13th International Conference on
Structural Health Monitoring of Intelligent Infrastructure
DOI: 10.3217/978-3-99161-057-1-125

Lightweight vision fundamental model-based structural surface crack segmentation using model distillation

Yapeng Guo[1], Shunlong Li[2]
[1]School of Transportation Science and Engineering, Harbin Institute of Technology, 150090 Harbin, China
email: guoyapeng@hit.edu.cn, lishunlong@hit.edu.cn

ABSTRACT: Vision fundamental models demonstrate considerable competitiveness in structural surface crack segmentation due to their strong generalization ability. Vision fundamental models improve the fitting capacity for various objects by increasing image encoder complexity. However, for crack segmentation, the excessive number of these parameters leads to slow running speeds and large space occupation. This paper presents a lightweight Segment Anything Model (SAM)-based crack segmentation method using model distillation technology, aiming for consistent crack image embedding. Firstly, end-to-end automatic crack segmentation is achieved by modifying the SAM model through the addition of a crack segmentation head. Secondly, model distillation is employed to transfer features from the heavy-parametric encoder in SAM with minimal loss. Comparative analysis of cutting-edge crack segmentation techniques across eight frequently utilized datasets demonstrates their effectiveness and precision. The findings reveal the potential of mobile deployment of civil structure damage identification based on vision fundamental models.

KEY WORDS: Crack segmentation, vision fundamental model, model distillation, lightweight, deep learning, bridge damage

1 INTRODUCTION

Crack is a kind of critical apparent damages in civil structures, and crack identification serves as a fundamental basis for evaluating structural condition and determining maintenance strategies [1,2]. Currently, structural surface crack identification heavily relies on visual inspection by engineers, which suffers from limitations such as low efficiency and subjectivity. To address this issue, researchers have started to replace human eye inspection with visible light cameras, enabling the automatic identification of cracks through the acquisition of structural visual images and the design of corresponding algorithms. Common methods of image acquisition include portable digital cameras, fixed monitoring systems, smartphones, unmanned aerial vehicles, climbing or underwater robots, etc., which significantly reduce the cost and risk associated with obtaining apparent structural information [3,4].

After obtaining structural visual images, researchers employ digital image processing algorithms for structural surface crack identification. Considering the impressive advancements of deep learning (DL) in various fields, there is a gradual shift within the field of crack identification towards automatic feature extraction utilizing deep learning models [5,6]. Researchers have employed DL-based object classification algorithms to classify multiple patches cut from crack images to judge the presence or absence of cracks, achieving patch-wise crack identification [7-9]. Another important direction in this type of research is using DL-based object detection algorithms to find crack locations in crack images automatically, enabling box-wise crack identification [10-13]. However, the accuracy of patch-wise and box-wise crack identification may not meet the requirements for assessing the apparent condition of structures. As a result, DL-based pixel-wise crack identification (i.e., crack segmentation), has emerged to address this limitation.

DL-based crack segmentation methods can be categorized into two main groups: convolutional neural network (CNN)-based and transformer-based. CNN-based crack segmentation mainly employs "encoder-decoder" architecture, including fully convolutional network (FCN) and its variants, as well as generative adversarial network (GAN). For FCN, Li et al. [14] proposed using FCN for pixel-to-pixel segmentation of various damages in civil structures. Chen and Jahanshahi [15] developed a rotation-invariant FCN to explicitly consider the rotational invariance of crack images. Hoskere et al. [16,17] introduced a FCN-based multi-class semantic segmentation approach using multi-task learning, which achieved better results than training multiple tasks independently for multi-type structural materials and defects. To address the high noise and background interference in pavement crack images, Huyan et al. [18] established the CrackU-net framework with a modification on U-net, which also addressed the false-positive crack detection issue. Jiang et al. [19] applied attention mechanisms to U-net for detecting corrosion defects in steel box girders. Liu et al. [20] introduced a framework for concrete crack segmentation and quantitative calculation that considers the weight of crack boundaries. Xiang et al. [21] proposed a crack image augmentation method using active learning to enhance the accuracy of crack segmentation methods. Nguyen et al. [22] discussed the influence of different training loss functions using U-net on different crack datasets. Xu et al. [23] proposed a limited-supervised deep learning framework for damage segmentation (including cracks) using meta learning based on U-net.

For GAN, Zhang et al. [24] aimed to address the severe imbalance between cracks and backgrounds with a crack-patch-only GAN framework. Kim et al. [25] tackled the issue of data

DOI: 10.3217/978-3-99161-057-1-125

scarcity in detecting cracks in steel structures using laser thermography for data augmentation by employing GAN. Similarly, to overcome the issue of limited training data, Ma et al. [26], Jin et al. [27], Li and Zhao [28], as well as Zhang et al. [29], utilized various variants of GAN to generate various crack images.

Transformer models have demonstrated significant progress in vision-based crack segmentation. Shamsabadi et al. [30] introduced Vision Transformer (ViT) into this area and attained higher detection accuracy in asphalt and concrete surface crack segmentation than CNNs. Wang and Su [31] developed a multi-level structure Transformer as an encoder to output multi-level features and fuse different levels of features. Ding et al. [32] analyzed the characteristics of crack recognition and proposed a boundary refinement Transformer for automatic segmentation of crack images obtained by drones. Guo et al. [33] used Swin Transformer to encode road crack images and employed UperNet to generate segmentation results. Tong et al. [34] combined Dempster-Shafer theory and Transformer network to construct a crack segmentation framework considering uncertainty and proposed a corresponding training strategy. Zhang et al. [35] proposed a segmentation Transformer framework called ShuttleNet v2, which can detect not only cracks but also multiple other diseases simultaneously. Furthermore, the amalgamation of CNNs' local modeling capability and Transformers' global modeling capability to build more powerful crack segmentation models is also an important research direction. Zhou et al. [36] fused Swin Transformer blocks and inverse residual blocks based on Deeplab v3 plus framework and combined channel attention mechanism to improve crack segmentation accuracy.

