VL-Con: Vision-Language Dataset for Deep Learning-based Construction Monitoring Applications

Shun-Hsiang Hsu[1], Junryu Fu[2] and Mani Golparvar-Fard[3]

[1]PhD student of Civil and Environmental Engineering, University of Illinois Urbana-Champaign, USA
[2]MSc of Civil and Environmental Engineering, University of Illinois Urbana-Champaign, USA
[3]Professor of Civil Eng., Computer Science, and Technology Entrepreneurship, University of Illinois Urbana-Champaign, USA

hsus2@illinois.edu, junryuf2@illinois.edu, mgolpar@illinois.edu

Abstract -

Recently, vision-language research has gained significant interest by successfully connecting visual concepts to natural language, advancing computer vision-based construction monitoring using a wide variety of text queries. While vision language models demonstrate high capability, performance degradation can be expected when adapting the model to the ever-changing construction scenarios. In contrast to the source image-text pairs, it is more challenging to cover the multitude of potentially involved objects and their naming conventions for construction activities. To bridge the domain gap, this study aims to collect construction-specific image-text pairs of building elements and related site work based on the ASTM Uniformat II. The image-text pairs of 641 activities in Uniformat are retrieved from the LAION-5B dataset based on the image and text embeddings using CLIP with two different prompts. Then, the collected images are labeled at the image level to conclude the requirements of the vision-language datasets for further development. Based on the results, a vision-language dataset, VL-Con, consisting of image-text pairs for construction monitoring applications is proposed with the aid of a construction semantic predictor and prompt engineering. The proposed VL-Con dataset can be accessed through `https://github.com/huhuman/VL-Con`.

Keywords -

Vision-Language Dataset; Construction Monitoring; Foundation Model

1 Introduction

Vision tasks such as scene understanding and object recognition have been widely studied for automation in construction monitoring [1]. A significant amount of reality capture data is collected to generate actionable insights for construction monitoring [2]. Supervised learning models, such as Mask R-CNN [3] and SegFormer [4], have been predominantly adopted in the past decade [5], necessitating large-scale annotated image datasets with hard-coded indexes of the target objects. The results can be utilized to compare with BIM or 4D BIM for progress monitoring. However, for recognition at a finer level (e.g., image segmentation), labeling images is extremely expensive, and such large-scale datasets still remain unavailable in the industry [6]. Without details in project schedules in today's Virtual Design and Construction (VDC) and Project Control practices, the lack of formal definitions around what needs to be segmented in photos causes the gap between different construction monitoring applications.

Applications based on supervised learning methods are limited to pre-defined classes and require extensive post-processing to derive actionable insights. Integrating knowledge from different models or enhancing their capability to broaden the recognition scope proves to be a challenging task. While the AEC/O research community has been focusing on these application-driven challenges, the AI community has worked on developing foundation models, such as MAE [7] and GPT-4 [8], respectively for vision and language understanding. In this research line, emphasis has been placed on self-supervised techniques, allowing models to learn from large-scale data without any annotation. Well-trained foundational models can be seamlessly adapted to generate fine-grained segmentation masks for various objects [9] and to create a chat agent capable of handling diverse requests [10]. The robustness and generalization of learned knowledge enable the applications for a wide range of downstream tasks.

Since large language models (LLMs) have achieved remarkable success in language understanding through self-supervised learning from the billion-scale training corpus, the robustness of text embeddings has turbocharged the development of open-ended vision systems by incorporating arbitrary text queries. Vision-language models have gained increasing interest in enhancing vision understanding by aligning image features with text embeddings in the latent space. Radford et al. [11] proposed the vision-language model, CLIP, using linear projection to map the outputs of the image and text encoders through contrastive pre-training. CLIP under natural language supervision has

shown promising zero-shot transfer performance on various public image datasets. Peng et al. [12] used CLIP features to segment 3D points according to their similarities to text queries, enabling open-vocabulary scene understanding. Tsai et al. [13] fine-tuned the CLIP model to translate construction safety-related semantics in images to textual information for site inspectors. These studies have demonstrated a more applicable user interface that facilitates the direct extraction of actionable insights, potentially advancing computer vision-based applications for construction monitoring.

