2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

EVLOD: Ensemble Vision-Language Open-vocabulary Detection for Construction Site Object Recognition

Yongdong Wang[1], Runze Xiao[1], Jun Younes Louhi Kasahara[1], Shota Chikushi[1,2], Keiji Nagatani[1,3], Atsushi Yamashita[1], and Hajime Asama[1]

***Abstract*— The construction industry faces severe labor shortages, driving the need for robotic automation solutions. However, effective deployment of construction robots requires robust environmental perception capabilities, particularly accurate identification of diverse objects in complex, dynamic construction environments. Closed-set object detection methods are limited to predefined categories, proving inadequate for the highly varied object types encountered on construction sites. This paper introduces EVLOD (Ensemble Vision-Language Open-vocabulary Detection), an ensemble framework that integrates multiple state-of-the-art vision-language models to enable open-vocabulary object detection in construction scenarios. EVLOD employs a voting-based fusion strategy that combines predictions from GroundingDINO and GroundingDINO-CLIP detectors, utilizing their complementary strengths while mitigating individual model weaknesses. The ensemble approach incorporates confidence voting, object name voting, and bounding box voting to produce reliable detections with reduced false positives. Evaluated on a comprehensive dataset of 825 Unmanned Aerial Vehicle (UAV)-captured construction images with 5,020 annotated objects, EVLOD achieves an Average Precision (AP) of 0.49 when Intersection over Union (IoU) equals 0.5, representing a 36.1% improvement over the best-performing baseline. The method effectively reduces detection noise from 5,495 to 3,232 detections. Qualitative analysis reveals primary limitations in detecting small-scale objects and low-contrast elements.**

I. Introduction

The construction industry is currently experiencing a severe labor shortage [1]. The adoption of digital transformation and robotic automation technologies offers promising solutions to this challenge [2], [3]. To operate effectively in complex and dynamic construction environments, construction robots must possess robust environmental perception capabilities, allowing them to accurately identify and interpret various on-site objects and conditions [4], [5]. Traditional approaches have largely focused on isolated atomic tasks such as excavation and loading [6], [7]. However, integrating multiple perception and decision-making techniques can substantially alleviate the burden on human operators, thereby addressing labor shortages.

The rapid advancement of Large Language Models (LLMs) has introduced transformative opportunities in robotics [8], [9]. Utilizing extensive general-purpose knowledge, LLMs enable robots to interpret natural language commands and reduce reliance on human labor through task decomposition and allocation [9]–[11]. Nevertheless, construction sites are significantly more complex than typical application scenarios [12], requiring enhanced precision and comprehensive environmental understanding. A key to enabling autonomous execution of complex robotic tasks is equipping LLMs with accurate situational awareness, which can be facilitated by converting visual data from construction sites into structured textual descriptions [13], thereby enhancing LLMs' understanding of the environment.

Early work in construction-site visual perception focused on specific object detection tasks such as worker identification [14], [15], recognition of personal protective equipment [16], [17], and monitoring of construction machinery [4], [18]. These conventional closed-set object detection approaches are limited to predefined object categories [19], [20], which is insufficient given the diversity of objects present on construction sites. More recently, the emergence of vision-language pre-trained models has opened up new possibilities for open-vocabulary detection [21], [22]. The CLIP model, which learns joint image-text representations through contrastive learning [21], has demonstrated strong zero-shot performance across multiple visual tasks. Building on CLIP, GroundingDINO [23] incorporates textual prompts into object detection, enabling open-vocabulary detection.

In construction applications, Cai et al. [24] integrated GroundingDINO into the automatic recognition of prefabricated construction components, validating the feasibility of text-guided detection across several construction datasets. Bang et al. [13] proposed a method to generate contextual information from UAV imagery, enabling automated textual descriptions of construction resources. However, recent evaluation studies [25] have highlighted that single open-vocabulary models often struggle in complex construction environments. Although datasets like SODA [26], CIS [27], and MOCS [28] provide key benchmarks for evaluating open-vocabulary detection, research on integrating multiple model strategies remains limited.

