IDENTIFYING HAZARDS IN CONSTRUCTION SITES USING DEEP LEARNING-BASED MULTIMODAL WITH CCTV DATA

Dai Quoc Tran
Global Engineering Institute for Ultimate Society, Sungkyunkwan University, South Korea

Yuntae Jeon & Seongwoo Son
Department of Global Smart City, Sungkyunkwan University, South Korea

Minsoo Park*
Sungkyun AI Research Institute, Sungkyunkwan University, South Korea

Seunghee Park*
School of Civil, Architectural Engineering and Landscape Architecture, Sungkyunkwan University, South Korea

**Corresponding authors*

ABSTRACT: *The use of closed-circuit television (CCTV) for safety monitoring is crucial for reducing accidents in construction sites. However, the majority of currently proposed approaches utilize single detection models without considering the context of CCTV video inputs. In this study, a multimodal detection, and depth map estimation algorithm utilizing deep learning is proposed. In addition, the point cloud of the test site is acquired using a terrestrial laser scanning scanner, and the detected object's coordinates are projected into global coordinates using a homography matrix. Consequently, the effectiveness of the proposed monitoring system is enhanced by the visualization of the entire monitored scene. In addition, to validate our proposed method, a synthetic dataset of construction site accidents is simulated with Twinmotion. These scenarios are then evaluated with the proposed method to determine its precision and speed of inference. Lastly, the actual construction site, equipped with multiple CCTV cameras, is utilized for system deployment and visualization. As a result, the proposed method demonstrated its robustness in detecting potential hazards on a construction site, as well as its real-time detection speed.*

KEYWORDS*: deep learning, multimodal, multiCCTV, synthetic data, pointcloud*

1. INTRODUCTION

Construction sites are dynamic and complex environments that pose significant safety risks, resulting in an high rate of accidents and fatalities worldwide (Abdelhamid & Everett, 2000). Consequently, implementing effective safety monitoring measures is vital for reducing such incidents. The use of closed-circuit television (CCTV) cameras for safety monitoring on construction sites has played a crucial role in mitigating risks. Despite this, the full potential of CCTV data is often underutilized, primarily due to the majority of existing approaches employing single detection models without considering the full context of CCTV video inputs (Park et al., 2022, 2023; Tran et al., 2020). In response to this issue, this study proposes a novel and robust system that incorporates a multimodal detection and depth map estimation algorithm, utilizing the power of deep learning. The distinctiveness of our approach lies in the context-aware analysis, providing a more comprehensive understanding of the potential hazards present within the dynamic environments of construction sites. Furthermore, our proposed method goes a step further by leveraging terrestrial laser scanning technology to acquire the point cloud of the test site and utilizing a homography matrix to project the detected object's coordinates into global coordinates. This step enhances the overall monitoring system's effectiveness by visualizing the entire monitored scene, thus providing a bird eye view of the potential hazards. A crucial aspect of any newly proposed system is rigorous validation. For our method, we have created a synthetic dataset of construction site accidents using Twinmotion, a high-powered graphic software. This dataset provides a range of simulated scenarios to test the precision and inference speed of our proposed method, ensuring its reliability and robustness in varied contexts.

Finally, we further validate our method by deploying it on an actual construction site equipped with multiple CCTV cameras, moving beyond simulations to a real-world setting. This on-site implementation allowed us to assess the practicality of our system and its ability to function optimally in an uncontrolled, real-world environment. The results from both simulation and real-world deployment demonstrate our proposed method's

Referee List (DOI: 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Dai Quoc Tran, Yuntae Jeon, Seongwoo Son, Minsoo Park, Seunghee Park, *Identifying Hazards in Construction Sites Using Deep Learning-Based Multimodal with CCTV Data*, pp. 625-633, © 2023 Author(s), CC BY NC 4.0, DOI 10.36253/979-12-215-0289-3.61

robustness in detecting potential hazards on a construction site. This paper aims to highlight the potential of multimodal detection approaches in enhancing construction site safety measures, moving towards a future where such hazards can be preemptively detected and effectively mitigated. In the following sections, we will provide a detailed explanation of our proposed method, its development, and the validation process. We will also present the results and implications of this study, demonstrating how a deep learning-based multimodal approach can be used for safety monitoring in the construction industry.

