Construction Industry Vision Alberta Dataset (CIVAD): Developing a Comprehensive Object Detection Dataset for Diverse Construction Applications

Mohamed Sabek[1], [2], Qipei Mei[1], [3], Gaang Lee[1], [4], Ali Golabchi[1], [5], and Vicente Gonzalez[1], [6]
1 Infrastructure and Human Tech Lab (https://www.iht-lab.com/), Department of Civil and Environmental Engineering, Faculty of Engineering, Donadeo Innovation Centre for Engineering, University of Alberta, 116 Street NW, Edmonton T6G 1H9, Alberta, Canada
2 sabek@ualberta.ca , **3** qipei.mei@ualberta.ca , **4** gaang@ualberta.ca , **5** alireza1@ualberta.ca ,**6** vagonzal@ualberta.ca

Abstract

Integrating computer vision technologies into the construction industry has the potential to revolutionize site monitoring, safety management, and quality control. However, a critical gap remains in the availability of specialized datasets tailored to construction sites' distinct conditions and complexities while including sufficient classes representing most items in construction sites. Existing Computer Vision (CV) models often rely on generic training datasets, which limit their effectiveness for specific construction-monitoring-related tasks. Consequently, there is a pressing need for comprehensive domain-specific datasets that can capture the full spectrum of construction-related objects and activities.

This study addresses this gap by developing a foundational training dataset called the Construction Industry Vision Alberta Dataset (CIVAD), specifically designed for CV applications in the construction sector. Our dataset included over 50 classes with more than 86,905 images of different objects, such as tools, machinery, safety equipment, and construction materials, to support diverse CV tasks. It utilizes a combination of web scraping, inclusion of existing open-source datasets, and direct data captured from construction sites.

A set of novel methods, such as semiauto-labeling with advanced models, such as Grounded SAM and Grounding DINO, were used with our custom algorithms. These models were utilized to process parts of the dataset imagery with humans in the loop. This approach facilitated an efficient and accurate dataset creation process. The CIVAD dataset and methods employed in this study represent a significant step forward in integrating CV technologies across various construction-related applications.

Keywords – Construction Industry, Computer Vision, Foundational Training Dataset, Object Detection, Auto-labeling, YOLO v11, Site Monitoring

1 Introduction

The construction industry is one of the largest and most complex sectors in the world and has a significant impact on the global economy [1]. It builds and maintains physical infrastructure, including roads, bridges, buildings, and other critical structures that support modern society. It remains a highly diverse and dynamic sector characterized by inherent challenges, including constantly evolving project environments, diverse stakeholder requirements, and the interplay of numerous variables and uncertainties that directly impact project outcomes [2]. These complexities require innovative solutions to enhance efficiency, ensure safety, and maintain quality standards under ever-changing conditions.

Although digital technologies have gradually penetrated the construction industry, they are yet to catalyze a fundamental transformation in their practices. The industry remains manual and fragmented, grappling with longstanding inefficiencies and challenges that underscore the nascent stage of it is digital evolution and highlight the gap in integrating Industry 4.0 [3]. However, in recent years, the construction industry has experienced several changes, primarily fueled by the growing implementation of artificial intelligence technologies, particularly in subfields such as computer vision.

CV has the potential to revolutionize various aspects of construction [4], including site monitoring, progress tracking, and safety compliance monitoring. CV technologies use cameras and other sensors to capture and analyze visual data, thereby providing valuable insights and information that can be used to improve construction processes and outcomes [12]. Despite the potential of CV technologies, their adoption in the construction industry has been limited. One of the most notable barriers is the lack of standardized datasets. Developing field-applicable computer vision models entails datasets tailored to the unique and dynamic conditions of construction sites, including diverse construction-related objects and environmental variability [13].

However, creating these datasets is challenging. First, collecting and annotating large amounts of visual data is labor-intensive and costly, and requires specialized knowledge of construction processes and CV techniques. In addition, the fragmented nature of the construction industry, in which projects often vary widely in scale, location, and standards, makes it difficult to develop universal datasets that replicate the complex environments of construction sites [4].

