Enhancing Water-Ponding Detection in Housekeeping Using Generative AI Inpainting–Based Synthetic Data

Dinh Quang Duy[1], Zhu Xuming[1], Jing Tian[2], Yang Miang Goh[1]

[1]Department of the Built Environment, College of Design and Engineering, National University of Singapore, 4 Architecture Drive, 117566, Singapore
[2]NUS-ISS, National University of Singapore, 25 Heng Mui Keng Terrace, 119615, Singapore
e1583361@u.nus.edu, e0688540@u.nus.edu, tianjing@nus.edu.sg, bdggym@nus.edu.sg

Abstract -

Water ponding or puddle is one of the key factors contributing to poor housekeeping conditions on construction sites, which can significantly hinder safety, operational efficiency, and creating favourable conditions for mosquito breeding. Traditional housekeeping monitoring methods, which rely on manual observations, are often ineffective. Moreover, although computer vision has shown promise in automating housekeeping safety monitoring, its effectiveness is frequently limited by the difficulty of collecting diverse and representative datasets. To address this challenge, we introduce a framework that integrates the strengths of zero-shot models with diffusion-based inpainting to generate synthetic images in a controlled and realistic manner that accurately reflects on-site conditions. The original water ponding dataset, combined with the synthetic data generated using the proposed method, is used to train the object detection model. By incorporating the synthetic housekeeping data into the original dataset, the model's detection performance improved substantially, yielding 8.0% increase in mAP@50–95 and 9.1% increase in mAP@50. This underscores the potential of synthetic datasets to enhance detection performance not only for water ponding but also for other housekeeping-related objects, thereby contributing to improved safety on construction sites.

Keywords –

Diffusion Model, Image inpainting, Zero-shot models, Safety Monitoring, Housekeeping.

1 Introduction

According to estimates from the World Health Organization (WHO) and the International Labour Organization (ILO), work-related diseases and injuries cause approximately 1.9 million deaths annually, with construction sites being among the highest-risk workplaces contributing substantially to occupational fatalities[1][2]. Good housekeeping on construction sites, characterized by maintaining an organized, clean, and hazard-free environment [3], is a fundamental factor in enhancing safety and reducing workplace accidents. Therefore, maintaining good housekeeping conditions on construction sites is essential to ensure safety, operational efficiency, and unobstructed task performance by workers. The current practice of housekeeping observation primarily depends on manual supervision by site personnel. While this method is simple to apply on construction sites, manual inspection remains inefficient due to its reliance on human judgment, requiring substantial time and effort [4]. Thus, the development of an automated and effective method for maintaining good housekeeping remains essential for improving safety inspections on construction sites.

Recent advances in computer vision (CV) have highlighted deep learning (DL) as a powerful and reliable approach for automating the detection of housekeeping issues on construction sites. Although beneficial, There are limited studies applied the advantages of DL for housekeeping monitoring [3][5]. For example, Lim et al. [5] proposed a self-supervised learning method to classify construction site images into good and poor housekeeping categories, achieving state-of-the-art performance. Although effective, limited in their capacity to identify the underlying causes of poor housekeeping. Moreover, this model can struggle to generalize to unseen construction sites due to annotation bias, as the training data are typically collected and manually labelled from limited areas views, failing to capture the variability inherent in diverse site environments. This demonstrates the demand for developing more robust and adaptable methods to detect and localize factors contributing to poor housekeeping in construction sites.

Object detection has emerged as a promising approach for advancing the automation of object recognition in construction site environments [6]. Owing to their strong detection and localization capabilities,

