, which is the same as the object detection object, the construction of the visual attribute detection dataset can be carried out with reference to the construction of the object detection dataset. Currently, commonly used object detection datasets include: SODA, CHV, MOCS, ACID, AIMD and other datasets [9]. Among them, the SODA dataset has more categories compared with other datasets. Therefore, it is more suitable for the construction of the visual attribute dataset for the construction site. Then, based on the SODA dataset, we constructed a new dataset VALD dataset (Visual Attribute Learning Datasets, VALD) for the object hitting construction safety accident scenario. The original 15 categories were expanded to include the category image of

construction machinery, for a total of 16 categories, covering common objects that appear frequently in the construction site. In addition, 5000 manually photographed and collected images were added on the basis of retaining some SODA datasets to form a new VALD dataset, which contains 11,980 images.

The object-strike scenario refers to a typical risk scenario on construction sites in which unsafe object states, such as falling, striking, collapse, or rollover, may cause harm to workers or surrounding structures.By selecting object impact scenarios and extracting visual attribute labels of object safety hazards based on real accident reports, a visual attribute detection dataset was labeled. The LabelImg software, an open-source image annotation tool developed in Python, was used for labeling. Detailed annotation content is shown in Table 3 below, which also illustrates the potential safety hazards of common objects in the scenario. Four visual attribute labels were included: falling, impact, collapse, and rollover. The image data subdivided by attribute is shown in Figure 6. Furthermore, to avoid semantic conflicts and enhance discriminative ability, this study adopted a dominant risk priority strategy, using single-label annotation. In complex safety emergency responses, the most urgent and deadliest dominant risks are prioritized. For example, although a hook may have multiple possible object safety hazards, including falling and impact, falling is currently the more significant visual attribute for a hook in the air. Simultaneously, to avoid semantic conflicts and enhance discriminative ability, this study only labeled the dominant visual risk for each instance.

Visual attribute annotation was performed on the VALD object detection dataset to obtain the VALD visual attribute detection dataset. Partial annotation results are shown in Figure 7 and the specific image format is shown in Figure 8. As can be seen from the figures, the dataset is constructed around object impact scenarios, with a relatively rich variety of object types and diverse dangerous scenes. Furthermore, the instance attributes of the four labels are roughly the same, and the image size distribution is also relatively wide, indicating strong generalization ability for models trained on this dataset.

Table 3. Attribute annotation references and possible unsafe aspects of common objects.

Object	Potential Hazard	Possible Visual Unsafe Attributes
board	Sliding when unsecured; sharp edges causing puncture; unstable stacking leading to collapse	Fall, Strike, Collapse,Rollover
wood	Sliding during handling; long pieces rolling due to imbalance; sharp ends causing impact	Fall, Strike, Collapse
rebar	Slipping and falling; sharp ends causing puncture; support failure leading to collapse	Fall Strike, Collapse
brick	Overstacking causing collapse; falling during handling; impact injuries	Fall, Strike, Collapse
scaffold	Structural instability causing collapse; connector failure; plank detachment and falling	Collapse, Break, Fall
handcart	Overloading causing toppling; cargo slipping; uncontrolled movement causing impact	Rollover, Fall, Strike
cutter	Blade breakage causing flying debris; uncontrolled rebound during operation	Strike
hopper	Overloading during filling causing rollover; discharge blockage leading to instability; structural breakage and leakage	Fall, Rollover
hook	Hook breakage or detachment; suspended load falling; uncontrolled swinging causing impact	Fall, Strike
fence	Unstable foundation causing toppling; damage from external impact	Rollover, Strike
slogan	Improper installation causing falling; strong wind causing toppling; metal edges causing impact	Fall, Rollover, Strike
con_machine	Robotic arm breakage; hopper overturn; pipe blockage causing rupture	Rollover, Strike

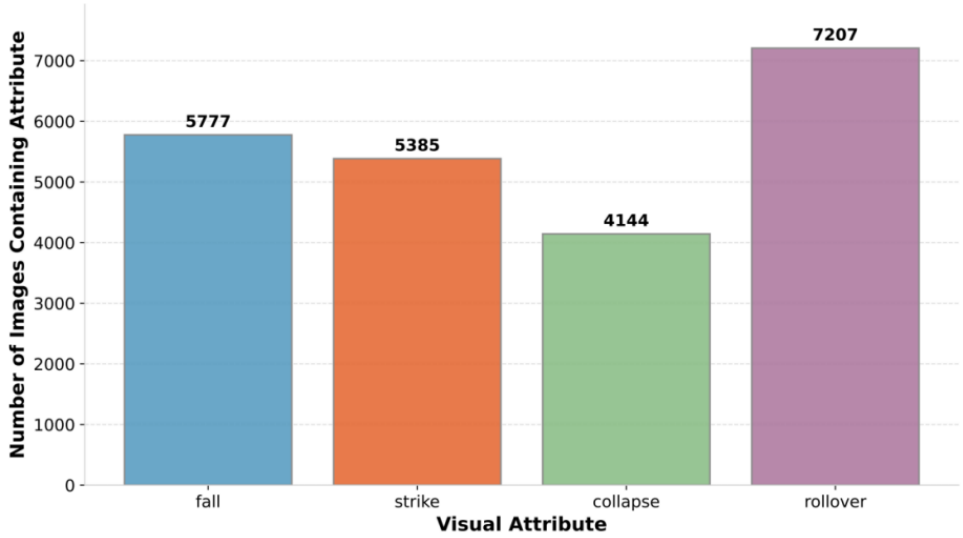

Figure 6. Image segmentation data of attributes.

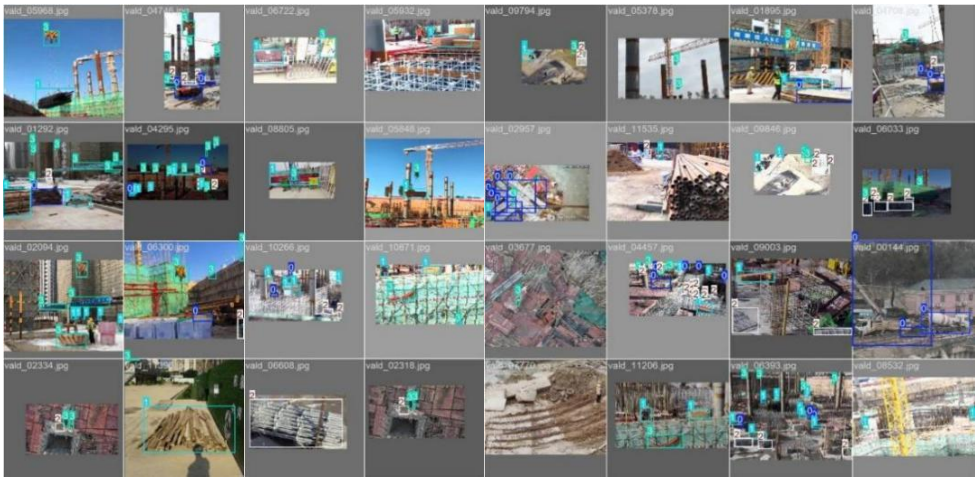

Figure 7. Part of the annotated results of the VALD visual attribute detection dataset.

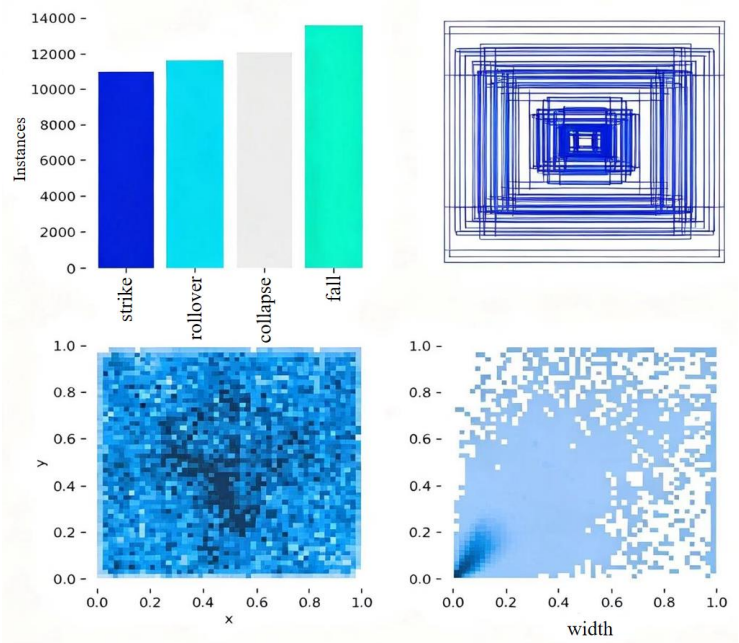

Figure 8. The number of labels and image formats for each category in the VALD visual attribute detection dataset.