With the emergence of vision fundamental models, the inherent paradigm of object segmentation has been disrupted. These vision fundamental models, characterized by a massive number of network parameters and extensive training data, exhibit unprecedented robust generalization capabilities, allowing for precise segmentation of most common objects in zero-shot and few-shot forms [37]. However, when applied to crack segmentation tasks, vision fundamental models face two primary issues: (1) the need for specific prompts during application or a lack of semantic information for automatic segmentation; (2) the excessive number of network parameters in vision fundamental models leads to slow segmentation speeds and deployment difficulties in hardware-constrained environments.

This paper's primary goal is to significantly lighten the vision fundamental model while preserving its strong generalization ability, to achieve precise and efficient segmentation of cracks. To achieve this objective, this paper proposes the following two innovative approaches: (1) modifying SAM structure by adding a crack segmentation head to incorporate semantic information for automatic segmentation, and (2) utilizing model distillation techniques to substantially reduce the parameters of SAM and significantly improve its running speed, with only acceptable loss in segmentation accuracy.

This paper's primary contributions are twofold: (1) it represents an early attempt to apply vision fundamental models to automatic crack segmentation, providing a feasible approach for the application of such models in civil engineering, thereby offering valuable reference results for future research; (2) it verifies the feasibility of lightweighting crack segmentation networks based on vision fundamental models, enabling effective transfer of the powerful generalization ability of these models under hardware-constrained conditions.

The remaining content of this paper are structured as follows. Section 2 provides an overview of the advancements in DL-based crack segmentation. Section 3 delves into the intricate framework of the proposed lightweight vision fundamental model-based crack segmentation approach. Section 4 outlines the implementation specifics. Section 5 offers the testing results under both full supervision and limited supervision, as well as the results evaluated on hardware-constrained platforms. Lastly, Section 6 concludes the paper.

2 METHODOLOGY

The proposed lightweight vision fundamental model-based crack segmentation method comprises a lightweight crack encoder and a crack segmentation head (shown in Figure 1). The former extracts the robust features of the crack image to generate crack image embeddings, while the latter uses high-quality embedding to complete pixel-level crack segmentation. The lightweight crack encoder's initialization weight originates from the SAM original heavy-parametric vision fundamental model through the utilization of model distillation technology (using common object segmentation dataset), the distillation objective is set to minimize the embedding difference of the image after the encoder. Finally, the crack segmentation model proposed here undergoes fine-tuning using the crack dataset.

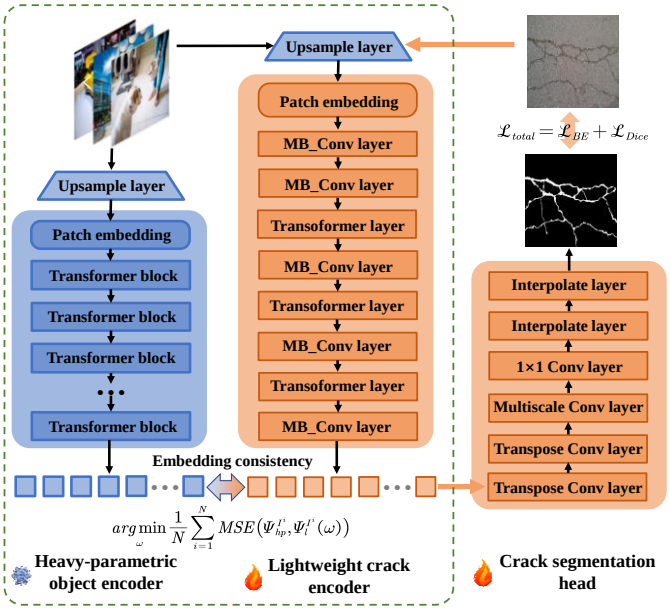

Figure 1. Overall architecture of the lightweight vision fundamental model-based crack segmentation method.

2.1 *Lightweight crack encoder using model distillation*

The heavy parameter encoder in SAM adopts the ViT model. To ensure the uniformity of the architecture, this study uses the lightweight TinyViT [38] as the crack encoder. TinyViT adopts a hierarchical vision transformer, serving as the foundational architecture, which can better integrate multiscale features for downstream tasks. TinyViT comprises four stages, leading to a gradual reduction in the resolution of the feature map. Each

DOI: 10.3217/978-3-99161-057-1-125

stage includes a feature extraction (FE) layer and a down-sampling (DS) layer. The FE layer of stage 1 adopts MB_Conv, and stages 2-4 adopt transformer. The DS layer of all stages adopts the MB_Conv. TinyViT has been efficiently designed from three aspects: the sliding window mechanism corresponding to the FE layer (limiting the transformer attention mechanism to the window to reduce the computational complexity), the hierarchical design mechanism corresponding to the DS layer (taking advantage of the CNN to save the amount of calculation while extending the window attention to the global) and the model size control mechanism (customizing models of different sizes by adjusting the model control parameters).

To make full use of the effective information in the vision fundamental model SAM, this paper proposes to transfer the features in the heavy-parametric encoder of SAM to the proposed lightweight crack encoder by means of model distillation. That is, SAM is used as the teacher model, and the student model is employed as the proposed model. By setting the optimization goal, the knowledge in the teacher model is transmitted to the student model as lossless as possible.