object detection models have been successfully applied by researchers to identify construction-related objects on site [7][8]. Although effective, the application of object detection models in housekeeping monitoring remains limited mostly due to their heavy reliance on large volumes of high-quality data to achieve high performance, the collection of which is often labour-intensive, time-consuming, and prone to human error. To address this challenge, numerous researchers have utilized various techniques to generate synthetic data that simulate real-world conditions, aiming to enhance the performance of object detection models. For example, Song et al. [6] enhanced object detection performance for equipment monitoring by leveraging zero-shot models to generate synthetic data from site-specific construction equipment datasets and construction site images. Specifically, regions of interest corresponding to construction equipment are segmented from site-specific equipment datasets, then scaled and randomly composited onto construction site background images. Although significant improving the performance of the object detection model, random pasting can result in unrealistic data, which can adversely affect the object detection model. Meanwhile, Lee et al. [9] introduced a novel approach that employs generative AI to augment image data, thereby improving hazard detection performance. Although this method ensures the quality and realism of the generated synthetic data and significantly enhances the performance of object detection models, it can be financially expensive, as generating large volumes of training images through generative AI requires substantial monetary resources. Overall, current synthetic data generation methods face two key challenges: (1) limited realism in methods such as random pasting, and (2) considerable financial costs in generative AI-based methods. These limitations can hinder the effectiveness of synthetic data generation approaches, thereby highlighting the need to develop more comprehensive methods that balance cost-effectiveness with improvements in model performance.

To address the limitations of current state-of-the-art synthetic data generation approaches, this study proposes an integrated framework that combines zero-shot DL models with generative AI to improve both the accuracy and realism of synthetic data for housekeeping object detection, with a particular focus on water ponding. The key contribution of this research lies in the integration of prompt engineering with generative AI, which enables the realistic generation of water ponding images for model training and performance enhancement. Moreover, the integration of zero-shot models to guide generative AI enables accurate and efficient water ponding synthesis without requiring manual labelling, thereby ensuring high-quality and scalable synthetic data generation that enhances the performance of the object detection model. Finally, the study evaluates the effectiveness of synthetic data and the significance of original data by training on the state-of-the-art YOLOv11 [10] and analysing their comparative performance. Our main contributions are:

- We propose an integrated conceptual framework that combines zero-shot DL models with generative AI for synthetic data generation in housekeeping object detection.
- We incorporate prompt engineering and zero-shot guidance to improve the accuracy and realism of synthetic data generation.
- We evaluated the effectiveness of synthetic data and the role of original data through comparative experiments using YOLOv11.

The remainder of this paper is organized as follows. **Section 2** presents the background of the techniques employed in this study, followed by the proposed methodology in **Section 3**. After that, **Section 4** illustrates the experimental results, and finally **Section 5** concludes and presents the future work.

2 Background

2.1 Generative AI

Generative AI is a subfield of artificial intelligence that enables machines to generate novel content, such as images, through the understanding and reproduction of patterns learned from existing datasets [11]. There are various generative AI techniques, including text-to-image, image-to-image and inpainting, each designed to create or modify visual content based on different types of input data. Numerous studies have applied text-to-image, image-to-image to generate synthetic data with the aim of enhancing the performance of object detection models in the construction industry [17][18]. Despite its effectiveness, limited research has explored the use of inpainting for synthetic data generation. Diffusion-based inpainting generates missing image regions from textual prompts, producing realistic content consistent with the prompt and surrounding context. The limited exploration of this technique is likely due to its requirement for a binary segmentation mask as input, which is typically produced by supervised learning models that depend on large, labelled datasets, making the process labour-intensive and time-consuming. Hence, there is a clear need to develop more efficient and robust approaches to overcome this limitation.

2.2 Grounded SAM

Grounded Segment Anything (Grounded-SAM) [14] is a zero-shot segmentation framework that addresses the labelling limitations of traditional supervised DL by integrating Grounding DINO [15] and Segment

Anything (SAM) [16] into a unified pipeline. The overview of the framework that integrates Grounding DINO and SAM is illustrated in **Fig. 1**. In particular, Grounding DINO builds upon DINO [17] for visual representation learning and GLIP [18] for language understanding, integrating both modalities within a transformer-based framework. Through self-attention and cross-attention mechanisms, the model establishes strong correlations between visual and textual features. A language-guided query selection module further aligns these features with textual prompts to achieve accurate object localization, and a cross-modality decoder subsequently fuses image and text information via cross-attention for grounding results. As such, by inputting the image and textual prompt, the model accurately detects the corresponding objects without the need for additional training and labelling.

