Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

GENVISION: ENHANCING CONSTRUCTION SAFETY MONITORING WITH SYNTHETIC IMAGE GENERATION

Jiuyi Xu[1], Meida Chen[2], and Yangming Shi[3]

[1]Robotics Program, Colorado School of Mines, Golden, CO, USA
[2]USC Institute for Creative Technologies, Los Angeles, CA, USA
[3]Dept. of Civil and Environmental Eng. & Robotics Program, Colorado School of Mines, Golden, CO, USA

ABSTRACT

The development of object detection models for construction safety is often limited by the availability of high-quality, annotated datasets. This study explores the use of synthetic images generated by DALL·E 3 to supplement or partially replace real data in training YOLOv8 for detecting construction-related objects. We compare three dataset configurations: real-only, synthetic-only, and a mixed set of real and synthetic images. Experimental results show that the mixed dataset consistently outperforms the other two across all evaluation metrics, including precision, recall, IoU, and mAP@0.5. Notably, detection performance for occluded or ambiguous objects such as safety helmets and vests improves with synthetic data augmentation. While the synthetic-only model shows reasonable accuracy, domain differences limit its effectiveness when used alone. These findings suggest that high-quality synthetic data can reduce reliance on real-world data and enhance model generalization, offering a scalable approach for improving construction site safety monitoring systems.

1 INTRODUCTION

Recently, many powerful deep learning models in object detection tasks have been proposed (Zou et al. 2023). These models bring growing applications in safety-critical domains such as construction. Automated construction safety monitoring systems rely on accurate and efficient detection models to identify hazards, assess personal protective equipment (PPE) compliance, and detect unsafe behaviors (Kanan et al. 2018; Awolusi et al. 2018). However, the development of such systems is often hindered by the scarcity of large-scale, high-quality annotated datasets. Challenges such as privacy concerns, site variability, and the labor-intensive nature of manual labeling make data collection in construction environments particularly difficult.

To mitigate this limitation, researchers have explored data augmentation techniques to expand training datasets artificially. While augmentation methods such as flipping, rotation, and color adjustments improve model generalization, they are inherently constrained by the representational scope of the original dataset (Shorten and Khoshgoftaar 2019). Recent advancements in generative AI, including models like DALL·E 3 (OpenAI 2021) and Stable Diffusion (Rombach et al. 2022), offer a promising alternative by generating synthetic images that closely resemble real-world scenarios (Eigenschink et al. 2023). These synthetic images can supplement existing real datasets or, in some cases, be used to construct standalone training data.

In this paper, we investigate the impact of synthetic image generation for training object detection models in the context of construction safety. Specifically, we evaluate the performance of YOLOv8 under three different fine-tuning configurations: (1) fine-tuning with 3,000 real-world images, (2) finetuning with 3,000 synthetic images generated via DALL·E 3 with 30 carefully designed prompts, and (3) fine-tuning

with a balanced combination of 1,500 real and 1,500 synthetic images. Through these experiments, we aim to assess whether synthetic images can replace or supplement real data to improve detection performance.

The results of our experiments demonstrate that synthetic data can effectively enhance the performance of object detection models when used in combination with real data. The mixed dataset (real + synthetic) configuration consistently outperformed both the Real-Only and Synthetic-Only setups across all evaluation metrics, including precision, recall, IoU, mIoU, and mAP@0.5. Notably, improvements were observed in the detection of classes such as safety helmets and safety vests—objects that often appear blurred or occluded in real-world images—suggesting that synthetic data introduces beneficial variation in appearance, pose, and environmental context. Although models trained exclusively on synthetic data performed reasonably well, they exhibited a performance gap likely due to domain discrepancies between synthetic and real images. These findings suggest that while synthetic data alone may not yet fully substitute for real data, it plays a critical role in augmenting and diversifying training datasets, thereby improving model generalization and robustness in complex construction environments.

This paper contributes to the research related to exploring generative AI for data-centric deep learning and shows practical implications for the development of scalable, cost-effective construction safety monitoring systems. By demonstrating that high-quality synthetic data can partially replace real data without sacrificing model performance, this paper supports a promising pathway toward reducing data dependency in safety-critical applications.

2 RELATED WORK

2.1 Construction Safety and Object Detection

Recent advancements in computer vision, particularly object detection, have significantly contributed to improving construction site safety. These technologies enable real-time monitoring of dynamic environments, helping identify potential hazards such as unauthorized personnel, falling objects, or unsafe behaviors. Deep learning-based object detection models, including YOLO (You Only Look Once) (Redmon et al. 2016; Redmon and Farhadi 2017; Redmon and Farhadi 2018; Bochkovskiy et al. 2020; Jocher 2020; Li et al. 2022; Wang et al. 2023; Jocher et al. 2023), Faster R-CNN (Ren et al. 2015), and SSD (Liu et al. 2016), have been widely used in construction sites to detect critical safety elements such as personal protective equipment (PPE) - helmets, safety vests, and masks - as well as to monitor worker behaviors indicative of potential hazards, including slipping, fall, tripping risks (Liu et al. 2025; Fang et al. 2018).

