GPT-based Logic Reasoning for Hazard Identification in Construction Site using CCTV Data

Dai Quoc Tran[1], Yuntae Jeon[2], Minsoo Park[3], and Seunghee Park[4]

[1]Global Engineering Institute for Ultimate Society, Sungkyunkwan University, South Korea
[2]Department of Global Smart City, Sungkyunkwan University, South Korea
[3]Sungkyun AI Research Institute, Sungkyunkwan University, South Korea
[4]School of Civil, Architectural Engineering and Landscape Architecture, Sungkyunkwan University, South Korea

daitran@skku.edu, jyt0131@g.skku.edu, pms5343@skku.edu, shparkpc@skku.edu

Abstract -

The applications of deep learning-based robust surveillance are vital for improving safety at construction sites, with closed-circuit television (CCTV) systems serving as a pivotal tool in achieving this goal. Despite the recent progress in state-of-the-art deep learning models, the task of hazard identification remains a persistent difficulty due to the complexity of the working environment. This paper presents a novel end-to-end pipeline termed *"Image-to-Hazard"* that aims to address the disparity between individual single-model predictions. The pipeline incorporates multimodal inputs and uses logical reasoning to establish connections. The pipeline integrates a model based on GPT architecture from the OpenAI API, encompassing various tasks such as detection, depth estimation, danger identification, and logical reasoning. Firstly, an actual video dataset was obtained from construction sites and annotated. Subsequently, customized object detection models were trained and optimized. Afterward, a thorough extraction of visual features was conducted by utilizing pre-trained models for tasks such as semantic segmentation and depth estimation. Subsequently, prompt engineering was conducted to seamlessly include the input of visual feature information, and these structures were integrated into OpenAI GPT-based models to enhance their capacity for logical reasoning. As a result, a proposed approach showed its robustness in integrating the GPT-based model and vision model for automated hazard identification and management at construction sites.

Keywords -

GPT; Logic Reasoning; Hazard Identification; CCTV; Safety Management

1 Introduction

Construction sites have constantly been identified as one of the most hazardous work environments worldwide [1]. According to the construction accident data from the Republic of Korea, there were an average of 538 deaths per year between 2010 and 2020 [2]. These fatalities accounted for around 27.9% of all accidents across ten different industries, making it the industry with the highest accident rate [2]. In 2021, the Occupational Safety and Health Administration (OSHA) [3] recorded a fatality rate of 12.3 deaths per 100,000 full-time equivalent workers in the construction industry. It is essential to have a thorough grasp of the various dangers involved, including falls, electrocution, and exposure to hazardous materials, to improve worker safety. Closed-circuit television (CCTV) cameras are widely acknowledged as important tools for monitoring safety on construction sites. CCTV cameras facilitate the identification of possible dangers, surveillance of adherence to safety protocols, and examination of incidents by capturing visual evidence of actions. However, there is not much research on synchronizing the multi-visions model with NLP for logical reasoning. With the advancement of the generative pre-trained transformer (GPT) model currently, the GPT model acts as a strong tool for combining knowledge and providing insight into a given scenario [4]. Manual monitoring of large-scale CCTV channels in construction applications can result in the failure to identify potential hazards. The GPT model enables the identification of hazards not only through visual information but also from a linguistic perspective. Subsequently, the safety report can be generated automatically. This study aims to bridge the gap between separate single-model predictions using multimodal connective logical reasoning, whereby *Image-to-Hazard*, a novel end-to-end pipeline that uses a GPT-based model from the OpenAI API, is proposed for detection, depth estimation, hazard identification, and logic reasoning for safety monitoring at construction sites. The contributions of this study can be summarized as follows:

1. A novel end-to-end pipeline using a GPT-based model from the OpenAI API is proposed for detection, segmentation, depth estimation, hazard identification, and logical reasoning for safety monitoring at construction sites.

2. Detection models are trained on real-life CCTV

datasets.