The SAM, developed by Meta is a large-scale foundation model trained on billions of masks across millions of images, capable of segmenting objects without task-specific retraining by utilizing various input prompts such as points and bounding boxes. The image encoder employs the ViT-H architecture to extract features from the input image, representing them as 256-channel embeddings. In parallel, the prompt encoder transforms different types of input prompts into prompt embeddings. These embeddings are then fused in the mask decoder using a cross-attention mechanism, which enables the model to effectively attend to image regions associated with the given prompts. Thus, by using the bounding box of the object obtained from the Grounding DINO model, SAM effectively segments the corresponding objects within the image without requiring additional training or manual annotation.

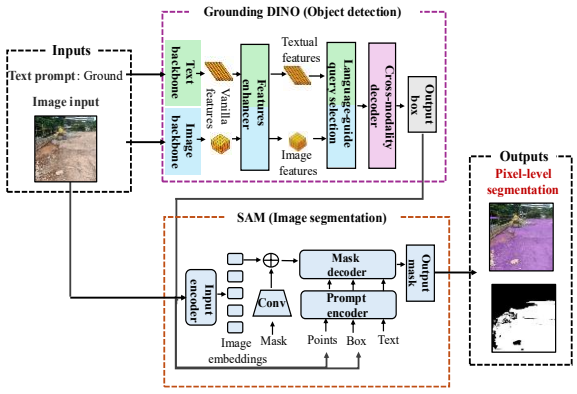


Figure 1. The overview of the integration of Grounding Dino and SAM

3 Methodology

To improve object detection performance for detecting water ponding in construction sites, this study developed an automated approach for generating synthetic water ponding images using a diffusion-based inpainting model. The overview of proposed method is illustrated in **Fig. 2**. Beginning with image collection and preprocessing, the ground regions in the images are subsequently detected and segmented using Grounded SAM. Finally, the Stable Diffusion model, a form of generative AI, combined with prompt engineering fills the segmented areas with realistic water ponding, enabling the generation of high-quality synthetic images for model training.

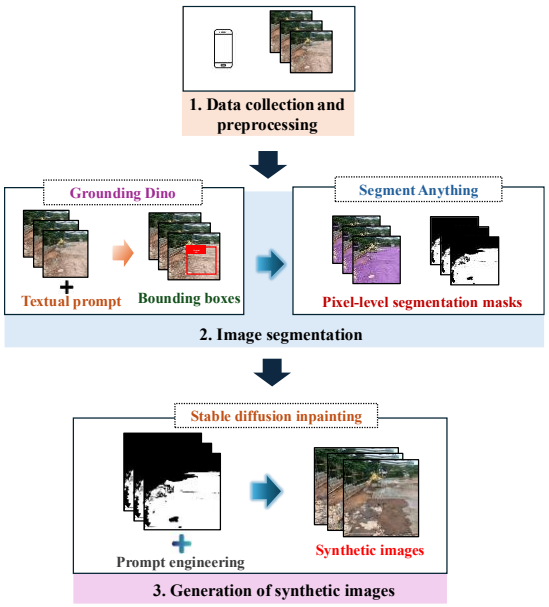


Figure 2. The overview of the proposed method

3.1 Data collection and preprocessing

This section describes the data acquisition and preprocessing procedures designed to collect and processed the construction site data used for generating synthetic images. To begin with, data acquisition is conducted using mobile phones to capture images of on-site activities. These devices enable flexible data collection from diverse viewpoints and under varying lighting conditions across complex construction environments, facilitating the generation of diverse synthetic data for enhancing model robustness and generalization. However, since manual image collection by site personnel often produces low-quality or blurry images, this study applies a Laplacian filter [19] to identify and remove such images, ensuring that only clear images are used for subsequent processing. Given $\boldsymbol{P(x, y)}$ is the image intensity, where $\boldsymbol{x}$ and $\boldsymbol{y}$ denote the horizontal and vertical pixel coordinates are, respectively, the Laplacian value $\boldsymbol{\nabla^2 P}$ for each pixel to assess image sharpness by measuring the rate of intensity variation across neighbouring pixels, as defined by:

$$\boldsymbol{\nabla^2 P} = \frac{\boldsymbol{\partial^2 P}}{\boldsymbol{\partial x^2}} + \frac{\boldsymbol{\partial^2 P}}{\boldsymbol{\partial y^2}} \qquad (1)$$

