Iterative Annotation for Road Damage Detection Using Human-in-the-Loop with a Vision and Language Model

Ryuichi Tomiya, Tristan Hascoet, Ryoichi Takashima, and Tetsuya Takiguchi

Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, 657-8501, Japan
236x049x@stu.kobe-u.ac.jp

Abstract. In this work, we investigated a human-in-the-loop approach along with vision and language models for the effective annotation of target road images for developing an object detection model at a limited cost. This work demonstrated that using our method allows for a reduction in the workload of the dataset construction of a target's road damage detection.

Keywords: human-in-the-loop, road damage detection, vision and language model.

1 Introduction

Road infrastructure monitoring that inspects and diagnoses deteriorating roads is a crucial challenge. Due to financial constraints and a shortage of personnel, there is a growing demand for such maintenance to be carried out at a limited cost. Recent advancements in computer vision have facilitated the automation of inspection and diagnosis, leading to increased efficiency and stability in operations [1]. The development of computer vision models requires a labeled dataset with manual annotation. However, there is a contradiction between the goal of deep learning to reduce human labor and the fact that a large amount of annotation is necessary for model development. To address this contradiction, a human-in-the-loop approach is being considered. This framework significantly reduces the human annotation burden, enabling the efficient and continuous updating of models through iterative processes of training data collection and parameter updates [2].

Concurrently, numerous studies have been published on vision and language models that combine computer vision and natural language understanding. Pretrained vision and language models with large-scale datasets have been made publicly available. These models, through the combination of language features and image features in training, possess zero-shot recognition, enabling them to even understand images of specific categories without being explicitly trained on those categories. Furthermore, when using a model with zero-shot recognition on a particular dataset, further accuracy improvement is expected through finetuning on limited samples (few-shot learning) [3–5].

The effectiveness of the human-in-the-loop approach depends on the few-shot ability of the model used. We think that integrating the vision & language model with high few-shot ability into the human-in-the-loop approach is beneficial, and we investigated its contributions. We utilized the road infrastructure monitoring task as a suitable challenge for our approach. In this work, we investigated a human-in-the-loop approach along with a vision and language model for the effective annotation of a target road's images for developing an object detection model that can be carried out at a limited cost.

2 Related Work

In this work, we utilized the vision and language model GLIP [4] as the baseline object detection model. GLIP unifies the object detection and phrase grounding tasks to learn an object-level, language-aware, and semantic-rich visual representation. GLIP demonstrates high accuracy through fine-tuning on datasets from various tasks.

Furthermore, the human-in-the-loop concept is also being applied in real-world situations. Miao et al. have proposed an efficient method to improve the accuracy of a classification model for wildlife monitoring [2]. Additionally, Adhikari et al. proposed the Iterative Bounding Box Annotation method as an efficient application of human-in-the-loop for constructing datasets for object detection models [6]. They propose a semi-automatic method for efficient bounding box annotation. However, they did not evaluate the accuracy when training on an efficiently constructed dataset.

Arya et al. have provided the RDD2022 dataset for road damage detection. This dataset consists of 47,420 road images from six countries that serve as images for road infrastructure monitoring [7]. Information regarding the types of damage and bounding box annotations is provided for each image. Assuming that these images are not annotated, the cost of annotating them manually can be automatically calculated from the information of the ground truth in the dataset. This work investigates whether Iterative Bounding Box Annotation can be adapted to practical road damage detection tasks using the vision and language model GLIP. It also investigates the accuracy of the model when data collected with efficient methods is used for training.

