ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0455
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Training Dataset Generation through Generative AI for Multi-Modal Safety Monitoring in Construction

Insoo Jeong[1]*, Junghoon Kim[2], Seungmo Lim[3], Jeongbin Hwang[4], Seokho Chi[5]

[1] *Graduate Student, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea,* E-mail address: scout1120@snu.ac.kr
[2] *Graduate Student, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea,* E-mail address: pable01@snu.ac.kr
[3] *Graduate Student, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea,* E-mail address: lssmm818@snu.ac.kr
[4] *Senior Researcher, Institute of Construction and Environmental Engineering (ICEE), Republic of Korea,* E-mail address: jb.hwang@hotmail.com
[5] *Professor, Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea; Adjunct Professor, the Institute of Construction and Environmental Engineering (ICEE), Republic of Korea,* E-mail address: shchi@snu.ac.kr

Abstract: In the construction industry, known for its dynamic and hazardous environments, there exists a crucial demand for effective safety incident prevention. Traditional approaches to monitoring on-site safety, despite their importance, suffer from being laborious and heavily reliant on subjective, paper-based reports, which results in inefficiencies and fragmented data. Additionally, the incorporation of computer vision technologies for automated safety monitoring encounters a significant obstacle due to the lack of suitable training datasets. This challenge is due to the rare availability of safety accident images or videos and concerns over security and privacy violations. Consequently, this paper explores an innovative method to address the shortage of safety-related datasets in the construction sector by employing generative artificial intelligence (AI), specifically focusing on the Stable Diffusion model. Utilizing real-world construction accident scenarios, this method aims to generate photorealistic images to enrich training datasets for safety surveillance applications using computer vision. By systematically generating accident prompts, employing static prompts in empirical experiments, and compiling datasets with Stable Diffusion, this research bypasses the constraints of conventional data collection techniques in construction safety. The diversity and realism of the produced images hold considerable promise for tasks such as object detection and action recognition, thus improving safety measures. This study proposes future avenues for broadening scenario coverage, refining the prompt generation process, and merging artificial datasets with machine learning models for superior safety monitoring.

Key words: Generative Artificial Intelligence, Construction Safety Monitoring, Stable Diffusion, Data Augmentation, Image Generation

1. INTRODUCTION

The construction industry has continuously suffered from safety incidents due to its dynamic and harsh environmental conditions. According to the Korean Government's Ministry of Employment and Labor (MOEL), the construction industry recorded the highest number of fatalities at 539, compared to other industries, with the number of occupational injuries rising to 31,245 over the past few years [1]. To deal with these issues, it's crucial to protect on-site workers and decrease safety risks across projects. Solutions such as managerial on-site inspections and preemptive removal of hazardous materials can be

implemented. However, while these processes may provide useful information, they often depend heavily on human labor, making them time-intensive and cost-inefficient [2]. Moreover, inspection procedures typically result in paper-based checklists and reports, which produce subjective and fragmented data. This leads to significant hurdles for safety management[3].

To surmount these shortcomings, many researchers have conducted studies to harness a wide spectrum of information technology (IT) to automatically monitor hazards in the field and respond adequately. In particular, computer vision technology, which emulates human vision, has attracted attention for its ability to efficiently inspect multiple hazardous materials, situations, and behaviors. With these advantages, various vision-based monitoring methodologies have been introduced to enhance safety in the field. For instance, Wu et al. [4] developed a one-stage convolutional neural network (CNN) system to automatically monitor whether workers are wearing hard hats and to classify their roles based on the color of these hard hats. Yu et al. [5] proposed a non-intrusive method for monitoring workers' physical fatigue by employing a 3D motion capture algorithm, aimed at ensuring occupational health and safety. Despite the numerous approaches utilizing computer vision technologies for safety monitoring, a significant limitation of vision-powered models is their need for appropriate training datasets. However, acquiring safety accident datasets is still difficult due to the limited availability of images or videos that capture safety accidents and the hesitation to distribute this data publicly, driven by concerns over security and violations of personal privacy. Consequently, existing studies have either relied on simulated environments [6] to collect experimental data or have recently turned to creating synthetic datasets to replicate real construction scenes [7]–[9]. Even so, simulating actual construction sites to gather training data requires additional time and financial resources. Furthermore, existing virtual datasets tend to have a rigid quality, limiting their effectiveness in replicating the varied and dynamic accident scenarios found in real construction settings. This rigidity contrasts sharply with the inherently dynamic and complex nature of construction sites, where the diversity and unpredictability of situations can far exceed the capacity of synthetic simulation to adapt, given the time and effort invested.