Despite the recent breakthrough of the vision-language models, their performance on the construction scenes remains untapped, leaving room for potential improvements in the current vision system. Considering that most vision-language models are trained with a wide variety of image-text pairs instead of specific domain knowledge, the models may not contain enough construction knowledge to be adapted to downstream tasks for construction monitoring. To address the issue, this paper leverages the CLIP model to collect images according to the text embeddings of the building construction activities in the Uniformat standard. The Uniformat standard encompasses eight main activities related to building construction. The `clip-retrieval` [14] API is employed to retrieve the top ten search results from the LAION-5B dataset [15]. To provide the model with additional textual clues regarding construction semantics, a second round of image collection is conducted using the prompt template "`A photo of {activity} in construction`" to observe any noticeable improvements. Consequently, about 6200 images of 641 categories are collected in each round. Each retrieved image is reviewed whether it is correct or is within the construction context. Zero-shot performance of CLIP on the construction image classification dataset [16] is presented.

The results show that even using the prompt template instead of contextless activity names, the CLIP model is still not good enough at retrieving correct images for the target activity as well as zero-shot classification. Improving the prompts in this way only ensures the retrieval of construction-related content and not for the correct category. To further improve the dataset, strategies aiming for enhancing text prompts and visual construction semantics to obtain more accurate images are proposed. Similar to [11], where different context prompts are ensembled to enhance the zero-shot performance, three different prompt templates are used to increase the opportunities of retrieving the correct images of the work activities. Additionally, a construction semantic classifier is developed to remove non-construction images by estimating the score of how an image is construction-related. Both of the two strategies expedite the image collection of the proposed VL-Con dataset and help ensure the quality of the collected images.

Based on the enhanced image retrieval through CLIP API, the images of different activities are collected and reviewed, and another round of manually image collection on web is conducted to finalize our dataset. Only 142 of the 641 activities have additional 5 images from the manual collection because some activities are considered ambiguous (e.g., specialty and supplementary activities) or cover too broad range of definitions (e.g., high-level activities). Note that the reason could also be why the previous collections fail to find proper images to some activities. By comparing the collected images of different activities at different round through large vision-language model and our manual collection, the gap in image features is significant and highlights the need of more construction-specific image-text pairs to improve in-domain knowledge. To adapt the large-scale vision-language models to construction monitoring applications, such datasets containing construction knowledge would be required. As a preliminary and pioneer work, the proposed VL-Con dataset is publicly available through `https://github.com/huhuman/VL-Con`.

2 Related work

2.1 Computer vision in construction monitoring

In construction progress monitoring, reality capture data were collected to be compared against BIM or 4D BIM for tracking element changes and confirming project schedules. Pal et al. [17] estimated the completion percentage of building construction progress in combination of site images, reconstructed point clouds, and BIMs. These efforts have been constrained by low LOD across various model disciplines of BIM and a lack of details in project schedules. Jung et al. [18] proposed a transformer model to ensure the consistency among project schedule data by aligning them with Uniformat classifications. Núñez-Morales et al. [19] generated synthetic images from high-LOD BIMs to help develop supervised learning models recognize different under-construction elements as the well-annotated datasets satisfying a certain quantity and quality to train the models are not available. Despite transfer learning from another pre-trained supervised model with large-scale datasets, Lin et al. [20] presented that the dataset bias may prevent the model from learning construction-specific contents, and the transfer-learning Faster R-CNN even had poorer performance than the one trained from scratch on bridge defects.