To address the limitations of single-model detectors, ensemble learning has been shown to improve performance in machine learning and conventional (closed-set) vision tasks [29], [30]. However, its use in open-vocabulary object detection has received comparatively less attention; most contemporary approaches adopt single-model designs [31], [32]. CLIP-based few-shot recognition methods [33], [34]

[1]Yongdong Wang, Runze Xiao, Jun Younes Louhi Kasahara, Shota Chikushi, Keiji Nagatani, Atsushi Yamashita, and Hajime Asama are with The University of Tokyo, Tokyo 113-8656, Japan
[2]Shota Chikushi is also with Kindai University, Hiroshima 739-2116, Japan
[3]Keiji Nagatani is also with The University of Tsukuba, Ibaraki 305-0006, Japan
Correspondence to wangyongdong@robot.t.u-tokyo.ac.jp

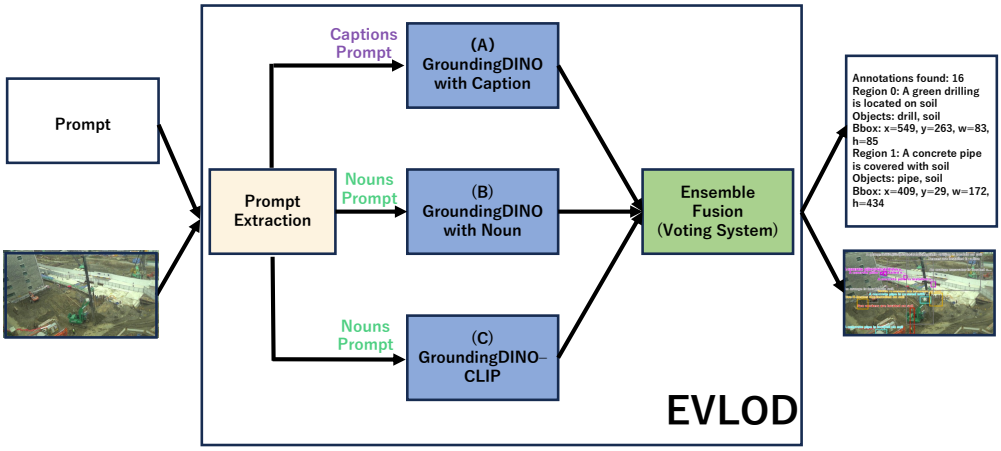

Fig. 1. Overview of the EVLOD ensemble framework for open-vocabulary object detection in construction sites. The system consists of three parallel detection branches: (A) GroundingDINO with complete image captions for sentence-level semantic understanding, (B) GroundingDINO with extracted noun for keyword-based detection, and (C) GroundingDINO-CLIP integration for two-stage detection. All detection candidates are processed through a voting-based fusion mechanism that performs confidence voting, object name voting, and bounding box voting to produce reliable final detections with reduced false positives.

have shown promising transferability in detecting temporary construction-related objects, offering valuable insights for vision-language applications in this domain. Moreover, the advancement of multimodal fusion techniques [35] and attention mechanisms [36], [37] provides a robust technical foundation for developing more powerful ensemble detection frameworks.

In this paper, we propose EVLOD (Ensemble Vision-Language Open-vocabulary Detection), as illustrated in Fig. 1, an ensemble-based vision-language framework tailored for object recognition in construction environments. By integrating predictions from several advanced models, including GroundingDINO and GroundingDINO-CLIP, EVLOD achieves a notable improvement in detection performance. Specifically, it improves the Average Precision (AP) metric from 0.36 to 0.49 when IoU equals 0.5, representing a 36.1% gain, while also achieving higher precision than the best-performing single baseline model.

II. Proposed Method

This section details the EVLOD method proposed for enhancing open-vocabulary object detection. EVLOD addresses misdetections using a voting-based strategy, significantly improving accuracy through a majority voting mechanism.

A. Overall Framework

The core idea of EVLOD is to enhance detection reliability through parallel detection and collaborative decision-making by multiple independent detectors. As illustrated in Fig. 1, the method comprises three hierarchical algorithms: the main algorithm handles the overall process control, the voting fusion algorithm manages grouping and decision-making based on detection results, and the weighted voting algorithm performs the actual fusion computation.