To overcome these limitations, this paper aims to utilize generative artificial intelligence (AI) to transform the circumstances of construction accidents into training datasets using text-to-image models, specifically Stable Diffusion [10]. Stable Diffusion stands out for its ability to produce photorealistic images from concise text descriptions, known as 'prompts'. [11]. By incorporating detailed object or background information into the prompt, the model is capable of producing closely aligned images with the desired outcome, offering exceptional flexibility and accuracy in visual data generation. Thus, this study applies real construction accident cases and their specific circumstances, inputting them into Stable Diffusion to generate artificial images of construction accidents. The methodology proceeds with the following steps: (1) accident prompts generation, (2) static prompts using empirical trials, and (3) generative AI-powered dataset generation. The primary objective of this study is to validate the qualitative performance of the generated images and evaluate their potential as training datasets for object detection or action recognition tasks. By creating scenarios of real-world construction accidents, this research addresses the significant challenge of the scarcity of datasets related to construction safety. In addition, by employing Stable Diffusion to generate photorealistic images from textual inputs, this methodology offers practical solutions to enhance efficiency in on-site safety management.

2. LITERATURE REVIEW

Safety monitoring through computer vision technology is an actively researched field, with numerous researchers having developed a variety of vision-based techniques to identify various construction hazards from on-site videos. In recent years, the research trend has expanded to include not only traditional deep learning-based algorithms for enhanced accuracy and efficiency but also more advanced models like transformers. For instance, Ding et al. [2], developed a system that integrates a Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) to automatically identify workers' unsafe activities. Moreover, Fang et al. [12] proposed an automated monitoring framework to check the usage of Personal Protective Equipment (PPE) by steeplejacks preparing for aerial work on exterior walls, aiming to prevent falling accidents. Li et al. [13] developed a real-time detection system for helmets, recognizing failures in helmet usage at construction sites. Cheng et al. [14] combined re-identification (ReID) and PPE classification to track workers' site movements and safety. Recently, Fang et al. [15] utilized matching algorithms that connect people’s unsafe behavior with semantic safety rules. Yang et al. [16] developed a Spatial Temporal Relation Transformer (STR-

Transformer), which extracts spatial and temporal features of work behaviors from video streams and identifies unsafe actions in construction projects.

Despite the significant advancements, the development of machine learning algorithms has continued to face challenges due to the scarcity of comprehensive and high-quality training datasets. Fundamentally, the development of reliable models has required a wealth of realistic images, a need that has posed significant challenges in gathering datasets related to safety incidents. These limitations have presented significant barriers to implementing vision-based safety research in real-world scenarios. To solve this problem, researchers have often simulated safety accident scenes and acquired data for training deep learning models, but recently, various methods for building synthetic training datasets have been implemented, thereby addressing the critical gap in data availability. For example, Neuhausen et al. [17] utilized 3D computer graphics software to model virtual construction site scenarios and evaluated the tracking performance of construction workers. Likewise, Lee et al. [9] introduced the Unity game engine to build synthetic data of construction scenarios, facilitating the simulation of diverse safety monitoring cases. Jeong et al. [7] integrated real-world images with synthetic models to inspect tower cranes' working behaviors in installing curtain walls. Previous research has verified the valuable assistance in alleviating the scarcity of real-world datasets; however, current synthetic datasets still showed limited situational simulation and could not reflect the diverse safety accident scenarios that occur at construction sites.