2. BACKGROUND

This section provides the necessary background on the key aspects of our methodology, namely object detection, depth estimation, and multimodal synchronization. These three components collectively constitute the core of our proposed method and are fundamental in understanding the context-aware safety monitoring approach.

2.1.Object Detection

Object detection, as a fundamental component of computer vision, has undergone significant developments over the past decade, thanks to the advancements in deep learning and convolutional neural networks (CNNs) (Li et al., 2021). This involves identifying and locating objects within images or video feeds. In construction sites, object detection can identify critical elements such as workers, machinery, tools, and other potential hazards, thereby playing a vital role in safety monitoring. However, traditional object detection models typically operate independently, failing to incorporate the broader context of a scene (Jeon et al., 2023; Tran et al., 2022). These models often struggle with complex environments like construction sites, where multiple objects interact dynamically, and understanding these interactions is crucial for effective hazard detection. This limitation forms the motivation for our study, aiming to integrate a higher level of contextual understanding into object detection models using deep learning algorithms. In this research, two state of the art object detection models are utilized: Yolov8 (Redmon et al., 2016) and RTMDet (Lyu et al., 2022). These two object detectors are trained and validated in actual CCTV images and utilized for incorporation with other models.

2.2.Depth Estimation

For understanding context of input image, spatial information is crucial, therefore, a depth estimation model MiDAS (Ranftl et al., 2020) is utilized. Depth estimation refers to the task of determining the relative distance of objects within a scene from the viewpoint of the camera. It is a crucial component in understanding three-dimensional spaces from two-dimensional images or video feeds, providing invaluable information about the positioning and interaction of objects within a scene. In construction sites, depth estimation can enhance the understanding of spatial relations among various elements, such as the proximity of a worker to a moving machine, thereby aiding in detecting potentially hazardous situations. This paper incorporates depth estimation into our proposed multimodal approach, further improving the context-awareness of the system.

2.3.Multimodal Synchronization

Incorporating multimodal synchronization into safety monitoring has the potential to significantly enhance the effectiveness of hazard detection. For example, a hazard that is not visible from one camera angle may be clearly observable from another. Similarly, certain hazardous situations may only be identifiable when considering multiple factors, such as the positioning and motion of various objects, which could be obtained from different data sources. In our proposed method, we aim to utilize multimodal synchronization to integrate object detection and depth estimation data, along with point cloud data obtained through terrestrial laser scanning. This synchronization allows for the projection of detected objects into global coordinates, enhancing the visualization of the entire monitored scene and thus the system's overall effectiveness. In the subsequent sections, we will elaborate on the implementation of these components in our proposed method and demonstrate their effectiveness in enhancing construction site safety monitoring.

3. PROSED APPROACH

3.1.Proposed approach

Figure 1 depicts the proposed method as follows: First, the input image is predicted using object detectors that have been pretrained. In this study, an object detector that can infer six classes is described in detail, along with the training procedure and training dataset. In addition, the MiDAS model is used to estimate the depth map using a previously trained depth estimation model. The Euclidean distance between objects is then estimated and visualized with the given depth and bounding box coordinates, after which each object's depth map is extracted. This research considers the distance between construction machines and employees, as well as the module to estimate whether or not a worker is in the danger zone. In addition, work-related personal protective apparatus is trained and implied.

