Zero-shot Learning-based Polygon Mask Generation for Construction Objects

Taegeon Kim[1], Minkyu Koo[1], Jeongho Hyeon[1], and Hongjo Kim[1]

[1]Department of Civil & Environmental, Yonsei University, South Korea
E-mail: ktg9655@yonsei.ac.kr, kmk0119804@yonsei.ac.kr, hyeon9404@yonsei.ac.kr, hongjo@yonsei.ac.kr (Corresponding author)

Abstract

For construction site monitoring, the use of segmentation-based computer vision technology has been proposed. In such environments, the main technical challenge is the generation of data for training the segmentation model. The training data for a segmentation model involves polygon annotation of objects within an image, which is a time-consuming task. To address this issue, this study proposes a new approach that uses the YOLOv8 object detection model to predict bounding box labels and inputs these into a Segment Anything Model (SAM) to automatically generate polygon label data. The performance of the YOLOv8 model exceeded 80%, and the automatic generation of polygon labels through SAM resulted in an IoU range of 55-86%, producing high-quality mask label data. This approach significantly reduces the time, labor, and cost associated with the labeling process.

Keywords – Polygon label generation, Instance segmentation, Zero-shot learning

1 Introduction

The construction industry is recognized globally as one of the most hazardous sectors, with a high incidence of injuries and fatalities [1]. Global statistics indicate that the fatality and injury rates in the construction industry are three and two times higher, respectively, than the average for other industries [2].

Faced with this high rate of accidents, the construction industry is progressively adopting advanced digital technologies such as Digital Twins (DT), Building Information Modeling (BIM), Artificial Intelligence (AI), the Internet of Things (IoT), and Smart Vision (SV) to improve efficiency, productivity, accuracy, and safety[3]. The introduction of these technologies represents a continuous effort to transition from traditional industrial practices and manufacturing methods to autonomous smart systems [4], and the construction industry has innovated its work processes through the digitalization of project management processes, gradually improving competitiveness [5].

Traditional methods of construction site monitoring often involve manual inspections and assessments, which are not only time-consuming but also prone to human error[6]. The application of computer vision technologies for site monitoring has emerged as a pivotal tool [7,8], significantly enhancing safety and progress management. Computer Vision-based systems offer a more efficient, accurate, and real-time alternative to traditional monitoring methods[9]. However, the efficacy of such systems is heavily reliant on the quality of the training data used to develop them, particularly in the context of object segmentation models [10].

The process of generating polygon annotations for the training of segmentation models has historically been a labor-intensive and time-consuming task [6]. These annotations are crucial for teaching models to accurately identify and segment various objects in a construction setting, such as equipment and personnel. The challenge is further compounded when adapting deep learning models to new domains, a process that traditionally requires extensive manual data annotation[6,11].

Recent advancements in deep learning have seen the exploration of few-shot learning and domain adaptation methods [12,13]. These techniques aim to reduce the reliance on large volumes of manually annotated training data when adapting models to new domains. However, the performance of segmentation models trained using these methods remains suboptimal. There is a notable lack of research focused on the preparation of polygon annotations for construction objects, despite the potential benefits they offer, such as monitoring personal protective equipment (PPE) compliance.

Addressing this gap, this research proposes an innovative method for the automatic generation of polygon masks, which serve as training data for segmentation models. This method leverages the capabilities of an instance segmentation model, utilizing detection results (bounding boxes) as inputs to fully automate the polygon mask generation process. Specifically, the study employs the "You Only Look Once version 8" (YOLOv8) [14] for predicting bounding boxes of construction objects. These bounding boxes are then used as prompts for the Segment Anything Model (SAM)[15], which generates the polygon masks.

To validate the effectiveness of the proposed method, experiments were conducted using the Moving Objects in Construction Sites (MOCS) dataset [16]. This dataset encompasses a comprehensive range of construction objects including workers, various types of vehicles, and equipment. The experimental design involved training the model with 19,404 images and testing it with a separate set of 4,000 images. The results indicate that the polygon masks generated by the method are comparable to those produced through manual human annotation, with an Intersection over Union (IoU) deviation ranging from 10.8% to 34.6%.