For example, Kim et al. (2023) proposed a deep learning-based framework for detecting safety risk factors using YOLOv5 and YOLOv8, demonstrating the system's potential for improving the work process, quality control, and progress management in addition to safety management. Similarly, Wang et al. (2021) and Li et al. (2020) employed deep convolutional neural networks to detect workers' personal protective equipment (PPE), ensuring compliance and minimizing risk. Moreover, Jeelani et al. (2021) demonstrated the use of real-time object tracking to prevent potential equipment–worker collisions on construction sites. These studies all emphasize the practical applicability of these systems in dynamic construction environments.

2.2 Data Scarcity and Synthetic Data Generation

Collecting large-scale, annotated datasets from real-world construction sites poses significant challenges. Factors such as safety concerns, restricted access, variable site conditions, and the presence of heavy machinery often limit the feasibility of continuous data collection (Xu et al. 2024; Cuypers et al. 2021). Due to privacy issues and the high cost of manual labeling, obtaining diverse, high-quality annotated datasets remains a major bottleneck. This scarcity of real data has motivated researchers to explore alternative approaches, such as synthetic data generation, to supplement training datasets and improve model performance in construction-specific vision tasks (Neuhausen et al. 2020; Hong et al. 2021; Barrera-Animas and Davila Delgado 2023).

Generative models are increasingly popular in computer vision due to their ability to synthesize high-fidelity and contextually rich images. Two large families of these models—Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and diffusion models (Ho et al. 2020)—have demonstrated strong potential in augmenting datasets, particularly for tasks where real-world data collection is costly or limited. For example, Kim and Yi (2024) have explored the use of generative models to generate training image data for deep neural networks (DNNs) to enhance safety monitoring on construction sites. However, the application of generative models for synthetic data generation remains relatively underexplored and more experiments are needed to fully understand their potential and limitations on construction safety.

3 YOLOV8

YOLOv8 (Jocher et al. 2023), released by Ultralytics in 2023, represents a fundamental redesign of the YOLO architecture. It introduces several novel components and methodological advancements, particularly suited for deployment in dynamic and cluttered environments such as construction sites. This section provides a detailed overview of the YOLOv8 framework, particularly in the context of its suitability for construction safety applications.

3.1 Architecture

YOLOv8 adopts a fully convolutional, anchor-free architecture that departs from traditional anchor-based approaches employed in earlier versions (e.g., YOLOv3–YOLOv5 (Redmon and Farhadi 2018; Bochkovskiy et al. 2020; Jocher 2020)). Anchor-free detection simplifies the model by removing the need for predefined anchor box sizes and aspect ratios. Instead, object center points are directly regressed from feature maps, allowing for improved generalization to variable object scales and shapes. This modification is particularly advantageous in construction settings, where objects such as helmets, tools, and machinery may exhibit significant intra-class variability. The overall structure of YOLOv8 is shown in Figure 1.

The backbone of YOLOv8 is constructed using C2f (Concatenate-to-Fusion) modules, which improve feature extraction efficiency by enabling feature reuse and enhanced gradient flow. This module is a modification of the Cross Stage Partial (CSP) structure used in YOLOv5, wherein feature maps are partitioned, processed separately, and later fused. YOLOv8's C2f modules introduce additional skip connections that preserve spatial detail while maintaining low computational overhead.

The neck of YOLOv8 combines elements from Path Aggregation Networks (PANet) (Liu et al. 2018) and Feature Pyramid Networks (FPN) (Lin et al. 2017) to enable multi-scale feature fusion. This is critical for detecting small and large objects in the same scene, which is often required in construction environments, where object scales can vary dramatically across views.

The detection head in YOLOv8 employs a decoupled structure that separates classification and bounding box regression branches. This architectural choice aligns with recent findings in object detection literature, where decoupling these tasks leads to improved learning dynamics and reduced task interference, ultimately enhancing both localization accuracy and classification confidence.

3.2 Loss Functions and Optimization

YOLOv8 utilizes a composite loss function (1) consisting of three principal components:

- Box regression loss (2, where b is the predicted bounding box and b^{gt} is the ground truth box): Based on the Complete Intersection over Union (CIoU) metric, which integrates overlap area, center distance, and aspect ratio differences between predicted and ground truth boxes.
- Objectness loss (3, where $y \in \{0,1\}$ indicates object presence and p is the predicted objectness score): A binary cross-entropy loss applied to the objectness score to determine the presence of an object within a predicted box.