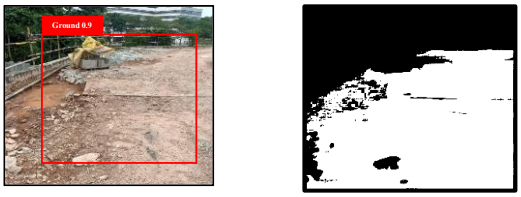
To evaluate overall image quality, the variance of the Laplacian is used:

$$\mathrm{Var}(\boldsymbol{P}) = \frac{\boldsymbol{1}}{\boldsymbol{M \times N}} \sum_{\boldsymbol{x=1}}^{\boldsymbol{M}} \sum_{\boldsymbol{y=1}}^{\boldsymbol{N}} \left(\boldsymbol{P(x, y)} - \overline{\boldsymbol{P}}\right)^{\boldsymbol{2}} \qquad (2)$$

where M and N are image height and width, respectively. $\bar{P}$ denotes the mean Laplacian value. If the variance of the Laplacian within images is lower than a predefined threshold of 150, which is empirically selected, these images are regarded as blurred and is therefore eliminated from the dataset. Finally, the high-quality images are classified into water ponding-related good and poor housekeeping categories. In our previous study [5], Swin Transformer achieved 84% classification accuracy on a large-scale construction image dataset. This pretrained model is therefore employed in the present study to ensure robust housekeeping classification and support the generation of high-quality synthetic data.

3.2 Image segmentation

This section presents the image segmentation framework designed to identify ground regions in good housekeeping images for generating synthetic poor housekeeping data through water ponding synthesis. As discussed in **Section 2.2**, Grounded SAM is employed, as this model enables effective segmentation of previously unseen objects without the need for additional annotations or model retraining. To start, the textual prompt "Ground" and the good housekeeping images are provided as inputs to the model. Grounding DINO first performs object detection to generate bounding boxes representing potential ground regions in the images, as illustrated in **Fig. 3(a)**. To improve the robustness of the model, only the bounding boxes achieving the highest Intersection over Union (IoU) scores, representing the most accurate alignment between predicted and true regions, are preserved. This process ensures the accurate identification of ground regions in good housekeeping images, enabling accurate pixel-level segmentation by SAM in the subsequent stage. Following object detection, SAM performs object segmentation to generate pixel-level masks of the identified ground regions. Based on the bounding box coordinates predicted by Grounding DINO, SAM performs segmentation to effectively obtain pixel-level masks corresponding to the detected ground regions, as illustrated in **Fig. 3(b)**.



(a) The result of Grounding Dino

(b) The result of SAM

Figure 3. The pixel-level mask result

3.3 Generation of the synthetic poor housekeeping data

This section presents a diffusion-based inpainting approach for generating synthetic water ponding housekeeping images to enhance the performance of the object detection model. By leveraging prompt engineering and the Stable Diffusion model which is an open-source model, the method enables controlled and realistic synthesis of water ponding within the segmented ground regions. Beginning with the binary mask generated from SAM, contour analysis [20] is applied to identify the largest continuous regions within each mask, ensuring precise selection of ground areas for generative filling while removing isolated regions that may lead to unrealistic artifact during synthetic generation. The example of clustered region removals is illustrated in **Fig. 4**.

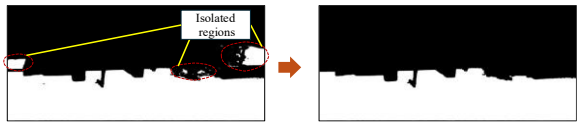


Figure 4. Removal of isolated regions

Subsequently, the Stable Diffusion model is employed to synthesize water ponding within the white regions of the mask. The model takes three inputs: the original image, the corresponding binary mask, and a textual prompt guiding the generation process. In this study, both image-guided prompts, which direct the model to generate desired water ponding features, and negative prompts, which constrain undesired artifacts or unrealistic elements during synthesis, are utilized. The design of the prompts used in this study is summarized in **Table 1**.