This paper proposes a visual attribute-based paradigm for detecting unsafe object states, focusing on four typical risks at construction sites: falls, impacts, collapses, and rollovers. This paradigm shifts from "identifying objects" to identifying risks. Unlike the traditional two-stage process of identifying objects first and then attributes, this method directly focuses on where the unsafety lies, enabling the model to learn the semantic association between object features and risk attributes during training. This allows it to capture more critical contextual clues in complex working conditions. Visual attribute detection can directly quantify visual information related to danger, such as the relative distance between personnel and the danger edge, and changes in posture, allowing potential risks to be exposed earlier. Through end-to-end attribute learning, the model eliminates intermediate steps, improving detection robustness and reducing false positives and false negatives caused by factors such as occlusion and cluttered backgrounds. Furthermore, unsafe attributes such as falls and collapses have cross-object category universality, giving the model stronger transferability and allowing it to generalize to unseen scenes and new object types. Overall, the visual attribute-based detection paradigm advances the vision system from "identifying what is seen" to "locating hazards," providing more accurate and scalable intelligent monitoring capabilities for construction safety.

4. Results

4.1. Model training and evaluation

This study randomly divided the VALD dataset into training, testing, and validation sets in an 8:1:1 ratio. The YOLOv11 model was trained and validated for object impact scenarios, including object detection and visual attribute detection. Object detection extracted the target category of the object; visual attribute detection extracted the safety risks of objects at the construction site, revealing the potential hazards of objects in their current state. A confusion matrix was used for evaluation, and precision, recall, mAP, and F1 score were used as metrics to assess experimental performance. The specific calculation process is shown below.

$$precision=\frac{TP}{TP+FP} \tag{5}$$

$$Recall=\frac{TP}{TP+FN} \tag{6}$$

$$AP=\int_0^1 p(r)dr \tag{7}$$

$$mAP=\frac{1}{N}\sum_{n}^{i=1} AP \tag{8}$$

First, the YOLOv11 object detection model and visual attribute detection model were trained separately to verify the effectiveness of the VALD dataset in both tasks. As training progressed, the training and validation losses of both models gradually decreased and stabilized, indicating good convergence. Simultaneously, performance metrics continuously improved and eventually stabilized, demonstrating continuous performance enhancement, as shown in Figures 9 and 10. Experimental results show that in the object detection task, the model achieved 81.10% mAP@50, 88.00% recall, and

an F1 score of 0.79; in the visual attribute detection task, the model achieved 84.80% mAP@50, 91% recall, and an F1 score of 0.82. These results demonstrate that the VALD dataset effectively supports both object detection and visual attribute detection tasks, enabling the model to achieve good detection results, as shown in Figures 11 and 12.

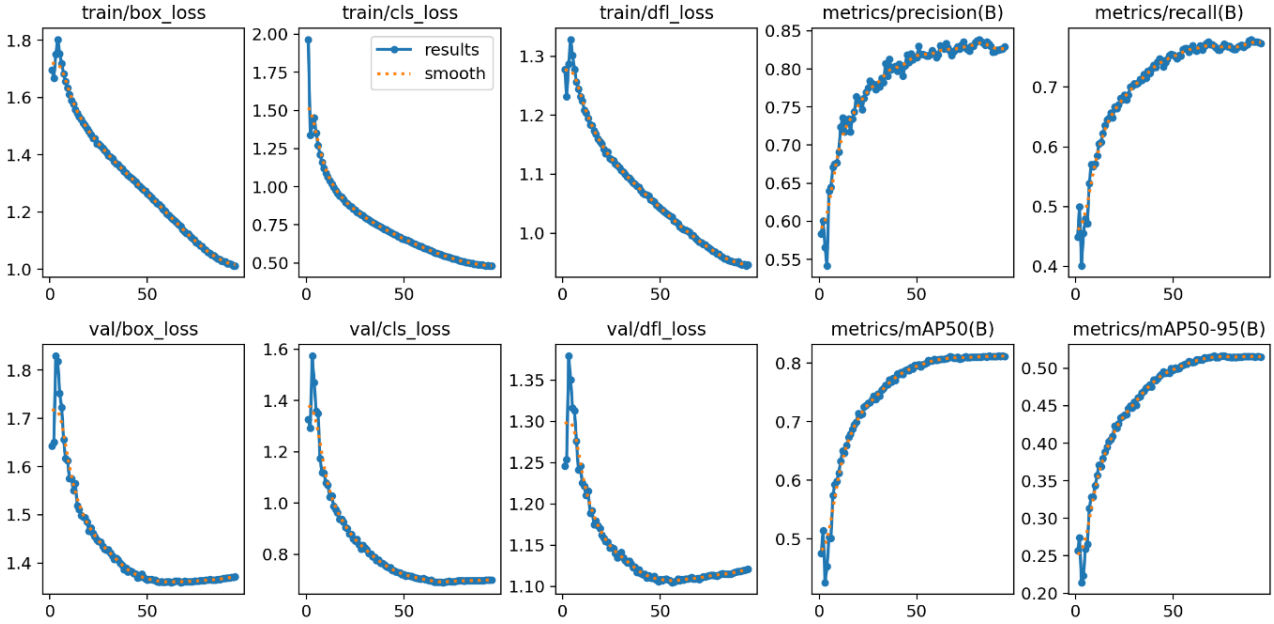

Figure 9. Training and validation results of the object detection model.

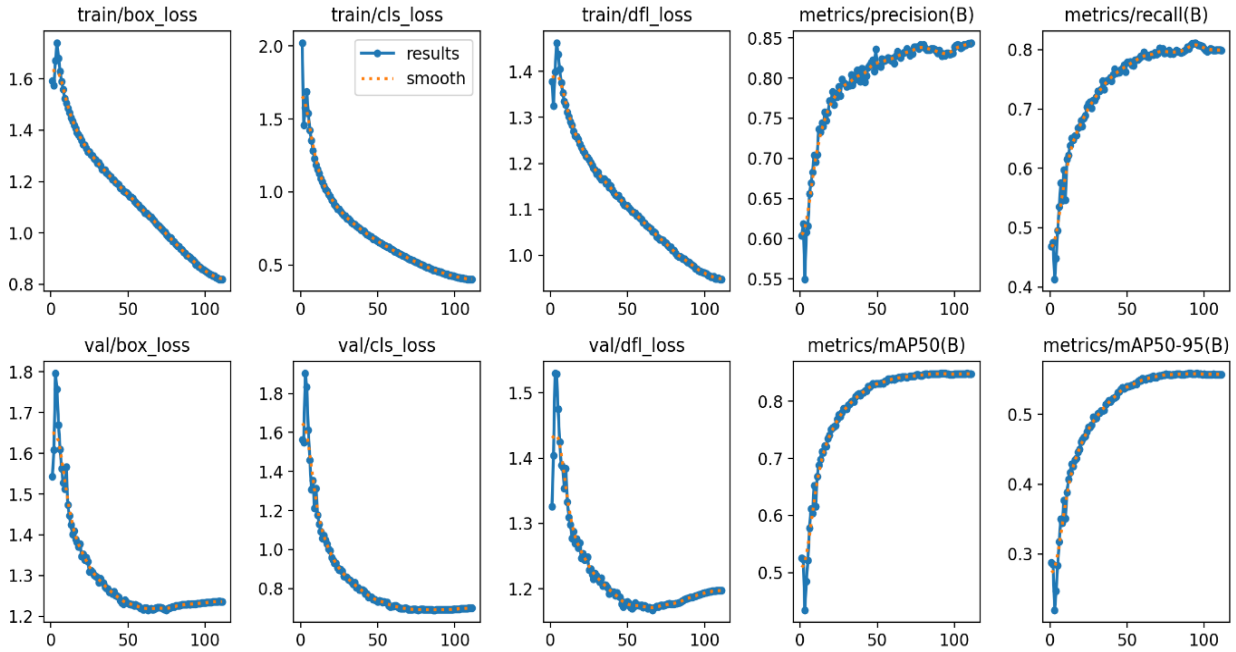

Figure 10. Training and validation results of the visual attribute detection model.

