Text-enhanced Label-efficient Automated Bridge Defect Semantic Segmentation from Inspection Images

Shengyi Wang, M.S., S.M.ASCE[1]; and Nora El-Gohary, A.M.ASCE[2]

[1]Ph.D. Student, Dept. of Civil and Environmental Engineering, Univ. of Illinois at Urbana-Champaign, Urbana, IL. Email: shengyi4@illinois.edu
[2]Professor, Dept. of Civil and Environmental Engineering, Univ. of Illinois at Urbana-Champaign, Urbana, IL. Email: gohary@illinois.edu

ABSTRACT

The utilization and integration of unmanned aerial vehicles (UAVs) and computer vision technologies in recent automated bridge inspection methodologies has shown advancement in capturing and analyzing images to enhance the efficiency and safety of bridge inspection. However, information extraction from inspection images collected on-site remains challenging. First, although extensive research efforts have focused on segmenting defects from images, the localization and segmentation performance is limited due to complex backgrounds and irregular defect shapes in images. Second, precise pixel-level annotation of defect masks is labor-intensive and time-consuming, which underscores the need for a label-efficient method for defect segmentation. To address these gaps, this paper proposes a deep learning-based method to extract and segment different types of bridge defects from on-site inspection images using a label-efficient way, which leverages corresponding text descriptions, the Grounding DINO (DETR with Improved deNoising anchOr boxes) object detection model, and the segment anything model (SAM). This paper discusses the proposed method and its performance results. The experimental results show that the method can efficiently extract and segment various bridge defects, which would support automated bridge inspection.

INTRODUCTION

Transportation agencies in the United States face the challenge of proactively managing the nation's aging civil infrastructure. Meanwhile, a growing concern for bridges is marked by declines in operational efficiency, delays in recovery operations, and deterioration. The United States has over 617,000 bridges, of which nearly half are older than 50 years. Given the current pace of funding, it would take 50 years to cover the $125 billion needed for rehabilitation (ASCE 2021). These statistics underscore the need for more efficient bridge inspection and maintenance systems to ensure safety and optimal use of the limited resources.

Recently, unmanned aerial vehicles (UAVs) and computer vision inspection have been utilized and integrated to detect defects from bridge inspection images, ranging from traditional approaches (e.g., statistical methods, binarization methods, machine learning-based models) to deep learning techniques (Hüthwohl et al. 2019, Munawar et al. 2021, Zheng et al. 2022). However, there are two major knowledge gaps that exist. First, current techniques have limitations when applied to an automated bridge inspection environment because irregular shapes, lighting conditions during image capturing, or different backgrounds can affect the performance of the

results (Kang et al. 2020). Second, there is a lack of a label-efficient method for defect segmentation. Current deep learning-based defect segmentation relies on supervised learning, which requires extensive and accurate pixel-level labeled images (Bianchi and Hebdon 2022, Savino and Tondolo 2023, Wang and El-Gohary 2024). The image labeling process is time-consuming and labor-intensive. Besides that, the model trained on labeled data may not perform well on previously unseen data.

To address these challenges, this paper proposes a deep learning-based method to extract and segment different types of bridge defects from on-site inspection images using a label-efficient way, which leverages corresponding text descriptions, the Grounding DINO (DETR with Improved deNoising anchOr boxes) (Liu et al. 2023) object detection model, and the segment anything model (SAM) (Kirillov et al., 2023). This paper discusses the image data acquired for evaluation, the proposed method, and its performance results on segmenting spalling and exposed rebar. To the best of the authors' knowledge, it is the first attempt to segment bridge defects leveraging text prompts.

BACKGROUND

Grounding DINO. Unlike traditional object detection algorithms which can only detect predefined object classes, Grounding DINO is the state-of-the-art model for open-set object detection. Grounding DINO utilizes the shifted windows (Swin) transformer (Liu et al. 2021) as the image feature extraction backbone and bidirectional encoder representations from transformers (BERT) (Devlin et al. 2019) for textual feature embedding. It also performs multi-stage multi-modal fusion. Such strategies effectively improve its performance on object detection on rare (i.e., few-shot) or unseen (i.e., zero-shot) object classes.

SAM. The development of large language models has changed the field of natural language processing significantly. Parallel to this, the recent development of SAM has marked a milestone in computer vision. SAM introduces a prompt-based mechanism, where the prompt can be several points, bounding boxes, and so on. The model can provide a segmentation mask based on the prompts. SAM employs the vision transformer (ViT) (Dosovitskiy et al. 2020) as the backbone and was pretrained on a large dataset, the Segment Anything 1 Billion (SA-1B), which contains 11M images and 1.1 billion masks. It demonstrated great success in various tasks in the general domain, including image annotation, image inpainting, and object tracking (Zhang et al. 2023). However, the potential of SAM remains unexplored in the civil infrastructure domain.

PROPOSED DEEP LEARNING-BASED METHOD FOR LABEL-EFFICIENT BRIDGE DEFECT SEGMENTATION

This paper proposes a deep learning-based label-efficient method to automatically identify and segment bridge defects from bridge inspection images. The proposed method includes three primary steps: (1) data collection and annotation; (2) method framework; and (3) evaluation.