It is assumed that the output image of the i^{th} image I^i after entering the SAM's heavy-parametric encoder is embedded as $\Psi_{hp}^{I^i}$, and the output image of the proposed lightweight encoder is embedded as $\Psi_{l}^{I^i}$. The crack segmentation head of the proposed framework directly uses image embedding as input, so it is not necessary to minimize the segmentation error after adding the crack segmentation head to SAM, but only to minimize the difference between the two embeddings [39] (illustrated in Equation (1)), where N represents the total count of training crack images required for model distillation, MSE is the least square error function, and ω is the trainable weight parameters of the proposed lightweight encoder.

$$\arg\min_{\omega} \frac{1}{N}\sum_{i=1}^{N} MSE\left(\Psi_{hp}^{I^i}, \Psi_{l}^{I^i}(\omega)\right) \quad (1)$$

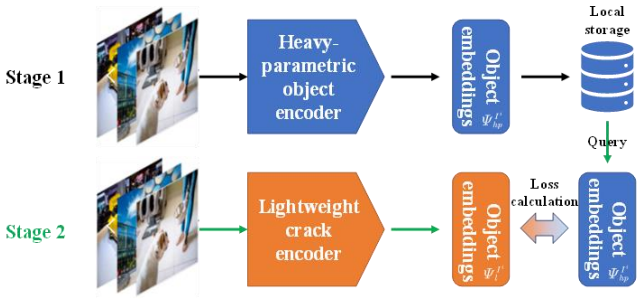

Figure 2. Processing procedure of model distillation.

Specifically, as shown in Figure 2, there are two stages during model distillation, the encoder parameters in the original SAM are frozen (untrainable), while the proposed lightweight encoder parameters are trainable. In stage 1 (embedding preparation), the image to be trained is input into the original SAM in advance to obtain the corresponding image embedding $\Psi_{hp}^{I^i}$ and then stored locally. In stage 2 (lightweight encoder training), the image is fed into the proposed lightweight encoder to obtain the image embedding $\Psi_{l}^{I^i}$, $\Psi_{hp}^{I^i}$ can be queried from the local storage. The optimization goal of training can be directly calculated. Taking such a training strategy will greatly reduce the time and cost of training while ensuring the distillation effect.

2.2 Crack segmentation head

To assist object segmentation, SAM employs prompts through the integration of incorporating prompt encoder and mask decoder into image embedding process. This encompasses the handling of image and prompt embedding, and output tokens subsequent to the image encoder. However, since the crack segmentation task in this study is automatic and does not require prompt input, the latter part of SAM needs to be modified. To address this, this paper introduces a crack segmentation head to fulfill the necessary functions.

The crack segmentation head is composed of several key components, including two transposed convolution layers, a multiscale convolution layer, a convolution layer with the size of 1×1, and two interpolation layers (shown in Figure 1). The transposed convolution layer is designed to increase the resolution of the encoded crack image embedding by a factor of 2, thereby restoring spatial information. The multiscale convolution layer is utilized to leverage feature fusion at different scales, enabling the model to learn information in various ranges around crucial pixels through backpropagation gradient. Figure 3 illustrates the detailed architecture of the multiscale convolution. According to the channel dimension, features from *n* channels of the transpose convolutional layer are partitioned into *k* groups. To retain features at the current scale, a 1×1 convolution is applied to the first group. For the remaining *k*-1 groups, 3×3 convolutions with varying dilation rates are utilized to capture features at different scales, where the dilation rate is determined by the number of groups minus 1 [40]. The 1×1 convolutional layer is responsible for integrating information from different channels and adjusting the output dimension accordingly. Finally, the interpolation layer further upsamples the output to fine-tune the output dimension.

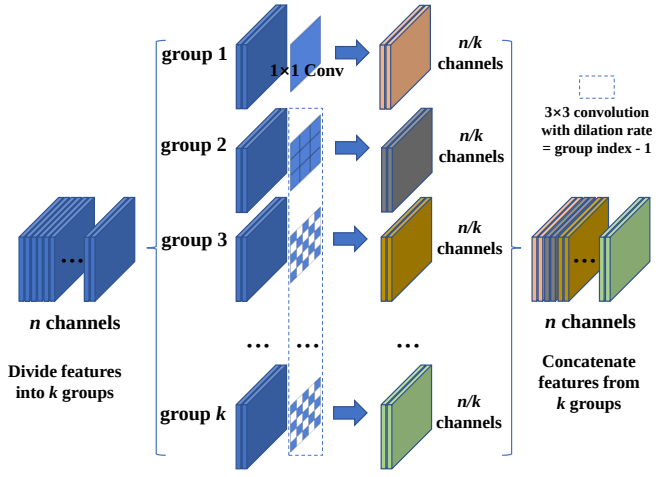

Figure 3. Architecture of the multiscale convolution.

