ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0791
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Computer Vision-based Construction Hazard Detection via Data Augmentation Approach using Generative-AI

WooWon Jo[1], YeJun Lee[2], Daegyo Jung[3], HyunJung Park[4], JungHo Jeon[5]*

[1] *Department of Architectural Engineering, Researcher, Pusan National University, South Korea,* E-mail address: jww2864@pusan.ac.kr
[2] *Department of Architectural Engineering, Researcher, Pusan National University, South Korea,* E-mail address: lyj6720@pusan.ac.kr
[3] *Department of Architectural Engineering, Ph.D. student, Pusan National University, South Korea,* E-mail address: 94-09-03@pusan.ac.kr
[4] *Department of Architecture, Professor, Silla University, South Korea,* E-mail address: phj@silla.ac.kr
[5] *Department of Architectural Engineering, Assistant Professor, Pusan National University, South Korea,* E-mail address: junghojeon@pusan.ac.kr

Abstract: Construction industry records poor safety records annually due to a large number of injuries and accidents on construction jobsite. In order to improve existing safety performance, object detection approaches have been extensively studied using vision-sensing techniques and deep learning algorithms. Unfortunately, an insufficient number of datasets (e.g., images) and challenges that reside in manually collecting quality datasets constitute a significant hurdle in fully deploying object recognition approaches in real construction sites. Although advanced technologies (e.g., virtual reality) have attempted to address such challenges, they have achieved limited success because they still rely on labor-intensive work. A promising alternative is to adopt generative AI-based data augmentation methods attributed to their efficiency in creating realistic visual datasets and proven performance. However, there remain critical knowledge gaps on how such alternatives can be effectively employed by safety managers on real construction sites in terms of practicability and applications. In this context, this study establishes a framework that can identify effective strategies for improving object detection performance (e.g., accuracy) using generative AI technologies. The outcome of this study will contribute to providing guidelines and best practices for practitioners as well as researchers by exploring different generative AI-driven augmentation approaches and comparing the corresponding results in a quantitative manner.

Key words: construction safety, vision sensing, object detection, generative AI, data augmentation

1. INTRODUCTION

Construction industry has poor safety records, with a large number of injuries and fatalities on construction jobsites. To prevent such accidents and improve existing safety performance, monitoring unsafe workplace conditions and workers' risky behaviors has become a critical component in the construction safety management domain [1]. The underlying rationale is that the monitoring approach allows for recognizing unsafe conditions and unsafe acts—which were identified as two leading causes of accidents—and establishing corrective actions to prevent unexpected accidents in the workplace [2].

In the current practice, the safety monitoring process heavily relies on safety managers' manual observations, which are time-consuming, inaccurate, and subjective [3]. Also, the dynamic nature of construction jobsites further aggravates the challenges in current practice. In order to address the challenges, computer vision technology has been widely used for continuous and automated safety monitoring because it can extract information (e.g., locations of hazards) from captured data (e.g., images); automate construction processes (e.g., visual inspection); and holds the potential to be applied for various tasks [4]. In particular, within the computer vision domain, improving the performance of object detection, which aims to recognize the presence of a particular object (e.g., hazard) within an image, remains an open challenge due to many factors (e.g., the number/quality of imagery data and selected algorithms) contributing to the performance. Among many of these factors, the number of imagery data used to train the model is of particular interest because it was found that increasing training data led to better detection performance and generalization capability [5]. Unfortunately, collecting a large number of high-quality imagery data used to train and test the object-detecting model is a challenging task because it usually relies on researchers' significant manual efforts (e.g., taking photos and collecting images via web search).

A promising alternative is to utilize generative artificial intelligence (Gen-AI) technologies that can create auditory, visual, and textual content based on user queries [6]. For example, the text-to-image Gen-AI technique can take user prompts as input, understand the context, and generate corresponding images in an automated and efficient manner. Several studies have demonstrated the feasibility of employing a text-to-image approach to augment imagery data and improve corresponding object detection performance [7].