Data Collection and Annotation. A total of 3,000 image-text pairs were collected from the Washington Department of Transportation (WSDOT) (Wang and El-Gohary 2023), and 20 images were randomly selected where their corresponding text descriptions contained spalling or exposed rebar. The ground truth masks for those images were annotated for evaluation.

Method Framework. This study utilized an approach combining the capabilities of the Grounding DINO and the SAM models. Fig. 1 illustrates the framework of this proposed method. The Grounding DINO model, pretrained on the Object365 dataset, was employed for localization. SAM, having been pretrained on an extensive dataset of one billion masks, was leveraged for detailed object segmentation. SAM only focuses on segmentation based on prompts without predicting semantic labels to the identified objects. The Grounding DINO model provides a prompt in the form of a bounding box, which is coupled with class information derived from textual input.

Note: DINO = DETR with Improved deNoising anchOr boxes, SAM = segment anything model

Figure 1. Proposed method framework for bridge defect segmentation.

Model Evaluation. The evaluation of bridge defect segmentation was conducted by comparing the segmentation results to the ground truth using four key metrics: precision, recall, F-1 measure, and Intersection-over-Union (IoU). Precision is defined as the correctly segmented pixels over the total predicted pixels, indicating the accuracy of positive predictions. Recall measures the proportion of correctly identified pixels against the total pixels that should have been identified, representing the model's ability to find all relevant cases. The F-1 measure, or Dice score, is the harmonic mean of precision and recall. The IoU metric measures the overlap between the predicted and ground truth segmentation.

PRELIMINARY EXPERIMENTAL RESULTS AND DISCUSSION

The proposed method was evaluated using a testing dataset of 20 annotated images. The performance results are summarized in Table 1. Fig. 2 shows some examples of predicted segmentation output and ground truth annotation, where the text prompt of the first row is "spalling", and the rest are "exposed rebar". Overall, the model achieved mean precision, recall, F-1 measure, and IoU of 0.565, 0.833, 0.603, and 0.495, respectively, on the testing dataset. As per Table 1 and Fig. 2, the model showed good performance in identifying and segmenting spalling areas (row 1). However, for exposed rebar, it showed high recall (0.949) but low precision (0.397), suggesting that many non-rebar pixels were incorrectly identified as exposed rebar. As shown in Fig. 2, although the Grounding DINO model provides a relatively accurate bounding box, the SAM model also segments rust (row 2), pipes, and spacing area (row 3) surrounding them as part of the exposed rebar. The experiments were carried out on a Windows 11 system with the Intel(R) 11th Gen Intel(R) Core (TM) i9-11900KF @ 3.50GHz CPU, 32.0 GB RAM, and NVIDIA GeForce GTX 3090 GPU (Graphics Processing Unit).

Table 1. Model performance on bridge defect segmentation.

Class	Precision	Recall	F-1(Dice)	IoU
Spalling	0.733	0.717	0.694	0.606
Exposed rebar	0.397	0.949	0.513	0.384
Mean	0.565	0.833	0.603	0.495