2 Literature Review

Deep learning (DL) has been widely employed in diverse fields such as computer vision, natural language processing, and robotics. Furthermore, deep learning methods have demonstrated resilience in improving safety surveillance at construction sites [1]. Integrating deep learning algorithms into CCTV systems enhances the ability to identify and mitigate potential safety issues at building sites [5]. Deep learning algorithms provide a range of approaches to improve safety monitoring using CCTV systems. As stated by [6], 400,000 photos are taken throughout the construction stages. Furthermore, CCTV systems are built at nearly every construction site and are consistently employed to monitor the situation. Safety management, progress tracking, and quality inspections can benefit from the use of high-resolution photos, videos, and algorithms that rely on deep learning techniques. Object recognition and tracking have garnered considerable interest among various deep-learning approaches [7, 8]. Deep learning-based object detection can be categorized into two main types: one-stage and two-stage detectors [9]. Notable architectures in this field include Fast R-CNN (fast region-based convolutional neural network) [10] and YOLO (you-only-look-once) [11]. These two object detection methods are used based on unique research aims, taking into account the trade-off between accuracy and inference time. Real-time detection is essential in monitoring construction site safety using CCTV footage, requiring a careful balance between accuracy and inference speed. Some common applications of deep learning in construction site, such as: Nath *et al.* employed the YOLO architecture [11] to create three distinct models for identifying worker personal protective equipment (PPE). The constructed models were verified using a dataset that was created specifically for this purpose, called *Picto-v3*. The technique, designed using the YOLO architecture for one-stage detection, exhibited an inference speed of around 13 frames per second (FPS), which is close to real-time. Following that, several studies have endeavored to improve the efficacy of the YOLO framework in monitoring safety at building sites. Park and colleagues [12] enhanced the YOLO architecture by incorporating an attention mechanism, resulting in SOC-YOLO. This modification aimed to enhance the detection accuracy of small and overlapping worker images. SOC-YOLO is a technique that employs advanced methods such as distance intersection over union (DIoU), non-maximum suppression (NMS), weighted triplet attention, expansion feature levels, and a soft pool. It has proven to be effective in detecting small and overlapping targets (workers) at complex construction sites. When conducting studies with SOC-YOLO, the average precision (AP) for small items showed an increase from 67.52% to 73.88% mAP for minute objects, and from 74.56% to 77.57% for overlapping objects. This demonstrates the practicality of SOC-YOLO in safety monitoring. Currently, the research paradigm is shifting from single-model to multimodal development, highlighting the necessity of considering the context of a given image or video alongside object detection, as quantifying objects cannot rely solely on a single model. By incorporating various types of models, including detection, segmentation, and depth estimation, it is possible to develop a context-aware model that encompasses a broader understanding of the visual data. Chern *et al.* [13] proposed modularized context-aware safety monitoring for fall accidents. Using the CCTV far-field monitoring dataset YUD-COSAv2, the approach involves training detection models and subsequently combining them with segmentation and depth map models to create a context-aware model. This model takes the scenario into account rather than relying solely on detection results, enabling a more informed and comprehensive inference of results. This model was able to differentiate workers by height and could apply different PPE compliance rules, whereby average precisions of 78.50% and 86.22% were obtained, respectively. Overall, these studies focus on single-model development, and the improvement is derived from improved datasets, customized models, or ensemble models. However, a multimodal understanding of scenarios from input images or videos is still lacking. Chen *et al.* [14] developed a framework consisting of three modules for: (1) automated processing of regulatory rules and transformation of sentences into computable graph structure representations; (2) combining two object detection and pose estimation models to represent scene information; and (3) automated reasoning of hazard notification from the above two output graphs. The proposed framework, which extracts safety rules through feature engineering, was effective in identifying individuals operating a grinder. Khan et al. [15] identified mobile scaffolding and workers using Mask R-CNN. This study proposed a correlation-based approach for mobile scaffold safety monitoring and the detection of unsafe worker behaviors. Mask R-CNN was used to classify and segment worker tasks and an object correlation detection (OCD) module was used to detect unsafe behaviors. Subsequently, safety rules were used to determine whether the scenario was safe based on the detection results. The test results exhibited 85-97% precision and recall for class-1 (safe behavior) and 91-65% precision and recall for class-2 (unsafe behavior). An overall accuracy of 86-96% confirms the Mask R-CNN-based OCD module's applicability in the construction environment.

