Proceedings of Digital Frontiers in Buildings and Infrastructure International Conference Series

DFBI 2025 aims to encourage the international exchange of innovative ideas between researchers from academia and industry. In addition to knowledge dissemination, the conference offers a valuable platform for professional networking, particularly benefiting university professors, graduate students, and postdoctoral researchers.

Research Article

Knowledge-enhanced Vision-Language Models for Few-Shot Object Detection in Construction Site

Zhaoxin Zhang [1], Yantao Yu [1, 2]

1. Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China
2. HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen 518045, China
Correspondence: ceyantao@ust.hk

Abstract (250 words) Style Name

Visual understanding of complex construction site objects is critical for project safety management and worker-robot collaboration within the construction domain. However, deploying deep learning algorithms on construction sites presents significant challenges due to high data annotation costs, substantial computational requirements, and the absence of large-scale training datasets. While large-scale pre-trained multimodal foundation models have shown success in natural language understanding and visual recognition, their application in construction safety management remains limited because of the need for domain-specific knowledge. To address these challenges, this paper proposes a knowledge-enhanced multimodal learning approach for few-shot object detection in construction scenarios. The proposed method comprises two components: (1) leveraging existing semantic knowledge in the construction domain to detect potential objects in construction scenes using a template matching approach; and (2) introducing a multimodal image semantic recognition method that integrates visual and textual knowledge specific to the construction field. We evaluate our approach on the AIMDataset. The results demonstrate that, without network training or large-scale construction samples, our method can achieve effective object detection under few-shot conditions using only existing models and a small amount of provided visual and textual knowledge. This approach highlights its potential for applications in construction scenarios.

Keywords: Construction sites, object recognition, knowledge-enhanced, vision-language models, few-shot learning

Cite this article:
Lastname, F. M., & Lastname, F. M. (Year). Title of the article. Title of the Journal, Volume 2025, Page range: 168-177

Highlights

- A training-free, knowledge-enhanced object detection algorithm for construction scenarios was developed.
- A two-stage detection framework was introduced to perform object localization and classification.
- The method achieved competitive detection performance on the public AIMDataset.

1 Introduction

Visual understanding of large-scale machinery and objects on construction sites supports automated monitoring systems, enhancing operational control. It enables personnel to assess resource efficiency (e.g., direct work efficiency, hourly productivity) and diagnose productivity losses (Rezazadeh Azar, 2017; Roberts et al., 2020; Yang et al., 2015). Continuous monitoring also detects hazardous objects and unsafe behaviors, including automated recognition of worker–machinery interactions (Seo et al., 2015). Camera-based image analysis facilitates vision-based remote oversight (Ham et al., 2016; Kim et al., 2019). This enhances safety, quality, speed, and profitability. Thus, timely and accurate monitoring of machinery and facilities is vital for effective project control.

Most state-of-the-art methods rely on traditional deep learning, which demands large, precisely labeled datasets (e.g., object types, locations) for high-performance vision algorithms (Paneru & Jeelani, 2021). This process is time-consuming, costly, and labor-intensive, complicating the creation of extensive, high-quality image datasets (Kim, 2020; Kim & Chi, 2017; Liu & Golparvar-Fard, 2015; Y. Wang et al., 2019). The challenge intensifies with machinery variability across construction phases, requiring constant dataset updates. In practice, vision-based monitoring faces three key data challenges: (1) privacy limits data availability and sharing; (2) data collection and labeling disrupt workflows (Teizer, 2015); and (3) dynamic site conditions vary by time and location (Paneru & Jeelani, 2021). Hence, computer vision algorithms requiring fewer samples and shorter training times are preferred for construction applications.

Few-shot object detection achieves promising results in recognizing novel objects from limited samples. These methods typically train a base model on a large dataset of base classes, fine-tune it on a small labeled support set of novel classes, and evaluate it on a test set containing those classes. Research has largely focused on extracting meta-knowledge from the base dataset, which critically influences detection performance (Zhou et al., 2020), partly due to limited understanding of optimal base dataset construction. Additionally, collecting and annotating base datasets poses further challenges in construction applications.

Foundation and multimodal models (e.g., BERT, GPT, CLIP) show strong potential for feature transfer and generalization across diverse tasks (Brown et al., 2020; Devlin et al., 2019; Radford et al., 2021). These models support vision-language pretraining tailored to construction, enhancing safety management. Their development underpins specialized models by providing rich linguistic and visual knowledge for construction-target detection. However, their application in construction safety remains limited due to the need for domain-specific expertise.