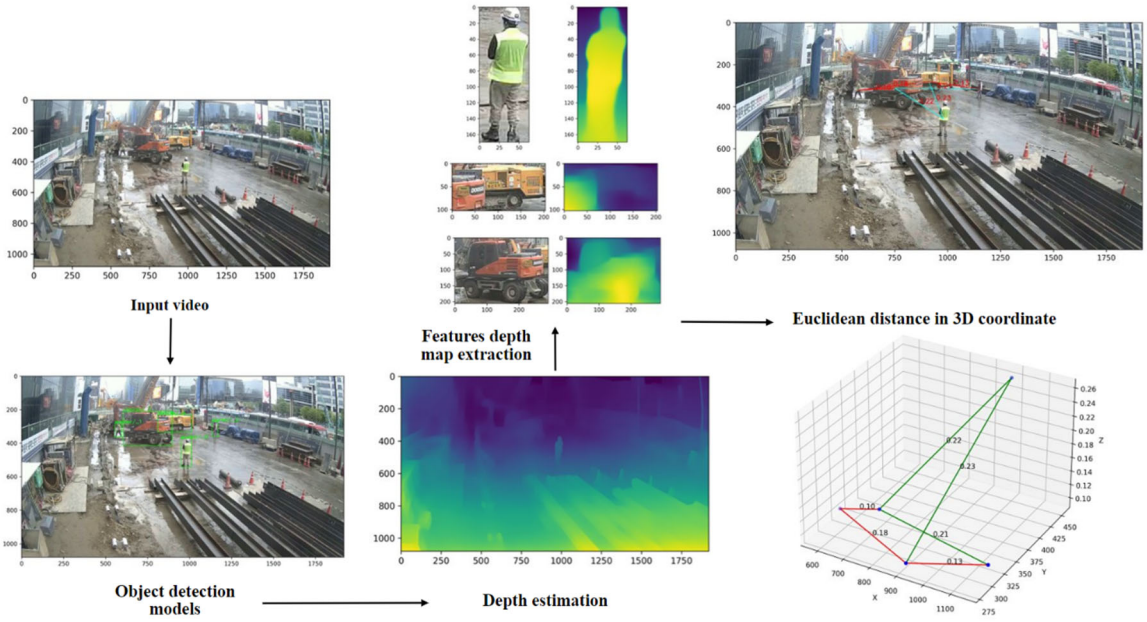

Figure 1. Proposed Approach

3.2.Dataset Acquisition

The training dataset contains bounding boxes objects from 6 classes: *normal worker, signalman, harness, hardhat, mixer truck and excavator*. As visualized in Figure 2, each object is labeled in detail from actual CCTV footage. Figure 3 presented a training and testing dataset classes distribution.