2 Related Work

Computer vision technologies are increasingly applied in the construction industry to help perform specific tasks [5], such as monitoring and tracking construction equipment at construction sites. This cannot be achieved without proper training datasets to train the CV models on these activities, which is essential for building any CV model. The training datasets help the model to understand it as a deployment environment, which in this case is a construction site. To support these efforts, researchers have developed several domain-specific datasets that can be used to train and evaluate CV models in construction-related applications. This study examines four widely utilized and cited publicly available construction datasets: MOCS [6], SODA [7], ACID [8], and CIS [9], highlighting their contributions and limitations and demonstrating how these shortcomings highlight the need for the development of more comprehensive and versatile datasets.

The MOCS (Moving Objects in Construction Sites) dataset contains roughly 41,668 images sourced from 174 construction sites. It consists of 13 classes of moving objects commonly found in construction environments, including workers, various types of cranes, vehicles, and machinery. Although MOCS has proven valuable for safety analysis, activity recognition, and productivity monitoring, the limited number of object classes restricts it is applicability to a narrow range of elements commonly encountered in traditional construction settings. In addition, it focuses on dynamic entities and overlooks the material, environmental, and structural aspects, making the trained models less adaptable to complex or unconventional construction scenarios.

Although MOCS has broader coverage, the Site Object Detection Dataset (SODA) features 19,846 images across 15 classes, offering a more balanced approach to object selection. However, the overall scale and scope of the dataset remain limited because of the relatively small number of classes and limited extent of scene variability. Models trained merely on SODA are not sufficiently generalizable to cover many items found within the construction environment, such as different types of construction equipment tools and materials, because of the limited scope of the SODA dataset in terms of the number of classes. A significant limitation of both the MOCS and SODA datasets is their geographic bias, as most of the annotated data were collected exclusively from construction sites in Asian countries. This narrow scope restricts the dataset's ability to generalize to North American or European contexts, where differences in color schemes, signage, safety banner languages, and weather conditions, such as snow, can significantly affect the performance and accuracy of the trained models. Another existing dataset is the Alberta Construction Image Dataset (ACID), which consists of 10,000 images across 10 classes with instance segmentation and image captioning. It primarily focuses on heavy construction equipment. Despite it is usefulness in machinery identification and detection, ACID is narrow and has only a few categories, limiting it is applicability to more integrated applications in construction [23].

Another important dataset is the Construction Instance Segmentation (CIS) dataset, which has many unique features that differentiate it from other construction-specific datasets. The dataset contained 50,000 images with over 100,000 instances annotated across 10 different classes. One of the most notable aspects is the inclusion of Prefabricated Building Construction (PBC) objects such as precast components (PCs) and PC delivery trucks, which are required for computer vision-based studies on PBC management. Moreover, the dataset distinguishes between workers who wore safety helmets and those who did not, making helmet-wearing detection more efficient than separately detecting helmets and workers. The CIS dataset spans a large set of construction sites, viewpoints, and weather, providing a variety of construction-specific object segmentation, helmet detection, and large instance segmentation data, thereby addressing gaps in the current PBC-related object segmentation research. However, the CIS dataset still suffers from the same limitations that exist in all the previously discussed datasets, which is the limited number of classes included in these datasets, which makes them only cover a specific portion of machinery, tools, workers, and materials existing in traditional construction sites.

These limitations make it challenging to build and train accurate CV models that are capable of handling the diverse needs of traditional construction sites. These sites typically involve multiple vehicle types, various tools, numerous workers, and assorted materials that must be tracked and classified. Consequently, tasks beyond equipment-related recognition, such as integrated site monitoring, schedule compliance checks, and resource allocation, are incredibly challenging without broader scene coverage.

Although merging existing datasets to mitigate their individual gaps could be considered, in practice, combining them with CV models such as YOLO is difficult because of potential issues such as data misalignment, differing annotation formats, inconsistent image resolutions, varied augmentation strategies, and ensuring that all datasets use the same classes [10].