This research makes significant contributions to the field of computer vision in construction engineering. Firstly, it presents a method to automatically generate polygon annotations without the need for human involvement, thus streamlining the training process for segmentation models. Secondly, it demonstrates the quality and viability of automatically generated segmentation masks derived from detection results.

2 Related work

In the field of computer vision, the development of deep learning models heavily relies on the quality and quantity of training data. Particularly, segmentation models require pixel-level labeling data, a process that consumes approximately ten times more resources in terms of labor and cost compared to bounding box labeling for object detection models [17]. In this experience, the preparation of polygon masks takes ten to twenty times more efforts than the preparation of bounding box labels. This challenge becomes more complex in dynamic environments such as construction sites, where the constant movement of equipment and machinery necessitates diverse data collection. To overcome these challenges, the construction domain has conducted active research in various ways such as synthetic data generation, zero-shot or few-shot learning, and domain adaptation.

2.1 Synthetic data generation

Synthetic data has emerged as a key solution to alleviate the time and labor burdens associated with preparing data for model training. There has been research utilizing synthetic data for visual data analysis in infrastructure management, automating the data collection process to address labor-intensive and time-consuming issues [18]. Additionally, studies have been conducted on acquiring scaffolding point cloud data through Mobile Laser Scanning (MLS) and utilizing it for training data in construction sites [19].

2.2 One-shot or few-shot learning

One of the key challenges in the advancement of segmentation model, especially supervised learning models, is effectively recognizing object with limited training data. Many models require a substantial amount of training data for each class, but obtaining sufficient data for certain classes can often be challenging. To address this issue, new learning paradigms such as few-shot and one-shot learning methods [20,21] have been proposed. These approaches utilize knowledge from instances of various classes to enable effective learning even with a small number of instances, aiming to overcome additional challenges such as classifying instances of classes that have not been encountered before [22].

2.3 Domain Adaptation

Domain adaptation addresses performance degradation due to differences in data distribution between source and target domains. Enhancing model performance by adding data from the target domain similar to the source domain has been explored. Particularly, self-training techniques, which involve using a trained model to predict labels on unlabeled data and utilizing it as training data, have been researched to improve the generalization capabilities of models [6].

2.4 Knowledge gap of previous studies

Despite the advancements in these technologies, the need for human annotation remains a crucial element in the training process of segmentation models, especially when the model is applied to a new target domain. Human annotation ensures the quality and accuracy of data, playing a vital role in reflecting the complexity and diversity of construction environments in the data. Based on this context, this study proposes a new methodology that can automatically generate training data, emulating the characteristics of human-annotated data.

The methodology developed in this research utilizes SAM to automatically generate high-quality training data comparable to human annotation. This presents an opportunity to effectively enhance the performance of deep learning models, particularly in complex and dynamic environments like construction sites. The automated data generation process can significantly replace the time-consuming and costly tasks performed by human annotators, providing rich training data in a faster and more cost-efficient manner.

3 Proposed Method

3.1 Overview of the proposed method

To automatically generate polygon masks of construction objects, the proposed method, as shown in Fig. 1, is divided into 2 steps, as follows:

1) Training the YOLOv8 object detection model using the MOCS training dataset to predict bounding boxes on the test dataset.

2) Polygon mask generation using the Segment Anything Model (SAM)

3.1.1 Object detection model training and bounding box prediction

You Only Look Once version 8 (YOLOv8) [14,23] is the latest model that enables fast and accurate object detection, similar to the human visual system. This model performs the process of classifying objects within an image and determining their location information through a single inference. Among the YOLO series, YOLOv8 has established itself as the preferred architecture in applications requiring fast inference speeds by providing the highest mAP performance and inference speed on the Microsoft Common Object in Context (MS COCO) dataset [24].

Despite these capabilities, performance degradation occurs due to visual differences between the training domain and the intended application domain (target domain), which is more pronounced in complex environments like construction sites. Therefore, to maximize the model's performance for a specific domain, it is essential to optimize or retrain the model's weights for the target domain data. In this study, the YOLOv8 model pretrained on the COCO dataset was retrained on the MOCS dataset. The re-trained YOLOv8 model was used to predict bounding boxes of target construction object classes (see Fig.2 listing the target object classes).