3 Proposed Approach

This section describes the deep learning and self-developed models used in the proposed approach. The proposed approach is illustrated in Figure 1 and comprises four major phases: data acquisition, model inference, synchronization, and logic reasoning. During the data acquisition phase, an actual CCTV dataset is acquired and labeled to train the detection model. The input image is then used for inference in multiple types of deep learning models, including object detection, and depth-map estimation. The information from the detected objects is then synchronized using multimodal synchronization modules. Finally, the GPT-based model is used for hazard-scenario logic reasoning using the developed auto-generating prompt structure. Each phase of the proposed approach is essential to contributing to the overall objective. The data-acquisition phase ensures that the model has access to a representative, high-quality dataset. The inference phase enables the model to recognize and categorize the objects within the input image. The synchronization phase ensures that the information regarding the detected objects is consistent across models. The reasoning phase enables the model to generate a logic-based hazard scenario that can be utilized to prevent catastrophic events. The proposed approach is a promising novel method for hazard detection and prevention with the potential to improve worker and property safety at construction sites.

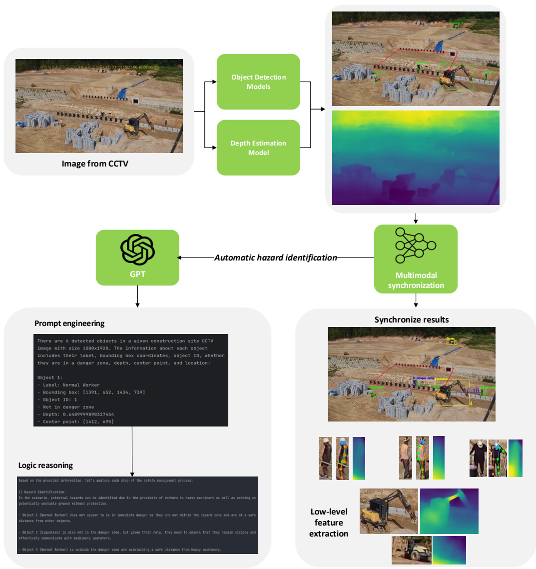

Figure 1. Image-to-Hazard.

3.1 Object Detection Modules

A construction site is a complex environment containing numerous objects that must be detected and analyzed. The focus of this study is to classify and identify primary target objects on a construction site, including construction vehicles, workers, and signalman. The specific class names are listed below.

1. Construction vehicle with eight classes *(Excavator, Dozer, Forklift, Dump Truck, Mixer Truck, Cargo Truck, Scissor Lift, Crane).*

2. Signalman detection with two classes *(Worker, Signalman).*

The exploratory data analysis is presented in Figure 2. Numerous detection models can be used to detect these objects. In this study, the state-of-the-art one-stage object detector Yolov8X was used to train and finetune the detection model. To enhance the diversity of the training samples, the default data for training all the models in this study were augmented using Yolov8. The augmentation technique used was "mosaic," which combines four different training images into one in a mosaic-like pattern. This approach helps to improve the variation and representation of the training data. Subsequently, a unique augmentation technique specific to YOLOv5, known as 'CopyPaste', was employed to randomly select an object from one image and paste it onto another, thereby enhancing the complexity of the image data. The model then applied random affine transformations, including scaling, shearing, and rotation, to the images. The ranges of these transformations were predefined. Following these steps, another YOLOv5 specific technique called 'MixUp' was used to blend two images and their labels to generate a new, more complex image. Following 'MixUp,' the mask information was removed from the data, and then a series of augmentations from the Albumentations library were performed. These transformations include techniques such as Blur, MedianBlur, conversion to grayscale, and contrast-limited adaptive histogram equalization, each applied with a certain probability. The experiment was conducted using CentOS Linux 8 with two GTX A6000 graphics processor units, each with 48 GB of memory. The model was developed using PyTorch, based on the MMYolo [16] library. The mean Intersection over Union (mIoU) was used to evaluate the model. mIoU is a frequently used metric in computer vision for evaluating the efficacy of object detection and segmentation algorithms. This metric quantifies the amount of overlap between the predicted bounding boxes (or segmentation masks) and ground truth bounding boxes (or masks), indicating the accuracy of model prediction.