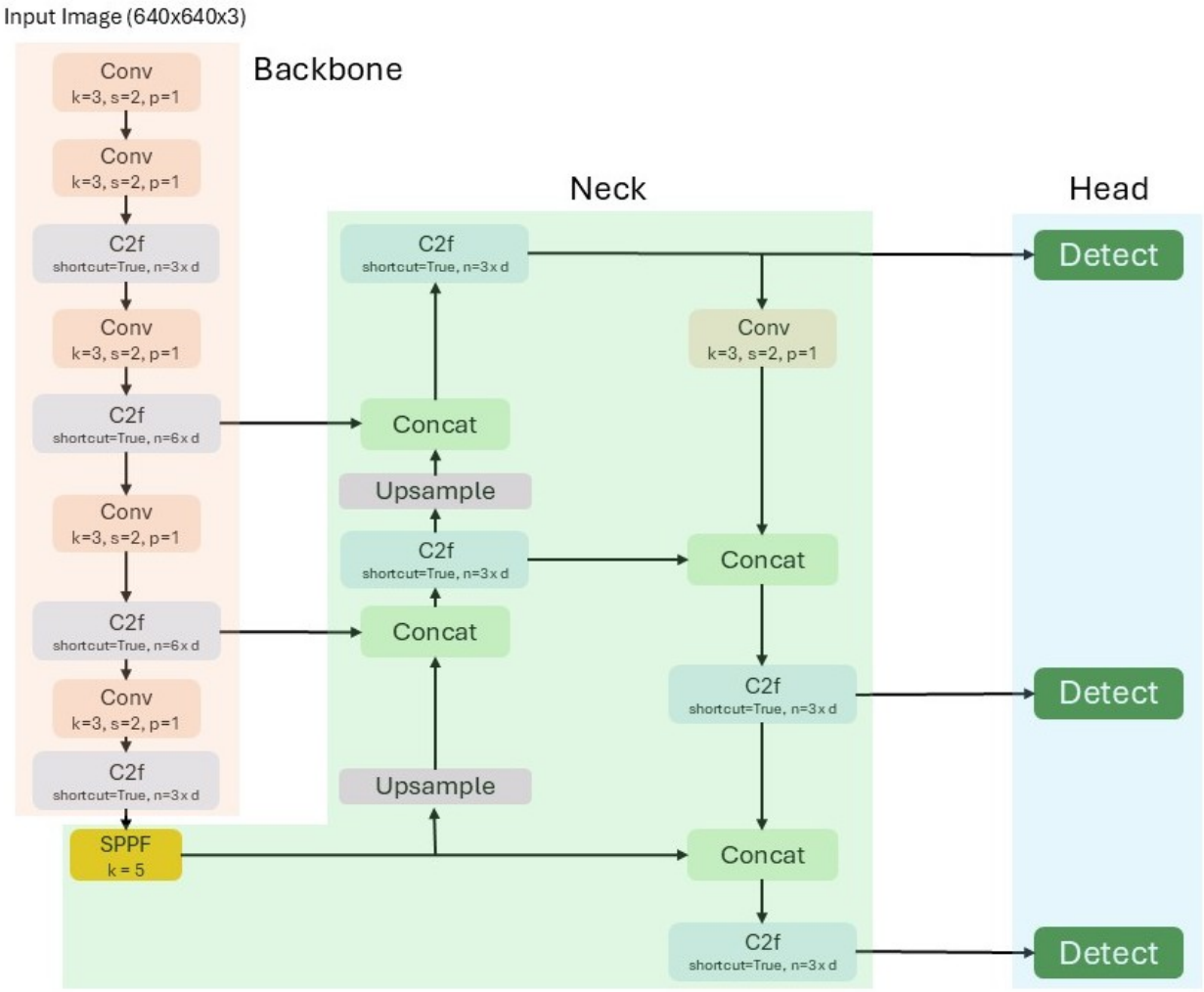

Figure 1: The general architecture of YOLOv8.

- Classification loss (4, where C is the number of classes, y_c is the ground truth for class c, and p_c is the predicted probability for class c): A sigmoid-based binary cross-entropy loss allowing for multi-label classification within a single bounding box.

$$\mathscr{L}_{\text{YOLOv8}} = \lambda_{\text{box}} \cdot \mathscr{L}_{\text{CIoU}} + \lambda_{\text{obj}} \cdot \mathscr{L}_{\text{obj}} + \lambda_{\text{cls}} \cdot \mathscr{L}_{\text{cls}} \tag{1}$$

$$\mathscr{L}_{\text{CIoU}} = 1 - \text{CIoU}(b, b^{gt}) \tag{2}$$

$$\mathscr{L}_{\text{obj}} = -\left[y \log(p) + (1-y)\log(1-p)\right] \tag{3}$$

$$\mathscr{L}_{\text{cls}} = -\sum_{c=1}^{C} \left[y_c \log(p_c) + (1-y_c)\log(1-p_c)\right] \tag{4}$$

This loss formulation allows YOLOv8 to handle ambiguous and partially occluded objects—conditions frequently encountered in construction environments. Furthermore, YOLOv8 benefits from modern training

strategies, including label smoothing, learning rate warm-up, gradient accumulation, and mixed-precision training via Automatic Mixed Precision (AMP), all of which contribute to faster convergence and improved generalization.