In this paper, our goal is to integrate visual and textual knowledge specific to the construction domain to develop a multimodal object detection method tailored for construction scenarios, referred to as knowledge-enhanced automatic detection. We propose a few-shot learning approach based on a knowledge-enhanced vision-language model for the localization and identification of large-scale machinery on construction sites. Specifically, this work makes two major contributions: (1) Localization of Construction Machinery: We propose a template-matching-based method for locating construction machinery. By leveraging the powerful feature extraction and similarity computation capabilities of existing self-supervised visual models, this method matches given visual knowledge of construction machinery with environmental visual targets to locate the machinery. The approach does not require retraining network parameters and relies solely on existing vision foundation models. (2) Classification of Construction Machinery: We construct a similarity distribution matrix infused with construction domain knowledge to enhance the association between the semantic representation of machinery target images and textual descriptions. This enables semantic classification of construction machinery. This study demonstrates the potential of using foundation models for few-shot learning in the recognition of objects on construction sites. The proposed method shows promise in enabling the identification of large-scale construction machinery using open-source text labels, handcrafted image prompts, and a small number of target images.

The paper is structured as follows: Section 2 reviews object detection and classification methods in construction scenarios, few-shot learning approaches, and vision-language foundation models. Section 3 presents the proposed algorithm. Section 4 validates the algorithm through extensive experiments on a publicly available construction machinery dataset. Section 5 discusses the limitations of this work and potential future improvements. Section 6 concludes with a summary of the contributions.

2 Literature Review

2.1 Object Detection and Classification in Construction

Deep learning algorithms now dominate various industry applications, particularly object detection, which includes two main types: two-stage detectors (e.g., Faster R-CNN) that generate region proposals before classification, and single-stage detectors (e.g., YOLO) that unify localization and classification (Redmon & Farhadi, 2018; Ren et al., 2017).

2.2 Few-shot Learning Methods

Recent applications in construction include facade defect detection via contrastive few-shot learning (Cui et al., 2022), structural damage classification using meta-learning (Xu et al., 2021), and few-shot object detection for emerging targets (Kim & Chi, 2021). Other examples include CLIP-based recognition of temporary objects (Liang et al., 2024) and detection of fall-related site objects for compliance checks (X. Wang & El-Gohary, 2024).

2.3 Few-Shot Learning Based on Foundation Models

Leveraging vast data and computational resources, foundation models have achieved major breakthroughs in computer vision and NLP. CLIP, trained on 400 million image-text pairs, demonstrates strong zero-shot image classification (Radford et al., 2021) . DINOv2 learns universal visual features via patch-level self-supervised learning (Oquab et al., 2023). Tip-Adapter (Zhang et al., 2022) enhances CLIP's zero-shot predictions via visual caching, where embeddings from a small image set serve as "keys" in a cache model. During inference, the test image is encoded using the same CLIP Image Encoder, and the label is predicted based on the most similar cached feature. Tip-X (Udandarao et al., 2023) incorporates image-text similarity to boost few-shot performance.

3 Methodology

In this section, we first outline the formulation of the problem we are addressing, followed by a detailed description of each component of our proposed method.

3.1 Algorithm Overview

Our knowledge-enhanced detection method follows a two-stage process. In the first stage, the image and a template target are input, and a self-supervised encoder extracts features for target localization using feature matching. In the second stage, a CLIP-based zero-training prototype cache model is used for semantic classification of construction machinery. Unlike traditional few-shot detection, our method does not require meta-training on base classes, treating all target dataset classes as new.