Inspired by LLMs, vision foundation models using self-supervised learning techniques have shown promising results and become popular alternatives. He et al. [7] proposed masked autoencoders (MAE) following the idea of masked language modeling to pre-train the large vision model with a wide variety of image data without labels. Kirillov et al. [9] proposed the large-scale dataset of 1B

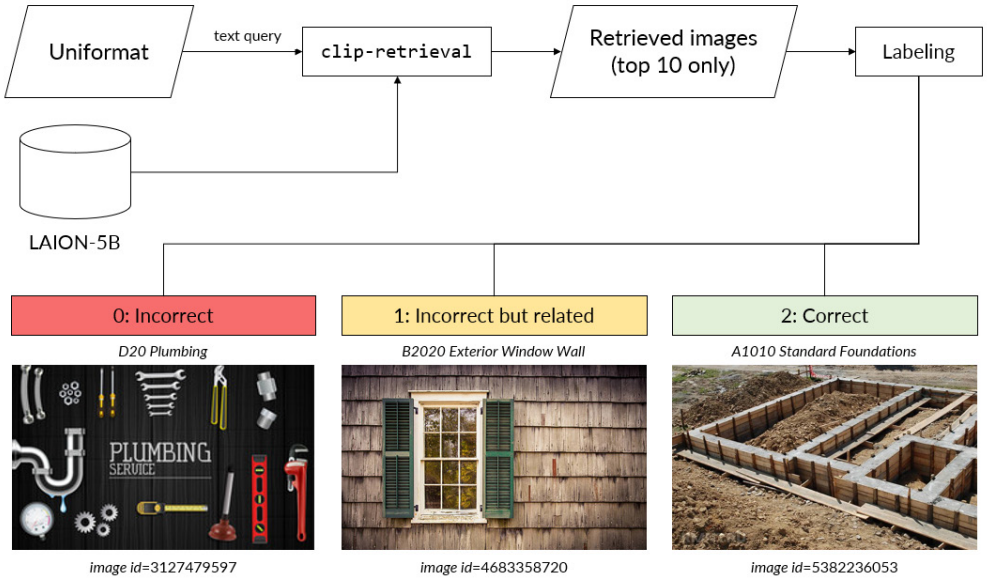

Figure 1. Overview of the image collection and labeling

Table 1. The label distributions of the two collected image sets

Label	A	B	C	D	E	F	G	Total
				"{activity}"				
Correct	18.6%	48.8%	62.6%	28.4%	21.7%	33.0%	43.2%	37.7%
Incorrect but related	30.4%	18.8%	23.1%	10.2%	15.1%	11.7%	14.3%	16.5%
Incorrect	51.0%	32.4%	14.3%	61.4%	63.2%	55.3%	42.5%	45.8%
Total	494	738	854	1677	503	528	1352	6146
				"{activity} in construction"				
Correct	26.0%	50.1%	53.4%	26.1%	21.5%	25.5%	44.1%	36.4%
Incorrect but related	54.7%	34.3%	40.1%	29.0%	24.2%	34.5%	26.0%	32.6%
Incorrect	19.3%	15.7%	6.5%	44.9%	54.2%	40.0%	29.9%	31.1%
Total	494	738	854	1677	503	528	1352	6212

masks and 11M images to develop the segment anything model (SAM) using MAE pre-trained vision transformer as the image encoder for class-agnostic segmentation. Taking advantage of the robustness of such vision foundation models, Ahmadi et al. [21] combined SAM with U-Net to enhance crack detection in concrete. Ge et al. [22] fine-tuned SAM for crack segmentation to improve cross-dataset generalization.

However, a number of challenges still hinder the computer vision-based applications for construction monitoring, including (1) lack of available ground truth segmentation for relevant physical assets in reality capture datasets and (2) lack of formal definitions around what needs to be segmented in pictures in the first place. Vision-language models that take the advantage of robust language understanding have broadened vision understanding to alleviate the limitations. The extracted image features reflecting the context relationship can be more easily adapted to various construction scenes.