3.3 Model Scaling and Inference

YOLOv8 is released in five scalable model variants: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large), and YOLOv8x (extra-large). These variants differ in the number of parameters, floating-point operations per second (FLOPs), and backbone depth, thereby providing a flexible trade-off between computational complexity and detection performance. Table 1 summarizes the key characteristics of each model. In our experiments, due to limited computational resources, we restrict our fine-tuning to the available pretrained YOLOv8s model.

Table 1: Comparison between different sizes of YOLOv8 model.

Model	Parameters (M)	FLOPs (B)	COCO mAP@0.5	Inference Speed (A100)
YOLOv8n	~3.2M	~8.7	~37.3%	~2.2 ms/image
YOLOv8s	~11.2M	~28.6	~44.9%	~3.0 ms/image
YOLOv8m	~25.9M	~78.9	~50.2%	~4.6 ms/image
YOLOv8l	~43.7M	~165.2	~52.9%	~6.2 ms/image
YOLOv8x	~68.2M	~257.8	~53.9%	~8.2 ms/image

4 RESEARCH METHODOLOGY - DATA GENERATION

4.1 Ground Dataset

Due to the high cost of collecting data on personal protective equipment on the construction site, we selected 3,500 labeled images from publicly available datasets on Kaggle and Roboflow as our real dataset. The ground dataset includes 3,000 training images and 500 testing images. Each image contains annotations for multiple object categories relevant to safety compliance, including *Person*, *Safety Helmet*, *Safety Vest*, and *Safety Cone*. We focus on these four object classes due to their direct relevance to construction site safety monitoring. These classes correspond to primary safety compliance indicators (e.g., PPE adherence and site demarcation). While other site hazards involve conditions (e.g., exposed wires, open trenches) or actions (e.g., running, falling), these are more complex to annotate reliably across datasets and may require video-based behavior recognition. Our scope in this work is to validate object-based compliance indicators, with future research targeting action- and condition-based hazard detection. The dataset reflects a range of real-world construction site conditions, with variations in lighting, occlusion, and object scale. All annotations follow the YOLO format (<class_id>, <x_center>, <y_center>, <width>, <height>), specifying bounding boxes and class labels. Before training, we also applied standard preprocessing steps, including image resizing to 640x640 and normalization. We show some examples in Figure 2.

4.2 Data Generation - DALL·E 3

To generate the synthetic dataset to validate if the generative model can help with data limitation issues on construction safety, we used DALL·E 3 (Betker et al. 2023) to generate 640x640 images focused on the same four object categories with annotations following the YOLO format for our experiments specifically.

DALL·E 3 is a cutting-edge text-to-image generation model developed by OpenAI. As the third iteration of the DALL·E series (OpenAI 2021; Ramesh et al. 2021; Ramesh et al. 2022), DALL·E 3 demonstrates significantly improved semantic understanding and visual fidelity compared to DALL·E 1 and DALL·E 2. It is built upon GPT-4 (Achiam et al. 2023) architecture and is capable of translating complex textual

(a) Ground data sample 1 (b) Ground data sample 2 (c) Ground data sample 3

Figure 2: Some ground data examples.

descriptions into photorealistic or contextually rich images, making it highly suitable for creating synthetic datasets for computer vision applications such as object detection and PPE compliance monitoring.

Since DALL·E 3 cannot directly output annotations in YOLO format, we designed some structured and consistent prompts, then used the automatic Roboflow annotation tool to label the generated images and save them in YOLO format. We designed 30 distinct prompts, each used to generate 100 images, resulting in a total of 3,000 images that comprise our synthetic dataset. All the prompts used are listed in appx A. We also show three synthetic generated data samples in Figure 3.

(a) Synthetic data sample 1 (b) Synthetic data sample 2 (c) Synthetic data sample 3

Figure 3: Some synthetic data examples.

A known limitation of using DALL·E 3 for synthetic image generation is the tendency to generate subject-centric images with shallow depth and minimal background clutter. This diverges from the far-field characteristics of real construction surveillance footage. However, we deliberately structured our prompts (see appx A) to include environmental context (e.g., cranes, equipment, distant workers) and diverse viewing angles (e.g., aerial, wide-angle). Despite these efforts, current generative models often struggle to fully emulate complex far-field scenes. We acknowledge this limitation and plan to incorporate far-field domain adaptation techniques or 3D simulation-based rendering (e.g., BlenderProc (Denninger et al. 2023) or UnrealCV (Qiu et al. 2017)) in future work.

5 EXPERIMENT

5.1 Experiment Settings

We conducted three fine-tuning experiments on the pretrained models using 3,000 ground images, 1,500 ground images and 1,500 synthetic images, 3,000 synthetic images, respectively. For all experiments,

the fine-tuned model was evaluated on the same test set of 500 real construction site images to maintain consistency. Evaluation metrics included Precision, Recall, Intersection over Union (IoU), mean IoU (mIoU), mean Average Precision at IoU=0.5 (mAP@0.5).