In recent times, generative text-to-image models like DALL-E [18] and Stable Diffusion have emerged as groundbreaking alternatives, revolutionizing the field of image generation. These models employed deep learning techniques to produce high-quality, varied images from textual descriptions, presenting a new paradigm in content creation that merges art with technology. Built on the diffusion model framework, which excels in creating photorealistic images, Stable Diffusion represents a significant departure from traditional generative adversarial networks (GANs) [19]. Instead of relying on the competitive dynamics between a generator and a discriminator, diffusion models transform a random noise distribution into one that mirrors the training data through a two-phase process. This includes a forward phase, where noise is progressively added to the data until it becomes entirely random, followed by a reverse phase, where the model learns to remove this noise, effectively generating coherent images from textual inputs. The innovation of Stable Diffusion lies in its ability to adjust the generative process to specific textual descriptions, facilitated by the integration of a transformer-based text encoder. This encoder processes the inputs and guides the diffusion process to create images that closely align with complex narratives. Thus, Stable Diffusion combines the image generation prowess of diffusion models with the advanced natural language processing capabilities of transformers. This study leverages the Stable Diffusion model to interpret complex real-world accident scenarios and generate corresponding images. In this process, it demonstrates the model's utility in creating accurate visual representations from detailed textual descriptions, underscoring its potential across various applications.

3. RESEARCH METHODOLOGY

This research aims to generate an artificial dataset by incorporating real-world accident scenarios into the Stable Diffusion model. As illustrated in Figure 1, the methodology consists of three main steps: (1) generating accident prompts from real accident cases; (2) combining static prompts through empirical trials; and (3) generating the dataset with Stable Diffusion. The details of each step are further explained in the following sections.

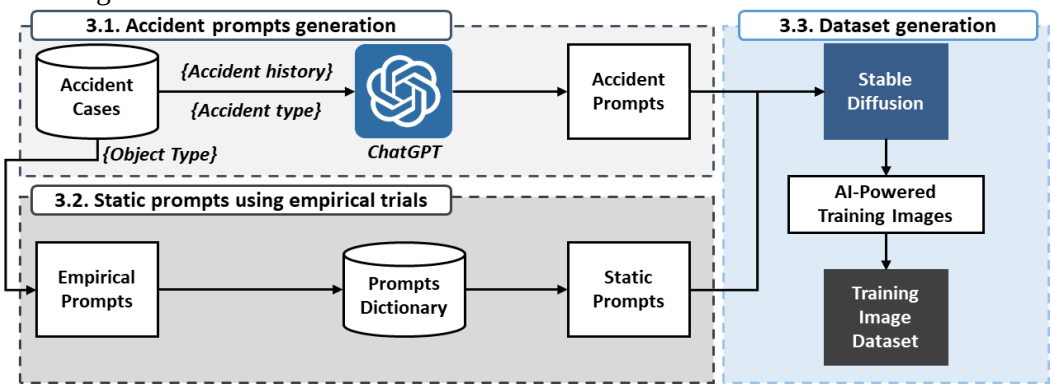


Figure 1. Overview of the research methodology.

3.1. Accident prompts generation

The leading focus of this section is to explain a novel technique for generating 'positive prompts' aimed at improving safety situations in construction environments through the application of Stable Diffusion models. This initiative is grounded in the systematic analysis of real-world accident cases meticulously documented in the Korean Government's Construction Safety Management Integrated Information System (CSI) [20]. The dataset encompasses a comprehensive range of parameters, including the names of construction projects, geographical locations, precise timings of incidents, the nature of construction activities, categorizations of accidents (e.g., falls from heights, slips, impacts from objects, etc.), accompanied by detailed narratives that describe the circumstances preceding each incident (hereinafter referred to as 'accident history information').