2.2 Vision-language in construction monitoring

Before vision-language foundation models emerge, research has been focused on interpreting construction images in the form of natural language. For example, image captioning that can directly generate actionable insights for construction monitoring has been studied in the past decade. Without the robust text embeddings from foundation models, creating new and meaningful textual labels or captions of various construction scenes is the core hindrance. Xiao et al. [23] proposed the image captioning dataset for common construction machines and their activities. Liu et al. [24] proposed the image captioning dataset of five construction activities with the details of worker actions and safety gears. Zhai et al. [25] created the image captioning dataset for perceiving unsafe behav-

ior of workers in construction.

The reviewed image captioning methods mostly adopted the encoder-decoder architecture to perform image-to-text translation, where CNN models were used as the image encoder, and RNN models were used as the text decoder. Bang and Kim [26] extracted features of object regions from drone images through Faster R-CNN as the image encoder and decode the features using LSTM to produce dense captioning. Wang et al. [27] used Mask R-CNN as the image encoder and LSTM with the attention layer as the text decoder for construction works, including masonry, reinforcement steel bar tying, and tiling. The adopted single-modal models were only trained with their proposed datasets to connect the representations across vision and language. As the studies focused on specific scopes and scenarios, the learned knowledge of their fully supervised models was limited to the adopted datasets.

The limitations of the encoder-decoder architecture made the applications difficult to be scaled. In contrast, Radford et al. [11] proposed the dual-encoder model, CLIP, to first jointly train text and vision encoders with numerous image-text pairs of a wide range of cases. By bridging multi-modal understanding through natural language supervision, the vision-language model was capable of handling various scenarios with more robust image embeddings. A text decoder can be specifically trained for a downstream task that needs text generation [28]. As being a promising alternative, the feasibility analysis of the vision-language foundation model for construction monitoring applications is needed to explore and validate model's understanding of construction contexts.

3 Vision-Language Understanding of Construction Context

3.1 Data collection and labeling

This paper retrieves the corresponding images using the Uniformat work item as the text query from the LAION dataset through their `clip-retrieval` [14] API (see Figure 1). The default parameter values are adopted to search and rank the images, including aesthetic scoring. Additionally, the keyword, "*in construction*", is prepended to the original names as text queries to collect the images in a second run. This is expected to provide more semantic clues of construction and help improve the performance because some of the names are not exclusive in the construction industry, and the model does not specifically learn to recognize them.

After that, this paper manually review every image and classify them into three groups: (1) **correct** - the image represents the corresponding activity, (2) **incorrect but related** - the image does not indicate the corresponding activity but contains construction semantics, (3) **incorrect** - the image is not related to any construction activities. Figure 1 illustrates the examples of the defined three groups.

Table 1 presents the summary of the dataset over eight different main activities. As presented, only about one-third of the images are correctly retrieved for the given construction activities. Despite the increase in the ratio of the class of incorrect but related as shown in Figure 2 when using the prompt "`{activity} in construction`", the overall accuracy is not significantly improved. As a result, construction-specific image-text pairs are needed to enhance the construction knowledge for developing a more robust foundation model in the construction domain.

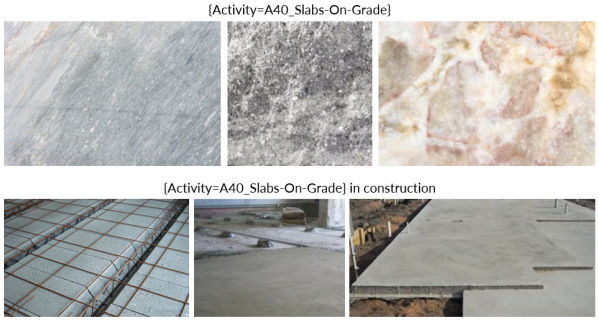

Figure 2. Image retrieval results with and without *"in construction"*

3.2 Zero-shot performance on construction images

The zero-shot performance of the CLIP model on construction images is evaluated by the BCS dataset [16], which contains about 212,000 photos of buildings and construction sites for classification. To evaluate the understanding of construction contexts, 104,484 images of 54 categories for construction sites in the BCS dataset are employed to perform the zero-shot classification. The image numbers of different classes are *Bridge* (6752), *Site fence* (4980), *Wood floor* (4808), *Ordinary Door* (4568), etc. As stated in [11], the same prompt for zero-shot transfer to existing image classification datasets, "`A photo of a {label}.`", is used to wrap the inputs instead of using contextless class names.