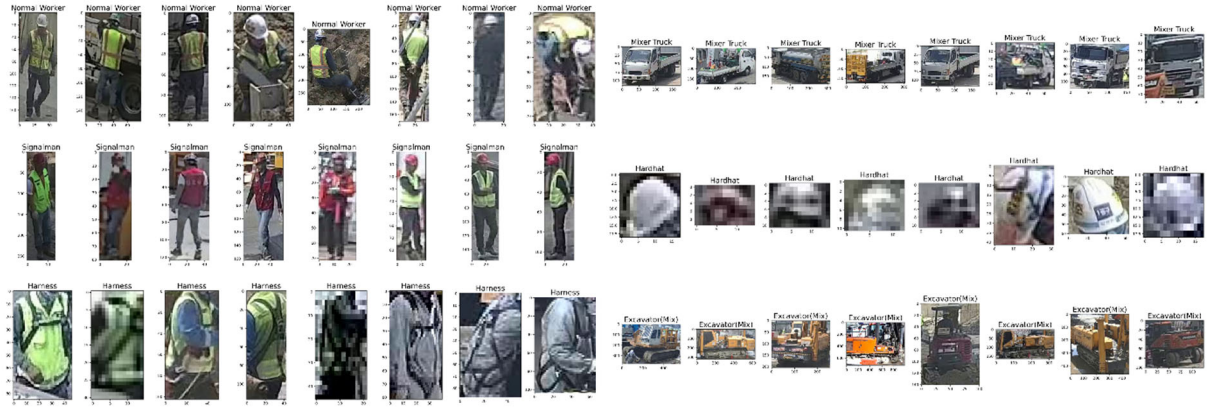

Figure 2. Dataset visualization

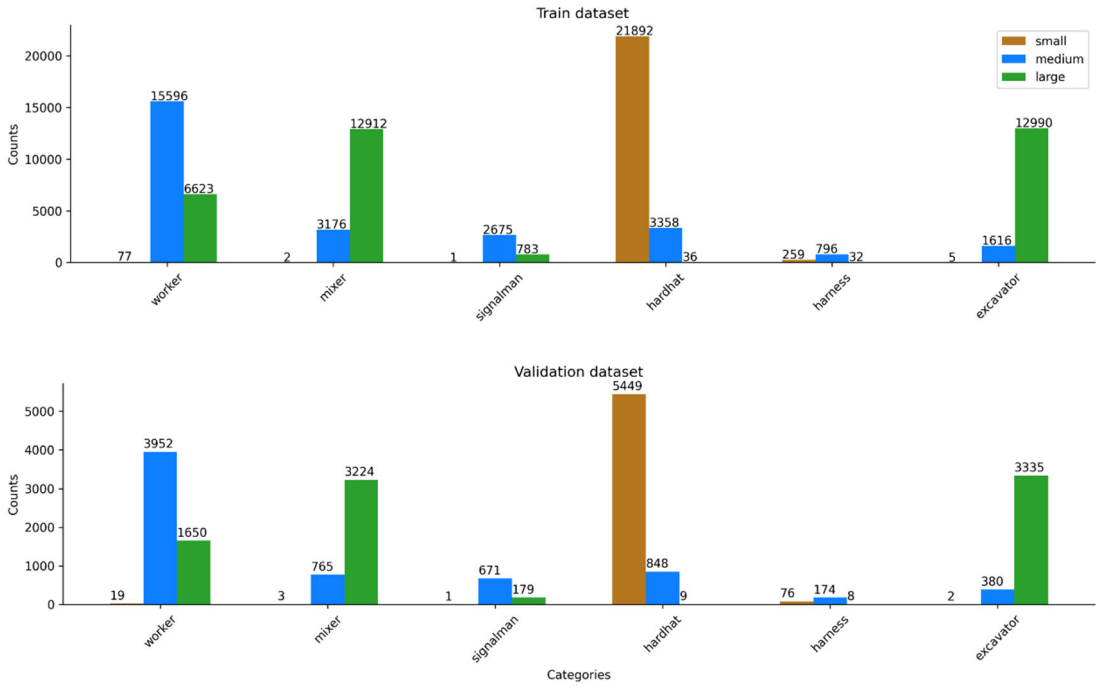

Figure 3. Training and testing dataset distribution. The y-axis depicts the total of each class in training and testing datasets. The x-axis lists the distinct class names or labels in dataset. In our case, these are: “worker”, “mixer”, “signalman”, “hardhat”, “harness”, and “excavator”.

For evaluate model performance, the detected objects are categorized into 3 types: "small", "medium", and "large" follows a specific definition based on the number of pixels occupied by an object in an image. Specifically, objects are categorized as follows:

- "Small": Objects occupying 0 to 1024 pixels, which translates to dimensions up to 32 × 32 pixels.
- "Medium": Objects occupying between 1024 and 9216 pixels, which corresponds to dimensions between 32 × 32 and 96 × 96 pixels.
- "Large": Objects occupying more than 9216 pixels, equating to dimensions of 96 × 96 pixels or larger.

With these categorizations, our dataset reflects an essential aspect of object detection tasks: dealing with objects of varying sizes. The balance or imbalance of different categories may significantly affect the predictive performance of an object detection model. In our dataset, categories such as "mixer" and "excavator" contain a substantial number of "medium" and "large" objects. Conversely, the "hardhat" category is predominantly composed of "small" objects. This imbalance suggests that an object detection model trained on this data might develop a bias toward detecting "medium" or "large" objects and underperform on "small" objects.