To transform this raw accident history information into practical text 'prompts', the authors leveraged the capabilities of ChatGPT [21] to examine and interpret the textual accident reports. This process entails the structuring of accident history information into specific formats, as exemplified in Figure 2 [22], to facilitate the effective generation of 'prompts' that are customized for the Stable Diffusion model. For example, the accident case 'a worker slipped from the work platform while descending from the scaffolding and fell 2 meters down' was changed into standardized prompt: *'Afternoon masonry work completion, a worker slips and falls 2 meters down from the scaffolding during descent, observed by a bystander, ultra-realistic, high-detail capture of the moment of slip, dramatic contrast, focused on the expressions of both the worker and observer, construction site ambiance, (scaffolding:1.3), (moment of fall:1.5), (worker's uniform:1.2), (observer's reaction:1.3), cinematic lighting, 4K resolution, (emergency situation:1.25), (alert:1.3), (safety gear:1.1)'* In this prompt, the numerical values represent the relative importance or weight of each element, indicating their significance in the context of fall accidents. The text description, on the other hand, outlines critical conditions or situations commonly associated with fall incidents on construction sites.

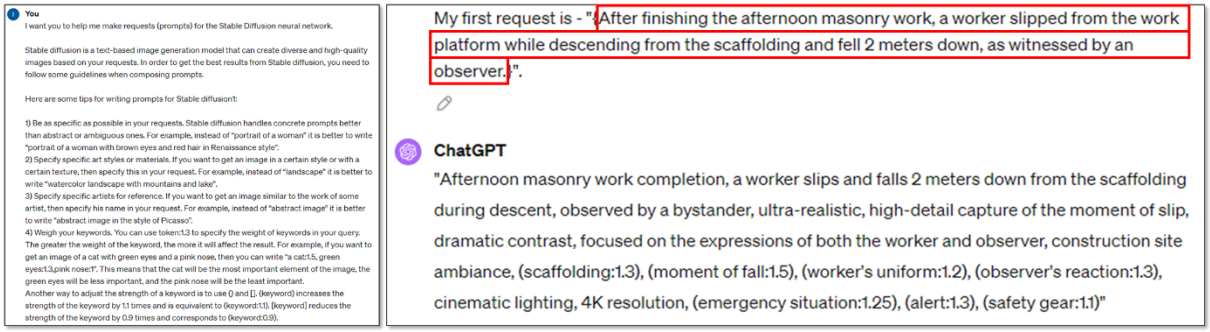
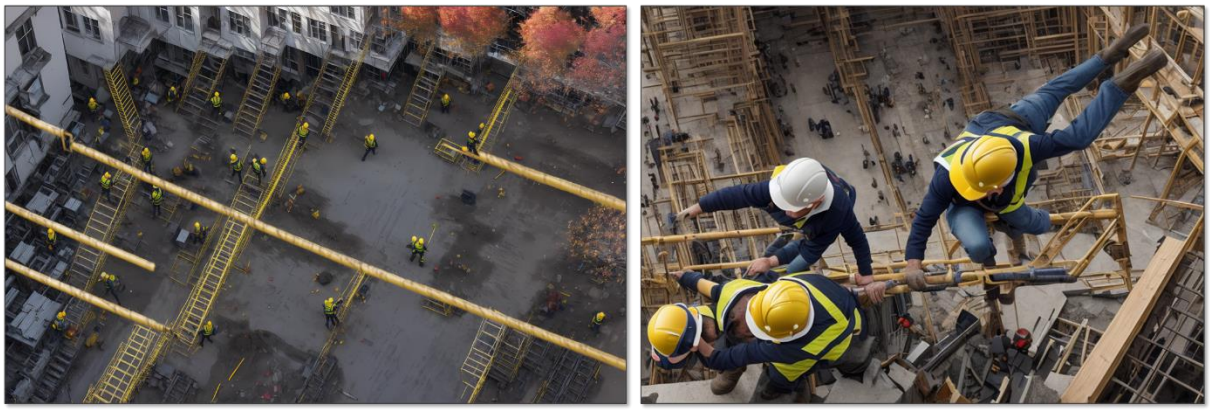


Figure 2. Examples of generating accident prompts with ChatGPT.