Cheng et al. [16] has achieved the top-1 accuracy of up to 94.7% on the dataset by a fully supervised model while zero-shot CLIP underperforms by over 35%. The CLIP model only achieves the top-1 accuracy of 59.39% and top-3 accuracy of 81.41%. Figure 3 presents the zero-shot accuracy distribution over different categories. Among all the categories, the model has the highest accuracy of 99.40% for *Site vest* and the lowest accuracy of 0% for *U-steel*. The *U-steel* images are mostly misclassified into other steel-related classes, such as *Sheet steel* and *Angle steel*. The situation infers that the lack of construction-

specific knowledge limits the model to only recognize general contents. The professional terms with only minor differences significantly confuse the model. The evidence can be found as the top-3 accuracy is increased by over 20%.

4 VL-Con: Vision-Language Dataset for Construction Monitoring

4.1 Requirements of vision-language datasets

Based on the CLIP's understanding of construction context, construction-specific vision-language datasets are required to enhance the construction knowledge. Through manually inspecting the collected images, Figure 4 presents that potential causes of the poor performance on image retrieval. The requirements of the vision-language dataset preparation reflecting the issues for construction monitoring are summarized as follows:

1. Ambiguous description/name of work activity

   Although construction knowledge is required to understand the semantics of the activities, some of their names are too ambiguous to be easily interpreted even by people with construction backgrounds. The names become more abstract when the corresponding activities are at a higher level because they need to cover various children's items. To keep the simplicity, only a few words are used to define the whole scope, leading to failures in capturing the semantics behind the words without any additional context. For example, in the adopted Uniformat system, *Floor Construction* (B1010), *Roof Construction* (B1020), and *Stairs* (B1080) are all sub-items of *Superstructure* (B10). Figure 4(a) demonstrates that the CLIP model fails to retrieve correct images of the superstructure but is capable of recognizing other common words like floor and roof. Consequently, prompt engineering is needed to include more context for CLIP to retrieve more correct images.

2. Missing photos of work activity

   Though the examples of image retrieval show correct semantics, more construction-related and progress-detailed images are expected. Most of the activities suffer from missing photos in the training data because the construction activity or corresponding assemblies contain numerous components and steps. The issue limits the model to return only the photos of finished states or irrelevant content. Figure 4(b) shows that this is especially true for any categories associated with "supplementary" or "specialties". For example, *Exterior Wall Specialties* (B2090) include below-grade egress, window wells, and any kind of finished product tangent to the exterior wall [29]. The exhibited semantics overlapped with other categories such as *Exterior Fabrications* (B2080.70) leading to failures in differentiating between each other, whereas exterior fabrications are more about column covers or decorative finishing directly applied onto the wall. This phenomenon can also be attributed to the ambiguous nature of the activity names. Without specific images that demonstrate the difference between such similar sub-categories, even engineers could be confused with the definitions.

3. Searching preference for construction needs

   In the adopted CLIP image retrieval, the image quality can be determined by the aesthetic predictor, ensuring the retrieved images are closer to what users are expecting. A photo of a document copy or procedure diagram may get a low score and be ranked behind because such an image contains less vision information, meaning little object information is included. For example, Figure 4(c) shows the image retrieval results of heat generation under different aesthetic score thresholds and average weights. As searching preference could significantly impact image retrieval, the construction-specific preference can be developed to help retrieve more correct images of our interest to build the vision-language dataset.