4. Experimental

4.1.Quantitative

Table 1. Object detectors performance.

	mAP	mAP_50	mAP_75	mAP_s	mAP_m	mAP_l	Flops
YOLOv8X	0.689	0.831	0.741	0.123	0.733	0.811	0.129T
RTMDet	0.628	0.776	0.688	0.059	0.641	0.756	79.964G

In order to quantify the results, the mean average precision (mAP) is employed as an evaluation metric. The mAP is a widely recognized indicator used to quantitatively assess the performance of object identification models. The research provides a comprehensive explanation of the mAP (Zhao et al., 2019). As presented in Table 1, YOLOv8X and RTMDet, with regard to their performance metrics, a notable variance in their effectiveness becomes clear. The YOLOv8X model exhibits superior precision with a mAP score of 0.689, which considerably exceeds the 0.628 mAP score of the RTMDet model. This indicates an overall higher rate of accurate detections by the YOLOv8X model. Further disparity can be observed at varying Intersection over Union (IoU) thresholds. The mAP_50 and mAP_75 scores, which represent the mAP values computed at IoU thresholds of 0.50 and 0.75 respectively, demonstrate a superior adaptability of the YOLOv8X model to changes in detection difficulty levels. When checking the mAP values across different object sizes, denoted by mAP_s (small), mAP_m (medium), and mAP_l (large), the YOLOv8X model continues to display superior performance. The model's proficiency in detecting small objects is particularly noteworthy, with a score of 0.123 compared to the RTMDet's score of 0.059. Nevertheless, it is crucial to consider the computational complexity of the models. RTMDet has a significant advantage in this regard, with a computational demand of 79.964G Flops, markedly lower than the YOLOv8X's 0.129T (or 129,000G) Flops. This positions RTMDet as a more feasible option for applications with limited computational resources, despite its inferior mAP performance. While the YOLOv8X model outperforms RTMDet in terms of object detection performance across various metrics, the latter's significantly lower computational demand may make it a more suitable candidate for resource-constrained applications. The selection between these two models, therefore, necessitates careful consideration of the balance between performance efficiency and resource utilization, contingent on the specific requirements of the application.

4.2.Qualitative

As mentioned in previous sections, after detecting objects, the depth map is estimated and calculating the distance between worker and construction vehicles. As can be seen from Figure 4, by utilizing spatial information, the distance can be estimated and from that, the necessary warning can be conducted.

Figure 4. Distance estimation using depth estimation and object detection.

Figure 5 showed another application of the proposed approach by identifying which workers are in the danger area. The danger area is defined by the safety officer, and when detected worker violate that area, the number of violated cases will be shown and reported directly to safety manager. Along with detecting worker, a PPE detection models also consider to utilized, as can be seen from Figure 6, both hardhat and harness is detected for checking. To reduce the false positive, we remove detected PPE outside of the worker detected area, as can be seen in Algorithm 1.