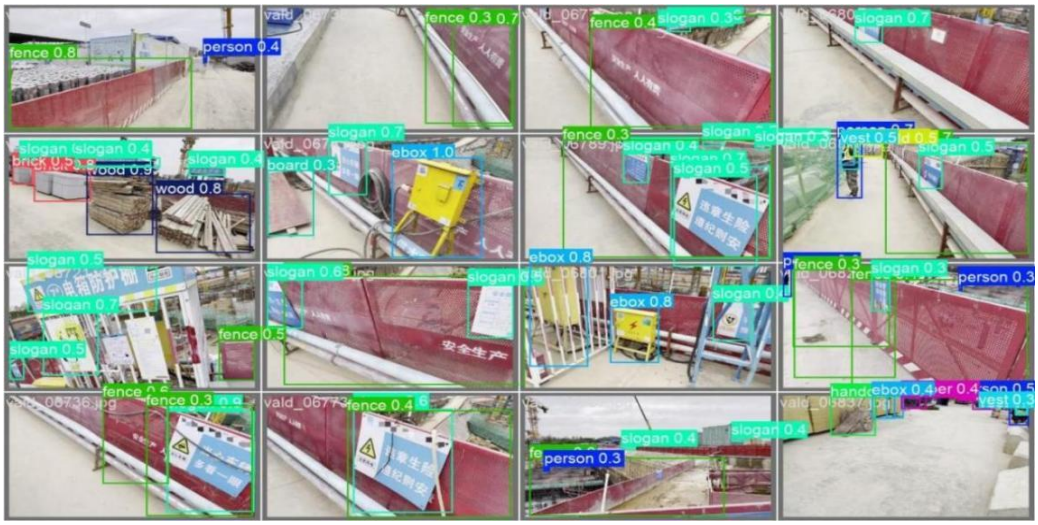

Figure 11. Validation of the target detection model.

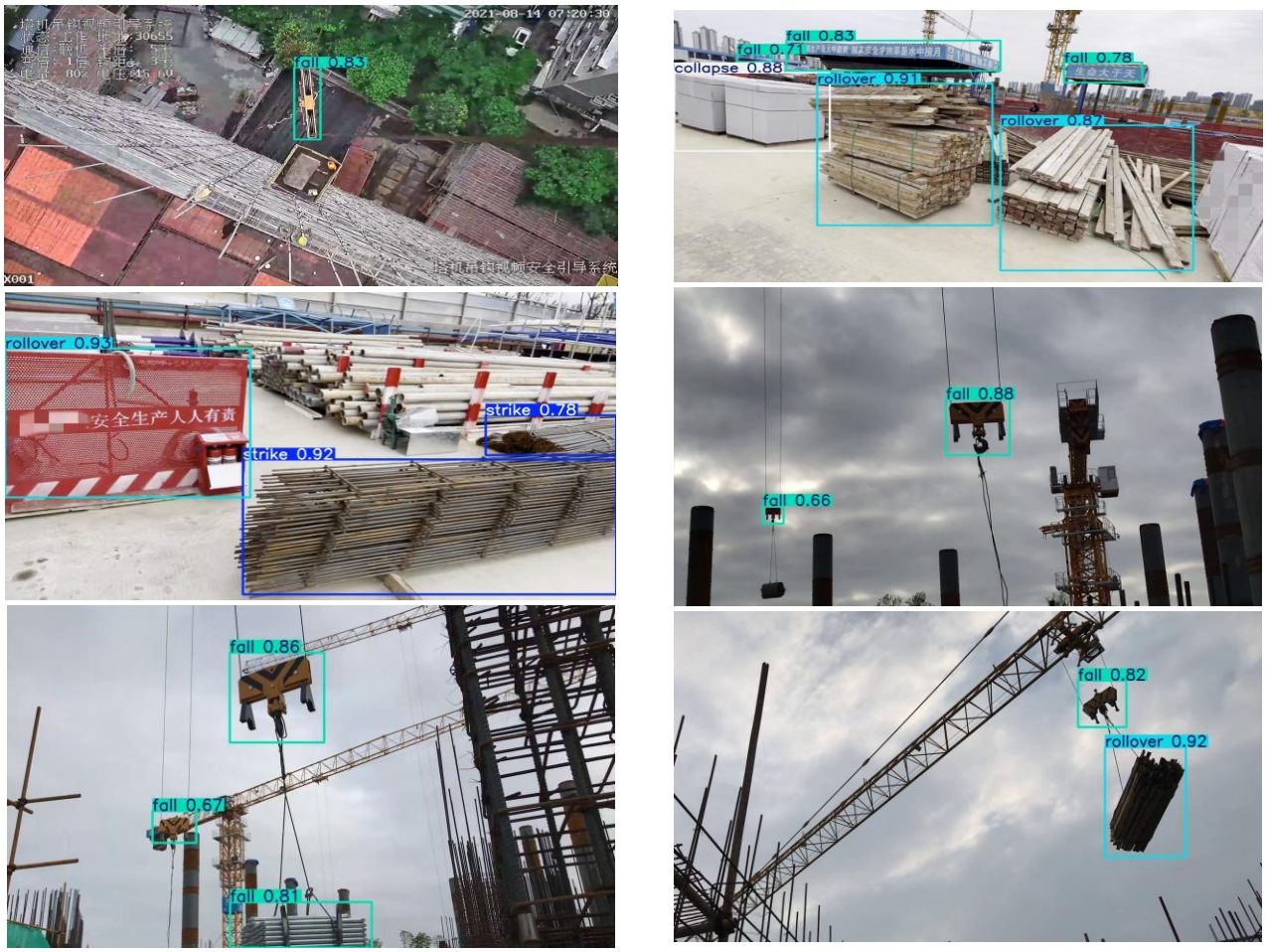

Figure 12. Performance validation of the visual attribute detection model.

4.2. Ablation experiment

To systematically quantify the contribution of the attention mechanism in the YOLOv11 visual attribute detection model proposed in this study, we removed the C2PSA spatial attention module from the backbone of the original YOLOv11 architecture and replaced it with a C3k2 module without an attention mechanism, while keeping the remaining network structure unchanged. The ablation model with the

attention component removed, namely YOLOv11_noattn, was then compared with the complete baseline model, YOLOv11_attn. All experiments were conducted under the same training settings and on an independent test set to ensure the rigor of the evaluation. As training progressed, the training loss continued to decrease and eventually stabilized, indicating that the model had sufficiently learned the features of the training data, as shown in Figure 13. Meanwhile, the validation loss decreased rapidly in the early stage and then entered a stable phase in the later stage, suggesting that the model's generalization ability gradually stabilized. Correspondingly, performance metrics such as precision, recall, mAP@50, and mAP@50-95 showed a steady upward trend and eventually converged. This indicates that the model's performance in the visual attribute detection task continued to improve and finally reached a relatively stable optimal state.

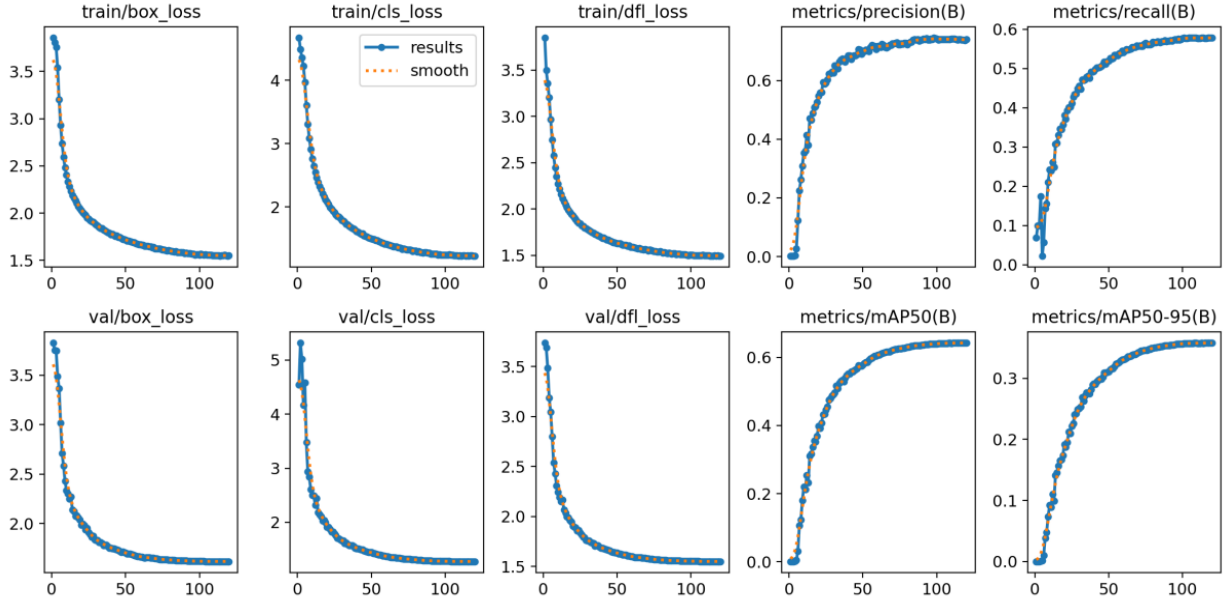

Figure 13. The training and validation results of the YOLOv11_noattn visual attribute detection model on the VALD dataset.

