Research on bridge disease recognition algorithm based on SAM and YOLOv8

Weiqiang Liu[a], Di Wu[b,c], Zhenyu Chen[*b,c]
[a]Guangzhou Road Affairs Center, Guangzhou 510635, Guangdong, China; [b]Guangdong Jianke Traffic Engineering Quality Inspection Center Co., Ltd., Guangzhou 510530, Guangdong, China; [c]Guangdong Provincial Transportation Infrastructure Intelligent Detection Engineering Technology Research Center, Guangzhou 510530, Guangdong, China

ABSTRACT

In this paper, a new method for bridge disease image segmentation is introduced, in which the data set includes exp_rebar, breakage, patch and joint. The proposed method uses the YOLOv8 model to partition the region of disease interest, which serves as the cue input of the Segment Anything Model (SAM) and the high-quality HQ-SAM pre-trained large model, and performs automatic and accurate segmentation based on this. In this study, three evaluation indexes including accuracy, recall rate and F1 score were used to quantify the accuracy of segmentation results of YOLOv8, YOLOv8+SAM and YOLOv8+HQ-SAM models. The results show that the SAM model performs better than the other two models, showing higher segmentation accuracy and overall performance. Although HQ-SAM is improved by SAM, the more complex network architecture did not achieve the expected gain on the dataset in this paper. The YOLOv8+SAM model proposed in this paper provides a new technical direction for the intelligent recognition of bridge diseases.

Keywords: SAM, HQ-SAM, YOLOv8, bridge disease, image segmentation, LPFMs

1. INTRODUCTION

In recent years, with the rapid development of deep learning algorithms, the research on bridge disease image recognition has also been rapidly promoted[1-3]. However, the use of deep neural networks for bridge disease intelligent recognition also has certain limitations. Firstly, due to different bridge structures and their disease characteristics, deep neural networks often need to be customized for specific scenarios, and their generalization ability is limited. Secondly, training a network model with complex structure and remarkable recognition effect requires a large amount of disease image data, especially accurate disease information labeled by professional engineers. This process is time-consuming and laborious.

With the rapid development of Large-scale Pretrained Foundation Models (LPFMs) in the field of artificial intelligence, more and more tasks have been better performed by fine-tuning on LPFMs[4,5]. LPFMs refers to a general model trained with large amounts of data, which learns more basic and general representation capabilities and can be transferred to different domains, so that different downstream tasks can be easily fine-tuned based on such a general model[6]. In April 2023, Segment Anything Model (SAM) was proposed as a basic large model in the field of natural image segmentation[7]. SAM is trained on a large-scale dataset SA-1B[7] based on Vision Transformer (ViT)[8] model, which makes the model have strong generalization ability and achieves excellent results in the segmentation of natural images by combining three different prompt modes: point prompt, frame prompt and text prompt[9]. However, although the powerful ability of SAM in natural image segmentation provides a new perspective for intelligent recognition of bridge disease, the effect of direct application of SAM in bridge disease image segmentation is not satisfactory. There are two main reasons for this. The first is the limitation of training data set. In the training process, SAM lacks bridge disease images, and the edges of bridge disease images are usually fuzzy, which is quite different from the clear edges in natural images. Secondly, the characteristics of the model prompt, that is, the prompt mode of SAM has a significant impact on the segmentation results, and only by choosing the prompt mode properly can the potential of SAM be fully utilized[10,11].

In response to these two problems, this paper proposed a new bridge disease recognition method: by training a small part of samples by YOLOv8[12,13], the image is roughly divided into the boundaries of various diseases, and the region and

*1069320910@qq.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133952I · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048877

corresponding labels are used as the hint of SAM model, combined with the excellent segmentation ability of SAM large model, the hint is divided, and more accurate disease region is obtained. This method can solve the drawbacks of SAM in the intelligent recognition of bridge diseases, and further exert the segmentation ability of SAM large model.