Table 1: The prompts designed for generating synthetic water ponding–related poor housekeeping images.

Prompt: "small shallow puddles on concrete ground on the foreground, thin water film, micro or small ripples, Fresnel reflections, darker wet concrete, micro-pitting on concrete, subtle roughness, photorealistic, correct perspective, natural shadows, high detail, high dynamic range"
Negative prompt: "people, text, watermark, logo, blurry, soft, noisy, grain, cartoon, painting, plastic sheen, rubbery, over sharpened edges, repeated textures, artifacts"

By integrating both prompts and negative prompts into the model, the corresponding white regions in the images are synthesized in a controlled manner that enhances visual realism while reducing unwanted artifacts. **Fig. 5** illustrates examples of synthetic images generated by the Stable Diffusion model guided by the designed prompts. Therefore, the synthetic data can be effectively integrated with real-world data for training object detection models, significantly improving detection accuracy and enhancing model generalization across different background conditions within the construction site.

Figure 5. Water ponding synthetic images generated by the proposed framework

4 Experimental results

4.1 Dataset and experimental setup

To validate the effectiveness of the proposed diffusion-based inpainting framework, both the synthetic and actual datasets are utilized to train the object detection model. Specifically, after the data collection and preprocessing stages, the dataset comprised 1011 water ponding–related good housekeeping images and 484 poor housekeeping images. Based on the good housekeeping images, 2000 synthetic water ponding images are automatically generated using the proposed method. Stable Diffusion XL was employed for the image generation process, with a total generation time of approximately 5 hours. These synthetic images, along with the actual poor housekeeping images, are manually annotated to ensure labelling accuracy and consistency. Subsequently, these annotated images are utilized to train the object detection model. To examine the impact of the synthetic dataset, a series of incremental training experiments is conducted. First, the state-of-the-art YOLOv11 [10] object detection model is trained using only the original dataset, consisting of 400 images for training and 84 images for validation, which serves as the baseline. Subsequently, synthetic water ponding images generated by the proposed method are progressively added to the training set in three stages: an additional 500 synthetic images, then another 500, and finally 1000 more. This stepwise strategy enables a systematic assessment of how increasing volumes of synthetic data influence detection performance.

All experimental procedures are performed on a workstation configured with a 14th Gen Intel® Core™ i9-14900K processor and dual NVIDIA GeForce RTX 4090 GPUs (24 GB VRAM each).

4.2 Performance metrics

The model is evaluated using the mean Average Precision (mAP) metric, a standard measure for object

detection performance, including mAP@50 and mAP@50-95. The mAP is computed based on the area under the Precision-Recall (PR) curve for each class and then averaged across all classes. Formally, the Average Precision (AP) for a given class is defined as:

$$\mathrm{AP} = \int_0^1 p(r)dr \qquad (3)$$

where p(r) is the precision as a function of recall r. Meanwhile, mAP@50–95 is computed as the average AP across multiple IoU thresholds ranging from 0.50 to 0.95 in increments of 0.05, providing a more stringent assessment of both detection accuracy and localization precision.

4.3 Results

The effectiveness of the synthetic image dataset is validated through a comparative evaluation of the model's performance when trained on combined datasets of synthetic and original images versus training on the original labelled dataset alone.

Table 2: Experimental results of different datasets trained on YOLOV11

	mAP@50-95	**mAP@50**
Original images	40.5	64.2
Original & 500 synthetic images	46.8	67.4
Original & 1000 synthetic images	**48.5**	**73.3**
Original & 2000 synthetic images	48.4	71.6

Table. 2 and **Fig. 6** present the experimental results and performance comparison of YOLOv11 trained on different datasets. The results show that incorporating synthetic images consistently enhances detection performance relative to training solely on the original dataset. The baseline model trained on original images achieved mAP@50-95 of 40.5 and mAP@50 of 64.2. After adding 500 synthetic images, performance increased by 6.3% to 46.8 (mAP@50-95) and by 3.2% to 67.4 (mAP@50). Subsequently, incorporating 1,000 synthetic images further improved performance, by 8.0% in mAP@50-95 and 9.1% in mAP@50, yielding the highest mAP@50 value of 73.3 and an mAP@50-95 of 48.5 compared to the baseline. Although using 2,000 synthetic images produced a comparable mAP@50-95 of 48.4, the mAP@50 slightly decreased to 71.6, indicating that incorporating 1000 synthetic images provides the most substantial overall performance gain.