Algorithm 1 EVLOD: Ensemble Voting for Open-Vocabulary Detection

Require: Image I, captions $\mathcal{C}$, weights (w_A, w_B, w_C), voting threshold $V_{min} = 2$
Ensure: Final detections $\mathcal{D}$
Prompt Preparation:
$\mathcal{T}_{cap} \leftarrow$ clean captions from $\mathcal{C}$
$\mathcal{Q}_{noun} \leftarrow$ extract nouns from $\mathcal{C}$
Multi-Method Detection:
$\mathcal{B}_A \leftarrow$ GroundingDINO$(I, \mathcal{T}_{cap})$ with weight w_A
$\mathcal{B}_B \leftarrow$ GroundingDINO$(I, \mathcal{Q}_{noun})$ with weight w_B
$\mathcal{B}_C \leftarrow$ GroundingDINO-CLIP$(I, \mathcal{Q}_{noun})$ with weight w_C
Ensemble Voting:
$\mathcal{U} \leftarrow \mathcal{B}_A \cup \mathcal{B}_B \cup \mathcal{B}_C$
$\mathcal{D} \leftarrow$ VotingFusion$(\mathcal{U}, V_{min})$ ▷ Algorithm 2
return $\mathcal{D}$

Algorithm 1 outlines the complete EVLOD detection workflow. The process begins with preprocessing of textual prompts, including cleaning image captions and extracting key nouns. This is followed by three parallel detection branches with distinct strategies and inputs. Branch A uses GroundingDINO to process complete image captions, utilizing sentence-level semantics for object detection. Branch B also employs GroundingDINO but focuses on extracted nouns, achieving efficient detection via keyword matching. Branch C integrates the GroundingDINO Region Proposal Network with CLIP-based semantic scoring to perform two-stage detection using extracted noun inputs. Each branch employs distinct textual inputs and detection strategies to optimize performance in different scenarios. The main algorithm collects all candidate detections and forwards them to the voting fusion module to determine the final detection results.

Algorithm 2 Voting-Based Detection Fusion

Require: Detection candidates $\mathcal{U}$, voting threshold V_{min}
Ensure: Voted detections $\mathcal{D}$
Group overlapping detections by IoU ≥ 0.4: $\mathcal{G} = \{\mathcal{G}_k\}$
$\mathcal{D} \leftarrow \emptyset$
for all group $\mathcal{G}_k \in \mathcal{G}$ **do**
  if $|\mathcal{G}_k| \geq V_{min}$ **then** ▷ Voting threshold check
    $(b^*, s^*, y^*) \leftarrow$ WEIGHTEDVOTING($\mathcal{G}_k$) ▷ Algorithm 3
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(b^*, s^*, y^*)\}$
  end if
end for
Apply NMS and confidence filtering to $\mathcal{D}$
return $\mathcal{D}$

Algorithm 3 Weighted Voting for Detection Group

Require: Detection group $\mathcal{G} = \{(b_i, s_i, y_i, w_i)\}$
Ensure: Fused detection (b^*, s^*, y^*)
Confidence Voting:
$s^* \leftarrow \sum_i w_i \cdot s_i$ ▷ Weighted confidence sum
Object Name Voting:
$votes \leftarrow \{\}$
for all $(b_i, s_i, y_i, w_i) \in \mathcal{G}$ **do**
  $votes[y_i] \leftarrow votes[y_i] + w_i$
end for
$y^* \leftarrow \arg\max_y votes[y]$ ▷ Majority voting
Box Coordinate Voting:
$b^* \leftarrow \frac{\sum_i w_i \cdot b_i}{\sum_i w_i}$ ▷ Method-weighted average
return (b^*, s^*, y^*)

B. Group-Level Decision

This module handles the grouping of spatially overlapping detections and filters out unreliable results, retaining only those supported by multiple detection methods. Algorithm 2 introduces EVLOD's core innovation: a voting-based fusion mechanism. It begins by grouping overlapping detections using an IoU threshold, with each group corresponding to multiple detections of the same object. A critical step is the voting threshold check: only groups containing at least $V_{\min}$ detections from distinct methods (where $V_{\min} = 2$) are considered reliable. This effectively filters out spurious detections generated by a single method, ensuring that final outputs reflect multi-method consensus. For groups passing the voting threshold, the algorithm invokes the weighted voting module for detailed fusion. Non-maximum suppression and confidence-based filtering are then applied to produce refined detection outputs.