2. CONSTRUCTION OF BRIDGE DISEASE DATA SET

2.1 Standardized image acquisition

Bridge apparent disease image can represent the current state of bridge, so the acquisition of disease image is the premise of the whole intelligent recognition algorithm. Bridge disease images are collected from HarmonyOS image sensor BIDS mounted on bridge inspection vehicles, as shown in Figure 1. The CMOS model of the sensor is IMX707Y, the pixel resolution is 5000w (8192*6144), and the acquisition accuracy range is (0.05 mm-1.0 mm). Based on this, this paper establishes a standardized image to ensure the quality of acquisition and unify the image size and resolution, which is convenient for subsequent training.

Figure 1. HarmonyOS image sensor BIDS to collect bridge diseases on site.

The samples used in this experiment are divided into four categories: exp_rebar, breakage, patch and joint. Since SAM and HQ-SAM are both pre-trained large models and only need YOLOv8 to roughly divide the region of interest as the prompt input of the large model, the number of samples for YOLOv8 training this time is only 100. Among them, the number of diseases is 445, and the distribution of various diseases is detailed in Table 1:

Table 1. Detailed statistics of various target diseases.

	exp_rebar	Breakage	Patch	Joint
Amount	116	133	101	95
Percentage	26.07%	29.89%	22.69%	21.35%

Note: A disease image can contain multiple diseases

3. MODEL BUILDING

3.1 SAM network structure

SAM is an innovative breakthrough in the field of computer vision, mainly used for image segmentation tasks. SAM uses a concise structure, and by training on a large dataset SA-1B, the model has unprecedented generalization ability. The SA-1B dataset has high quality and rich diversity, which provides strong support for SAM's high segmentation capability. The structure of SAM is shown in Figure 2.

The three core modules of SAM are the VIT-based image encoder, the prompt encoder that can input multiple prompts, and the lightweight mask decoder. First, the structure of the image encoder is a ViT-based Masked autoencoders (MAE), which is used to extract image embeddings from high-resolution image inputs and input them into the mask decoder together with the prompt embeddings after feature extraction. The processing of mask hints is to sum the image features element by element after convolution. Finally, the mask decoder predicts the segmentation mask by fusing the image embeddings and hint embeddings provided by the encoder.

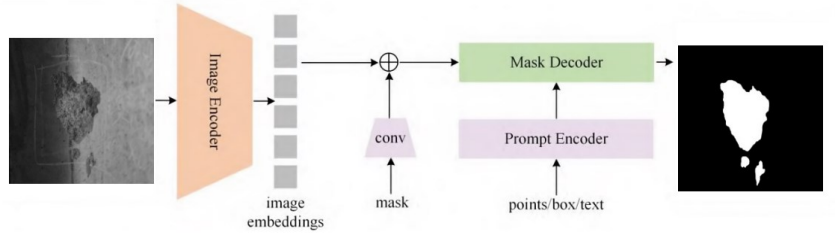

Figure 2. SMA structure overview

3.2 HQ-SAM network structure

HQ-SAM[14] is an improved model of SAM, and its network structure is closely connected with SAM. HQ-SAM reuses most of SAM's pre-trained model weights, adding only two new key components, the High-Quality Output Token and the Global-local Feature Fusion, as shown in Figure 3. Building the HQ-SAM architecture requires designing a learnable HQ-Output Token based on SAM, and then inputting Prompt Tokens, Output Tokens, and HQ-Output tokens together into SAM's mask decoder. Instead of just reusing SAM's mask decoder functionality, the HQ-Output Token runs on a fine feature set to achieve accurate mask details. The mask decoder features of SAM and the early and late Feature maps of ViT encoder are fused together by the global-local Feature Fusion component, so that HQ-SAM has both Global semantic context and local fine-grained features. Therefore, HQ-SAM can be seen as a high-quality zero-sample segmentation model evolved from SAM, with negligible additional model parameters and computational costs.