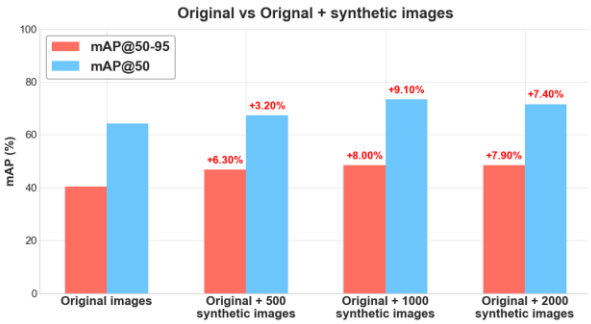


Figure 6. Performance comparison between the original dataset and the original dataset enhanced with synthetic images

4.4 Qualitative visualization results

To highlight the improvement in object detection performance contributed by the synthetic data, a visualization experiment is conducted using real-world images of water ponding–related poor housekeeping conditions collected from construction sites. As illustrated in **Fig. 7**, the visual comparison shows that the fine-tuned model trained on both original and synthetic data achieves higher detection accuracy and greater confidence in identifying water ponding across diverse construction site conditions, whereas the baseline model trained only on original data sometimes misses detections or produces lower confidence scores. This improvement can be attributed to the added data diversity and contextual variation introduced by the synthetic images, which enhances the model's capability to detect water ponding contributing to poor housekeeping conditions on construction sites.

With original dataset

With combination of original and synthetic dataset

Figure 7. Inference results on real-world water ponding-related poor housekeeping images

5 Conclusion and future work

In conclusion, this study demonstrates that synthetic data can effectively enhance the performance of models used to detect water ponding that causes the poor housekeeping conditions on construction sites. By leveraging zero-shot models and diffusion-based inpainting, the synthetic water ponding–related poor housekeeping images are generated in a controlled and realistic manner that accurately reflects on-site conditions. Moreover, this method does not require additional financial expenditure, thereby addressing a key limitation of existing studies that rely on paid generative AI services, while simultaneously enhancing the accuracy and generalizability of the object detection model compared with training on the original dataset alone.

Although effective, several limitations remain. The first limitation concerns the manual selection of text prompts used as inputs to the Grounding DINO model. Although this manual selection produces correct results for the construction site dataset examined in this study, the chosen prompts may not generalize reliably to other environments or broader housekeeping scenarios. Moreover, the Grounded-SAM segmentation stage was adopted as a practical enabling module and was not separately benchmarked using segmentation quality metrics or alternative ground-detection methods. Future work should investigate how segmentation accuracy and failure cases affect inpainting realism and downstream detection performance. The second drawback is the significant computational cost imposed by diffusion-based inpainting, as these models require considerable time and resources to produce high-quality synthetic images. Therefore, future work will first focus on developing a system capable of automatically and accurately generating optimal prompts for Grounding DINO, which is expected to substantially improve the robustness of the proposed approach. Finally, future work will also aim to optimize the diffusion-based inpainting process to reduce computational demand while maintaining high synthesis quality. This advancement is expected to transform the proposed method into a more practical and deployable solution for real construction site applications.

Acknowledgement

This research is supported by NUS GAP funding (GAP502024-04-12).

References

[1] "WHO/ILO: Almost 2 million people die from work-related causes each year." Accessed: Sept. 24, 2025. [Online]. Available: https://www.who.int/news/item/17-09-2021-who-ilo-almost-2-million-people-die-from-work-related-causes-each-year

[2] Y. Halabi *et al.*, "Causal factors and risk assessment of fall accidents in the U.S. construction industry: A comprehensive data analysis (2000–2020)," *Safety Science*, vol. 146, p. 105537, Feb. 2022, doi: 10.1016/j.ssci.2021.105537.