C. Fusion Computation

This module integrates multiple detections within each group to produce unified outputs via weighted voting. Algorithm 3 applies three types of weighted voting to integrate detection information.

The fusion process employs confidence voting to compute the weighted sum of normalized confidences, where the unified confidence score is calculated as

$$s^* = \sum_i w_i \cdot s_i, \tag{1}$$

where s^* denotes the fused confidence score, w_i denotes the method weight, and s_i denotes the normalized confidence.

$$y^* = \arg\max_y \sum_{i\,:\,y_i = y} w_i, \tag{2}$$

where y^* is the selected object class name, y_i represents the class name of the i-th detection, and the condition $i : y_i = y$ indicates all detections with class name y. Bounding box voting averages coordinates in proportion to method weights through

$$b^* = \frac{\sum_i w_i \cdot b_i}{\sum_i w_i}. \tag{3}$$

where b^* denotes the fused bounding box coordinates and b_i represents the bounding box coordinates of the i-th detection.

III. EXPERIMENTS

A. Experimental Setup

We evaluate our method on 825 UAV-captured construction site images from the ConstrUAV dataset [13], as illustrated in Fig. 2, comprising 5,020 annotated ground truth instances. The dataset originates from several real-world construction projects, with image resolutions of 960×540, and includes diverse construction elements such as workers, equipment, and materials.

The weights for EVLOD's ensemble components are tuned based on validation set performance: GroundingDINO-Caption is assigned a weight of 0.4, while both GroundingDINO-Noun and GroundingDINO-CLIP are weighted at 0.3. Key parameters are set as follows: a minimum consensus threshold of 2 to ensure agreement across multiple methods; a merging IoU threshold of 0.4 for controlling detection grouping; a Non-Maximum Suppression (NMS) IoU threshold of 0.6 to suppress duplicate detections; and a confidence threshold of 0.15 to filter low-quality outputs.

The evaluation metrics are Average Precision when IoU equals 0.5 and Recall when IoU equals 0.5. We also report the total number of detections and inference time. All experiments are conducted on an NVIDIA RTX 4090 GPU.

B. Experimental Results

Table I compares the performance of EVLOD against major baselines. EVLOD achieves an AP@0.5 of 0.49, representing a 36.1% improvement over the strongest baseline, GroundingDINO (0.36), demonstrating the effectiveness of our ensemble strategy for open-vocabulary detection tasks.

EVLOD produces 3,232 detections, which is 41.2% fewer than GroundingDINO's 5,495, suggesting that our method effectively reduces false positives and improves precision. Although the recall (0.31) is slightly lower than that of GroundingDINO (0.39), the considerable gain in precision is more critical for applications in construction robotics. The inference time is 3.83 seconds, longer than that of a single model, but this overhead is acceptable given the performance benefits.

The YOLO-CLIP method performs the poorest, achieving an AP@0.5 of only 0.07 and generating 21,144 detections,

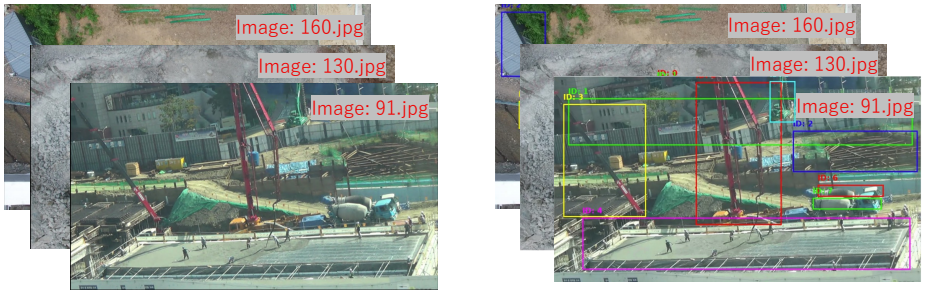

Image: 91.jpg Annotations found: 8

Region 0: A red crane truck is located on soil
Objects: crane
Bbox: x=371, y=15, w=224, h=374

Region 1: Fences are installed in a line
Objects: fence
Bbox: x=37, y=57, w=901, h=123

...