The results show that the performance of the YOLOv11_attn visual attribute detection model suffered a significant loss when the attention mechanism was removed. Specifically, the YOLOv11_noattn visual attribute detection model achieved a detection precision of 63.70% mAP@50, a recall of 57.40%, and an F1 score of 0.645. Detailed comparison results are shown in Table 4.

Table 4. Comparison of relevant evaluation Indicators of the YOLOv11_attn and YOLOv11_noattn visual attribute detection models.

Model	mAP@50	Recall(R)	F1score
YOLOv11_attn (Baseline)	84.8%	91.0%	0.82
YOLOv11_noattn	63.7%	57.4%	0.645
Performance Change	−21.1%	−33.6%	−0.175

This comparison strongly demonstrates the indispensable role of the attention mechanism in the YOLOv11 visual attribute detection model. The model exhibits significant missed detections in complex

scenes after removing the attention module, indicating its crucial role in capturing cross-regional semantic associations. This is also a key factor in the high performance of the baseline model, as shown in Figure 14. Although YOLOv11_noattn achieves extreme lightweighting, the decline in various performance metrics proves that removing the attention module is an unreasonable strategy for the visual attribute detection task in this study.

Figure 14. Verification of the performance of the YOLOv11_noattn visual attribute detection model on the VALD dataset.

4.3. Visualization of the attention mechanism

This paper employs Grad-CAM interpretability technology to reveal the decision-making basis of YOLOv11 in safety risk identification, transforming it from a black box model into an interpretable system. By extracting feature activations and gradients through forward and backward hooks, a visualized heatmap can intuitively show the key areas the model focuses on when predicting unsafe states, providing a basis for model debugging and optimization.

As shown in Figure 15a–d, in complex dynamic scenes such as hoisting operations, the interpretability results indicate that the model has strong contextual feature perception capability. When identifying suspended components at height and determining whether they are associated with “falling” or “collision” risks, the high-response regions are not limited to the suspended object itself, but also cover key structures such as the crane boom, cable force-transmission path, and high-altitude work area. This reflects the model’s ability to associate visual features across different scales and regions. Whether feature response or gradient-guided methods are used, the high-response regions remain relatively consistent, indicating that the model can jointly utilize visual cues related to object state, environmental structure, and spatial relationships during prediction. Based on this global scene perception and contextual integration capability, the model can achieve relatively stable visual attribute recognition under real-world conditions with noise, occlusion, and significant scale variation. The interpretability analysis shows the key regions attended to by the model during risk attribute prediction and further

illustrates the applicability of visual attribute detection in complex interactive behaviors and dynamic construction scenes.

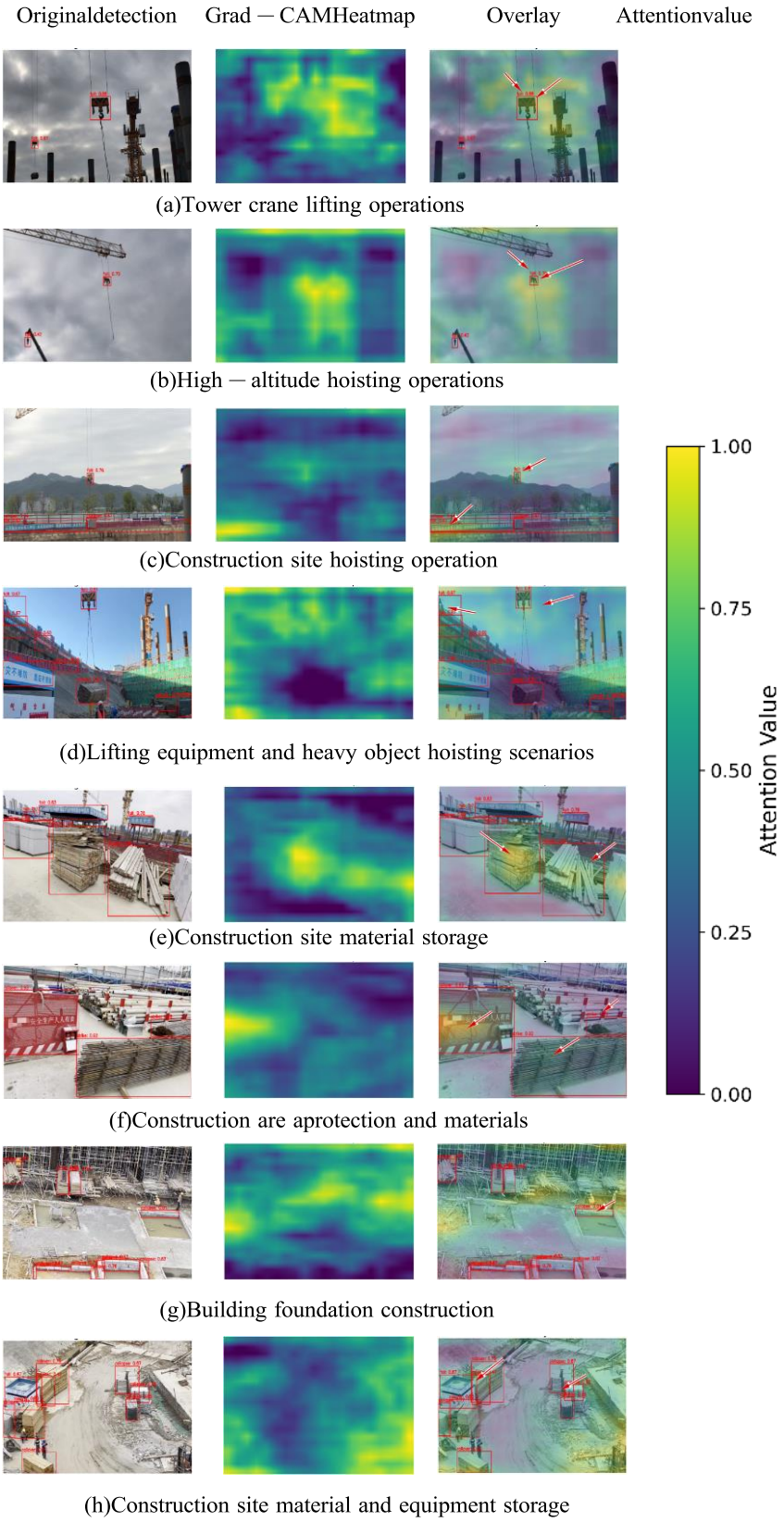

Figure 15. Visualization results of attention mechanisms. **(a–d)** Visual effects in complex dynamic scenarios such as lifting and hoisting; **(e–h)** Visualized effect of the scenario where materials are stacked on an unsafe ground.

Compared to high-altitude hoisting operations, the model also demonstrates strong semantic understanding and risk assessment capabilities in scenarios involving unsafe material stacking on the ground. As shown in Figure 15e–h, the model can not only identify scattered materials such as rebars and planks, but also determine whether they are in an unsafe stacking state. The attention visualization results show that the model's response regions are not limited to the material texture itself, but are mainly concentrated on structural features such as stacking edges, tilted shapes, and irregular contact surfaces with the ground, as well as the spatial relationship between material locations and worker movement paths. This indicates that when identifying the "unsafe stacking" attribute, the model can use risk-related visual information such as stacking form, space occupation, and work-flow interference, rather than relying solely on material appearance.

This attention pattern suggests that the model can extract safety-risk-related contextual features in complex construction scenes. It not only identifies material categories, but also responds to observable information such as stacking state, workspace occupation, and worker passage relationships. Compared with traditional methods that rely mainly on local features, the proposed method can integrate multiple visual cues, including environmental layout, workspace conditions, and worker activities, for risk attribute recognition, which is more consistent with the need for comprehensive on-site safety assessment. The attention visualization results provide intuitive evidence for understanding the model's feature utilization and offer interpretability support for construction-site risk warning.