However, from a perspective of construction safety management, there remain three gaps in the current knowledge base to attain the ultimate goal of fully detecting construction hazards on jobsites based on an approach consisting of text-to-image data augmentation and computer vision-based object detection. First, despite the importance of fall hazards that were identified as a primary contributor to injuries and fatalities, a limited number of studies focused on augmenting fall imagery data (i.e., a scene containing fall hazards) using Gen-AI and investigated the feasibility of using them for advanced hazard detection. Second, it remains unclear how imagery data can be efficiently augmented using what types of prompt engineering techniques. In other words, the relationship between the user's input that is fed into the text-to-image algorithm and the quality of outcome produced by Gen-AI has not been thoroughly studied. Third, it is unknown how the augmented imagery data contributes to the object (i.e., fall hazard) detection performance.

To bridge the gaps, this paper aims to assess the feasibility of using Gen-AI (i.e., text-to-image) to augment imagery data and investigate how it contributes to object detection performance based on a methodology consisting of the following three steps. First, two sets of imagery datasets are constructed based on web search and text-to-image approach, respectively; each image illustrates a construction site containing a fall hazard (i.e., floor opening) without protection (e.g., safety net). Second, all the imagery data points are labeled for object detection experiments. Third, the fall hazard-detecting model is developed using the YOLOv8n algorithm. The results of this paper demonstrate the effectiveness of using Gen-AI technology to augment a large number of imagery data, revealing that they can improve object detection performance. The findings are expected to contribute to advancing Gen-AI and computer vision-centered construction safety monitoring and improve existing safety performance.

2. REVIEW OF RELATED STUDIES

2.1. Computer Vision and Object Detection

Computer vision—an interdisciplinary field encompassing varying theories and techniques—aims to help computers understand the contents of imagery data collected by sensors (e.g., cameras) [8]. It can be used for object detection, classification, visual tracking, image segmentation, etc. Among the above various applications, the goal of object detection is to identify a semantic object of interest (e.g., hazard) within an image based on a well-established workflow (e.g., data acquisition, preprocessing, and feature extraction). Attributed to its unique advantages (e.g., automated extraction of relevant information), object detection approaches have been widely adopted within the construction safety domain to detect on-site hazards (e.g., falls), assess the absence of personal protective equipment (PPE) on workers, and monitor the unsafe behavior of construction workers [9–11]. It was found that many factors (e.g., size of datasets, model architecture, and quality of data labeling) contribute to the performance of object detection. Among them, the size of the dataset used to train and test the object detector (model) is critical to achieving an acceptable level of detection performance, as demonstrated by previous studies [10]. However, collecting a sufficient number of image data is a time-consuming and labor-intensive task as it mostly counts on researchers' manual efforts (e.g., taking photos). This highlights the need to adopt data augmentation approaches that will be explained in the following section.

2.2. Data Augmentation

Data augmentation, a crucial technique in computer vision and many other fields, artificially expands and diversifies a dataset by creating new data points based on existing ones (or from scratch) to enhance the model's performance [12]. Taking image data as an example, augmentation methods can be classified into several types, such as geometric transformations (e.g., random rotation, cropping, and flipping), photometric transformations (e.g., changing contrast and brightness), etc. [13]. In the construction safety field, several studies augmented image data to construct a large-size dataset and achieved better performance in detecting various objects (e.g., workers, pipes, and excavators) [14,15]. However, existing data augmentation approaches often fail to maintain semantic information, still require significant human involvement and guidance, and do not always lead to enhanced performance [16]. To address the limitations in existing augmentation approaches, recent research has started to harness Gen-AI approaches for image data augmentation, and their effectiveness has been continuously studied and reported [17,18].

2.3. Generative Artificial Intelligence (Gen-AI)

Gen-AI is a subfield of artificial intelligence that is capable of creating even novel outputs in response to user-provided requests (prompts) [19]. It encompasses a spectrum of approaches, such as text-to-text, text-to-image, and text-to-video, all of which generate outputs based on corresponding textual inputs. Particularly, the text-to-image method transforms textual descriptions into visual representations through several technical components (e.g., deep learning algorithms and encoder-decoder architecture). Attributed to its unique advantages and potential, it has been widely adopted for architectural design, personalized education, and the creation of challenging test cases for software [20,21]. Some studies attempted to leverage the text-to-image approach for imagery data augmentation to improve object detection and classification performance [7]. Reviewing the related studies revealed that the quality of the generated images largely depends on the selected tools (e.g., Dall-E and Midjourney) and the prompts designed by the users. A prompt that comprises multiple components (e.g., word and sentence structure) is critical in achieving quality outputs as it serves as guidance that models can follow and learn. Well-known prompting techniques include *n*-shot prompting (where *n* defines the number of examples provided to the input template), structured prompting, and resampling [22]. Since there is no universal agreement that one specific prompting technique outperforms others for the given research task, identifying the optimal prompting techniques remains a critical challenge that deserves continuous research efforts.