A combined loss ($\mathcal{L}_{total}$) comprising binary entropy ($\mathcal{L}_{BE}$) and Dice ($\mathcal{L}_{dice}$) is set as the objective of fine-tuning the proposed method, as illustrated in Equations (2)-(4). Here, *N* denotes the image's pixel number, while y_i and p_i respectively

DOI: 10.3217/978-3-99161-057-1-125

denote the annotated label and predicted value for the i^{th} pixel. Additionally, Y and $\hat{Y}$ represent the crack mask annotations and predictions, respectively.

$$\mathcal{L}_{total} = \mathcal{L}_{BE} + \mathcal{L}_{dice} \tag{2}$$

$$\mathcal{L}_{BE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) \tag{3}$$

$$\mathcal{L}_{dice} = 1 - \frac{2\left|\hat{Y} \cap Y\right| + 1}{\left|\hat{Y}\right| + |Y| + 1} \tag{4}$$

3 IMPLEMENTATION DETAILS

3.1 Dataset

Distillation on common object dataset SAM's powerful segmentation generalization ability comes from the large-scale segmentation dataset SA-1B. The dataset is generated by SAM's data engine and is divided into three stages: preliminary manual, semi-automated, and fully automated. During the initial stage, SAM aids the annotator to annotate the mask, akin to the traditional interactive object segmentation scenario. During the second stage, SAM is capable of autonomously producing the mask for certain objects by proposing their potential locations, and the annotator annotates the remaining objects, which helps to increase the diversity of the mask. In the last stage, the regular grid prompt SAM of the foreground point is employed to produce approximately one hundred high-quality masks for each image. Ultimately, SA-1B generates over 1 billion object masks across 11 million images. Since the proposed crack segmentation framework uses a lightweight encoder, the parameters are much smaller than the original SAM encoder. Therefore, 0.1% of the SA-1B dataset (11,000) is randomly sampled according to the literature results as the training dataset of the model distillation [39].

Fine-tuning on crack dataset Lately, crack recognition field have achieved significant progressions, and several crack segmentation datasets have been released to the research community. While many studies have trained and evaluated models on specific datasets, there is a lack of comprehensive testing across multiple datasets, which hinders our understanding of the generalization capabilities of crack segmentation models [41]. To address this limitation, this paper selected eight influential datasets for evaluating the proposed method. These datasets include CFD, Crack500, Cracktree200, DeepCrack, EugenMiller, GAPs, Rissbilder, and Volker (referred to as Datasets 1-8) [41]. These datasets exhibit significant variations in terms of structural materials, structural parts, image quality, and quantity, thereby enabling an effective evaluation of the generalization abilities of crack segmentation models. Figure 4 offers a summary of the quantity of training and testing images contained within each of the eight datasets. The entire set of training images, amounting to 7754, is partitioned into training subset (90%) and validation subset (10%). The testing images, collectively referred to as the testing subset, are employed to evaluate the segmentation methods proposed in this paper.

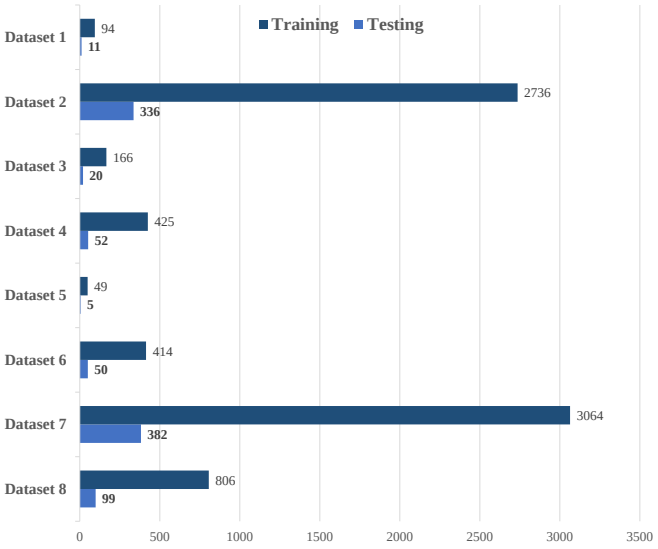

Figure 4. Numbers of training and testing images of eight datasets.

3.2 Distillation and fine-tuning strategy

During the model distillation, the image embedding vector of the training image through the SAM heavy-parametric encoder has been calculated in advance and saved to the local storage. The training image only needs to go through the lightweight encoder to obtain the new image embedding vector, and then read the previously saved SAM embedding vector and calculate mean squared error (MSE). During the training, a single GPU was used, the batch size was 2, and a total of 50,000 iterations were performed. The MSE value obtained by the final convergence was 0.977, indicating the effectiveness and accuracy of the model distillation.

Throughout the proposed crack segmentation model fine-tuning, the initial learning rate was established as 1e-5. The total training iteration count was determined as 24,000. A multi-step learning rate change approach was employed, where it was reduced to 0.1 times the previous value at the 16,000th and 20,000th iterations, respectively. The fine-tuning process implemented an early stopping strategy. The parameters at this iteration were taken as the final model weights. In addition, the batch size was 2.

The configuration utilized for fine-tuning and evaluating the proposed method comprised an Intel Xeon(R) E5-2620 v4 central processing unit (CPU), complemented by a robust Nvidia RTX 3090 graphics processing unit (GPU) offering 24GB memory. Additionally, the system was equipped with an ample 128GB of memory.

4 RESULTS AND DISCUSSIONS

The quantitative and qualitative results of the proposed lightweight crack segmentation method under full supervision and limited supervision conditions are illustrated in this section, and the state-of-the-art CNN-based method Deeplab v3 plus (with MobileNet v3) [42] and transformer-based Segformer [43] are employed to be comparison. Additionally, the weight file space occupation, the running speed on different hardware

DOI: 10.3217/978-3-99161-057-1-125

platforms and the possibility of mobile deployment of the proposed method are discussed.

4.1 Testing results under full supervision

Full supervision refers to training on the entire training subset (6,978 pairs of crack images and masks) and testing on the entire test subset (955 crack images). This condition facilitates the transfer of feature parameters from vision fundamental models to the domain of crack segmentation, enabling exploration of the segmentation accuracy limits across different models.