3.1.2 Automated polygon mask generation with predicted bounding box

This study utilizes the Segment Anything Model (SAM), an instance segmentation model, which serves as a foundational model in the field of computer vision, aiming for the universality like that of ChatGPT. This model was trained on the 1.1 billion SA-1B dataset and possesses the capability to perform segmentation of various objects through simple prompt input [25]. SAM can process various forms of prompts, including masks, bounding boxes, points, and text, enabling the automatic generation of polygon label data [15,17]. The proposed method leverages SAM to efficiently generate high-quality polygon label data within construction site environments. Specifically, the bounding box prediction information for construction site objects predicted by the YOLOv8 model is used as input prompts, generating accurate polygon label data. This methodology automates the generation of polygon mask data in complex and dynamic construction site environments by integrating prompt-based systems with the latest computer vision model, SAM, replacing labor-intensive data labeling tasks in the target domain.

4 Experiments

4.1 Experimental Settings

4.1.1 Computer & Datasets

The experiments were conducted on systems equipped with 4 NVIDIA GeForce RTX 4090 GPUs running on Ubuntu 20.04 operating system. The Moving Objects in Construction Sites (MOCS) dataset was utilized, comprising 19,404 training images and 4,000 validation images. All target classes (Worker, Static crane, Hanging head, Crane, Roller, Bulldozer, Excavator, Truck, Loader, Pump truck, Concrete mixer, Pile driving, Other vehicle) were used. This data was used to train a YOLOv8 object detection model, which was then utilized to predict bounding boxes on the MOCS validation dataset composed of 4,000 images. The MOCS dataset includes publicly available images for training, validation, and testing; however, bounding boxes and polygon annotations are only provided for the training and validation datasets, which limits the use of the test dataset. Therefore, the performance of the fully trained model was evaluated using the validation dataset. The predicted bounding boxes and original images were inputted into SAM to generate polygon mask label data, and the IoU of these extracted mask labels was calculated. To compare the accuracy of the generated mask label data, ground truth bounding boxes and original images were inputted into SAM to calculate the IoU, and finally, the differences between the two results were compared.

4.1.2 Model selection & hyperparameters

In this experiment, the largest 'x' model of the YOLOv8, pre-trained on the COCO dataset, was used. Although this model has a large number of parameters, resulting in longer training times, it was selected for its outstanding performance on the COCO dataset. Additionally, a threshold of 0.5 was set for the bounding boxes predicted by the detection model, and the extracted bounding boxes were used as prompt values for SAM.

4.1.3 Model training

The hyperparameters set for YOLOv8 training included an image size of 1280*720 HD, 300 epochs, a batch size of 16, and a learning rate of 0.01. The epochs were adjusted to 300 to correspond with the quality of the data. Additionally, data augmentation techniques such as Mixup and Copy-Paste were incorporated into the training process.

5 Experimental Results

5.1 Detections results of YOLOv8

As shown in Table 1, the mAP results, which are the primary performance metrics for the YOLOv8 detection model trained on the MOCS dataset. The target class that exhibited the highest performance was 'Excavator' with mAP of 94.2%, while the lowest performing target class was 'Crane' with mAP of 77.7%. Despite a considerable number of instances in the Worker class as shown Fig. 2, which could have led to concerns about model bias, the model still achieved high performance.