4.4. Comparative experiment

For the visual attribute detection task, YOLOv8 and Faster-RCNN were selected as comparison models to further evaluate the performance of YOLOv11 on the VALD dataset. YOLOv8 belongs to the same YOLO series as YOLOv11 and was used to compare the performance differences between one-stage detectors, while Faster-RCNN is a classical two-stage detector commonly used as a benchmark in object detection tasks. Comparing these representative models under the same dataset and experimental settings provides a more objective evaluation of YOLOv11 for visual attribute detection in construction scenes. The experimental results are shown in Table 5.

Table 5. Comparison of YOLOv11 with relevant baseline models.

Model	Precision	Recall	F1-score
YOLOv11	0.848	0.910	0.820
YOLOv8	0.810	0.708	0.756
Faster-RCNN	0.781	0.691	0.733

The results show that the YOLOv8 model has a detection precision of 0.81, a recall of 0.708, and an F1-score of 0.756, while the Faster-RCNN model has a detection precision of 0.781, a recall of 0.691, and an F1-score of 0.733. The comparison results demonstrate that YOLOv11 outperforms both YOLOv8 and Faster-RCNN in all evaluation metrics, with a detection precision improvement of approximately 3.8% compared to YOLOv8 and approximately 6.7% compared to Faster-RCNN; the F1-score improvements are 6.4% and 8.7%, respectively. The experimental results indicate that YOLOv11 achieves better detection performance in visual attribute detection tasks and has stronger target recognition capabilities in construction safety-related scenarios. As shown in Figure 16.

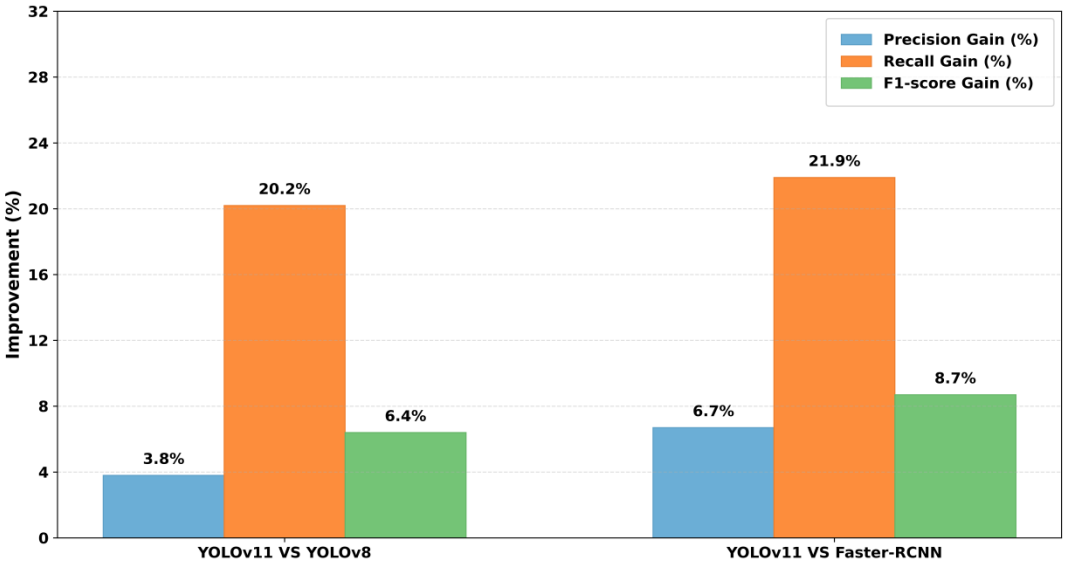

Figure 16. Comparison of performance indicators between YOLOv11 and the baseline model.

This study employs a single-label annotation strategy, assigning only the most significant risk attribute to each instance to eliminate semantic conflicts and enhance the model's identification of dominant instability modes. As shown in Figure 17, experiments demonstrate that the model can characterize visually dominant risks, and its prediction logic highly relies on visual cues such as geometric shape and support relationships, rather than simple category statistics. This indicates that the model has successfully learned state-related representations and possesses a deep understanding of structural physical instability states.

Figure 17. Visual properties of the same object in different spaces.

For the visual attribute detection task, to compare the performance of the YOLOv11 model on the VALD dataset, visual attribute labels were annotated on the existing SODA dataset, and the YOLOv11 model was trained for the visual attribute detection task. The results, shown in Figure 18, indicate that the model achieved a detection precision of 76.6% mAP@50, a recall of 87.00%, and an F1 score of 0.75. Comparing the training and validation results on existing datasets, YOLOv11 achieved a detection precision of 84.8% mAP@50 on the VALD visual attribute detection dataset, an improvement of 8.2%; a recall of 91%, an improvement of 4%; and an F1 score of 0.82, an improvement of 0.07. These results demonstrate that the visual attribute detection model performs better on the VALD dataset, which is constructed around specific object-hitting scenarios.

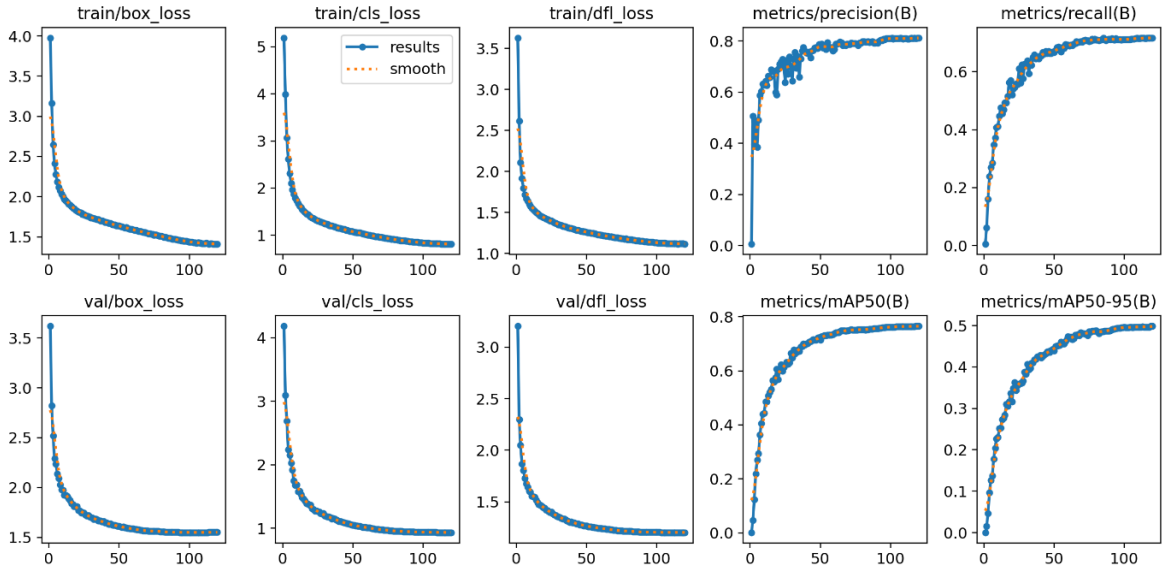

Figure 18. The training and validation results of the YOLOv11 model on the attribute detection task of the SODA dataset.

Experimental results show that YOLOv11 performs well in object detection and visual attribute detection tasks on the VALD dataset. Compared with previous models [46], YOLOv11 has achieved significant improvements in multiple evaluation metrics, especially in complex scenes and high-density target environments, demonstrating higher detection accuracy and robustness. In addition, the visualization results of the attention mechanism further verify that the method can effectively mine deep semantic information and achieve accurate identification of unsafe states of objects, providing reliable technical support for automated safety monitoring of construction sites.

4.5. Visual unsafety detection of objects

Visual attribute detection is fundamental to risk calculation at construction sites, and its accuracy directly impacts the reliability of subsequent risk assessments. To verify the generalization ability of the method in real-world environments, this paper uses a YOLOv11 visual attribute network model to identify unsafe object states based on actual construction site images. Even under highly unstructured conditions such as haphazardly stacked materials, severe occlusion, frequent lighting changes, and complex background structures, the model can still stably identify risky objects such as “hanging hooks,” “unsecured materials,” and “unstable stacked components.” This demonstrates that the model not only learns the appearance features of objects but also masters “hazard attribute patterns” that can be transferred across scenarios, maintaining consistent recognition under different construction sites, viewing angles, and lighting conditions, providing reliable input for risk assessment.

As shown in Figure 19. The detection results show that the model accurately highlights potentially hazardous targets regardless of whether they are large components at close range, small targets at high altitudes or distant locations, or against different backgrounds such as dense steel reinforcement or open ground. This good generalization ability allows subsequent risk calculations to perform quantitative inference based on stable visual input, avoiding fluctuations or misjudgments caused by scene changes. Overall, the robust performance of visual attribute detection significantly improves the comprehensiveness

of hazard identification and the reliability of early warning, enabling the system to adapt to real construction scenarios across multiple sites, stages, and environments, and achieving closed-loop support from visual perception to safety decision-making.