Table 1 elucidates the testing Dice scores (×100%) for eight datasets under full supervision, as well as the total parameter numbers of different algorithms and their GPU memory occupancy during training (with an input size of 1024×1024×3 and a batch size of 2). In Table 1, SAM represents the crack segmentation method based on vision fundamental models that employs the heavy-parametric encoder from the original SAM. Compared to Deeplab v3 plus, the proposed method significantly improves the Dice score by 13.6, and although the parameter count increases to twice that amount, the required GPU memory during training decreases to 36%, which is the most direct assessment of algorithm training expenses. Compared to Segformer, the proposed method is on par in terms of accuracy, but with a 45% reduction in parameter count and a 37% reduction in required GPU memory. Compared to the original SAM-based method, the proposed approach experiences a 5.4 decrease in Dice score, but with a parameter count reduced to 7% and at least an 85% reduction in required GPU memory. In summary, the proposed method not only maintains segmentation accuracy in comparison to other cutting-edge methods but also significantly reduces training costs.

Table 1. Testing Dice scores (×100%) on eight datasets, total parameter numbers and GPU memory occupation under full supervision

Method		DL v3p	Segformer	SAM	Proposed
Dataset	1	52.6	51.8	68.5	56.1
	2	54.9	68.4	71.1	65.6
	3	24.3	24.4	37.6	24.4
	4	69.6	72.8	79.9	68.8
	5	36.2	56.1	57.7	53
	6	25.8	28.9	45.4	44.77
	7	30.9	50.1	54.9	50.8
	8	60	67.5	75.5	68.7
Average		44.4	58	63.4	58
Params		3.2M	13.6M	89.8M	6.2M
Mem		10.3G	10.1G	>24G	3.8G

Figure 5 displays representative results of crack segmentation using different methods, where each row corresponds to a representative crack image from each dataset, and each column represents the original image, ground truth, and the test results using Deeplab v3 plus, Segformer, the original SAM-based method, and the proposed method, respectively. Whether they are concrete or asphalt surface cracks, whether they are dot-like, strip-like, or mesh-like, the crack segmentation results of methods based on vision fundamental models are superior in terms of integrity and connectivity compared to CNN-based and transformer-based. While the proposed method's segmentation effect is slightly lacking in local detail handling compared to the original SAM-based crack segmentation method, it exhibits evident advancement over other methods, demonstrating the proposed method's efficacy in enhancing accuracy.

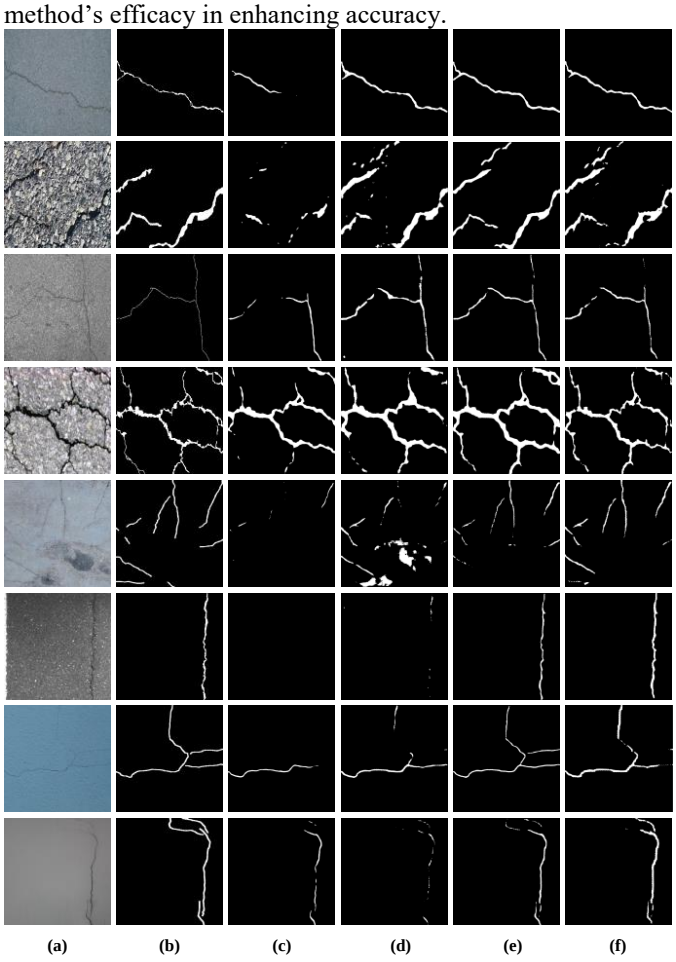

Figure 5. Representative testing results: (a) raw image, (b) annotation, (c) DL v3p, (d) Segformer, (e) SAM, (f) the proposed method.

4.2 Testing results under limited supervision

In contrast to prior CNN-based and transformer-based object segmentation approaches, the primary advantage of object segmentation methods grounded on vision fundamental models is their robust generalization capability. This means that they can achieve high segmentation accuracy with minimal domain-specific supervision information. Hence, this part showcases the proposed method's strong generalization ability by comparing the test accuracy of different methods under two limited supervision cases (1%-shot and one-shot). The 1%-shot case refers to training with 1% of the training subset of full supervision (77 crack image-mask pairs), while the one-shot case refers to training with only one crack image-mask pair per dataset (a total of 8). These two cases provide an extreme test of the generalization capabilities of different methods.

The testing Dice scores of different algorithms under limited supervision are illustrated in Table 2. Consistent with

DOI: 10.3217/978-3-99161-057-1-125

theoretical analysis, the crack segmentation method based on vision fundamental models significantly outperforms those based on CNNs and transformers in terms of segmentation accuracy in both the 1%-shot and one-shot cases.