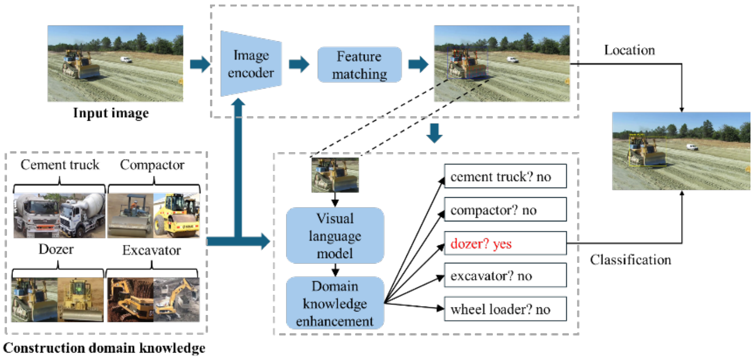

Figure 1 Knowledge-Enhanced Few-Shot Object Detection Method

3.2 Construction Machinery Localization Method

We illustrate the flow of the proposed method using two given construction machinery images as an example, as shown in Figure 2. First, two template construction machinery images and one target construction machinery image are provided. After passing through an image encoder, feature embeddings are generated for each, with pre-defined mask regions for the template images. We use the self-supervised image encoder DINOv2 to extract visual features from the template image x_r and the target image x_t. The encoder computes the output features of the images using self-supervised network parameters, yielding the feature embeddings f_r for the template image and f_t for the target image, where $f_r, f_t \in \mathbb{R}^{H\times W\times C}$. We then compute the feature similarity between the two, with the similarity matrix calculated as follows:

$$SIM = F(x_r) \cdot F(x_t)^T \tag{1}$$

where, SIM represents the feature similarity matrix between the template and target images, and F denotes the pretrained image encoder. $F(x_r) = f_r, F(x_t) = f_t$.

After obtaining the similarity matrix between the template image and the target image, we use the Hungarian algorithm to obtain the most similar coordinates between the target image and the template image. In this case, we treat the similarity matrix SIM as a cost matrix, and the goal is to find the optimal matching pairs between the template image and the target image. Specifically, we define the feature points of the template image as f_r^i and the feature points of the target image as f_t^i. By matching the features of the mask region in the template image with the features in the target image, we obtain the location region B of the construction machinery target in the target image. We define the position information of the feature points P in the target image as $\{(p_i^x, p_i^y)\}_{i=1}^N$, where N is the number of feature points in the construction machinery region of the target image. This relationship can be mathematically defined as follows:

$$B = \begin{cases} min(p_i^x) & s.t. p \in P \\ min(p_i^y) & s.t. p \in P \\ max(p_i^x) & s.t. p \in P \\ max(p_i^y) & s.t. p \in P \end{cases} \tag{2}$$

Then, by converting the obtained feature point coordinates into the real coordinates of the construction image, we obtain the real position information of the construction target.

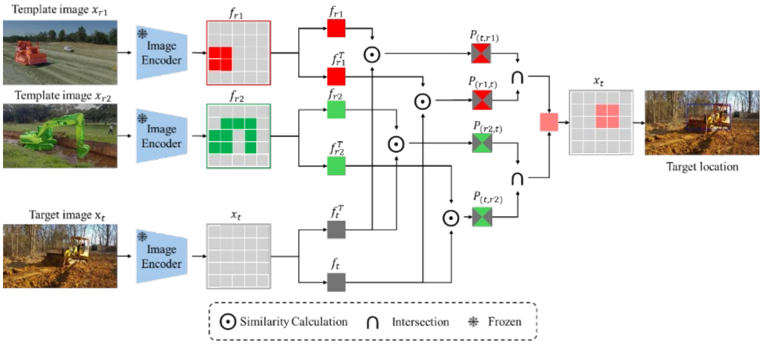

Figure 2 Construction Machinery Localization Method with Integrated Domain Knowledge

3.3 Construction Machinery Category Classification

The CLIP Image Encoder encodes the image into I_1 , and "Zero-shot Classification" generates $Logit_{zeroshot} \in \mathbb{R}^{1\times N}$, where N represents the number of object classes. In this study, we utilize Tip-Adapter to construct a few-shot learning method for classifying construction machinery. First, we build a support set for the few-shot learning method, where the image embeddings and their corresponding class labels are stored as "key-value" pairs, as shown in Figure 3. Let N be the number of categories in the target dataset and K the number of images per class. The total number of images in the Support Set is $n = K \times N$. These n supported images are encoded by the Image Encoder into image embeddings $I_S \in \mathbb{R}^{n\times C}$. The labels of the support images are encoded as One-Hot vectors $L_S \in \mathbb{R}^{n\times N}$.

$$I_S = VisualEncoder(\mathcal{D}_S) \tag{3}$$

$$L_S = OneHot(Labels) \tag{4}$$

The Image Cache is created by storing the image embeddings I_S and their corresponding class labels L_S as "key-value" pairs. The image features I_S serve as the predictive weights when using the Image Cache, while L_S provides the true class labels associated with the Image Cache.