3. METHODOLOGY

3.1. Data Collection and Augmentation

Two sets of datasets were constructed based on real images and text-to-video approach, respectively. For the first dataset, the images that illustrate a construction site containing a fall hazard were collected from search engines (e.g., Google) using various combinations of keywords (e.g., opening, fall hazard, and construction site). Both English and Korean were used for the search process. Due to diverse forms of fall hazards (e.g., open trenches and wall openings), the authors focused on images that display real construction sites where openings without protection (e.g., guardrails and safety nets) could be observed on a floor. In addition, the images lacking clarity and exhibiting significant blurring were removed. As a result, the first dataset consisted of a total of 100 real images.

The second dataset was developed by testing several prompt engineering techniques—that focus on strategically designing user queries to achieve quality outputs—and leveraging the text-to-image approach. Of many publicly available text-to-image platforms (e.g., Midjourney, Dall-E, and Adobe Firefly), Midjourney was employed due to its proven capabilities of creating realistic scenes and unique features [23]. For example, it can refine resulting outcomes through iterative processes and explore different interpretations of the user prompt. In order to maximize the quality of images created by Midjourney, the authors tested three prompt engineering techniques (*n*-shot prompting, structured prompting, and Chain-of-Thought) in an exploratory manner. This preliminary experiment revealed that structured prompting, a method that organizes user queries into predefined templates, was the most effective approach in generating realistic scenes depicting floor openings lacking proper safety measures (Figure 1). As a result, the second dataset comprised 1,000 augmented images; the rationale for creating larger data points was that one of the objectives of this paper was to observe the hazard detection performance as the number of augmented data increases while training the object detection model.

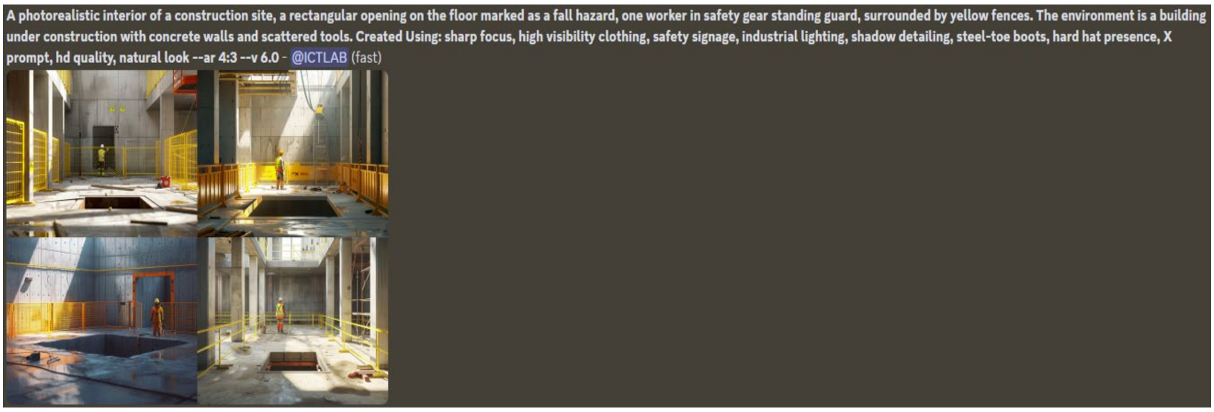
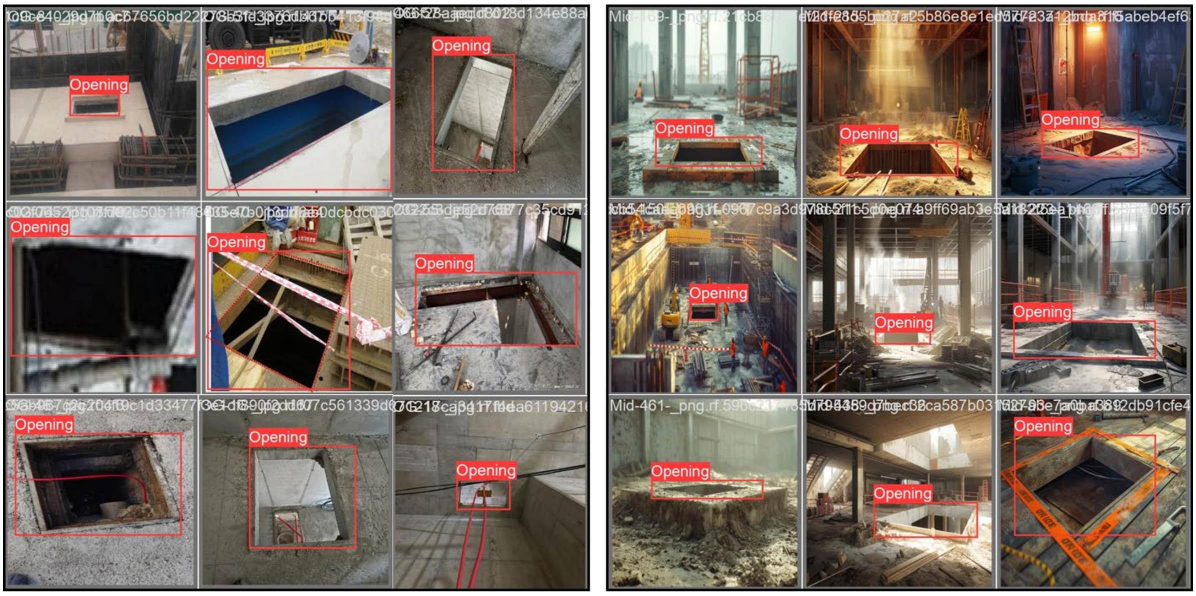