3 Iterative Bounding Box Annotation Using GLIP

Figure 1 shows the annotation method employed in this work. In the first step, all unlabeled images are input into GLIP along with the category names to be detected, producing output for bounding boxes, class labels, and scores. The second step is sorting the images in the order of annotation and selecting 50 images. Two sorting methods were used: one based on arranging images in the descending order of the highest detection scores in the image and the other using a random order. The third step, which follows the procedure of [6], involves having a human annotator review and manually correct the bounding boxes and

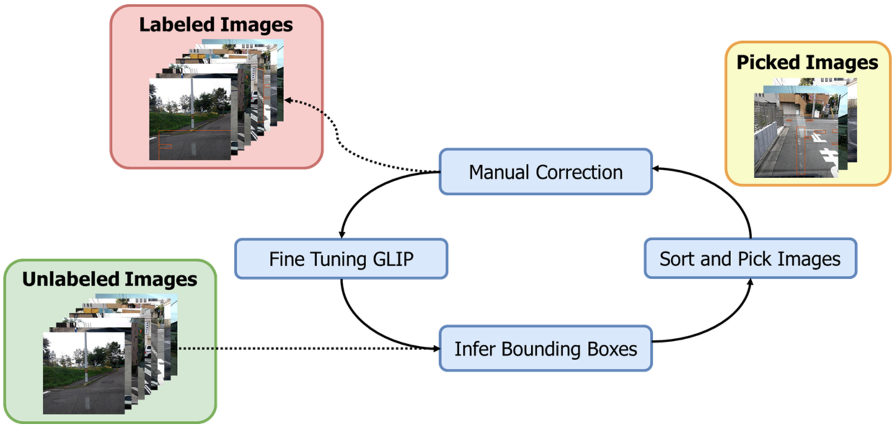

Fig. 1. Method for dataset construction using human-in-the-loop. The loop continues until annotations are provided for all unlabeled images.

class labels of the selected 50 images. Incorrectly predicted boxes are removed, wrongly labeled classes are corrected, and new boxes are drawn, if needed. The corrected images are accumulated as labeled images. The fourth step involves fine-tuning GLIP using the accumulated labeled images, which were collected as training data through the corrections made so far. These four steps are repeated until annotation is completed for all the images in the unlabeled images. This method aims to generate a fully-labeled image dataset for object detection through iterative loops while reducing the workload for a human annotator.

The estimation method for workload is adopted from Adhikari et al.'s approach [8], following the methodology used in [6]. The workload is equivalent to $\#corrections$ of the following equation.

$$\#additions = (\#\, of\, true\, objects) \times (1 - recall) \tag{1}$$

$$\#removals = (\#\, of\, all\, detections) \times (1 - precision) \tag{2}$$

$$\#corrections = \#additions + \#removals \tag{3}$$

The information from ground truth is compared with the detection results, and a detection with a partial overlap exceeding 50% (IoU>0.5) is taken as the correct detection. We estimate the workload from the above equations.

4 Results of Annotation and Road Damage Detection

4.1 Effectiveness of Iterative Annotation

Figure 2 shows the progression of our method's experiments on 1,000 images of the RDD2022 Japan dataset using the GLIP-Large model. We used 1,000 images from the RDD2022 Japan dataset for annotation purposes, 1,313 images for validation, and 1,313 images for evaluation which were not used in training and validation. The blue circle and the green star represent the amount of ground truth and manual correction in terms of numbers of bounding boxes as a function of the number of images, respectively. The left graph shows the results of annotations conducted in the order in which images were randomly shuffled from the dataset. On the other hand, the right graph shows the results of annotations performed in a descending order of rearranging images based on the highest score predicted by the detection model for each image. Both show that the workload has been reduced using our method.