During the prediction phase, a query image is encoded as I_Q. The final image classification result, $Logit_{vis}$, is computed by comparing the query image's embedding with the entries in the Image Cache.

$$A_{vis} = \exp\left(1 - I_Q {I_S}^T\right) \tag{5}$$

$$Logit_{vis} = A_{vis} L_S + I_Q W^T \tag{6}$$

where, W means the classifier weight matrix.

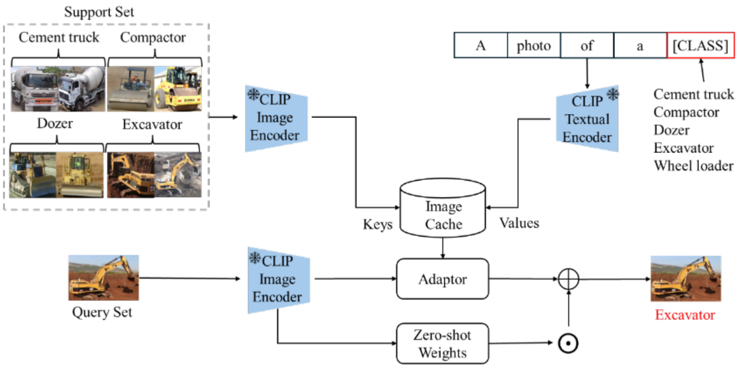

Figure 3 Method for Classifying Construction Machinery Categories

4 Experimental results and analysis

4.1 Datasets and Experimental Details

Two experiments were conducted to validate the proposed method: one for few-shot learning and the other for zero-shot learning. Zero-shot learning relies on pre-trained weights for object recognition without prior domain-specific knowledge, while few-shot learning uses a small number of samples. The zero-shot method in this paper is applied only to object classification, with object localization relying on the proposed domain-knowledge-enhanced approach. The AIMDataset (Xiao & Kang, 2021), a popular construction benchmark, was used for training and testing. The dataset contains 3462 samples, distributed as follows: cement truck (613), compactor (799), dozer (787), excavator (453), and wheel loader (810). The model was tested in various k-shot scenarios (0, 8, 16, 32, 64), and performance was evaluated using the mAP metric (Everingham et al., 2010).

Table 1 Number of images per class

Cement truck	compactor	dozer	excavator	Wheel loader	SUM
613	799	787	453	810	3462

4.2 Performance of the Proposed Method

Table 2 shows experimental results for different few-shot detection scenarios (k = 0, 8, 16, 32, 64). The model achieved a mean Average Precision (mAP) of 60.74% in the 64-shot scenario, detecting all construction equipment, with accuracies of 73.70% for "Compactor," 73.18% for "Cement truck," 57.53% for "Dozer," 54.25% for "Excavator," and 45.05% for "Wheel loader." mAP improved as the number of images increased. However, detection errors were observed. For instance, recognition accuracy for "Dozer" was lower in the 16-shot scenario than in the 8-shot scenario, and the 32-shot scenario had lower accuracy than the 16-shot one. This may be due to the modality gap between image and text (Udandarao et al., 2023), as more samples did not always improve model performance.

Table 2 Experimental Results of the Proposed Method

	0-shot	8-shot	16-shot	32-shot	64-shot
Compactor	2.04	31.53	55.65	46.78	73.70
Cement truck	71.13	75.65	75.03	73.59	73.18
Dozer	52.32	59.04	38.86	68.79	57.53
Excavator	71.06	33.22	70.01	45.22	54.25
Wheel loader	35.47	43.60	57.14	46.03	45.05
Total mAP	46.40	48.61	59.34	56.08	60.74

The proposed few-shot learning method showed a significant positive impact in the experiments. As shown in Table 1, under the same conditions, few-shot learning consistently outperformed zero-shot learning for new classes. These results support the notion that multimodal foundation models with zero-shot capabilities struggle to adapt to specialized industries like construction. This suggests that few-shot learning not only reduces training data requirements but also enables rapid learning from limited data, making it suitable for real-world construction environments with diverse objects.

5 Discussion and limitation

The study demonstrated that the proposed method leverages multimodal foundation models for few-shot learning in construction machinery detection. It first learns feature knowledge from template images to localize construction machinery and then distinguishes categorical attributes using image and textual knowledge. Figure 4 shows that the model performs well, even with varied shapes and colors of heavy equipment. These results suggest that the method effectively utilizes pretrained foundation models and domain-specific knowledge to represent construction scenes.