Figure 1. Example of structured prompting to generate augmented image data

3.2. Data Annotation

A total of 1,100 images (100 real images and 1,000 augmented images) were labeled using Roboflow, which is a reliable platform that streamlines the process of object labeling in imagery data. During the labeling process, the bounding boxes were manually drawn in the area of individual images by assigning one class (i.e., opening) representing fall hazards. Note that all imagery data included no more than one class, which confirms the existence of a single opening on the construction site within each image. Figure 2 presents sample figures of labeled real images and augmented images.



(a) Real image data (b) Augmented image data

Figure 2. Illustrative examples of dataset images

3.3. Implementation of YOLO Model

After labeling the image dataset, an object detection model that can recognize and localize the fall hazard was trained using the YOLOv8n (You Only Look Once, version 8 nano) algorithm due to its proven performance and computing efficiency [24,25]. Some studies found that YOLOv8n has a faster detection speed compared to different versions (e.g., YOLOv8x and YOLOv8m) [26]. The entire dataset was randomly split between training, validation, and testing with a 70/15/15 ratio. When training the model, hyperparameters (e.g., epoch and learning rate) that can be configured by users were optimized by initiating the training process with default values. The optimal hyperparameters were determined through an exploratory approach by testing various parameter combinations and changing the settings, ensuring the prevention of both overfitting and underfitting. In addition, when the augmented images were used, they were not used for testing since the focus of this study is on how the augmented fall hazard data can impact the resulting detection performance that can be observed in the testing set. For example, when the augmented 100 images were used along with the 100 real data, the augmented data remained on training and validation sets. Google Colab and associated libraries were used for the implementation.

4. RESULTS AND DISCUSSIONS

To evaluate and compare the fall hazard detection performance using the optimized YOLOv8n model, four well-established metrics (precision, recall, mAP50, and mAP50-95) were adopted. Precision is the percentage of correctly identified objects with respect to the total number of objects that are identified as positive, as shown in Eq. (1). Recall is defined as the ratio of correctly recognized objects to the total number of actual objects, as shown in Eq. (2).

$$Precision = TP/(TP + FP) \quad (1)$$

$$Recall = TP/(TP + FN) \quad (2)$$

where *TP* indicates the number of true-positive, *FP* represents the number of false-positive, and *FN* signifies the number of false-negative.