3.2. Static prompts using empirical trials

This section is dedicated to the development of static prompts that have a significant impact on the image quality generated by Stable Diffusion. In this research, a distinction is made between two types of static prompts: positive prompts and negative prompts. This distinction is crucial as, traditionally, the creation of text-to-image prompts involves both positive and negative prompts to accurately generate desired outcomes. Positive prompts in diffusion text-to-image models guide the generation towards desired attributes or themes, specifying what should be included in the image. Negative prompts, conversely, instruct the model on what to avoid, helping to exclude unwanted elements or characteristics from the generated image. The selection of these prompts is based on extensive empirical trials to evaluate their effectiveness before and after their application.

The absence of either type of prompt can result in the production of inferior images. For instance, in Figure 3, the image on the left displays a viewpoint that is too distant, failing to accurately represent the desired accident scene. Conversely, the image on the right suffers from incorrect anatomical form and proportions of the worker, making the image unrecognizable and unusable. Therefore, by adding *'{full body angle}'* to the positive prompt and *'{bad anatomy, bad proportions}'* to the negative prompt, it is possible to avoid such defects in the generated data. Accordingly, our research team has developed a customized prompt dictionary specifically designed for construction site scenarios. By applying this specialized dictionary with Stable Diffusion, it has been possible to significantly enhance the quality of the generated images



Without positive prompts: {full body angle} *Without negative prompts: {bad anatomy, bad proportions}*

Figure 3. Generated images without static prompts.

3.3. Dataset generation with Stable Diffusion

The purpose of this section is to generate images using the prompts defined in the previous steps. For this task, the Stable Diffusion Web UI project [23] was adopted, providing access to the Stable Diffusion model through a user-friendly interface. This Web UI offers advanced features like facial restoration correction, image upscaling, and textual inversion, enabling the personalization of prompts for generating highly specific and detailed images. To enhance quality of outputs, the research team used a model checkpoint that was pretrained on realistic architecture [24]. It begins with the selection of a pretrained checkpoint. Subsequently, users input the positive and negative prompts, as identified in previous sections, into their designated fields. This is followed by the adjustment of hyperparameters such as batch size, upscaling factors, and sampling methods. After these steps are completed, users can proceed to generate the images.