3 Addressing Shortcomings in Construction Datasets: Why CIVAD Is Necessary

Although pioneering datasets, such as MOCS, SODA, ACID, and CIS, have advanced CV research in construction, they share limitations that hinder broader industry adoption, highlighting the need for a new dataset. This dataset captures the unique complexities of construction sites, including variable lighting, occlusions, and diverse camera viewpoints. Unlike static environments, construction sites often face unpredictable conditions [11]. Additional factors, such as local safety regulations and specialized machinery in places such as Canada, further underscore the need for more comprehensive construction-specific datasets.

In response to these limitations, we created a new dataset called the Construction Industry Vision Alberta Dataset (CIVAD), which offers a more versatile training option for researchers looking into training computer vision models to implement in different types of applications within the construction industry. These applications can include productivity estimations, vehicle tracking and monitoring, safety monitoring of construction individuals, and identification of potential bottlenecks in different construction processes. The CIVAD contains more than 86,905 images across 50 classes, and Figure 1 shows a sample of the dataset. In terms of class numbers, diversity, and the general number of annotated images, CIVAD covers much more than any other available computer vision dataset, specifically for construction applications, as listed in Table 1.

Figure 1- a sample of the CIVAD Dataset shows different classes

Table 1. A comparison between CIVAD and other datasets

	Dataset name				
Feature/Metric	**MOCS**	**SODA**	**ACID**	**CIS**	**CIVAD**
Total Images	41,668	19,846	10,000	50,000	86,905
Number of Classes	13	15	10	10	50
Annotation Types	Bounding Boxes, Masks	Bounding Boxes	Bounding Boxes, Limited Segmentation	Bounding Boxes, Masks	Bounding Boxes, Multiple Formats
Regional Scope	Asia	Asia	Canada	Asia	Worldwide
Objects Covered	Moving Objects	Site Objects	Heavy Machinery	PBC Objects, Helmet Detection	Workers, Tools, Machinery, Materials, Safety Equipment
Specialized Classes	Vehicles, Workers	Equipment, Tools	Machinery	Workers (Helmet Detection), PBC	Machinery, Workers, Safety Equipment, Tools, Materials
Format ready to Fintune Vision-Language Models	No	No	No	No	Yes
Application Areas	Safety Analysis, Activity Recognition	Object Detection	Machinery Identification	PBC Management, Safety Compliance	Safety Monitoring, Productivity Analysis, Resource Management
Unique Features	Focus on dynamic objects	Balance between object types	Heavy machinery emphasis	Helmet detection, PBC-specific data	Comprehensive data diversity
Data Collection Method	Site Collection	Site Collection	Open-source, Site Collection	Site Collection	Web Scraping, Site Collection, Augmented

CIVAD is unique because it is data diverse, which includes machinery, tools, workers, safety equipment, materials, and infrastructure components; therefore, it broadly represents the construction environment. To improve the robustness of the trained CV model on CIVAD, the dataset covers a wide range of weather conditions, lighting conditions, and different camera angles, with different types of augmentations applied to the dataset to expand it is size and versatility. It is also important to mention that our dataset offers training data in more than five different formats, ranging from JSON, TFRecord, and Pascal VOC to the YOLO format in TXT. This diversity is particularly advantageous for fine-tuning Vision-Language Models (VLMs) and vision foundation models, such as Microsoft's Florence-2 [15], which rely on large and variable datasets to learn robust representations.

By incorporating multiple object types under different environmental scenarios, CIVAD helps these models capture subtle visual cues that generic datasets may overlook, such as the unique features of construction equipment or the contextual interplay of workers, machinery, and materials. VLMs demonstrate notable zero-shot detection abilities and can detect and classify objects without explicit training data [16]. CIVAD enables researchers and practitioners to continuously refine the accuracy and generalization of VLM in construction environments, allowing the development of real-world applications. CIVAD fills a gap in the literature by providing a standardized and comprehensive dataset that bridges the gap between academic research and practical implementation.

This can accelerate innovation in computer vision technology and encourage the development of resilient AI models that can reshape construction safety, productivity, and efficiency.

4 CIVAD Dataset Creation Process

The CIVAD dataset was created using a well-defined process with different steps and techniques to efficiently and accurately develop this large dataset. This includes new approaches for making this dataset as efficient as possible, as shown in Figure 2, which shows all steps from data collection to the final dataset composition.