The mAP is computed by averaging the average precision values of the corresponding class; mAP50 indicates that the overall detection is correct if predicted bounding boxes overlap with the ground truth by at least 50%, and mAP50-95 represents that the overall detection performance is determined based on multiple intersection-over-union (IoU) from 50% to 95%. Table 1 illustrates the detection performance as the number of augmented data increases while training and testing the model. Note that the initial epoch value was set to 500; it increases until no further accuracy improvement is observed.

Table 1. Comparison of object detection performance

Case No.	Number of real data	Number of augmented data	Precision	Recall	mAP50	mAP50-95	Epoch
1		0	0.816	0.824	0.794	0.396	184
2		100	0.857	0.781	0.855	0.552	135
3		200	0.924	0.809	0.853	0.568	192
4		300	0.852	0.853	0.854	0.583	115
5		400	0.931	0.779	0.879	0.59	105
6	100	500	0.944	0.914	0.953	0.737	201
7		600	0.924	0.841	0.900	0.674	196
8		700	0.922	0.877	0.900	0.689	306
9		800	0.938	0.882	0.917	0.723	278
10		900	0.938	0.908	0.946	0.748	351
11		1,000	0.932	0.904	0.943	0.738	216

When 100 real images were used without the involvement of augmented images (Case No. 1), the model achieved a precision of 81.6%, recall of 82.4%, mAP50 of 79.4%, and mAP50-95 of 39.6%. As the number of augmented data increases, the performance continuously changes with some fluctuations. Note that the mAP50 is used as the primary measure of object detection performance. In Case No. 2, where 100 imagery data generated by Midjourney was used to further enhance the model with 100 existing real images, the performance increased from 79.4% to 85.5%, as indicated by mAP50 values. When 500 augmented data was used in Case No. 6, the model achieved the highest performance of 95.3%. Further increasing the augmented data points did not lead to a significant improvement. For example, although all the augmented imagery data was used in Case No. 11, the mAP50 of 94.3% was observed, which is lower than that of Case No. 6. Figure 3 presents the results by focusing on three measures of mAP50, precision, and recall.