Table 1 displays model detection performance for construction vehicles and signalman datasets. In this study, the mIoUs of small, medium, and large objects are considered in the analysis. Comparative studies with other SOTA models were not considered because the objective

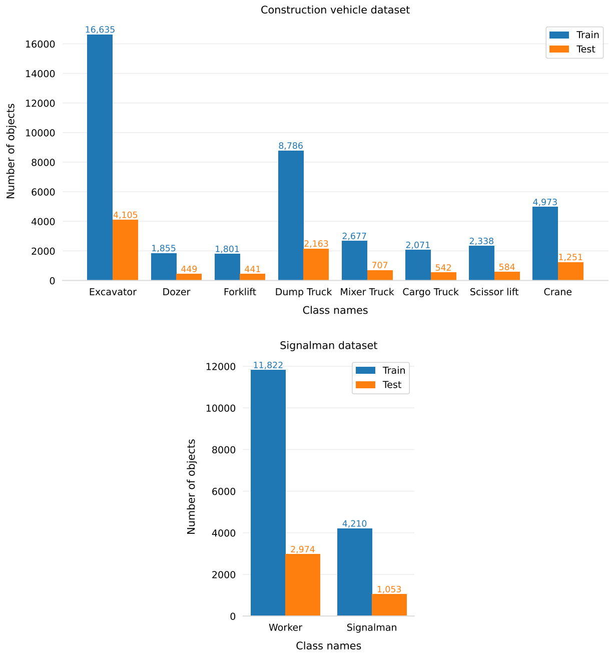

Figure 2. Dataset class distribution.

of this research was to develop end-to-end multimodal logic reasoning. Objects that occupy 0 to 1024 pixels (32 × 32 pixels) are considered 'small'; objects that occupy between 1024 and 9216 pixels (32 × 32 to 96 × 96 pixels) are considered 'medium'; and objects that occupy more than 9216 pixels (96 × 96 pixels or larger) are considered 'large.'

The construction vehicle detection model revealed a significant size-dependent performance, with smaller objects proving to be more challenging for the model to detect. The model exhibited a precision of 0.377 for small objects, which increased substantially to 0.763 for medium objects and peaked at 0.921 for large objects. This steep increase suggests that the model is particularly adept at detecting larger construction vehicles, possibly because of their distinct high-contrast features that are easier to discern at larger scales. The signalman detection model demonstrated comparatively consistent performance across object sizes, with precision measures ranging from 0.652 for small objects to 0.765 for medium objects, and then slightly increasing to 0.838 for large objects. Unlike the other models, this model did not show a steep size-dependent performance gradient, which may be attributed to the distinct characteristics of the signalmen, making them easier to detect irrespective of their size. The discrepancies in model performance across different sizes and categories indicate strengths and limitations. These models are currently more effective in detecting larger objects. Smaller objects prove to be a common challenge in visualization, possibly because of their indistinguishable features at smaller scales. This highlights the need for further research to enhance the precision of object detection models, particularly for small- and medium-sized objects.

Table 1. Detection model performance.

Model	**IoU=0.5:0.95↑**		
	Small	**Medium**	**Large**
Construction vehicle	0.377	0.763	0.921
Signalman	0.652	0.765	0.838

3.2 Human Pose Estimation

The overlapping area of the upper body precludes the exact location of the worker from being extracted if only the bounding box from the object detection model is utilized. In numerous instances, the bounding box of the detected worker does not encompass the entire body; consequently, the ankle coordinates cannot be estimated. For estimating the ankle midpoint of the detected worker, the well-known pre-trained model HRNet [17] utilizes keypoint extraction. Figure 3 presents examples of the algorithm. The detected person bounding boxes are extracted first, followed by the application of a pose estimation model, to estimate all the key points of the body. In some instances, ankle points cannot be detected because of overlapping objects, and the ankle midpoint is estimated using an estimation ratio between upper body keypoint.