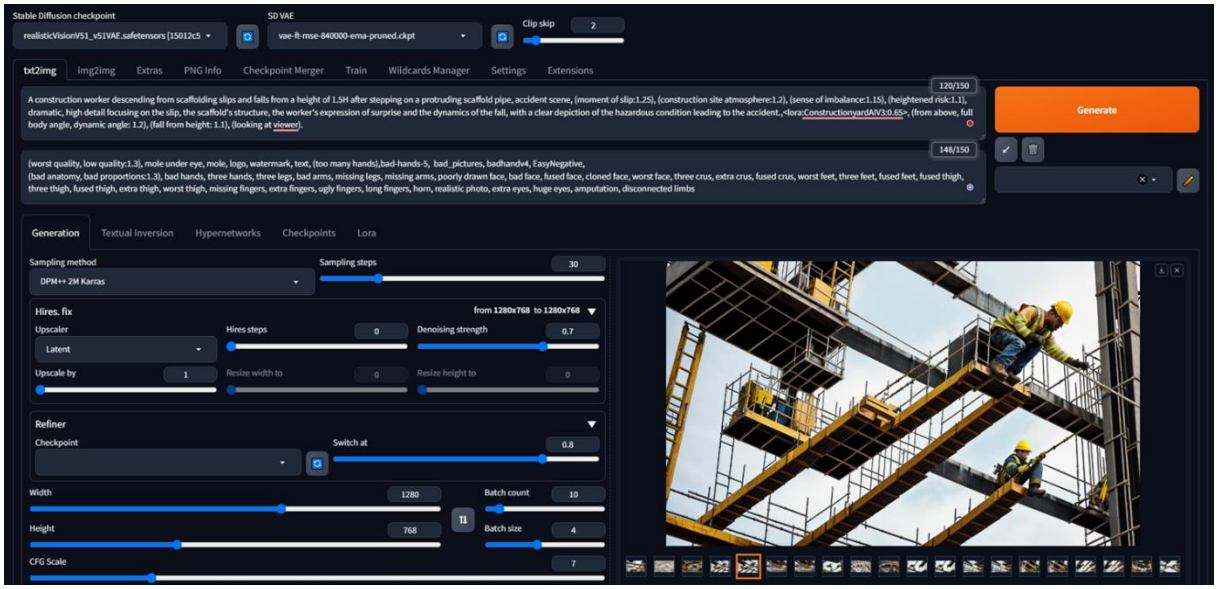


Figure 4. Examples of inputting prompts and adjusting parameters in the Stable Diffusion Web UI.

4. PRELIMINARY RESULTS AND DISCUSSION

To validate the proposed methodology, 'fall accident' cases from the CSI database were used. A comprehensive analysis of 2,324 fall accident scenarios was conducted using the ChatGPT (GPT-4 model), which aided in the generation of prompts for image synthesis. This process yielded 300 artificial images with a resolution of 1280×720 pixels, after applying optimized settings. These images, exemplified in Figure 5, were produced employing the Stable Diffusion v1.5 model, utilizing 100 reverse diffusion steps and a classifier-free guidance (CFG) scale ranging from 8 to 14. Additionally, a face detection and restoration model was employed to enhance and align faces within the generated images, ensuring realism and consistency, as demonstrated in Figure 5.

Figure 6 showcases the efficiency of the proposed methodological pipeline in synthesizing images that accurately depict 'fall hazards,' achieving quality comparable to that of traditional 3D-based game engine models more efficiently. Stable Diffusion has proven its ability to generate images that not only demonstrate each fall hazard scenario with a high degree of realism but also closely simulate actual real-world situations. For example, in depicting a 'fall hazard' scenario, details such as the posture of workers operating at heights without safety harnesses, the distance from unprotected edges, and the absence of necessary safety equipment were prominently featured, reflecting real-world conditions. Furthermore, the model showed exceptional flexibility in representing various aspects of construction sites, including camera angles, compositions, on-site objects, and project statuses. This adaptability stemmed from its ability to adjust these elements freely based on textual inputs, enabling the creation of images that closely resemble real-life scenarios. Moreover, by refining viewpoint-related prompts, images providing varied perspectives on the same hazard scenarios were produced, significantly enhancing the dataset's value and utility.

While the results are promising, there is potential for further research to explore the integration of these artificial images with real-world data for training and validating machine learning models. Such an endeavor could significantly enhance their capability to recognize and predict fall hazards in construction settings.