4.2 Prompt engineering and construction semantic classification

To improve image retrieval of the Uniformat activities, three different prompts with more details are used: (1) `"A photo of {activity}, a type of building construction activity"`, (2) `"A photo of {full-activities-hierarchy}"`, and (3) `"A photo of {activity}, revit"`. The first and second one is to replace the previous *"in construction"* keyword with more specific definitions of the activities. The final one is taking advantage of the exclusive word in the construction industry to force the retrieved images to be related to construction while the images are mostly about virtual scenes. The enhanced prompts can help collect more potentially correct image-text pairs as presented in our final proposed VL-Con dataset.

For the semantic predictor, the original image retrieval already employed the aesthetic predictor to ensure the images contain more useful information and significant object appearance instead of diagrams and flow charts. This paper trained a ResNet-18 model for binary construction semantic classification with the previously labeled dataset of Uniformat categories, followed by the prompt `"{activity} in construction"`. After removing duplicate images, nearly 5000 images were then separated into an 8 to 2 ratio for training and validation sets. The

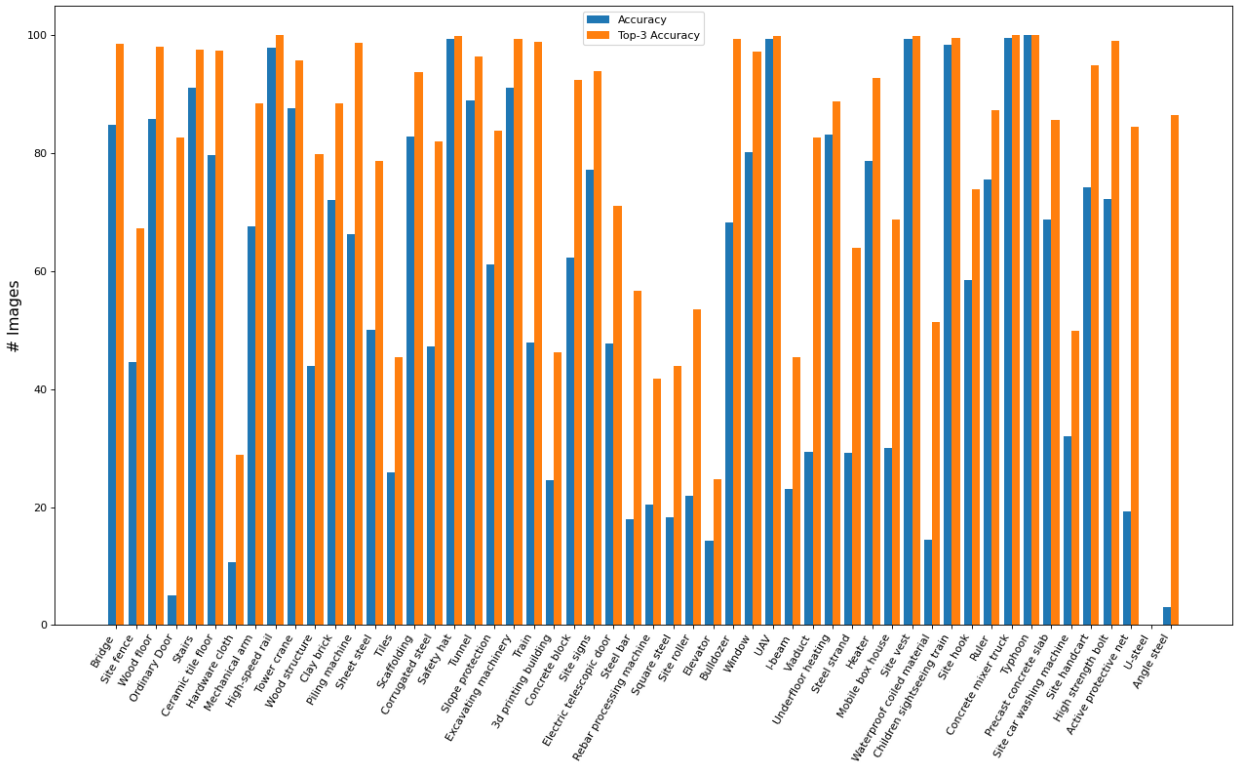

Figure 3. The zero-shot top-1 and top-3 accuracy of different classes

Figure 4. The examples for the requirements of vision-language datasets

images were all resized to 224 by 224, and the training set was augmented with random crop and random horizontal flip. The model after 25 epochs achieves the accuracy of 73.63% on the validation set with a wide range of images associated with various categories in the Uniformat. Figure 5 demonstrates the model predictions and scores of the construction semantics of the images. One noticeable feature of this very simple trained model is to identify diagram-like images or ”unrealistic images” and filter more realistic scenes such as those of construction sites.

With the two proposed strategies, another three rounds of image collection are conducted, and the retrieved images are firstly filtered by the proposed construction semantic predictor. After that, manual inspection is still required to finally complete the vision-language dataset for enhancing construction knowledge of foundation models. Note that because some of the activities may not exist in the LAION-5B dataset, the image retrieval possibly fail to find any correct images of those activities. In that case, another image sources will be needed in the future to acquire the corresponding photos of the activi-