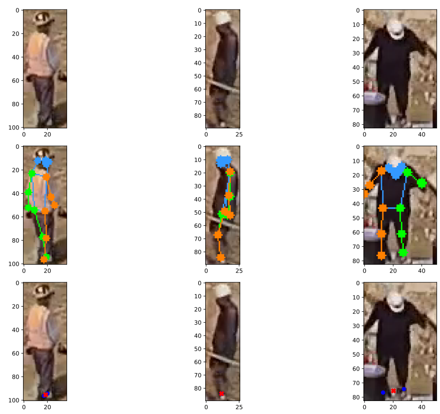

Figure 3. Actual dataset.

3.3 Depth Estimation Module

Objects detected by a detection model can be extracted from image coordinates without considering spatial information. A monocular CCTV depth-estimation map can contribute to spatial information analysis. Therefore, the trained MiDaS [18] model was used to extract depth data

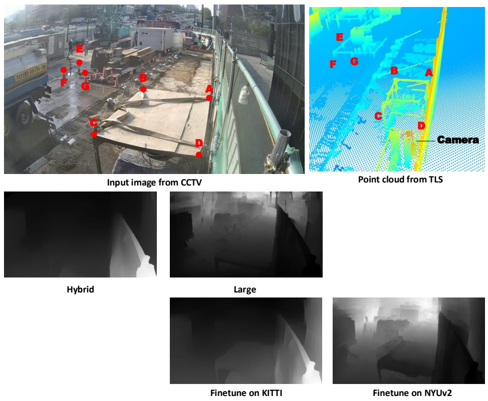

Figure 4. Comparison of depth map estimation results using point cloud.

from the CCTV video. Using the x and y coordinates in the image and depth z, the Euclidean distance in three dimensions can be estimated. As shown in Figure 4, the depth estimation results from monocular CCTV were then compared with the actual point cloud data from terrestrial laser scanning (TLS). The four trained models were used to compare the results. The MiDaS model finetuned on NYUv2 dataset robustly estimated depth using the monocular CCTV image. In this study, the trained MiDaS, finetuned on the NYUv2 dataset, was selected to estimate the depth map at the construction site.

3.4 Prompt Engineering

For logical reasoning, GPT is used. Currently, GPT-based models perform phenomenally in the natural language processing (NLP) domain [4]. GPT-based models are a type of large-language model (LLM) trained using a technique called generative pre-training. GPT-based models are typically trained on a massive dataset of text and can be used for a variety of NLP tasks, such as text generation, translation, and question-answering. With this large-scale dataset, the model can understand extremely complex text structures and provide meaningful answers. Therefore, in this research, GPT-based models were utilized for the analysis of hazard scenarios, using the extracted multimodal visual information.

An important technique is *"Prompt engineering"*, which is the process of designing and optimizing inputs (prompts) to a model, such as GPT, in a way that maximizes the quality and relevance of its outputs. This is a crucial part of using AI language models because the way a question or task is framed can significantly affect the model's response. Object detection models identify and classify objects within images typically by providing a bounding box around the detected object and a label indicating the object. This output can be used as an input prompt for a language model such as GPT for further analysis, reasoning, or narrative generation.

An ontology model can be used to define entities at a construction site (such as workers, vehicles, and equipment), their properties (location, PPE status, and moving direction), and the relationships between these entities. Once the ontology model is developed, it can be used to generate prompts for the GPT model.

Given:

- D : the set of detected objects.
- d_i : a detected object in D, which is a tuple $(label, bbox, object\_id, in\_danger\_zone, depth, center\_point)$.
- P : the set of object distances.
- p_i : a pair of objects in P, which is a tuple $(object1, object2, distance)$.