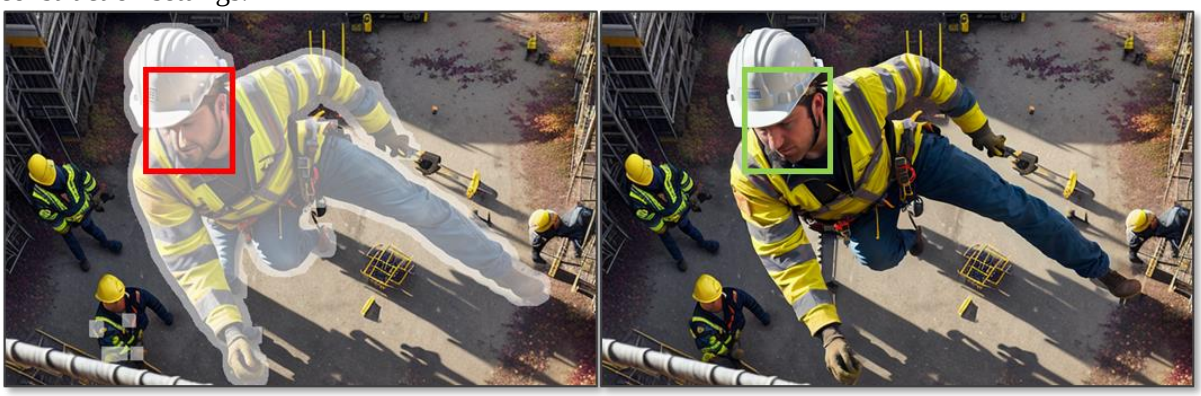


Figure 5. Examples of face detection and restoration model.

Figure 6. Examples of generated training images ‘fall hazard’.

5. CONCLUSION

This paper introduces a novel approach to address the scarcity of safety-related datasets in the construction industry through the innovative use of generative AI, specifically the Stable Diffusion

model. By employing real-world accident scenarios as inputs for generating photorealistic images, the authors have effectively highlighted the capability of this technology to produce valuable training datasets. These datasets are notable for their diversity and realism, providing significant contributions to the enhancement of safety monitoring and management practices within construction environments

The methodology, which includes generating accident prompts, strategically employing static prompts in empirical trials, and creating datasets with Stable Diffusion, has been shown to successfully navigate the challenges associated with conventional data collection methods. The reliable quality of the generated images and their suitability for tasks such as object detection or action recognition underscore the benefits of leveraging generative AI for improving safety protocols.

In future, the research will aim to broaden the spectrum of scenarios explored, fine-tune the prompt generation process to achieve even more precise image synthesis, and utilize the produced datasets to further training machine-learning models dedicated to safety monitoring. This endeavor not only promises to advance our understanding and application of generative AI in the field of construction safety but also sets the stage for significant improvements in preventive measures and safety management strategies.

ACKNOWLEGEMENTS

This work was supported by the National R&D Project for Smart Construction Technology (23SMIP-A158708-04) funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure, and Transport. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00241758 and No. 2021R1A2C2003696).

REFERENCES

[1] Ministry of Employment and Labor, "2022 Industrial Accident Statistics," 2022. Accessed: Feb. 13, 2024 [Online]. Available: https://www.moel.go.kr/policy/policydata/view.do?bbs_seq=20230300058.

[2] L. Ding, W. Fang, H. Luo, P. E. D. Love, B. Zhong, and X. Ouyang, "A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory," *Autom. Constr.*, vol. 86, pp. 118–124, 2018, doi: https://doi.org/10.1016/j.autcon.2017.11.002.

[3] Q. Xu, H. Y. Chong, and P. C. Liao, "Collaborative information integration for construction safety monitoring," *Autom. Constr.*, vol. 102, no. January, pp. 120–134, 2019, doi: 10.1016/j.autcon.2019.02.004.

[4] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset," *Autom. Constr.*, vol. 106, no. July, p. 102894, 2019, doi: 10.1016/j.autcon.2019.102894.

[5] Y. Yu, H. Li, X. Yang, L. Kong, X. Luo, and A. Y. L. Wong, "An automatic and non-invasive physical fatigue assessment method for construction workers," *Autom. Constr.*, vol. 103, no. August 2018, pp. 1–12, 2019, doi: 10.1016/j.autcon.2019.02.020.

[6] K. Yang, C. R. Ahn, M. C. Vuran, and S. S. Aria, "Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit," *Autom. Constr.*, vol. 68, pp. 194–202, 2016, doi: 10.1016/j.autcon.2016.04.007.