Algorithm 1. Detecting PPE

```
Lqsxw=#Lpdjh#L#
Rxwsxw=#Glvsod|hg#erxqglqj#er{hv#ri#ghwhfwhg#vdihw|#htxlsphqw#rq#shuvrqv#
#
4=#surfhgxuh#VDIHW\bHTXLSPHQWbGHWHFWLRQ+L,#
5=#####shuvrqPrgho#?0#Lqlwldol}h#suhwudlqhg#prgho#iru#shuvrq#ghwhfwlrq#
6=#####vdihw|HtxlsphqwPrgho#?0#Lqlwldol}h#rxu#prgho#iru#kdugkdw#dqg#kduqhvv#ghwhfwlrq#
7=######
8=#####shuvrqErxqglqjEr{hv#?0#shuvrqPrgho1ghwhfw+L,#
9=#####vdihw|HtxlsphqwErxqglqjEr{hv#?0#vdihw|HtxlsphqwPrgho1ghwhfw+L,#
:=######
;=#####iru#hdfk#shuvrqEr{#lq#shuvrqErxqglqjEr{hv#gr#
<=#########iru#hdfk#htxlsphqwEr{#lq#vdihw|HtxlsphqwErxqglqjEr{hv#gr#
43=#############LrX#?0#FDOFXODWHbLRX+shuvrqEr{/#htxlsphqwEr{,#
44=##############
45=#############li#LrX#A#318#wkhq#
```

```
46=##################GLVSOD\bERXQGLQJbER[+htxlsphqwEr{/#L,#
47=##############hqg#li#
48=#########hqg#iru#
49=#####hqg#iru#
4:=#hqg#surfhgxuh#
```

Figure 5. Detect workers in the danger area

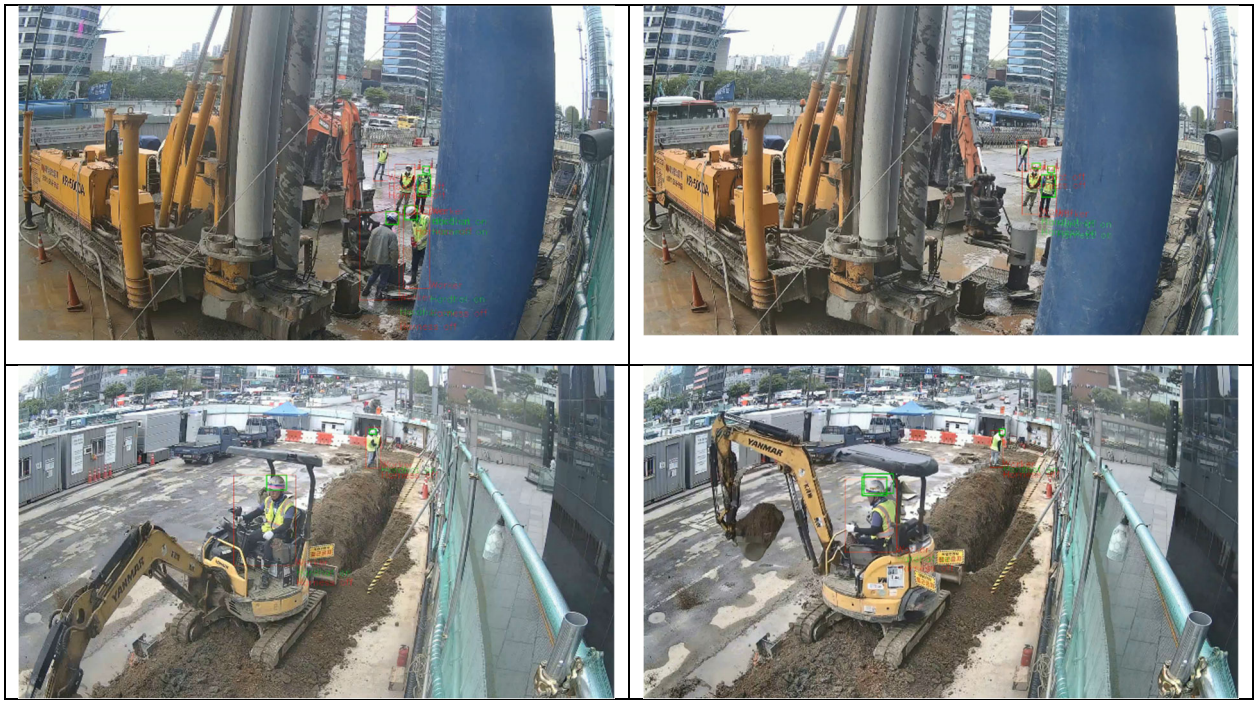

Figure 6. PPE detection

Finally, for developing multi-CCTV and bird eye view (BEV) synchronization. Two cases of examples are conducted as follows: first, by utilizing TwinMotion, the 4 CCTV channel is developed and used as an input for detection models, with homography matrix, the BEV is shown in Figure 7.