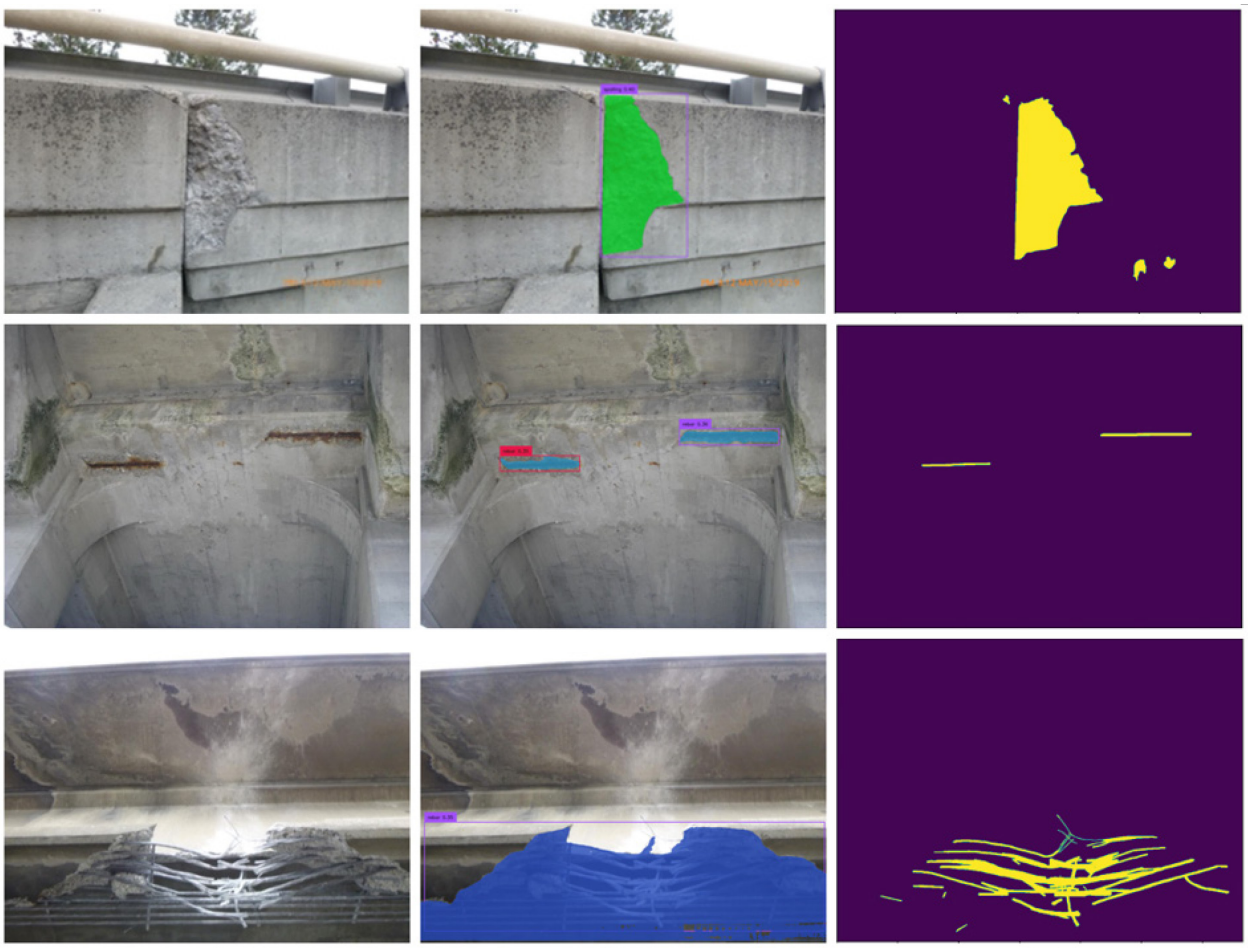

Figure 2. Examples of (a) original images, (b) predicted masks, and (c) ground truth masks.

CONCLUSION AND FUTURE WORK

In this paper, a deep learning-based method for segmenting bridge defects (i.e., spalling and exposed rebar) leveraging text descriptions was proposed. The proposed method was evaluated on a testing dataset of 20 images with ground truth annotation. The method achieved a mean precision, recall, F-1 measure, and IoU of 0.565, 0.833, 0.603, and 0.495, respectively, which indicates its potential in supporting label-efficient automated bridge defect segmentation. Two main limitations of the work are acknowledged. First, the number of images and classes for evaluation is limited, which cannot represent all the cases in bridge inspection. Second, there is a need for improvement to enhance the accuracy of detection and segmentation. In their future work, the authors plan to address the aforementioned limitations by expanding the size of the data, extending the work to more classes, performing additional experiments on other pretrained models, and exploring more advanced model architectures.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Foundation (NSF). This material is based on work supported by the NSF under Grant No. 2305883. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

ASCE (2021). "Infrastructure report card." from https://infrastructurereportcard.org/.

Bianchi, E., and Hebdon, M. (2022). "Visual structural inspection datasets." *Autom. Constr.*, 139, 104299.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929*.

Hüthwohl, P., Lu, R., and Brilakis, I. (2019). "Multi-classifier for reinforced concrete bridge defects." *Autom. Constr.*, 105, 102824.

Kang, D., Benipal, S. S., Gopal, D. L., and Cha, Y.-J. (2020). "Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning." *Autom. Constr.*, 118, 103291.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). "Segment Anything."

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." *arXiv preprint arXiv:2103.14030*.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. (2023). "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection." *arXiv preprint arXiv: 2303.05499*.

Munawar, H. S., Hammad, A. W., Haddad, A., Soares, C. A. P., and Waller, S. T. (2021). "Image-based crack detection methods: A review." *Infrastructures*, 6(8), 115.

Savino, P., and Tondolo, F. (2023). "Civil infrastructure defect assessment using pixel-wise segmentation based on deep learning." *J. Civ. Struct. Health Monit.*, 13(1), 35-48.

Wang, S., and El-Gohary, N. (2023). "Automated bridge inspection image interpretation based on vision-language pre-training." *Proc., 2023 ASCE International Conference on Computing in Civil Engineering*, Corvallis, OR, June 25-28, 2023.

Wang, S., and El-Gohary, N. (2024). "Semantic Segmentation of Bridge Components from Various Real Scene Inspection Images." *Proc., 2024 ASCE CI & Construction Research Congress (CRC) Joint Conference*, Des Moines, IA, March 20-23, 2024.

Zhang, C., Puspitasari, F. D., Zheng, S., Li, C., Qiao, Y., Kang, T., Shan, X., Zhang, C., Qin, C., Rameau, F., Lee, L.-H., Bae, S.-H., and Hong, C. S. (2023). "A Survey on Segment Anything Model (SAM): Vision Foundation Model Meets Prompt Engineering." *arXiv preprint arXiv:2306.06211*.

Zheng, Y., Gao, Y., Lu, S., and Mosalam, K. M. (2022). "Multistage semisupervised active learning framework for crack identification, segmentation, and measurement of bridges." *Comput.-Aided Civ. Infrastruct. Eng.*, 37(9), 1089–1108.