We define:

- A function $f(d_i)$ to generate the description for a detected object d_i.
- A function $g(p_i)$ to generate the description for a pair of objects p_i.

Then, the function $h(D, P)$ used to generate the sentence for the GPT model is defined as

$$h(D, P) = \bigcup_{d_i \in D} f(d_i) \cup \bigcup_{p_i \in P} g(p_i)$$

Algorithm 1 illustrates the conversion of visually detected features into a GPT-based model.

Using the aforementioned ontology models, the generated prompt is then input into the GPT-based model for logical reasoning.

4 Experiment and Discussion

This Section presents sample hazard scenarios and analyzes the input and output of logical reasoning from the GPT-based model. Currently, OpenAI supports API for GPT-based models up to GPT4.

As shown in Figure 5 and Figure 6, after extracting object information using detection models, such as draw a danger area, provide a reason for danger area, estimate the Euclidean distance in 3D between "Normal Worker" and "Vehicle," extract depth and segmentation map. The GPT model infers the hazard scenarios in this input image based on three main questions, as follows:

1. Identify potential hazards between objects.
2. Detail the specific risks associated with these potential hazards.

Algorithm 1 Generating descriptive sentences for object detection

```
procedure FORMAT_OBJECT_INFO(obj)
    if not obj['in_danger_zone'] then
        danger_zone_status ← 'Not in'
    else
        danger_zone_status ← 'In'
    end if
    info ← format_info(obj, danger_zone_status)    ▷ Formatting the information about the object
    if obj['is_signalman'] then
        append_signalman_info(info, obj)    ▷ Appending signalman information if present
    end if
    return join_info(info)    ▷ Joining all information into a single string
end procedure
procedure FORMAT_DISTANCES(distances)
    descriptions ← format_pairwise_distances(distances)    ▷ Formatting the pairwise distances
    append_safe_distance_info(descriptions)    ▷ Appending safe distance information
    return join_descriptions(descriptions)    ▷ Joining all descriptions into a single string
end procedure
procedure GENERATE_SENTENCE(detected_object_info_danger, object_distances)
    object_descriptions ← format_all_objects_info(detected_object_info_danger)    ▷ Formatting all object information
    num_objects ← count_objects(detected_object_info_danger)
    if object_distances is not empty then
        distances_text ← format_distances(object_distances)
        base_text ← format_base_text(num_objects, objects_text, distances_text)
    else
        base_text ← format_base_text_no_distances(num_objects, objects_text)
    end if
    additional_text ← format_additional_text()
    return base_text + additional_text
end procedure
```

3. Highlight the safety management approach for reducing risks.

Considering the visual information extraction input, the GPT4 model provides feedback regarding hazard scenarios in a given image, as follows:

- Potential hazards: Object 1 (normal worker) and Object 5 (excavator): The normalized distance between Object 1 and Object 5 is 0.159, which is less than the safe distance threshold of 0.2. This indicates that the worker is too close to the excavator. Object 4 (normal worker) and Object 5 (excavator): The normalized distance between Object 4 and Object 5 is 0.275, which passes the safe distance threshold, but is close enough to warrant attention.

- Analyze the "Risks associated with these potential hazards": The primary risk in both of these scenarios is injury due to the close proximity of workers to heavy machinery. These injuries can occur in a variety of ways, such as: (1) Being struck by moving parts of the machinery (e.g., the bucket of the excavator). (2) Being caught between machinery and another object. (3) Slips, trips, or falls due to uneven or unstable ground near the machinery.

- Safety management approach for reducing risks.

Clearly, with multimodal inference, different aspects such as worker type, vehicle name, location, and 3D distance between objects, are used to extract visual information. Thus, the GPT model can perform logical reasoning much more effectively. The construction site is a complex environment. Without a tailored training model and considering all possible hazards, the pretrained single-detection model can meet the challenge of understanding context-aware scenarios.