Figure 19. Visual attribute detection results.

5. Case study: object-striking accident scenario

This section uses visual attribute detection results and construction site risk calculation methods as a typical application case scenario, taking object impact as an example, to realize spatial and temporal risk assessment and data visualization at the construction site, demonstrating the application effect of the risk assessment system in construction safety management.

For specific engineering scenarios, the specific expression for risk calculation methods at construction sites is difficult to determine. Furthermore, since this stage aims to verify the feasibility of visual attribute detection results in risk quantification, it was decided to use numerical simulation for calculation, quantifying the safety accident risks under multiple hazard sources in construction scenarios. In the risk assessment stage, visual attribute detection results are not merely triggering conditions, but rather serve as input information for the risk assessment model in risk probability calculation. Specifically, to clarify the relationship between visual attribute detection and numerical risk simulation, the detection confidence was introduced into the triggering probability of each hazard source. For the i-th hazard source, the baseline triggering probability is p_i^0, the visual attribute detection confidence is c_i, and the weather influence factor is w_i. The adjusted triggering probability is defined as:

$$p_i = p_i^0 \cdot c_i \cdot w_i \tag{9}$$

where c_i represents the model confidence in identifying the visual unsafety attribute. ByteTrack was used to extract the real-time worker position. If the distance between the worker and the hazard

source,$d(x_t, h_i)$ is less than or equal to the influence radius r_i, the hazard source contributes to the risk at the current time step; otherwise, its contribution is set to zero:

$$R_i(t) = \begin{cases} p_i, d(x_t, h_i) \le r_i \\ 0, d(x_t, h_i) > r_i \end{cases} \quad (10)$$

When multiple hazard sources exist at the same time, the overall risk is calculated as:

$$R(t)=1-\prod_{i=1}^{n}(1-R_i(t)) \quad (11)$$

The system constructs a two-dimensional discrete model of the construction plane with specific dimensions. Four types of heterogeneous hazards (impact, rollover, collapse, and fall) are defined, each parameterized by a quintuple (type, location, foundation trigger probability, radius of influence, and weather influence factor). In the preliminary experiment, the worker agent is simulated using a random walk model (initial position and velocity control). Its trajectory within the boundary is generated by a normal distribution perturbation and constrained within the construction plane boundary. In the actual task, the worker coordinates are extracted using the ByteTrack target tracking algorithm. Then, dynamic evaluation is achieved through multi-period stochastic simulation, and the results are output as a CSV data file, recording the conditional probabilities of the four types of risks within each time window. The data format includes time window labels and the corresponding four columns of risk probability values, facilitating subsequent statistical analysis. Table 6 shows the input and output parameters of the risk assessment model. By setting and inputting three types of variables—total parameters, hazard source parameters, and worker parameters—the model can output the risk values and dynamic changes of each safety accident within the construction plane in the spatial and temporal dimensions.

Table 6. Input and output parameters of the risk assessment model.

Parameter Type	Model Input	Model Output
Global Parameters	Construction area size (m)	Risk values of various accidents within the spatial–temporal domain of the construction site and their dynamic evolution
	Simulation duration (s)	
	Probability of extreme weather (‰)	
	Data recording interval (s)	
Hazard Source	Category label	
	Coordinate position	
	Baseline triggering probability (‰)	
	Influence radius (m)	
	Extreme-weather influence factor	
Worker	Coordinate information including initial position and real-time location	

In addition, the system will automatically trigger corresponding alarms based on real-time monitoring results and notify on-site workers and safety management personnel through sound, light, or other means. The risk assessment demonstration will visually present the on-site risk status through heat maps and charts, allowing managers to quickly understand the location of high-risk areas and tasks, thereby concentrating safety management resources on key monitoring of these areas. Figure 20 shows the risk values and risk heat maps for various categories in the pre-experiment.

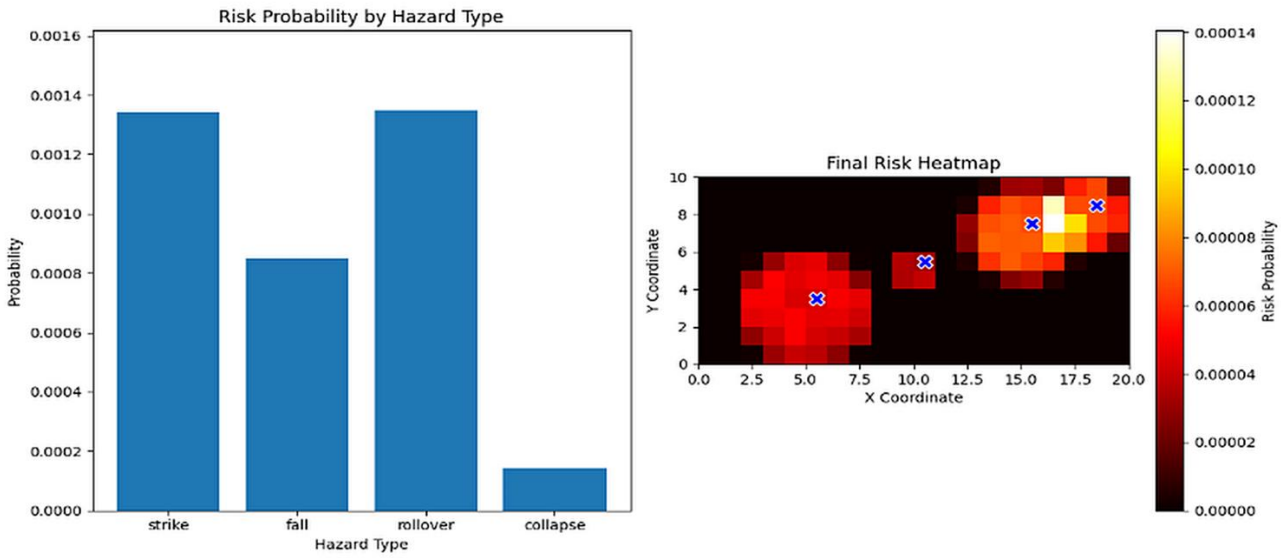

Figure 20. Risk values and risk heatmaps for various categories.

This experiment, based on the visual attribute detection of objects, achieved a quantitative assessment of construction site risks through numerical simulation. Figure 21, a risk heat map, presents the distribution of regional hazard levels formed by the superposition of multiple safety accident risks. Color represents risk intensity, and the radius of influence of object accidents is simultaneously marked in the figure. The results show that high-risk areas at the construction site are mainly concentrated in the hook operation area. Although unsafe conditions (potential hazards) were detected in other areas, no fatal or injury-related safety accidents occurred because the workers' activity trajectories did not intersect with the radius of influence of the object accidents.

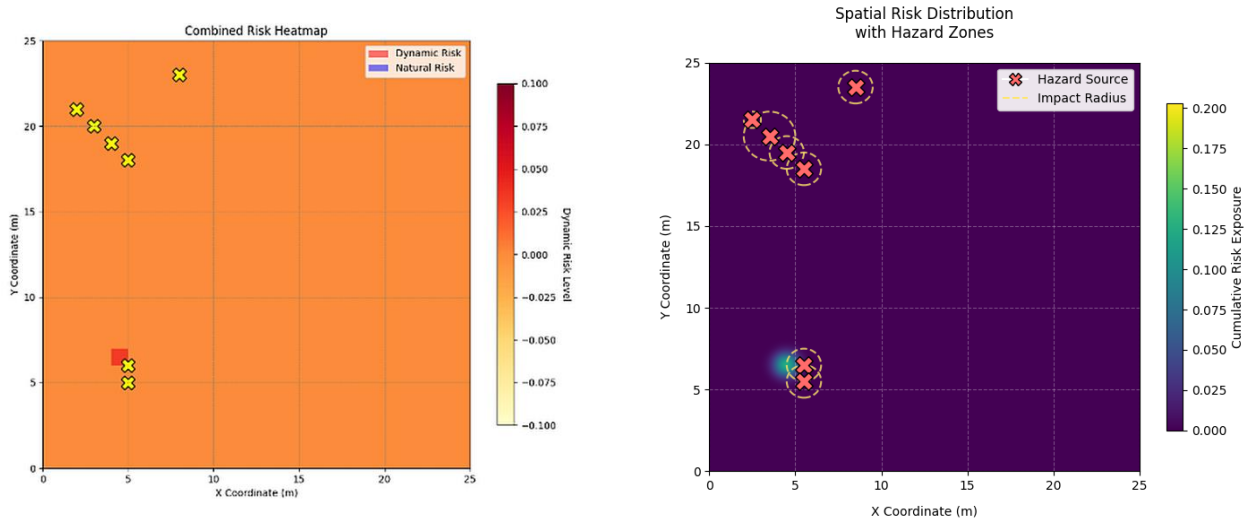

Figure 21. Risk heat map within the construction plane.