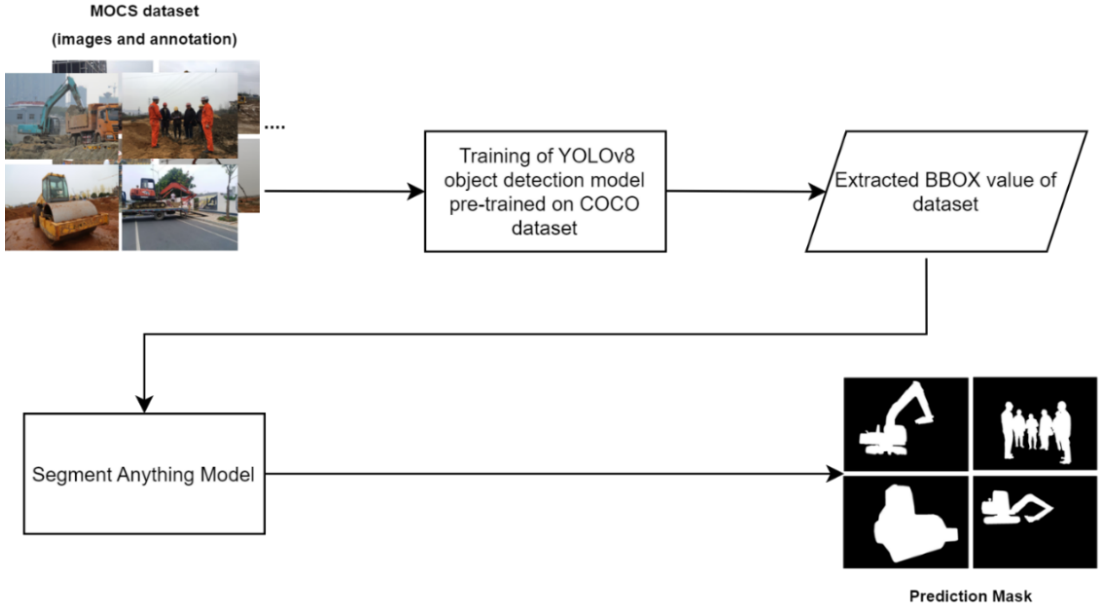


Figure 1. Overview of proposed method

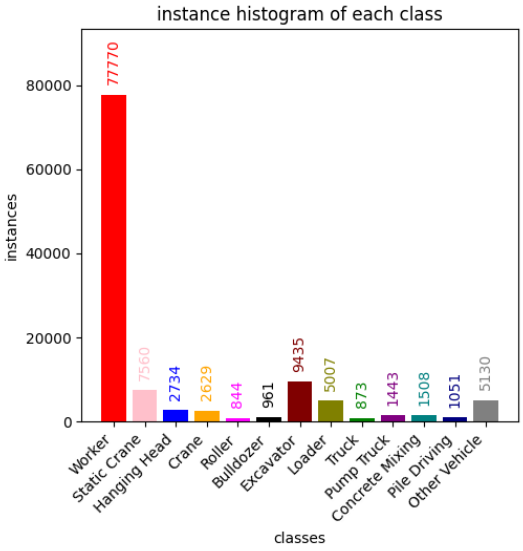


Figure 2. Number of instances each class in the MOCS dataset

5.2 SAM Result

Table 2 shows the results of the Segment Anything Model. The IoU values for masks generated by inputting predicted bounding boxes and original images into SAM were calculated and compared with those of masks created using actual ground truth bounding boxes. The most substantial discrepancy was observed in the 'Concrete Mixer' class with approximately 21%, and the smallest discrepancy was in the 'Bulldozer' class with about 3%. These figures provide crucial information on how accurately SAM classifies and labels objects across various classes.

These results indicate that SAM can generate polygon labels with relatively higher accuracy for certain classes of objects, while exhibiting lower accuracy for others. This suggests a critical interplay between the performance of the SAM algorithm and characteristics of objects such as their features, shapes, sizes, and colors. Understanding this interaction is vital in improving the algorithm of SAM and enhancing its detection and labeling performance for specific classes.