All fine-tuning experiments were conducted using the Ultralytics YOLOv8 framework, implemented in Python 3.10 and PyTorch 2.1. We utilized the YOLOv8s model as a balance between detection performance and computational efficiency. Although YOLOv11 has recently been released, YOLOv8 was selected for our study due to its demonstrated stability, comprehensive documentation, and validated performance across a wide range of real-world object detection tasks. Currently, YOLOv11 lacks publicly available pretrained weights and has not yet been used in many evaluations in the context of construction safety monitoring. Future work may include a comparative analysis incorporating YOLOv11 once its performance and compatibility with construction-specific datasets are more thoroughly established. To fine-tune the pretrained YOLOv8s model for our experiments, we used the official training configuration provided by Ultralytics, with data augmentation explicitly disabled to maintain consistency and control over input data variability. The fine-tuning process was performed on a high-performance computing setup equipped with two 32 GB NVIDIA V100 GPUs and running Rocky Linux 8. The fine-tuning was conducted using a batch size of 16, an image resolution of 640×640 pixels, and a total of 100 epochs. Automatic Mixed Precision (AMP) was enabled to optimize GPU memory usage and training speed. The model was initialized with pretrained weights. An initial learning rate of 0.01 with cosine learning rate scheduling was used, alongside a weight decay of 0.0005 and momentum of 0.937 to enhance optimization stability. We also adopted early stopping with a patience of 100 epochs to stabilize model performance.

To decrease the effect of the randomness in training, each experimental condition was run independently five times using different random seeds. The final reported performance metrics—Precision, Recall, IoU, mIoU and mAP@0.5—represent the mean values across these five runs. The evaluation metrics are introduced in detail in appx B

6 RESULTS & DISCUSSION

Our experimental results are summarized in Table 2 and 3. The results highlight the benefits of using synthetic data to replace part of the real images in the training dataset for object detection tasks in construction environments. Among the three experiment settings, the model trained with a mixed dataset of real and synthetic images consistently outperformed the Real-Only and Synthetic-Only configurations across all evaluation metrics. This performance suggests that synthetic images, when generated with sufficient visual realism and contextual diversity, serve as effective data augmentation tools. They help reduce overfitting and introduce variation in object appearances, poses, and lighting conditions that may not be fully captured in limited real-world datasets. The Safety Helmet and Safety Vest objects, which are often blurred or partially occluded in real images, benefited most from synthetic augmentation.

Table 2: The general performance of each fine-tuned model.

	Precision	Recall	IoU	mIoU	mAP@0.5
Real-Only	0.84	0.78	0.66	0.61	0.81
Real + Synthetic	0.85	0.81	0.66	0.62	0.85
Synthetic-Only	0.75	0.70	0.59	0.54	0.76

The Synthetic-Only model, while demonstrating reasonable performance, lagged behind the models fine-tuned by the real images, likely due to domain gaps between synthetic and real images' distributions. Although DALL·E-3 produced highly detailed and contextually rich scenes, subtle discrepancies in texture, background clutter, and object realism can lead to reduced generalizability when evaluated on real-world images. This observation is consistent with prior findings in domain adaptation and synthetic-to-real transfer learning research.

Table 3: Per-Class mAP@0.5 of each fine-tuned model.

Class	Real-Only	Real + Synthetic	Synthetic-Only
Person	0.88	0.91	0.79
Safety Helmet	0.76	0.82	0.71
Safety Vest	0.79	0.85	0.70
Safety Cone	0.81	0.84	0.74

Furthermore, the per-class performance also proves the idea that synthetic data is most valuable when it augments real data rather than replaces it. The superior performance of the mixed training method supports that strategically balancing real and synthetic data can yield robust models with enhanced generalization across diverse object types and scenes.

7 CONLCUSION

This study demonstrates the potential of using high-quality synthetic images generated by DALL·E 3 to improve object detection models for construction safety monitoring. Among the three fine-tuning configurations tested, the mixed dataset—combining real and synthetic images—achieved the best performance across all evaluation metrics, particularly improving detection of occluded or ambiguous objects like safety helmets and vests. While synthetic-only training showed reasonable results, domain discrepancies limited its effectiveness when used in isolation. Additionally, manually designed prompts may not capture the full range of real-world variability. These limitations suggest that synthetic data is best used to supplement—not fully replace—real images. Future work will focus on improving the diversity and realism of synthetic images through automated prompt generation and domain adaptation, as well as scaling evaluations across more object classes and deployment scenarios. Overall, this approach offers a promising, cost-efficient path toward building more robust and generalizable safety monitoring systems in construction and other high-risk environments.

ACKNOWLEDGMENTS

The authors acknowledge the use of generative models to create synthetic images for this study. These artificially generated images were used to augment the dataset and support the evaluation of the proposed method. The generative process was carefully controlled to maintain relevance and quality, and the synthetic data was used solely for research purposes in alignment with ethical guidelines.

A DESIGNED PROMPTS

The designed prompts for generating synthetic images are shown in table 4.