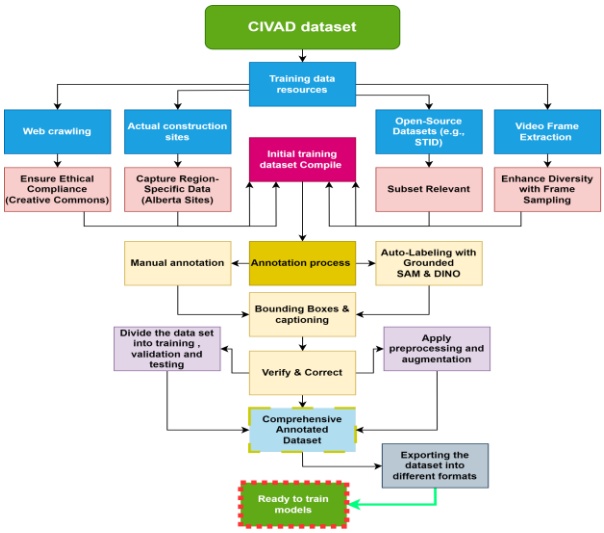

Figure 2. CIVAD dataset creation steps

4.1 included classes

CIVAD currently includes 50 object classes organized into five major categories, as shown in Table 2.

Table 2. CIVAD classes according to their categories

Categories	Classes names				
Machinery	Articulated Haulers	Bulldozer	Concrete Mixer	Concrete Pump	Dump Truck
	Excavator	Forklift	Mobile Crane	Motor Graders	Pile Boring Machine
	Scissor Lift	Sheepsfoot Roller	Skid Steer Loader	Smooth Roller	Static Crane
	Tower Crane	Wheel Loader			
Workers	White Hat Engineers	Red Hat Firefighter	Yellow Hat Worker	Green Hat Safety Officer	-
Tools	Bolts	Wrench	Trowel	Bucket	Cutter
	Drill	Grinder	Hammer	Knife	Saw
	Shovel	Spanner	Tacker		
Safety Equipment	Safety Barrier	Safety Cone	Safety Helmet	Safety Shoes	Safety Sign
	Safety Vest	Traffic Barrel	Welding Helmets	Safety Goggles	
Materials	Wood Plank	Concrete Block	-	-	-

4.2 Image Collection

The process of collecting images for the CIVAD dataset was carefully designed to strike a balance between scale, diversity, and realism, all of which have been recognized as critical factors in developing strong and generalizable computer vision models [27]. For example, large and varied datasets, such as COCO [14], have demonstrated that increasing both the quantity and variety of training samples can significantly improve a model's ability to generalize to real-world scenarios. Additionally, realism ensures that data capture authentic conditions and further refines model performance when transitioning from laboratory to field applications [28].
The CIVAD dataset creation methodology utilizes a combination of approaches, including web scraping, subsetting open-source datasets, and manual data capture to assemble a large, varied, and contextually relevant dataset that aligns with the requirements of the construction industry.

By encompassing a wide range of conditions and subjects, CIVAD aims to mirror the complexity and unpredictability inherent in real construction sites, thereby enhancing the robustness and transferability of the trained models. All images and video sources used were ethically compliant (i.e., open access or licensed under Creative Commons). During scraping, specific criteria were used to ensure that the retrieved images contained relevant construction activity, machinery, workers, tools, and materials. The systematic analysis of many online archives, such as Pixabay [26], has enabled the collection of a large pool of high-resolution visuals that represent real-world construction environments. Frame extraction from videos was used to increase the size and diversity of the dataset.

Individual frames were extracted from high-resolution videos at regular intervals during construction activities. This technique increases the total number of images and brings about natural variations in lighting, angles, and viewpoints, as well as worker/machine interaction, making the photos more realistic and providing rich supervision [29]. Frame extraction ensures that the dataset contains the dynamic and ever-changing nature of construction environments while replicating real-world scenarios better than static imagery alone [30]. We also included a subset of 3,600 pictures from an existing open-source dataset called the Small Tool Image Database (STID), which is under Creative Commons license 4.0 [17]. The STID dataset consists of 12 classes of small tools used in the construction industry, which we used as a base to include these types of tools in our dataset and then expanded it with more images for each class.