[7] I. Jeong, J. Hwang, J. Kim, S. Chi, B.-G. Hwang, and J. Kim, "Vision-Based Productivity Monitoring of Tower Crane Operations during Curtain Wall Installation Using a Database-Free Approach," *J. Comput. Civ. Eng.*, vol. 37, no. 4, pp. 1–14, 2023, doi: 10.1061/jccee5.cpeng-5105.

[8] J. Hwang, J. Kim, S. Chi, and J. Seo, "Development of training image database using web crawling for vision-based site monitoring," *Autom. Constr.*, vol. 135, p. 104141, 2022, doi: https://doi.org/10.1016/j.autcon.2022.104141.

[9] H. Lee, J. Jeon, D. Lee, C. Park, J. Kim, and D. Lee, "Game engine-driven synthetic data generation for computer vision-based safety monitoring of construction workers," *Autom. Constr.*, vol. 155, no. August, p. 105060, 2023, doi: 10.1016/j.autcon.2023.105060.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 10674–10685, 2022, doi: 10.1109/CVPR52688.2022.01042.

[11] H. Moreno, A. Gómez, S. Altares-López, A. Ribeiro, and D. Andújar, "Analysis of Stable Diffusion-derived fake weeds performance for training Convolutional Neural Networks," *Comput. Electron. Agric.*, vol. 214, no. August, 2023, doi: 10.1016/j.compag.2023.108324.

[12] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, and C. Li, "Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment," *Autom. Constr.*, vol. 93, no. May, pp. 148–164, 2018, doi: 10.1016/j.autcon.2018.05.022.

[13] Y. Li, H. Wei, Z. Han, J. Huang, and W. Wang, "Deep Learning-Based Safety Helmet Detection in Engineering Management Based on Convolutional Neural Networks," *Adv. Civ. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/9703560.

[14] J. C. P. Cheng, P. K. Y. Wong, H. Luo, M. Wang, and P. H. Leung, "Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification," *Autom. Constr.*, vol. 139, no. May, p. 104312, 2022, doi: 10.1016/j.autcon.2022.104312.

[15] W. Fang, P. E. D. Love, L. Ding, S. Xu, T. Kong, and H. Li, "Computer Vision and Deep Learning to Manage Safety in Construction: Matching Images of Unsafe Behavior and Semantic Rules," *IEEE Trans. Eng. Manag.*, vol. 70, no. 12, pp. 4120–4132, 2023, doi: 10.1109/TEM.2021.3093166.

[16] M. Yang *et al.*, "Transformer-based deep learning model and video dataset for unsafe action identification in construction projects," *Autom. Constr.*, vol. 146, no. June 2022, p. 104703, 2023, doi: 10.1016/j.autcon.2022.104703.

[17] M. Neuhausen, P. Herbers, and M. König, "Using synthetic data to improve and evaluate the tracking performance of construction workers on site," *Appl. Sci.*, vol. 10, no. 14, 2020, doi: 10.3390/app10144948.

[18] A. Ramesh *et al.*, "Zero-Shot Text-to-Image Generation," *Proc. Mach. Learn. Res.*, vol. 139, pp. 8821–8831, 2021.

[19] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[20] "Construction Safety Management Integrated Information," Accessed: Feb. 13, [Online]. Available: https://www.csi.go.kr/index.do.

[21] "ChatGPT." Accessed: Feb. 13, [Online]. Available: https://chat.openai.com/.

[22] "Stable Diffusion generation." Accessed: Feb. 13, [Online]. Available: https://github.com/salire123/Chatgpt-Prompt-by-salire/blob/main/Stable-Diffusion-generation.md.

[23] "Stable Diffusion Webui," Accessed: Feb. 13, [Online]. Available: https://github.com/AUTOMATIC1111/stable-diffusion-webui/.

[24] "Realisitc Architecture." Accessed: Feb. 13, [Online]. Available: https://civitai.com/models/53493/constructionyardai-konyconi.