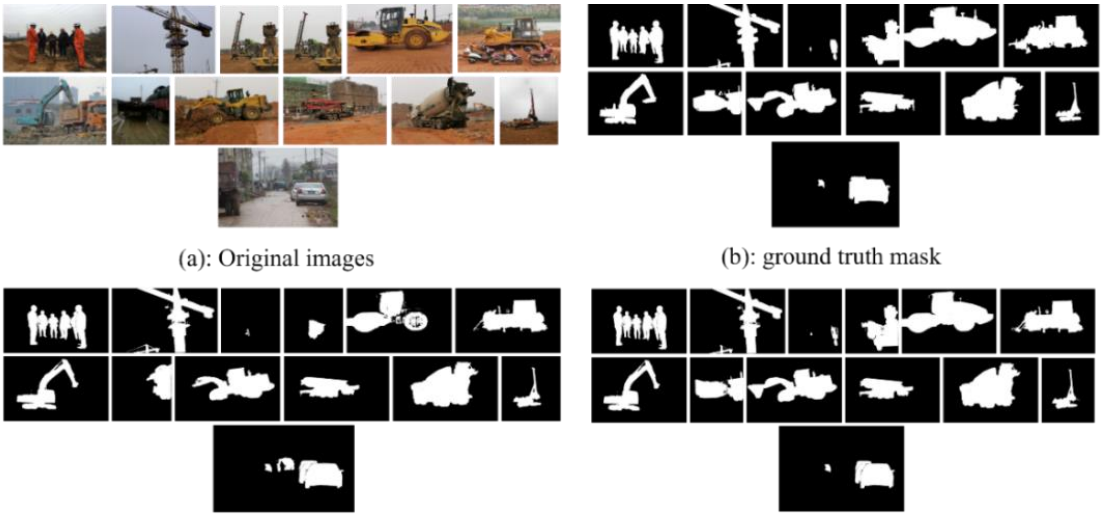


Figure 3. Original images and generated polygon labels

6 Discussion

Performance of YOLOv8 to generate bounding box prompts for SAM: In this study, the performance of the YOLOv8 object detection model was evaluated based on the mean Average Precision@50 (mAP@50) metric, showed an impressive result of 85.8% in the average mAP across all classes. Additionally, the mAP performance for each class was observed to range between 77% and 94%, further emphasizing the model's excellence (see Table 2). According to the study by Xuehui et al. [16], when other models such as YOLOv3 and Faster R-CNN were applied to the MOCS test dataset, they showed relatively low performances with mAP at 9.99% and 66.59%, respectively. However, the direct comparison between the previous methods [16] and our method is not possible as the test datasets were different (This study used the MOCS validation dataset because the ground truth of the MOCS test dataset is not publicly available).

Quality of generated polygon annotations: Furthermore, the Intersection over Union (IOU) of polygon mask labels generated by inputting predicted bounding boxes to SAM was measured between 55% to 86% among the target classes. In contrast, when actual ground truth bounding boxes were input into SAM, the IOU was evaluated to be between 65% and 89%. There was a performance difference of 5% to 22% between polygon label data generated by these two methods.

In the study by Chern et al. [26], a dataset having a single target class per image was used, and the model's Pseudo Label (P.L), Refined Pseudo Label (Refined P.L), and Feature Pyramid Networks (FPN) performance for target classes Background, Dump truck, Excavator, Mixer truck, Roller ranged between 48.48% and 78.76%. This demonstrated that the IOU performance of mask labels generated by SAM for all classes was higher.

These results illustrate the competitiveness of the proposed system to generate target domain annotations in a fully automated manner, reducing the labor and cost involved in bounding box labeling. These results offer the possibility of automatic training data generation in complex environments like construction sites.

As shown in Fig. 3, instances were identified where ground truth labeling for certain objects was inaccurately applied. In some cases, polygon annotations generated by SAM were more accurate than the original annotations. These factors are considered to have contributed to the observed decline in overall performance.

6.1 Limitation & Future study

Several key limitations were identified in the process of automatically generating polygon mask label data. Firstly, it was revealed that the detection performance for small objects with a low number of pixels within the image was low. If small objects' bounding boxes were not accurately detected, SAM was not able to generate polygon masks for those objects. Secondly, SAM experienced difficulties in generating accurate mask labels in areas where the object's features are similar to the background. For example, as shown in Fig. 4, the generated polygon annotations by SAM are more accurate compared to the annotations by humans. This was particularly pronounced in situations involving objects partially obscured by other objects. While SAM generates masks without considering the obscured parts, ground truth data includes the obscured portions of objects in the polygon masks. Since the exact shape of objects hidden behind obstructions cannot be known, the predicted polygons by SAM are more accurate than the original masks in these

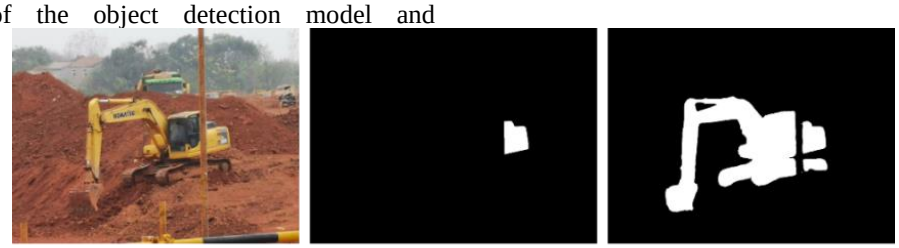
respects. Additionally, some ground truth annotations were represented differently from the actual objects' appearance. These factors are deemed to have negatively impacted the overall performance of SAM.