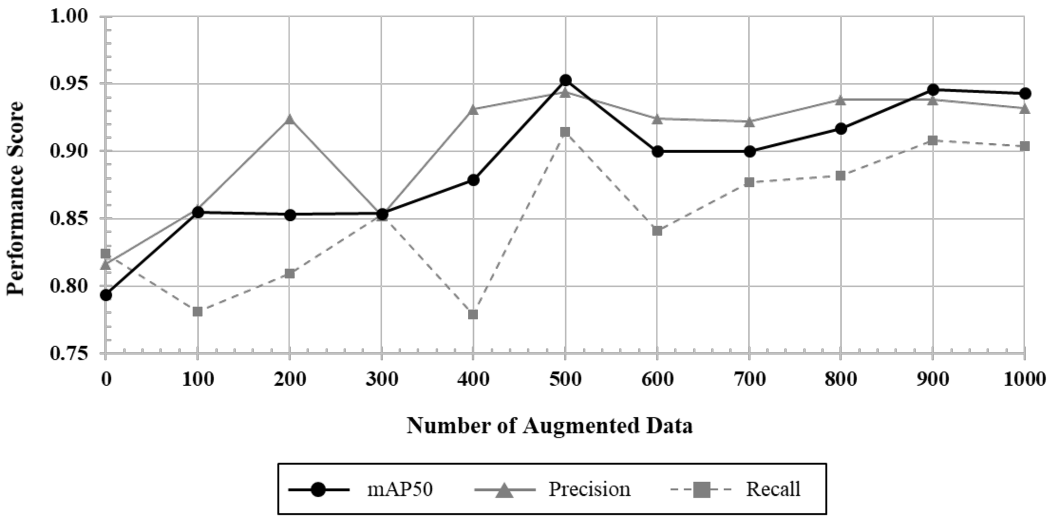


Figure 3. Performance improvement results

An in-depth analysis of the results led to the following three observations. First, the use of augmented imagery data produced by the text-to-image tool resulted in an improvement in fall hazard detection performance in all cases. An average mAP50 of 90.0% was recorded for case No.2-No.11 and mAP50 of 79.4% was found in No. 1. Although the sample size was not sufficient, this suggests that a data augmentation approach based on Gen-AI can be used as a reliable method to address the existing limitations that reside in the size of the training set used to develop the detection model, and increase the object detection performance. Second, it was found that the number of augmented data points and the resulting performance do not shape a linear relationship, as demonstrated by continuous fluctuations in Figure 3. It is likely that such a result was due to variations that could be found among augmented images in terms of quality, which partially contributed to the model training process and the resulting performance. In other words, some realistically generated images could have resulted in accuracy improvement, while others negatively impacted the performance. This highlights the future research efforts on investigating relationships among the quality of images generated by Gen-AI, the number of fused data, and the object detection performance. Third, after examining some of the predictions made by the YOLO model (Figure 4), it was found that the model performed well in identifying fall hazards for rectangular-shaped openings compared to irregular-shaped openings, as indicated by confidence scores, which represent the model's certainty in its predictions. For example, the confidence score of 0.6 was recorded in the third example in Figure 4, while relatively high scores of 0.8 and 0.9 were found in the first and fourth examples. The model's performance might be biased towards rectangular fall openings due to the prevalence of such shapes in the training data. Similarly, as seen in the second case of Figure 4, when the image exhibited relatively low brightness, the performance significantly dropped, as indicated by a confidence score of 0.5.

Figure 4. Examples of predictions in the test data

5. CONCLUSION

This paper investigated the feasibility of using the Gen-AI approach to augment imagery data and improve computer vision-based object detection performance. Two types of datasets were constructed by collecting real images and using text-to-image tools (i.e., Midjourney), respectively. Imagery data were labeled, and the YOLOv8n algorithm was used to develop a fall-hazard detection model. The testing of the proposed methodology produced several main findings. First, using augmented images generated by Gen-AI along with real images led to performance improvement. Particularly, when 500 augmented images and 100 real imagers were jointly used as input to train the YOLOv8n model, the maximum performance (mAP50) of 95.3% was observed. Second, testing different sizes of augmented images revealed that the overall performance does not directly increase in proportion to the number of input data. This suggests that when utilizing the Gen-AI technique for object detection, identifying an optimal number of augmented data points is critical for various research tasks and applications. Third, the detection model can show a preference for specific types of objects (e.g., rectangular-shaped) due to the overrepresentation of such examples in the training data. In future research, the authors will diversify the classes (e.g., hazard, equipment, and human) that the computer vision model can detect, explore more prompt engineering techniques, and test different computer vision algorithms.

REFERENCES

[1] J. Park, K. Kim, Y.K. Cho, Framework of Automated Construction-Safety Monitoring Using Cloud-Enabled BIM and BLE Mobile Tracking Sensors, J. Constr. Eng. Manage. 143 (2017) 05016019. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001223.

[2] P. Mitropoulos, T.S. Abdelhamid, G.A. Howell, Systems Model of Construction Accident Causation, J. Constr. Eng. Manage. 131 (2005) 816–825. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:7(816).

[3] S. Taneja, B. Akinci, J.H. Garrett, L. Soibelman, E. Ergen, A. Pradhan, P. Tang, M. Berges, G. Atasoy, X. Liu, S.M. Shahandashti, E.B. Anil, Sensing and Field Data Capture for Construction and Facility Operations, J. Constr. Eng. Manage. 137 (2011) 870–881. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000332.

[4] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Advanced Engineering Informatics 29 (2015) 239–251. https://doi.org/10.1016/j.aei.2015.02.001.

[5] Y. Liu, P. Sun, N. Wergeles, Y. Shang, A survey and performance evaluation of deep learning methods for small object detection, Expert Systems with Applications 172 (2021) 114602. https://doi.org/10.1016/j.eswa.2021.114602.

[6] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, L. Chen, Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration, Journal of Information Technology Case and Application Research 25 (2023) 277–304. https://doi.org/10.1080/15228053.2023.2233814.

[7] Y. Ge, J. Xu, B.N. Zhao, N. Joshi, L. Itti, V. Vineet, Beyond Generation: Harnessing Text to Image Models for Object Detection and Segmentation, (2023). https://doi.org/10.48550/ARXIV.2309.05956.