ties. To ensure certain quantity and quality of the dataset, the dataset additionally include 5 images for each of 142 activities manually collected through the web. The final VL-Con dataset are publicly available and scalable, allowing other researchers to add more image-text pairs to any of Uniformat categories.

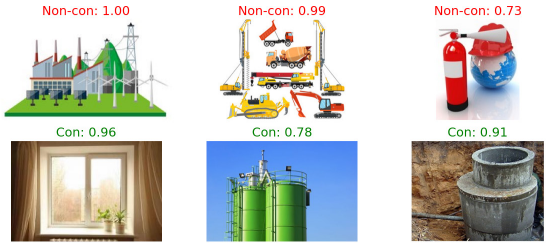

Figure 5. Construction semantic classification

5 Conclusions

This paper conducted preliminary analysis of vision-language understanding of CLIP to construction scenes with regards to Uniformat. The images from the LAION-5B dataset were retrieved using the CLIP model to assess its applicability of understanding construction scenes. Upon detailed review of the retrieved images, the limitations of the existing vision language model were identified: appropriate prompt to maximize the likely result, insufficient images that precisely describe all the activities in Uniformat, and inability to contextualize construction scenes from images. To address these limitations, two strategies, prompt engineering and a semantic classifier of construction scenes, were proposed to complete the vision-language dataset for construction monitoring. Another manual image collection is also conducted to further enhance the dataset. The final VL-Con dataset was released to provide the basis for further method development and model training benchmark. As a pioneer work for adapting vision-language models to construction monitoring applications, the dataset can be scaled by adding more images for any of the activities, boosting the construction-specific foundation models.

References

[1] Shuai Tang, Dominic Roberts, and Mani Golparvar-Fard. Human-object interaction recognition for automatic construction site safety inspection. *Automation in Construction*, 120:103356, 2020.

[2] Youngjib Ham, Kevin K Han, Jacob J Lin, and Mani Golparvar-Fard. Visual monitoring of civil infrastructure systems via camera-equipped unmanned aerial vehicles (uavs): a review of related works. *Visualization in Engineering*, 4(1):1–8, 2016.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[4] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[5] Shun-Hsiang Hsu, Ting-Wei Chang, and Chia-Ming Chang. Impacts of label quality on performance of steel fatigue crack recognition using deep learning-based image segmentation. *Smart Structures and Systems*, 29(1):207, 2022.

[6] Yeji Hong, Somin Park, Hongjo Kim, and Hyoungkwan Kim. Synthetic data generation using building information models. *Automation in Construction*, 130:103871, 2021.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[8] OpenAI et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2021.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[10] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023.