This inclusion was motivated by the need to represent small, handheld tools that are critical to everyday construction tasks yet are often underrepresented in general-purpose image datasets. Incorporating STID data helped ensure that these tools, which can significantly influence safety, productivity, and site operations, were adequately covered in the CIVAD. Also, incorporating a pre-annotated subset of 3,600 images from the STID (Small Tool Image Database) significantly reduced the time and effort needed to include 12 classes of small tools in our dataset, sparing us from annotating all such data from scratch. However, many publicly available construction image datasets originate from or predominantly replicate Asian construction environments, which can differ markedly from those found in Canada, the U.S., and Europe; for instance, in the types of equipment used, signage design, and specific safety protocols.

Therefore, we conducted manual data capture at construction sites in Alberta, Canada, where we captured approximately 3,158 pictures representing approximately 10% of the CIVAD without augmentation, to address these regional discrepancies and enhance the applicability of the dataset in Canadian contexts. This approach ensures that our dataset reflects region-specific factors such as environmental conditions and safety standards, which may not be adequately represented in datasets with predominantly Asian or non-Canadian content. Researchers obtain permission from project managers at active construction sites, ranging from commercial developments to infrastructure projects, and adhere to strict guidelines to respect property rights and protect sensitive information. As a result, the collected images more accurately represented the construction machinery, materials, and safety measures typical of North American settings. This emphasis on North American conditions, combined with the diverse set of tools, weather conditions, and site layouts in the dataset, boosts the overall realism and transferability of computer vision applications across regions, with construction practices similar to those in Canada, the U.S., and potentially even parts of Europe.

The final dataset comprised 86,905 images, which were expanded from 32,380 images using different augmentation techniques and strategically split into training, validation, and test sets to support model development and evaluation: a training set of 81,849 images (94%), a validation set of 3,632 images (4%), and a test set of 1,424 images (2%) of the total dataset.

4.3 The Annotation Process

To annotate the CIVAD dataset, the Roboflow platform [18] was utilized to perform accurate and efficient construction image labeling. Because of the large size of the CIVAD dataset (over 32,380 images without augmentation), a semiautomatic annotation approach was used to ensure accuracy and efficiency. Manual annotation and integration of cutting-edge automated tools were used in a process that combined these approaches. Manual annotation was performed to ensure that key objects, such as machinery, tools, workers, and safety equipment, were adequately identified and labeled by applying bounding boxes to the objects of interest in the images to achieve accuracy in object localization. For annotation acceleration and dataset scale management, semi-automatic labeling techniques have used state-of-the-art foundational computer vision models such as Grounded SAM [19] and Grounding DINO [20], which generate automatic predictions of bounding box coordinates.

These models proved helpful for adding annotations to identifiable object classes based on their training data; however, not all objects were correctly identified. Therefore, a thorough manual review was performed to ensure that all the relevant classes were properly labeled in each image. This hybrid approach reduced the annotation time as much as possible without sacrificing the accuracy across the entire dataset. The result was a comprehensive dataset of 32,380 images annotated with over 58,789 instances representing 50 object classes. On average, there are 2.8 annotations per image, demonstrating the complexity and variety of construction site visuals. It's worth noting that the resolution of the images in the dataset ranged from 4 megapixels to 1 megapixel, and the median image ratio for the dataset was 1280x900 pixels, with an average of 1.08 megapixels throughout the whole dataset. Figure 3 shows the number of individual annotation count distributions for images in the CIVAD dataset.

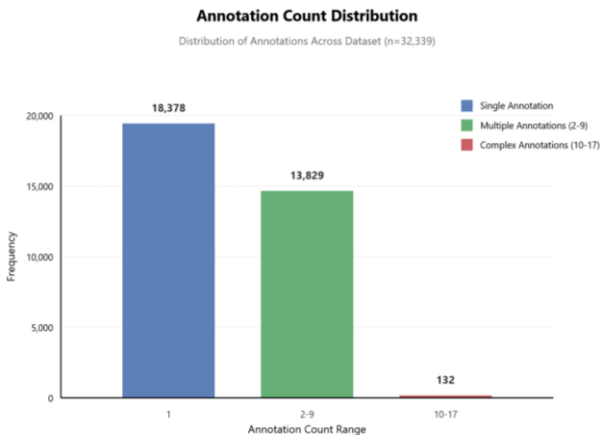

Figure 3. Annotation count distributions for images in the CIVAD dataset.