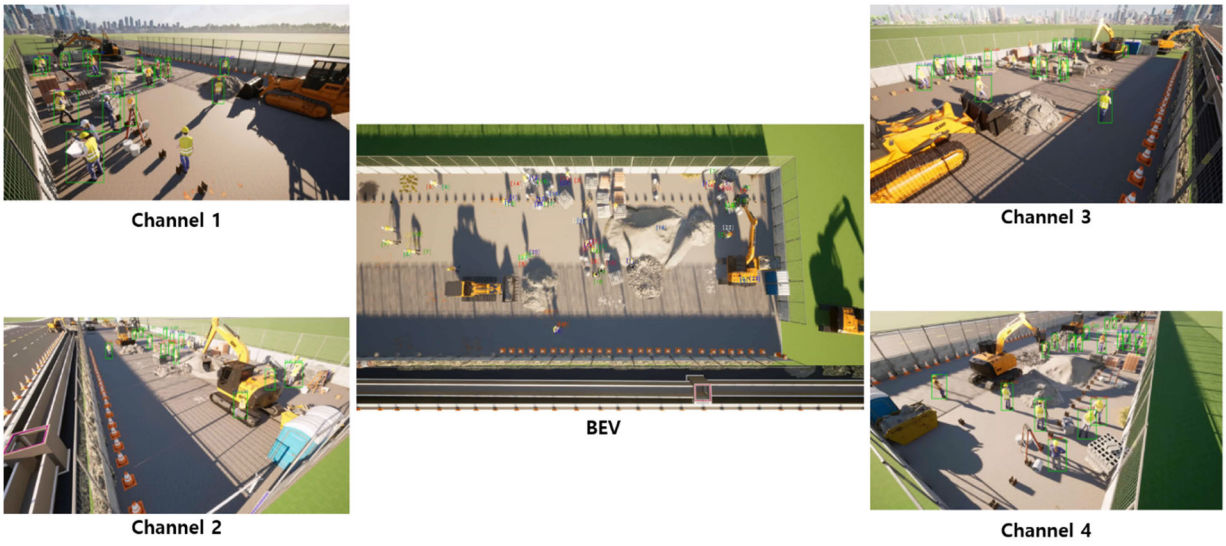

Figure 7. Multi CCTV with BEV in Twinmotion

Similarly with above example, but in actual construction site, it difficult to obtain BEV image, that why TLS is utilized for scanning and from that BEV can be estimated as shown in Figure 8 and 9. However, by only projecting all detected object into BEV, the exact ID of objects vary channels to channels. Therefore, the application of multi object detection and tracking can be used for future research. The expected output is given multi-channel CCTV, the output is the BEV with exact number and ID of detected objects. This study mostly emphasizes qualitative experiments. In order to obtain quantitative results, the forthcoming experiment will be undertaken by establishing an indoor environment and thereafter measuring the error projection using a meter.