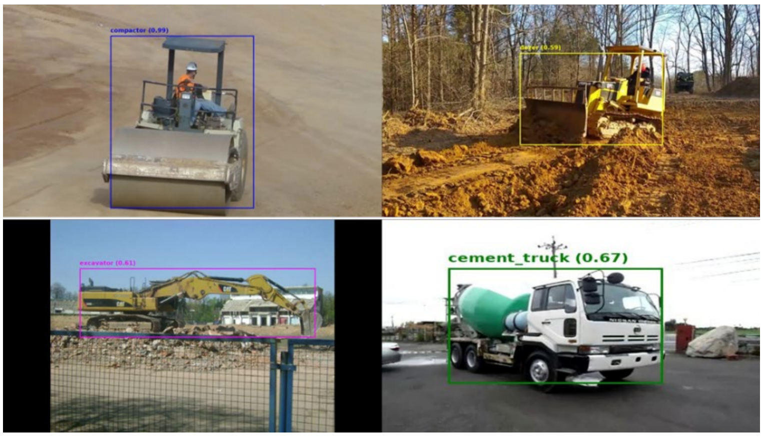

Figure 4 Recognition results of the proposed method

Few-shot models may struggle with target localization. For instance, a model may detect a construction object (e.g., "compactor") but fail to localize it accurately (Figure 5(a)). To address this, we propose using a self-similarity algorithm to differentiate target from background features. By leveraging self-supervised learning, the localization module removes non-target points without fine-tuning. While the matched target should ideally show the highest similarity, resolution differences between template and target images can complicate this. An image pyramid method can extract more detailed information. This study focuses on single-target recognition, with detection of multiple instances in the same category as future work.

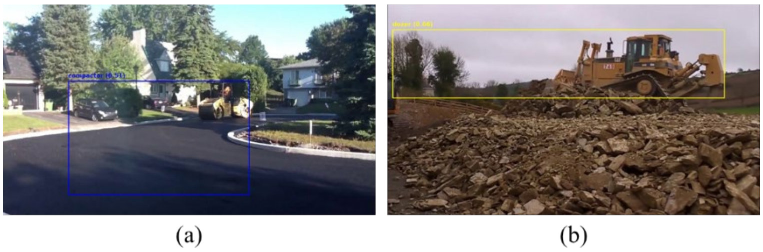

Figure 5 Example of Misidentification in Construction Machinery Localization

6 Conclusion

This paper proposes a Few-shot Object Recognition Method for Construction Site Scenes using a Knowledge-enhanced Vision-Language Multimodal Model, covering machinery localization and classification. In a 64-shot scenario, the detector achieves 60.74% accuracy, while zero-shot learning is limited to 46.4%. The method leverages a visual foundation model for machinery localization without retraining and uses CLIP for text-based semantic classification, reducing data labeling costs and improving the system's applicability on construction sites. The study introduces a domain-enhanced multimodal few-shot learning approach, providing quantitative results to demonstrate its feasibility. These findings lay the foundation for future research in vision-based construction monitoring and other areas, such as safety monitoring and human-machine collaboration.

Acknowledgements

None

Funding

This work was supported by the National Natural Science Foundation of China [grant number NSFC22EG19]; the Research Grants Council (Hong Kong) [grant number 26208323, C6044-23GF].

Data Availability Statement

The data mentioned in the article is publicly available AIMDataset and does not involve any private datasets.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020-December*.

Cui, Z., Wang, Q., Guo, J., & Lu, N. (2022). Few-shot classification of façade defects based on extensible classifier and contrastive learning. *Automation in Construction*, *141*. https://doi.org/10.1016/j.autcon.2022.104381

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4

Ham, Y., Han, K. K., Lin, J. J., & Golparvar-Fard, M. (2016). Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works. In *Visualization in Engineering* (Vol. 4, Issue 1). https://doi.org/10.1186/s40327-015-0029-z

Kim, J. (2020). Visual Analytics for Operation-Level Construction Monitoring and Documentation: State-of-the-Art Technologies, Research Challenges, and Future Directions. In *Frontiers in Built Environment* (Vol. 6). https://doi.org/10.3389/fbuil.2020.575738