[13] Wei Lun Tsai, Jacob J Lin, and Shang-Hsien Hsieh. Generating construction safety observations via clip-based image-language embedding. In *European Conference on Computer Vision*, pages 366–381. Springer, 2022. doi:10.1007/978-3-031-25082-8_24.

[14] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. `https://github.com/rom1504/clip-retrieval`, 2022.

[15] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[16] Xuanhao Cheng, Mingming Jia, and Jian He. A large-scale dataset of buildings and construction sites. *Computer-Aided Civil and Infrastructure Engineering*, 2023.

[17] Aritra Pal, Jacob J Lin, Shang-Hsien Hsieh, and Mani Golparvar-Fard. Activity-level construction progress monitoring through semantic segmentation of 3d-informed orthographic images. *Automation in Construction*, 157:105157, 2024.

[18] Yoonhwa Jung, Julia Hockenmaier, and Mani Golparvar-Fard. Transformer language model for mapping construction schedule activities to uniformat categories. *Automation in Construction*, 157: 105183, 2024.

[19] Juan D Núñez-Morales, SHUN-HSIANG HSU, Amir Ibrahim, and Mani Golparvar-Fard. Reality-enhanced synthetic image training dataset for computer vision construction monitoring. *Proceedings of International Structural Engineering and Construction*, 10(1):CON–29, 2023.

[20] Jacob J Lin, Amir Ibrahim, Shubham Sarwade, and Mani Golparvar-Fard. Bridge inspection with aerial robots: Automating the entire pipeline of visual data capture, 3d mapping, defect detection, analysis, and reporting. *Journal of Computing in Civil Engineering*, 35(2):04020064, 2021.

[21] Mohsen Ahmadi, Ahmad Gholizadeh Lonbar, Abbas Sharifi, Ali Tarlani Beris, Mohammadsadegh Nouri, and Amir Sharifzadeh Javidi. Application of segment anything model for civil infrastructure defect assessment. *arXiv preprint arXiv:2304.12600*, 2023.

[22] Kang Ge, Chen Wang, and Yutao Guo. Fine-tune vision foundation model for crack segmentation in civil infrastructures. *arXiv preprint arXiv:2312.04233*, 2023.

[23] Bo Xiao, Yiheng Wang, and Shih-Chung Kang. Deep learning image captioning in construction management: A feasibility study. *Journal of Construction Engineering and Management*, 148, 7 2022. ISSN 0733-9364. doi:10.1061/(asce)co.1943-7862.0002297.

[24] Huan Liu, Guangbin Wang, Ting Huang, Ping He, Martin Skitmore, and Xiaochun Luo. Manifesting construction activity scenes via image captioning. *Automation in Construction*, 119:103334, 2020. ISSN 0926-5805. doi:https://doi.org/10.1016/j.autcon.2020.103334.

[25] Peichen Zhai, Junjie Wang, and Lite Zhang. Extracting worker unsafe behaviors from construction images using image captioning with deep learning–based attention mechanism. *Journal of Construction Engineering and Management*, 149(2):04022164, 2023. doi:https://doi.org/10.1061/JCEMD4.COENG-12096.

[26] Seongdeok Bang and Hyoungkwan Kim. Context-based information generation for managing uav-acquired data using image captioning. *Automation in Construction*, 112, 4 2020. ISSN 09265805. doi:10.1016/j.autcon.2020.103116.

[27] Yiheng Wang, Bo Xiao, Ahmed Bouferguene, Mohamed Al-Hussein, and Heng Li. Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Advanced Engineering Informatics*, 53:101699, 2022. doi:10.1016/j.aei.2022.101699.

[28] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.

[29] Jim Bedrick, Will Ikerd, and Jan Reinhardt. Level of development (lod) specification. `https://bimforum.org/resource/lod-level-of-development-lod-specification/`, 2022.