Figure 8. Multi CCTV with BEV in actual construction site

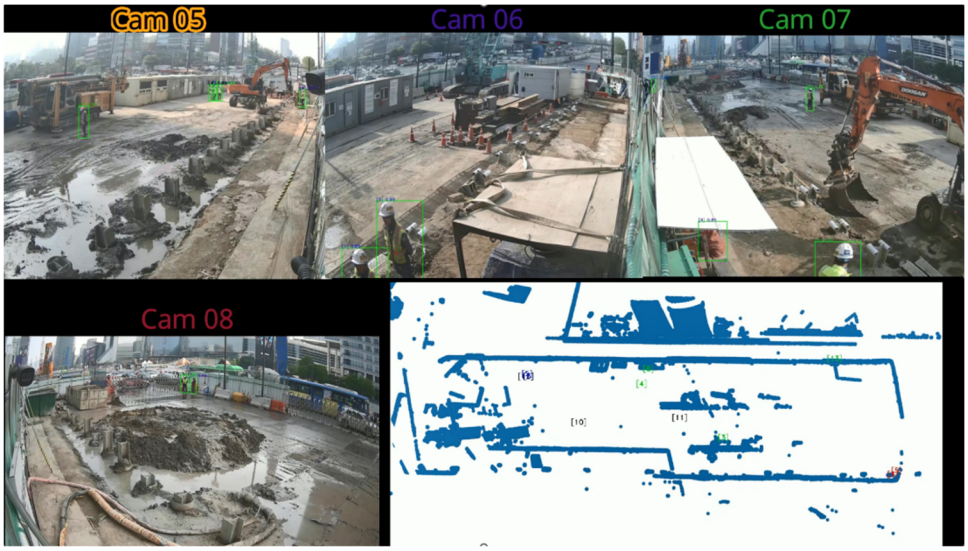

Figure 9. Multi CCTV with BEV in actual construction site

5. CONCLUSION

By adopting a multimodal detection approach, the study proposes a depth map estimation algorithm designed to enhance the contextual understanding of video inputs from CCTV. The proposed method uses terrestrial laser scanning to generate a point cloud of the test site and leverages a homography matrix to project detected objects into global coordinates. With object detection models used in the proposed method, a detailed analysis of YOLOv8X and RTMDet was conducted. YOLOv8X exhibited superior precision across various measures, including overall mAP, mAP at varying IoU thresholds, and mAP across different object sizes. However, the RTMDet model was identified as more resource-efficient, demanding significantly fewer computational resources despite its lower mAP performance. This research also presented some use cases of multimodel detections, it proves that context-aware approach in safety monitoring is important and should be considered for further research.

6. ACKNOWLEDGEMENT

This research was supported by a grant from the Korean Government (MSIT) to the Research Foundation of Korea (NRF) [RS-2023-00250166]. In addition, this research was supported by a grant [2022-MOIS38-002 (RS-2022-ND630021)] from the Ministry of Interior and Safety (MOIS)'s project for the development of accident prevention technology for vulnerable groups.

REFERENCES

Abdelhamid, T. S., & Everett, J. G. (2000). Identifying Root Causes of Construction Accidents. *Journal of Construction Engineering and Management*, *126*(1), 52–60. https://doi.org/10.1061/(ASCE)0733-9364(2000)126:1(52)

Jeon, Y., Tran, D. Q., Park, M., & Park, S. (2023). Leveraging Future Trajectory Prediction for Multi-Camera People Tracking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5398–5407.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*.

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., & Chen, K. (2022). Rtmdet: An empirical study of designing real-time object detectors. *arXiv Preprint arXiv:2212.07784*.

Park, M., Tran, D. Q., Bak, J., & Park, S. (2022). Advanced wildfire detection using generative adversarial network-based augmented datasets and weakly supervised object localization. *International Journal of Applied Earth Observation and Geoinformation*, *114*.

Park, M., Tran, D. Q., Bak, J., & Park, S. (2023). Small and overlapping worker detection at construction sites. *Automation in Construction*, *151*, 104856. https://doi.org/10.1016/j.autcon.2023.104856

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). 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.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Tran, D. Q., Park, M., Jeon, Y., Bak, J., & Park, S. (2022). Forest-Fire Response System Using Deep-Learning-Based Approaches With CCTV Images and Weather Data. *IEEE Access*, *10*, 66061–66071. https://doi.org/10.1109/ACCESS.2022.3184707

Tran, D. Q., Park, M., Jung, D., & Park, S. (2020). Damage-Map Estimation Using UAV Images and Deep Learning Algorithms for Disaster Management System. *Remote Sensing*, *12*(24), 4169.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865