[3] K. Sun, Z. Shao, Y. M. Goh, J. Tian, and V. J. L. Gan, "Change detection network for construction housekeeping using feature fusion and large vision models," *Automation in Construction*, vol. 172, p. 106038, Apr. 2025, doi: 10.1016/j.autcon.2025.106038.

[4] Y. M. Goh, J. Tian, and E. Y. T. Chian, "Management of safe distancing on construction sites during COVID-19: A smart real-time monitoring system," *Comput Ind Eng*, vol. 163, p. 107847, Jan. 2022, doi: 10.1016/j.cie.2021.107847.

[5] Y. G. Lim, J. Wu, Y. M. Goh, J. Tian, and V. Gan, "Automated classification of 'cluttered' construction housekeeping images through supervised and self-supervised feature representation learning," *Automation in Construction*, vol. 156, p. 105095, Dec. 2023, doi: 10.1016/j.autcon.2023.105095.

[6] S. Song, J. Hong, J. Jeoung, J. Ahn, and T. Hong, "Data-centric enhancement of site-specific automated construction equipment detection in wide-angle site images," *Automation in Construction*, vol. 179, p. 106483, Nov. 2025, doi: 10.1016/j.autcon.2025.106483.

[7] D. Gil and G. Lee, "Zero-shot monitoring of construction workers' personal protective

equipment based on image captioning," *Automation in Construction*, vol. 164, p. 105470, Aug. 2024, doi: 10.1016/j.autcon.2024.105470.

[8] V. S. K. Delhi, R. Sankarlal, and A. Thomas, "Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques," *Front. Built Environ.*, vol. 6, Sept. 2020, doi: 10.3389/fbuil.2020.00136.

[9] Y. Lee, G. Kang, J. Kim, S. Yoon, and J. Jeon, "Generative AI-driven data augmentation for enhanced construction hazard detection," *Automation in Construction*, vol. 177, p. 106317, Sept. 2025, doi: 10.1016/j.autcon.2025.106317.

[10] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," Oct. 23, 2024, *arXiv*: arXiv:2410.17725. doi: 10.48550/arXiv.2410.17725.

[11] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "Generative artificial intelligence: a systematic review and applications," *Multimed Tools Appl*, vol. 84, no. 21, pp. 23661–23700, Aug. 2024, doi: 10.1007/s11042-024-20016-1.

[12] Y. Lee, G. Kang, J. Kim, S. Yoon, and J. Jeon, "Generative AI-driven data augmentation for enhanced construction hazard detection," *Automation in Construction*, vol. 177, p. 106317, Sept. 2025, doi: 10.1016/j.autcon.2025.106317.

[13] H. Kim and J.-S. Yi, "Image generation of hazardous situations in construction sites using text-to-image generative model for training deep neural networks," *Automation in Construction*, vol. 166, p. 105615, Oct. 2024, doi: 10.1016/j.autcon.2024.105615.

[14] T. Ren *et al.*, "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks," Jan. 25, 2024, *arXiv*: arXiv:2401.14159. doi: 10.48550/arXiv.2401.14159.

[15] S. Liu *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," July 19, 2024, *arXiv*: arXiv:2303.05499. doi: 10.48550/arXiv.2303.05499.

[16] A. Kirillov *et al.*, "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 3992–4003. doi: 10.1109/ICCV51070.2023.00371.

[17] H. Zhang *et al.*, "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection," July 11, 2022, *arXiv*: arXiv:2203.03605. doi: 10.48550/arXiv.2203.03605.

[18] L. H. Li *et al.*, "Grounded Language-Image Pre-training," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, June 2022, pp. 10955–10965. doi: 10.1109/CVPR52688.2022.01069.

[19] "Laplacian Filter - an overview | ScienceDirect Topics." Accessed: Oct. 16, 2025. [Online]. Available: https://www-sciencedirect-com.libproxy1.nus.edu.sg/topics/engineering/laplacian-filter

[20] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and Texture Analysis for Image Segmentation," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 7–27, June 2001, doi: 10.1023/A:1011174803800.