Table 2. Testing Dice scores (×100%) on eight datasets under limited supervision

Data size		1%-shot			
Method		DL v3p	Segformer	SAM	Proposed
Dataset	1	39.6	47.2	54.2	46.8
	2	53.7	60	65.8	57.4
	3	23	17.3	21.1	24
	4	66.4	65.6	70.5	63.2
	5	23.4	44.8	48.1	38.2
	6	9.7	23	32.9	19
	7	26.7	47	48.3	50.3
	8	54.5	64.2	66.8	66.1
Average		40.4	52.5	56.3	52.8
Data size		one-shot			
Method		DL v3p	Segformer	SAM	Proposed
Dataset	1	35	47	44.8	49.2
	2	21.4	22.4	34.1	25
	3	20.1	22.2	16.5	20.3
	4	48	51.4	59.3	67.1
	5	46.1	48.7	54.2	53.9
	6	29.3	34.8	34.7	24.2
	7	24	36.1	39.3	42.8
	8	48.3	58.9	58	56.8
Average		27.4	34.3	39.9	38

Notably, in the one-shot case, the proposed method improved the Dice score by 3.7 compared to Segformer and only decreased by 1.9 compared to the original SAM-based method. Considering that the proposed method has fewer parameters and lower training costs, it is substantiated that the approach can reach high precision with maintaining operational efficiency. Former transformer models necessitated a substantial volume of supervised training data to attain elevated segmentation precision, although vision fundamental models are also based on transformer architectures. Preliminary judgments can also be made based on experimental results, the vision fundamental model-based crack segmentation method demonstrates good generalization capability even under extremely limited supervision conditions.

4.3 Deployment of the proposed method

Although the vision fundamental model has strong segmentation generalization ability, its deployment difficulty and cost are high, which aligns with the primary research concentration of this paper. Therefore, this subsection deploys the original SAM-based and the proposed lightweight vision fundamental model-based crack segmentation models on different hardware platforms to illustrate the advantages of the latter.

The weight file space occupancy serves as a metric for gauging the complexity of the model, encompassing all parameters and configurations stored on the disk. Table 3 demonstrates that the space occupancy of SAM-based models based on backbones of different sizes is 404MB, 1230MB, and 2665MB, respectively, while the proposed model is only 70MB (17% of the minimum SAM-based).

Table 3. Weight file space occupancy and running speed of the proposed method

Method		Space occupancy	Running speed	
SAM	ViT-B	404MB	GPU	0.107s
			x86	5.498s
			arm	/
	ViT-L	1230MB	/	/
	ViT-H	2665MB	/	/
Proposed		70MB	GPU	0.016s
			x86	0.637s
			arm	2.245s

The running speed is the most direct indicator to measure the complexity of a model. Although it is affected by factors such as code implementation, the relative speed of different models can still be compared after controlling variables. This paper tests on three common hardware platforms (GPU, x86 CPU and arm CPU), and the results are shown in Table 3. With 1024 × 1024 images as input, the SAM-based model takes 0.107 s and 5.498 s on GPU and x86 CPU respectively. Because the model is too complex to be deployed on the arm CPU used in this experiment. Meanwhile, the proposed model consumes 0.016 s and 0.637 s on GPU and x86 CPU (15% and 12% of SAM-based, respectively), and 2.245s on arm CPU (twice as fast as SAM-based on x86 CPU).

It is worth noting that the arm CPU used in this experiment is Kirin 970, a consumer and low-cost chip released six years ago. Compared with the current mobile phone CPUs, the performance difference is huge. Employing the most cutting-edge chip would significantly enhance the performance of the proposed method. Figure 6 illustrates the exemplary testing outcomes of the deployed proposed method on a mobile phone.