[8] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep Learning for Computer Vision: A Brief Review, Computational Intelligence and Neuroscience 2018 (2018) 1–13. https://doi.org/10.1155/2018/7068349.

[9] B. Yang, B. Zhang, Q. Zhang, Z. Wang, M. Dong, T. Fang, Automatic detection of falling hazard from surveillance videos based on computer vision and building information modeling, Structure and Infrastructure Engineering 18 (2022) 1049–1063. https://doi.org/10.1080/15732479.2022.2039217.

[10] B.E. Mneymneh, M. Abbas, H. Khoury, Automated Hardhat Detection for Construction Safety Applications, Procedia Engineering 196 (2017) 895–902. https://doi.org/10.1016/j.proeng.2017.08.022.

[11] T. Kong, W. Fang, P.E.D. Love, H. Luo, S. Xu, H. Li, Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction, Advanced Engineering Informatics 50 (2021) 101400. https://doi.org/10.1016/j.aei.2021.101400.

[12] D.A. Van Dyk, X.-L. Meng, The Art of Data Augmentation, Journal of Computational and Graphical Statistics 10 (2001) 1–50. https://doi.org/10.1198/10618600152418584.

[13] A. Mumuni, F. Mumuni, Data augmentation: A comprehensive survey of modern approaches, Array 16 (2022) 100258. https://doi.org/10.1016/j.array.2022.100258.
[14] S. Bang, F. Baek, S. Park, W. Kim, H. Kim, Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques, Automation in Construction 115 (2020) 103198. https://doi.org/10.1016/j.autcon.2020.103198.
[15] K.M. Rashid, J. Louis, Times-series data augmentation and deep learning for construction equipment activity recognition, Advanced Engineering Informatics 42 (2019) 100944. https://doi.org/10.1016/j.aei.2019.100944.
[16] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, F. Shen, Image Data Augmentation for Deep Learning: A Survey, (2022). https://doi.org/10.48550/ARXIV.2204.08610.
[17] A. Biswas, N. Md Abdullah Al, A. Imran, A.T. Sejuty, F. Fairooz, S. Puppala, S. Talukder, Generative Adversarial Networks for Data Augmentation, in: B. Zheng, S. Andrei, M.K. Sarker, K.D. Gupta (Eds.), Data Driven Approaches on Medical Imaging, Springer Nature Switzerland, Cham, 2023: pp. 159–177. https://doi.org/10.1007/978-3-031-47772-0_8.
[18] H. Ohno, Auto-encoder-based generative models for data augmentation on regression problems, Soft Comput 24 (2020) 7999–8009. https://doi.org/10.1007/s00500-019-04094-0.
[19] D. BaiDoo-Anu, L. Owusu Ansah, Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning, Journal of AI 7 (2023) 52–62. https://doi.org/10.61969/jai.1337500.
[20] Z.M. Albaghajati, D.M. Bettaieb, R.B. Malek, Exploring text-to-image application in architectural design: insights and implications, Archit. Struct. Constr. 3 (2023) 475–497. https://doi.org/10.1007/s44150-023-00103-x.
[21] C. Cao, Z. Ding, G.-G. Lee, J. Jiao, J. Lin, X. Zhai, Elucidating STEM Concepts through Generative AI: A Multi-modal Exploration of Analogical Reasoning, (2023). https://doi.org/10.48550/ARXIV.2308.10454.
[22] L. Reynolds, K. McDonell, Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 2021: pp. 1–7. https://doi.org/10.1145/3411763.3451760.
[23] A. Borji, Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2, (2022). https://doi.org/10.48550/ARXIV.2210.00586.
[24] J. Terven, D. Cordova-Esparza, A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS, (2023). https://doi.org/10.48550/ARXIV.2304.00501.
[25] T. Chen, Q. Zeng, Research on Bubble Detection Based on Improved YOLOv8n, IEEE Access 12 (2024) 9659–9668. https://doi.org/10.1109/ACCESS.2024.3353196.
[26] Y. Huang, W. Tan, L. Li, L. Wu, WFRE-YOLOv8s: A New Type of Defect Detector for Steel Surfaces, Coatings 13 (2023) 2011. https://doi.org/10.3390/coatings13122011.