6. Discussion

6.1. Technical contributions and effectiveness

In complex construction environments, safety management demands high real-time performance and efficiency in monitoring. Traditional methods relying on manual inspections suffer from limitations such

as high workload, strong subjectivity, and poor real-time performance, making it difficult to support refined safety management, especially in complex tasks such as determining unsafe conditions of objects. Although existing research has attempted to improve the real-time performance of monitoring using artificial intelligence methods such as computer vision, these methods are mostly limited to surface information of objects and lack an understanding of deep semantics. Visual attributes, as a general description of underlying features, are considered key to bridging the semantic gap; however, this field has previously faced problems such as unclear definitions of key visual attributes and a lack of relevant datasets and management methods. Therefore, this study introduces visual attributes into the determination of unsafe conditions of objects, exploring its application potential in this task. Through empirical analysis of object impact scenarios, this study provides new ideas for the intelligent determination of unsafe conditions and construction safety management. Specifically, the introduction of visual attribute detection mainly solves the following key problems:

(1) Regarding the definition of visual attributes of unsafe states of objects: Addressing the lack of a systematic definition of the visual characteristics of "unsafe states of objects" in traditional construction safety research, this paper proposes the higher-order visual attribute concept of "object unsafety." This concept not only characterizes the potential risk state of objects under specific environmental and spatiotemporal conditions but also provides structured semantic support for visual model recognition. By integrating real construction accident cases, this study constructs a fine-grained visual attribute labeling system, achieving a multi-dimensional description of risky objects on construction sites. The results show that the visual attributes of unsafe states can serve as a key mediating variable connecting "perception layer features" and "management layer semantics," providing a theoretical foundation for the subsequent construction of risk perception models.

(2) Regarding the definition of visual attributes of unsafe states of objects: To address the challenge of learning and identifying structural safety-related features from visual data streams, this paper proposes an efficient visual attribute perception model based on the YOLOv11 architecture. Using the VALD dataset for construction and training, experiments validate the model's robustness and generalization ability in complex construction scenarios. The study reveals that the model not only accurately identifies object categories but also effectively captures fine-grained differences in their unsafe states, achieving joint reasoning of behavior-environment-structure at the feature layer. Attention visualization results show that the model's focus expands from local textures to semantically critical regions such as crane booms, load-bearing links, and ground supports, indicating that the model has learned to discern risks from structural relationships rather than local appearance—a crucial skill for complex construction scenarios.

(3) Modeling the Mapping Mechanism Between Visual Perception Features and Engineering Safety Semantics: This paper proposes a semantic-layer mapping mechanism framework to explore how to transform visual perception features into interpretable engineering safety semantics. By modeling the coupling between visual attributes, object relationships, and risk semantics, a cross-layer mapping from low-level pixel features to high-level semantic concepts is achieved. This mechanism can embed the visual features output by the model into a safety knowledge system, thereby enabling interpretable identification and semantic reasoning of unsafe states. This mapping process from perception to semantics not only broadens the application boundaries of visual intelligence in construction safety management but also provides ideas and references for building knowledge-driven safety perception systems in the future.

While YOLOv11's performance on the SODA dataset demonstrates its feasibility in visual attribute detection tasks and its ability to identify desired visual attributes, several issues remain: the scene selection in the dataset is not uniform, and some images suffer from object occlusion, lighting variations, and complex backgrounds, leading to decreased accuracy in visual attribute detection. Furthermore, different construction scenarios may involve diverse accident types, resulting in varying degrees of object hazard.

To address these issues, this study optimized the dataset construction for specific object impact scenarios. For example, it improved the dataset's annotation standards, clearly defining attribute definitions and boundaries to reduce annotation errors. Simultaneously, it optimized the selection of scenarios for the dataset, ensuring consistency of samples within construction scenarios, thereby improving dataset quality and enabling the model to exhibit higher robustness and accuracy on the VALD dataset.

Experimental results demonstrate that constructing a new VALD dataset, collecting more specialized data, and performing high-quality annotation are effective. By increasing scene consistency and the number of precisely annotated samples, especially attribute data in complex scenes, the accuracy of visual attribute recognition can be improved. Furthermore, the automated method for extracting visual attributes from real accident reports ensures that the definition boundaries of each attribute have a mapping space on the image dataset. This helps the model better understand and learn the relationships between various attributes, thereby improving overall recognition accuracy and robustness.

Furthermore, the risk assessment model proposed in this study is currently constructed based on certain rules, and the parameter settings mainly refer to engineering experience and safety specifications. It has not yet been systematically calibrated and validated using real accident data. Although the case study validated the feasibility of the technical approach, the model's early warning accuracy and the rationality of its risk thresholds still need quantitative evaluation through backtracking analysis of real accident cases. Future research will collect more real construction safety accident cases, establish a statistical correlation between risk values and accident occurrences, optimize model parameters, and improve the practicality and reliability of risk assessment.

6.2. Limitations and future research

While the proposed method demonstrates high accuracy and effectiveness in identifying and managing unsafe conditions of objects, there is still room for expansion of its application. Validation on a dataset focusing on object impact scenarios within construction settings shows promising results based on visual attribute detection. However, the effectiveness of visual attribute detection is somewhat limited for image data from other construction scenarios.

At present, the semantic content of visual attributes still has considerable room for expansion. For visual attributes, their definitions directly determine the scope of application of the proposed method. In addition, one important advantage of visual attributes lies in their transferability and scalability: they can be recognized by visual attribute detection models, extracted from accident texts using large language models, and integrated with other safety management frameworks. Future research can construct datasets covering more object types, construction scenarios, and risk situations to further improve the generalization ability of trained models. Meanwhile, the semantic content of visual attributes can be further expanded by developing more fine-grained and implicit visual attribute concepts and detection

frameworks. For complex construction scenes where the same object may involve multiple potential risks, future studies may also explore multi-label visual attribute annotation and detection methods to more comprehensively represent risk co-occurrence. On this basis, further research can be conducted on visual attribute transfer learning and its application in construction safety management, thereby improving the intelligence level of construction safety management.

Meanwhile, future research can further expand the semantic content of visual attributes by introducing more fine-grained and engineering-interpretable attribute concepts. For example, “support instability” can be used as a visual attribute to describe abnormal states of scaffolds, formwork, or temporary support structures, representing potential risks such as tilting, loose connections, or uneven loading. In addition, “pathway occupation” can be introduced to describe the obstruction of worker movement paths caused by material stacking, equipment placement, or component storage. Such attributes can further strengthen the correspondence between visual detection results and construction safety management semantics.

7. Conclusion

To improve the efficiency and objectivity of construction safety management, this paper proposes a method for detecting unsafe object states based on the YOLOv11 visual attribute network and constructs a dynamic risk assessment system. The proposed approach addresses the limitations of traditional manual inspections, such as strong subjectivity and difficulty in identifying potential risks.

First, unsafe object states are clearly defined, and an application framework of visual attributes for construction safety is established. Visual attribute labels related to object safety are extracted from accident reports using the large language model GLM-4, forming a structured visual attribute label set through text processing and semantic disambiguation. Based on an object impact scenario, a VALD dataset is constructed, containing 11,980 images, 16 object categories, and four visual attributes, and is used to train and validate the YOLOv11 model. Experimental results demonstrate that the proposed method achieves satisfactory performance in terms of precision, recall, and F1 score.

Furthermore, a risk assessment system is developed using the visual attribute detection results, and a case study verifies the feasibility and effectiveness of the proposed method in real-word construction safety management scenarios. Overall, this study provides a practical visual inspection solution for construction safety management and shows potential for extension to other construction scenarios. This paper uses a large language model for translation and polishing.

Data availability statement

No supplementary or additional data were generated in this study.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools solely for the purpose of improving language expression and readability. Specifically, the authors used ChatGPT/Grammarly for language polishing, drafting, rewriting, and idea development in limited sections of the manuscript. The authors take full responsibility for the entire content of the manuscript, and confirm that all scientific data, analyses, conclusions, and intellectual contributions remain entirely their own.