...

Region 6: A white concrete truck is located on soil
Objects: truck
Bbox: x=690, y=284, w=172, h=31

Region 7: A blue concrete truck is located on soil
Objects: truck
Bbox: x=677, y=318, w=180, h=30

Fig. 2. Representative samples from the ConstrUAV dataset [13] illustrating the annotation process and prompt generation. Left: original UAV-captured construction site images. Right: ground truth annotations with bounding boxes for diverse objects including workers, equipment, and materials. Below each image pair: annotation information with Caption Prompts (complete descriptive sentences) and Noun Prompts (extracted key object terms) used for different detection branches in EVLOD.

TABLE I
PERFORMANCE COMPARISON ON UAV-CONSTRUCTION BENCHMARK.

Method	AP@0.5 ↑	Recall@0.5 ↑	Dets	Time (s/img) ↓
GroundingDINO [24]	0.36	**0.39**	5,495	0.99
YOLO-CLIP [38]	0.07	0.28	21,144	**0.08**
Ensemble (Ours)	**0.49**	0.31	3,232	3.83

indicating severe over-detection. Despite having the fastest inference time (0.08 seconds), the detection quality is inadequate for practical deployment.

C. Qualitative Analysis

Fig. 3 illustrates EVLOD's detection performance and limitations. Image 187 (P:0.88, R:0.64) exhibits missed detections of gray-colored objects with low visual contrast, while Image 793 (P:1.00, R:0.57) shows missed detections of small worker instances. The high precision in both cases indicates accurate detection of large-scale objects, while the moderate recall reflects systematic failures with low-contrast elements and small-scale objects respectively.

Image 420 (P:0.00, R:0.00) represents complete detection failure due to insufficient illumination and object scales below the detection threshold. This demonstrates the method's sensitivity to lighting conditions and scale limitations.

IV. CONCLUSION AND FUTURE WORK

This paper presented EVLOD, an integrated vision-language open-vocabulary object detection framework designed for construction site environments. By aggregating predictions from multiple detectors, including GroundingDINO and GroundingDINO combined with CLIP, and by employing a voting-based ensemble decision strategy, EVLOD significantly enhanced detection performance. When evaluated on a dataset comprising 825 UAV-captured construction images, EVLOD increased the AP@0.5 from 0.36, which was the best-performing baseline, to 0.49. This corresponds to a 36.1% relative improvement. The voting mechanism effectively reduced false positives, decreasing the number of predicted detections from 5,495 to 3,232, thereby improving precision. Future work will focus on enhancing small-scale object detection through multi-scale feature fusion and improving robustness to illumination variations. In addition, future work will investigate model compression, dynamic routing, and more efficient ensemble methods to narrow the gap between EVLOD's detection accuracy and its suitability for near real-time deployment on autonomous construction robots.

ACKNOWLEDGMENT

This work is supported by JST [Moonshot Research and Development], Grant Number [JPMJMS2032].

REFERENCES

[1] Policy Bureau, Ministry of Land, Infrastructure, Transport and Tourism (MLIT), "Summary of the white paper on land, infrastructure, transport and tourism in japan, 2024," Ministry of Land, Infrastructure,

Fig. 3. Qualitative detection results of EVLOD on construction site images. Each row shows (a) original image, (b) ground truth annotations, and (c) EVLOD results with precision (P) and recall (R) metrics. Image 187 demonstrates successful detection of large-scale objects while missing low-contrast gray elements. Image 793 shows accurate detection of equipment and materials but fails to detect small worker instances. Image 420 represents a complete failure case due to poor illumination and small object scale.

Transport and Tourism, Japan, Tech. Rep., 2024. [Online]. Available: https://www.mlit.go.jp/statistics/content/001855598.pdf

[2] T. Bock, "The future of construction automation: Technological disruption and the upcoming ubiquity of robotics," *Automation in construction*, vol. 59, pp. 113–121, 2015.

[3] Y. Pan and L. Zhang, "Roles of artificial intelligence in construction engineering and management: A critical review and future trends," *Automation in Construction*, vol. 122, p. 103517, 2021.