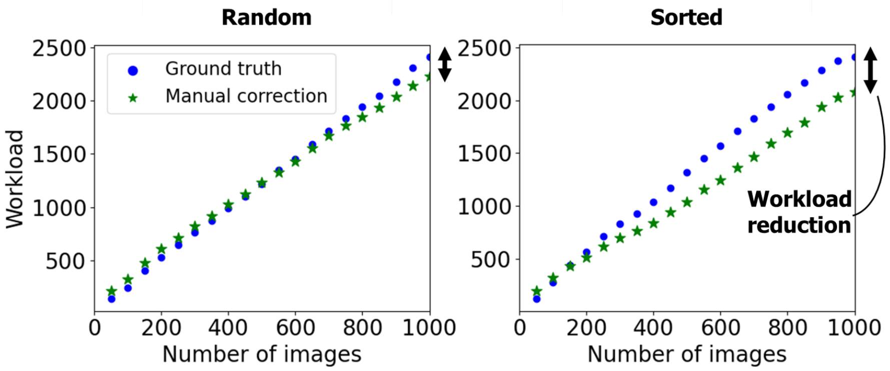

Fig. 2. An example of the effect of the order of iterative annotation. The figures show the cumulative number of ground truth boxes and the manual corrections required in the dataset. These numbers are equivalent to workload. The images are annotated by sorting them in random order (left) and in the descending order of the highest detection scores in the image (right). Workload reduction is better when sorted in the descending order of the highest detection scores in the image.

Table 1 shows the workload reduction for annotations averaged over three trials. Both methods effectively reduced the workload, but it was demonstrated that the approach of checking detections in the descending order of scores is superior. This might be because, from an early iteration, the boxes to be detected were more likely to be correct. Additionally, using instances of false positives at high scores as negative examples during training might have played a crucial role. As a result, the detection model improved its performance over subsequent iterations, leading to more accurate detections. In this experiment, we utilized a labeled dataset, and among the 1,000 images annotated this time, only 27 images did not contain any damage.

However, considering practical scenarios, such as checking long-duration videos where many frames do not contain any damage, it becomes necessary to inspect numerous images without damage to confirm if there is some damage. In such situations, confirming images input to GLIP in descending order of

Table 1. Comparison of workload and workload reduction [%] using iterative annotation.

	workload	workload reduction
Manual	$2,410$	-
Iterative (random)	$2,217.33$	7.99
Iterative (sorted)	$2,096.67$	13.00

scores allows for the early acquisition of damage information, potentially leading to improved accuracy. Therefore, in practical scenarios, it is conceivable that the advantages of confirming in descending score order could be more pronounced.

4.2 Accuracy Using the Human-in-the-Loop Approach

Table 2 shows the Average Precision (AP) achieved when training with data collected using each approach. There is no significant difference in AP when using manually collected data for training with any approach. We found there is validity in utilizing data collected using the human-in-the-loop approach for model training. The high AP for Manhole Cover and Crosswalk Blur can be attributed to the consistent shapes and features present in almost every image. On the other hand, the lower AP for Crack types may be due to variations in length and depth across different images. The training data consisting of 1,000 images might not be sufficient to thoroughly capture these diverse characteristics, leading to lower accuracy in the detection of Crack types.

Table 2. Comparison of Average Precision (AP) [%] in detection using GLIP trained on 1,000 annotated images. (AC = Alligator Crack, WLB = White Line Blur, PH = Pothole, LC = Longitudinal Crack, MC = Manhole Cover, TC = Transverse Crack, CB = Crosswalk Blur)

	AC	WLB	PH	LC	MC	TC	CB	mean
Manual	60.15	63.37	53.07	42.04	83.15	40.80	84.25	60.98
Iterative (random)	60.90	61.92	51.53	43.00	81.28	43.16	82.31	60.59
Iterative (sorted)	60.91	62.01	51.89	42.66	82.28	41.99	83.54	60.75

Figure 3 shows detection examples. The left image shows two detections where both position and class are accurately predicted. The center image shows an example where the damage location is correct, but the classification is incorrect. The right image shows an issue with position detection. The detection results are indicated by blue boxes, and the ground truth is represented by orange boxes. A human annotator needs to remove the blue box and add the orange box.