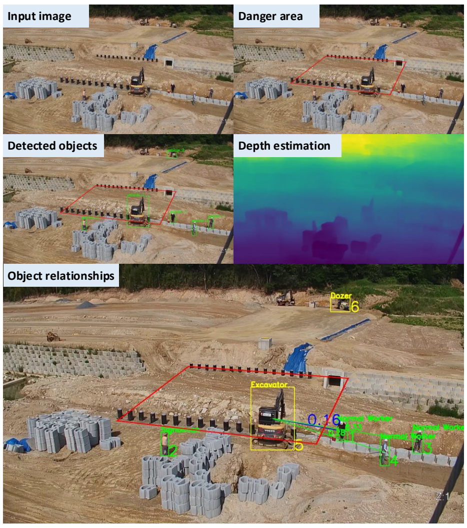

Figure 5. Step-by-step model detection.

In Figure 7 and 8, the report identifies potential hazards and risks in a construction site scenario, indicating close proximity of workers to heavy machinery and an unstable slope. Notably, Object 2 (normal worker) is dangerously close to Object 7 (excavator), with a normalized distance of 0.196, which is less than the safe limit of 0.2. Additionally, Object 3 (normal worker) is within the danger zone, being close to an unstable slope. These situations raise significant risks, including physical injury or fatality for Object 2, from possible contact with the excavator and potential slope collapse hazards for Object 3. Several safety strategies have been proposed to mitigate these risks. These strategies include enforcing safe distances between workers and machinery, particularly for Object 2; evacuating and restricting access to the danger zone around the unstable slope where Object 3 is located; continuous site monitoring with the CCTV system, possibly

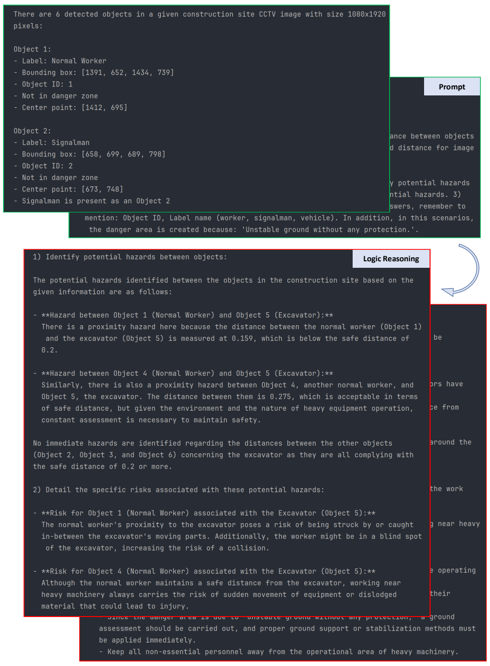

Figure 6. GPT logic reasoning.

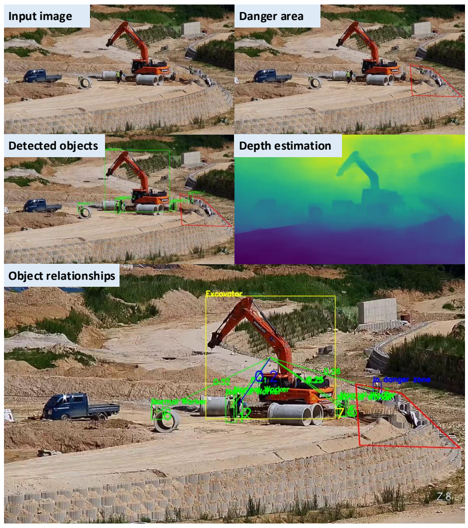

Figure 7. Step-by-step model detection.

supplemented with automated alerts for danger zone intrusions; slope stabilization procedures; mandatory use of PPE for all workers; and regular safety training sessions. This integrated approach aims to enhance worker safety and reduce accident risk.