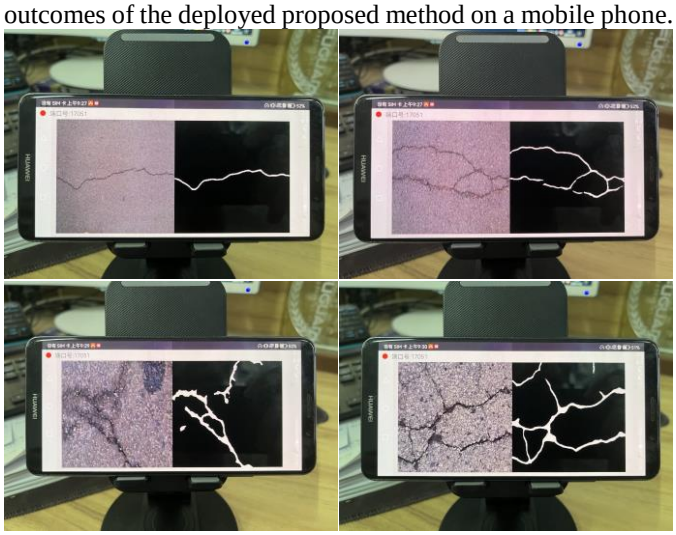

Figure 6. Representative testing results on mobile phones

DOI: 10.3217/978-3-99161-057-1-125

5 CONCLUSIONS

A novel crack segmentation method using lightweight vision fundamental model is developed in this paper. The approach incorporates model distillation techniques to substantially decrease the model parameters and enhance operational speed while maintaining the robust generalization capabilities of the vision fundamental model to the greatest extent possible. In light of the findings, the following conclusions can be inferred: (1) By substituting the ViT encoder in the original SAM with the lightweight encoder TinyViT and using model distillation techniques with image embedding consistency as the optimization goal, effective transfer of the vision fundamental model's generalization ability is achieved. (2) Under full supervision, the proposed method surpasses current cutting-edge methods based on non-vision fundamental models, achieving a segmentation Dice score of 58.0. Moreover, relative to the original SAM, the model's parameter count is reduced to 7%, and the required GPU memory is decreased to 15%, with only a 5.4 decrease in Dice score. (3) Under limited supervision, the proposed method comprehensively surpasses methods based on non-vision fundamental models in terms of segmentation accuracy and algorithmic efficiency, with Dice scores reaching 52.8 (1%-shot) and 38.0 (one-shot). Furthermore, as the degree of available supervision information decreases, the proposed method demonstrates a heightened advantage, resulting in a diminished disparity with respect to the original SAM. (4) The proposed method achieves a sixfold and eightfold acceleration on GPU and x86 CPU, respectively, compared to the original SAM, and has been successfully deployed on cost-effective ARM CPUs.

The crack segmentation method developed from lightweight vision fundamental model serves as a reference for the efficient application of vision fundamental models in the field of automatic identification of civil engineering damages. Nevertheless, there remains potential for enhancing the detailed recovery of crack identification outcomes in this study. Future work will concentrate on incorporating crack boundary constraints into the loss function and bolstering the post-processing methodologies within the crack segmentation framework to enhance the precision of crack detail identification. To facilitate practical applications, follow-up research should further develop a quantitative measurement and evaluation module for crack dimensions.

ACKNOWLEDGMENTS

This study received financial support from National Key Research and Development Program of China [2024YFC3015201], National Natural Science Foundation of China (NSFC) [52408324, U22A20230 and 52278299] and Natural Science Foundation of Heilongjiang Province of China [LH2024E056].

REFERENCES

[1] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, P. Fieguth, A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure, Advanced Engineering Informatics 29 (2) (2015), pp. 196-210.

[2] J. Guo, P. Liu, B. Xiao, L. Deng, Q. Wang, Surface defect detection of civil structures using images: Review from data perspective, Automation in Construction 158 (2024), p. 105186.

[3] C.M. Yeum, S.J. Dyke, Vision-Based Automated Crack Detection for Bridge Inspection, Computer-Aided Civil and Infrastructure Engineering 30 (10) (2015), pp. 759-770.

[4] Y. Xu, Y. Bao, Y. Zhang, H. Li, Attribute-based structural damage identification by few-shot meta learning with inter-class knowledge transfer, Structural Health Monitoring 20 (4) (2020), pp. 1494-1517.

[5] Y. Hou, Q. Li, C. Zhang, G. Lu, Z. Ye, Y. Chen, L. Wang, D. Cao, The State-of-the-Art Review on Applications of Intrusive Sensing, Image Processing Techniques, and Machine Learning Methods in Pavement Monitoring and Analysis, Engineering 7 (6) (2021), pp. 845-856.

[6] B.F. Spencer, V. Hoskere, Y. Narazaki, Advances in Computer Vision-Based Civil Infrastructure Inspection and Monitoring, Engineering 5 (2) (2019), pp. 199-222.

[7] Y. Que, Y. Dai, X. Ji, A. Kwan Leung, Z. Chen, Z. Jiang, Y. Tang, Automatic classification of asphalt pavement cracks using a novel integrated generative adversarial networks and improved VGG model, Engineering Structures 277 (2023), p. 115406.

[8] Y. Xu, Y. Bao, J. Chen, W. Zuo, H. Li, Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images, Structural Health Monitoring 18 (3) (2018), pp. 653-674.

[9] L. Chen, H. Yao, J. Fu, C. Tai Ng, The classification and localization of crack using lightweight convolutional neural network with CBAM, Engineering Structures 275 (2023), p. 115291.

[10] X. Ye, T. Jin, C. Yun, A review on deep learning-based structural health monitoring of civil infrastructures, Smart Structures and Systems 24 (5) (2019), pp. 567-585.

[11] H. Zhang, K. Gao, H. Huang, S. Hou, J. Li, G. Wu, Fully decouple convolutional network for damage detection of rebars in RC beams, Engineering Structures 285 (2023), p. 116023.

[12] P. Wu, A. Liu, J. Fu, X. Ye, Y. Zhao, Autonomous surface crack identification of concrete structures based on an improved one-stage object detection algorithm, Engineering Structures 272 (2022), p. 114962.

[13] L. Chen, W. Chen, L. Wang, C. Zhai, X. Hu, L. Sun, Y. Tian, X. Huang, L. Jiang, Convolutional neural networks (CNNs)-based multi-category damage detection and recognition of high-speed rail (HSR) reinforced concrete (RC) bridges using test images, Engineering Structures 276 (2023), p. 115306.

[14] S. Li, X. Zhao, G. Zhou, Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network, Computer-Aided Civil and Infrastructure Engineering 34 (7) (2019), pp. 616-634.

[15] F.-C. Chen, M.R. Jahanshahi, ARF-Crack: rotation invariant deep fully convolutional network for pixel-level crack detection, Machine Vision and Applications 31 (6) (2020), p. 47.

[16] V. Hoskere, Y. Narazaki, T.A. Hoang, B.F. Spencer Jr, MaDnet: multi-task semantic segmentation of multiple types of structural materials and damage in images of civil infrastructure, Journal of Civil Structural Health Monitoring 10 (5) (2020), pp. 757-773.