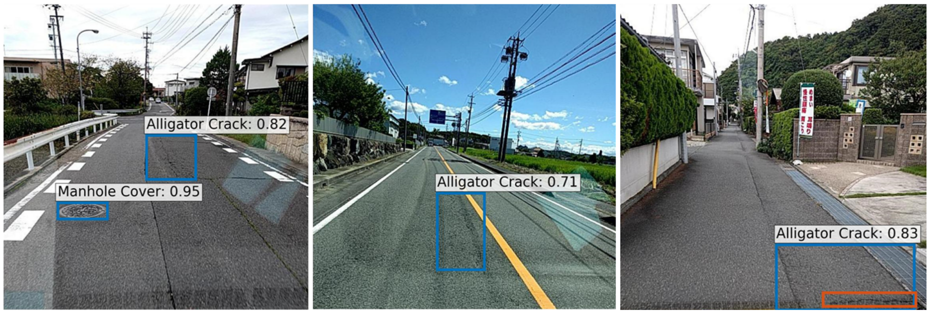

Fig. 3. Example of detection results using GLIP trained by iterative annotation. The left image shows the correct detection and classification of a manhole cover and an alligator crack, with both position and class accurately predicted. The center image shows a case where the damage location is correct, but the classification is incorrect. Although predicted as an Alligator Crack, it is a Longitudinal Crack. The right image shows an issue with position detection and classification. The correct prediction should be a Transverse Crack at the orange position (lower right). All images are in RDD2022 Japan dataset [7].

Following the approach of Adhikari et al. [8] and using the label information in the dataset, we calculated the workload for box corrections. The iterative bounding box annotation using GLIP demonstrated its contribution to reducing the workload in constructing datasets for road damage detection. Additionally, it was found that confirming detections in descending order of scores is superior to confirming them in random order.

5 Conclusions

This work demonstrated that using human-in-the-loop and a vision-language model allows for a reduction in the workload of dataset construction. Through a comparative experiment on the order of image confirmation, it was found that the approach utilizing scores output by the model is superior. This will prove useful in developing object detection models for target roads at a limited cost. Furthermore, it was shown that using reduced workload data as training data does not result in a change in accuracy compared to fully manually annotated data.

In the future, when adopting this method for constructing datasets in actual infrastructure monitoring, we need to investigate the extent to which it contributes to reducing the workload for human annotators. Additionally, it is essential to explore efficient methodologies considering factors such as the contribution to larger dataset construction when increasing the number of images from 1,000, the number of images to be confirmed in each correction, and the training time and computational costs of the model.

Our final goal is to propose a system for developing an efficient and highly-accurate road damage detection model. We will also investigate methods to enhance the model's accuracy while minimizing annotation costs. This will involve careful consideration of data labeling within road images to contribute to accuracy improvement, with a focus on the selection of the type of damage that requires annotation.

References

1. Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., Omata, H.: Road damage detection and classification using deep neural networks with smartphone images. In: Computer-Aided Civil and Infrastructure Engineering, vol. 33, no. 12, pp. 1127-1141. (2018)
2. Miao, Z., Liu, Z., Gaynor, K. M., Palmer, M. S., Yu, S. X., Getz, W. M.: Iterative human and automated identification of wildlife images. In: Nature Machine Intelligence, vol. 3, no. 10, pp. 885-895. (2021)
3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748-8763. (2021)
4. Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J. N., Chang, K. W., Gao, J.: Grounded Language-Image Pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10965-10975. (2022)
5. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1780-1790. (2021)
6. Adhikari, B., Huttunen, H.: Iterative Bounding Box Annotation for Object Detection. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4040-4046. (2021)
7. Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., Sekimoto, Y.: RDD2022: A multi-national image dataset for automatic Road Damage Detection. arXiv preprint arXiv:2209.08538. (2022)
8. Adhikari, B., Peltomaki, J., Puura, J., Huttunen, H.: Faster Bounding Box Annotation for Object Detection in Indoor Scenes. In: 2018 7th European Workshop on Visual Information Processing (EUVIP), pp. 1-6. (2018)