The attributes of this dataset, such as high-resolution images, diverse object classes, and realistic coverage, make it ready to train robust CV models for object detection and tracking tasks. Higher-resolution images in CIVAD capture subtle details, such as tool features and safety gear, thereby improving accuracy in complex environments. A variety of machinery, tools, and contexts help models learn richer features and generalize more effectively. The combination of manual precision with automated tools leads to annotations with a sufficient scale and the required quality accuracy to serve a range of computer vision applications in the construction industry, as proven by the high accuracy of the trained models on the dataset, which will be discussed in the next section.

4.4 The Applied Preprocessing and Augmentations Process on CIVAD

A series of preprocessing techniques was applied to standardize the images and enhance their usability in preparing the dataset for use in computer vision tasks. We began by performing an auto-orientation step to correct the image alignment by ensuring a consistent orientation based on the EXIF metadata, thereby preventing annotation mismatches caused by misaligned pixel coordinates. This preprocessing step eliminates common issues in computer vision pipelines, such as misrotation, ensuring that the images in the CIVAD dataset are uniformly oriented, have accurate annotations, and are ready for training [25]. All images were then resized to 640 × 640 pixels, which is a standard input size requirement for YOLO-based models [21], to ensure compatibility with state-of-the-art object detection frameworks and reduce computational overhead. Various augmentation techniques have been used to create camera augmentations for simulating conditions commonly seen in real-world construction sites, further diversifying the dataset and improving the model's generalization ability.

The following augmentations were applied:

- Crop: Variations in the scale and focal point of the objects were achieved by cropping the images with 0% minimum zoom and 20% maximum zoom, which gave us good results in our tests.
- Rotation: To simulate tilted or angled camera views, random rotations between -15° and +15° were performed.
- Exposure: Changes in lighting conditions were applied, by changing lighting conditions within a range of -15% to +15%, reflecting the effects of natural and artificial lighting on construction sites.

Augmentation was strategically applied to expand CIVAD from 32,380 original images to 86,905 images, significantly enhancing the dataset diversity without incurring additional annotation costs. The rationale for this augmentation is to simulate real-world variations, such as lighting changes, camera angles, and object perspectives, which are crucial for training robust and generalizable computer vision models, thus reflecting realistic construction site scenarios more accurately.

5 Trained Model Performance Evaluation

In this section, we discuss the training of a YOLOv11-M model [24] on CIVAD's 50 classes to confirm the dataset's quality and alignment with our research goals. The dataset was divided into training (81,849 images), validation (3,632 images), and test (1, 424 images) sets. Training was performed over 300 epochs with a batch size of 16, an initial learning rate of 0.001, and stochastic gradient descent optimization on an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM [22], which required approximately 44 hours to complete. By applying preprocessing and strategic augmentation, we increased the dataset from the original size of 32,380 images to 86,905 images. The rationale behind this augmentation was to simulate real-world variability, such as changes in lighting, camera angles, and environmental conditions, thus enhancing the robustness and generalization capability of the trained computer vision models. To comprehensively evaluate the model's performance, we utilized the mean Average Precision (mAP), Precision, and Recall, which are recognized metrics for evaluating object detection model performance [31]. These metrics reflect the accuracy of the model and can be generalized across diverse construction scenarios, validating both the dataset size and appropriateness of augmentation techniques for simulating real-world variations.

These metrics demonstrate that the model is capable of accurately detecting and classifying objects from the 50 object classes of workers, machinery, tools, materials, and safety equipment across the dataset, as shown in Figure 4, which shows the trained model's performance across some of the CIVAD classes.