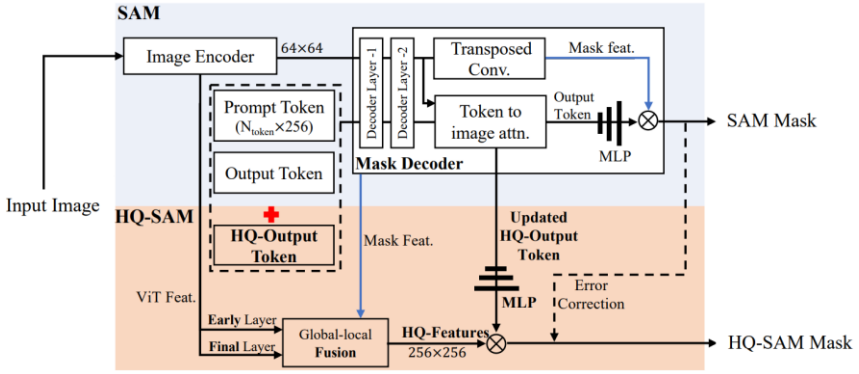

Figure 3. Network structure of SMA and HQ-SAM.

3.3 Overall architecture: YOLOv8 combines SAM and HQ-SAM

The core purpose of this study is to fuse YOLOv8 with SAM models (including SAM and HQ-SAM). YOLOv8 has extremely high detection speed and accuracy, which can achieve fast target identification while maintaining high accuracy. The algorithm adopts end-to-end network design, simplifies the detection process, and makes the process from input image to output detection result more efficient. In addition, it also has strong generalization ability, which can show good performance in different scenarios and data sets. Therefore, YOLOv8 is used to train a small number of bridge disease samples, and it is most suitable to outline the region of interest in the disease image, and its excellent detection speed can greatly improve the efficiency of recognition. SAM and HQ-SAM are multi-functional and powerful pre-trained large models tailored for segmentation tasks, using 11 million images and more than 1 billion masks for training, with super zero sample and small sample generalization capabilities.

By combining YOLOv8 with two SAM models, the detection speed and segmentation accuracy of the whole recognition process can be greatly improved. Figure 4 shows the flow chart of the combination of the two, first using the YOLOv8 model to generate the predictive segmentation bounding boxes that act as the base input for the SAM and HQ-SAM models and act as the region of Interest (ROI). By focusing the model's attention on the relevant parts of the image, these regions help to achieve accurate segmentation of bridge diseases.

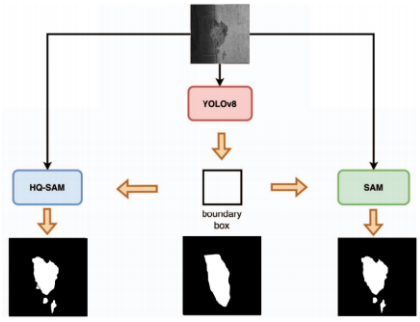

Figure 4. The overall architecture of YOLO, SAM, and HQ-SAM.

4. ANALYSIS OF MODEL TRAINING RESULTS

4.1 Test environment and design

In the implementation of this experiment, Python is used as the core programming language and PyTorch framework is used to build the network model. To improve training efficiency, we deployed CUDA12.1 to speed up the training process. The experimental hardware environment consisted of an i7-14700KF CPU with 3.40GHz and an RTX4090 GPU with 24GB of video memory.

4.2 Model training results

Figure 5 randomly shows the recognition effects of four disease images, including the real mask, SAM (including HQ-SAM and SAM), and the mask predicted by the YOLOv8 model. It can be clearly observed from the figure that both SAM models (HQ-SAM and SAM) show excellent performance in identifying different disease types, and their predicted masks are significantly better than those of YOLOv8. However, it should be noted that YOLOv8 is only used to generate hints for the SAM model, so YOLOv8 is expected to perform poorly when comparing segmentation performance with the SAM model. This paper includes the segmentation results of YOLOv8 to confirm the expected conjecture, and shows how to use the hints generated by YOLOv8 to achieve fully automated processing in the SAM model. In order to further explore these results, we performed a computational analysis of YOLOv8+SAM and YOLOv8+HQ-SAM.