7 Conclusion

This study presents the automated polygon annotation generation method using YOLOv8 for an object detector and SAM for an annotation generator. The experimental results showed promising results, with quality annotations comparable to the ground truth, as shown in Figure 3. These results provide important implications for the accuracy and reliability of generating polygon label data in complex environments like construction sites. Therefore, future research should focus on improving the performance of the object detection model and optimizing the SAM model. Through this, more accurate and reliable generation of polygon label data is expected, contributing to the expansion of the application scope in the fields of deep learning and computer vision.

Acknowledgment

This research was conducted by the support of the "2023 Yonsei University Future-Leading Research Initiative (No. 2023-22-0114)" and the "National R&D Project for Smart Construction Technology (No. RS-2020-KA156488)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation.



(e): Incorrect labeling of Ground truth

(f): Occlusion problem

Figure 4. Original image (Left), Ground truth mask (Center), SAM predicted mask (Right)

Table 1. mAP performance for each class of object detection

Other Model	**Target Class**												
	Worker	Static crane	Hanging head	Crane	Roller	Bulldozer	Excavator	Truck	Loader	Pump truck	Concrete mixer	Pile driving	Other vehicle
Yolov8 object detection model (Ours)	91.3	80.7	80.5	77.7	90.5	91.3	94.2	87.0	83.3	86.8	89.5	80.8	80.9

Table 2. SAM result on MOCS dataset

Segment Anything	**Target Class**

Model		Worker	Static crane	Hanging head	Crane	Roller	Bulldozer	Excavator	Truck	Loader	Pump truck	Concrete mixer	Pile driving	Other vehicle
mIoU	Predicted BBOX	78.5	65.6	73.0	68.2	86.7	83.2	76.3	80.5	83.2	66.3	65.9	55.8	59.9
	GT BBOX	83.1	79.7	78.6	77.1	89.2	86.3	79.8	85.9	86.8	75.7	86.6	65.4	77.4

References

[1] M. Abbas, B.E. Mneymneh, H. Khoury, Assessing on-site construction personnel hazard perception in a Middle Eastern developing country: An interactive graphical approach, Safety Science 103 (2018) 183–196. https://doi.org/10.1016/j.ssci.2017.10.026.

[2] S. Teran, H. Blecker, K. Scruggs, J. García Hernández, B. Rahke, Promoting adoption of fall prevention measures among Latino workers and residential contractors: Formative research findings, American Journal of Industrial Medicine 58 (2015) 870–879. https://doi.org/10.1002/ajim.22480.

[3] F. Craveiro, J.P. Duarte, H. Bartolo, P.J. Bartolo, Additive manufacturing as an enabling technology for digital construction: A perspective on Construction 4.0, Automation in Construction 103 (2019) 251–267. https://doi.org/10.1016/j.autcon.2019.03.011.

[4] S.K. Baduge, S. Thilakarathna, J.S. Perera, M. Arashpour, P. Sharafi, B. Teodosio, A. Shringi, P. Mendis, Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications, Automation in Construction 141 (2022) 104440. https://doi.org/10.1016/j.autcon.2022.104440.

[5] Economies | Free Full-Text | Industry 4.0 for the Construction Industry: Review of Management Perspective, (n.d.). https://www.mdpi.com/2227-7099/7/3/68 (accessed March 13, 2024).

[6] Y. Hong, W.-C. Chern, T.V. Nguyen, H. Cai, H. Kim, Semi-supervised domain adaptation for segmentation models on different monitoring settings, Automation in Construction 149 (2023) 104773. https://doi.org/10.1016/j.autcon.2023.104773.

[7] R. Bai, M. Wang, Z. Zhang, J. Lu, F. Shen, Automated Construction Site Monitoring Based on Improved YOLOv8-seg Instance Segmentation Algorithm, IEEE Access 11 (2023) 139082–139096. https://doi.org/10.1109/ACCESS.2023.3340895.