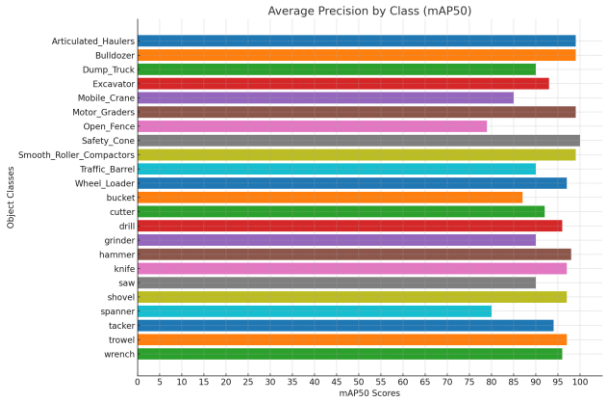

Figure 4. YOLOv11-M training results on common classes from the CIVAD dataset

6 Conclusion

To support the digital transformation of the construction industry, this study introduces some promising annotation techniques along with the CIVAD dataset, which comprises more than 86,905 images across 50 classes, capturing only a subset of the existing items in traditional construction sites. However, this dataset bridges critical gaps in existing datasets and provides vast diversity and applicability. It offers great flexibility through multiple annotation file formats and compatibility with various deep-learning frameworks, including YOLO and Vision-Language Models. CIVAD has great potential for use in safety monitoring, productivity analysis, and automated resource management, which can help improve the efficiency and safety of construction sites. We anticipate that these advancements will foster new opportunities for construction automation and that CIVAD will continue to catalyze innovation in computer vision applications related to the construction industry for many years.

7 Future Work

The CIVAD dataset provides a robust foundation for increasing the number of existing computer vision applications in construction. However, the evolution of the dataset is necessary to meet changing requirements and extend it is scope.

Future plans include:

- Expansion of Dataset Scope: We will increase the number of object classes from 50 to over 100 and the dataset size to more than 120,000 images. The dataset will include more sophisticated tools, machinery, materials, and safety equipment to better represent the different environments at construction sites. This is important for extending the generalization capabilities of the dataset.
- Enhancing Dataset Diversity: Introducing rare and region-specific scenarios of construction, such as nighttime activities, extreme weather, and geographic diversity, to enhance model robustness and generalization.
- Support for Vision-Language Models (VLMs): We will introduce additional captioning, metadata, and annotations tailored for VLM development. This multimodal focus will enable automated documentation, semantic scene understanding, and other tasks that require synergy between textual and visual data.

These points will expand the CIVAD dataset as a valuable tool to help researchers overcome barriers that hinder the integration of modern technologies, such as computer vision, into traditional industries, including the construction industry. This will enable researchers to build more reliable computer vision models in the construction industry.

Disclaimer: Specific portions of this dataset will be available upon request. Please contact Dr. Vicente Gonzalez at vagonzal@ualberta.ca to submit any requests.

References

[1] Alaloul W. S., Musarat M. A., Rabbani M. B. A., Altaf M., Alzubi K. M., and Al Salaheen M. Assessment of economic sustainability in the construction sector: Evidence from three developed countries (the USA, China, and the UK). Sustainability, 14(10):6326, 2022.

[2] Osobajo O., Oke A., Omotayo T., and Obi L. Systematic review of circular economy research in construction industry. Smart and Sustainable Built Environment, 11(1):39-64, 2020.

[3] Costin A., Hu H., and Medlock R. Building information modeling for bridges and structures: Outcomes and lessons learned from the steel bridge industry. Transportation Research Record Journal of the Transportation Research Board, 267:576-586, 2021.