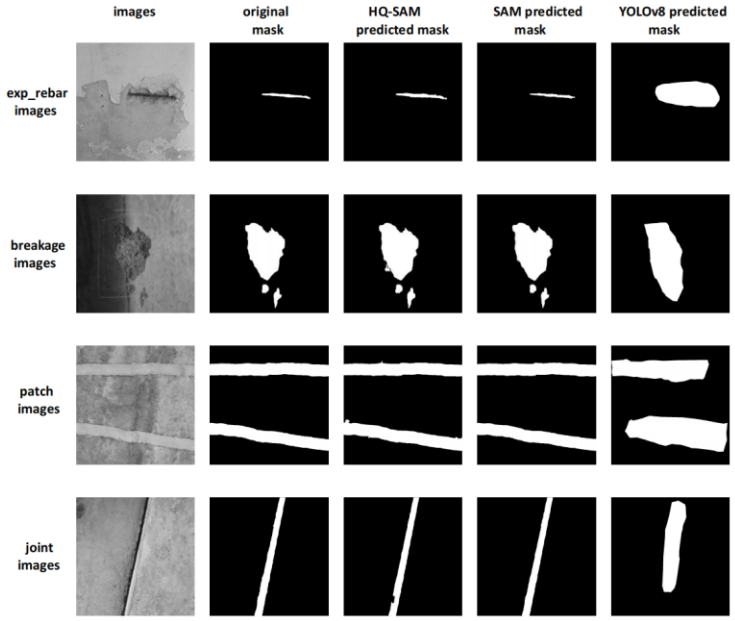

Figure 5. The recognition effect of the three algorithms.

In order to quantify the segmentation effect of the three methods, the evaluation indexes used in this paper include Precision, Recall and F1-score[15]. It can be seen from the data in Table 2 that the evaluation indicators of YOLOv8 are indeed not ideal, which is consistent with the segmentation effect shown in Figure 5. However, the main purpose of YOLOV8 is only to find out the general area of the disease, so there is no need to go into details here. This paper focuses on the analysis of HQ-SAM and SAM evaluation indicators. In various data sets, SAM model shows strong segmentation ability, most of the data is ahead of HQ-SAM, and a few are close to it. exp_rebar is one of the most important diseases in Bridges. The Precision of HQ-SAM on exp_rebar is 0.8772, Recall is 0.8327, and F1-score is 0.8544, while the indexes of SAM model are 0.8767 in sequence. 0.8524 and 0.8644, where Recall and F1-score are both better than HQ-SAM, while Precision is slightly lower than HQ-SAM. However, in terms of breakage disease, SAM was superior to HQ-SAM, and the improvement of various indexes was 5.28%, 3.68% and 4.47%, respectively. Finally, there is little difference between patch and joint, and its recognition effect is better than exp_rebar, especially for patch, each evaluation index is above 0.9.

Table 2. Comparison of evaluation indexes of various diseases.

Algorithm	Type	Precision	Recall	F1-score
HQ-SAM	exp_rebar	0.8772	0.8327	0.8544
	Breakage	0.8524	0.8435	0.8479
	Patch	0.9379	0.9142	0.9259
	Joint	0.8974	0.8625	0.8796
SAM	exp_rebar	0.8767	0.8524	0.8644
	Breakage	0.8974	0.8745	0.8858
	Patch	0.9465	0.9124	0.9291
	Joint	0.8812	0.8797	0.8804
YOLOv8	exp_rebar	0.4143	0.8771	0.5628
	Breakage	0.5238	0.7032	0.6004
	Patch	0.6485	0.6851	0.6663
	Joint	0.5074	0.6481	0.5692

In summary, by comparing the segmentation effect of SAM and HQ-SAM on four bridge disease data sets, SAM is slightly superior to HQ-SAM and more suitable as the main algorithm for bridge disease segmentation. Although HQ-SAM is an improvement of SAM algorithm, there may be deviation due to the deeper detailed information provided by HQ-Output Token. As a result, there are more wrong features in feature fusion than SAM, resulting in lower segmentation effect, which can be seen from Figure 5.