5 Conclusion

This study addresses the challenges of identifying hazards in complex and unpredictable construction site environments, with the support of advanced deep-learning models. To overcome these obstacles, we propose a novel end-to-end pipeline, *Image-to-Hazard*, designed to bridge the gap between separate single-model predictions using multimodal and logical reasoning. This pipeline integrates detection, depth estimation, hazard identification, and logical reasoning, by employing a GPT-based model from OpenAI API for safety monitoring at construction sites. A large-scale video dataset was obtained from actual construction sites and labeled. Subsequently, a large-scale dataset was utilized to train and optimize tailored object detection models. Pre-trained models for semantic segmentation and depth estimation were utilized to generate a comprehensive visual feature extraction dataset. Visual feature information was then integrated with prompt structures and input into OpenAI GPT-based models for logical reasoning. In conclusion, this study demonstrated that the *Image-to-Hazard-Scenarios* pipeline, which combines multimodal data for context-aware hazard identification, was successful in enhancing safety monitoring at construction sites. The method was validated using actual CCTV footage, quantifying its performance and leading to the development of a deployment procedure for its implementation in practical construction site settings. This finding underscores the significant potential of employing integrated multimodal approaches in improving the safety of complex working environments, such as construction sites. Future research should focus on refining this pipeline and exploring its applicability to other similarly complex settings using real-time inference.

Acknowledgements

This research was supported by a grant [2022-MOIS38-002 (RS-2022-ND630021)] from the Ministry of Interior and Safety (MOIS)'s project for proactive technology development safety accidents for vulnerable groups and facilities, and this research was supported by a grant from the Korean Government (MSIT) to the NRF [RS-2023-00250166]. This work is financially supported by Korea Ministry of Land, Infrastructure and Transport(MOLIT) as Innovative Talent Education Program for Smart City.

References

[1] Fan Zhang, Hasan Fleyeh, Xinru Wang, and Minghui Lu. Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99:238–248, 2019.

[2] Large Accident Case Book, 2022. [Online; accessed 2022-07-18].

Figure 8. GPT logic reasoning.

[3] U.S. Bureau of Labor Statistics. https://www.bls.gov, 2023.

[4] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.

[5] JoonOh Seo, SangUk Han, SangHyun Lee, and Hyoungkwan Kim. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, 29(2):239–251, 2015.

[6] Kevin K Han and Mani Golparvar-Fard. Potential of big visual data and building information modeling for construction performance analytics: An exploratory study. *Automation in Construction*, 73: 184–198, 2017.

[7] Dai Quoc Tran, Minsoo Park, Daekyo Jung, and Seunghee Park. Damage-map estimation using uav images and deep learning algorithms for disaster management system. *Remote Sensing*, 12(24):4169, 2020.

[8] Dai Quoc Tran, Minsoo Park, Yuntae Jeon, Jinyeong Bak, and Seunghee Park. Forest-fire response system using deep-learning-based approaches with cctv images and weather data. *IEEE Access*, 10:66061–66071, 2022. doi:10.1109/ACCESS.2022.3184707.

[9] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023.

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[12] Minsoo Park, Dai Quoc Tran, Jinyeong Bak, and Seunghee Park. Small and overlapping worker detection at construction sites. *Automation in Construction*, 151:104856, 2023. ISSN 0926-5805. doi:https://doi.org/10.1016/j.autcon.2023.104856. URL `https://www.sciencedirect.com/science/article/pii/S0926580523001164`.

[13] Wei-Chih Chern, Jeongho Hyeon, Tam V Nguyen, Vijayan K Asari, and Hongjo Kim. Context-aware safety assessment system for far-field monitoring. *Automation in Construction*, 149:104779, 2023.

[14] Shi Chen, Kazuyuki Demachi, and Feiyan Dong. Graph-based linguistic and visual information integration for on-site occupational hazards identification. *Automation in Construction*, 137:104191, 2022.

[15] Numan Khan, Muhammad Rakeh Saleem, Doyeop Lee, Man-Woo Park, and Chansik Park. Utilizing safety rule correlation for mobile scaffolds monitoring leveraging deep convolution neural networks. *Computers in Industry*, 129:103448, 2021.

[16] MMYOLO Contributors. MMYOLO: OpenMMLab YOLO series toolbox and benchmark. `https://github.com/open-mmlab/mmyolo`, 2022.

[17] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[18] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637, 2020.