[4] Pal A. and Hsieh S. H. Deep-learning-based visual data analytics for smart construction management. Automation in Construction, 131, 2021.
[5] Jiang Z. and Messner J. I. Computer vision applications in construction and asset management phases: Literature review. Journal of Information Technology in Construction, 28, 2023.
[6] Xuehui A., Li Z., Zuguang L., Chengzhi W., Pengfei L., and Zhiwei L. Dataset and benchmark for detecting moving objects in construction sites. Automation in Construction, 122, 2021.
[7] Duan R., Deng H., Tian M., Deng Y., and Lin J. SODA: Large-scale open-site object detection dataset for deep learning in construction. Automation in Construction, 142:104499, 2022.
[8] Xiao B. and Kang S. C. Development of an image dataset of construction machines for deep-learning object detection. Journal of Computing in Civil Engineering, 35(2):05020005, 2021.
[9] Yan X., Zhang H., Wu Y., Lin C., and Liu S. Construction instance segmentation (CIS) dataset for deep learning-based computer vision. Automation in Construction, 156:105083, 2023.
[10] Zhao J., Ou M., Xue L., Cui Y., Wu S., and Chen G. Joining datasets via data augmentation in the label space for neural networks. ArXiv, abs/2106.09260, 2021.
[11] Schuldt S., Nicholson M. R., Adams Y. A., and Delorit J. D. Weather-related construction delays in a changing climate: A systematic state-of-the-art review. Sustainability, 2021.
[12] Guo B. H., Zou Y., Fang Y., Goh Y., and Zou P. X. Computer vision technologies for safety science and management in construction: A critical review and future research directions. Safety Science, 135:105130, 2021.
[13] Salimi M., Loni M., Afshar S., Sirjani M., and Cicchetti A. ConstScene: Dataset and model for advancing robust semantic segmentation in construction environments. ArXiv, abs/2312.16516, 2023.
[14] Lin T. Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., and Zitnick C. L. Microsoft Coco: Common objects in the context. In Proceedings of the 13th European Conference on Computer Vision (ECCV), pages 740-755, Zurich, Switzerland, 2014.
[15] Xiao B., Wu H., Xu W., Dai X., Hu H., Lu Y., and Yuan L. Florence-2: Advancing a unified representation for various vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
[16] Al Rahhal M. M., Bazi Y., Elgibreen H., and Zuair M. Vision-language models for zero-shot classification of remote-sensing images. Applied Sciences, 13(22):12462, 2023.
[17] Lee K., Jeon C., and Shin D. H. Small-tool image database and object detection approach for indoor construction site safety. KSCE Journal of Civil Engineering, 27(3):930-939, 2023.
[18] Roboflow. Computer vision tools for developers and enterprises. On-line: https://roboflow.com/, Accessed: 19/12/2024.
[19] Ren T., Liu S., Zeng A., Lin J., Li K., Cao H., and Zhang L. Grounded Sam: Assembling open-world models for diverse visual tasks. ArXiv, abs/2401.10968, 2024.
[20] Liu S., Zeng Z., Ren T., Li F., Zhang H., Yang J., and Zhang L. Grounding Dino: Marrying a Dino with grounded pre-training for open-set object detection. In Proceedings of the European Conference on Computer Vision (ECCV), pages 38-55, 2024.
[21] Jiang P., Ergu D., Liu F., Cai Y., and Ma B. Review of the YOLO algorithm developments. Procedia Computer Science, 199, 2022.
[22] NVIDIA. NVIDIA GeForce RTX 4090 graphics card. On-line: https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4090/, Accessed: 19/12/2024.
[23] Park S. and Kim S. Semantic segmentation of heavy construction equipment based on point-cloud data. Buildings, 14(8):2393, 2024.
[24] Ultralytics. YOLO11, NEW. Ultralytics YOLO Docs. On-line: https://docs.ultralytics.com/models/yolo11/#overview, Accessed: 02/01/2025.
[25] Fritz J. Image rotation and Exif Data. On-line: https://www.jonathanfritz.ca/2016/07/25/image-rotation-and-exif-data/, Accessed: 01/01/2025.
[26] Pixabay open-source archive. On-line: https://pixabay.com/, Accessed: 05/01/2025.
[27] Sun P. et al. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
[28] Michaelis C., Bethge M., and Ecker A. S. A broad dataset is all you need for one-shot object detection. ArXiv, abs/2011.04267, 2020.
[29] Feng Y., Zhang J., Li G., Togo R., Maeda K., Ogawa T., and Haseyama M. A novel frame-selection metric for video inpainting was used to enhance the urban feature extraction. Sensors, 2024.
[30] Bonetto E., Xu C., and Ahmad A. GRADE: Generating realistic animated dynamic environments for robotics research. ArXiv, abs/2310.17041, 2023.
[31] Padilla R., Netto S. L., and da Silva E. A. Survey of performance metrics for object-detection algorithms. In Proceedings of the International Conference on Systems, Signals, and Image Processing (IWSSIP), pages 237-242, 2020.