5. CONCLUSION

In this paper, we propose to use the YOLOv8 model to generate disease ROI as the prompt input of the large model SAM, and only a small number of samples can be used to complete high-precision segmentation, and the results are worthy of recognition. In the experiment, we conducted a comprehensive evaluation of three different models, namely SAM, HQ-SAM and YOLOv8, for segmentation testing on the bridge disease dataset containing exp_rebar, breakage, patch and joint. Our results show that in most cases, the SAM model consistently outperforms the other two models, HQ-SAM and YOLOv8. With higher accuracy, recall and F1 scores, SAM models demonstrate superior segmentation accuracy and overall performance. This can be attributed to the model's advanced architecture, which combines convolutional neural networks and attention mechanisms to efficiently learn and represent complex patterns in disease images. Therefore, the method of YOLOv8+SAM can realize the accurate segmentation of bridge diseases under a small number of samples, providing a new technical direction for the intelligent recognition of bridge diseases.

However, it must be acknowledged that this study has some limitations, as the evaluation was limited to three specific models and other state-of-the-art models may have been excluded. In addition, the data sets used in the study may not fully represent the diversity of bridge disease images, and future studies should incorporate more diverse data sets for experimental analysis.

REFERENCES

[1] Zhang, J., Qian, S. and Tan, C., "Automated bridge surface crack detection and segmentation using computer vision-based deep learning model," Eng. Appl. Artif. Intell. 115, 105225 (2022).

[2] Fu, H., Meng, D., Li, W., et al., "Bridge crack semantic segmentation based on improved Deeplabv3+," Journal of Marine Science and Engineering, 9(6), 671 (2021).

[3] Lee, J. S., Park, J. and Ryu, Y. M., "Semantic segmentation of bridge components based on hierarchical point cloud model," Automation in Construction, 130, 103847 (2021).

[4] Wang, X., Chen, G., Qian, G., et al., "Large-scale multi-modal pre-trained models: A comprehensive survey," Machine Intelligence Research, 1-36 (2023).

[5] Bommasani, R., Hudson, D. A., Adeli, E., et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, (2021).

[6] Zhou, C., Li, Q., Li, C., et al., "A comprehensive survey on pretrained foundation models: A history from bert to Chatgpt," arXiv preprint arXiv:2302.09419, (2023).

[7] Kirillov, A., Mintun, E., Ravi, N., et al., "Segment anything," arXiv preprint arXiv 2304.02643, (2023).

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, (2020).

[9] Wang, D., et al., "Samrs: Scaling-up remote sensing segmentation dataset with segment anything model," Advances in Neural Information Processing Systems, 36, (2024).

[10] Ji, G. P., Fan, D. P., Xu, P., et al., "SAM struggles in concealed scenes—empirical study on 'Segment Anything'," Science China Information Sciences, 66(12), (2023).

[11] Xie, Z., Guan, B., Jiang, W., et al., "PA-SAM: prompt adapter SAM for high-quality image segmentation," arXiv preprint arXiv:2401.13051, (2024).

[12] Sohan, M., Sai Ram, T., Reddy, R., et al., "A review on yolov8 and its advancements," International Conference on Data Intelligence and Cognitive Informatics, Springer, Singapore, 529-545 (2024).

[13] Hussain, M., "Yolov1 to v8: Unveiling each variant—a comprehensive review of yolo," IEEE Access, 12, 42816-42833 (2024).

[14] Ke, L., Ye, M., Danelljan, M., et al., "Segment anything in high quality," Advances in Neural Information Processing Systems, 36, (2024).

[15] Wang, Z., Wang, E. and Zhu, Y., "Image segmentation evaluation: a survey of methods," Artificial Intelligence Review, 53(8), 5637-5674 (2020).