[8] Z. Wang, Y. Zhang, K.M. Mosalam, Y. Gao, S.-L. Huang, Deep semantic segmentation for visual understanding on construction sites, Computer-Aided Civil and Infrastructure Engineering 37 (2022) 145–162. https://doi.org/10.1111/mice.12701.

[9] J. Cho, K. Kim, Detection of moving objects in multi-complex environments using selective attention networks (SANet), Automation in Construction 155 (2023) 105066. https://doi.org/10.1016/j.autcon.2023.105066.

[10] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, E. Herrera-Viedma, Data set quality in Machine Learning: Consistency measure based on Group Decision Making, Applied Soft Computing 106 (2021) 107366. https://doi.org/10.1016/j.asoc.2021.107366.

[11] N.A. Koohbanani, B. Unnikrishnan, S.A. Khurram, P. Krishnaswamy, N. Rajpoot, Self-Path: Self-Supervision for Classification of Pathology Images With Limited Annotations, IEEE Transactions on Medical Imaging 40 (2021) 2845–2856. https://doi.org/10.1109/TMI.2021.3056023.

[12] Y. Gao, H. Li, W. Fu, Few-shot learning for image-based bridge damage detection, Engineering Applications of Artificial Intelligence 126 (2023) 107078. https://doi.org/10.1016/j.engappai.2023.107078.

[13] H. Wang, M. Xu, B. Ni, W. Zhang, Learning to Combine: Knowledge Aggregation for Multi-Source Domain Adaptation, (2020). https://doi.org/10.48550/arXiv.2007.08801.[14] G. Jocher, A. Chaurasia, J. Qiu, YOLO by Ultralytics, (2023). https://github.com/ultralytics/ultralytics (accessed December 28, 2023).

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment Anything, (2023).http://arxiv.org/abs/2304.02643 (accessed July 27, 2023).

[16] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, L. Zhiwei, Dataset and benchmark for detecting moving objects in construction sites, Automation in Construction 122 (2021) 103482. https://doi.org/10.1016/j.autcon.2020.103482.

[17] M. Gröger, V. Borisov, G. Kasneci, BoxShrink: From Bounding Boxes to Segmentation Masks, (2022). http://arxiv.org/abs/2208.03142 (accessed December 27, 2023).

[18] H. Murtaza, M. Ahmed, N.F. Khan, G. Murtaza, S. Zafar, A. Bano, Synthetic data generation: State of the art in health care domain, Computer Science Review 48 (2023) 100546. https://doi.org/10.1016/j.cosrev.2023.100546.

[19] J. Kim, J. Kim, Y. Kim, H. Kim, 3D reconstruction of large-scale scaffolds with synthetic data generation and an upsampling adversarial network, Automation in

Construction 156 (2023) 105108. https://doi.org/10.1016/j.autcon.2023.105108.

[20] Li Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Trans. Pattern Anal. Machine Intell. 28 (2006) 594–611. https://doi.org/10.1109/TPAMI.2006.79.

[21] S. Ravi, H. Larochelle, Optimization as a Model for Few-Shot Learning, in: 2016. https://openreview.net/forum?id=rJY0-Kcll (accessed December 28, 2023).

[22] W. Wang, V.W. Zheng, H. Yu, C. Miao, A Survey of Zero-Shot Learning: Settings, Methods, and Applications, ACM Trans. Intell. Syst. Technol. 10 (2019) 1–37. https://doi.org/10.1145/3293318.

[23] D. Reis, J. Kupec, J. Hong, A. Daoudi, Real-Time Flying Object Detection with YOLOv8, (2023). http://arxiv.org/abs/2305.09972 (accessed December 27, 2023).

[24] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common Objects in Context, (2015). http://arxiv.org/abs/1405.0312 (accessed February 13, 2024).

[25] C. Zhang, L. Liu, Y. Cui, G. Huang, W. Lin, Y. Yang, Y. Hu, A Comprehensive Survey on Segment Anything Model for Vision and Beyond, (2023). http://arxiv.org/abs/2305.08196 (accessed December 28, 2023).

[26] W.-C. Chern, T. Kim, T.V. Nguyen, V.K. Asari, H. Kim, Self-supervised sub-category exploration for Pseudo label generation, Automation in Construction 151 (2023) 104862. https://doi.org/10.1016/j.autcon.2023.104862.