B EVALUATION METRICS

B.1 Precision

Precision evaluates the accuracy of positive predictions made by the object detection model. It is defined as the ratio of correctly predicted positive samples (true positives, TP) to all samples predicted as positive, including both true positives and false positives (FP). A higher precision indicates fewer false positives.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Table 4: Different prompts used for generating synthetic images.

1	"An outdoor construction site on a sunny day, with workers wearing and not wearing safety helmets and vests, safety cones along the path, realistic lighting and tools in the background, 640x640."
2	"Construction zone at night with artificial lighting, workers partially compliant with PPE, safety cones reflecting light, realistic textures, 640x640."
3	"Indoor construction area in a warehouse, 2–3 people working, some missing safety gear, cones near structural materials, 640x640."
4	"Roadwork construction site under cloudy weather, workers standing near safety cones, wearing mixed PPE, photorealistic setting, 640x640."
5	"Construction workers in a high-rise building site during sunset, some wearing safety helmets and vests, others not, safety cones marking areas, wide-angle view, 640x640."
6	"One construction worker close-up wearing a vest but no helmet, with safety cone nearby, blurred background, 640x640."
7	"Three workers collaborating at a construction zone, one not wearing a helmet, cones marking boundaries, daylight environment, 640x640."
8	"Group of 5 workers actively working, some fully geared, some non-compliant, safety cones and construction tools visible, realistic textures, 640x640."
9	"Two people walking on a construction site, wearing safety vests, one missing helmet, cones near sidewalk, industrial background, 640x640."
10	"Ten workers scattered across a construction area, various PPE conditions, cranes in the background, multiple cones on the ground, 640x640."
11	"Aerial view of a construction site with workers wearing different combinations of PPE, safety cones placed in zones, natural lighting, 640x640."
12	"Side-view of a construction worker placing a cone, wearing vest but no helmet, industrial background with equipment, 640x640."
13	"Front view of two construction workers talking, one without safety vest, cones behind them, clear site layout, 640x640."
14	"Wide-angle shot of an active construction site, multiple workers, mixed PPE use, cones forming a path, 640x640."
15	"Close-up photo of a safety helmet and vest placed on the ground, a worker standing nearby without PPE, cone in distance, 640x640."
16	"Street construction zone with heavy traffic nearby, workers in high-visibility vests, some missing helmets, cones separating lanes, 640x640."
17	"Construction site inside a tunnel, low lighting, workers wearing vests, some with no helmets, cones placed for safety, 640x640."
18	"High-rise scaffolding with one worker visible, wearing vest but no helmet, safety cone placed at entry, cloudy sky background, 640x640."
19	"Suburban road under maintenance, two workers near cones, only one wearing proper PPE, small construction machinery in background, 640x640."
20	"Bridge construction site with several workers, partial PPE usage, orange safety cones, water in background, 640x640."
21	"Construction workers handling a jackhammer, one wearing helmet and vest, the other not, cones around the tool area, 640x640."
22	"Worker operating a small bulldozer near cones, not wearing a helmet, another person standing nearby with PPE, open space, 640x640."
23	"Workers lifting pipes in a trench, with cones marking the danger zone, one missing a vest, realistic shadow and lighting, 640x640."
24	"Two workers painting lane lines, cones spread out, both wearing vests but only one has a helmet, road construction environment, 640x640."
25	"Scaffolding construction with worker climbing, wearing vest and helmet, another below missing helmet, cones below the structure, 640x640."
26	"Diverse group of male and female construction workers, mixed PPE compliance, safety cones in foreground, multicultural team, 640x640."
27	"Female construction worker in vest and helmet, standing next to another worker without PPE, cones on the ground, modern construction site, 640x640."
28	"Construction workers of different ethnic backgrounds, in different PPE combinations, safety cones organized around them, inclusive team, 640x640."
29	"A worker walking past a safety cone without any PPE, others in background fully geared, construction hazard zone, 640x640."
30	"Worker climbing on unsafe structure with no helmet, safety cone nearby, others wearing proper PPE observing, realism emphasized, 640x640."

B.2 Recall

Recall measures the ability of the model to identify all relevant objects in the dataset. It is calculated as the ratio of correctly predicted positive samples (TP) to the total number of actual positive samples, including both true positives and false negatives (FN). A higher recall indicates fewer missed detections.

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

B.3 Intersection over Union (IoU & mIoU)

IoU quantifies the overlap between the predicted bounding box and the ground-truth bounding box. It is defined as the area of intersection divided by the area of union between the two boxes. For a dataset with multiple detection instances, mean IoU (mIoU) is used to report the average IoU across all predictions.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{7}$$

$$mIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \tag{8}$$

B.4 mean Average Precision (mAP)

mAP is one of the most commonly used evaluation metrics for object detection. It summarizes the precision-recall curve by calculating the area under the curve for each class and then taking the mean across all classes. In this study, we report mAP at a fixed IoU threshold of 0.5, meaning that a prediction is considered correct if its IoU with the ground truth is at least 0.5.

$$mAP@0.5 = \frac{1}{C}\sum_{c=1}^{C} AP_c^{@0.5} \tag{9}$$

REFERENCES

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, *et al*. 2023. "Gpt-4 Technical Report". Technical report, OpenAI.