Kim, J., & Chi, S. (2017). Adaptive Detector and Tracker on Construction Sites Using Functional Integration and Online Learning. *Journal of Computing in Civil Engineering*, *31*(5). https://doi.org/10.1061/(asce)cp.1943-5487.0000677

Kim, J., & Chi, S. (2021). A few-shot learning approach for database-free vision-based monitoring on construction sites. *Automation in Construction*, *124*. https://doi.org/10.1016/j.autcon.2021.103566

Kim, J., Ham, Y., Chung, Y., & Chi, S. (2019). Systematic Camera Placement Framework for Operation-Level Visual Monitoring on Construction Jobsites. *Journal of Construction Engineering and Management*, *145*(4). https://doi.org/10.1061/(asce)co.1943-7862.0001636

Liang, Y., Vadakkepat, P., Chua, D. K. H., Wang, S., Li, Z., & Zhang, S. (2024). Recognizing temporary construction site objects using CLIP-based few-shot learning and multi-modal prototypes. *Automation in Construction*, *165*, 105542. https://doi.org/https://doi.org/10.1016/j.autcon.2024.105542

Liu, K., & Golparvar-Fard, M. (2015). Crowdsourcing Construction Activity Analysis from Jobsite Video Streams. *Journal of Construction Engineering and Management*, *141*(11). https://doi.org/10.1061/(asce)co.1943-7862.0001010

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2023). *DINOv2: Learning Robust Visual Features without Supervision*.

Paneru, S., & Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities & challenges. In *Automation in Construction* (Vol. 132). https://doi.org/10.1016/j.autcon.2021.103940

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of Machine Learning Research*, *139*.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. In *arXiv preprint arXiv:1804.02767*. http://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rezazadeh Azar, E. (2017). Semantic Annotation of Videos from Equipment-Intensive Construction Operations by Shot Recognition and Probabilistic Reasoning. *Journal of Computing in Civil Engineering*, *31*(5). https://doi.org/10.1061/(asce)cp.1943-5487.0000693

Roberts, D., Torres Calderon, W., Tang, S., & Golparvar-Fard, M. (2020). Vision-Based Construction Worker Activity Analysis Informed by Body Posture. *Journal of Computing in Civil Engineering*, *34*(4). https://doi.org/10.1061/(asce)cp.1943-5487.0000898

Seo, J., Han, S., Lee, S., & Kim, H. (2015). Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, *29*(2). https://doi.org/10.1016/j.aei.2015.02.001

Teizer, J. (2015). Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Advanced Engineering Informatics*, *29*(2). https://doi.org/10.1016/j.aei.2015.03.006

Udandarao, V., Gupta, A., & Albanie, S. (2023). SuS-X: Training-Free Name-Only Transfer of Vision-Language Models. *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV51070.2023.00257

Wang, X., & El-Gohary, N. (2024). Few-shot object detection and attribute recognition from construction site images for improved field compliance. *Automation in Construction*, *167*, 105539. https://doi.org/https://doi.org/10.1016/j.autcon.2024.105539

Wang, Y., Liao, P. C., Zhang, C., Ren, Y., Sun, X., & Tang, P. (2019). Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety. *Advanced Engineering Informatics*, *42*. https://doi.org/10.1016/j.aei.2019.101001

Xiao, B., & Kang, S.-C. (2021). Development of an Image Data Set of Construction Machines for Deep Learning Object Detection. *Journal of Computing in Civil Engineering*, *35*(2). https://doi.org/10.1061/(asce)cp.1943-5487.0000945

Xu, Y., Bao, Y., Zhang, Y., & Li, H. (2021). Attribute-based structural damage identification by few-shot meta learning with inter-class knowledge transfer. *Structural Health Monitoring*, *20*(4). https://doi.org/10.1177/1475921720921135

Yang, J., Park, M. W., Vela, P. A., & Golparvar-Fard, M. (2015). Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. *Advanced Engineering Informatics*, *29*(2). https://doi.org/10.1016/j.aei.2015.01.011

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., & Li, H. (2022). Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13695 LNCS*. https://doi.org/10.1007/978-3-031-19833-5_29

Zhou, L., Cui, P., Jia, X., Yang, S., & Tian, Q. (2020). Learning to Select Base Classes for Few-Shot Classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR42600.2020.00468

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and do not reflect the views of the Architecture, Buildings, Construction and Cities (ABC2) Journal and/or its editor(s). DFBI Journal and/or its editor(s) disclaim any responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.