Acknowledgments

The authors would like to acknowledge the support by National Science Foundation of China (52308314), the support by Guangdong Basic and Applied Basic Research Foundation (2023A1515030169) and the support by Guangzhou Science and Technology Project (2025A04J5216).

Authors' contribution

Conceptualization, Y.D.; methodology, Z.D. and H.Z.; software, Z.D. and H.Z.; validation, Z.D. and H.Z.; formal analysis, Z.D., H.Z. and V.J.G.; investigation, Z.D. and H.Z.; resources, Y.D.; data curation, H.Z.; writing—original draft preparation, Z.D. and H.D.; writing—review and editing, Y.D., H.D. and V.J.G.; visualization, Z.D.; supervision, Y.D., H.D. and V.J.G.; project administration, Y.D.; funding acquisition, Y.D. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Hengyun Zhang is employee of CCCC Fourth Harbor Engineering Institute Co., Ltd. The author declares that this affiliation does not influence the design, analysis, interpretation, or presentation of the results reported in this manuscript. The author declares no other conflicts of interest.

References

[1] Fang W, Love PE, Ding L, Xu S, Kong T, *et al.* Computer vision and deep learning to manage safety in construction: matching images of unsafe behavior and semantic rules. *IEEE Trans. Eng. Manage.* 2023, 70(12):4120–4132.

[2] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, *et al.* Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* 2017, 123(1):32–73.

[3] Liu L, Wiliem A, Chen S, Lovell BC. What is the best way for extracting meaningful attributes from pictures? *Pattern Recognit.* 2017, 64:314–326.

[4] Zhang C, Wu T, Zhang Y, Zhao B, Wang T, *et al.* Deep semantic-aware network for zero-shot visual urban perception. *Int. J. Mach. Learn. Cybern.* 2022, 13(5):1197–1211.

[5] Wu X, Li Y, Long J, Zhang S, Wan S, *et al.* A remote-vision-based safety helmet and harness monitoring system based on attribute knowledge modeling. *Remote Sens.* 2023, 15(2):347.

[6] Alipour M, Harris DK. Increasing the robustness of material-specific deep learning models for crack detection across different materials. *Eng. Struct.* 2020, 206:110157.

[7] Guo BH, Zou Y, Fang Y, Goh YM, Zou PXW, *et al.* Computer vision technologies for safety science and management in construction: A critical review and future research directions. *Saf. Sci.* 2021, 135:105130.

[8] Chi S, Caldas CH. Automated object identification using optical video cameras on construction sites: automated object identification using optical video cameras. *Comput.-Aided Civ. Infrastruct. Eng.* 2011, 26(5):368–380.

[9] Soltani MM, Zhu Z, Hammad A. Automated annotation for visual recognition of construction resources using synthetic images. *Autom. Constr.* 2016, 62:14–23.

[10] Duan R, Deng H, Tian M, Deng Y, Lin J. SODA: a large-scale open site object detection dataset for deep learning in construction. *Autom. Constr.* 2022, 142:104499.

[11] Yang Q, Shi W, Chen J, Lin W. Deep convolution neural network-based transfer learning method for civil infrastructure crack detection. *Autom. Constr.* 2020, 116:103199.

[12] Lowe DG. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, Corfu, Greece, September 20–27, 1999, pp. 1150–1157.

[13] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, USA, June 20–25, 2005, pp. 886–893.

[14] Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* 2008, 110(3):346–359.

[15] Kim H, Kim H, Hong YW, Byun H. Detecting construction equipment using a region-based fully convolutional network and transfer learning. *J. Comput. Civ. Eng.* 2018, 32(2):04017082.

[16] Li Z, Zhang A, Han F, Zhu J, Wang Y. Worker abnormal behavior recognition based on spatio-temporal graph convolution and attention model. *Electronics* 2023, 12(13):2915.

[17] Yang J, Arif O, Vela PA, Teizer J, Shi Z. Tracking multiple workers on construction sites using video cameras. *Adv. Eng. Inform.* 2010, 24(4):428–434.

[18] Park MW, Brilakis I. Continuous localization of construction workers via integration of detection and tracking. *Autom. Constr.* 2016, 72:129–142.

[19] Teizer J, Vela PA. Personnel tracking on construction sites using video cameras. *Adv. Eng. Inform.* 2009, 23(4):452–462.

[20] Li F, Chen Y, Hu M, Luo M, Wang G. Helmet-wearing tracking detection based on StrongSORT. *Sensors* 2023, 23(3):1682.

[21] Fang Q, Li H, Luo X, Ding L, Rose TM, *et al.* A deep learning-based method for detecting non-certified work on construction sites. *Adv. Eng. Inform.* 2018, 35:56–68.

[22] Chi S, Caldas CH. Image-based safety assessment: automated spatial safety risk identification of earthmoving and surface mining activities. *J. Constr. Eng. Manage.* 2012, 138(3):341–351.

[23] Zhu Z, Park MW, Koch C, Soltani M, Hammad A, *et al.* Predicting movements of onsite workers and mobile equipment for enhancing construction site safety. *Autom. Const.* 2016, 68:95–101.

[24] Farha YA, Richard A, Gall J. When will you do what?—Anticipating temporal occurrences of activities. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA, June 18–22, 2018, pp. 5343–5352.

[25] Smeulders A, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 2000, 22(12):1349–1380.

[26] Schwartz G, Nishino K. Automatically discovering local visual material attributes. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, June 2015, pp. 3565–3573.

[27] Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, *et al.* Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision (ECCV 2010)*, Heraklion, Greece, September 5–11, 2010, pp. 15–29.

[28] Han Y, Yang Y, Ma Z, Shen H, Sebe N, *et al.* Image attribute adaptation. *IEEE Trans. Multimedia* 2014, 16(4):1115–1126.

[29] Gibson JJ. *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin, 1966.
[30] Yunlong LI, Linbo Q, Longmei HA, Yuchen W. Survey on visual affordance research. *Comput. Eng. Appl.* 2022, 58(18):1.
[31] Norman DA. *The Design of Everyday Things*. New York: Basic Books, 2013.
[32] Tomé A, Kuipers M, Pinheiro T, Nunes M, Heitor T. Space–use analysis through computer vision. *Autom. Constr.* 2015, 57:80–97.
[33] Fang Q, Li H, Luo X, Ding L, Luo H, *et al.* Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. *Autom. Constr.* 2018, 93:148–164.
[34] Luo H, Wang M, Wong P, Cheng J. Full body pose estimation of construction equipment using computer vision and deep learning techniques. *Autom. Constr.* 2020, 110:103016.
[35] Wang X, Zhu Z. Vision-based hand signal recognition in construction: a feasibility study. *Autom. Constr.* 2021, 125:103625.
[36] Stilla U, Xu Y. Change detection of urban objects using 3D point clouds: a review. *ISPRS J. Photogramm. Remote Sens.* 2023, 197:228–255.
[37] Piyathilaka L, Kodagoda S. Affordance-map: mapping human context in 3D scenes using cost-sensitive SVM and virtual human models. In *Proceedings of the 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Zhuhai, China, December 6–9, 2015, pp. 2035–2040.
[38] Roy A, Todorovic S. A multi-scale CNN for affordance segmentation in RGB images. In *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, Amsterdam, The Netherlands, October 11–14, 2016, pp. 186–201.
[39] Nguyen A, Kanoulas D, Caldwell DG, Tsagarakis NG. Detecting object affordances with convolutional neural networks. In *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, Republic of Korea, October 9–14, 2016, pp. 2765–2770.
[40] Han K, Wang Y, Chen H, Chen X, Guo J, *et al.* A survey on visual transformer. *arXiv* 2020, arXiv:2012.12556
[41] Deng H, Fu K, Yu B, Li H, Duan R, *et al*, Enabling high-level worker-centric semantic understanding of onsite images using visual language models with attention mechanism and beam search strategy. *Buildings* 2025,15(6):959.
[42] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, USA, December 4–9, 2017, pp. 5998–6008.
[43] Ali M L, Zhang Z. The YOLO Framework: a comprehensive review of evolution, applications, and benchmarks in object detection. *Computers* 2024, 13(12):336.
[44] Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancement. *arXiv* 2024, arXiv:2410.17725.
[45] Ghosh A. YOLO11: redefining real-time object detection. 2024, Available: https://learnopencv.com/yolo11/ (accessed on 15 January 2025).
[46] Wang X, Cui W, Tao Y, Shi T. Research on deep learning-based object detection algorithm in construction sites. *Eng. Lett.* 2025, 33(1):1.