Awolusi, I., E. Marks, and M. Hallowell. 2018. "Wearable Technology for Personalized Construction Safety Monitoring and Trending: Review of Applicable Devices". *Automation in Construction* 85:96–106 https://doi.org/https://doi.org/10.1016/j.autcon.2017.10.010.

Barrera-Animas, A. Y., and J. M. Davila Delgado. 2023. "Generating Real-world-like Labelled Synthetic Datasets for Construction Site Applications". *Automation in Construction* 151:104850 https://doi.org/https://doi.org/10.1016/j.autcon.2023.104850.

Betker, J., G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, *et al*. 2023. "Improving Image Generation with Better Captions". https://cdn.openai.com/papers/dall-e-3.pdf, accessed 8th April 2025.

Bochkovskiy, A., C.-Y. Wang, and H.-Y. M. Liao. 2020. "Yolov4: Optimal Speed and Accuracy of Object Detection". *arXiv preprint arXiv:2004.10934*.

Cuypers, S., M. Bassier, and M. Vergauwen. 2021. "Deep Learning on Construction Sites: A Case Study of Sparse Data Learning Techniques for Rebar Segmentation". *Sensors* 21(16) https://doi.org/10.3390/s21165428.

Denninger, M., D. Winkelbauer, M. Sundermeyer, W. Boerdijk, M. Knauer, K. H. Strobl, *et al*. 2023. "BlenderProc2: A Procedural Pipeline for Photorealistic Rendering". *Journal of Open Source Software* 8(82):4901 https://doi.org/10.21105/joss.04901.

Eigenschink, P., T. Reutterer, S. Vamosi, R. Vamosi, C. Sun, and K. Kalcher. 2023. "Deep Generative Models for Synthetic Data: A Survey". *IEEE Access* 11:47304–47320.

Fang, Q., H. Li, X. Luo, L. Ding, H. Luo, T. M. Rose *et al*. 2018. "Detecting Non-hardhat-use by A Deep Learning Method from Far-field Surveillance Videos". *Automation in Construction* 85:1–9 https://doi.org/https://doi.org/10.1016/j.autcon.2017.09.018.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, *et al*. 2014. "Generative adversarial nets". In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 2672–2680.

Ho, J., A. Jain, and P. Abbeel. 2020. "Denoising Diffusion Probabilistic Models". *Advances in Neural Information Processing Systems* 33:6840–6851.

Hong, Y., S. Park, H. Kim, and H. Kim. 2021. "Synthetic Data Generation using Building Information Models". *Automation in Construction* 130:103871 https://doi.org/https://doi.org/10.1016/j.autcon.2021.103871.

Jeelani, I., K. Asadi, H. Ramshankar, K. Han, and A. Albert. 2021. "Real-time Vision-based Worker Localization & Hazard Detection for Construction". *Automation in Construction* 121:103448 https://doi.org/https://doi.org/10.1016/j.autcon.2020.103448.

Jocher, G. 2020. "Ultralytics YOLOv5". Technical report, Ultralytics.

Jocher, G., A. Chaurasia, and J. Qiu. 2023. "Ultralytics YOLOv8". Technical report, Ultralytics.

Kanan, R., O. Elhassan, and R. Bensalem. 2018. "An IoT-based Autonomous System for Workers' Safety in Construction Sites with Real-time Alarming, Monitoring, and Positioning Strategies". *Automation in Construction* 88:73–86 https://doi.org/https://doi.org/10.1016/j.autcon.2017.12.033.

Kim, H., and J.-S. Yi. 2024. "Image Generation of Hazardous Situations in Construction Sites using Text-to-image Generative Model for Training Deep Neural Networks". *Automation in Construction* 166:105615 https://doi.org/https://doi.org/10.1016/j.autcon.2024.105615.

Kim, K., K. Kim, and S. Jeong. 2023. "Application of YOLO v5 and v8 for Recognition of Safety Risk Factors at Construction Sites". *Sustainability* 15(20):15179.

Li, C., L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, *et al*. 2022. "YOLOv6: A Single-stage Object Detection Framework for Industrial Applications". *arXiv preprint arXiv:2209.02976*.

Li, Y., H. Wei, Z. Han, J. Huang, and W. Wang. 2020. "Deep Learning-based Safety Helmet Detection in Engineering Management based on Convolutional Neural Networks". *Advances in Civil Engineering* 2020(1):9703560.

Lin, T.-Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature Pyramid Networks for Object Detection". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Liu, S., L. Qi, H. Qin, J. Shi, and J. Jia. 2018. "Path Aggregation Network for Instance Segmentation". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu *et al*. 2016. "Ssd: Single Shot Multibox Detector". In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37. Springer.