[17] Y. Narazaki, V. Hoskere, T.A. Hoang, Y. Fujino, A. Sakurai, B.F. Spencer Jr, Vision-based automated bridge component recognition with high-level scene consistency, Computer-Aided Civil and Infrastructure Engineering 35 (5) (2020), pp. 465-482.

[18] J. Huyan, W. Li, S. Tighe, Z. Xu, J. Zhai, CrackU-net: A novel deep convolutional neural network for pixelwise pavement crack detection, Structural Control and Health Monitoring 27 (8) (2020), p. e2551.

[19] F. Jiang, Y. Ding, Y. Song, F. Geng, Z. Wang, Automatic pixel-level detection and measurement of corrosion-related damages in dim steel box girders using Fusion-Attention-U-net, Journal of Civil Structural Health Monitoring 13 (1) (2023), pp. 199-217.

[20] G. Liu, W. Ding, J. Shu, A. Strauss, Y. Duan, Two-Stream Boundary-Aware Neural Network for Concrete Crack Segmentation and Quantification, Structural Control and Health Monitoring 2023 (2023), p. 3301106.

[21] Z. Xiang, X. He, Y. Zou, H. Jing, An active learning method for crack detection based on subset searching and weighted sampling, Structural Health Monitoring (2023), p. 14759217231183661.

[22] Q.D. Nguyen, H.-T. Thai, Crack segmentation of imbalanced data: The role of loss functions, Engineering Structures 297 (2023), p. 116988.

[23] Y. Xu, Y. Fan, Y. Bao, H. Li, Task-aware meta-learning paradigm for universal structural damage segmentation using limited images, Engineering Structures 284 (2023), p. 115917.

[24] K. Zhang, Y. Zhang, H.D. Cheng, CrackGAN: Pavement Crack Detection Using Partially Accurate Ground Truths Based on Generative Adversarial Learning, IEEE Transactions on Intelligent Transportation Systems 22 (2) (2021), pp. 1306-1319.

DOI: 10.3217/978-3-99161-057-1-125

[25] C. Kim, S. Hwang, H. Sohn, Weld crack detection and quantification using laser thermography, mask R-CNN, and CycleGAN, Automation in Construction 143 (2022), p. 104568.
[26] D. Ma, H. Fang, N. Wang, C. Zhang, J. Dong, H. Hu, Automatic Detection and Counting System for Pavement Cracks Based on PCGAN and YOLO-MF, IEEE Transactions on Intelligent Transportation Systems 23 (11) (2022), pp. 22166-22178.
[27] T. Jin, X.W. Ye, Z.X. Li, Establishment and evaluation of conditional GAN-based image dataset for semantic segmentation of structural cracks, Engineering Structures 285 (2023), p. 116058.
[28] S. Li, X. Zhao, High-resolution concrete damage image synthesis using conditional generative adversarial network, Automation in Construction 147 (2023), p. 104739.
[29] T. Zhang, D. Wang, A. Mullins, Y. Lu, Integrated APC-GAN and AttuNet Framework for Automated Pavement Crack Pixel-Level Segmentation: A New Solution to Small Training Datasets, IEEE Transactions on Intelligent Transportation Systems 24 (4) (2023), pp. 4474-4481.
[30] E. Asadi Shamsabadi, C. Xu, A.S. Rao, T. Nguyen, T. Ngo, D. Dias-da-Costa, Vision transformer-based autonomous crack detection on asphalt and concrete surfaces, Automation in Construction 140 (2022), p. 104316.
[31] W. Wang, C. Su, Automatic concrete crack segmentation model based on transformer, Automation in Construction 139 (2022), p. 104275.
[32] W. Ding, H. Yang, K. Yu, J. Shu, Crack detection and quantification for concrete structures using UAV and transformer, Automation in Construction 152 (2023), p. 104929.
[33] F. Guo, J. Liu, C. Lv, H. Yu, A novel transformer-based network with attention mechanism for automatic pavement crack detection, Construction and Building Materials 391 (2023), p. 131852.
[34] Z. Tong, T. Ma, W. Zhang, J. Huyan, Evidential transformer for pavement distress segmentation, Computer-Aided Civil and Infrastructure Engineering n/a (n/a) (2023).
[35] H. Zhang, A.A. Zhang, A. He, Z. Dong, Y. Liu, Pixel-level detection of multiple pavement distresses and surface design features with ShuttleNetV2, Structural Health Monitoring (2023), p. 14759217231183656.
[36] Z. Zhou, J. Zhang, C. Gong, Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network, Computer-Aided Civil and Infrastructure Engineering n/a (n/a) (2023).
[37] 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, arXiv e-prints (2023), p. arXiv:2304.02643.
[38] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, L. Yuan, TinyViT: Fast Pretraining Distillation for Small Vision Transformers, arXiv e-prints (2022), p. arXiv:2207.10666.
[39] C. Zhang, D. Han, Y. Qiao, J.U. Kim, S.-H. Bae, S. Lee, C.S. Hong, Faster Segment Anything: Towards Lightweight SAM for Mobile Applications, arXiv e-prints (2023), p. arXiv:2306.14289.
[40] J. Zhang, X. Chen, Z. Qiu, M. Yang, Y. Hu, J. Liu, Hard Exudate Segmentation Supplemented by Super-Resolution with Multi-scale Attention Fusion Module, 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2022, pp. 1375-1380.
[41] E. Bianchi, M. Hebdon, Development of Extendable Open-Source Structural Inspection Datasets, Journal of Computing in Civil Engineering 36 (6) (2022), p. 04022039.
[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, arXiv e-prints (2018), p. arXiv:1802.02611.
[43] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, arXiv e-prints (2021), p. arXiv:2105.15203.