[4] J. Kim, Y. Ham, Y. Chung, and S. Chi, "Camera placement optimization for vision-based monitoring on construction sites," in *Proceedings of the International Symposium on Automation and Robotics in Construction (IAARC)*, 2018.

[5] S. Bang, H. Kim, and H. Kim, "Uav-based automatic generation of high-resolution panorama at a construction site with a focus on preprocessing for image stitching," *Automation in construction*, vol. 84, pp. 70–80, 2017.

[6] C. Haas, M. Skibniewski, and E. Budny, "Robotics in civil engineering," *Computer-Aided Civil and Infrastructure Engineering*, vol. 10, no. 5, pp. 371–381, 1995.

[7] J. Zhang, H. Luo, and J. Xu, "Towards fully bim-enabled building automation and robotics: A perspective of lifecycle information flow," *Computers in Industry*, vol. 135, p. 103570, 2022.

[8] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[9] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International conference on machine learning*. PMLR, 2022, pp. 9118–9147.

[10] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," *arXiv preprint arXiv:2209.11302*, 2022.

[11] Y. Wang, R. Xiao, J. Y. L. Kasahara, R. Yajima, K. Nagatani, A. Yamashita, and H. Asama, "Dart-llm: Dependency-aware multi-robot task decomposition and execution using large language models," *arXiv preprint arXiv:2411.09022*, 2024.

[12] J. Teizer, "Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 225–238, 2015.

[13] S. Bang and H. Kim, "Context-based information generation for managing uav-acquired data using image captioning," *Automation in Construction*, vol. 112, p. 103116, 2020.

[14] W. Fang, L. Ding, B. Zhong, P. E. Love, and H. Luo, "Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach," *Advanced Engineering Informatics*, vol. 37, pp. 139–149, 2018.

[15] H. Son, H. Choi, H. Seong, and C. Kim, "Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks," *Automation in Construction*,

vol. 99, pp. 27–38, 2019.
[16] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," *Automation in construction*, vol. 112, p. 103085, 2020.
[17] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset," *Automation in construction*, vol. 106, p. 102894, 2019.
[18] D. Kim, M. Liu, S. Lee, and V. R. Kamat, "Remote proximity monitoring between mobile construction resources using camera-mounted uavs," *Automation in Construction*, vol. 99, pp. 168–182, 2019.
[19] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
[20] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.
[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceeding International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
[22] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.
[23] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.
[24] R. Cai, Z. Guo, X. Chen, J. Li, Y. Tan, and J. Tang, "Automatic identification of integrated construction elements using open-set object detection based on image and text modality fusion," *Advanced Engineering Informatics*, vol. 64, p. 103075, 2025.
[25] A. Abdalwhab, A. Imran, S. Heydarian, I. Iordanova, and D. St-Onge, "Are open-vocabulary models ready for detection of mep elements on construction sites," *arXiv preprint arXiv:2501.09267*, 2025.
[26] R. Duan, H. Deng, M. Tian, Y. Deng, and J. Lin, "Soda: A large-scale open site object detection dataset for deep learning in construction," *Automation in Construction*, vol. 142, p. 104499, 2022.
[27] X. Yan, H. Zhang, Y. Wu, C. Lin, and S. Liu, "Construction instance segmentation (cis) dataset for deep learning-based computer vision," *Automation in Construction*, vol. 156, p. 105083, 2023.
[28] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, "Dataset and benchmark for detecting moving objects in construction sites," *Automation in Construction*, vol. 122, p. 103482, 2021.
[29] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.
[30] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review*, vol. 33, no. 1, pp. 1–39, 2010.
[31] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European conference on computer vision*. Springer, 2022, pp. 350–368.
[32] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple open-vocabulary object detection," in *European conference on computer vision*. Springer, 2022, pp. 728–755.
[33] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
[34] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.
[35] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[38] S. Matsuzaki, T. Sugino, K. Tanaka, Z. Sha, S. Nakaoka, S. Yoshizawa, and K. Shintani, "Clip-loc: Multi-modal landmark association for global localization in object-based maps," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 673–13 679.