Liu, Z., J. Xu, C. W. K. Suen, M. Chen, Z. Zou, and Y. Shi. 2025. "Egocentric Camera-based Method for Detecting Static Hazardous Objects on Construction Sites". *Automation in Construction* 172:106048 https://doi.org/https://doi.org/10.1016/j.autcon.2025.106048.

Neuhausen, M., P. Herbers, and M. König. 2020. "Using Synthetic Data to Improve and Evaluate the Tracking Performance of Construction Workers On Site". *Applied Sciences* 10(14):4948.

OpenAI 2021. "DALL·E: Creating Images from Text". Technical report, OpenAI.

Qiu, W., F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim *et al*. 2017. "UnrealCV: Virtual Worlds for Computer Vision". In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, 1221–1224. New York, NY, USA: Association for Computing Machinery https://doi.org/10.1145/3123266.3129396.

Ramesh, A., P. Dhariwal, A. Nichol, C. Chu, and M. Chen. 2022. "Hierarchical Text-conditional Image Generation with Clip Latents". *arXiv preprint arXiv:2204.06125*.

Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, *et al*. 2021. "Zero-shot Text-to-image Generation". In *International Conference on Machine Learning*, 8821–8831. Pmlr.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. "You Only Look Once: Unified, Real-time Object Detection". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Redmon, J., and A. Farhadi. 2017. "YOLO9000: Better, Faster, Stronger". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7263–7271.

Redmon, J., and A. Farhadi. 2018. "Yolov3: An Incremental Improvement". *arXiv preprint arXiv:1804.02767*.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. "Faster R-CNN: towards real-time object detection with region proposal networks". In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 91–99.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. "High-resolution Image Synthesis with Latent Diffusion Models". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shorten, C., and T. M. Khoshgoftaar. 2019. "A Survey on Image Data Augmentation for Deep Learning". *Journal of Big Data* 6(1):1–48.

Wang, C.-Y., A. Bochkovskiy, and H.-Y. M. Liao. 2023. "YOLOv7: Trainable Bag-of-freebies Sets New State-of-the-art for Real-time Object Detectors". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7464–7475.

Wang, Z., Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, and Y. Zhao. 2021. "Fast Personal Protective Equipment Detection for Real Construction Sites using Deep Learning Approaches". *Sensors* 21(10):3478.

Xu, J., M. Chen, A. Feng, Z. Yu, and Y. Shi. 2024. "Open-Vocabulary High-Resolution 3D (OVHR3D) Data Segmentation and Annotation Framework". *arXiv preprint arXiv:2412.06268*.

Zou, Z., K. Chen, Z. Shi, Y. Guo, and J. Ye. 2023. "Object Detection in 20 years: A Survey". *Proceedings of the IEEE* 111(3):257–276.

AUTHOR BIOGRAPHIES

JIUYI XU is a Ph.D. student in Robotics under the supervision of Dr. Yangming Shi at Colorado School of Mines in Golden, USA. His research interests include image generation and generative reinforcement learning. He holds an M.S. in Computer Science from the University of Southern California (USC) and a B.E. in Software Engineering from Dalian University of Technology (DLUT). He was a student research assistant at USC Institute for Creative Technologies, working on open-vocabulary object detection (OVOD) and open-vocabulary semantic segmentation (OVSS). Beyond research, Jiuyi is an active peer reviewer for the Journal of Computing in Civil Engineering and has contributed to multiple academic conferences. He is also a student member of IEEE, ACM, and ASCE. His email address is jiuyi_xu@mines.edu and his website is https://jiuyixu25.github.io.

MEIDA CHEN is a Research Scientist within the Geospatial Terrain Research Lab, focusing on 3D computer vision and synthetic training data generation. He received his Ph.D. degree from the Department of Civil and Environmental Engineering, University of Southern California in 2020, following his MS, Computer Science and MCM, Construction Management (Civil Engineering), also from USC. He joined USC Institute for Creative Technologies as a Research Assistant in 2017, and received two promotions: Research Associate (2020), then Senior Research Associate (2022). Alongside his role as a Research Scientist at ICT, Dr. Meida Chen is also a lecturer at the Department of Civil and Environmental Engineering, Viterbi School of Engineering, on the Introduction to Civil Engineering Graphics course. His e-mail address is mechen@ict.usc.edu and his website is https://www.linkedin.com/in/meida-chen-938a265b/.

YANGMING SHI is an Assistant Professor in the Department of Civil and Environmental Engineering and also a core faculty member in the robotics program at the Colorado School of Mines. He received his Ph.D. degree from the University of Florida under the supervision of Dr. Eric Du. He received his Master of Science degree from Texas A&M University and a bachelor's degree from Ningbo University, China. His research interests lie in Artificial Intelligence (AI), Virtual Reality (VR)/Augmented Reality (AR), Human-computer Interactions, Human-robot interactions, NeuroErgonomics, and Human Factors. His research vision is to develop data-driven solutions to improve human performance and augment human capabilities for future design, construction, and engineering process. His e-mail address is yangming.shi@mines.edu and his website is https://shiyangming.com/.